**Low Level Document**

**SQL-SQOOP-DATA-INGESTION-ETL-PROJECT**

Aman Yadav

Email: amanyadav18072000@gmail.com

**Contents**

1. Introduction

1.1. What is Low-Level design document?

1.2. Scope

2. Architecture

3. Architecture Description

3.1. AWS RDS

3.2. AWS EMR

3.3. AWS EC2 Apache Airflow

3.4. Data Insertion into Database

3.5. Export Data from AWS RDS to AWS EMR (Hadoop Cluster)

3.6. Scheduling

1. **Introduction**

**1.1. What is Low-Level design document?**

The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the

actual program code for Food Recommendation System. LLD describes the class diagrams with the

methods and relations between classes and program specs. It describes the modules so that the

programmer can directly code the program from the document.

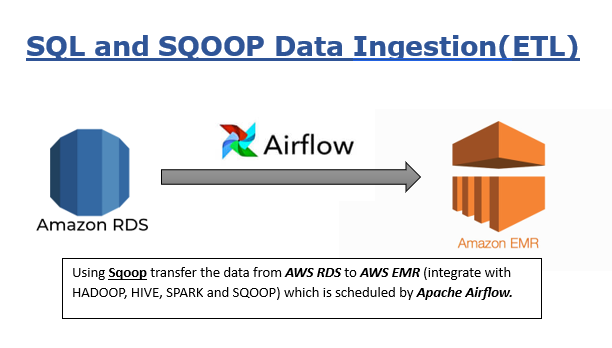
**1.2. Scope**

Low-level design (LLD) is a component-level design process that follows a step-bystep refinement process. This process can be used for designing data structures, required software

architecture, source code and ultimately, performance algorithms. Overall, the data organization

may be defined during requirement analysis and then refined during data design work

**2. Architecture**

****

*AWS EMR INTEGRATE WITH HADOOP CLUSTER*

*AWS RDS WITH MYSQL ENGINE*

*TRANSFER THE STRUCTURED E-COMMERCE DATA THROUGH SQOOP*

*APACHE AIRFLOW WHICH SCHEDULE THE WHOLE JOB ON DAILY BASIS*

OR

**3.Architecture Description**

**3.1 AWS RDS**

Set up the MySQL on AWS RDS by enable public access “Yes” and add security group so that it can connect to MySQL Workbench.

**3.2 AWS EMR**

Set up the AWS EMR integrate with Hadoop, spark, Hive and Sqoop and select the m5.xlarge machine for cluster.

And add one bootstrap action to download the MySQL connector into Hadoop cluster for Sqoop import command.

**3.3. AWS EC2 Apache Airflow**

In this step create another EC2 instance on AWS for Apache Airflow.

After creating the instance now install the Apache airflow and python.

**3.4. Data Insertion into Database**

In this step first create the E-commerce database and table schema which is available in the git hub repository: SQL filename ***TableSchema.sql.***

Then insert the data into respective tables which is available in the git hub repository: filename ***TableRecors.sql***

**3.5. Export Data from AWS RDS to AWS EMR (Hadoop Cluster)**

Now run the Sqoop command which available in the git hub repository with the file name ***E-commerce-dag.py.***

**3.6. Scheduling**

Now the most important step is to schedule the Sqoop job on daily basis which is handle by Apaches airflow running on AWS EC2 instance.