CHAPTER 1

INTRODUCTION TO PROJECT

Project Aim and Objective:

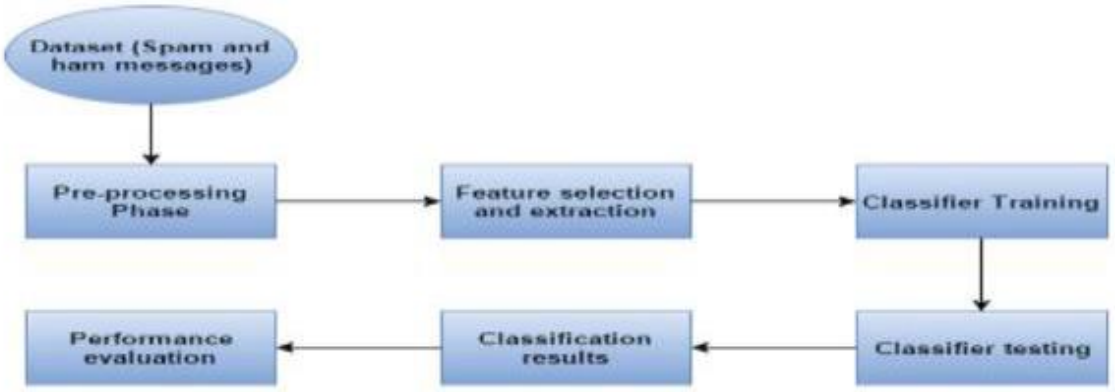
Providing facility of SPAM Detection to counter the fake mails as well as comparative study of the real time technology that can be used to get better and faster outcome.

Problem Statement:

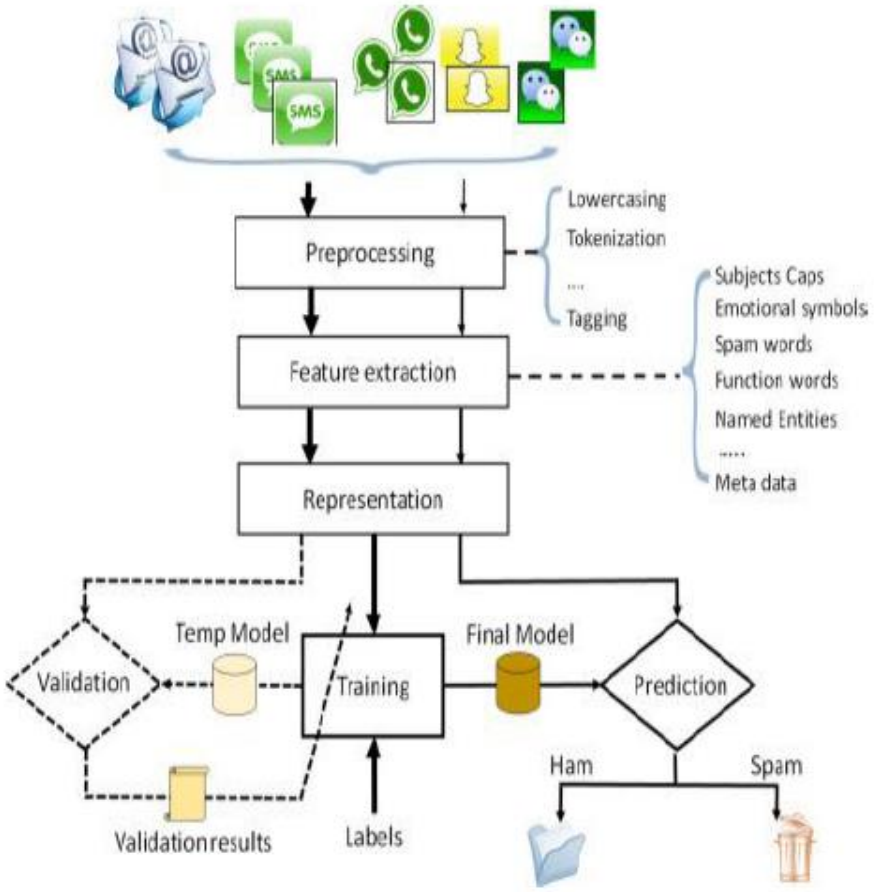
Abusive and unwanted words are initially unavoidable. But they can be removed. In emails, such mails are known as spam mails. For the convenience of the user, these mails should be identified and put into a different category. We will carry out this process with the help of Machine Learning and Big Data both. Later on we will compare them so that we can analyse.

Background of the Project

- **Organization of work element :**



- **Methodology:**



Software Requirements: Spark, HDFS, HADOOP

Hardware Requirements:

Processor: Intel(R) Core (TM) i5-7200U CPU @ 2.50GHz (4 CPUs), ~2.7GHz

Memory: 8192MB RAM

ROM: 1TB

GRAPHICS: 4 GB

CHAPTER 2

PRODUCT BACKLOG

1. **PRODUCT Backlog**

First of all, we'll be searching for datasets for spam detection. Then we'll be removing the spam mails with the help of Machine Learning and Big Data. Finally, we'll be comparing them in order to analyze the,.

PRODUCT BACKLOG							
			Batch 2018_2019				
SPRINT BACKLOG	US ID	BACKLOG ITEM			PRIORITY	RESPONSIBLE	REMARKS
		AS A/AN	I WANT TO	SO THAT			
1	US1/N1	Developer	Understand the problem	I can search for relevant solutions	1	AB+VB+US	
1	US1/N2	Developer	Understand user stories	I can search for dataset accordingly	1	AB+VB+US	
1	S81/US1	Researcher	Learn about Big Data	I Can get the basic knowledge	1	AB+VB+US	
1	S81/US2	Researcher	Data Collection	I Can get the basic knowledge	1	AB+VB+US	
1	S81/US3	Researcher	Filtering of Data	I Can Analyse the data	1	AB+VB+US	
1	S81/US4	Researcher	Pre-Installation Setup	I Can Analyse the data	1	AB+VB+US	
1	S81/US5	Researcher	Hadoop Installation	I Can Analyse the data	1	AB+VB+US	
1	S81/US6	Researcher	Verifying Data	I Can Analyse the data easily	1	AB+VB+US	
1	S81/US7	Researcher	Start HDFS	I Can continue with the project	1	AB+VB+US	
1	S81/US8	Researcher	Compilation and Execution of Process Units Program	I Can make all programs like word count etc	1	AB+VB+US	
2	S81/US9	Researcher	LEARNING Other Tech.	I Can Analyse the data easily	2	AB+VB+US	
2	S81/US10	Researcher	Do PIG Installation	I Can access PIG environment	2	AB+VB+US	
2	S81/US11	Researcher	Do HIVE Installation	I Can access HIVE environment	2	AB+VB+US	
2	S81/US12	Researcher	Do SPARK Installation	I Can access SPARK environment	2	AB+VB+US	
2	S81/US13	Researcher	Download related Papers & project	I Can get to know all other related work done	2	AB+VB+US	
2	S81/US14	Researcher	LEARNING ML	I Can implement in my project	2	AB+VB+US	
2	S81/US15	Researcher	Learning about various algo.	I can implement them	2	AB+VB+US	
2	S81/US16	Researcher	Implementation of naive bayes(ML)	I can use it in my project	2	AB+VB+US	
2	S81/US17	Researcher	Implementation of Decision Tree(ML)	I can use it in my project	2	AB+VB+US	
2	S81/US18	Researcher	Implementation of Logistic Regression(ML)	I can use it in my project	2	AB+VB+US	
3	S81/US19	Researcher	Implementation of naive bayes(BD)	I can use it in my project	3	AB+VB+US	
3	S81/US20	Researcher	Implementation of Decision Tree(BD)	I can use it in my project	3	AB+VB+US	
3	S81/US21	Researcher	Implementation of Logistic Regression(BD)	I can use it in my project	3	AB+VB+US	
3	S81/US22	Researcher	Accuracy Calculation(ML)	improve my performance	3	AB+VB+US	
3	S81/US23	Researcher	Accuracy Calculation(BD)	I can work on it	3	AB+VB+US	
3	S81/US24	Researcher	Comparative Analysis	I can Compare them	3	AB+VB+US	
3	S81/US25	Researcher	present results	I can show them	4	AB+VB+US	

