
Tapan P Gondaliya Hiren D Joshi Hardik J Joshi

Abstract

One of the new technologies recently evaluated is called the “Big Data”. Now a day’s in this digital data era everything is digital like an e-library, e-mail, e-shopping, e-ticket, e-payment, e-governance and many more. People used more and more websites for entertainment like a facebook, twitter, and youtube for video, photos, twits, and data downloads as well as uploads on the internet. Internet has stored a massive amount of data or information that is in the zeta or in Exabytes it is nothing but the Big Data. According to IDC in future the growth of data will never stop and it will become in 7910 Exabytes in end of year 2015. Big Data is basically in the format of uncompressed data so it is very large, complex and difficult to process in traditional data processing application. So in this kind of massive dataset it’s very difficult to visualize, analyze, search, storage and transfer the data for any of the organization or company. And these are the biggest challenges for big company to how to solve this kind of problem. Behind this paper our main motive is to describe the reality of big data, how can different big data with the traditional database, what are the different types of big data, characteristic of big data and in actual how its work with different tools and technology and how company can face these big challenges using these tools. Here we also describe the comparative study of different tools that is basically used for analyze, visualize, store and transfer a big data.

Keywords: Big Data, Data Warehouse, 4V’s, Big Data Tools, Visualization, Big Data Analysis, Big Data

Addresses

1. Introduction

We are living in digital era where we are awash in a flood of data. In last recent year the big data has buzzword for different kind of fields like a Computer Science and Information Technology, Science and Medical, Education and Entertainment, Health and Wealth, Tools Trends and Technology, Just think a moment world without a data! Now a day’s data is very important things without data nothing is possible. Big data is nothing but the simply data that expands the processing ability of conventional database system and the data is so large and moving very fast or does not fit in database architecture. According to other authors The hottest IT buzzword of year 2012 Big Data is a group of large and complex dataset that is includes the huge quantity of data, social media analytic and real time data as well. Big data has divided in to different characteristics like a 4V’s means the Velocity, Volume, Variety, and Veracity that terms we discussed later in this paper.

2. Big Data

‘BIG DATA’ is recently been arrived and applied in to dataset, big data growth is very large and even they don’t work with traditional database management system. According to the IDC Big Data is a new generation technology and architecture de-
New Big Things in Era of Digital Data...

signed to efficiently extract value from huge volume of wide variety of data. Regarding to Jason Bloom Big Data is a huge amount of both structured and unstructured types of data that is so large that it's difficult to process using traditional database and other software. Big data is a term for massive data sets having large, varied & complex type of structure with the challenges of analyzing, Storing and visualizing for extra processes or results. Big data is a kind of datasets that are not only big but also is in form of high variety with fast velocity, which is not easy to handle through traditional tools. Various sources of the big data are as under.

Big data is huge amount of digital data or information collect by the government, private companies, social web sites and many firms around us. Every day we create 2.5 quintillion bytes of data in this 90 percent of data is created in last two year alone. Big Data have different sources those produced a that kind of large amount of data and those sources are like a Social websites facebook, tweeter, google+, and many more that produced lots of data in one day and that data is in form of Video, Image, Textual, Audio and Other. Government as well as the Private Companies websites also generated a large amount of data. Some of the scientific devices and instruments, media and mobile devices, that also a one of the reason that produced a large amount of data.

3. Traditional Data

Traditional data is basically stored in the database or in the data warehousing. Example of traditional database is RDBMS, DBMS, SQL and more. Data
Warehouse is a set of tools and techniques to enable the collection of data from operational systems, integration, and management of that data into a central database and then the analysis, visualization, and other operations can be performed in a dashboard. Data warehouse is basically implemented in a single relational database system that serves as the centralized machine. Basically the data warehouse is the combination of four things that includes integrated, time variant, non-volatile, subject oriented for data support management in decision making process.

Figure 2: Traditional Data Warehouse

Now let we see how can differs both the data is in size, speed, technology and architecture wise here we summarized both of the data in tabular form.

