ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. I, Special Issue IV, December 2015 in association with BPC COLLEGE PIRAVOM
National Conference on Recent Trends in Data Mining (NCRTID-2015) - 10th & 11th December 2015

Big Data – A Modern Trend in Data Mining

Ms. Leeja Mathew
Assistant Professor,
Dept. of Computer Applications
B.P.C. College , Piravom
Ernakulam, Kerala
leejarejibpc@gmail.com

***Abstract*— Data Mining is known as knowledge discovery from data. But now a days there is a tremendous increase in data. So data mining tool is not sufficient to capture knowledge. Now come in the era of Big Data – an emerging growing dataset due to its volume , variety and velocity. This research paper discusses about the introduction and analysis of big data, various usages and applications of big data , tools used for big data analysis , solution providers of big data in India and contributed articles in big data.**

Keywords*— *Big Data, Apache Hadoop, Storm, R, MOA, Pegasus, GraphLab

I. INTRODUCTION

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume , velocity and variety.

According to Gartner Big data is defined as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

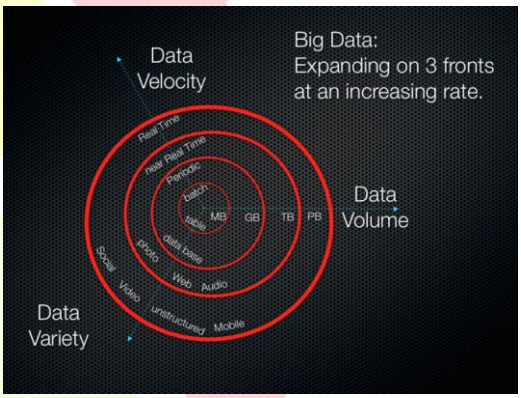

Fig. 1 Big Data

- Volume:- There is more data than ever before , its size continues increasing , but not the percent of data that our tool can process.
- Variety:- There are many different types of data , as text ,sensor data,audio,video,graph and more.
- Velocity:- Data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time.

Nowadays, there are two more V's :

- Variability:- There are changes in the structure of the data and how users want to interpret that data.
- Value:- Business values that give organization a compelling advantage , due to the ability to making decisions based in answering questions that were previously considered beyond reach.

Big data typically refers to the following types of data:

- Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data.
- Machine-generated /sensor data – includes Call Detail Records (—CDR‖), weblogs, smart meters,

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. I, Special Issue IV, December 2015 in association with BPC COLLEGE PIRAVOM
National Conference on Recent Trends in Data Mining (NCRTID-2015) - 10th & 11th December 2015

manufacturing sensors, equipment logs (often referred to as digital exhaust), and trading systems data.

- Social data – includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook.

The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44 xs between 2009 and 2020. The current report says that in the 9th year there are 30 crore users in twitter and there are 150 crore users in Facebook of their 11th year .But while it's often the most visible parameter, volume of data is not the only characteristic that matters. Big data can also be defined as "Big data is a large volume unstructured data which cannot be handled by standard database management systems like DBMS, RDBMS or ORDBMS".

II. Big Data Analysis

Big Data Analytics consists of 6Cs in the integrated Industry 4.0 and Cyber Physical Systems environment. 6C system that is consist of Connection (sensor and networks), Cloud (computing and data on demand), Cyber (model & memory), Content/context (meaning and correlation), Community (sharing & collaboration), and Customization (personalization and value). In this scenario and in order to provide useful insight to the factory management and gain correct content, data has to be processed with advanced tools (analytics and algorithms) to generate meaningful information. Considering the presence of visible and invisible issues in an industrial factory, the information generation algorithm has to capable of detecting and addressing invisible issues such as machine degradation, component wear, etc. in the factory floor.

Handling the three Vs helps organizations extract the value of Big Data. The value comes in turning the three Vs into the three is:

1) Informed intuition: predicting likely future occurrences and what course of actions is more likely to be successful.
2) Intelligence: looking at what is happening now in real time (or close to real time) and determining the action to take
3) Insight: reviewing what has happened and determining the action to take.