Fig- 2.1 PRODUCT Backlog

2. **Sprint Backlog-1**

SPRINT BACKLOG 1

US ID	USER STORY	TASK ID	TASKS	TM	STATUS (NOT STARTED / IN PROGRESS / COMPLETED)
SPRINT 1 - SPAM DETECTION					
SB1/N1	Understand the problem	SB1/01/T1	Breaking the problem and understanding each component	AB|VB|US	COMPLETED
		SB1/01/T2	Searching for relavent solutions	AB|VB|US	COMPLETED
		SB1/01/T3	Searching alternative approches	AB|VB|US	COMPLETED
		SB1/01/T4	Analysing similar projects	AB|VB|US	COMPLETED
SB1/N2	Understand user stories	SB1/02/T1	Looking for various use cases	AB|VB|US	COMPLETED
		SB1/02/T2	Discussion about user stories	AB|VB|US	COMPLETED
		SB1/02/T3	Searching for type of data required	AB|VB|US	COMPLETED
SB1/US1	LEARNING BIG DATA	SB1/01/T2	Learning about Big Data Hadoop	AB|VB|US	COMPLETED
		SB1/01/T3	Learning about Big Data HDFS	AB|VB|US	COMPLETED
		SB1/01/T4	Learning about Big Data Mapreduce	AB|VB|US	COMPLETED
		SB1/01/T5	Learning about Big Data Echosystem	AB|VB|US	COMPLETED
SB1/US2	DATA COLLECTION	SB1/02/T1	Downloading datasets from github	AB|VB|US	COMPLETED
		SB1/02/T2	Downloading datasets from kaggle	AB|VB|US	COMPLETED
		SB1/02/T3	Downloading datasets from popular dataset sites	AB|VB|US	COMPLETED
SB1/US3	FILTERING OF DATA	SB1/03/T1	Data cleaning	US	COMPLETED
		SB1/03/T2	Applying Transformations	AB|US	COMPLETED
		SB1/03/T3	Removing Null values	VB	COMPLETED
		SB1/03/T4	Data screening	AB	COMPLETED
SB1/US4	Pre-installation Setup	SB1/04/T1	Linux platform	VB	COMPLETED
		SB1/04/T2	Installing Java	US	COMPLETED
		SB1/04/T3	setting up PATH	US	COMPLETED
		SB1/04/T4	Download and extract Hadoop	VB	COMPLETED
		SB1/04/T5	Setting Up Hadoop	AB	COMPLETED
SB1/US5	HADOOP INSTALLATION	SB1/04/T1	Installing single node cluster	AB	COMPLETED
		SB1/04/T2	Installing SSH localhost	AB	COMPLETED
		SB1/04/T3	Setting up the paths	VB	COMPLETED
		SB1/04/T4	Disable the ipv6	US	COMPLETED
		SB1/04/T5	changing Bashrc file	AB	COMPLETED
SB1/US6	Verifying Hadoop Installation	SB1/06/T1	Hadoop Configuration	AB	COMPLETED
		SB1/06/T2	Verifying Hadoop dfs	AB	COMPLETED
		SB1/06/T3	Verifying Yarn Script	AB	COMPLETED
		SB1/06/T4	Accessing Hadoop on Browser	VB|US	COMPLETED
SB1/US7	HDFS	SB1/07/T2	Starting HDFS	VB|US	COMPLETED
		SB1/07/T3	Listing Files in HDFS	VB|US	COMPLETED
		SB1/07/T4	Inserting Data into HDFS	VB|US	COMPLETED
		SB1/07/T5	Retrieving Data from HDFS	VB|US	COMPLETED
SB1/US8	Compilation and Execution of Process Units Program	SB1/08/T1	PROG-1(Word Count)	VB|US	COMPLETED
		SB1/08/T2	PROG-2(Line Count)	VB|US	COMPLETED
		SB1/08/T3	PROG-3(Case Study)	VB|US	COMPLETED
		SB1/08/T4	PROG-4(Case Study)	VB|US	COMPLETED

Fig-2.2 Sprint Backlog 1

3. **Sprint Backlog-2**

SPRINT BACKLOG 2						
US ID	**USER STORY**	**TASK ID**	**TASKS**	**TM**	**STATUS (NOT STARTED / IN PROGRESS / COMPLETED)**	**ESTIMATED DATE OF TASK COMPLETION**
SPRINT 2 - SPAM DETECTION						
SB2/US9	LEARNING Other Tech.	SB2/D1/T1	Learning about Pig	AB-VB-US	COMPLETED	
		SB2/D1/T2	Learning about Hive	AB-VB-US	COMPLETED	
		SB2/D1/T3	Learning about Hbase	AB-VB-US	COMPLETED	
		SB2/D1/T4	Learning about various Algo	AB-VB-US	COMPLETED	
		SB2/D1/T5	Learning about Basic Programs	AB-VB-US	COMPLETED	
SB2/US10	PIG INSTALLATION	SB2/D1/T1	Download Apache PIG	AB-VB-US	COMPLETED	
		SB2/D1/T2	Installing PIG	AB-VB-US	COMPLETED	
		SB2/D1/T3	Configured Apache PIG	AB-VB-US	COMPLETED	
		SB2/D1/T4	Verifying Apache PIG	AB-VB-US	COMPLETED	
SB2/US11	HIVE INSTALLATION	SB2/D2/T1	Download Apache HIVE	AB-VB-US	COMPLETED	
		SB2/D2/T2	Installing HIVE	AB-VB-US	COMPLETED	
		SB2/D2/T3	Configured Apache HIVE	AB-VB-US	COMPLETED	
		SB2/D2/T4	Verifying Apache HIVE	AB-VB-US	COMPLETED	
SB2/US12	SPARK INSTALLATION	SB2/D3/T1	Download Apache SPARK	AB-VB-US	COMPLETED	
		SB2/D3/T2	Installing SPARK	AB-VB-US	COMPLETED	
		SB2/D3/T3	Configured Apache SPARK	AB-VB-US	COMPLETED	
		SB2/D3/T4	Verifying Apache SPARK	AB-VB-US	COMPLETED	
SB2/US13	Download related Papers & project	SB2/D4/T1	Designing of GUI	AB-VB-US	COMPLETED	
		SB2/D4/T2	Finding modules required	AB-VB-US	COMPLETED	
		SB2/D4/T3	Downloading and Installing the modules	AB-VB-US	COMPLETED	
SB2/US14	LEARNING BD	SB2/D5/T1	Studing naïve bayes model	AB-VB-US	COMPLETED	
		SB2/D5/T2	Studing decision tree	AB-VB-US	COMPLETED	
		SB2/D5/T3	Studing Logistic Regression model	AB-VB-US	COMPLETED	
SB2/US16	Naive Bae Model BD	SB2/D6/T1	Studying Naive Bayes Model	AB-VB-US	COMPLETED	
		SB2/D6/T2	Implementing Naive Bayes Model	AB-VB-US	COMPLETED	
		SB2/D6/T3	Implementing Naive Bayes Model on research data	AB-VB-US	COMPLETED	
SB2/US17	Decision Tree BD	SB2/D7/T1	Studying decision tree	AB-VB-US	COMPLETED	
		SB2/D7/T2	Implementing decision tree	AB-VB-US	IN PROGRESS	
		SB2/D7/T3	Implementing decision tree on research data	AB-VB-US	IN PROGRESS	
SB2/US18	Logistic Regression BD	SB2/D8/T1	Studying Logistic Regression	AB-VB-US	COMPLETED	
		SB2/D8/T2	Implementing logistic regression	AB-VB-US	COMPLETED	
		SB2/D8/T3	Implementing decision tree on research data	AB-VB-US	COMPLETED	

Fig-2.3 Sprint Backlog 2

CHAPTER 3

TECHNOLOGY APPLIED AND PROJECT MANAGEMENT

Brief Description of Technologies

a) What is Hadoop?

