



Big Data: Concept, Challenges and Management Tools

Ranjana Bahri

Research Scholar, Punjab Technical University, Kapurthala
Assistant Professor in KCLIMT, Jalandhar,
Punjab, India

Abstract: We are at the beginning of a big data era when data is generated at an incredible speed from everywhere. Computing has become global, number of devices like cell phones, smart phones, laptops, personal sensors are creating countless new digital oceans of information. A few years ago we talked about data storage in megabytes and gigabytes but now a days huge amount of data found on internet which is close to 500 billion gigabytes[1]. According to IBM 2.5 quintillion bytes of data are generated daily. Data is generated at unbelievable speed from satellites, social networking sites, videos uploaded on YouTube. All the people around the world acts as a sensor and generate data at every second either it is our location shown on Facebook through gprs signals or our movement by gathering data from our mobile phones or smartphone devices[2]. So Big Data is a term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization of Big Data[4].

Keywords: Big data, Hadoop, Big Table, map reduce.

I. WHAT EXACTLY BIG DATA IS?

Big Data is relatively a new concept and a lot of definitions have been given to it by researchers, organizations and individuals. As far back as 2001, industry analyst Doughty Laney (currently with Gartner), articulated the mainstream of definition of Big Data as the three Vs; Volume, Velocity and Variety. At SAS, SAS considered two additional dimensions when thinking about Big Data: the Variability and Complexity [6]. Oracle defined Big Data in terms of four Vs – Volume, Velocity, Variety and Value [7]. **Oguntimilehin Adefine** Big Data in terms of five Vs and a C[4]. The five Vs are: Volume, Velocity, Variety, Variability, Value and complexity

- **Volume:** the size of data. In the modern era of technology it has become very difficult to talk about data volume in any absolute sense. As technology grows, numbers get quickly outdated, so it is better to think about volume in a relative sense instead[4]. The following bifurcation gives details about volume of Big Data in relation to the fast growth of data. It also analyzes Big Data in the current environment of enterprises and technologies.
- **Fast growth of Data:** Unstructured data is growing more rapidly. This type of data includes all human information like tweets, facebook data, geospatial maps, medical records and images, high definition videos etc. In most of the organizations 70 to 80% of data is in unstructured form and it is very difficult to analyze that data. **Table 1** describes the rapidly grown data in different organizations.

Table 1: Rapid growth of unstructured data in various companies

Organization	Amount Of Data Produced
YouTube[11]	(i) 300 hours of videos are uploaded to YouTube every minute. (ii) Each month, more than 1 billion unique users access YouTube (iii) Every day people watch hundreds of millions of hours on YouTube and generate billions of views. The number of hours people are watching on YouTube each month is up 50% year over year
Facebook	(i) In every 20 minutes 3

[12]	million message sent (ii) 1 million links are shared in every 20 minutes. (iii) total number of monthly active users 1,310,000,000 (iv) Average number of photos uploaded per day 205.
Twitter[13]	(i)The site has over 645,750,000 users. (ii) The site generates 175 million tweets per day
Google+[4]	1 billion account has been created
Instagram[4]	Users share 40million photos per day
Linkedin[14]	(i) Total number of Linked users 313,000,000

- **Velocity:** data is generating at unpredictable speed. The rate at which data is being received and has to be acted upon is becoming much more real-time. While it is unlikely that any real analysis will have to be completed in the same time period, delays in execution will inevitably limit the effectiveness of campaigns, limit interventions or lead to sub-optimal processes[7].
- **Variety:** Today data comes in heterogeneous type means it includes both structured and unstructured data. Managing, merging and governing different varieties of data is a challenging task for organizations now a days.[4]
- **Variability:** In addition to the increasing velocities and varieties of data, amount of data flows is highly variable. We can observe this thing on social media as in table 1. [4]
- **Value:** It is the matter to think upon that are we really found good quality data or is that data have any value for our business [4]. Is after Big data has came in existence our exciting problems in industry has been solved or not?
- **Complexity:** Data become more complex because it comes from multiple sources. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages [4].

II. CHALLENGES IN BIG DATA

In computational sciences, Big Data is a critical issue that requires serious attention. There are only two or three main issues appear capable of making or breaking the promise of Big Data, and these are related to: solution approach, personal privacy and intellectual priority (IP). The first issue deals with technology, deployment and the organizational context, whereas the later two are concerned with nature and applicable use of Big Data. Other challenges of Big Data are heterogeneity and incompleteness, scale, timeliness; another closely related concern is data security [4]. Processing of BigData using existing technologies and methods is not possible. According to data analytics standard tools have not been designed to search and analyse large datasets. As a result, organizations encounter early challenges in creating, managing, and manipulating large datasets. Systems of data replication have also displayed some security weaknesses with respect to the generation of multiple copies.

2.1 Heterogeneity

Machine analysis algorithms expect homogenous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step to (or prior to) data analysis.

