Adapting Big Data Ecosystem for Landscape of Real World Applications

Jyotsna Talreja Wassan
University of Delhi, India

INTRODUCTION

Big data is revolutionizing world in the age of Internet. The wide variety of areas like online businesses, electronic health management, social networking, demographics, geographic information systems, online education etc. are gaining insight from *big data principles*. Big data is comprised of heterogeneous datasets which are too large to be handled by traditional relational database systems. An important reason for explosion of interest in big data is that it has become cheap to store volumes of data and there is a major rise in computation capacity.

To extract valuable patterns from big data, one needs to choose a right platform for capturing, organizing, searching and analyzing the context of voluminous data.

Data Management sytems adhering to big data, aim to add computing nodes to cater to increasing data volumes, automatic balancing of data between various nodes, and reducing the operational cost for functioning of distributing data over various nodes (Patel, 2016).

Various NoSQL data stores like Cassandra, MongoDB and Hadoop HBASE etc. are in use today to acquire, manage, store and query big data. NoSQL databases are inherently schema-less and permit records to have variable number of fields, making them distinct from other non-relational databases like hierarchical databases and object-oriented databases. These are highly scalable and well suited for dynamic data structures. NoSQL data is characterized by being basically available and eventually consistent.

The frameworks like MapReduce, Dryad etc. support processing of large amounts of data in parallel and hence the management of big data (Singh & Reddy, 2015). The technologies like GNU R and Apache MAHOUT are useful in exploring and analyzing big data for finding relevant valuable patterns. This article aims at giving an overview of Big Data Ecosystem comprising various big data platforms useful in today's competitive world.

BACKGROUND

In 1970's big meant megabytes, subsequently with the increasing data needs, it grew to gigabytes and terabytes and further to zettabytes with the increase in digital information. The traditional world of relational database systems like Oracle RDBMS etc. faced challenges in storing large quantities of data and needed to scale databases to data volumes beyond the storage and/or processing capabilities of a single large computer system. Many efforts have been made to store and manage data being generated from everywhere on the web. Several database management systems were proposed on the basis of master/slave, cluster computing or partitioning architecture like IBM DB2 partitioning, VoltTB etc.

However, the problems in reliance on shared facilities and resources (CPU, Disk, and Processors), scalability and complex administration limitations, augmented by lack of support for critical requirements, led to development of SHARED NOTHING architectures (Strauch, 2011; Lee, 2011) in 1980's. These systems focused on paral-

DOI: 10.4018/978-1-5225-2255-3.ch029

lel and distributed data computation and solved big data problems using parallel computations. By 90's, even these solutions faced challenges in running OLTP and queries due to data overload. To provide solutions to these problems, Google responded with its GFS (Dean & Ghemawat, 2004), followed by a powerful programming paradigm of MapReduce (Dean & Ghemawat, 2004).Thereafter a spectrum of new technologies emerged as the NoSQL movement stating a broad class of database management system to support increasing data storage and analytical requirements.

MAIN FOCUS

Major real world applications like health care, business analytics etc., operational on big data, cannot store or process all of the data on just one machine. The data must be stored, distributed or processed in parallel manner for computations to be completed efficiently. Various platforms are making *big data* management and processing more effective, forming the basis of current research theme in the era of *Big Data (*Gandomi & Haider, 2015). The main focus of this article is to discuss about NoSql Movement, big data platforms which could support processing of futuristic massive volumes of data in parallel and their applications.

BIG DATA ECOSYSTEM

Big data could be visualized in two aspects:

1. Big Data Storage and
2. Big Data Analytics.

Hence, Big data applications fall under two general categories:

1. Large-volume applications needing hundreds of terabytes of data to work on, or
2. Performance-intensive big data analytics applications needing computation.

Broadly big data ecosystem comprise of stacked layers of Storage (NoSQL stores), MapReduce and Query (SMAQ) as illustrated in Figure 1. SMAQ structured systems are typically open source and distributed. These systems are changing the landscape of big data processing to a broader class of users similarly as LAMP stack of Linux, Apache, MySQL and PHP changed the horizon of developing web applications (Dumbill, E., 2010).

The base idea is to store big data in parallel NoSQL data stores like MongoDB, Apache HBASE, and BigTable etc.

Above the storage layer, a layer is required to divide the stored big data set, and run it in parallel over many machine nodes. This distribution provides a solution to the issue of data being too

Figure 1. SMAQ stack for big data