1. ***A) Create a schema for your data from the raw data.***

***B) Using your schema, load the data from Azure Storage into Hive. Assume that the data is in a container called “employee”. Call the table, “employee”.***

1. ***Validate that the schema is correct using the “table view” tab.***
2. ***Inspect the data were loaded properly using the SQL query interface.***

CREATE EXTERNAL TABLE employee

(

Age int,

Attrition string,

BusinessTravel string,

DailyRate int,

Department string,

DistanceFrom int,

Education int,

EducationField string,

EmployeeCount int,

EmployeeNumber int,

EnvironmentSatisfaction int,

Gender string,

HourlyRate int,

JobInvolvement int,

JobLevel int,

JobRole string,

JobSatisfaction int,

MaritalStatus string,

MonthlyIncome int,

MonthlyRate int,

NumCompaniesWorked int,

Over18 string,

OverTime string,

PercentSalary int,

PerformanceRating int,

RelationshipSatisfaction int,

StandardHours int,

StockOptionLevel int,

TotalWorkingYears int,

TrainingTimesLastYear int,

WorkLifeBalance

YearsAtCompany string,

YearsInCurrentRole int,

YearsSinceLastPromotion int,

YearsWithCurrManager int

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

STORED AS TEXTFILE LOCATION '/employee'

TBLPROPERTIES("skip.header.line.count"="1");

1. ***a) Create a new table for your analysis called “employee\_sales”.***
2. ***Load the table “employee” into this table.***

***c) Select these columns: Attrition, Department, JobSatisfaction & MonthlyIncome.***

CREATE TABLE employee\_sales

(

Attrition string,

Department string,

JobSatisfaction int,

MonthlyIncome int

);

INSERT OVERWRITE TABLE employee\_sales

SELECT Attrition, Department, JobSatifaction, MonthlyIncome

FROM employee;

***2) Round the data found in the “MonthlyIncome” column to the nearest $1000. (HINT: the SQL function to round a number is ROUND(obs, -3))***

INSERT OVERWRITE TABLE employee\_sales

SELECT Attrition, Department, JobSatifaction, ROUND (MonthlyIncome, -3) AS MonthlyIncome

FROM employee\_sales;

***Check:***

SELECT \* FROM employee\_sales LIMIT 10;

***3) Filter the data to only look at those items in the “Sales Department”.***

INSERT OVERWRITE TABLE employee\_sales

SELECT \*

FROM employee\_sales

WHERE Department LIKE "%sales%";

***Check:***

SELECT \* FROM employee\_sales LIMIT 10;

***4) Order the data by “JobSatisfaction” from highest to lowest. (HINT: Use the DESC query)***

INSERT OVERWRITE TABLE employee\_sales

SELECT \*

FROM employee\_sales

ORDER BY JobSatisfaction DESC;