2.2 Scale

As we have examined above that volume or scale is the second major challenge of Big Data. Actually Size is directly related with term BIG and this data is grown rapidly. Managing large and rapidly increasing volumes of data has a challenging issue for many decades. Data volume is scaling faster that computer resources and CPU speeds are static. These unprecedented changes require us to rethink how we design, build and operate data processing components [5].

2.3 Timeliness

Timeliness is directly related with size, larger the size of data more time is required to process and analyze data. The best system is that which gives user data in correct form on right time[5].

2.4 Personal Privacy

In the era of Big Data our personal information that is stored and transmitted through ISPs, mobile network operators, supermarkets, local councils, medical and financial service organizations (e.g hospitals, banks, insurance and credit card agencies). Also information shared and stored on social networking sites like facebook, twitter etc. Privacy is an important issue for everyone. All wants to hide their personal information in order to avoid the misuse of that information. But as the Big Data is grown it is very difficult to achieve.

III. BIG DATA MANAGEMENT

New methods and tools to embed information into business processes — use cases, analytics solutions, optimization, work flows and simulations — are making insights more understandable and actionable. Respondents identified trend analysis, forecasting and standardized reporting as the most tools they use today. However, they also identified tools that will have greater value in 24 months like:[16].

1. Data visualization, such as dashboards and scorecards
2. Simulations and scenario development
3. Analytics applied within business processes
4. Advanced statistical techniques, such as regression analysis, discrete choice modeling and mathematical optimization.

Organizations expect the value from these emerging techniques to soar, making it possible for data-driven insights to be used at all levels of the organization. For example, GPS-enabled navigation devices can superimpose real-time traffic patterns and alerts onto navigation maps and suggest the best routes to drivers.

3.1 The architecture of Big Data: must be synchronized with the support infrastructure of the organization [5].Data which is generated from various sources like from machines or sensors is unstructured and messy in nature. Previously most of the organizations were unable to either capture or store this type of data with the available tools even they can't manage the data in a reasonable amount of time. However, the new Big Data technology improves performance, facilitates innovation in the products and services of business models, and provides decision making support. Properly managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can be applied in various complex scientific disciplines, including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry [5]. In this paper we briefly discuss data management tools.

3.2. Management Tools. As the computing technology grows, large volumes of data can be managed without requiring supercomputers and high cost. There are number of tools and techniques are available for Big data management, including Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort [5]. However, there is need to develop special tools and technologies that can store, access, and analyze large amounts of data in near-real time because Big Data differs from the traditional data and cannot be stored in a single machine. For Big Data, some of the most commonly used tools and techniques are Hadoop, MapReduce, and Big Table [5]. These innovations have redefined data management because they effectively process large amounts of data efficiently, cost effectively, and in a timely manner.

3.2.1 Hadoop. Hadoop is written in Java and is a top-level Apache project that started in 2006. It emphasizes discovery from the perspective of scalability and analysis to realize near-impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system. Presently, it is used on large amounts of data. With Hadoop, enterprises can harness data that was previously difficult to manage and analyze. Hadoop is used by approximately 63% of organizations to manage huge number of unstructured logs and events (Sys.con Media, 2011). In particular, Hadoop can process extremely large volumes of data with varying structures (or no structure at all). Hadoop is composed of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka; however, the most common components and well-known paradigms are Hadoop Distributed File System (HDFS) and MapReduce for Big Data. Table2 shows the hadoop components and its functionality in brief.[5].

Table 2: Hadoop components and their functionalities.

Hadoop Component	Functions
(1) HDFS	Storage and replication
(2) MapReduce	Distributed processing and fault tolerance
(3) HBASE	Fast read/write access
(4) HCatalog	Metadata
(5) Pig	Scripting

(6) Hive	SQL
(7) Oozie	Workflow scheduling and
(8) ZooKeeper	Coordination
(9) Kafka	Messaging and data integration
(10) Mahout	Machine learning

3.2.2 MapReduce[16]:-

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate keys. Many real world tasks are expressible in this model [16]. The computation takes a set of *input* key/value pairs, and produces a set of *output* key/value pairs. The user of the MapReduce library expresses the computation as two functions: *Map* and *Reduce*. *Map*, written by the user, takes an input pair and produces a set of *intermediate* key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key *I* and passes them to the *Reduce* function. The *Reduce* function, also written by the user, accepts an intermediate key *I* and a set of values for that key. It merges together these values to form a possibly smaller set of values. Typically just zero or one output value is produced per *Reduce* invocation. The intermediate values are supplied to the user's reduce function via an iterator. This allows us to handle lists of values that are too large to fit in memory.

