Big Data: Does it Call for Distributed File System

Komal Verma
Scholar
Amity University
Uttar Pradesh
Lucknow-226028

Rajiv Pandey, PhD
Member IEEE
Amity University
Uttar Pradesh
Lucknow-226028

Arpit Gupta
Scholar
Amity University
Uttar Pradesh
Lucknow-226028

ABSTRACT

Today, in order to support decision for strategic advantages alignment, companies' have started to realize the importance of using large data. It is being observed through different study cases that "Large data usually demands for faster processing". As a result, companies are now investing more in processing larger sets of data rather than investing in expensive algorithms. A larger amount of data gives a better inference for decision making but also working with it can create challenge due to processing limitations. In order to easily manage and use this large amount of content in a proper systematic manner, Big Data, HDFS & other file systems were being introduced.

Big data is used for 'larger data sets having more varied and complex structure, having problems in analyzing, visualizing and storing for further processing'. The process of examining such large amounts of data inorderto reveal hidden patterns and secret correlations is named as Big Data Analytics. The useful information for companies or organizations will help them in gaining richer and deeper insights and getting advantages over the competition. Implementation of this Big Data needs to be analyzed and executed as accurately as possible. This term paper give an overview about what Big Data is, its classification, challenges it faces, need for Distributed File System, Hadoop and its components i.e. Hadoop Distributed file System and Map Reduce, and application of HDFS in Cloud Computing

Keywords

Big Data,Hadoop,Cloud Computing

1. INTRODUCTION

About 5 exabytes (1018 bytes) of data was being created by human up-till 2003. Today this amount of information is created just in span of two days. In the year 2012, Digital Data has grown up to 2.72 zettabytes (approx.) but the sum is still multiplying [1][2]. Big data and data analytics are hot topics in both the popular and business world. Today, many organizations are collecting, storing, and analysing massive amounts of data daily. This data is referred to as "Big Data" due to its huge volume, the velocity with which it arrives, and the variety of forms it takes. Big data is making another era of choice bolster data administration. Indeed, even the quality in analysing unstructured data (e.g. email and records) has been surely known. What's new is the meeting up of advances in computer technology and software, new source of data (e.g., online networking), and business opportunity. This has made the present interest and opportunities in big data analytics.

Managing, storing, assessing and analysing the data in a proper and safe manner is being done by Hadoop. HDFS and Map Reduce -together as components of Hadoop- do the managing and storing purpose. Hadoop Distributed File System is basically a Distributed File System that is being designed to run on the commodity hardware. It is highly-tolerant and is designed to be work on low-cost hardware. It provides high through-put access to the data application data and is very much suitable for the large sized data sets.

Cloud computing technology is widely reorganized for IT companies which have been launching their own commercial products. These companies would be making cloud computing as their priority in future development. But to lead to such a vast data, it may create problem in handling it. Thus to manage and store the information, HDFS is used.

This term paper explores about the how the data is being managed and stored through Distributed File System. This paper is divided into five sections. Section 1: "Introduction", giving a brief description of the term paper. Section 2: "Big Data", describing about Big Data, transformative potential of Big Data in five Domains, characteristics of big data and the challenges it faces. Section 3:"Distributed File System" describing what is distributed file system".Section 4: "Hadoop", describing about what is Apache Hadoop, its purpose, its components and their procedures. Section 5: "K-Meansused for HDFS in Cloud Computing", explaining about the use of HDFS in cloud computing.

2. BIG DATA

2.1 Important Issue

Big Data applies to the information that is difficult to process or rather cannot be processed or analysed by traditional computing processes or tools. Data is growing exponentially due to number of reasons. Organization in educational, financial health related field and several other sectors are also capturing large amount of data continuously. Organizations are facing problems in analysing data, extracting required information or adding information to the data, as they have access to wealth information to maintain. They actually don't know how to extract valuable content out from it, sitting in its most raw form, i.e. semi-structured or even unstructured form. Considering the example of Facebook, there are about 1005 million accounts that are active monthly, using 70 languages, 160 billion photos being uploaded regularly, 140 billion friend connections are established, every day 35 billion pieces of content and 3.41 billion likes and comments are being posted. Keeping a track of this much amount of data regularly generated is difficulty. Thus to analyse Big Data, it requires a major step towards traditional data analysis, characterized by its components: variety, velocity, volume, veracity, variability and complexity.

