Big Data Processing and Visualization in the Context of Unstructured data set

A Thesis Submitted to School of Information Science

By: Temesgen Desalegn

Advisor: Million Meshesha (Ph.D.)

7/27/2016

DECLARATION

I, the undersigned, certify that this research is my original work and does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Name: Temesgen Desalegn

Signature: _____

This thesis has been submitted for examination with my approval as university advisor.

Advisor: Million Meshesha (Ph.D.)

Signature: _____

ADDIS ABABA UNIVERISTY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

Big Data Processing and Visualization in the Context of Unstructured data set

A Thesis submitted to the School of Graduates Studies of Addis Ababa University in Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Science

By Temesgen Desalegn

ADDIS ABABA UNIVERISTY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

Big Data Processing and Visualization in the Context of Unstructured data set

By: Temesgen Desalegn

Advisor: Million Meshesha (Ph.D.)

APPROVED BY

EXAMINING BOARD:

1. **Dr. Million Meshesha, Advisor** _____
2. **Dr. Dereje Teferi, Examiner** _____
3. **Dr. Wondwossen Mulugeta, Examiner** _____

To my beloved families!

Acknowledgements

"Commit to the **LORD** whatever you do, and **HE** will establish your plans." – Proverbs 16:3

First of all, I would like to express my deeper gratitude to my advisor, Dr. Million for his unreserved advice and humble supports from selection of research area throughout all activities of the thesis. Next, I would like to say thank you my colleagues and friends who were participating directly and indirectly in the path of the study by providing valuable ideas and suggestions. In particular, I want to appreciate Ato. Getinet Tibebu who had been encouraging and sharing burdens and thoughts.

Table of Contents

List of Figures

List of Tables

Acronyms

3Vs – Volume, Velocity and Variety

ACID – Atomicity, Consistency, Integrity and Durability

API – Application Programming Interface

BI – Business Intelligence

CAP – Consistency, Availability and Partition

DISC – Data Intensive Scalable Computing

DML – Data Modification Language

DRIP – Data Rich Information Poor

HDFS – Hadoop Distributed File System

HiPPO – Highest Paid Person's Opinion

HQL – Hive Query Language

IoT – Internet of Things

MPP – Massively Parallel Processing

NoSQL – Not only Structured Query Language

NSF – US National Science Foundation

RDBMS – Relational Database Management System

SSBI – Self-Service Business Intelligence

Abstract

Today, it is not uncommon to face data deluge that has brought challenges to every sector across all industries. The rate of data growth is exceeding currently available storage capacity as a result of data creation by everything which is connected to internet; in addition to human activities over cyberspace, Internet of Things (IoT) are playing crucial role in business activities by generating highly valuable information and insights that cannot be tapped otherwise. On the other hand, social networks has brought a platform that facilitates human interaction among themselves which is creating a room to everyone to produce huge data sets using computers and smart phones as well. Moreover, data creation rate in variety of formats is yielding real challenges to traditional technologies. Big Data processing and visualization is current challenge due to data growth with high velocity in variety of data type.

To tackle Big Data problems, the methodology applied is in detail investigation of current challenges, identification of technology frameworks and ecosystems, design solutions, implementation of the designed solution and test of implemented solution using Big Data set is taken place. Hadoop ecosystem which is starting point of technological shift from traditional technologies to more advanced and different has shown the change of data and technology landscape.

The result of experimentation has revealed that Big Data processing and visualization requires comprehensive framework and collaborative ecosystem. In addition, change of model of data storage and processing is changed to send process where data resides rather than bring data to process. Huge and complex data sets visualization is not possible to realize using accustomed set of technologies.

Key words: Big Data, Hadoop, MapReduce, NameNode, DataNode, JobTracker, TaskTracker, Hadoop Distributed File System, Visualization

Chapter One

Introduction

1.1. Background

Nowadays, more devices are coming to cyberspace with a lot of functionalities that provide services at different level, for instance, individual, group and community. Now, people are at a verge of simplifying life questions which could be expressed in terms of space and time. Interactions of Internet of Things (IoT) and people are generating data that cannot be left alone due to its value. World data growth in the years (2011 and 2012) was 90 percent of total human history of data creation [1]. Current infrastructure and applications are allowing human kind freedom for communication and doing activities in the format of digital data which was inconceivable some years ago. These all-encompassing facilities and capabilities are pouring data from different sources and directions to global data storage which is accumulated to about 1,800 EB (Exabyte) or 1.8 ZB (Zettabytes) [2]. The continuation of data accumulation, which is expected to be 50 times in 2020, at an alarming rate within a variety of formats makes it difficult for current practice of management of data.

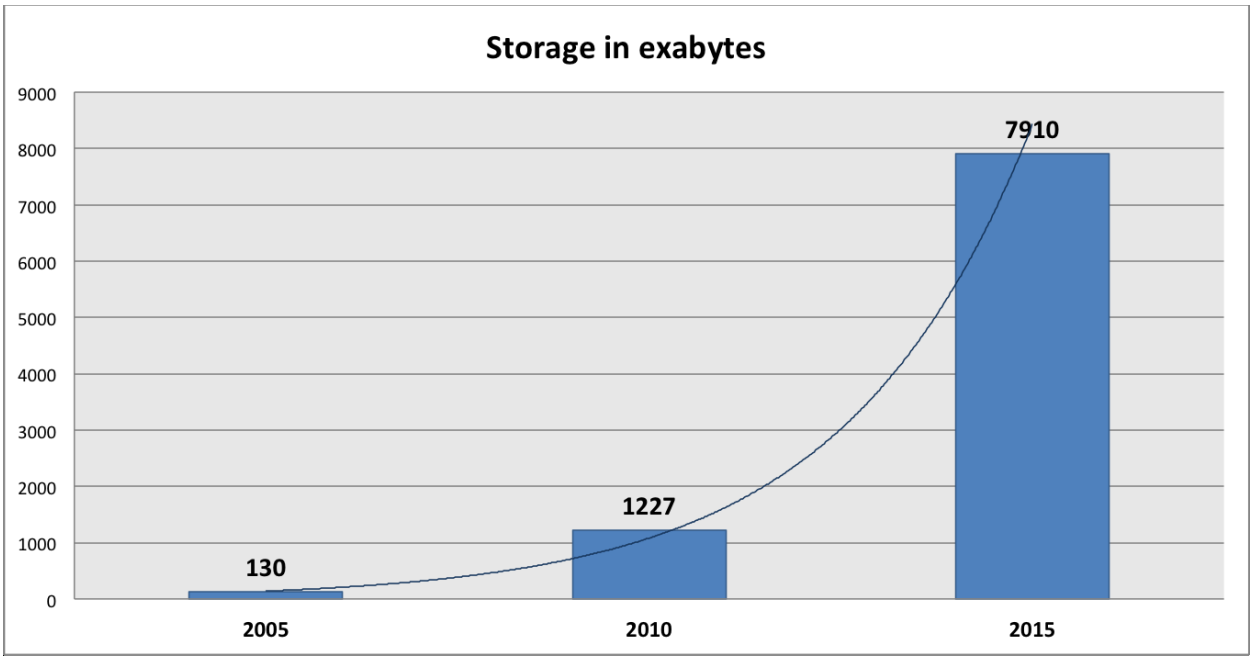


Fig. 1.1: Data growth over time [3]

As shown in Fig. 1.1, Big Data refers to the explosion of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, data is accruing at the rate of several gigabytes per day [4].

Volume is the key issue that is bringing challenge to current art of state of technologies with regard to storage capacities and accessibilities. It is critical for business organizations as well as scientific communities to get full picture of environments surrounding them in order to act or react in the way that enhances their outcomes. Highly competitive organizations are looking for more data to take advantage of competitive edge over their competitors. More importantly, the output of data driven decisions are by far greater than decisions that are dependent on intuitions or guts of individuals who is also known as Highest Paid Person's Opinion (HiPPO) in an organization. In addition, scientific researches become more dependent on accumulated data ever.

Creating data is now on finger print of everyone within few seconds [5]. Mobile devices, social medias, sensor embedded devices and others are the major means to create data instantaneously. Rapid velocity is creating a room for transient data which is short lived by nature. However, it should be channeled to drop its value to organization bed rock data accumulations. Data in motion gives insight for decision as well as actions.

Period of structured data repositories and management have been coming to end as technologies are advancing from time to time. In this challenging time for organizations, modality of data creation is shifted from employees to users or consumers. Practically, consumers or users are not willing to enter to limited set of columns or fields of data entry. Variety of data such as structured, semi structured and unstructured is streaming of data to organization so that organizations get opportunities to act upon it so as to extract insight and take decisions based on extracted insight.

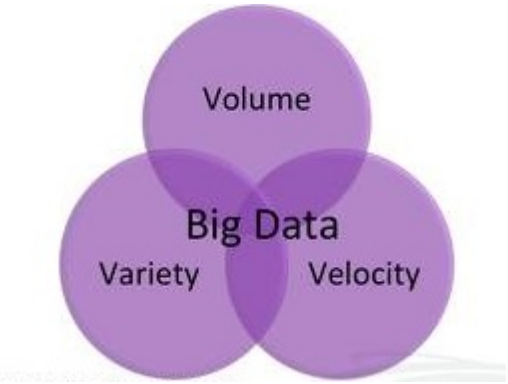


Fig 1.2: 3Vs of Big Data [6]

As shown in Fig. 1.2, combination of Volume, Velocity and Variety (3Vs) [7] are current challenges of every industry due to advancement or new model of communication and transaction. Globalization is pushing cross-cultural unification to modernize of societies those who were isolated or lagging behind for centuries from technological progress in developed countries. As a

result, peoples have started to consume and produce data in huge, with high frequency and in different format.

Extraction of values from big data is a challenging task that needs a means of processing and visualization so as to get in depth insight. The challenges that are creating and retrieving data especially unstructured data set is imposing to look for new way of processing and visualization. Furthermore, data growth becomes indistinguishable phenomena that initiates swift action to tap its benefits.

1.2. Statement of the Problem

Our planet is becoming host of data creation and repository at alarming rate. Everyone in everywhere is generating data from their day to day activities, communications, transactions, and so on. The main sources of data are human-generated digital footprints which comprises 70%, machine generated data and sensory data. Nowadays, organizations are overwhelmed by external data sources on the top of their internal data source which is called dark data. The challenges that organizations are facing is multifaceted such as the availability of big storage space (volume), the speed at which data creations takes place (velocity) and diversification of data types (variety). Tsunami of data from smarter planet whereby Internet of Things and people are introducing values extractions from big data [8].

In general, big data is immersed with wealth of new insights across all industries, expertise and life to provide and guide discoveries and innovations. Smart devices are the main actors in creations and utilization of big data that shifts tradition practices into modern life style and even research direction is reversed from ‘theory to data’ to ‘data to theory’ paradigm. It is estimated that total population and total mobile phones are approximately 6.8 billion and 6 billion

respectively [9]. Moreover, mobile applications have changed the way people think, live and transact in smart spaces and time.

However, traditional tools and techniques have no capabilities to serve big data in the way that accommodates three Vs (volume, velocity, variety). Data Rich Information Poor (DRIP) is a scenario where there are vast data but inadequate useful information for a given purpose. This is because of limitation of RDBMS, data warehouse and data analysis tools.

Even though there are researches that have been conducted on area of Big Data in general, there is no clearly solidified or identified agreement what Big Data is all about. Some of them are discussing traditional data mining under umbrella of Big Data; and others are dealing with samples and statistics to explain or explore Big Data. Actually there are some handful researches so far encountered to spot what is done in the area. As described in [10], Big Data components, challenges and opportunities is discussed in terms of seven dimensions, historical background, what's big data?, data collection, data analysis, data visualization, impact, human capital, and infrastructure & solutions. It indicates that Big Data and Analytics require all the above dimensions in today's business environment. Large scale web mining by utilizing Data Intensive Scalable Computing (DISC) system to extract information and models from web data necessitates traditional algorithms by putting power of parallelism [11]. DISC system is considered as one of powerful, fault tolerant and inexpensive to process large data sets. Shown in [12], design of conceptual big data adoption model within organizations by employing business case development, technical, organizational, and information privacy related processes. As indicated in [3], real time big data processing using Storm system is used instead of MapReduce which is appropriate for batch processing. Storm is distributed and fault tolerance system which achieves processing in collaboration with other tools such as Cassandra, Redis and Kafka over NoSQL.

Experimental researches on Big Data architecture, data processing using DISC System, real time data processing using Storm and conceptual framework have been studied in arena of Big Data; however, no research is found that shows experimental study on actual data set for Big Data processing and visualization using emerging Hadoop ecosystem to date. So in this research an attempt is made, to investigate and identify techniques and tools to process Big Data to extract values and visualize to audiences. The 1V (Volume) is experimented using Hadoop ecosystem in this study; however, remaining 2Vs (Variety and Velocity) requires further investigation.

1.3. Research Questions

In conducting this study, the following research questions are explored and addressed:

- What are the challenges and opportunities of big data?
- To what extent the current available toolset enables to process and visualize unstructured data sets of big data?
- How does data growth be handled?

1.4. Objective of the Study

The purpose of this research is to identify and investigate the means of big data processing and visualization using Hadoop ecosystem.

1.4.1. General Objective

The general objective of this study is to process and visualize unstructured data sets of Big Data so as to analyze execution time, memory requirement and fault tolerance.

1.4.2. Specific Objectives

The specific objectives of this study to fulfill general objective are:

- To identify big data technologies landscape

- To experiment unstructured data sets of Big Data using Hadoop ecosystem
- To explore means and impacts of big data visualization
- To recommend further study in the area

1.5. Scope and Limitation of the Study

The scope of this research is specifically conducting big data value extraction for Volume using available technology stacks that facilitates knowing the landscape of data processing technologies. At the same time, utilization of available technologies in an arena of data science might lead to in detail investigation of created capabilities to handle current problems.

The limitation of the study is that it does not encompass the other 2Vs; for instance Variety and Velocity are not experimented. In addition, large data size such as Terabytes and Petabytes are not experimented in this study. These factors need well equipped labs with clusters of machines in order to conduct complete experimentation on top of adequate timespan.

In this study, we are not dealing with clusters of machines to process Big Data in its full scale which requires high fund and collaboration of experts. While processing huge data sets, there are violations of privacy of individuals; and privacy is not considered under this study. In addition, security is becoming central focus of discussion and real challenges of everyone; nevertheless, we are not encompassing it in this study.

1.6. Methodology of the study

1.6.1. Literature review

Extensive literature review is conducted from books, journals, conference procedures and the internet in order to gain deeper understanding of Big Data landscape and its value proposition to

society at different level. It gives the spot of current data challenges as well as its application in wide areas.

1.6.2. Data sources

There is organization that provides free datasets or eBooks in different file formats such as text, pdf, epub e.t.c.: http://www.gutenberg.org

From the above source, philosophy category of five hundred books are downloaded which is text format data set that is used for experimentation as well as testing for big data technology stack implementation. Actually, the link has vast number, more than 50,000, of free eBooks with different categories which are easily and freely accessible but some other companies provide links as open data sources but they may require payment or processing in their custody and then they charge service fee. In addition, there are also claiming that they provide public open data set by requiring registration as citizens of a country.

1.6.3. Development and Processing Tools

Open source framework, Apache Hadoop Framework, is mostly customized for academic purpose and it is Hadoop Distributed File System (HDFS) which alleviates current processors limitation that processing capacity is at its ceiling point [13].

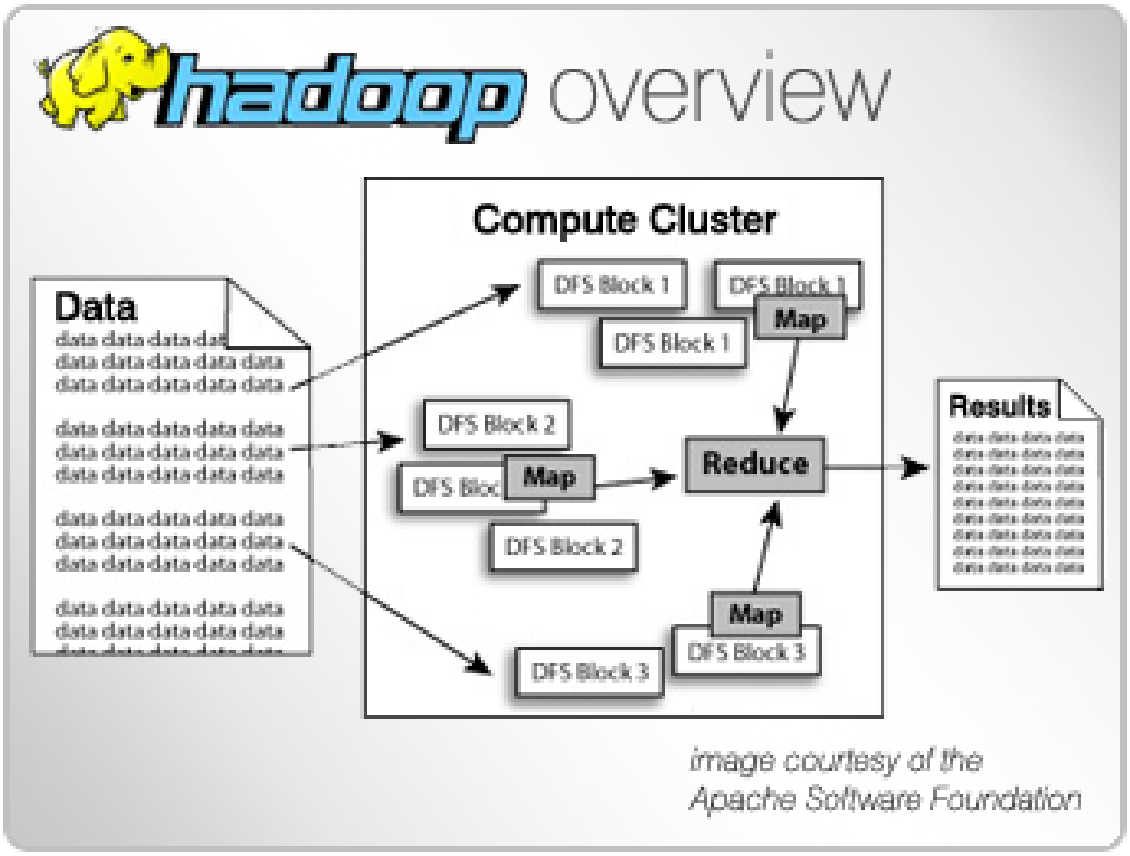


Fig. 1.3: Apache Hadoop Framework [14]

As shown in Fig. 1.3, Apache Hadoop Framework provides a platform so that data sets passes a number of phases from input or raw data stage to output or result stage.

Hadoop is open source platform which is used for data storage and processing of very large volumes of data at high speed with low costs. It is possible to build large scale distributed data processing system using commodity computers that lowers cost of computation. It is also possible to run Hadoop on single desktop or laptop for testing [15].

1.6.4. Visualization Tools

After processing data sets, the next step is converting the output or result of processing as input into visualization tools. Visualization is key companion of big data processing so that the result of

huge data set processing can easily be grasped by experts as well as others. Without using visualization tools, it requires a lot of effort and time to comprehend output of big data processing. There are few big data visualization tools that are powerful and with a capability of accommodating vast data elements within a single screen [16]. In fact, data sets of population, big data, cannot be easy task to present using customary visualization tools as its fitness for sample data set in case of limited variables. Tableau, visualization tool, is marketing leading and its application in wide industries including research makes it as a tool of choice to visualize in this study.

1.6.5. Evaluation Procedure

Studies in this area are at their infant stage; so, it is difficult to find common or conventional evaluation procedures that can be utilized for this study. So, results of the study are evaluated with major three dimensions; namely execution time, memory requirement and presentation. Firstly, the output of experimentation is evaluated in terms of time of execution that has taken to ingest, process and yield result. Secondly, memory utilization to process from client command to output file. Lastly, the output of experiment is required to fit for consumption by audiences so its easiness for presentation is taken in account. The result of evaluation is expected to show better parameters value in comparison with traditional data warehouse.

1.7. Significance of the Study

The output of the study benefits research communities as well as others such business, government, scientific communities. The global scenario is now utilizing big data as source of insight, bases of decisions and development; for instance, U.S. government is taking wide initiatives in big data projects as priority. In perusing innovations and values propositions within data as commodity, we

have opportunities and at the same time burden to make and bring all necessary resources to tackle current practice of data management so as to avoid falling behind in battles of globalization if we are not alert enough to go head to head. Moreover, the impact of this study is strong enough to show and motivate the immerse of values that are related to data creation, storage, processing and utilization by individuals, groups, organizations and government bodies who are interested in getting values of data for their decisions. As a matter of fact data silos, data errors and data governance are the main obstacles to provide timely analysis and decision for mega projects in developing countries. So, this study tries to give clue on value and commoditization of data.

1.8. Organization of the Thesis

The thesis is organized as follows:

- Chapter One discusses Background; Statement of the problem; Research questions; Objective of the study: General objective and Specific objective; Scope and Limitation of the study; Methodology of the study: Literature review, Data sources, Development tools Visualization Tools and Evaluation procedure; Significance of the study; and Organization of the thesis.
- Chapter Two discusses literature review particularly Big Data and its challenges; Tools and Framework: Hadoop, Hadoop Distributed File System, MapReduce, Data Processing; Data Visualization; and Related Work.
- Chapter Three discusses Data Collection and Design particularly Data Collection: Data Type/Nature, Data size and Data Sources; Planning of Technology stacks; Architecture of the system; Design: Approaches and Techniques, Design goals, Data analytics design and Data visualization design; and Implementation particularly Algorithms: Mapper algorithm and Reducer algorithm; and Visual components.

- Chapter Four discusses Experimentation and Results particularly Experimentation; and Results: Data processing and Data visualization.
- Chapter Five discusses Conclusion and Recommendation
- Appendices

Chapter Two

Literature Review

Companies are now overwhelmed by the blast of data flows from sensors, radio frequency identification or other devices which is either voluminous or unstructured to be processed using traditional practices. This real time data or information is consumed for new product development, service enhancements or ways to respond to changes in the environment [17]. Values embedded in stream of structured and unstructured data that answers most of questions that could not be raised by business yet due to technological limitations. Only 5% of data available in organization is utilized but 95% values of data have option to be tapped as big data technology advances. Anything that goes through digitalization speaks about who is utilizing it, how it is utilized further why it is utilized. Actually, BI has no capabilities to process data which is diverse, more granular, real time and iterative, demands organization to get in depth information from specific moment in time before changes happen. The old thinking is “too much data is bad thing” is reversed nowadays to “more is better” [18].