Example[16]

Consider the problem of counting the number of occurrences of each word in a large collection of documents. The user would write code similar to the following pseudo-code:

```
map(String key, String value):
// key: document name
// value: document contents
for each word w in value:
EmitIntermediate(w, "1");
reduce(String key, Iterator values):
// key: a word
// values: a list of counts
int result = 0;
for each v in values:
result += ParseInt(v);
Emit(AsString(result));
```

The map function emits each word plus an associated count of occurrences (just '1' in this simple example). The reduce function sums together all counts emitted for a particular word.

In addition, the user writes code to fill in a *mapreduce specification* object with the names of the input and output files, and optional tuning parameters. The user then invokes the *MapReduce* function, passing it the specification object. The user's code is linked together with the MapReduce library (implemented in C++).

3.2.3 Google Big Table[17]:

Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Finance. These applications place very different demands on Bigtable, both in terms of data size (from URLs to web pages to satellite imagery) and latency requirements (from backend bulk processing to real-time data serving). Despite these varied demands, Bigtable has successfully provided a flexible, high-performance solution for all of these Google products.

Data Model

A Bigtable is a sparse, distributed, persistent multidimensional sorted map. The map is indexed by a row key, column key, and a timestamp; each value in the map is an uninterpreted array of bytes.

(row:string, column:string, time:int64) →string

We settled on this data model after examining a variety of potential uses of a Bigtable-like system. As one concrete example that drove some of our design decisions, suppose we want to keep a copy of a large collection of web pages and related information that could be used by many different projects; let us call this particular table the *Webtable*. In *Webtable*, we would use URLs as row keys, various aspects of web pages as column names, and store the contents of the web pages in the contents: column under the timestamps when they were fetched[17].

IV. CONCLUSION

This paper presents the fundamental concepts of Big Data. These concepts include the increase in data in various organizations, and the role of Big Data in the current environment of enterprise and technology. There are uncountable applications and advantages of Big Data as some of them had been identified in this paper. It is to be noted that there are a lot of challenges facing the Big Data and in order to make optimal use of this modern discovery, users must be quite aware of these challenges so as to providing a measurable adjustment or solutions to them as quick as possible.

REFERENCES

- [1] <http://23.66.85.199/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf>
- [2] IEEE Conference on Visual Analytics Science & Technology 2012 : Seattle, Washington, USA, 14 - 19 October 2012 ; Proceedings / Giuseppe Santucci and Matthew Ward (eds.) - Piscataway, NJ : IEEE, 2012, S. 173-182. - ISBN 978-1-4673-4753-2
- [3] The Promise and Peril of Big Data by David Bollier Rapporteur
- [4] A Review of Big Data Management, Benefits and Challenges 1 Oguntimilehin A., 2 Ademola E.O. 1,2Department of Computer Science, AfeBabalola University, Ado-Ekiti, Nigeria
- [5] *Review Article:Big Data: Survey, Technologies, Opportunities, and Challenges* By Nawsher Khan,1,2Ibrar Yaqoob,1 Ibrahim AbakerTargio Hashem,1 Zakira Inayat,1,3WaleedKamaleldinMahmoud Ali,1 Muhammad Alam,4,5 Muhammad Shiraz,1and Abdullah Gani1
- [6] Mark Troester(2013), "Big Data Meets Big Data Analytics", www.sas.com/resources/.../WR46345.pdf, retrieved 10/02/14.
- [7] Oracle (2013), "Information Management and Big Data:A Reference Architecture", www.oracle.com/.../info-mgmt-big-data-r..., retrieved 20/03/14.
- [8] Intel, "Big Data Analytics," 2012, <http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insightspeer-research-report.pdf>.
- [9] Chris Deptula(2013), "With all of the Big Data Tools, what is the right one for me", www.openbi.com/blogs/chris%20Deptula, retrieved 08/02/14.
- [10] Oracle (2013), "Information Management and Big Data:A Reference Architecture", www.oracle.com/.../info-mgmt-big-data-r..., retrieved 20/03/14.
- [11] YouTube "YouTube Statistics Feb24, 2015"<https://www.youtube.com/yt/press/statistics.html>
- [12] Facebook, Facebook Statistics, jan 27,2015, <http://www.statisticbrain.com/facebook-statistics/>.
- [13] Twitter, "Twitter statistics," 2015, <http://www.statisticbrain.com/twitter-statistics/>.
- [14] "LinkedIn Statistics" , October 28th, 2014 <http://www.statisticbrain.com/linkedin-company-profile-and-statistics/>
- [15] Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems Leishi Zhang,University of Konstanz, Germany Andreas Stoffel,University of Konstanz, GermanyMichael Behrisch,University of Konstanz, Germany,Sebastian Mittelst "adt§ University of Konstanz, Germany,Tobias Schreck,University of Konstanz, Germany Ren "ePompl, Siemens AG,Stefan Weber, Siemens AG Holger Last,Siemens AG,Daniel Keim,University of Konstanz, Germany
- [16] <http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf>
- [17] <http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf>