Tribhuvan University

Texas International College

Final Year Internship Report

On

BACKEND DEVELOPER OF ANALYTICS ENGINE

At

Deerwalk Services Pvt. Ltd.

Under the Supervision of

Mr. Kumar Poudyal

HoD, Department of CSIT

Texas International College

Submitted To:

Department of Computer Science and Information Technology

Tribhuvan University

Texas International College

In partial fulfillment of the requirement for the Bachelor Degree in Computer Science and Information Technology

Submitted By:

Aakash Khadka (15561/074)

September, 2022

MENTOR'S RECOMMENDATION

I hereby recommend that this report prepared under my mentorship by **Aakash Khadka** entitled "**BACKEND DEVELOPER OF ANALYTICS ENGINE**" in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Information Technology to be processed for the evaluation.

—----------

Mr. Roshan Shrestha
Assocate Engineering Manager
Deerwalk Services Pvt. Ltd.

SUPERVISOR'S RECOMMENDATION

I hereby recommend that this internship report prepared under my supervision by **Aakash Khadka** (15561/074, Texas International College) entitled "**BACKEND DEVELOPER OF ANALYTICS ENGINE**" in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Information Technology of Tribhuvan University be processed for the evaluation.

Mr. Kumar Poudyal
Project Supervisor
HoD, Department of CSIT
Texas International College
Chabahil, Kathmandu

LETTER OF APPROVAL

This is to certify that this internship report prepared by Aakash Khadka (15561/074) in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Information Technology has been well studied. In our opinion it is satisfactory in the scope and quality as a internship report for the required degree.

Mr. Kumar Poudyal
Supervisor
Texas International College
Chabahil, Kathmandu

Mr. Kumar Poudyal
HoD, Department of CSIT
Texas International College
Chabahil, Kathmandu

Mr. Hikmat Rokaya
External Supervisor
CDCSIT, Tribhuvan University

ACKNOWLEDGEMENT

This internship would not have been possible without the support of many people. Firstly, I would like to sincerely express my gratitude to Deerwalk Service Pvt. Ltd. for providing this golden opportunity of internship. I would also like to thank my mentor **Mr. Roshan Shrestha** for continuously guiding and motivating me during my internship.

Similarly, I would also like to thank my internship supervisor and our BSc. CSIT Head, **Mr. Kumar Poudyal**, for providing all the required guidance. Finally, I would like to share my appreciation for all of our classmates from the batch of 2074 for their support and help. I hope that all of us will achieve more in our future endeavors.

Our parents were our first teachers and they have provided us with such a great exposure that has helped us bloom. Their precious suggestions and guidelines motivated me to work on this internship with great interest. I would like to thank my parents for their continuous support. Finally, I would like to thank all our friends, teachers, and everyone who helped me in this internship directly and indirectly.

Date: September 2022

Sincerely,
Aakash Khadka(15561/074)

ABSTRACT

This internship serves the purpose to record the details of my industrial training which was conducted in Deerwalk Services Pvt. Ltd., which provides Secure, flexible, scalable platform for the automated collection, enrichment and management of healthcare data. This report will cover the details of the internship at the Engine team of the plan analytics platform, the major objective was to gain real-world experience in Big data solutions, as well as get familiar with the technologies associated with it.The engine team which implements the business logic of the plan analytics platform presented chance to learn and implement data pipelines using technologies such as hadoop, elasticsearch, etc.

This internship provided a perfect opportunity to apply some of the knowledge that was gained from academics. In addition, the internship helped to develop soft skills like teamwork, communication, and flexibility. This report includes the tasks performed during the internship that involves various aspects of software development concerning big data solution.

Keywords: *Big Data, Plan Analytics, Hadoop, etc.*

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	: Application Programming Interface
AWS	: Amazon Web Services
BA	: Business Analysis
CEO	: Chief Executive Officer
CFO	: Chief Financial Officer
CGO	: Chief Governance Officer
COO	: Chief Operating Officer
CSIT	: Computer Science and Information Technology
DAS	: Deerwalk Analytics System
HIPAA	: Health Insurance Portability and Accountability Act
JSON	: Javascript Object Notation
NDA	: Non Disclosure Agreement
PA	Plan Analytics
PC	: Personal Computer
PHI	: Personal Health Information
QA	: Quality Assurance
R&D	: Research and Development
S3	: Simple Storage Service
SCR	Small Client Request
SDLC	: Software Development Life Cycle
SQL	: Structured Query Language
SVP	: Senior Vice President
VP	: Vice President
VPN	Virtual Private Network

CHAPTER 1: INTRODUCTION

1.1 Introduction

As an intern, I worked in Deerwalk Services Pvt. Ltd. as a backend developer. I worked in the Engine team of the plan analytics Department of the company which implements the business logic of plan analytics software using Big data technologies like Java, Hadoop(cascading), NoSQL(elasticsearch), etc. Plan analytics provides integrated informatics and actionable insights and savings recommendations, presented in an easy-to-use application that requires no special training for the population health data.

During the internship, I was tasked to learn about the PA Engine project and the technology stack used such as Cascading(Hadoop), elasticsearch, AWS, etc. The tasks were to implement simple data pipelines using cascading and write the output in elasticsearch. Also, I had to learn about the Hadoop jobs and their workflows and understand their usage. After learning the basics, tasks were assigned with increasing complexity. The tasks were assigned through deermine(project management software) using tickets along with their priority. Also, I was majorly involved in initiating and monitoring batch application processing of Hadoop jobs for different clients. Also, I needed to attend daily standup and End of Day meetings along with code reviews for every task done in a particular sprint.

1.2 Problem Statement

Everyday vast amounts of healthcare data is generated all around the world. The raw data is meaningless and very expensive to store without their proper utilization. Various big data technologies have also evolved to handle vast amounts of data to support decision making, study current trends, and make future predictions but utilizing big data solutions for healthcare data is not seen as a business opportunity in many parts of the world.
Thus as an emerging business opportunity, Deerwalk Services have developed PA Engine project to implement business logic on the healthcare data. So, I have joined PA Engine team in order to gain domain knowledge of the healthcare industry as well as the technologies associated with the Big data solution.

1.3 Objectives

The primary objectives of the internship program are:

- To gain real-world experience in utilizing Big data solutions for healthcare data.
- To learn and utilize big data technology stacks such as Hadoop, elasticsearch, AWS, etc.
- To create new or modify existing big data pipelines according to business requirements of the organization.

1.4 Scope and Limitation

The scope of the internship is:

- Attend daily Standup and End of Day meetings for daily updates on tasks and understanding of tasks. Also, perform code reviews for every change and attend code reviews for other team members.
- Attend feature discussion meeting with QA and BA team
- Understand the requirements of tasks from the deermine ticket
- Attend various training sessions organized by the company and team
- Develop and modify data pipelines of various Hadoop jobs and write the data to elasticsearch index

The limitation of the internship is:

- Due to Health Insurance Portability and Accountability Act (HIPAA) policy and Non-Disclosure Agreement (NDA), some restrictions are adopted while sharing information regarding clients, bug details, and core software details on which internship work was carried out.
- This report does not cover information about all the modules of the back end domain. It only covers technical details of two major features completed during the internship period, an overview of a bug, and a feature added.
- This report does not contain the project structure structures and layouts due to company policy.
- Due to time constraints for the internship, this report only describes the knowledge gained during the working hours of Deerwalk Services.

1.5 Report Organization

This internship report consists of four chapters altogether. The report has been organized in the order given as below;

Chapter 1: Introduction

It describes the project in detail with its scopes, limitations, objectives and problem statement.

Chapter 2: Organization Detail and Literature Review

It contains the details of an organization including the hierarchy, working domains as well as the description of the intern department. It also talks about related studies and literature review of the offered project.

Chapter 3: Internship Activities

It includes the roles and responsibilities of an intern during the internship period in an organization. It also contains a weekly log of all the activities done during internship along with details about the tasks and project in which the intern was assigned.

