DISEÑO, DESARROLLO E IMPLEMENTACIÓN DE UNA HERRAMIENTA ETL CON BASES DE DATOS NOSQL ORIENTADAS A GRAFOS PARA GENERAR CONSULTAS OLAP EN BIG-DATA

ALEJANDRO PERLAZA VILLALBA
JEIMY ANGEL NEUTA

ESCUELA TECNOLÓGICA INSTITUTO TÉCNICO CENTRAL
PROGRAMA DE INGENIERÍA DE SISTEMAS
BOGOTÁ D.C.
2016

DISEÑO, DESARROLLO E IMPLEMENTACIÓN DE UNA HERRAMIENTA ETL CON BASES DE DATOS NOSQL ORIENTADAS A GRAFOS PARA GENERAR CONSULTAS OLAP EN BIG-DATA

ALEJANDRO PERLAZA VILLALBA
JEIMY ANGEL NEUTA

PROYECTO DE GRADO

Docente Asesor
Ingeniero Rafael Thomas Bohórquez

ESCUELA TECNOLÓGICA INSTITUTO TÉCNICO CENTRAL
PROGRAMA DE INGENIERÍA DE SISTEMAS
BOGOTÁ D.C.
2016

NOTA DE ACEPTACIÓN:

_____
_____
_____
_____
_____
_____
_____

FIRMA DE ASESOR METODOLÓGICO

_____

Bogotá D.C, junio de 2016.

DEDICATORIA

Queremos dedicar este proyecto a nuestras madres, quienes fueron las personas que nos motivaron, nos apoyaron y estuvieron en todo este proceso de comienzo a fin con los buenos y malos momentos. Su infinito amor nos permitirá agregar otro peldaño de éxito a nuestra vida profesional.

A nuestros maestros quienes con su conocimiento y sabiduría nos orientaron durante este proceso aclarando dudas y brindandonos ideas que le aportaron sentido a nuestra idea original.

A nuestros amigos más cercanos, familiares y compañeros de clase que con sus palabras de aliento nos motivaron a seguir adelante.

AGRADECIMIENTOS

Queremos dar un especial agradecimiento a nuestro asesor de tesis Rafael Thomas Bohórquez por su paciencia, dedicación, sus aportes y en especial por su entera disposición y compromiso con este proyecto de grado.

También queremos agradecer a nuestro anterior asesor Fabio Dávila quien también nos brindo sus aportes y conocimiento que en gran medida nos sirvieron de gran ayuda.

A nuestro Decano de la Facultad el Ingeniero Sócrates Amador, infinitas gracias por brindarnos sus consejos y aportes.

Por último y no menos importante queremos agradecer a los demás docentes de la facultad que de alguna manera estuvieron dispuestos a brindarnos su ayuda, consejos y apoyo.

RESÚMEN

El presente documento contiene información acerca del desarrollo de una herramienta ETL que proporciona soluciones para procesar y analizar grandes volúmenes de información (BigData).

El de desarrollo de esta herramienta cuenta con las ultimas tendecias en tecnología tal como lo es el uso de las Bases de Datos NoSQL gráfica con el software Neo4j que permite consultar los datos procesados y visualizar las relaciones en forma de grafos. Otra tecnología que es tendencia en las páginas web, es el Diseño Web Adaptable o tambien conocido en ingles como Responsive Web Design (RWD). La combinacion de estas tecnologías junto con una arquitectura basada en Inteligencia de negocios proporciona una herrramienta interesante y de fácil uso sin necesidad que tengan un conocimiento experto para los usuarios administradores de bases de datos

CONTENIDO

TABLA DE ILUSTRACIONES

LISTA DE TABLAS

LISTA DE ANEXO

GLOSARIO

CSV: Comma-separated values, valores separados por columnas.

INDEXAR: ordenar una serie de datos de acuerdo a un patrón común con el propósito de facilitar su consulta y análisis.

GUI: Interfaz Gráfica de Usuario, conjunto de elementos que permite a un usuario interactuar con una aplicación o entre ella y otros programas.

META-DATA: datos que resumen o describen otros datos.

PERSISTENCIA: propiedad que presentan los datos de permanecer después de haber sido nominalmente eliminados.

PLATAFORMA: sistema capaz de lograr el funcionamiento de determinados módulos con los cuales es compatible.

SGBD (Sistema Gestor de Base de Datos): sistema que permite la definición de una base de datos así como también su modificación, eliminación y análisis.

POSTGRESQL: sistema de gestión de datos relacional orientado a objetos que puede usarse en todos los sistemas operativos.

OLAP: forma de procesamiento analítico que permite al usuario extraer datos y visualizarlos desde distintos puntos de vista.

NEO4J: base de datos estructurada en grafos

JAVA: lenguaje de programación que se orienta a objetos.

MAVEN: herramienta de gestión de proyectos en Java

JAVASCRIPT: lenguaje de programación orientado a objetos con funciones de primera clase.

D3JS: biblioteca de Java script para manipular documentos basados en datos.

INTRODUCCIÓN

El principal objetivo de analizar la información es permitir realizar rápidas búsquedas y obtener un reporte o informe que genere las predicciones de sus acciones para establecer y compartir una visión fiable que servirá de soporte para la toma de decisiones.

Ahora bien la información no es un simple conjunto de datos catalogados y agrupados por coincidencias, la información a la que se hace referencia es más profunda , compleja densa y difícil de procesar, uno de los ejemplos podría ser el proceso inversión de capital en un mercado de riesgo, pero si por algún motivo surgiera un heroico asesor de inversiones que pudiese interpretar el comportamiento de una o varias acciones en el mercado y advertir al incauto comprador que puede quedar en bancarrota luego de un mal movimiento de compra o venta, estos casos son fortuitos pero gracias al tratamiento, procesamiento e interpretación de grandes volúmenes de datos un sistema de información especializado si podría hacerlo, claro está con un grupo de expertos.

El presente proyecto de grado consta de la creación de una herramienta que utiliza los procesos ETL (Extract, Transform and Load) en conjunto con un modelo integral de soluciones BI y la metodología Kimball para el desarrollo de DataWarehose, este modelo se usa normalmente para el análisis de datos en empresas, entidades u organizaciones que manejan grandes volúmenes de información (BigData). El principal objetivo es trabajar con las últimas tecnologías en estos dos campos y lograr tratar una cantidad masiva de datos, que permita analizar la información de una forma interactiva, sencilla e intuitiva. Los procesos ETL inspeccionan o extraen datos de diferentes fuentes, los limpia y transforma en información útil, es decir, información estructurada, con los parámetros necesarios para simplificar el proceso de análisis posterior con algunas de las soluciones o herramientas que se usan en la inteligencia de negocios, así el análisis resulta más preciso y acorde a las necesidades requeridas. También, para el modelado de datos

se usó Neo4j, una herramienta de código libre NoSQL que permite montar una base de datos basada en grafos. Finalmente, la información resultante se visualizará con visor de Neo4j, ofertando una visión clara de los datos existentes y permitiendo interactuar con ellos para obtener la información que se requiera, cuya principal característica es el manejo de enormes cantidades de información.

Con el fin de ajustar los datos al desarrollo de la herramienta, éstos fueron escogidos de forma aleatoria, cumpliendo con un volumen masivo de información. Adicionalmente, que fuera capaz de soportar un proceso ETL, al terminar el proceso la información útil se almacena en una base de datos no relacional en grafo, siguiendo los conceptos de análisis de información para inteligencia de negocios se construyen cubos OLAP, estos cubos permiten ver la información desde diferentes formas al realizar una búsqueda o una consulta según como el usuario lo desee. Es decir muestra diferentes perspectivas de la información que se encuentra ya procesada y almacenada.

Estas herramientas o soluciones de la inteligencia de negocios, son una alternativa más dinámica de mostrar la información almacenada de forma abstracta en una manera más explícita para que un humano experto la interprete, haciendo evidentes los patrones para tomar decisiones en base a la información.

I. PLANTEAMIENTO DEL PROBLEMA

1.1 DESCRIPCIÓN DEL PROBLEMA

Hoy en día el auge de las tecnologías de la información y las comunicaciones (TICS), el crecimiento exponencial de los sitios web, las redes sociales y la infinidad de aplicaciones móviles, los usuarios proveen grandes volúmenes de datos, ya sean personales, comerciales, financieros, administrativos, entre otros. Cada vez se hace más complejo para una persona que no tiene conocimientos suficientes sobre el tema, poder almacenar y analizar de una forma fácil y rápida los datos que maneja su organización o entidad. De allí surge el concepto de BigData (Datos masivos), que no es más que toda aquella información que no puede ser procesada por un sistema o herramienta tradicional debido a su extenso volumen y complejidad.

Ya sea que los usuarios suministren la informacion por medio de las nuevas tecnologías como WhatsApp, correos electrónicos, encuestas, o sea extraida por otros medios tales como transacciones de datos, la biometría, E-marketing y web, es importante para las entidades y las grandes empresas que su infomacion tanto compleja como no tan compleja pueda ser analizada con facilidad y rapidez, incrementando así la productividad y los resultados en las labores cotidianas.

Últimamente se han implementado soluciones de la Inteligencia de negocios que analizan la información y facilitan la toma de decisiones. Comunmente estas soluciones se utilizan en conjunto con un proceso ETL (Extraccion, Transformación y Carga), lo que permite transformar en información útil la cantidad de datos disponible. Si a estos dos proceso se le agrega un motor de bases de datos en grafo que soporte BigData y un entorno agradable al usuario final, se obtendrá una herramienta capaz de organizar y estructurar los datos, soportar grandes volumenes de información, que además los analice y entregue resultados, esto hará

incrementar la productividad de las empresas o entidades con una manera más cómoda para los administradores de la informacioón.

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo desarrollar una herramienta ETL, con el fin de analizar fácilmente datos masivos para los procesos de inteligencia de negocios?

II. JUSTIFICACIÓN

La idea de implementar procesos ETL junto con las bases de datos orientadas a grafos bajo el esquema de las soluciones de inteligencia de negocios y bajo la metodología Kimball, surge de acuerdo a los nuevos modelos que se están implementando para los grandes volúmenes de información, las nuevas tendencias hacen que se recurra a nuevas alternativas para mejorar la eficacia y rapidez del análisis de datos masivos.

