UNIVERSIDAD CENTRAL DE VENEZUELA

FACULTAD DE CIENCIAS

ESCUELA DE COMPUTACIÓN

CENTRO DE INVESTIGACION EN SISTEMAS DE INFORMACIÓN

SOLUCIÓN DE BIG DATA QUE APOYE A LA FASE DE RECLUTAMIENTO DE LA GESTIÓN DEL TALENTO HUMANO EN EL ÁREA DE TECNOLOGÍA DE LA INFORMACIÓN

Trabajo Especial de Grado presentado ante la ilustre

Universidad Central de Venezuela por la

Br. María Gabriela De Freitas Padrón

Para optar al título de Licenciado en Computación

Tutores:

Profa. Brenda López

Profa. Concettina Di Vasta

Caracas, Octubre de 2016

ACTA

Quienes suscriben, miembros del Jurado designado por el Consejo de Escuela de Computación, para examinar el Trabajo Especial de Grado presentado por la Br. María Gabriela De Freitas Padrón C.I. 19.819.365, con el título "SOLUCIÓN DE BIG DATA QUE APOYA LA FASE DE RECLUTAMIENTO DE LA GESTIÓN DEL TALENTO HUMANO EN EL ÁREA DE TECNOLOGÍA DE LA INFORMACIÓN" a los fines de optar al título de Licenciada en Computación, dejan constancia de lo siguiente:

Dicho trabajo, leído por cada uno de los miembros del jurado, se fijó el día 19 de Octubre de 2016 , a las 2:00 p.m., para que su autora lo defendiera en forma pública en la Escuela de Computación, mediante una presentación oral de su contenido, luego de lo cual se respondieron las preguntas formuladas. Finalizada la defensa pública del Trabajo Especial de Grado, el jurado decidió aprobarlo con la nota de 20 puntos.

En fe de lo cual se levanta la presente Acta, en Caracas a los 19 días del mes de Octubre del año 2016.

Profa. Brenda López
Tutora

Profa. Concettina Di Vasta
Tutora

Profa. Ospina Mercy
Jurado Principal

Prof. Sandoval Franklin
Jurado Principal

DEDICATORIA

A Dios y a la Virgen, ya que gracias a ellos he logrado concluir mi ciclo universitario.

A mi Abuela, porque un buen día en el año 2010 le prometí que le dedicaría mis estudios. Sé que desde el cielo me has guiado y cuidado para que todo saliera bien durante mi ciclo universitario.

A mis Padres, porque ellos siempre estuvieron a mi lado brindándome su apoyo y consejos para hacer de mí una mejor persona, luchadora y perseverante.

A mi Novio por sus palabras y confianza, por su amor y por acompañarme en este camino para formarme profesionalmente.

A mi familia quienes han creído en mí siempre, dándome ejemplo de superación, humildad y sacrificio. A todos ustedes les dedico mi tesis, porque han fomentado en mí, el deseo de superación y triunfo en la vida.

A mis amigos y compañeros que de una u otra manera han contribuido para el logro de mis objetivos.

AGRADECIMIENTOS

Primero que nada a Dios, por las bendiciones y sabiduría recibida para alcanzar esta meta tan anhelada.

A mis padres quienes han creído en mí siempre, dándome ejemplos de superación, humildad y sacrificio; enseñándome a valorar todo lo que tengo y lo que he logrado. A ustedes, mis padres amados, les agradezco por haber inculcado en mi el deseo de superación y de triunfo en la vida. Este logro no es solo mío, es de ustedes por cada vez que desvelaron sus sueños para acompañarme a cumplir los míos, por cada vez que le pedían a Dios para iluminarme a cumplir mis metas. Como bien me dijeron, "Echarle ganas al día a día es Triunfar", la cual fue la frase inspiradora de este Trabajo especial de grado. Gracias por ser mis padres. Los amo!

A mi novio Enrique Buono gracias por tu paciencia y comprensión durante el recorrido de mi carrera universitaria. Gracias inmensas por tu amor incondicional, fuente de sabiduría, calma, consejos y apoyo en todo momento durante el desarrollo de este trabajo especial de grado, te amo y te amare por siempre!. A la Sra. Niurka quien estuvo siempre presente en el desarrollo de esta Tesis, muchas gracias!.

A mi Familia fuente de apoyo constante e incondicional en toda mi vida y más aun en mis años de carrera universitaria. En especial quiero expresar mi más grande agradecimiento a mi Tía Chila, que sin su ayuda no hubiese sido posible empezar a estudiar la carrera de la cual me enamore y desempeño hoy por hoy con tanto orgullo.

A Sofia Vallejo, tú quien has sido mi mano derecha durante el desarrollo de este Trabajo especial de grado; te agradezco por tu desinteresada ayuda, por brindarme tu mano cuando más la necesité, por aportarme tus conocimientos y experiencia. Te agradezco no solo por la ayuda brindada, sino por tu amistad y todos los buenos momentos que hemos convivido. Eres una gran persona, que Dios te Bendiga.

A mis profesores, por el apoyo que brindaron durante mi formación como profesional. A mis tutoras, la Profa. Concettina Di Vasta y la Profa. Brenda López. Gracias por creer en mí y apoyarme en el desarrollo de este Trabajo Especial de Grado que fue un reto tanto para ustedes como para mi persona. Gracias por sus orientaciones, persistencia, paciencia, dedicación, apoyo y motivación.

Muchas fueron las personas que en forma directa o indirecta y aun sin saberlo, me ayudaron, poniendo a mi disposición el valor incalculable de sus conocimientos, compartiendo mis dudas y ansiedades, o apoyándome e impulsándome para seguir adelante. A todas ellas, muchas gracias!.

La vida se encuentra inmersa a retos y uno de ellos para mí era la Universidad. Tras verme dentro de ella, me di cuenta que mas allá de ser un reto, era una base no solo para mi desarrollo como Licenciada en Computación, sino para lo que concierne a mi vida y mi futuro. Agradezco a la Universidad Central de Venezuela por abrirme las puertas para formarme profesionalmente dentro de la casa que vence las sombras. Por siempre estaré orgullosa de ser UCEVISTA.

UNIVERSIDAD CENTRAL DE VENEZUELA.
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN
CENTRO DE INVESTIGACIÓN EN SISTEMAS DE INFORMACIÓN

SOLUCIÓN DE BIG DATA QUE APOYE A LA FASE DE RECLUTAMIENTO DE LA GESTIÓN DEL TALENTO HUMANO EN EL ÁREA DE TECNOLOGÍA DE LA INFORMACIÓN

Autor: Br. María Gabriela De Freitas Padrón

Tutores: Profa. Brenda López

Profa. Concettina Di Vasta

Fecha: Caracas, Octubre de 2016

RESUMEN

Hoy en día las organizaciones usan constantemente las bolsas de empleo como medio para postular ofertas laborales, convirtiéndose en una herramienta francamente útil para el área de gestión de talento humano, ya que acorta el tiempo de reclutamiento de nuevos empleados para las organizaciones. En la actualidad los encargados de la gestión de talento humano en las organizaciones deben encontrar a los mejores candidatos de manera más rápida y eficaz. Por lo tanto es necesario contar con mecanismos que permitan procesar gran volumen de información a la misma velocidad que se está generando y poder integrarlas con otras fuentes de datos que son usadas tradicionalmente para la fase de reclutamiento, poderlas analizar y ayudar en la toma de decisiones y así encontrar al candidato ideal para la organización de manera más asertiva.

Con el objetivo de ofrecer mejoraras para solventar este inconveniente surge esta investigación que tiene como objetivo el desarrollo de una solución de Big Data para la generación de indicadores, los cuales podrán nutrir de información al área de Gestión de Talento Humano para apoyar en la toma de decisiones, partiendo de la adquisición de información referente al área de TI proporcionada por las bolsas de empleo tales como Empléate y Bumeran.

Palabras Clave: Gestión de talento humano, big data, bolsas de empleo, tecnología de la información, indicadores.

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

En la actualidad los encargados de la gestión de talento humano en las organizaciones deben encontrar a los mejores candidatos de manera más rápida y eficaz. Hoy en día las organizaciones usan constantemente las bolsas de empleo para postular sus ofertas laborales, convirtiéndose en una herramienta que facilita la fase de reclutamiento para el área de gestión de talento humano.

Se hace cada vez más necesario contar con mecanismos que puedan procesar gran volumen de información a la misma velocidad que se está generando en las ofertas laborales que las organizaciones realizan a través de las bolsas de empleo y poder integrarlas con las otras fuentes de datos que son usadas tradicionalmente para la fase de reclutamiento y selección. Por tal motivo se necesita contar con una solución que permita la unificación de todas las fuentes de datos, poderlas analizar y ayudar en la toma de decisiones y así encontrar al candidato ideal para la organización de manera más asertiva.

La principal tendencia tecnológica que permite unificar un gran volumen de datos generados por diversas fuentes, poderlas analizar y procesar rápidamente para brindarle información nueva y de valor a la organización es Big Data, la cual en este trabajo de investigación se plantea como una solución a la problemática que se presenta actualmente en la gestión de talento humano de las organizaciones.

El objetivo del presente trabajo especial de grado es desarrollar una solución de Big Data que apoye a la fase de reclutamiento de la Gestión del Talento Humano en el área de Tecnología de la Información.

A continuación, se presenta una breve descripción del contenido abarcado en los cuatro (4) capítulos del presente trabajo especial de grado:

PROBLEMA DE INVESTIGACIÓN. En este capítulo se plantea la situación actual y las dificultades asociadas al problema para el cual se presenta la solución. Además se plantean el objetivo general, los objetivos específicos, la justificación y el alcance de la solución.

MARCO CONCEPTUAL. En este capítulo se presentan las bases teóricas conceptuales y las herramientas tecnológicas que dan soporte al desarrollo de una solución de big data para la obtención de indicadores que apoyen a la gestión de talento humano específicamente en la fase de reclutamiento en el área de tecnología de la información.

MARCO METODOLÓGICO. En este capítulo se describe el método seleccionado para llevar a cabo el desarrollo de una solución de Big Data. La metodología presentada fue propuesta por (Krishnan, 2013).

MARCO APLICATIVO. En este capítulo se explican y describen las fases de trabajo que se aplicaron para el análisis, diseño y desarrollo de la solución propuesta, haciendo uso del método de desarrollo seleccionado.

Para finalizar se presentan las conclusiones y referencias bibliográficas que se utilizaron en el desarrollo del presente trabajo especial de grado.

CAPÍTULO 1
PROBLEMA DE INVESTIGACIÓN

1.1. Planteamiento del Problema

Según un estudio realizado por (Randstad, 2012) la segunda empresa de recursos humanos del mundo y la primera del mercado en la Península Ibérica, hoy en día las organizaciones consumen 72 horas laborales en identificar a los candidatos que más se ajusten a posiciones muy especializadas o críticas, roles específicos, competencias y habilidades requeridas, además de cualidades interpersonales que más se ajusten a las culturas organizacionales. Cabe destacar que el área de gestión de talento humano junto al área vacante de la organización, publican avisos y a su vez ubican posibles prospectos a través de Bolsas de Empleo, tales como Empléate y Bumeran.

Todo este contexto es un reto para las organizaciones que deben rápidamente dar respuesta a la operatividad y proceso productivo de las áreas que se quedan sin personal especializado, sea por despido, renuncia, ascenso, cambio de cargo o de área, o porque se crea una nueva vacante o una nueva división dentro de la organización, entre otros. El área de gestión de talento y las áreas afectadas deben ser capaces de cubrir las posiciones que estén allí abiertas a la brevedad posible, para evitar el menor efecto en la operación de la organización.

Cubrir las posiciones ejemplificadas anteriormente, requiere iniciar una fase de reclutamiento del personal, lo cual genera costos para las organizaciones (costos de captura y análisis de perfiles profesionales, costos de llamadas a prospectos entre otros costos indirectos), por tanto es necesario conseguir la mayor asertividad en el reclutamiento y la selección de los candidatos para reducir estos costos, además de que les interese permanecer en la organización. Un estudio realizado por la revista Forbes (Josh, 2013) nos muestra que la retención de talento en la organización les ahorra en el largo plazo.

La asertividad descrita anteriormente aumenta a medida que se tome en cuenta mayor información y más variables de estos candidatos, construyendo así un modelo que se nutra del análisis e historia de los perfiles requeridos en la organización. El resultado permite colocar al prospecto mejor calificado en una posición donde sus habilidades sean óptimamente utilizadas.

El problema que se presenta para las organizaciones actualmente es lograr manejar esa diversidad de información de manera oportuna para reducir los tiempos/costos y encontrar, atraer y retener las mejores personas del mercado para cubrir las vacantes en la organización. Un grupo de empleados capacitados se traduce en un aumento de productividad, pero encontrar estas personas no es una tarea fácil.

Las organizaciones hoy en día se han ido acoplando al gran avance tecnológico para ayudar a solventar dicha problemática, adaptándose a diferentes desafíos comentados anteriormente que van desde manipular la gran cantidad de datos adquirida de diversas fuentes, poder analizar todos los datos recolectados a una alta velocidad y ayudar al área de gestión de talento humano a la toma de decisiones en la fase de reclutamiento, logrando así no atrasar la operatividad de la organización. Con el uso de tecnologías como Big Data las organizaciones pueden enfrentar estos desafíos de una manera más rápida. Integrando diversas fuentes de datos, lo que conlleva a obtener un gran volumen de información, la cual se almacena y utiliza para determinar información asociada a los perfiles de empleados que la organización está buscando.

Con el fin de dar una solución al problema planteado, se establecen los siguientes objetivos para este trabajo especial de grado.

1.2. Objetivos

1.2.1. General

Desarrollar una solución de Big Data que apoye a la fase de reclutamiento de la Gestión del Talento Humano en el área de TI.

1.2.2. Específicos

- Analizar los requerimientos del usuario.
- Adquirir data de bolsas de empleo tales como Empléate y Bumeran relacionadas a TI.
- Organizar, procesar y almacenar los datos adquiridos de las bolsas de empleo.
- Elaborar indicadores que apoyen a la gestión de talento humano.
- Implementar una arquitectura de una solución de big data para apoyar a la fase de reclutamiento de la gestión de talento humano en el área de TI.
- Realizar pruebas unitarias e integrales de los módulos.

1.3. Solución Propuesta

Se propone una solución de Big Data que apoye a la fase de reclutamiento de la Gestión del Talento Humano en el área de TI.

Se emplea un proceso de captura de información de bolsas de empleo como Bumeran y Empléate a través de una araña web, el cual nos permite conocer toda la información referente a los cargos relacionados al área de TI. Dicho contenido una vez exportado del analizador web, es transformado de data no estructurada a estructurada a través de patrones, para ser almacenado en un repositorio y posteriormente poder obtener indicadores. Este proceso de captura de la información permite el análisis de resultados a partir de la obtención de indicadores de empleos haciendo uso de soluciones de big data. Para lo antes descrito, se plantea la siguiente arquitectura de solución, la cual puede observarse en la Figura 1.

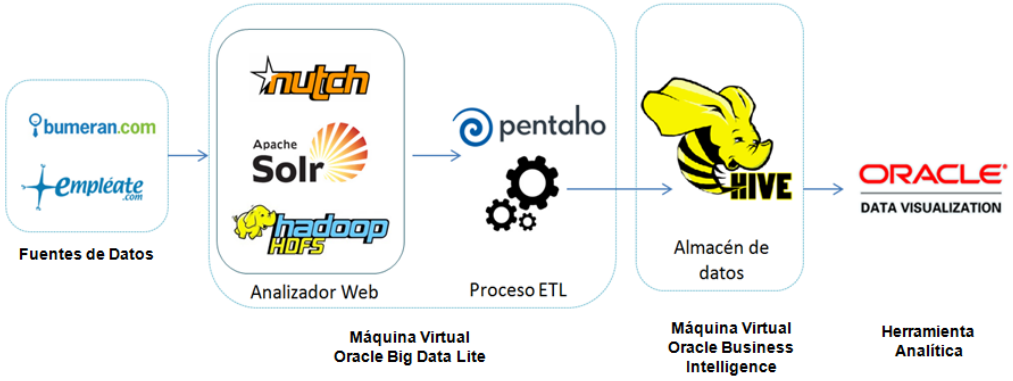


Figura 1: Arquitectura de Solución Propuesta

El primer componente de la arquitectura son las **fuentes de datos**, siendo estas las bolsas de empleo Bumeran y Empléate. Estas fuentes de datos se conectan a través de sus URLs con **Apache Nutch** quien es una araña web y se encarga de extraer todo el contenido de estos e indexarlo a **Solr** que a su vez son almacenados en **Hadoop HDFS**. Una vez que los datos se encuentran en HDFS se extraen y son tratados con la herramienta **Pentaho Data Integrator,** la cual nos apoya en el proceso de ETL, esto con el fin de determinar los patrones que poseen cada una de estas distintas bolsas empleo para que las empresas completen y carguen sus vacantes, y así poder limpiar y transformar data no estructurada

en data estructurada. Una vez que se determina el patrón y se separa el contenido útil para el análisis, es exportado y cargado en el almacén de datos **Apache Hive**, el cual nos va a permitir almacenar gran volumen de información y conectarlo con herramientas analíticas, en este estudio se hara uso de **Oracle Data Visualization** para explotar al máximo todo el contenido almacenado y proveer valor para el área de gestión de talento humano.

Para realizar esto posible se cuenta con dos máquinas virtuales separadas, una primera máquina virtual nombrada como **Oracle Big Data Lite** y otra máquina virtual nombrada como **Oracle Business Intelligence**.

1.4. Justificación de la Solución

Hoy en día el área de gestión de talento en las organizaciones deben analizar diversidad de información originada de muchas fuentes de datos lo que se ha convertido en un reto, ya que deben rápidamente dar respuesta a la operatividad y proceso productivo de las áreas que se quedan sin personal especializado, sea por despido, renuncia, ascenso, cambio de cargo o de área, o porque se crea una nueva vacante o una nueva división dentro de la organización, entre otros.

Cubrir las posiciones ejemplificadas anteriormente, requiere iniciar una fase de reclutamiento y selección del personal, lo cual genera costos para las organizaciones (costos de captura y análisis de perfiles profesionales, costos de llamadas a prospectos, costos por aplicar pruebas de aptitud y pruebas psicológicas, tiempo invertido en entrevistas por parte de todos los interesados, entre otros costos indirectos), por tanto es necesario conseguir la mayor asertividad en la selección de los candidatos para reducirlos. La asertividad descrita anteriormente aumenta a medida que se tome en cuenta mayor información y más variables de estos candidatos, construyendo así un modelo que se nutra del análisis e historia de los perfiles requeridos en la organización.

Con dicha solución las organizaciones pueden emplear la fase de reclutamiento de una manera más rápida y precisa, integrando diversas fuentes de datos, lo que conlleva a obtener un gran volumen de información de los candidatos postulados, la cual se almacenada y utilizada para determinar información asociada a los perfiles de empleados que la organización está buscando.

1.5. Alcance de la Solución

El desarrollo de esta solución de Big Data se limita a la captura de datos relacionadas a organizaciones del área de TI, provenientes de dos bolsas de empleo tales como Empléate y Bumeran en el territorio Venezolano. Logrando así la obtención de algunos indicadores que permitan al área de gestión de talento humano realizar de manera más sencilla la fase de reclutamiento adquiriendo un gran volumen de información asociada a los perfiles de los empleados que la organización esté buscando.

Adicional a esto, se capturó sólo información de las bolsas de empleo durante el período Mayo 2016 – Agosto 2016, por falta de recursos físicos tales como disco y memoria, los cuales son indispensables para poder implementar una solución de big data, que capture mayor cantidad de información. El desarrollo de este trabajo de investigación, se realizo en dos máquinas virtuales las cuales son nombradas como Oracle Big Data Lite y Oracle Business Intelligence.

CAPÍTULO 2
MARCO CONCEPTUAL

2.1. Talento humano

2.1.1. Definición del Talento Humano

El talento humano, es definido por (Balza, 2010), como el conjunto de conocimientos y actitudes de los individuos y grupos de trabajo en las organizaciones, pero también a sus habilidades, convicciones, aptitudes, valores, motivaciones y expectativas respecto a la organización, el trabajo y a la sociedad.

Según (Belén, 2015) el talento es la capacidad dinámica que deben desarrollar las organizaciones para tener éxito. Tanto el talento individual como el organizativo son fundamentales para la implementación de una excelente estrategia de negocios. Se puede decir que el talento en las personas implica un conjunto de conocimientos, habilidades, capacidades, motivaciones y actitudes puestas en práctica por un profesional comprometido con la organización. El talento no es innato, se puede desarrollar a través de los tres elementos claves que lo componen al mismo tiempo: la capacidad, el compromiso y la acción. Es decir, si el postulado tiene compromiso y actúa, pero no dispone de las capacidades necesarias, casi seguro que no alcanzará resultados, aunque haya tenido buenas intenciones. Si por el contrario, dispone de capacidades y actúa en el momento, pero no se compromete con el proyecto, puede que alcance resultados pero sin innovar. Puede ocurrir también que el postulado tenga las capacidades y el compromiso, pero cuando actúa ya ha pasado el momento, por la sencilla razón de que alguno de la competencia se le ha podido adelantar. Finalmente la gestión de talento humano en las organizaciones debe identificar en los candidatos los tres elementos del talento humano mencionados anteriormente, y a su vez apalancar sus habilidades y capacidades, en las metas estratégicas de la organización.

2.1.2. Definición de Gestión de Talento

Según el (Centro Nacional de Información de Calidad, 2011) la gestión de talento debería ser considerada como la principal inquietud de los Gerentes, Directores Generales y Responsables de recursos humanos de las organizaciones, ya que esta gestión es fundamental en un mercado competitivo y globalizado, donde las organizaciones tienden a diferenciarse por su capital humano.

La Gestión de Talento es clave para diferenciar a la organización en el mercado actual, este elemento diferenciador no siempre se consigue. En realidad, las organizaciones que logran llegar al éxito gracias a la gestión de talento es porque dedican su tiempo y esfuerzo a entender las tendencias del mercado laboral, cuando hablamos de tendencias tecnológicas en ámbito de TI se refiere a las siguientes según International Data Corporation (IDC): Big Data, Almacenamiento En La Nube, Movilidad, Internet De Las Cosas y Redes Sociales.

La Gestión de Talento Humano o Gestión Humana según (Veras & Cuello, 2005), es la forma como la organización libera, utiliza, desarrolla, motiva e implica todas las capacidades y el potencial de su personal, con miras a una mejora sistemática y permanente tanto de éste como de la propia organización. Su fin principal consiste en promover el desarrollo de las competencias de las personas por medio de una labor coordinada y de estrategia de mejoramiento continuo del conocimiento y el talento humano. La Gestión del Talento Humano es dinámica, interactiva e integral, de manera que permite tanto a la organización como a sus colaboradores crecer juntos y desarrollar al máximo sus potencialidades. El motor del cambio en las organizaciones está en las personas, ellas son las portadoras del conocimiento que permite la innovación y la adaptación permanente. Es por ello por lo que las tendencias de las mejores prácticas de Gestión Humana en las organizaciones modernas van dirigidas a formular programas de gestión de conocimientos y gestión de personal que contribuyan a generar una cultura de cambio en el proceder de los ejecutivos, los accionistas y los colaboradores y de esta manera alinear la productividad interna con un mundo en constante cambio y que sólo el individuo con su capacidad de análisis e innovación abordará.

Los elementos básicos de la Gestión de Talento Humano vienen dados por la planeación, la formación, la selección, el liderazgo, el seguimiento, la evaluación, la calidad y el bienestar del talento humano. La gestión de talento se lleva a cabo a través de fases que componen al sistema de gestión de talento, sin embargo en algunas organizaciones no se suele llevar a

cabo en su totalidad, debido a que parten de alguna fase intermedia y no desarrollan el procedimiento completo. En el siguiente tópico se estudiara en detalle cada una de las fases que componen al sistema de gestión de talento humano, haciendo énfasis en la fase de reclutamiento y selección por objetivo del estudio.

2.1.3. Sistema de Gestión de Talento

Según (Marcos, 2015) el sistema de gestión de talento se desarrolla en sucesivas fases y las organizaciones deben ser conscientes de que es un sistema complejo. Las fases que componen al sistema de gestión de talento son las siguientes, tal como se muestra en la figura 2:

Figura 2: Fases del sistema de gestión de talento

Fuente: Apoyo administrativo a la gestión de recursos humanos.

Reclutamiento y Selección: esta primera fase del sistema de gestión de talento comprende a las actividades dirigidas a cubrir las necesidades del personal y afrontar la carga productiva en la organización. El reclutamiento consiste según (López, 2003) en una acción dinámica y flexible cuyo objetivo es buscar y reunir el máximo número de candidatos en un tiempo determinado. Es importante que esta fase se realice correctamente, ya que si se recluta candidatos no aptos para las vacantes en la organización, se invertirá tiempos/costos y el rendimiento del empleado en la organización se verá afectado negativamente. En el proceso general de selección, el reclutamiento se realiza de una forma detallada. Se puede hablar de dos tipos de reclutamiento:

Reclutamiento externo: este tipo de reclutamiento se realiza utilizando fuentes externas a la organización para reunir candidatos potenciales. Como fuentes externas se puede citar algunas: anuncios en prensa, centros de enseñanza, oficinas de empleo, bolsas de empleo, entre otros. Este tipo de reclutamiento conlleva ciertas ventajas, como por ejemplo, se tiene

acceso a un número mayor de candidatos para cubrir una posición dentro de la organización, al no pertenecer a la organización y venir de otras culturas organizacionales aportan nuevas ideas o si el empleado no ha trabajado en otras organizaciones se tiende a adaptar más fácil a la organización.

El reclutamiento de personas a través de fuentes externas como las bolsas de empleo están solapando ya los canales tradicionales para el reclutamiento de personal, como por ejemplo a las oficinas de empleo, sobre todo porque el candidato tiene acceso a las oportunidades de trabajo de las organizaciones.

Reclutamiento interno: este tipo de reclutamiento se realiza utilizando fuentes internas para reunir a los posibles candidatos, y esto quiere decir, que el reclutamiento se realiza con personal dentro de la organización. Generalmente, este reclutamiento se debe a rotaciones o posibles traslados de personas, o bien a promociones de los trabajadores. Este tipo de reclutamiento conlleva ciertas ventajas, como por ejemplo, un costo más bajo, se conoce mejor a los candidatos y por tanto el proceso es más valioso, los candidatos ya conocen a la organización y esto ahorra tiempo respecto a su integración y adaptación.

Una vez culminada la fase de reclutamiento, solo faltaría la parte más objetiva, la selección de los mismos. Existen diversas técnicas de selección las cuales ayudan a escoger al mejor candidato, entre ellas se encuentran el análisis del currículum vítae, entrevistas, pruebas psicológicas, personalidad, motivación e intereses y habilidades sociales del candidato, pruebas prácticas profesionales y pruebas de conocimiento.

Se hace énfasis en conocer en detalle esta fase de reclutamiento y selección en la gestión de talento humano, principalmente porque a través del uso de tecnologías como Big Data se puede almacenar un gran volumen de datos proveniente de diversas fuentes. Permitiendo que en esta fase se realice un análisis profundo de los candidatos y poder determinar patrones asociados a los perfiles de empleados que la organización está buscando previo a seguir con las siguientes fases del sistema de gestión de talento, como lo son, las fases de evaluación del desempeño, desarrollo de personas y finalmente la fase de retención de talento. Luego se continúa con la segunda fase del sistema de gestión de talento, la cual es la evaluación del desempeño.

Evaluación del desempeño: esta segunda fase del sistema de gestión de talento se realiza a través de la medición del grado de cumplimiento de los objetivos y la generación del valor de los empleados. Algunas organizaciones incluyen sistemas de seguimiento del desempeño del personal pero pocas trabajan en la evaluación del potencial para poder medir e identificar el desempeño del talento.

Desarrollo de personas: esta tercera fase del sistema de gestión de talento comprende aquellos esfuerzos que realiza la organización para desarrollar profesionalmente a sus empleados a través de entrenamientos, cursos o certificaciones para ayudarlos a crecer profesionalmente. Estas iniciativas han cobrado protagonismo en los últimos años en las organizaciones que más apuestan por la gestión del talento humano.

Retención de talento: esta cuarta y última fase del sistema de gestión de talento se basa fundamentalmente en desarrollar una buena política salarial y otras iniciativas relacionadas con la generación de compromiso y motivación por parte del empleado, contribuyendo a una relación sólida entre la organización y el mismo.

Con el tiempo los métodos de reclutamiento se han ampliado drásticamente. Según (Giacomelli, 2009) el 60% de las personas que buscan trabajo utilizan internet en especial las bolsas de empleo como principal recurso, por lo cual las postulaciones por vacante están en constante ascenso, de ahí que el problema actual de los reclutadores no sea la cantidad, sino la calidad de los CV recibidos, según un estudio de la bolsa de trabajo por Internet Bumeran.com México.

2.2. Bolsas de Empleo

Las bolsas de empleo, son una de las tantas maneras que existen en la actualidad para buscar empleo y encontrar candidatos que puedan ocupar las vacantes en la organización. No es el medio más fácil de conseguir empleo, ya que sin duda el contacto directo con la organización que tiene la necesidad o vacante, junto con la previa recomendación y referencia con el área encargada de la selección, es el sistema más expedito para que una persona se de a conocer.

Según (Gastélum & Campas, 2009) las bolsas de trabajo tuvieron sus orígenes en las universidades. Fueron diseñadas para buscar empleo a sus egresados. Cada facultad, poseía

una base de datos, con los nombres de las personas egresadas de sus aulas. Por ende, las ponían en contacto, con aquellas organizaciones que buscaban personal correspondiente a esa carrera.

Por ende, las bolsas de trabajo, eran el lugar "físico", si es que se puede llamar así, donde las universidades actuaban de interlocutoras, entre las organizaciones que ofrecían el puesto de trabajo y la demanda, que vendrían siendo los alumnos egresados de aquella universidad. Sin embargo, hoy en día, el concepto de bolsas de empleo se ha ampliado. Según (Coordinación Empresarial, 2014) una bolsa de empleo es un espacio en el cual se registran una serie de ofertas laborales, y a su vez lo mismo sucede con las personas que buscan empleo. Las bolsas de trabajo pueden ser públicas o privadas, en función del puesto al cual se oposite, y en el caso de las bolsas privadas existen dos tipos más: las propias de las organizaciones, y los portales en los cuales se da un cúmulo de oferta y demanda de puestos de trabajo y la bolsa de trabajo ejerce una función de intermediario.

- **Bolsas de empleo públicas**

En estas bolsas, se ofertan puestos de trabajo para organizaciones públicas, es decir, pertenecientes al estado, y en las cuáles los beneficios se destinan a gastos públicos. En este tipo de bolsas, se ofrece por parte de las organizaciones públicas información sobre convocatorias para opositar, requisitos mínimos, etc. En este caso no se habla de intermediario puesto que la bolsa es publicada y actualizada por la misma parte contratante.

- **Bolsas de empleo privadas**

Estas bolsas de empleo se basan en un portal web en el cual acceden dos tipos de usuarios, por un lado las organizaciones que necesitan cubrir un puesto de trabajo, y por otro lado aquellas personas que están en busca del empleo. En este caso, el portal ejerce de intermediario, conectando los empleados con los puestos vacantes. Y todo esto funciona gracias al proceso de selección que realizan dichos portales, empezando por el registro de las ofertas laborales con los requisitos pertinentes al puesto de trabajo y por otro lado los registros de los perfiles de usuarios, en el cual deberán introducir su curriculum vitae para poder ser clasificado.

- **Bolsas de empleo pertenecientes a organizaciones**

Las organizaciones suelen utilizar bolsas de empleo de trabajo propias, en las cuales se encuentra un registro interno de todos los candidatos entrevistados, por entrevistar, y cuáles han sido sus valoraciones. Esta es una herramienta útil ya que facilita el registro de posibles empleados, y el posible acceso a ellos, ayuda a organizar todas aquellas personas que puedan encajar dentro de la organización en una base de datos, de fácil acceso, y que está ordenada con datos de interés para la organización y no por un portal ajeno.

En la actualidad, las bolsas de empleo, prestan un valioso servicio para aquellas personas que están en la búsqueda de un puesto laboral. Asimismo, aquellas personas que desean cambiar de trabajo, también pueden hallar aquel que sea de su gusto. La ventaja principal de las bolsas de empleo reside en la facilidad de acceso a las ofertas de trabajo, cada usuario puede acceder a una gran cantidad de ofertas de trabajo y postularse a diferentes solicitudes al mismo tiempo. Y a su vez, en las bolsas de empleo normalmente suele haber publicidad de escuelas de formación, en las cuales pueden proporcionar la formación necesaria para el cargo. Por otro lado, el hecho de que sea una herramienta gratuita hace que cada persona interesada pueda encontrar el empleo, pueda registrarse en aquellos sitios que más le interesen y a su vez pueda tener acceso a más cantidad de ofertas.

Por otra parte, la gran desventaja de las bolsas de empleo reside en que, como suele pasar en los portales de empleo que hacen de intermediario entre una parte solicitante y otra solicitadora, pues acaban dándose una sobrecarga de solicitantes. Según (Coordinación Empresarial, 2014) una bolsa de empleo es una herramienta francamente útil, ya que acorta mucho el tiempo de reclutamiento de nuevos empleados para las organizaciones, y por otro lado, los empleados pueden encontrar una gran cantidad de ofertas a las que puede presentar su candidatura a través de *clicks*, con lo cual convierten la búsqueda de empleo en un proceso mucho más sencillo, real y tangible.

Como parte de este trabajo de investigación, fue necesario el análisis de la data contenida en bolsas de empleo, en este sentido, a continuación se presenta toda la información relacionada con las Arañas Web, mediante las cuales, se podrá extraer la mayor cantidad de data de estas fuentes de datos.

2.3. Arañas Web

Según (Stecanella & Bonanata, 2015) los llamados en ingles *Web Crawler* o Araña Web son programas que inspeccionan las páginas web de forma metódica y automatizada. Uno de los usos más frecuentes consiste en crear una copia de todas las páginas web visitadas para su procesado posterior por un motor de búsqueda que indexa las páginas proporcionando un sistema de búsqueda rápido y eficiente.

Las Arañas Web son programas que utilizan diversos algoritmos para extraer datos a partir de páginas web. Por ejemplo el título de un documento embebido en la página, el contenido de una página web, o incluso metadatos utilizados por redes sociales. Estos programas, a partir de unos enlaces base que se indican previamente a la ejecución, obtienen sucesivamente nuevos enlaces con los que continuar el rastreo mientras se cumplan las condiciones establecidas por el usuario. Estos programas se configuran mediante diversos parámetros de entrada y devuelven como salida los archivos encontrados.

2.3.1. Componentes

Aunque existan diferentes tipos de Arañas Web, y cada programa pueda ofrecer utilidades propias, se pueden destacar unos componentes básicos comunes según (Trupti, Ravindra D, & Dharmik, 2014) los cuales se enumeran a continuación:

- **Frontera de la Araña:** su función es albergar la lista de enlaces pendientes de ser rastreados, y seleccionar, cada vez, el próximo enlace que se descargará. Una vez se ha descargado el contenido de una página web, es probable que se hayan encontrado nuevos enlaces (que cumplan los filtros), los cuales serán, a su vez, añadidos a la frontera y quedarán a la espera de ser seleccionados para ser descargados.
- **Descargador de páginas:** este componente se encarga de descargar las páginas web que le indica la frontera. Funciona como un cliente HTTP que envía peticiones e interpreta las respuestas del servidor. Dependiendo de los parámetros indicados a la araña, este componente puede cancelar la descarga de un enlace si excede en tiempo o en tamaño.
- **Repositorio:** es el espacio en el que se guardan los archivos descargados.

