New Big Things in Era of Digital Data : 'Big Data' & Big Data Challenges with its Solution Using Different Tools

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 Exabyte's 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 Exabyte's 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 a 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 4'Vs 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-

10th International CALIBER-2015
HP University and IIAS, Shimla, Himachal Pradesh, India
March 12-14, 2015

© INFLIBNET Centre, Gandhinagar, Gujarat, India

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.

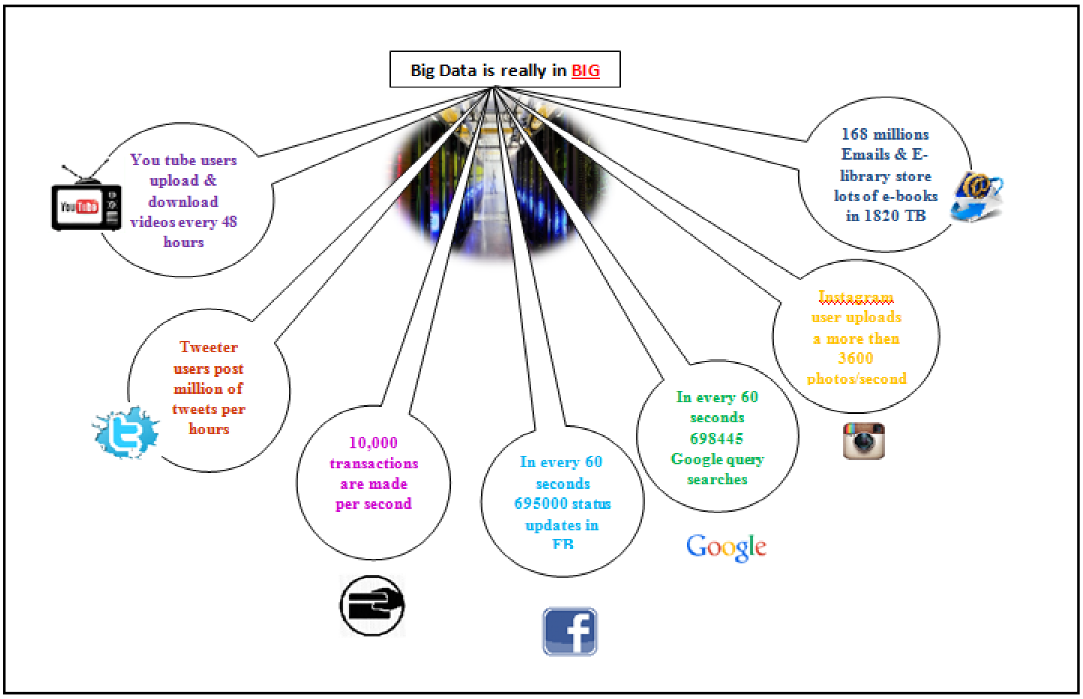

Big Data is really in BIG
You tube users upload & download videos every 48 hours
168 millions Emails & E-library store lots of e-books in 1820 TB
Tweeter users post million of tweets per hours
Instagram user uploads a more then 3600 photos/second
10,000 transactions are made per second
In every 60 seconds 695000 status updates in FB
In every 60 seconds 698445 Google query searches
Google

Figure 1: Source Where Generate the Big Data

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 is 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 system, integration and management of that data into a central database and then the analysis, visualization and other operation can perform 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.