Las nuevas tecnologías exigen que la información se encuentre bien estructurada pero son tantos los datos que se encuentran en la red o hay tantas fuentes de donde se pueden obtener que resulta casi imposible realizar un buen análisis con exactitud y rapidez. Para ello se utilizan los procesos ETL, permite a un usuario identificar y definir qué parámetros son los que verdaderamente necesita, es decir, el tipo de dato, encabezados, formato, entre otros y exportarlos a un texto plano, desde allí analizar este nuevo conjunto de datos ya procesado y estructurado resulta más eficiente para la toma de decisiones porque se enfoca en lo que el usuario ha definido, sin perder tiempo en lo que no se necesita. Ya sea para una empresa u organización, este proceso no se limita solo al comercio, también pueden ser usado en otros campos de la vida cotidiana como por ejemplo los accidentes de tránsito, es decir que este proceso es general aunque comúnmente se use para datos del área de los negocios.

Cuando la información ya está procesada y se convierte en información útil, se almacena en una base de datos no relacional, se usan comúnmente cuando la información es masiva, Neo4j es uno de los motores de datos que se basa en un modelo de datos no relacional NoSQL, dando solución a los problemas de escalabilidad, heterogeneidad y rendimiento e implementa JAVA por lo que resulta sencillo incorporarlo al desarrollo de la herramienta.

Por último una arquitectura basada en herramientas o soluciones de inteligencia de negocios provee recursos para cualquier tipo de análisis de datos ya sea datos de accidentes, comercio, banca u otros, al igual que los procesos ETL su uso puede aplicarse en cualquier campo. Para este proyecto se implementó la creación de cubos e hipercubos OLAP, los cuales permiten pasar de una visión estática a una dinámica, cambiando las dimensiones de consultas. Las soluciones de Inteligencia de Negocios varían según las necesidades de los usuarios, sus perfiles y su necesidad de información, por ello se escogió la creación de los cubos OLAP, para dar un ejemplo del alcance de estas herramientas de análisis para datos masivos.

III. OBJETIVOS

3.1 OBJETIVO GENERAL

Diseñar e implementar una herramienta ETL con bases de datos NoSQL orientadas a grafos para generar consultas OLAP en big-data.

3.2 OBJETIVOS ESPECÍFICOS

- Utilizar un origen de datos de fuente abierta para desarrollar los diferentes componentes de software y librerías propuestas en el proyecto.

- Implementar procesos ETL con las librerías desarrolladas sobre la base de datos Neo4J y con el origen de datos que se seleccione.

- Construir un componente web que permita implementar las librerías desarrolladas para comprobar su funcionamiento.

- Utilizar el modelo Kimball para la creacion de Data warehouse y Datamart.

- Implementar Cypher como caracterización de la información previamente analizada con kimball.

IV. DELIMITACIÓN

4.1 DELIMITACION TEMPORAL

El presente proyecto está enmarcado en tiempo comprendido entre el segundo semestre del año 2015 y el primer semestre del año 2016.

4.2 DELIMITACION FINANCIERA

Durante la ejecución del presente proyecto se ejecutó el siguiente presupuesto

Item	Características	Cantidad	Precio Unitario
Estación de trabajo para desarrollo de software	Procesador AMD x8 RAM 16 GB 1 TB de disco	1	$1.800.000
Computador portátil	Intel Core i5 Ram 16 GB 500 Gb de disco	2	$1.200.000
Resma de papel	Para documentos de revisión.	2	$9.000
Impresión / papelería Conexión a internet	Incluye empaste e impresión 2 Líneas tipo hogar	- 24 meses	$120.000 $68.000
Transportes/ Combustible	Desplazamiento a reuniones	-	$250.000
TOTAL			**$7.732.000**

Tabla 1 Presupuesto

4.3 ALCANCE

El alcance del presente proyecto es desarrollar una aplicación web en java, para ejecutar procesos ETL sobre fuentes de datos estructuradas en archivos planos, para implementar almacenes de datos y cubos OLAP sobre Neo4J.

4.4 LIMITACIONES

- Para el presente proyecto se usaron datos abiertos expuestos por el gobierno de los Estados Unidos a través del sitio web https://www.data.gov.

- La fuente de datos procesada esta disponible en la URL https://catalog.data.gov/dataset/traffic-violations-56dda.

V. FACTOR DIFERENCIAL O VALOR AGREGADO

El factor diferencial del proyecto se enfoca en la implementación de una nueva estructura de datos sobre grafos , utilizando como tecnología base Neo4j además de construir un conjunto de librerías construidas sobre Java además y una interface web para comprobar su funcionamiento (ya que oficialmente no se dispone de ninguna), todo esto para brindar a la comunidad nuevas herramientas y un punto de partida a desarrolladores para que incursionen en el mundo del Big Data y la inteligencia de negocios ya que el poder está en la información y no en los datos. Además la aplicación contiene un diseño web adaptable, conocido en inglés como Responsive Web Design (**RWD**), es decir la página web se adapta al dispositivo en el que se este visualizando. Esta característica permite competir en un mercado cuya tendencia es ir en aumento: la navegación móvil.

VI. MARCO REFERENCIAL

6.1 ESTADO DEL ARTE

Las diferentes herramientas que existen en la actualidad para el análisis de informacion varian de acuerdo a su utilidad, dentro de las herramientas ETL más comunes están :

- ✓ Pentaho Kettle: Esta basado en meta datos, posee una GUI para acelerar los procesos. La compañía Pentaho empezó operaciones en el año 2001. Tiene una comunidad activa de usuarios grande, alrededor de 13,500 usuarios. Funciona utilizando Java y presenta como ventaja el ser una solución multiplataforma.(BUSTILLOS, 2014)

- ✓ Talend : Es una herramienta OpenSource. Usa un enfoque hacia la generación de código para la manipulación de información y posee una GUI implementada en Eclipse RC. Lanzó su primera versión en el año 2006. Genera código en Java o Scripts en Pearl que pueden ser implementados en servidores que lo soporten. Cuenta con una gran variedad de testimonios por parte de compañías importantes. (BUSTILLOS, 2014)

- ✓ Informatica Power Center : Informatica tiene una muy buena suite empresarial de integración de datos. Fue fundada en el año de 1993. Líder actual del sector Data Integration (Gartner Dataquest). Tiene alrededor de 2600 clientes, entre los cuales figuran Bancos como Grupo BBVA, organizaciones Gubernamentales, etc. La compañía se enfoca meramente en soluciones para la integración de datos. (BUSTILLOS, 2014)

- ✓ Inabplex Inaport : Fundado en Reino Unido desde el año 2004 para satisfacer la migración de información hacia distintas soluciones CRM y software contable como Sage y Goldmine. (BUSTILLOS, 2014)

- ✓ IBM Cognos Data Manager : Proporciona funciones dimensionales de extracción, transformación y carga (ETL) para conseguir una inteligencia empresarial de alto rendimiento. Se puede integrar con la GUI de IBM Data Manager Designer para diseñar y crear prototipos También, se pueden ejecutar compilaciones y secuencias de trabajos en sistemas remotos desde un sistema de entorno de diseño de Data Manager. Data Manager Engine se tiene que instalar en un sistema UNIX o Linux. (BUSTILLOS, 2014)

- ✓ Oracle Warehouse Bulde : La opción empresarial ETL (Enterprise ETL Option) para Warehouse Builder es una opción que puede ser adquirida con Oracle Warehouse Builder como parte de la edición empresarial del motor de base de datos. Permite ejecutar cargas de datos usando métodos rápidos y eficientes tales como el Oracle Data Pump y transportable tablespaces. Permite prever el efecto que puedan tener los cambios que se hagan en cualquier lugar de los metadatos del sistema ETL Es posible generar un modelo para configurar los ambientes de desarrollo, pruebas y producción a niveles separados. (BUSTILLOS, 2014)

- ✓ Microsoft Integration Services : Puede extraer y transformar datos de diversos orígenes como archivos de datos XML, archivos planos y orígenes de datos relacionales y, después, cargar los datos en uno o varios destinos. Se pueden realizar tareas de migración fácilmente usando tareas visuales. Si se desea crear nueva funcionalidad, se pueden crear scripts en c# o VB Puede conseguir conectividad mediante CLI vía DLLs tipo ensamblador. (BUSTILLOS, 2014)

	FACILIDAD DE USO	SOPORTE	VELOCIDAD	CALIDAD DE DATA	MONITOREO	CONECTIVIDAD
PENTAHO KETTLE	GUI fácil de usar dentro de las alternativas OpenSource	EU, Reino Unido y consultorias asociadas	Al requerir Java Database Conector disminuye la velocidad de transacción	Ofrece herramientas para DQ dentro de la GUI, sentencias SQL personalizables herramientas JavaScript y REGEX paradepurar la información.	Tiene herramientas practicas de monitoreo y registro histórico	Varias BD, archivos planos, xml, Excel, servicios web.
TALEND	GUI basado en un add-on para Eclipse RC	De paga en EU	Requiere configuarion específica y manual con conocimiento previo de la data a utilizar.	Ofrece herramientas para DQ dentro de la GUI, sentencias SQL personalizables usando Java.	Tiene herramientas practicas de monitoreo y registro histórico	Varias BD, archivos planos, xml, Excel, servicios web. Necesita JDBC para conexión.
INFORMATICA POWER CENTER	GUI facil de usar pero requiere de entrenamiento	Mundial via web y consultoria	Rapida gracias a PushDown no permite volver a un estado anterior.	Ofrece DQ a traves de otro producto llamado Informatica Data Quality	Tiene herramientas practicas y extensivas de monitoreo y registro histórico	Varias BD, archivos planos, xml, Excel, servicios web puede exportar como servicio web
INABPLEX INAPORT	Se concecta directamente a la CRM de importación	Mundial via web y consultoria	Conexión proporcional a la velocidad del CRM	Debido a la restriccion del origen de información se puede realizar tareas de DQ dentro de la misma.	Tiene herramientas practicas de monitoreo y registro histórico	Cualquier conexión ODBC, MSSQL, OUTLOOK, ACT, EXCEL
IBM COGNOS DATA MANAGER	Se puede integrar con la GUI de IBM Data Manager Designer pero es un modulo aparte	Mundial contratando en paquete	Rapido si se trabaja con DB2 con distintos manejadores de BD disminuye la velocidad	Mediante Cognos Data Manager se pueden incorporar herramientas para DQ	Maneja registro de históricos.	Cualquier conexión ODBC, DB2, para importación a DB2, cubos de información T1MAP.
ORACLE WAREHOUSE BULDER	Facil si se almacena en BD de Oracle con la herramienta Data Pump pero no ofrece compatibilidad con otras BD.	Via local Oracle Latinoamerica	Proporcional al servicio Oracle donde se este trabajando	Permite DQ mediante el uso de Oracle Warehouse Builder Data profiling Features	Tiene herramientas prácticas y extensivas de monitoreo y registro histórico.	Solamente compatible con BD Oracle
MICROSOFT INTEGRATION SERVICES	Se pueden realizar tareas de migracion facilmente usando tareas visuales	Via plataforma TechNet	Proporcional al servicio MSSQL donde se este trabajando	Requiere de SQL Server Data Quality Services para ofrecer DQ	Tiene herramientas prácticas y extensivas de monitoreo y registro histórico.	Bases de Datos SQL SERVER, ACCESS, ADO.NET.