2.3.2. Clasificación

Las Arañas Web se pueden clasificar en cuatro categorías de acuerdo con el modo en el que se ejecutan.

- Arañas Web centradas: su objetivo principal es descargar páginas relacionadas con un tema concreto. Por tanto, es necesario hallar la relevancia respecto del tema de cada página que se visita para determinar si será descargada. El principal problema de este método radica en conocer qué páginas serán similares al objetivo sin haberlas descargado. La ventaja es que minimiza considerablemente el número de páginas que se visitan, por lo que es el más adecuado si no se dispone de un hardware que permita realizar búsquedas avanzadas y se pretenden encontrar archivos pertenecientes a un determinado tema. La primera opción para manejar este método consiste en ejecutar una araña web centrada por definición, y que, mediante un algoritmo, escoja de la frontera, con prioridad, las páginas con más relevancia respecto del tema en el que se centra la araña web. Otra opción consiste en ejecutar una araña web tradicional añadiéndole filtros. Por ejemplo, al emplear el filtro ".gif", la araña web se centrara en descargar las imágenes con formato GIF que estén contenidas en los enlaces base.
- Arañas Web incrementales: un rastreador se puede ejecutar en modo incremental, una vez que haya finalizado, para mantener los archivos del repositorio actualizados. La diferencia radica en que en el modo tradicional se repetiría todo el proceso de rastreo para que los resultados estén actualizados, mientras que en éste, la actualización de resultados se realiza de forma incremental visitando las páginas frecuentemente, mediante un algoritmo que indica qué páginas hay que actualizar. Según (Edwards, McCurley, & Tomlin, 2011) existen diversas medidas para estimar el momento de realizar las actualizaciones, tales como el tiempo que ha transcurrido desde que el contenido de dicho enlace fue modificado en la página web de origen.
- Arañas Web paralelas: debido al notable volumen de enlaces que alberga la frontera, estos programas son muy paralelizables. Dicha paralelización resulta necesaria cuando se realiza una búsqueda con una cantidad de enlaces desmesurada. Este modo se emplea en diversos motores de búsqueda, ya que no son arañas webs centrados en un tema, sino que realizan un rastreo general para analizar la mayor parte posible de la Web y de esta forma poder devolver al usuario de forma precisa las páginas web que más se asemejen con la consulta que se haya introducido.
- Arañas distribuidas: son las arañas web que siguen el modelo de computación distribuida mediante su ejecución, la cual permite que un programa se ejecute en diversas áreas que no están conectadas físicamente. Este método, junto con la paralelización, posibilita la ejecución de grandes búsquedas sin poseer físicamente un equipo con la suficiente capacidad para ello. Se emplea en diversos motores de búsqueda para realizar consultas

de una gran dimensión de forma paralela en un tiempo viable. Desde hace años, las arañas webs más importantes (GoogleBot, Yahoo!, Slurp…) lo han empleado.

2.3.3. Programas de Arañas Webs

Según (Journal of Telecommunications, 2013) existen diversos programas que han sido desarrollados para ejecutar técnicas de arañas web y que están catalogados como software libre. No obstante, las características de mucho de esos programas son similares. A continuación se analizan cuatro arañas web que poseen distintas características:

- **Web Crawler de IBM:** se incluye en Analytics for Apache Hadoop, que es la versión en la nube de la máquina virtual ofrecida por IBM (IBM Infosphere BigInsights for Hadoop 3.0).Sencillo de utilizar, sin embargo, esa sencillez implica que no se pueda realizar una ejecución realmente personalizada.
- **Httrack:** según (HTTrack Website Copier, 2007) es un programa (software libre) multiplataforma, desarrollado en C y que posee una interfaz gráfica para su ejecución. Ofrece numerosos parámetros de configuración, por lo que se puede ajustar la búsqueda a los requerimientos de un trabajo correcto. Existen diversas opciones de prioridades para los enlaces: priorizar los archivos HTML o los que no lo sean, priorizar páginas del mismo dominio, directorio, subdirectorios o directorios superiores. Además, se puede elegir el número de procesos (enlaces) que se ejecutan (rastrean) en paralelo. Debido a que guarda una copia local de los archivos que rastrea (útil para posteriormente tratar los datos de forma offline) se considera una araña web adecuado para ejecuciones en modo centrado, ya que resultaría inviable descargar en un equipo local todos los archivos de una búsqueda con un nivel de profundidad elevado y sin filtros.
- **GNU Wget:** según (GNU operating System, 2010) además de ser un comando de consola con numerosos parámetros, se puede ejecutar en un script de forma que posteriormente se puedan manejar otros programas sobre los datos que el comando ha descargado de Internet. Está desarrollado en C, y es, además de multiplataforma, software libre. Posibilita la opción de descargar recursivamente enlaces de una página web, hasta descargar la página web al completo. De manera similar al Httrack, guarda los archivos en local, y es más complejo de utilizar que el Web Crawler de IBM, pero es más personalizable.

Entre los diferentes programas de Araña Web, se encuentra Apache Nutch, el mismo será descrito a profundidad, dado que, para el presente trabajo de investigación será usado como parte de la solución propuesta.

2.3.4. Apache Nutch

Según (Apache Nutch, 2014) es una araña web software libre altamente extensible y escalable. Su objetivo final es implementar un motor de búsqueda para la web que permita fácilmente rastrear e indexar páginas web para luego realizar búsquedas sobre las mismas.

En este trabajo especial de grado, se hará uso de Apache Nutch según la clasificación de araña web, como una araña web centrada. Ya que su objetivo principal será descargar páginas webs relacionadas con un tema en especifico.

2.3.4.1. Arquitectura de Apache Nutch

Nutch se divide naturalmente en dos partes, la araña web y el motor de búsqueda. La araña web lo que hace es descargar las páginas e introducirlas en el índice donde luego el motor de búsqueda realiza sus consultas, el motor de búsqueda es básicamente una instancia de Solr. A su vez, tanto la araña web como la instancia a Solr son altamente parelizables para su ejecución en clústeres mediante Hadoop o Solr Cloud.

2.3.4.2. La Araña Web de Nutch

El sistema de araña web de Nutch está manejado por una familia de herramientas de diversas estructuras de datos. Entre estas se encuentran la base de datos web, un conjunto de segmentos y el índice, tal como aparece a continuación en la Figura 3.

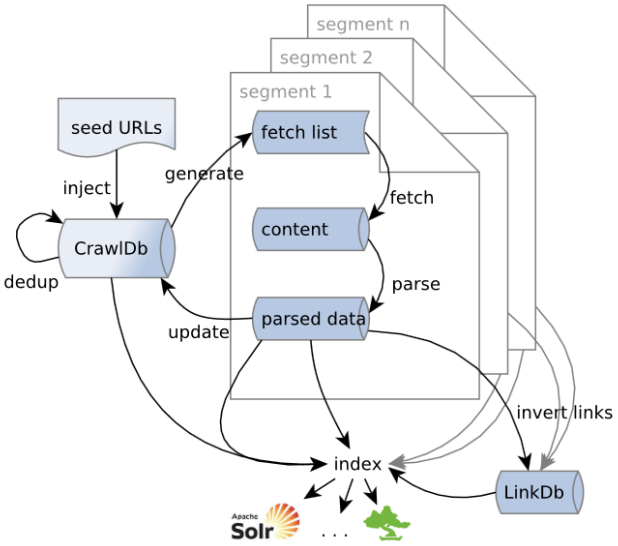


Figura 3: Herramienta Nutch

Fuente: (Nagel, 2014)

La base de datos web (**CrawlDb**) es una estructura de datos persistente que tiene como objetivo mantener la estructura y propiedades del grafo web que está siendo rastreado. La CrawlDb no cumple ningún rol durante la búsqueda de la información, sino que guarda dos tipos de entidades: páginas y links.

Una **página** representa una página web y es indexada por su URL y un hash MD5 de sus contenidos. Además también se guarda otra información acerca de la página incluyendo la cantidad de links en la misma, información acerca de cuándo la página fue rastreada y el puntaje de la página basado en la cantidad de links que apuntan a la misma.

Un **link** representa una URL encontrada dentro de una página que apunta hacia otra distinta. En el grafo web de la CrawlDb un nodo es una página y un vértice es un link.

Un **Segmento** es un conjunto de páginas rastreadas e indexadas en una misma corrida del programa. La lista de búsqueda de un segmento es la lista de URLS de páginas que la araña web debe descargar y es generada a partir de la CrawlDb. La información de las páginas guardadas en un segmento se vuelve obsoleta una vez que todas las páginas han sido nuevamente rastreadas. El período de tiempo para un re-rastreo es configurable.

Un **Índice** se crea a partir de la unión de todos los segmentos.

La Araña Web de Nutch es un proceso cíclico que sigue los siguientes pasos: la araña web genera una lista de búsquedas de urls a rastrear a partir de la CrawlDb, las páginas correspondientes a dichas urls se descargan y se guardan en segmentos y dichos segmentos se indexan a motores de búsqueda, la araña web actualiza la CrawlDb con los nuevos links que fueron encontrados, la araña web genera una nueva lista de busqueda y el ciclo se repite.

La información que arroje la Araña Web se puede transformar en indicadores que ayudan a la toma de decisiones de acuerdo a un análisis que necesiten realizar a una actividad específica, integrando diversos URLS. Con base a lo anteriormente descrito, se presenta el concepto de indicador.

2.4. Indicador

Para (Beltrán Jaramillo, 2006), es la relación entre variables que permite observar la situación actual y las tendencias de cambio generadas con respecto a los objetivos previstos. Representando de esta manera factores para establecer el logro y el cumplimiento de estos objetivos, apoyar el proceso de toma de decisiones dentro de una organización.

Dos de las características más relevantes de un indicador son las siguientes:

- Su pertinencia, fiabilidad, precisión y comparabilidad.
- Su capacidad para resumir la información sin deformarla.

2.4.1. Tipos de Indicadores

En la Figura 4 se observan los diferentes tipos de indicadores:

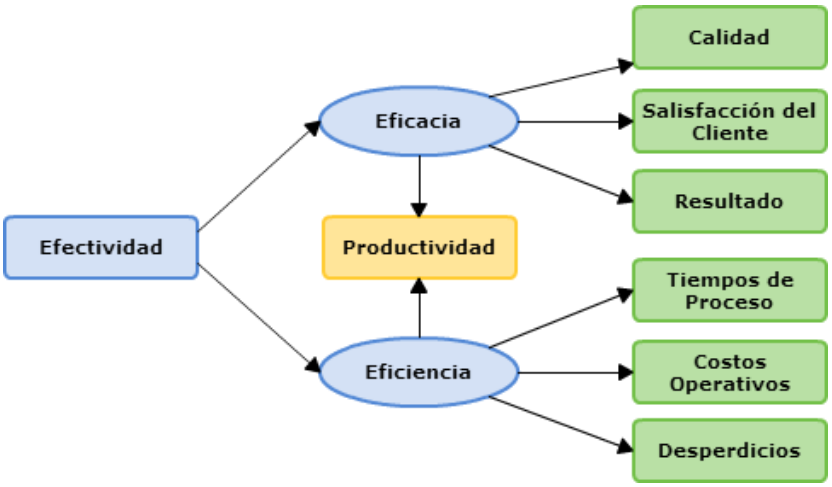


Figura 4: Tipos de Indicadores

Fuente: (Beltrán Jaramillo, 2006)

- **Indicadores de Eficacia:** miden el logro de los resultados propuestos.
- **Indicadores de Eficiencia:** miden el nivel de ejecución del proceso, cómo se hicieron las cosas y el rendimiento de los recursos.
- **Indicadores de Efectividad:** miden la relación entre la eficiencia y la eficacia.

Para (Beltrán Jaramillo, 2006), una organización puede optar por ser eficiente y aprovechar sus recursos con el objetivo de prestar sus servicios sin considerar los medios utilizados. O puede ser eficaz, buscando la satisfaccion de los clientes a partir de estos servicios. Sin embargo, expone que para lograr el éxito no basta ser eficaz o eficiente, sino ser efectivos en cada uno de los procesos y tener un correcto desempeño, es decir, la combinacion de ambas.

Cabe destacar que, la definición, el cálculo y comparación de indicadores que permitan controlar la gestión, son requeridos en un sistema de informacion eficaz; de esta manera, se percibe al sistema no solo como recolector y procesador de datos, sino como generador de la informacion necesaria para la toma de decisiones dentro de la organización. En este sentido a continuación se describen los sistemas de informacion.

2.5. Sistemas de Información

2.5.1. Definición

Según (Laudon & Laudon, Sistemas de Informacion Gerencial, 2004) un sistema de información es un conjunto de componentes interrelacionados que recolectan (o recuperan), procesan, almacenan y distribuyen información para apoyar los procesos de toma de decisiones y de control dentro de una organización. Además de apoyar la toma de decisiones, la coordinación y el control, los sistemas de información también pueden ayudar a los gerentes y trabajadores a analizar problemas, visualizar asuntos complejos y crear productos nuevos. Los mismos contienen información acerca de gente, lugares y cosas importantes dentro de la organización o en el entorno en que se desenvuelve. Por lo tanto por información se entienden a los datos que se han moldeado en una forma significativa y útil para los seres humanos.

El principal propósito de un sistema de información es el procesamiento de información en todas sus etapas: recolección, organización, almacenamiento, proceso y despliegue, desde información primaria, procesada e interpretada hasta el conocimiento.

Los principales componentes de un Sistema de Información son los siguientes:

- Datos: unidad mínima de un Sistema de Información, que se utiliza para alimentar programas y producir información.

- Hardware: según (RAE, 2014) es un conjunto de componentes que integran la parte material de una computadora.

- Software: según (RAE, 2014)es un conjunto de programas, instrucciones y reglas informáticas para ejecutar ciertas tareas en una computadora.

Siguiendo las líneas anteriores, y de acuerdo a su función se distinguen seis tipos de sistemas de información: los sistemas de información ejecutivos (ESS), sistema de soporte de decisiones (DSS), sistemas de información de gestión (MIS), sistemas de gestión de conocimiento (KWS), sistemas de oficinas (OfficeS) y sistemas de procesamiento de

transacciones(TPS). A continuación se estudiará en detalle los tipos de sistemas de información nombrados anteriormente.

2.5.2. Tipos de Sistemas de Información

Los sistemas de información se desarrollan con diferentes objetivos y sobre todo según las necesidades que presente la organización. En la siguiente Figura 5 se muestra la pirámide organizativa, donde cada uno de los niveles representa un tipo de sistema de información específico para brindarles apoyo y simplificar las tareas y tomas de decisiones.

Figura 5: Pirámide de los diferentes tipos de sistemas de información

- **Sistemas de Procesamiento de Transacciones**

Este tipo de sistema de información (TPS, Transaction processing system) recolecta, almacena, modifica y recupera toda la información generada por las transacciones producidas en una organización. Una transacción es un evento que genera o modifica los datos que se encuentran eventualmente almacenados en un sistema de información. Según (Kendall, 2005) son creados para procesar grandes cantidades de datos relacionados con transacciones rutinarias del negocio, como las nóminas y los inventarios. Un TPS elimina la molestia que representa la realización de transacciones operativas necesarias y reduce el tiempo que una vez fue requerido para llevarlas a cabo de manera manual.

- **Sistemas del nivel de conocimiento de la organización**

Existen dos tipos de sistemas en este nivel, los sistemas de automatización de la oficina (OAS, *Office Automation Systems*) que consisten en aplicaciones destinadas a ayudar al trabajo administrativo diario de una organización, forman parte de este tipo de sistemas: los procesadores de textos, las hojas de cálculo, los editores de presentaciones, los clientes de correo electrónico, entre otros.

Los sistemas de trabajo del conocimiento (KWS, *Knowledge Work Systems*) dan soporte a los trabajadores profesionales, tales como científicos, ingenieros y doctores, ayudándoles a crear nuevos conocimientos que contribuyan a mejorar la organización.

- **Sistemas de información gerencial**

Los sistemas de información gerencial (MIS, *Management Information Systems*) tienen como propósito general cooperar a la correcta interacción entre los usuarios y las computadoras. Producen información que es usada para la toma de decisiones, basándose en los sistemas de procesamiento de transacciones. En otras palabras, dan soporte a un espectro más amplio de tareas organizacionales que los sistemas de procesamiento de transacciones, incluyendo el análisis de decisiones y la toma de decisiones.

- **Sistemas de apoyo a la toma de decisiones**

Los sistemas de apoyo a la toma de decisiones (DSS, Decision Support Systems) son sistemas de información interactivos que ayudan al tomador de decisiones a utilizar datos y modelos para resolver problemas. Estos sistemas se ajustan más al gusto de la organización que los utiliza que a los sistemas de información gerencial tradicionales. En ocasiones se hace referencia a ellos como sistemas que se enfocan en la inteligencia de negocio.

- **Sistemas de soporte a ejecutivos**

Los sistemas de soporte a ejecutivos (ESS, Executive Support Systems) se encuentran en el nivel estratégico de la administración, ayudan a los ejecutivos a organizar sus interacciones

con el ambiente externo proporcionando datos resumidos, indicadores, gráficos, entre otros. Estos se apoyan en la información generada por los TPS y los MIS.

Además de la clasificación de sistemas de información, también se cuenta con dos tipos de procesamientos, el procesamiento de transacciones y el procesamiento de analíticos en línea. A continuación se estudiará cada uno de ellos y sus diferencias.

2.5.3. Clasificación de sistemas de información según el procesamiento de datos

- **Procesamiento de Transacciones en Línea (OLTP)**

OLTP representa toda aquella información transaccional que genera la organización en su accionar diario, además de las fuentes externas con las que puedan disponer. Entre los OLTP más habituales, que pueden existir en cualquier organización, se encuentran: archivos de texto, hojas de cálculos y base de datos transaccionales entre otros.

Según (Sinnexus, 2007) los sistemas OLTP se definen como bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado con un commit, o invalidado con un rollback) y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales.

Los sistemas OLTP se caracterizan por:

- El acceso está optimizado para tareas frecuentes de lectura y escritura. Por ejemplo, la enorme cantidad de transacciones que tienen que soportar las BD de bancos o supermercados diariamente.
- Los datos se estructuran según el nivel de aplicación (según sus siglas en inglés, ERP (*Enterprise Resource Planning*) o CRM (según sus siglas en inglés*, Costumer Relationship Management*) implantado, sistema de información departamental, entre otros).
- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos (es común la falta de compatibilidad y la existencia de islas de datos).

- El historial de datos suele limitarse a los datos actuales o recientes.

- **Procesamiento Analítico en Línea (OLAP)**

Según (Sinnexus, 2012) los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos, entre otros. Este sistema es típico de los Datamarts.

Los sistemas OLAP se caracterizan por:

- El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo, normalmente de dos a cinco años.
- Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga denominado ETL.

Una vez que ya hemos conocido la definición de un sistema de información, los tipos de sistema de información que existen y a su vez a que se refiere el procesamiento OLAP y OLTP. Es importante conocer una pieza fundamental dentro de los sistemas de información, la cual es, la Inteligencia de Negocios.

2.6. Inteligencia de Negocio

Conocida en inglés como *Business Intelligence* (BI), es la habilidad para transformar los datos en información, y esta información ser transformada en conocimiento para optimizar el proceso de toma de decisiones en las organizaciones. (Loshin, 2012)

- **Características de una Solución de Inteligencia de Negocio**

Según (Cano, 2007), son soluciones que deben tener las siguientes características:

- Visión unificada de los datos.
- Permitir la creación personalizada de informes y consultas.
- Proveer de vistas gráficas e interactivas para la presentación de información con funciones analíticas.
- Tener capacidad de procesamiento de grandes volúmenes de datos.

2.6.1.1. Arquitectura de una Solución de Inteligencia de Negocio

En la Figura 6, se observan los que para Cano (2007) son los componentes que conforman una solución de inteligencia de negocio:

Figura 6: Arquitectura de una Solución de Inteligencia de Negocio

Fuente: (Cano, Business Intelligence: Competir con Información, 2007)

A continuación se profundiza en cada uno de estos componentes:

- **Fuentes de Datos**

Son las que alimentan de información al repositorio de datos. Pueden ser internas cuando pertenecen totalmente a la institución o externas cuando se trata de datos comprados a terceros.

- **Procesos de Extracción, Transformación y Carga**

Comúnmente conocidos como ETL por sus siglas en inglés (Extract, Transformation and Load). Debido a que se pueden tener fuentes de datos heterogéneas, estos procesos son los encargados de extraer, limpiar y realizar las transformaciones necesarias a los datos para cargarlos en un repositorio. Los procesos de ETL están compuestos por tres subprocesos:

- **Proceso de Extracción:** consiste en obtener los datos que se encuentran localizados en diferentes fuentes de datos.
- **Proceso de Transformación:** los datos procedentes de las distintas fuentes pueden referenciar la misma información, pero puede existir inconsistencia de formato o nombramiento. Estos escenarios requieren que se realicen transformaciones regidas por reglas de negocio que tienen como objetivo evitar inconsistencias en los datos para integrarlos y posteriormente cargados.
- **Proceso de Carga:** en este punto se cargan los datos transformados al almacén de datos, considerando los escenarios en los que se elimina la información anteriormente almacenada en el repositorio o si se mantiene un historial.

- **Área de Integración de Datos**

Para (Cano, 2007), es todo componente que actúe como puente entre los sistemas origen y destino, cumpliendo algunas funciones:

- Integrar las diversas fuentes de datos en un solo repositorio, ya que, normalmente la información que se tiene en los sistemas transaccionales no está preparada para la toma de decisiones.
- Al modelar un proceso de negocio en que no se necesite toda la data de la institución, se construye un repositorio que contenga el subconjunto relevante al proceso.
- Ser un espacio temporal y volátil, sobre el que se ejecutarán los procesos de ETL.

- Se usa para hacer una extracción rápida de las fuentes de datos y almacenarlos temporalmente mientras es llevado a cabo el proceso de optimización, para luego cargarlos al almacén de datos.

- **Almacén de Datos**

Es una colección de datos orientada a un determinado ámbito y que tiene como características ser integrada, no volátil y variable en el tiempo. (Imhoff & Galemmo, 2003).

Son la base de los sistemas de apoyo a la toma de decisiones. Surgen a partir de la necesidad de simplificar los inconvenientes asociados al acceso a la información, y en consecuencia lograr la optimización de los procesos de análisis y diseño de estrategias por parte de los directivos o analistas de la institución (Inmon W. H., 1996).

- **Características de un Almacén de Datos**

En la Figura 7 se observan las características de un almacén de datos:

Figura 7: Características de un Almacén de Datos

Fuente: (Inmon, 1996)

- **Orientado a Temas:** está orientado hacia los procesos de negocio básicos de la institución y las entidades que en ellos intervienen. Centrándose en obtener únicamente los datos asociados a estos procesos a partir del entorno operacional.
- **Integrado:** las diferentes fuentes de datos se encuentran integradas y almacenadas en un mismo repositorio, por lo que la inconsistencia existente en los sistemas operacionales debe ser eliminada.
- **Variable en el Tiempo:** los datos almacenados se enfocan en un periodo de estudio, y son extraídos y archivados desde los sistemas operacionales, lo que los hace históricos. Al modificarse este periodo, los datos anteriores se mantienen, con el fin de hacer comparaciones y generar conocimiento.

- **No Volátil:** la información almacenada es permanente, es decir, no modificable. A diferencia de un sistema operacional, donde se realizan tareas de lectura, inserción y modificación de forma regular, en un almacén de datos solo se realiza una operación de carga que inserta múltiples datos y posteriormente se realizan operaciones de lectura, con la finalidad de realizar análisis y estudios.

- **Objetivos de un Almacén de Datos**

Los principales objetivos asociados a un almacén de datos son los siguientes:

- Reunir y consolidar las bases de datos que se mantienen en los diferentes departamentos o áreas funcionales de una institución.
- Dar soporte a las necesidades cambiantes que se presentan en el negocio, y así planear mejor las conductas y actividades a realizar.
- Mejorar la productividad de las organizaciones a partir de los estudios realizados con la información almacenada en el almacén de datos.
- Asegurar la calidad y eficiencia de las decisiones tomadas en las organizaciones.
- Permitir un acceso fácil y flexible a la información. Para ello se hace uso de un modelo dimensional en el que se adopta un esquema de representación.

- **Modelo Dimensional**

Según Kimball & Ross (2002), es una método para el diseño lógico de un almacén de datos. Se basa en tener los datos organizados en torno a hechos que son descritos con precisión en mayor o menor nivel de detalle por un conjunto de tablas dimensión. Los conceptos básicos asociados a un modelo dimensional son presentados a continuación:

- **Dimensión:** representa las diferentes formas de visualizar la información que se encuentra asociada a un hecho de acuerdo al nivel de detalle utilizado. Se representa físicamente con una Tabla Dimensión (Ver Figura 8) que se encuentra compuesta por una clave primaria y un conjunto de atributos descriptivos que permiten dar sentido a lo almacenado. Algunos tipos de tabla dimensión son: dimensión conformada, dimensión degenerada, dimensión role-playing, entre otros.

Figura 8: Tabla Dimensión

Fuente: (Kimball & Ross, The Data Warehouse Toolkit, 2002)

- **Jerarquía:** Es una relación en cascada de uno a muchos y está asociada a la ubicación de un atributo con respecto a otro. Puede representarse como un árbol (Ver Figura 9) en el que la raíz es nivel mayor que aporta menos detalle y las hojas son el menor nivel y aportan más detalle. (Kimball & Ross, 2002)

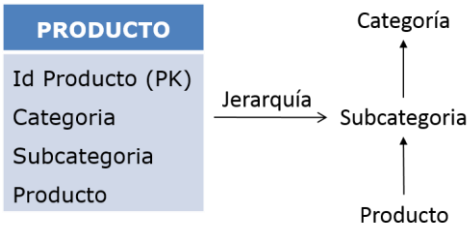


Figura 9: Jerarquía

Fuente: (Kimball & Ross, The Data Warehouse Toolkit, 2002)

- **Granularidad:** Se refiere al nivel de detalle de los datos dentro del almacén de datos. A mayor nivel de granularidad se tiene menos detalle de los datos y a menor nivel de granularidad se tiene mayor detalle. (Ver Figura 10)

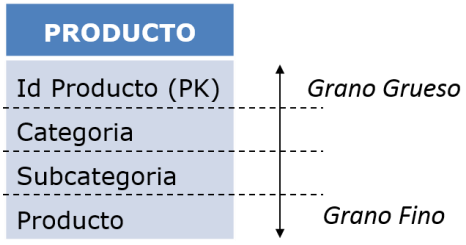


Figura 10: Granularidad

Fuente: (Kimball & Ross, The Data Warehouse Toolkit, 2002)

- **Hecho:** Es una medición del negocio distinta a un atributo, tiene carácter dinámico con el fin de realizar estudios. Físicamente se representan dentro de una Tabla de Hechos

(Ver Figura 11), que es el centro del modelo y es una relación multiclave en la que cada una de las claves referenciadas está relacionada con una dimensión y la unión de las mismas compone la clave primaria de la tabla. Algunos tipos de tablas de hechos son: transaccional, de foto acumulada, de foto periódica, entre otros.

Figura 11: Tabla de Hechos

Fuente: (Kimball & Ross, The Data Warehouse Toolkit, 2002)

- **Esquema:** Según Kimball & Ross (2002), es la representación genérica de un modelo multidimensional en una base de datos relacional, donde una tabla de hechos está unida a varias dimensiones. Existen diferentes tipos de esquemas: la estrella en que las dimensiones están des-normalizadas, el copo de nieve en que las dimensiones están en tercera forma normal y el esquema constelación que consiste en la unión de esquemas estrella y/o copo de nieve con común. (Ver Figura 12)

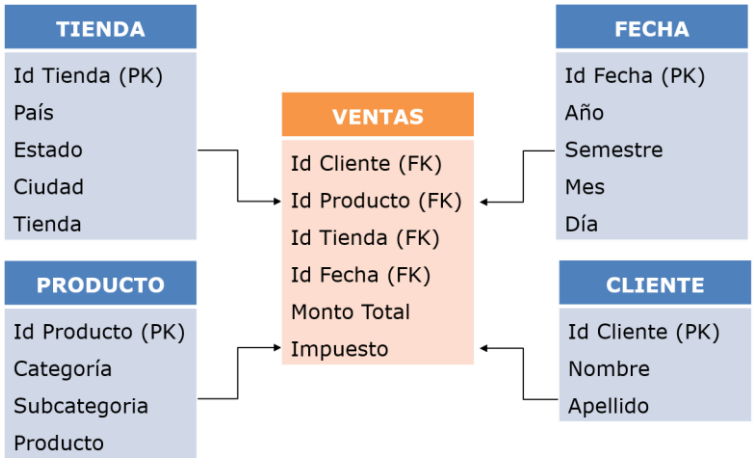


Figura 12: Esquema Estrella

Fuente: (Kimball & Ross, The Data Warehouse Toolkit, 2002)

- **Herramientas Analíticas**

Se trata de herramientas que tienen el objetivo de asistir en el análisis y la presentación de los datos. Según Cano (2007), las herramientas para el desarrollo de soluciones analíticas deben tener los siguientes componentes:

- **Herramientas de usuario final de consultas e informes:** empleadas por usuarios finales para crear informes para ellos mismos o para otros, no deben requerir programación y debe permitir personalización
- **Herramientas de construcción de cuadros de mando:** permiten a los usuarios finales ver información crítica para el rendimiento.
- **Generadores de informes:** utilizados por desarrolladores profesionales para crear informes estándar para grupos, departamentos o la institución.
- **Herramientas OLAP:** permiten a los usuarios finales tratar la información de forma multidimensional para explorarla desde distintas perspectivas y periodos de tiempo.

Opcionalmente también pueden tener los siguientes componentes:

- **Herramientas de planificación, modelización y consolidación:** permiten a los analistas y a los usuarios finales crear planes de negocio y simulaciones.
- **Herramientas de Datamining:** permiten a estadísticos o analistas de negocio crear modelos estadísticos de las actividades de los negocios.

En la **Tabla 1** se listan algunas herramientas de uso comercial y de código abierto.

Tabla 1: Herramientas de Inteligencia de Negocios

Uso Comercial	Código Abierto
– Oracle Data Visualization	– Pentaho
– Business Objects (SAP)	– SpagoBI
– MicroStrategy	– Jedox

A continuación se detallan las herramientas de Oracle Data Visualization y Pentaho, seleccionadas para el desarrollo del proceso de análisis realizado en este trabajo de grado.

- ***Oracle Data Visualization***

Según (Olmo, 2016) Oracle Data Visualization es una herramienta de análisis de datos para las organizaciones, la cual permite trabajar en entornos híbridos (hardware o software propios o en la nube) para satisfacer tanto necesidades analíticas específicas de los

departamentos técnicos y corporativos, como las de flexibilidad y usabilidad que demandan el resto de las áreas de las organizaciones.

Esta herramienta de oracle no dispone de un cuadro de mando predefinido, permitiéndole así al usuario incorporar y mezclar archivos, documentos o cualquier información que quiera analizar procedentes de diferentes fuentes de datos.

Según (Expansión.com, 2016) el cual es un portal de noticias nos indica que Oracle Data Visualization fue construida con los siguientes principios:

- Visualizaciones más inteligentes: permite visualizar los datos con solo arrastrar y soltar atributos en la pantalla. En función de la naturaleza de estos, Data Visualization muestra la visualización recomendable de forma automática.
- Visualizaciones conectadas: los datos y las visualizaciones están conectadas por defecto, de modo que, cuando se marca un dato en un gráfico, aparecen automáticamente resaltados en otros relacionados, mostrando correlaciones y patrones ocultos.
- Compatible con cualquier tipo de archivo: permite cargar ficheros y bases de datos, mezclarlos con los previamente cargados o con datos corporativos de forma sencilla e intuitiva.

A continuación se muestra en la figura 13 una interfaz de usuario de la herramienta Oracle Data Visualization.

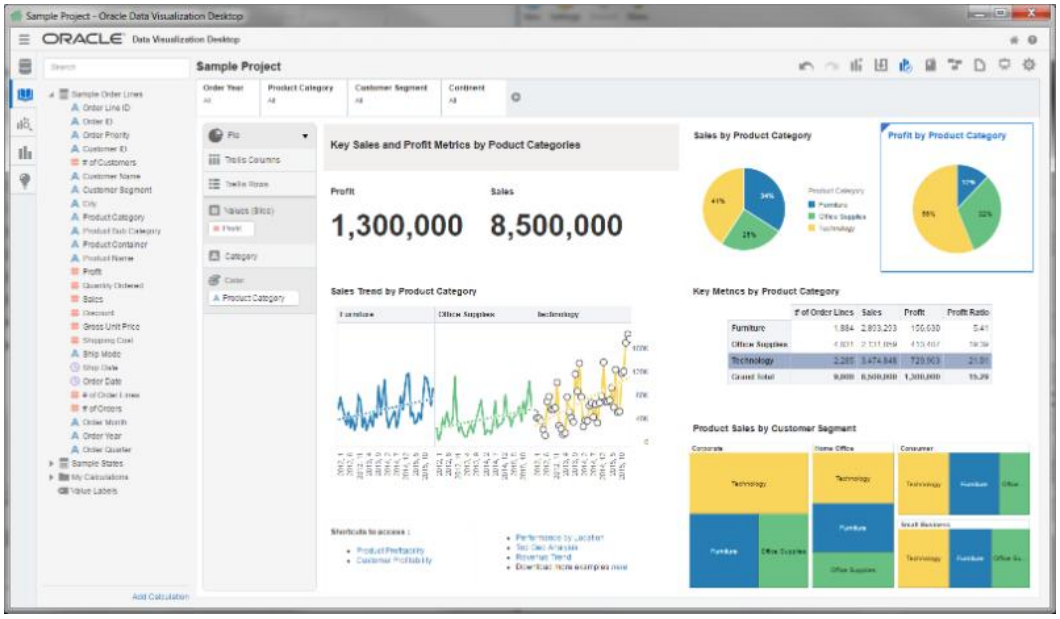


Figura 13: Interfaz gráfica de Oracle Data Visualization

- ***Pentaho***

Según Pentaho (2012), es una plataforma de orientada a soluciones, centrada en procesos e incluye todos los principales componentes requeridos, para la gestión y toma de decisiones organizacionales. Está compuesta por diferentes programas que satisfacen los requerimientos fundamentales de una solución de inteligencia de negocio, ofreciendo alternativas para la gestión y análisis de la información, incluyendo el análisis multidimensional OLAP, presentación de informes, creación de cuadros de mando, entre otros.

La plataforma de Pentaho consta de dos ediciones:

- **Pentaho Community Edition:** tiene como objetivo la contribución de nuevas funcionalidades, pruebas e innovación
- **Pentaho Enterprise Edition:** posee características adicionales que no se encuentran en la edición de la comunidad. Es un software certificado, listo para ambientes de producción, posee módulos exclusivos y facilidades de uso; además incluye soporte técnico oficial.

Ambas ediciones cuentan con herramientas que soportan el desarrollo de soluciones de inteligencia de negocios. A continuación se explican algunas de estas herramientas.

- ***Pentaho Data Integration (PDI)***

Herramienta que permite extraer, limpiar e integrar la información disponible en aplicaciones y bases de datos separadas y ponerla en manos del usuario, proyectando consistencia. También es conocido como Kettle y posee las siguientes aplicaciones:

- **Spoon:** herramienta grafica que permite diseñar procesos ETL. Esta herramienta soporta conexión con diversas fuentes de datos y permite transformar los datos necesarios para cargarlos dentro de la base de datos destino. En la Figura 14 se observa la interfaz de esta herramienta.
- **Pan:** ejecuta transformaciones diseñadas en el Spoon.
- **Chef:** herramienta para ejecutar trabajos complejos, que automatizan los procesos de actualización de la base datos.
- **Kitchen:** herramienta que ayuda a ejecutar los trabajos por lotes, permitiendo iniciar y controlar fácilmente el proceso ETL.
- **Carte:** servidor web que permite la supervisión remota del proceso ETL.