Apache Hadoop is an Open Source software framework for storage and large-scale processing of data-sets on a clusters of commodity hardware. It is an Open source Data Management software framework with scale-out storage and distributed processing. It is being built and used by a global community of contributors and users.

Apache Hadoop has been originated from Google's Whitepapers:

1. Apache HDFS is derived from GFS (Google File System).
2. Apache MapReduce s derived from Google MapReduce
3. Apache HBase is derived from Google BigTable.

Though Google has only provided the Whitepapers, without any implementation, around 90-95% of the architecture presented in these Whitepapers is applied in these three Java-based Apache projects.

HDFS and MapReduce are the two major components of Hadoop, where HDFS is from the 'Infrastructural' point of view and MapReduce is from the 'Programming' aspect. Though HDFS is at present a subproject of Apache Hadoop, it was formally developed as an infrastructure for the Apache Nutch web search engine project.

To understand the magic behind the scalability of Hadoop from one-node cluster to a thousand-nodes cluster (Yahoo! has 4,500-node cluster managing 40 petabytes of enterprise data), we need to first understand Hadoop's file system, that is, HDFS (Hadoop Distributed File System).

b) What is HDFS (Hadoop Distributed File System)?

HDFS is a distributed and scalable file system designed for storing very large files with streaming data access patterns, running clusters on commodity hardware.Though it has many similarities with existing traditional distributed file systems, there are noticeable differences between these.

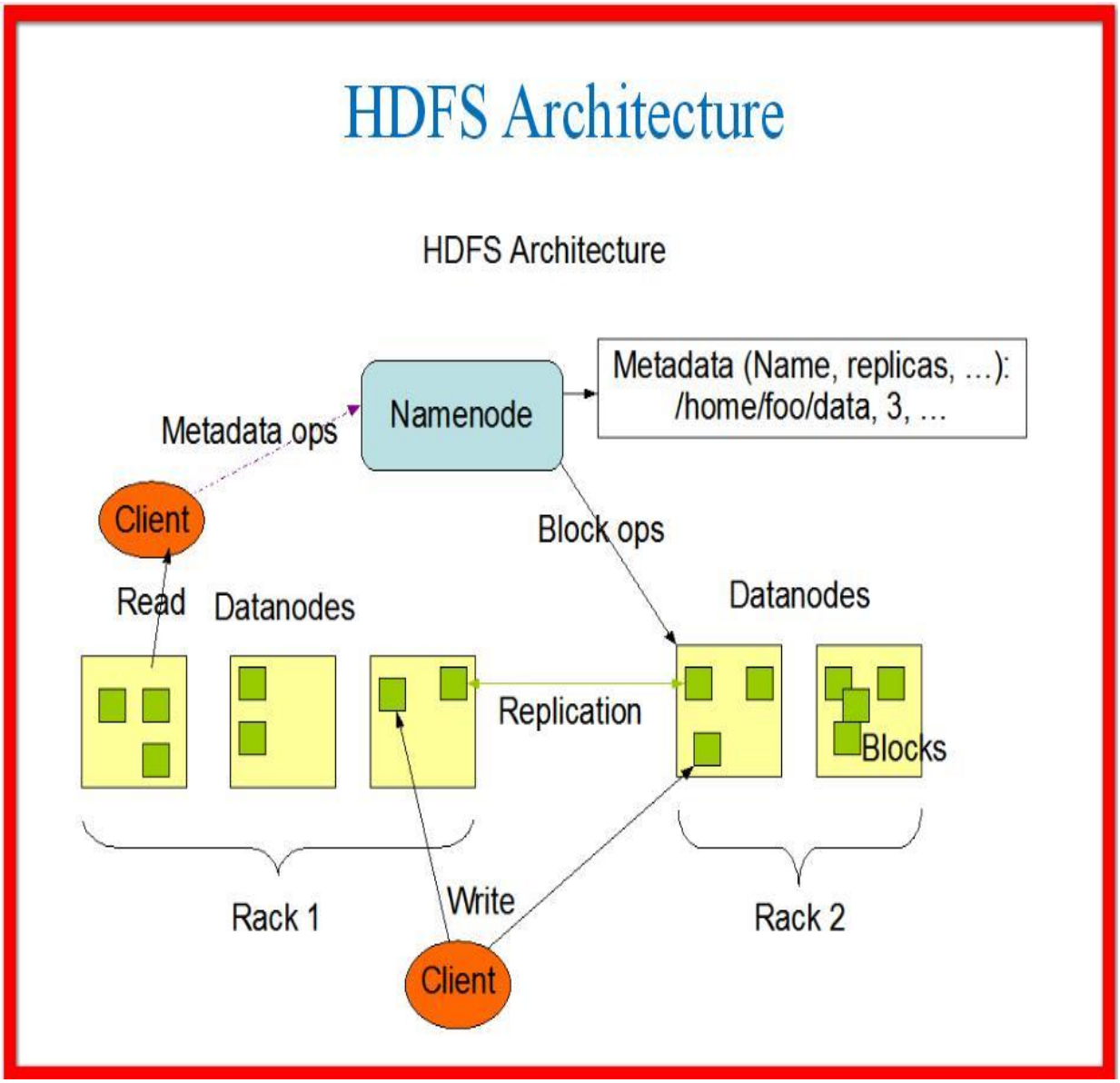


FIGURE- 3.1 Hadoop Distributed File System Architecture

Goals/Objectives behind HDFS:

1. Large Data Sets.
2. Write Once, Read Many Model.
3. Streaming Data Access.
4. Commodity Hardware.
5. Data Replication and Fault Tolerance.
6. High Throughput.

c) HADOOP ECOSYSTEM

Hadoop Ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problems.