Table 1: Comparative Study of Traditional vs. Big Data

<table>
<thead>
<tr>
<th>Data Generated</th>
<th>Traditional Data</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data generated in Enterprise level or it is include traditional data.</td>
<td>Data generated in Outside and Enterprise level.</td>
<td></td>
</tr>
<tr>
<td>2. Traditional data sources are include ERP transaction data, CRM data, Web transaction data, Financial data</td>
<td>Non-traditional data sources are include Social Media, Log data, Device data, Sensor data, Video, Images</td>
<td></td>
</tr>
<tr>
<td>3. Data store in gigabytes or terabytes</td>
<td>Data store in Petabytes, Zettabytes, Exabytes</td>
<td></td>
</tr>
<tr>
<td>4. Data stored in a form of structured and unstructured</td>
<td>Data stored in a form of structured, unstructured and semi-structured</td>
<td></td>
</tr>
<tr>
<td>5. Data managed in centralized form</td>
<td>Data manage in physically distributed form</td>
<td></td>
</tr>
<tr>
<td>6. By default stable and interrelationship</td>
<td>Unknown Relationship</td>
<td></td>
</tr>
<tr>
<td>7. Specialized high level software as well as hardware used</td>
<td>Inexpensive commodity boxes in cluster mode</td>
<td></td>
</tr>
</tbody>
</table>

4. Types of Big Data

Big Data is basically divided into three different parts namely is called the structure, unstructured and Semi-structured data.
4.1 Structured Data

When we talking about the structured data that is basically stored in the RDBMS Data Warehouse in formal way. Data is in grouped in the form of rows and columns. Now a day's total 10% of structured data available around us. Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets. Structured data gives names to each field in a database and defines the relationships between the fields. Example of structured data is RDBMS (ERP and CRM), Data Warehousing, Microsoft Project plan File (.MPP File).

4.2 Un Structured Data

Data generated through human language including text and numeric value with or without delimitation punctuation or metadata. Data which can't be stored in raw and column format just like a video and audio data, streaming data, pictographically data is called the unstructured data. Massively these kinds of data more and more generating now a day's its 80% this kind of data generating in last 2 years. Refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and symbol as well. Example of Unstructured data is Video, Audio, Text Message, Blogs, Email, Social Media, Click Stream Weather Pattern, Location Coordinates, and Sensor Data.

4.3 Semi-Structured Data

Semi-structured data is one kind of structured data that does not conformed its formal structured of that data model is called the semi structured data. This type of data currently 10% existing and example of this kind of data is RSS Feeds and XML Formats Data.

5. Big Data 4V's

In the article of Colin White the Big data comes in many shapes and sizes and length and Regarding to the IBM data scientists big data is divided into mainly four parts or we can say that 4V’s: 1. volume, 2. variant...
3. Variety
Variety refers to the origin of data that needs to be processed. Variety is a very important characteristic for the big data points of view and mainly variety means a different format of data that do not ground themselves to storage in Structured Relational Database Management System. That basically includes big list of the data such as text data, email, audio, video, picture, moving image, device data, sensor data, etc. In some research studies this is estimated to account for 90% or more then data is in organizations.

4. Volume
Volume refers to the amount of data that needs to be processed. Basically the second V is a Volume; this is basically the amount of data generated by company, organizations or individuals. Big data means amount of data in a terabytes or petabytes. So it is the most immediate challenge of big data, as it requires a scalable storage and support for complex, distributed queries across multipule data sources.

5. Velocity
Velocity refers to the speed at which new big data is generated and the speed at which data moves around us. Just think that in a second how many transactions was done, in a second how many photos will be post by the Instagram users, in a second how many videos will be downloading or uploads. Now big data technology provides us to analyze the data while it is being generated, without ever put it into databases.

6. Veracity
Veracity is commonly refers to the messiness or trustworthy kind of data. Big data has many types, quality and accuracy are less but big data and analytics technology now provide us to work with these types of data. The volumes often make up for the lack of quality or accuracy.

