IT Training, IT Staffing and Solution Services

Website: www.itexps.com
Email: info@itexps.com **Phone:** 847-350-9034
Address: 951 N Plum Grove Rd, Suite A, Schaumburg IL 60173

BIG DATA Project

Project Objective

The main objective of this project is to provide students an opportunity to leverage their learnings of the BIG DATA class, and apply their skills into real life problems.

Project Requirements

- Students are encouraged to work in a Group Setting
- Students are encouraged to use appropriate filenames, table names and column names
- Students should follow the given records and data **AS-IS** [DO NOT change the case, space, duplicate data, and empty rows ...etc) , because these values are used to create a complex situation for this project
- Students are encouraged to use any technologies (example: Linux Command, SHELL, R, Python, PIG and HIVE) to solve the given problem.
- Project is open book , open discussion, so you are allowed to use all the resources
- This document contains some of references URLs and command statements that would provide a quick help while solving the problem

Customer Data File

CID	Name	Email
1	Tom	Tom@abc.com
2	tim	Tim@xyz.com
3	Marry	Marry@cnn.com
3	Marry	Marry@cnn.com
4	Mike	Mike@abc.com
5	Mike	M@xya.com
6	Kevin	Kevin@abc.com
7	Ron	ron@abc.com
8	Rina	
9	Drew	drew@abc.com
9	Drew	drew@abc.com
10	Maya	maya@cnn.com
11	marry	mary@abc.com
12	Rina	Rinaabc.com
13	Rina	Rina@abc

Hint:

1) Create File : Customer, with Pipe (|) delimited

2) Using PIG eliminate duplicate data
3) Using PIG eliminate the empty rows
4) Store the file to HDFS

Transaction data file

CID	OID	ODate	OPrice
1	1	02/10/1999	200
1	2	02/10/1999	555
5	3	03/10/1999	34
6	4	04/10/1999	100
7	5	04/10/1999	300.50
8	6	05/10/1999	5000.55
2	7	06/10/2000	434
2	8	06/10/2000	445
2	9	07/10/2000	44
3	10	07/11/2000	555
4	11	08/11/2001	600
5	12	09/11/2001	500

Problem Statement

Answer the following Questions:

Be Creative **these are not one LAB**, so don't try to solve using one program, it may not work well. You can use any tools/techniques that you prefer to answer these questions [example: Linux commands, PIG, Python, SHELL, HIVE, Excel and R]. You may need to create multiple programs to answer the following questions.

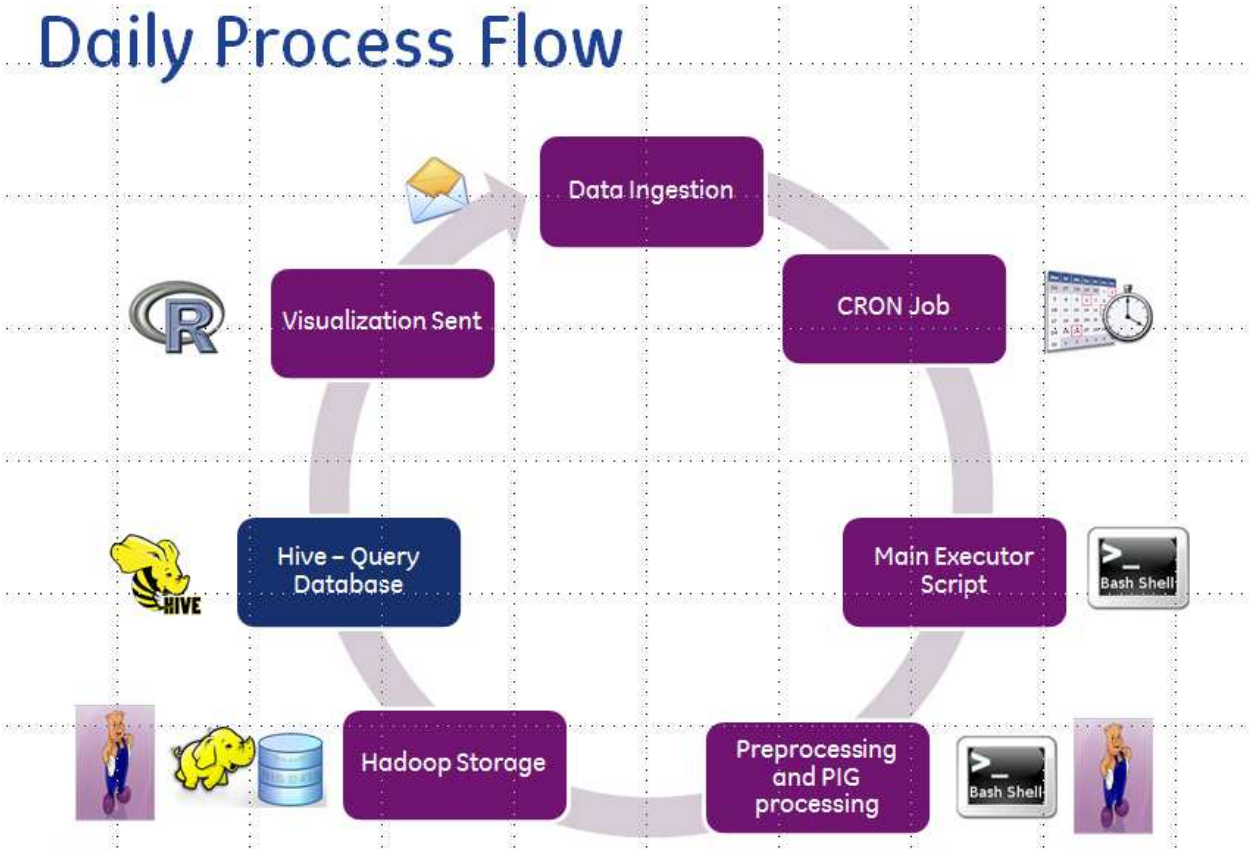
**** Be Calm ….and don't Give up….you will learn a lot by keep trying….**

