With its speed and variety of information treated would Big Data end Data warehousing?*

Kamagate Azoumana**
Universidad Simón Bolívar, Barranquilla

Recibido: 4 de septiembre de 2013 Aceptado: 27 de noviembre de 2013

Con la velocidad y la variedad de los datos que trata, el *Big Data* pondrá fin a los de almacén de datos (*Data warehousing*)?

Key words:
Big Data, Data warehouse, Hadoop, NoSQL, Data Integration.

Abstract
This article is aimed at key decision makers, database administrators, integrated software vendors as well as those people who face the challenge of analyzing large quantities of data generated by a business or an information system resource. This has been written with the assumption of a basic understanding of the covered concepts like Big Data or Data warehousing. Are not Big Data and all the technologies that are related, such as Hadoop, a new opportunity for the world of data integration and the data warehousing?

Palabras clave:
Big Data, Almacén de datos, Hadoop, NoSQL, Integración de Datos.

Resumen
Este artículo está escrito para los tomadores de decisiones, los administradores de bases de datos, proveedores de *software* integrado y otras personas que se enfrentan al reto de hacer el análisis de grandes cantidades de datos generados por un sistema de información o de negocios. Ha sido escrito con la asunción de una comprensión básica de los conceptos tratados como *Big Data*, *Data warehouse*. ¿*Big Data* y todas las tecnologías que se relacionan, como Hadoop, no son una nueva oportunidad para el mundo de la integración de datos y el almacén de datos?

Referencia de este artículo (APA): Kamagate, A. (2014). With its speed and variety of information treated would Big Data end Data warehousing? En Revista *Educación y Humanismo, 16*(26), 122-128.

* Este artículo es producto del proyecto de investigación Estudio del impacto de data warehouse y Business Intelligence en el desempeño de las empresas colombianas.

** Magíster en Tecnología de Información. Docente investigador en Ingeniería de Sistemas de la Universidad Simón Bolívar. Grupo de investigación Ingebiocaribe. Correo electrónico: kazoumana@unisimonbolivar.edu.co

Introduction

The Oxford dictionary defines Big Data as data set too large and complex to manipulate or to question with standard methods or tools. Big Data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it. The value of Big Data to an organization falls into two categories: analytical use and enabling new products. Big Data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analyzing shoppers' transactions and social and geographical data. Being able to process every item of data in reasonable time, removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports.

A data warehouse is a computer database that collects, integrates and stores an organization's data with the aim of producing accurate and timely management information and supporting data analysis. It explains the importance of good data warehousing and covers the process of building such a specialized database using open source technologies.

The universally accepted definition of a data warehousing developed by Bill Inmon in the 1980s is "a subject-oriented, integrated, time variant and non-volatile collection of data used in strategic decision making". Data warehousing acts as the central point of data integration, which is the first step toward turning data into information. Due to this enterprise focus, it serves the following purposes: First, it delivers a common view of enterprise data, regardless of how it may later be used by the consumers. Since it is the common view of data for the business consumers, it supports the flexibility in how the data is later interpreted (analyzed). The data warehouse produces a stable source of historical information that is constant, consistent, and reliable for any consumer. According to the definitions there are some similarities between Big Data and data warehousing.

What does Big Data look like?

As an all term catch "Big Data" can be pretty nebulous, in the same way that the term "cloud" covers diverse technologies. Input data into big data systems could be chatter from social networks, web server logs traffic flow sensors, satellite imagery, broadcast audio streams, banking transactions, Rock music MP3s, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, financial market data and the list goes on. Are these all really the same thing?

To clarify matters, the three "V's" of *volume, velocity, and variety* are commonly used to characterize different aspects of Big Data. They're helpful lenses through which the nature of data and software platforms available to be used can be seen and understood. You most likely will

contend with each of the "V's" to one degree or another.

The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics. Having more data beats out having better models: simple bits of math can be unreasonably effective given large amounts of data. If you could run that forecast taking into account 300 factors rather than 6, could you predict demand better.

Many companies already have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it. Assuming that the volumes of data are larger than those conventional, relational database infrastructures can cope with; processing options which break down broadly into a choice between massively parallel processing Architectures, data warehousing or databases such as Greenplum and Apache Hadoop-based solutions.

At its core, Hadoop is a platform for distributing computing problems across a number of servers. First developed and released as open source by Yahoo, it implements the MapReduce approach pioneered by Google in compiling its search indexes. Hadoop's MapReduce involves distributing a dataset among multiple servers and operating on the data: the "map" stage. The partial results are then recombined: the "reduce" stage Volume:

One of the most well-known Hadoop users is Facebook, whose model follows this pattern. A MySQL database stores the core data. This is then reflected into Hadoop, where computations occur, such as making recommendations for you, based on your friends' interests. Facebook then transfers the results back into MySQL, for use in pages served to users.