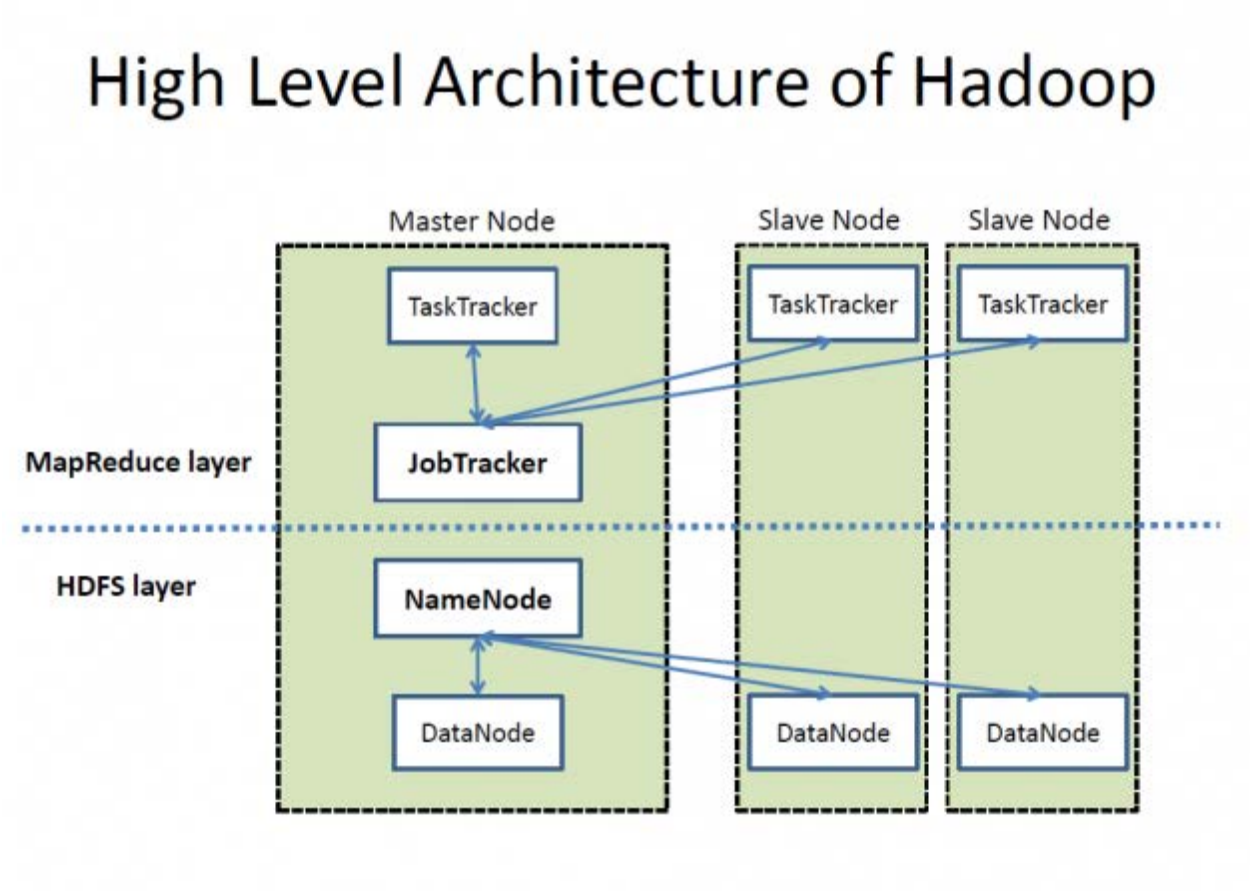


Fig. 2.1: High Level Hadoop Architecture [19]

Files are distributed and taken in Hadoop HDFS to store across all computers in the Hadoop cluster(s). A file will be chopped into smaller blocks with size greater than or equal 64MB and distributed over other nodes in order to assure replications and fault tolerance. For instance, whenever one or more node(s) fail(s), the chunk of a file or data in failed node(s) will be replicated to other nodes. So there will never be data loss. As shown in Fig. 2.1, Hadoop HDFS comprises NameNode as master node, secondary NameNode as checkpoint and DataNode as slave node which stores actual data [15].

2.1. Big data and Its Challenges

Streaming of data from different sources is accumulating large volume of data with variety of forms at high speed, velocity. Big data is different from “lotsa data” or “massive data” in the way that it should incorporate all Vs (volume, variety and velocity) in order to be treated as big data. It is not analyzed in its totality so it is required to pass through multiple steps like data extraction, data filtration, data transformation and data analysis. The difference between small and big data is goal, location, data structure and content, data preparation, longevity, measurement, reproducibility, stakes, introspection and analysis. These dimensions help to distinguish big data from small data [20].

Sources of data are Call logs, mobile-banking transactions, online user generated content such as blog posts and tweets, online searches, satellite images and so on. Insight from big data narrows gap between information and time. Big data is industrial revolution of data. Data is taken as raw good with little “intent and capacity” [21]. However, big data is not without challenges as it is under development stage. It has data challenges (Volume, Velocity, Variety), process challenge (display complex analytics on mobile devices) and management challenges (data privacy, security, governance, ethics) [9]. On the other hand, benefits of big data which are by far overwhelming such as its application to medicine (e.g. flu trend analysis by Google), climate change, food safety, science, business, technology, manufacturing, financial markets, cyber security, etc. There are a number of giant companies like E-bay, Facebook, Wal-mart, Yahoo, Google … who are implementing as well as enhancing big data technologies; additionally they have started selling big data as services for small and medium sized companies [22].

Business firms that use big data are apart from traditional analytics implemented in a way that by focusing on data flow, depending on data scientists and process and product developers rather than

data analysts, and analytics is become part of core business. Commonly used tools like MPP Databases, Apache Hadoop Framework or Internet and Storage System provides capabilities to load, store and query large datasets in near real time. Moreover, it executes advanced analytics which is developed under information ecosystem. Big data analytics tools are considered as next generation of IT processes and systems that is designed for insight but not just for automation [17].

In most organizations, information is taken as boarding spring of successful business activities and similarly they give full attention for every drop of information by considering it as life blood of business activities as product or service to be sold to customers. So they are exercising information management practices as a means of managing available information for innovation and decision making. However, current trend of data overwhelm from inside as well as outside of organization is creating burden on processing capability of data as usual. The era of big data puts thinking of data storage rather than value creation from stored datasets to deliver innovative solutions and at the same time coping changing environment in the way that enhances organization's competitive position in an industry. Even to the existent, some organizations are focusing their effort only on lighting of operation of current activities rather than enabling business as well as differentiating their services or products. Big data technologies, especially open source framework such as Apache Hadoop Framework, are creating conducive environment for processing large volume of data with low cost and high speed and it has capability to scale out as capacity of the organization grows to accommodate more data sources. It scales horizontally without need of rework in order to scale up processing capability of already in placed system [23].

In near real time, data is required to be processed to extract insight and achieve tangible time value of data in transit. More data does not allow us to see more rather it allows us to see new, better

and different. Big data as resource or tool helps to advance society; moreover, it supports to address recurring global challenges such as energy, environment, drought, poverty and so on [23].

The benefit of big data is awesome in all paths of life despite the obstacles and the risks, the potential value of Big Data is inestimable ... NSF aims to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to:

- accelerate the progress of scientific discovery and innovation
- lead to new fields of inquiry that would not otherwise be possible
- encourage the development of new data analytic tools and algorithms
- facilitate scalable, accessible, and sustainable data infrastructure
- increase understanding of human and social processes and interactions
- promote economic growth and improved health and quality of life

The new knowledge, tools, practices, and infrastructures produced will enable breakthrough discoveries and innovation in science, engineering, medicine, commerce, education, and national security. However, finding and using standards and measurement for big data analysis is limited due to its infant stage; so big data analysts especially data scientists are developing a set of strategies as well as tools to align data with meaning and reality. On the other hand, in small data defining “control” is common practice; practically, groups are divided into control and test groups. But defining “control” group in big data is impractical because data analysts have no controls over big data. In addition, experiment results are difficult to repeat with given population [20].

Testing hypothesis using big data resources could lead to false confirmation; so, forcing big data to answer specific question is act of self-deceive which might produce wrong conclusion.

Moreover, retesting results pass through hectic and long path that require a lot of resources and time. Finally, confirmation of the result could not be as expected due to a number of factors. As matter of fact, big data projects are done or processed without help of statistical or analytical software packages. On the contrary, human beings are better in processing large information, organizing and visualizing it as appropriate. For instance, "… we humans have a long-term memory capacity in the petabyte range and that we process many thousands of thoughts each day. In addition, we obtain new information continuously and rapidly in many formats (visual, auditory, olfactory, proprioceptive, and gustatory)." [20].

Smart machines are nowadays inseparable human companions in every aspect of endeavor which ranges from dressing to spacecraft. As Internet of Things, smart homes, smart cities are getting broad bases and controlling business sectors and research centers. They are becoming sources of high volume data at very high speed as result they have surpassed data generation by digital human foot print. This phenomenon is stressing the importance of complete suite of tools to analyze and extract relevant value so as to arrive important decisions. Even though big data is the answer, it is difficult to formulate the questions [24].

2.2. Tools and Framework

2.2.1. Hadoop

Hadoop is a framework that comprised of a number of components for its proper functioning and returning intended results. As shown in Fig. 2.2, major components are NameNode, secondary NameNode, DataNode, JobTracker and TaskTracker. Each of these components has well designed to accomplish certain task in general. NameNode is the master or brain of the whole Hadoop system and its main duties are tracking address of all stored files, listening heartbeat message of all DataNodes, manages schedules of JobTracker, holds information about inter rack status and so on. Secondary NameNode is taken as backup node which takes snapshot of NameNode in order to restore normal functioning after its failure. DataNode is a slave node where the data is deposited and data manipulation takes place before aggregation activities started. JobTracker is the one that orchestrates all tasks to be carried out throughout across task assigned nodes. TaskTracker is a slave by its very nature and its responsibility is carrying out ordered task to be performed at low level which is individual nodes or commodity machines where data is stored [25].

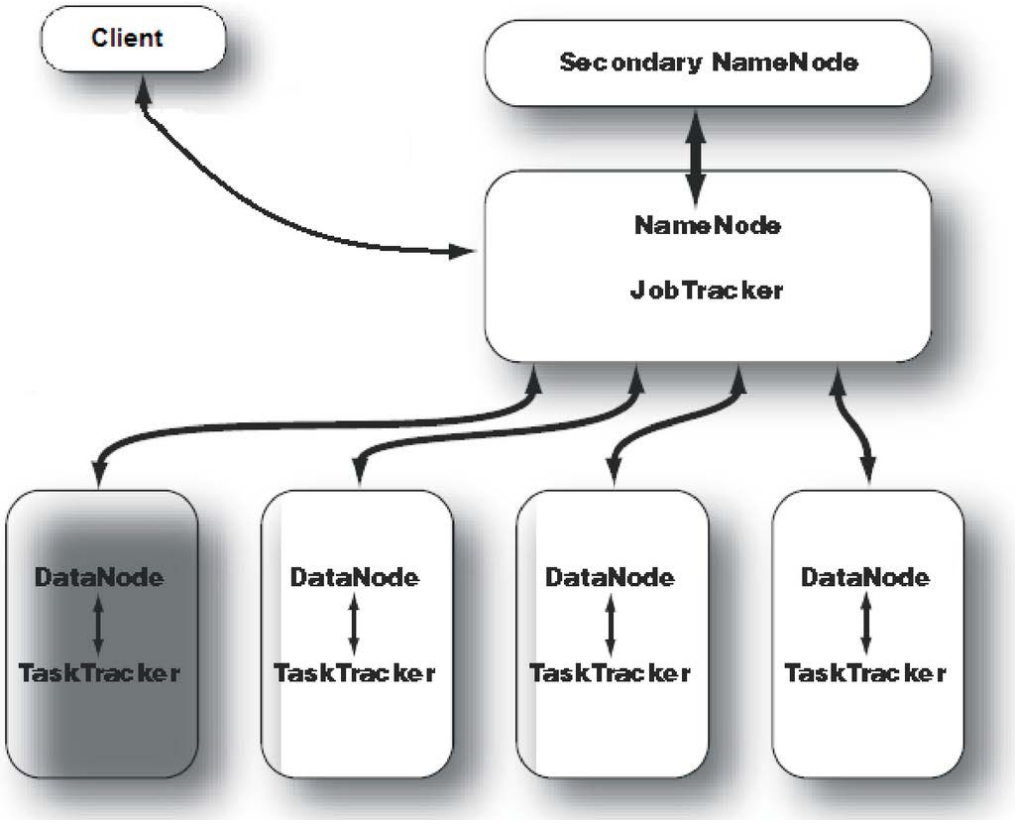


Fig. 2.2: Hadoop Components [26]

Hadoop is also an ecosystem which consists of a set of related projects that are implemented to facilitate customization based on experience and expertise of organizations. The major projects are Hadoop Streaming which enables script writing for those who are familiar on script languages, Hadoop Hive which provides SQL writing capabilities for those who are working with SQL languages, Hadoop Pig which is purely procedural language that supports data pipeline scenarios and Hadoop HBase that stands with real time data retrieval rather than batch processing. On top of these, Hadoop Distributed File System and MapReduce are the major projects that can be taken as backbone of the ecosystem [27].

In general, Hadoop MapReduce architecture provides an environment where parallel processing is done in large set of commodity nodes. Each node is a single unit of machine which executes

assigned task in full responsibilities without depending on other machines for its execution. As mentioned above, Hadoop MapReduce framework is purely software solution for current limitation of space and processing capacity. Instead of putting single machine with vast space and high speed like super computer which is actually very expensive; additionally, it demands top expertise to setup and for ongoing operations as well, there comes very cheap solution that can be implemented with reasonable investment. Return on investment of new big data technologies is amazingly high in terms of insight that may be extracted from processing untapped, unstructured, data set due to traditional technological limitation. From overall dataset 90% of data is unstructured data and it is rich with insights that can shape usual practices of every industry to modern way of accomplishing activities or achieving objectives [28].

The limitations of traditional RDBMS [7] and analysis tools are mainly scalability challenge which means as the size of data increases their retrieving and manipulation do not be scaling up to proportionally. In addition, schema oriented data storage and manipulation has been becoming bottleneck for diversification of data set processing. The foundational building blocks of traditional technologies such as data warehousing, transactional databases, ETLs, business intelligence etc. is directly related with structured data [27]. So, its application for semi-structured and unstructured data would be very laborious. Even though there are a number of attempts to alleviate scalability limitations of these technologies, their ceiling point to embrace changes is not elastic enough [29]. Moreover, ACID (Atomicity, Consistency, Integrity and Durability) [30] property of relational databases is not relaxing to elasticity for growth of data. Transactional nature of relational databases which is all transaction processing shall be committed at once or fail altogether makes it strict rule to be abided [3].

On the other hand, CAP (Consistency, Availability and Partition) [30] theorem has brought a room to achieve two of the three CAP tolerances by taking a context as guiding principle. It is impossible to secure all of the three tolerance variables in distributed computing environment. Especially, when big data is taken as platform to process large dataset at high speed of variety data in distributed setting, it is of sure there is trade off among CAP tolerance variables to properly achieve all of them once. As the theorem indicates, there is always compromise between Consistency and Availability in distributed computing situation. Whenever Consistency is given priority to achieve right response for requesters, Availability is sacrificed in terms of speed of response for requesters. On the contrary, if Availability is given priority in distributed computing environment, Consistency is relaxed to ensure uptime of the system [3].

In a nutshell, Hadoop has changed the landscape of data analytics in the way that eases data processing regardless of structure of data, with high performance and fault tolerant manner.

2.2.2. Hadoop Distributed File System (HDFS)

File storage structure has been changed to maintain distributed file storage along with ensuring fault tolerance. In 2004 [31], Google started to change algorithm in order to boost its search capability by indexing whole files in the internet. As result, it has released white paper on Google File System which was initiated new file system, Hadoop Distributed File System, to be developed by open source community. It is a mechanism to handle large files in distributed manner over multiple of nodes in the form of chunks that each chunk will be replicated as per set replication factor at time of configuration. Whenever there is failure of one or more nodes, data will be moved from failed nodes to active nodes where accommodation space is available. In addition, it creates

an environment where horizontal scaling is easily achieved to scale out to hundreds of thousands of commodity machines [25].

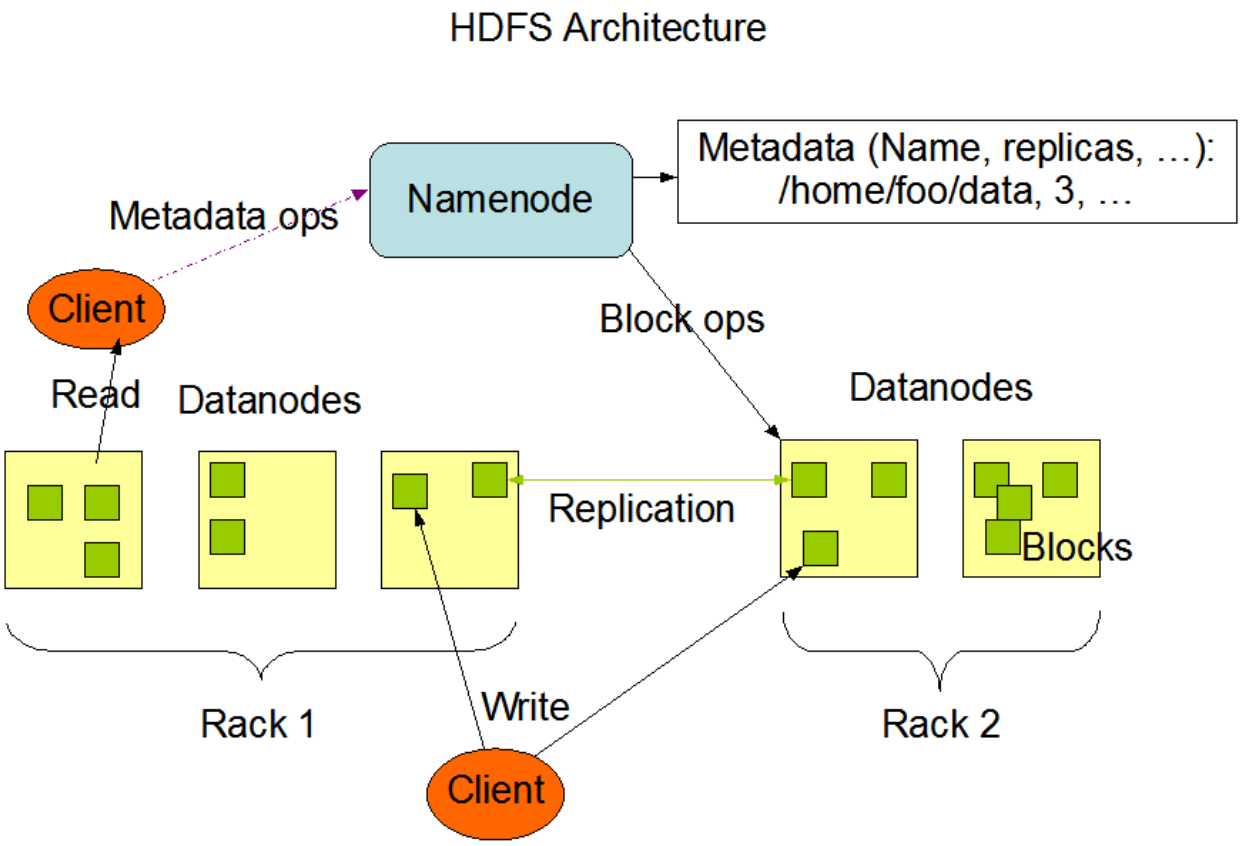


Fig. 2.3: HDFS architectural view [32]

In addition, as shown in Fig. 2.3, HDFS is becoming the center of architectural change for current computational practice by improving performance of latency and throughput. The impact of performance improvement, at level of software (Hadoop Framework) rather than hardware, is attracting giant companies like facebook, google, yahoo etc. so as to adopt the principle and practice as well. It enhances read/write operations of local file chunks by moving computation to where data is stored. It handles very large files which be gigabytes or more by reading or writing

sequentially to/from nodes therefore there is no need to bring data to memory in order to manipulate so the role of primary memory is becoming insignificant [12].

2.2.3. MapReduce

MapReduce is a programing model for processing large scale datasets in a single pass in clusters of thousands of nodes by assuring fault tolerance and it supports two types of functions for different purpose of duties [33]. Map Task is a function which is used to allocate data to nodes based on replication factor set. On the other hand, Reduce Task is also a function for aggregation of data results according to request initiated by client.

Even though Map Task and Reduce Task are two functions that are clearly visible to all parties, there are other functions in between Map Task and Reduce Task to play a role for supportive activities such as splitting, sorting, shuffling etc. Map Task depends on split function before distributing chunks of a file to nodes as per replication factor. In the same fashion, Reduce Task is heavily reliance on shuffle and sort functions in order to aggregate the result. Split function accomplishes the task of chopping file into preset size of chunks so that Map Task will able to send these chunks to a designated nodes after gathering information for free space availability. As shown in Fig. 2.4, Mappers create key/value pair for all coming chunks while storing them. Shuffle function, in addition, is responsible for taking input from Mappers and categorizing keys based on their groups. Sort function plays a role of sorting keys according their values before Reducers take in. Finally, Reducers combine similar keys and aggregate their values at each node which is local disk where the data resides.