Chapter 4: Conclusion and Learning Outcomes

It is the final section of the report and it talks about the things that intern learn during within the internship period in the organization

CHAPTER 2: ORGANIZATION DETAILS AND LITERATURE REVIEW

2.1. Introduction to Organization

Deerwalk Services Pvt. Ltd. recently acquired by Cedar Gate Technologies is a revolutionary Population Health Management, Data Management, and Healthcare Analytics software company. Deerwalk was created by former executives of several well-known healthcare analytics companies, including D2Hawkeye, Kanawha Health Solutions, etc. Deerwalk's senior executive team has more than a century of combined healthcare IT and analytics experience at the C-Suite level of healthcare businesses.

Deerwalk not only brings excellent products and services to our clients but some of the brightest healthcare analytics subject matter expertise in the industry. Deerwalk's understanding and experience in the healthcare analytics space are unparalleled, offering Big Data and NoSQL products and services to Third Party Administrators, Brokers & Consultants, Employers & Providers, Health Plans, Accountable Care Organizations, Care Management & Wellness firms.

Table 2.1: Organization Profile

Organization	Deerwalk Services Pvt. Ltd
Address	Sifal, Kathmandu
Office hour	9:00 AM – 6:00 PM
Working hour	8 hours per day
Holidays	Saturday and Sunday
Website	https://www.deerwalk.com/

2.2. Organizational Hierarchy

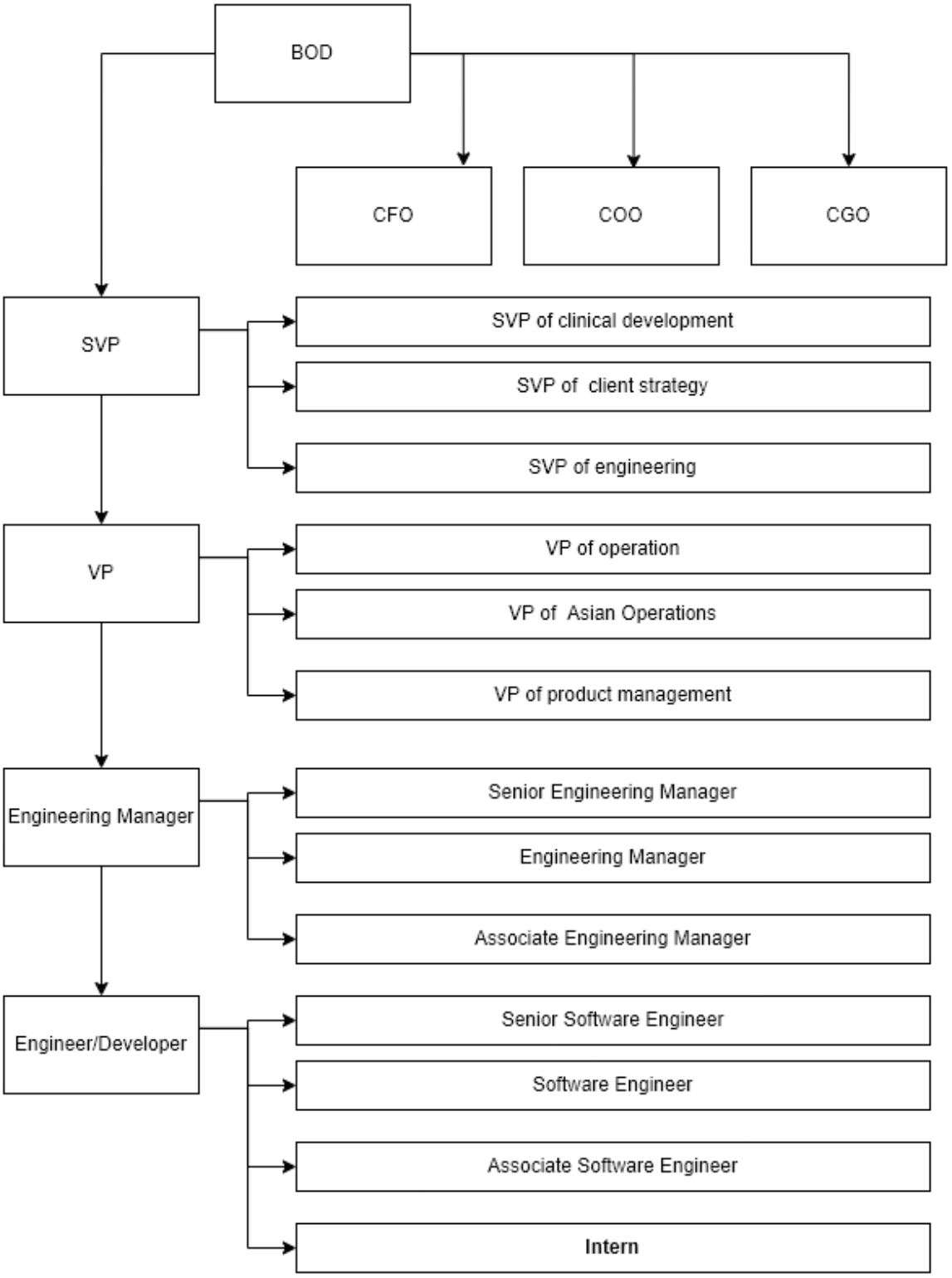

Figure 2.1: Organization Hierarchy

- **Board of Directors:** Board of Directors in Deerwalk includes the Chairman, CEO, President, CFO, and Director. CEO works closely with the Deerwalk team in the US and Nepal directing the operations and technology teams. President leads the business development, healthcare consulting, and strategy for Deerwalk.

- **COO:** COO is responsible for the daily operation of the company and routinely reports to the highest-ranking executive, usually the chief executive officer.

- **CGO**: CGO is responsible to work across key areas that drive growth, which include marketing, sales, research & development, and finance, to create and implement a longer-term vision and enterprise-wide execution of growth-generating strategies for the daily operation of the company and routinely reports to the highest-ranking executive, usually the chief executive officer.

- **CFO**: CFO is responsible for financial planning and record-keeping, as well as financial reporting to higher management.

- **SVP and VP**: Deerwalk includes Senior Vice President for multiple fields. There is a senior vice president for clinical development, sales and marketing, and consulting services. All of them are responsible for strengthening Deerwalk in its respective fields.
- **Engineering Manager:** Engineering managers are responsible to handle individual teams and plan the various tasks in sprints. They track the tasks in each team and are responsible to hire people to their team.
- **Engineer/Developer:** Engineers and developers are the foundation of the team. They work based on the sprint plan and complete individual tasks assigned to them.

2.3. Working Domains of Organization

The working domains of the organization are:

- Enterprise Data Management: processing and utilization
- Analytics: Healthcare Benefits and value-based care
- Population Health: Care Management, Care Coordination, Clinical Decision Support
- Payment Technology: Bundles, Capitation
- Services: Managed, Consulting

2.4. Description of Intern Department/Unit

Deerwalk Plan Analytics develops software solution that provides integrated informatics and actionable insights and savings recommendations, presented in an easy-to-use application that requires no special training. By housing current and historical data, the Plan Analytics supports period-to-period comparisons and trending analysis. The teams involved in Plan Analytics department are:

2.4.1. Engine Team

The Engine Team is responsible for processing and adding business logic to the data and storing it in the elasticsearch index. The data varies from textual, numerical, geospatial, structured, or unstructured. The engine team mostly interacts with DAS Team and the Engine QA team.

2.4.2. DAS Team

Deerwalk Analytics System or DAS is responsible for taking the data from the engine and mapping it into fields that allow the data to be presentable to the front-end. DAS is responsible for performing elasticsearch queries on the data and providing a meaningful output through APIS and is used by the front-end teams. Teams associated with DAS are Engine for data input, the DAS QA team, and the Front-end team to map the incoming data to displayable User Defined Functions.

2.4.4. Front-End Team

The front-End team takes the JSON output from DAS and displays it in an interactive form and displays it in presentable form of either Bar, Charts, Tables, Reports, etc. The front-End team also allows the clients to provide various serviced logic to the data and helps the client to exactly the type of data they desire. Teams associated directly with the Front-End Team are DAS Team and its QA Team.

