Diseño de Aplicaciones de Inteligencia de Negocios usando la Tecnología Big Data

Lilian Judith Sandoval

Lcda. en Administración de Empresas, Docente Investigadora, Escuela de Ingeniería en Computación, ITCA-FEPADE Sede Central. Email: lilian.sandoval@itca.edu.sv

Resumen

El presente artículo trata sobre el estudio de una nueva tecnología llamada Big Data para manejo de información, en sustitución de sistemas tradicionales de gestión de datos. Se hace un análisis de sus componentes y del nuevo software a utilizar. Específicamente, se propone la forma para realizar el diseño de una aplicación de inteligencia de negocios para la toma de decisiones a nivel gerencial, utilizando herramientas Big Data.

Palabras clave

Big data, inteligencia empresarial, bases de datos relacionales, diseño de sistemas.

Abstract

The present article concerns about the knowledge of a new technology called Big Data to manage information, replacing traditional systems for data management. An analysis of its components and new software to use is done. Specifically, the way to perform the design of business intelligence applications is proposed, and all this for the decision making at management level making use of Big Data tools.

Keywords

Big data, business intelligence, relational databases, systems design.

Introducción

Con la expansión de las redes sociales ha surgido la necesidad de manejar volúmenes de información gigantescos y variados, que las bases de datos tradicionales ya no pueden soportar. Además, de lo complicado que se ha vuelto el proceso de búsqueda de información, se ha hecho necesario pensar en estructuras de datos completamente distintas, donde la limitación de espacio no fuera más un problema.

Siempre que hacemos una búsqueda en Internet, enviamos un email, usamos un teléfono móvil, actualizamos una red social, usamos una tarjeta de crédito, activamos el GPS, hacemos uso de un seguro o hacemos una compra en línea, dejamos detrás una montaña de datos, huellas digitales y registros que ofrecen una información muy valiosa para las empresas.

Así es como se ha creado la nueva tecnología Big Data, para manejo de volúmenes de datos e interpretación de ellos para diferentes propósitos.

INTELIGENCIA DE NEGOCIOS (BI)

Es una colección de estrategias y aspectos relevantes enfocada a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización.

Las herramientas de Inteligencia de Negocios, se basan en la utilización de un sistema de información de inteligencia que se conforma con distintos datos extraídos de la información de producción, finanzas u otro tipo de información relacionada con la empresa o sus diferentes ámbitos.

La vida o el periodo de éxito de un software de Inteligencia de Negocios, dependerá únicamente del éxito de su uso en beneficio de la empresa. Si la empresa es capaz de incrementar su nivel económico, administrativo y sus decisiones mejoran la actuación de sus miembros, el software de inteligencia de negocios seguirá presente por mucho tiempo; en caso contrario, será sustituido por otro que aporte mejores y más precisos resultados.

Recepción: 29/01/2016 - Aceptación: 15/06/2016

Este documento es propiedad de la Escuela Especializada en Ingeniería ITCA-FEPADE. Derechos Reservados.

Las herramientas de Inteligencia Analítica posibilitan el modelado de las representaciones basadas en consultas para crear un Cuadro de Mando Integral que sirve de base para la presentación de informes.

De acuerdo a su nivel de complejidad se pueden clasificar las soluciones de Inteligencia de Negocios en:

- Informes predefinidos
- Informes a la medida
- Consultas (Query) / Cubos OLAP (On-Line Analytic Processing)
- Alertas
- Análisis estadístico
- Pronósticos (Forecasting)
- Modelado Predictivo o Minería de Datos (Data Mining)
- Optimización
- Minería de Procesos

FUNDAMENTOS DE BIG DATA

Big Data maneja conjuntos de datos enormes que crecen tan rápido que se vuelve muy difícil manipular y analizar a una granularidad tal donde los procesos colapsan.

Esta nueva tecnología no solo viene a resolver los problemas de almacenamiento y gestión que plantean las redes sociales, sino que también auxilia a otros sectores que también presentaban las mismas dificultades como el científico, el médico, el mercadológico, entre otros.

