Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 195 (2015) 1890 – 1897

Procedia Social and Behavioral Sciences

World Conference on Technology, Innovation and Entrepreneurship

Hadoop Ecosystem and Its Analysis on Tweets

Can Uzunkaya[a], Tolga Ensari[a], Yusuf Kavurucu[b]

[a]*Department of Computer Engineering, Istanbul University, Istanbul, Turkey*
[b]*Department of Computer Engineering, Turkish Naval Academy, Istanbul, Turkey*

Abstract

Hadoop is Java based programming framework for distributed storage and processing of large data sets on commodity hardware. It is developed by Apache Software Foundation as open source framework. Hadoop basically has two main components. First one is Hadoop Distributed File System (HDFS) for distributed storage and second part is MapReduce for distributed processing. HDFS is a file system which builds on the existing file system. It is Java-based sub project of Apache Hadoop. HDFS provides scalable and reliable data storage on commandity hardware. A master/slave architecture is used by HDFS. In this architecture, HDFS has a single NameNode and more than one DataNodes. The NameNode manages the file system and stores the metadata. It acts like a file manager on HDFS. Because all files and directories are represented on the NameNode. DataNodes stores the part of data. A file is splited into one or more blocks (default 64MB or 128MB) and that blocks are stored in DataNodes. MapReduce is a programming model which is used for processing and generating large data sets with a parallel, distributed algorithm on a cluster. A MapReduce job generally splits the input data set into independent blocks which are processed by the map tasks in a completely parallel manner. First step is mapping of data set in MapReduce architecture. The framework sorts the outputs of the mapping process, which are then input to the second step is reduce task. Input and the output of the job are stored in a file-system. The MapReduce framework consists of two process which are JobTracker and TaskTracker. The JobTracker manages the resources that are TaskTracker. The TaskTracker is a processing node in the cluster. It accepts several tasks like map reduce and shuffle from a Job Tracker. Twitter4J is an unofficial Java library for the Twitter application programming interface. It is integrated Java application with the all Twitter services. This paper focuses on Hadoop and its ecosystem and implementation Hadoop based platform for analyzing on collected tweets. The regarding analyzed results are transferred to graphical charts which is showed on a web page.

Peer-review under responsibility of Istanbul Univeristy.

Keywords: Hadoop, MapReduce, Hadoop Distributed File System (HDFS), Twitter4j , Big Data

* Corresponding author: Tel: +90 5353053897
E-mail address: ensari@istanbul.edu.tr

Peer-review under responsibility of Istanbul Univeristy.
doi:10.1016/j.sbspro.2015.06.429

1. Introduction

Data management, processing and storing processes are becoming more difficult with the increased use of digital technology. Because the amount of data increases day by day on the world. This result many company looked for a solution to solve of regarding processes on petabytes of data. The problems are often repeated that the big data problems are that relational databases cannot scale to process the massive volumes of data. The traditional systems are not enough for this solution. In these day Hadoop is often used for data-intensive computing.

Hadoop was created by two Yahoo employee who are Doug Cutting and Mike Cafarella in 2005. It is developed to support distribution for the Nutch search engine project. After the development and dissemination, Hadoop is a registered trademark of the Apache Software Foundation. The Apache Hadoop Java based programming framework that allows for distributed storage and processing of large data sets on commodity hardware. It is designed to reliable and cost effective scale up from single servers to thousands of machines, each offering local computation and storage. The Hadoop Ecosystem has a several Hadoop-related projects. That may use for different purposes. However the main purpose is easy to writing code and simply create a project cost effectively. Some of the Hadoop-related projects are pig, hive, hbase, impala, zookeeper, sqoop, mahout. The Hadoop basically has two main components that are Hadoop Distributed File System (HDFS) and MapReduce programming model.