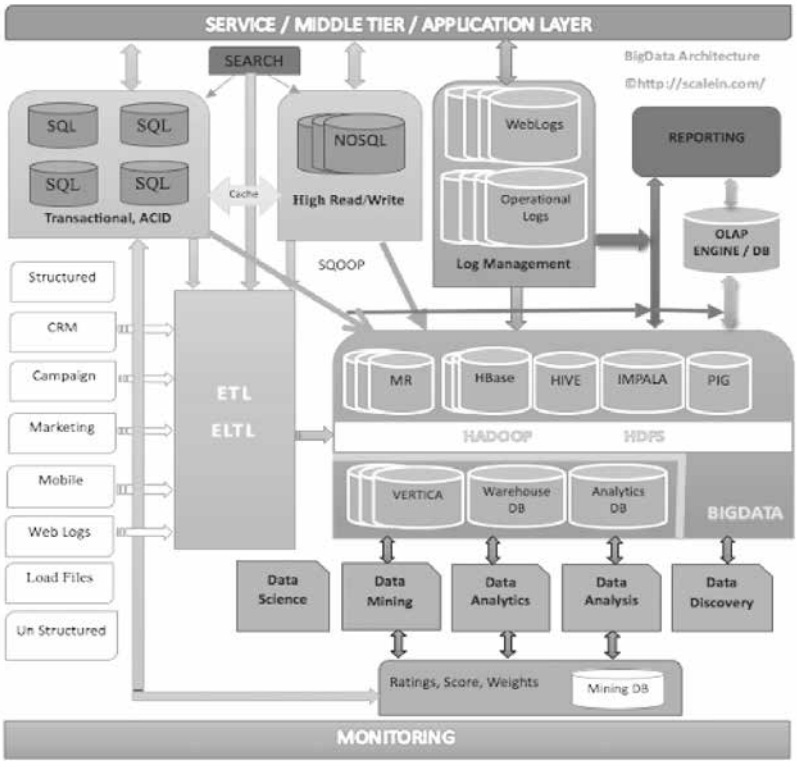
Velocity

The importance of data's velocity**,** the increasing rate at which data flows into an organization, has followed a similar pattern to that of volume. Problems previously restricted to segments of industry are now presenting themselves in a much broader setting. Specialized companies such as financial traders have long turned to systems that cope with fast moving data to their advantage. Now it's our turn. Product categories for handling streaming data divide into established proprietary products such as IBM's InfoSphere and less polished streams which are still emergent open source frameworks originating in the web industry: Twitter's Storm and Yahoo S4. These databases form part of an umbrella category known as NoSQL, used when relational models are not the right fit.

Variety

Rarely does data present itself in a form perfectly ordered and ready for processing. A common theme in Big Data systems is that the source data is diverse, and does not fall into neat relational structures. It could be a text from the social networks, an image data, a raw feed directly from a sensor source. None of these things come readily available for integration into an existing application.

Architecture of Big Data



What does data warehouse look like?

A data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, online analytical processing (OLAP) and data mining capabilities, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users. According to Inmon data warehousing is subject-oriented, integrated, time-variant, and a non-volatile collection of data in support of management's decision making process. It provides information from historical perspective e.g. past 5-10 years. Every key structure contains either implicitly or an explicit element of time.

Integrated Data Warehousing is built by integrating multiple heterogeneous sources.

Integration Data Proprocessing is applied to ensure consistency.

Subject-oriented Data Warehousing is organized around such areas such as sales, product and customer-based information.

It focuses on modeling and analysis of data for decision makers.

Non-volatile data cannot be updated. Data warehousing requires two operations in data, accessing Initial loading of data, access of data. From a conceptual perspective, data warehousing store data snapshots and additional data collected from a variety of source systems. Data warehouses encompass a variety of subject areas.

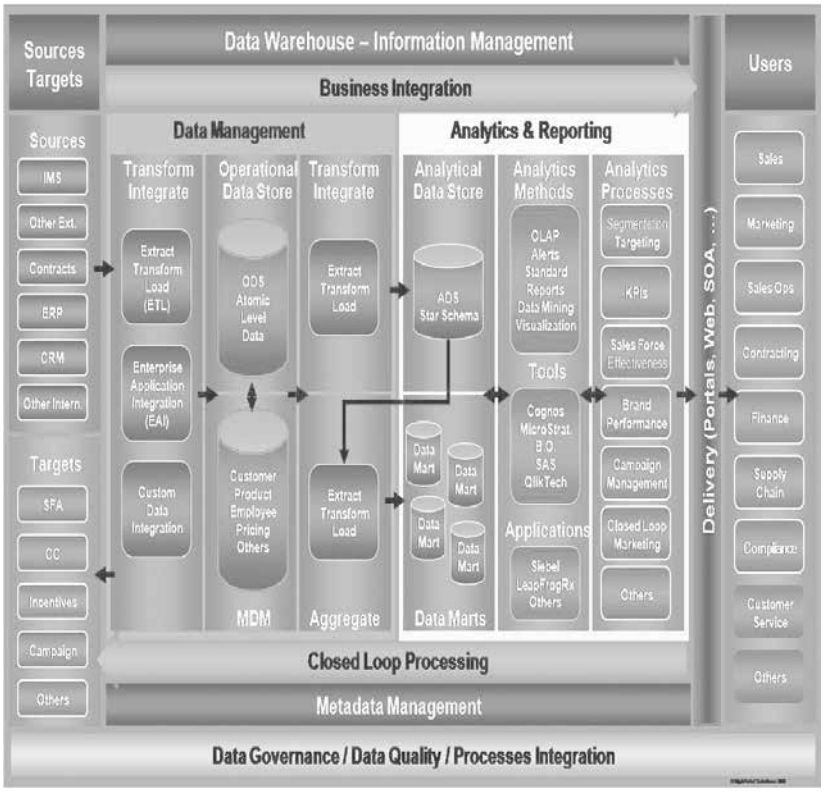
Each of these source systems could store the same data in different formats, with different editing rules and different value lists. For example, gender code could be represented in three separate systems as male/female 0/1, and M/F respectively; dates might be stored in a year/ month/day, month/day/year, or dd/mm/yy format. In the United States "03062015" could represent March 6, 2015, while in the United Kingdom it might represent June 3, 2015.