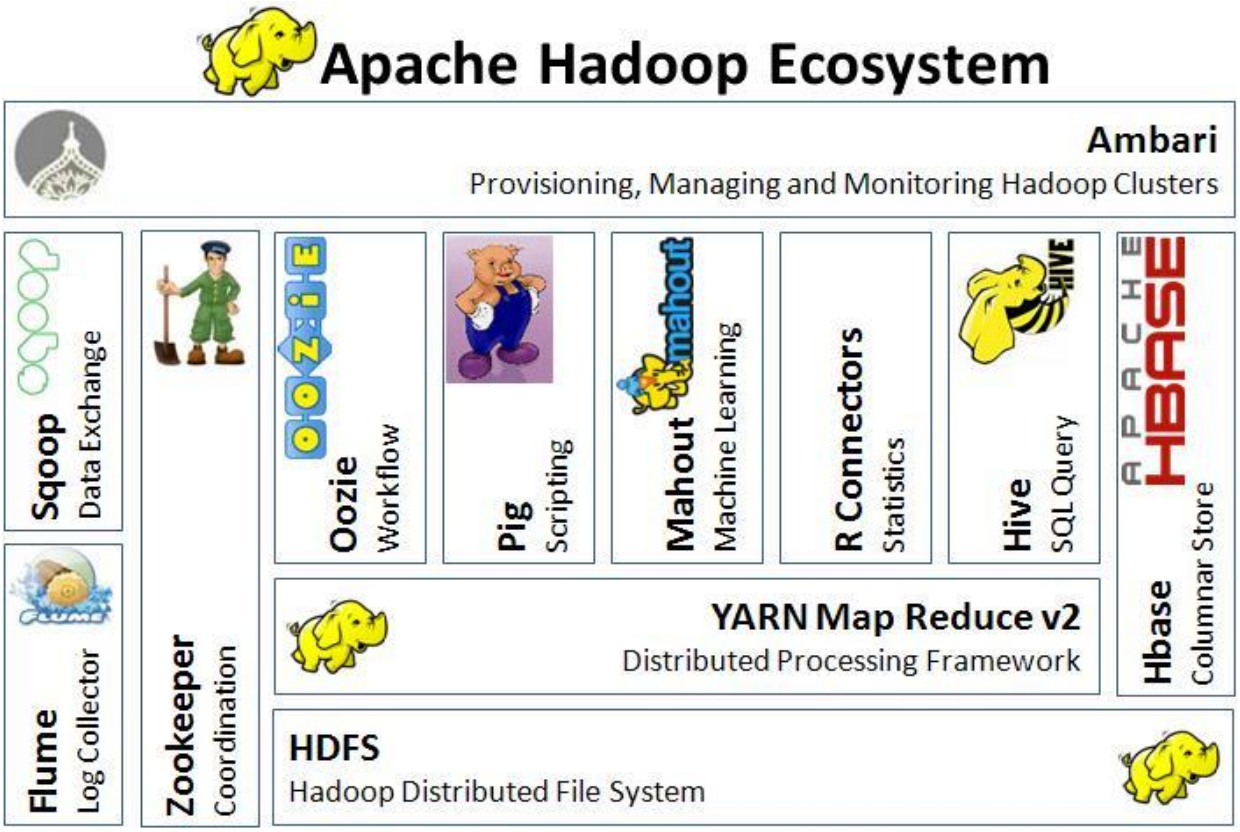


FIGURE-3.2 Hadoop Ecosystem

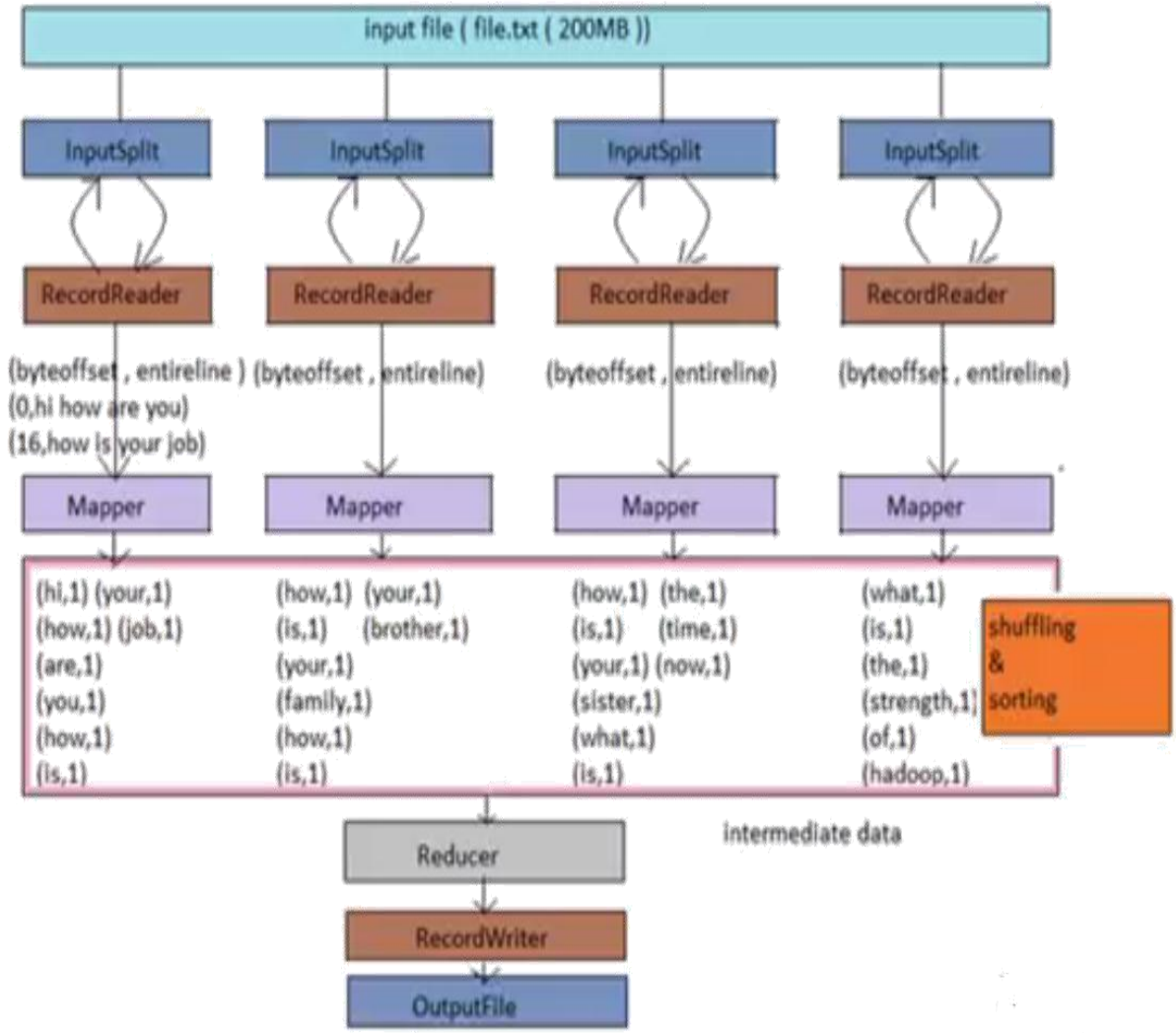
- **HDFS**->*Hadoop Distributed File System*
  - **YARN**->*Yet Another Resource Negotiator*
  - **MapReduce**->*Data processing using programming*
  - **Spark**-> In-memory Data Processing
  - **PIG, HIVE**->*Data Processing Services using Query (SQL-like)*
  - **HBase**->*NoSQL Database*
  - **Mahout, Spark MLlib** ->*Machine Learning*
  - **Apache Drill**->*SQL on Hadoop*
  - **Zookeeper**->*Managing Cluster*
  - **Oozie**->*Job Scheduling*
  - **Flume, Sqoop**->*Data Ingesting Services*
  - **Solr & Lucene** ->*Searching & Indexing*
  - **Ambari** ->*Provision, Monitor and Maintain cluster*

2.1 MAPREDUCE :

MapReduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside Hadoop environment.

- In a MapReduce program, **Map() and Reduce()** are two functions.
    1. The **Map function** performs actions like filtering, grouping and sorting.
    2. While **Reduce function** aggregates and summarizes the result produced by map function.
    3. The result generated by the Map function is a key value pair (K, V) which acts as the input for Reduce function.

MapReduce Flow Chart :



Machine Learning:

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.

Some machine learning methods

Machine learning algorithms are often categorized as supervised or unsupervised.

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- In contrast, **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- **Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired

labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

Project management:

Project management is the application of processes, methods, knowledge, skills and experience to achieve the project objectives. General. A project is a unique, transient endeavor, undertaken to achieve planned objectives, which could be defined in terms of outputs, outcomes or benefits.

Project management is the practice of initiating, planning, executing, controlling, and closing the work of a team to achieve specific goals and meet specific success criteria at the specified time. A project is a temporary endeavor designed to produce a unique product, service or result with a defined beginning and end undertaken to meet unique goals and objectives, typically to bring about beneficial change or added value. The temporary nature of projects stands in contrast with business as usual, which are repetitive, permanent, or semi-permanent functional activities to produce products or services. In practice, the management of such distinct production approaches requires the development of distinct technical skills and management strategies.

Software project management

Software project management is the art and science of planning and leading software projects. It is a sub-discipline of project management in which software projects are planned, implemented, monitored and controlled.

The job pattern of an IT company engaged in software development can be seen split in two parts:

- Software Creation
- Software Project Management

A project is well-defined task, which is a collection of several operations done in order to achieve a goal (for example, software development and delivery). A Project can be characterized as:

- Every project may have a unique and distinct goal.
- Project is not routine activity or day-to-day operations.
- Project comes with a start time and end time.
- Project ends when its goal is achieved hence it is a temporary phase in the lifetime of an organization.
- Project needs adequate resources in terms of time, manpower, finance, material and knowledge-bank.

Software Project

A Software Project is the complete procedure of software development from requirement gathering to testing and maintenance, carried out according to the execution methodologies, in a specified period of time to achieve intended software product.