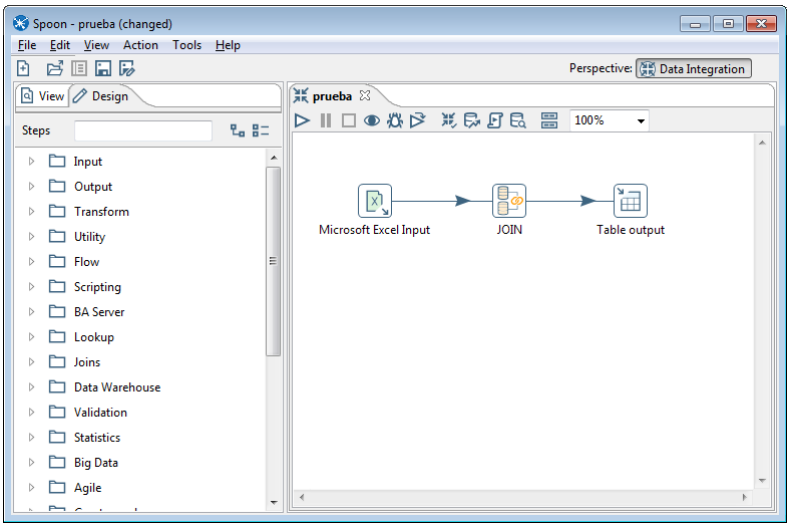


Figura 14: Interfaz gráfica de Spoon

Para el diseño y construcción de procesos ETL, la herramienta Pentaho Data Integration posee una serie de elementos clave:

- **Transformación (Transformation):** herramienta que permite realizar una variedad de tareas con los datos, moviéndolos de un lugar a otro. Su rol principal es extraer datos de diversas fuentes de datos, transformarlos de una representación a otra y cargarlos en fuentes de salida. Consiste de un número separado de acciones llamadas steps relacionadas a través de un flujo llamado hop.
- **Pasos (Steps):** Son las acciones que conforman a una transformación, los cuales son diseñados con una función específica.
- **Flujos (Hops):** representación gráfica del flujo de datos entre los steps que conforman una transformación o un job.
- **Trabajo (Job):** componente con el que se define una secuencia de actividades que brinda un orden de ejecución, por lo tanto son usados para crear un control de flujo. Generalmente está conformado por una serie de transformaciones que se desean ejecutar en un orden específico.

➢ ***Pentaho Analysis Services***

Es un servidor que permite realiza procesamiento analítico en línea (OLAP). Soporta el lenguaje de consulta MDX y lenguaje XML para análisis y especificaciones. A través del uso de las tablas dinámicas generadas, el usuario puede navegar por los datos ajustando la visión de los mismos modificando los filtros y añadiendo o quitando campos de agregación. Pentaho Analysis está compuesto por un servidor y las siguientes herramientas cliente:

- **Schema Workbench:** herramienta que permite llevar a cabo la construcción de esquemas multidimensionales. Además, permite al usuario publicar el esquema generado en el servidor de BI, para que sea utilizado en el desarrollo de análisis y reportes. En la Figura 15 se observa la interfaz de esta herramienta.
- **Aggregation Designer:** herramienta que simplifica la creación y despliegue de tablas de agregación que mejoran el rendimiento de los cubos OLAP.
- **Administration Console:** herramienta cliente, donde se pueden crear nuevas vistas de análisis y reportes.

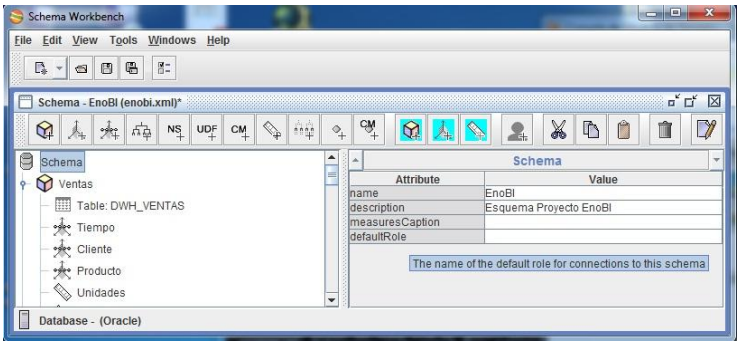


Figura 15: Pentaho Schema Workbench

➢ ***Pentaho Reporting***

Aplicación utilizada para presentar el contenido de BI de alto nivel a los usuarios finales, tal como se observa en la Figura 16. Este contenido es casi siempre de carácter gráfico, y proporciona cierto grado de interactividad que permite al usuario navegar por contenidos más detallados y así medir, monitorear y gestionar el rendimiento más efectivamente.

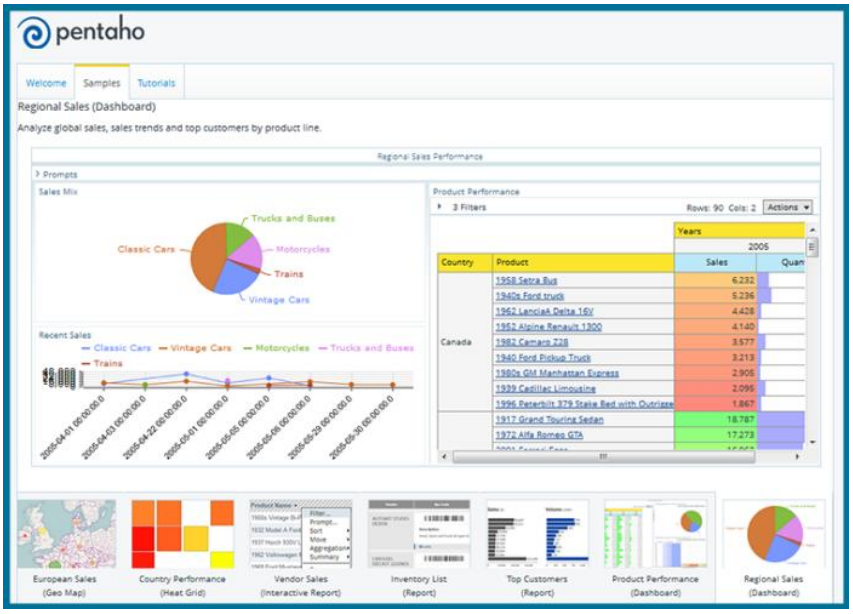


Figura 16: Pentaho Dashboards

Finalmente, luego de conocer la definición, características, componentes y herramientas tecnológicas de una solución de inteligencia de negocios es importante destacar una tendencia tecnológica que actualmente ha estado ayudando a diversos sectores organizacionales a manipular una gran cantidad de datos, los cuales generan mucha información, permitiendo analizarla rápidamente y así poder ayudar de manera oportuna en la toma de decisiones. A continuación se estudiará esta tendencia tecnológica, la cual es nombrada como Big Data

2.7. Big Data

En nuestros días, es un hecho incuestionable la gran cantidad de información que se genera cada segundo en nuestro planeta. Dicha información puede ser de diferentes tipos de datos estructurados, semiestructurados o no estructurados. El análisis inteligente (y la mayoría de las veces en tiempo real) de estos tipos de datos puede ser un requisito innegable para las organizaciones. Como consecuencia de ello han surgido en los últimos años términos como el de big data. Según (Dutcher, 2014) no existe una definición estándar de big data pero se le han propuesto 43 definiciones al respecto. Mientras llega esa definición universal se comentan en esta investigación algunas de las más utilizadas. Según (IDC, 2012), Big Data es una nueva generación de tecnologías y arquitecturas diseñadas para extraer valor económico de grandes volúmenes de datos estructurados, semi-estructurados o no

estructurados, habilitando una captura, identificación y/o análisis a alta velocidad. Para entender mejor que significa big data, se explicará detalladamente las 5 V que la componen:

- **Volumen:** El volumen de datos nos puede indicar desde datos de identificación del cliente, el sexo, la edad, el origen geográfico, su dispositivo de navegación, qué páginas visita en nuestra web, el tiempo que permanece en ellas, con qué frecuencia visita nuestra página, los clicks que hace en ella, el histórico de su navegación, el idioma de navegación que usa. Los datos son ilimitados.

  Este gran volumen de información generada requiere una correcta orientación de las estrategias para filtrar la información necesaria y proporcionar así, también, un ahorro de tiempo.

- **Variedad:** La información digital es variada: texto, gráficos, imágenes, vídeos, presentaciones multimedia, juegos, entre otros, en diferentes formatos y tipos de archivos y provenientes, además, de variados dispositivos como smartphones, tabletas, portátiles, ordenadores de sobremesa.

  Cuanto más amplias sean las fuentes de información, más factible se presenta la posibilidad de expandir el mercado hacia nuevos usuarios y localizar nuevos segmentos de público potencial.

  Pero para que toda esta variedad sea útil, hay que poder estructurarla de la mejor forma posible, con herramientas que nos ayuden al registro, análisis e interpretación de todos estos datos que se obtienen en tiempo real.

- **Velocidad:** Una de las características intrínsecas del medio digital es que todo ocurre a mucha velocidad y tiene que ser capturado a la misma velocidad que es generado, e incluso en tiempo real: lo que hoy es trending topic en Twitter… mañana puede ya no serlo, un comentario negativo sobre una marca en Facebook, puede desencadenar un terremoto en cuestión de minutos.

- **Veracidad:** Es imprescindible poder medir y evaluar la veracidad y la autenticidad de los datos que se obtienen mediante Big Data, para así poderlo aprovechar al máximo. La veracidad de los datos obtenidos es clave para poder ofrecer a nuestros clientes y usuarios productos que realmente se adapten a sus gustos y que den respuestas a sus necesidades.

- **Valor:** Es imprescindible generar valor de negocio con los datos que nos proporciona el Big Data. Todas las empresas generan, trabajan y gestionan multitud de datos, pero la clave está en cómo obtener la mejor información, el mejor valor y conocimiento, para sacar la mayor rentabilidad posible.

A continuación se observará algunos tópicos que nos permitirán entender fácilmente todo lo relacionado a Big Data, comenzando por los tipos de datos descritos anteriormente.

2.7.1. Tipos de Datos

- **Datos Estructurados**

Según (Factor Humano Formación, 2014) son aquellos datos que tienen bien definido su longitud y su formato. Suelen ser fechas, números, cadenas de caracteres y están almacenados en tablas. En las organizaciones estos datos se encontrará en información obtenida a partir de CRM, ERP etc. Estos datos también suelen estar guardado en un almacén de datos si contienen mucha información y si la organización no genera tal cantidad de datos tendrán una base de datos relacional. Para consultar estos datos se realizan mediante consultas SQL.

- **Datos semiestructurados**

Según (Factor Humano Formación, 2014) los datos semiestructurados son una mezcla de los estructurados y no estructurados, es decir, estos datos siguen una especie de estructura implícita, pero no tan regular como para poder ser gestionada y automatizada como la información estructurada.

- **Datos no estructurados**

Según (Factor Humano Formación, 2014) es información que se encuentra en el formato tal y como fueron recolectados y que puede tener cualquier estructura. Se puede encontrar en formatos como: texto, imagen, video, comentarios en redes sociales o logs de aplicaciones en redes sociales son buenos ejemplos de datos no estructurados.

Para poder comprender los tipos de datos explicados anteriormente, a continuación en la Tabla 2, se muestran con ejemplos sencillos la aplicabilidad de dichos tipos de datos.

Tabla 2: Ejemplos Tipos de Datos

Datos Estructurados	Datos Semiestructurados	Datos No Estructurados
Fichas de clientes Fecha de nacimiento Nombre Dirección Transacciones en un mes Puntos de compra	**Correos Electrónicos** Parte estructurada: destinatario, receptores, tema Parte no estructurada: cuerpo del mensaje	**Persona a Persona** Comunicaciones a través de las redes sociales **Persona a Máquina** Dispositivos médicos Comercio electrónico Ordenadores Móviles **Máquina a Máquina** Sensores, dispositivos GPS, cámaras de seguridad

Una vez comprendido los tipos de datos como lo son datos estructurados, semiestructurados y no estructurados. Se puede entender fácilmente las definiciones propuestas de big data. Como objetivo de estudio se cita una definición propuesta por Gartner la cual es una de las empresas más grande de consultoría e investigación de las tecnologías de información.

2.7.2. Definición de Big Data

Como se menciono anteriormente, actualmente no existe una definición concreta de que es big data, sin embargo una de las aproximaciones más completas de Big Data es la facilitada por (Gartner, 2012) la cual cita lo siguiente "*Son activos de información caracterizado por su*

alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones."

Como se puede observar existen diversas definiciones de Big Data todas con cierto parecido, pero que en conjunto puede producir confusión sobre el término. La definición más clara no tiene que hacer referencia a nuevas tecnologías, la definición de Big Data es la que se centra en el tratamiento y análisis de gran volumen de información, de diversos tipos de datos, donde lo importante es el procesamiento rápido y el valor que se obtiene luego de analizar tanta información.

Lo descrito anteriormente por Gartner permite decir que una de las características principales de un sistema Big Data es el de trabajar con los 3 tipos de datos explicados anteriormente, permitiendo aumentar la variedad de la información. De esta forma también se induce que el sistema podría almacenar y trabajar con un gran de volumen de datos. A continuación se estudiará cual es la arquitectura de Big Data.

2.7.3. Arquitectura Big Data

Según (Hurwitz, Nugent, & Halper, 2013) hemos pasado de una época en la cual una organización debía implementar una base de datos para satisfacer una necesidad de un proyecto específico, pero a medida que los datos iban creciendo y las fuentes de datos iban variando en el tiempo, convirtiéndose en el combustible del crecimiento y la innovación, se necesitaba contar con una arquitectura subyacente para soportar los requerimientos de crecimiento.

Antes de profundizar en la arquitectura, es importante tener en cuenta las actividades funcionales en un ciclo de big data. En la Figura 17 se ilustra como primer paso la **captura** de los datos, y luego la **organización** e **integración**. Después que se lleva a cabo con éxito lo mencionado anteriormente, los datos pueden ser **analizados**. Por último, a través del análisis de los resultados se toman **acciones**, como por ejemplo el apoyo en la toma de decisiones de la organización para dar respuesta a una problemática.

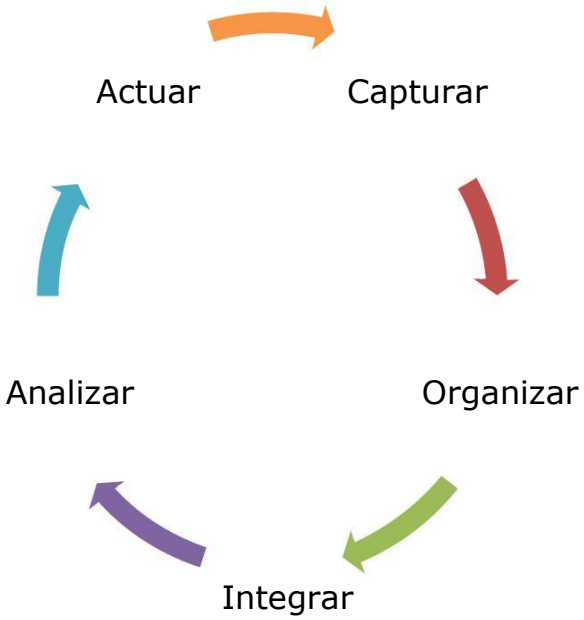


Figura 17: Ciclo Funcional de Big Data

Fuente: Big Data for Dummies. (Hurwitz, Nugent, & Halper, 2013)

Es un tema importante la validación de las fuentes de datos. Si las organizaciones combinan varias fuentes de datos, es fundamental tener la capacidad de validar que estas fuentes tienen sentido para resolver la problemática. Además, ciertas fuentes de datos pueden contener información sensible, por lo que se debe poner en práctica un nivel suficiente de seguridad. Por supuesto, antes de cualquier combinación de fuentes de datos primero que nada se debe conocer si ellas tratan de resolver la problemática que se está presentando en la organización, lo cual permitirá determinar qué tipos de datos se necesitarán y la arquitectura que se diseñará.

Adicional a esto, es importante apoyar el desempeño requerido, ya que atender la problemática planteada en la organización dependerá también de la velocidad con la que se realice el análisis de los datos. Existen análisis que se realizan en tiempo real y es inevitable almacenar gran cantidad de datos, por lo cual la arquitectura debe tener la capacidad de soportarlos, también debe manejar la redundancia para que esté protegido de la latencia imprevista que pueda ocurrir como el tiempo de inactividad.

Una arquitectura de big data robusta debe incluir una variedad de servicios que permitan a las organizaciones hacer uso de las fuentes de datos de una manera rápida y eficaz. Para ayudar a dar sentido a esto, se presenta en la Figura 18 los componentes de una arquitectura de big data, y la relación que existe entre ellos. Es importante señalar que para

la implementación de una arquitectura de big data pudiesen no tener todos los componentes que serán mencionados a continuación, esto dependerá de la aplicabilidad que se le vaya a dar a la solución.

Figura 18: Arquitectura de Big Data

Fuente: Big Data for Dummies. (Hurwitz, Nugent, & Halper, 2013)

- **Interfaces y Alimentación:** Antes de entrar en la pila tecnológica de big data, se puede observar que a ambos lados de la arquitectura existen indicaciones de interfaces que se alimentan desde o hacia datos gestionados internamente y fuentes de datos externas. Para entender cómo funciona big data en el mundo real, es importante empezar por comprender esta necesidad. De hecho, lo que hace que big data sea tan amplia es el hecho de que se basa en primera instancia, en recoger un montón de datos de muchas fuentes. Por lo tanto, las interfaces de programación de aplicaciones (API) son un elemento básico para cualquier arquitectura de big data. Además, considerando que existen las interfaces en todos los niveles y entre todas las capas de la pila para poder manejar esta gran cantidad de datos.

- **Infraestructura Física Redundante:** La infraestructura física es fundamental para el funcionamiento y la escalabilidad de una arquitectura de big data. Considerando también que en dichas arquitecturas para manejar grandes volúmenes de datos, la redundancia es importante porque se está tratando con gran cantidad datos de diversas fuentes. De hecho,

sin la disponibilidad de infraestructuras físicas robustas, big data probablemente no se habría convertido en una tendencia tan importante. Para soportar un volumen inesperado o impredecible de los datos, una infraestructura física para big data tiene que ser diferente a la de los datos tradicionales. La infraestructura física se basa en un modelo de computación distribuido. Esto significa que los datos pueden ser almacenados físicamente en muchos lugares diferentes y pueden ser unidos entre sí a través de redes.

- **Infraestructura de Seguridad:** El análisis de grandes volúmenes de datos es importante en una organización, pero más aún es asegurar todos esos datos. Por ejemplo, en una organización de salud, es probable que se desee utilizar aplicaciones de big data para determinar los cambios en las necesidades que presenta el paciente. Estos deben ser protegidos tanto para satisfacer los requisitos de cumplimiento de la organización como para proteger la privacidad de los pacientes. Por lo tanto se deben asignar políticas de seguridad para poder ver los datos y en qué circunstancias se les permite hacerlo. Siendo capaz de verificar la identidad de los usuarios, así como proteger la identidad de los pacientes. Este tipo de requisitos de seguridad es importante en una arquitectura de big data.

- **Fuentes de datos operacionales:** Cuando se piensa en grandes volúmenes de datos, es importante entender que hay que incorporar todas las fuentes de datos que le darán una visión completa a la problemática que presenta la organización. Esas fuentes de datos pueden ser de diversos tipos, como datos estructurados, datos no estructurados y datos semiestructurados.

- **Organización de los datos operacionales:** una cantidad creciente de datos provienen de una variedad de fuentes que no se encuentran organizadas o no son fáciles de agrupar y distribuir, incluidos los datos que provienen de máquinas o sensores, de las redes sociales o aplicaciones internas de la organización, entre otros. Es aquí donde se incluyen todas las herramientas capaces de ayudar a manipular, organizar y almacenar esa gran cantidad de datos adquiridas de fuentes operacionales para luego poder ser trabajados de manera más sencilla.

- **Analíticos tradicionales o avanzados:** para poder ayudar a solventar la problemática de la organización y crear indicadores que permitan darle sentido a los datos, se requiere de muchos enfoques diferentes dentro del análisis. Algunos análisis utilizarán un almacén de datos tradicional conectándose directamente a fuentes como data marts y bases de datos

analíticas, mientras que otros análisis aprovecharán herramientas para el análisis predictivos. Este componente es opcional en algunas arquitecturas de big data, todo dependerá del problema presentado para la cual big data es una solución.

- **Datawarehouses y Data Marts analíticos:** después de haber organizado todos los datos que se adquirieron de diversas fuentes, a menudo se selecciona un subconjunto de esos datos los cuales revelan algún tipo de patrón y se cargan en datawarehouses o data marts para que estén disponible para el negocio.

- **Visualización y Reportes:** Las organizaciones habitualmente se han basado en la creación de informes para dar sentido a los datos generados por sus aplicaciones internas, desde las cifras de ventas mensuales hasta las proyecciones de crecimiento. Con la aparición de nuevas fuentes y el incremento cada vez mayor de los datos, big data ha empezado a cambiar la forma en que se gestionan y se utilizan los datos, manejando una nueva generación de herramientas para ayudar a la toma de decisiones y a comprender verdaderamente el impacto no sólo de la colección de esta gran cantidad de datos, sino también cómo ellos ayudan a la solución del problema en la organización.

- **Aplicaciones de Big Data:** Algunas de las aplicaciones emergentes están en áreas como la salud, la gestión de fabricación, gestión de tráfico automovilístico, y así sucesivamente. Estas aéreas de aplicabilidad de big data se basan en grandes volúmenes, velocidades y variedades de datos para transformar el comportamiento de un mercado. En la asistencia sanitaria, una aplicación de big data podría ser capaz de monitorear a través de sensores a los bebés prematuros para determinar cuándo es necesaria una intervención quirúrgica. En la fabricación, una aplicación de big data puede ser utilizada para prevenir que una máquina se apague durante un ciclo de producción. Una aplicación de big data para el tráfico automovilístico puede ser utilizada para reducir el número de atascos de tráfico en las carreteras de la ciudad, disminuir los accidentes, ahorrar combustible y reducir la contaminación.

2.7.4. Herramientas Tecnológicas

En el mercado actualmente existen una gran cantidad de herramientas para el desarrollo de una solución de Big Data, a continuación listaremos algunas de las herramientas más usadas que componen una arquitectura de big data, según la arquitectura propuesta por (Hurwitz, Nugent, & Halper, 2013). Cada uno de los componentes de la arquitectura de big data que

se estará trabajando se muestran en las siguientes figuras resaltando en color amarillo el componente de la arquitectura en el cual nos enfocamos.

- **Fuentes de Datos Operacionales:**

Figura 19: Componente-Fuentes de Datos Operacionales

Fuente: (Hurwitz, Nugent, & Halper, 2013)

Algunas fuentes de datos operacionales del cual una arquitectura de big data puede adquirir datos para su posterior trabajo y análisis son las siguientes: sistemas transaccionales, ERP (por sus siglas en inglés, *Enterprise Resource Planning*), CRM (por sus siglas en inglés, *Customer relationship management*), SCM (por sus siglas en inglés *Supply Chain anagement*), redes sociales, correos electrónicos, señales de sensores, bolsas de empleo, puntos de venta, cajeros automáticos, páginas webs, blogs, telefonía móvil, logs, archivos XML, archivos JSON, comercios electrónicos, documentos de Word, Excel, PDF, imágenes y videos, entre otros.

- **Organización de los datos operacionales:**

Figura 20: Componente – Organización de los datos operacionales

Fuente: (Hurwitz, Nugent, & Halper, 2013)

Algunas de las herramientas tecnológicas que ayudan a manipular, organizar y almacenar la gran cantidad de datos adquiridas de las fuentes operacionales son las siguientes: Hadoop, MapReduce, HDFS, Apache Hive, Apache Flume, Pig, Sqoop, Cloudera, Hortonworks, Avro, Hbase, DataStax, Data Integrator, Cognos Decisionstream, MongoDB, Kettle, Oracle NoSQL, Apache Cassandra entre otras.

Por objetivo de estudio, se describen las siguientes herramientas tecnológicas sobre este componente de la arquitectura de big data:

- Map Reduce

Según (Mario, 2014) MapReduce es una herramienta que permite procesar cantidades masivas de datos en paralelo mediante un algoritmo que se ejecuta en los nodos de un clúster de una arquitectura distribuida. Fue inicialmente desarrollado por Google para reemplazar el algoritmo indexado de páginas web en su motor de búsqueda en 2004. Este algoritmo está compuesto por dos pasos, el primer paso se denomina Map() y consiste en el

"mapeo" de los datos efectuando operaciones de filtrado, ordenación y agregación con el fin de ir minorando el volumen de información. El segundo paso es Reduce () consiste en el procesado del resultado del paso anterior para generar un nuevo conjunto de datos resultados de la operación y que por lo general reduce el número de particiones.

A continuación se aplica en la figura 21 de manera de ejemplificada el proceso de MapReduce.

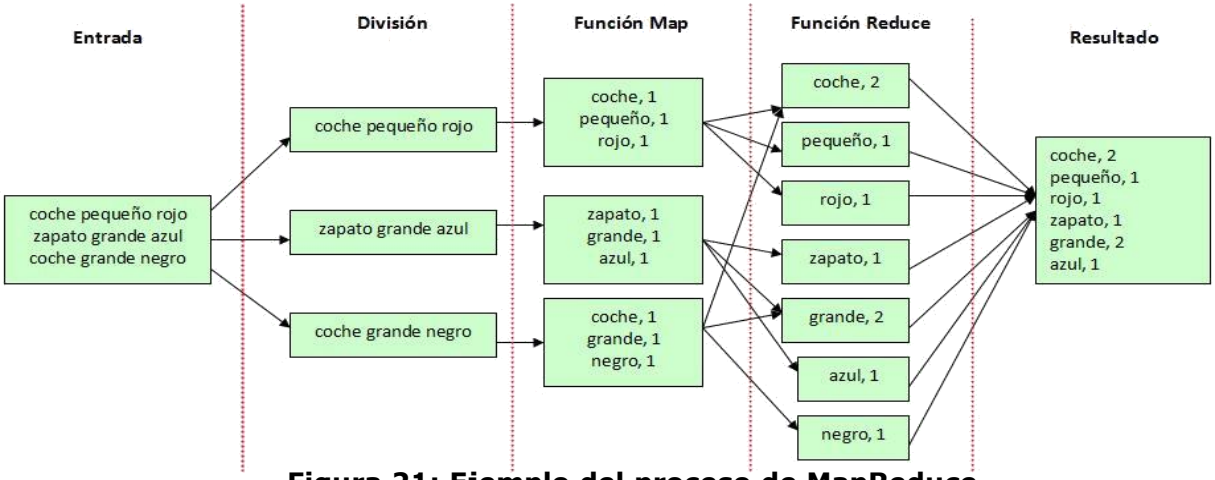


Figura 21: Ejemplo del proceso de MapReduce
Fuente: Qué se esconde detrás de "big data". (Montero, 2015)

- Hadoop

Apache Hadoop es una solución de software libre diseñada para el tratamiento de hasta exabytes de datos distribuidos en múltiples nodos. Hadoop se ha convertido en un estándar sobre el que se desarrollan herramientas comerciales por compañías tradicionales. La solución de Hadoop se basa en un desarrollo de Google del año 2009 denominado MapReduce, el cual fue explicado anteriormente. Cabe destacar que Hadoop no es un programa en sí, es decir, no se puede descargar un programa denominado Hadoop directamente, ya que Hadoop es un ecosistema de productos bajo el paraguas de Apache la cual es una organización no lucrativa creada para dar soporte a los proyectos de software bajo la denominación Apache, incluyendo el servidor HTTP Apache. Hadoop se caracteriza por ser económico, escalable o adaptable, eficiente a la hora de realizar trabajos en forma paralela y confiable ya que mantiene automáticamente copias de los datos en nodos para la prevención de fallos.

- Hadoop Distributed File System (HDFS)

Es un sistema de archivos que trata de recopilar toda la información posible. Se puede definir como un sistema de archivos distribuidos, escalable y portátil escrito en el lenguaje Java para el framework de Hadoop. Según (Hadoop, 2013) es un sistema de archivos distribuido diseñado para ejecutarse en hardware. Posee muchas similitudes con sistemas de archivos distribuidos existentes. HDFS es altamente tolerante a fallos y está diseñado para ser implementado en hardware de bajo costo, proporcionando alto rendimiento de acceso a datos de la aplicación y es adecuado para aplicaciones que tienen grandes volúmenes de datos. Fue construido como infraestructura para un proyecto de motor de búsqueda también construido por Apache. HDFS está diseñada para el procesamiento por lotes en lugar de la utilización interactiva de los usuarios. HDFS ha sido diseñado para ser fácil de transportar de una plataforma a otra. Esto facilita la adopción generalizada de HDFS como plataforma de almacenamiento para un gran conjunto de aplicaciones.

- **Datawarehouses y Data Marts analíticos:**

Figura 22: Componente- Datawarehouses y Data Marts Analíticos
Fuente: (Hurwitz, Nugent, & Halper, 2013)

Algunas de las herramientas tecnológicas que funcionan como repositorio para cargar la data previamente trabajada que mediante su organización revelan algún tipo de patrón y puedan estar disponibles para todo el negocio son las siguientes: Oracle Database, PostgreSQL, MySQL, DB2, Interbase, Sybase, Microsoft SQL Server, Apache Hive entre otros.

Por objetivo de estudio, se describen las siguientes herramientas tecnológicas relacionadas al componente de datawarehouses y data marts de la arquitectura de big data:

- Apache Hive

Según (Hive Software Foundation, 2014) Apache Hive es una infraestructura de almacenamiento de datos construida sobre Hadoop para proporcionar agrupación, consulta, y análisis de datos. Apache Hive soporta el análisis de grandes conjuntos de datos almacenados bajo HDFS de Hadoop y en sistemas compatibles como el sistema de archivos Amazon S3. Ofrece un lenguaje de consultas basado en SQL llamado HiveQL5 con esquemas para leer y convertir consultas de forma transparente en MapReduce.

- Oracle Database

Oracle Database es un sistema de gestión de bases de datos relacionales (RDBMS). Originalmente desarrollado 1997 por Lawrence Ellison y otros desarrolladores. Es uno de los motores de bases de datos relacionales más confiables y ampliamente usados. Está construido en torno a un marco de bases de datos relacional en la que los objetos de datos se pueden acceder directamente por los usuarios a través del lenguaje de consulta estructurado (SQL). Es totalmente escalable, estable y con soporte multiplataforma. Trabaja bajo el paradigma cliente/servidor para la gestión de bases de datos. Para el desarrollo de consultas en bases de datos Oracle se suele utilizar el lenguaje SQL o el lenguaje PL/SQL, el cual es un lenguaje bastante potente para tratar y gestionar la base de datos. Este SMBD se puede instalar en Windows, Linux y sistemas operativos basados en Unix.

- PostgreSQL

Según (Martinez, 2010) es un sistema manejador de bases de datos objeto-relacional, distribuido bajo licencia BSD y desarrollado bajo la filosofía de código abierto. Es el sistema manejador de bases de datos (SMBD) de código abierto más potente del mercado y, en sus

últimas versiones, no tiene nada que envidiarle a otras bases de datos comerciales. Utiliza un modelo cliente/servidor y usa multiprocesos, en vez de multihilos, para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto, y el sistema continuará funcionando. PostgreSQL funciona muy bien con grandes volúmenes de datos y una alta concurrencia de usuarios en el sistema.

- MySQL

Según (Casillas & Perez) es un sistema manejador de base de datos relacional, multiplataforma, muy conocido, y ampliamente usado por su simplicidad, estabilidad, notable rendimiento y su libre distribución en internet bajo licencia GNU/GPL. Aunque carece de algunas características avanzadas, disponibles en otros SMBD del mercado, es una opción atractiva tanto para aplicaciones comerciales, como de entretenimiento, precisamente por su facilidad de uso y capacidad para un rápido despliegue.

Analíticos tradicionales y avanzados:

Figura 23: Componente- Analíticos tradicionales o avanzados

Fuente: (Hurwitz, Nugent, & Halper, 2013)

Algunas de las herramientas tecnológicas que permiten tener diversos enfoques analíticos de los datos almacenados son las siguientes: Open Source R, IBM SPSS Statistics, StatSoft Statistica, Matlab, SAS Enterprice Miner, Rapid Miner/Analytics Comercial Edition, Python,Weka, Rapid Miner/Analytics Free Edition, entre otras.

A continuación se describe una de las mejores herramientas de analítica y estadística según (Data Mining Consulting SAC, 2013) relacionada al componente de analíticos tradicionales y avanzados de la arquitectura de big data:

- Open Source R

Según (The R Foundation, 2013) R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas.

Al igual que S, se trata de un lenguaje de programación, lo que permite que los usuarios lo extiendan definiendo sus propias funciones. De hecho, gran parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionalmente exigentes es posible desarrollar bibliotecas en C o C++ que se cargan dinámicamente. Los usuarios más avanzados pueden también manipular los objetos de R directamente desde código desarrollado en C. R también puede extenderse a través de paquetes desarrollados por su comunidad de usuarios.

R hereda de S su orientación a objetos. Además de integrarse con distintas bases de datos, existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python. Otra de las características de R es su capacidad gráfica, que permite generar gráficos con alta calidad. R posee su propio formato para la documentación basado en LaTeX.

R también puede usarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas específicas tales como GNU Octave y su equivalente comercial, MATLAB. Se ha desarrollado una interfaz, RWeka para interactuar con Weka que permite leer y escribir ficheros en el formato *arff* y enriquecer R con los algoritmos de minería de datos de dicha plataforma.

- **Visualización y Reportes:**

Figura 24: Componente Visualización y Reportes

Fuente: (Hurwitz, Nugent, & Halper, 2013)

Algunas de las herramientas tecnológicas que ayudan a la visualización de los datos previamente trabajados en cada uno de los componentes de la arquitectura de big data son las siguientes: Pentaho, Palo BI Suite, Oracle Business Inteligence, Oracle Data Visualization, Tableau, Qlik View, Gephi, Many Eyes, Quadrigram, Nodebox, JasperSoft, Microstrategy, HTML5, D3, NodeJS, RESTful Api, Java, Google Fusion Table, entre otras.

Por objetivo de estudio, se describen las siguientes herramientas tecnológicas relacionadas al componente de visualización y reportes de la arquitectura de big data:

- Pentaho y Oracle Data Visualization

Las cuales fueron explicadas anteriormente

- NodeJS

Node es un intérprete Javascript del lado del servidor de código abierto que cambia la noción de cómo debería trabajar un servidor. Su meta es permitir a un programador construir aplicaciones altamente escalables y escribir código que maneje decenas de miles de conexiones simultáneas en una sólo una máquina física. Node.js puede ser combinado con una base de datos documental (por ejemplo, MongoDB o CouchDB) y JSON lo que permite desarrollar en un ambiente de desarrollo JavaScript unificado. Con la adaptación de los patrones para desarrollo del lado del servidor tales como MVC y sus variantes MVP, MVVM, etc. Node.js facilita la reutilización de código del mismo modelo de interfaz entre el lado del cliente y el lado del servidor.

