VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANASANGAMA, BELAGAVI-590 018, KARNATAKA.

AN INTERNSHIP

REPORT ON

"Working on Hadoop Cluster and DevOps"

A Case Study – Employee Database Payroll Management System

Submitted in the partial fulfilment of requirement for the award of Degree

B.E. in Computer Science & Engineering

PROJECT ASSOCIATES

JEEVAN H K	**4BD20CS040**
NITISH S M	**4BD20CS064**
SHRIKAR H M	**4BD20CS090**
DARSHAN A HIREMATH	**4BD20CS124**

INTERNSHIP GUIDE

Prof. Waseem Khan M.Tech.,
Assistant Professor
Department of CS&E,
B.I.E.T., Davangere

EXTERNAL GUIDE

Mr. Santosh Navale,
Ex-Huawei | Freelance Trainer
Bangalore

Department of Computer Science and Engineering

Bapuji Institute of Engineering and Technology

Davanagere-577004

2023-24

Bapuji Institute of Engineering and Technology

Davanagere - 577004

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that **JEEVAN H K,** bearing USN **4BD20CS040** respectively of Computer Science and Engineering department have satisfactorily submitted the Internship Project Report entitled "Working on Hadoop Cluster and DevOps with Employee Database and Payroll Management System" in the partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering (B.E.) in Computer Science & Engineering, under the VTU during the academic year 2023-24.

INTERNSHIP GUIDES

Prof. Waseem Khan M.Tech.,
Internal Guide

Mr. Santosh Navale
External Guide

Dr. Abdul Razak M S Ph.D.
Internship Co-ordinator

13/05/24

Dr. Nirmala C R Ph.D.
Head of the Department

PRINCIPAL
Bapuji Institute of Engineering & Technology
DAVANGERE.

Dr. H B Aravind Ph.D.
Principal

External Examination

Name of the Examiners	Signature with Date
1. Dr. Arun Kumar GH	1. 30/05/24
2. Dr. Naveen Kumar KR	2. 30/5/24

ACKNOWLEDGEMENT

Salutations to our beloved and highly esteemed institute**, "BAPUJI INSTITUTE OF ENGINEERING AND TECHNOLOGY"** for having well qualified staff and lab furnished with necessary equipment's.

Foremost, We would like to express our sincere gratitude to **Mr. Santosh Navale**, Ex-Huaweiand Freelance trainer, for his guidance and knowledge sharing throughout his journey to carryout the internship project work successfully.

We express our sincere thanks to our guide **Prof. Waseem Khan** for giving us constant encouragement, support and valuable guidance throughout the internship without those stable guidance this internship project would not have been achieved.

We express whole hearted gratitude to **Dr. Nirmala C R,** H.O.D of Computer Science and Engineering Department. We wish to thank her for making our task easy by providing her valuable help and encouragement.

We express whole hearted gratitude to our Internship Coordinator **Dr. Abdul Razak M S,** We wish to acknowledge him, who made our task easy, by providing with his valuable help and encouragement.

We also express our whole hearted gratitude to our principal, **Dr. H B Aravind** for his moral support and encouragement.

We would like to extend our gratitude to all the staff of **Computer Science and Engineering Department** for their help and support rendered to us We have benefited a lot from the feedback, suggestions given by them.

Jeevan H K (4BD20CS040)

Nitisth S M (4BD20CS064)

Shrikar H M(4BD20CS090)

Darshan A Hiremath (4BD20CS124)

Vision and Mission of the Department

Vision

"To be a Centre-of-excellence by imbibing state-of-the-art technology in the field of Computer Science and Engineering, thereby enabling students to excel professionally and be ethical."

Mission

M1	Adapting best teaching and learning techniques that cultivates Questioning and Reasoning culture among the students.
M2	Creating collaborative learning environment that ignites the critical thinking in students and leading to the innovation.
M3	Establishing Industry Institute relationship to bridge the skill gap and make them industry ready and relevant.
M4	Mentoring students to be socially responsible by inculcating ethical and moral values.

Program Educational Outcomes (PEOs)

PEO 1	To apply skills acquired in the discipline of Computer Science and Engineering for solving societal and industrial problems with apt technology intervention.
PEO 2	To continue their career in industry/academia or to pursue higher studies and research.
PEO 3	To become successful entrepreneurs, innovators to design and develop software products and services that meets the societal, technical and business challenges.
PEO 4	To work in the diversified environment by acquiring leadership qualities with effective communication skills accompanied by professional and ethical values.

Program Specific Outcomes (PSOs)

PSO 1	Analyze and develop solutions for problems that are complex in nature by applying the knowledge acquired from the core subjects of this program.
PSO 2	Ability to develop Secure, Scalable, Resilient and distributed applications for industry and societal requirements.
PSO 3	Ability to learn and apply the concepts and construct of emerging technologies like Artificial Intelligence, Machine learning, Deep learning, Big Data Analytics, IoT, Cloud Computing, etc for any real time problems.

ABSTRACT

Agile is a method of software development that aims to deliver functional software consistently through brief iterations. Employee Database Payroll Management System project is implemented in terms of agile to deliver in sprints. Team has collected requirements and created product backlog. Sprint planning has been done by team; user stories delivered in sprints. Database size estimations and peak operations size has been identified. The Employee Database Payroll Management System is a critical application used by organisations ,institutions to manage their employees salary. Aim of our work is to implement DevOps practices for a Employee Database Payroll Management System. The project involved the creation of a continuous integration and continuous deployment pipeline for the Employee Database Payroll Management System. The pipeline included various stages such as code compilation, testing, packaging and deployment. The pipeline was implemented using popular DevOps tools such as Git, Jenkins and cypress. To ensure the quality of the Employee Database Payroll Management System, several automated tests were integrated into the pipeline using cypress. These tests included unit tests, integration tests and acceptance tests. The pipeline was also configured to trigger automatic builds and deployments whenever changes were made to the source code repository. The project demonstrated the benefits of DevOps practices in improving the software development and deployment processes.

