Customer Retention Strategy

Capstone Project - III

Table of Content

1. Business Challenge/Requirement

Customer retention and acquisition strategies are on top of every organization's agenda. To offer better customer service and boost loyalty, a company has to invest in a state-of-the-art CRM tool. In pursuit of these goals, every organization implements CRM as a strategy that integrates the concepts of data mining and data warehousing. The data collected through the CRM helps the leadership team make actionable decisions in real time. It helps them build and retain long-term and profitable relationships with customers.

FutureCart Inc. is a hypothetical leading retail company with an omnipresence in India with more than 5000 retail stores and hypermarkets across and e-commerce in the country.

The company has formed a dedicated team to handle after-sales services. The team is entrusted with the responsibility to address customer complaints and delight them - and eventually increase brand loyalty

Below is an abstract of end to end process:

- The company has multiple contact centers across India to provide support service to their customers
- Customers can reach out to the care team over different communication channels depending on their preference and convenience: Calls, Chat, or Email.
- CCR (Customer Care Representative) registers the complaint by collecting all the necessary details -- which is called a **case**
- A case can have a status -- open or closed
- Each case can belong to a category and sub-category. This category and sub-category will determine case priority. Depending on the priority key, CCR has an SLA (in hours) to close the case within the SLA hours
- Once a case is closed, the customer is sent a survey link to rate the overall experience of interacting with the contact center representative
- The customer can take a survey or leave it unattended. The customer can rate the experience on a scale of 1-10 on various questions
- Survey response is captured for that particular case

The data collected through complete CRM process is used by the company for analysis. The analytics team working on this data captures the below KPIs to further enhance and optimize the CRM process.

KPIs (Both on real-time data and batch-processed data)

- Total numbers of cases
- Total open cases in the last 1 hour
- Total closed cases in the last 1 hour
- Total priority cases

- Total positive/negative responses in the last 1 hour
- Total number of surveys in the last 1 hour
- Total open cases in a day/week/month
- Total closed cases in a day/week/month
- Total positive/negative responses in a day/week/month
- Total number of surveys in a day/week/month

Real-time KPIs

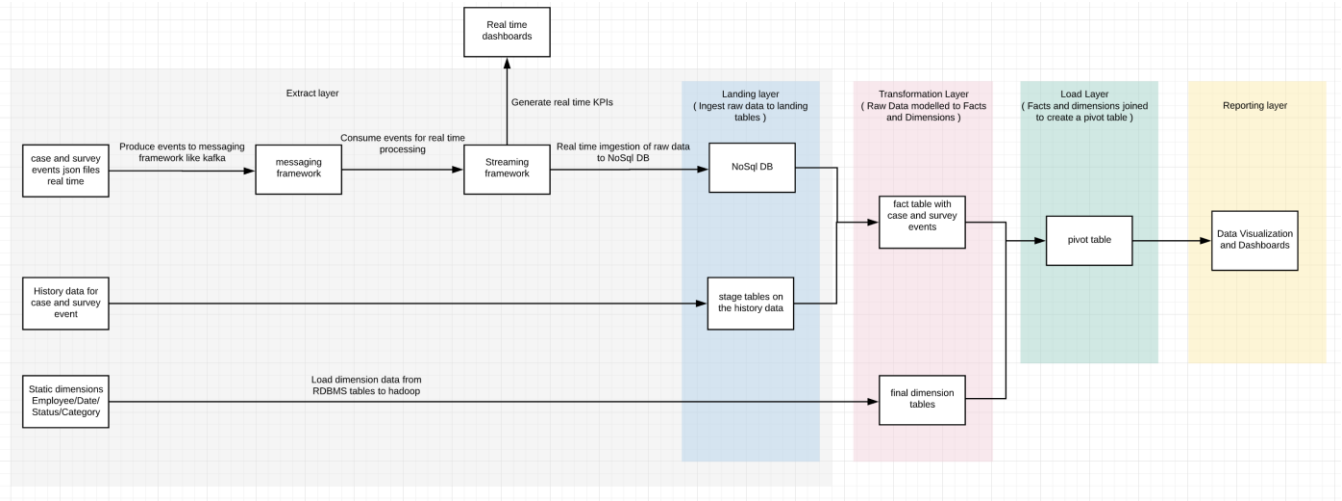
- Total numbers of cases that are open and closed out of the number of cases received
- Total number of cases received based on priority and severity

2. The Goal of the Project

Below are some of the high-level technical and non-technical goals for this project:

- Get an overall understanding of the CRM domain
- Learn the **fundamentals** & standards of ETL and data warehousing
- Real-time and batch ingestion of data from multiple sources to Big Data storage like Hive/ Cassandra /HDFS using Kafka and Spark
- Data cleansing/wrangling/transformation using Hive and Spark
- Lambda architecture where data can be processed in both batch and real-time
- Reporting KPIs