Data can come from a variety of sources (typically both internal and external to an organization) and in a variety of types. With the explosion of sensors, smart devices as well as social networking, data in an enterprise has become complex because it includes not only structured traditional relational data, but also semi-structured and unstructured data as shown in fig 2.

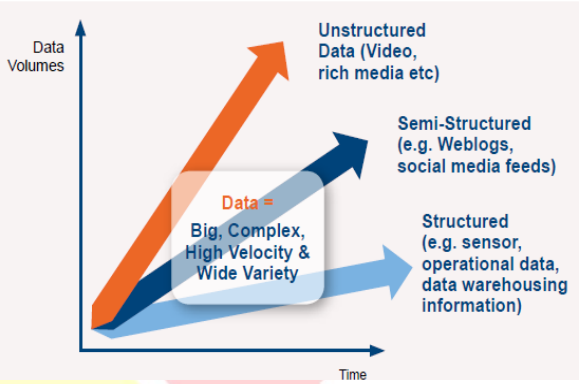

Fig. 2 Types of data

Structured data: This type describes data which is grouped into a relational scheme (e.g., rows and columns within a standard database). The data configuration and consistency allows it to respond to simple queries to arrive at usable information, based on an organization's parameters and operational needs.

Semi-structured data: This is a form of structured data that does not conform to an explicit and fixed schema. The data is inherently self-describing and contains tags or other markers to enforce hierarchies of records and fields within the data. Examples include weblogs and social media feeds.

Unstructured data: This type of data consists of formats which cannot easily be indexed into relational tables for analysis or querying. Examples include images, audio and video files.

III. Users And Applications Of Big Data

i) Science and Research

When the Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy. Continuing at a rate of about 200 GB per night, SDSS has amassed, SDSS has amassed more than 140 terabytes of information. When the Large Synoptic Survey Telescope, successor to SDSS, comes online in 2016 it is anticipated to acquire that amount of data every five days. Decoding the human genome originally took 10 years to process, now it can be achieved in less than a day: the DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100 times cheaper than the reduction in cost predicted by Moore's Law. The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.

ii) Government

- In 2012, the Obama administration announced the Big Data Research and Development Initiative, to explore how big

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. I, Special Issue IV, December 2015 in association with BPC COLLEGE PIRAVOM
National Conference on Recent Trends in Data Mining (NCRTID-2015) - 10th & 11th December 2015

data could be used to address important problems faced by the government. The initiative is composed of 84 different big data programs spread across six departments. Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign.

- Big data analysis was, in parts, responsible for the BJP and its allies to win a highly successful Indian General Election 2014.
- Kerala Water Authority Uses IBM Big Data & Analytics Technology for Seamless Water Distribution in Thiruvananthapuram: In September 2014: IBM (NYSE: IBM) announced that Kerala Water Authority (KWA), Government of Kerala, India is using IBM Analytics and Mobility solutions to analyze, monitor and manage water distribution in the city of Thiruvananthapuram. With the solutions, KWA aims to achieve 100 percent success in equitable water supply with the ability to monitor and flag irregularities in water usage using sensors and intelligent meters.
- Open engagement — through the Big Data Working Group and other groups such as the Australian Tax Office's Data Analytics Centre of Excellence, agency stakeholders in big data and its related technologies will be able to engage with industry, academia, non-government organizations and other interested parties locally and internationally. These engagements will help to build knowledge, spark ideas, generate growth and better inform decisions and solutions that meet the needs of the government, both on a national and local level.

iii) Private Sector

- **eBay.com** uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB datawarehouse(http://www.itnews.com.au/News/342615,inside-ebay8217s- 90pb-data-warehouse.aspx)
- **Amazon.com** handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- **Walmart** handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.
- **Facebook** handles 50 billion photos from its user base.
- **FICO** Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide. The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.
- **Windermere Real Estate** uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

In addition to the above the potential of Big Data to identify business trends, predict and manage disease outbreaks, combat crime, evolve population health management strategies, and manage road traffic has contributed to its acceptance as a powerful tool for greater operational efficiency, cost reduction, and reduced risk.