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 About the Company/ Resource Person

iSans Technologies is into software consulting which handles system migration and maintenance of legacy systems as consultants, also into data analysis of the systems to create dash boards usingbusiness intelligence tools. It provides a platform for the students that bridge the gap in the transition phase from academics to workplace.

Santosh Navale is an entrepreneur leader with Twenty plus years of experience in data analytics, bigdata, development of innovative products and solutions and traditional databases in Telecom and Financial verticals. He has good knowledge in SDLC, agile & lean methodologies, continuous integration and continuous delivery and cloud services. He was working with Huawei in a technical architect role. He is one of the co-founders of Fresher Profiles Private Limited, he held director Technology Strategy position. He Holds degrees BE and MBL (National Law School).

1.2 Agile Methodology

Agile is an iterative approach of project management and software development that helps team members deliver significance to their clients more rapidly and stress-free. An agile team produces work in small, digestible increments as opposed to placing all of their eggs in one massive "big bang" launch. Due to the regular evaluation of needs, plans, and results, teams have a technique for responding quickly to change.

The Manifesto for Agile Software Development: The programmers describe a novel approach to creating software as well as 4 crucial characteristics they believe should take precedence over other factors. As they put it, agile software development teams should value:

- Individuals and interactions over processes and tools.
- Working software over comprehensive documentation.
- Customer collaboration over contract negotiation.
- Responding to change over following a plan.

1.3 DevOps

The most effective way to describe DevOps is as a team effort to create, develop, and quickly deliver secure software. With automation, teamwork, quick feedback, and iterative improvement, DevOps principles allow software development (dev) and operations (ops) teams to expedite delivery.

Stemming from an Agile approach to software development, a DevOps process expands on the cross-functional approach of building and shipping applications in a faster and more iterative manner. In adopting a DevOps development process, one can make a decision to improve the flow and value delivery of their application by encouraging a more collaborative environment at all stages of the development cycle.

DevOps represents a change in mindset for IT culture. In building on top of Agile, lean practices, and systems theory, DevOps focuses on incremental development and rapid delivery of software. Success relies on the ability to create a culture of accountability, improved collaboration, empathy, and joint responsibility for business outcomes.

1.4 Hadoop Concepts

Hadoop is an open source software programming framework for storing a large amount of data and performing the computation. Its framework is based on Java programming with some native code in C and shell scripts.

Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets**.**

Hadoop has several key features that make it well-suited for big data processing:

- Distributed Storage: Hadoop stores large data sets across multiple machines, allowing for the storage and processing of extremely large amounts of data.
- Scalability: Hadoop can scale from a single server to thousands of machines, making it easy to add more capacity as needed.
- Fault-Tolerance: Hadoop is designed to be highly fault-tolerant, meaning it can continue to operate even in the presence of hardware failures.
- Data locality: Hadoop provides data locality feature, where the data is stored on the same node where it will be processed, this feature helps to reduce the network traffic and improve the performance
- High Availability: Hadoop provides High Availability feature, which helps to make sure that the data is always available and is not lost.
- Flexible Data Processing: Hadoop's MapReduce programming model allows for the processing of data in a distributed fashion, making it easy to implement a wide variety of data processing tasks.

- Data Integrity: Hadoop provides built-in checksum feature, which helps to ensure that the data stored is consistent and correct.
- Data Replication: Hadoop provides data replication feature, which helps to replicate the data across the cluster for fault tolerance.
- Data Compression: Hadoop provides built-in data compression feature, which helps to reduce the storage space and improve the performance.
- YARN: A resource management platform that allows multiple data processing engines like real-time streaming, batch processing, and interactive SQL, to run and process data stored in HDFS.

Hadoop has two main components:

- HDFS (Hadoop Distributed File System): This is the storage component of Hadoop, which allows for the storage of large amounts of data across multiple machines. It is designed to work with commodity hardware, which makes it cost-effective.
- YARN (Yet Another Resource Negotiator): This is the resource management component of Hadoop, which manages the allocation of resources (such as CPU and memory) for processing the data stored in HDFS.

Some common frameworks of Hadoop

- Hive- It uses HiveQL for data structuring and for writing complicated MapReduce in HDFS.
- Spark- It contains a Machine Learning Library (ML lib) for providing enhanced machine learning and is widely used for data processing. It also supports Java, Python, and Scala.
- Pig- It has Pig Latin, a SQL-Like language and performs data transformation of unstructured data.
- Drill- It consists of user-defined functions and is used for data exploration.
- Storm- It allows real-time processing and streaming of data.

Hadoop framework is made up of the following modules:

1. Hadoop MapReduce- a MapReduce programming model for handling and processing large data.
2. Hadoop Distributed File System- distributed files in clusters among nodes.
   Hadoop YARN- a platform which manages computing resources.
3. Hadoop Common- it contains packages and libraries which are used for other modules.

CHAPTER 2

TASK PERFORMED

2.1 Task Performed During Weeks 1 and 2

Learning about waterfall and agile methods

- Understood the requirements of the project " Employee Database Payroll Management System " andcreate a requirement list for waterfall and agile methods.
- Understood the requirements of the project " Employee Database Payroll Management System " andcreate a requirement list for agile methods.

Creation of story card in Trello

- Creating story cards in Trello.
- Move the cards to Doing, Done based on progress.

Product Backlog

- Understood the “Employee Database Payroll Management System” requirement and create a productbacklog using the template

sprint backlog

- Created sprint backlog from product backlog using the same template.

Git Installation

- Download and install Tortoise GIT.
- Execute Git bash commands.

Creating Repository in GitHub

- Create a new repository in GitHub for “Employee Database Payroll Management System”.
- Add "Heath Insurance Management System" code to repository using Git Bash.

Installing Jenkins

- Pull the project from GitHub into Jenkins.
- Create a Freestyle Project in Jenkins to copy files to \htdocs folder.
- Create a Freestyle Project in Jenkins run cypress automation script.