6. Big Data Challenges
Big Data is basically in the format of uncompressed data so it is very large, complex and difficult to process in traditional data processing application. So in this kind of huge dataset it’s very difficult to visualize, analyze, search, storage and transfer the data for any of the company or organization. And last but not least one of the most important challenges is privacy and security. And these are the biggest challenges for big company to how to solve this kind of problem. So here we provide the some of the tools or techniques which will be useful to different organization to face the different challenges like a analyze the big data we have a different analysis tools, for visualize the big data we have a different visualization tools for store the big data we here describe the storing tools & transferring tools as well. All these tools we mentioned here with its different characteristics in a tabular form.
6.1 Big data Analysis & Analysis Tools

Big data analysis is one of the biggest problems for the big data point of view and most of the companies suffering from this problem. Big data analytics is the one kind of process to examine a big data to find out hidden patterns, unknown correlations and other useful information that can be used to make superior decisions. With big data analytics, scientists and others can analyze massive volumes of data that conventional analytics and business intelligence. Big data analytics is used for analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights.

<table>
<thead>
<tr>
<th>Analysis Tools Name</th>
<th>Features</th>
<th>Work with /Support</th>
<th>OS Support</th>
<th>Free/Paid</th>
<th>Official Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td>GridGain</td>
<td>Memory processing for fast analysis of real-time data</td>
<td>HDFS</td>
<td>Windows, Linux, OS X</td>
<td>Free: GIT HUB</td>
<td><a href="http://www.gridgain.com">http://www.gridgain.com</a></td>
</tr>
<tr>
<td>Jaspersoft</td>
<td>Real-time analytics, Fast and Easy Integrate all your data, Easy to use</td>
<td>Work with any big data store HADOOP, MongoDB, Cloudera</td>
<td>Windows, Linux</td>
<td>Free</td>
<td><a href="http://www.jaspersoft.com">http://www.jaspersoft.com</a></td>
</tr>
<tr>
<td>Upsight</td>
<td>Enterprise-grade analytics, Unlimited Data Storage</td>
<td>Hadoop</td>
<td>Windows, Linux</td>
<td>Free</td>
<td><a href="http://www.up">http://www.up</a> sight.com</td>
</tr>
<tr>
<td>Infochimps Cloud</td>
<td>Log and Mobile data analysis, Fraud detection &amp; Risk analysis, Integrated with different data sources, CRM, Solutions</td>
<td>NoSQL DB &amp; Hadoop Cluster</td>
<td>Windows, Linux</td>
<td>Paid</td>
<td><a href="http://www.infochimps.com">http://www.infochimps.com</a></td>
</tr>
</tbody>
</table>

6.2 Bigdata Visualization & Visualization Tools
“Picture Is Worth A Thousand Words”

Past studies show how the brain processes images 60,000s faster than the text. Visualization is basically the presentation of data in a format of graphics or in pictorial. Visualization provides a facility to people visualize their data and represent it in charts and maps to understand information more easily and quickly. Visualizing Big Data is one of the biggest challenges in the field of the digital data and so many companies are suffering from this challenge. So here we describe some of the tools that will very helpful to companies or the developers/scientists to face this kind of challenges.
# Table 2: Comparative Study of Visualization, Transfer & Storage Tools