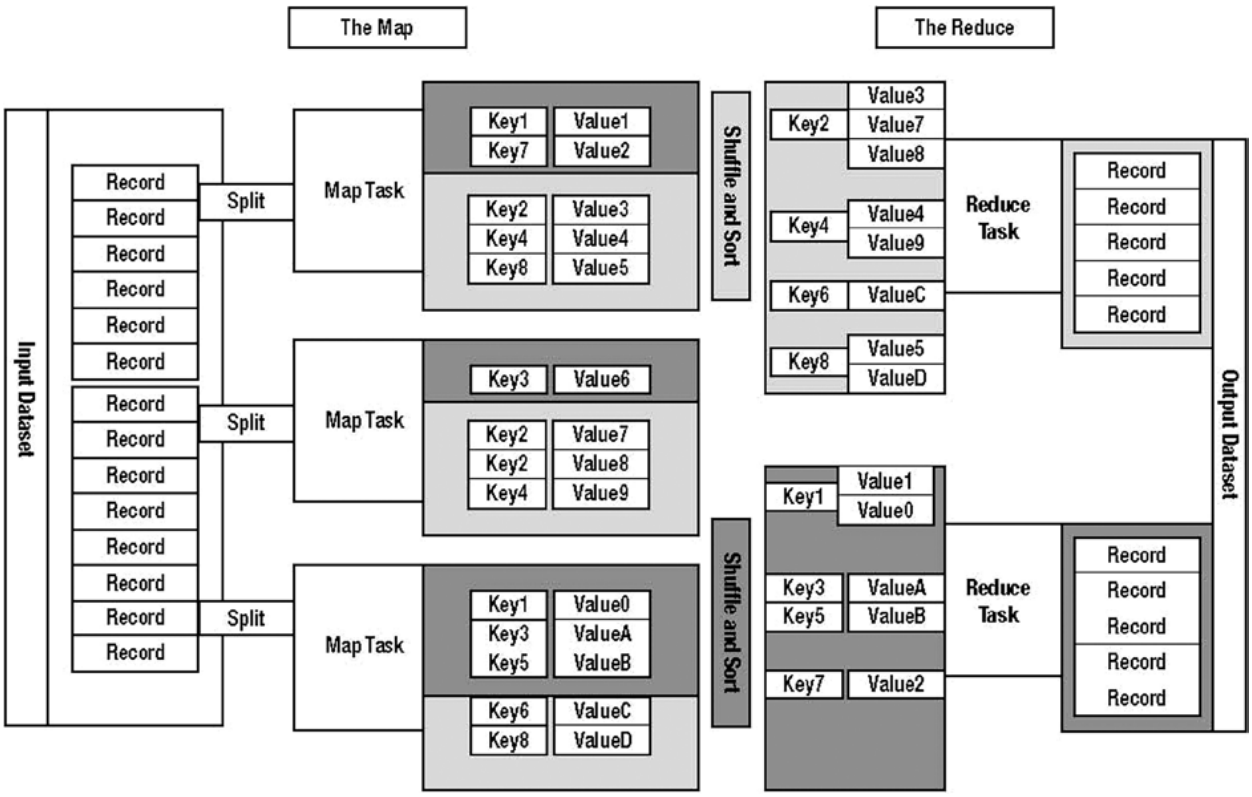


Fig. 2.4: MapReduce framework [34]

2.2.4. Data Processing (Technology Stack)

In big data technology stack scenarios as depicted in Fig. 2.6 below, style of data processing is shifted from retrieval of data from hard disk and sending it primary memory for processing to sending computation where data resides. This is great innovation for petascale data in order to avoid disk access and network traffic bottlenecks so that results will be achieved in reasonable time.

Major data processing paradigm shifts has been brought through implementation of MapReduce framework on top of Hadoop Distributed File System. Even though this has reduced burden of data transfer and manipulation to the level of uniformity in dealing big data, it has still challenge

in terms of generality to the specialists in the field by forcing them to know programming language implementation and its complexity.

Java [35] is the programming language which has been used to implement as an open source code and it is customizable by interested parties in order to adopt wherever Hadoop is used as a means to process big data. A lot of companies as well as expert communities are adopting Hadoop ecosystem and then adapting it to their own favorite environment by adding projects as depicted in Fig. 2.6. For instance, Microsoft is one of big providers of Big Data products as well as services but it has adopted Hadoop for big data storage and processing so its projects are totally dependent on Java libraries as foundation. Other programming and scripting languages are becoming part of Hadoop ecosystem as plugin onto MapReduce framework so that working on Hadoop is made as simple as performing usual projects using those languages. To mention some of these languages: Python, Ruby, SQL-like languages, Script-like languages etc. all of them run on the top of MapReduce framework.

Apache Hive [33] is a project that act like data warehouse for Hive Query Language (HQL) which provides for users a capability to process data using SQL-like language. In general, it abstracts details of MapReduce implementation such that users can inject their task into MapReduce without delving how it functions. The tasks are either sending data for storage or retrieval specific result after processing data from a set of nodes, commodity hardware. Actually, Hive queries are converted into Hadoop Jobs to run whether Map Task or Reduce Task which does not mean that rational database structure is imposed on MapReduce framework rather HQL queries are interpreted as task so that users will not be forced to write Map or Reduce Task programs to achieve data analysis objectives. Even if HQL is SQL-like language, it has additional features that are completely dissimilar to SQL queries, for example structs, maps (key/value pairs) and array.

Apache Pig [33] is a scripting language that eases to write jobs and send as MapReduce jobs so as to be executed against Hadoop. It is a platform which is openly extensible for data loading, manipulating and transforming by using scripting language is called Pig Latin. It supports complex and sophisticated data manipulation though it is simple scripting language.

SQOOP [7] is one of highest projects that is used to link relational database and Hadoop projects together. So it facilitates data movement from relational databases, structured data, to Hadoop, schema less or unstructured data, and vice versa. It is plug and play extensible framework that helps developers to program through the SQOOP application programming interface (API) so as to add new connectors.

Apache HCatalog [30] has a role to abstract data view from HDFS files stored in Hadoop into tabular form. It provides integrated abstraction form for all other projects that relay on tabular structure of data view. For instance, Pig and Hive use this abstraction in order to reduce complexity of reading data from HDFS. Despite the fact that HDFS can be any data format and stored anyplace in the cluster, HCatalog gives a means for mapping to file formats and locations into tabular view of the data. In addition, it is open and extensible for proprietary file formats.

HBase [36] is a project that supports the functionality of NoSQL (Not only SQL) database on the top of HDFS. It is a storage of large column that could be limitless number of columns along with billions of rows that facilitate fast access to huge datasets or large tables which is sparsely stored. It has a functionality of Data Modification Language (DML) [37] which supports inserts, updates and deletes; however, Hadoop by its nature it is a write once and read many or infinite times. In spite of its rational database nature, it does not provide full features of relational databases such as typed columns, security, enhanced data programmability and query language capabilities.

Flume [38] is a framework that handles streaming data of events besides the batch processing system nature of Hadoop ecosystem. It ingests incoming data stream into stages: collect, aggregate and shifts large volumes of data before commits to HDFS. It has major components – client, source, channel, sink and destination so that events flow through all components from client to destination.

Apache Mahout [39] is a machine learning and its overall goal is developing scalable machine learning libraries which is implemented on the top of Hadoop using MapReduce framework. Currently, it is based on four use cases: recommendation mining is used as core for recommendation engine, clustering is used to group documents based on related topics, classification is an algorithm that consumes already classified documents so as to classify new documents and frequent item set mining is a means of understanding bucketed items together.

Ambari, Oozie and Zookeeper are supporting tools for Hadoop ecosystem to work efficiently and effectively data analyzing process. Ambari [40] is a system center to Hadoop ecosystem for provisioning, operational insight and management of cluster. Oozie [41] is a scheduling application for Hadoop that manages chain of events, processing or processes which must be initiated and completed at specific time interval. Zookeeper [42], on the other hand, is used to support to manage and store configuration information.

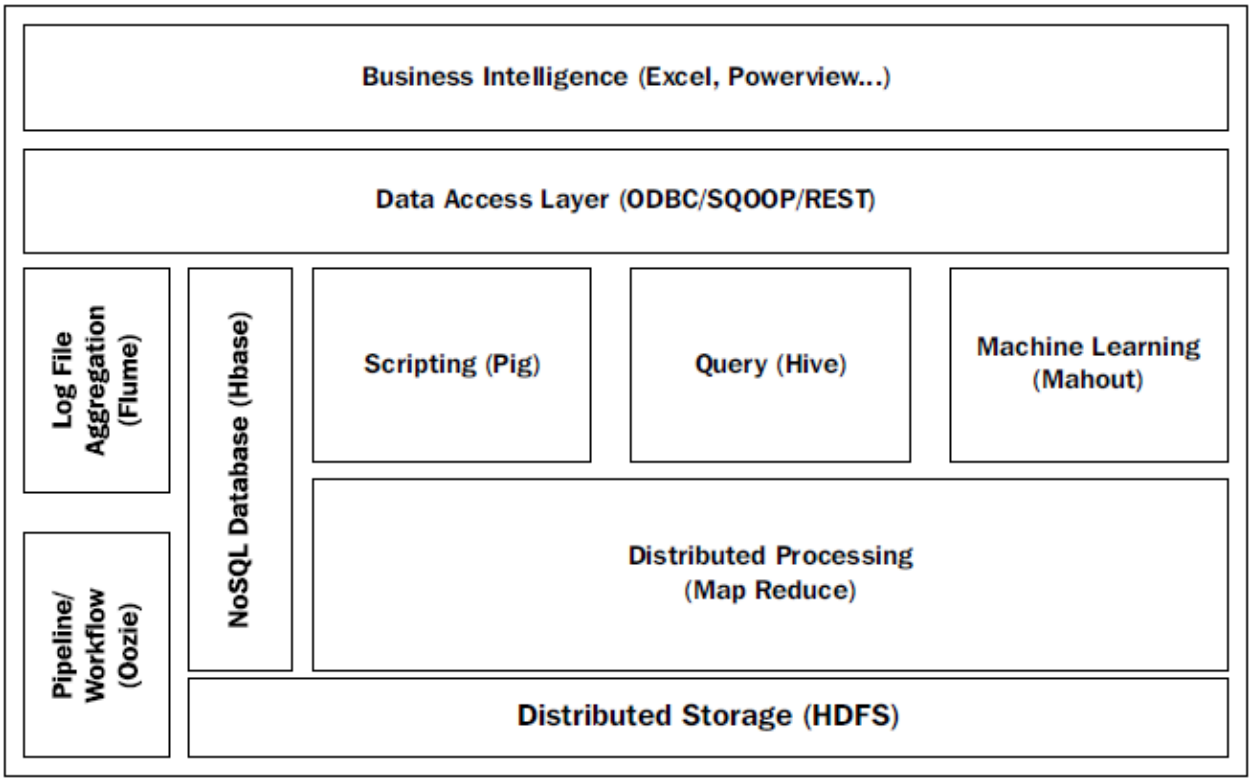


Fig. 2.5: Hadoop ecosystem [43]

2.2.5. Data Visualization (Presentation)

Using visualizations to express or communicate ideas is one of preexisting or oldest practice of human being before start of written materials. It is the most primitive means of communication which is dated back 3,000 B.C. because vision is the first and most used form of communication. Moreover, it is single important faculty of sense which helps to process and grasp huge information as compared to others. Beginning cave drawings to modern charts have been playing major roles in conveying pertinent information among people, organizations and others. The benefits of visualization is multidimensional in terms of vast volume of information at a time and its space usage is very small when compared to tables and textual data. Visualization system provides deep understanding independent of any language which is enlarges grasping capabilities of complexity of information. As Tufte [44] said, "Graphical excellence is that which gives to the viewer the

greatest number of ideas in the shortest time with the least ink in the smallest space." Actually, data visualization has two targets which are explanatory and exploratory for its users; where explanatory shows direct information that viewer begins with specific question whereas exploratory firstly presents information and then encourages viewer to generate questions from presented information [45].

Presentation of big data, requires exploratory type of visualization, differs from traditional Business Intelligence (BI), highly dependent on explanatory visualization type, in many different ways. Traditional BI tools have been focusing on models and reports that are consumable by few highly trained data analysts and executives. These models and reports are narrower in scope and additionally relays on historical and internal data only. As the size or volume of data bursts their embracing power would be shrink in much proportion so organizations decision making based on vast data keeping aside velocity and variety would be limited. Even variety and velocity present more challenges for data visualization by traditional BI technology stacks. In addition, it takes weeks or months to generate reports and dashboards from which necessary figures, issue static, rearview reports are pulled for executives and employees by highly trained data experts [46].

Emergence of big data has brought opportunities to create and utilize a number of self-serving data visualization tools into existence as shown in Fig. 2.6. These tools provide enormous options to all levels of users so that they can consume appropriate information regardless of time and space which is ubiquitous visualization of data. One of the tools is Self-Service Business Intelligence (SSBI) visualization platform which improves accessibility through smartphones, tablets, notebooks, laptops, desktops and so on. It provides a capability to "mash up" data from a number of sources: click stream, social media, log files, videos, and more. So users are able to analyze and visualize in real time with their high performing desktops as well as mobile devices in order to get

insight for their business. And it is supplied by TIBCO software which is second largest data discovery vendor in the world [46].

Limitation of visualization processing and display especially for big data depends on a number of factors such as nature of data, processing capacity of the machine, screen size and its resolution. The data items to be visualized has impact on type of visual means such as bar chart, pie chart, scatterplot, bubble chart, boxplot, heat maps, and others which have inbuilt size of accommodation. However, it is not without solution for this challenges; for instance, data analytics plays its role in reducing data size and complexity to the point that the level of appropriate information consumable by intended audiences [44].

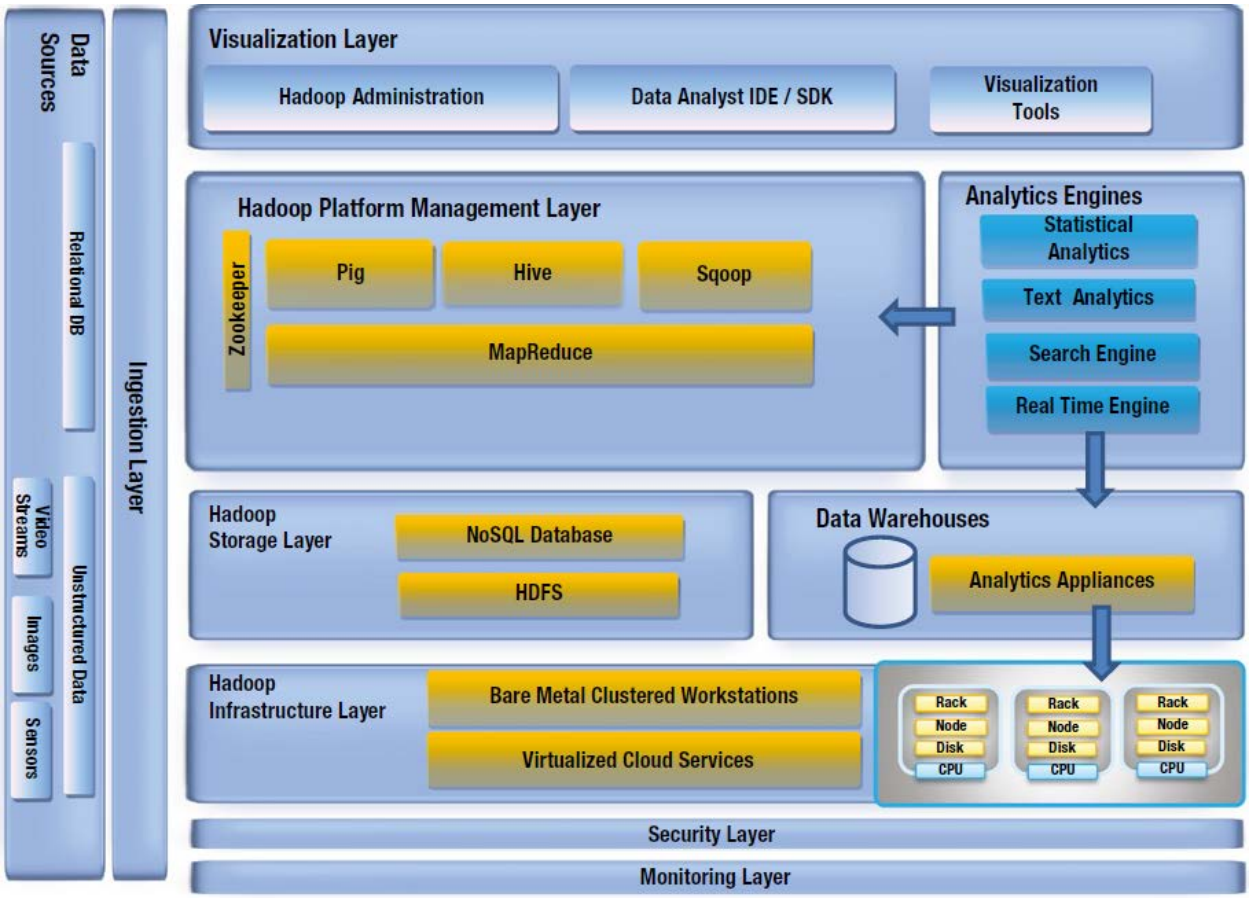


Fig. 2.6: big data architecture [1]

2.3. Related Works

As described in [10], Big Data components, challenges and opportunities is discussed to review the evolution and current state of Big Data in terms of seven dimensions, historical background, what's big data?, data collection, data analysis, data visualization, impact, human capital, and infrastructure & solutions. It surveys and distills literatures in order to know the effects of Big Data in business environment. The study is clearly showing the rewards of Big Data not only in business environments but also in everyday life activities of individuals. In general, it is conceptually indicating that Big Data and Analytics require all the seven dimensions in today's business environment.

Large scale web mining by utilizing Data Intensive Scalable Computing (DISC) System to extract information and models from web data necessitates traditional algorithms by putting power of parallelism [11]. DISC system is considered as one of powerful, fault tolerant and inexpensive to process large data sets even though it has limited computing primitive. The study has tackled three classical problems in Web mining: finding similar items from a bag of Web pages, content distribution from Web 2.0 to users through graph matching and suggesting new articles from stream in real time.

As indicated in [12], the study deals with design of conceptual Big Data adoption Model by exploring Big Data solution adoption within organizations. The methodology used is multi-case study research by interviewing practitioners of Big Data in telecommunication and energy utility sectors. Its result was a strategy development phase, a knowledge development phase, a pilot/test-case phase and a fine tuning phase are followed by organizations to implement Big Data solution.

As indicated in [3], high speed real time Big Data processing using Storm system is used instead of MapReduce which is appropriate for batch processing. Storm is distributed and fault tolerance system which achieves processing in collaboration with other tools such as Cassandra, Redis and Kafka over NoSQL. The study is also proposed system architecture that support to process Twitter and Bitly streams of data.

The research in [47] explains effects of Big Data, expressed by 3Vs, analytics on organizations' value creation. As per the study, data growth rate is becoming enormously high which has forced organizations to look for new technologies to handle it economically. Case study methodology is used to confirm value creation in organizations using Big Data analytics. The finding shows that Big Data analytics might create value in two ways: improving transaction efficiency and supporting innovation.

As explored in [48], spring of data sources are stretching limits of traditional data management so as to extract unused sources to gain more insights. To realize these values, organizations need to consider architectural expansion to accommodate new technologies on top of traditional architecture. According to the research, additional requirements have to be elicited on the basis of new data behavior to design reference architecture by combining several data management components. The reference architecture is built on traditional enterprise data warehouse architecture using evolutionary approach.

Literature review, related works and knowledge of researcher show that researches are conducted as conceptual studies in general to date. The studies indicate a real and current challenges of flood of data from a varying sources in different formats with high frequency; and they are showing potential benefits of Big Data implementation in organizations using qualitative methodology. But

the challenges are not addressed using experimental study for implementation and usage problems that are faced by all levels yet. In this study, Big Data processing and visualization particularly on unstructured data sets are conducted taking into account Volume of Big Data.

Chapter Three
Data Collection and Design

3.1. Data Collection

There are open access data for public use and research community to carry their activities without any associated fee even though now companies whose line of business (LOB) as data broker are flourishing in the data market. These companies are collecting demographic data for sale to all interested companies so as to increase their customer base.

3.1.1. Data Type/Nature

Big data is a type of diversified data that cannot be forced to align a certain format or confront standards and practice of an organization. In addition, in a big data scenario, data is short lived in terms of value that could be extracted from it for decision making or actions. Data has to be connected with other data sets to be most valuable to yield accurate insight.

Most big data research projects deal with behavioral aspect of data rather than pursuing veracity of data. This is because data creation is decentralized to individuals who are expressing their activities and whereabouts using various technologies like facebook, twitter, google+, pinterest, instagram and so on. So, truthfulness of data from individual data creator cannot be verified by any means. As a matter of fact, big data analytics projects are concentrating on behavioral aspects of data.

Categorization or sorting of data for management or manipulation is not a simple task. On the top of unstructured data sets, there are other types of data sets that are structured data which are transactional or machine generated data sets and semi structured data sets which are generated

from social media sources. Free text, for example books, is one of unstructured data sets that require high computational resources to preprocess and process with multiple steps so as to come up with final result.

In this study, unstructured data sets, free text, particularly more than five hundred eBooks of philosophy category is used for processing. The books are from a number of languages such as English, French, Chinese, Germany, Greece etc. and they are zip files that are being extracted while processing them in Hadoop framework.

3.1.2. Data Size

Current limitation of data storage and retrieval is triggering for further innovation in the form of software rather than adding more storage and processing speed for hardware. Commodity hardware machines are gaining bases in large companies instead of cutting edge super computers in order to solve complex data problems. As the size of data is sky rocketing from time to time from every direction, it is possible to say everything is generating data from every corner, about hundreds of Exabyte in a day. More importantly, the philosophy of the value of data overtime is reversed in the way that every drop of data should be tracked so as to tap its value because data is "new oil" [49].

Big data as target population can be varying size to be processed for a given analysis which depends on available sources and resources to analyzer(s). Analyzer might have hundreds of thousands clusters of commodity hardware machines which could be grouped into thousands of racks that will provide greater deep insight which shape business operations and strategy as well. Therefore, there is no upper ceiling for determining how vast data sets are good enough in order to get better or reliable insight so that business may utilize to act up on. More data, in general,

provides better and different perspective to see in depth; for instance take a sample of a population to test statistical significance of a certain treat but the sample might not yield practical significance for aforementioned test result whereas population as a whole might give real statistical significance for which the sample could not be revealed. Similarly, big data with its wealthy revelation is able to provide statistical significance regardless of practical significance.

Lower limit of data, on the other hand, in terms of size might be impossible to determine however either the composition of 3Vs [7] or Terabyte of transactional data that traditional data processing technology such as RDBMS [33] or data warehouse tools are incapable to store as well as process effectively rather their upper limit is bounded in the range of Gigabytes. It is not all about the size that determine minimum amount of data to be considered in order to be called big data, it is also important to know variety of data that determine its fate as big data as well. Another major factor that influence shift of technology in addition to volume and variety is velocity – the rate at which data creation takes place and its pouring into global data accumulation which brings challenge to traditional technology stack to tackle speed of data processing.

For purpose of this research, one data type, unstructured data sets, is used with size of 210MB as experimentation point of implementation for processing words in more than five hundred documents in a category of philosophy, Africa, Language Education, Wars and Science Fiction. It is final testing and to make it presentable through visualization.