- HTML5

HTML5 (*HyperText Markup Language*, versión 5) es la quinta revisión importante del lenguaje básico de la W*orld Wide Web*, HTML. HTML5 especifica dos variantes de sintaxis para HTML: un HTML clásico (text/html), la variante conocida como HTML5 y una variante XHTML conocida como sintaxis XHTML5 que deberá ser servida como XML. Esta es la primera vez que HTML y XHTML se han desarrollado en paralelo. Todavía se encuentra en modo experimental, lo cual indica la misma W3C, aunque ya es usado por múltiples desarrolladores web por sus avances, mejoras y ventajas. Al no ser reconocido en viejas versiones de navegadores por sus nuevas etiquetas, se recomienda al usuario común actualizar a la versión más nueva, para poder disfrutar de todo el potencial que provee HTML5. Según (w3schools, 2014) el desarrollo de este lenguaje de marcado es regulado por el Consorcio W3C.

- **Aplicaciones de Big Data:**

Figura 25: Componente- Aplicaciones de Big Data

Fuente: (Hurwitz, Nugent, & Halper, 2013)

El uso de herramientas tecnológicas en este componente de la arquitectura, dependerá del sector o departamento dentro de una organización que haga uso de la solución de big data para solventar algún problema en particular. Cuando se habla de sector se está hablando del sector bancario, sector de la salud, sector de telecomunicaciones, entre otros y cuando se habla de departamentos dentro de la organización pudiesen ser el departamento de mercadeo, de gestión de talento humano, de ventas, entre otros.

2.4.6. Importancia de Big Data

Según (IDC, 2012) los mercados y las organizaciones están viviendo una transformación de base tecnológica y social cuya principal derivada es el crecimiento exponencial de datos tanto dentro como fuera de los sistemas organizacionales. De hecho, según IDC el volumen de los datos digitales alcanzara 35 Zbytes, 44 veces más que en el 2009. Este crecimiento se caracteriza principalmente por datos no estructurados. Big Data supone un proceso de

cambio en la organización, no solo desde la perspectiva tecnológica sino principalmente desde la de negocio.

La generación de grandes volúmenes de datos de diversas fuentes integrados con la data almacenada de la organización actualmente, habilita no solo una mayor comprensión de los negocios, sino también proporciona la capacidad de crecer en cuanto a los servicios que se ofrecen hoy en día.

De hecho, Big Data representa una oportunidad importante para que desde los departamentos de tecnología se impacte el negocio de la organización de forma realmente significativa. Los altos cargos de tecnología de las organizaciones deben liderar la adopción de un conjunto de tecnologías como Big Data y Analíticos para extraer valor de los datos generados por fuentes externas. Big Data habilita la extracción de valor para las organizaciones a partir de grandes volúmenes de datos con una alta variabilidad mediante la adquisición y análisis de datos a una alta velocidad.

2.4.7. Beneficios e Inconvenientes de usar Big Data

A continuación se citan los beneficios e inconvenientes más relevantes que han sido extraídos de un artículo publicado por Eureka-Startups (Vauzza, 2013).

Beneficios:

- Análisis de Redes Sociales: determinar los círculos sociales de los clientes a partir de interacciones telefónicas y redes sociales online genera una visión completa de los clientes, identificando el papel que desempeñan en sus círculos y su grado de influencia.
- Marketing Viral: detecta clientes más influyentes, para maximizar la difusión de productos y servicios (conocer mejor al cliente y al mercado en las redes sociales).
- Anticipación de problemas: un sistema predictivo de análisis y cruce de información nos permite poder anticiparnos a posibles problemas que pueden surgir en el futuro, como por ejemplo una predicción de riesgo de catástrofes.
- Análisis de Seguridad: analítica proactiva que permite la reducción de riesgos y pérdida frente a fraudes.

- Permite detectar patrones complejos de fraude en tiempo real analizando los datos históricos, el patrón de uso de información de geolocalización, análisis de transacciones y operaciones sospechosas.
- Una analítica avanzada que analice todos los informes y datos de diversidad de fuentes, ayudando a la toma de decisiones, reduciendo los riesgos y descubriendo información que antes podría estar oculta o no parecía relevante.
- Reducción de tiempos de procesamiento.

Inconvenientes:

No obstante no hay que olvidarse de los inconvenientes de Big Data. Siendo el principal de ellos el proceso de adopción de Big Data: software y hardware necesario y costo. Pero además existen otros de menor peso como:

- Preparación a personal especializado en Big Data para formar al Científico de Datos genera un gasto en formación.
- Filtrado de la información adquirida de fuentes externas (no todos los datos son información relevante).
- Costos de implementación muy altos.

A parte de estos, hay que considerar un gran inconveniente antes de realizar un proyecto de Big Data y que es darse cuenta si es realmente útil para la organización hacer una inversión en una implementación de Big Data, a pesar de todos los beneficios que me pueden proporcionar la implementación de la misma.

CAPÍTULO 3
MARCO METODOLÓGICO

En este capítulo se describe el método seleccionado para llevar a cabo el desarrollo de una solución de Big Data.

3.1. Metodología de desarrollo

La metodología presentada a continuación fue propuesta por (Krishnan, 2013) la cual consiste de cuatro (4) etapas: recolección o recopilación, carga, transformación y extracción de datos, para este caso el autor le llama enfoque de procesamiento de Big Data.

- Etapa 1: Recolección o Recopilación de Fuentes de datos de Big Data

En la primera etapa los datos son recibidos de diferentes orígenes o fuentes, pueden ser: páginas web, redes sociales, máquina a máquina (M2M), transacciones, biometría y generados por el ser humano los cuales son producto de correos electrónicos, grabaciones de voz, documentos en papel, entre otros. Esta clasificación de los datos puede estar en formatos estructurados, y/o no estructurados y/o semi estructurados. La velocidad de adquisición y volumen de los datos depende del origen o fuente de los datos.

- Etapa 2: Carga de datos de Big Data

En esta etapa los datos se cargan en el repositorio aplicando el concepto de metadatos (datos que describen otros datos). Es de aclarar que los metadatos son información que describe características de cualquier dato, como el nombre, ubicación, importancia percibida, la calidad y sus relaciones con otros objetos de datos que la organización considere digno de la gestión.

Se busca vincular los datos entre el conjunto de datos estructurados y no estructurados con metadatos. Se debe transformar los datos no estructurados en datos estructurados.

- Etapa 3: Transformación de datos de Big Data

En este punto los datos se transforman para ser cargados en el almacén de datos tradicional mediante la aplicación de las reglas del negocio y el procesamiento de los datos.

El resultado de esta etapa es tener la información transformada en formato de Clave Valor, ej: Nombre: Juan, donde Nombre es la Clave y Juan es el valor.

- Etapa 4: Presentación de los datos de Big Data

Para efectos de estudio, se decidió realizar una adaptación del nombre de esta etapa de la metodología propuesta por (Krishnan, 2013), se realizó el cambio de "*Extracción de los datos de Big Data*" a "*Presentación de los datos de Big Data*".

El objetivo de esta etapa de la metodología, es obtener indicadores generados por la información manipulada desde la adquisición hasta el almacenamiento en el repositorio de datos tradicional para su posterior análisis. Esto va a permitir la generación de informes operativos, indicadores, para la posible visualización de ellos.

CAPÍTULO 4
MARCO APLICATIVO

En este capítulo se detallan los pasos que se siguieron para el desarrollo de este trabajo especial de grado siguiendo la metodología propuesta por (Krishnan, 2013). Por efectos de estudio, se decidió realizar una adaptación a la metodología la cual fue explicada anteriormente. Durante el desarrollo de este proyecto se contó con un equipo de desarrollo conformado por un único integrante.

4.1. Proyecto

El presente Trabajo Especial de Grado está basado en el desarrollo de una solución de big data para la obtención de indicadores que apoyen al área de gestión de talento humano específicamente en la fase de reclutamiento en el área de tecnología de la información.

4.2. Etapas del Proyecto

El proyecto se dividió en 3 etapas siguiendo la metodología de (Krishnan, 2013). Por motivos de estudio se decidió hacer una adaptación, y unir la etapa 1 y la etapa 2 en una sola etapa:

1. Etapa 1 y 2: Recolección o Recopilación de Fuentes de Datos de Big Data y Carga de Datos de Big Data.

2. Etapa 3:Transformacion de Datos de Big Data

3. Etapa 4: Presentación de los datos de Big Data

A continuación se explica el detalle el desarrollo de cada una de estas etapas que conforman la metodología aplicada en este trabajo de investigación.

4.2.1. Etapa 1 y 2 : Recolección o Recopilación de Fuentes de Datos de Big Data y Carga de datos de Big Data

En estas etapas se hace uso de herramientas como Apache Nutch, Apache Solr y HDFS para trabajar con la Araña Web. El propósito es obtener todo el contenido de las páginas webs de empleate y bumeran relacionados a cargos de Tecnología, haciendo uso de Apache Nutch y será almacenado en HDFS y automáticamente indexado con Solr.

Se cuenta con dos máquinas virtuales separadas, una primera máquina virtual nombrada como **Oracle Big Data Lite**, la cual entre sus componentes se encuentra Cloudera Manager 4.0 y en la misma se tienen las herramientas Open Source con las cuales se puede trabajar una arquitectura completa de Big Data. Seguidamente se tiene otra máquina virtual nombrada como **Oracle Business Intelligence**, la misma provee herramientas que permiten realizar análisis en profundidad de los datos almacenados, permitiendo a su vez poder trabajar con componentes propios de una arquitectura de Big Data.

Para ejecutar la Araña Web se siguen los siguientes pasos dentro de la máquina virtual Oracle Big Data Lite:

1. En la figura 26 se muestra como se ejecuta el comando *"gedit seed.txt"* para abrir el editor de texto en nuestro caso se uso gedit y se crea un nuevo archivo donde se van a definir las URLs con las cuales se va a trabajar con la Araña Web, en nuestro caso se llama seed.txt.

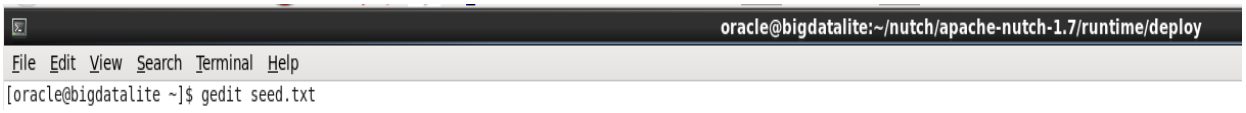


Figura 26: Comando para crear un archivo txt

En el archivo seed.txt se cargan todas las urls que van a ser usadas dentro de la araña web. Se puede ver en la figura 27 un ejemplo de los urls que se desean rastrear de la bolsa de empleo Bumeran con cargos relacionados al área de TI. Y una vez que ya ha sido cargado se guarda y cierra el archivo. Del mismo modo que fue cargado los URLs de Bumeran se realiza con la bolsa de empleo Empleate.

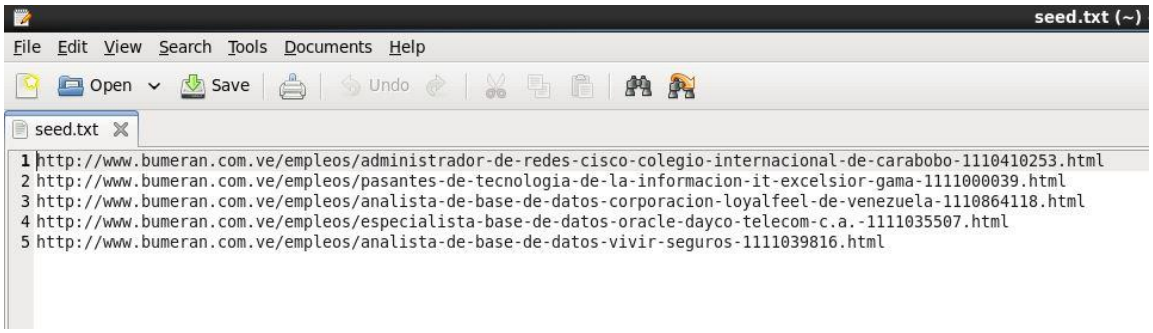


Figura 27: Incorporación de URLs usados dentro de la Araña Web

Seguidamente se carga dicho archivo seed.txt dentro del Hadoop File System (HDFS) para poder iniciar con la Araña Web seleccionada. Se carga dicho archivo dentro de una carpeta en HDFS llamada urls. La misma contendrá todos los urls de cada una de las bolsas de empleo con las que se trabaja a lo largo del proyecto. En la figura 28 se puede mostrar el paso a paso de la carga en HDFS.

Figura 28: Carga de archivo seed.txt en HDFS y verificación de la misma.

Se ejecutan los comandos *"hadoop fs –put seed.txt urls"* para realizar la carga del archivo seed.txt dentro de HDFS, específicamente en la carpeta urls. Seguidamente se utilizó el comando *"hadoop fs –cat urls/seed.txt"* para verificar que el contenido se ha copiado satisfactoriamente dentro del HDFS.

Una vez que ya se encuentra todos los urls que se quieren trabajar dentro de la Araña Web se empieza a activar los servicios en la máquina virtual relacionados con Solr, el cual permite realizar la indexación de los datos dentro de HDFS, esto se ve reflejado en la figura 29.

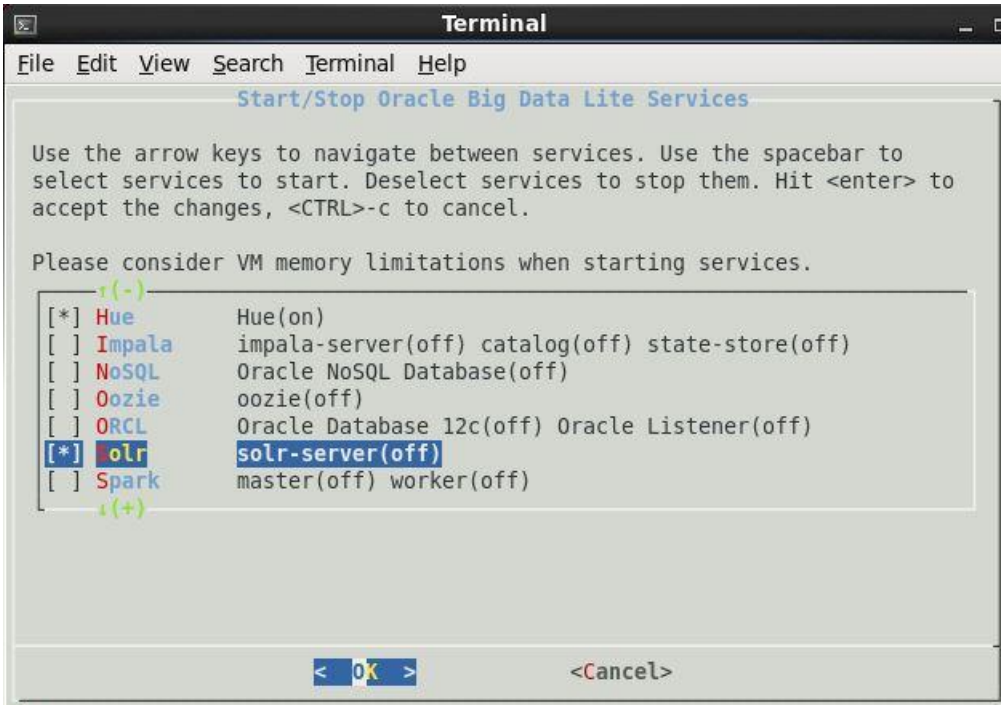


Figura 29: Terminal de Servicios dentro de la Máquina Virtual Oracle Big Data Lite.

Seguidamente se usa Apache Nutch para iniciar la Araña Web de Bolsas de Empleo. Dentro de la Máquina Virtual Oracle Big Data Lite se encuentra un Job MapReduce listo para ser ejecutado dentro de Apache Nutch, la función de dicho job, es hacer el filtrado de la data proveniente de la araña web con sentencias MapReduce. Para ejecutarlo se hace uso del siguiente código en cónsola *"cd nutch/apache-nutch-1.7/runtime/deploy"*, el cual permite iniciar Nutch; seguidamente ejecuta el Job MapReduce con el siguiente comando en cónsola *"hadoop jar apache-nutch-1.7.0.job org.apache.nutch.crawl.Crawl urls –dir CRAWL –depth 10 –topN 10 –solr http://localhost:8983/solr/collection1"*. Detallando el comando a ejecutar, se tiene que, en color rosado se identifica el Job MapReduce que ejecuta la Araña Web; en verde se encuentra definida tanto la ubicación de los urls, como los parámetros Depth y el TopN los cuales limitan la profundidad de la inspección a ser realizada por la Araña web en el análisis de las sub páginas asociadas al url principal. Por último en color azul, se define la ubicación del servicio de Solr. En la figura 30 se muestra la ejecución de los pasos descritos anteriormente.

Es importante señalar que la Araña Web se ejecuta en una máquina virtual y no en un clúster real. Por ello el proceso MapReduce toma aproximadamente 2 horas en poder extraer todo el contenido de dichos urls.

```
[oracle@bigdatalite deploy]$ hadoop jar apache-nutch-1.7.job org.apache.nutch.cr
awl.Crawl urls -dir CRAWL -depth 10 -topN 10 -solr http://localhost:8983/solr/co
llection1
16/08/14 10:59:02 INFO crawl.Crawl: crawl started in: CRAWL
16/08/14 10:59:02 INFO crawl.Crawl: rootUrlDir = urls
16/08/14 10:59:02 INFO crawl.Crawl: threads = 10
16/08/14 10:59:02 INFO crawl.Crawl: depth = 10
16/08/14 10:59:02 INFO crawl.Crawl: solrUrl=http://localhost:8983/solr/collection1
16/08/14 10:59:02 INFO crawl.Crawl: topN = 10
16/08/14 10:59:03 INFO crawl.Injector: Injector: starting at 2016-08-14 10:59:03
16/08/14 10:59:03 INFO crawl.Injector: Injector: crawlDb: CRAWL/crawldb
16/08/14 10:59:03 INFO crawl.Injector: Injector: urlDir: urls
16/08/14 10:59:03 INFO Configuration.deprecation: mapred.temp.dir is deprecated. Instead, use mapreduce.cluster.temp.dir
16/08/14 10:59:03 INFO crawl.Injector: Injector: Converting injected urls to crawl db entries.
16/08/14 10:59:04 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/08/14 10:59:06 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/08/14 11:00:32 INFO mapred.FileInputFormat: Total input paths to process : 1
16/08/14 11:00:53 INFO mapreduce.JobSubmitter: number of splits:2
16/08/14 11:00:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1471086684193_0113
16/08/14 11:01:02 INFO impl.YarnClientImpl: Submitted application application_1471086684193_0113
16/08/14 11:01:02 INFO mapreduce.Job: The url to track the job: http://bigdatalite.localdomain:8088/proxy/application_1471086684193_0113/
16/08/14 11:01:02 INFO mapreduce.Job: Running job: job_1471086684193_0113
16/08/14 11:03:57 INFO mapreduce.Job: Job job_1471086684193_0113 running in uber mode : false
16/08/14 11:03:57 INFO mapreduce.Job:  map 0% reduce 0%
16/08/14 11:06:37 INFO mapreduce.Job:  map 83% reduce 0%
16/08/14 11:06:39 INFO mapreduce.Job:  map 100% reduce 0%
16/08/14 11:07:06 INFO mapreduce.Job:  map 100% reduce 100%
16/08/14 11:07:07 INFO mapreduce.Job: Job job_1471086684193_0113 completed successfully
16/08/14 11:07:08 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=687
                FILE: Number of bytes written=330500
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1015
                HDFS: Number of bytes written=797
                HDFS: Number of read operations=9
```

Figura 30: Ejecución en cónsola de la Araña Web

Una vez que finaliza la ejecución de la Araña Web, en cónsola aparece un mensaje, tal como se muestra en la figura 31.

```
                Spilled Records=162
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=184
                CPU time spent (ms)=2900
                Physical memory (bytes) snapshot=738062336
                Virtual memory (bytes) snapshot=2650591232
                Total committed heap usage (bytes)=529530880
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
INFO solr.SolrDeleteDuplicates: SolrDeleteDuplicates: finished
INFO crawl.Crawl: crawl finished: CRAWL
[oracle@bigdatalite deploy]$
```

Figura 31: Culminación de Araña Web

Ahora se procede a revisar en Solr cual fue la indexación que realizó la Araña Web y el contenido que nos arrojó de los urls de Empléate y Bumeran. Para ejecutar Solr se coloca

en un navegador web la siguiente ruta *"localhost:8983/solr/#/".* Una vez en Solr, se selecciona en Core Selector la opción de Collection1. En el mismo se encuentra todo el contenido que se pudo capturar por la Araña Web. Se puede visualizar lo descrito anteriormente en la siguiente figura 32.

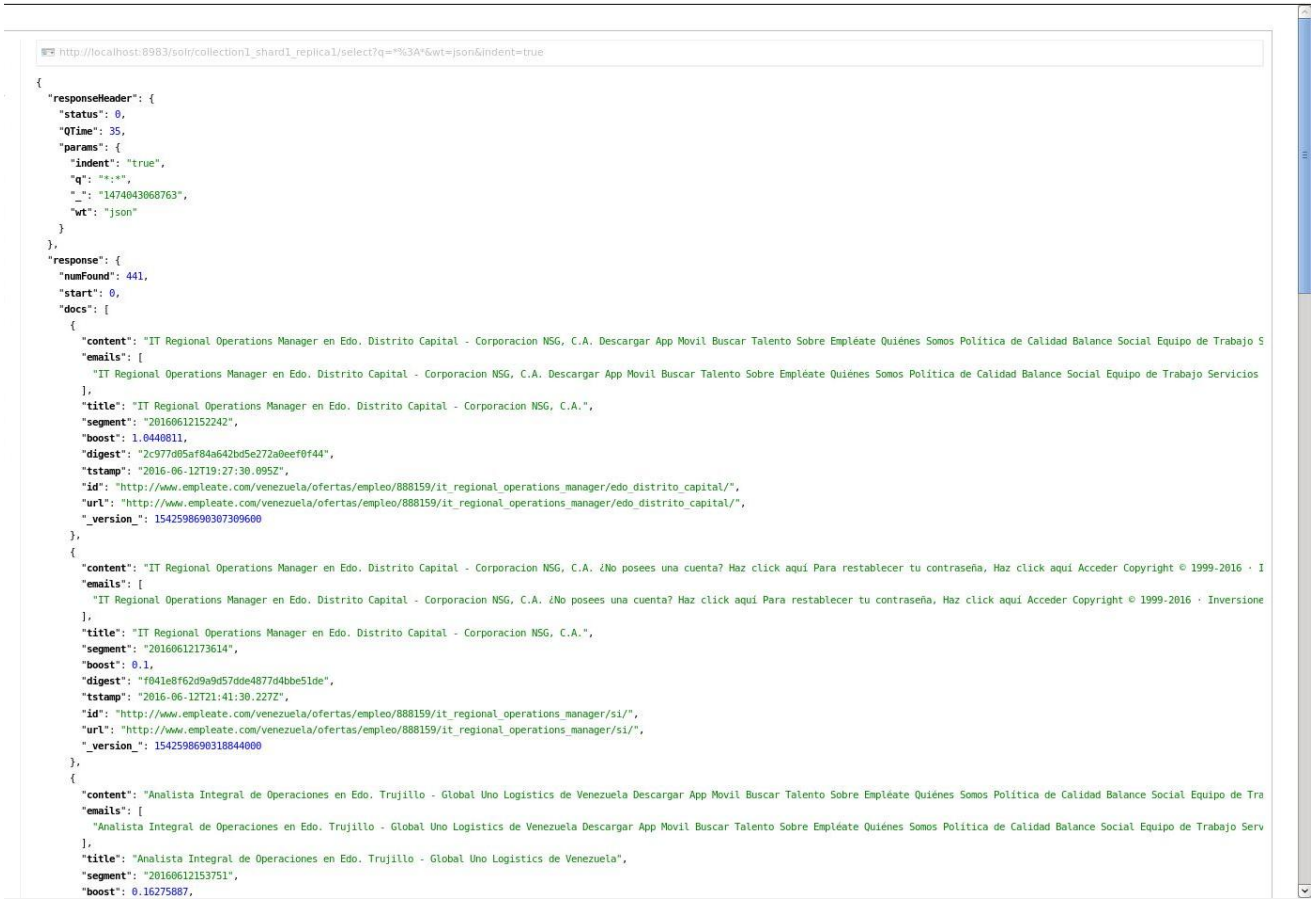


Figura 32: Contenido en Solr

En la figura anterior también se observa la cantidad de Urls y subpáginas obtenidas de la Araña Web a las bolsas de empleo bumeran y empléate, en el campo numfound nos señala un total de 441 urls. Adicional a esto también se obtiene información relacionada al contenido de la página web, el título, el url y el tmstamp del día que fue extraída la información.

Una vez que se revisa el contenido resultante de la Araña Web se procede a exportar el contenido almacenado en HDFS usando la interfaz gráfica HUE. El contenido se exporta en formato XLS. En la figura 33 se puede observar el contenido tal como se encuentra almacenado en HDFS y concuerda con la cantidad de urls adquiridos de la Araña Web.

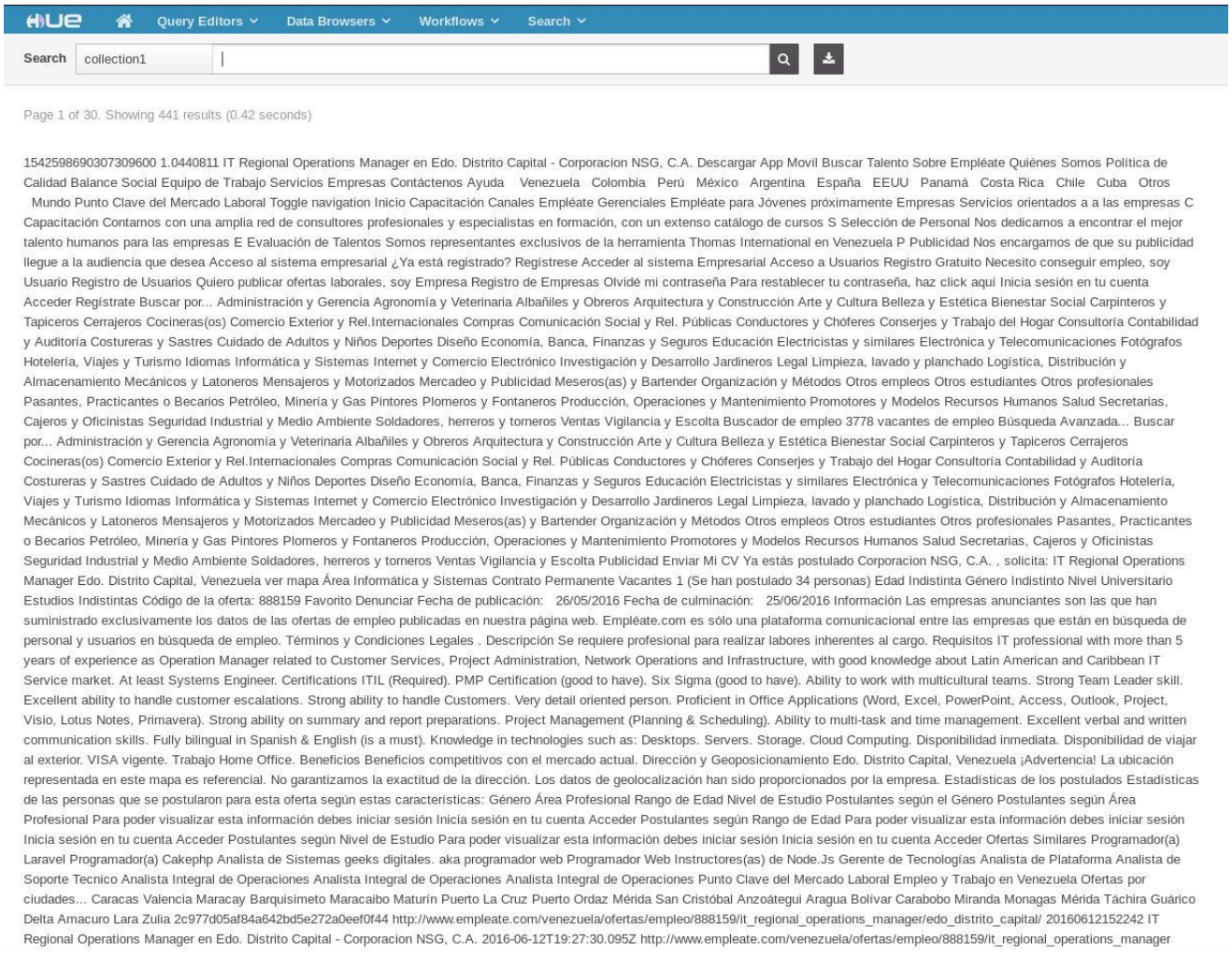


Figura 33: Contenido Almacenado en HDFS

Seguidamente luego de ser exportadas las 30 páginas del contenido adquirido de la Araña Web, se visualiza en formato XLS y se hace énfasis en el Content de las páginas webs extraídas. Ya que en el mismo se encuentra toda la información relacionada al cargo que las empresas postularon en las bolsas como Empléate y Bumeran, en la siguientes figuras 34 y 35 se muestra el contenido ya exportado de ambas bolsas de empleo en formato XLS.

De la Bolsa de empleo Empléate se extrajo un total de 40 cargos, mientras que en Bumeran se logró extraer un total de 39 cargos. Sumando así un total de 79 vacantes de empleo para analizar.

title	url	segment	tstamp	content	id	_version_	boost	emails	digest
ADMINISTRADOR BASE DE DATOS En Person to Person Análisis,	http://www.bu	20160814114	2016-08-14T1	ADMINISTRA	http://www.bu	1,54268E+18	0.16872957	[u'ADMINISTF	19089b7139361bc4168273842c6037c6
Administrador de Base de Datos En Farmatodo, C.A. - 20 de junio d	http://www.bu	20160814114	2016-08-14T1	Administrado	http://www.bu	1,54268E+18	0.1598599	[u'Administra	9e83269c843d18825c9008f0f71d4385
Administrador de Base de Datos En IMPORTANTE EMPRESA DE S	http://www.bu	20160814114	2016-08-14T1	Administrado	http://www.bu	1,54268E+18	0.17130919	[u'Administra	b0045e6103c4f5e812a11ebcdca92843
Administrador de Redes En Manaplas S.A - 03 de agosto de 2016 - I	http://www.bu	20160814122	2016-08-14T1	Administrado	http://www.bu	1,54268E+18	0.15632048	[u'Administra	0897988df8766f9af0d485de757830a5
Analista de Soporte Técnico En SOPORTE SPI, C. A. - 26 de julio d	http://www.bu	20160814120	2016-08-14T1	Analista de S	http://www.bu	1,54268E+18	0.15213197	[u'Analista de	0661a4df067f20e9dfb1984892432fe9
Analista de Soporte Técnico En SOPORTE SPI, C. A. - 26 de julio d	http://www.bu	20160814122	2016-08-14T1	Analista de S	http://www.bu	1,54268E+18	0.1524025	[u'Analista de	ff1d71e6de51897f4f41b999522a65d2
Analista de Soporte Tecnico En SOPORTE SPI, C. A. - 04 de agost	http://www.bu	20160814122	2016-08-14T1	Analista de S	http://www.bu	1,54268E+18	0.15229736	[u'Analista de	35a1c62d25268e77cd24c41c91e0f09b
Analista de Soporte Técnico En Banco Activo - 01 de agosto de 201	http://www.bu	20160813172	2016-08-13T2	Analista de S	http://www.bu	1,5426E+18	0.15858497	[u'Analista de	bc379138e362a1fa9267b00a9a6c79bc
Analista de Soporte Técnico En Instituto Médico La Floresta - 20 de	http://www.bu	20160813172	2016-08-13T2	Analista de S	http://www.bu	1,54268E+18	0.15655607	[u'Analista de	22509f1084104aaf1d017af1ce9f0131
Analista de Soporte Tecnico En INVERSIONES COLODRA, C.A. - 0	http://www.bu	20160813170	2016-08-13T2	Analista de S	http://www.bu	1,54268E+18	0.15534712	[u'Analista de	b2aa14110c1fff53f50e2c6416bec39e
Analista Integral de Sistema En CORPORACION LOYALFEEL DE V	http://www.bu	20160814114	2016-08-14T1	Analista Integ	http://www.bu	1,54268E+18	0.15966062	[u'Analista Int	d537356c9e3e08bc0ba6468be41622c1
Auditor de Sistemas Senior En Vivir Seguros - 20 de junio de 2016 -	http://www.bu	20160814120	2016-08-14T1	Auditor de Si	http://www.bu	1,54268E+18	0.15598978	[u'Auditor de	6a4046820652109570a24c85db1fa966
Consultor Especialista de Infraestructura IT En Manapro Consultores	http://www.bu	20160813170	2016-08-13T2	Consultor Es	http://www.bu	1,5426E+18	0.15969197	[u'Consultor E	31e81a933c2b7090d843d266c274cfc6
Consultor Redes & Telecom En DIGITALES 2008, S.A. - 11 de julio d	http://www.bu	20160813170	2016-08-13T2	Consultor Re	http://www.bu	1,5426E+18	0.16458972	[u'Consultor F	eadb90da058280983cea2e7e53fc6f77
Consultor Redes y Telecom En DIGITALES 2008, S.A. - 26 de julio d	http://www.bu	20160813170	2016-08-13T2	Consultor Re	http://www.bu	1,5426E+18	0.16458972	[u'Consultor F	270a15d40b7b5cd69a2f765893bf4d30
Consultor Routing and Switching En ADVANTEL CONSULTORES - (	http://www.bu	20160813170	2016-08-13T2	Consultor Ro	http://www.bu	1,5426E+18	0.16523454	[u'Consultor F	d0263738cbf14be9abe9e64b9b173358
Consultor Senior de Desarrollo En OLTP VOICE SYSTEM - 30 de jur	http://www.bu	20160813174	2016-08-13T2	Consultor Se	http://www.bu	1,5426E+18	0.13565373	[u'Consultor S	a799ccb24afc1bae63a5543df55b27e6
CONSULTOR SENIOR DE TECNOLOGIA En LETO IT CONSULTING	http://www.bu	20160813152	2016-08-13T1	CONSULTOR	http://www.bu	1,5426E+18	1.1258761	[u'CONSULT(	cd0037ac3ed03dbfbd8c81b12a919352
Coordinador de Sistemas y TI En CORPORACION LOYALFEEL DE	http://www.bu	20160814114	2016-08-14T1	Coordinador c	http://www.bu	1,54268E+18	0.16094306	[u'Coordinado	3e25dcc956899e875be7e2e8de8bf96f
DESARROLLADOR & PROGRAMADOR WEB En Soluciones Integr	http://www.bu	20160813172	2016-08-13T2	DESARROLL	http://www.bu	1,5426E+18	0.14837316	[u'DESARRO	6e4ec3a999b6fae9cb6a5178cd39c026
Desarrollador / Programador de Sistemas Informáticos En Siaca Ser	http://www.bu	20160814114	2016-08-14T1	Desarrollador	http://www.bu	1,54268E+18	0.15586345	[u'Desarrollad	a73ed2404b765bb7e7f59dae439f4f75
DESARROLLADOR DE SISTEMAS SENIOR En Stefanelli & Asocia	http://www.bu	20160813174	2016-08-13T2	DESARROLL	http://www.bu	1,5426E+18	0.14381833	[u'DESARRO	829ecea5e42962a75b933e9508d2640d
Desarrollador Odoo Sr. (Free Lance) En 4Geeks - 29 de junio de 201	http://www.bu	20160813174	2016-08-13T2	Desarrollador	http://www.bu	1,5426E+18	0.14678246	[u'Desarrollad	3437f875ec2d2e315118b2606143a9fc
Desarrollador Php Laravel Sr. (Free Lance) En 4Geeks - 12 de agost	http://www.bu	20160813152	2016-08-13T1	Desarrollador	http://www.bu	1,5426E+18	1.1099317	[u'Desarrollad	5341309c430ad7a6f46ba6c742d9e5e5