Tabla 2 Tabla comparativa entre las herramientas ETL mas utilizadas en la actualidad (BUSTILLOS, 2014)

6.2 MARCO CONCEPTUAL

6.2.1 BIGDATA

Conceptos clave para comprender BigData

El término desde que apareció por el MGI (McKinsey Global Insitute) en Junio de 2011 define Big Data como:

"El conjunto de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos".[1]

Dentro de las deficniciones mas completas de Big Data se encuentra la de (Gartner, 2012), ésta afirma que:

"Son activos de información caracterizados 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."[2]

Según Wikipedia Big Data se aplica a: *"Un conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable".* El informe de TicBeat (2012) puntualiza a Big Data como: "la enorme cantidad de datos que desde hace unos años se genera constantemente a partir de cualquier actividad." ; más adelante dicho informe recalca que: " BigData bien entendido en la búsqueda del mejor camino para aprovechar dicha avalancha de datos". [3] Según el artículo "In Perspective" de Fidelity Worldwide Investment2 (2012) Big Data es: "El término inglés que designa los conjuntos de datos de gran

[1] (MGI McKinsey Global Institute, Junio 2011)
[2] (Gartner, 2012)
[3] (TicBeat, 2012)

tamaño y generalmente desestructurados que resultan difíciles de manejar usando las aplicaciones de bases de datos convencionales". [4]

Al notar este sin fin de deficiciones parecidas entre si, se puede concluir que BigData consta del tratamiento y análisis de grandes volúmenes de información.

Es necesario precisar las diferencias entre dato, información y conocimiento. Un dato es un elemento primario de información que por sí solo es irrelevante para la toma de decisiones, un número de teléfono o un nombre de una persona, son datos, que sin un propósito o utilidad no sirven para nada. La información por el contrario es un conjunto de datos procesados y que tiene relevancia o propósito y que por lo tanto son de utilidad para la toma de decisiones, por último el conocimiento es una mezcla de experiencias, valores, información y know-How que aplicaran los conocedores de este para la toma de decisiones.

Asi las cosas, el proposito de BigData es convertir la informacion en conocimiento util para la toma de desiciones en una organización o empresa.

[4] (Fidelity Worldwide Investment2, 2012)

6.2.2 SOLUCIÓN BI BODEGA DE DATOS DATA WAREHOUSE

Inteligencia de Negocios BI (Bussiness Intelligence)

Existen infinidad de formas de analizar datos, una de las mas utilizadas en la actualidad si se habla de BigData es sin duda las herramientas que ofrece una

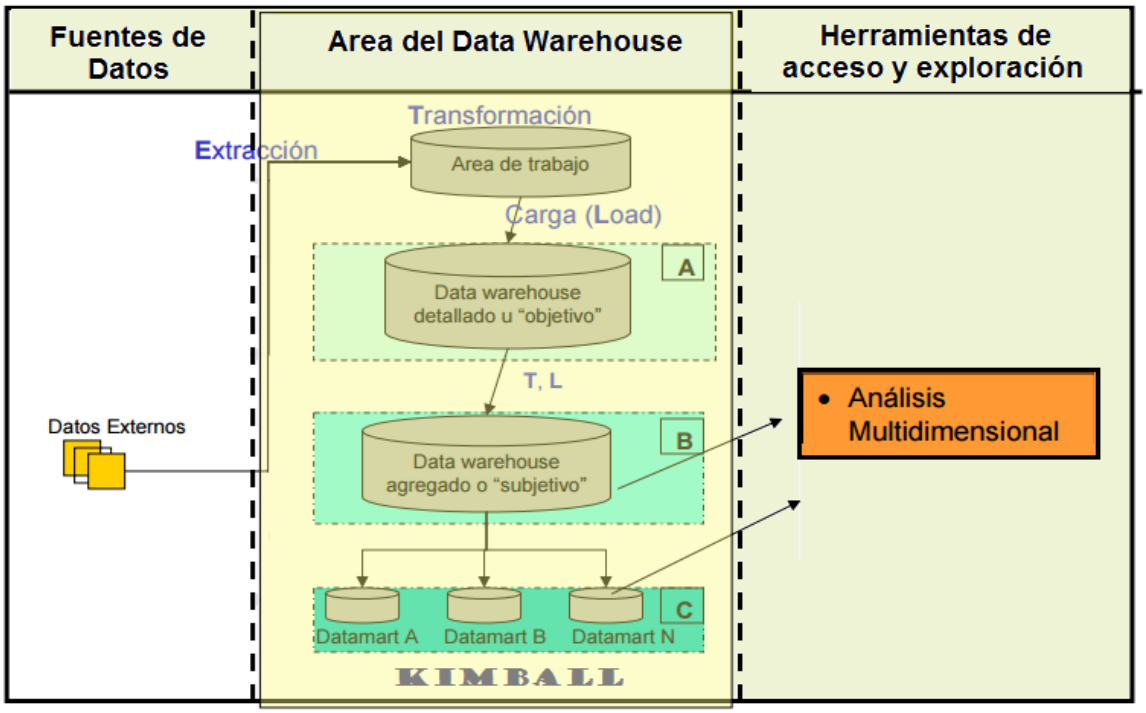


Ilustración 0.1 Arquitectura de la Solución BI Adaptada de:
http://www.econ.uba.ar/sistemas/materias/715/echinkes/material/BI_clase_4_El_Data_warehouse.pdf

solucion BI (ver Ilustración I), asociándolo directamente a las tecnologías de la información, podemos definir Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, análisis OLAP...) o para su análisis y conversión en conocimiento soporte a la toma de decisiones sobre el negocio. Estas herramientas permiten el análisis de datos masivos, facilita la toma de desiciones, de una manera rápida y eficaz.

Componentes de una Solucion BI

Los componentes de una solucion BI (Bussines Inteligence), son tres:

1. **Fuentes de Datos:** Para el diseño es importante tener presentes tres preguntas básicas:
    - ¿Cuál es la información requerida para gestionar y tomar decisiones?.
    - ¿Cuál debe ser el formato y composición de los datos a utilizar?.
    - ¿Y de dónde proceden esos datos y cuál es la disponibilidad y periodicidad requerida?.

2. **Construcción y alimentación del datawarehouse y/o de los datamarts:** Conceptualmente un datawarehouse es una base de datos corporativa que replica los datos transaccionales una vez seleccionados, depurados y especialmente estructurados para actividades de query y reporting. Un datamart (o mercado de datos) es una base de datos especializada, departamental, orientada a satisfacer las necesidades específicas de un grupo particular de usuarios (en otras palabras, un datawarehouse departamental, normalmente subconjunto del corporativo con transformaciones específicas para el área a la que va dirigido). En otras palabras es una base de datos donde se van a almacenar los datos ya estructurados.

3. **Herramientas de explotación de la información:** las herramientas de explotación BI que permiten extraer la información del BI según sea la necesidad o el tipo de información requerida.

    - **Reporting empresarial :** Herramienta de recopilación y presentación de datos, informes y listados procedentes de los sistemas ERP.

- **Análisis OLAP :** (On-Line Analytical Processing). Herramienta que tiene como objetivo agilizar la consulta de grandes cantidades de datos proporcionando un acceso multidimensional a los datos.
- **Análisis visual.** : Es la interpretación de los datos mediante componentes gráficos como barras histogramas, grafos, entre otros.
- **Análisis predictivo** : Agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos.
- **Cuadro de mando analítico** (EIS tradicionales) : Herramienta que permite la creación de reportes de manera gráfica con indicadores clave para la gestión. Facilita el acceso a la información corporativa con el objetivo de mejorar la toma de decisiones y que se caracteriza por ser muy rápida y visual y por ser extremadamente sencilla de utilizar, sin necesidad de tener conocimientos técnicos.
- **Cuadro de mando integral o estratégico** : Permite establecer y monitorizar y establecer los objetivos de una empresa de sus diferentes áreas o unidades por medio de informes que dan soporte a la estrategia de la empresa.
- **Análisis avanzado (Minería de datos – DSS)** : Proceso que intenta descubrir patrones y tendencias en grandes volúmenes de conjuntos de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.
- **Gestión del rendimiento. (Corporate Performace Management - CPM) :** Se refiere a las herramientas informáticas para gestionar el rendimiento de una empresa basándose en metodologías, métricas, procesos y sistemas necesarios para monitorizar. Estos programas están basados en su origen sobre herramientas tipo Business Intelligence (BI). Muchos analistas no reconocen aún una diferencia tecnológica entre ambos campos. Generalmente se reconoce que el CPM se dedica sobre todos a temas como

presupuesto financiero, consolidación legal, pero también a la gestión de riesgos o al balanced scorecard.