3.1.3. Data Sources

There are plenty of data sources for public use which have been availed by different bodies like governments, non-governmental organizations, corporations and the like. Even world number one companies whose business totally dependent on data especially social data are selling data to other

companies who are interested in 360 degree view of their customers. One of these companies is providing free data set but it is limited set in terms of size and completeness. For instance in [50], Twitter has provided application programing interface (API) so that the research communities are able to ingest to their specific task. In Ethiopia, there is no such practice of availing open sources standard data access provisioning which might encourage more and better innovation among research communities.

From free available data sources for public use, mainly one data source will be used to conduct this research work as indicated in section 1.6.2. It consists of thousands of free books in a number of formats but text format book is appropriate for words in a document processing task. As a matter of fact words in a document processing for large sized books are not simple task which could be very difficult to handle by a single individual that might take longer time to finish it.

Individual book with plenty of pages that may require a series of steps to process in order to generate indexes for all words accordingly. However, HDFS file structure and MapReduce framework provides a means to manipulate in speed. In the first place, Map function distributes data to nodes by chunking a file into preset block size. And then Reduce function will sort and shuffle all words by individual word category after that Reduce function summarizes as a final single result.

3.2. Planning of Technology Stacks

The technology stacks that are used to demonstrate this research experiments based on available limited resources. Actually, big data projects implementation demands huge investment in terms of human resources, fund, space and other resources. For instance [51], Facebook is one of heavy implementer of big data technologies for its data processing activity. As depicted in Fig. 3.1, it has

hundreds of thousands of clusters of machines in its data centers which are managed by highly skilled experts. It processes petabytes of data every single day which is indeed collected from Facebook users who are uploading images, videos, sharing events, comments and the likes so that Facebook Company is able to use this information to advertise targeting, suggesting friends of friends to be connected and so forth.

Implementation of required components such as Hadoop single node cluster and its related java SDK to experiment is carried out in single machine. At the same time, external hard disk with 1TB has been used to install Hadoop framework and store experimental data so that data can be replicated to DataNodes and then processed locally. Furthermore, NameNode and JobTracker nodes are used to submit, control and monitor job execution.

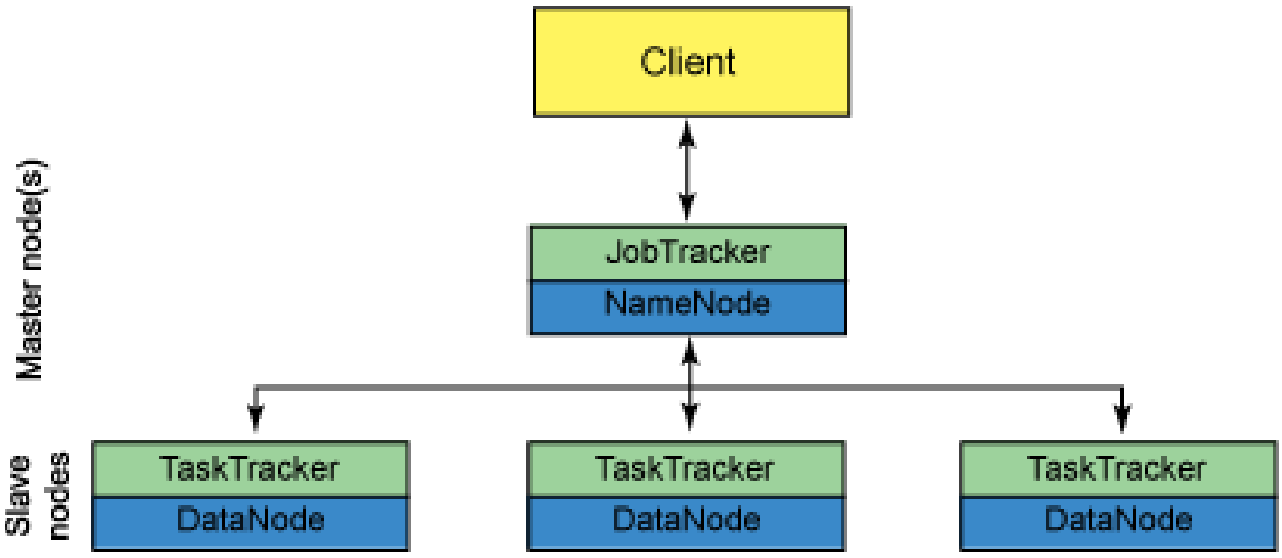


Fig. 3.1: General Architecture of Hadoop Framework [52]

3.3. Architecture of the System

The major components of the system comprise Hadoop framework and MapReduce framework. Hadoop framework library has been implemented as a bed platform so that other tools are able to

run on top of it. MapReduce framework, on top of Hadoop framework, utilizes classes to execute its functions.

As shown in Fig. 3.2, the architecture of this research implementation bases on single node cluster of Apache Hadoop 2.6.1 framework which encompasses two nodes; namely, NameNode which takes command from client and assigns task to DataNode, and DataNode where actual data processing takes place. In addition, it integrates with projects like Pig and Hive so that it has a capability to analyze and present data in a way that easily understandable.

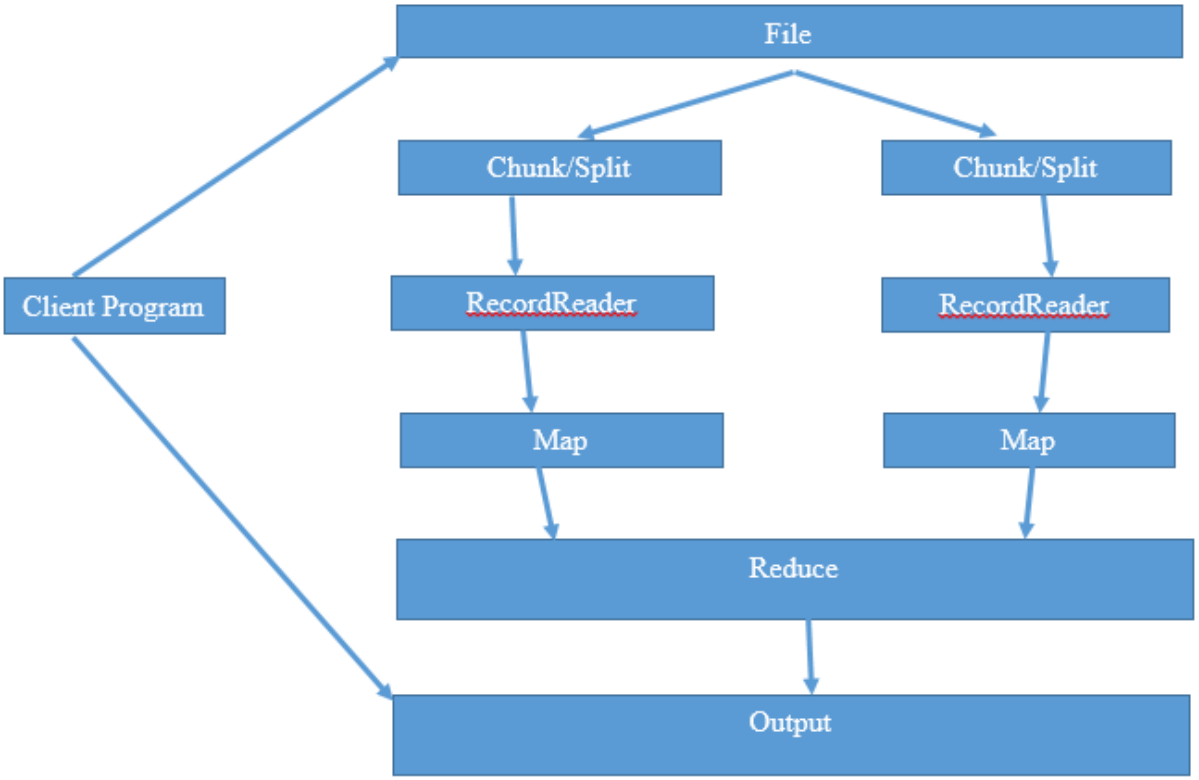


Fig. 3.2: Single node cluster architecture

3.4. Design

Implementing big data technology stacks require careful planning and management works to be executed. As number of clusters increases, clusters is organized into racks and number of nodes in a cluster depends on file size and storage space of each node. Complexity of design for Hadoop ecosystem is directly proportional to file size to be stored, storage space of nodes, size of clusters and number of racks to group number of clusters. Even though Master nodes and Jobtracker nodes are not part of clusters as well as racks, they play a major role in controlling, managing, scheduling jobs etc. of all clusters and racks being outside of both clusters and racks.

The larger file size that reduces the number of nodes in a cluster which directly affects the efficiency of Map task processing ability by slowing because total file breakdown into chunks become heavy. On the contrary, as file size becomes smaller which is least size of 64 MBs the number of nodes in a cluster will be higher. Similarly, performance of Map task becomes faster than larger file sizes; however, Reduce task gets much load in processing or aggregating data from a number of nodes.

3.4.1. Design Goal

The goal of design is developing Hadoop ecosystem environment to test big data technology stack on unstructured data sets or words in documents processing so as to differentiate from traditional relational data technology stack. In doing so, test environment will be setup using Apache Hadoop library on the top of Ubuntu Linux operating system which is native operation platform for Hadoop ecosystem. In general, the main approach that is employed to conduct this research is experimenting a set of big data technology stack by performing data processing and visualization.

3.4.2. Experimental Procedure

As starting point to conduct implementation and testing this research, the following step by step approach is taken into considerations. First, Ubuntu Linux operating system has been installed on a single machine which is used as platform for other big data technology stack to run. Second, Hadoop 2v has been downloaded, installed and configured as pseudo NameNode, JobTracker, TaskTracker and DataNode. It is pseudo nodes because whenever Hadoop framework is installed and configured on single machine, it acts as two nodes which is divided into master and slave nodes. In real scenarios, Hadoop implementation requires two or more machines in order to separate master node which comprises NameNode and JobTracker, and slave nodes that consists of DataNode and TaskTracker nodes. Third, ingesting data using client program from command line to start processing is done on MapReduce framework. Finally, appropriate visualizations tools is used to present analysis result as comprehensive as possible by reducing complexity and information overload.

3.4.3. Data Analytics Design

As presented in Algorithm 4.1, the Mapper algorithm is used to accomplish three major tasks. The first task is taking files submitted by client program(s) and then chopping into blocks of chunks according to preset size, 64MB, of chunks. The second task is creating key/value pairs of the chunks. Finally, the third task is replicating chunks to nodes where enough space is assured.

As seen in Algorithm 4.2, the Reducer algorithm, on the other hand, performs four main tasks. The first task is going to nodes where target chunks are stored and then retrieving key/value pairs, sorts these key/value pairs and aggregate them locally. The second task is collecting from all nodes aggregated key/value pairs after that shuffles and sorts, and aggregates them as intermediate result

to the next level of processing. The third task is taking intermediate key/value pairs to produce final result by shuffling, sorting and aggregating, and the fourth task is responding to client program final result.

Algorithm 4.1 Map function [31]
The mapper emits an intermediate key-value pair for each word in a document. 1: class Mapper 2: method Map(docid a, doc d) 3: for all term t ∈ doc d do 4: Emit(term t, count 1)

Map (k1, v1) → list (k2, v2) ---------- mapper algorithm stores chunks in the form of key/value pairs.

Algorithm 4.2 Reduce function [31]
The reducer sums up all counts for each word. 1: class Reducer 2: method Reduce(term t, counts [c1, c2, . . .]) 3: sum ← 0 4: for all count c ∈ counts [c1, c2, . . .] do 5: sum ← sum + c 6: Emit(term t, count sum)

Reduce (k2, list (v2)) → list (v2) ---------- reducer algorithm processes data locally to merge intermediate result.

As point of start, the experiment deals with words of all books by counting individual word throughout the books and summing up their total number. The books have been chopped into chunks to be stored in DataNodes using Mapper function which provides key/value pairs of words.

That means every word is mapped its name as a ‘key’ and ‘one (or 1)’ as value at time of split process. After splitting is done, TaskTracker starts placing these chunks to designed memory addresses in the format of Hadoop Distributed File System (HDFS). In the return, NameNode registers the address of all chunks which is actually metadata of chunks that explains detail about each chunk. In addition, TaskTracker reports back to JobTracker about its accomplishment and any failure if there is any.

On the other hand, whenever there is job submission to MapReduce framework to process stored files, Reducer function brings required result by aggregating from a number of nodes. NameNode and JobTracker coordinate the processing of submitted jobs into Hadoop platform by assigning tasks to TaskTracker which in return facilitates Reducer function to sort, shuffle and aggregate locally and then return result to next level. TaskTracker is required to update status of tasks in specified time interval; for instance, as heartbeat messages to JobTracker which is responsible to scheduling and coordinating tasks across nodes. Reducer function performs activities locally, which means process to data paradigm where data resides, such as reading blocks of data, combining similar key/value pairs and sorting.

3.4.4. Data Visualization Design

Even if huge set of data yields immerse knowledge to make decisions, without data visualizations tools the values of data would be difficult to be realized [53]. The difficulty of snatching out meaning or insight from big data will increase by many folds without data visualization tools in place. So, the importance of data visualization is like hand and glove for big data scenarios because it does not make sense just long processing data for the sake of analysis.

Data visualization is dependent on a number of factors to be effective which could hinders its utilization as well as applicability for desired purpose unless properly considered in detail at the time of design [54]. The major ingredients of data visualization elements are screen size, screen resolution, data nature and machine capacity. Screen size creates a room to accommodate more data elements and at the time it creates comfort to explore as well as navigate. Screen resolution, on the other hand, provides an ability to see or visualize clearly all data sets in terms of inter data elements and intra data elements as well. Data nature puts burden for visualization tools as it is more unstructured, variety in type and huge in amount. Finally, machine capacity plays major role in processing and presenting to end user. In a nut shell, data visualization constraints can be expressed in terms of a formula below.

Data Visualization = screen size + screen resolution + data nature + machine capacity

In this study, data visualization is considering all the above factors to ensure data presentation through smart devices (mobiles and tablets) as well as computers (desktop and laptop). As it is well known that smart devices have limited capabilities in terms of screen size and machine capacity; however, data visualization is encompassing this limitation by adding interactivity by hovering cursor over data elements. On the other hand, computers are de facto standards for any data visualization design so the design data first targets computers and then smart devices become part of it.

3.5. Algorithms

Although Hadoop framework provides foundational functionalities that are consumed by MapReduce libraries to accomplish big data processing, the major algorithms that are required to ingest data sets into Hadoop Distributed File System are Mapper and Reducer functionalities to be implemented for specific problems. As a matter of fact, problems for different context and purpose

of the business demands the choice of approaches as well as unique implementation or application of algorithms for specific situation. For instance, two completely unrelated data natures such as data of IoT and social network data need separate treatment of them.

3.5.1. Mapper Algorithm

Every data set which could be structured, semi-structured or unstructured data has to be split into preset size of chunks by default 64MB for which HDFS minimum support is 64MB or 128MB. Mapper algorithm makes use of a number of libraries from Hadoop Distributed File System, e.g. RecordReader function in one to one alignment, which provides input file split so that Map function will use it to produce intermediate result for the next level of processing. Actually, Map function produces key/value pairs of a given chunk according to algorithm designed as per nature of data set to be processed. A file can be spilt into a number of splits or chunks and one split is directly processed by one Map function.

In this research, we make use of four Mapper algorithms to experiment words in all documents processing from the documents. In the case of words in documenst processing problem, Map algorithm breaks every line of statements of a document as key and its line numbers as value and then subsequently each line of statements will be broken down into words which is taken as key and its value is assigned to '1' (one). In general, all Map functions that are assigned data set processing task run in parallel which is monitored and controlled by JobTracker and TaskTracker in cascading. If there are failed tasks, MapReduce framework reinitiates tasks again to ensure fault-tolerance.

The mapper algorithm follows the following step by step procedure:

1: Hadoop reads files

2: files are chopped as per preset size into chunks

3: chunk is allocated to specific DataNode

4: individual chunk/split is read by RecordReader a line of statement with corresponding line number at a time

5: a line of statement is tokenized by Map function into a word and 1 (one) as a value

6: output key/value pair to intermediate result set

3.5.2. Reducer Algorithm

On the other hand, Reducer algorithm plays a great role in aggregating values of a key by summing or combining set of values from a single or multiple Map functions. Reduce function depends on two major libraries from Hadoop are shuffle and sort functions which take intermediate output from Map function as input to shuffle and sort the same keys together so that Reduce function can easily combine or aggregate values of each key. A single Reduce function, most of the time, is implemented for all Map functions output aggregation.

The implementation of Reducer algorithm in this research considers single Reduce function in order to aggregate outputs of Map functions. Intermediate outputs from shuffle and sort functions, these functions are libraries of the framework, is directly processed by Reduce function. The single file output which is generated by Reduce function will be taken to visualization to present result in convenient for human interpretation.

The reducer algorithm follows the following step by step procedure:

1: MapReduce library shuffles and sorts intermediate result

2: For each word, its value is aggregated

3: Hadoop writes key/values of aggregated word to Hadoop Distributed File System file

4: Output file is saved to local file system

3.6. Visual Components

The results of huge data sets processing might not be comprehensible in the way as traditional data visualization tools and practice does. It needs new way of presenting information within usual screen size and pixels density as simple as easily understandable by all level of experts to accomplish their day to day activities. Data visualization gives a flavor in grasping overall trends of events and changes across horizon. Most important factors in designing visualization components in big data scenarios is knowledge of elements of data and relationship among data sets.

The output of Reduce function is raw result by itself which requires appropriate data formatting and arrangement so as to be consumable by target audiences. So, it is necessary to select suitable visualization tools for specific data type. The chosen visualization tool to this research is Tableau Big Data Visualization tool which is highly powerful in the area of Business Intelligence and now big data visualization is incorporated as a service into a Public Edition Tableau Desktop Version 9.3 [55].

Chapter Four
Experimentation and Results

The experiment of this research is applied on unstructured data set which is to show application of big data technology stacks in scenarios of Single Hadoop Node Setup. The setup of Single Hadoop Node but it is pseudo distributed Hadoop framework which acts as fully distributed cluster of nodes which comprises all nodes; such as NameNode, DataNode, Secondary DataNode, JobTracker and TaskTracker that are important to process experimental data sets.

4.1. Experimentation

All documents are chopped into four chunks so that one of chunks or splits is being processed on each of DataNodes. Totally, four DataNodes are used to process the data sets; however, the reducer is aggregating output of mappers using single node, one of four DataNodes, in order to return final result to specified location. Mapper firstly ingests a split using RecordReader library which provides every single line of statements as key/value pairs.

Hadoop ecosystem is coming up with great advantages for current limitation of computation by enhancing processing speed and storage capacity. As shown in Fig. 4.1, it just took four minutes and eighteen seconds which is totally 278 seconds to process more than five hundred sixty books of 205MB with total of greater than 40 million words.

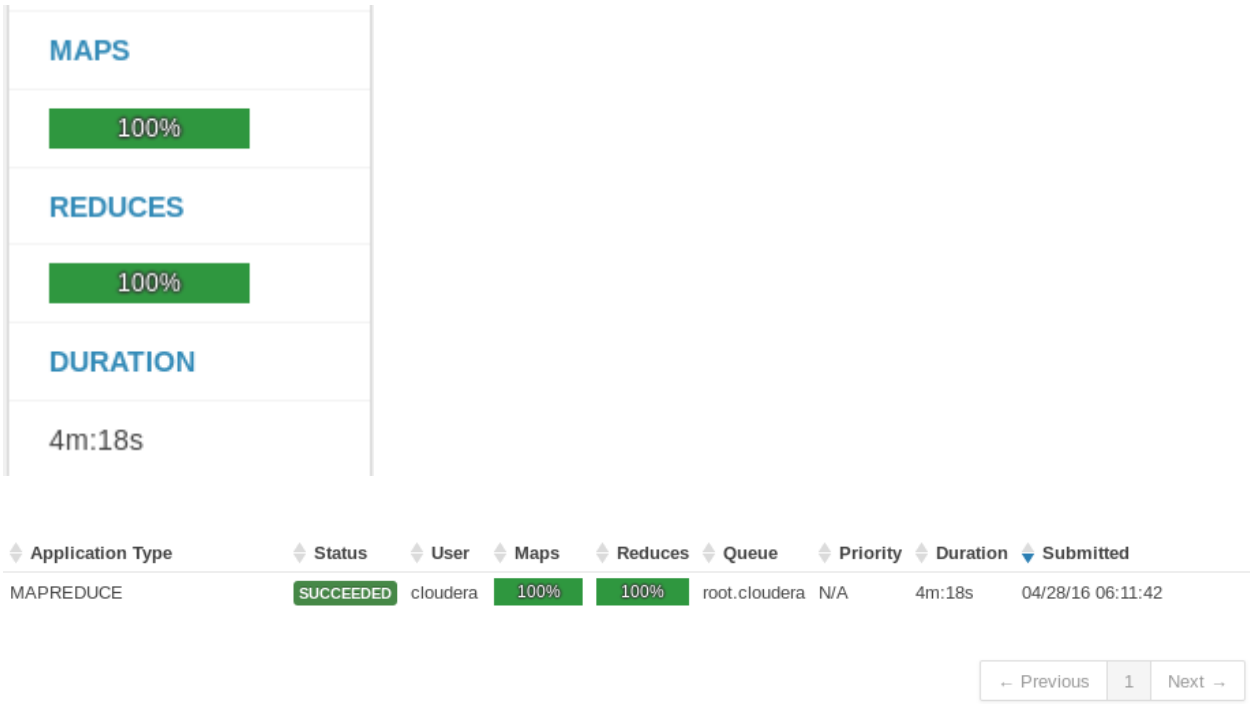


Fig. 4.1: MapReduce execution duration

The documents processing functions of MapReduce framework was running in three separate modes such as Mappers and Reducer functions mainly; in addition, sort and shuffle functions play critical role in facilitating huge data set to be shuffled and sorted within short period of time. The following Table 4.1 shows the role of JobTracker, TaskTrackers, Mappers and Reducer.