2.4.5. QA Team

Each of the previous teams has its own QA team and its task is to check and maintain the stability of the product/ services. Based on the feature, they build scenarios and test the feature for stability and quality. They assign any kind of feature defect, bug, or support for a specific feature to its appropriate team to check/ fix the issues. QA teams maintain the functionality of the products so that the product is ready to deploy

CHAPTER 3: INTERNSHIP ACTIVITIES

3.1 Roles and Responsibilities

The various roles and responsibilities are:

- Attend meetings organized by mentor and team members
- Complete tasks assigned by manager/mentor in the PA Engine project
- Support in daily deliverables of senior members of the team by learning the standard process used.
- Learn technologies such as Cascading, elasticsearch, AWS etc.
- Keep the Personal Health Information(PHI) and other client related data confidential
- Keep the PC provided by the company secure from damage and other users. Also, use the PC only for company related tasks only.
- Show a Positive attitude and eagerness to learn new things

3.2 Weekly Log

The following is the weekly log of tasks that I did during the internship period from the day of my hiring to the end of 3 months contract i.e., 9th May, 2022 to 8th August, 2022.

Table 3.1: Weekly Log

Week	**Activities Performed**
One (9th May, 2022 to 13th May, 2022)	● Nepal Employee Orientation, introduction to team members and different teams. ● Assigned new hire checklist, assigned mentor, brief description of company and rules and regulation by mentor. ● Attended HIPPA course and pass the HIPAA test. ● Completed US healthcare course. ● Completed secure coding training 2021.
Two (16th May, 2022 to 20th May, 2022)	● Understand Deerwalk SDLC process. ● Practiced Basic SQL Queries. ● Enrolled on Deerwalk product development manuals. ● Practiced Cascading and elasticsearch queries. ● Learned and practiced Cascading word count, split pipe, functions, filter etc

Three (23th May, 2022 to 27th May, 2022)	● Completed Plan Analytics course ● Setup hadoop, VPN, elasticsearch in PC ● Learned about existing hadoop jobs ● Ran my own hadoop job and practiced pipeline functions of cascading ● Read cascading documentation
Four (30th May, 2022 to 3rd June, 2022)	● Read product development manual document ● Further practice and understand cascading from documentation ● Learned about how to write hadoop job to elasticsearch ● Attended Communication training
Five (6th June, 2022 to 10th June, 2022)	● Attended code reviews and hotfix meetings ● Got access to project and ran local jobs ● Learned about some modules in PA Engine ● Involved in data processing of some clients and learned the process ● Learned about datasearch and visit admission modules
Six (13th June, 2022 to 17th June, 2022)	● Involved in data processing of multiple clients ● Completed all demo tasks and reviewed with mentor ● Practiced elasticsearch query ● Learned about some hadoop jobs ● Learned about visit admission and readmission modules
Seven (20th June, 2022 to 24th June, 2022)	● Involved in data processing of clients and for testing ● Re-configured PC to install java, hadoop, redis, elasticsearch due to hardware change ● Re-Ran the local hadoop jobs of PA Engine ● Learned about a demo SCR task
Eighth (27th June, 2022 to 1st July, 2022)	● Worked on multiple SCR tasks to expose fields in multiple jobs. ● Performed testing and processing of jobs for the changes ● Performed processing to test the changes made

	with SCR
Nine (4th July, 2022 to 8th July, 2022)	● Learned to perform backup processing and backup input and processing data. ● Perform some data processing for some clients ● Performed local processing with new changed data ● Learned about SCR tasks from previous sprints ● Learned to perform debugging issues in local as well as production processing.
Ten (11th July, 2022 to 15th July, 2022)	● Attended code reviews of other team members ● Performed many data processing tasks. ● Had first meeting with QA for followup on tasks ● Performed local processing with new changed data ● Prepared a TDD document for an SCR task and got it reviewed by mentor.
Eleven (18th July, 2022 to 22th July, 2022)	● Learned about many modules watching videos of previous knowledge sharing sessions. ● Performed multiple backup processing due to failures. ● Joined knowledge sharing session to learn about modules ● Update multiple layout files for the extract jobs in code and excel sheet also.
Twelve (25th July, 2022 to 29th July, 2022)	● Understanding tasks for the next sprint ● Requirement understanding of features with QA ● Performed coding for exposing fields from multiple tables and did testing of the feature ● Performed code review and batch processing to test the feature
Thirteen (1th August, 2022 to 5th August, 2022)	● Prepared a unit test for feature and show it to mentor ● Requirement discussion and understanding of task for skipping jobs and stop creating directory checking particular flags ● Performed coding and testing for the feature ● Performed code review for the task ● Attended hotfix and code review meetings of team members

3.3. Description of the Project Involved During Internship

3.3.1. PA Engine

I worked on a PA Engine project that is developed and managed by the Analytics Engine team of the Plan Analytics Department. The PA Engine project follows agile methodology to develop various modules.The core business logic is implemented on raw american healthcare data. The data is provided by the data team as parquet or csv format or the data is used from the Care Manager database which collects data in real time and provides the data in parquet or csv file. The logic is implemented using hadoop jobs that contain cascading pipelines which perform various modifications of data such as eligibility data, medical data, pharmacy data etc. The raw data is passed through many jobs in order to create meaningful health information. The hadoop jobs store the data in the hdfs directory, which is then again used by another job as input. The sequence of jobs are required to process the huge amount of data without affecting the delay in processing. The health information of members is then written in the elasticsearch index through various pipelines of various jobs through batch processing. Then, the elasticsearch index is utilized by the DAS team to create APIs for data access by the front-end team.

The sprint is of 4 weeks where the development takes place in the first 3 weeks and the last week is for regression testing. The tasks are assigned through deermine which is the in house software to assign tickets for features, bugs, support tickets, etc. The data pipelines are developed using the cascading framework on top of hadoop which is based in Java programming language. The application tool is Maven to build the project and manage the dependencies required for the PA Engine. Also, python is used for managing the jobs as a dictionary and to write various scripts to support the hadoop job. Redis is used for storing the intermediate values for comparison between various data in different pipelines. AWS is one of the integral part of data sharing in the PA Engine project. Many s3 buckets are used to receive input data for data processing and also to store the data backup by the engine team and also store the processing backup in case any hadoop jobs causes error or hadoop cluster fails due to any errors. Amazon EMR is used as the big data platform for application processing, it creates instances to run hadoop jobs and various data storage during processing, which is managed by OM software which is

inhouse software of deerwalk. While most of the data is written in flat format in the elasticsearch index, protobuf is used to store nested values in the elasticsearch index.

The technical details of the project are:

Table 3.2: Technical details of the project

Main Technology	Hadoop
Framework	Cascading
Programming Language	Java
Application Tool	Maven
Other Technologies	● Elasticsearch ● AWS ● Python ● Redis ● Protobuf
Project Management Methodology	Agile
Project Management Tool	Deermine(Inhouse Software)
Version Management	Amazon Code Commit
Length of sprint	4 weeks