Cypress

- Writing Automation script for the “Employee Database Payroll Management

System” project.Linux Command

- Learning about the Linux Commands

Hadoop Installation

- Installing Hadoop and setting up the cluster consisting of 3 nodes

HDFS

- Understanding the HDFS Architecture and Components

2.2 Task Performed During Week 3 and 4

YARN

- Understanding the YARN Architecture and its Components
- Learning about Resource Manager and Node Manager

MapReduce

- Understanding Distributed MapReduce Framework
- Learning MapReduce Terminologies

Apache Hive

- Installing Apache Hive
- Creating Tables in the Hive

HBase

- Installation of HBase
- Understanding the HBase Architecture

Spark

- Understanding Spark Architecture and Components
- Working in Spark Shell

CHAPTER 3

SYSTEM REQUIREMENTS

3.1 Tools and Technologies Identified

3.1.1 SOFTWARE REQUIREMENTS

Software: XAMPP

OS: platform independent

Front end: HTML, CSS and JavaScript

Programming language: PHP and SQL

Database: MySQL

3.1.2 HARDWARE REQUIREMENTS

Processor: intel pentium dual core, 1.7 ghz

RAM: 512 MB

Hard disk: 20 GB or more

3.1.3 TOOLS IDENTIFIED

- Git
- Trello
- VS Code
- Jenkins
- Cypress
- Xampp
- MySQL workbench

CHAPTER 4

SYSTEM DESIGN

4.1 System Topology

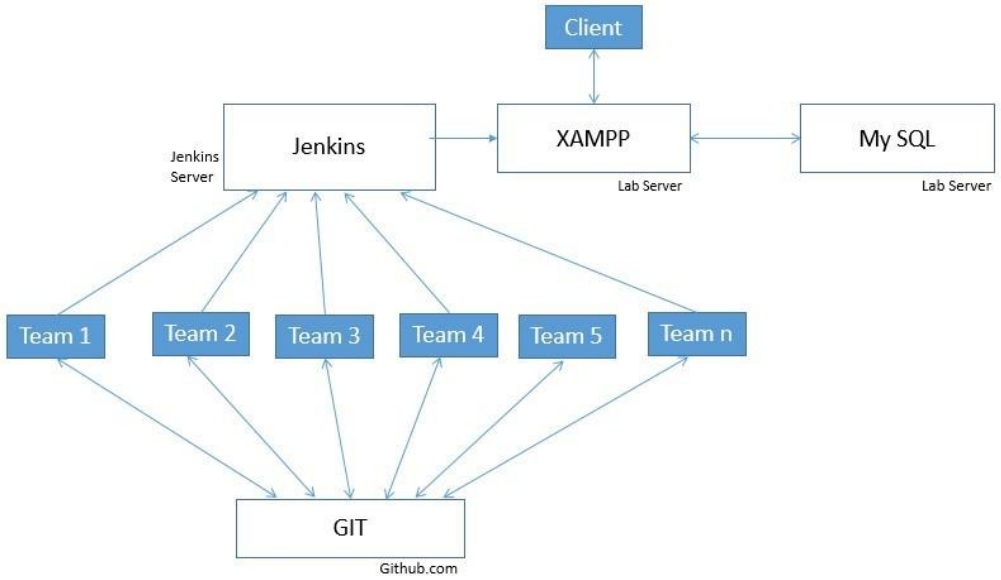

Fig 4.1 System Topology

Figure 4.1 shows the system architecture of the project.

4.2 Flow Diagram

Fig 4.2 Flow Diagram

Figure 4.2 describes the process flow of the project.

CHAPTER 5

METHODOLOGY

5.1 Description of the project work

The main objective of the project is to know fundamental concepts and can work on Agile methodology and DevOps frameworks, to gain a broad understanding of build cycles. The project flows as users create multiple projects in Jenkins to fetch the code, build, prepare datain the database, run the automation tests, and deploy the code webserver. To accomplish this,we have worked on the activities and tasks like Requirement analysis, User Story creation, and Storyboard on Trello, Writing automation tests in Cypress, create projects in Jenkins.

5.2 Steps to be followed

The following steps are used for each activity under each method:

Method	Activity	Task Description
Waterfall and agile method	Requirement Analysis	Understand the " Employee Database Payroll Management System " requirements and create a requirement list for waterfall and agile methods
		Understand the " Health Insurance Management System" requirement and create a product backlog using the template
		Create a sprint backlog from the product backlog using the same template.
		Create Storyboard in Trello for sprint handling
	Sprint Preparation, Design and Implementation	Conduct sprint planning meeting and pick-up the stories for sprint 1
		Create user stories with acceptance criteria
		Daily Standup Meeting
		Create story cards in Trello in To Do status
		Move the cards to Doing, Done based on progress
		Table size and average row size calculation using MySQL commands

DevOps Concepts and Implementation	Plan phase	Analyze Requirements and prepare list
		Write detail design for all requirements
	Code Version control and Static Code Check Review	Create a GitHub account or use existing
		Download and install Tortoise GIT Download https://git-scm.com/downloads
		Perform or execute Git bash commands
		Create a new repository in GitHub for " Health Insurance Management System"
		Add Employee Database Payroll Management System " code to repository using Git Bash
		Install MYSQL workbench
	Build Setup Dev, Test Environment	Install XAMPP in local for dev and test environments
		Install / Check Mysql or MariaDB
		Create all tables in DB
		Install XAMPP in cloud for production environment
		Install Cypress in local for system test environment
		Install Jenkins in local
	Testing	Write system Test cases using test case sheet
		Execute all test cases manually
		Write Automation scripts for all test cases
		Analyze and Submit the Test Report
	Environment Setup	Create a Freestyle Project in Jenkins to prepare data base environment
	Release	Create a Freestyle Project in Jenkins to fetch files from git fetch
	Deploy	Create a Freestyle Project in Jenkins to copy files to \htdocs folder
	Run Cypress Test	Create a Freestyle Project in Jenkins run cypress automation script