Data warehouses involve a long-term effort and are usually built in an incremental fashion. In addition to adding new subject areas, at each iteration, the breadth of data content of existing subject areas is usually increased as users expand their analysis and their underlying data requirements. Users and applications can directly use the data warehouse to perform their analysis. Alternately, a subset of the warehoused data, often relating to a specific line-of business and/ or a specific functional area, can be exported to another, smaller data warehouse, commonly referred to as a data mart. Besides integrating and cleansing an organization's data for better analysis, one of the benefits of building a data warehouse, is that the effort initially spent to populate it with complete and accurate data content, further benefits any data marts that is sourced from the data warehouse.

The Multipurpose Nature of the Data Warehouse

Hopefully by now, you have a good under-

Architecture of Data warehouse



standing of the role data warehousing plays in your Business intelligent environment.

It not only serves as the integration point for your operational data, it must also serve as the distribution point of data to the hands of the various business users. If the data warehouse is to act as a stable and permanent repository of historical data for use in your strategic BI applications, it should have the following characteristics:

It should be enterprise focused. The data warehouse should be the starting point for all data marts and analytical applications; thus, it will be used by multiple departments, maybe even multiple companies or subdivisions.

Data warehouse vs. Big Data

I think the data from Big Data represents an additional source of data that must be integrated in order to provide decision makers with a single version of the truth. We're talking about Big Data and no information is needed by policy makers today than yesterday, than reliable methods to transform data into information.

The Integrated Data Hub, is a set of best practices and reference architecture, as well as advice in terms of data modeling to carry out this problematic data transformation, big or not, for reliable information for decision making.

Volume, Velocity and Variety are additional constraints that the data warehouse architect must take into account. If your data architecture is able to handle data whose structure may vary, and fast enough to handle the new data flow; thus, I do not think the volume is a problem. The new databases are quite capable of handling large volumes of data in real time. Whether Oracle, IBM and Microsoft have Appliances or not, there are no technical constraints that may prevent us today to treat these real-time Big Data. Do not forget that besides being able to deliver integrated and validated, the Data Warehouse is supposed to deliver comparisons with historical data. To do this, we will always need a place to store these time slots and apply business rules that calculate indicators for decision-making.

Conclution

Big Data and all the technologies that are related, such as Hadoop, NoSQL and other, are a new opportunity for world data integration. Big Data represents an additional source of data that must be integrated in order to provide decision makers with a single version of the truth. In summary, the Data Warehouse is not dying, but evolving and adapting. The new Data Warehouse will be modified, in which structured and unstructured data are stored and managed where it is the most coherent version using the best technology and especially by adopting an extended architecture and well-coordinated as shown in the architecture of Big Data.

References

Christian Borgelt Department of Knowtedge Processing and Language Engineering-School of Computer Science, Otto-

vonGuerlcke-University of Magdeburg 2005. McGraw-Hill, 2000.

College student dropout syndrome. *American Educational Research Journal*, 35-64.

Faculty of Computer and Slovenia Information Science, University of Liubliana. Orange, fruitful and fun. http://wvw.ailab.si/orange, 2007.

Hurwitz, J. S., Fern Halper, A. F. (2013). *Big Data for Dummies*. Published by John Wiley & Sons, Inc.

Imhoff, C., Galemmo, N., Geiger, J. G. (2003). *Mastering Data warehouse Design Published simultaneously in Canada*.

Jansen, M. (October, 2006). *Building data warehouses using open source technologies*.

Kenneth, C. & Jane, P. (2005). *Administración de la Información y toma de decisiones, Resúmenes de los principales capítulos del libro, Management Information Systems Organization and Technology*. Documento. Chile: Universidad de Taparaca.

Kimball, R. & Margy, R. (2002). *The Data Warehouse Toolkit The Complete Guide to Dimensional Modefing*. McGraw-Hill. Second Editan.

Núñez, F. A., Lugones, F. A. (2001). *Modelos de Negocios en Internet visión poscrisis*. McGraw-Hill. 384 p.

Rakotomalala. Tanagra project. http://chiroubte.univ- lyon2.fr/ricco/tanagra/en/tanagra.html 2007.

Zadrozny, P. & Raghu Kodali, R. (2012). Big Data Analytics Using Splunk.