Big Data Problem, Technologies and Solutions

Hoda Ahmed Abdelhafez
Suez Canal University, Egypt

INTRODUCTION

Big data challenge is about how to process and analysis terabytes and petabytes of data from multiple sources like social media interactions, real-time sensory data feeds and others as well as extract meaningful value of it. Big data describes a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for timely and accurate decision making (Sagiroglu & Sinanc, 2013; Bertolucci, 2013; SAS, 2012).

IDC (International Data Corporation) defined big data as a "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis". Big data can include transactional data, warehoused data, metadata, healthcare, video and media/entertainment, in addition to social media solutions such as Facebook and Twitter (Gantz and Reinsel, 2011). Another definition is: big data means data that is *too big* (means that organizations must deal with petabyte-scale collections of data), *too fast* (it must be processed quickly-for example, to perform fraud detection at a point of sale), or *too hard* (means that is difficult to process or needs some kind of analysis) for existing tools to process (Wielki, 2013; Madden, 2012).

When big data is distilled and analyzed in combination with traditional data, organizations can develop insightful understanding of their business. Therefore, new technologies (such as Hadoop and MapReduce) are emerging that enable organizations to handle large amounts of unstructured data, making the prospect of gaining insight both feasible and cost-effective (Wielki, 2013; Oracle, 2012 a)

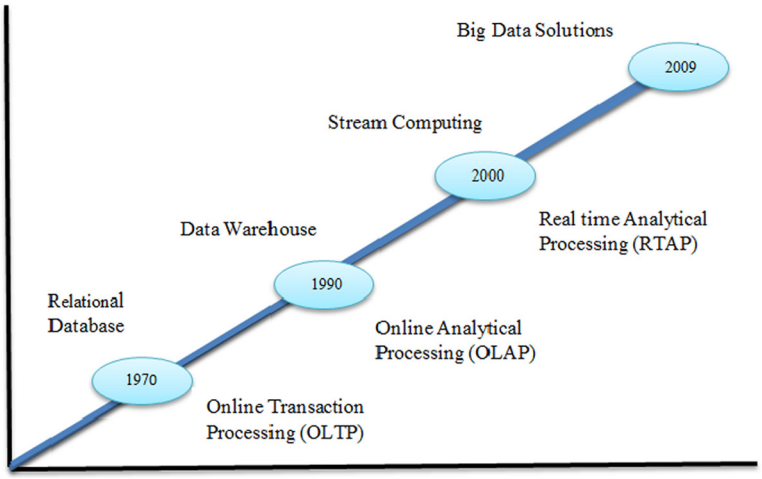
BACKGROUND

Relational database systems were used in 1970s and the structure query language (SQL) had been the way of dealing with data structure in a relational form as shown in Figure 1. The explosion of mainframes and personal desktops has created data warehouse that can easily manage data from multiple databases. The development of data warehouse in 1990s introduced thousands of applications across industries domains such as purchasing, shipping, enterprise resource planning (ERP) and supply chain management (SCM). In 2001 the XML (Extensible Markup Language) technology was born and then the demand for content management systems was lead to analyze unstructured and semi-structured data in the enterprises. Today, the advent of Internet creates and distributes multiple formats that explored all types of data. The ability to manage volume, velocity and variety of data and find analytical ways to provide better information at precisely time needs this is the evolution called big data (Zikopoulos, et al., 2012).

Big data represents large data sets whose size ranging from a few dozen terabytes to many petabytes of data in a single data set. Consider Web logs (the Web log records millions of visits a day), retail data (thousands of stores, tens of thousands of products, and millions of customers as well as billions of individual transactions in a year) and

DOI: 10.4018/978-1-4666-5202-6.ch031

Figure 1. From relational database to big data



cell phone database (It stores time and location every 15 seconds for each of a few million phones) as sources of big data. Recently a wide variety of technologies such as Hadoop and MapReduce has been developed and adapted to aggregate, manipulate, analyze, and visualize big data in order to help the organizations to derive value from big data (Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh, & Byers, 2011).

In addition to NoSQL or "Not Only SQL" database which overcomes the scaling limitations of relational database and manages unstructured data as well as distributing the work across multiple locations. NoSQL databases are schema-free design, therefore they enable applications to quickly upgrade the structure of data without table rewrites and allow data integrity and validity at data management layer. NoSQL systems are replicating and partitioning data over many servers to support a large number of simple read/write operations per second (Cattell, 2011).

HADOOP AND MAPREDUCE

The Apache Hadoop is defined as an open source software framework that supports data-intensive distributed applications. It enables applications to work with thousands of independent computers and petabytes of data. In hadoop the processing and querying of big data are applicable on large clusters of commodity hardware (Data Science Series, 2012). The first use of Hadoop was Google through Google file system (GFS) and the MapReduce programming mechanism. Hadoop has become widely deployed for massively parallel computation and distributed file systems in a cloud environment. Hadoop has allowed the largest Web properties (Yahoo!, LinkedIn, Facebook, etc.) to store and analyze any data in near real-time at a fraction of the cost that traditional data management and data warehouse approaches could even contemplate. Hadoop has two main parts: Hadoop distributed file system (HDFS) as a storage layer and Hadoop MapReduce as a processing layer (Kala Karun & Chitharanjan, 2013; Ha Lee, et al., 2011; Zikopoulos, et al., 2012; Kobielus, 2012).

Hadoop distributed file system (HDFS) helps Hadoop cluster to scale into hundreds or thousands nodes. The data into hadoop cluster is broken into blocks and the blocks are distributed in the cluster nodes. The input data in HDFS is executed by MapReduce, and the results are written back in the HDFS. The HDFS nodes are: DataNode to store the datablocks of the files in HDFS, NameNode containing the metadata, and two Map Reduce nodes (Job Tracker and Secondary Name Node). To protect the data, HDFS ensures data