Customer Request	DE normalize the tables used in DBMS projects	Provide the requirement analysis and Design
	Change the title	Change heading in index.php
		Commit and push in Git thru git bash
		Observe the task execution Jenkins
		Observe the changed Heading in the portal
HDFS	Hadoop Cluster Preparation HDFS Installation HDFS Read write	Setup all 3 nodes in cluster Install NN, DN, RM, NM Execute HDFS commands and Yarn commands Check the NN and RM url
Hive	Hive Setup Hive Commands	Hive Nodes install Create DB and run hive ql commands Create DB for DBMS project and insert data related to project
Spark and HBase	Setup Spark and Hbase Commands	Installation and Command Execution
Power BI	Setup and Installation Reports for DBMS project	Setup and MySQL Driver connection HDFS driver connection Hive driver connection
Linux	Setup WSL/Linux terminal	Execute Linux commands related to HDFS

Table 5.2: Task and Activity Table

5.2.1 Waterfall Method

Sl. No	Requirements	Description	Requirement Type	Priority	Responsible
1	Admin login	Provide an interface for the admin to enter admin id and password	GUI	High	Darshan H M
		Redirects to Admin Dashboard	GUI, Back-end	High	
2	Department	Provide an interface for the admin to add new departments	GUI, Back-end	High	Shrikhar M
		Provides an interface for the admin to delete existing departments	GUI, Back-end	Low	
		Provides an interface for the admin to view department list	GUI	Medium	
		Provides an interface for the admin to update existing department details	GUI, Back-end	Medium	
3	Employee	Provide an interface for the employee to enter id and password	GUI	High	Nitish S M
		Provide an interface for the employee to edit details	GUI, Back-end	Medium	
		Provide an interface for the employee to remove details	GUI, Back-end	Low	
4	Payment	Provide an interface for admin to make payment to respective employee	GUI, Back-end	High	Jeevan H K
		Provide an interface to generate pay slip for each employee	GUI	High	
5	Salary	Provide an interface to set salary for employee working in department	GUI, Back-end	High	Jeevan H K

Table 5.2.1: Waterfall Method table

5.2.2 Agile Method

Sl no	Requirements	Description	User Stories	Acceptance Criteria	Requirement Type	Priority
1.	Admin Login	Provides an interface for the admin to enter admin id and password	As an admin I want to login and access and look into user and gas statistics for gas booking	Validate the credentials	GUI	High
		Redirects to Admin Dashboard	System should validate the admin credentials and redirect to admin dashboard	I should be logged in to the Emplyee database management	GUI, Back-end	High
2.	Department	Provide an interface for the admin to add new departments	As an admin I want add department data	Entered new departments should be displayed in the forms	GUI, Back-end, DB	High
		Provides an interface for the admin to delete existing departments	As an admin I want delete existing departments	Deleten existing departments should be displayed in the forms	GUI, Back-end	low
		Provides an interface for the admin to view department list	As an admin I want view department list	View department list	GUI, Back-end	Medium
		Provides an interface for the admin to update existing department details	As an admin I want update existing department details	Update the department details	GUI	Medium
3.	Employee	Provide an interface for the employee to enter id and password	Interface of the employee to be enter id and password	Entered employee id and password that should be available in database	GUI	High

		Provide an interface for the employee to edit details	Interface for the employee to edit details	Allows the employee to edit details	GUI, Back-end	Medium
		Provide an interface for the employee to remove details	Interface for the employee to remove details	Allows the employee to remove details	GUI, Back-end	Low
4.	Payment	Provide an interface for admin to make payment to respective employee	Interface for admin to make payment to respective employee		GUI, Back-end	HIgh
		Provide an interface to generate pay slip for each employee	interface to generate pay slip for each employee		GUI	High
5.	Salary	Provide an interface to set salary for employee working in department	As an admin I want set salary for employee working in department	Validate the credentials	GUI, Back-end, DB	High

Table 5.2.2 : Agile Method table

5.2.3 Cypress Automated Testing

- Create a new spec file inside e2e folder of Cypress Automation folder.
- Write the test cases in the new spec file.
- Open command prompt and execute “npx cypress open” command.
- Select E2E testing.
- Select the spec file to run the test cases.
- Verify the results.

5.2.4 Jenkins CI/CD Implementation

1.Job to pull code from GitHub:

- Open Jenkins and create a new item with name “Employee Database Payroll Management System”. Select free style project.
- Go to Source code management and select git. Specify the url for GitHubrepository. Edit the default branch name to “main”.
- Save the job.

2. Job to Deploy code into Apache server:

- Open Jenkins and create a new item with name “Employee Database Payroll Management System” . Select free style project.
- Inside build projects select “Build after other projects are built”.
- For ‘projects to watch’ select ‘Employee Database Payroll Management System’ job.
- Inside ‘Build steps’ select ‘execute windows batch commands’ and add below commands

```
mkdir C:\xampp\htdocs\files1
copy C:\ProgramData\Jenkins\.jenkins\workspace\ Employee Database Payroll Management System \C:\xampp\htdocs\files1
```

- Save the job

3. Job to run Automated Testing:

- Open Jenkins and create a new item with name "Employee Database Payroll Management System Test". Select free style project.
- Inside build projects select "Build after other projects are built".
- For 'projects to watch' select 'Employee Database Payroll Management System Deploy' job.
- Inside 'Build steps' select 'execute windows batch commands' and add below commands

```
set CYPRESS_RUN_BINARY=C:\Users\BIET\AppData\Local\Cypress\Cac
he\10.9.0\Cypress\Cypress.exe
cd /d E:\CypressAutomation
npx cypress run --browser chrome --spec
```