Figura 34: Cargos obtenidos de la bolsa de empleo Bumeran

title	url	segment	tstamp	content	id	_version_	boost	emails	digest
Coordinador c	http://www.er	20160810222	2016-08-11T0	Coordinador c	http://www.er	1,5426E+18	0.3270037	[u'Coordinado	834dbe094f0e239d8f4d161a382e1576
Auditor (a) de	http://www.er	20160810222	2016-08-11T0	Auditor (a) de	http://www.er	1,5426E+18	0.3270037	[u'Auditor (a)	349653c10f9bb98009254c1a3f0583ed
Coordinador c	http://www.er	20160810223	2016-08-11T0	Coordinador c	http://www.er	1,5426E+18	0.3279111	[u'Coordinado	ad254b991e4a1288353247e7f0084be5
Administrado	http://www.er	20160810223	2016-08-11T0	Administrado	http://www.er	1,5426E+18	0.32781222	[u'Administra	c97bfae893bf4cc7437066f8d156740d
Administrado	http://www.er	20160810223	2016-08-11T0	Administrado	http://www.er	1,5426E+18	0.3277153	[u'Administra	75bd2e572ce3f1251f6297ea1c8c9b92
Ingenieros(as	http://www.er	20160810223	2016-08-11T0	Ingenieros(as	http://www.er	1,5426E+18	0.3277153	[u'Ingenieros(	85ee238112f403d7bdffefa4e487a4ac
Supervisor de	http://www.er	20160810223	2016-08-11T0	Supervisor de	http://www.er	1,5426E+18	0.3277153	[u'Supervisor	71c474905bcd7bf6c2ad4e608d9b4bf3
Analista Prog	http://www.er	20160810223	2016-08-11T0	Analista Prog	http://www.er	1,5426E+18	0.3277153	[u'Analista Pr	aeeb33c80bbddb9ffd46276d0f865fc8
Ingeniero Pro	http://www.er	20160810203	2016-08-11T0	Ingeniero Pro	http://www.er	1,5426E+18	0.326661	[u'Ingeniero P	c68e6145db217abcdf8eb7fe7546ec26
Consultor de	http://www.er	20160812195	2016-08-13T0	Consultor de	http://www.er	1,5426E+18	1.0440307	[u'Consultor c	c4f7c6000716d112771efd9d173bb0ce
Coordinador(a	http://www.er	20160810201	2016-08-11T0	Coordinador(a	http://www.er	1,5426E+18	1.0440307	[u'Coordinado	9d6fa74d38e403a5461eefdc3571df12
Instructores(a	http://www.er	20160810201	2016-08-11T0	Instructores(a	http://www.er	1,5426E+18	1.0441294	[u'Instructore	09e034f9fe42061d75e02cd22d8f52e0
Gerente de Á	http://www.er	20160810201	2016-08-11T0	Gerente de Á	http://www.er	1,5426E+18	1.0440307	[u'Gerente de	e614320835b25a2ce633e8b5e5f1f5ba
Instructores (	http://www.er	20160810201	2016-08-11T0	Instructores (	http://www.er	1,5426E+18	1.0441294	[u'Instructore	3d77644888b5e660ca5c7614927eeecb
Analista Func	http://www.er	20160810230	2016-08-11T0	Analista Func	http://www.er	1,5426E+18	0.15018643	[u'Analista Fu	fb1e3702c8a93083985b8c53dabb7daf
Responsable	http://www.	2,0161E+13	2016-08-11T(	Responsable	http://www.	1,54E+18	0,326661	[u'Responsal	8ad4b63932591a57671d1e799a78d180
Analistas y P	http://www.	2,0161E+13	2016-08-11T(	Analistas y P	http://www.	1,54E+18	0,3270037	[u'Analistas	af4921747b1adb5f4bd1faf98da2762d
Consultores	http://www.	2,0161E+13	2016-08-11T(	Consultores	http://www.	1,54E+18	0,3270037	[u'Consultor	44a9cf23f6592b694b1e4f8d3bfa3684
Analista Inte	http://www.	2,0161E+13	2016-06-12T2	Analista Inte	http://www.	1,54E+18	0,07375497	[u'Analista Ir	8b35c24712194255b75fee57ac614be8
Analista Inte	http://www.	2,0161E+13	2016-06-12T1	Analista Inte	http://www.	1,54E+18	0,16275887	[u'Analista Ir	b1096a0921de2e46342db2b13015e55a
Analista de F	http://www.	2,0161E+13	2016-06-12T1	Analista de F	http://www.	1,54E+18	0,16275887	[u'Analista d	50291d2b13c0efe468b540ebbd792c7d
Instructores(	http://www.	2,0161E+13	2016-06-12T1	Instructores(	http://www.	1,54E+18	0,16275887	[u'Instructor	accf68f711570631858fe278185f5f58
Programador	http://www.	2,0161E+13	2016-06-12T1	Programador	http://www.	1,54E+18	0,16329029	[u'Programa	afb6851837f7433fc60af001c15391f9
Gerente de T	http://www.	2,0161E+13	2016-06-12T1	Gerente de T	http://www.	1,54E+18	0,16275887	[u'Gerente d	a6ceca1bd514c22aaf6b6f6f759592c7

Figura 35: Cargos obtenidos de la bolsa de empleo Empléate

4.2.2. Etapa 3: Transformación de Datos de Big Data

Una vez que los datos son exportados se procede a estudiar los patrones relacionados a la manipulación de la data dentro del registro Content por cada una de las bolsas de empleo. Se encontraron patrones distintos por cada bolsa de empleo, lo cual conllevo a que se trabajara por separado cada una ellas dentro de la herramienta Pentaho Data Integrator.

Una vez conocida cual es la información que se necesita extraer del content de cada una de las bolsas de empleo, se procede a trabajar con Pentaho Data Integrator, la cual nos va

ayudar a la transformación de data no estructurada a data estructurada. Es importante resaltar según (Inquidia Consulting, 2015) la conexión entre Pentaho Data Integrator y Hive aún no se encuentra implementada, motivo por el cual, se exporta la data resultante de las transformaciones en formato xls para luego ser insertadas en Hive.

A continuación se detallan los procesos de transformación para cada una de las dimensiones y tabla de hecho mencionados anteriormente.

4.2.2.1. Transformación Información de Cargo de la data arrojada por la bolsa de empleo Bumeran

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados a los cargos por parte de la bolsa de empleo Bumeran.

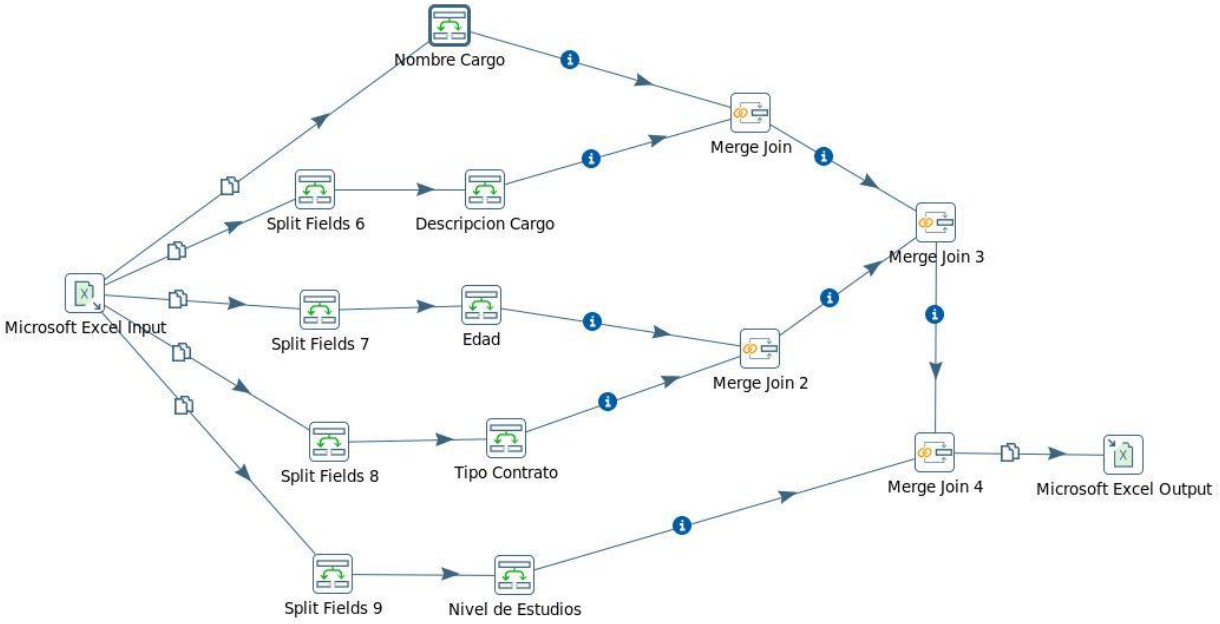


Figura 36: Transformación Cargo de la bolsa de empleo Bumeran

4.2.2.2. Transformación Información Empresa de la data arrojada por la bolsa de empleo Bumeran

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados a los empresas por parte de la bolsa de empleo Bumeran.

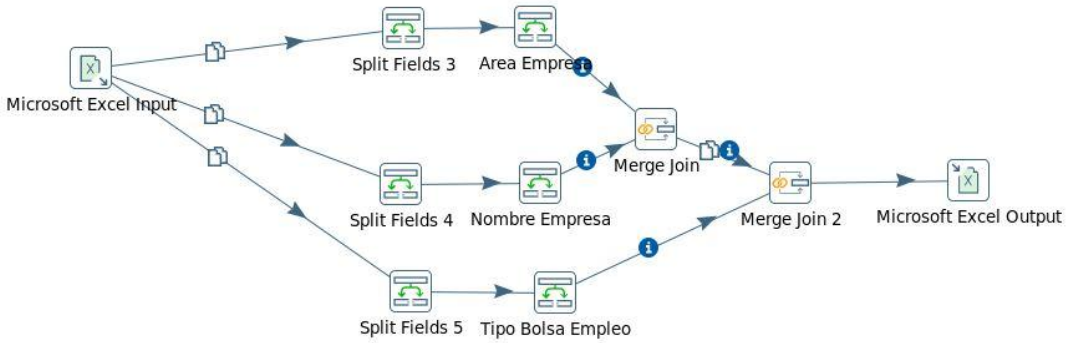


Figura 37: Transformación Empresa de la bolsa de empleo Bumeran

4.2.2.3. Transformación Información Tiempo de la data arrojada por la bolsa de empleo Bumeran

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados a Tiempo por parte de la bolsa de empleo Bumeran.

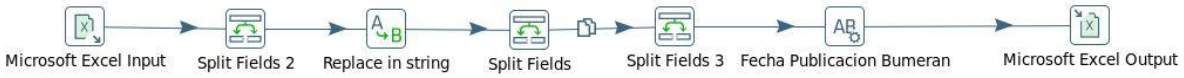


Figura 38: Transformación Tiempo de la bolsa de empleo Bumeran

4.2.2.4. Transformación Información Ubicación de la data arrojada por la bolsa de empleo Bumeran

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados a las ubicaciones por parte de la bolsa de empleo Bumeran.

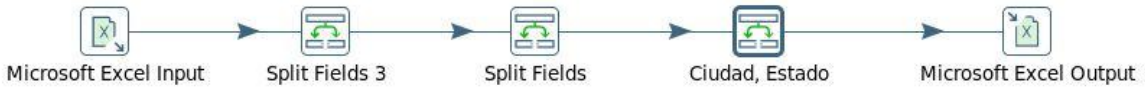


Figura 39: Transformación Ubicación de la bolsa de empleo Bumeran

4.2.2.5. Transformación Información Cargo de la data arrojada por la bolsa de empleo Empléate.

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados a los cargos por parte de la bolsa de empleo Empléate.

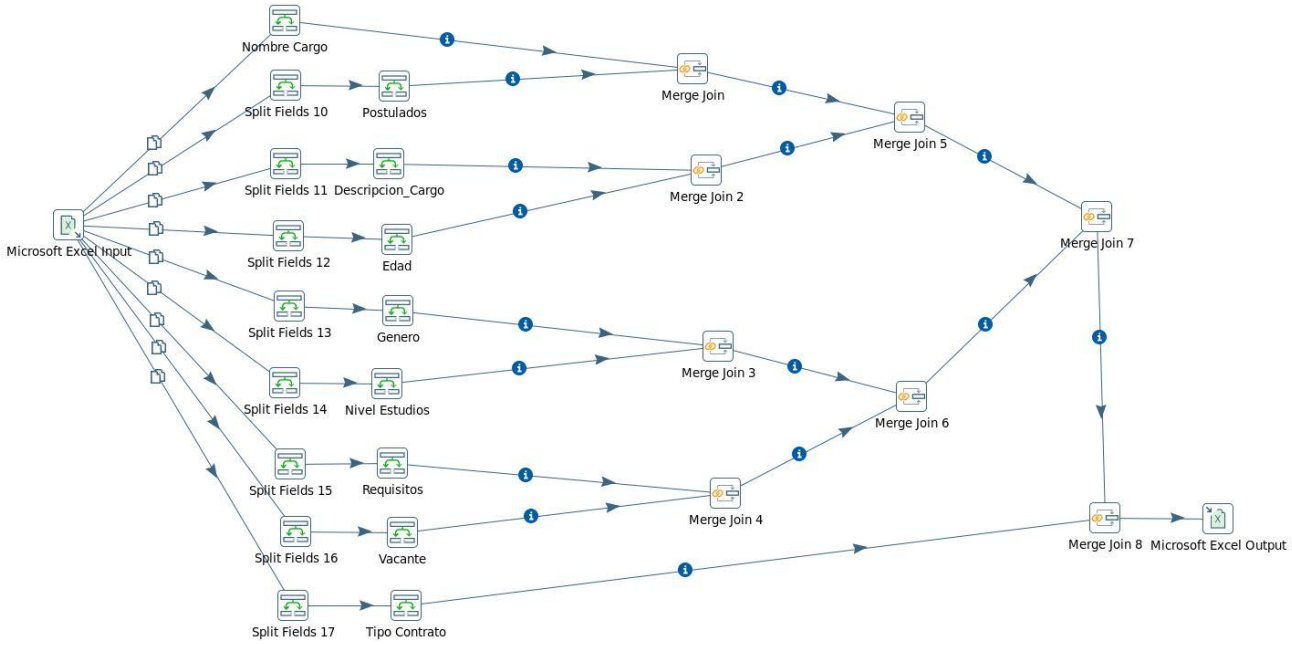


Figura 40: Transformación Cargo de la bolsa de empleo Empléate

4.2.2.6. Transformación Información Empresa de la data arrojada por la bolsa de empleo Empleate.

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados a las empresas por parte de la bolsa de empleo Empleate.

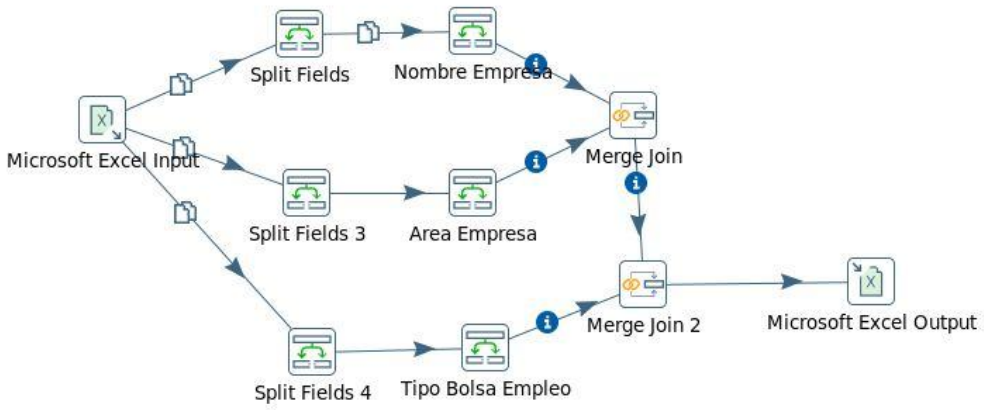


Figura 41: Transformación Empresa de la bolsa de empleo Empléate

4.2.2.7. Transformación Información Tiempo de la data arrojada por la bolsa de empleo Empléate.

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados al tiempo por parte de la bolsa de empleo Empléate.

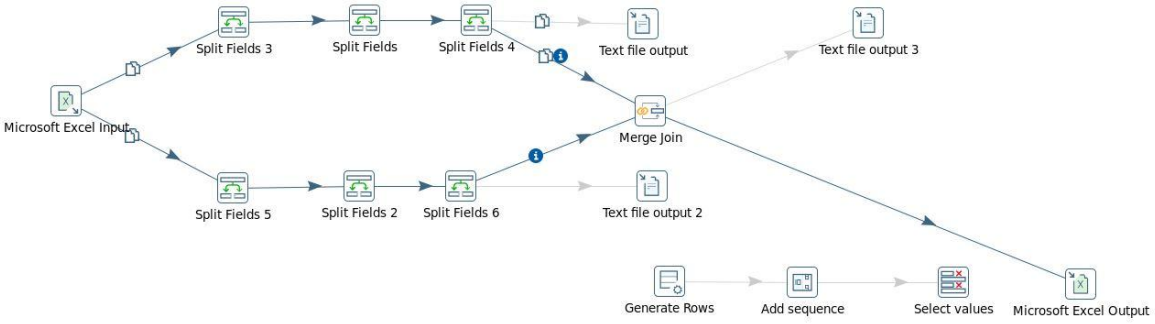


Figura 42: Transformación Tiempo de la bolsa de empleo Empleate

4.2.2.8. Transformación Información Ubicación de la data arrojada por la bolsa de empleo Empléate.

Se muestra la transformación de data no estructurada a data estructurada que se realizó para extraer todos los valores relacionados a la Ubicación por parte de la bolsa de empleo Empleate.

Figura 43: Transformación Ubicación de la bolsa de empleo Empléate

Una vez que se tiene todas las transformaciones por separado de las bolsas de empleo, se procede a unirlas y generar la tabla de hechos. A continuación se listan unas figuras relacionado a lo descrito anteriormente.

4.2.2.9. Transformación Tabla de Hechos

Se procede a realizar la unión de las transformaciones relacionadas a las informaciones de las bolsas de empleo empléate y bumeran, se puede visualizar en la siguiente figura 44.

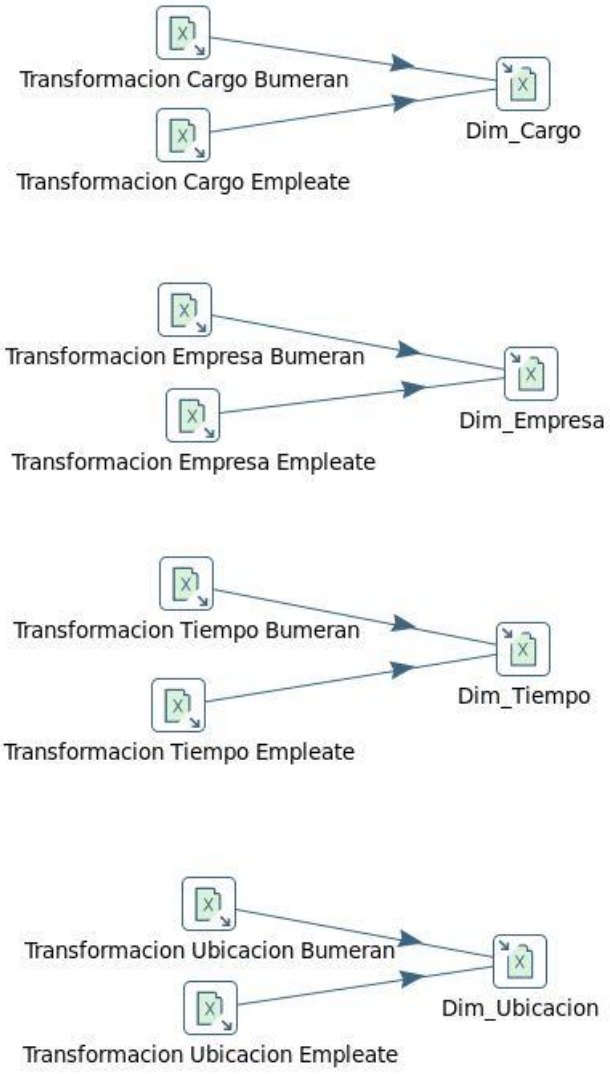


Figura 44: Unión Transformaciones Bolsas de Empleo

Seguidamente se procede a realizar la transformación de la tabla de hechos junto a las métricas correspondientes, se puede visualizar en la siguiente figura 45.

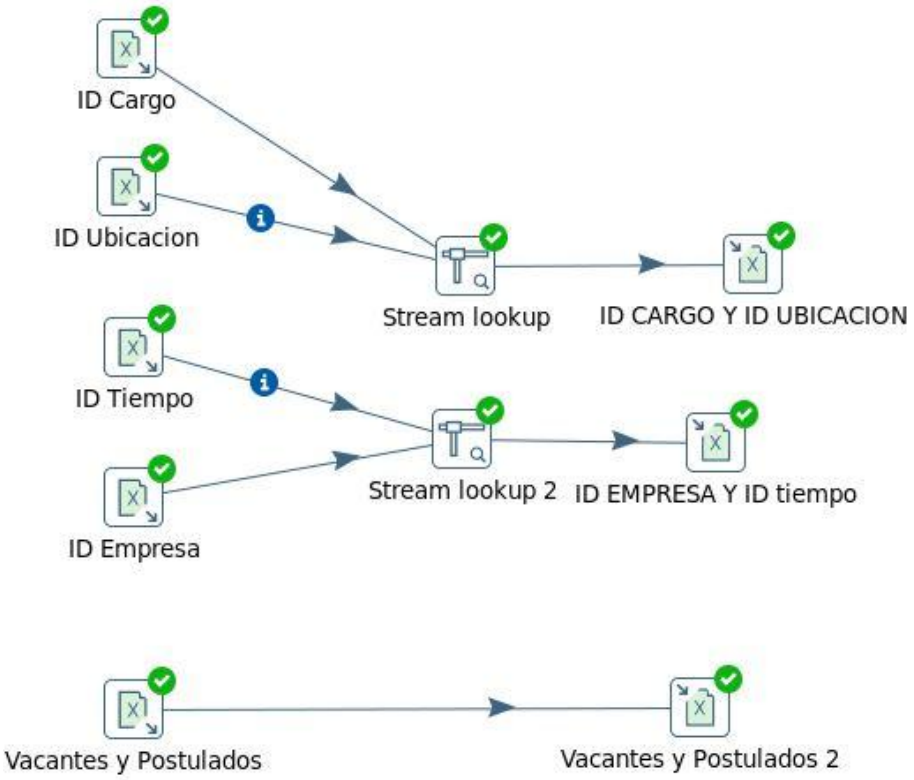


Figura 45: Transformación Tabla de Hechos

Una vez que se tiene toda la data transformada para cada una de las dimensiones y la tabla de hechos. Se procede a exportar el contenido en formato CSV para posteriormente ser insertado en el almacén de datos Hive. A continuación se procede con la etapa de presentación de los datos de Big Data.

4.2.3. Etapa 4 Presentación de los datos de Big Data

4.2.3.1. Creación de las Dimensiones con Apache Hive

El análisis de la data resultante de la araña web nos permitió definir las dimensiones y la tabla de hechos tomando la data resultante de las transformaciones previamente realizadas. A continuación se muestra en la figura 46 el modelo dimensional correspondiente a este estudio.

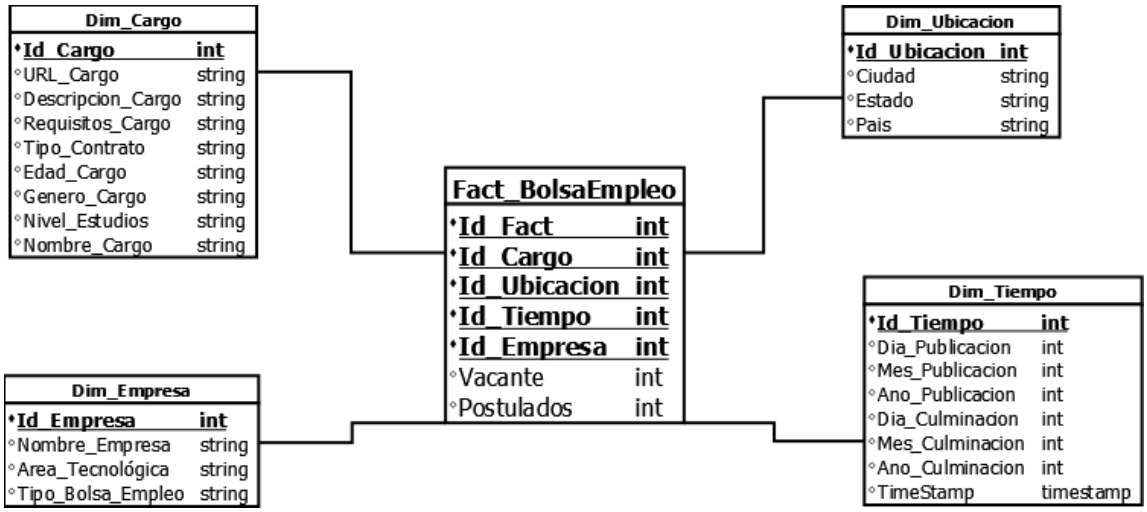


Figura 46: Modelo Dimensional

A continuación se muestra una tabla con la descripción de cada una de las dimensiones:

Tabla 3: Dimensiones

Dimensión	Descripción
Tiempo	Contiene las fechas en que se ha realizado la publicación y fecha de clausura de la publicación, dividida en día, mes y año. Al igual contiene el timestamp de la fecha en que fue visitado el sitio
Empresa	Contiene la información asociada a la empresa.
Cargo	Almacena la información asociada a los cargos postulados por las empresas.
Ubicación	Contiene información relacionada a la ubicación de la empresa.

Seguidamente, se crearon archivos CSV con la información asociada a cada una de las dimensiones que serian desplegadas en Hive. En este momento se trabaja con la segunda máquina virtual Oracle Business Intelligence y se asegura que los servicios de Cloudera Manager Activos antes de ejecutar comandos en consola.

Una vez que los servicios se encuentran activos, se continua la siguiente guía (REPO CODIGO, 2015) se ejecuta el siguiente comando *"sudo –u hdfs hadoop fs –mkdir /user/tesis"* para crear dentro de HDFS un directorio llamado tesis, luego se procede a asignarle permisología de root, por lo cual se ejecuta el siguiente comando *"sudo –u hdfs hadoop fs –chown root /user/tesis"* tal como se muestra en la siguiente figura 47.

```
sudo -u hdfs hadoop fs -mkdir /user/tesis
sudo -u hdfs -chown root /user/tesis
sudo -u hdfs hadoop -chown root /user/tesis
clear
sudo -u hdfs hadoop fs -chown root /user/tesis
```

Figura 47: Creando Directorios en HDFS

Una vez que ya se tenían los archivos CSV, se procedió a importarlos en hive de la siguiente manera: se realiza un put de los archivos a cargar en hdfs en la carpeta creada, se accede al directorio donde se encuentran los archivos y se ejecuta el siguiente código tal como muestra la siguiente figura 48.

```
cd Desktop/
cd Dimensiones\ Tesis/
ls
sudo hadoop fs -put Tabla_Cargo.csv /user/tesis/
sudo hadoop fs -put Tabla_Empresa.csv /user/tesis/
sudo hadoop fs -put Tabla_Fact.csv /user/tesis/
sudo hadoop fs -put Tabla_Tiempo.csv /user/tesis/
sudo hadoop fs -put Tabla_Ubicacion.csv /user/tesis/
sudo hadoop fs -ls /user/tesis
sudo -u hdfs hadoop fs -chown admin /user/tesis
```

Figura 48: Copiando archivos al directorio creado en HDFS

Luego se crean las dimensiones en Hive a partir de los archivos ya cargados en HDFS. Donde, primero se conecta a HUE, interfaz gráfica de HDFS.

- Creación de la Dimensión Cargo

Se selecciona Metastore y se crea la dimensión cargo a partir del archivo CSV previamente cargado, tal como se muestran en las siguientes figuras. En la figura 49 se muestra la importación del archivo previamente cargado con los datos de la dimensión, en la figura 50 se selecciona el tipo de separador de las columnas del archivo importado y por ultimo en la figura 51 se muestran los tipos de datos pertenecientes a las columnas.

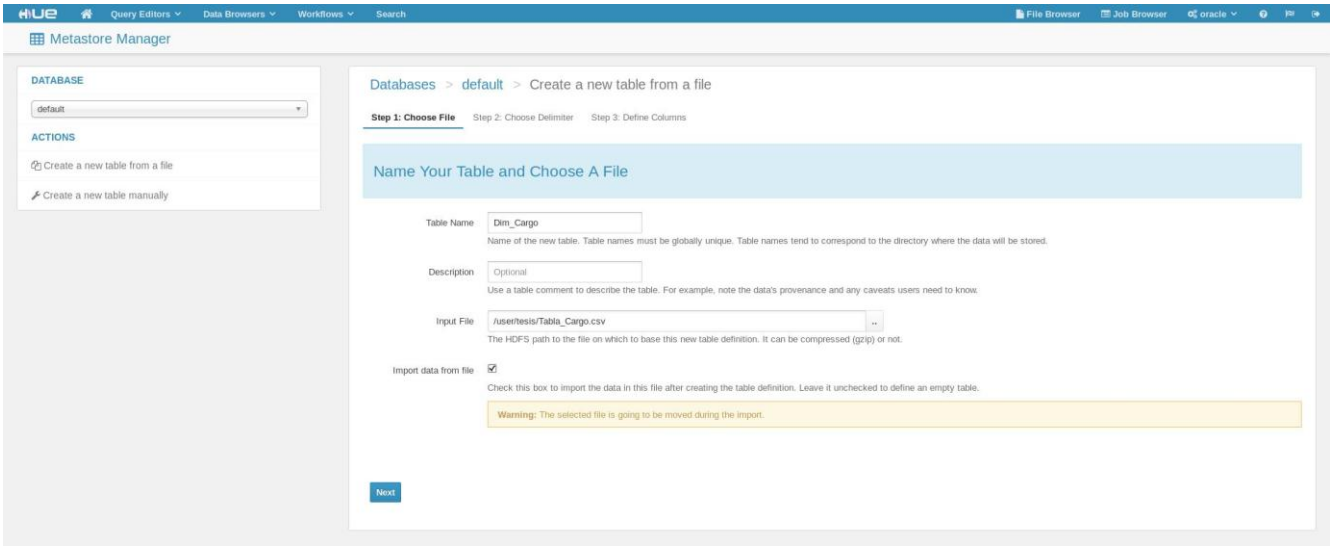


Figura 49: Creación Dimensión Cargo

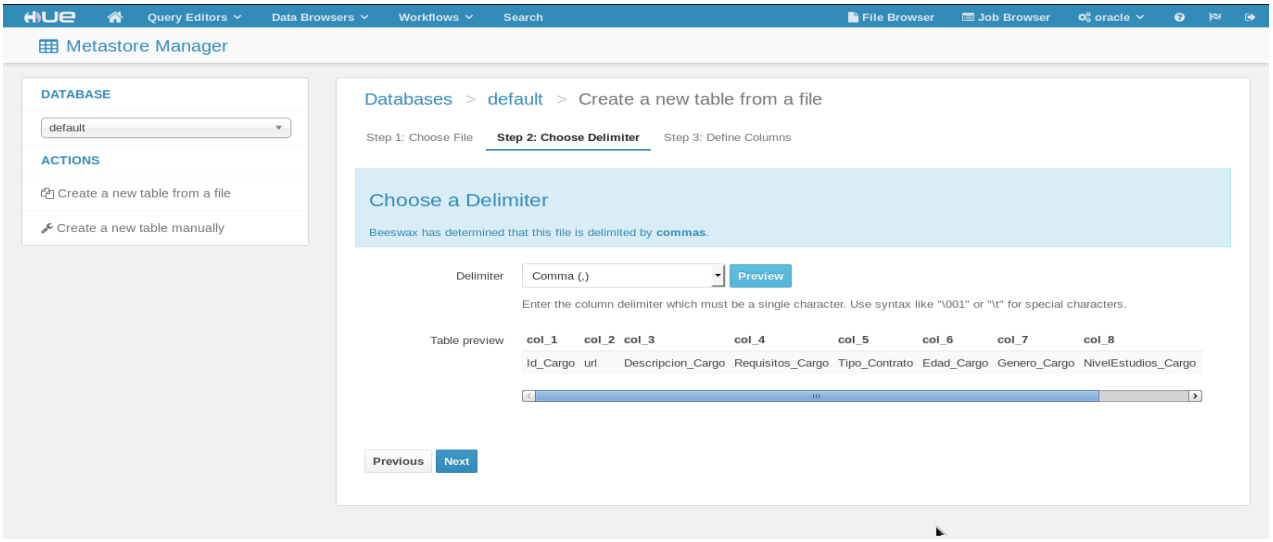


Figura 50: Delimitador de Columnas Dimensión Cargo

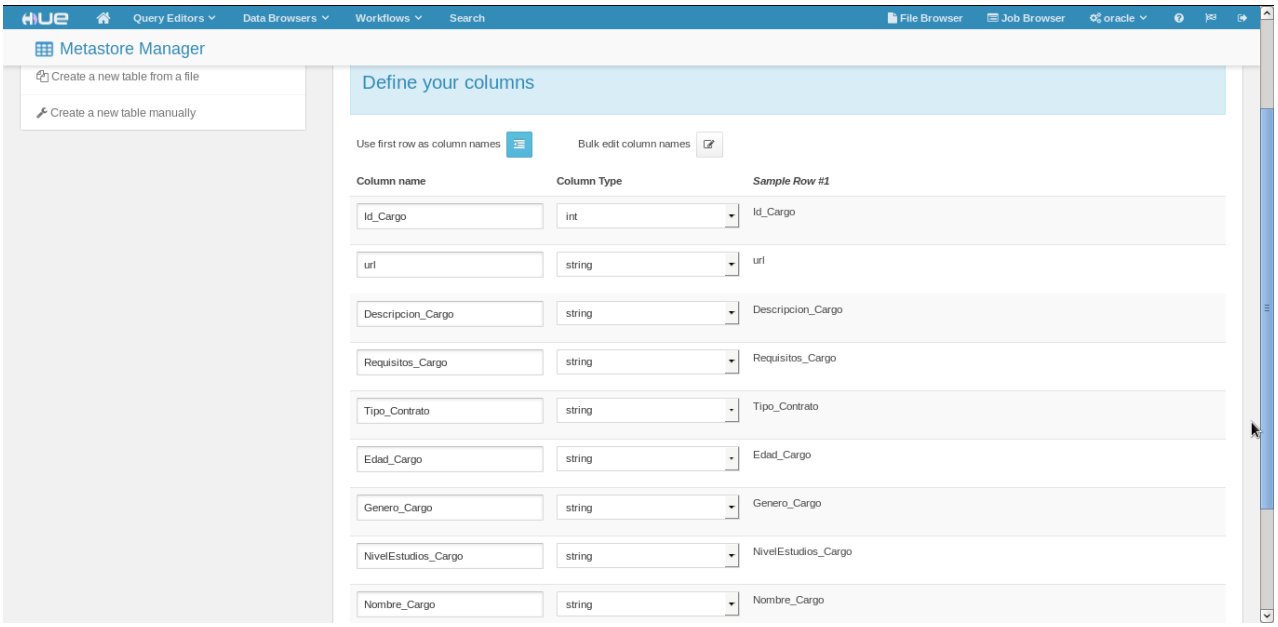


Figura 51: Tipo de Datos de columnas Dimensión Cargo

- Creación de la Dimensión Empresa

En la figura 52 se muestra la importación del archivo CSV previamente cargado con los datos de la dimensión, en la figura 53 se selecciona el tipo de separador de las columnas del archivo importado y por ultimo en la figura 54 se muestran los tipos de datos pertenecientes a las columnas.