2.2 Transformative Potential of Big Data in Five Domains

McKinsey Global Institute specified the potential of big data in five main topics [2]:

- **Public sector:**
  - creating transparency by available related data
  - discovering needs and improving performance
- **Personal location data:**
  - smarter routing
  - geo targeted advertising or emergency response
- **Retail:**
  - store and analysing variety and price of products
  - their placement design
- **Healthcare:**
  - clinical decision
  - keeping record of individual analytics applied for patient profile
- **Manufacturing:**
  - supporting sale
  - developing production operations

2.2.Characteristics of Big Data

"Legacy systems will remain necessary for specific high-value, low-volume workloads, and compliment the use of Hadoop - optimizing the data management structure in the organization by putting the right Big Data workloads in the right systems"[13].

Volume, variety, velocity, veracity, variability and complexity, together define Big Data. These have actually created the need for a new class of capabilities to argument the way things are done and managed to provide better machine and human understand ability and control over the existing domains and the ability to response to act on them. Big Data starts with large-volume, Heterogeneous, an autonomous source with distributed and centralized control and seeks to explore complex and evolving relationships among data [3], HACE theorem.

- **Volume**- Data on the earth is growing exponentially due to many reasons. Volume possess the greatest challenge and opportunity as Big Data can help many organization in understanding people better and allocated resources effectively. However, traditional computing methods face difficulties in handling the un-scalable data and its magnitude. The data amount has changed from terabyte to petabyte and now to zettabytes.

- **Velocity**- velocity in Big Data also raises many issues with the rate with which it is flowing in many organizations exceeding the capacity of the organization systems. Storing and analysing data in time can be quite challenging. For time limited processes, big data is used, as it streams order to provide maximum value to the organizations.

- **Variety**- variety of data to be processed is getting diverse. Different variety of data type is difficult to be characterised and process by using traditional processes. Data that is received for different sources can be structured, unstructured or semi-structured. Structured data is a fully tagged and easily sorted data easy to interpret in process. The unstructured data is a raw data, random and difficult to analyse. While the semi-structured data the combination of both types of data (structured and unstructured).

- **Veracity**- Since numerous business pioneers don't believe the data that they use to decide. Therefore, building up trust in big data exhibits a tremendous test as the mixed bag and number of sources develops.

- **Variability**- this factor could be a problem for those who analyse data. As sometime, due to some inconsistency, there may he hampering of data thus creating difficulty in managing and analysing data.

- **Complexity**- there may be cases when it may become difficult to manage such a large volume and variety of data, so that could be linked to each other in order to grasp information from different places easily. This situation is call 'complexity' of Big Data.

2.3.Challenges Faced by Big Data

The challenges thatare encountered in storage and computing of big data [4]:

a) Solution based on cloud computting-

The common issue to be focussed on is the security of data and confidential company documents. This can be done by ERP, CRM, Document Management, Data Warehouses and Business Intelligence. These solutions offer the companies much flexibility and cost savings opportunities as compared to those offered by more traditional solutions. These solutions even raise dimensions that are related to the security and overall management of big data.

b) Big Data utility and Understanding-

It is an apprehensive task for most industries and companies that are dealing with big data just in order to understand the data that is available to be used and processes, determining the best way how to use the data based on the companies' industry, strategy, and tactics. The analysis of these data is to be done on regular bases so as to avoid loss of data as the data landscape change constantly with ever-increasing rate.

c) Archiving and Disposal of Big Data-

Since big data is larger in context to volume and structured variety, it will lose its value to current time decision-making over-time. Thus, it is necessary to use new, helpful and safe tools and technologies for analysing, storing and accessing data without sacrificing the effectiveness of using big data to current business needs.

d) New, Complex and Continuously Emerging Technologies-

Since day-by-day new technology is being introduced in order to utilize and maintain Big Data, it is important for the companies to know and understand these technologies at an ever-accelerating pace. Also get engage potentially with these technology provider and partner rather than with their earlier ones. Like with all technologies, entering into the world Big Data and remain engaged with the data, processing, storing and analysing content, is quite expensive. And it is necessary to maintain a balance with the business needsassociated to big data with associated costs of entering the big data world.

3. DISTRIBUTED FILE SYSTEM

Today, storing each day record on the hardware might be quite hectic. Similarly assign a particular form of information over an internet might take a longer period of time just reading the whole content as the data is roughly stored in the hardware. To store the data properly in assessing data in as least time as possible, a system known as distributed file system was being introduced; in this system the whole problem is being divided in number of sub-problems. The problems are being processed parallel executing the particular assigned problem. In the end the result of all the sub-problems are being summed up and the result is being supplied to the user.