MapReduce is a programming model which is used for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Generally it follows the divide, process and merge steps. MapReduce programming model works by the processing into two phases which are map and reduce phase. Each phase has key-value pairs as input and output. They are specified by user. The map function that processes a key/value pair to generate a set of intermediate key/value pairs. The reduce function that merges all intermediate values associated with the same intermediate key. The MapReduce framework consists of two process which are JobTracker and TaskTracker. The JobTracker manages the resources that are TaskTracker. The TaskTracker is a processing node in the cluster. MapReduce programs are inherently parallel, thus putting very large-scale data analysis into the hands of anyone with enough machines at their disposal. MapReduce programs are suitable for parallel computing for large-scale data analysis.

Hadoop Distributed File System (HDFS) is a distributed, scalable, and portable file system that runs on large clusters of commodity machines. The HDFS allows users to have a single addressable namespace so it is easier to manage the data and spread across many hundreds or thousands of servers, creating a single large file system. HDFS provides the streaming data access for efficient data process. A master/slave architecture is used by HDFS also it has the concept of a block. In this architecture, a single NameNode and more than one DataNode blocks are used. The NameNode manages the file system and stores the metadata. DataNodes stores the part of data. In twitter analyses application is implemented with MapReduce algorithm. Data are stored in HDFS and the regarding analyzed results are transferred to graphical charts which is showed on a web page.

2. Hadoop and It's Architectural Environments

Hadoop is a Java based programming framework for writing and running distributed applications that process large amounts of data on commodity hardware. Hadoop was created by two Yahoo employees in 2005. Doug Cutting and Mike Cafarella developed to support distribution for the Nutch search engine project. Cutting named this project as Hadoop that means Cutting's son's toy elephant. After development Hadoop is a registered trademark of the Apache Software Foundation. Mainly it has two basis components which are MapReduce programming model and Hadoop Distributed File System. They constitute the Hadoop architecture and the best way for distributed processing of large scale of data. With the scalable ability to a Hadoop cluster can be expanded by adding new servers or resources without having to data lost and additional cost for the transport of data. Hadoop provides the cost effective storage and processing of large volumes of data. Another powerful aspect of Hadoop is fault tolerant. This means secure and nonstop data processing. The data is not store on one node. When the node is lost, the system redirects work to another location of the data and continues processing without missing a beat. On the other hand this works like backup systems.

3. Hadoop Architecture and Design

The Hadoop framework purposes a reliable, scalable and shared storage with distributed computation on large scale data sets. Therefore ıt uses two main component which are MapReduce programming model for distrusted processing and Hadoop Distributed File System (HDFS) for distrusted storage. Both architectures suitable to working with high capacity data sets. Hadoop automatically handles data replication and node failure. Thus it is cheaper than other conventional system for reliable data storage. Hadoop is suitable for bulk process on big data. Because it has layer structure. This architectural does not work well for low-dimensional data. For this reason Hadoop should use on high scale data. Otherwise it extends the processing time for easy analyzing of data. Hadoop has three type of installation modes which are standalone (or local) mode, pseudo-distributed mode and fully distributed mode.

3.1. MapReduce

The MapReduce functionality is designed as a tool for deep data analysis, the transformation of very large data sets and based on the concept of parallel programming with high speed. MapReduce programming model works by the processing into two phases which are map and reduce phase. Each phase has key-value pairs as input and output.
The map and reduce function processing rules are defined by programmer. Hadoop divides the original data segment to process into fixed size pieces called input splits, or just splits. One mapper is defined for one splits by the Hadoop Firstly the key and value pairs are generated by map function. Each key and values are sorted in mappers. The values are merged when the key are same. Then the combiner summations value for all unique key for each distributed mapper. This output represents intermediate key value pair and these values moved to Reduce phase as input. If the node running the map task fails then a new instance of the Mapper will be started on another machine, operating on the same data. The reduce function still does not start until all intermediate data has been transferred and sorted. The reduce function combines the values for a key. Reducer task has 3 primary phases: shuffle, sort and reduce. The only time nodes communicate with other node is at the shuffle step. Shuffling is the process of transferring data from the mappers to the reducers. In the sort phase, the framework groups Reducer inputs by keys in this phase. The shuffle and sort phases occur simultaneously, while mapper outputs are being fetched they are combined. MapReduce phase includes two types of nodes for the control the job execution process. First one is the JobTracker which manages all the jobs on the system by the scheduling task to run on tasktrackers.It determines the execution plan to process on tasktrackers, assigns nodes to different tasks and monitors all tasks. When the task fail, the JobTracker will automatically relaunch the task. There is only one JobTracker exists for a Hadoop cluster and it is run on a servers as a master node of the cluster. Second one is the TaskTrackers. They mean worker or slave nodes which are run task then send progress report to JobTracker. Also it known as DataNode.