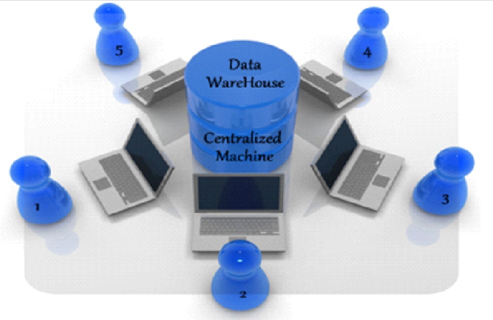


Figure 2: Traditional Data Warehouse

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

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

4. Types of Big Data

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

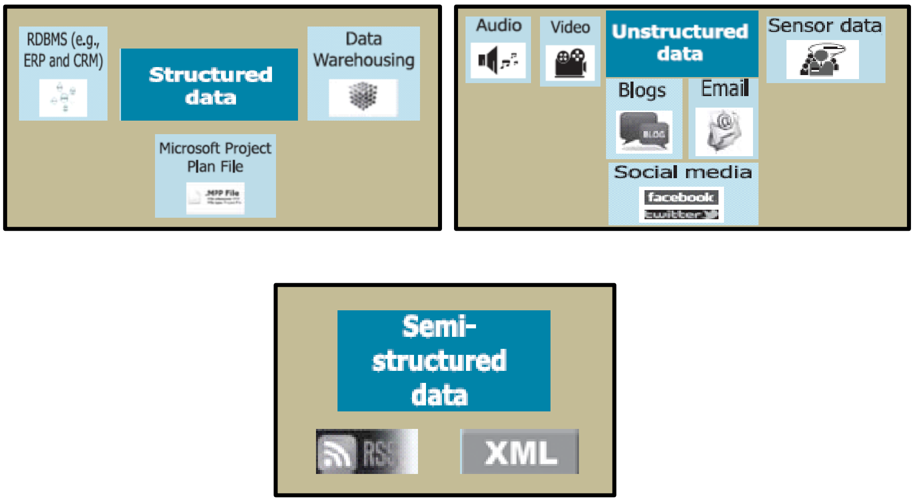


Figure 3: Type of Data

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 predefined data model or is not organized in a predefined 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.variety, 3.velocity and 4.veracity. These four characteristics are well explained as under with its diagrams.

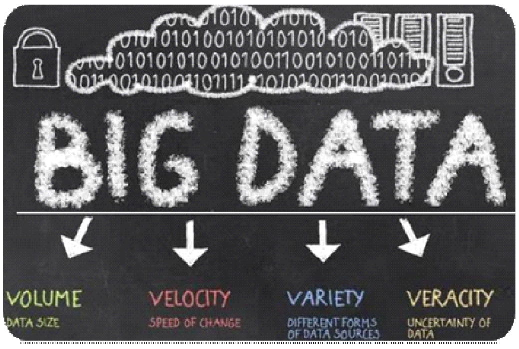


Figure 4: Big Data Characteristics

5.1 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, devise data, sensor data, etc. In some research studies this is estimated to account for 90% or more then data is in organizations.

5.2 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 multiple data sources.

5.3 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 moment 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.

5.4 Veracity

Veracity is commonly refers to the messiness or trustworthy kind of data. Bid 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. [3] 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 2: Comparative Study of Big Data Analysis Tools

Analysis Tools Name	Features	Work with /Support	OS Support	Free/Paid	Official Websites
GridGain	• Memory processing for fast analysis of real-time data	HDFS	Windows, Linux, OS X	Free:- GIT HUB	http://www.gridgain.com
Jaspersoft	• Real-time analytics • Fast and Easy Integrate all your data • Easy to use	Work with any big data store HADOOP, MongoDB, Cloudera	Windows, Linux	Free	http://www.jaspersoft.com
Upsight	• enterprise-grade analytics • Unlimited Data Storage	Hadoop	Windows, Linux	Free Analytics & Storage Paid	http://www.upsight.com
Karmasphere	• Analytics a big data using Hadoop • Sql data explorer • Support 250 Hadoop algorithm • SAS,SPSS,R Analytics model	Hadoop	Windows, Linux	Paid	http://www.fico.com
Infochimps Cloud	• Log and Mobile data analysis • Fraud detection & Risk analysis • Integrated with different data sources • CRM Solutions	NoSQL DB & Hadoop Cluster	Windows, Linux	Paid	http://www.infochimps.com

