Semester internship Report

Thomson Reuters semester internship report

Akhil Gattu

Data Engineering Intern

Document Version 0.1
Date of issue: 22 June 2023

Contents

1. Introduction

1.1 About

Thomson Reuters is a Canadian multinational conglomerate, it has products in 4 main divisions such as: Legal, Reuters News Agency, Tax & Accounting, and Government departments. The company is a globally recognized provider of business information services, data analytics, and solutions for professionals across various industries. The projects listed in the report are done as a part of the Data and Analytics team with a focus on improving customer service.

1.2 Tools Used

Programming languages: Python, SQL

Data ingestion/extraction: ETL tools (Extract, transform and load)

Visualization: Power BI, Tableau

1.3 Technology Used

Machine learning: Natural Language Processing, Large Language Models, Time series forecasting

Data warehouse: Snowflake

Cloud: AWS

1.4 Methodology

The usual methodology to deliver projects is as follows: The stakeholder presents a business problem, then a “Proof of concept” (POC) is made with an initial solution to the problem by the engineering team, then the stakeholder reviews and suggests solutions accordingly. After the suggested solutions are incorporated, the solution is deployed to production and the stakeholder delivers the solution to the business leaders which helps the company’s business.

2. Projects

Following are the projects implemented in the tenure of the internship with the corresponding business objectives attached. In the report below, MDS refers to “My Data Space”, which is a database in TR’s (Thomson Reuters) snowflake warehouse.

2.1 Chat interface analysis

2.1.1 Business objective

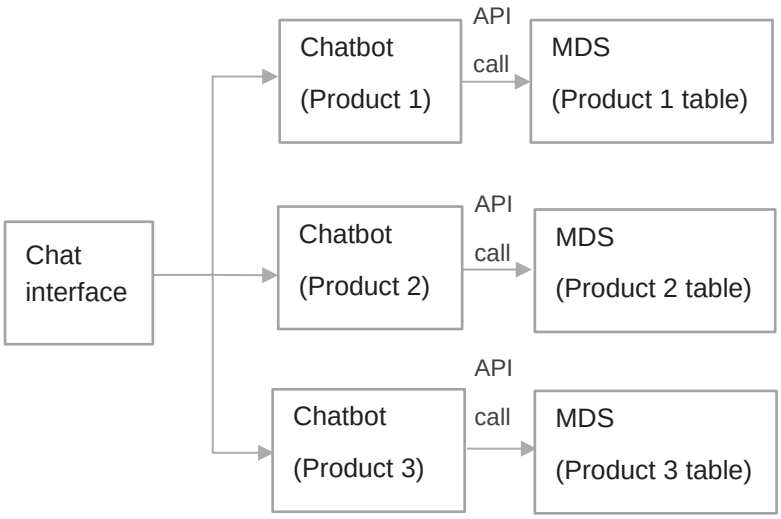
Thomson Reuters intends to improvise their customer experience/support service. Currently, customers have the option of using self-serving capabilities of various product platforms to assist their issues. In case the self-serving capabilities do not provide the required solution, customers can interact with chatbots for further assistance.

Chatbots can resolve a lot of issues, in case they are unable to do so, they direct to a chat engine which becomes an interface for customers to directly interact with live people (agents/employees). Thus, the company would like to analyze the flow of chats from customers to agents and various metrics associated with it.

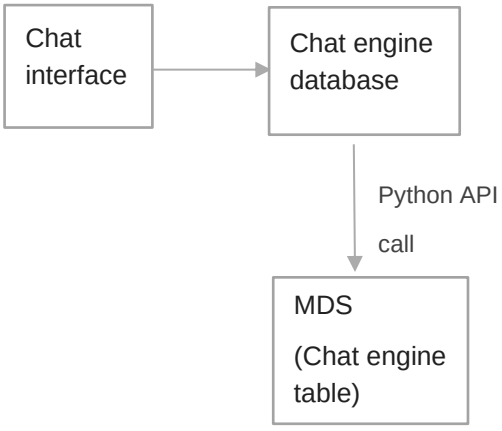
To generate the required metrics, the first step was to establish a data architecture and later process and extract data using the architecture. In this project, the required data has been extracted using batch processing algorithm (since loading a huge amount of data at once isn't feasible, it is loaded batchwise in chunks of 1000/10000).