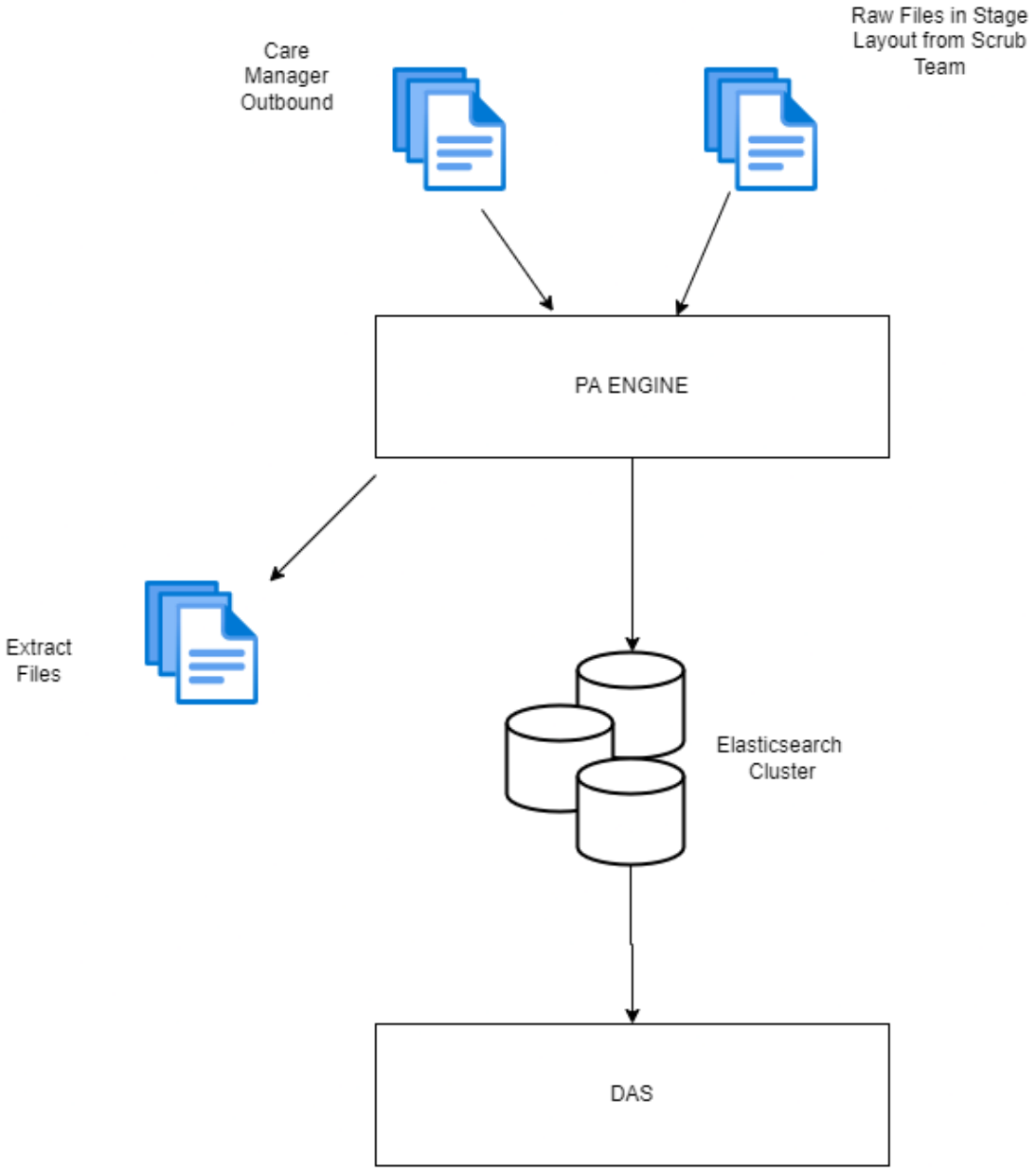

Figure 3.1: Basic Workflow of PA Engine

The PA Engine project receives data from 2 sources, The Scrub team provides data in parquet or csv format. While the Care Manager Team collects data in real time and stores it in their database. Then the data is again provided to the PA Engine in the parquet or csv format in appropriate standard layout used by the PA Engine project. The PA Engine then processes the data and writes the output in elasticsearch index or it generates extract files that are stored in s3 bucket which can be used by the client for other purposes. The

elasticsearch index is then utilized by the DAS team to generate various APIs to be used in the web application by the Front team.

Some of the major Hadoop jobs of the PA Engine project are:

- Format Validation

  The format validation jobs support Pipe delimited and parquet format data. It checks if the data is in the correct data format or not. Otherwise it stops application processing.

- Em Validation

  It removes eligibility missing records from input data. The members are eligible if they fall within the date range required for calculation, which can be possible if they are insured or paying for the service.

- DataSearch

  Search and filter client-provided data from elasticsearch functionality. It adds functionality of drill down from reports. Also, custom fields like age, weights, etc are added.

- Engine Extracts

  Applying business logic and generating extracts on a given layout. The data can be uploaded to s3 bucket.

- Third party integration

  Integrate third party engines such as neural network for Risk scores, and machine learning in membersearch

- Elasticsearch jobs

  Define mapping for the data in the elasticsearch index and create analyzers for the various fields.

3.4. Tasks / Activities Performed

The various tasks and activities performed during the internship are:

3.4.1. Orientation and team introduction

1. Orientation
   a. Team overview and work areas of the team by mentor
   b. Added to 'R&D Daily Update' Teams chat group where daily updates needed to be written after the end of the day.

2. Introduction to the team
   a. Introduction to developer team including the team leads
   b. Introduction to QA team

3. Human Resource Tool
   a. Punch in before 9 AM and punch out after 6 PM
   b. Leave options such as work from home, sick leave, floating leave and inventory management were explained by mentor

4. Email Setup
   a. Setup email signature with phone, email and position using a cedargate email provided by IT team
   b. Learned about the email etiquette on how to address, compose and effective email throughout the organization for effective communication

5. Microsoft Teams/ Teams Emails Setup
   a. Setup teams with the cedargate email in the PC and added to other chat groups of the team.
   b. Added contacts to team account and communicated with the team members via chat

6. Work from home guidelines
   a. Explained about the timing of work from home along with the various Dos and Don'ts

7. Company Introduction
   a. Overview of the company along with its work areas and its various teams were explained by the mentor
   b. Browsed Deerwalk site to learn more about services provided by Deerwalk.

8. Deerlms overview
   a. Learned about the overview of Deerlms site where various training courses are available to the employees of Deerwalk where you can earn points for every course completed

9. HIPAA Guidelines
   a. Provided with high-level HIPAA overview by mentor and also informed about Dos and Don'ts with organization data and infrastructure
   b. Guided how to learn HIPAA course via Deerlms and provided with a deadline to pass the HIPAA exam by the HR team.
   c. Also provided with guidelines on any confusion and concerns about the course and HIPAA overall

10. Addition to group email
   a. Added to team group chat for communication such as Developer chat group, Engine chat group etc.
   b. Added to unit/ product distribution list
   c. Described about team group email/distribution list
   d. Create a gmail account using cedargate email

3.4.2. US Health Care and HIPAA course

1. HIPAA Course
   a. Complete HIPAA course via Deerlms website where various questions regarding the American health care data and its privacy policy were asked
   b. Gave HIPAA test via deerlms and scored 80% to pass the test

c. Explained about effort log procedure by the mentor via various tickets for different purposes such as meetings, coding, testing, etc.

2. US Healthcare Course
   a. Understand terms used in US Healthcare
   b. Completed US Healthcare course from deerlms.

3.4.3. Product/ Unit Specific training

1. Product Overview
   a. Product overview was given by the mentor where the tools and technologies were explained along with basic working principles of Product
   b. Deerwalk data flow was explained about how data is parsed from source and processed throughout the Plan analytics platform.
   c. Inter product dependencies were also explained where roles of each team were explained

2. Secure coding training 2021
   a. Enroll for the course 'Secure Coding Training 2021' from deerlms.
   b. Also explained about the principles to keep the PC secure and best practice to prevent malware and unauthorized access

3.4.3. SDLC and Dev Manual

1. Deerwalk SDLC Process

   Read through SDLC document that explains the process being followed in Deerwalk

2. Deerwalk Product Development Manual

   Go through the development manual and get familiar with the processes involved by various members of team for software development and document them to be

understood in future

3. Basic SQL

   Read through all the slides of basic SQL courses where concepts of joins, subquery, etc. were explained.

3.4.4. Demo project

1. Installation
   a. Install IntelliJ Idea Ultimate version
   b. Install java 1.8, elasticsearch 7.2
   c. Install Hadoop 3.2 and cascading 3 and configure various files such as mapred-site.xml, hdfs-site.xml, core-site.xml and also prepare client directory with various configuration files.

2. Learn Basic technologies for Demo Project
   a. Learn Hadoop, cascading from a slide and various internet resources for better understanding
   b. Learn Elasticsearch with a major focus on querying the data.
   c. Learn how to read/write data from hdfs/to elasticsearch using the cascading API of Hadoop.