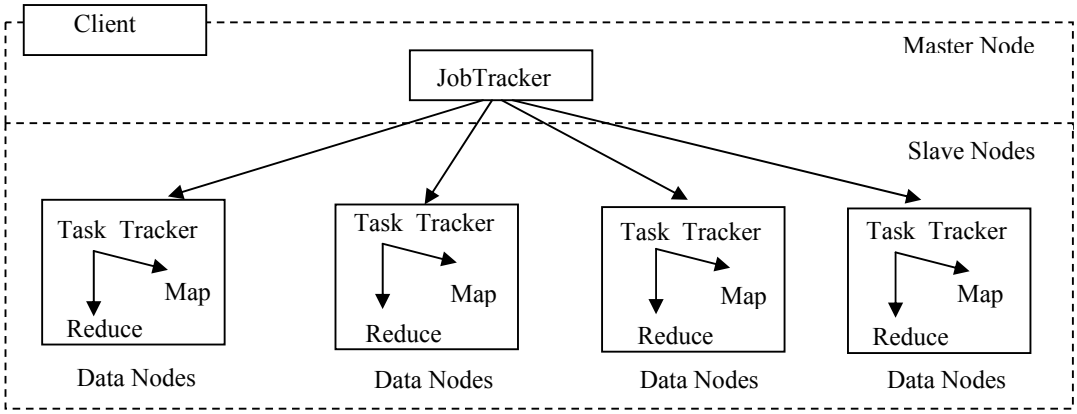

Figure 1. Interaction between Jobtracker and TaskTrackers

3.2. Hadoop Distributed File System(HDFS)

HDFS is a file system which is written in Java and developed by Apache . It is designed for storing very large files with streaming data access on the cluster. HDFS sits on top of a native file system. It splits the all data into blocks and distributes the chunks on community servers. Each block is replicated multiple times and replicas are stored on different nodes. This provides detection of faults and quick, automatic recovery. Hadoop HDFS allows users to have a single addressable namespace so it is easier to manage the data and spread across many hundreds or thousands of servers, creating a single large file system.