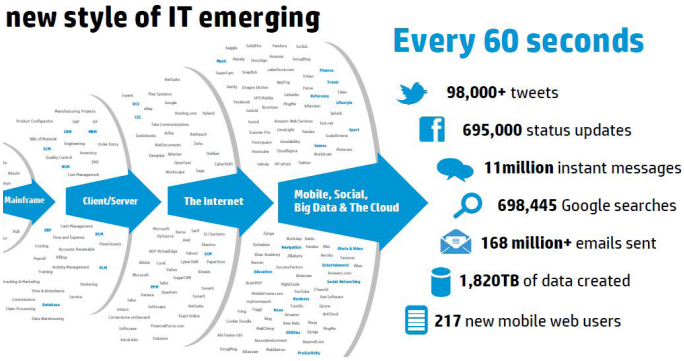

Fig. 1. Velocidad de la información.

A. Tipos de datos

Existen tres tipos de datos en el ambiente de Big Data:

- ✓ Estructurados: son datos que tienen bien definido su tipo, ya sea texto, numérico, fecha, etc. Por lo general estos datos son almacenados en tablas en un sistema de base de datos relacional.
- ✓ No estructurados: son datos que conservan el formato con el que 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, etc.
- ✓ Semiestructurados: son datos que no se limitan a tipos determinados, pero que contiene marcadores para separar los diferentes elementos. Es 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. Algunos ejemplos son HTML, XML y JSON.

Fig. 2. Características de Big Data.

B. Almacenamiento NoSQL

El término NoSQL significa Not Only SQL y son sistemas de almacenamiento que no cumplen con el esquema entidad-relación. Proveen un sistema de almacenamiento mucho más flexible y concurrente y permiten manipular grandes cantidades de información de manera mucho más rápida que las bases de datos relacionales.

Existen cuatro tipos de almacenamiento NoSQL:

- **Almacenamiento Clave-Valor (Key-Value):** son sistemas de almacenamiento donde se accede al dato a partir de una clave única. Los valores son aislados e independientes entre ellos y no son interpretados por el sistema. Pueden ser enteros, caracteres u objetos. Por otro lado, este sistema de almacenamiento carece de una estructura de datos clara y establecida, por lo que no requiere un formateo de los datos muy estricto. Son útiles para operaciones

simples basadas en claves. Un ejemplo es el aumento de velocidad de carga de un sitio web que puede utilizar diferentes perfiles de usuario, teniendo mapeados los archivos que hay que incluir según el Id de usuario y que han sido calculados con anterioridad. Cassandra es la tecnología de almacenamiento Clave-Valor más reconocida por los usuarios.

- **Almacenamiento Documental:** bases de datos con este sistema de almacenamiento guardan un gran parecido con las bases de datos Clave-Valor, diferenciándose en el dato que guardan. Si en la anterior no requería una estructura de datos concreta, en este caso sí se guardan datos semiestructurados. Estos datos pasan a llamarse documentos, y pueden estar formateados en XML, JSON o en el formato que acepte la misma base de datos. Un ejemplo de este tipo de almacenamiento es un blog: se almacena el autor, la fecha, el título, el resumen y el contenido del post. CouchDB o MongoDB son las bases de datos documentales más conocidas.

- **Almacenamiento en Grafo:** las bases de datos en grafo rompen con la idea de tablas y se basan en la teoría de grafos, donde se establece que la información son los nodos y las relaciones entre la información son las aristas. Relacionan grandes cantidades de datos que pueden ser muy variables. Por ejemplo, los nodos pueden contener objetos, variables y atributos diferentes unos de los otros. Las uniones se sustituyen por recorridos a través del Grafo y se guarda una lista de adyacencias entre los nodos. Un ejemplo es el Facebook, donde cada usuario es un nodo que puede tener aristas de amistad con otros usuarios, o aristas de publicación con nodos de contenidos. Soluciones como Neo4J y GraphDB son las más conocidas dentro de las bases de datos en Grafo.