3. Demo project task
   a. Created a demo cascading project of following tasks:
      - Do the word count of random texts or lyrics.
      - Demerge a single pipe to mPipe and fPipe on the basis of gender “M” and “F”.
      - Calculate and store age on the basis of provided DOB in source file.
      - From multiple records having the same SSN, take only one tuple with the longest name.
      - Filter out the records if ssn is not 9 digits. Source Fields must contain: ssn, name.

- From a record of mbr_id and paid_amount fields, if there are multiple records of the same mbr_id, then sum up the paid_amount fields.

b. Presented it to leads, mentor, and manager
c. Accommodate any feedbacks/ enhancements suggested by leads, mentor, and manager

4. Demo Hadoop job
   a. Study a simple PA Hadoop job about how the pipelines are created and the input directories for the pipe. Also, the major focus was on dependencies of previous jobs that passed out the output as input for that particular input.
   b. Create own PA Hadoop job having:
      i. A job config file that has details such as input directory, output directory and the input and output layout files
      ii. Register job in the job dictionary with its loader name
      iii. Create a pipeline in the loader class making the required changes by operations such as grouping, buffer, etc.
   c. Accommodate any feedbacks/ enhancements suggested by leads, mentor, and manager

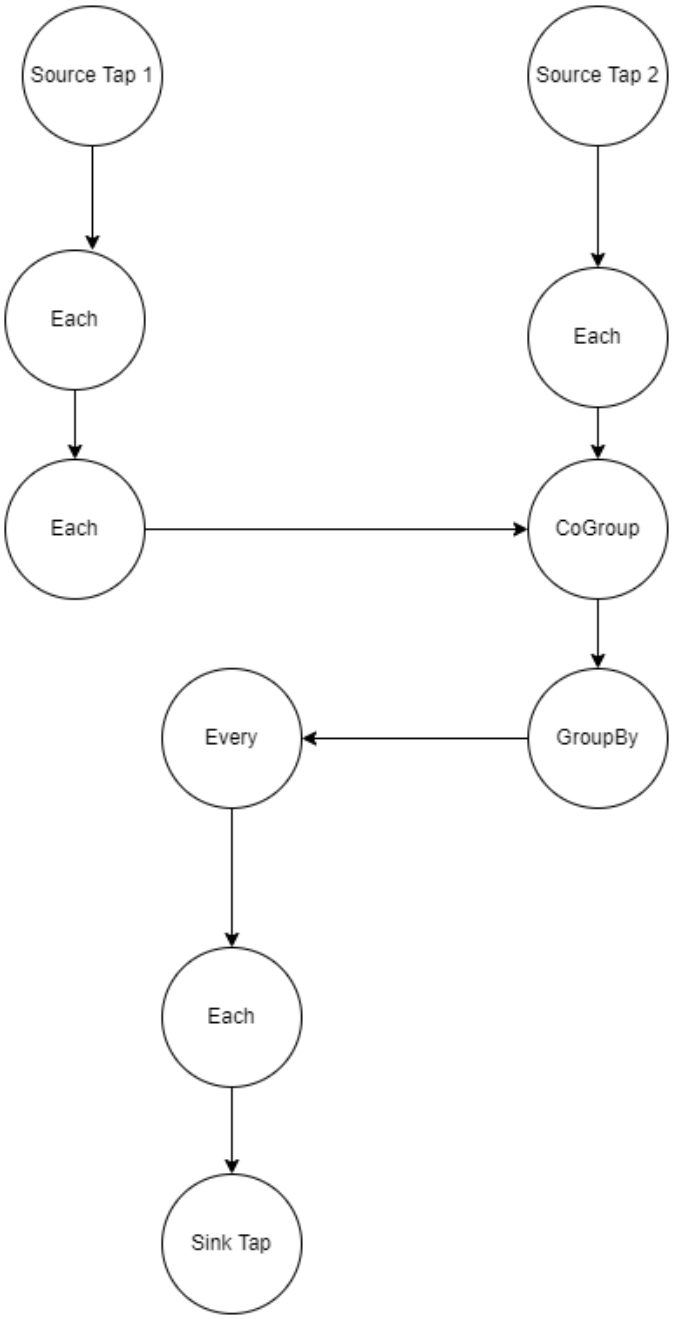

Figure 3.2: Basic Cascading Pipeline

3.4.5. Create Feature/ Bug/Support Branch

1. Create a separate branch to test bug and features
2. Create a new branch from a branch and verify it with the source branch
3. Write a verification email and provide the sample

3.4.6. Learn Git pull request

1. Get familiar with git pull request
   a. Send a sample pull request
   b. Send a sample pull request email
   c. Resolve merge conflict
2. Get familiar with code check-in
   a. Learn about the cause of the issue and impact areas
   b. Send a test code checkin email to mentor, lead and manager

3.4.7. Prepare Technical Design Document

1. Prepare a Technical Design Document of sample feature
   a. Prepare a Technical Design Document following the format provided in the Dev manual including:
      i. various test cases that need to be performed
      ii. Impact analysis on existing code
      iii. Files that needs to be changed.
   b. Send the sample Technical Design Document to lead and manager for signoff after completion
   c. Accommodate any feedbacks/ enhancements suggested by leads, mentor, and manager

3.4.8. Requirement Review, Analysis and Estimation

1. Get familiar with requirements review, analysis and estimation of features where time for requirement analysis, understanding, coding, testing and code review are included in a sprint plan
2. Performed requirements review and analysis of test feature that was assigned to me in sprint plan
3. Provided the detailed estimation of a small client request where the effort was minimal.

3.4.9. Development Environment Setup

1. Project Access
   a. Got access to repository in amazon code commit and cloned the PA Engine project on my local PC and built the project using maven :
   mvn clean package -Dskiptests

2. Learn about Project
   a. Learned Engine overall workflow by going through the jobs and their configurations.
   b. Learn how to read/write data from hdfs/to elasticsearch from the already available elasticsearch job in the PA Engine.
   c. Understanding AWS and its uses in PA engine in various jobs as well as in the processing of the jobs.

3. 4.10. Learn and perform Data Processing

1. Learn Application processing process of various clients by the help of mentor and other team members by involving in meetings.
2. Involved in Application processing task by following the steps below:
   i. Get a processing ticket along with input s3 location of the data from the team lead.
   ii. Get list of client configurations for that particular client

iii. Initiate processing after getting the processing ticket using the required number of nodes, specific regions, master storage etc. in the Processing dashboard of deerwalk. The application allocates storage locations and processing framework automatically.

iv. switch to specific development or feature branch using git checkout and build the project using maven

v. load the data into hdfs from the input s3 bucket into client specific folder

vi. add the various configurations to the client_config file that is required by that specific client or the configurations required to test a feature of a bug.

vii. select the jobs that are required to run for particular feature or for particular client

viii. Mail about initiation of the processing to all member of plan analytics department

ix. Monitor the processing in case it gets halted due to some error in code or due to technical error.

x. After completion of processing mail with the snapshot details for the other teams to test the results.