Name	**Maps Total**	**Reduces Total**	**Total**
File Bytes Read	477411456	477411366	954822822
File Bytes Written	955269776	477523045	1432792821
File Large Read Ops	0	0	0
File Read Ops	0	0	0
File Write Ops	0	0	0
Hdfs Bytes Read	214732711	0	214732711
Hdfs Bytes Written	0	3106396	3106396

Hdfs Large Read Ops	0	0	0
Hdfs Read Ops	12	3	15
Hdfs Write Ops	0	2	2

Table 4.1: MapReduce file system

As it is seen in Table 4.1, 477411456 Bytes (455 MB) and 477411366 Bytes (455 MB) of regular file data are read by Mapper and Reducer function respectively, totally 954822822 Bytes (911 MB). And 955269776 Bytes (911 MB) and 477523045 Bytes (455 MB) of regular file data are written by Mapper and Reducer functions respectively, totally 1432792821 Bytes (1 GB). On the other hand, 214732711 Bytes (205 MB) of HDFS data are read by Mapper function and 3106396 Bytes (3 MB) of HDFS data are written by Reducer function. There are 12 and 3 HDFS read operations of Mapper and Reducer functions respectively. There are also 2 HDFS write operations of Reducer function.

Name	**Total**
Data Local Maps	4
Total Launched Maps	4
Total Launched Reduces	1

Table 4.2: MapReduce Job

As shown in Table 4.2, JobTracker, TaskTracker, Shuffle and Sort, InputFormat and OutputFormat, a set of execution and parameters have been taken place with respect to Mapper and Reducer functions. For instance in JobTracker, there are four Data Local Maps, four Total Launched Maps and one Total Launched Reduce. In TaskTracker as shown in Table 4.4, there are four Merged Maps and four Shuffled Maps for Reduce. As shown in Table 4.3, Input and Output

format have been 214732211 Bytes (205 MB) read and 3106396 Bytes (3 MB) written Map and Reduce respectively.

Name	Maps Total	Reduces Total	Total
Bytes Read	214732211	0	214732211
Bytes Written	0	3106396	3106396

Table 4.3: Input and Output Format

4.2. Results

4.2.1. Data Processing

The results of experimentation show in Table 4.4 that CPU milliseconds for Maps function is 158960, Reduce function is 140000 and it is totally 298960; GC milliseconds for Maps function is 5175, for Reduce function 748 and it is totally 5923; Physical Memory Bytes for Maps function is 2006507520 (~2 GB), for Reduce function is 807206912 (~1 GB) and it is totally 2813714432 (~3 GB); and Virtual Memory Bytes for Maps function is 6229786624 (~6 GB), for Reduce function is 1566797824 (~1.5 GB) and it is totally 7796584448 (~7.5 GB) which indicate big data processing requires least resource utilization in terms of time and computation as it is compared to resource requirements of transactional data warehouse data set processing. Hadoop ecosystem has provided tremendous capabilities to ingest and process huge data sets with a least resource requirements. As shown in Table 4.4, the time and memory to compute for specified data set were manifested by its execution and generated result.

Name	Maps Total	Reduces Total	Total
Cpu Milliseconds	158960	140000	298960

Gc Time Millis	5175	748	5923
Map Output Bytes	389234094	0	389234094
Map Output Records	44088633	0	44088633
Merged Map Outputs	0	4	4
Physical Memory Bytes	2006507520	807206912	2813714432
Reduce Shuffle Bytes	0	477411384	477411384
Shuffled Maps	0	4	4
Virtual Memory Bytes	6229786624	1566797824	7796584448

Table 4.4: MapReduce Task

In particular, the result of experiment is manifesting that Hadoop ecosystem is processing unstructured data sets of Big Data in inexpensive and with high throughput. As shown in Table 4.5, IO Error, Wrong Length, Wrong Map and Wrong Reduce are all zero; so, the Big Data processing using Hadoop ecosystem is fault tolerant and reliable.

Name	**Maps Total**	**Reduces Total**	**Total**
Bad Id	0	0	0
Connection	0	0	0
Io Error	0	0	0
Wrong Length	0	0	0
Wrong Map	0	0	0
Wrong Reduce	0	0	0

Table 4.5: Shuffle Errors

4.2.2. Data Visualization

Even though experimentation has generated a single raw file which is output of Reduce function, the need of data visualization of the same file is a must so that the result of processing could easily be understood to grasp the information in appropriate and consumable format. However, the availability of visualization platforms for big data is just handful, i.e., there are very few companies as providers of visualization components. Actually, big data technology sets are now emerging on the top of traditional business intelligence technologies so big data visualization toolsets are at their infant stage. As shown below charts, the output file of MapReduce framework processing is converted into interactive charts using Tableau visualization platform. Tableau is one of few great visualization tools that embraces whole output of big data processing result without breaking down into a set of files so as to visualize. In addition, its charts are consumable through all devices regardless of their screen size or pixel density because it provides interactivity by allowing to hover mouse cursor in order to spot value of an element of interest.

All charts or graphs show results of processing in different forms but they are conveying the same information by accommodating capability to figure out the content or value of a single element of data. Horizontal Bar chart, Treemap, Pie Chart, Highlight Table, Stacked Bar Chart, Circle Views Chart, Bubble Chart, Box-and-Whisker plot, Heat Map and Packed Bubbles Chart are used to present MapReduce framework processing results. Each of them has given interactivity as well as elegant format of information presentation capabilities by enhancing information consumption for all audiences. More importantly, such information presentation for advanced users creates a room for further exploration and analysis.

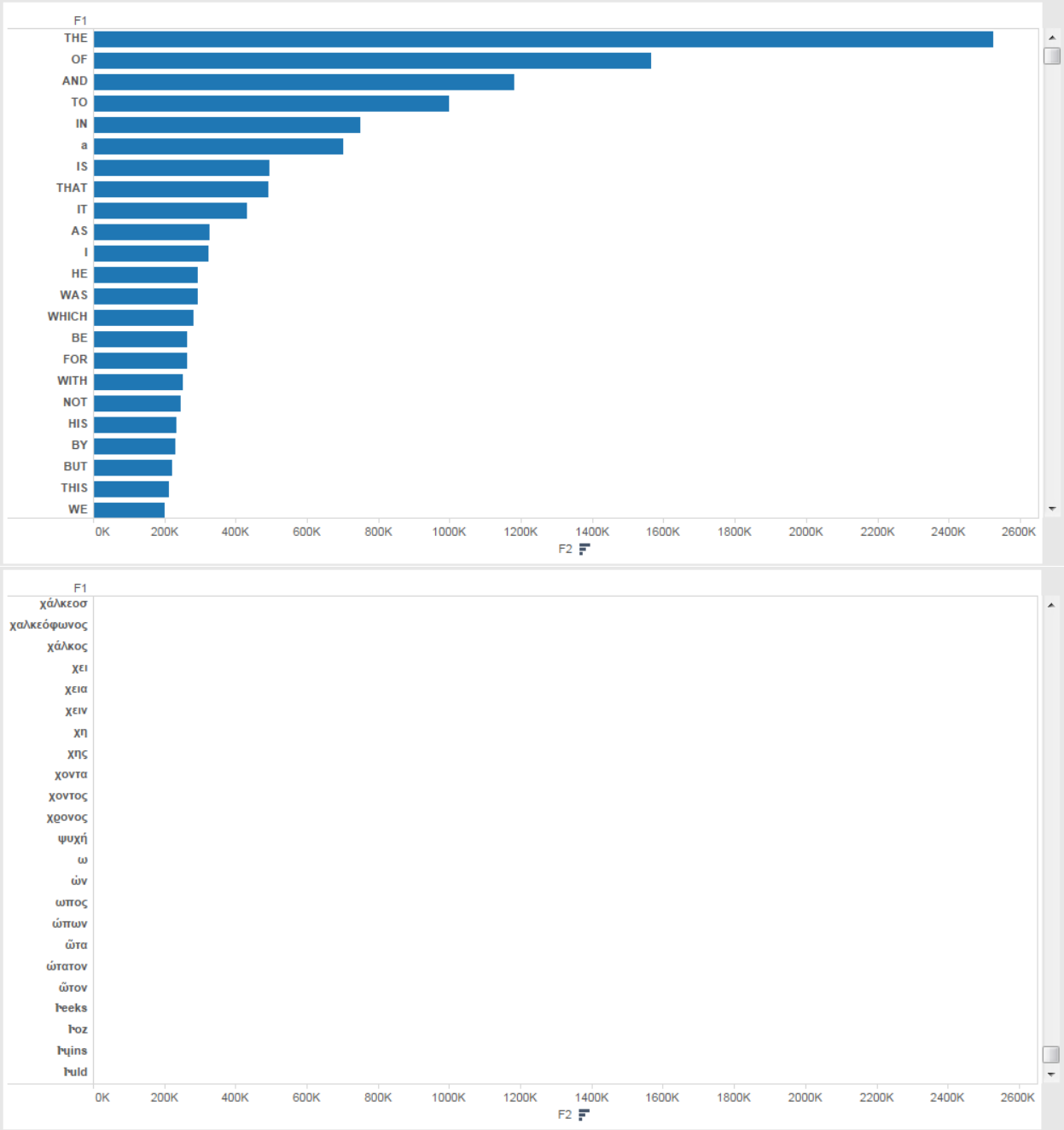


Fig. 4.2: Horizontal Bar chart

As Horizontal Bar chart indicates, the count of a word is shown by length of a bar by demonstrating its relative value from other words. For instance, words are listed on vertical axis and values of each word is placed on horizontal axis and the graph is interactive enough to display specific value of a word by hovering over it.

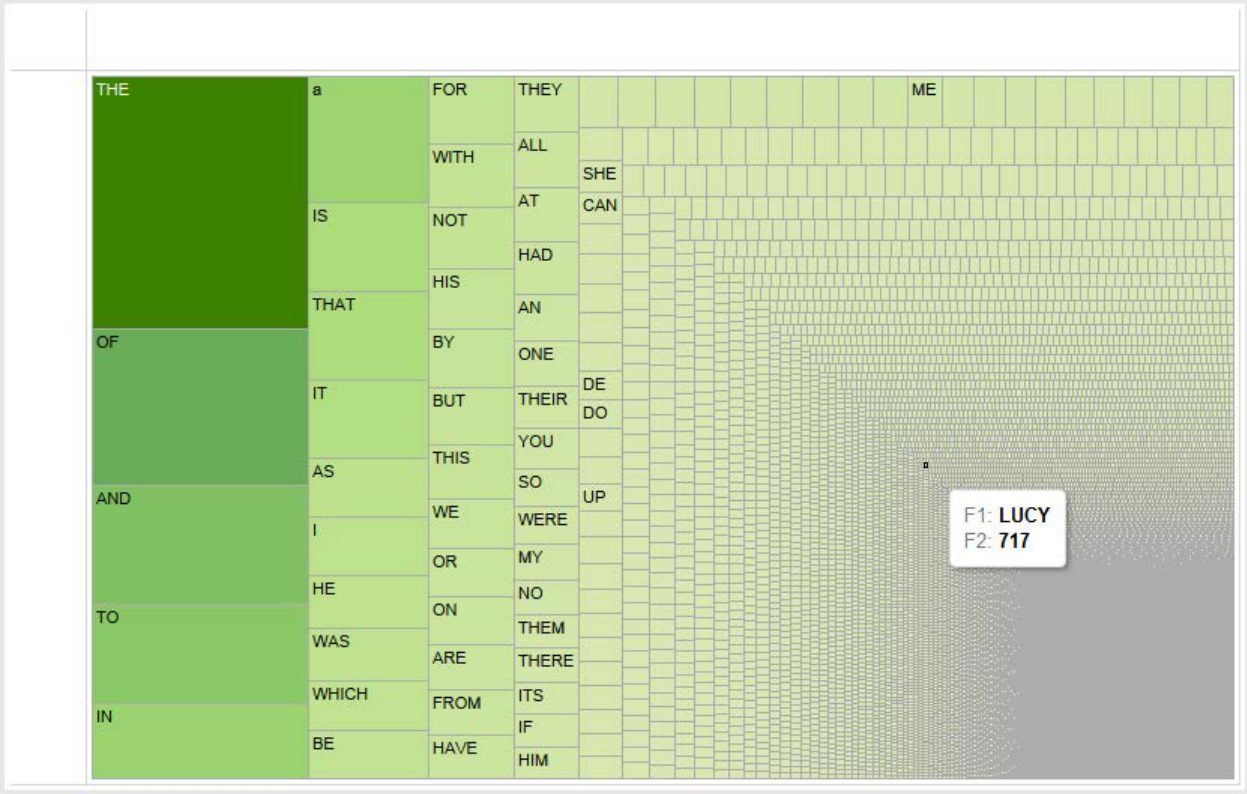


Fig. 4.3: Treemap

In the Teemap chart, the size of a block and its color intensity shows that word's count relative value, i.e., as the count of a word increases its block size becomes larger and color intensity becomes brighter. Word 'LUCY' is counted 717 times which is indicated with small black square box that is found hovering over in the area.

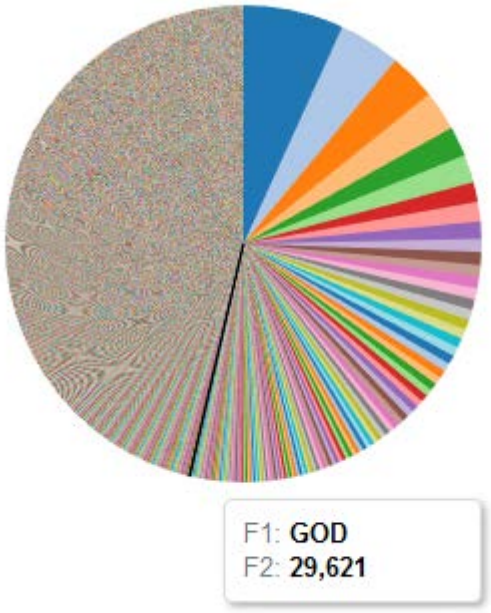
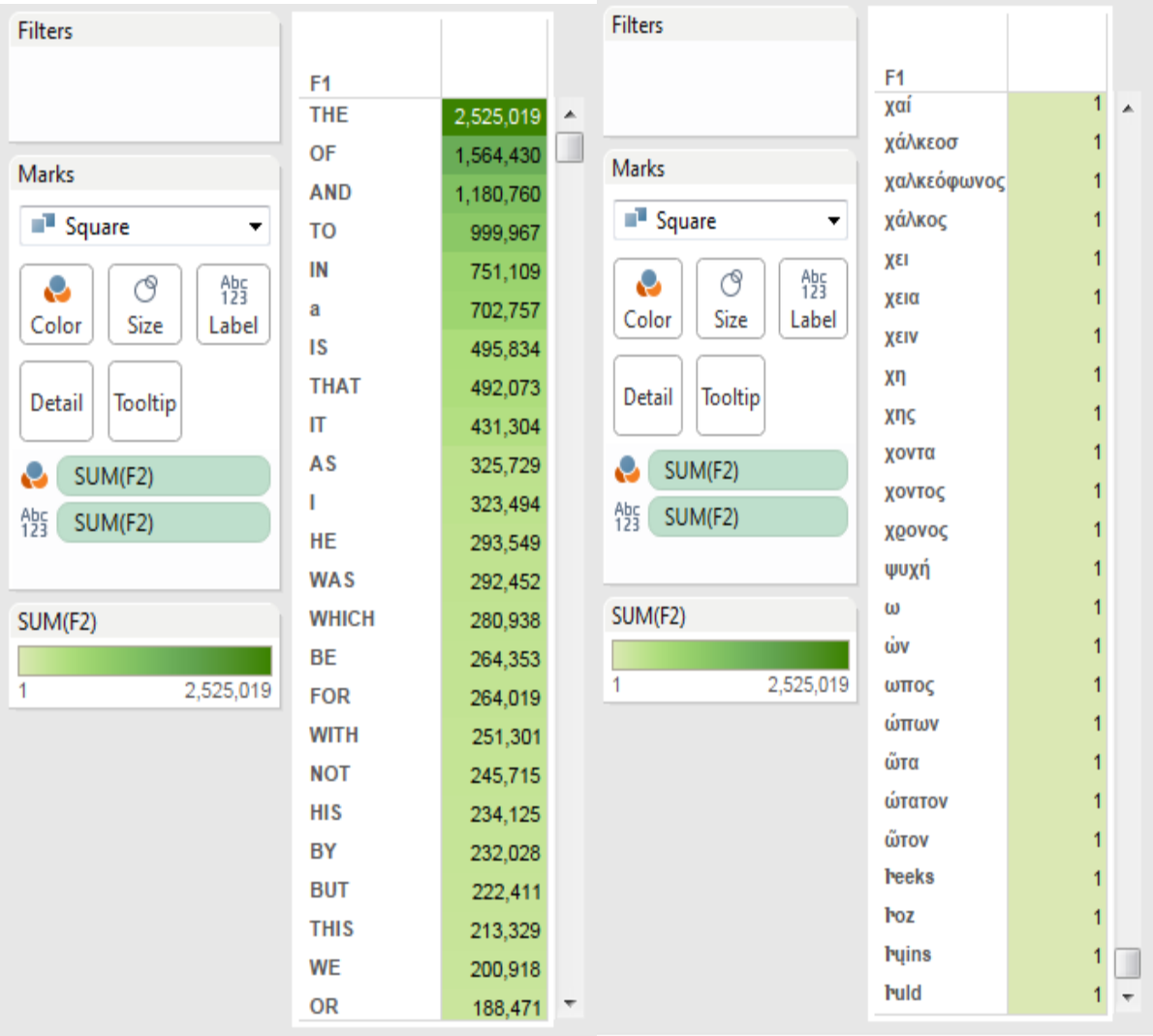


Fig. 4.4: Pie Chart

Similarly, Pie chart displays value of each word using different coloring schema composition to indicate relative value of a word. As shown word ‘GOD’ is counted 29,621 times as indicated by black strip from center of the circle to down circumference.



Graph a. high value words Graph b. low value words

Fig. 4.5: Highlight Table

In Highlight Table, values or counts of words along with color intensity but there is no distinct demarcation between two consecutive values. As shown in graph a & b, the value of words are exhaustively listed in descending order.

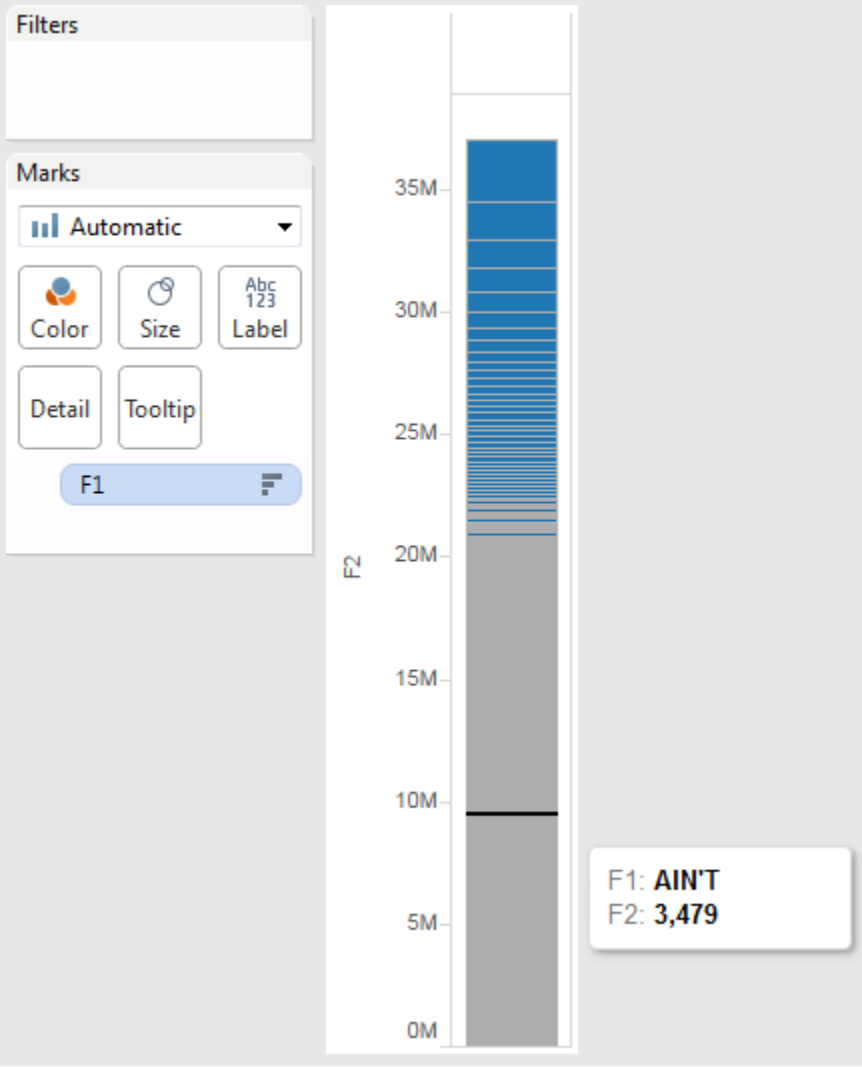


Fig. 4.6: Stacked Bar Chart

In Stacked Bar chart, there is differentiation of values of word count using color intensity and height of bars except its width. Only vertical axis is used to show size of word count and shade of size of the color indicates the word itself. For instance, the value of word 'AIN'T' is 3,479.

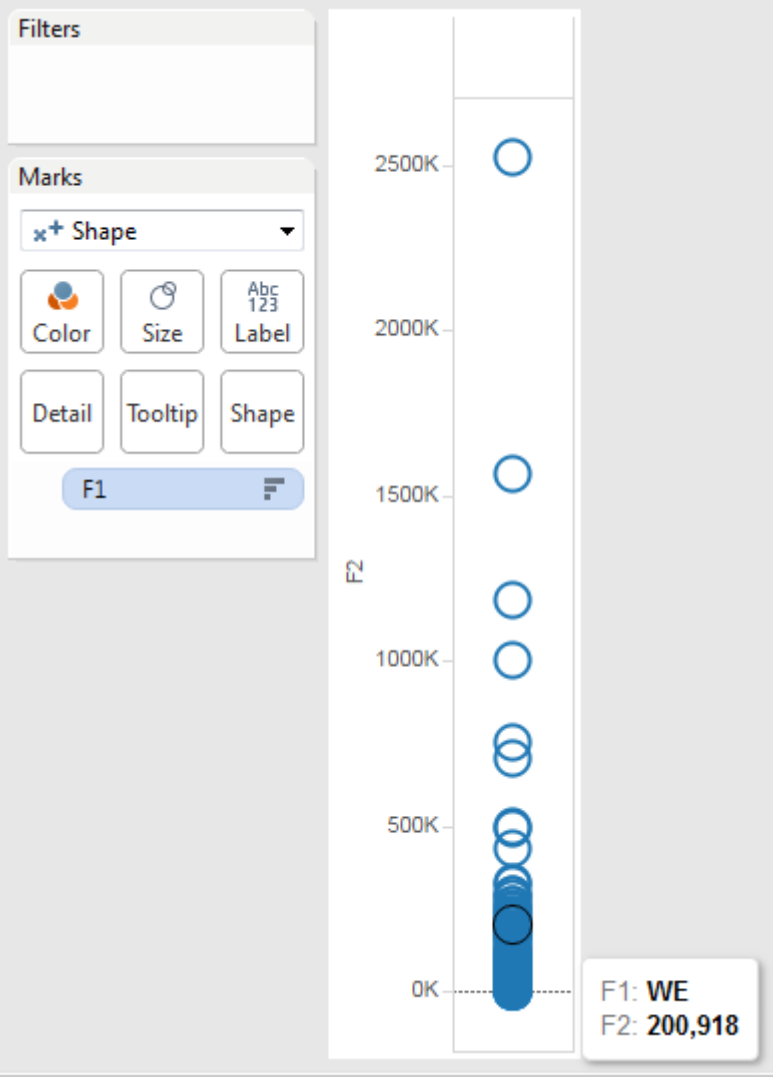


Fig. 4.7: Circle Views Chart

The Circle Views chart displays data elements in circles on vertical axis within its relative position; as shown word 'WE' is counted 200,918 times which falls in range of between 0K and 500K.