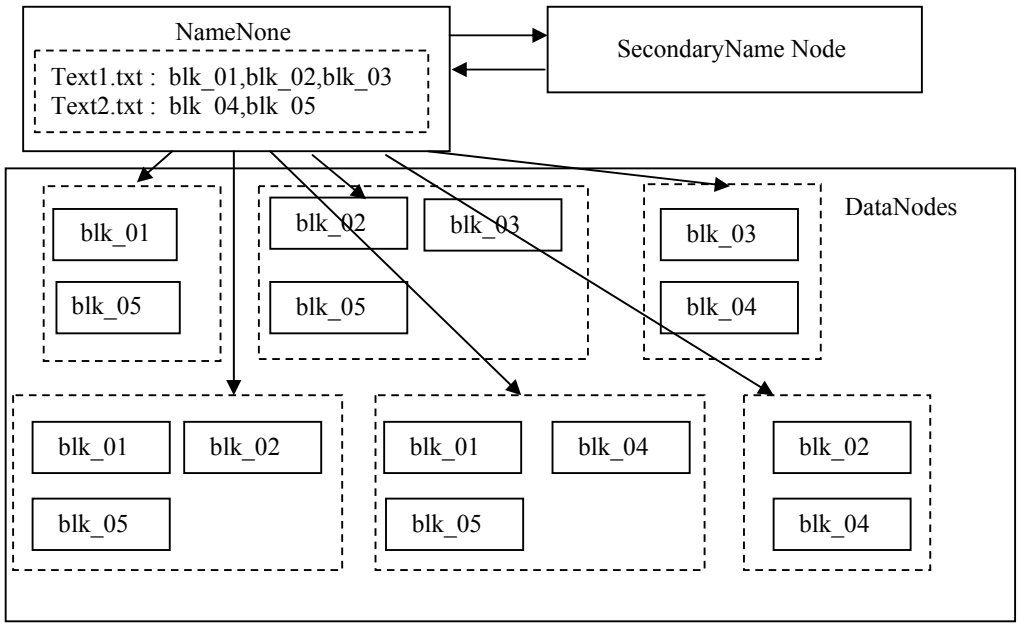

Figure 2: The Relation Between Nodes on HDFS

Hadoop architecture does not support for an append operation and it is based write once read many architecture. A file once created, written and closed need not be changed. This assumption provides the coherency and enables high throughput data access. Applications need streaming access to data. The main purpose of HDFS is batch processing rather than interactive use by users. It focuses on high throughput of data access rather than low latency of data access.

HDFS concept is represented as master slave architecture which has a single NameNode and more than one DataNodes. The NameNode manages the file system and hold all of its metadata in RAM. It acts like a file manager on HDFS. The namenode knows the data nodes on which all the blocks for a given file are located. DataNodes are the worker node of the file system. A file is splited into one or more blocks (default 64MB or 128MB) and that blocks are stored in DataNodes. SecondaryNameNode communicates with the NameNode to take checkpoints of the HDFS metadata at intervals defined by the cluster configuration but it is not a backup of NameNode. When necessary, the check pointed image is read by the primary NameNode. It is usually run on a different server than the primary NameNode. It provides monitoring the state of the cluster HDFS and each cluster has one Secondary Name Node.

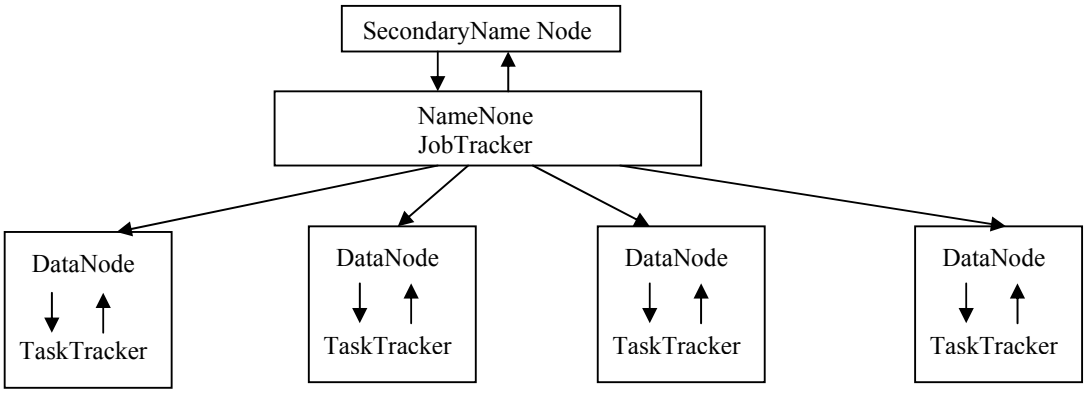

Figure 3: Topology of a Typical Hadoop Cluster

4. The Other Projects in Hadoop Ecosystem

The Hadoop ecosystem bases core components (HDFS and MapReduce) of Hadoop to the extended family of proprietary and open source tools. These projects are not meant to all be used together so needs to be determined well then suitable Hadoop project should selected for our main project. The main purpose of other Hadoop projects, making Hadoop more usable for the need on different project with core functionality and speed in Hadoop. Some Hadoop related projects are described below.