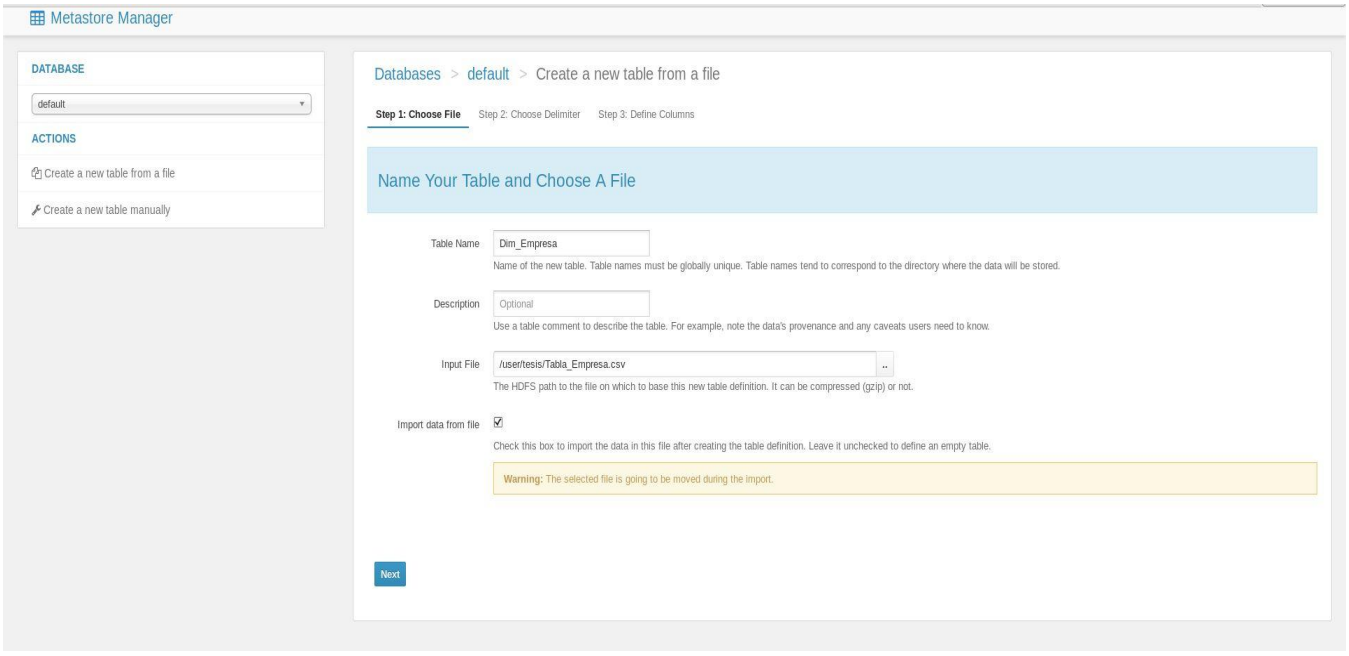


Figura 52: Creación Dimensión Empresa

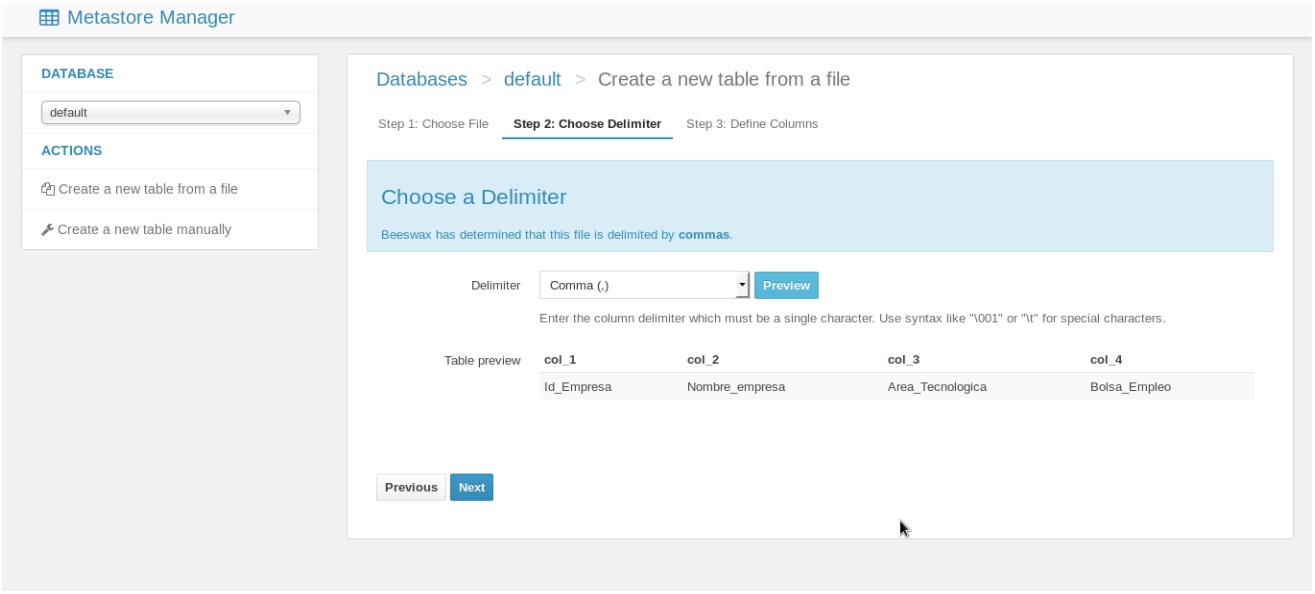


Figura 53: Delimitador de Columnas Dimensión Empresa

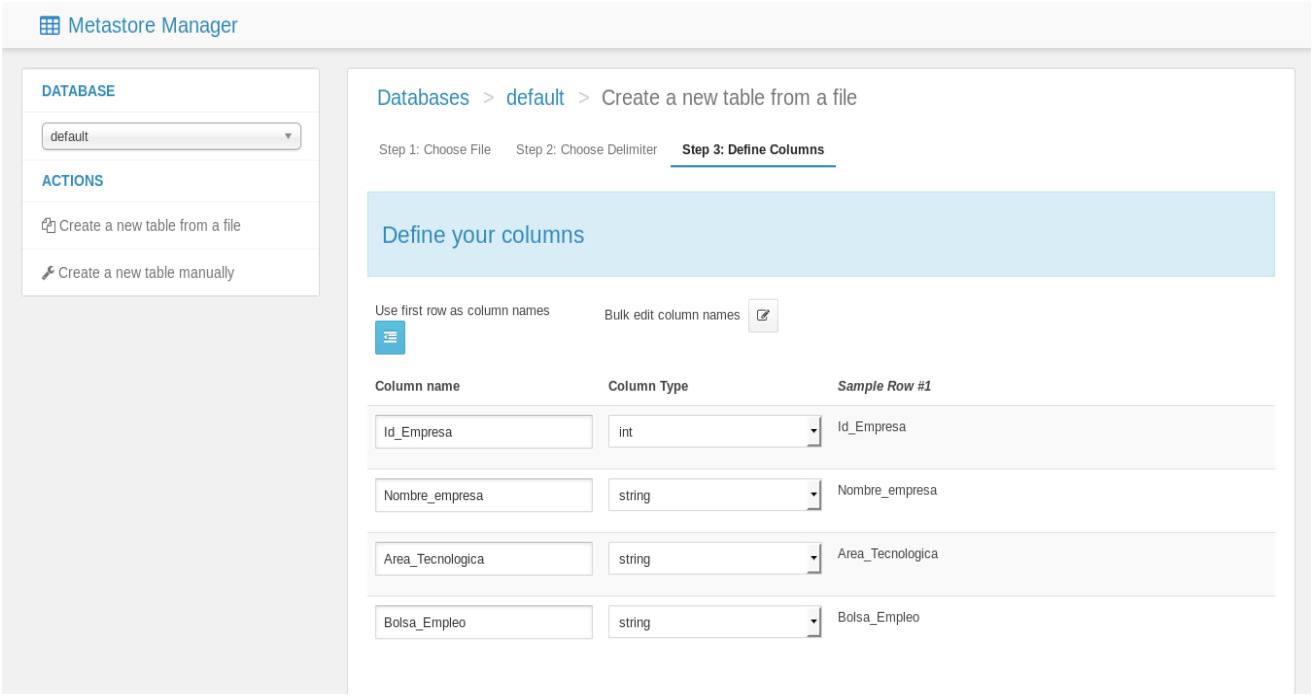


Figura 54: Tipo de Datos de columnas Dimensión Empresa

- Creación de la Dimensión Ubicación

En la figura 55 se muestra la importación del archivo CSV previamente cargado con los datos de la dimensión, en la figura 56 se selecciona el tipo de separador de las columnas del archivo importado y por ultimo en la figura 57 se muestran los tipos de datos pertenecientes a las columnas.

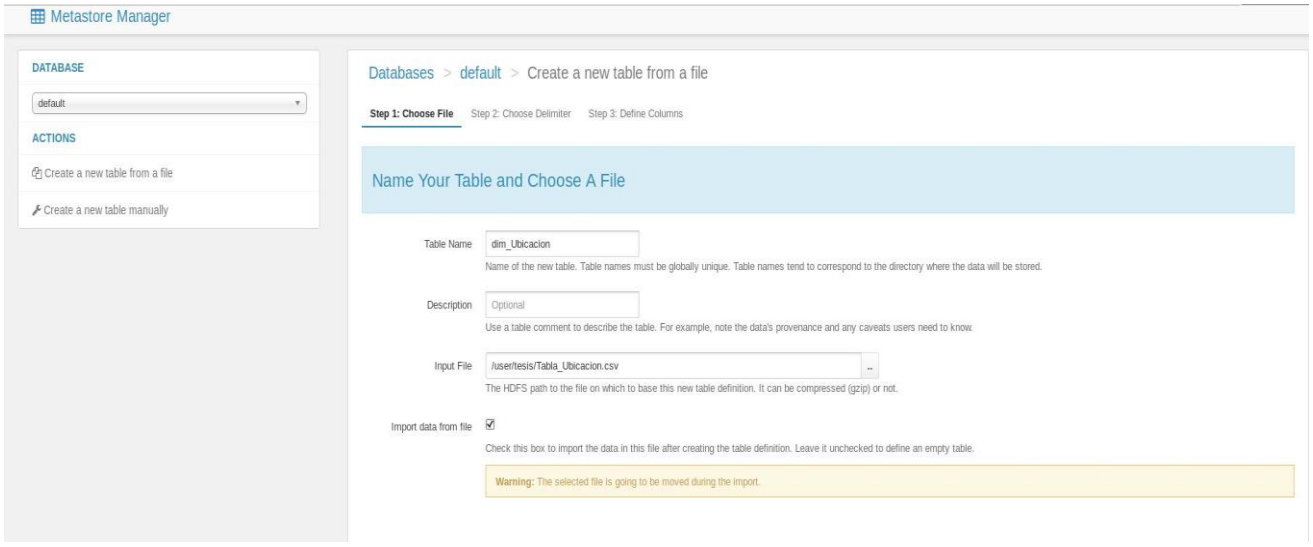


Figura 55: Creación Dimensión Ubicación

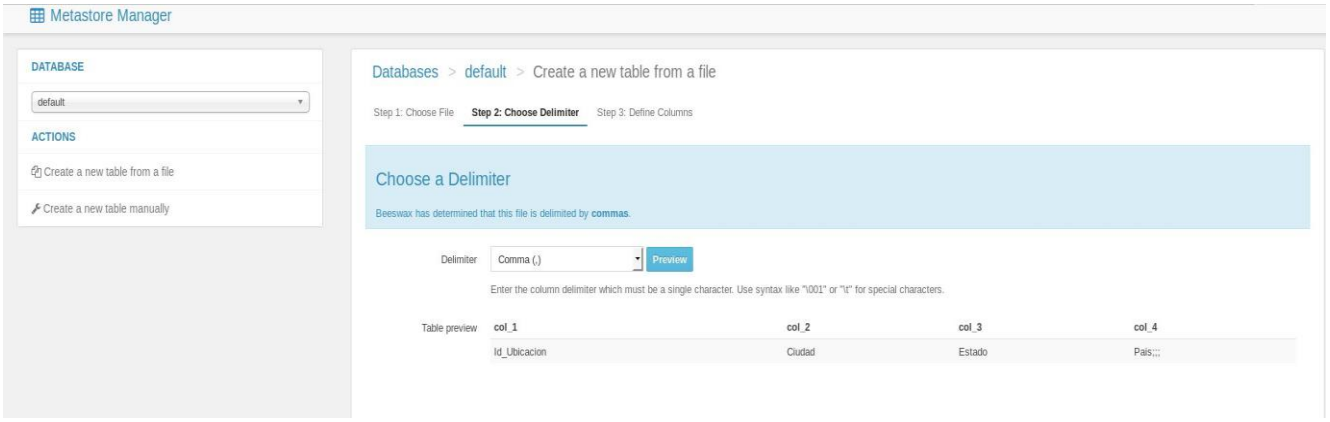


Figura 56: Delimitador de Columnas Dimensión Ubicación

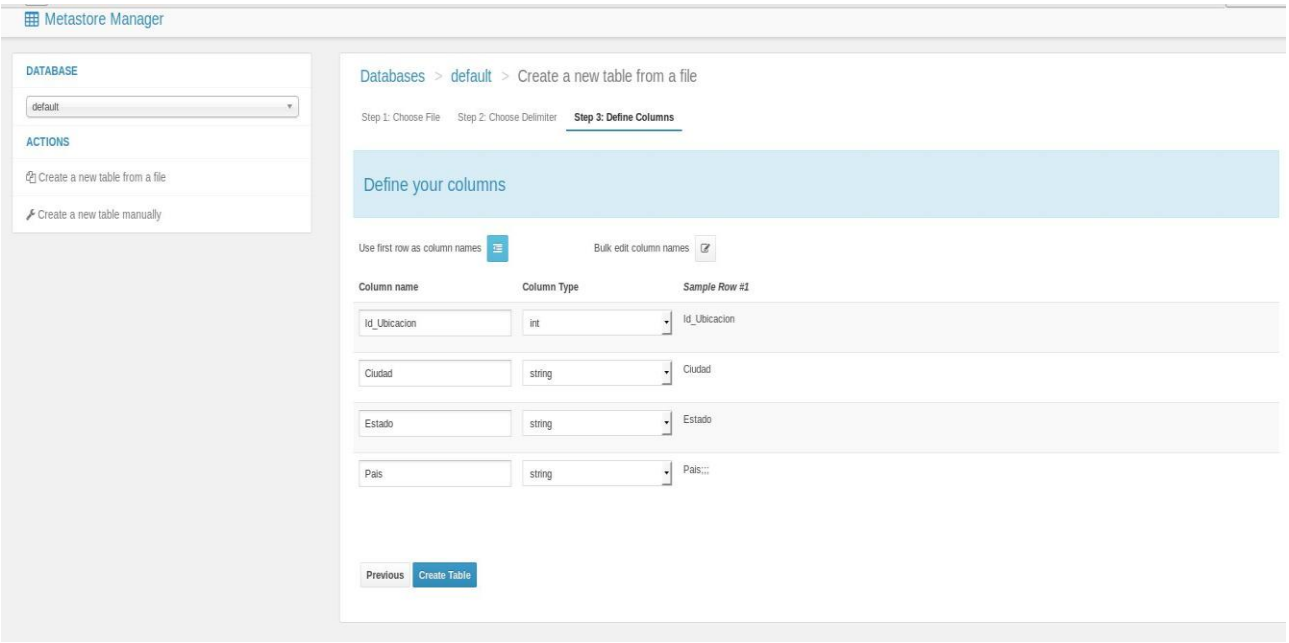


Figura 57: Tipo de Datos Dimensión Ubicación

- Creación de la Dimensión Tiempo

En la figura 58 se muestra la importación del archivo previamente cargado con los datos de la dimensión, en la figura 59 se selecciona el tipo de separador de las columnas del archivo importado y por ultimo en la figura 60 se muestran los tipos de datos pertenecientes a las columnas.

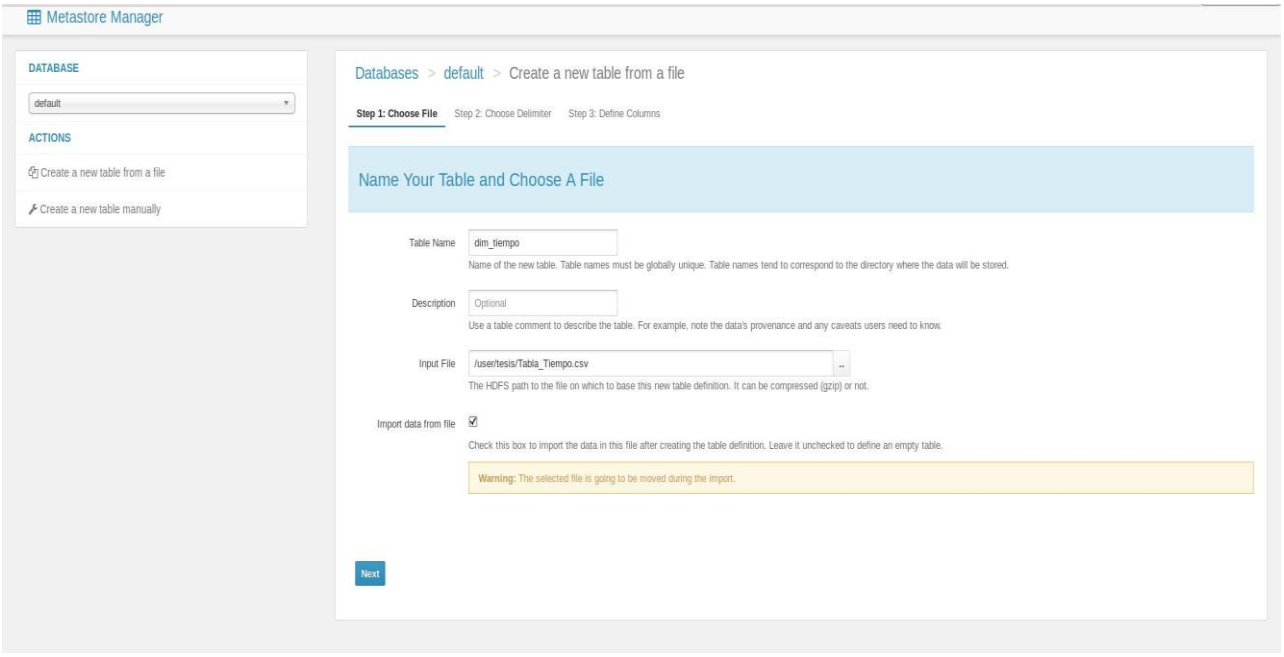


Figura 58: Creación Dimensión Tiempo

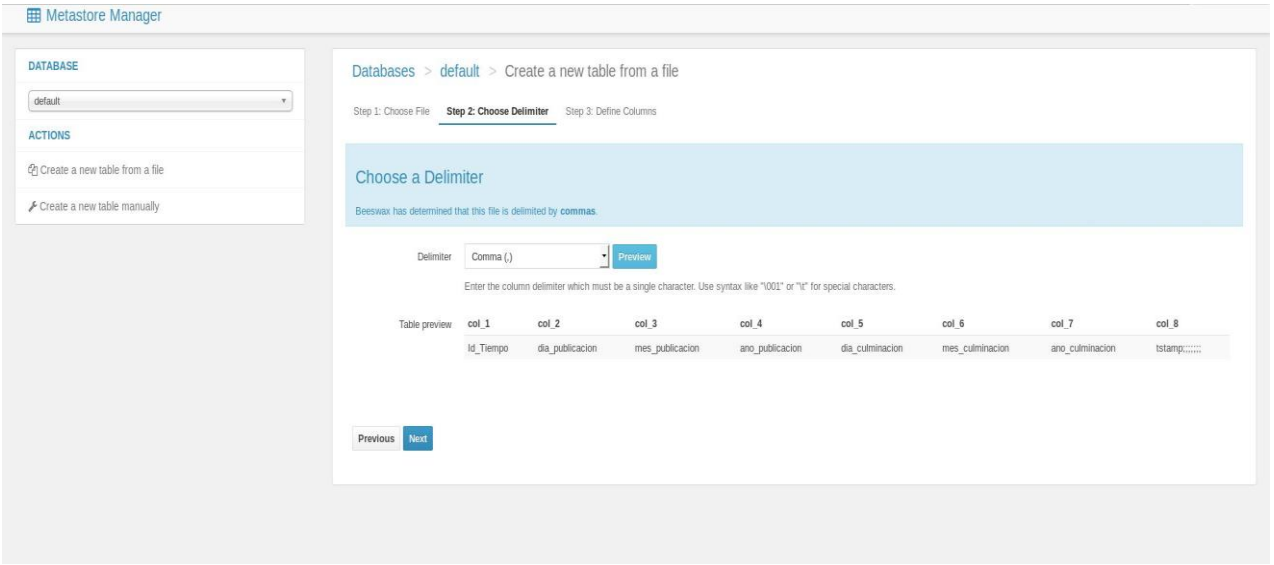


Figura 59: Delimitador de Columnas Dimensión Tiempo

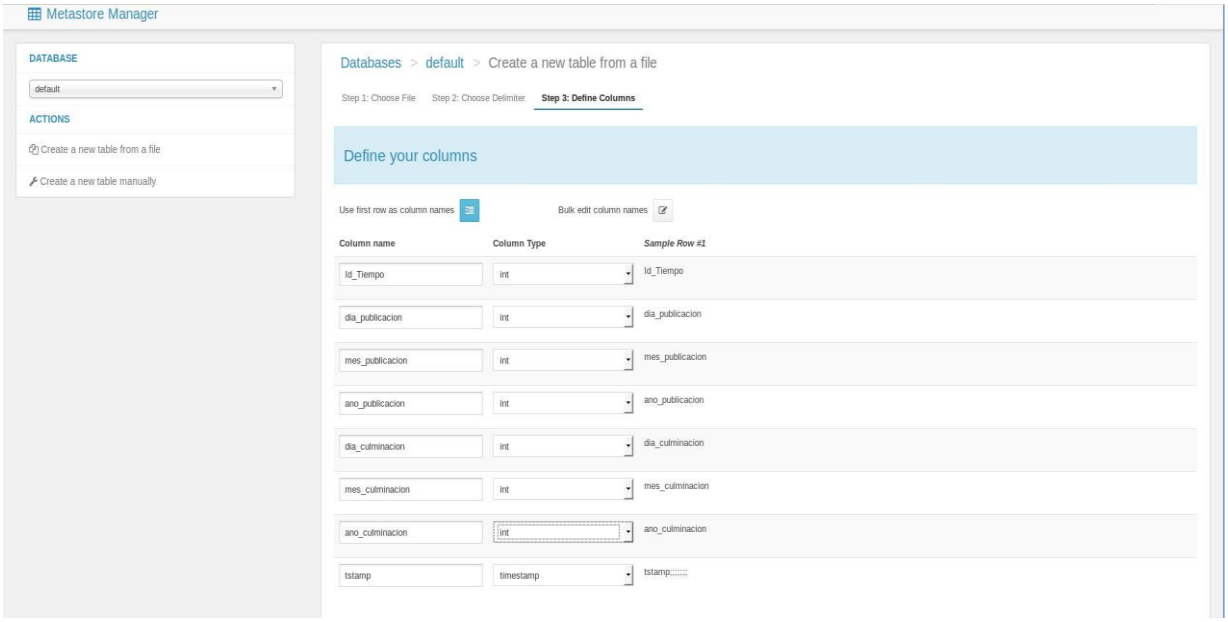


Figura 60: Tipo de Datos Dimensión Tiempo

- Creación de la Tabla de Hechos Bolsa de Empleo

En la figura 61 se muestra la importación del archivo CSV previamente cargado con los datos de la dimensión, en la figura 62 se selecciona el tipo de separador de las columnas del archivo importado y por ultimo en la figura 63 se muestran los tipos de datos pertenecientes a las columnas.

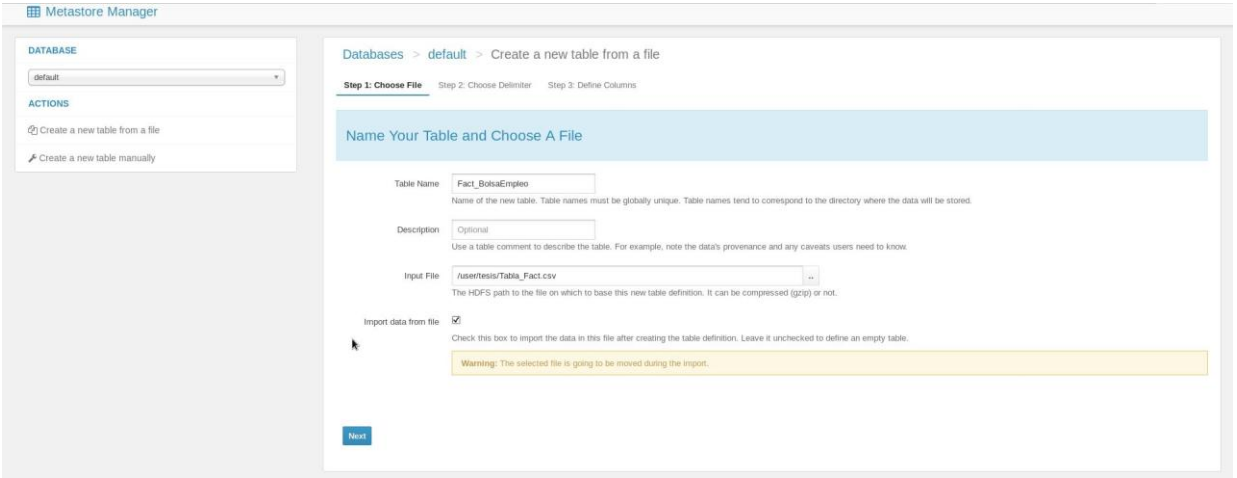


Figura 61: Creación Tabla de Hechos

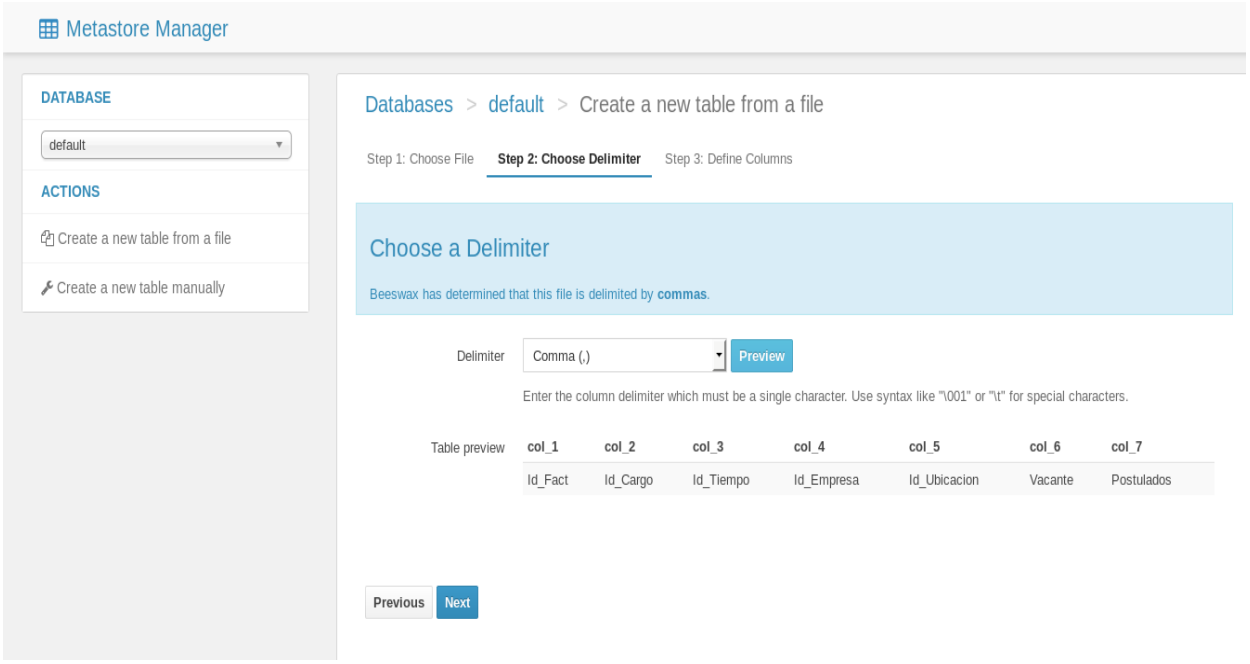


Figura 62: Delimitador de Columnas Tabla de Hechos

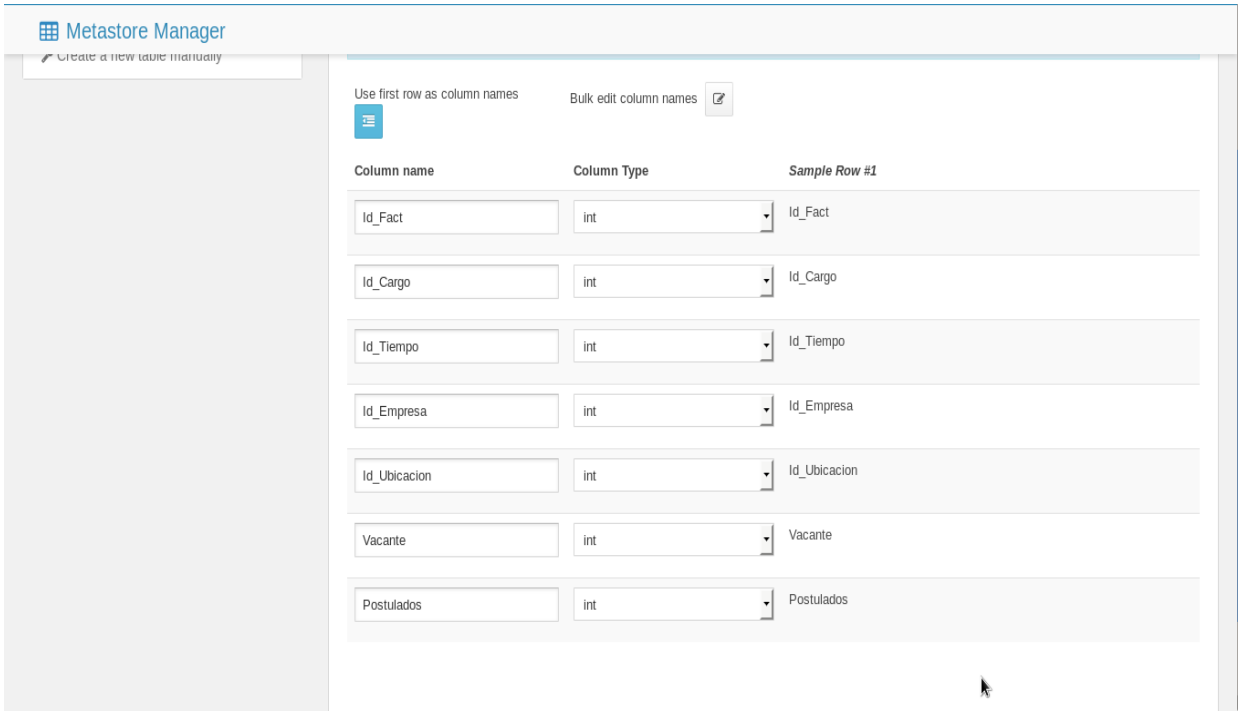


Figura 63: Tipo de datos Tabla de Hechos

4.2.3.2. Inserción en las Dimensiones y Tabla de Hechos.

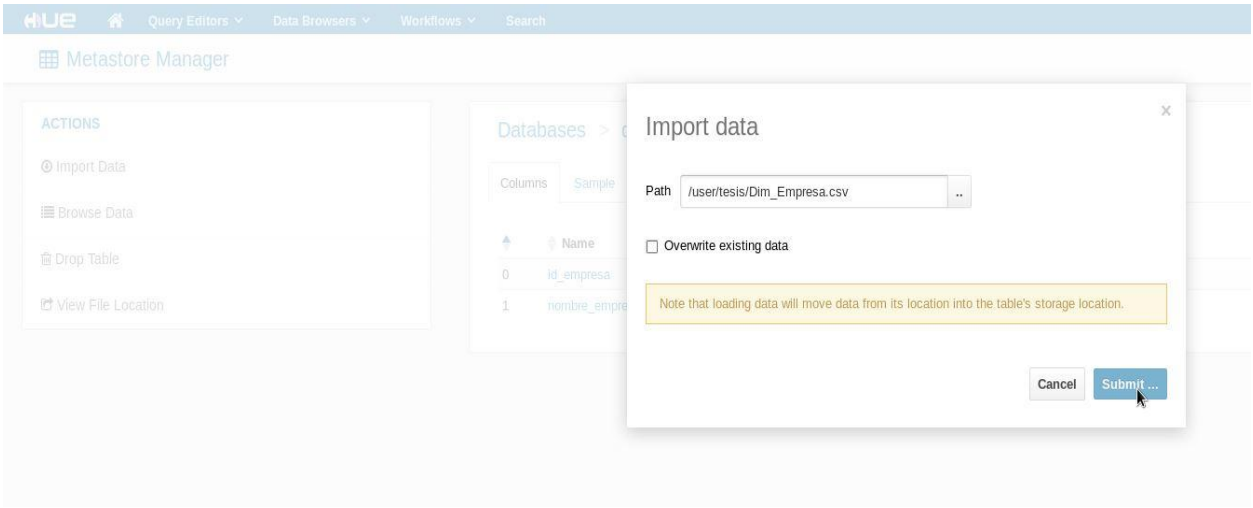
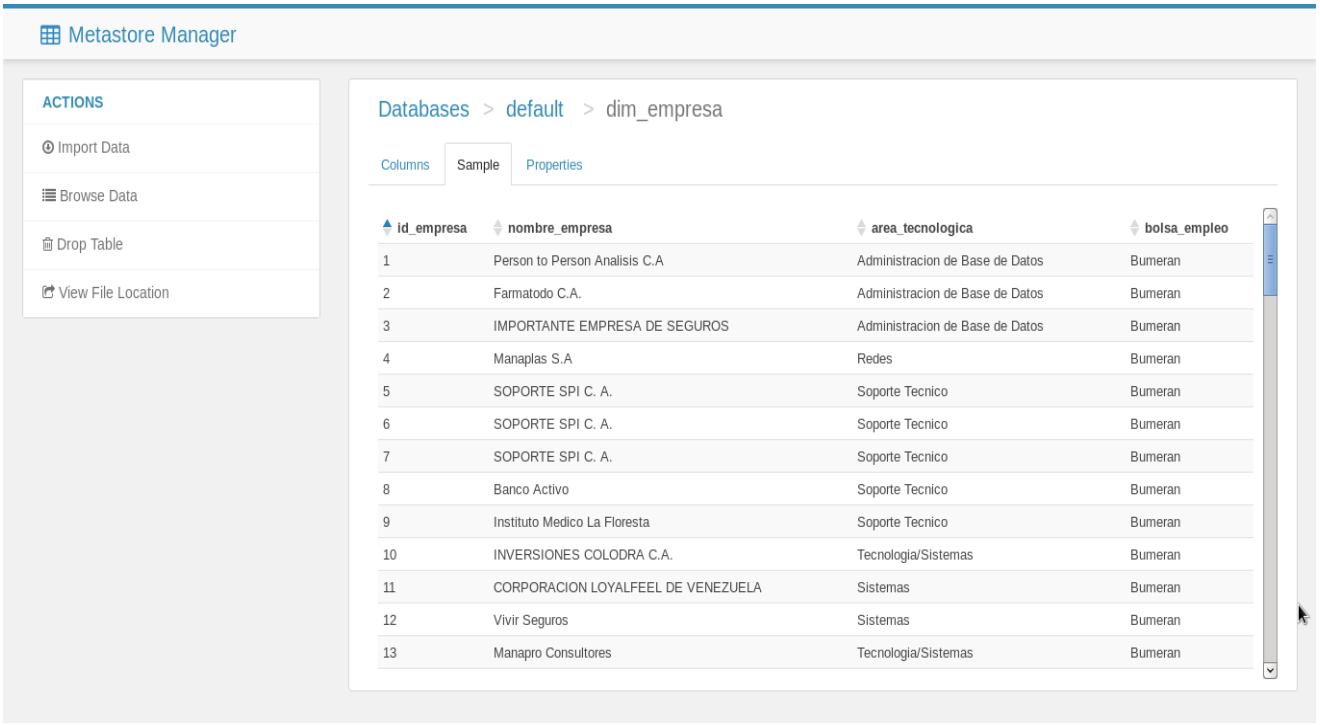


