Big Data and Its Impact on Business Revolution - A Review
Monika Yadav¹, Neha Goyal*²
²Student & Computer Science & Banasthali University
Rajasthan, India

Abstract — Adaptation of Big Data proves practically beneficial to many organizations. This paper will include study of Big Data by taking brief of data mining and data analytics, importance, applications, parallel processing and sources of big data into consideration. Various surveys on big data demonstrate the practical applicability of big data approach in various company or organization. Various challenges and factors influencing big data implementation have been discussed.

Keywords— Applications, Big Data, Data Analytics, Data Mining, Importance, Parallel Processing, Sources, Surveys.

I. INTRODUCTION

Big Data is named as Big because it mainly represents large-volume, complex, growing data sets with multiple, autonomous sources. The big data may be created by handheld device, social network, internet of things, multimedia, and many other new applications that all have the characteristics of volume, velocity, and variety. Data is growing fast with fast development of networking, data storage, and the data collection capacity. Increase in usage of social networking sites like Facebook, Twitter etc. in all domains including science and engineering domains, any social issue and every field of life.

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Various definitions are given for Big Data. Out of it a well-known definition known as 3Vs is given by Laney [1] which defines Big Data in terms of volume, velocity and variety. This definition of Laney defines 3Vs as large data size, rapidly created data and multiple data types collected from different sources respectively. Later veracity, validity, value, variability, venue, vocabulary and vagueness were added to this definition in order to make some complement explanation of big data[2].

Various available Big Data Analysis frameworks are DOT, GLADE, Starfish, ODT-MDC, MRAM, CBDMASP, SODSS, BDAF, HACE, Hadoop, CUDA, Strom, Pregel, MLPACK, RADOOP etc. In order to satisfy large demands of computing power and storage, the cloud computing technologies are widely used on these platforms. Parallel computing is one of the important future trends to make the data analytics work for big data, and consequently the technologies of cloud computing, Hadoop, and map-reduce will play the important roles for the big data analytics.

II. BIG DATA: DATA ANALYTICS AND DATA MINING

A few efficient methods [3], such as sampling, data condensation, density-based approaches, grid-based approaches, divide and conquer, incremental learning, and distributed computing, have been presented for analysing large scale data i.e. Big Data. The Knowledge Discovery Database KDD process is vital to deal with big data which is summarized into three parts: Data Input, Data Analysis and Output. Fig.1 shows the process of knowledge discovery in database.

A. Data Input

The gathering, selection, pre-processing, and transformation operators are in the input part.

1) Gathering: It deals with gathering raw data from different sources.

2) Selection: This operator determines which kind of data was required for data analysis and selects the relevant information from the gathered data or databases. It will integrate the target data gathered from different data sources.

3) Pre-processing: This operator aimed at making data useful by detecting, cleaning and filtering the unnecessary, inconsistent, and incomplete data.

4) Transformation: It will transform the number of different data formats into data-mining capable format. The methods like dimensional reduction, sampling, coding, or transformation for reducing the complexity and downsizing the data scale to make the data useful for data
analysis part are usually employed in the transformation

It can be expected that these operators may affect the analytics result of KDD, be it positive or negative. In summary, the systematic solutions are usually to reduce the complexity of data to accelerate the computation time of KDD and to improve the accuracy of the analytics result.

B. Data Analysis

It is responsible for finding the hidden patterns/rules/information from the data. The term data mining is used by most of researchers to describe raw data into information or knowledge. The data extraction, data cleaning, data integration, data transformation, and data reduction operators can be regarded as the pre-processing processes of data analysis.

Several efficient analysis methods were presented to accelerate the computation time or to reduce the memory cost for the KDD process. Some of them are clustering, classification, Association rules, sequential patterns etc.

Various solutions have been presented for the big data analytics which can be divided into processing/compute, storage and analytics. Hadoop [9], Nvidia CUDA [10], or Twitter Storm [11] is used for processing, Titan or HDFS is used for storage and MLPACK [12] or Mahout [13] is used for analytics. Various data mining algorithm are DBDC, PKM, CloudVista, MFCUDA, BDCAC, SVMGA, Quantum SVM, SPC, FPC, and DPC etc.

![Fig. 1 The process of knowledge discovery in databases](image)

C. Output the result

Evaluation and interpretation are two vital operators of the output.

III. VARIOUS CRITERIA OF BIG DATA

A. Importance of Big Data[14]

The importance of big data doesn’t revolve around amount of data, but what can be done with it. Data can be taken from any source and analysed in order to reduction of cost and time. Data analysis also results in new product development, optimized offerings and smart decision making. Various business related tasks that can be accomplished via combination of big data with high–powered analytics are described as follows.

- On the basis of customer’s buying habits coupon can be generated at the point of sale.
- Quick recalculation of risk portfolios.
- Root causes of failures and issues detection.
- Detection of defects in near-real time.
- Fraudulent behaviour detection.

