International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Big Data vs. Data Warehousing

Mason White
Graduate Student, Regis University, 3333 Regis Blvd, Denver, CO 80221, USA

ABSTRACT

In today's data driven world the terms Big Data and data warehousing get thrown around a lot. Both data warehousing and Big Data are two complex and seemingly similar concepts. However, the two concepts could not be more different. This short paper is meant to compare and contrast Big Data and data warehousing in an attempt to clarify the differences between the two in plain English. Above all else this work will attempt to fill a gap that compares and contrasts what appears to be two similar concepts.

Keywords: Big Data, Data Warehousing, Hadoop, Inmon, Big Data vs. Data Warehousing

I. INTRODUCTION

Today's world is ruled by data. Without data an organization or corporation simply cannot function. However, having data is not enough. If an organization or corporation does not have an adequate means to makes sense out of the data they might as well not even have the data at all. In the struggle to find better ways to harness and make sense out of data organizations and corporations often rely on Big Data analytics, data warehousing, or a combination of the two to help make sense out of the stored data.

II. PROBLEM STATEMENT

Big data and data warehousing seem very similar at first glance. They both hold large amounts of data and they both allow users to essentially gain something useful out of the stored data. Due to the complexity and the perceived similarities of the two concepts the differences between Big Data and data warehouses are often confused to a point where Big Data is sometimes seen as a replacement for data warehousing. However, this is not the case. To say that Big Data is a replacement for data warehousing is to say apples are a replacement for oranges. To understand the differences between the two concepts one must first understand what the two concepts are.

III. DATA WAREHOUSING

A data warehouse is a repository of potentially valuable past and current data that is used to support decision makers [1][3]. A data warehouse can best be described as a melting pot for data. At its heart, a data warehouse is an architecture [2]. In other words, it is a way of organizing data. There are two common architectures for implementing a data warehouse. An organization or corporation can choose to implement either the Inmon approach or the Kimball approach [1].

Both the Kimball and the Inmon approach to data warehousing are two different ways of accomplishing the same goal, to provide decision makers with the right data to make a proper decision. However, they are both governed by different design philosophies that go beyond the scope of this paper. Therefore, only the Inmon approach will be explored to illustrate the concept of data warehousing.

A. ETL

Before one can understand the Inmon architecture one must first understand how data is integrated into a data warehouse. Data in a data warehouses comes from many other storage systems. In order to upload the data into the warehouse a process known as ETL has to take place. ETL stands for Extract, Transfer, and Load. The ETL process consists of extracting data from an OLTP source, transforming the extracted data to match the data warehouse's schema, and lastly loading the transformed data into the data warehouse [3]. ETL is the costliest steps in developing a data warehouse. Up to 70% of the data warehouse's development time can be spent on ETL alone [4].

B. Inmon Approach

The Inmon approach uses what is known as the top-down approach [1]. In other words, data is extracted from the original sources, transformed, and loaded into the data warehouse. From the data warehouse the data is again extracted and loaded into the data marts. Fig.1 is the usual depiction of the flow of an Inmon data warehouse. In practice there can be as many data sources and data marts as needed to accomplish the system's purpose.

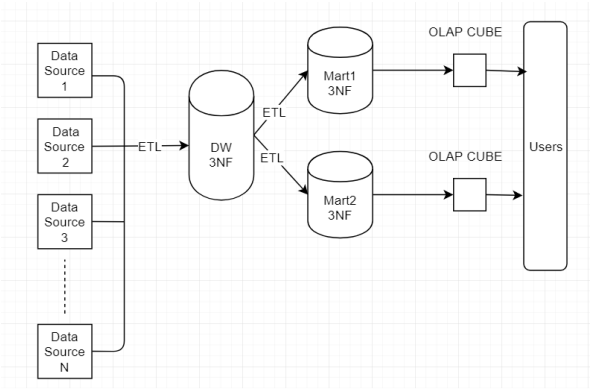

Figure 1. Inmon DW Diagram.

The Inmon architecture has an organization wide focus [5]. Inmon describes his data warehouse approach as a "single version of the truth" [5]. By using the Inmon approach decision makers can rest assure that they are using the most accurate information that reflects most if not the entire organization or corporation. In this approach the data warehouse is a relational database implemented in a traditional RDBMS (Relational DataBase Management System) like MySQL, SQL-Server, Oracle, DB2, or the like.

There are a few drawbacks to the Inmon architecture though. The Inmon data warehouse is modelled using the ER technique [4]. The purpose of modelling the data warehouse using the ER technique is to seek to eliminate redundant data [6]. However, this approach will require a high level of expertise to model the data warehouse as well as a high up-front time and cost commitment [4].

IV. BIG DATA

Big Data is an IT buzzword that means different things to different people. Big Data is governed by the 3Vs which are Volume, Velocity, and Veracity. Volume is the very large amount of data that is generated, Velocity is the analysis speed of the data stream, and Veracity is the quality of the data [7]. Some works like to expand the concept of 3Vs to 5Vs by adding Value and Variety [8]. Whether one considers Big Data to be governed by 3Vs or 5Vs, Big Data is the analysis of large volumes of high quality data.

A. Data Structure

Big Data is usually linked to unstructured data. This means that there is no relation between pieces of data. As such unstructured data cannot live in a relational database. Unstructured data can be thought of as Twitter post, E-mails, cell phone records, and so on. However, Big Data can have any structure [9]. For example, data can be structured, semi-structured, or unstructured [9]. It should never be assumed that Big Data is only unstructured data.

B. Hadoop

Normally, when a person refers to Big Data they are usually referring to Hadoop [2]. In today's world Big Data is synonymous with Hadoop and vice versa. Hadoop at its core allows for the distributed processing of large data sets set across a cluster of computers [10].

Hadoop is not a single technology, instead Hadoop should be thought of as a software ecosystem. The core of the ecosystem is Common, the HDFS (Hadoop Distributed File System), YARN, and MapReduce [10]. However, software packages like Pig, Hive, Spark, Cassandra, HBase, and so on are also very common.

In terms of this paper, Hive is a very interesting Hadoop component as it is a data warehouse tool. Hive provides a SQL like language and a relational model [8]. Hive is a system that sits on top of Hadoop that is designed to process structured data [11]. However, Hive should not be considered a relational database [11]. Hive demonstrates that there are differences between Big Data and data warehousing. Due to the nature of Hive it can be seen certain applications can use both Big Data and data warehousing.

V. CONCLUSION

In conclusion, this paper has been a brief overview of data warehousing and Big Data. Big data is not and can never be a replacement for data warehousing. Though both concepts do seem similar at first glance, they are as similar to each other as apples are to oranges. The best way to think of Big Data is as a technology that analyses data where a data warehouse is an architecture [2]. Technologies like Hive not only demonstrate that Big Data cannot be a replacement for data warehouses but that Big Data and data warehousing can be used together.

VI. REFERENCES

[1]. Sharda, R., Delen, D., & Turban, E (2014). Business intelligence and analytics: Systems for decision support. Boston, Mass. u.a: Pearson

[2]. Inmon, B. (2013, November 7). Big Data implementation vs. data warehousing. Retrieved from http://www.b-eye-network.com/view/17017

[3]. El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. Journal of King Saud University-Computer and Information Sciences, 23(2), 91-104.

[4]. Jukic, N. (2006). Modeling strategies and alternatives for data warehousing projects. Communications of the ACM, 49(4), 83-88.

[5]. Inmon, W. H. 2010. "A tale of two architectures" Retrieved from http://scholar.googleusercontent.com/scholar?q=cache:xb2NbTsepawJ:scholar.google.com/+a+tale+of+two+architectures&hl=en&as_sdt=0,44 Accessed

[6]. Kimball, R (1997, August 2). A dimensional modeling manifesto. Retrieved from https://www.kimballgroup.com/1997/08/a-dimensional-modeling-manifesto/

[7]. Rahm, E. (2016). Big Data analytics. it-Information Technology, 58(4), 155-156.

[8]. Aloysius, A., & Prakash, A. A. (2018). Architecture design for Hadoop No-SQL and Hive. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, 3(1), 1069-1077

[9]. Sultana, A. (2018). Unraveling the Data Structures of Big data, the HDFS Architecture and Importance of Data Replication in HDFS.

[10]. Apache Hadoop. Available at http://hadoop.apache.org

[11]. Hive-Introduction. Available at https://www.tutorialspoint.com/hive/hive_introduction.htm