- **Almacenamiento Orientado a Columnas:** este sistema de almacenamiento es similar al Documental. Su modelo de datos es definido como "un mapa de datos multidimensional poco denso, distribuido y persistente. Se orienta a almacenar datos con tendencia a escalar horizontalmente, por lo que permite guardar diferentes atributos y objetos bajo una misma clave". A diferencia del Documental y el Key-Value, en este caso podremos almacenar varios atributos y objetos, pero no serán interpretables directamente por el sistema. Permite agrupar columnas en familias y guardar la información cronológicamente, mejorando el rendimiento. Esta tecnología se utiliza en casos de contar 100 o más atributos por clave. Su precursor es BigTable de Google, pero han aparecido nuevas soluciones como HBase o HyperTable.

MINERÍA 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. Utiliza el análisis matemático para deducir los patrones recurrentes y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiados datos.

Los modelos de Minería de Datos se pueden aplicar en escenarios como los siguientes:

- Pronósticos de ventas
- Cálculo del riesgo en el lanzamiento de productos nuevos
- Análisis de comportamiento del mercado
- Predicción de posibles cambios de tendencias.
- Otros

Relación de la Minería de Datos y Big Data

La diferencia fundamental de la Minería de Datos y Big Data es la velocidad de resolución para analizar y resolver las situaciones en tiempo real en todos aquellos ámbitos donde se manejen datos complejos, como en la banca e instituciones financieras y de seguros, investigación de mercados, medicina, educación, biología, procesos industriales, telecomunicaciones, transacciones por Internet y todo lo que tenga relación con él.

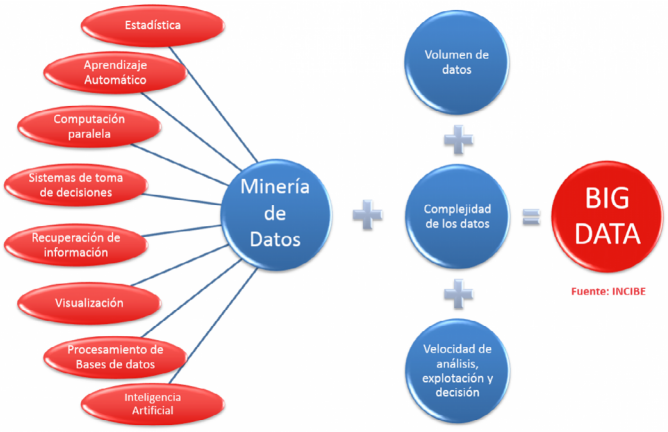

Fig. 3. Componentes de Big Data.

A. Apache Hadoop

Hadoop es un framework que permite el procesamiento distribuido de grandes volúmenes de datos a través de clusters de computadoras que utilizan modelos de programación sencilla. Está diseñado para escalar de servidores individuales a miles de computadoras alrededor del mundo. En lugar de depender del hardware para ofrecer alta disponibilidad, el mismo Hadoop está diseñado para detectar y manejar las fallas en la capa de aplicación, por lo que la entrega de un servicio de alta disponibilidad en la parte superior de un grupo de servidores, cada uno de los cuales puede ser propenso a fallos.

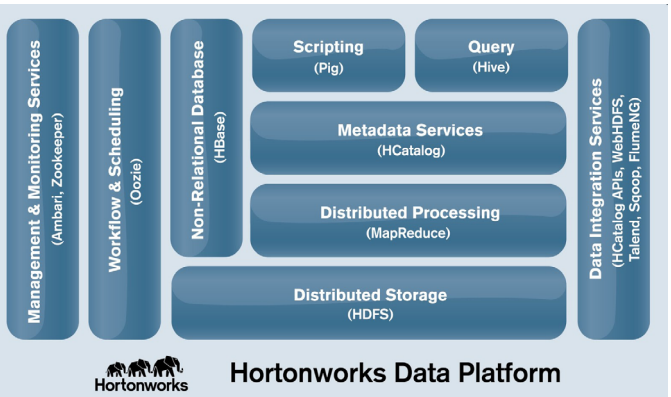

Fig. 4. Plataforma de Big Data.