- **Previsiones :** Conjeturar lo que va a suceder a través de la interpretación de indicios o señales; ver con anticipación; preparar medios para futuras contingencias
- **Reglas de negocio :** Describe las políticas, normas, operaciones, definiciones y restricciones presentes en una organización y que son de vital importancia para alcanzar los objetivos misionales.
- **Dashboards :** Es una interfaz donde el usuario puede administrar el equipo y/o software (tablero de instrumentos).
- **Integración de datos ETL :** (Extract, Transform and Load). En español Extraer, transformar y cargar, es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos y cargarlos en otra base de datos para analizar o en otro sistema operacional para apoyar un proceso de negocio. [5]

6.2.3 ADQUISICION DE DATOS

Fuentes de Datos

Hoy en dia hay más fuentes y tipos de datos que nunca y con ello el volumen de datos aumenta, integrar estos datos y convertirlos en nformación útil depende de una herramienta que permita manipular estas fuentes y transformarlas, dentro de las fuentes es posible encontrar datos estructurados, no estructurados o semiestructurados.

Las fuentes de datos pueden ser tal como se muestran en la Ilustración 2.

[5] (Ibermática, 2007)

Bases de Datos Relacionales	• MySQL • PostgresSQL • SQLite • Microsoft SQL Server	• Oracle • IBM DB2 • Teradata
Bases de Datos Analíticas	• HPE Verrtica • Greenplum • Amazon Redshoft	• SAP HANA • Teradata • Netezza
Bases de Datos NoSQL	• Cassandra • MongoDB • Apache Hbase	• CouchDB • Neo4j
Otras Fuentes	• JSON • XML • Amazon S3 • Excel • Splunk • Fixed-width Files • HL7 data • Java Messages Service (JMS) • Bussinnes Applications (ERP,CRM)	• Text Files • CSV Files • Rss Feeds • Avro • Email Messages • Google Analytics

Ilustración 0.2 Tipos de fuentes de datos , Adaptado de http://www.pentaho.com/data-sources-for-business-analytics

La herramienta desarrollada usa una fuente de datos en formato CSV aunque tambien permite la carga de datos desde un Archivo de texto.

Tipos de Datos

Dentro de los BigData se manejan tres tipos de datos.

- ✓ Datos estructurados (Structured Data): Datos que tienen bien definidos su longitud y su formato. Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y las hojas de cálculo.
- ✓ Datos no estructurados (Unstructured Data): Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, e-mails o documentos de texto.
- ✓ Datos semiestructurados (Semistructured Data): Datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos

semiestructurados que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. Un ejemplo es el HTML, el XML, CSV, o el JSON.[6]

La herramienta desarrollada cuenta con datos de prueba semiestructurados en formato CSV.

6.2.4 CONSTRUCCION DEL DATA WAREHOUSE

Para la construccion del DW según la Solucion BI se deben realizar los procesos ETL los cuales se exponen a continuación:

Conceptos de un Proceso ETL

Según Wikipedia se dice lo siguiente de las herramientas ETL: “ETL son las siglas en inglés de Extraer, Transformar y Cargar (Extract, Transform and Load). Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio”.

Los procesos ETL también se pueden utilizar para la integración con sistemas heredados (aplicaciones antiguas existentes en las organizaciones que se han de integrar con los nuevos aplicativos, por ejemplo, ERP´s. La tecnología utilizada en dichas aplicaciones puede hacer dificil la integración con los nuevos programas).

Fases de un proceso ETL

Las distintas fases o secuencias de un proceso ETL son las siguientes:

- Extracción de los datos desde uno o varios sistemas fuente.

[6] (García, 2012 -2013)

- Transformación de dichos datos, es decir, posibilidad de reformatear y limpiar estos datos cuando sea necesario.
- Carga de dichos datos en otro lugar o base de datos, un data mart o un data warehouse, con el objeto de analizarlos o apoyar un proceso de negocio.

Beneficios de los procesos ETL

A cualquier empresa u organización le beneficia poner en marcha un proceso ETL para mover y transformar los datos que maneja por los siguientes motivos:

Poder crear una Master Data Management, es decir, un repositorio central estandarizado de todos los datos de la organización. Por ejemplo, si tenemos un objeto cliente en una base de datos de créditos y otro objeto cliente en la base de datos de tarjetas de crédito, lo que haría el Master sería definir, de forma concreta e inequívoca, un registro cliente único con su nombre y apellidos para la organización.

Posibilita a los directivos tomar decisiones estratégicas basadas en el análisis de los datos cargados en las bases nuevas y actualizadas: la dama mart o data warehouse.

Sirve para integrar sistemas. Las organizaciones crecen de forma orgánica y cada vez se van agregando más fuentes de datos. Esto provoca que comience a surgir nuevas necesidades, como por ejemplo integrar los datos de un banking on line con los datos antiguos de un sistema legacy.

Poder tener una visión global de todos los datos consolidados en una data warehouse. Por ejemplo, crear una estrategia de marketing basándose en el análisis de los datos anteriores. [7]

[7] (Data, 2013)

Descripciín del Proceso ETL aplicado a la herramienta Desarrollada

Un proceso ETL le permite a un usuario identificar y definir qué parámetros son los que verdaderamente necesita, es decir, el tipo de dato, encabezados, formato, entre otros y exportarlos a un texto plano para su posterior análisis. Cuando el proceso de identificar y definir los parametros esta completo (Extraer), se procede a realizar una estandarizacion de los datos como por ejemplo omitir un campo nulo; de alli se transorma para cargarlos en una base de datos, un Dashboard o un Datamarth. Para el caso de esta investigacion se uso una base de datos en grafos Neo4j, este tipo de base de datos permite la carga de datos masivos lo cual se acomoda al desarrollo de la herramienta, por ello no solo se transorman los datos al lenguaje de la base de datos (cyper) sino el archivo que contiene los datos, es decir el tamaño. Después de realizar este proceso (Transformar), se procede al proceso final (Carga), es decir se ejecutan los scripts que el usuario construyó a partir de la herramienta. Estos procesos hacen que el análisis de la informacion sea mas limpio, rápido y exacto. Al tener los datos ya estructurados de cierta manera se pueden hacer uso de las herramientas de analisis que hacen posible que la toma de deciciones sea más.

6.2.5 ALIMENTACIÓN DEL DATA WAREHOUSE

Almacenamiento para BigData

Una base de datos con capacidad de almacenar datos masivos es lo que se necesita para el desarrollo optimo de la herramienta, dentro de las más opcionadas estan las bases de datos no relacionales o NoQSL y dentro de este tipm de bases de datos se encuentran las orientadas a grafos que permiten una visualizacion de los datos almacenados. Para este proyecto de grado se implemento el uso de la herramienta Neo4j, tanto la base de datos como el motor seleccionado se desctiben a continuación.

Bases de Datos No Relacionales

El término NoSQL se usó por primera vez en 1998 por el señor Carlo Strozzi. Consistía en una base de datos *open-source* ligera que no ofrecía un interface SQL en otras palabras fue una alternativa a los sistemas clásicos relacionales tanto en estructura como en la forma en que se relacionan los datos. En las bases de datos NoSQL no se utiliza este tipo de lenguaje estándar, existen otros tipos de lenguaje de ahí el acrónimo "No sólo SQL" es decir no se maneja un estándar específico. Otras ventajas con las que cuenta este tipo de base de datos es que son fácilmente escalables, se conoce como escalabilidad horizontal ofrecen mínimos tiempos de consulta y pueden trabajar con grandes volúmenes de datos (*Big Data*), su forma de almacenamiento es más ágil y dinámica. Los *Big Data* son realmente la parte novedosa de las bases de datos NoSQL.

Es por ello que empresas como Google, Amazon, Facebook o Twitter hacen uso de las bases de datos NoSQL para sus aplicaciones web, necesitan de una respuesta rápida en tiempo real a las peticiones de los datos y que el rendimiento fuera capaz de superar otros niveles, otra ventaja es cuando de asegurar la consistencia de los datos se trata, ya que las NoSQL son muy flexibles en ocupar menos tiempo asegurando dicha consistencia, en pocas palabras estos sistemas son óptimos en agregar y recuperar grandes volúmenes de información.

Ventajas de NoSQL

- ✓ Se ejecutan en máquinas con pocos recursos: Estos sistemas, a diferencia de los sistemas basados en SQL, no requieren de apenas computación, por lo que se pueden montar en máquinas de un coste más reducido.
- ✓ Escalabilidad horizontal: Para mejorar el rendimiento de estos sistemas simplemente se consigue añadiendo más nodos, con la única operación de indicar al sistema cuáles son los nodos que están disponibles.
- ✓ Pueden manejar gran cantidad de datos: Esto es debido a que utiliza una estructura distribuida, en muchos casos mediante tablas Hash.

- ✓ No genera cuellos de botella: El principal problema de los sistemas SQL es que necesitan transcribir cada sentencia para poder ser ejecutada, y cada sentencia compleja requiere además de un nivel de ejecución aún más complejo, lo que constituye un punto de entrada en común, que ante muchas peticiones puede ralentizar el sistema.[8]

Asi como existen diferentes tipos de bases de datos relacionales, las NoSQL dentro de los más utilizados están:

- ✓ Base de Datos Clave Valor
- ✓ Base de Datos Documental
- ✓ Base de Datos en Grafos

Las Bases en Datos en Grafo la información se representa como nodos de un grafo y sus relaciones con las aristas del mismo, de manera que se puede hacer uso de la teoría de grafos para recorrerla. Algunos ejemplos de este tipo son Neo4j, InfoGrid o Virtuoso.

Asi como las bases de datos relacionales tienen su terminología, Estos son algunos términos importantes de bases de datos en grafos son:

✓ Nodo: Un nodo es un equipo en red que ofrece algunos tipo de servicio (por lo general un servicio de computación), almacenamiento local, y el acceso a un conjunto de datos o archivos mucho más grandes distribuida almacenar.

✓ Clusters: Tal como se utiliza en la tierra NoSQL, un cluster es un conjunto de nodos que constituyen una sola unidad. Dependiendo de la base de datos, un cluster puede ser un conjunto de nodos en un estante particular en el centro de datos o nodos que están en la misma fila como otra nodos.