<table>
<thead>
<tr>
<th>Visualization Tools Name</th>
<th>Features</th>
<th>Web/Tools</th>
<th>OS Support</th>
<th>Free/Paid</th>
<th>Official Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLOT</td>
<td>• Supports lines, plots, filled areas in any combination</td>
<td>Web Browser Based Support: IE, Mozilla, Safari, Opera</td>
<td>Windows</td>
<td>Free</td>
<td><a href="http://www.flotcharts.org">www.flotcharts.org</a></td>
</tr>
<tr>
<td>Q Research Software</td>
<td>• Used for research and data visualization</td>
<td>Tools Available</td>
<td>Windows</td>
<td>Paid</td>
<td><a href="http://www.q-researchsoftware.com">www.q-researchsoftware.com</a></td>
</tr>
<tr>
<td>Lumify</td>
<td>• Used for analysis and visualization</td>
<td>Web Based</td>
<td>Open Source</td>
<td></td>
<td><a href="http://lumify.io">http://lumify.io</a></td>
</tr>
<tr>
<td>Polymaps</td>
<td>• Polymaps is a kind of JavaScript library that provide us a slippy maps in a style of google maps, modest map &amp; Open Layers</td>
<td>Web Based</td>
<td>Open Source</td>
<td></td>
<td><a href="http://polymaps.org">http://polymaps.org</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Storage Tools Name</th>
<th>• Features</th>
<th>Web/Tools</th>
<th>OS Support</th>
<th>Free/Paid</th>
<th>Official Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hbase</td>
<td>• Store non-relational data for Hadoop, linear &amp; modular scalability, consistent reads and writes and automatic failover support</td>
<td>Tools Java Application based</td>
<td>OS Independent</td>
<td>Free</td>
<td><a href="http://hbase.apache.org">http://hbase.apache.org</a></td>
</tr>
<tr>
<td>MongoDB</td>
<td>• Support humongous databases, full index support, replication &amp; high availability, used for document oriented storage</td>
<td>Tools Available</td>
<td>Windows, Linux OSX, Solaris</td>
<td>Free</td>
<td><a href="http://www.mongodb.org">http://www.mongodb.org</a></td>
</tr>
<tr>
<td>Sqoop</td>
<td>• Transfer data between Hadoop and RDBMS, Used for bulk data transfer</td>
<td>Tools Os Independent</td>
<td></td>
<td></td>
<td><a href="http://sqoop.apache.org/">http://sqoop.apache.org/</a></td>
</tr>
<tr>
<td>Flume</td>
<td>• Aggregates and transfers log data from HDFS, robust and fault-tolerant</td>
<td>Tools Java Based</td>
<td>Windows, Linux, OS X</td>
<td>Free</td>
<td><a href="https://cwiki.apache.org/confluence/display/FLUME/Home">https://cwiki.apache.org/confluence/display/FLUME/Home</a></td>
</tr>
</tbody>
</table>

- 503 -
10th International CABLIBER 2015

New Big Things in Era of Digital Data...

Figure 10: SAS Screenshot

Figure 11: Lumify Screenshot

Figure 12: Polymaps Screenshot

Figure 13: Flot Screenshot

Figure 14: Q Research Software

Figure 15: MongoDB Screenshot
7. Bigdata to Address National Priorities

Now a day's Big Data are everywhere or each and every field generated this kind of massive data here we take some of the example of different fields.

Figure 16: HBase Tools Screenshot

Figure 17: Health & Wealth, E-Library & E-Education, Transportation, Science & Technology, Robotics & Smart system, Internet, Disaster & Emergency, Cyber Security
8. Conclusion

In this paper we described actual meaning of the Big Data, and how the big data is different than the traditional data, as well as the different types and characteristics of the big data. After here we mentioned the different challenges of the Big Data for organization or company point of view & we also discussed organization how to face these challenges using the different big data tools. In the last we described the comparative study of the different tools and its screenshot. Researcher point of view almost all of the challenges were solved regarding to the Big Data but last not least privacy and security is the still biggest challenges to the big data researchers.

References

New Big Things in Era of Digital Data...

41. https://jira.mongodb.org/browse/JAVA-86

About Authors

Mr. Tapan P Gondaliya, Research Scholar, School Of Computer Science, RK University, Rajkot, Gujarat. Email: tgondaliya@acm.org

Dr. Hiren D Joshi, Associate Professor, School Of Computer Science, Dr.Babasaheb Ambedkar Open University, Ahmadabad, Gujarat. Email: hiren.joshi@baou.edu.in

Mr. Hardik J. Joshi, Assistant Professor, School of Computer Science, Gujarat University, Ahmadabad, Gujarat. Email: joshee@acm.org