B. Applications of Big Data[14]

Big data is useful almost in every field of life. Each organization can take benefit from big data usage. Following are benefits of big data usage to various organizations:

1) Education: In education field big data analysis played a major role for measuring student progress, identifying at-risk students and implementing a better system for evaluation and support of teachers and principals.

2) Banking: Adaptation of big data technologies seems to be extremely useful in banking field in order to manage large amount of information collected from variety of sources. Big data technology is also comes with new and innovative way to manage data and minimise risk, fraud while maintaining regulatory compliance. Big data brings big insights, but it also requires advance analytical skills.

3) Retail: Retailers utilizes big data technology to find out best way for building customer relationship, effective way for handling transactions and most strategic way for handling business.

4) Health Care: Big data analytics can be used for analyzing patient records, treatment plans and prescription information. Patient care can be improved if big data is managed effectively. Big data techniques also provide quick and accurate results which is most critical requirement in the field of health care.
Health care providers can uncover hidden insights when big data is managed effectively which results in patient care improvement.

5) **Manufacturing:** Quality and output processes that are key in today’s highly competitive market can be boosted with the help of big data methodologies. Adaptation to big data techniques results in faster problem solution and more agile business decisions.

6) **Government:** Government agencies are utilizing big data analytics for managing utilities, running agencies, dealing with traffic congestion or preventing crime.

C. Sources of Big Data

Big data comes from variety of sources. Some of the most popular data sources are mentioned here.

1) **Streaming Data:** This category includes data that comes from a web of connected devices and further this data can be refined by implementing big data analysis.

2) **Social media data:** The data on social interactions is an increasingly attractive set of information, particularly for marketing, sales and support functions. Various social sites like Facebook, Twitter and LinkedIn etc. contribute major amount to big data. This data generally collected in unstructured or semi structured forms. Therefore it poses unique challenge for consumption and analysis.

3) **Publicly available sources:** Massive amounts of data are available through open data sources like the US government’s data.gov, the CIA World Facebook or the European Union Open Data Portal.

D. Big Data and Parallel processing

Parallel processing plays important role for managing big data. In order to achieve faster result, higher throughput and valuable result, parallel processing proves to be very helpful and challenging. Various affordable open source, distributed big data platforms, such as Hadoop are there for achieving such tasks via cloud computing.

Software written in SQL are running in parallel for more than twenty years but with big data, programmers finds SQL an unfamiliar and restrictive way to write code as they are interested in building programs on parallel models. The parallel programming framework called Map Reduce has gained lot of popularity for parallel programming. The emergence of platforms that provide both SQL and MapReduce interfaces within a single runtime environment plays important role in parallel processing. These are especially useful when they support parallel access to both database tables and file system files from either language. Examples of these frameworks include the commercial Greenplum system, the commercial Aster Data system (which provides SQL and MapReduce over database tables), and the open-source Hive framework from Facebook (which provides a SQL-like language over files, layered on the open-source Hadoop MapReduce engine.)

Following are some issues in programming multithreaded Big Data applications [15]:

1) **Consider High Performance Storage Technologies:** RAID 5 can be used for solving big data problems. Data striping and data caching features make it fit into big data requirement. Data striping introduces less seek-time delay because multiple drives can read a file’s data on different physical devices simultaneously and data caching results in faster read because of stripping and data get cached for even faster future access.

2) **Make Threading Configurable:** Benchmarking functionality to evaluate the performance of each piece of the application can be used to make threading configurable. The configuration can be automatic or manual depending on skills of installer.

3) **Different Parts of Systems Have Varied Performances:** Data retrieval and number crunching can be used to deal with this issue. For data retrieval, an approach of where parallel threads that are synchronized manage data retrieval should be considered. Data can be queued and cached based on a system of heuristics. Then, the modules that need the data can retrieve it without the waits that would have otherwise been experienced.

4) **Optimize memory Allocation:** Two good solutions that should be considered for optimizing memory allocation are a third-party memory management library such as Hoard and implementing own memory management solution which will eliminates the lock contention issue.
IV. SURVEYS ON BIG DATA

A. Big Success with Big Data by Accenture 2014[16]:

In organizations that are using big data today, user report overwhelming satisfaction with their results, according to a new Accenture Analytics survey, and see big data as a catalyst for their transformation as digital enterprises. Key findings emerging from the research cluster around these themes. Accenture Analytics surveyed more than 1000 respondents from companies operating across seven industries and headquartered in 19 countries that had completed at least one big data projects, respondents from companies that had not completed at least one big data installation were not included in the result. Fig.2. shows the various sources of Big Data measured in survey.

Annotations used in Fig.2. are as follows:

1) Sources of Big Data:

LD: Large data files.
AAA: Advanced analytics or analysis.
VT: Data from visualization tools.
SN: Data from social networks.
UD: Unstructured Data (e.g. video, open text, voice).
G: Geospatial /location information.
SMM: Social media/Monitoring/Mapping.
T: Telematics.
ULF: Unstructured data/log files/ free text.

2) Challenges in implementing big data:

S: Security.
B: Budget.
T1: lack of big data implementation talent.
T2: lack of talent for big data and analytics on an ongoing basis.
I: Integration with existing systems.