Need of software project management

Software is said to be an intangible product. Software development is a kind of all new stream in world business and there's very little experience in building software products. Most software products are tailor made to fit client's requirements. The most important is that the underlying technology changes and advances so frequently and rapidly that experience of one product may not be applied to the other one. All such business and environmental

constraints bring risk in software development hence it is essential to manage software projects efficiently.

The image above shows triple constraints for software projects. It is an essential part of software organization to deliver quality product, keeping the cost within client's budget constrain and deliver the project as per scheduled. There are several factors, both internal and external, which may impact this triple constrain triangle. Any of three factor can severely impact the other two.

Therefore, software project management is essential to incorporate user requirements along with budget and time constraints.

Software Project Manager

A software project manager is a person who undertakes the responsibility of executing the software project. Software project manager is thoroughly aware of all the phases of SDLC that the software would go through. Project manager may never directly involve in producing the end product but he controls and manages the activities involved in production.

A project manager closely monitors the development process, prepares and executes various plans, arranges necessary and adequate resources, maintains communication among all team members in order to address issues of cost, budget, resources, time, quality and customer satisfaction.

Let us see few responsibilities that a project manager shoulders -

Managing People

- Act as project leader
- Liaison with stakeholders
- Managing human resources
- Setting up reporting hierarchy etc.

Managing Project

- Defining and setting up project scope
- Managing project management activities
- Monitoring progress and performance
- Risk analysis at every phase
- Take necessary step to avoid or come out of problems
- Act as project spokesperson

Software Management Activities

Software project management comprises of a number of activities, which contains planning of project, deciding scope of software product, estimation of cost in various terms, scheduling of tasks and events, and resource management. Project management activities may include:

- **Project Planning**
- **Scope Management**
- **Project Estimation**

Project Planning

Software project planning is task, which is performed before the production of software actually starts. It is there for the software production but involves no concrete activity that has any direction connection with software production; rather it is a set of multiple processes, which facilitates software production. Project planning may include the following:

Scope Management

It defines the scope of project; this includes all the activities, process need to be done in order to make a deliverable software product. Scope management is essential because it creates boundaries of the project by clearly defining what would be done in the project and what

would not be done. This makes project to contain limited and quantifiable tasks, which can easily be documented and in turn avoids cost and time overrun.

During Project Scope management, it is necessary to -

- Define the scope
- Decide its verification and control
- Divide the project into various smaller parts for ease of management.
- Verify the scope
- Control the scope by incorporating changes to the scope

Project Estimation

For an effective management accurate estimation of various measures is a must. With correct estimation managers can manage and control the project more efficiently and effectively.

Project estimation may involve the following:

- **Software size estimation**

  Software size may be estimated either in terms of KLOC (Kilo Line of Code) or by calculating number of function points in the software. Lines of code depend upon coding practices and Function points vary according to the user or software requirement.

- **Effort estimation**

  The managers estimate efforts in terms of personnel requirement and man-hour required to produce the software. For effort estimation software size should be known. This can either be derived by managers' experience, organization's historical data or software size can be converted into efforts by using some standard formulae.

- **Time estimation**

  Once size and efforts are estimated, the time required to produce the software can be estimated. An effort required is segregated into sub categories as per the requirement specifications and interdependency of various components of software. Software tasks

are divided into smaller tasks, activities or events by Work Breakthrough Structure (WBS). The tasks are scheduled on day-to-day basis or in calendar months.

The sum of time required to complete all tasks in hours or days is the total time invested to complete the project.

- **Cost estimation**

  This might be considered as the most difficult of all because it depends on more elements than any of the previous ones. For estimating project cost, it is required to consider -

  - Size of software
  - Software quality
  - Hardware
  - Additional software or tools, licenses etc.
  - Skilled personnel with task-specific skills
  - Travel involved
  - Communication
  - Training and support

Project Estimation Techniques

We discussed various parameters involving project estimation such as size, effort, time and cost. Project manager can estimate the listed factors using two broadly recognized techniques

Decomposition Technique

This technique assumes the software as a product of various compositions.

There are two main models -

- **Line of Code** Estimation is done on behalf of number of line of codes in the software product.
- **Function Points** Estimation is done on behalf of number of function points in the software product.

Empirical Estimation Technique

This technique uses empirically derived formulae to make estimation. These formulae are based on LOC or FPs.

- **Putnam Model**

  This model is made by Lawrence H. Putnam, which is based on Norden's frequency distribution (Rayleigh curve). Putnam model maps time and efforts required with software size.

- **COCOMO**

  COCOMO stands for COnstructiveCOstMOdel, developed by Barry W. Boehm. It divides the software product into three categories of software: organic, semi-detached and embedded.

Project Scheduling

Project Scheduling in a project refers to roadmap of all activities to be done with specified order and within time slot allotted to each activity. Project managers tend to define various tasks, and project milestones and they arrange them keeping various factors in mind. They look for tasks lie in critical path in the schedule, which are necessary to complete in specific manner and strictly within the time allocated. Arrangement of tasks which lies out of critical path are less likely to impact over all schedule of the project.

For scheduling a project, it is necessary to -

- Break down the project tasks into smaller, manageable form
- Find out various tasks and correlate them
- Estimate time frame required for each task
- Divide time into work-units
- Assign adequate number of work-units for each task
- Calculate total time required for the project from start to finish

Resource management

All elements used to develop a software product may be assumed as resource for that project. This may include human resource, productive tools and software libraries.

The resources are available in limited quantity and stay in the organization as a pool of assets. The shortage of resources hampers the development of project and it can lag behind the schedule. Allocating extra resources increases development cost in the end. It is therefore necessary to estimate and allocate adequate resources for the project.

Resource management includes -

- Defining proper organization project by creating a project team and allocating responsibilities to each team member
- Determining resources required at a particular stage and their availability
- Manage Resources by generating resource request when they are required and de-allocating them when they are no more needed.

Project Risk Management

Risk management involves all activities pertaining to identification, analysing and making provision for predictable and non-predictable risks in the project. Risk may include the following:

- Experienced staff leaving the project and new staff coming in.
- Change in organizational management.
- Requirement change or misinterpreting requirement.
- Under-estimation of required time and resources.
- Technological changes, environmental changes, business competition.

Risk Management Process

There are following activities involved in risk management process:

- **Identification -** Make note of all possible risks, which may occur in the project.
- **Categorize -** Categorize known risks into high, medium and low risk intensity as per their possible impact on the project.

- **Manage -** Analyze the probability of occurrence of risks at various phases. Make plan to avoid or face risks. Attempt to minimize their side-effects.
- **Monitor -** Closely monitor the potential risks and their early symptoms. Also monitor the effects of steps taken to mitigate or avoid them.

Project Execution & Monitoring

In this phase, the tasks described in project plans are executed according to their schedules.

Execution needs monitoring in order to check whether everything is going according to the plan. Monitoring is observing to check the probability of risk and taking measures to address the risk or report the status of various tasks.

These measures include -

- **Activity Monitoring -** All activities scheduled within some task can be monitored on day-to-day basis. When all activities in a task are completed, it is considered as complete.
- **Status Reports -** The reports contain status of activities and tasks completed within a given time frame, generally a week. Status can be marked as finished, pending or work-in-progress etc.
- **Milestones Checklist -** Every project is divided into multiple phases where major tasks are performed (milestones) based on the phases of SDLC. This milestone checklist is prepared once every few weeks and reports the status of milestones.

Project Communication Management

Effective communication plays vital role in the success of a project. It bridges gaps between client and the organization, among the team members as well as other stake holders in the project such as hardware suppliers.

Communication can be oral or written. Communication management process may have the following steps:

- **Planning** - This step includes the identifications of all the stakeholders in the project and the mode of communication among them. It also considers if any additional communication facilities are required.

- **Sharing** - After determining various aspects of planning, manager focuses on sharing correct information with the correct person on correct time. This keeps every one involved the project up to date with project progress and its status.
- **Feedback** - Project managers use various measures and feedback mechanism and create status and performance reports. This mechanism ensures that input from various stakeholders is coming to the project manager as their feedback.
- **Closure** - At the end of each major event, end of a phase of SDLC or end of the project itself, administrative closure is formally announced to update every stakeholder by sending email, by distributing a hardcopy of document or by other mean of effective communication.