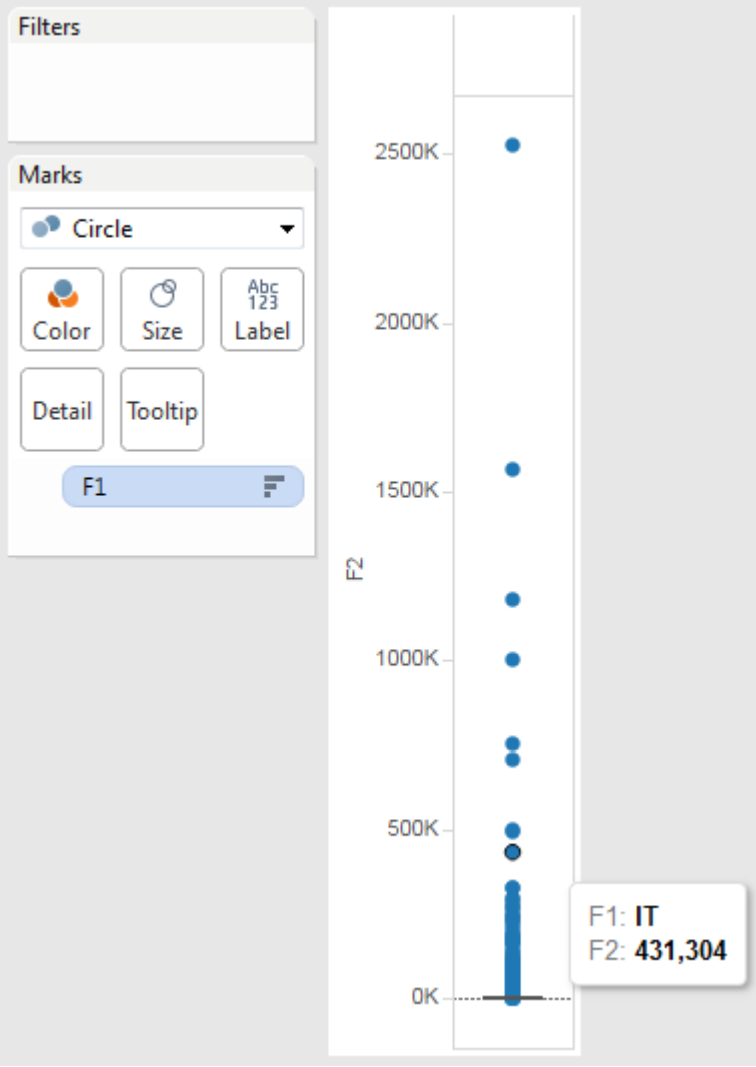
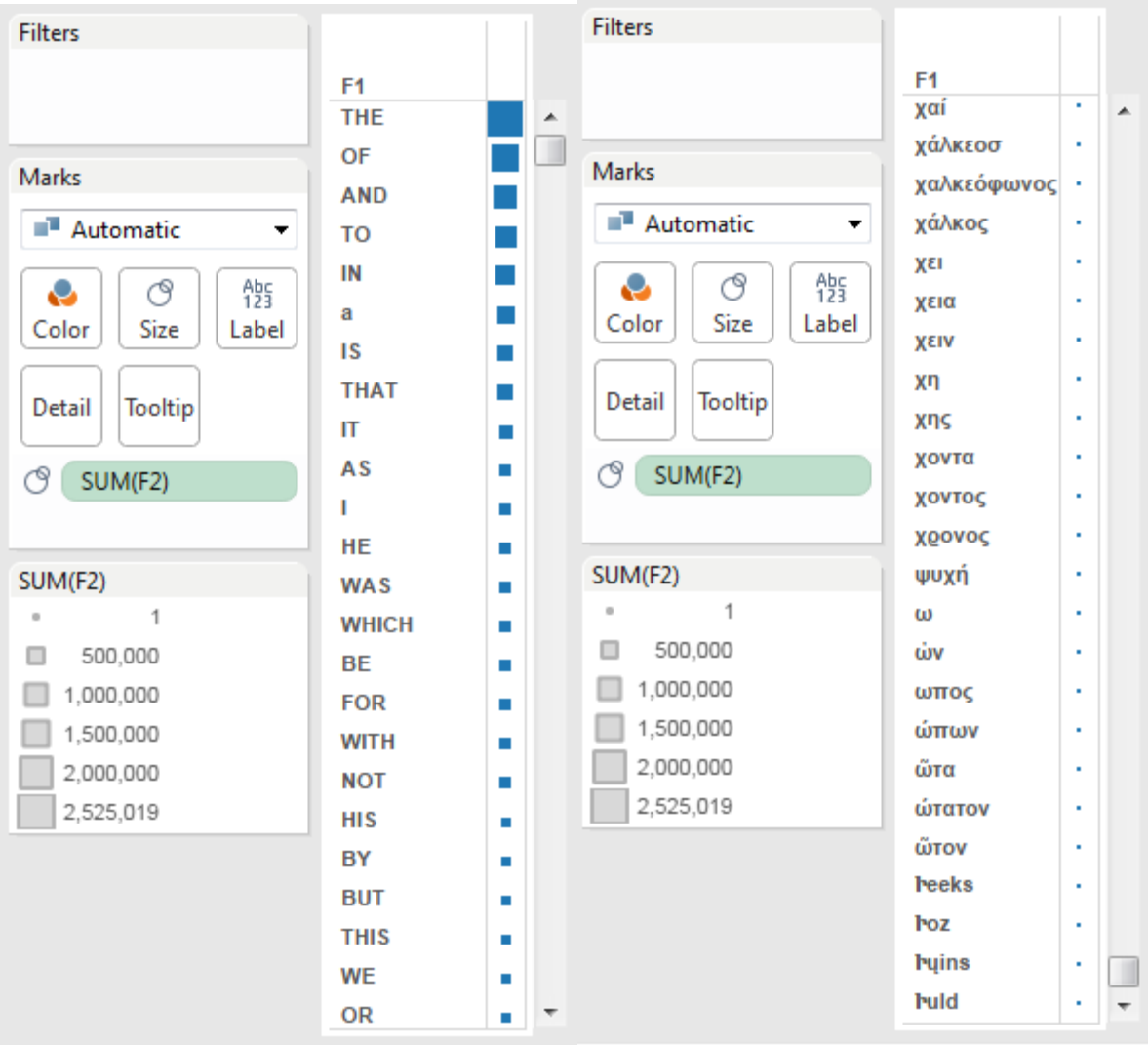


Fig. 4.9: Box-and-Whisker plot

Vertical axis is used to display value of word count but words are represented with color filled circles; for instance, word 'IT' is counted 431,304 times as highlighted with black circle in range of between 0K and 500K.



Graph a. high value words Graph b. low value words

Fig. 4.10: Heat Map

The value of a word in Heat Map is square in proportion to size of count, i.e., larger square shows bigger count of a word in opposite smaller square shows few word counts. As shown in Graph a & b, the size of a square is directly proportional to the value of word.

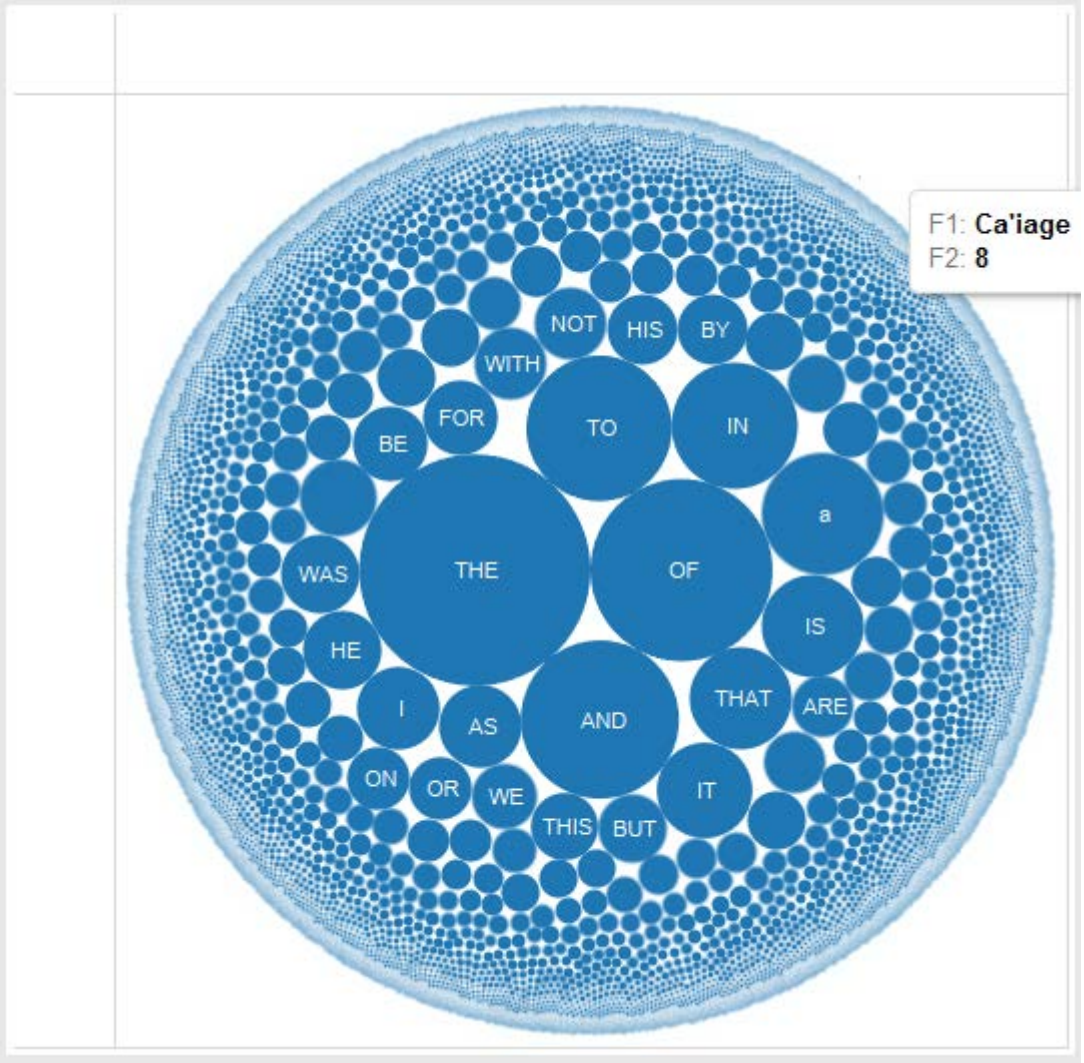


Fig. 4.11: Packed Bubbles Chart

This chart, Packed Bubbles, has the most fascinating way of presentation of word and its count in proportion to value or size of count. The bigger counts are placed in the center of the circle and few counts are scattered on circumference of the circle. For example, word 'Ca'iage' is counted 8 times and placed inside the circle near to circumference as indicated with black spot.

Chapter Five
Conclusion and Recommendation

5.1. Conclusion

Big data processing and visualization is a challenge that needs new way of tackling which is otherwise cannot be solved with current practice of data management because data deluge and data creation frequency in varieties of formats are inevitable scenarios. The approach that is employed in this study to undertake these challenges are reviewing problem areas in detail, followed by designing solution, then implementation of designed solution after that testing implemented solution using big data sets. As a result shows, Hadoop ecosystem provides platform to process unstructured data sets of Big Data in cheap, fault tolerant and high speed. The achievement of the study expounds next generation of IT in areas of data storage, processing and visualization. Especially, reliability and computational power does not need scale up in terms of hardware and processor capacities. Therefore, Big Data processing and visualization challenges are able to handle using software solutions rather than in placing specialized machines with increased hardware and processing capabilities.

This study has strong points to be raised for practical study in the area. These are data dimension and technological dimensions which indicates glimpse of light that sheds for upcoming challenges how to confront and extract insights from huge unstructured data sets. As we have seen in the study, it is possible to manage big data regardless of size and nature of data. However, full scale experimentation on all data types including multimedia have not been carried out in this study, due to time and resources constraints, which can be researched further. Apart these, the points that require further investigation and study are fully distributed environments or clustered machines to

exploit full potential by processing Terabytes and Petabytes of data sets of big data in general and its specific application for decision making by implementing revealed insights.

5.2. Recommendation

Big Data has enormous potential and benefits at every level of societies which can be considered as an eye opener to new discoveries and innovations. Now, it is not only possible to study populations as whole without looking for samples and its representativeness but also it becomes common to forecast or trend analysis of unimaginable situations. So, it is important to study further impacts of structured and semi-structured data sets by accommodating Velocity data along with power of parallelism computation in fully distributed setup.

As new area of study, it is recommended further studies in specific contexts to identify wealth of benefits and cautions. The main points to consider to dig deep wealth of Big Data are:

- Data management
- Data retrieval
- The need for well-equipped experimental lab with clusters of machines

References

[1] N. S. a. H. Shah, Big Data Application Architecture Q&A, 2013.

[2] S. O. B. B. a. L. H. Mike Kinkead, Big Data and Analytics, 2011.

[3] T. Chardonnens, "Big Data analytics on high velocity streams Specific use cases with Storm," 2013.

[4] J. Yan, Big Data, Bigger Opportunities; Data.gov, U.S. General Services Administration, 2013.

[5] B. Franks, TAMING THE BIG DATA TIDAL WAVE Finding Opportunities in Huge Data Streams with Advanced Analyltics, 2012.

[6] M. Ludloff, "A Big Data Showdown: How many V's do we really need? Three!," Insight Voices, 17 January 2013. [Online]. Available: http://blog.patternbuilders.com/2013/01/17/big-data-showdown-how-many-vs-do-we-really-need/. [Accessed 1 October 2014].

[7] Y. W. T.-S. C. a. X. L. Han Hu, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE,* 2014.

[8] P. Z. a. R. Kodali, Big Data Analytics Using Splunk, 2013.

[9] R. Akerkar, Big Data Computing; CRC Press, 2014.

[10] A. Z. Santovena, "Big Data: Evolution, Component, Challenges and Opportunities," 2013.

[11] G. D. F. Morales, "Big Data and theWeb: Algorithms for Data Intensive Scalable Computing," 2012.

[12] B. Verheij, "THE PROCESS OF BIG DATA SOLUTION ADOPTION An exploratory study within the Dutch telecome and energy utility sector," 2013.

[13] C. I. C. Study, "How Cisco IT Built Big Data Platform to Transform Data Management," *Cisco IT*, 2013.

[14] HARM GEERLINGS, "Big data whither the Enterprise Data Warehouse/Relational Data Management System?," OCEANBI, 31 July 2014. [Online]. Available:

http://www.oceanbi.com/big-data-whither-enterprise-data-warehouserelational-data-management-system/. [Accessed 25 March 2015].

[15] A. Collins, Big Data Doom, 2014.

[16] A. Lurie, "39 Data Visualization Tools for Big Data," ProfitBricks The IaaS - Company, 13 February 2014. [Online]. Available: https://blog.profitbricks.com/39-data-visualization-tools-for-big-data/. [Accessed 1 March 2016].

[17] P. B. a. R. B. Thomas H. Davenport, How 'Big Data' is Different; MITSloan Management Review, 2012.

[18] J. S.-S. a. I. Willson, Big Data: Big Opportunities to Create Business Value; Leadership Council for Information Advantage, 2011.

[19] S. P. Bappalige, "An introduction to Apache Hadoop for big data," IBM's Systems & Technology Group, 26 August 2014. [Online]. Available: https://opensource.com/life/14/8/intro-apache-hadoop-big-data. [Accessed 31 December 2014].

[20] K. Cukier, Interviewee, *Big data is better data.* [Interview]. 13 July 2015.

[21] Global Pulse White Paper, "“Big Data for Development: Opportunities & Challenges”," *Global Pulse* , 2012.

[22] D. Krishna, Big Data, 2010.

[23] An Oracle White Paper, "“Information Management and Big Data A Reference Architecture”," *An Oracle*, 2013.

[24] J. J. Berman, PRINCIPLES OF BIG DATA Preparing, Sharing, and Analyzing Complex Information, 2013.

[25] a. P. (. W. Ted Garcia, Analysis of Big Data Technologies and Methods - Query Large Web Public RDF Datasets on Amazon Cloud Using Hadoop and Open Source Parsers, 2013.

[26] M. C. a. R. d. V. Alfredo Cuzzocrea, "An Effective and Efficient MapReduce Algorithm for Computing BFS-Based Traversals of Large-Scale RDF Graphs," *Algorithms*, vol. 9, no. 1, 2015 .

[27] F. C. P. A. A. a. M. V. Elena Geanina ULARU, "Perspectives on Big Data and Big Data Analytics," *IEEE,* 2012.

[28] S. S. a. M. M. Gaber, "Large Scale and Big Data; Processing and Management," *IEEE,* 2014.

[29] M. H. W. A. H. A. L. D. S. A. a. M. A. C. Katarina Grolinger, "Challenges for MapReduce in Big Data," *2014.*

[30] M. S. a. J. V. Sameer Wadkar, Pro Apache Hadoop, 2014.

[31] J. D. a. S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Google,* 2004.

[32] "HDFS Architecture Guide," Apache, 08 April 2013. [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. [Accessed 25 January 2015].

[33] M. H. W. A. H. a. A. L. Katarina Grolinger, "Challenges for MapReduce in Big Data," *IEEE,* 2014.

[34] Tangient LLC, "MapReduce," Tangient LLC, 25 July 2011. [Online]. Available: https://hadooptutorial.wikispaces.com/MapReduce?responseToken=0752c09b2c2a07fc9544884e34a335629. [Accessed 2 March 2015].

[35] R. K. M. a. F. J. Carson Kai-Sang Leung, "Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data," *IEEE,* 2014.

[36] A. D. B. a. A. O'Driscoll, "A big data methodology for categorising technical support requests using Hadoop and Mahout," *Journal of Big Data*, 2014.

[37] C. White, "Big Data and Advanced Analytics Technologies and Use Cases," *BIresearch,* 2013.

[38] T. White, Hadoop The Definitive Guide, 2012.

[39] A. K. a. S. Lu, "A System Architecture for Running Big Data Workflows in the Cloud," *IEEE,* 2014.

[40] Media, O'Reilly, Big Data Now, 2012.

[41] D. A. Sathi, Big Data Analytics Disruptive Technologies for Changing the Game, 2012.

[42] A. N. D. F. H. a. M. K. Judith Hurwitz, Big Data FOR DUMMIES, 2013.

[43] D. Sarkar, Microsoft SQL Server 2012 with Hadoop Integrate data between Apache Hadoop and SQL Server 2012 and provide business intelligence on the heterogeneous data, 2013.

[44] T. A. Keahey, "Using visualization to understand big data," *IBM Software Business Analytics,* 2013.

[45] W. Paper, "Principles of Data Visualization - What We See in a Visual".

[46] "Big Data Visualization: Turning Big Data Into Big Insights," *Intel IT Center White Paper | Big Data Visualization,* 2013.

[47] N. MOUTHAAN, "EFFECTS OF BIG DATA ANALYTICS ON ORGANIZATIONS' VALUE CREATION," 2012.

[48] M. Maier, "Towards a Big Data Reference Architecture," 2013.

[49] Global Pulse, "Big Data for Development: Challenges & Opportunities," *Global Pulse,* 2012.

[50] R. Z. a. A. H. Seth C. Lewis, "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods," *Journal of Broadcasting & Electronic Media,* 2013.

[51] I. S. Rubinstein, "Big Data: The End of Privacy or a New Beginning?," *International Data Privacy Law Advance,* 2013.

[52] M. T. J. a. M. Nelson, "Moving ahead with Hadoop YARN," IBM, 02 July 2013. [Online]. Available: http://www.ibm.com/developerworks/library/bd-hadoopyarn/. [Accessed 25 January 2015].

[53] J. G. Wolff, "Bg Data and the SP Theory of Intelligence," *IEEE,* 2014.

[54] K. B. Carter, Actionable Intelligence A Guide to Delivering Business Results with Big Data Fast, 2014.

[55] P. Warden, Big Data Glossary, 2011.