3) Variation in Big Data definitions:

LDF: Large data files.
AAA: Advanced analytics or analysis.
VT: Data from visualization tools.

Effect on decision making by big data adaptation:

C: Executives surveyed said that big data has changed decision-making in their organizations.
CF: Expect it will do so over the next two years.
CIE: of executives agree that data analysis is undermining the credibility of intuition or experience.
DCIE: Disagree that data analysis is undermining the credibility of intuition or experience.

4) Various companies views on adaptation of Big Data:

EQ: Big data will revolutionize business operations in the same way the Internet did.
DC: Big data will dramatically change the way they do business.
CP: agree that “companies that do not embrace big data will lose their competitive position and may even face extinction. Fig.3 shows the Big Data competitive significance as observed during Accenture 2014 survey.

Annotations used in Fig.3 are as follows:

5) Big Data competitive Significance:

RB: Big data will revolutionize the way we do business to a degree similar to the advent of the internet in 1990s.
TABLE I
BIG SUCCESS WITH BIG DATA: A SURVEY BY ACCENTURE IN 2014

<table>
<thead>
<tr>
<th>Challenges in implementing Big Data</th>
<th>Variation in big data definition</th>
<th>Effect on decision making by big data adaptation</th>
<th>Companies views on big data adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Factors</td>
<td>Percentage</td>
<td>Key Factors</td>
<td>Percentage</td>
</tr>
<tr>
<td>S</td>
<td>51</td>
<td>LDF</td>
<td>65</td>
</tr>
<tr>
<td>B</td>
<td>47</td>
<td>AAA</td>
<td>60</td>
</tr>
<tr>
<td>T1</td>
<td>41</td>
<td>VT</td>
<td>50</td>
</tr>
<tr>
<td>T2</td>
<td>37</td>
<td>-</td>
<td>DCIE 21</td>
</tr>
<tr>
<td>I</td>
<td>35</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

DC: Big data will dramatically change the way we do business in the future.

E: Companies that do not embrace big data will lose their competitive position and may even face extinction.

CA: We feel we are ahead of our peers in using big data and this creates a competitive advantage for us.

Table 2. shows the most popular data resources and most popular big data use cases as measured by TIBCO Jaspersoft survey.

TABLE II
BIG DATA SURVEY BY TIBCO JASPERSOFT

<table>
<thead>
<tr>
<th>Most popular Data Resources</th>
<th>Most popular Big Data Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Factors</td>
<td>Key Factors</td>
</tr>
<tr>
<td>CRM</td>
<td>40</td>
</tr>
<tr>
<td>Financial Applications</td>
<td>38</td>
</tr>
<tr>
<td>Retail POS</td>
<td>15</td>
</tr>
<tr>
<td>Supply Chain Management</td>
<td>14</td>
</tr>
<tr>
<td>Human Capital Management</td>
<td>12</td>
</tr>
<tr>
<td>Product Life Cycle Management</td>
<td>11</td>
</tr>
<tr>
<td>Support Case Logic</td>
<td>10</td>
</tr>
<tr>
<td>Bug Tracking</td>
<td>9</td>
</tr>
<tr>
<td>E-Commerce</td>
<td>27</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
</tr>
<tr>
<td>-</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 4 shows the reason for “No Plans” for Big Data in various organizations according to Jaspersoft survey.

![Fig. 4 Reasons for "No Plans" with Big Data.](image)

Annotation used in Fig. 4 are as follows:

1) **Reasons for “No Plans” with Big Data are diminishing in importance:**

   SR: Data is structured and relational.
   UBD: Don’t understand Big Data.
   A: Doesn’t apply to my applications.
   BJ: No business justification.

2) **Most Popular Data Stores:**

   Survey shows the most popular data stores used by respondent as Relational Databases (56%), MongoDB (23%), Analytic Database (14%), Hadoop HDFS (12%), and Apache Hive (4%).

**V. CONCLUSIONS**

Big Data approach collect data from variety of resources and analyse it to enable cost reduction, time reduction and smart decision making. Big Data has great significance in various fields like banking, education, health caring, manufacturing and retail. The parallel programming framework called Map Reduce and parallel processing platform Hadoop proves to be very helpful in managing big data and to achieve faster result, higher throughput and valuable result.

Big Data surveys “Big Success with Big Data” by Accenture and “Big Data Survey” by TIBCO Jaspersoft, sees big data as catalyst for their transformation as digital enterprises. These surveys shows the statistics of adaptation of big data approach by considering various factors like challenging implementing Big Data, variations in big data definitions, effect on decision making, various companies views data resources and use cases.

Big Data is an emerging field and various concepts and surveys mentioned in this research paper shows the statistics of adaptation of big data approach by considering various factors like challenging implementing Big Data, Variations in Big Data definitions, Effect on decision making, highlight the various facts about Big Data which will help the organization in managing Big Data and attracts them towards adaptation of big data approach.

**REFERENCES**