- Save the job

CHAPTER 6

RESULTS AND DISCUSSIONS

6.1 SNAPSHOTS

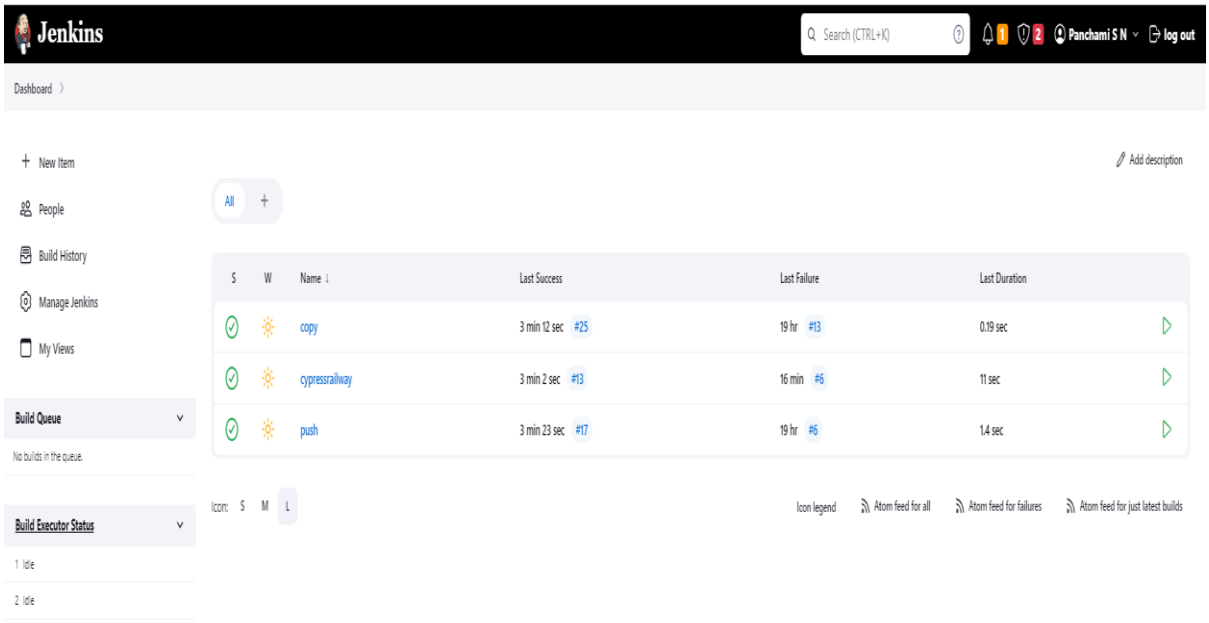

Fig 6.1 : Jenkins Dash Board

The figure 6.1 shows the free style projects executed on Jenkins dashboard

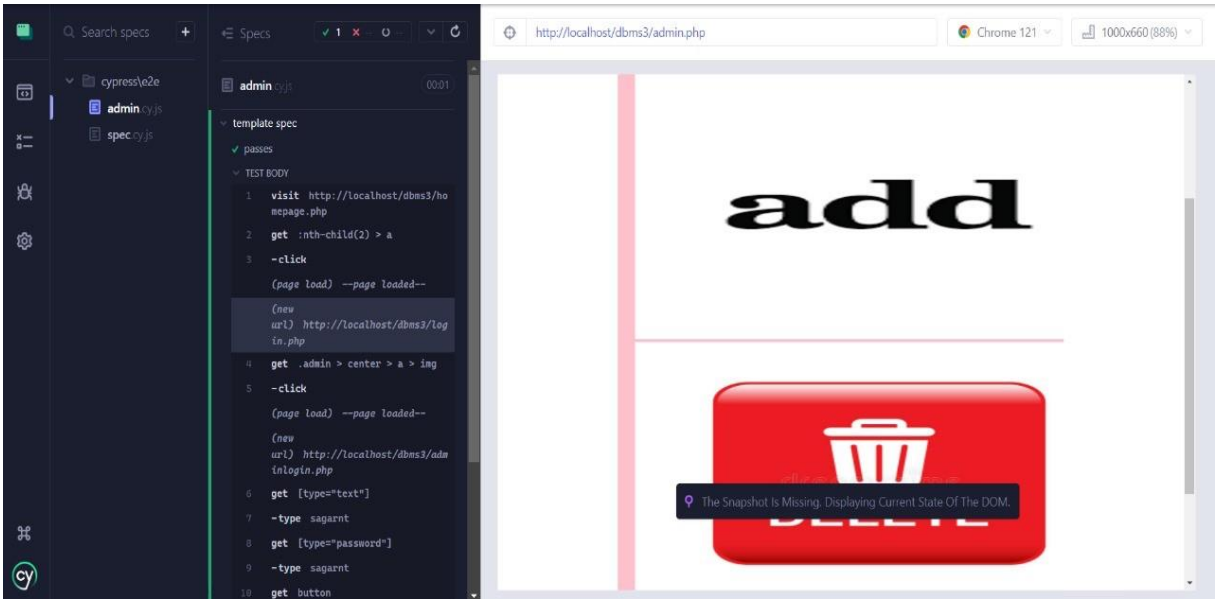

Fig 6.2: Figure shows the automation test results in Cypress