6.2 Bigdata Visualization & Visualization Tools

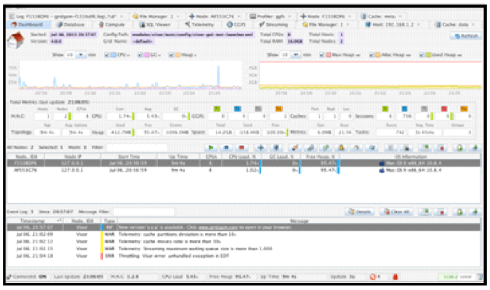


Figure 5: GreedGain Tools

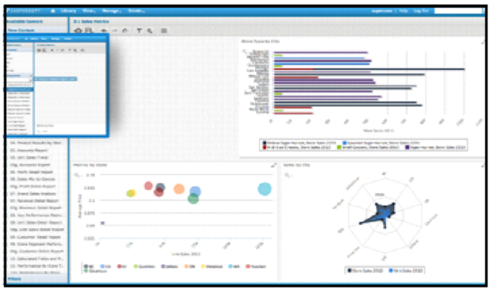


Figure 6: Jaspersoft Tools

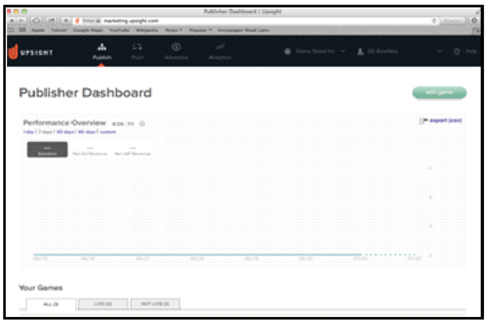


Figure 7: Upsight Tools

Figure 8: Karmasphere Tools

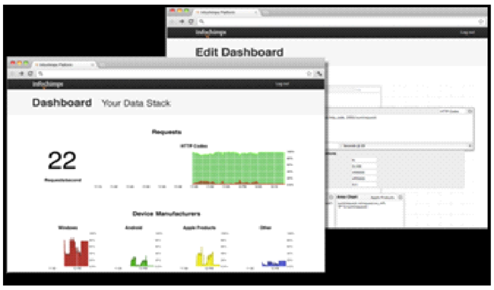


Figure 9: Infochimps Cloud Tools

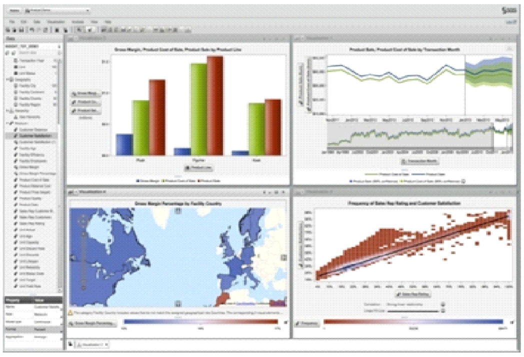
“Picture Is Worth A Thousand Words”

Past studies show how the brain processes images 60,000s fasters 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 there data and represent it in a charts and maps to understand information more easily and quickly. Visualize the Big Data is a one of the biggest challenges in the field of the digital data and so many company suffering from this challenges so here we describe some of the tools that will very helpful to companies or the developers/scientists for face this kind of challenges.

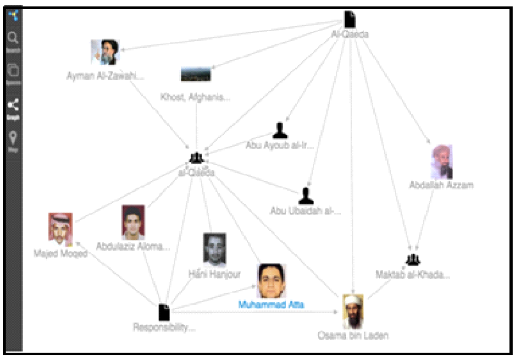
Table 2: Comparative study of Visualization, Transfer & Storage Tools

Visualization Tools Name	Features	Web/Tools	OS Support	Free/ Paid	Official Websites
FLOT	• Supports lines, plots, filled areas in any combination • Direct canvas access for drawing • Data points, interactive charts, stacked charts, panning and zooming	Web Browser Based Support:- IE, Mozila, Safari, Opera	Windows	Free	www.flotcharts.org
SAS Visual Analytics	• Used For Visualization, Analysis, and Reporting • Easy integration of manipulate user data • Create web based interactive report	Web & Tools Available	Windows	Paid	www.sas.com/en_us/software/ business-intelligence/visual-analytics.html
Q Research Software	• Used for research and data visualization • Export to Word, Excel and PowerPoint in graphical format •Multiple chart supported •Updatable with real data •Histograms & Scatter Plots	Tools Available	Windows	Paid	www.q-researchsoftware.com
Lumify	• Used for Analysis and visualization • Provide a facility to user to Search, Graph visualization, link analysis, multimedia analysis	Web Based windows	Open Source		http://lumify.io
Polymaps	• Polymaps is a kind of JavaScript library that provide us a slippy maps in a style of google maps, modest map & Open Layers • Polymaps is used in many different fields like Pale Down, satellite, statehood, Midnight Commander, Population Density, K Means Clustering, Internet Usage windows	Web Based	Open Source		http://polymaps.org