Appendices

Data set size

Name	Symbol	Value
Kilobyte	KB	2^{10} = 1,024
Megabyte	MB	2^{20} = 1,048,576
Gigabyte	GB	2^{30} = 1,073,741,824
Terabyte	TB	2^{40} = 1,099,511,627,776
Petabyte	PB	2^{50} = 1,125,899,906,842,624
Exabyte	EB	2^{60} = 1,152,921,504,606,846,976
Zettabyte	ZB	2^{70} = 1,180,591,620,717,411,303,424
Yottabyte	YB	2^{80} = 1,208,925,819,614,629,174,706,176

Source Code

```
import java.io.IOException;

import java.util.*;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.conf.*;

import org.apache.hadoop.io.*;

import org.apache.hadoop.mapreduce.*;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class WordCount {

 public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);

    private Text word = new Text();

    public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {

        String line = value.toString();

        StringTokenizer tokenizer = new StringTokenizer(line);

        while (tokenizer.hasMoreTokens()) {

            word.set(tokenizer.nextToken());

            context.write(word, one);

        }

    }

 }
```

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
   public void reduce(Text key, Iterator<IntWritable> values, Context context)
    throws IOException, InterruptedException {
      int count = 0;
      while (values.hasNext()) {
         count += values.next().get();
      }
      context.write(key, new IntWritable(count));
   }
}