3. Data Flow Architecture/Process Flow



4. Dataset Explanation and Schema

We have three types of data sources:

1) Data for which static/dimension tables to be created in MySQL
2) Historical data of 6 days for cases and survey events to be created in MySQL
3) Real-time data for the current date for cases and survey events in JSON files

4.1 Data for which static/dimension tables to be created in MySQL

We have the below datasets present in HDFS at location - /bigdatapgp/common_folder/projects/futurecart/batchdata/ which can act as dimensions:

futurecart_calendar_details.txt - Calendar details for the company

column Name	Data type	Column description	sample value
calendar_date	date,	Calendar date in yyyy-mm-dd format	2011-02-20
date_desc	varchar(50)	Calendar date in words	Sunday, February 20, 2011
week_day_nbr	smallint	Number of days in a week	2
week_number	smallint	Week number of the year	4
week_name	varchar(50)	Week name	Week 04
year_week_number	int	Week number with year	201104
month_number	smallint	Month number in the year	1
month_name	varchar(50)	Month name	february
quarter_number	smallint	Quarter number in the year	1
quarter_name	varchar(50)	Quarter name	Q1
half_year_number	smallint	Half-year number in the year	1
half_year_name	varchar(50)	Half-year name	1st Half
geo_region_cd	char(2)	Geographic region code	US

futurecart_call_center_details.txt – Contact/Call center details for the company

column Name	Data type	Column description	sample value
call_center_id	varchar(10)	Unique identifier for a call center	C-101
call_center_vendor	varchar(50)	Vendor company name which is handling the call center	Concentrix
location	varchar(50)	Call center location	New york
country	varchar(50)	Call center country	US

futurecart_case_category_details.txt - Category details of a case event

column Name	Data type	Column description	sample value
category_key	varchar(10)	Unique identifier for a case category	CAT1
sub_category_key	varchar(10)	Unique identifier for a case sub category	SCAT1
category_description	varchar(50)	Category description	Subscription
sub_category_description	varchar(50)	Subcategory description	Renewal
priority	varchar(10)	Priority key	P1

futurecart_case_country_details.txt - Country details

column Name	Data type	Column description	sample value
id	int	Unique identifier for a country	4
Name	varchar(75)	Country name	India
Alpha_2	varchar(2)	Country short name 2 chars	IN
Alpha_3	varchar(2)	Country short name 3 chars	IND

futurecart_case_priority_details.txt - Priority details of a case

column Name	Data type	Column description	sample value
Priority_key	varchar(5)	Unique identifier for a case priority	P1
priority	varchar (20)	Priority level	Highest
severity	varchar (100)	Severity level	critical
SLA	varchar (100)	SLA in HOURS for the priority and severity combination	1

futurecart_employee_details.txt - Employee details of the company

column Name	Data type	Column description	sample value
emp_key	Int	Unique ID of an employee	10001
first_name	varchar	First name	Georgi
last_name	varchar	Last name	Facello
email	varchar	email	Georgi.Facello01@testmail.com
gender	varchar	gender	M
ldap	varchar	User id	5941CF7D
hire_date	Date	Hire date	2014-04-06
manager	varchar	Manager key	455246

futurecart_product_details.txt - Product details of the company

column Name	Data type	Column description	sample value
product_id	varchar	Unique id for a product	26355
department	varchar	Department description	GROCERY
brand	varchar	Brand description	Private
commodity_desc	varchar	Commodity description	COOKIES/CONES
sub_commodity_desc	varchar	Subcommodity description	SPECImALTY COOKIES

futurecart_survey_question_details.txt - Question details for the survey

column Name	Data type	Column description	sample value
question_id	varchar	Unique id for a survey question	Q1
question_desc	varchar	Question text	How would you rate your overall experience with the customer support process?
response_type	varchar	Response type (scale or options)	Scale
range	varchar	Scale range if the response type is scale else NA	1-10
negative_response_range	varchar	Scale range to qualify a survey response as negative	1-4
neutral_response_range	varchar	Scale range to qualify a survey response as neutral	5-7
positive_response_range	varchar	Scale range to qualify a survey response as positive	8-10

4.2 Historical data of 6 days for cases and survey events to be created in MySQL

We have the below datasets present in HDFS at location - /bigdatapgp/common_folder/projects/futurecart/batchdata/ that has the historical data