[8] (acenswhitepaper, 2014)

✓ Sharding: Sharding (también llamado particionamiento horizontal) implica la partición de la base de datos sobre el valor de algunos campo. Esto se hace por algunas bases de datos NoSQL para igualar la cantidad de datos entre los nodos. Usted puede hacer sharding directamente por hash el valor llave o por el equilibrio de carga - La dirección de cada nodo para redistribuir sus datos a otro nodo menos muy cargado.

✓ Replicación: replicación es el mecanismo que proporciona disponibilidad de base de datos. Las porciones de una base de datos se escriben a varios nodos de modo que si un nodo falla, otro nodo contiene una réplica de los datos del nodo que ha fallado. Para más detalles sobre replicación.

✓ ÁCIDO: significa atomicidad, consistencia, aislamiento y durabilidad. ACID es una consigna en los sistemas transaccionales (tales como los utilizados en la banca y comercio electrónico) y necesario para cualquier sistema de registro.

✓ BASE: significa básicamente disponibles, el estado blando, y finalmente consistente. Básicamente disponible indica que la base de datos puede no estar disponible 24/7. Estado Soft implica que el estado de la base de datos puede ser inconsistente; si cambia su dirección de correo electrónico del trabajo, por ejemplo, un amigo puede no ver que la información de inmediato. Eventualmente consistente, sin embargo, significa que tu amigo vea el tiempo la información de su correo electrónico modificado.[9]

[9] (BROOKS, 2014)

NEO4J

Neo4j es una base datos orientada a grafos en la cual la información es dirigida de un nodo a otro, es capaz de representar cualquier tipo de datos de una forma mucho más comprensible.

Como se mencionó anteriormente, las unidades fundamentales para un gráfico son los nodos y las relaciones. Los nodos se utilizan en la representación de entidades y son ideales para datos complejos y conectados.

Patrocinado por Neo Technology, Neo4j es una base de datos NoSQL gráfico de código abierto implementado en Java y Scala. Con el desarrollo de partida en 2003, ha sido a disposición del público desde el año 2007. El código fuente y seguimiento de problemas están disponibles en GitHub, con el apoyo fácilmente disponibles en desbordamiento de pila y el grupo Neo4j Google. Neo4j se utiliza hoy en día por cientos de miles de empresas y organizaciones en casi todas las industrias. Los casos de uso incluyen contactos, gestión de redes, análisis de software, la investigación científica, enrutamiento, gestión organizacional y de proyectos, recomendaciones, redes sociales y más.

Neo4j implementa la Propiedad Gráfico Modelo de manera eficiente hasta el nivel de almacenamiento.[10]

Neo4j se caracteriza por las siguientes cualidades:

- Materializa relaciones desde el momento justo en que se están creando
- Representación eficaz de los nodos y sus relaciones
- Todas las relaciones realizadas en Neo4j tienen la misma relevancia y hay gran velocidad al materializar nuevas relaciones o cuando surgen nuevas necesidades

[10] (Neo Technology, Inc., s.f.)

- Almacenamiento compacto y almacenamiento en caché de memoria para gráficos, conduciendo a una eficaz ampliación de los nodos correspondiente a una base de datos moderada.
- Entre los Beneficios de las bases de datos en grafos se encuentran el alto desempeño y disponibilidad
- Escalable: 32 miles de millones de Nodos, 32 miles de millones de Relaciones, 64 miles de millones de Propiedades
- Escrito en la parte superior de la JVM

Sumit Gupta en su libro Neo4j Essentials está de acuerdo con que los datos sin relaciones no sirven de nada, si no es posible establecer relaciones entre las entidades el uso será mínimo o ninguno.

ALEKSA VUKOTIC[11] en su libro Neo4j in action aduce la baja capacidad que tiene el modelo relacional con respecto al modelo de datos interconectados, toda vez que no solo se limita a la interacción de las cosas sino también a las diferencias y similitudes de éstas conexiones. Las bases de datos gráficas, ofrecen mucha claridad y sobre todo la visibilidad de los datos expresados. Neo4j supera el mundo NeoQSL. Neo4j proporciona soporte de bases de datos tradicionales y la magnitud de las órdenes supera en rendimiento a los tradicionales.

6.2.6 HERRAMIENTAS DE EXPLOTACIÓN DE LA INFORMACIÓN

Modelado con Sistemas OLAP

Concepto de OLAP

OLAP es el acrónimo en inglés de (On-Line Analytical Processing). Es una solucion del campo de la inteligencia de negocios, cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Es un metodo para buscar en los datos

[11] (VUKOTIC, Aleksa; NICKI , Watt ;, 2015)

de diferentes maneras. Con OLAP los datos son clasificados en diferentes dimensiones, las que pueden ser vistas unas con otras en cualquier combinación para obterner diferentes análisis de los datos que contienen.

Utilidades OLAP

Las utilidades que presentan las aplicaciones OLAP son:

- ✓ Tienen accesso a grandes cantidades de datos, por ejemplo varios años de datos en una bodega.
- ✓ Analizan las relaciones entre muchos tipos de elementos.
- ✓ Involucran datos agregados.
- ✓ Comparan datos agregados a traves de periodos jerarquicos, mensualmente, trimestralmente, anualmente, etc.
- ✓ Presentan los datos en diferentes perspectivas.
- ✓ Involucran cálculos complejos entre elementos de datos.
- ✓ Puede responder con rápiez a las consultas de los usuarios de manera que los agentes o analistas puedan seguir un proceso de analítico o de desición sin verse impedidos por el sistema.

Beneficios que presenta OLAP

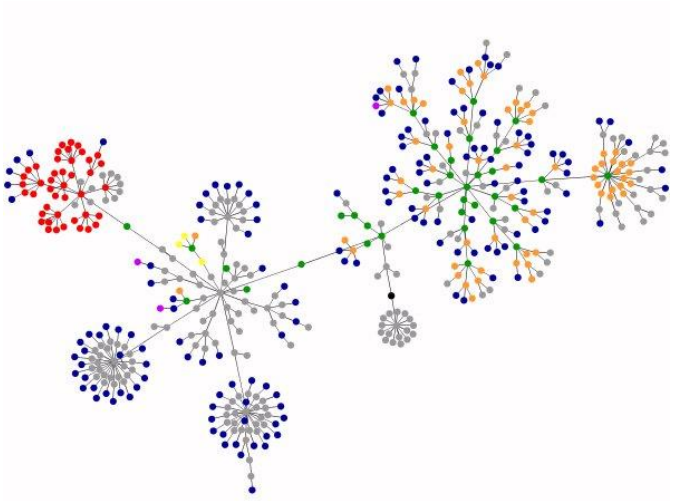
- ✓ Es de fácil uso y acceso flexible para el usuario.
- ✓ Los datos estan organizados en varias dimensiones lo que permite que los usuarios hagan un mejor análisis.
- ✓ Ahorro generado por la productividad de personal altamente profesional y costoso que usa permanentemente software y sistemas de información.
- ✓ Permite encontrar la historia en los datos.
- ✓ Genera ciertas ventajas competitivas. [12]

[12] (Tangient LLC, 2016)

Visualizacion de los Datos (Neo4j)

Neo4j fuera de ser un motor de bases de datos permite mostrar los datos en forma gráfica. Las bases de datos gráficas son el antídoto perfecto ante el crecimiento desbordante de los datos. La gran cantidad de información, dispositivos y usuarios hacen que las tecnologías tradicionales no puedan gestionar tantos datos. La flexibilidad, rendimiento y escalabilidad de Neo4j permite gestionar, monitorizar y optimizar todo tipo de redes físicas y virtuales pese a la gran cantidad de datos.

Neo4j cuenta con un visor que permite mostrar los datos en forma de grafos (Ilustración 3).



lustración 0.3 Ejemplo del esquema relacional de los datos en Neo4j.

Para esta seccion se ha tomado como referencia el segundo capítulo del libro”Interactive Data Visualization” de Scott Murray.

Al tratar grandes volúmens de información se hace preciso utilizar métodos para obtner la visualización correcta de los datos. Resulta más sencillo interpretar datos en grafos que en cifras, es más cómodo y menos tedioso, aún más para datos masivos. La visualización es un proceso de mapear información textual hacia

información visual. El mapeo depende dela interpretacion y la forma de expresar la información por medio de propiedades visuales. Cuanto más compleja la información , más complejo sera el mapeo de los datos.

Para dar un mayor valor a las visualizaciones es necesario tener varias vistas de los datos, de este modo poder iteractuar con la información. Las visualizaciones iteractivas hacen que los usuarios puedan explorar la información por ellos mismos, de este modo llegar a concusiones mas acertadas.

Cyper

Cyper es el lenguaje para crear consultas en Neo4j. Es un lenguaje inspirado en SQL para describir que queremos seleccionar, insertar, actualizar o borrar de una base de datos basada en grafos sin describir a nivel fisico como hacerlo. Algunas de las consultas básicas son:

Escribiendo cláusulas:

- ✓ Crear Nodos en función de una etiqueta y con unas propiedades.

```
CREATE (nodo:Etiqueta {propiedad1:"prop1",propiedad2="prop2"})
```

- ✓ Crear Relaciones entre nodos ya existentes.

```
MATCH node1,node2 CREATE(node1)-[:REL_TYPE]->(node2)
```

- ✓ Buscar nodos en función de una etiqueta.

```
MATCH (node:Label) RETURN node
```

- ✓ Buscar nodos filtrando según sus propiedades.

```
MATCH (n) WHERE n.propiedad1 = 'valor' RETURN n
```

- ✓ Borrar nodos y relaciones.

```
MATCH (n) OPTIONAL MATCH (n)-[r]-() DELETE n,r
```

✓ Cargar un vsc y crear nodos a partir de él

```
USING PERIODIC COMMIT 1000
LOAD CSV WITH HEADERS FROM 'file:/ejemplo.csv' AS line
FIELDTERMINATOR';' WITH line
CREATE (n:Nodos{ propiedad:line.cabecera})
```

[13]

D3.js

Es una librería para crear visualizacoines de datos. D3 es la abreviacion de Data-Driven-Documents. El autor es Mike Bostock con aportaciones de otros desarrolladores, Se encuentra disponible de forma gratuita en GitHub. Se lanzo bajo licencia BSD por lo que se puede utilizar, modificar y adaptar el código para cualquier uso. Su propósito es crear visualizaciones que expliquen la información.