Big Data incluye los siguientes módulos:

- *Hadoop Common:* está conformado por las utilidades comunes que apoyan los otros módulos de Hadoop.
- *Hadoop Distributed File System (HDFS):* es un sistema de archivos distribuido que permite el acceso de alto rendimiento a los datos de la aplicación.
- *Hadoop YARN:* es un framework para la planificación de tareas y gestión de recursos de clúster.
- *Hadoop MapReduce:* es un sistema basado en YARN para el procesamiento paralelo de grandes conjuntos de datos.

B. Base de Datos

- *HBase:* es una base de datos columnar (column-oriented database) que se ejecuta en HDFS. Es una base de datos distribuida y usa el concepto de BigTable que permite escalar casi linealmente con solo agregar más servidores. HBase permite que muchos atributos sean agrupados llamándolos familias de columnas, de tal manera que los elementos de una familia de columnas son almacenados en un solo conjunto. Eso es distinto a las bases de datos relacionales orientadas a filas, donde todas las columnas de una fila dada son almacenadas en conjunto.
- *Apache Cassandra:* es una base de datos NoSQL distribuida y basada en un modelo de almacenamiento de "Clave-Valor", de código abierto que está escrita en Java. Permite grandes volúmenes de datos en forma distribuida. Por ejemplo, lo usa Twitter para su plataforma. Su objetivo principal es la escalabilidad lineal y la disponibilidad. La arquitectura distribuida de Cassandra está basada en una serie de nodos iguales que se comunican con un protocolo P2P, con lo que la redundancia es máxima. Cassandra ofrece soporte robusto para múltiples centros de datos, con la replicación asincrónica sin necesidad de un servidor maestro que permite operaciones de baja latencia para todos los clientes.
- *MongoDB* (de la palabra en inglés "humongous" que significa enorme): es un sistema de base de datos NoSQL orientado a documentos, desarrollado bajo el concepto de código abierto. Forma parte de la nueva familia de sistemas de base de datos NoSQL. Guarda estructuras de datos en documentos tipo JSON con un esquema dinámico (MongoDB llama ese formato BSON), haciendo que la integración de los datos en ciertas aplicaciones sea más fácil y rápida. El desarrollo de MongoDB empezó en octubre de 2007 por la compañía de software 10gen. Esta base de datos se utiliza mucho en la industria y MTV Network, Craiglist y Foursquare.

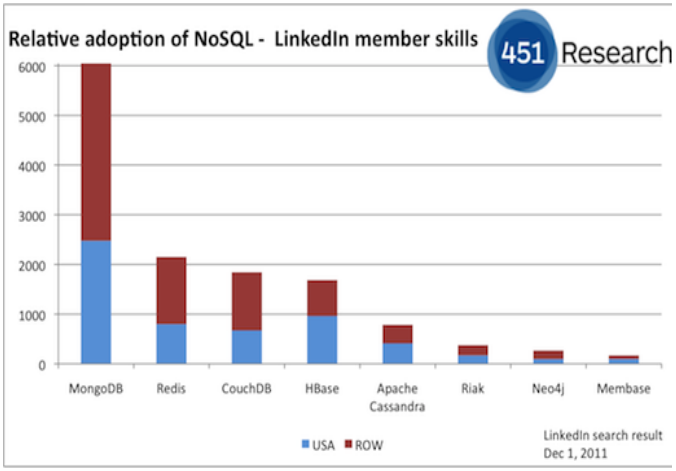

Fig. 5. Comparativo de la aceptación de MongoDB con las demás bases de datos Big Data.

C. Hive: Estructura Data Warehouse

Hive es una infraestructura de Data Warehouse que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar a SQL llamado Hive Query Language (HQL). Estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el cluster de Hadoop.

Hive abre el gran ecosistema Hadoop Datos para no programadores debido a sus capacidades de tipo SQL y la funcionalidad de la base de datos similares. A menudo se describe como una infraestructura de almacenamiento de datos construida sobre Hadoop. Esta es una declaración verdadera parcialmente - ya que se pueden transformar los datos de origen en un esquema en estrella - pero es más sobre el diseño de la tecnología cuando se crea un hecho de tablas y tablas de dimensiones.