Conventional businesses like retail chains, banks, telecom operators, media houses and insurance companies, and government organizations which generate large quantities of transactional data through their operations are potential beneficiaries of Big Data. For example, the government may leverage big data projects to monitor and analyze the effectiveness of its expenditure on service sectors such as healthcare, education, and agriculture. However, the paucity of trained manpower hampers Big Data initiatives.

IV. Tools Used For Big Data Analysis

The Big Data phenomenon is intrinsically related to the open source software revolution. Large companies as Facebook, Yahoo!, Twitter, LinkedIn benefit and contribute working on open source projects. Big Data infrastructure deals with Hadoop, and other related software as:

- **Apache Hadoop [7]**: software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.
- Apache Hadoop related projects [17]: Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others.
- Apache S4 [14]: platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time.
- Storm [16]: software for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. I, Special Issue IV, December 2015 in association with BPC COLLEGE PIRAVOM
National Conference on Recent Trends in Data Mining (NCRTID-2015) - 10th & 11th December 2015

In Big Data Mining, there are many open source initiatives. The most popular are the following:

- Apache Mahout [8]: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.
- R [15]: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets.
- MOA [9]: Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The streams framework [10] provides an environment for defining and running stream processes using simple XML based definitions and is able to use MOA, Android and Storm. SAMOA [6] is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.
- Vowpal Wabbit [12]: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning.

More specific to Big Graph mining we found the following open source tools:

- Pegasus [11]: big graph mining system built on top of MapReduce. It allows to find patterns and anomalies in massive real-world graphs. See the paper by U. Kang and Christos Faloutsos in this issue.
- GraphLab [13]: high-level graph-parallel system built without using MapReduce. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices.

V.SOLUTION PROVIDERS OF BIG DATA IN INDIA

- **Bodhtree |Hyderabad** : The company's business domain expertise, coupled with rich technical competencies, enables it to define a Big Data strategy for an organisation, integrate Big Data into the organisation's overall IT roadmap, or architect and implement a solution. Bodhtree addresses some of the critical business challenges for global customers in the areas of customer analytics, natural language processing for social sentiment analysis, predictive analysis, log analysis and machine data analysis for MTBF, security and compliance validation. The offerings comprise Oracle Endeca services, Hadoop services, Google Big Query for SMBs, analytics and KTLO for Big Data stacks and patterns, and more.

- **Cloudera India** More organisations manage Big Data using Cloudera's Hadoop solutions, the company's 100 per cent open source and enterprise-ready distribution of Hadoop and related projects, than all other Hadoop distributions combined, claims the company. It's the most comprehensive Hadoop distribution available, and coupled with Cloudera Manager software and Cloudera support, it provides stability, integration and ease-of-use.With Cloudera Enterprise RTD and Cloudera Enterprise RTQ, the company has dramatically increased Hadoop's effectiveness by delivering on the promise of real-time operations. It has done this by making Apache HBase and Cloudera Impala technologies manageable and fully supported, overcoming Hadoop's past interactivity limitations, and ushering in a new era of speed-of-thought interactive analysis on data stored in a Hadoop cluster.

- **CSC India |Noida :** The CSC and Hortonworks alliance brings a wealth of expertise and benefits to organisations deploying enterprise-ready Apache Hadoop. This alliance enables CSC to resell subscription support for the Hortonworks Data Platform (HDP), the industry's only 100 per cent open source data platform powered by Apache Hadoop. CSC's solutions extend the reach of HDP to enterprises around the world by offering a supported Apache Hadoop platform to its enterprise customer base. Together, the CSC and Hortonworks alliance offers industry and technology expertise to enterprises deploying a next generation data architecture using Apache Hadoop, providing a fully tested, certified enterprise-ready Apache Hadoop solution to CSC's customers and clients.