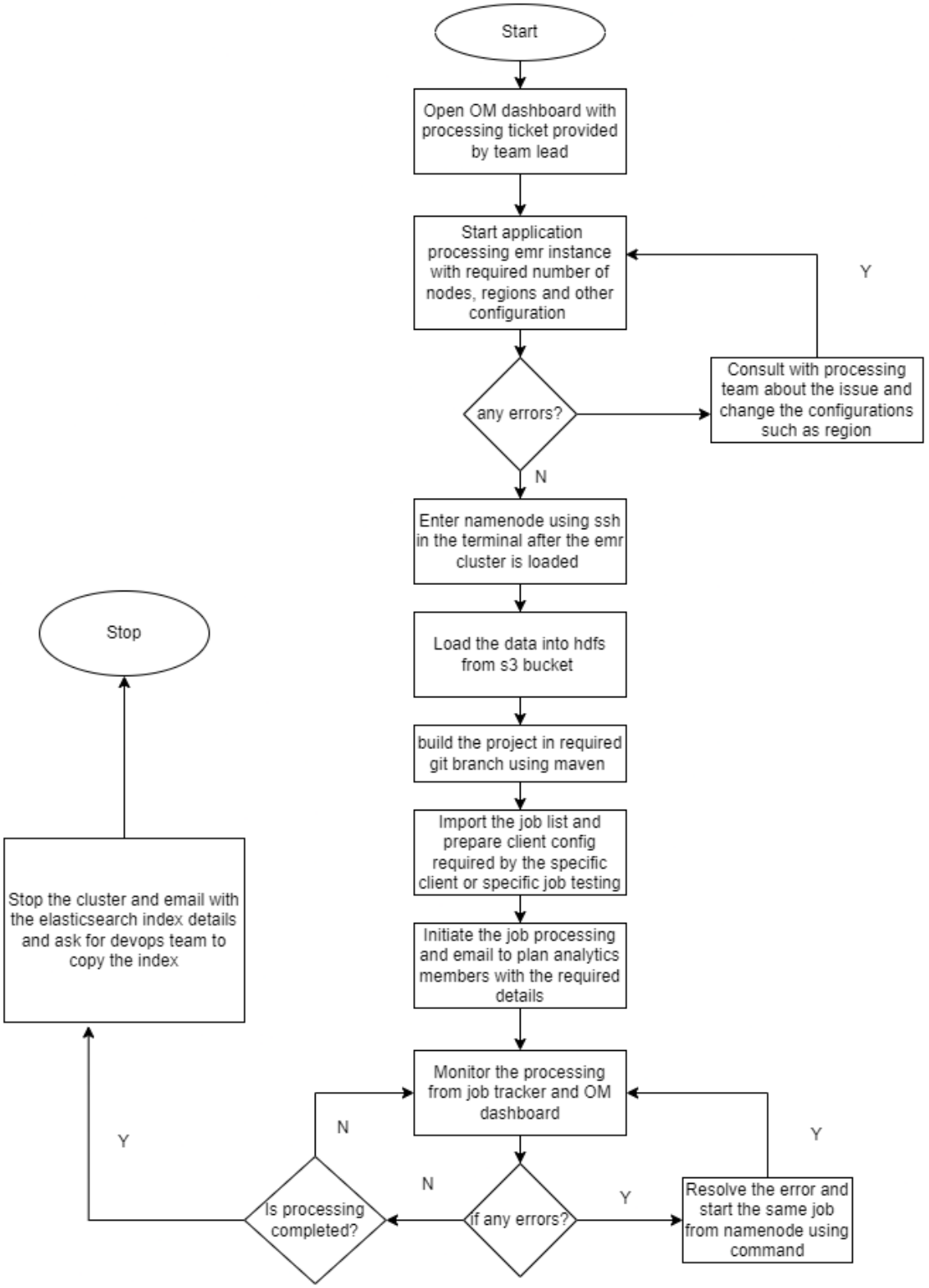

Figure 3.3: Data Processing Steps

3.4.11. Minor Bug fixes/ SCR

1. Learning SCR
   a. Learn SCR processes where requirements such as exposing fields, adding analyzers to the fields, making the fields common across various jobs.

2. Start SCR

   Start being involved in SCR code changes by looking into previous commits and documents prepared by other team members for the similar task.

3. Assigned minor maintenance bugs of minimum effort supports
   a. Assigned 4 minor bugs with ticket details and provided with the fixes by the mentor
   b. Performed code review with the team after making the changes before sending the first merge request mail

3.4.12. Check-in mail and deermine update

1. Have the check-in mail reviewed with the mentor after completing the tasks provided by the mentor
2. Update the determine with the details of the changes by following the steps mentioned in the Dev manual

3.4.13. Feature Development

1. Feature Discussion with QA and BA team
   a. Attend the scheduled meeting involving development team, QA and BA team members.
   b. Carefully understand the requirements of the feature along with the impact areas of the required changes.
   c. Raise any issue that the feature might affect an already existing feature or any modules.

2. Understanding and Development
    a. Understand the module where the development is to be done and understand its dependencies with other modules.
    b. Make necessary changes to the module or add new module as required by the feature

3. Testing and Code review
    a. Make unit tests and integration tests for the module and test it with provided test data with the feature ticket.
    b. Perform code review for the code with team members and optimize the code if any required after review.
    c. Write a merge request mail to team members to verify the code changes
    d. Merge the feature branch to the major branch and prepare a code checkin mail for the management team and other teams.

4. Application processing and testing
    a. After completion of data processing(batch processing), get the elasticsearch index where the output of processing is stored.
    b. Test the changes that you made by querying the index with the required fields of that particular module.