public static void main(String[] args) throws Exception {
   Configuration conf = new Configuration();
      Job job = new Job(conf, "wordcount");
   job.setOutputKeyClass(Text.class);
   job.setOutputValueClass(IntWritable.class);
   job.setMapperClass(Map.class);
   job.setReducerClass(Reduce.class);
   job.setInputFormatClass(TextInputFormat.class);
   job.setOutputFormatClass(TextOutputFormat.class);
   FileInputFormat.addInputPath(job, new Path(args[0]));
   FileOutputFormat.setOutputPath(job, new Path(args[1]));
   job.waitForCompletion(true);
}
}
```

List of books for Experimentation

The following are lists of books are used for experimentation. The books are composed of a number of categories such as literatures or language education, science fictions, Africa, war and philosophy.

No.	Author(s)	Title	Remark
1	Frederick Douglass	The Narrative of the Life of Frederick Douglass An American Slave	
2	René Descartes	A Discourse on Method	
3	Karl Marx and Friedrich Engels	The Communist Manifesto	
4	Norman Coombs	The Black Experience in America The Immigrant Heritage of America	
5	Edgar Rice Burroughs	Tarzan of the Apes	
6	Frederick Douglass	Collected Articles of Frederick Douglass	
7	Edgar Rice Burroughs	Jungle Tales of Tarzan	
8	Frederick Douglass	My Bondage and My Freedom	
9	Harriet Beecher Stowe	Uncle Tom's Cabin	
10	Various	The Martin Luther King, Jr. Day, 1995, Memorial Issue	
11	Lao-Tse	Tao Teh King	
12	Joseph Conrad	Heart of Darkness	
13	Charles Alexander Eastman	The Soul of the Indian An Interpretation	
14	W.E.B. Du Bois	The Souls of Black Folk	
15	Phillis Wheatley	Religious and Moral Poems	
16	P. J. PROUDHON	THE EVOLUTION OF CAPITALISM SYSTEM OF ECONOMICAL CONTRADICTIONS OR, THE PHILOSOPHY OF MISERY	
17	CHARLES W. CHESNUTT	THE HOUSE BEHIND THE CEDARS	
18	Harriet E. Wilson	Our nigor, Sketches from the Life of a Free Black, In A Two-Story White House, North	

19	JOHN BUCHAN	PRESTER JOHN	
20	William James	The Varieties of Religious Experience	
21	BERTRAND RUSSELL	PROPOSED ROADS TO FREEDOM	
22	H. Rider Haggard	ALLAN QUATERMAIN	
23	John Dewey	DEMOCRACY AND EDUCATION	
24	Owen Wister	PHILOSOPHY 4	
25	Epictetus	THE GOLDEN SAYINGS OF EPICTETUS	
26	Benedict de Spinoza	Benedict de Spinoza, THE ETHICS I	
27	Benedict de Spinoza	Benedict de Spinoza, THE ETHICS II	
28	Benedict de Spinoza	Benedict de Spinoza, THE ETHICS III	
29	Benedict de Spinoza	Benedict de Spinoza, THE ETHICS IV	
30	Benedict de Spinoza	Benedict de Spinoza, THE ETHICS V	
31	Baruch Spinoza	A Theologico-Political Treatise Part I	
32	Baruch Spinoza	A Theologico-Political Treatise Part II	
33	Baruch Spinoza	On the Improvement of the Understanding (Treatise on the Emendation of the Intellect)	
34	David Livingstone	MISSIONARY TRAVELS AND RESEARCHES IN SOUTH AFRICA	
35	Xenophon	THE SYMPOSIUM	
36	H. Rider Haggard	NADA THE LILY	
37	Edouard le Roy	A NEW PHILOSOPHY: HENRI BERGSON	
38	C. F. VOLNEY	THE RUINS, OR, MEDITATION ON THE REVOLUTIONS OF EMPIRES: AND THE LAW OF NATURE	
39	Olive Schreiner	THE STORY OF AN AFRICAN FARM	
40	Solomon Tshekisho Plaatje	Native Life in South Africa, Before and Since the European War and the Boer Rebellion	
41	John Fiske	THE UNSEEN WORLD AND OTHER ESSAYS	
42	Plato	PROTAGORAS	
43	Plato	EUTHYDEMUS	

44	Plato	SYMPOSIUM	
45	Plato	CRATYLUS	
46	Plato	EUTHYPHRO	
47	Plato	APOLOGY	
48	Plato	CRITO	
49	Plato	LESSER HIPPIAS	
50	Olive Gilbert	NARRATIVE OF SOJOURNER TRUTH	
51	a Platonic Imitator	ERYXIAS	
52	Plato	MENEXENUS	
53	Plato	PARMENIDES	
54	H. RIDER HAGGARD	CHILD OF STORM	
55	H. RIDER HAGGARD	FINISHED	
56	Plato	THEAETETUS	
57	Plato	SOPHIST	
58	Aristotle	THE POETICS OF ARISTOTLE	
59	Various	STORIES BY ENGLISH AUTHORS AFRICA	
60	Friedrich Nietzsche	THUS SPAKE ZARATHUSTRA A BOOK FOR ALL AND NONE	
61	William Wells Brown	Clotel; or, The President's Daughter	
62	O. A. BROWNSON	THE AMERICAN REPUBLIC: ITS CONSTITUTION, TENDENCIES, AND DESTINY	
63	THOMAS LOVE PEACOCK	CROTCHET CASTLE	
64	The Raven	THE WORKS OF EDGAR ALLAN POE	
65	H. RIDER HAGGARD	KING SOLOMON'S MINES	
66	Joel Chandler Harris	Uncle Remus: His Songs and His Sayings	
67	Booker T. Washington	UP FROM SLAVERY: AN AUTOBIOGRAPHY	
68	Aristotle	The Categories	
69	Voltaire	LETTERS ON ENGLAND	

70	ERSTER TEIL	SIDDHARTHA Eine indische Dichtung von Hermann Hesse	
71	Bertrand Russell	THE ANALYSIS OF MIND	
72	H. Rider Haggard	MAIWA'S REVENGE OR THE WAR OF THE LITTLE HAND	
73	H. Rider Haggard	ALLAN'S WIFE	
74	H. Rider Haggard	HUNTER QUATERMAIN'S STORY	
75	John Galsworthy	ESSAYS ON CENSORSHIP AND ART	
76	Richard Harding Davis	Notes of a War Correspondent	
77	Plutarch	Essays and Miscellanies The Complete Works Volume 3	
78	Arthur Conan Doyle	The Great Boer War	
79	H. Rider Haggard	She	
80	Samuel White Baker	In the Heart of Africa	
81	John Hanning Speke	The Discovery of the Source of the Nile	
82	Raphael Sabatini	The Sea-Hawk	
83	George Bernard Shaw	Man And Superman	
84	Jules Verne	Five Weeks in a Balloon Journeys and Discoveries in Africa by Three Englishmen	
85	Abraham Cowley	Essays	
86	Michel de Montaigne	The Essays of Montaigne, Volume 3	
87	Count Lyof N. Tolstoi	On the Significance of Science and Art from What to Do?	
88	Thomas Paine	The Writings of Thomas Paine, Volume IV. 1794-1796.	
89	Victor Appleton	Tom Swift and his Electric Rifle	
90	Benedict de Spinoza	The Ethics	

91	J. H. Patterson	The Man-eaters of Tsavo and Other East African Adventures	
92	Walter Horatio Pater	Plato and Platonism	
93	Henry Lindlahr	Nature Cure	
94	Immanuel Kant	The Critique of Pure Reason	
95	David Hume	An Enquiry Concerning the Principles of Morals	
96	Henri Bergson	Laughter: An Essay on the Meaning of the Comic	
97	Rene Descartes	The Principles of Philosophy	
98	J. B. Bury	The Idea of Progress An Inquiry Into Its Origin And Growth	
99	David Hume	Dialogues Concerning Natural Religion	
100	David Hume	A Treatise of Human Nature	
101	Grant Allen	An African Millionaire Episodes in the Life of the Illustrious Colonel Clay	
102	George Berkeley	A Treatise Concerning the Principles of Human Knowledge	
103	George Berkeley	Three Dialogues between Hylas and Philonous in Opposition to Sceptics and Atheists	
104	David Starr Jordan	The Philosophy of Despair	
105	Lewis Carroll	The Game of Logic	
106	William James	Pragmatism A New Name for Some Old Ways of Thinking	
107	Henry M. Stanley	How I Found Livingstone	
108	H. Rider Haggard	Allan and the Holy Flower	
109	J. Alexander Gunn	Modern French Philosophy: A Study Of The Development Since Comte	
110	Dom	PoPHILO	
111	Francis Bacon	The Advancement of Learning	
112	Max Pearson Cushing	Baron d'Holbach A Study of Eighteenth Century Radicalism in France	
113	James E. Talmage	The Story of "Mormonism"	
114	Various	Literary and Philosophical Essays	
115	Friedrich Nietzsche	Thoughts out of Season (Part One)	

116	John Stuart Mill	Considerations on Representative Government	
117	Immanuel Kant	The Critique of Practical Reason	
118	Immanuel Kant	The Metaphysical Elements of Ethics	
119	W.E. Burghardt Du Bois	The Conservation of Races	
120	William Osmer	A Dissertation on Horses	
121	J. Alexander Gunn	Bergson and His Philosophy	
122	H. Rider Haggard	She and Allan	
123	Bertrand Russell	The Problems of Philosophy	
124	Herbert Spencer	The Philosophy of Style	
125	Mary H. Kingsley	Travels in West Africa	
126	Dewitt H. Parker	The Principles Of Aesthetics	
127	St. George Stock	Deductive Logic	
128	Aristotle	Politics A Treatise on Government	
129	Aristotle	The Poetics	
130	Thomas Wentworth Higginson	Army Life in a Black Regiment	
131	Frederich Schiller	The Philosophical Letters	
132	Frederich Schiller	The Project Gutenberg Works of Frederich Schiller in English	
133	Rabindranath Tagore	Sadhana The Realisation of Life	
135	Richard F. Burton	First Footsteps in East Africa or, an Exploration of Harar	
136	Purnananda Chakravartin	The Tattva-Muktavali	
137	Emilie Kip Baker	Short Stories and Selections for Use in the Secondary Schools	
138	Swami Abhedananda	Reincarnation	
139	St. George Stock	A Little Book of Stoicism	
140	S. M. Dubnow	Jewish History	

141	Finley Peter Dunne	Mr. Dooley's Philosophy	
142	Selected and arranged by William Patten	The Junior Classics Volume 8 Animal and Nature Stories	
143	Jacob Feis	Shakspere And Montaigne	
144	Delia Bacon	The Philosophy of the Plays of Shakspere Unfolded	
145	P. H. Ditchfield	Books Fatal to Their Authors	
146	Paul Henri Thiery (Baron D'Holbach)	The System of Nature, Volume 1	
147	Emile Faguet	Initiation into Philosophy	
148	Benedetto Croce	Aesthetic as Science of Expression and General Linguistic	
149	Various	Stories Worth Rereading	
150	Charles Francis Adams	Tis Sixty Years Since	
151	Sarah H. Bradford	Harriet, The Moses of Her People	
152	George Tucker	A Voyage to the Moon	
153	James M. Beck	The Constitution of the United States A Brief Study of the Genesis, Formulation and Political Philosophy of the Constitution	
154	Thomas Taylor	Introduction to the Philosophy and Writings of Plato	
155	John Stuart Mill	Autobiography	
156	Thomas Clarkson	An Essay on the Slavery and Commerce of the Human Species, Particularly the African	
157	Thomas Clarkson	The History of the Rise, Progress and Accomplishment of the Abolition of the African Slave-Trade, by the British Parliament (1839)	
158	Epictetus	A Selection from the Discourses of Epictetus With the Encheiridion	
159	Robinson and Overton	Life, Letters, and Epicurean Philosophy of Ninon de L'Enclos, the Celebrated Beauty of the Seventeenth Century	
160	Alpha of the Plough (Alfred	Pebbles on the Shore	

	George Gardiner)		
161	Arthur Schopenhauer	The Art of Literature	
162	Thomas E. Willson	Ancient and Modern Physics	
163	Arthur Schopenhauer	The Essays of Arthur Schopenhauer; Religion, A Dialogue, Etc.	
164	Edward Howard Griggs	The Soul of Democracy The Philosophy Of The World War In Relation To Human Liberty	
165	Charles Waddell Chesnutt	Frederick Douglass A Biography	
166	James Weldon Johnson	The Autobiography of an Ex-Colored Man	
167	Harriet Jacobs (AKA Linda Brent)	Incidents in the Life of a Slave Girl Written by Herself	
168	Charles Waddell Chesnutt	The Wife of his Youth and Other Stories of the Color Line, and Selected Essays	
169	Richard Falckenberg	History Of Modern Philosophy From Nicolas of Cusa to the Present Time	
170	Edward A. Johnson	History of Negro Soldiers in the Spanish-American War, and Other Items of Interest	
171	E.W. Hornung	No Hero	
172	Harriet Beecher Stowe	Uncle Tom's Cabin, Young Folks' Edition	
173	John Stuart Mill	Utilitarianism	
174	Edited by Mrs. M. H. Adams	Small Means and Great Ends	
175	Rev. W. Lucas Collins	Cicero Ancient Classics for English Readers	
176	Charles W. Chesnutt	The Conjure Woman	
177	George Willis Cooke	George Eliot; A Critical Study of Her Life, Writings & Philosophy	
178	Work Projects Administration	Slave Narratives: Arkansas Narratives Arkansas Narratives, Part 6	
179	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves Kentucky Narratives	

180	Albert G. Mackey	The Symbolism of Freemasonry	
181	Arthur Schopenhauer	Essays of Schopenhauer	
182	William James	A Pluralistic Universe Hibbert Lectures at Manchester College on the Present Situation in Philosophy	
183	Edited by James Weldon Johnson	The Book of American Negro Poetry	
184	George Grote	Review of the Work of Mr John Stuart Mill Entitled, 'Examination of Sir William Hamilton's Philosophy.'	
185	John Stuart Mill	Essays on some unsettled Questions of Political Economy	
186	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves Mississippi Narratives	
187	Inazo Nitobe	Bushido, the Soul of Japan	
188	Benjamin Brawley	A Social History of the American Negro Being a History of the Negro Problem in the United States. Including A History And Study Of The Republic Of Liberia	
189	Arthur Christopher Benson	Father Payne	
190	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves Florida Narratives	
191	Various	The German Classics of The Nineteenth and Twentieth Centuries, Vol. VII. Masterpieces of German Literature Translated into English. In Twenty Volumes	
192	Frances E.W. Harper	Iola Leroy Shadows Uplifted	
193	Dante Alighieri	The Banquet (Il Convito)	
194	George Robert Stow Mead	Simon Magus	
195	J. G. Holland	The Mistress of the Manse	
196	Frances Reynolds	An Enquiry Concerning the Principles of Taste, and of the Origin of our Ideas of Beauty, etc.	

197	Frank R. Stockton	Amos Kilbright; His Adscititious Experiences	
198	Henry Jones	Browning as a Philosophical and Religious Teacher	
199	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves: Indiana Narratives	
200	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves: Volume IV, Georgia Narratives, Part 1	
201	Various	The Worlds Greatest Books, Volume XIII. Religion and Philosophy	
202	Various	The Journal of Negro History, Vol. I. Jan. 1916	
203	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves: Volume II, Arkansas Narratives, Part 2	
204	Arachne	Cobwebs of Thought	
205	Various	The Talking Beasts	
206	Richard William Church	Bacon English Men Of Letters, Edited By John Morley	
207	Various	Readings on Fascism and National Socialism	
208	Three Initiates	The Kybalion A Study of The Hermetic Philosophy of Ancient Egypt and Greece	
209	Boethius	The Consolation of Philosophy	
210	Abel J. Jones	Rudolph Eucken	
211	Winston Spencer Churchill	London to Ladysmith via Pretoria	
212	John Frederick Helvetius	The Golden Calf, Which the World Adores, and Desires	
213	Marcus Tullius Cicero	Academica	
214	Ida B. Wells-Barnett	Southern Horrors Lynch Law in All Its Phases	
215	Ida B. Wells-Barnett	Mob Rule in New Orleans Robert Charles and His Fight to Death, the Story of His Life, Burning Human Beings Alive, Other Lynching Statistics	
216	Ida B. Wells-Barnett	The Red Record Tabulated Statistics and Alleged Causes of Lynching in the United States	
217	Various	The Unity of Civilization	

218	Booker T. Washington, et al.	The Negro Problem	
219	William Wells Brown	The Narrative of William W. Brown, a Fugitive Slave	
220	W. E. B. Du Bois	Darkwater Voices From Within The Veil	
221	William Still	The Underground Railroad A Record Of Facts, Authentic Narratives, Letters, &C., Narrating The Hardships, Hair-Breadth Escapes And Death Struggles Of The Slaves In Their Efforts For Freedom, As Related By Themselves And Others, Or Witnessed By The Author.	
222	W. E. B. Du Bois	The Quest of the Silver Fleece A Novel	
223	Herbert Spencer, Henry Fawcett, Frederic Harrison and Other Distinguished Authors	John Stuart Mill; His Life and Works	
224	W.E.B. Du Bois	The Negro	
225	Henry Bibb	Narrative of the Life and Adventures of Henry Bibb, an American Slave, Written by Himself	
226	Olaudah Equiano	The Interesting Narrative of the Life of Olaudah Equiano, Or Gustavus Vassa, The African Written By Himself	
227	Henry Rogers	Reason and Faith; Their Claims and Conflicts From The Edinburgh Review, October 1849, Volume 90, No. CLXXXII	
228	Pierre Besnier	A Philosophicall Essay for the Reunion of the Languages Or, The Art of Knowing All by the Mastery of One	
229	Gale and Polden, Limited	A Handbook of the Boer War	
230	George W. Williams	History of the Negro Race in America From 1619 to 1880. Vol 1 Negroes as Slaves, as Soldiers, and as Citizens	

231	Edward Moore	Edward Caldwell Moore Outline of the History of Christian Thought Since Kant	
232	George Stuart Fullerton	An Introduction to Philosophy	
233	William Playfair	An Inquiry into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations. Designed To Shew How The Prosperity Of The British Empire May Be Prolonged	
234	W. Allison Sweeney	History of the American Negro in the Great World War His Splendid Record in the Battle Zones of Europe; Including a Resume of His Past Services to his Country in the Wars of the Revolution, of 1812, the War of Rebellion, the Indian Wars on the Frontier, the Spanish-American War, and the Late Imbroglio With Mexico	
235	George Santayana	Some Turns of Thought in Modern Philosophy Five Essays	
236	Thomas Henry Huxley	Lay Sermons, Addresses and Reviews	
237	Ibn Tufail	The Improvement of Human Reason Exhibited in the Life of Hai Ebn Yokdhan	
238	John-Stuart Mill	Auguste Comte and Positivism	
239	W. Tudor Jones	An Interpretation of Rudolf Eucken's Philosophy	
240	Charles Sotheran	Percy Bysshe Shelley as a Philosopher and Reformer	
241	David Livingstone	The Last Journals of David Livingstone, in Central Africa, from 1865 to His Death, Volume II (of 2), 1869-1873 Continued By A Narrative Of His Last Moments And Sufferings, Obtained From His Faithful Servants Chuma And Susi	
242	Laurence Oliphant	Fashionable Philosophy and Other Sketches	
243	Martin R. Delany	The Condition, Elevation, Emigration, and Destiny of the Colored People of the United States	
245	Lady Mary Wortley Montague	Letters of the Right Honourable Lady M--y W--y M--e Written during Her Travels in Europe, Asia	

		and Africa to Persons of Distinction, Men of Letters, &c. in Different Parts of Europe	
246	Mary Mills Patrick	Sextus Empiricus and Greek Scepticism	
247	W. E. B. Du Bois	The Suppression of the African Slave Trade to the United States of America 1638-1870	
248	George Santayana	Winds Of Doctrine Studies in Contemporary Opinion	
249	Tito Vignoli	Myth and Science An Essay	
250	Paul Laurence Dunbar	The Sport of the Gods	
251	Myrtle Reed	The Spinster Book	
252	Friedrich Nietzsche	Homer and Classical Philology	
253	Blaise Pascal	Pascal's Pensees	
254	James Froude	Froude's Essays in Literature and History With Introduction by Hilaire Belloc	
255	Paul Laurence Dunbar	The Complete Poems of Paul Laurence Dunbar	
256	Carveth Read	Logic Deductive and Inductive	
257	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves Georgia Narratives, Part 4	
258	J. Edward Mercer	Nature Mysticism	
259	Voltaire	Voltaire's Philosophical Dictionary	
260	Sir Sidney Lee	Shakespeare and the Modern Stage with Other Essays	
261	Grant Allen	Post-Prandial Philosophy	
262	Alfred North Whitehead	The Concept of Nature The Tarner Lectures Delivered in Trinity College, November 1919	
263	Edmund Beecher Wilson	Biology A lecture delivered at Columbia University in the series on Science, Philosophy and Art November 20, 1907	
264	Various	Slave Narratives Vol. XIV. South Carolina, Part 1 A Folk History of Slavery in the United States From Interviews with Former Slaves.	
265	Zora Hurston and Langston Hughes	The Mule-Bone: A Comedy of Negro Life in Three Acts	

266	Work Projects Administration	Slave Narratives: a Folk History of Slavery in the United States From Interviews with Former Slaves Arkansas Narratives Part 3	
267	Giordano Bruno	The Heroic Enthusiasts,(1 of 2) (Gli Eroici Furori) An Ethical Poem	
268	Giordano Bruno	The Heroic Enthusiast, Part II (Gli Eroici Furori) An Ethical Poem	
269	James J. Walsh	Old-Time Makers of Medicine The Story of The Students And Teachers of the Sciences Related to Medicine During the Middle Ages	
270	John Marshall	A Short History of Greek Philosophy	
271	Morris J. MacGregor Jr.	Integration of the Armed Forces, 1940-1965	
272	Various	The Journal of Negro History, Volume 2, 1917	
273	James Anthony Froude	Short Studies on Great Subjects	
274	G.E. Partridge	The Psychology of Nations A Contribution to the Philosophy of History	
275	Matthew A. Henson	A Negro Explorer at the North Pole	
276	John T. McCutcheon	In Africa Hunting Adventures in the Big Game Country	
277	William Hone	The Queen's Matrimonial Ladder A National Toy, With Fourteen Step Scenes; and Illustrations in Verse, With Eighteen other Cuts	
278	John Cowper Powys	The Complex Vision	
279	Hastings Rashdall	Philosophy and Religion Six Lectures Delivered at Cambridge	
280	John Dee	The Mathematicall Praeface to Elements of Geometrie of Euclid of Megara	
281	George W. Clark	The Liberty Minstrel	
282	Rudolf Schmid	The Theories of Darwin and Their Relation to Philosophy, Religion, and Morality	
283	Joel Chandler Harris	Uncle Remus and Brer Rabbit	
284	John Abercrombie	The Philosophy of the Moral Feelings	

285	Ernst Haeckel, J. Arthur Thomson and August Weismann	Evolution in Modern Thought	
286	Arthur Shearly Cripps	Cinderella in the South Twenty-Five South African Tales	
287	Various	Slave Narratives: a Folk History of Slavery in the United States From Interviews with Former Slaves, North Carolina Narratives, Part 1	
288	Rabindranath Tagore	Creative Unity	
289	Alexander Philip	Essays Towards a Theory of Knowledge	
290	Confucius	The Sayings Of Confucius	
291	Friedrich Max Mueller	The Silesian Horseherd - Questions of the Hour	
292	Joel Chandler Harris	Nights With Uncle Remus	
293	Various	The Wit and Humor of America, Volume IX (of X)	
294	Henry A. Beers	Four Americans Roosevelt, Hawthorne, Emerson, Whitman	
295	William George Hooper	Aether and Gravitation	
296	Ernest Dunlop Swinton	The Defence of Duffer's Drift	
297	Elizabeth Keckley	Behind the Scenes or, Thirty years a slave, and Four Years in the White House	
298	Various	The World's Greatest Books--Volume 14--Philosophy and Economics	
299	Friedrich Nietzsche.	The Case Of Wagner, Nietzsche Contra Wagner, and Selected Aphorisms.	
300	Josephine Paterson and Loretta Zderad	Humanistic Nursing	
301	Ralph Barton Perry	The Approach to Philosophy	
302	John Dewey	Moral Principles in Education	
303	Jacob Abbott	Rollo's Philosophy. [Air]	

304	Bertrand Russell	Mysticism and Logic and Other Essays	
305	Various	The World's Greatest Books - Volume 15 – Science	
306	John M. Robertson	Montaigne and Shakspere	
307	Frank R. Stockton	A Chosen Few Short Stories	
308	Leslie Stephen	The English Utilitarians, Volume II (of 3)	
309	Edgar Wallace	The Keepers of the King's Peace	
310	Andrew T. Still	Philosophy of Osteopathy	
311	John Stuart Mill	A System Of Logic, Ratiocinative And Inductive (Vol. 1 of 2)	
312	Booker T. Washington	The Future of the American Negro	
313	William Somerset Maugham	The Hero	
314	Thomas W. Talley	Negro Folk Rhymes Wise and Otherwise: With a Study	
315	W. Somerset Maugham	The Explorer	
316	Andrew Preston Peabody	A Manual of Moral Philosophy	
317	Benjamin Franklin Cocker	Christianity and Greek Philosophy or, the relation between spontaneous and reflective thought in Greece and the positive teaching of Christ and His Apostles	
318	Leslie Stephen	The English Utilitarians, Volume I.	
319	Frederick Engels	Feuerbach: The roots of the socialist philosophy	
320	Isaac Husik	A History of Mediaeval Jewish Philosophy	
321	John Stuart Mill	A System Of Logic, Ratiocinative And Inductive	
322	Francis J. (Francis Joseph) Reynolds, Allen L. (Allen Leon) Churchill, and Francis Trevelyan Miller	The Story of the Great War, Volume V (of 8) Battle of Jutland Bank; Russian Offensive; Kut-El-Amara; East Africa; Verdun; The Great Somme Drive; United States and Belligerents; Summary of Two Years' War	
323	Augustin Calmet	The Phantom World or, The philosophy of spirits, apparitions, &c, &c.	

324	Ralph Waldo Emerson	Nature	
325	Herbert Spencer	Essays: Scientific, Political, & Speculative, Vol. I	
326	Daniel G. Brinton	The Religious Sentiment Its Source and Aim: A Contribution to the Science and Philosophy of Religion	
327	Nathaniel Sands	The Philosophy of Teaching The Teacher, The Pupil, The School	
328	Henry More	Democritus Platonissans	
329	Work Projects Administration	Slave Narratives: a Folk History of Slavery in the United States From Interviews with Former Slaves Texas Narratives, Part 1	
330	Arthur W. Robinson	God and the World A Survey of Thought	
331	Stephen H. Carpenter	The Philosophy of Evolution and The Metaphysical Basis of Science	
332	John S. Hart	In the School-Room Chapters in the Philosophy of Education	
333	Baruch de Spinoza	The Philosophy of Spinoza	
334	Work Projects Administration	Slave Narratives: a Folk History of Slavery in the United States From Interviews with Former Slaves, North Carolina Narratives, Part 2	
335	W. E. Burghardt Du Bois	The Conservation of Races The American Negro Academy. Occasional Papers No. 2	
336	John Emerich Edward Dalberg-Acton	The History of Freedom	
337	Various	The Upward Path A Reader For Colored Children	
338	William Minto	Logic, Inductive and Deductive	
339	Ferruccio Busoni	Sketch of a New Esthetic of Music	
340	W. James King	The Natural Philosophy of William Gilbert and His Predecessors	
341	Roscoe Pound	An Introduction to the Philosophy of Law	
342	William James	Essays in Radical Empiricism	
343	Antonio Labriola	Essays on the Materialistic Conception of History	
345	William Butler Yeats	Ideas of Good and Evil	

346	G.A. Henty	The Young Colonists A Story of the Zulu and Boer Wars	
347	W. T. Stace	A Critical History of Greek Philosophy	
348	Thomas Belden Butler	The Philosophy of the Weather And a Guide to Its Changes	
349	Andrew Lang	The Grey Fairy Book	
350	John Dewey, Addison W. Moore, Harold Chapman Brown, George H. Mead, Boyd H. Bode, Henry Waldgrave, Stuart James, Hayden Tufts, Horace M. Kallen	Creative Intelligence Essays in the Pragmatic Attitude	
351	Epiphanius Wilson	The Wisdom of Confucius with Critical and Biographical Sketches	
352	Madhava Acharya	The Sarva-Darsana-Samgraha Review of the Different Systems of Hindu Philosophy	
353	Alfred H. Lloyd	The Will to Doubt An essay in philosophy for the general thinker	
354	Alfred William Benn	History of Modern Philosophy	
355	John A. Widtsoe	Joseph Smith as Scientist A Contribution to Mormon Philosophy	
356	George Santayana	Character and Opinion in the United States	
357	Ray Stannard Baker	Following the Color Line an account of Negro citizenship in the American democracy	
358	John Stuart Mill	On Liberty	
359	Charles Sumner	White Slavery in the Barbary States	
360	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves: Volume X, Missouri Narratives	
361	Work Projects Administration	Slave Narratives: a Folk History of Slavery in the United States From Interviews with Former Slaves: Volume XVI, Texas Narratives, Part 4	

362	John Stuart Mill	A System of Logic: Ratiocinative and Inductive 7th Edition, Vol. I	
363	John Stuart Mill	A System of Logic: Ratiocinative and Inductive 7th Edition, Vol. II	
364	George Robert Stowe Mead	Apollonius of Tyana, the Philosopher-Reformer of the First Century A.D.	
365	Edgar Wallace	Sanders of the River	
366	George Santayana	Three Philosophical Poets Lucretius, Dante, and Goethe	
367	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 1 (of 10) From "The Works of Voltaire - A Contemporary Version"	
368	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 2 (of 10) From "The Works of Voltaire - A Contemporary Version"	
369	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 3 (of 10) From "The Works of Voltaire - A Contemporary Version"	
370	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 5 (of 10) From "The Works of Voltaire - A Contemporary Version"	
371	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 6 (of 10) From "The Works of Voltaire - A Contemporary Version"	
372	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 7 (of 10) From "The Works of Voltaire - A Contemporary Version"	
373	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 8 (of 10) From "The Works of Voltaire - A Contemporary Version"	
374	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 9 (of 10) From "The Works of Voltaire - A Contemporary Version"	
375	Francois-Marie Arouet (AKA Voltaire)	A Philosophical Dictionary, Volume 10 (of 10) From "The Works of Voltaire - A Contemporary Version"	
376	Peter Coffey	Ontology or the Theory of Being	
377	Sigmund Freud	Reflections on War and Death	

378	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves: Volume I, Alabama Narratives	
379	Dorothy B. Porter	The Negro in the United States; a selected bibliography. Compiled by Dorothy B. Porter	
380	Work Projects Administration	Slave Narratives: A Folk History of Slavery in the United States From Interviews with Former Slaves: Volume XIV, South Carolina Narratives, Part 3	
381	David Hume	Essays	
382	Victor Cousin	Lectures on the true, the beautiful and the good	
383	Mrs. C. E. Humphry	A Word to Women	
384	George Woodward Warder	The Universe a Vast Electric Organism	
385	Jane Haldimand Marcet and Thomas P. Jones	Conversations on Natural Philosophy, in which the Elements of that Science are Familiarly Explained	
386	R. E. Barrett	Treading the Narrow Way	
387	Austin Holyoak	Ludicrous Aspects Of Christianity A Response To The Challenge Of The Bishop Of Manchester	
388	Percy FitzPatrick	Jock of the Bushveld	
389	Bertrand Russell	Our Knowledge of the External World as a Field for Scientific Method in Philosophy	
390	Arthur, comte de Gobineau	The Moral and Intellectual Diversity of Races With Particular Reference to Their Respective Influence in the Civil and Political History of Mankind	
391	Various	Zanzibar Tales Told by natives of the East Coast of Africa	
392	Heinrich Heine	The Prose Writings of Heinrich Heine	
393	C. Lloyd Morgan	Spencer's Philosophy of Science The Herbert Spencer Lecture Delivered at the Museum 7 November, 1913	
394	J. A. Warder	American Pomology Apples	
395	Myrtle Reed	The Myrtle Reed Cook Book	

396	Joseph Priestley	Heads of Lectures on a Course of Experimental Philosophy: Particularly Including Chemistry	
397	Charles Knowlton	Fruits of Philosophy A Treatise on the Population Question	
398	Various	International Congress of Arts and Science, Volume I Philosophy and Metaphysics	
399	Friedrich Schlegel	The Philosophy of History, Vol. 1 of 2	
400	Arthur Schopenhauer	The World As Will And Idea (Vol. 1 of 3)	
401	Various	The philosophy of B*rtr*nd R*ss*ll	
402	Auguste Sabatier	Outlines of a Philosophy of Religion based on Psychology and History	
403	Lydia Maria Child	The Freedmen's Book	
404	John Albert Macy	The Critical Game	
405	Georg Wilhelm Friedrich Hegel	Hegel's Philosophy of Mind	
406	William de Witt Hyde	The Five Great Philosophies of Life	
407	Edward Clodd	Pioneers of Evolution from Thales to Huxley With an Intermediate Chapter on the Causes of Arrest of the Movement	
408	Auguste Comte	The philosophy of mathematics	
409	Henri Poincaré	The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method	
410	George Berkeley	The Works of George Berkeley. Vol. 1 of 4.	
411	Epictetus	The Teaching of Epictetus Being the 'Encheiridion of Epictetus,' with Selections from the 'Dissertations' and 'Fragments'	
412	Joseph Dietzgen	The Positive Outcome of Philosophy The Nature of Human Brain Work. Letters on Logic.	
413	Herbert Spencer	Illustrations of Universal Progress A Series of Discussions	
414	Edgar Saltus	The Philosophy of Disenchantment	
415	John Dewey	Reconstruction in Philosophy	
416	Arthur Schopenhauer	The World As Will And Idea (Vol. 2 of 3)	

417	George Grote	Plato and the Other Companions of Sokrates, 3rd ed. Volume I (of 4)	
418	John Dewey	Studies in Logical Theory	
419	Lyford Paterson Edwards	The Transformation of Early Christianity from an Eschatological to a Socialized Movement A Dissertation Submitted to the Faculty of the Graduate School of Arts and Literature in Candidacy for the Degree of Doctor of Philosophy	
420	William A. Smith	Lectures on the Philosophy and Practice of Slavery As Exhibited in the Institution of Domestic Slavery in the United States, with the Duties of Masters to Slaves	
421	Emile Durkheim	The Elementary Forms of the Religious Life	
422	Albert Schwegler	A History of Philosophy in Epitome	
423	J. Castell Hopkins Murat Halstead	South Africa and the Boer-British War, Volume I Comprising a History of South Africa and its people, including the war of 1899 and 1900	
424	Muhammad Iqbal	The Development of Metaphysics in Persia A Contribution to the History of Muslim Philosophy	
425	John Dewey	German philosophy and politics	
426	Will Durant	Philosophy and The Social Problem	
427	Walter Leon Hess	Feline Philosophy	
428	Winston Churchill	My African Journey	
429	Thomas Brown	Lectures on the Philosophy of the Human Mind (Vol. 1 of 3)	
430	Frederick von Schlegel	The philosophy of life, and philosophy of language, in a course of lectures	
431	John Elliot Drinkwater Bethune	The Life of Galileo Galilei, with Illustrations of the Advancement of Experimental Philosophy Life of Kepler	
432	Hans Driesch	The Science and Philosophy of the Organism The Gifford Lectures Delivered Before the University of Aberdeen in the Year 1907	
433	Joseph Trapp	Lectures on Poetry Read in the Schools of Natural Philosophy at Oxford	
434	Solomon Northup	Twelve Years a Slave Narrative of Solomon Northup, a Citizen of New-York, Kidnapped in	

		Washington City in 1841, and Rescued in 1853, from a Cotton Plantation near the Red River in Louisiana	
435	Arnold Haultain	Of Walks and Walking Tours An Attempt to find a Philosophy and a Creed	
436	Georg Hegel	The Introduction to Hegel's Philosophy of Fine Arts Translated from the German with Notes and Prefatory Essay	
437	Remy de Gourmont	The Natural Philosophy of Love	
438	J. H. Ward	Gospel Philosophy Showing the Absurdities of Infidelity, and the Harmony of the Gospel with Science and History	
439	Ernst Haeckel	The Wonders of Life A Popular Study of Biological Philosophy	
440	Yue-Gwan Chen	Synthesis of 2-methyl-4-selenoquinazolone, 2-phenylbenzoselenazole, and its derivatives Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Faculty of Pure Science of Columbia University	
441	Remy De Gourmont	Philosophic Nights In Paris Being selections from Promenades Philosophiques	
442	William Jackson	The Philosophy of Natural Theology An Essay in confutation of the scepticism of the present day	
443	Blaise de Vigenère	A Discovrse of Fire and Salt (A Discourse of Fire and Salt) Discovering Many Secret Mysteries as well Philosophicall, as Theologicall (Discovering Many Secret Mysteries as well Philosophical, as Theological)	
445	John Ayrton Paris	Philosophy in Sport Made Science in Earnest Being an Attempt to Illustrate the First Principles of Natural Philosophy by the Aid of Popular Toys and Sports	
446	A. D. Lindsay	The Philosophy of Immanuel Kant	
447	Benjamin Franklin	The Complete Works in Philosophy, Politics and Morals of the late Dr. Benjamin Franklin, [Vol 1 of 3]	

448	Benjamin Franklin	The Complete Works in Philosophy, Politics and Morals of the late Dr. Benjamin Franklin, [Vol 2 of 3]	
449	Benjamin Franklin	The Complete Works in Philosophy, Politics and Morals of the late Dr. Benjamin Franklin, [Vol 3 of 3]	
450	Friedrich von Schlegel	The Philosophy of History, Vol. 2 of 2	
451	George Santayana	Egotism in German Philosophy	
452	Emil Edward Kusel	Humanitarian Philosophy, 4th Edition	
453	Jaime Luciano Balmes	Fundamental Philosophy, Vol. 2 (of 2)	
454	H. L. Mencken	The Philosophy of Friedrich Nietzsche	
455	Lycurgus A. Wilson	Outlines of Mormon Philosophy or the Answers Given by the Gospel, as Revealed Through the Prophet Joseph Smith, to the Questions of Life	
456	Newell Dwight Hillis	German Atrocities Their Nature and Philosophy	
457	John Buchan and Henry Newbolt	Days to Remember The British Empire in the Great War	
458	Martin F. Tupper	Proverbial Philosophy The First and Second Series	
459	T. H. Pasley	The Philosophy Which Shows the Physiology of Mesmerism and Explains the Phenomenon of Clairvoyance	
460	Murray Leinster	Juju	
461	H. G. (Herbert George) Wells	The Time Machine	
462	H. G. Wells	The War of the Worlds	
463	Robert Louis Stevenson	Dr. Jekyll and Mr. Hyde	
464	Robert Louis Stevenson	The Strange Case Of Dr. Jekyll And Mr. Hyde	
465	Edgar Rice Burroughs	A Princess of Mars	
466	Mary Wollstonecraft	Frankenstein or The Modern Prometheus	

	(Godwin) Shelley		
467	Mark Twain (Samuel Clemens)	A Connecticut Yankee in King Arthur's Court, Complete	
468	Arthur Conan Doyle	H. G. Wells	
469	H. G. Wells	The Island of Doctor Moreau	
470	Jules Verne	Twenty Thousand Leagues under the Sea	
471	Ayn Rand	Anthem	
472	Jules Verne	The Mysterious Island	
473	Mark Twain (Samuel Clemens)	Mark Twain's Speeches	
474	H. G. Wells	The Invisible Man	
475	Samuel Johnson	Preface to a Dictionary of the English Language	
476	Frederick Marryat	Mr. Midshipman Easy	
477	H. G. Wells	Mankind in the Making	
478	Achilles Rose	Napoleon's Campaign in Russia Anno 1812	
479	Theodore Roosevelt	The Naval War of 1812 Or The History of the United States Navy during the Last War with Great Britain to Which Is Appended an Account of the Battle of New Orleans	
480	Edward Sapir	Language An Introduction to the Study of Speech	
481	Arthur Quiller-Couch	On the Art of Writing Lectures delivered in the University of Cambridge 1913-1914	
482	N. A. Belcourt	Bilingualism Address delivered before the Quebec Canadian Club, at Quebec, Tuesday, March 28th, 1916	
483	George Grote Count Philippe-Paul de Segur	The Two Great Retreats of History	
484	William Lawrence	The Autobiography of Sergeant William Lawrence A Hero of the Peninsular and Waterloo Campaigns	
485	Isaac Asimov	Youth	
486	Philip Kindred Dick	Second Variety	
487	Anonymous	Child of the Regiment	

488	Sylvia Jacobs	The Pilot and the Bushman	
489	Stephen Barr	I Am A Nucleus	
490	Kris Neville	Fresh Air Fiend	
491	Evelyn E. Smith	The Man Outside	
492	Michael Shaara	Citizen Jell	
493	Jim Harmon	The Spicy Sound of Success	
494	Fritz Leiber	Dr. Kometevsky's Day	
495	Robert Silverberg	Birds of a Feather	
496	W. T. Haggert	Lex	
497	Alfred Coppel	Double Standard	
498	Fritz Leiber	Time In the Round	
499	J.F. Bone	Survival Type	
500	Dean Evans	Not a Creature Was Stirring	
501	Alan Arkin	People Soup	
502	Charles V. de Vet	Growing Up On Big Muddy	
503	Frank Quattrocchi	Sea Legs	
504	Edgar Pangborn	Angel's Egg	
505	Evelyn E. Smith	The Ignoble Savages	
506	Winston Marks	...So They Baked a Cake	
507	Louis Newman	License to Steal	
508	L.J. Stecher	Man in a Quandary	
509	Fritz Leiber	Bullet With His Name	
510	William W. Stuart	A Husband for My Wife	
511	Fritz Leiber	A Pail of Air	
512	L.J. Stecher	Perfect Answer	
513	Frank M. Robinson	The Reluctant Heroes	
514	Fritz Leiber	Kreativity For Kats	
515	Daniel F. Galouye	The Chasers	
516	Keith Laumer	Doorstep	
517	Fritz Leiber	The Last Letter	
518	Bernard Wolfe	Self Portrait	
519	Fritz Leiber	The Big Engine	
520	Patrick Fahy	Bad Memory	

521	Edgar Rice Burroughs	Warlord of Mars	
522	Edward Bellamy	Looking Backwards from 2000 to 1887	
523	Jack London	The Iron Heel	
524	William Hope Hodgson	The House on the Borderland	
525	Philip Kindred Dick	Beyond Lies the Wub	
526	Philip K. Dick	Beyond the Door	
527	Howard Phillips Lovecraft	The Shunned House	
528	Philip Kindred Dick	The Eyes Have It	
529	Philip K. Dick	The Variable Man	
530	Mary Shelley	Frankenstein or, The Modern Prometheus	
531	H. P. Lovecraft	The Dunwich Horror	
532	Ray Bradbury	A Little Journey	
533	Frederik Pohl	My Lady Greensleeves	
534	Jim Harmon	Break a Leg	
535	Alan E. Nourse	Prime Difference	
536	R. DeWitt Miller	Swenson, Dispatcher	
537	Phyllis Sterling Smith	What is Posat?	
538	Kris Neville	Voyage to Far N'jurd	
539	Jim Harmon	No Substitutions	
540	Edgar Pangborn	The Music Master of Babylon	
541	Kris Neville	Hunt the Hunter	
542	L.J. Stecher	An Elephant for the Prinkip	
543	Jack Sharkey	The Business, As Usual	
545	Tom Purdom	Sordman the Protector	
546	Kris Neville	Moral Equivalent	
547	James Stamers	Dumbwaiter	
548	Marshall King	Beach Scene	
549	Patrick Fahy	Bad Memory	
550	Jim Harmon	Blueblood	
551	C.C. MacApp	The Drug	
552	C.M. Kornbluth	With These Hands	
553	Donald Colvin	The Celestial Hammerlock	

554	Frank Banta	Handyman	
555	Magnus Ludens	The Long, Silvery Day	
556	Andrew Fetler	Cry Snooker	
557	Henry Slesar	The Stuff	
558	Sydney Van Scyoc	Shatter the Wall	
559	Joseph Farrell	Security Plan	
560	Algis Budrys	The Rag and Bone Men	
561	R. A. Lafferty	Aloys	