- **EMC 2|New Delhi***:* EMC Isilon scale-out storage solutions for Hadoop combine EMC Isilon scale-out network attached storage (NAS) and EMC Greenplum HD with EMC consulting, training, and support for powerful analytics capabilities on a flexible, efficient data storage platform with native Hadoop integration. Isilon scale-out storage solutions for Hadoop marry the simplicity of Isilon scale-out NAS storage with the leading-edge analytics tools of Greenplum HD, resulting in a highly integrated, one-stop Hadoop solution that allows you to quickly extract new insight from your data. The unique integration of Isilon scale-out NAS with the Hadoop Distributed File System (HDFS) also eliminates the need for the resource-intensive

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. I, Special Issue IV, December 2015 in association with BPC COLLEGE PIRAVOM
National Conference on Recent Trends in Data Mining (NCRTID-2015) - 10th & 11th December 2015

task of importing and exporting data into and out of Hadoop.

- **Hitachi Data Systems|Mumbai:** Solutions from Hitachi Data Systems (HDS) include capitalising on the experience the company has gained building the machines and social infrastructure that generate much of Big Data: power plants, trains, medical equipment—Hitachi builds them all. One can take advantage of the deep industry expertise of HDS in managing and processing vast amounts of disparate data. One can also benefit from its vertical expertise to capitalise on Big Data. Clients can apply HDS's holistic perspective on Big Data, which the company uses to drive analytics innovation across all the Hitachi businesses.
- **IBM India|Bengaluru :** IBM is unique in having developed an enterprise class Big Data platform that allows you to address the full spectrum of Big Data business challenges. IBM is perhaps the only vendor with this broad and balanced view of Big Data. The benefit is pre-integration of its components to reduce your implementation time and cost. It uses Hadoop-based analytics processes and analyses any data type across commodity server clusters.
- **Oracle India Pvt Ltd|Gurgaon**:Oracle's Big Data strategy is centered on the idea that you can extend your current enterprise information architecture to incorporate big data. The Oracle Big Data Appliance is an engineered system that combines optimised hardware with a combination of open source.Software and specialised software developed by Oracle to deliver a complete, easy-to-deploy solution for acquiring, organising and loading Big Data into a database. With Big Data Connectors, the solution is tightly integrated with Oracle Exadata and Oracle Database. The Big Data Connectors deliver a high-performance Hadoop to Oracle Database integration solution and also enable optimised analysis using Oracle's distribution of open source R directly on Hadoop data. By providing efficient connectivity, Big Data Connectors enables analysis of all data in the enterprise—both structured and unstructured.Once data has been loaded from Oracle Big Data Appliance into Oracle Database or Oracle Exadata, Oracle Exalytics can be used to deliver a wealth of information to the business analyst. Oracle Exalytics is an engineered system providing speed-of-thought data access for the business community. It is optimised to run Oracle Business Intelligence Enterprise Edition with in-memory aggregation capabilities built into the system. Oracle Big Data Appliance, in conjunction with Oracle Exadata Database Machine and the Oracle Exalytics Business Intelligence Machine, delivers everything customers need to acquire, organise, analyse and maximise the value of Big Data within their enterprise.
- **Sesame Technologies |Calicut** : Sesame claims to be the first company in Kerala to provide solutions in Hadoop, Hive and Mongo DB. The company has a good team dedicated to providing Hadoop-based solutions. It can set up offshore development centres (ODCs) in a 'Build Operate and Transfer' mode and also on a 'Build and Operate' basis for clients who need Hadoop and Big Data-based solutions.
- **Teradata India | Mumbai** : Teradata provides a single source for all things related to Apache Hadoop. Teradata not only integrates with Hadoop, but also delivers engineered and supported Hortonworks appliances that go beyond standard Hadoop.The Teradata portfolio for Hadoop provides customers with the most trusted and flexible Hadoop product platforms in next-eneration data architecture with new best-in-class services, training and customer support.Clients can use Teradata as an exclusive, secure gateway into their Hadoop systems by only allowing access through Teradata SQL-H.
- **Trendwise Analytics |Bengaluru** : Trendwise Analytics is a premier global provider of Hadoop-based solutions, cloud computing and Big Data analytics. The company was recently recognised as one of the most promising Big Data companies in India. Trendwise Analytics has some premier partners like Jaspersoft, Cloudera, HPCC Systems, Hortonworks & Pingar, Datastax and MapRTechnology. Some of its satisfied clients are EGR, Appcara, Electronique Microtech, Cirrus, and Wiley. The company's other services include advanced analytics, social media analytics and mobile dashboards.