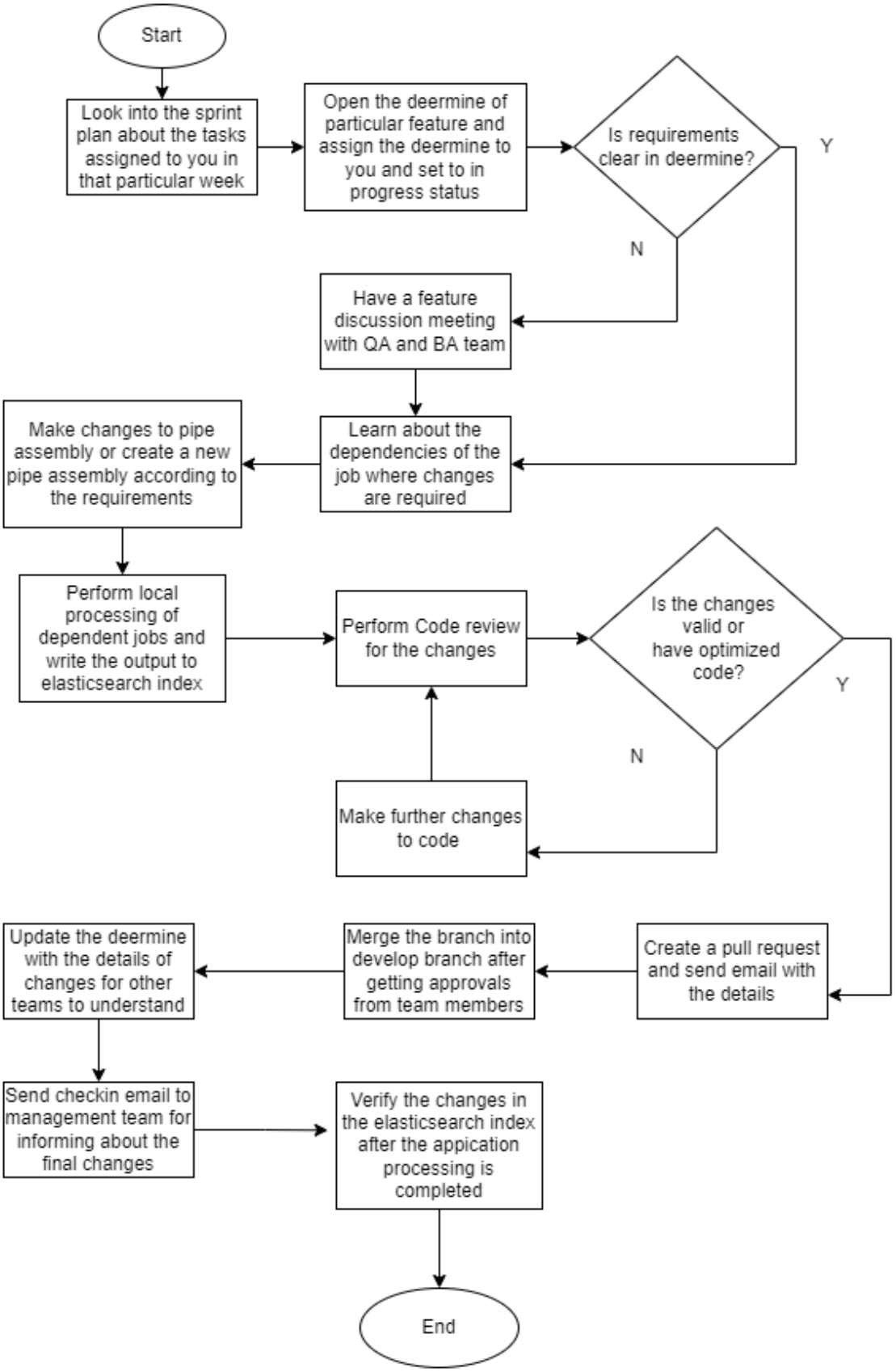

Figure 3.4: Steps in feature development

3.4.14. Sample Code Prototypes

- Hadoop job with Cascading pipe

```
public class Main {

    public static void main(String[] args) {

        String sourcePath =
"src/main/resources/cascading/example/function/function_source.txt";

        String sinkPath = "target/cascading/function/function_sink.txt";

        Tap sourceTap = new FileTap(new TextDelimited(new Fields("id",
"full_name"), ";"), sourcePath);

        Tap sinkTap = new FileTap(new TextDelimited(true, ";"), sinkPath,
SinkMode.REPLACE);

        Pipe pipe = new Pipe("nameSplitPipe");

        Fields newFields = new Fields("first_name", "middle_name", "last_name");

        pipe = new Each(pipe, new Insert(newFields, "", "", ""), Fields.ALL);

        pipe = new Each(

                pipe

                , newFields.append(new Fields("full_name")) //full_name field
appended in newFields

                , new NameSplitFunction()

                , Fields.REPLACE

        );

        pipe = new Discard(pipe, new Fields("full_name"));
```

```
        FlowDef flowDef = FlowDef
                .flowDef()
                .addSource(pipe, sourceTap)
                .addTailSink(pipe, sinkTap);
        Flow flow = new LocalFlowConnector().connect(flowDef);
        flow.complete();
        System.out.println("Process completed.\nPlease visit: \n" + sinkPath);
    }
```

- Cascading function

```
    public class NameSplitFunction extends BaseOperation implements Function {
        public NameSplitFunction() {
            super(Fields.ARGS);
        }

        @Override
        public void operate(FlowProcess flowProcess, FunctionCall functionCall) {
            TupleEntry tupleEntry = functionCall.getArguments();
            TupleEntry tuple = new TupleEntry(tupleEntry);
            String[] name = tupleEntry.getString("full_name").split("\\s+");
    //logic applied
```

```
        tuple.setString("first_name", name[0]);

        String middleName = "";

        for(int i = 1;i<name.length -1 ; i++){

            middleName = middleName + name[i] + " ";

        }

        tuple.setString("middle_name", middleName.trim());

        tuple.setString("last_name", name[name.length - 1]);

        functionCall.getOutputCollector().add(tuple);  //sends resulting tuple to the
pipe stream

    }

}
```

CHAPTER 4: CONCLUSION AND LEARNING OUTCOMES

4.1. Conclusion

The three months of internship period at Deerwalk Services has been very much valuable and fruitful to gain practical knowledge of Software Development and Big Data Solutions. It was a wonderful opportunity to test my knowledge and skills in the real world. The internship period provided me with a clear idea of various steps that are required to convert a business requirement into a working part of the software. Not only the technical aspect of the sector but also I learned about corporate culture and team collaboration. The internship has also helped me to develop my interpersonal skills like speaking and writing skills and time management for performing my day-to-day tasks also.

During my internship, I could relate many things of the development process to that I have learned academically in the BSC CSIT course. I feel that the internship provided me with a great platform to learn and grow professionally utilizing my already existing knowledge and learning great many things with many friendly people in the company.

4.2. Learning Outcome

During my internship, the technical tasks have helped me to learn the best practices of programming and debugging skills. I also gained hands-on experience with the various BigData technologies such as the Hadoop platform and Enterprise Batch processing which I had gained theoretical knowledge from BSC. CSIT course.

Apart from the technical knowledge, the time management skills and interpersonal skills gained from the internship will definitely help me in other areas of my life. I learned that teamwork and hard work with proper planning are necessary to accomplish any complex task of software development. The internship also helped to grow my professional network after working with many people in the organization.

REFERENCES

Deerwalk. (n.d.). Retrieved July 13, 2022, from https://www.deerwalk.com/ .

Deerwalk plan analytics. (n.d.). Retrieved July 26, 2022, from https://www.deerwalk.com/pdf/plan-analytics.pdf

(OCR), O. for C. R. (2022, June 29). *Hipaa Home*. HHS.gov. Retrieved September 13, 2022, from https://www.hhs.gov/hipaa/index.html

Platform. Cedar Gate Technologies. (2022, January 19). Retrieved August 02, 2022, from https://www.cedargate.com/platform/

.