The figure 6.2 shows that execution status of cypress test of the project

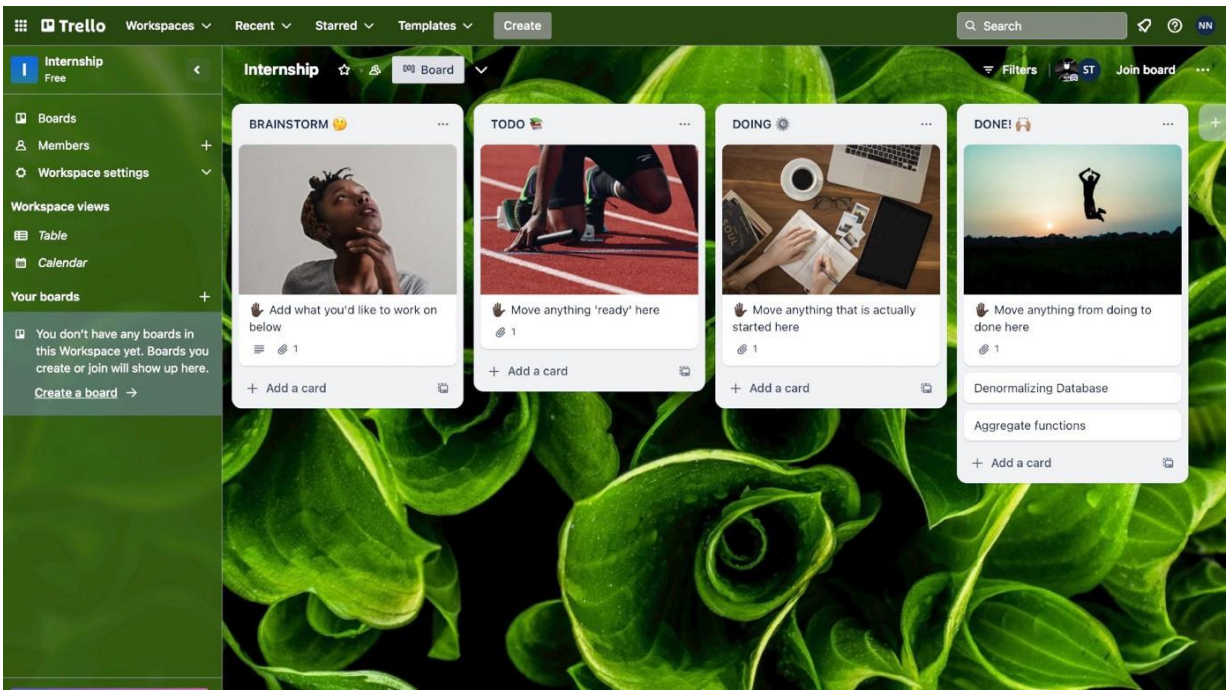

Fig 6.3: Figure shows the Story Board on Trello

The figure 6.3 shows story board on Trello

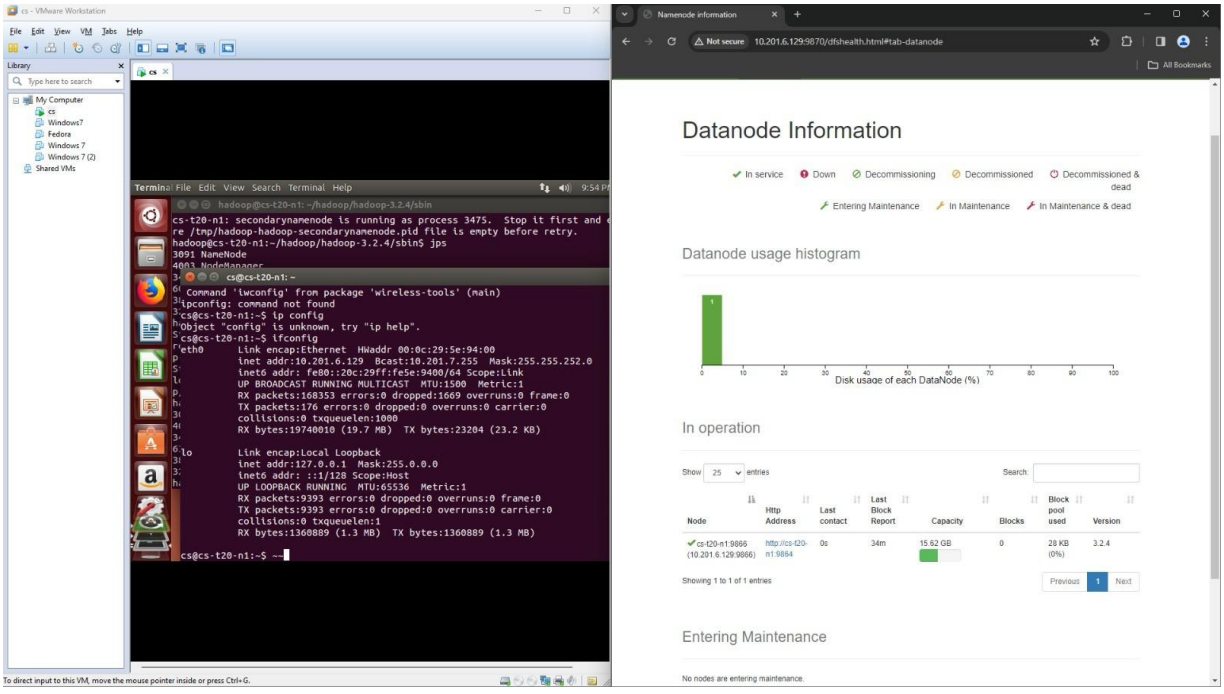

Table 6.5: Data node availability on Hadoop

