# Data Transformation

### load.project()

### \*1. Create a schema for the data from the raw data. Load the data into Hive. Call the table, “employee”.

CREATE EXTERNAL TABLE employee

(

Age int,

Attrition string,

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 string,

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");

## \*2. Create a new table for your analysis called “employee\_sales”.

CREATE TABLE employee\_sales

(

Attrition string,

Department string,

JobSatisfaction int,

MonthlyIncome int,

);

#### \*a) Load the table “employee” into this table.

INSERT OVERWRITE TABLE employee\_sales

SELECT Attrition, Department, JobSatisfaction, MonthlyIncome

FROM employee

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

INSERT OVERWRITE TABLE employee\_sales

SELECT attrition, Department, JobSatisfaction, MonthlyIncome

FROM employee\_sales

### \*3. 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

### \*4. Filter the data to only look at those items in the “Sales Department”.

SELECT \*

FROM employee\_sales

WHERE Department LIKE "%Sales%"

### \*5. Order the data by “JobSatisfaction” from highest to lowest.

SELECT \*

FROM employee\_sales

ORDER BY JobSatisfaction DESC