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



Figue10: SAS Screenshot



Figue11: Lumify Screenshot

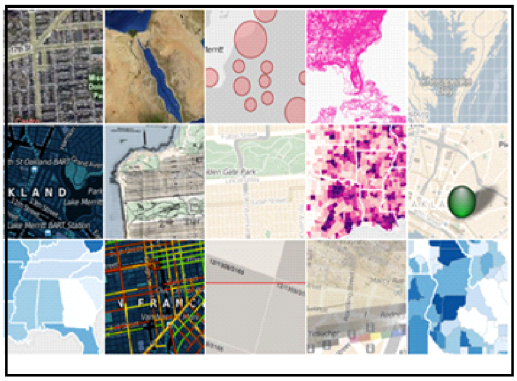


Figure 12: Polymaps Screenshot

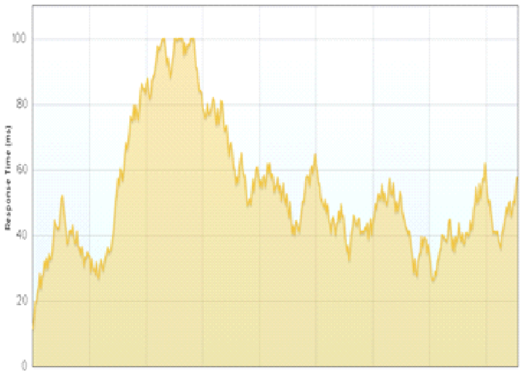
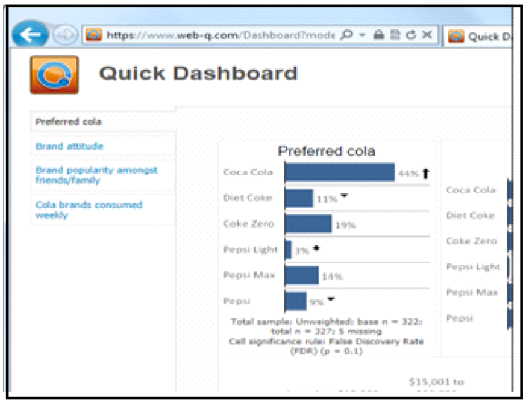


Figure 13: Flot Screenshot



https://www.web-q.com/Dashboard?mode

Quick D

Quick Dashboard

Preferred cola

Brand attitude

Brand popularity amongst friends/family

Cola brands consumed weekly

Preferred cola

Coca Cola	44%
Diet Coke	11%
Coke Zero	19%
Pepsi Light	3%
Pepsi Max	14%
Pepsi	9%

Total sample: Unweighted; base n = 322; total n = 327; 5 missing
Cell significance rule: False Discovery Rate (FDR) (p = 0.1)

Figure 14: Q Research Software

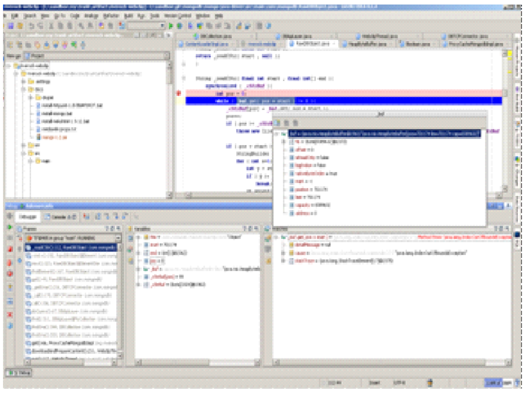
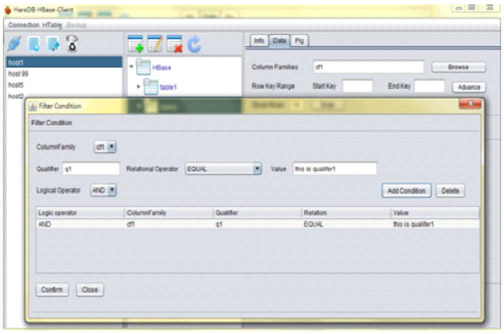


Figure 15: MongoDB Screenshot



Column Families: cf1

Row Key Range Start Key: End Key:

Filter Condition

ColumnFamily: cf1

Qualifier: q1 Relational Operator: EQUAL Value: this is qualifer1

Logical Operator: AND

Logic operator	ColumnFamily	Qualifier	Relation	Value
AND	cf1	q1	EQUAL	this is qualifer1

Figure 16: HBase Tools 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 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

