**Assignment: Assignment #2  
Name: Kenatu Bruce  
Student ID: C0770979  
Course ID: DAT1030**

Assignment Objective: Data Transform Activity  
For the business problem you selected, you will extract data for further analysis using a series of transformations, including: select, mutate, filter, and arrange - **Human Resources Analysis - Data**

1. **Create Schema**

CREATE EXTERNAL TABLE employee  
(  
Age int,  
Attrition int,  
BusinessTravel string,  
DailyRate int,   
Department string,  
DistanceFromHome 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 char,  
OverTime string,  
PercentSalaryHike int,  
PerformanceRating int,  
RelationshipSatisfaction int,  
StandardHours int,  
StockOptionLevel int,  
TotalWorkingYears int,  
TrainingTimesLastYear int,  
WorkLifeBalance int,  
YearsAtCompany int,  
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. **Verify Data Load**

SELECT \* FROM employee LIMIT 10;

1. **Upload file to Azure Storage**

hadoop fs -copyFromLocal WA\_Fn-UseC\_-HR-Employee-Attrition.csv /example/data/employee/hrdata.csv

1. **Create a new table for your analysis called “employee\_sales”.**

CREATE TABLE employee\_sales  
(  
Age int,  
Attrition int,  
BusinessTravel string,  
DailyRate int,   
Department string,  
DistanceFromHome 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 char,  
OverTime string,  
PercentSalaryHike int,  
PerformanceRating int,  
RelationshipSatisfaction int,  
StandardHours int,  
StockOptionLevel int,  
TotalWorkingYears int,  
TrainingTimesLastYear int,  
WorkLifeBalance int,  
YearsAtCompany int,  
YearsInCurrentRole int,  
YearsSinceLastPromotion int,  
YearsWithCurrManager int  
);

* 1. **Load the table “employee” into this table.**
  2. **Select these columns: Attrition, Department, JobSatisfaction & MonthlyIncome**

INSERT OVERWRITE TABLE employee\_sales

SELECT Attrition, Department, JobSatisfaction, MonthlyIncome

FROM employee;

1. **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, JobSatisfaction, ROUND (MonthlyIncome, -3) as MonthlyIncome

FROM employee\_sales;

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

INSERT OVERWRITE TABLE employee\_sales

SELECT \*

FROM employee\_sales  
WHERE LOWER (Department) LIKE “%sales department%”;

1. **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;

.