
BiG
DATA

Capstone Project

Case Study - 1

Email Marketing Campaign

Objective:

Analyze Email Marketing Campaigns of a Magazine Publisher .

Data Availability:

- Data available in the form of a csv file (/home/data/CampaignData_full.csv).
- Data for 2010 and 2011.

Data Dictionary:

- Solicitation history and outcome
- Solicitation details
- Demographics information about the individual being solicited
- Household information for the individual solicited

Case Study - 1

Email Marketing Campaign

Reports Requirements

1. **Find the Click to Open Rate (CTOR)**

A. Overall CTOR (use CLICK_FLG and OPEN_FLG column)
B. CTOR by Gender (use I1_GNDR_CODE column)
C. CTOR by Time of the day (use mailed_date column)
D. CTOR by Day of the week (use mailed_date column)
E. CTOR by Month (use mailed_date column)
F. CTOR by Lead's Income Group (use TRW_INCOME_CD_V4 column)
G. CTOR by Lead's Ethnicity (use ASIAN_CD column)
H. CTOR by Lead's Household Status (use I1_INDIV_HHLD_STATUS_CODE column)

This information should be represented in Tableau/Power BI charts (bar/pie/anything relevant) which should then be shown on a dashboard.

Case Study - 1

Email Marketing Campaign

Reports Requirements

2. Household Members Information

A. Find count of leads with information about members of their household. If a lead has information about 3 members, and another has information about 2 members and another has none, then the answer to this question is 2. (use statcd_hh_mem1 - statcd_hh_mem8 column)

B. Find count of total number of household members information is available. For example, if a lead has 3 household members, and another has about 2 members, and the other has none, then the total count of household members is 3+2+0 = 5.

C. Find count of household members by type (Head of Household, Spouse etc.).

D. %age of household members type. For example, if there are 5 Head of Household, 10 Spouse and 85 in the other categories, then the %age of Spouses is 10.

Case Study - 1

Email Marketing Campaign

Reports Requirements

2. Household Members Information

E. How many known households have children?
(use PRESENCE_OF_CHLDRN column)

F. Overall, how many children are there?
(use NUMBER_OF_CHLDRN_18_OR_LESS column and PRESENCE_OF_CHLDRN)

G. How many of the children are male and how many are female?
(use GNDR_OF_CHLDRN_0_3 - GNDR_OF_CHLDRN_13_18 column)

Case Study - 1

Email Marketing Campaign

Workflow Requirements

- Data flow from source to be copied to HDFS
- Data from HDFS to be loaded to Pig for filtering and transformations
- Final output from Pig to be stored in HDFS
- Data from HDFS to be loaded to Hive for finding solutions for above mentioned problem in earlier slide
- Connect to tableau
  <IP: 52.4.16.124 need to connect to 54.174.252.76 for hiveserver2>
  <IP: 52.3.237.208 need to connect to 52.3.237.208 for hiveserver2>
- Create Dashboard

Case Study - 1

Email Marketing Campaign

High Level Design

Information about Source:

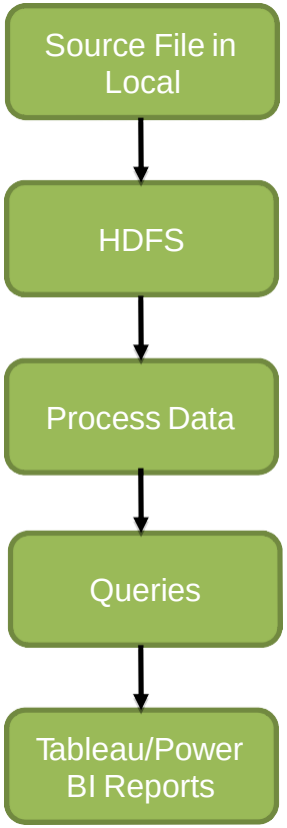
- Data in the form a csv file is stored on the local file system which needs to be moved to HDFS.
- Certain data transformations need to be implemented.
- Certain pre-calculations need to be implemented.
- Reports should be displayed in a dashboard.