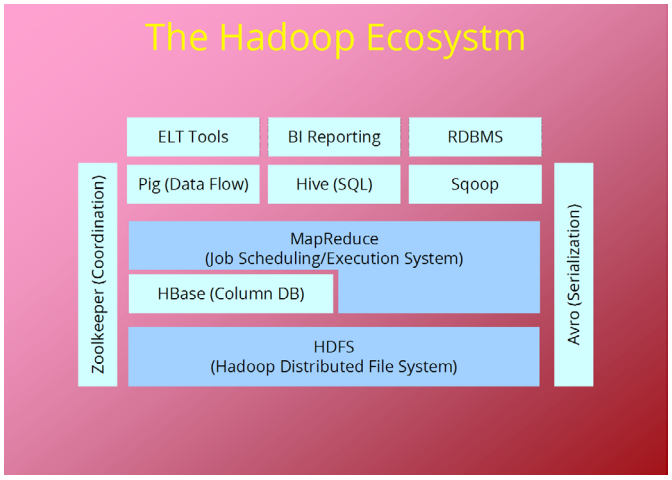

Fig. 6. El ecosistema Hadoop.

D. Pentaho: Extracción, Transformación y Carga de Datos (ETL)

Pentaho BI Suite es un conjunto de programas libres para generar inteligencia de negocios. Incluye herramientas integradas para generar informes, Minería de Datos, ETL y otros.

ETL es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos o "data warehouse" para analizar o en otro sistema operacional para apoyar un proceso de negocio.

- *Extracción:* la primera parte del proceso ETL consiste en extraer los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir bases de datos no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

- *Transformación:* esta fase aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos.

- *Carga:* esta fase es el momento en el cual los datos de la fase anterior **(transformación)** son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunas bases de datos se sobrescribe la información antigua con nuevos datos. Los **data warehouse** mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un valor a lo largo del tiempo. La fase de carga interactúa directamente con la base de datos de destino. Al realizar esta operación se aplicarán todas las restricciones y triggers que se hayan definido.

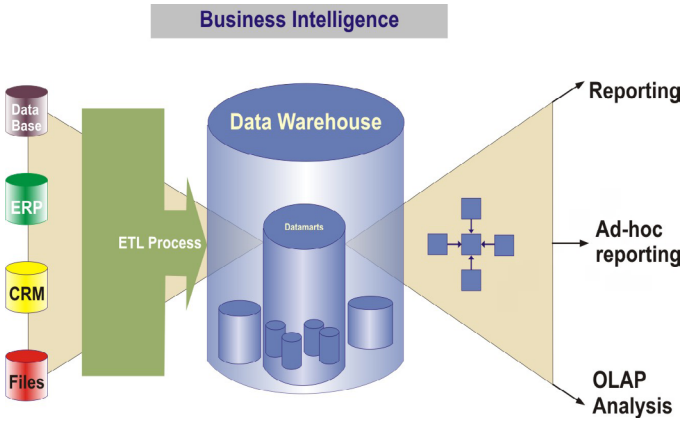

Fig. 7. Proceso de extracción, transformación y carga de datos.

E. Panel de Inteligencia de Negocios (Dashboards)

En la tecnología de la información, un panel de inteligencia de negocios o panel de control es una interfaz

de usuario que, pareciéndose un poco el tablero de un automóvil, organiza y presenta la información de una manera que es fácil de leer. Hasta cierto punto, la mayoría de las interfaces gráficas de usuario se asemejan a un tablero de instrumentos. Sin embargo, algunos desarrolladores de productos emplean conscientemente esta metáfora para que el usuario reconozca al instante la similitud. Es una herramienta de visualización de datos que muestra el estado actual de métricas e indicadores clave de rendimiento para una empresa en una sola pantalla. Las características esenciales de un producto de tablero de mandos de BI incluyen una interfaz personalizable y la capacidad de reflejar datos en tiempo real de múltiples fuentes.