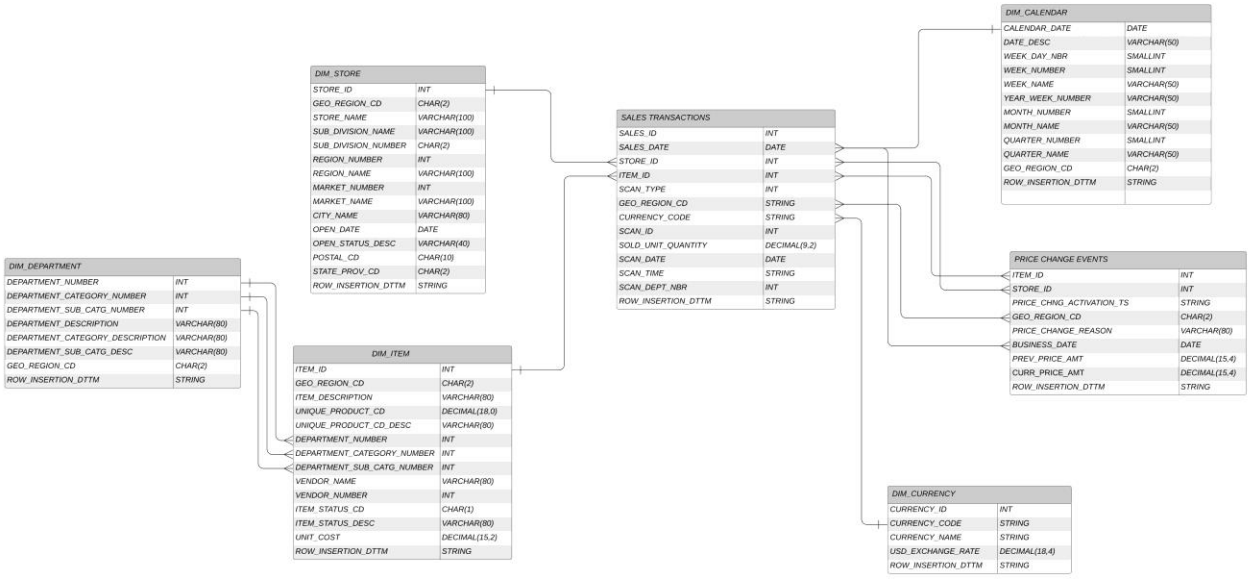
futurecart_case_details.txt – Case details of the company

column Name	Data type	Column description	sample value
case_no	varchar	Unique ID of a case	2024
create_timestamp	varchar	Case create timestamp	2020-04-20 01:01:29
last_modified_timestamp	varchar	Case last modified timestamp	2020-04-20 01:01:29
created_employee_key	varchar	Employee key who created the case	274649
call_center_id	varchar	Call center id where case is logged and handled	C-104
status	varchar	Current status of the case	Open
category	varchar	Category key of the case	CAT1
sub_category	varchar	Subcategory key of the case	S CAT1
communication_mode	varchar	Mode of communication	Email
country_cd	varchar	Country code	PY
product_code	varchar	Product code	997719

futurecart_case_survey_details.txt – survey details of the cases closed

column Name	Data type	Column description	sample value
survey_id	varchar	Unique ID of a survey	S-1000
Case_no	varchar	Case number for which survey has been filled	130114
survey_timestamp	varchar	Survey taken timestamp	2020-04-20 01:01:29
Q1	varchar	Q1 response	2
Q2	varchar	Q2 response	7
Q3	varchar	Q3 response	3
Q4	varchar	Q4 response	N
Q5	varchar	Q5 response	7

4.3 Relation between different datasets:



4.4 Real-time data for the current date for cases and survey events in JSON files

Data sources for real-time processing is present at the HDFS location – /bigdatapgp/common_folder/project_futurecart/realtimedata/realtime_simulator.py

Copy the file - realtime_simulator.py to your webconsole and generate real-time data.
This real-time simulator script which will generate JSON files in the corresponding directory of below two events every 15 seconds:

- futurecart_case_event
- futurecart_survey_event

Commands to execute:

hdfs dfs -get /bigdatapgp/common_folder/project_futurecart/realtimedata
python2 realtimedata/realtime_simulator.py --outputLocation path_of_directory

For example, `python2 realtimedata/realtime_simulator.py --outputLocation /mnt/bigdatapgp/edureka_921625/futurecart/realtime/`

This script will create two directories, i.e., case & survey under which the corresponding JSON files are generated

JSON formats:

Generated JSON files with the naming convention as mentioned below:

<data type>_data_<epochtimestamp>.json
Data type can be 'case' or 'survey'

Sample file names :
survey_data_1592422939.json
case_data_1592422939.json

Case JSON format :


```
[
 {
  "status": "Open",
  "category": "CAT3",
  "sub_category": "SCAT14",
  "last_modified_timestamp": "2020-06-17 18:42:19",
  "case_no": "600999",
  "create_timestamp": "2020-06-17 18:42:19",
  "created_employee_key": "240604",
  "call_center_id": "C-116",
  "product_code": "9829787",
  "country_cd": "PR",
  "communication_mode": "Chat"
 },
 {
  "status": "Open",
  "category": "CAT3",
  "sub_category": "SCAT14",
  "last_modified_timestamp": "2020-06-17 18:42:19",
  "case_no": "601000",
  "create_timestamp": "2020-06-17 18:42:19",
  "created_employee_key": "215285",
  "call_center_id": "C-114",
  "product_code": "12457101",
  "country_cd": "EE",
  "communication_mode": "Call"
 }
]
```