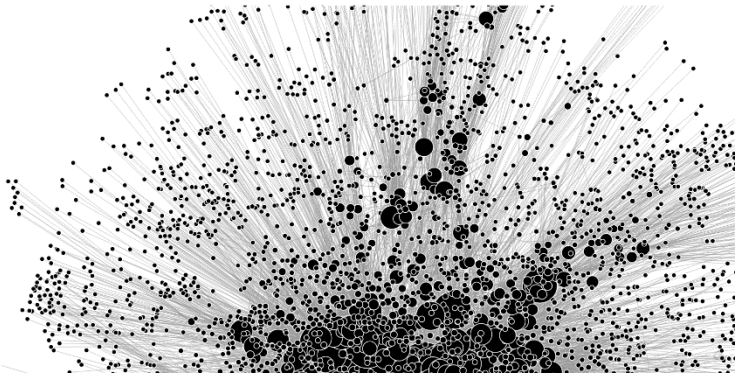


Ilustración 0.4 Ejemplo nodos conectados con D3 Disponible en http://alanahryding.me/blog/?p=618

D3 se caracteriza por facilitar la generación y manipulación de documentos web con datos cargándolos en la memoria del navegador, los enlaza a elementos de la página web y creando nuevos elementos. Permite hacer transiciones de elementos entre estados en respuesta a las iteracciones del usuario. Una de sus desventajas es que no soporta navegadores antiguos, esto para incentivar a los usuarios a tener

[13] (neo4j, 2016)

las últimas versiones de los navegadores continuar con el crecimiento de la tecnología web.

D3 no oculta los datos, se ejecuta desde el lado del cliente para que pueda generar la visualización por ello es necesario enviar los datos.[14]

[14] (Carrasco, 2015)

VII. METODOLOGÍAS PARA EL DESARROLLO DEL PROYECTO

7.1 MODELO CASCADA

Es el modelo más básico. Es de finales de los 70.. El desarrollo software se realiza a través de una secuencia simple de fases. Cada fase tiene un conjunto de metas y actividades bien definidas. Las flechas verdes muestran el flujo normal de avance del ciclo de vida. Las flechas azules discontinuas muestran la realimentación entre fases. [15]

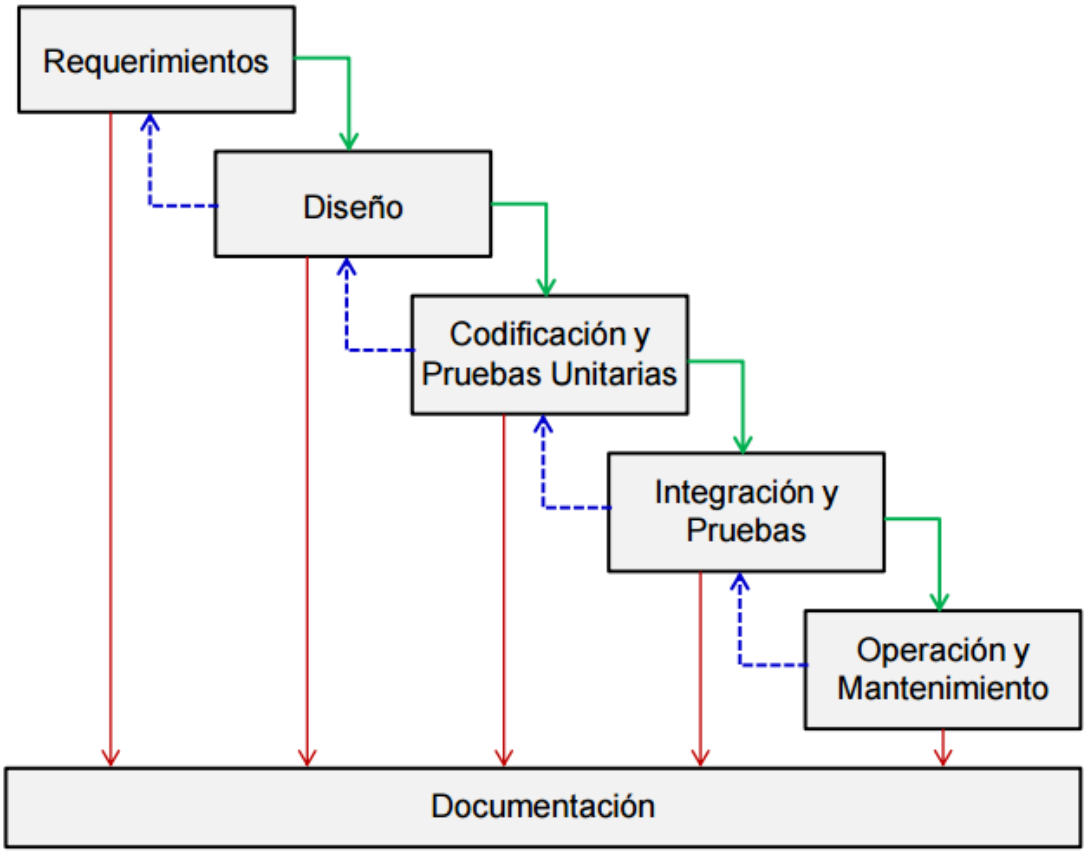


Ilustración 0.5 Arquitectura Modelo Cascada Disponible en : http://arantxa.ii.uam.es/~proyectos/teoria/C5_Proyectos%20de%20desarrollo%20software.pdf

[15] (Tapias, 2014)

7.2 KIMBALL

El componenete importante en la arquitectura Kimball es como se aplica a la inteligencia de negocios, Ralph Kimball es el pionero de esta metodología, introdujo la inteligencia en la industria de almacenamiento de datos y esta basada en un ciclo de vida dimensional, por ello se mantiene dentro de las favoritas cuando se construye un Datawarehouse, el cual es el principal objetivo de kimball.

Como primera medida Kimball se basa en los siguientes pasos para la construcción de una solución de DW/BI (Datawarehouse/Business Intelligence)

1. Planificación
2. Análisis de Requerimientos
3. Modelado Dimensional
4. Diseño Físico
5. Diseño del sistema de Extracción, Transformación y Carga (ETL).
6. Especificación y desarrollo de aplicaciones de BI

Las tareas de esta metodología (ciclo de vida) se muestran en la Ilustracion 6.

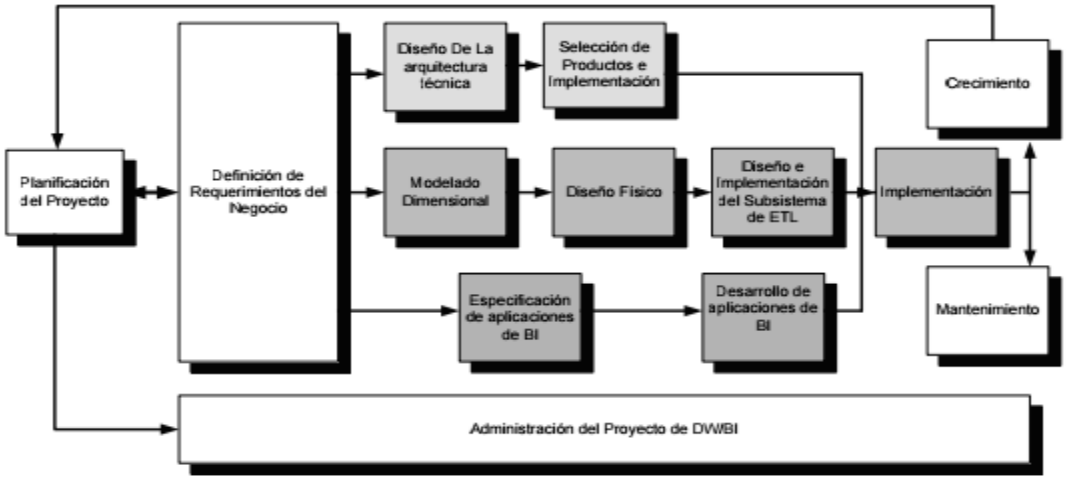


Ilustración 0.6 Tareas de la metodología de Kimball, denominada Business Dimensional Lifecycle (Kimball et al 98, 08, Mundy & Thornthwaite 06)

En la fase de **PLANIFICACIÖN** se busca principalmente la deficinion y el alcance del proecto Datawarehouse, esta fase indica el escenario donde va a surgir la necesidad de lDW (entender las necesidades del negocio), el nivel de planificación (identificar y programar las tareas que se deben llevar a cabo), la identiidad de este mismo, el personal (asignar la carga de trabajo a los recursos), el desarrollo del proyecto ((elaborar un documento final donde se represente el plan del proyecto), el seguimiento y por último el control (Monitoreo) que se va a llevar de este (procesos, actividades, rastreo de problemas, etc).

La fase de **REQUERIMIENTO DE NEGOCIO** es la más importante y es determinante para que este proceso sea todo un éxito, aquí se debe tener en claro los factores claves que van a guiar el negocio, se debe aprender lo necesario sobre el negocio, la industria, los competidores y los clientes a quien va dirigido.

Con respecto a los requerimientos fundamentales de esta fase en gran medida es saber muy bien a que personal se va a entrevistar oara el negocio, esto implica tener que examinar detalladamente el organigrama de la organización, ya que es claro que en este proceso, encontramos al directivo responsable de toma de decisines estratégicas, a los administradores quenes se encargan de buscar alternativas estratégicas para aplicarlas, el personal de sistemas que son ese conjunto de personas que realmente hace un seguimiento más exhaustivo sobre los problemas informáticos y los datos que existes y la gente que se necesita entrevistar, ya que con estas entrevistas se puede identificar los faactores claves que guian al negocio, se puede identificar temas analíticos que agrupan requerimientos de un tema en común.