Figura 64: Importar Data Dimensión Empresa



id_empresa	nombre_empresa	area_tecnologica	bolsa_empleo
1	Person to Person Analisis C.A	Administracion de Base de Datos	Bumeran
2	Farmatodo C.A.	Administracion de Base de Datos	Bumeran
3	IMPORTANTE EMPRESA DE SEGUROS	Administracion de Base de Datos	Bumeran
4	Manaplas S.A	Redes	Bumeran
5	SOPORTE SPI C. A.	Soporte Tecnico	Bumeran
6	SOPORTE SPI C. A.	Soporte Tecnico	Bumeran
7	SOPORTE SPI C. A.	Soporte Tecnico	Bumeran
8	Banco Activo	Soporte Tecnico	Bumeran
9	Instituto Medico La Floresta	Soporte Tecnico	Bumeran
10	INVERSIONES COLODRA C.A.	Tecnologia/Sistemas	Bumeran
11	CORPORACION LOYALFEEL DE VENEZUELA	Sistemas	Bumeran
12	Vivir Seguros	Sistemas	Bumeran
13	Manapro Consultores	Tecnologia/Sistemas	Bumeran

Figura 65: Visualización del contenido insertado en la Dimensión Empresa

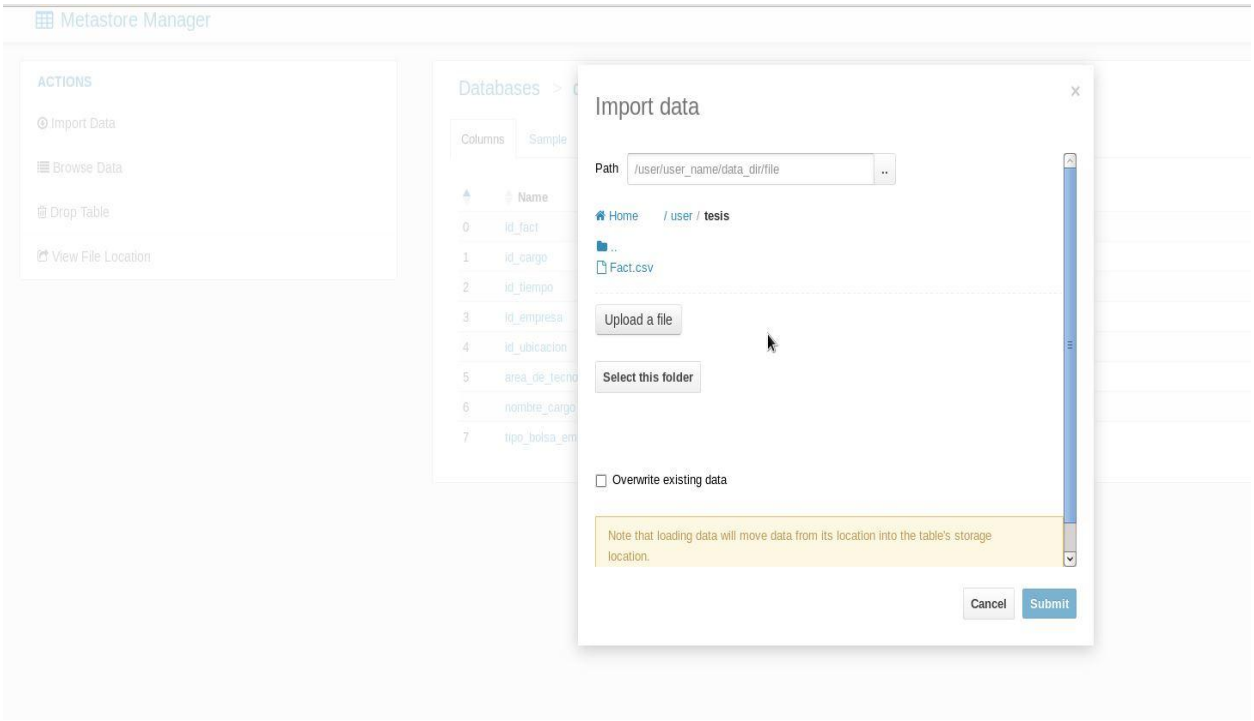
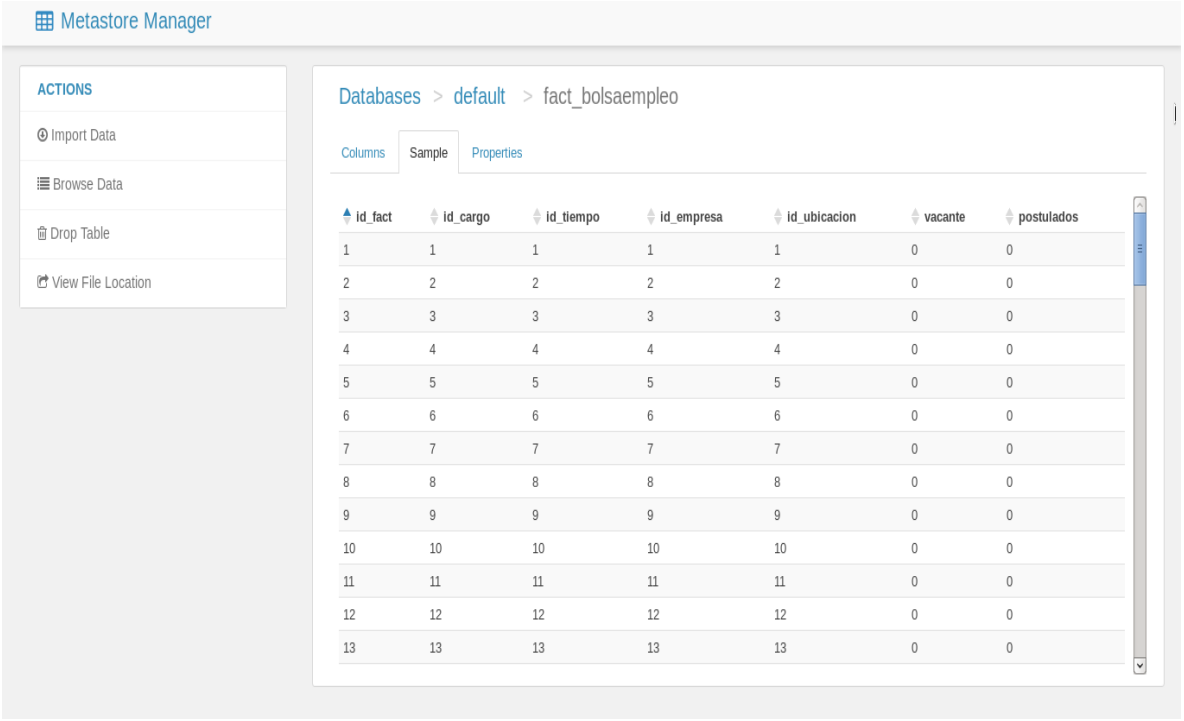


Figura 66: Importar Data Tabla de Hechos



id_fact	id_cargo	id_tiempo	id_empresa	id_ubicacion	vacante	postulados
1	1	1	1	1	0	0
2	2	2	2	2	0	0
3	3	3	3	3	0	0
4	4	4	4	4	0	0
5	5	5	5	5	0	0
6	6	6	6	6	0	0
7	7	7	7	7	0	0
8	8	8	8	8	0	0
9	9	9	9	9	0	0
10	10	10	10	10	0	0
11	11	11	11	11	0	0
12	12	12	12	12	0	0
13	13	13	13	13	0	0

Figura 67: Visualización del contenido insertado en la Tabla de Hechos

Una vez que el contenido se encuentra ya cargado en las dimensiones y en la tabla de hecho, se procede a utilizar herramientas analíticas para obtener valor a toda la cantidad de información que ha sido almacenada y a su vez proveer información al área de gestión de talento humano relacionados a los cargos en el área de TI.

La solución de big data implementada, es utilizada por el área de gestión de talento humano, con el fin de obtener una herramienta que apoye a la fase de reclutamiento y selección en el área de TI.

Esta solución se enfoca en obtener indicadores, generados de data no estructurada, relacionados a las publicaciones de las organizaciones en las bolsas de empleo tales como Empléate y Bumeran. Basados principalmente en capturar la mayor cantidad de información de aquellas organizaciones relacionadas a Tecnología de la Información. A partir de esta captura de información de las bolsas de empleo, se puede obtener indicadores que permitan al área de gestión de talento apoyar directamente a la fase de reclutamiento.

Para facilitar al área de gestión de talento humano la obtención de información rápida, precisa y de calidad para apoyar a la fase de reclutamiento en el área de TI específicamente, se definen algunas variables cuantitativas a utilizar:

- Cargos más solicitados y buscados
- Empresas con mayor cantidad de postulados
- Bolsa de empleo más demandada
- Rango de edad más solicitado por las empresas

Asimismo, para poder desarrollar los indicadores y reportes referentes a los requerimientos contemplados para la solución de inteligencia de negocios propuesta, es necesario especificar cómo se puede obtener lo deseado. Partiendo de lo anterior, en la Tabla 4 se realiza un análisis de los indicadores, especificando las fórmulas utilizadas para calcularlos.

Las unidades de medida utilizadas fueron Cantidad (#) y Cantidad (%):

Tabla 4: Descripción de Indicadores

Proceso de Negocio	Nombre del Reporte	Nombre del Indicador	Forma de Cálculo	Unidad de Medida	Frecuencia de Medición	Criterios de Clasificación	Forma de Representación
Evaluación del Mercado Laboral en el área de TI en Venezuela	**Top 5 de Cargos más solicitados y buscados**	Cantidad Postulados	$\sum POSTULADOS$	#	Mensual	Por Año Por Mes Por Nombre Cargo	Dos Gráficos de Barras Apiladas
		Cantidad Vacantes	$\sum VACANTES$	#			
	Top 5 de empresas con mayor cantidad de postulados	Cantidad Postulados	$\sum POSTULADOS$	#	Mensual	Por Año Por Mes Por Nombre Empresa Por Nombre Cargo	Gráfico de Barras Apiladas y Tabla Pivote
	Rangos de Edades por los 5 cargos más solicitados	Cantidad de cargos vacantes por Rango de Edad	$\sum VACANTES$	#	Mensual	Por Año Por Mes Por Edad Por Nombre Cargo	Gráfico de Barras
	Tipo de Genero por los 5 cargos más solicitados	Cantidad de cargos vacantes agrupada por Genero	$\sum VACANTES$	#	Mensual	Por Año Por Mes Por Genero Por Nombre Cargo	Gráfico de Barras

	Bolsa de Empleo más demandada por cantidad de vacantes	Cantidad de vacantes agrupados por Bolsa de Empleo	$\sum VACANTES$	#, %	Mensual	Por Año Por Mes Por Bolsa de Empleo	Gráfico de Barras Apiladas y Grafico de Tarta
	Niveles de estudio de los 5 cargos con mayor cantidad de vacantes	Cantidad de vacantes agrupados por Niveles de Estudio	$\sum VACANTES$	#	Mensual	Por Año Por Mes Por Nivel de Estudio Por Cargo	Gráfico de Barras Apiladas
	Estados en Venezuela con Solicitud y Búsqueda de Cargos en TI	Cantidad de postulados y vacantes por estados	$\sum POSTULADOS$	#	Mensual	Por País Por Estado	Mapa
			$\sum VACANTES$	#			

En esta última etapa de la metodología se analizan los indicadores generados por la cantidad de datos previamente trabajada con herramientas como SolR, Nutch, Hive y Pentaho. Para esta etapa se hace uso de la herramienta Oracle Data Visualization Desktop, la cual va a permitir conectar con el repositorio Hive que se encuentra en la máquina virtual Oracle Business Intelligence.

Para empezar, se inicia el Oracle Data Visualization Desktop y se selecciona el tipo de conexión que se desee realizar, tal como se muestra en la figura 68. Una vez que hemos seleccionado el tipo de conexión, en este caso sería desde Base de datos, se colocan todos los datos para hacer posible la conexión directamente con Hive, tal como se muestra en la figura 69. Luego de esto a través de sentencias SQL se extrae todo el contenido almacenado en Hive por dimensión, tal como se muestra en el ejemplo de la figura 70. Por último y no menos importante se realizó el modelo estrella asociado, logrando así la conexión de las dimensiones con la tabla de hechos, tal como se muestra en la figura 71.

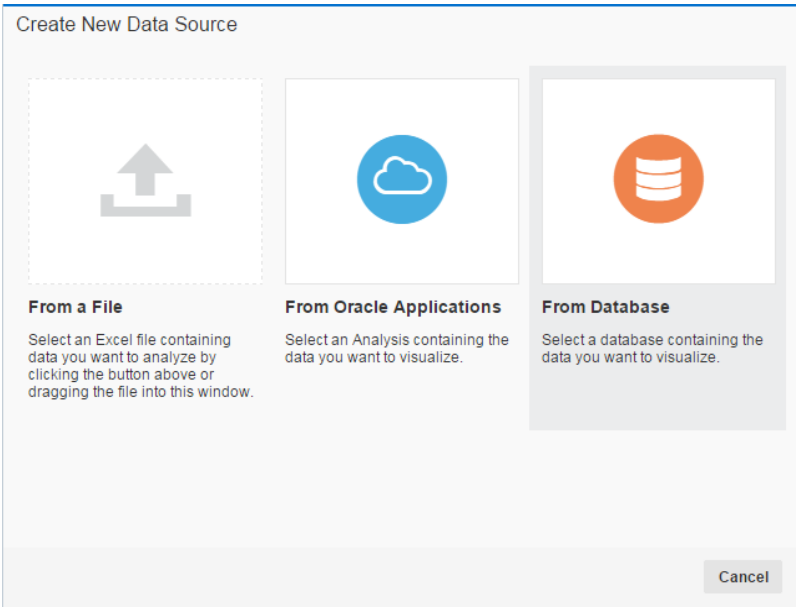


Figura 68: Tipo de Conexión

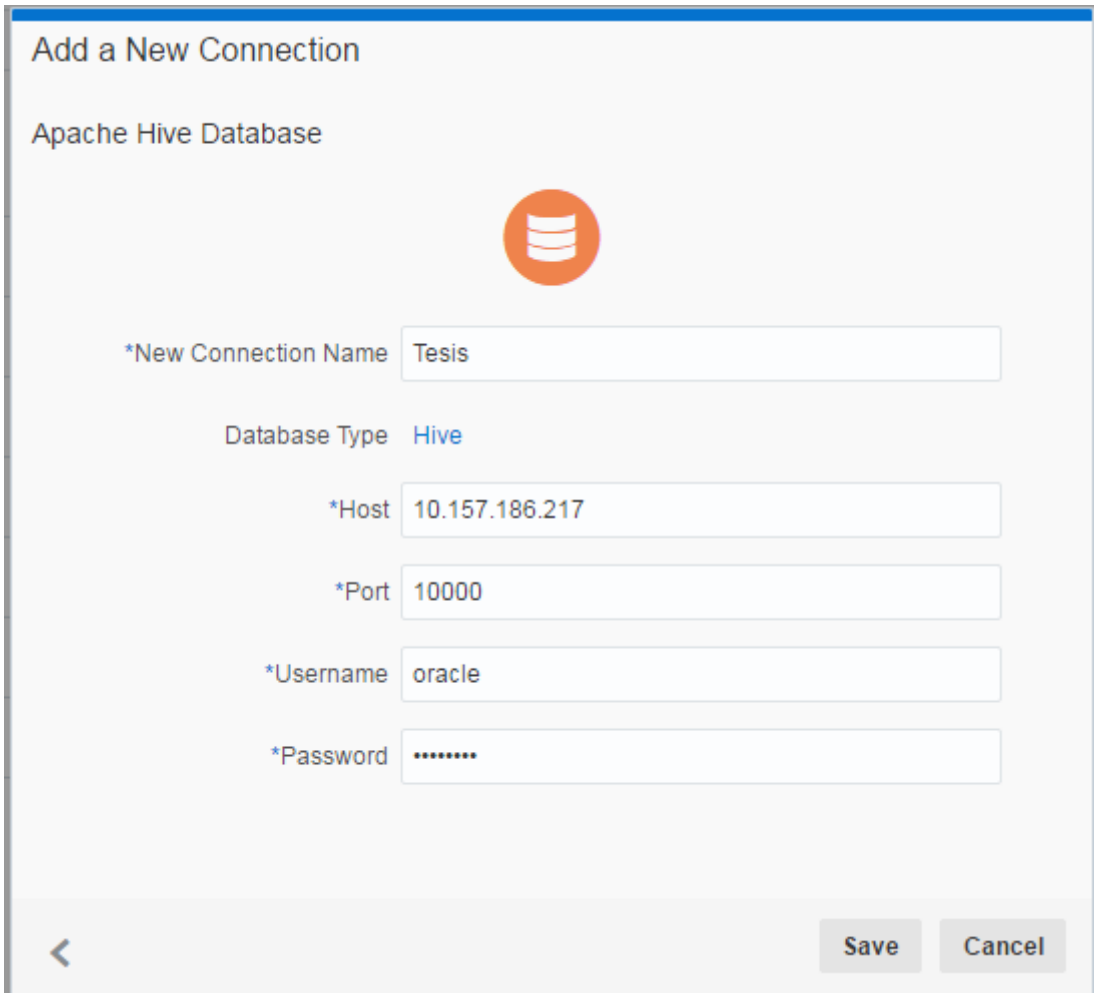


Figura 69: Conexión con Hive

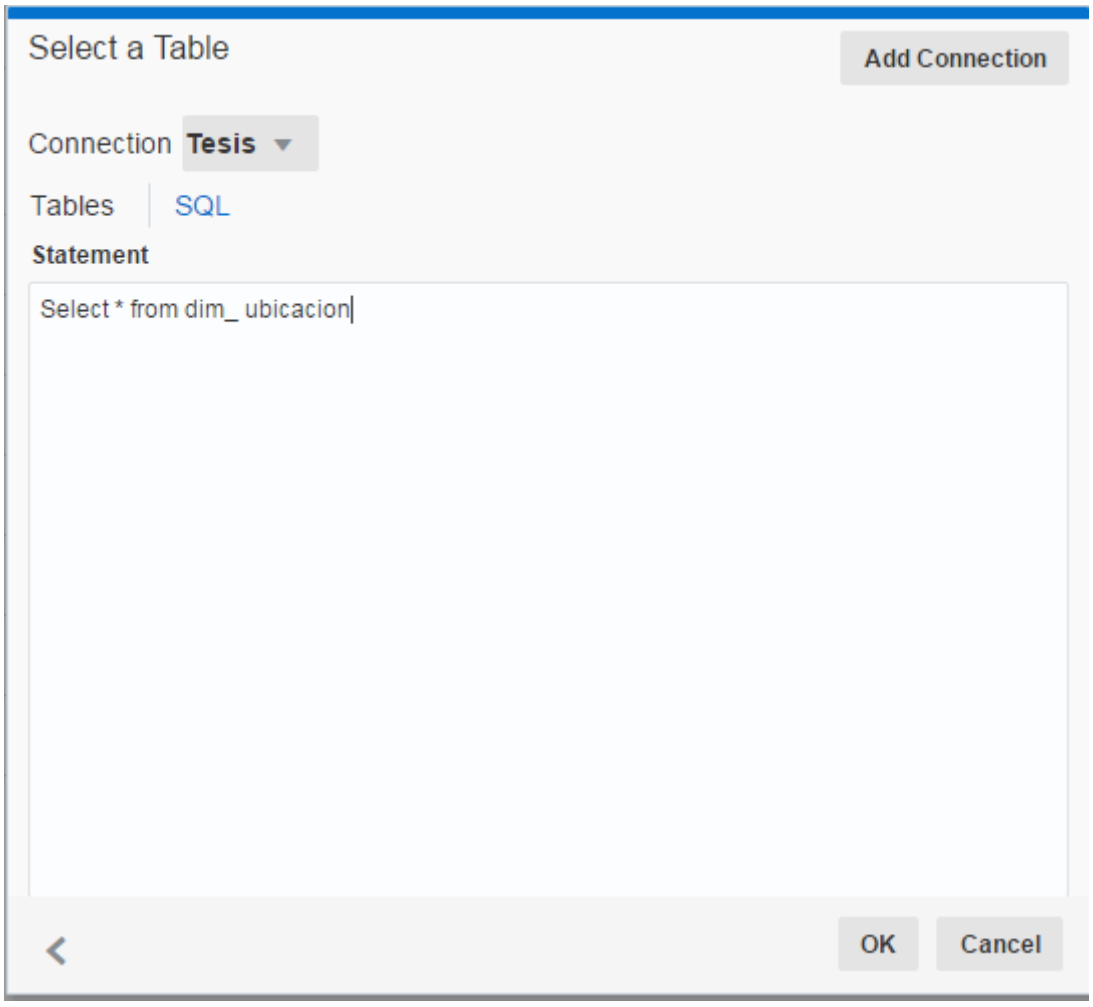


Figura 70: Ejemplo carga de datos de Hive a Oracle Data Visualization Desktop

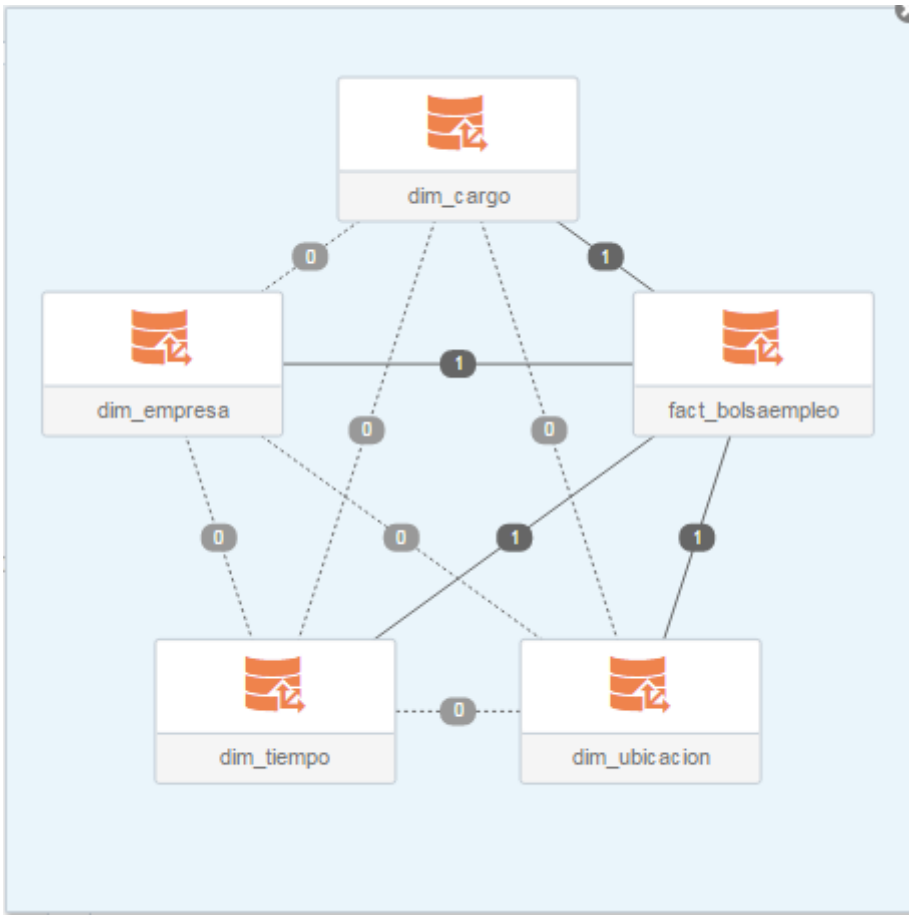


Figura 71: Modelo Estrella

Los reportes que se visualizan a continuación corresponden con los indicadores explicados anteriormente:

1. **Top 5 de Cargos más solicitados:** en la siguiente figura 72 se puede observar el top 5 de los cargos que poseen mayor cantidad de vacantes. Se puede visualizar que el cargo que posee mayor cantidad de vacante y ocupa el primer lugar del Top 5 es el de "*Desarrollador y Programador Web*" con una cantidad de 15 vacantes, mientras que un "*Administrador de Redes*" ocupa el último lugar con 5 vacantes. Esto significa que el cargo más solicitado en las bolsas de empleo en Venezuela por las organizaciones durante el periodo Junio 2016 – Agosto 2016 es el de "*Desarrollador y Programador Web*".

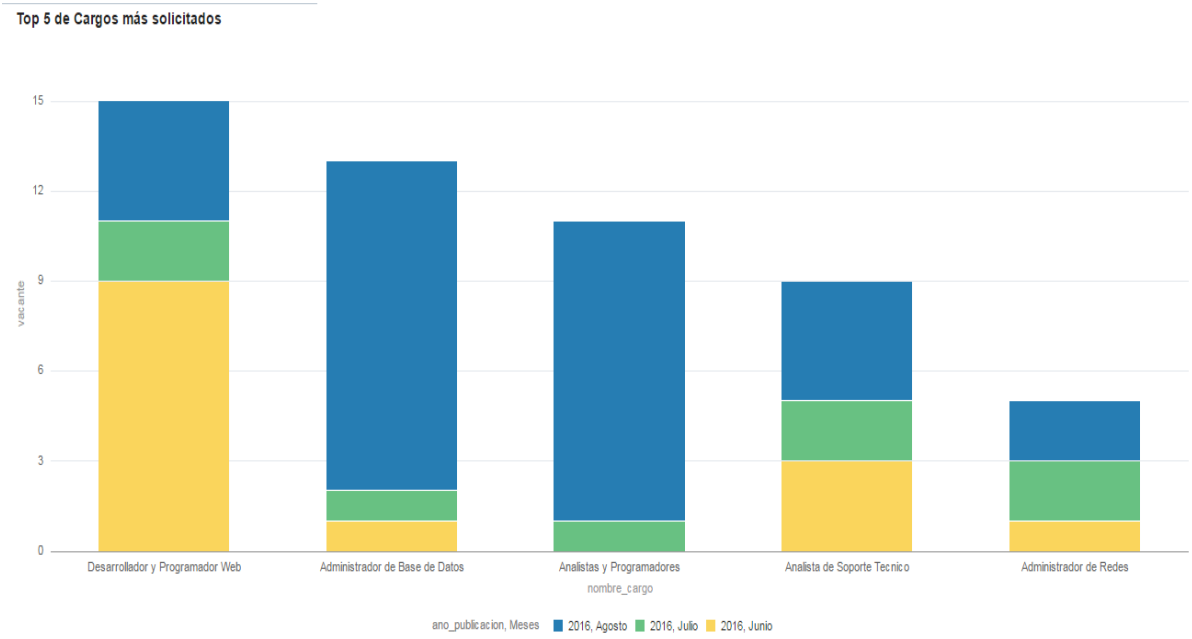


Figura 72: Top 5 de Cargos más solicitados

2. **Top 5 de Cargos más buscador:** en la siguiente figura 73 se puede observar el top 5 de los cargos que poseen mayor cantidad de postulados. Se puede visualizar que el cargo que posee mayor cantidad de postulados y ocupa el primer lugar del Top 5 es el de "*Analista de Soporte Tecnico*" con una cantidad de 53 postulados, mientras que un "*Administrador de Base de Datos*" ocupa el último lugar con 25 postulados. Esto significa que el cargo mas demandado en las bolsas de empleo en Venezuela por las

organizaciones durante el periodo Junio 2016 - Agosto 2016 es el de "*Analista de Soporte Tecnico*".

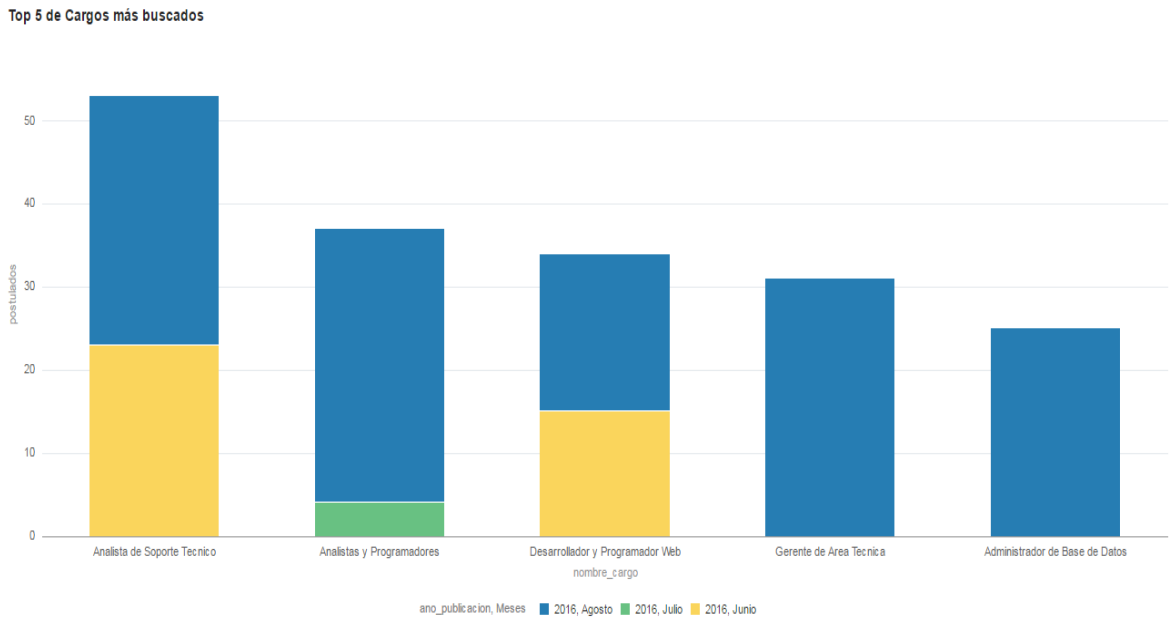


Figura 73: Top 5 de Cargos más buscador

3. **Top 5 de empresas con mayor cantidad de postulados:** en las figuras 74 y 75 se puede observar el top 5 de las empresas con mayor cantidad de postulados. Se puede visualizar que la empresa que posee mayor cantidad de postulados y ocupa el primer lugar del Top 5 es "*Empresa de Tecnología*" con una cantidad de 102 postulados, mientras que un "*Global Uno Logistic de Venezuela*" ocupa el último lugar con 29 postulados. Esto significa que la empresa con mayor cantidad de postulados en las bolsas de empleo en Venezuela durante el periodo Junio 2016 - Agosto 2016 es el de "*Empresa de Tecnología*".

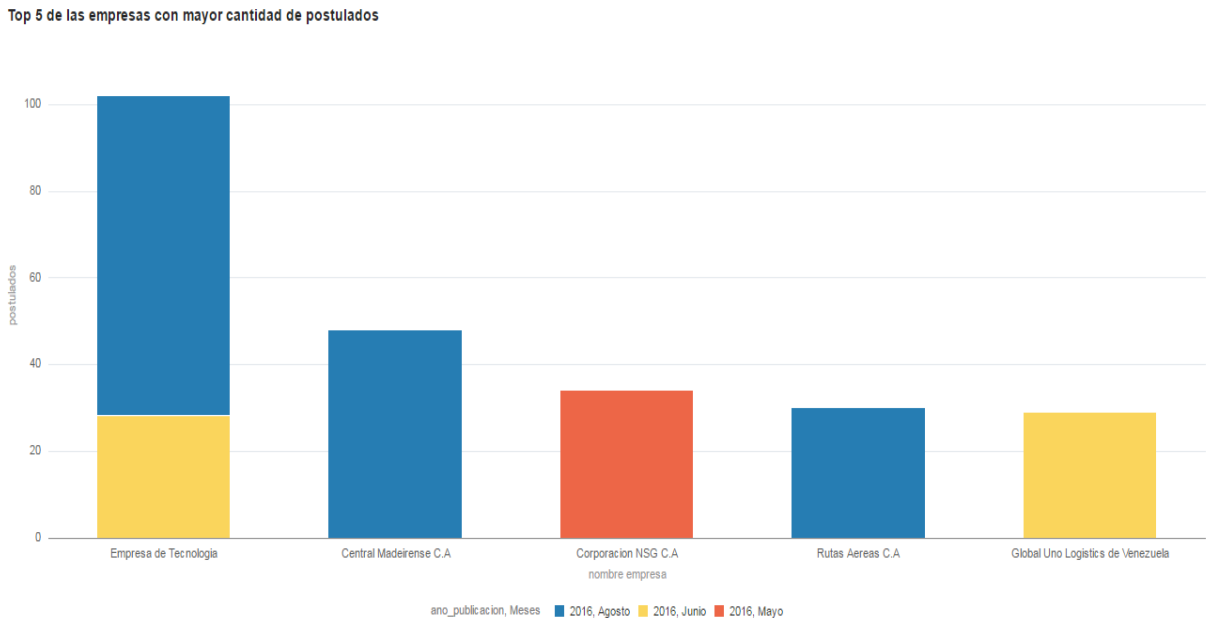


Figura 74: Top 5 de empresas con mayor cantidad de postulados

Top 5 de las empresas con mayor cantidad de postulados

				postulados
IT Regional Operations Manager	Corporacion NSG C.A	2016	Mayo	34
Gerente de Area Tecnica	Empresa de Tecnologia	2016	Agosto	31
Analista de Soporte Tecnico	Rutas Aereas C.A	2016	Agosto	30
Analista Integral de Operaciones	Global Uno Logistics de Venezuela	2016	Junio	29
Coordinador de Tecnologia	Empresa de Tecnologia	2016	Agosto	29
Gerente de Tecnologias	Empresa de Tecnologia	2016	Junio	28
Ingenieros(as)	Central Madeirense C.A	2016	Agosto	19
Auditor (a) de Tecnologia de la Informacion	Central Madeirense C.A	2016	Agosto	15
Administrador de Base de Datos	Central Madeirense C.A	2016	Agosto	14
Consultor de Seguridad de la Informacion	Empresa de Tecnologia	2016	Agosto	14

ano_publicacion, Meses 2016, Mayo 2016, Agosto 2016, Junio

Figura 75: Tabla-Top 5 de empresas con mayor cantidad de postulados

4. **Rangos de Edades por los 5 cargos más solicitados:** en la figura 76 se puede observar los rangos de edades mas demandados por el top 5 de los cargos mas solicitados en TI. Se puede visualizar que para el periodo de Junio 2016 – Agosto 2016 las organizaciones estan en la busqueda de talentos en *"Desarrollador y Programador web"* de edades variadas o "Indistintas". Se podría concluir que ellos no consideran la edad como limitante para laborar como Desarrollador y Programador web .