1. Tapan P. Gondaliya, Dr. Hiren D. Joshi. June 2014. “Big Data challenges and Hadoop as one of the solution of big data with its Modules”. IJSER. ISSN 2229-5518. Volume 5. Issue 6. June-2014.
2. Nov 2012. “Big Data”. Infocomm Technology Roadmap 2012. http://www.ida.gov.sg/InfocommLandscape/Technology/Technology-Roadmap.
3. http://en.wikipedia.org/wiki/Big_data
4. Zhanpeng Huang, Pan Hui, Christoph Peylo. July 2014. “When Augmented Reality Meets Big Data”. arXiv.
5. Dylan Maltby. Oct-2011. “Big Data Analytics”. ASIST 2011.
6. Nada Elgendy, Ahmed Elragal. 2015. “Big Data Analytics: A Literature Review Paper”, Advances in Data Mining. Applications and Theoretical Aspects Lecture Notes in Computer Science. Springer. Volume 8557. pp 214-227.
7. Edd Dumbill. Jan 2012. “An introduction to the big data landscape”. radar.oreilly.com
8. Carl W. Olofson, Dan Vesset. August 2012. “Big Data: Trends, Strategies, and SAP Technology”. IDC Analyze the future. White Paper.
9. Jason Bloomberg. Jan 2013. “The Big Data Long Tail”. http://www.devx.com/blog/the-big-data-long-tail.html
10. Sagiroglu S, Sinanc D. May 2013. “Big Data: A Review”. CTS 2013 IEEE Conference. ISBN: 978-1-4673-6403-4.
11. Bill Franks, Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman, JorisMeys, Andrie de Vries, Mark Gardener, Dr. Murray Logan, Michael J. Crawley, Deborah J. Rumsey, Johannes Ledolter, Stephane Tuffery, Dean Abbott. 2013. “Introducing Big Data Analytics and Predictive Modeling”. Wrox Certified Big Data Analyst (WCBDA). Wiley Publishers.
12. Shilpa, Manjit Kaur. March 2014. “Big Data Visualization tool with Advancement of Challenges”. IJARCSSE. Vol-4. Issue 3. ISSN: 2277 128X.
13. Vangie Beal. “Structured data”. Webopedia.com.
14. Michael Walker. Dec 2012. Blog. “Structured vs. Unstructured Data: The Rise of Data Anarchy”. datasciencecentral.com
15. Creative Inc. 2012. “Big Data the Next Big Things”. Nasscom (New Delhi)
16. http://en.wikipedia.org/wiki/Unstructured_data
17. Colin White. Jan 2012. Article. “What Is Big Data and Why Do We Need It?” Technology Transfer.
18. http://www.ibmbigdatahub.com/infographic/four-vs-big-data

19. Bill Vorhies. October 2013. “How Many “V”s in Big Data – The Characteristics that Define Big Data”. Data Magnum.

20. Canada Health Infoway. 2013. Emerging Technology Series. “Big Data Analytics in Health”.

21. Michael C. Daconta. Jan 2014. Blog. “Is Hadoop the death of data warehousing?” Reality Check.

22. Ahmed Banafa. Blog. “Small Data vs. Big Data: Back to the basics”. Linkedin.

23. David Floyer. Nov 2014. “Enterprise Big Data”. wikibon.org.

24. Bernard Marr, March 2014, “Big Data: The 5 Vs Everyone Must Know”, linkedin.com

25. http://www.sas.com/en_us/insights/analytics/big-data-analytics.html

26. http://searchbusinessanalytics.techtarget.com/definition/big-data analytics

27. William Toll. March 2014. Blog. “Top 45 Big Data Tools for Developers”. profitbricks.com

28. http://www.gridgain.com/products/management-console-2/

29. http://www.zdnet.com/article/jaspersoft-6-0-packs-in-more-dashboard-tools-for-developers/

30. https://help.gamesalad.com/hc/en-us/articles/202528736-7-9-Setting-up-Upsight-PlayHaven-in-your-GameSalad-game

31. http://image.slidesharecdn.com/finalplatforapresentationmike-george-131030175824-phpapp02/95/ga-project-4-student-presentation-platfora-21-638.jpg?cb=1383267818

32. http://siliconangle.com/blog/2012/04/10/infochimps-rounds-out-platform-with-dashpot/platform-dashboard-graphic/

33. http://www.datameer.com/product/data-visualization.html

34. http://www.sas.com/en_us/insights/big-data/data-visualization.html

35. JAN 2014. BLOG. “Lumify”. http://lumify.io/blog/2014/01/21/what-is-lumify.

36. http://polymaps.org

37. http://www.flotcharts.org

38. Jan. March 2012. Blog. “Share a Dashboard”. http://blog.q-researchsoftware.com.

39. http://hbase.apache.org

40. http://www.softpedia.com/get/Internet/Servers/Database-Utils/HareDB-HBase-Client.shtml

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

42. http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.html

43. Howard Wactlar. June 2012. “Big Data R&D Initiative”. NIST Big Data Meeting. National Science Foundation (NSF).

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