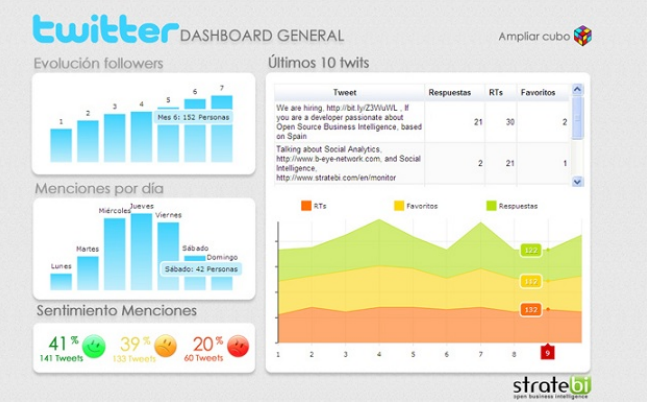

Fig. 8. Cuadros de mando de Facebook y Twitter.

Conclusiones

Se realizó una investigación sobre los diferentes términos de Big Data y el impacto que en la actualidad está teniendo esta nueva tecnología y la necesidad de introducirse en ella, ya que en poco tiempo ha abarcado la mayoría de los ámbitos de la sociedad.

Se presentaron las diferentes herramientas que integran la plataforma Big Data, que incluyen el sistema distribuido de archivos, las diferentes bases de datos, la estructura de data warehouse, el sistema de extracción, transformación y carga de datos y los dashboards. Todos, necesarios para diseñar una aplicación de Inteligencia de Negocios.

Referencias

LIBROS

[1] V. Mayer Schonberger, K. Cukier. Big Data: La Revolución de los Datos Masivos. Editorial Turner. 2013.

[2] O'Really Media Inc. Big Data Now. Kindle Edition. 2012.

ARTÍCULOS PRESENTADOS EN CONFERENCIAS

[3] L. J. Sandoval. "Tools for Design of Knowledge Management Systems Based on Business Intelligence" en Proc. "Proceedings of the 2014 IEEE Central America and Panama Convention (CONCAPAN XXXIV)". 2014. IEEE.

[4] R. Hecht, S. Jablonski."NoSQL Evaluation, a use case oriented survey». International Conference on Cloud and Service Computing. 2011.

REPORTES TÉCNICOS

[5] D. Lopez García. Analysis of the possibilities of use of Big Data in organizations. (2012-2013).

TUTORIALES

[6] Microsoft Developer Network. Tutorial Básico de Minería de Datos.
Disponible: http://msdn.microsoft.com/es-es/library/ms167167.aspx

[7] IBM Developer Works. Que es Big Data?
Disponible:https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

[8] Brandchats. Tipos de datos que comprende el Big Data.
Disponible: http://www.brandchats.com/7-tipos-de-datos-que-comprende-el-big-data/

[9] NoSQL.
Disponible: http://nosmoke.cycle-it.com/2014/03/31/nosql/

[10] Apache. Hive.
Disponible: https://hive.apache.org/

[11] IBM. Hive Warehouse.
Disponible: http://www.ibm.com/developerworks/library/bd-hivewarehouse/

[12] Incibe. Minería de Datos.
Disponible: https://www.incibe.es/blogs/post/Empresas/BlogSeguridad/Articulo_y_comentarios/mineria_datos_big_data_seguridad

[13] Fundación Big Data. Big Data y la Ciberseguridad: El Nuevo Futuro.
Disponible: http://fundacionbigdata.org/category/articulos_big_data/

[14] Wikipedia. Dashboards.Management Information Systems.
Disponible: https://en.wikipedia.org/wiki/Dashboard_(management_information_systems)

[15] Pentaho. Tutorial de Integración de Datos.
Disponible: http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+(Kettle)+Tutorial

[16] Wikipedia. ETL.
Disponible: https://es.wikipedia.org/wiki/Extract,_transform_and_load

[17] The Apache Software Foundation. Apache Hadoop.
Disponible: https://hadoop.apache.org/

[18] IBM developerWorks. ¿Qué es Big Data?
Disponible:https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Este documento es propiedad de la Escuela Especializada en Ingeniería ITCA-FEPADE. Derechos Reservados.