VI. Contributed Articles

Here is selected two contributions that together shows very significant state-of-the-art research in Big Data Mining, and that provides a broad overview of the field and its forecast to the future.

- **Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies** : This paper performs the session identification in log files using Hadoop in a distributed cluster. Apache Hadoop Mapreduce a data processing platform is used in pseudo distributed mode and in fully distributed mode. The framework effectively identifies the session utilized by the web surfer to recognize the unique users and pages accessed by the users. The identified session is analyzed in R to produce a statistical report based on total count of visit per day. The results are compared with non-hadoop approach a java environment, and it results in a better time efficiency, storage and processing speed of the proposed work.

- **Big Graph Mining: Algorithms and discoveries :** The patterns and anomalies in very large graphs with billions of nodes and edges efficiently is a problem. Big graphs are

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. I, Special Issue IV, December 2015 in association with BPC COLLEGE PIRAVOM
National Conference on Recent Trends in Data Mining (NCRTID-2015) - 10th & 11th December 2015

everywhere, ranging from social networks and mobile call networks to biological networks and the World Wide Web. This paper presents an overview of mining big graph, focusing in the use of the PEGASUS tool, showing some finding in the Web Graph and Titter social network. The paper gives inspirational future research directions for big graph mining.

VII. CONCLUSION

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. Big data is largely used in government, public and private sectors. The challenges in using big data are the massive spike and the noise available in data. Finally the Big Data is becoming the new frontier for scientific data research and for business applications.

References

[1] ^"a b c d e f"Data, data everywhere" (http://www.economist.com/node/15557443). The Economist.

[2] ^ "Community cleverness required" (http://www.nature.com/nature/journal/v455/n7209/full/455001a.html).

[3] Nature 455 (7209): 1. 4 September 2008. doi:10.1038/455001a (http://dx.doi.org/10.1038%2F455001a).

[4] ^ "IBM What is big data? — Bringing big data to the enterprise" (http://www.ibm.com/big-data/us/en/) www.ibm.com

[5] ^ "The FOUR Vs of Big Data" (http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-bigdata.jpg

[6] SAMOA, http://samoa-project.net, 2013.

[7] Apache Hadoop, http://hadoop.apache.org

[8] Apache Mahout, http://mahout.apache.org.

[9] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa. cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[10] C. Bockermann and H. Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012.

[11] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS: Mining Billion-Scale Graphs in the Cloud. 2012.

[12] J. Langford. Vowpal Wabbit, http://hunch.net/˜vw/, 2011. [21] D. J. Leinweber. Stupid Data Miner Tricks: Overfitting the S&P 500. The Journal of Investing, 16:15- 22, 2007.

[13] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In Conference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, California, July 2010.

[14] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed Stream Computing Platform. In ICDM Workshops, pages 170–177, 2010.

[15] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0

[16] Storm, http://storm-project.net.

[17] P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Companies,Incorporated, 2011.

[18]https://youtu.be/Sr1y5Qm2K0o