|  |
| --- |
| **Big Data Processing Training, R&D** |
| Power by Data Cloud Lab |
|  |
| [Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Big data was originally associated with three key concepts: volume, variety, and velocity.] |
|  |
|  |
|  |
|  |

Data Set – 1M Data:

1. Healthcare\_ [Record – 46935]
2. Weather-history - [Record – 4573]
3. World Demography - [Record – 5000]
4. Census Tracts 2010 - [Record -21
5. Animal\_Services\_Intake\_Data - [Record -187594]
6. Average\_Daily\_Traffic\_Counts - [Record -1280]
7. Acciental\_Durg\_Related\_Death - [Record -5106]
8. Retails Store - [Record – 182728]

customer12435,category\_59,Departments\_7,orders\_68883,products\_1345,order\_items\_99999

1. Popular\_Baby\_Names - [Record – 46935]
2. SAT\_\_College\_Board\_\_2010\_School\_Level\_Results - Total Data [Record -461]
3. Sales\_Tax\_Rates - [Record -1911]
4. Restaurants [Record -1328]
5. Transportation : 34\_drivers , 17076\_truck\_event\_text\_partition , 1768\_timesheet - [Record -18878]
6. Acciental\_Durg\_Related\_Death - [Record -5106]
7. Census Tracts 2010 - [Record -216]
8. Employees\_Salary - [Record – 824]
9. Customer\_transactional\_spending - [Record – 60000]
10. Customer\_Order - [Record – 1000]
11. Employees\_Salary - [Record – 824]

Power by: Software Linux, Hadoop Big Data, Hive & Power BI)

Case Study 01: Healthcare [Record – 46935]

Raw Data (Date, Sex, Diseases, Age) :

12/10/1950,M,Diabetes,78

12/10/1984,F,PCOS,67

712/11/1940,M,Fever,90

12/12/1950,F,Cold,88

12/13/1960,M,Blood Pressure,76

Result :

Blood Pressure,5215

Cold,5215

Diabetes,5215

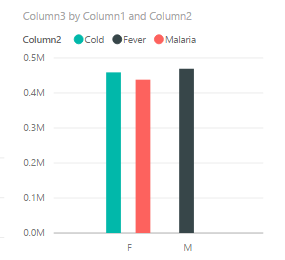
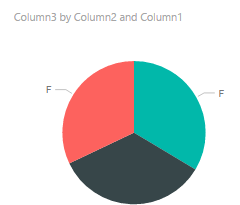
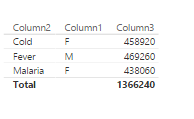
Fever,15645

Malaria,5215

PCOS,5215

Swine Flu,5215

Data Visualizations:



Backend Data Process by HiveQL command:

select diseases, count(\*) from health group by diseases;

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

Query ID = hduser\_20200125220715\_338a065f-f176-4464-b03e-28fb18dc66f5

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes): , set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers: , set hive.exec.reducers.max=<number>

In order to set a constant number of reducers: , set mapreduce.job.reduces=<number>

Job running in-process (local Hadoop) , 2020-01-25 22:07:18,630 Stage-1 map = 100%, reduce = 100%

Ended Job = job\_local171670995\_0001, Moving data to local directory /home/hduser/Dataset

MapReduce Jobs Launched: , Stage-Stage-1: HDFS Read: 2336322 HDFS Write: 0 SUCCESS, Total MapReduce CPU Time Spent: 0 msec, OK

Time taken: 3.617 seconds