4. HADOOP

4.1 Introduction

Hadoop is a top-level Apache project in the Apache Software Foundation that's written in Java. It a fast-growing big-data processing platform is being defined as "the distributed processing of large set of data across clusters of commodity servers, being enabled an open source software project"[6]. It is being intended to scale up from single server to those of thousands of machines in a sector, with a high capacity of detecting and correcting faults, so that the data can be assessed and stored in best possible manner.

Developed by Doug Cutting, Cloudera's Chief Architect and the Chairman of the Apache Software Foundation, Hadoop was born out of need as data from the web exploded, and it grew to such a large extend that it became difficult for the traditional methods to handle the situation [7]. Hadoop was initially inspired by papers published by Google sketching out, its way to deal with oversee managing an overwhelming slide of data, and has following turned into the certainty of standard for storing, processing and analysing a few hundreds of terabytes, and even petabytes of data.

Apache Hadoop is purely an open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, restrictive hardware and different systems to process and store data, Hadoop enables distributed parallel processing of large volume of data across reasonable servers that both store and process the data, and can scale without any limitation. Hadoop is not about high speed response times, real-time warehousing, or blazing transactional speed; it speaks the truth revelation and making the once close unthinkable conceivable from a versatility and examination point of view.

Hadoop has the ability to handle all types of data whether being structured, unstructured, pictures, communications records, audio files, log files, or email - regardless of its original pattern. Even when different variety of data is being stored in unrelated systems, it is still possible to store it all into Hadoop cluster without any prior need for a schema.

A traditional Hadoop cluster consists of a few basic components [8]:

- Hundreds of subordinate nodes that provide both Task Tracker and Data Node functionality.
- The Checkpoint node, which is a secondary NameNode that manages the on-disk representation of the Name Node metadata.
- The NameNode, which manages the Hadoop Distributed File System namespace.
- The Job Tracker node, which manages all jobs submitted to the Hadoop cluster and facilitates job and task scheduling.

These nodes performm all of the real work done by the cluster. As Data Nodes, they store all the blocks of data that make up the file system, and they serve I/O requests. As Task Trackers, they perform the job tasks assigned to them by the Job Tracker.

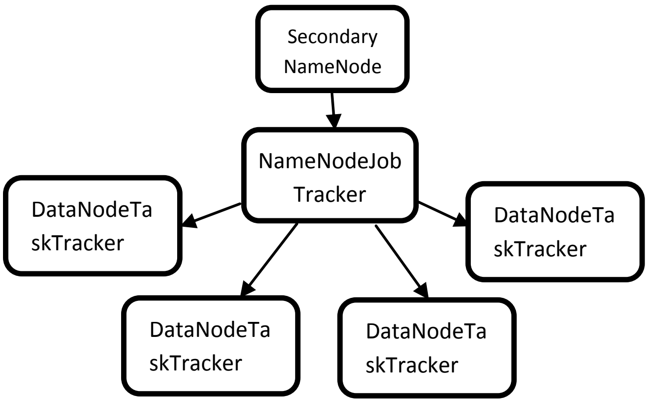

Fig.1 Hadoop Architecture [9]

4.2 Components of Hadoop

Apache Hadoop is divided into subprojects:

- *Map Reduce* – Map Reduce is considered to be the Heart of Hadoop. It is these programming paradigms that allow the distribution of sub-problems to several hundreds or thousands of servers across the Hadoop cluster.
- The term Map Reduce refers to separate and distinct processes broadly classified as Map and Reduce. The first job known as Map refers to the distribution of the problem into several small set of sub-problems which are processed parallely at the same time. There are cases where the sub-set are to be divided in small set for easy and fast processing. These small individual sub-problem/elements are broken down in tuples.
- The next job is Reduce. In this section of Hadoop, the output that is being received from the data tuples in the Map job is combined to form small set tuples. As the name suggests, Map process is always followed by Reduce process.
- In the Hadoop cluster, a Map Reduce program is referred to as a job. A job is executed by then breaking it down into a number of small pieces called tasks. An application submits a job to a specified node in a Hadoop cluster, which is running a daemon called the Job Tracker. This Job Tracker interact with the NameNode to find out where the data to be processed exists in the clusters, and sequently divides the job into Map and Reduce job. These tasks are scheduled on the nodes in the cluster where the data exists. This is called data locality. In Hadoop cluster, a set of continually running daemon, referred to as Task Tracker agents, monitor the status of each task. If while executing the task a failure

occurs, the Task Tracker reports this to the Job Tracker, which will then reschedule that task on another node in the cluster.