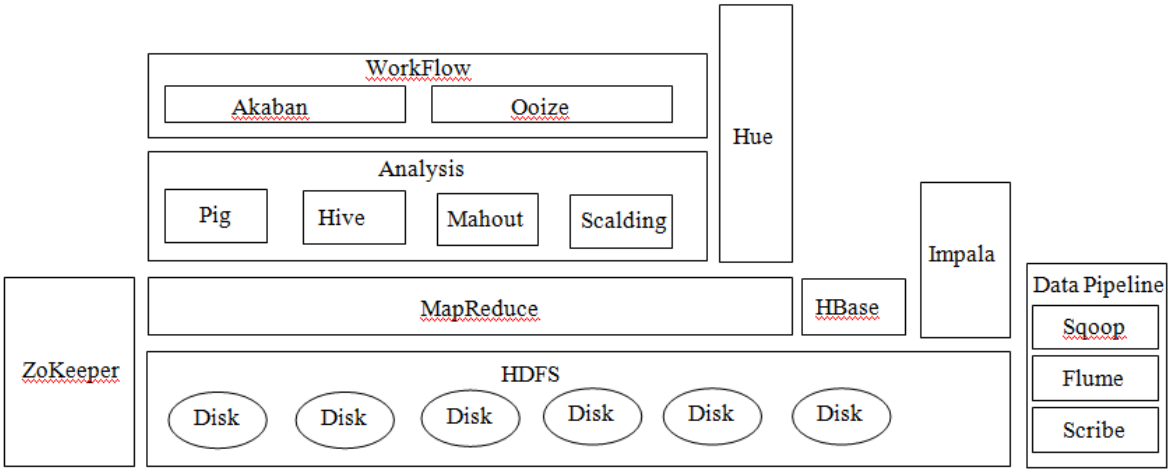

Figure 4: A section of Hadoop Ecosystem

Pig is a platform for analyzing large data sets. It is an alternative abstraction on top of MapReduce programming model. It uses a special data flow scripting languages as Pig Latin. The expressions are used like Sql. It takes the Pig Latin scripts and converts it into MapReduce jobs. Because at the sometimes the writing of MapReduce code can be difficult and it may take time. So it provides a lot of effort optimizing Java MapReduce code and saving time with writing Pig Latin scripts. Pig is not suitable for all data processing or analyzing tasks, it is designed for batch processing of large data sets.

Hive is a data warehousing package built on top of Hadoop and SQL-like access for data in HDFS and other Hadoop input sources. Firstly it is developed by Facebook to processing large amount of user and log data then it is a Hadoop subproject with many contributors. It uses the query language HiveQL which is the similar to Sql and it supports SELECT, JOIN, GROUP BY etc. Through this project HiveQL queries are turned into MapReduce jobs.

Mahout is a project of the Apache to build a scalable machine learning Java library. It bases the Hadoop techniques for many of the implementations. Mahout has several algorithms which are recommendations, collective filtering classification, categorization, clustering, frequent itemset mining, parallel frequent pattern mining. Mahout library can be executed in a distributed fashion and have been written to be executable in MapReduce programming language.

HBase is scalable, distributed and column oriented an open source non relational database built on top of HDFS. It is developed by Apache for real time read and write random access to very large datasets. Hbase is significant for an enterprise data hub(EDH) and its design caters to applications that require fast, random access to significant data sets. HBase provides full consistency, high availability, scale out architecture, transparently and efficiently scale out data across machines in the cluster.

