Customer Retention Strategy

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*

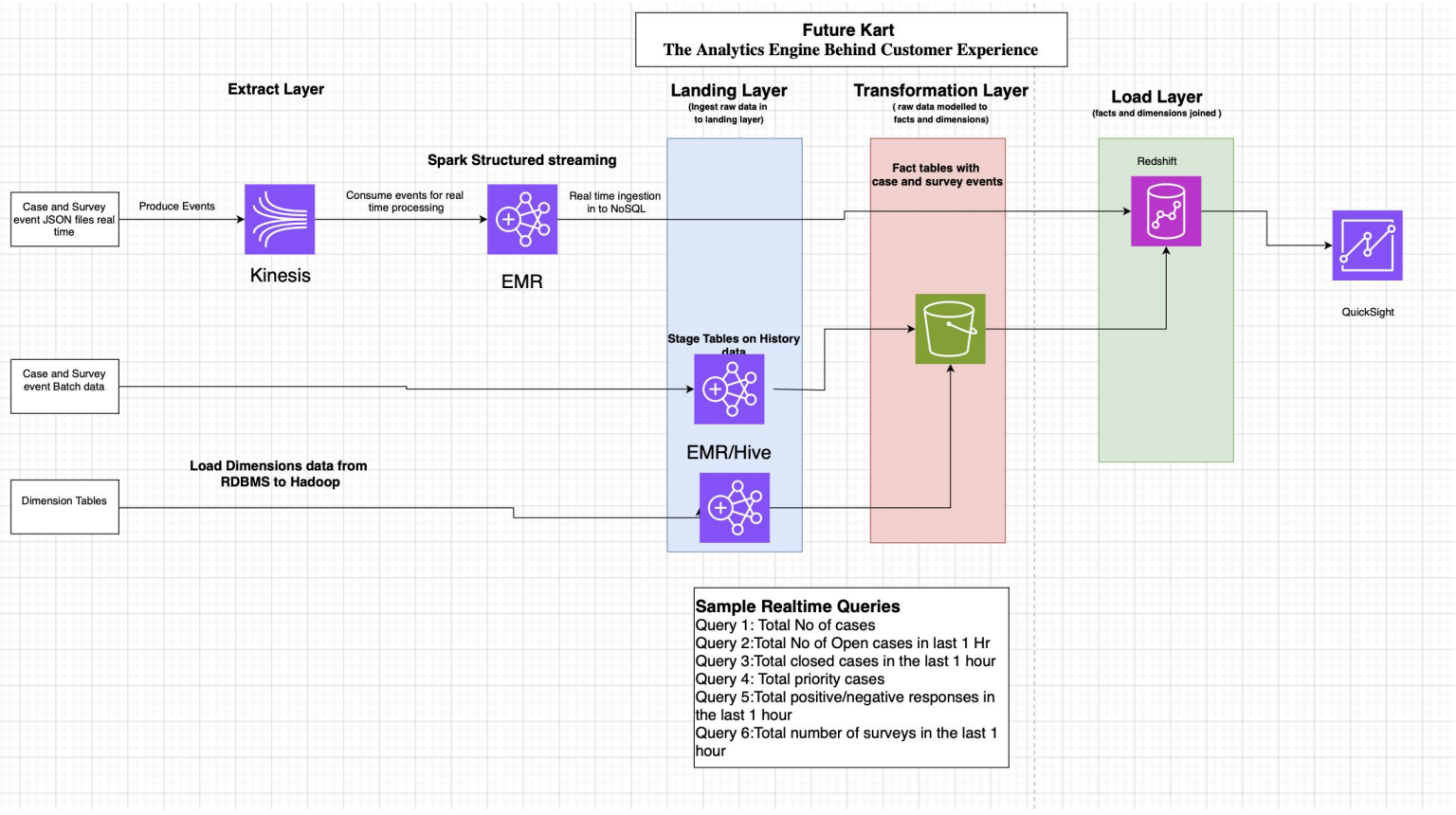
- *Total number of cases received based on priority and severity*

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/ DynamoDB /HDFS using Kinesis 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(Key Performance Indicators)*

3. Data Flow Architecture/Process Flow



4. Dataset Explanation and Schema

We have three types of data sources:

- ***Data for which static/dimension tables to be created in MySQL***
- ***Historical data of 10 days for cases and survey events to created in Hive(JSON)***
- ***Real-time data for the current date for cases and survey events in JSON format***

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

We have the below datasets -

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 e 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*

4.2Historical data of 10 days for cases and survey events to be created in hive

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 01:01:2
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

4.3Real-time data for the current date for cases and survey events in JSON files

Copy the file - stream_to_kinesis.py to your VM and generate real-time data. This real-time simulator script which will generate JSON data and writes it to the kinesis stream of your choice. Change the script to point it to your stream.

- ***futurecart_case_event***
- ***futurecart_survey_event***

JSON formats:

Sample data :

case_data:

[{

"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 Data:

[{

"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 the dimension datasets shared. Create a database in an EC2 instance with MySQL 5.6 server and create all the required dimension tables.***

2. ***Perform batch ingestion from MySQL to Hive tables for static dimensions.***

3. ***Historical data for the last 10 days is generated using the script (generate_historical_data.py) which generates json data for cases and surveys. Load this data in HDFS and create historical tables for cases and surveys.***

4. ***Generate new cases and survey events in json format from the python script and send them to kinesis stream***

5. *Create an application that will consume and process real-time data.*

6. *Load the realtime data coming from kinesis in to redshift as fact tables. If the incoming record from stream is a "Case" data load it into case table else load it in to survey table*

7. *Load the historical and dimension tables from hive in to s3 using spark with dataframes. Load these s3 data in to reshift tables*

8. *Create the above queries on the redshift tables.*

9. *Create dashboard for some of the queries in quicksight with redshift as the source*