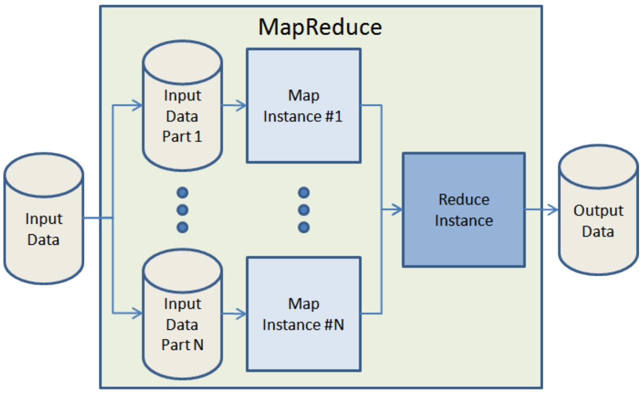

Fig. 2 Map Reduce Architecture [9]

- *Hadoop Distributed File System (HDFS)* – In the Hadoop cluster, data is broken down into a number of smaller Blocks and then distributed thought the cluster. In this way, the map and reduce functions can be executed on smaller subset of a larger data sets, and this provides the scalability that is needed for Big Data processing.

Though being one of the most popular Distributed file system had few drawbacks that were being solved by Hadoop [10].

- HDFS makes the processing of data faster.
- Distributed file system could store the data for a limited storage space, Hadoop solved the problem by providing a large amount of storage data
- In HDFS, it offers disappointment portecture; while in case of Distributed file system, there might be the chances of failure of the portecture.

HDFS is master/slave architecture. In it, all the Hadoop's data placement logic is being managed by a special server called NameNode (master). The server keeps track of all the data file that are in HDFS, such as where the file is created, where the block of contents are being stored, and more. DataNodes manages storage that is attached to the nodes that they run on. All the details of the NameNode are being stored in the memory which makes storing, manipulation or reading request in a quick response time. The server component that you use for the NameNode has very little chances of failure as compared to the rest of the servers in the Hadoop. Not only this, there is a regular backup of the data that is stored in the NameNode [11]. When you request of a Hadoop job, to reads the data and starts the Map-Reduce task, Hadoop call the NameNode server which holds the data the needs to be processed. Initially, a file is split into small blocks and nowthese blocks are stored in a group of DataNodes. The NameNode executesall of the file system namespace processes such as closing, opening, and renaming directories and files. While the DataNodes are responsible for providing read and write requests from the client of file system. The interaction is done basically when the data is being stored on different servers in the Hadoop cluster. This greatly reduces communication to the NameNode during job execution, which helps to remove the scalability of the solution.

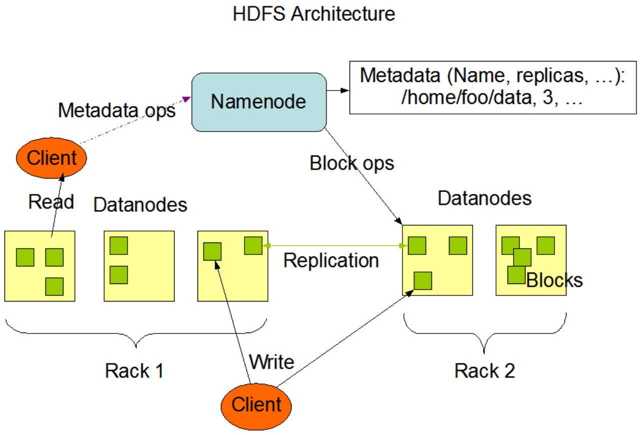

Fig. 5 Architecture of HDFS [9]

Hadoop is supplemented by an environment of Apache projects, such as Pig, Hive and Zookeeper, which extend the value of Hadoop and improve its usability. Due to the cost-effectiveness, scalability and streamlined architectures, Hadoop has changed the financial aspects and the flow of large scale computing, having a remarkable influence based on its four salient characteristics.

The cost advantage of Hadoop is thatin light of the way that it depends in an inside redundant data structure and is conveyed on industry standard servers as opposed to lavish particular information stockpiling frameworks, you can bear to store information not already practical.