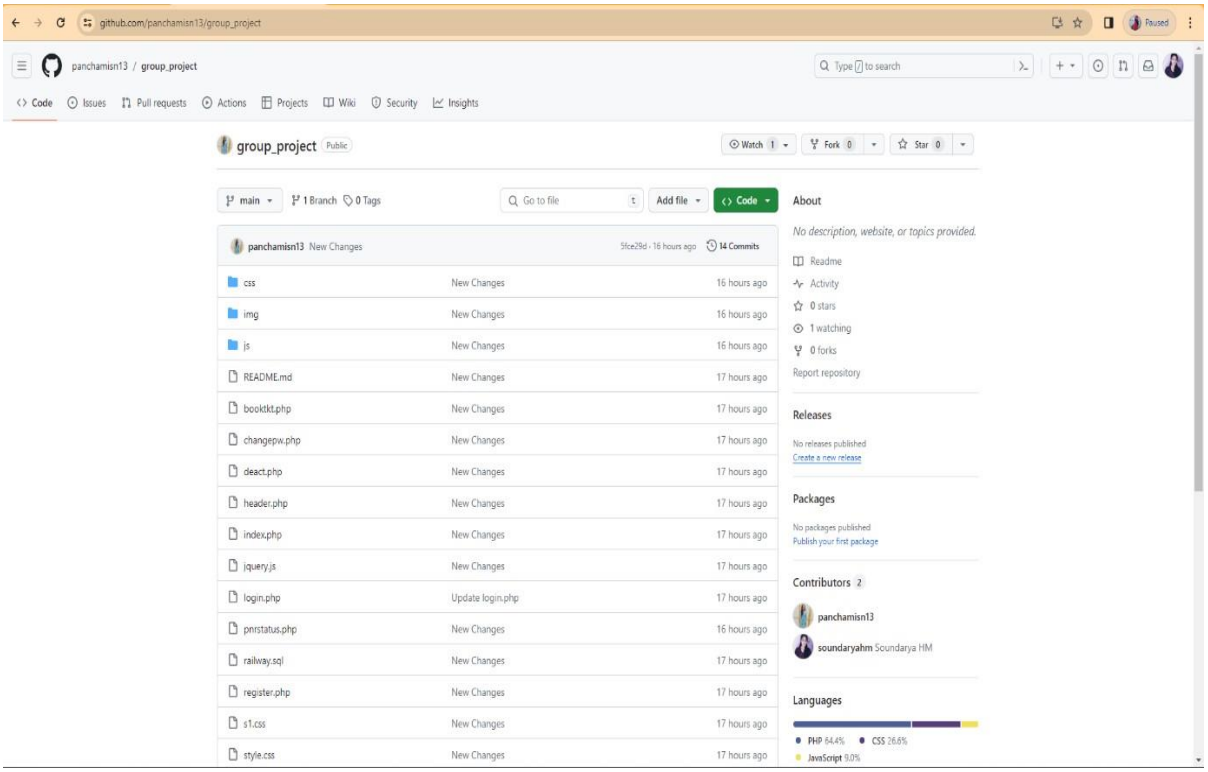

Fig 6.6: Commits on repository

The Figure 6.6 describes commits made by teammates.

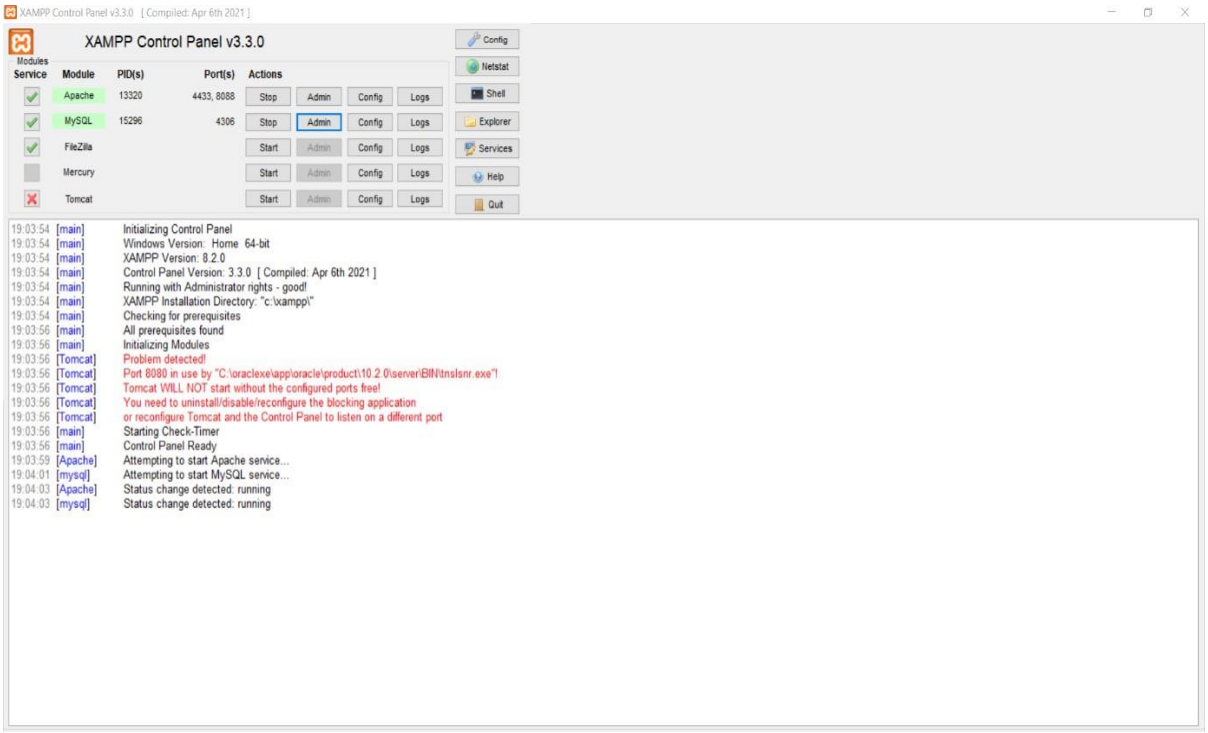

Fig 6.7: Apache and MySQL server

The Figure 6.7 describes initialization of Apache and MySQL server using Xampp.