After closure, the team moves to next phase or project.

Configuration Management

Configuration management is a process of tracking and controlling the changes in software in terms of the requirements, design, functions and development of the product.

IEEE defines it as "the process of identifying and defining the items in the system, controlling the change of these items throughout their life cycle, recording and reporting the status of items and change requests, and verifying the completeness and correctness of items".

Generally, once the SRS is finalized there is less chance of requirement of changes from user. If they occur, the changes are addressed only with prior approval of higher management, as there is a possibility of cost and time overrun.

Project management Tools:

Project management required tools to manage the work , time and resources. At present many of the software are available for project management. Some of the popular software tools are as follows.

01. <u>Trello</u>

Trello is an project management tool, instead this app is a free visual way to to glance at the entire project with a single view. With Trello you can organise cards, these cards can be your thoughts, conversations and to-do lists and be placed on a board for everyone to collaborate on.

02. Basecamp

Basecamp is the granddaddy of project management apps. Basecamp is considered the leading project management tool around. It boost a simple and easy to use interface to collaborate with your team and client. It allows you to create multiple projects and setup discussions, write to-do lists, manage files, create and share documents, and organise dates for scheduling.

03. Teamwork Projects

Teamwork Projects is the ultimate productivity tool to manage projects with your team. Teamwork allows you to keep all your projects, tasks and files all in one place and easily collaborate with a team. Teamwork helps you to visualise the entire project through a marked calendar and gantt chart and setup reporting. Teamwork supports file management with Google Drive, Box.com and Dropbox. As well as integration with leading apps such as third party accounting software and customer support apps.

04. Resource Guru

Billed as the "simple way to schedule people, equipment and other resources", Resource Guru is a streamlined resource scheduling and leave management tool that’s designed to keep your projects on track. You can plan your team's workloads, receive daily booking reminders, report on KPIs, and more. Apple, Saatchi & Saatchi and Deloitte are among some of the cloud-based team calendar’s heavyweight customers.

05. Active Collab

ActiveCollab recently released its new version 5.0. The new revamped app is now more powerful and focused project management tool. It offers team collaborating features, task management, time tracking and importing expenses. One of the biggest asset of ActiveCollab is it offers invoicing features. You are able to track payments and expenses and have invoices paid directly within ActiveCollab with PayPal, and other credit card payments.

06. Zoho Projects

Zoho offers a wide range of business software including Projects. Zoho Projects is an proficient tool to project plan and project coordinator from start to finish. It boost all the features you need for project management with some advance features including reporting, integration with Google Apps and Dropbox, bug tracking, setup Wiki Pages to build a repository of information, forums and more.

07. Jira

Jira is specifically targeted for software development teams. Jira offers abilities to raise issues and bugs. Jira makes it real easy to track bugs and see which issues are still outstanding and how much time was spent on each task. Jira offer other products including Confluence a document collaboration tool, and HipChat a team chat and video and file sharing platform and other products.

08. Asana

Asana is the easiest way for teams to track their work so everyone knows who's doing what, by when. With tasks, projects, conversations and dashboards, Asana keeps your work organized, and teammates accountable so you can move work forward faster. Asana also lets you keep track of your work wherever you are with mobile apps for both iOS and Android.

09. Podio

Podio is a ever growing tool to organise and communication tool for any business. Podio allows you to personalise this platform to fit your business needs. Besides being able to communicate with a team, setup task management, use as a file storage system, like a traditional project management app, Podio can be an internal intranet for all your colleagues and departments to interact.

10. Freedcamp

Whatever your project may be, either setting up an event, a web project or organising a wedding, Freedcamp helps you organise and plan effectively. Freedcamp has an organised dashboard to view the entire project at a glance. You can easily setup tasks, use sticky notes to visually setup tasks and organise them into the calendar. Freedcamp provides advance add-

ons for high level business use including CRM, invoicing, issue tracking and setting up wiki pages.

11. Wrike

Wrike is advance application to help you work smarter. By making sure you are always staying on track and ensure you have the adequate resources to finish on time and on budget.Setting up tasks, engage your team and integrate with your business tools including Google Apps, Microsoft Excel, Dropbox and many more is so easy with Wrike.

PO and Their Relevance to project

PO1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the solution of complex engineering problems.

In this project creation process engineering knowledge of the software engineering and Electronics engineering have been applied. we have used software engineering , HTML,xml, java , android , java script , php , j2ee, data base , oracle , my sql , mango and other programming language and database to the project. We have applied all above engineering subjects in our projects.

PO2: Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

In our projects we have identified an problem , once verified by the client we have worked to identify the solution using all of our theoretical and practical knowledge.

PO3: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

In the project development we have applied Integrated Development Environment IDE for the rapid development of the code, used web server for the software development.

PO6: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

In 1961 , the Conference of Engineering Societies of Western Europe and the United States of America defined "professional engineer" as follows.

A professional engineer is competent by virtue of his/her fundamental education and training to apply the scientific method and outlook to the analysis and solution of engineering problems. He/she is able to assume personal responsibility for the development and application of engineering science and knowledge, notably in research, design, construction, manufacturing, superintending, managing and in the education of the engineer. His/her work is predominantly intellectual and varied and not of a routine mental or physical character. It requires the exercise of original thought and judgement and the ability to supervise the technical and administrative work of others. His/her education will have been such as to make him/her capable of closely and continuously following progress in his/her branch of engineering science by consulting newly published works on a worldwide basis, assimilating such information and applying it independently. He/she is thus placed in a position to make contributions to the development of engineering science or its applications. His/her education and training will have been such that he/she will have acquired a broad and general appreciation of the engineering sciences as well as thorough insight into the special features of his/her own branch. In due time he/she will be able to give authoritative technical advice and to assume responsibility for the direction of important tasks in his/her branch.

PO7: Environment and sustainability: Understand the impact of the professional engineering solutions in and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

> Sustainability is the ability to continue a defined behavior indefinitely. Sometimes environmental, social and economic are termed to be the three pillars of sustainability.

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice

The ethics of engineers and the fundamental principles for Engineers are as follows.

Engineers uphold and advance the integrity, honor and dignity of the engineering profession by:

I. Using their knowledge and skill for the enhancement of human welfare;

II..Being honest and impartial, and servicing with fidelity the public, their employers and clients;

III. Striving to increase the competence and prestige of the engineering profession; and

IV. Supporting the professional and technical societies of their disciplines.

PO9. Individual and team work: Function effectively as an individual and as a member or leader in diverse teams, and in multidisciplinary settings.

To work successful in team a team member must have following capabilities.

1. The Ability to Listen

it is important to listen to one another's ideas. Too often in a business setting, you have a group of people simply waiting for their turn to speak, not paying one iota of attention to the persons on their left or right. So it is a good teamwork skill to have the ability to listen

2. Check Your Ego

This isn't saying abandon your ego all together, because that isn't healthy. But leaving your ego at the door temporarily is a very important team work skill. The reason this is so essential is because there is always someone better than you at something, no matter how brilliant you are.

3. Critique

By critique, I mean constructive criticism. Be able to give others constructive criticism and be able to listen to others critique your ideas and work. There shouldn't be any offense taken to constructive criticism. You all want to succeed, and this is a vital step in doing so.

4. Delegation

The mentality must be applied to teamwork. Delegate roles to those who do them best.

5. Show Respect

If you and another person happen to be paired up and can't stand each other, you can still put that aside for a couple of hours, treat each other civilly, and complete the tasks at hand. You may even overcome the dislike toward one another.

6. Be Helpful

This is simple. If one of your teammates does not understand an idea, discussion, or task that is being completed, take the necessary time to explain it to them and work with them. There are no weak links when everyone helps one another. Some take longer to learn than others, but that doesn't mean that they are of less intelligence. If in a meeting someone asks a question because they don't understand, don't frown at them. Just answer the questions patiently and concisely.