5. HDFS IN CLOUD COMPUTING

Cloud computing is the concept of outsourcing of PC administration. It has a capability of providing on demand networking resources. It builds a virtual group of resources such as network, storage, central processing unit and memory. Being an immerging technology, cloud computing plays an important role in providing a network access and supports on-going open source cloud services.

The main purpose of cloud computing is providing quick, easy and adaptable access to computing resources. Cloud computing is also responsible in providing IT services that plays an important role providing computing, communication and storage resources in a safe environment based on a services as fast as possible which are virtually provided via internet platform[12].

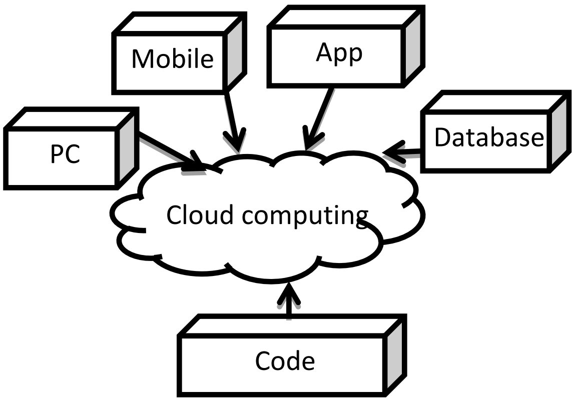

Fig. 4 General Structure of Cloud Computing [12]

Cluster algorithm is an important factor in data mining, especially in case of such a large system of data computing.

The division of data characteristics is the vital step in the storage and security of cloudcomputing. Cluster analysis includes a number of algorithms such as K-means. Optimizing the distance between the data and centre point is the main objective of K- means algorithm. Higher similarity of objects in the same cluster and least similarity between the cluster types arefew rule that clustering meets through such algorithm. It means the idea of "High cohesion, Low coupling" in software engineering. K-means algorithm, Map and Reduce are being separately performed in each turn. In the Hadoop platform, along with storing such a large data and managing the data file, HDFS even records the data nodes, where they are distributed and then gain results from them. The task of Map function is to work out the distance between each recorder and centre site, and re-mark the type which it belongs to. The input should be all wait Clustering Data and the Clustering Centre in the last round, also the Data Clustering<key, value>. On the basis of calculated results from Map function, the task of Reduce function is to calculatenew Clustering Centre and send it to the entire node,and update the results in the HDFS before the next iteration, until convergence.

Steps to work out Parallel K-means algorithm [12]:

Step 1: Select K samples arbitrarily as initial centre ID;

Step 2: Iterate and perform Map and Reduce;

- Each site receives cluster centre from centre site;
- Calculate sample size from every local cluster, and send it to centre site;
- Calculate new Global Clustered, and send it to each site.

Step 3: Repeat until convergence. Consequently, we can use K-means clustering method to separate the data which are similar each other. Combining with the Map/Reduce in the Hadoop, we can distribute the storage and use the data in cloud computing. The structure is shown as

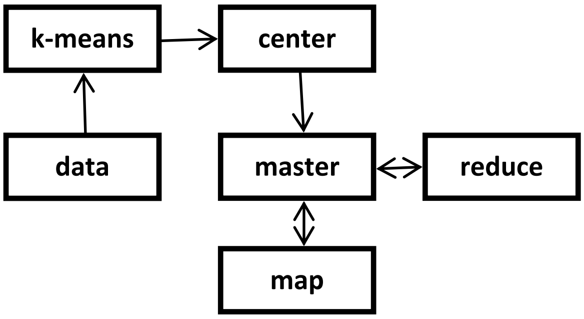

Fig 5: structure of combing K-means with the Map/Reduce in Hadoop

According to the figure, we can divide the process into five simple steps:

- In cloud computing, data will be distributed in a number blocks by using the K-means algorithm in order to ensure that data in each block have higher similarity.
- We redistribute the block by Centre controlled and then we allot the respective pointer to the each distributed data to make the Map/Reduce operation become more effective.
- Then we should notice to the Master Management about the location of these small block of data to facilitate the Master to assign the tasks.
- Master now splits the data point to Map to operate them.

At the end, Map return the intermediate value to Master then let Reduce operate. This procedure, which contrasts with the past calculation without utilizing the K-means, has more effect on the data distribution. Prior to the Map/Reduce, we utilize the K-intends to order the data sorts which is more viable to deal with the data classification.