A partit de los análisis se pueden construir Bus Matrix (En inglés), esta base es una herramienta metodológica llamada matriz de procesos / Dimenciones. En esta matriz se denota que en sus filas tiene identificado llos procesos del negocio, en sus columnas tiene las dimenciones también identificadas.

En la fase de **MODELADO DIMENSIONAL** se empieza por una matriz donde determina las dimensiones de cada indicador para ir asi especificando los diferentes detalles dentro de cada concepto del negocio, sin dejar de un lado la parte dinámica e iteractiva. El proceso de este modelo dimensional consiste en cuatro pasos

- ✓ Proceso de Negocio: Area a organizar, esta decisión es tomada del area de dirección y únicamente depende del análisis y los temas analíticos de la etapa anterior.
- ✓ Nivel de granularidad: Aquí lo importante es especificar el nivel de lo que se esta requiriendo, aquí todo depende de los requerimientos solicitados en el negocio.
- ✓ Dimensiones: Surgen a raiz de la granularidad y la matriz de procesos / dimensiones.
- ✓ Medidas y Tablas de los hechos: En este último paso, hay que identificar las medidas que vas surgiendo en los proceso del negocios. Las medidas habitualmente se vinculan con el niverl de granularidad.

En la fase de **DISEÑO FÏSICO** se centra todo en la selección que se hace de las estructuras para soportar el diseño. Uuno de los elementos principales de este proces son los estándares de la base de datos. Lo que es la indexación y las estrategias se determinan en este proceo.

La fase de **DISEÑO DEL SISTEMA ETL**, el sistema ETL se moldea perfectamente en esta etapa, ya que su función es extraer, transformar y cargar los datos permitiendo asi la alimentación de una base de datos en este caso DW, permite extraer, consolidar la informacion y finalmente cargar, o grabar esta información en el DW, en un formato que sea acorde para su utilización.[16]

[16] (Rivadera, 2010)

VIII. DISEÑO E IMPLEMENTACIÓN DE UNA HERRAMIENTA ETL CON BASES DE DATOS NOSQL ORIENTADAS A GRAFOS PARA GENERAR CONSULTAS OLAP EN BIG-DATA

8.1 ANÁLISIS DE REQUERIMIENTOS

1. Documento Requerimientos Funcionales y No Funcionales (DCR)

Dentro de la fase de Análisis de Requerimientos se contruyeron tanto los funcionales como los no funcionales en forma gereral. (Ver anexo 1).

2. Diagramas de Flujo (DFD)

El documento anexo Diagramas de Flujo (DFD), contiene el esquema donde se representan los algoritomos o procesos implementados en la herramienta desarrollada. (Ver anexo 2).

3. Cronograma General de Actividades

Este documento contiene las fechas desde la definición del proyecto, las fechas en que se realizaron cada una de las fases de la metodología, es decir, Analisis, Diseño, Desarrollo, Pruebas e Implementacion con cada una de las tareas que se realizaron en cada fase. (Ver anexo 4).

8.2 DISEÑO DEL SISTEMA

1. Documento de Diseño del Software (SDD)

El Docuemtno de Diseño de Software (Ver anexo 3) contiene los siguentes artefactos:

Diseño Arquitectónico

- Estilo arquitectónico

Diseño de Datos

- Diagrama Entidad Relación
- Modelado de Grafos
- Diagrama de clases (Este Diagrama debido a su complejidad se entrega dentro de el CD anexo a este documento)
- Diagrama de Casos de Uso
- Diseño DW y DataMart

GUI

- Diseño de Interfaces

2. Documento de Diccionario de Datos (DR)

Este Documento es un catálogo de los elementos en un sistema. Estos elementos se centran alrededor de los datos y la forma en que están estructurados para satisfacer los requerimientos de los usuarios y las necesidades de la organización. Aquí se encuentra la lista de todos los elementos que forman parte del flujo de datos en todo el sistema. (Ver Anexo 5)

8.3 DESARROLLO

1. Código Fuente de aplicación y base de datosDocumentación Técnica del Código (javadoc). Ver Anexo 6

8.4 PRUEBAS

1. Documentación de problemas significativos y su respectiva solución (Guion de Pruebas) Ver Anexo 7.

8.5 IMPLEMENTACIÓN

1. Documento o Manual de Usuario

Dentro de este documento se encuentrar las Instrucciones de uso para el usuario (Ver Anexo 8)

2. Manual Técnico

Describe los requerimientos de hardware y software utilizado para el desarrollo de la herramienta ETL. (Ver anexo 9)

IX. CONCLUSIONES

De acuerdo a la información recolectada y los resultados obenidos en el desarrollo de la herramienta ETL se llegaron a varias conclusiones, una de las más importantes es el hecho de que existen herramientas en la actualidad que permiten el procesamiento y analisis para datos masivos. Sin embargo, basados en las bases de datos NoSQL orientadas a grafos e integrando el software de Neo4j, no existe una aplicación que utilice el proceso ETL y el modelado de los cubos OLAP para que puedan ser visualizados posteriormente en los visores de Neo4j de una forma sencilla sin ser un experto en la materia. Dicho esto, Neo4j en conjunto con la aplicación desarrollada permite extraer, transformar, cargar, analizar y visualizar los datos de cualquier BigData de una manera más sencillla, cómoda para usuarios, sin conocimientos previos en lenguaje Cypher.

Como segunda conclusion, es inevitable el crecimiento diario de la información en la actualidad. El uso constante de internet en donde los mismos usuarios son alimentadores de grandes bodegas de datos, hace que existan nuevos métodos para depurar los datos, ayudados por nuevas tecnologías o soluciones es posible realizar una consulta en tiempo real. Las herramientas que se utilizaron en el desarrollo de la herramienta ETL fueron escogidas de acuerdo a las tendencias del mercado actual y pensando en crear algo novedoso, agradable y útil al usuario administrador que se encarga de estos datos.

En tercer lugar, al crear dimensiones de cubos e hipercubos se pueden consultar datos en cualquier manera, es decir, dependiendo de los requerimientos de búsqueda, un usuario podrá llegar a la consulta que necesita escogiendo los atriutos, asociaciandolos a su preferencia, de esta manera los datos consultados serán más exactos y acordes a sus necesidades. Es decir le da libertad al usuario de seleccionar un atributo asociarlo con otro atributo o con alguna otra dimension.

X. RECOMENDACONES

Dentro de las recomendaciones cabe destacar la importancia de seleccionar una buena metodología de trabajo, ya que dentro del desarrollo pueden surgir inconvenientes que pueden atrasar el plan inicial, mantener un cronograma y seguir unas pautas permite tener mas orden y tener un adecuado progreso en el desarrollo del proyecto.

XI. BIBLIOGRAFÍA

acenswhitepaper, 2014. *Bases de datos NoSQL. Qué son y tipos que nos podemos encontrar.* [En línea] Available at: http://www.acens.com/wp-content /images/2014/02/bbdd-nosql-wp-acens.pdf [Último acceso: 17 junio 2016].

BROOKS, C., 2014. Enterprise NoSQL for Dummies. En: Estados Unidos: MarkLogic.

BUSTILLOS, J., 2014. *Comparativas Herramientas ETL.* [En línea] Available at: http://myslide.es/software/comparativa-herramientas-etl.html [Último acceso: 17 junio 2016].

Carrasco, J. R., 2015. *Modelado y Visualización de Datos con Datos Masivos.* [En línea] Available at:http://oa.upm.es/37361/7/PFC_JORGE_RAMIREZ_CARRASCO _2015.pdf[Último acceso: 17 junio 2016].

Data, P., 2013. *El valor de la gestión de datos. Procesos ETL: Definición, Características, Beneficios y Retos.* [En línea] Available at: http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/ bid/312584/Procesos-ETL-Definici-n-Caracter-sticas-Beneficios-y-Retos [Último acceso: 17 junio 2016].

Fidelity Worldwide Investment2, 2012. *In Perspective.* [En línea] Available at: https://www.fondosfidelity.es/static/pdfs/informesfondos /Fidelity _ArgInvSXXI_Big Data_ Sept12_ES.pdf

García, D. L., 2012 -2013. *Análisis de las posibilidades de uso de Big Data en las Organizaciones.* [En línea] Available at: http://bucserver01.unican .es/xmlui/bitstream/handle/10902/4528/TFM%20-%20David% 20L%C3%B3pez %20Garc%C3%ADaS.pdf?sequence=1[Último acceso: 17 junio 2016].

Gartner, 2012. *"The Importance of 'Big Data': A Definition".* [En línea].

Ibermática, 2007. *Business Intelligence : El conocimiento Compartido.* [En línea] Available at: https://churriwifi.files.wordpress.com/2009/11/business-intelligence-ibermatica.pdf[Último acceso: 17 junio 2016].

MGI McKinsey Global Institute, Junio 2011. *Big Data: The next frontier for innovation, competition and opportunity",* s.l.: Fidelity .

Neo Technology, Inc., s.f. *Neo4j.* [En línea] Available at: http://neo4j.com/developer/graph-database/[Último acceso: 1 Agosto 2015].

neo4j, 2016. *developer-manual.* [En línea]Available at: http://neo4j.com/docs/developer-manual/current/#query-syntax[Último acceso: 17 junio 2016].

Rivadera, G. R., 2010. *La Metodología Kimball para el diseño de almacenes de datos (Data warehouses),* s.l.: s.n.