7. Question One Another
If someone brings up a topic of discussion and a solution to this topic, question them. Respectfully question, don't badger. Rather, ask them how it will work, why it will work over the long-run, and how everyone else can implement the idea.

8. Participation
Have the entire team encourage shy people to engage in the topics of discussion. Don't demand it, but make them realize that you really want to hear their ideas.

9. Rational Debate
Bad ideas are bad for teams. Spirited, friendly, rational debate is where facts come forward, ideas are born, and quality rises to the top.

10. Set The Right Environment
Try to make the space in which your team is assembled as comfortable, relaxing, and inviting as possible. You do not want your team to be tense and with frayed nerves.

PO 10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: Project management and finance: Demonstrate knowledge and understanding of the engineering management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

Project management is the application of processes, methods, knowledge, skills and experience to achieve the project objectives. In general project is a unique, transient endeavour, undertaken to achieve planned objectives, which could be defined in terms of outputs, outcomes or benefits.

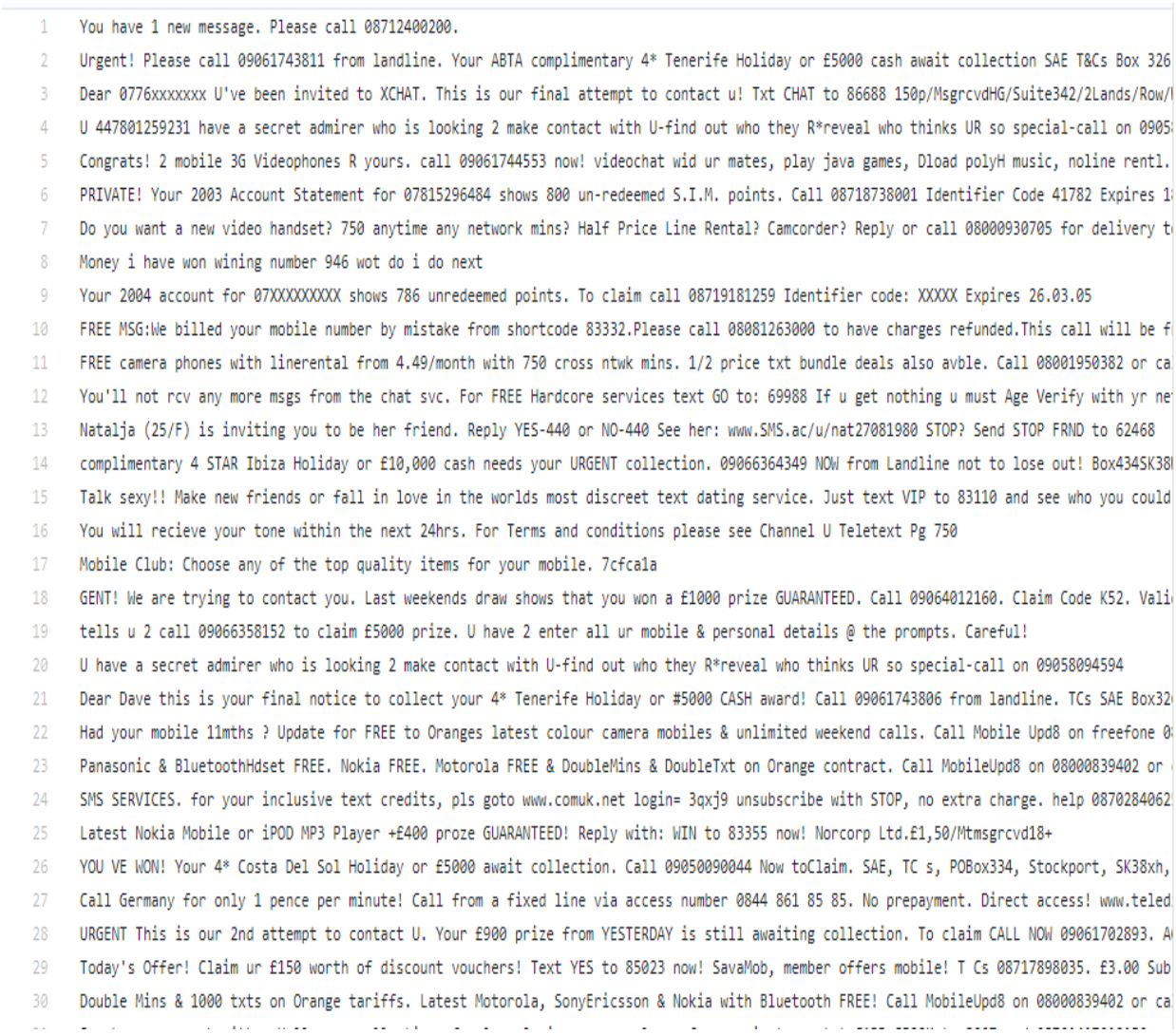
PO12: Life-long learning: Recognize the need for and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Life Long Learning means is the provision or use of both formal and informal learning opportunities throughout people's lives in order to foster the continuous development and improvement of the knowledge and skills needed for employment and personal fulfillment

CHAPTER 4

PROJECT IMPLEMENTATION

1. Sprint Backlog-1



```
You have 1 new message. Please call 08712400200.
Urgent! Please call 09061743811 from landline. Your ABTA complimentary 4* Tenerife Holiday or £5000 cash await collection SAE T&Cs Box 326
Dear 0776xxxxxxx U've been invited to XCHAT. This is our final attempt to contact u! Txt CHAT to 86688 150p/MsgrcvdHG/Suite342/2Lands/Row/
U 447801259231 have a secret admirer who is looking 2 make contact with U-find out who they R*reveal who thinks UR so special-call on 0905
Congrats! 2 mobile 3G Videophones R yours. call 09061744553 now! videochat wid ur mates, play java games, Dload polyH music, noline rentl.
PRIVATE! Your 2003 Account Statement for 07815296484 shows 800 un-redeemed S.I.M. points. Call 08718738001 Identifier Code 41782 Expires 1
Do you want a new video handset? 750 anytime any network mins? Half Price Line Rental? Camcorder? Reply or call 08000930705 for delivery t
Money i have won wining number 946 wot do i do next
Your 2004 account for 07XXXXXXXXX shows 786 unredeemed points. To claim call 08719181259 Identifier code: XXXXX Expires 26.03.05
FREE MSG:We billed your mobile number by mistake from shortcode 83332.Please call 08081263000 to have charges refunded.This call will be f
FREE camera phones with linerental from 4.49/month with 750 cross ntwk mins. 1/2 price txt bundle deals also avble. Call 08001950382 or ca
You'll not rcv any more msgs from the chat svc. For FREE Hardcore services text GO to: 69988 If u get nothing u must Age Verify with yr ne
Natalja (25/F) is inviting you to be her friend. Reply YES-440 or NO-440 See her: www.SMS.ac/u/nat27081980 STOP? Send STOP FRND to 62468
complimentary 4 STAR Ibiza Holiday or £10,000 cash needs your URGENT collection. 09066364349 NOW from Landline not to lose out! Box434SK38
Talk sexy!! Make new friends or fall in love in the worlds most discreet text dating service. Just text VIP to 83110 and see who you could
You will recieve your tone within the next 24hrs. For Terms and conditions please see Channel U Teletext Pg 750
Mobile Club: Choose any of the top quality items for your mobile. 7cfca1a
GENT! We are trying to contact you. Last weekends draw shows that you won a £1000 prize GUARANTEED. Call 09064012160. Claim Code K52. Vali
tells u 2 call 09066358152 to claim £5000 prize. U have 2 enter all ur mobile & personal details @ the prompts. Careful!
U have a secret admirer who is looking 2 make contact with U-find out who they R*reveal who thinks UR so special-call on 09058094594
Dear Dave this is your final notice to collect your 4* Tenerife Holiday or #5000 CASH award! Call 09061743806 from landline. TCs SAE Box32
Had your mobile 11mths ? Update for FREE to Oranges latest colour camera mobiles & unlimited weekend calls. Call Mobile Upd8 on freefone 0
Panasonic & BluetoothHdset FREE. Nokia FREE. Motorola FREE & DoubleMins & DoubleTxt on Orange contract. Call MobileUpd8 on 08000839402 or
SMS SERVICES. for your inclusive text credits, pls goto www.comuk.net login= 3qxj9 unsubscribe with STOP, no extra charge. help 0870284062
Latest Nokia Mobile or iPOD MP3 Player +£400 proze GUARANTEED! Reply with: WIN to 83355 now! Norcorp Ltd.£1,50/Mtmsgrcvd18+
YOU VE WON! Your 4* Costa Del Sol Holiday or £5000 await collection. Call 09050090044 Now toClaim. SAE, TC s, POBox334, Stockport, SK38xh,
Call Germany for only 1 pence per minute! Call from a fixed line via access number 0844 861 85 85. No prepayment. Direct access! www.teled
URGENT This is our 2nd attempt to contact U. Your £900 prize from YESTERDAY is still awaiting collection. To claim CALL NOW 09061702893. A
Today's Offer! Claim ur £150 worth of discount vouchers! Text YES to 85023 now! SavaMob, member offers mobile! T Cs 08717898035. £3.00 Sub
Double Mins & 1000 txts on Orange tariffs. Latest Motorola, SonyEricsson & Nokia with Bluetooth FREE! Call MobileUpd8 on 08000839402 or ca
```