6. CONCLUSION

Big Data demonstrates multiple features that are usually not visible in small relational datasets.The multi dimensional view of any big data demands a storage and computing mechanism that enables efficient storage and optimal and enhanced presentation of the analysed result set. The storage of big data is supported by distributed framework called HDFS. The distributed storage demands a quick efficient processing and analysis of data. This is achieved using the concept of map reduce. This paper has outlined the distributed storage framework of HDFS and the processing concept of Map Reduce. The paper also outlines Kmeans clustering and data analysis algorithm describing its relevance to Map reduces.

7. REFERENCES

[1] http://hpccsystems.com/ , last success 11.03.2013

[2] J. Manyika, M. chui, B. Brown, J. Bughin, R Dobbs,C. Roxburgh and A. H. Byers, "Big Data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011, http://www.mckinsey.com/~/media/McKinsey/dotcom/Insight%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI big data full report.ashx

[3] Xindong wu, Xingquan Zhu, Gong-Qing and Wel Ding, "Data Mining ith Big Data", IEEE Transaction on knowledge and data engineering, Vol 26, No. 1, January 2014.

[4] Navint, "Why is BIG Data Important?", a Navint Partners White Paper, May 2012 http://www.navint.comimages/Big.Data.pdf

[5] J, Manyika, M. Chui, B. Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, A.H. Byers, "Big Data: The next frontier of innovation, competition and productivity", McKinsey Global Institute, May, 2011 http://www.mckinsey.com/~/media/McKinsey/dotcom/Insight%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI big data full report.ashx

[6] http://en.wikipedia.org/wikiApache_Hadoop

[7] E. Geanina ULARU, F. Camelia PUICAN, A. APOSTU, M. VELICANU, "Perspective on Big Data and Bid Data Analytics", page no.- 9; April, 2012

[8] Prem Jain, Stewart Tate, "Big Data Networked Storage Solution for Hadoop", IBM ,Redpaper, http://www.ibm.com/redbooks

[9] Geert, " Big Data Too Big to Ignore", datacrunches,http://www.people.cs.kuleuven.be~joost.vennekens/DN/bigdata

[10] Hasan Mir, “Addressing Limitations of Distributed File System”

[11] DhrubaBorthaKur, “HDFS Architecture Guide”

a. page no.- 4

[12] R. M. Kharode, A.R. Deshmukh, “International Journal of Advanced Research in Computer Science and Software Engineering”, “Study of Hadoop Distributed File System in Cloud Computing”, Vol. 5, Issue 1, January,2015 http://www.ijarcsse.com

[13] http://www.informationweek.com/software/businessintelligance/sas-gets-hip-tp-hadoop-for-big-dta/240009035?pgno=2

[14] Apache Hadoop, http://Hadoop.apache.org/

[15] Introduction to Cloud Computing by Shang Juh Kao

[16] K. Bakshi, “Considerations for Big Data: Architecture and Approach”, Aerospace Conference IEEE, Big Sky Montana, March 2012

[17] Bernice Purcell, “Th emergence of big data technology and analytics”, journal of Technology Research, http://www.aabri.commanuscript/121219

[18] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, P. Zikopolos, “Understanding Big Data”, “Analytics for Enterprise class Hadoop and Streaming Data”

[19] S. Blazhievsky, N. Systems, “Introduction to Hadoop, Map Reduce and HDFS for Big Data Application, http://www.snia.org/sites/defaulteducation/tutorials/2013/fall/BigData/SergeBazhievsky_Introduction_to_Hadoop_MapReduce_v2.pdf

[20] Rajiv Pandey, Manoj Dhoundiyal, “Quantitative Evaluation of Big Data Categorical Variables Through R”, http://www.sciencedirect.com

[21] ExplainingBigData.https://www.youtube.com/watchv=7D1CQ_LOizA

[22] Five big data challengesAnd how to overcome them with visual analyticshttp://www.sas.com/resources/asset/five-big-data- challenges-article.pdf

[23] Dr. R. Pandey, N. Srivastava, Dr. S. Fatima, Extending R Boxplot analysis to Big Data in Education

[24] J. Nandimath, A. Patil, E. Banerjee, P. Kakade, S. Vaidya, “Big Data Analysis Using Apache Hadoop”

[25] Leskovec, Rajarman and Ullman, “ Mining of Massive Datasets, Map-Reduce”, Stanford University.