Table 1: Comparison of HBase and Traditional RDBMSs

Items	RDBMS	HBase
Data Layout	Row Oriented	Column Oriented
Transactions	Yes	Single Row Only
Query Language	SQL	Get/put/scan
Security	Authentication/Authorization	Kerberos
Indexes	On arbitary columns	Row-Key Only
Max Data Size	TBs	PB+
Read/Write Throughput Limits	1000s queries/seconds	Millions of queries/second

Sqoop is an open source Apache project and it is designed to transfer data between Apache Hadoop and structured datastores such as relational databases. Sqoop can import data from tables in a relational database to Hdfs can also be the opposite as export data from Hdfs to relational databases. Its imports operations can be used to populate tables in Hive or HBase. Data transfer takes place in parallel for fast performance and optimal system utilization.

Flume is developed by Cloudera and it supports distributed, highly reliable, configurable streaming data collection and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System. Particularly, Flume provides the collection high-volume log files from real time systems such as Web servers on cluster in real time for analysis.

5. Analysis of Tweets with Hadoop

Hadoop and its programming model MapReduce are great for batch oriented processing of large amounts of data sets. This data set can be collected from different locations, log, structured or unstructured data. At this point here can be considered; what should be my data set or how do we want to get a result from data. After you found the answer to these questions then can design the architecture for your need.

WeI want to create our data set by tweets. This data set can be low amount for Hadoop cluster. So we installed the Hadoop as standalone mode on Virtual Machine. In this study, the Twitter4j API is used to collecting tweets. Twitter4J is an unofficial Java library for the Twitter application programming interface. It is integrated Java application with the all Twitter services. We collected the tweets by using the searching word. The thread mechanism is used to collect sequential data from Twitter. For all thread cycle, these tweets are stored in local disk. Then a java job gets the file and sends the HDFS. All flows are shown as follows.

