
BiG
DATA

Case Study

Email Marketing Campaign

Objective:

Analyze Email Marketing Campaigns of a Magazine Publisher .

Data Availability:

- Data available in the form of a csv file (CampaignData.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

Note: Complete Data Dictionary will be uploaded in LMS.

Case Study

Email Marketing Campaign

Reports Requirements

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

A. Overall CTOR
B. CTOR by Gender
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

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 1.
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

Email Marketing Campaign

Reports Requirements

2. Household Members Information

E. How many known households have children?
F. Overall, how many children are there?
G. How many of the children are male and how many are female?

Case Study

Email Marketing Campaign

Workflow Requirements

- Data flow from source to final Reports table should be automated.

Case Study

Email Marketing Campaign

Things to keep in mind

- The Reports should not have any codes displayed. For example, it should display income group as J, K, L etc., but it should display the actual figures < $15,000 etc.
- Query performance has to be good.

Case Study

Email Marketing Campaign

High Level Design

Information about Source:

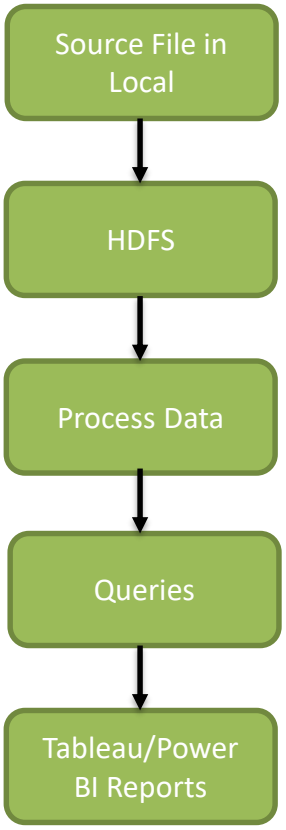
- Data in the form a csv file is stored on the local file system which needs to be moved to HDFS.
- Certain data transformations need to be implemented.
- Certain pre-calculations need to be implemented.
- Reports should be displayed in a dashboard.
- Data flow needs to be automated as much as possible.

Case Study

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.
5. Automate the entire flow.

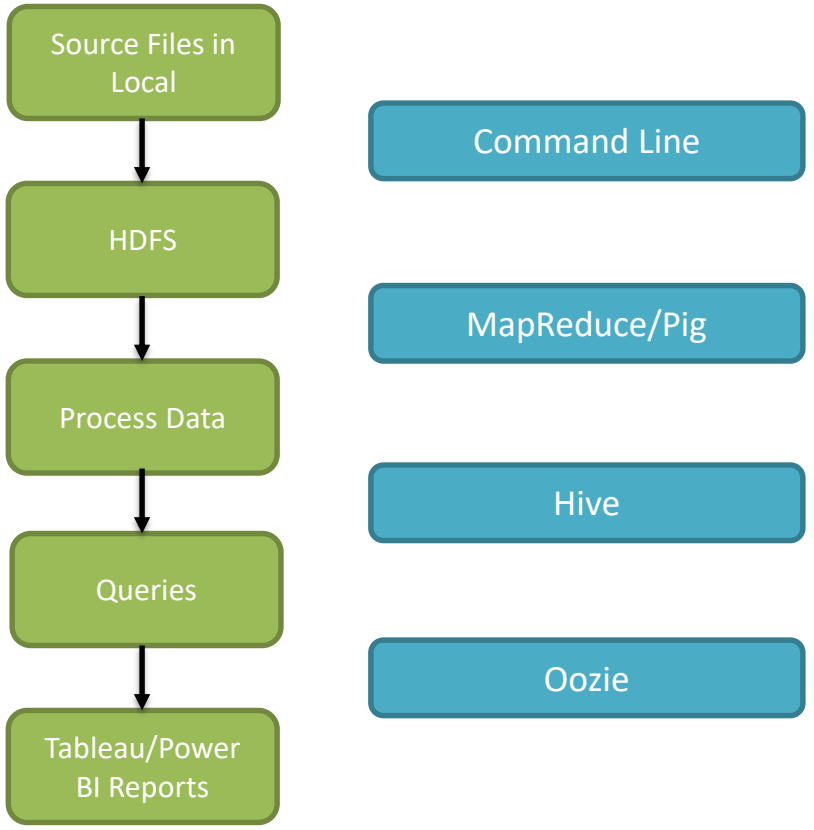


Case Study

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.**
5. **Automate the entire flow.**



Case Study

Email Marketing Campaign

Low Level Design

Loading source data into HDFS

- Write a shell script to move the source csv file exists to HDFS.
- This script should then be invoked from Oozie, only once.

Processing of source files

- Extraction of date and time from the mailed_date field.
  Example: Mon, 09/20/10 01:04 PM
- Certain reports involve finding sum of values of multiple fields, if they have values. These things might be easier to do in Java MapReduce (can be done in Pig as well).
- Since report should display the field description and not code values, it is better to add description along with the codes in the final data. So, we would need to decide how to implement this – hardcode or create configuration tables.

Case Study

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.

Automating the Workflow

- Use Oozie Workflows to run the Shell script to load files into HDFS, trigger the Java MapReduce jobs, run Pig Scripts and run Hive queries.

THANK YOU