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

9. Popular_Baby_Names - [Record – 46935]
10. SAT__College_Board__2010_School_Level_Results - Total Data [Record -461]
11. Sales_Tax_Rates - [Record -1911]
12. Restaurants [Record -1328]
13. Transportation : 34_drivers , 17076_truck_event_text_partition , 1768_timesheet - [Record - 18878]
14. Acciental_Durg_Related_Death - [Record -5106]
15. Census Tracts 2010 - [Record -216]
16. Employees_Salary - [Record – 824]
17. Customer_transactional_spending - [Record – 60000]
18. Customer_Order - [Record – 1000]
19. 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) :

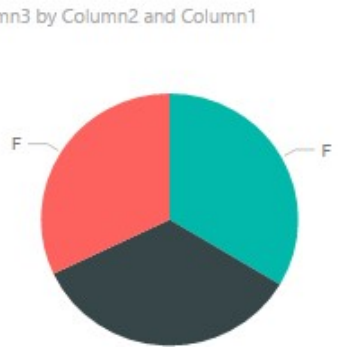
12/10/1950,M,Diabetes,78
12/10/1984,F,PCOS,67
712/11/1940,M,Fever,90
12/12/1950,F,Cold,88
12/13/1960,M,Blood Pressure,76

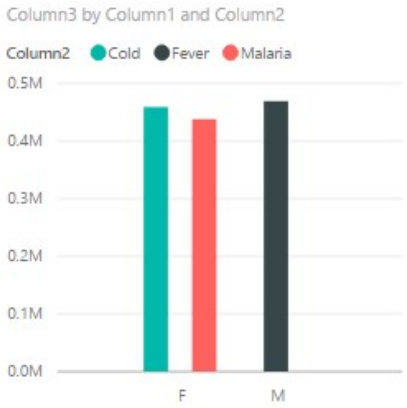
Result :

```
Blood Pressure,5215
Cold,5215
Diabetes,5215
Fever,15645
Malaria,5215
PCOS,5215
Swine Flu,5215
```

Data Visualizations:

Column2	Column1	Column3
Cold	F	458920
Fever	M	469260
Malaria	F	438060
Total		**1366240**





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
```