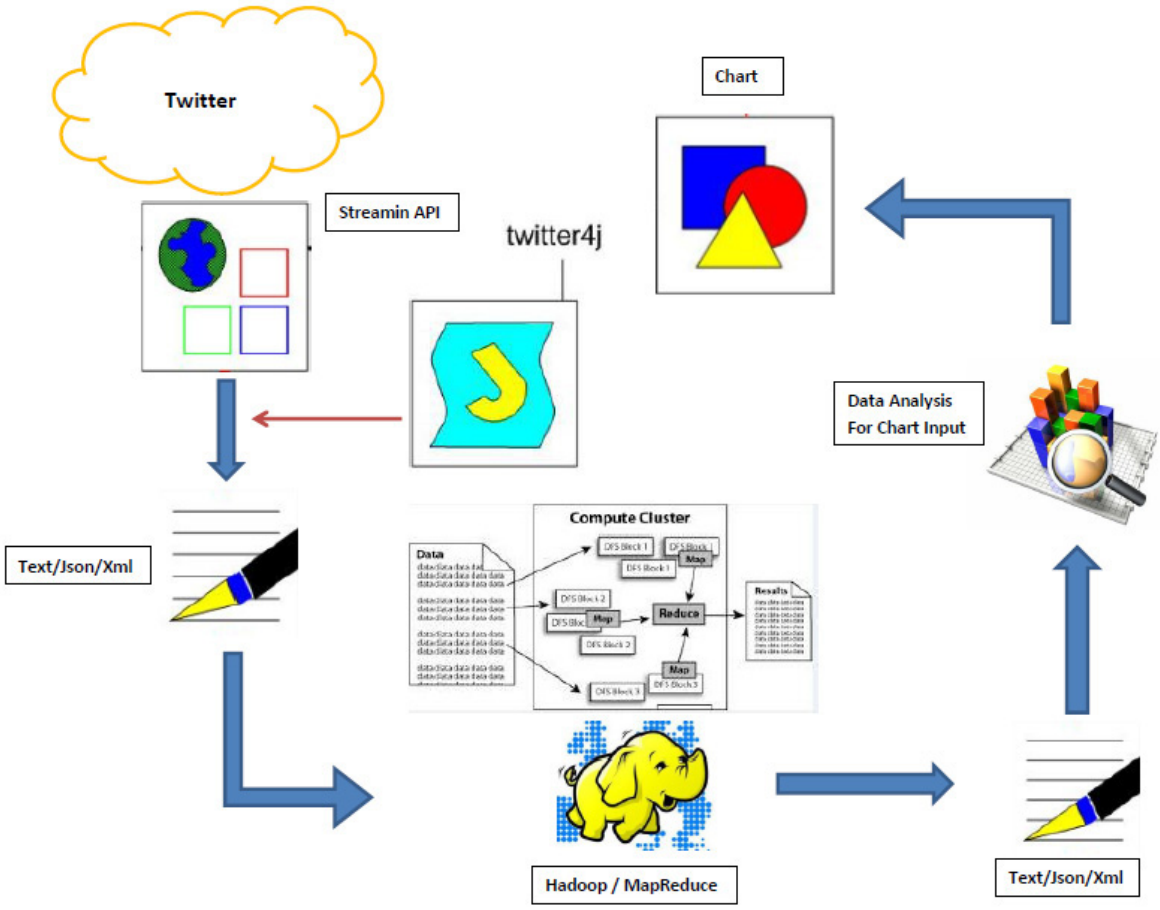

Figure 5: The Project Life Cycle

On the HDFS, The data is analyzed by written Map and Reduce algorithms. Then the result transferred from HDFS to Local disk. On the results data, some analyze can be done. Then the results shown on the charts from web page. This cycle takes place once every 15 min. The graph is updated depending on the analyze results for every cycle.

6. Conclusion

Hadoop is a framework for distributed storage and processing of high amount data sets on commodity hardware. Hadoop basically has two main components; Hadoop Distributed File System (HDFS) for distributed storage and MapReduce for distributed processing. Data processing speed and storage problems arise with emerging technologies so recently demand increased for big data solutions. The Hadoop and then projects on the site solutions for big data provide an open source, end-to-end scalable infrastructure. It is much faster and cheaper than traditional systems.

References

Big Data Analysis, http://thinkbiganalytics.com/leading_big_data_technologies/hadoop
Computer Networks, Tanenbaum, Wetherall , Fifth Edition
datasets, Procedia Computer Science 12, pp. 254-258.

E. Feller L. Ramakrishnan, C. Morin (2015), Performance and Energy Efficiency of big data applications in cloud environments: A Hadoop case study, Journal of Parallel and Distributed Computing,
E. Sivaraman and R. Manickachezian, (2014), High performance and fault tolerant distributed file system for big data storage and
G. Mataraci, (2013), Buyuk Veri, www.ereteam.com/buyuk_veri.asp
H. Ilter, (2013), Hadoop, HDFS ,MapReduce, www.devveri.com/hadoop-nedir
Hadoop Computing Solutions, www-01.ibm.com/software/data/infosphere/hadoop
Hadoop Core, www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html
Hadoop Ekosistemi, http://hortonworks.com/blog/category/industry-happenings/hadoop-ecosystem
Hadoop Fundamentals, http://en.wikipedia.org/wiki/Apache_Hadoop
Lam C., Hadoop in Action, Stamford,First Edition, Manning, 2011.
M. Edwards, A. Rambani, Y. Zhu, M. Musavi, (2012), Design of Hadoop-based framework for analytics of large synchrophasor
MapReduce Model, www.mapr.com/products/apache-hadoop
processing using hadoop, In Proceeding of International Conference on Intelligent Computing Applications, pp. 32-36.
Sammer E., Hadoop Operations, First Edition, 2012.
Venner J, Pro Hadoop, USA, First Edition, Apress, 2009.
What is Apache Hadoop?, www.hadoop.apache.org/
White T., Hadoop: The Definitive Guide,Third Edition, 2012.
X. Hou, A. Kumar, V. Varadharajan, (2014), Dynamic workload balancing for hadoop MapReduce, Proceeding of International Conference on Big data and Cloud Computing, pp. 56-62.
Yasser Ganjisaffar, Presentation, www.ics.uci.edu/~yganjisa