Case Study - 1

Email Marketing Campaign

Data Flow:

1. **Copy data into HDFS.**
2. **Apply transformations and perform calculations.**
3. **Queries to load data for reports.**
4. **Build Reports/Dashboards on Tableau/Power BI.**

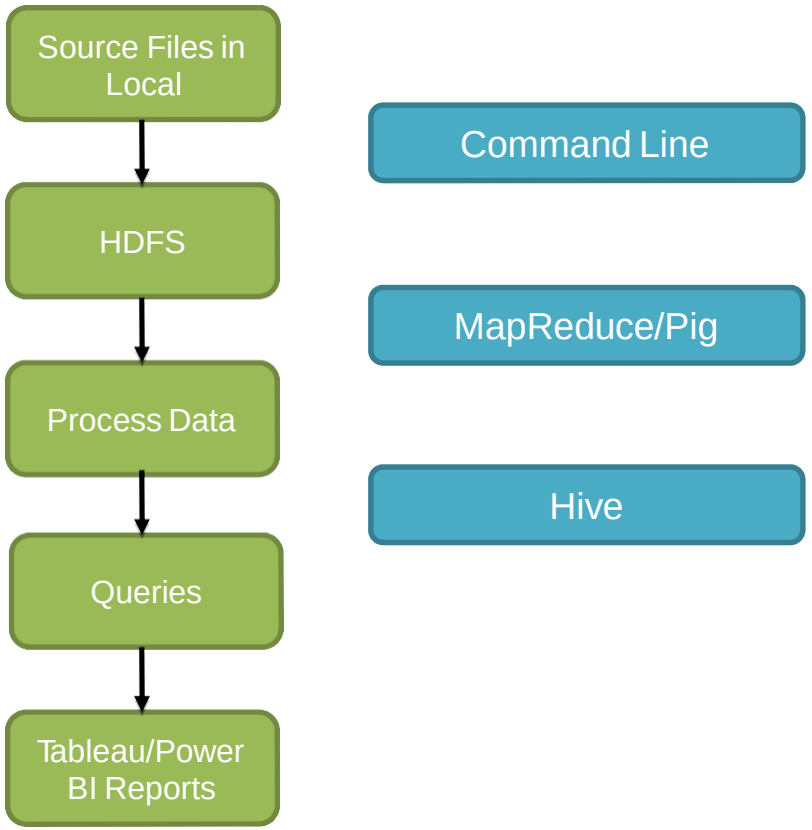


Case Study - 1

Email Marketing Campaign

Data Flow:

1. **Copy data into HDFS.**
2. **Apply transformations and perform calculations.**
3. **Queries to load data for reports.**
4. **Build Reports/Dashboards on Tableau/Power BI.**



Case Study - 1

Email Marketing Campaign

Low Level Design

Preparing Hive tables for Report queries

- Since performance of report queries is important, we can have the final table in denormalized form, so as to not involve any JOINS during querying.
- Although the reports do not have any time-series data as such, there is no need to partition the final table. But, it is a good practice to create partitioned tables.

Jigsaw Certification

Sample Questions

Which database does HIVE use, by default, for storing metadata about the Hive tables?

A. mysql
B. derby
C. PostGres
D. MongoDB

Jigsaw Certification

Sample Questions

Which database does HIVE use, by default, for storing metadata about the Hive tables?

A. mysql
B. derby
C. PostGres
D. MongoDB

Jigsaw Certification

Sample Questions

You need to transfer a subset of rows from a database table into HDFS. Consider the following scenario:
There is a database named COUNTRIES, which has a table named CITIES, which contains the details of all the cities/country around the world.
Now we need the details of all the cities which are in the country India.
Identify the correct syntax of Sqoop for the above scenario?

A. sqoop import --connect jdbc:mysql://localhost/countries --username username --password password --table cities --where "select * from cities where country ='india‘”

B. sqoop import --connect jdbc:mysql://localhost/countries --username username --password password --table cities --where "country = 'india'“

C. sqoop import --connect jdbc:mysql://localhost/ --username username --password password --table where countries.cities = 'india‘

D. sqoop import --connect jdbc:mysql://localhost/ --username usename --password password --table select * from countries where cities = 'india'

Jigsaw Certification

Sample Questions

You need to transfer a subset of rows from a database table into HDFS. Consider the following scenario:
There is a database named COUNTRIES, which has a table named CITIES, which contains the details of all the cities around the world.
Now we need the details of all the cities which are in the country India.
Identify the correct syntax of Sqoop for the above scenario?

A. sqoop import --connect jdbc:mysql://localhost/countries --username username --password password --table cities --where "select * from cities where country ='india'"

B. sqoop import --connect jdbc:mysql://localhost/countries --username username --password password --table cities --where "country = 'india'"

C. sqoop import --connect jdbc:mysql://localhost/ --username username --password password --table where countries.cities = 'india'

D. sqoop import --connect jdbc:mysql://localhost/ --username usename --password password --table select * from countries where cities = 'india'

RECAP

Case Study
Jigsaw Certification