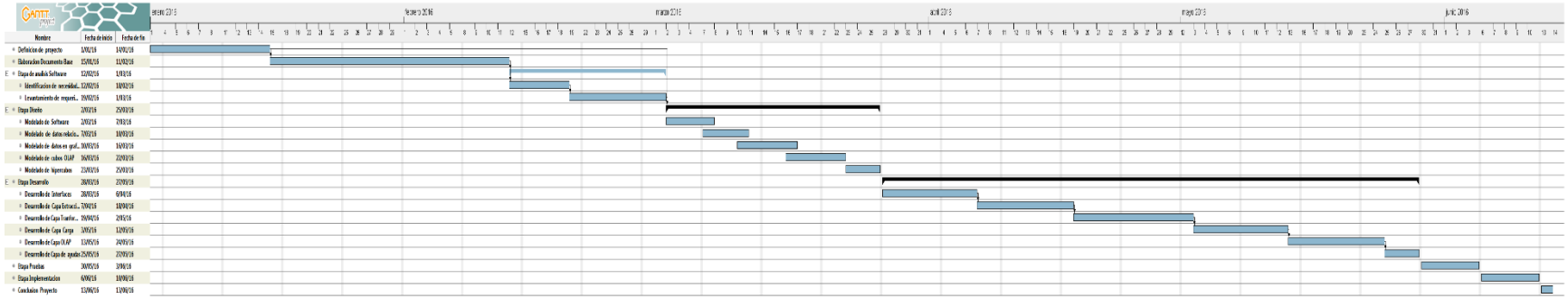
Tangient LLC, 2016. *McGus II.* [En línea] Available at: https://gusmc.wikispaces.com/3.2.1+Definiciones+y+conceptos[Último acceso: 17 junio 2016].

Tapias, D., 2014. *Proyectos de Desarrollo de Software.* [En línea] Available at: http://arantxa.ii.uam.es/~proyectos/teoria/ C5_Proyectos %20de%20desarrollo %20software.pdf[Último acceso: 17 junio 2016].TicBeat, 2012. s.l.: s.n.

VUKOTIC, Aleksa; NICKI , Watt ;, 2015. *Neo4j in action.* s.l.:Manning.

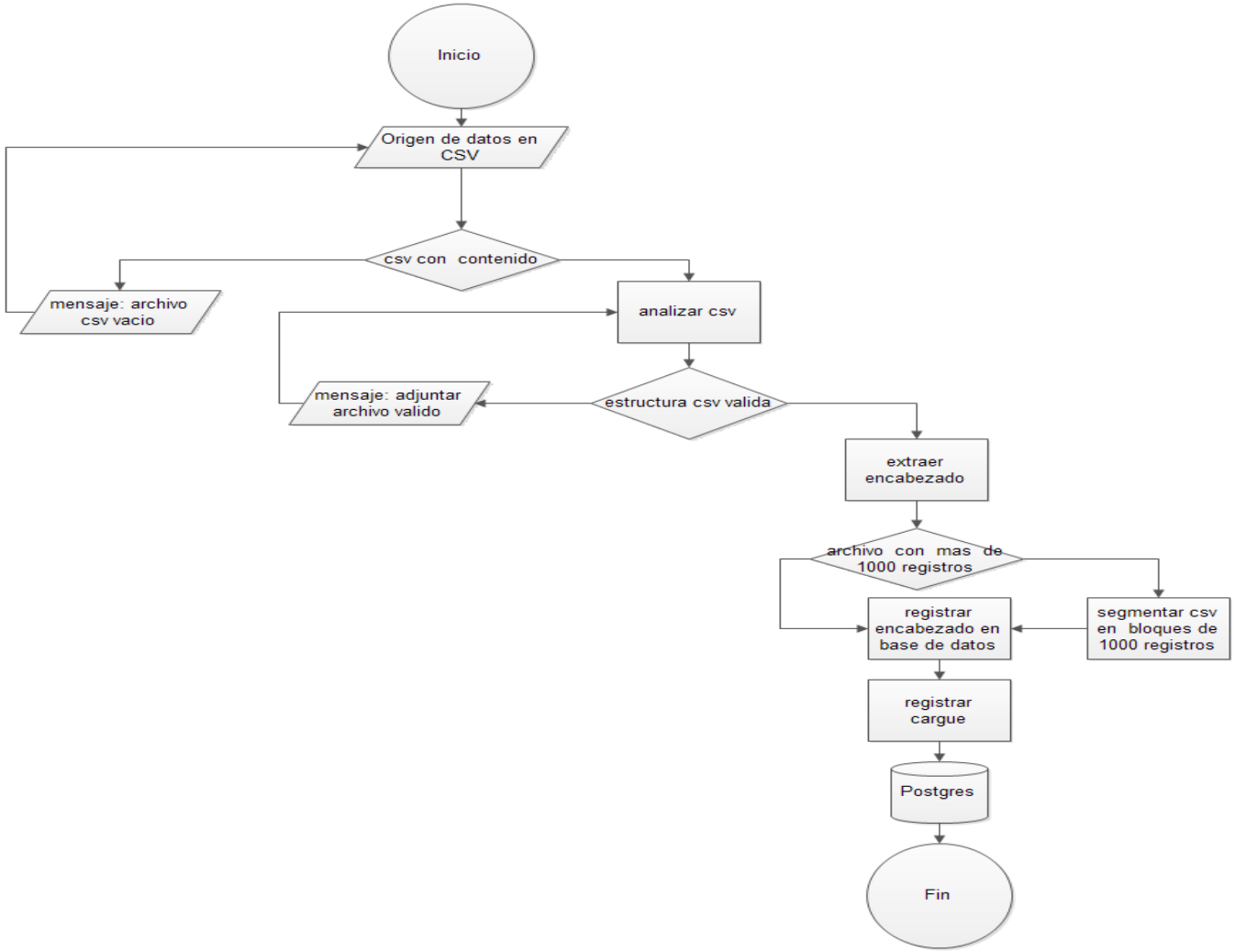
ANEXOS

ANEXO 3: Cronograma de Actividades

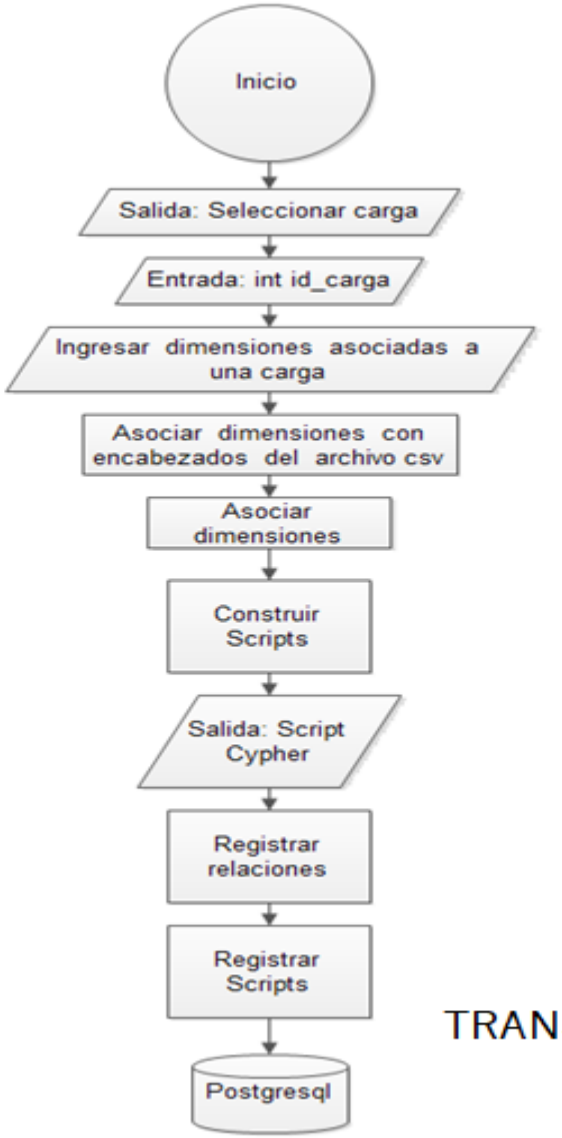


Nombre	Fecha de inicio	Fecha de fin
Definicion de proyecto	1/01/16	14/01/16
Elaboracion Documento Base	15/01/16	11/02/16
Etapa de analisis Software	12/02/16	1/03/16
Identificacion de necesidad...	12/02/16	18/02/16
Levantamiento de requeri...	19/02/16	1/03/16
Etapa Diseño	2/03/16	25/03/16
Modelado de Software	2/03/16	7/03/16
Modelado de datos relacio...	7/03/16	10/03/16
Modelado de datos en graf...	10/03/16	16/03/16
Modelado de cubos OLAP	16/03/16	22/03/16
Modelado de hipercubos	23/03/16	25/03/16
Etapa Desarrollo	28/03/16	27/05/16
Desarrollo de Interfaces	28/03/16	6/04/16
Desarrollo de Capa Extracci...	7/04/16	18/04/16
Desarrollo de Capa Tranfor...	19/04/16	2/05/16
Desarrollo de Capa Carga	3/05/16	12/05/16
Desarrollo de Capa OLAP	13/05/16	24/05/16
Desarrollo de Capa de ayudas	25/05/16	27/05/16
Etapa Pruebas	30/05/16	3/06/16
Etapa Implementacion	6/06/16	10/06/16
Conclusion Proyecto	13/06/16	13/06/16

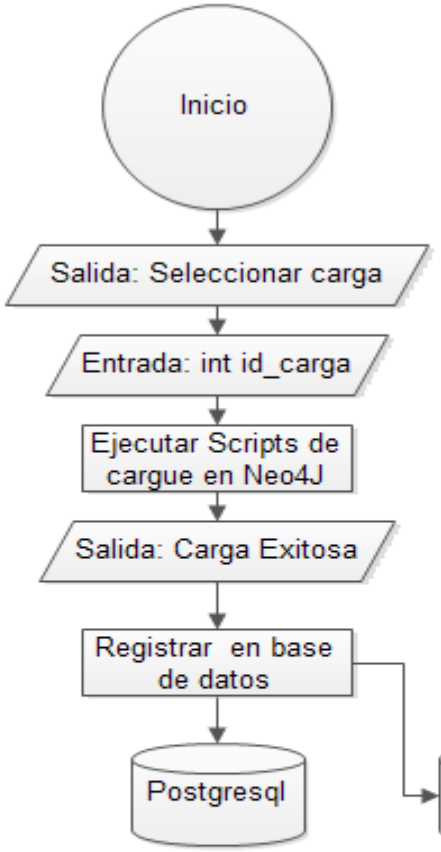
ANEXO 2 : Diagramas de Flujo (DFD)



ESTRACIÖN



TRANSFORMACION



CARGA