Survey JSON format:


```
[
  {
    "Q1": 9,
    "Q3": 1,
    "Q2": 8,
    "Q5": 3,
    "Q4": "N",
    "case_no": "600991",
    "survey_timestamp": "2020-06-17 19:42:04",
    "survey_id": "S-500014"
  },
  {
    "Q1": 8,
    "Q3": 9,
    "Q2": 1,
    "Q5": 1,
    "Q4": "N",
    "case_no": "600992",
    "survey_timestamp": "2020-06-17 19:42:04",
    "survey_id": "S-500015"
  }
]
```

5. Problem Statements/Tasks

The high-level task is to create a Data Mart on CRM data with a lambda architecture where we will ingest and process data in both batch and real time. We also want to enable reporting of KPIs in both batch and real time.

Technical tasks in details :

Refer data flow and architecture for additional reference:

1. Companies generally store transactional data in RDBMS because they provide faster read and write operations and support ACID properties. Hence, create MySQL tables for both the dimension and historical datasets shared.

   Note: Naming convention of the MySQL tables should be: your_cloudlab_username_dataset_name
   For example, if the Cloudlab username is edureka_396101 and the dataset is futurecart_employee_details.txt then the MySQL table name should be ***edureka_396101_futurecart_employee_details***
2. Perform batch ingestion from MySQL to Hive tables for static dimensions and historical data for the case and survey events.
3. Capture new cases and survey events from JSON files being written to a directory and produce them to a Kafka topic.
4. Create an application that will consume and process real-time data.
5. Once we have captured both batch and real-time data in stage tables, perform data modeling around business KPIs to create facts and dimension.
6. Join facts and dimensions and load pivot table.
7. Create tableau reports on the pivot table (Optional)
8. On the real-time feeds, develop a real-time analysis framework to create real-time KPIs and publish them to a dashboard. For this, you might have to join both real-time feeds and static MySQL tables.

6. Approach to Solve

- Identify tools and technologies
- Identify data for real-time and batch ingestions
- Build incremental/history ingestion load
- Model the data to facts and dimensions as per requirement
- Process the data and load facts and dimensions
- Create a pivot table for KPI reporting

7. Considerations/Assumptions

- To work with Hive, create a database named as your Cloudlab username and create tables within that. For example, if the Cloudlab username is edureka_396101, then the Hive database name should be ***edureka_396101,*** and all the hive tables will be created in this database
- For creating tables in NoSQL Databases, the table name should start with your username. For example, ***edureka_396101_futurecart_tablename***
- All the target tables that we develop should have an ORC storage format.
- All the target tables that we develop should have an additional row_insertion_dttm column, which will store the current timestamp.
- All the target tables that we develop should be partitioned on any date (CASE CREATE DATE /SURVEY date) columns if available.
- We will have real-time data for the current day in JSON files for case and survey events.
- There will be a simulator script in Python, which (if we run it) will start creating the JSON file for both case and survey events in a directory.
- If there are multiple surveys for a closed case, then we need to consider the survey with the earliest timestamp.
- Answers to survey questions will be divided into negative/neutral/positive responses in the final fact table depending on the range for every question available in the survey question dimension table.

8. Deliverables

- A fact table joined on the case and survey events
- A pivot table joined among the above fact table and other dimensions
- Real time analysis framework to monitor KPIs in real time

9. Business Benefits

After the solution is developed, a business can enjoy the below operational benefits.

- The company can track vital information/KPIs in real time, which will enable them to make actionable decisions
- With Data Mart created, the company can easily monitor its historical performance and check how it fared so far and what can be improved in the future.
- Data Mart can also be used for advanced data science.
- The solution will create a pivot table that can be integrated into any visualization tools such as Tableau to create reports. Or data analysts/business users can directly query the table

10. How to submit the project

- Create detailed documentation of the steps followed to complete this project.
- Create a directory named as your_cloudlab_username_project_name in your webconsole and place all the SQL scripts, sbt projects, codes, and jars used in this project along with the deliverables mentioned above.
  For example, if the Cloudlab username is edureka_396101 then directory should be named as ***edureka_396101_futurecart***

Or

You can even upload your code into your GitHub repository and share your repository with us.

11. Marks Allocation

- Creation of dimension tables [10 Marks]
- Creation of fact tables [20 Marks]
- Creation of pivot tables [30 Marks]
- Real-time data ingestion [10 Marks]
- Real-time KPIs [20 Marks]
- Detailed documentation [10 Marks]