APPENDIX

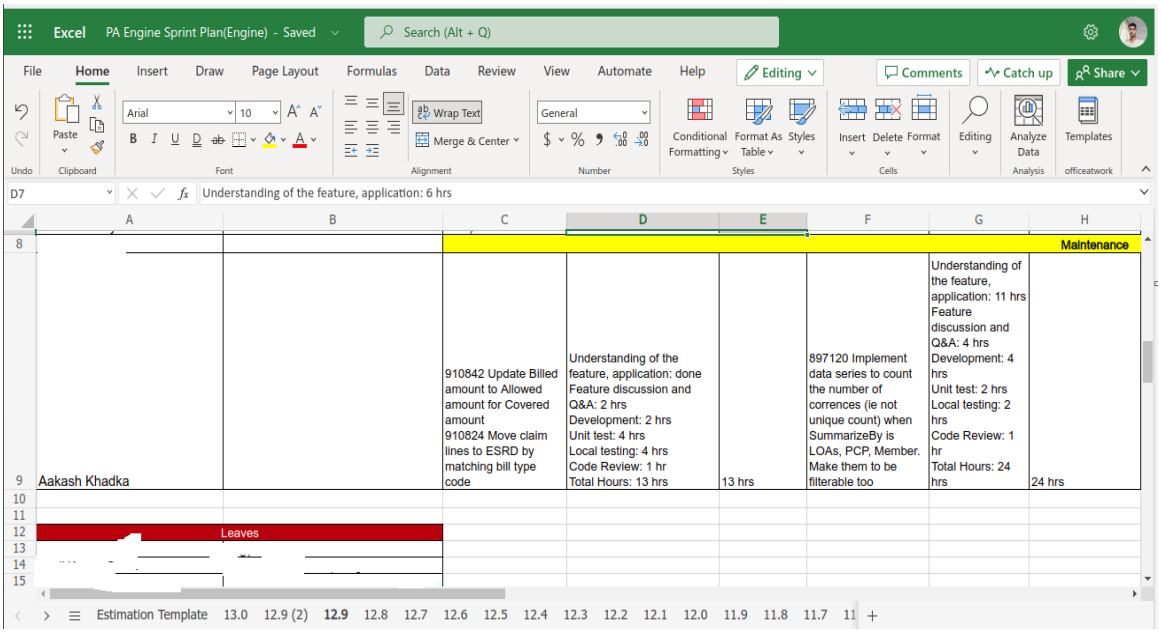

Figure A.1: Sprint Planning

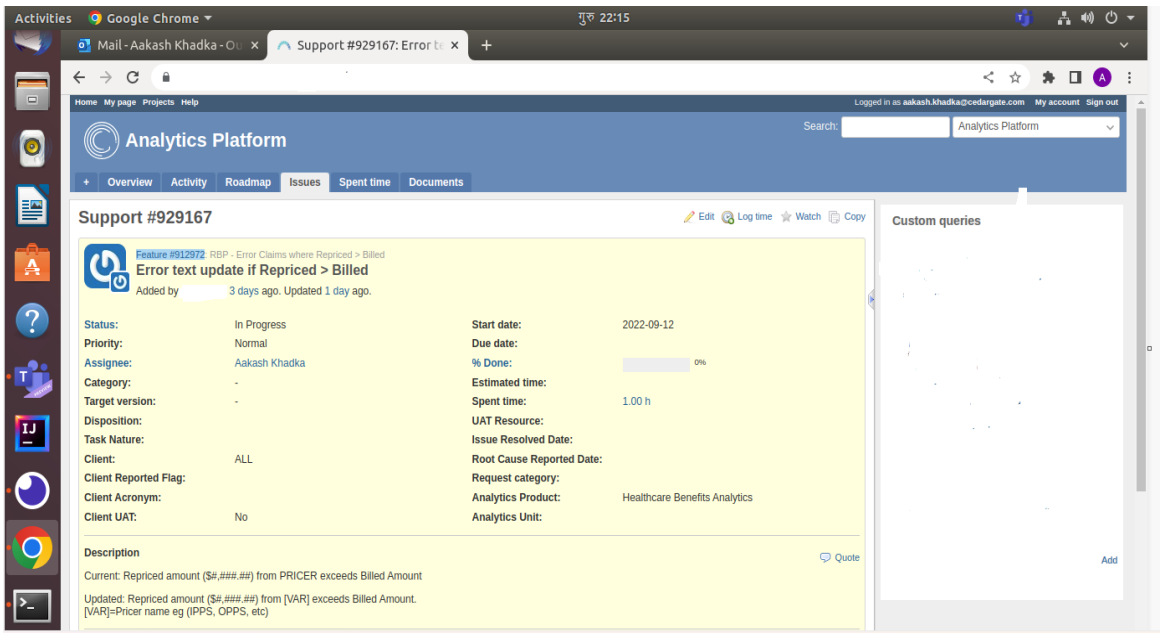

Figure A.2: Deermine ticket

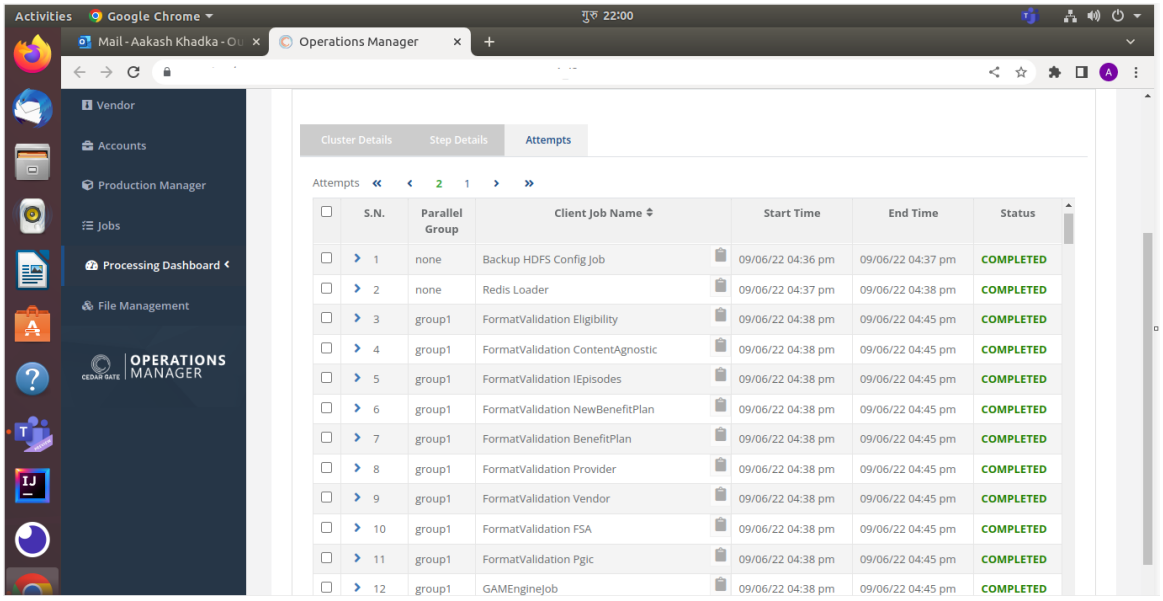

Figure A.3: Data Processing through OM Dashboard|

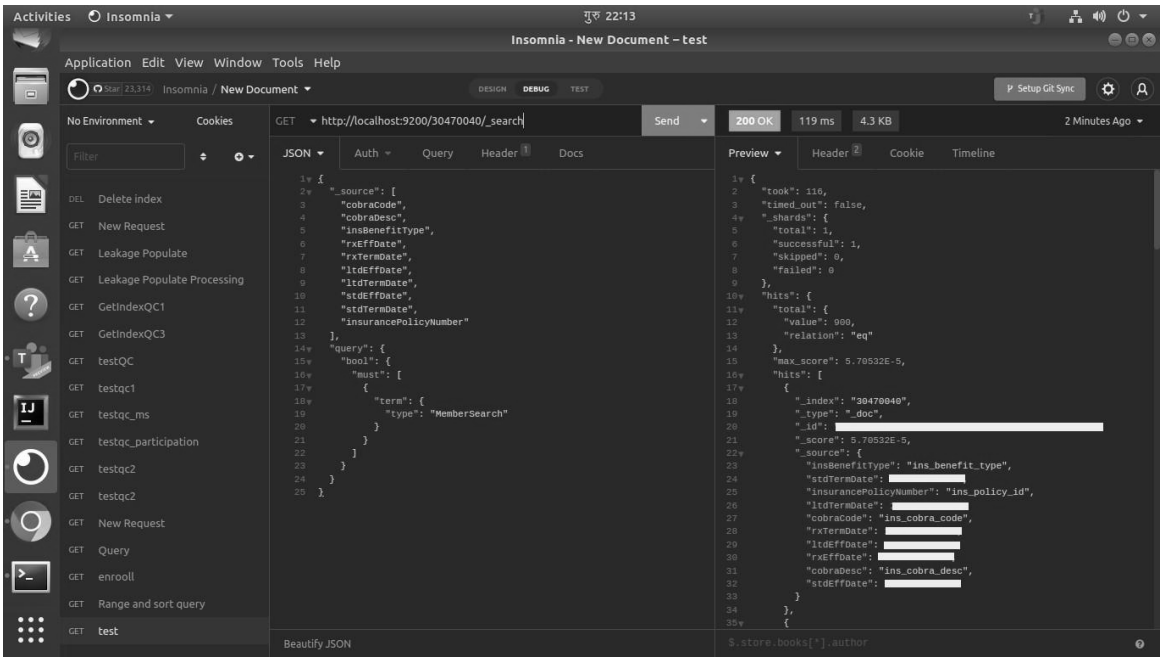

Figure A.4: Sample Elasticsearch query

Main.java

src > main > java > cascading > example > join > Main.java > Main > main

```
        * Join source1 and source2 on the basis of matching id.
        */

        String source1Path = "src/main/resources/cascading/example/join/join_source1.txt";
        String source2Path = "src/main/resources/cascading/example/join/join_source2.txt";
        String sinkPath = "target/cascading/join/join_sink.txt";

        Tap source1Tap = new FileTap(new TextDelimited(true, "|"), source1Path);
        Tap source2Tap = new FileTap(new TextDelimited(true, ";"), source2Path);
        Tap sinkTap = new FileTap(new TextDelimited(true, "|"), sinkPath, SinkMode.REPLACE);

        Pipe pipe1 = new Pipe("pipe1");
        Pipe pipe2 = new Pipe("pipe2");

        Fields commonFields = new Fields("id");
        Fields declaredFields = new Fields("id", "name", "phone", "id2", "address");
        Pipe finalPipe = new CoGroup(pipe1, commonFields, pipe2, commonFields, declaredFields, new LeftJoin());
        finalPipe = new Discard(finalPipe, new Fields("id2"));

        FlowDef flowDef = FlowDef
                .flowDef()
                .addSource(pipe1, source1Tap)
                .addSource(pipe2, source2Tap)
                .addTailSink(finalPipe, sinkTap);
        Flow flow = new LocalFlowConnector().connect(flowDef);
        flow.complete();

        System.out.println("Process completed.\nPlease visit: \n" + sinkPath);

    }
}
```

Figure A.5:Cascading Pipe

GenderFilter.java

src > main > java > cascading > example > filter > GenderFilter.java > ...

```
public class GenderFilter extends BaseOperation implements Filter {
    @Override
    public boolean isRemove(FlowProcess flowProcess, FilterCall filterCall) {
        TupleEntry tupleEntry = filterCall.getArguments();
        String gender = tupleEntry.getString("gender");
        return !(gender.equals("M") || gender.equals("F"));
    }
}
```

Figure A.6: Cascading Buffer