Figure- 4.1.1 Data Description

2. Sprint Backlog-2

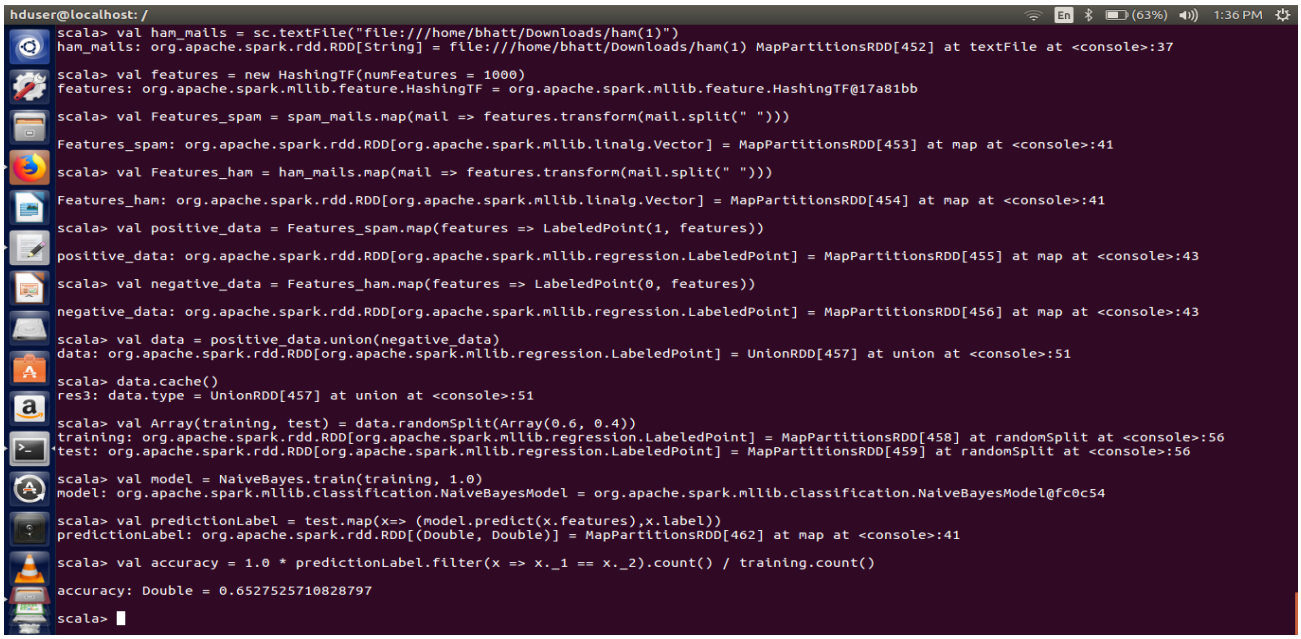


Figure- 4.2.1 Naïve Bayes

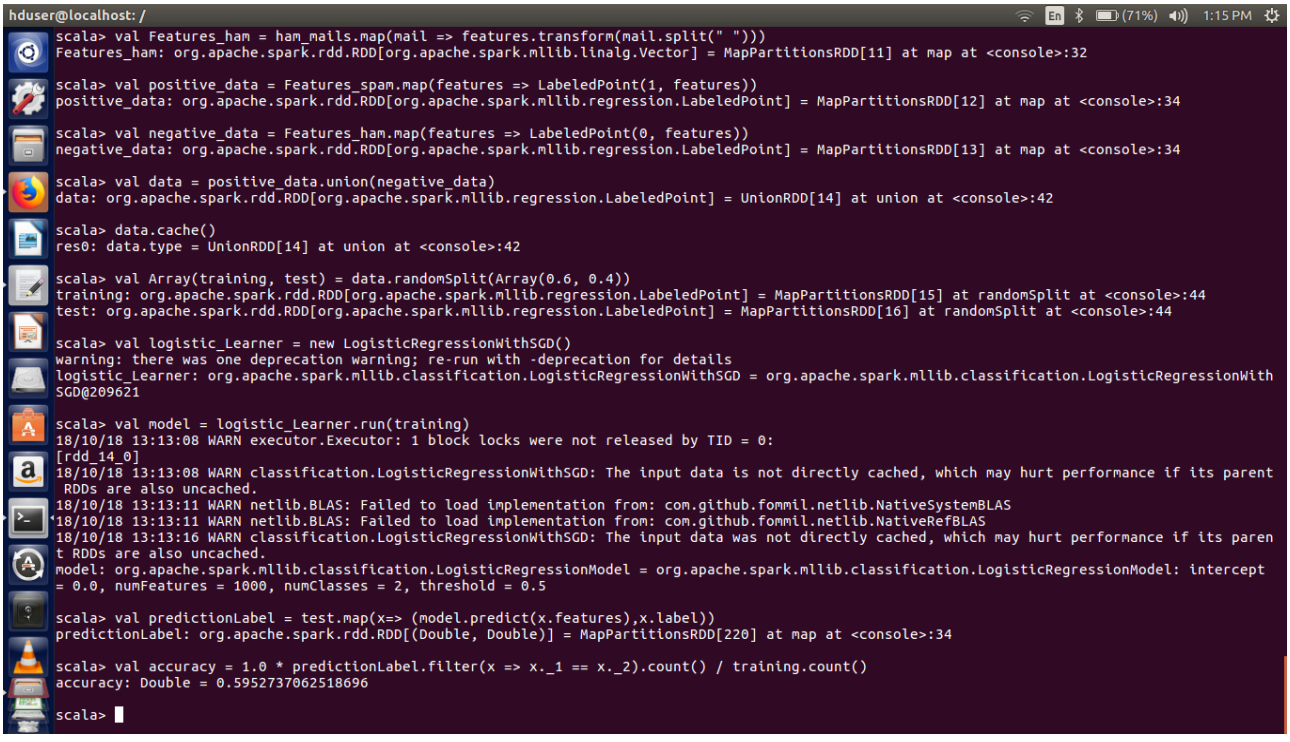


Figure- 4.2.2 Logistic Regression

CHAPTER 5

CONCLUSION

Results:

Found the accuracy between the different algorithms i.e. Naïve Bayes and Logistic Regression.

Conclusion:

Naïve Bayes is more accurate than Logistic Regression in our dataset.

Future Scope:

Implementation of the same in Machine Learning and research paper presentation.

ANNEXURES

References

https://acadgild.com/blog/building-spam-filtering-engine-using-spark-mllib

https://acadgild.com/blog/machine-learning-using-spark

https://www.kaggle.com/uciml/sms-spam-collection-dataset