Hadoop Project – Customer Rating Trends

- Sector = E-Retail Sector
- Customer = Amazon,Flipcart like retailer
- Product Base= 50 Million products
- Average active customer= 2 million daily

Description

- A well know US E-Retailer is business of selling different products across multiple segments. The business model is similar to that of Amazon, FlipCart and SnapDeal.

- On a Monthly basis for the sales that has happened , data is collected for product and no of review feedback comments.

- Like above we have around 200 million products with an average of 2 million review comments across all the products.

Input Data

- Input data format received on weekly basis:

File name:products_reveiw_08082016_240000.txt

- <BOOKS,998989>
- <HP_SERVER,2001>
- <APPLE_MOBILES,245678>
- <HP_LAPTOPS,12345>
- <MAX_CLOTHS,2345>
- <ARROW_CLOTHS,3456>
- <USPOLO_CLOTHS,53646>
- <DELL_LAPTOPS,3455>
- <LENOVO_LAPTOPS,3455>

....

Rating Calculation Formula

- In order to consider the product with higher positive reveiws , the overall reveiw no for each product code is calculated based on following criteria

- - Each product is given a numeric rating (5= Excellent, 4= Very Good, 3= Good , 2= Average ,1= Poor)
- - For Analytic only rating till level 3 (Excellent,Very Good,Good) is considered).
- - Rating no is an addition of all the above 3 types of rating across all feedback review comments

- Ex:
- User 1 Rating = <HP_LAPTOP,4>
- User 2 Rating = <HP_LAPTOP,3>
- User 3 Rating = <HP_LAPTOP,5>

- Rating Considered for HP_LAPTOP product

- <HP_LAPTOP,12> (12=4+3+5)

Phase 1: Data Load and Cleanup

Activities

- Start Oracle DB, Create a workspace and Create Project data related table and indexes.
- Payment gateway data in provided in zip form. Load data into the table.
- Run SQL query to check if the data is loaded successfully.

Phase 2: Data Migration into Hadoop

- 1) Configure Sqoop to use with Oracle DB by adding specific driver support.
- 2) Using Sqoop Impart feature move data from Oracle DB into Hadoop HDFS directory.
- 3) Configure Sqoop to run optimal Mappers by providing No of mappers in import cmd.

Phase 3: Apache Hive data load

- 1) Create a output table in Apache Hive for 1 dataset
- 2) Create a output table in Apache Hive for 2 dataset
- 3) Load above hive tables with data available in HDFS directories.
- 4) Create a Merge table and populate it by a INSERT-SELECT query.

Phase 4: HiveQL based data analytic

- 1) Create and run hive queries in shell as well as offline manner .

Analytic Reports:

- 1) Top 5 review products for each week across all segments.
- 2) Products with Maximum variation in weekly rating from user groups.