**Capstone Project**

**Storing Source Datasets to HDFS:**

1. hdfs dfs -put StudentCourseDetails.csv /user/ana002682/ArindamDataSource;
2. hdfs dfs -put StudentCourseCompletionStatus.csv/user/ana002682/ArindamDataSource;

**Setting Warehouse Location and creating Database for Project:**

* set hive.metastore.warehouse.dir = /user/ana002682/warehouse;
* create database ArindamDB;

**TASK 1 :**

1. **1.a)** Create two internal Hive tables for the provided dataset –
   1. **StudentCourseCompletionStatus** -
      1. Distribute **StudentCourseCompletionStatus** data randomly into 5 buckets based on **marks** column
      2. Table Schema - studentsid string, courseid string, examdate date, attendedstatus string, marks int , result string

**Query:-**

CREATE TABLE IF NOT EXISTS ArindamDB.course\_completion(

studentsid string,

courseid string,

examdate string,

attendedstatus string,

marks int ,

result string)

CLUSTERED BY (marks) INTO 5 BUCKETS

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',';

load data inpath '/user/ana002682/ArindamDataSource/StudentCourseCompletionStatus.csv' into table course\_completion;

**1.b) CourseDetails** -

* + 1. Split the **CourseDetails** table into two partitions using dynamic partitioning based on the **CourseType** column
    2. Table schema - coursetype string, courseid string, title string, competency string, complexity string

**Query:-**

CREATE TABLE IF NOT EXISTS ArindamDB.course\_details(

title string,

courseid string,

competency string,

complexity string)

PARTITIONED BY (coursetype string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',';

load data inpath '/user/ana002682/ArindamDataSource/StudentCourseDetails.csv' into table ArindamDB.course\_details;

**TASK 2:**

1. List the count of students qualified for various courses:

**Query:-** select count(\*) from ArindamDB.course\_completion where result='Qualified';

**Output:-**

Text

Description automatically generated

1. List the courses available.

**Query:-** select distinct coursetype from ArindamDB.course\_details;

**Output:**-

Text

Description automatically generated

1. Extract the following fields from **StudentCourseCompletionStatus** file from HDFS directory and save the following result set to HDFS output directory -
   1. StudentID, b. CourseTitle, c. Result

**Query:-** INSERT OVERWRITE DIRECTORY "/user/ana002682/ArindamOutputDirectory" ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT a.studentsid,b.title as coursetitle,result from ArindamDB.course\_completion a INNER JOIN ArindamDB.course\_details b ON a.courseid=b.courseid;

**Output:-**

A picture containing text

Description automatically generated