BIG DATA ASSIGENEMT

Name: Jebastin P
Emp ID: 2401171
Cohort Code: INTAIA25GCP003

Date: 09-05-2025

Question 1: Create a Hive table load and load the csv file.

Hive Query:

```
CREATE TABLE IF NOT EXISTS students_data (
id INT,
name STRING,
nationality STRING,
city STRING,
gender STRING,
age INT,
english.grade FLOAT,
math.grade FLOAT,
sciences.grade FLOAT,
language.grade FLOAT,
portfolio.rating FLOAT,
coverletter.rating FLOAT,
refletter.rating FLOAT )
PARTITIONED BY (city STRING)
CLUSTERED BY (id) INTO 4 BUCKETS
ROW FORMAT DELIMITED
```

FIELDS TERMINATED BY ”,”

STORED AS TEXTFILE;

Question 2: Create partition and bucket on the table.

Hive Query:

LOAD DATA INPATH ‘"C:\Users\2401171\Downloads\student-dataset.csv’ INTO TABLE students_data ;

Question 3: Show how many partitions created for the table.

Hive Query:

SHOW PARTITIONS student_data;

Question 4: Find top 3 students per city based on the average mark and rating.

Hive Query:

SELECT city, name, AVG((english.grade + math.grade + sciences.grade + lan guage.grade) / 4) AS avg_marks, AVG((portfolio.rating + coverletter.rating + r efletter.rating) / 3) AS avg_rating

FROM student_data

GROUP BY city, name

ORDER BY avg_marks DESC, avg_rating DESC

LIMIT 3;

Question 5: Identify the gender with the highest math mark in each city.

Hive Query:

SELECT city, gender, MAX(math.grade) AS highest_math_mark

```
FROM student_data
GROUP BY city, gender;
```