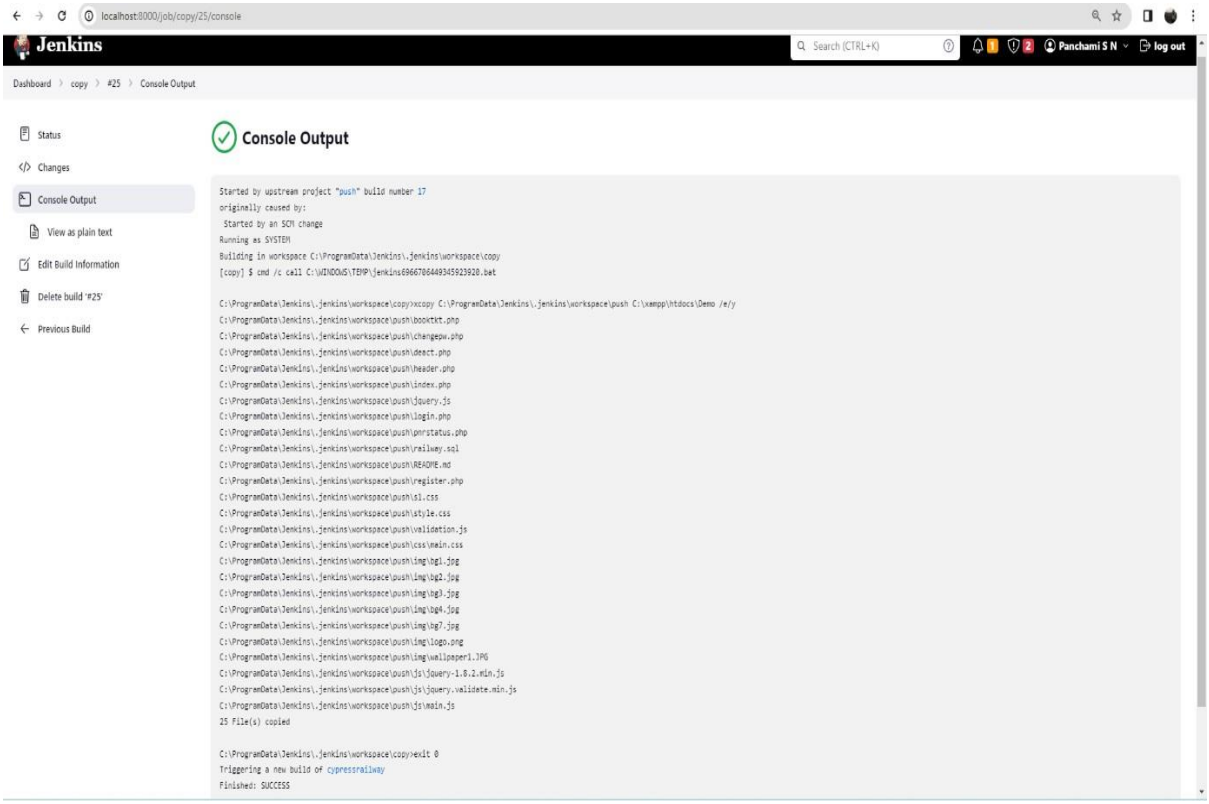

Fig 6.8: Employee Database Payroll Management System job in Jenkins

The Figure 6.8 describes the output of Jenkins job to pull code from GitHub.

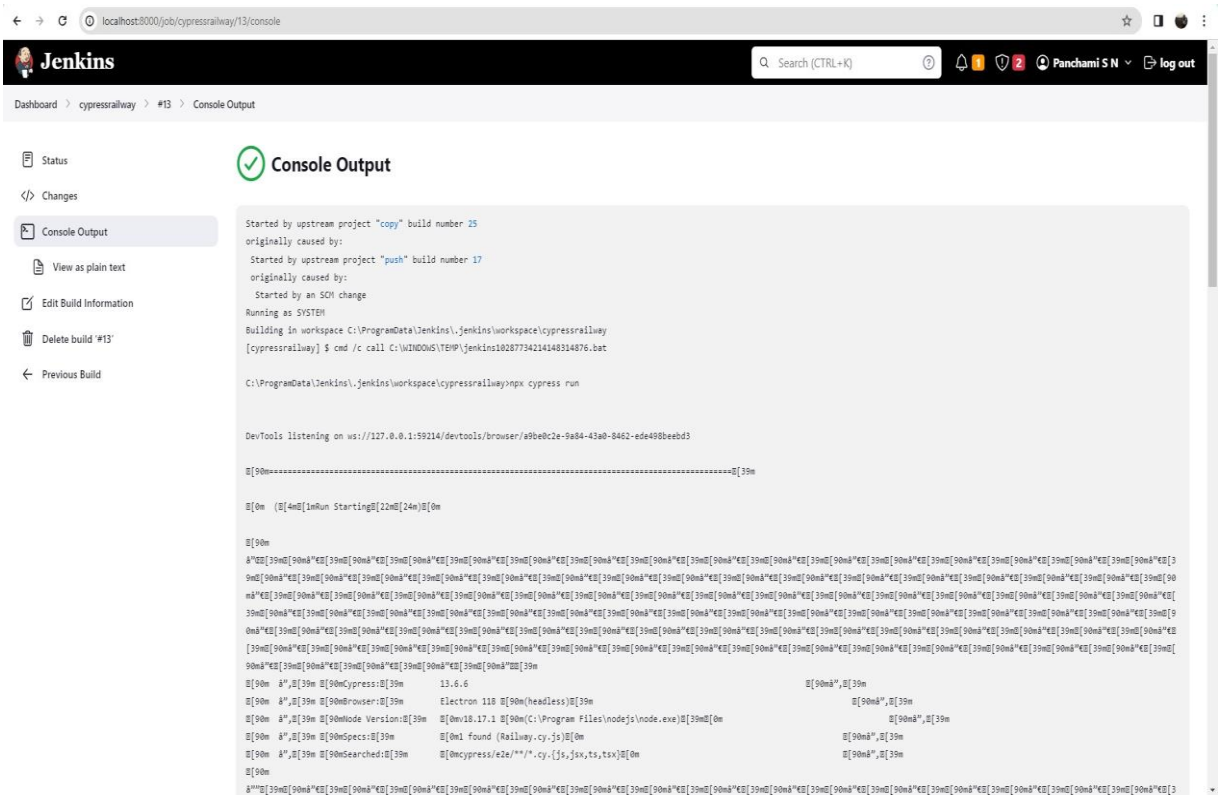

Fig 6.9: Employee Database Payroll Management System Deploying job in Jenkins

CONCLUSION

In this internship a comprehensive exploration of both Agile and Waterfall methodologies, offering valuable insights into the strengths and weaknesses of each approach. The experience of managing the team using Trello underscored the significance of effective communication and collaboration in project development. Integration of DevOps practices through Cypress, Jenkins. Additionally, the hands-on involvement in Hadoop installation, cluster node setup demonstrated a practical understanding of big data infrastructure and GitHub project pushing and version control. Furthermore, the integration of Cypress and Jenkins for project automation showcased the importance of streamlining development processes for efficiency. The installation of Hive and Spark, coupled with the creation of tables using Hive, underscored the practical application of big data technologies Overall, the experience served as a holistic learning journey, merging theoretical knowledge with hands-on application across various facets of modern software development practices.

REFERENCES

Links:

- https://www.jenkins.io/doc/book/installing/
- https://docs.cypress.io/guides/end-to-end-testing/writing-your-first-end-to-end-test
- https://www.apachefriends.org/download.html

Text books:

- DevOps for Beginners – Joseph Joyner
- Hadoop The Definitive Guide: Storage and Analysis at Internet Scale,4th Edition - Tom White
- Modern DevOps Practices: Implement and secure DevOps in the public cloud with cutting-edge tools,
- tips, tricks, and techniques – Gaurav Agarwal