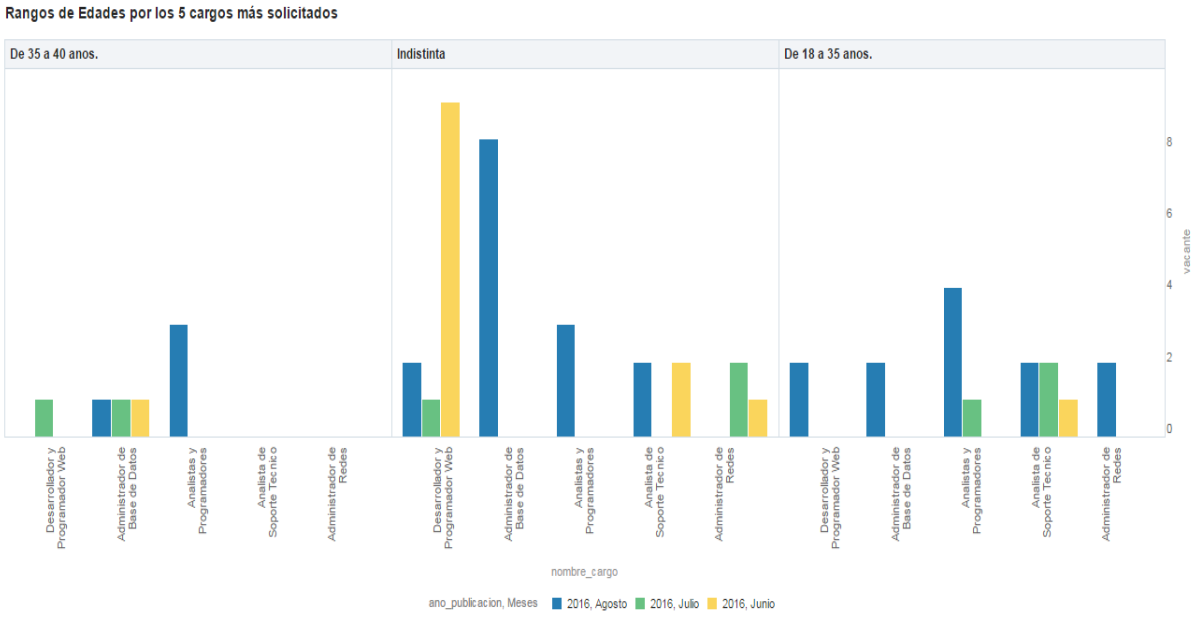


Figura 76: Rangos de Edades por los 5 cargos más solicitados

5. **Tipo de Genero por los 5 cargos más solicitados:** en la figura 77 se puede observar El tipo de genero mas demandado por las organizaciones durante el periodo Junio 2016 – Agosto 2016. Como resultado se tiene que la mayoria de las organizaciones están en la busqueda de talentos de ambos sexos, considerando importante el reclutamiento de los mismos en el mes de Agosto 2016.

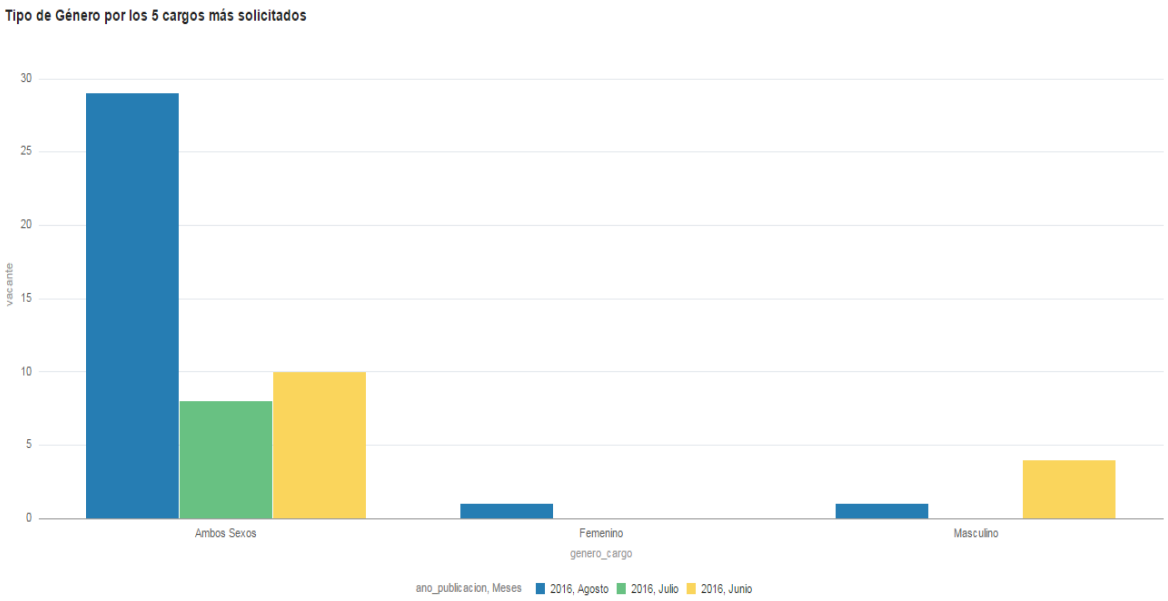


Figura 77: Tipo de Genero por los 5 cargos más solicitados

6. **Bolsa de Empleo más demandada:** en las figuras 78 y 79 se puede observar la bolsa mas usada por las organizaciones para postear sus vacantes. Se puede visualizar que se puede visualizar que la bolsa de empleo que tiene mayor generacion de demanda en el periodo Mayo 2016 - Agosto2016 es Empleate con un 68.8% de postulaciones por las organizaciones, mientras que Bumeran arrojo un resultado de 31.3%.

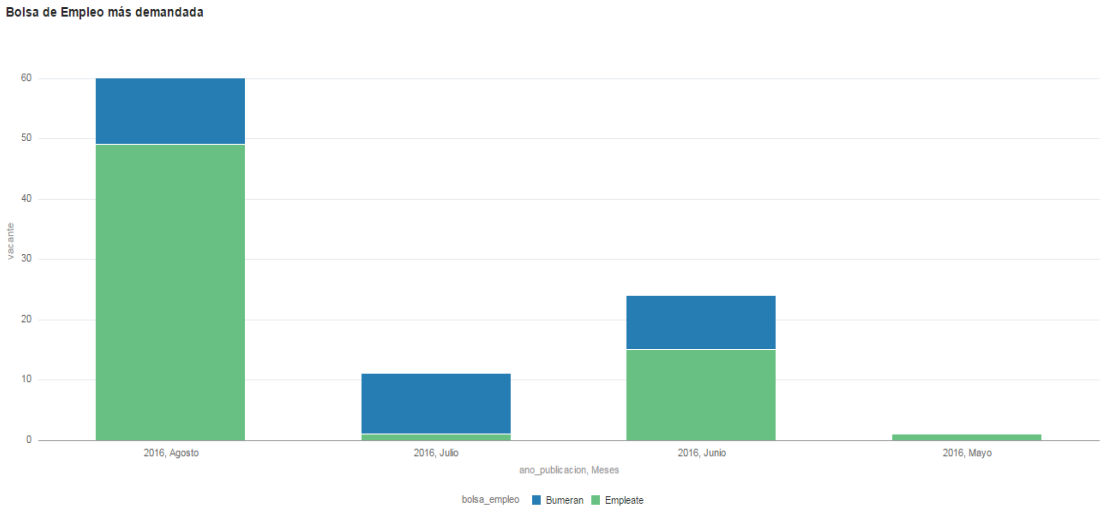


Figura 78: Bolsa de Empleo más demandada

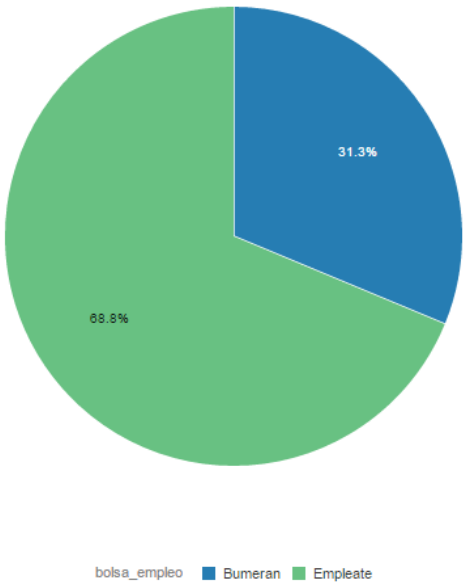


Figura 79:Tarta-Tipo Bolsa de Empleo más demandada

7. **Niveles de estudio de los 5 cargos más solicitados:** en la figura 80 se puede observar el nivel de estudios mas demandado por los top 5 de cargos durante el periodo Junio 2016 – Agosto 2016. Como resultado se tiene que la mayoria de las organizaciones estan en la busqueda de talentos con un nivel de estudios universitarios, no obstante de segundo lugar se considera el nivel de estudio tecnico, seguidamente el bachiller y por ultimo el terciario completo. Considerando importante el nivel de estudio Universitario para el reclutamiento de "Administradores de Bases de Datos" en el mes de Agosto 2016, mientras que en el mismo mes pero para "Analista y Programadores" prefieren el nivel de estudio Técnico .

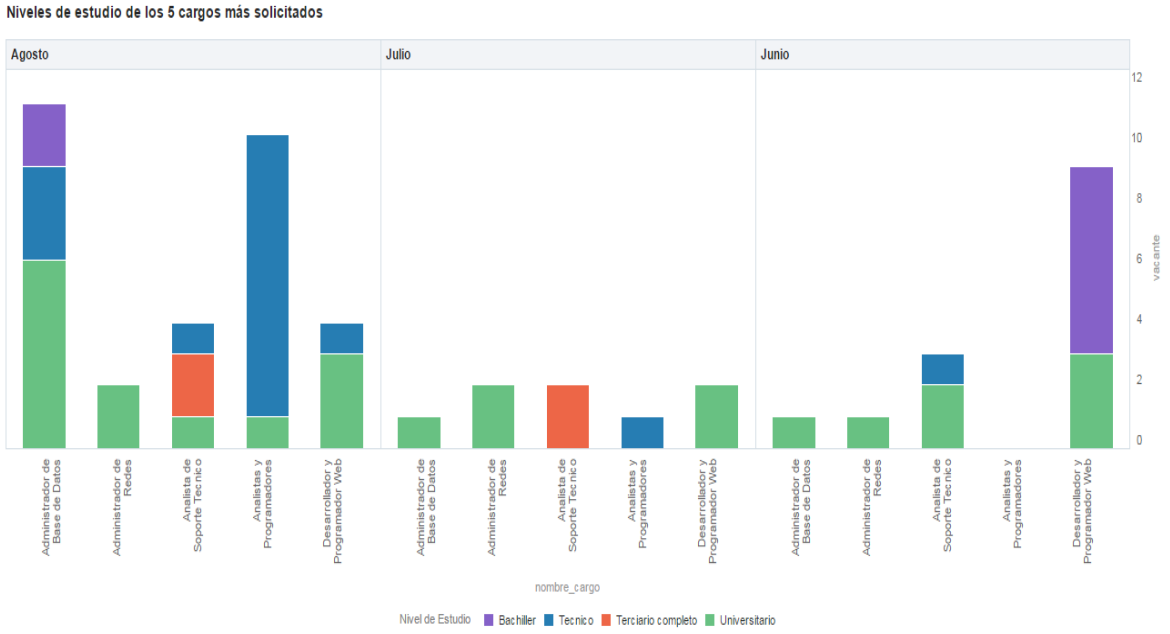


Figura 80: Niveles de estudio de los 5 cargos más solicitados

8. **Estados en Venezuela con Solicitud y Búsqueda de Cargos en TI:** en la figura 81 se puede observar los estados en Venezuela donde se encuentra solicitud y busqueda de vacantes por parte de las organizaciones en el area de TI. Como resultado se tiene que la mayoria de las organizaciones solicitan mayor cantidad de vacantes en el Estado Distrito Capital, mientras que la mayoria de las personas se postulan para empleos en el Estado Miranda.

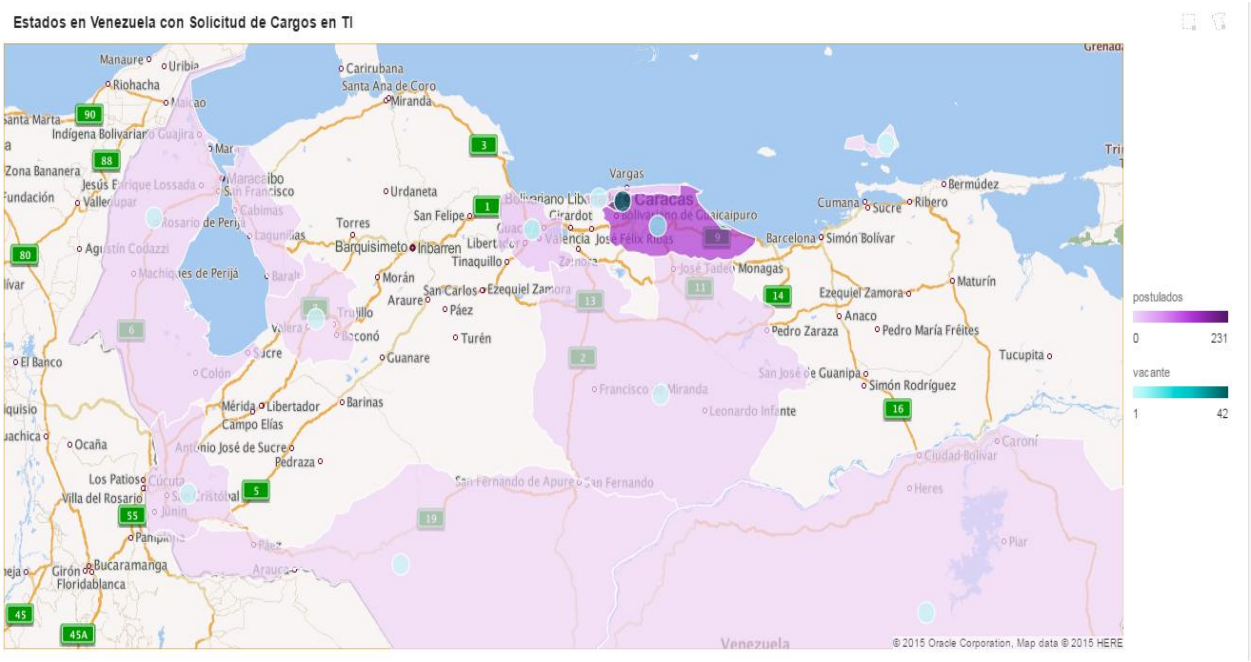


Figura 81: Estados en Venezuela con Solicitud y Búsqueda de Cargos en TI

CONCLUSIONES

Se cumplió satisfactoriamente el objetivo general del presente TEG, el cual consistió en el desarrollo de una solución de big data para apoyar la fase de reclutamiento de la gestión de talento humano en el área de TI, obteniendo así resultados confiables y oportunos para ayudar a la toma de decisiones significativas en base al reclutamiento de empleados para la organización. Se logró la obtención de una solución que cumple con los requerimientos generales planteados al principio del desarrollo de la misma.

En primera instancia, se realizó con éxito la extracción de todo el contenido posible de las organizaciones que solicitan vacantes a través de las bolsas de empleo tales como Empléate y Bumeran, específicamente del área de Tecnología de la Información.

Posteriormente, con el fin de desarrollar una solución de big data se utilizó una adaptación la metodología propuesta por (Krishnan, 2013) , apoyando todo el proceso de desarrollo iniciando con la transformación y manipulación de los datos usando Pentaho Data Integration, transformándolos de manera satisfactoria de data no estructurada a data estructurada y seguidamente se logró almacenar en el repositorio Apache Hive. Finalmente, a través de Oracle Data Visualization, se pudo realizar de manera exitosa la integración con Apache Hive, y poder construir en el mismo el modelo estrella para así poder visualizar una cantidad de indicadores resultantes de la data previamente manipulada.

Por medio de la realización de pruebas de aceptación sobre la arquitectura de big data implementada en este TEG se pudo obtener resultados positivos a lo que respecta a la tolerancia a fallas de ambas máquinas virtuales y la integración con los componentes que conforman la arquitectura de big data que brinda solución a la problemática descrita que presenta el área de gestión de talento humano en la fase de reclutamiento de empleados para la organización.

Una solución de big data como la obtenida en este trabajo genera grandes beneficios para el área de gestión de talento humano, facilitando la fase de reclutamiento en las organizaciones, ya que a través del análisis de los resultados obtenidos se apoya a la toma de decisiones dentro de la organización.

Recomendaciones

Con la finalidad de extender las funcionalidades creadas en esta solución de big data, se proponen las siguientes recomendaciones para trabajos futuros:

- Expandir la conexión de la araña web con otras bolsas de empleo, logrando así poder obtener mayor cantidad de fuentes de datos.
- Realizar la continuación de esta solución, con capacidades superiores en disco y memoria, logrando de esta manera que la información extraída de las bolsas de empleo sea abundante. Además permitir que el acceso a los datos y la exportación de los mismos se realice de manera más rápida y eficiente.

REFERENCIAS BIBLIOGRÁFICAS Y DÍGITALES

Acosta, A. (2011). *AgilUS: Construcción ágil de la Usabilidad.* Recuperado el 01 de Junio de 2015, de http://www.ciens.ucv.ve:8080/genasig/sites/interaccion-humano-comp/archivos/234_CLEI_Acosta_Paper.pdf

Acosta, A. (2011). *AgilUs: Un método ágil de desarrollo de software que incorpora la usabilidad.* Obtenido de Universidad Central de Venezuela: http://www.ciens.ucv.ve/escueladecomputacion/documentos/archivo/121

Alarcón, V. F. Desarrollo de sistemas de información: una metodología basada en el modelado.

Alcaraz, F., Espín, A., Martínez, A., & Alarcón, M. (1 de Octubre de 2006). *Revista Clínica Médica Familiar.* Recuperado el 15 de Junio de 2015, de Diseño de Cuestionarios para la recogida de información: metodología y limitaciones: http://www.revclinmedfam.com/PDFs/06409663226af2f3114485aa4e0a23b4.pdf

Alter, S. (1996). *Information Systems: A Management Perspective. 2nd Edition. .* California: The Benjamin / Cummings Publishing Company.

Apache Nutch. (2014). Obtenido de http://nutch.apache.org/

Argentina, M. d. (03 de Marzo de 2011). *Ministerio de Desarrollo Social de Argentina.* Recuperado el 03 de Junio de 2015, de Centro de Estudiantes: http://www.desarrollosocial.gob.ar/Uploads/i1/biblioteca/32.pdf

Balza. (2010). *Educación, Investigación y Aprendizaje. Una Hermeneusis desde el Pensamiento Complejo y Transdisciplinario.* Guárico: Gremial.

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., y otros. (2001). *Manifiesto por el Desarrollo Ágil de Software.* Recuperado el 21 de Abril de 2015, de http://agilemanifesto.org/iso/es/

Belén, M. I. (2015). *Apoyo administrativo a la gestión de recursos humanos.* España: Paraninfo.

Beltrán Jaramillo, J. M. (2006). *Indicadores de Gestión: Herramientas para lograr la competitividad. 2da Edición.* 3R Editores.

Bonnefoy, J. C. (Marzo de 2006). *Indicadores de Desempeño.* Obtenido de CEPAL (Comisión Económica para América Latina y el Caribe): http://www.cepal.org/ilpes/noticias/paginas/2/23992/Indicadores%20de%20Desempe%C3%B1o.pdf

Cano. (2007). *Business Intelligence: Competir con Información.*

Canós, J., Letelier, P., Sanchéz, E., & Penadés, C. (12 de Noviembre de 2003). *Ingeniería del Sofware y Sistemas de Información.* Recuperado el 05 de Junio de 2015,

de Metodologías Ágiles en el Desarrollo de Software: issi.dsic.upv.es/archives/f-1069167248521/actas.pdf

Caracheo, E. (28 de Enero de 2015). *Metodología del desarrollo para sistemas de información basados en Web.* Obtenido de http://ri.uaq.mx/handle/123456789/2394: http://ri.uaq.mx/bitstream/123456789/2394/1/RI001928.pdf

Casillas, S., & Perez, M. (s.f.). *Universidad Abierta de Cataluña*. Obtenido de http://ocw.uoc.edu/computer-science-technology-and-multimedia/bases-de-datos/bases-de-datos/P06_M2109_02151.pdf

Christiansen , R., & Gaete, E. (2014). *Sistemas de tipo OLTP.*

Colombia, S. d. (2014). *SIRAC*. Recuperado el 11 de Junio de 2015, de Indicadores: http://www.sirac.info/hospitales/html/indicadores.asp

Colombia, U. C. (2011). *Consideraciones para el diseño de un sistema de DESARROLLO ACADÉMICO Y EVALUACIÓN DEL DESEMPEÑO de profesores de la Universidad Católica de Colombia.* Bogotá: Dirección de Ediciones de la Universidad Católica de Colombia.

Colomina, E. (28 de Septiembre de 1998). *Universidad de Alicante.* Recuperado el 22 de Mayo de 2015, de Adopción de sistemas de información en las PYME: teoría y evidencia empírica: http://rua.ua.es/dspace/bitstream/10045/3393/1/Colomina%20Climent%2c%20Evaristo.pdf

Comunidad Rails. (2007). *Introducción a Ruby on Rails.* Recuperado el 17 de Abril de 2015, de https://librosweb.es/libro/introduccion_rails/

Coordinación Empresarial. (06 de Octubre de 2014). *El portal de la coordinacion empresarial*. Recuperado el 14 de Septiembre de 2016, de http://www.coordinacionempresarial.com/que-es-una-bolsa-de-trabajo/

Damián. (2010). *¿Qué es CSS3?* Recuperado el 24 de Marzo de 2015, de http://html5.dwebapps.com/que-es-css3/

Data Mining Consulting SAC. (2013). *KD Nuggets*. Obtenido de Los softwares mas usados para el analisis de datos y mineria de datos: http://www.peruanalitica.com/wp-content/uploads/2013/08/Software.jpg

Dutcher, J. (2014). *Berkeley School of Information*. Obtenido de What is Big Data: https://datascience.berkeley.edu/what-is-big-data/

Edwards, J., McCurley, K., & Tomlin, J. (2011). *An adaptive model for optimizing performance of an incremental web crawler. In Proceedings if the 10th international conference on World Wide Web.* ACM.

El Nacional. (9 de Mayo de 2016). Tecnología y ventas, las áreas con mayor demanda de profesionales. *Tecnología y ventas, las áreas con mayor demanda de profesionales* .

Empleos el Universal. (07 de Enero de 2013). *El Universal*. Recuperado el 14 de Septiembre de 2016, de El Universal: http://empleoseu.blogspot.com/2013/01/cuales-seran-los-puestos-de-trabajo-mas.html

Estrada, L. (8 de Junio de 2012). *El Desempeño Docente.* Obtenido de https://upload.wikimedia.org/wikipedia/commons/d/dd/IMPORTANCIA_DEL_DESEMPE%C3%91O_DOCENTE.pdf

Expansión.com. (26 de Abril de 2016). *Transformación Digital, por Oracle*. Obtenido de Especiales: http://www.expansion.com/especiales/oracle/noticias/2016/04/26/571f3154268e3e71638b4609.html

Factor Humano Formación. (2014). *Factor Humano Formación Escuela Internacional de Postgrado*. Obtenido de http://factorhumanoformacion.com/big-data-ii/

Ferrer, S. M. (2015). La pirámide de los diferentes tipos de sistemas de información. http://pertutatis.cat/la-piramide-de-los-diferentes-tipos-de-sistemas-de-informacion/.

Gartner. (2012). *The Importance of Big Data*.

Gastélum, G. S., & Campas, C. Q. (Junio de 2009). *Diseño de una bolsa de trabajo en una universidad.* Recuperado el 14 de Septiembre de 2016, de http://www.itson.mx/publicaciones/pacioli/Documents/no63/4b-siseno_de_bolsa_de_trabajo_modificado_2.pdf

Giacomelli, R. T. (Septiembre de 2009). *Las tecnologías de información y su aplicabilidad en el proceso de reclutamiento y seleccion.* Obtenido de http://www.spentamexico.org/v4-n2/4%282%29%2053-96.pdf

GNU operating System. (2010). *GNU Wget*. Obtenido de https://www.gnu.org/software/wget/

Grande, E., & Fernández, A. (2000). *Fundamentos y Técnicas de Investigación Comercial, 5ª ed.* Madrid: Esic.

Grant, N. (2003). El liderazgo de los estudiantes hoy. *New Horizon For Learning* .

Gravitar. (12 de Octubre de 2014). *Gravitar - Información sin límites*. Recuperado el 22 de julio de 2015, de Pentaho : http://gravitar.biz/pentaho/

Hadoop. (30 de Agosto de 2013). *Who we are. Apache Hadoop*. Obtenido de http://hadoop.apache.org/who.html

Hernández, G. A., Rodríguez, L. C., Mesa, S. D., & García, A. M. (s.f.). *Los 100 IFJ*. Obtenido de http://los100ifj.weebly.com/orientacioacuten-laboral.html

Hernández, G. (11 de Septiembre de 2011). *Slideshare.* Recuperado el 16 de Junio de 2015, de Escala de Likert: http://es.slideshare.net/gabriela_hernandez/escala-de-likert-9182198

Hernández, M., García , S., Abejón, N., & Zazo, M. (19 de Noviembre de 2010). *Universidad Autónoma de Madrid.* Recuperado el 11 de Junio de 2015, de Estudio de Encuestas: https://www.uam.es/personal_pdi/stmaria/jmurillo/InvestigacionEE/Presentaciones/Curso_10/ENCUESTA_Trabajo.pdf

Hive Software Foundation. (2014). *hive.apache.org*. Obtenido de http://www.hive.apache.org

HTTrack Website Copier. (2007). *HTTrack Website Copier - Offline Browser*. Obtenido de http://www.httrack.com/html/index.html

Hurwitz, J., Nugent, A., & Halper, D. (2013). *Big Data for Dummies.*

IDC. (2012). *Big Data: Un Mercado Emergente.* Obtenido de http://www.portalidc.com/resources/ponencias/big_data12/Resumen_Ejecutivo.pdf

Imhoff, C., & Galemmo, N. (2003). *Mastering DataWarehouse Desing: Relational and Dimensional Techniques.* New York: John Wiley & Sons.

Inmon, W. H. (1996). *Building the Data Warehouse.* New York: John Wiley & Sons.

Inquidia Consulting. (03 de Marzo de 2015). *Experimenting with Hive Insert/Update/Delete and Pentaho*. Obtenido de Experimenting with Hive Insert/Update/Delete and Pentaho: http://inquidia.com/news-and-info/experimenting-hive-insertupdatedelete-and-pentaho

Josh, B. (2013). Big Data in Human Resources: A World of Haves And Have-Nots. *Forbes* .

Journal of Telecommunications. (Enero de 2013). *Comparison of existing open-source tools for Web crawling and indexing of free Music*. Obtenido de http://es.scribd.com/doc/123153248/Comparison-of-existing-open-source-tools-for-web-crawling-and-indexing-of-free-Music.

jQuery. (2015). *jQuery user interface.* Recuperado el 24 de Abril de 2015, de https://jqueryui.com/

Juárez, J. J. (10 de Marzo de 2015). *Fundacion Bertelsmann*. Recuperado el 14 de Septiembre de 2016, de https://www.fundacionbertelsmann.org/es/home/orientacion-profesional-coordinada/orientacion-profesional-coordinada/que-es-la-orientacion-profesional/

Kendall, E. J. (2005). *Análisis y diseños de sistemas.* México: Pearson Educacion.

Kimball. (2002). *The Data Warehouse Toolkit.* Wiley Computer Publishing.

Kimball, & Caserta. (2008). The Data Warehouse ETL Toolkit (2da Ed.).

Kimball, R. (1998). *The Data Warehouse Lifecycle Toolkit.* New York: John Wiley & Sons.

Kimball, R. (2008). *The Data Warehouse Lifecycle Toolkit. 2nd Edition.* New York: John Wiley & Sons.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit.* New York: John Wiley & Sons.

Krishnan, K. (2013). *Data warehousing in the age of big data.* USA: Elsevier.

Laudon, K., & Laudon, J. (2004). *Sistemas de Información Gerencial (Versión en Español).* México: Prentice Hall.

Laudon, K., & Laudon, J. (2004). *Sistemas de Informacion Gerencial.* México: Pearson Prentice Hall.

López, M. J. (2003). *La Selección de Personal.* Madrid: FC Editorial.

Marcos, B. I. (2015). *Apoyo administrativo a la gestión de recursos humanos.* Madrid: Paraninfo.

Mario. (8 de Agosto de 2014). *https://sitiobigdata.com/que-es-mapreduce/*. Obtenido de https://sitiobigdata.com/que-es-mapreduce/

Martinez, R. (2010). *Portal en español sobre Postgresql*. Obtenido de http://www.postgresql.org.es/

Méndez del Río, L. Más allá del Business Intelligence: 16 experiencias de éxito.

Metodología Kimball. (30 de enero de 2014). Obtenido de Inteligencia de Negocio: http://inteligenciadenegociosval.blogspot.com/2014/01/metodologia-de-kimball.html

México, A. N. (2004). *Asociación Nacional de Universidades e Instituciones de Educación Superior de México .* México: Colección Biblioteca de la Educación Superior. Serie Investigaciones.

México, S. d. (20 de Septiembre de 2011). *SHCP*. Recuperado el 11 de Junio de 2015, de Indicadores de Desempeño: http://www.shcp.gob.mx/EGRESOS/sitio_pbr/progra_presupuestacion/Paginas/indicadores_des.aspx

Nagel, S. (18 de Noviembre de 2014). *Web Crawling with Apache Nutch.* Obtenido de http://events.linuxfoundation.org/sites/events/files/slides/aceu2014-snagel-web-crawling-nutch.pdf

Olmo, L. (21 de Abril de 2016). *TicBeat*. Obtenido de http://www.ticbeat.com/empresa-b2b/oracle-lanza-data-visualization-herramienta-de-analisis-de-datos/

Oracle. (s.f.). *Creating a Repository Using the Oracle Business Intelligence Administration Tool.* Obtenido de Oracle: http://www.oracle.com/webfolder/technetwork/tutorials/obe/fmw/bi/bi1113/biadmin11g_01/biadmin11g.htm#t7

Oracle. (s.f.). *Oracle Database 11g Standard Edition One.* Obtenido de Oracle: http://www.oracle.com/us/products/database/standard-edition-one/overview/index.html

Pacheco, J., Castañeda, W., & Caicedo, H. (2002). *Indicadores integrales de gestión.* Colombia: McGraw-Hill.

Padilla Sierra, G., & Ramos Tejeda, M. Psicología del aprendizaje.

Palacio, J. (16 de Octubre de 2006). El Modelo Scrum.

Pentaho. (2012). *PENTAHO*. Recuperado el 12 de julio de 2015, de http://www.pentaho.com/

Peña. (2006). *Tecnologías de la Informacion.* Mexico.

Pérez Lugo, J. (2002). *Monografías*. Recuperado el 11 de Junio de 2015, de Importancia del Liderazgo Directivo en el Desempeño Docente en la I y II Etapa de Educación Básica: http://www.monografias.com/trabajos13/lider/lider.shtml

RAE. (2014). *Hardware*. Obtenido de http://lema.rae.es/drae/?val=hardware

RAE. (2014). *Software*. Obtenido de http://lema.rae.es/drae/?val=software

Randstad. (27 de Noviembre de 2012). *reclutando*. Obtenido de ¿Cuánto tiempo debe durar un proceso de reclutamiento? : http://www.reclutando.net/%C2%BFcuanto-tiempo-debe-durar-un-proceso-de-reclutamiento/

REPO CODIGO. (7 de Mayo| de 2015). *Importar CSV a Hive*. Obtenido de www.repocodigo.blogspot.com/2015/05/importar-csv-hive.html

Rivadera, G. R. (s.f.). *La metodología de Kimball para el diseño de almacenes de .* Obtenido de http://www.ucasal.edu.ar/htm/ingenieria/cuadernos/archivos/5-p56-rivadera-formateado.pdf

Rivadera, G. (2010). *Ucasal.* Recuperado el 02 de Junio de 2015, de La metodología de Kimball para el diseño de almacenes de datos: http://www.ucasal.edu.ar/htm/ingenieria/cuadernos/archivos/5-p56-rivadera-formateado.pdf

Robbins, S. (2004). *Comportamiento Organizacional.* México: Prentice Hall.

Robbins, S., & DeCenzo, D. (2008). *Supervisión. Quinta Edición.* México: Prentice Hall.

Robles, J. (8 de Julio de 2010). *Generador de Encuestas.* Obtenido de https://prezi.com/ayachz5agskd/generador-de-encuestas/

Ruby Org. (2015). *Sitio en español de la organización Ruby*. Recuperado el 22 de Marzo de 2015, de https://www.ruby-lang.org/es/

Sánchez, J. (2003). *Manual de referencia de JavaScript.* Obtenido de http://www.jorgesanchez.net/web/javascript.pdf

Sánchez, J. (2012). *Servidores de Aplicaciones Web.* Obtenido de Implantación de Aplicaciones Web en Sistemas Informaticos de Red: http://www.jorgesanchez.net/web/iaw/iaw1.pdf

Scrum. (27 de Julio de 2015). *Scrum*. Obtenido de Acerca de Scrum: https://www.scrum.org/about

Sinnexus. (2007). *Bases de datos OLTP y OLAP*. Obtenido de http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx

Sinnexus. (2012). *Bases de datos OLTP y OLAP*. Recuperado el 25 de Mayo de 2015, de http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx

Stecanella, R., & Bonanata, J. (2015). *Busqueda de opiniones en medios de prensa.* Recuperado el 14 de 09 de 2016, de https://www.fing.edu.uy/inco/grupos/pln/prygrado/InformeProyectodeGradoPrensa.pdf

The R Foundation. (2013). *R-Project*. Obtenido de https://www.r-project.org/about.html

Trupti, U., Ravindra D, K., & Dharmik, R. (2014). *Study of Web Crawler and its Different Types* (Vol. 16). IOSR Journal of Computer Engineering.

Universidad Central de Venezuela, E. (10 de Diciembre de 2011). *Centro de Estudiantes de Computación.* Obtenido de Normativa de Evaluaciones de la Escuela de Computación: https://cecucv.wordpress.com/compilacion-legislativa/normativa-de-evaluaciones-de-la-escuela-de-computacion/

Universidad de Champagnat. (16 de Julio de 2002). *Gestiopolis*. Recuperado el 12 de Junio de 2015, de Encuesta, Cuestionario y Tipos de Pregunta: http://www.gestiopolis.com/encuesta-cuestionario-y-tipos-de-preguntas/

Vauzza. (2013). *Todo lo que necesitas saber sobre Big Data*. Obtenido de Eureka-Startups: http://www.eureka-startups.com/blog/2013/05/28/todo-lo-que-necesitas-saber-sobre-big-data/

Veras, M., & Cuello, C. (2005). *Prácticas de Gestión Humana en la República Dominicana.* República Dominicana: Universidad Intec.