1) Find How many total Customer records we have in file
2) Find How many **Unique** Customer records we have in file
3) Find How many Customers have made purchase
4) Find How many Unique Customers have provided correct email address
5) Using PIG program generate a files (joined output in a Pipe delimited file) with the following columns

CID	CNAME	CEMAIL	OID	ODDate	OPrice

6) Use dynamic variable (custom variable) and pass any customer ID and dump its associated records on the screen
7) Create a Hive table: **CustTransaction** points to the above file (you are allowed to adjust the file and directory path)
    a. List the total purchase amount based on the Customer ID
    b. Create three HIVE tables: **1999, 2000, and 2001** and store year specific transactions in it.
8) Copy the final output for year:**1999** in memory stick and bring it to Windows machine
9) Using R/Python – Read the **year:1999** file plot a graph for **CID** and **purchase** transaction amount
10) Using R/Python – Prepare Histograph for the Transaction amount
11) Using R /Python – Plot Box diagram for the Transaction amount and find Mean, Max, Min
12) Automate the project -> from data -> to process -> to visualize -> to email.

Try something like this:



** You are allowed to use different dataset, and but at the project end you should be able to show

Data -> Information ->Insight

And E2E – working Process

Hints:

To Run PIG File

pig –x local mypig.pig

To Run HIVE File

hive -f myive.sql

LINUX COMMANDS

Reference:

http://www.pixelbeat.org/cmdline.html

HDFS Commands:

Reference:

http://hadoop.apache.org/docs/r0.18.3/hdfs_shell.html

PIG Statement

- To Load File
  ```
  A = load 'NYSE_dividends' (exchange, symbol, date, dividends);
  ```
- To Filter the record
  ```
  A = filter A by dividends > 0;
  ```
- To convert particular record in upper case
  A = foreach A generate UPPER(symbol);
- Store final output
  store processed into **'processed'** using PigStorage**(',');**

```
-- count.pig
daily = load 'NYSE_daily' as (exchange, stock);
grpd  = group daily by stock;
cnt   = foreach grpd generate group, COUNT(daily);
```

```
A = load 'input1' as (x, y);
B = load 'input2' as (u, v);
C = load 'input3' as (e, f);
alpha = join A by x, B by u, C by e;
```

Self joins are supported, though the data must be loaded twice:

```
--selfjoin.pig
-- For each stock, find all dividends that increased between two dates
divs1     = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                date:chararray, dividends);
divs2     = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                date:chararray, dividends);
jnd       = join divs1 by symbol, divs2 by symbol;
increased = filter jnd by divs1::date < divs2::date and
                divs1::dividends < divs2::dividends;
```

PIG Commands Help:
http://chimera.labs.oreilly.com/books/1234000001811/ch05.html#join_basic

Hive Statement:

Hive Commands Help:

Create Table:

Create the table `retail_demo.products_dim_hive`:

```
CREATE TABLE retail_demo.products_dim_hive
(
  Product_ID      int,
  Category_ID     smallint,
  Price           double,
  Product_Name    string
)
  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
  STORED AS TEXTFILE
  LOCATION '/retail_demo/products_dim/';
```

Load Data into Table

```
LOAD DATA LOCAL INPATH '/retail_demo/order_lineitems/order_lineitems.tsv.gz' OVERWRITE
INTO TABLE retail_demo.order_lineitems_hive;
```

Join Two Tables

```
SELECT a.* FROM a JOIN b ON (a.id = b.id)
```

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Joins

Grouping

```
SELECT owner, COUNT(*)  FROM table GROUP  BY  owner;
```

Create multiple tables from one Table

```
FROM A
INSERT INTO TABLE B1 SELECT F1, F2
INSERT INTO TABLE B2 SELECT F1, F3
INSERT INTO TABLE B3 SELECT F2, F4;
```

Hive Commands:

https://cwiki.apache.org/confluence/display/Hive/Tutorial

http://hortonworks.com/wp-content/uploads/downloads/2013/08/Hortonworks.CheatSheet.SQLtoHive.pdf

R Hint

Read File

data = read.table("clipboard", sep = "|")

To read particular column from the data frame: data

Data$columnmae

Scatterplots

Simple Scatterplot

There are **many** ways to create a scatterplot in R. The basic function is **plot(*x*, *y*)**, where *x* and *y* are numeric vectors denoting the (x,y) points to plot.

```
# Simple Scatterplot
attach(mtcars)
plot(wt, mpg, main="Scatterplot Example",
   xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
```

Graph:

http://www.statmethods.net/graphs/scatterplot.html

https://www.harding.edu/fmccown/r/

> There are no secrets to success. It is the result of preparation, hard work, and learning from failure.
>
> **Colin Powell**