2.1.2 The data architecture

Chatbots:

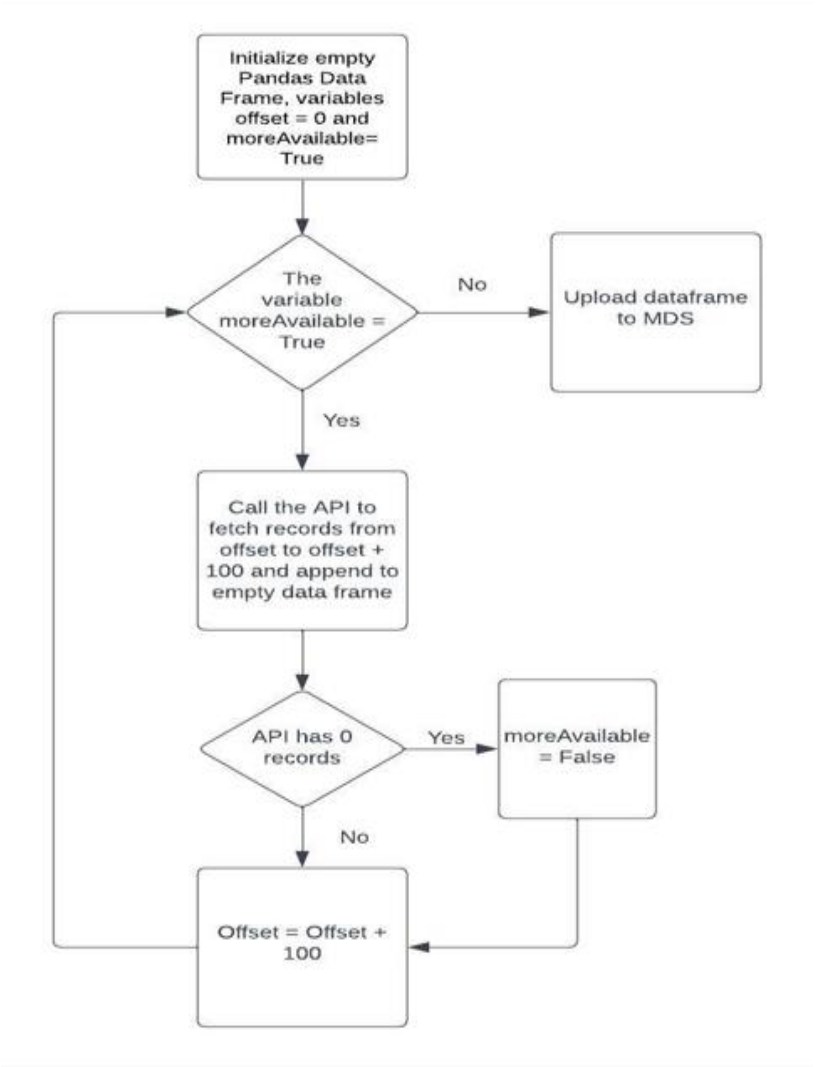


Chat engine:

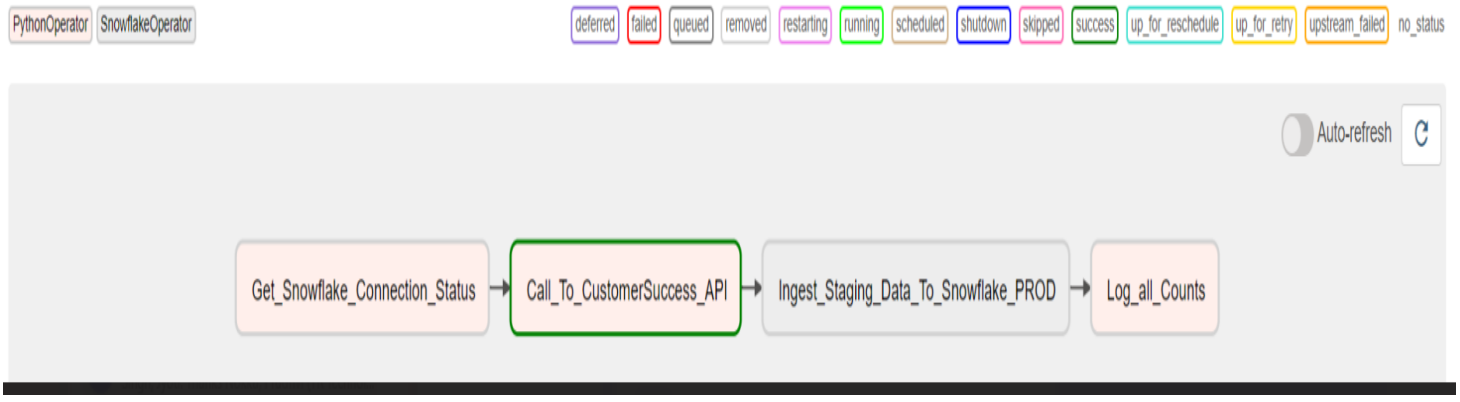


2.1.3 Data processing algorithm

(The flow of control can be automated using directed acyclic graphs (DAG) in Apache Airflow/run manually everyday):

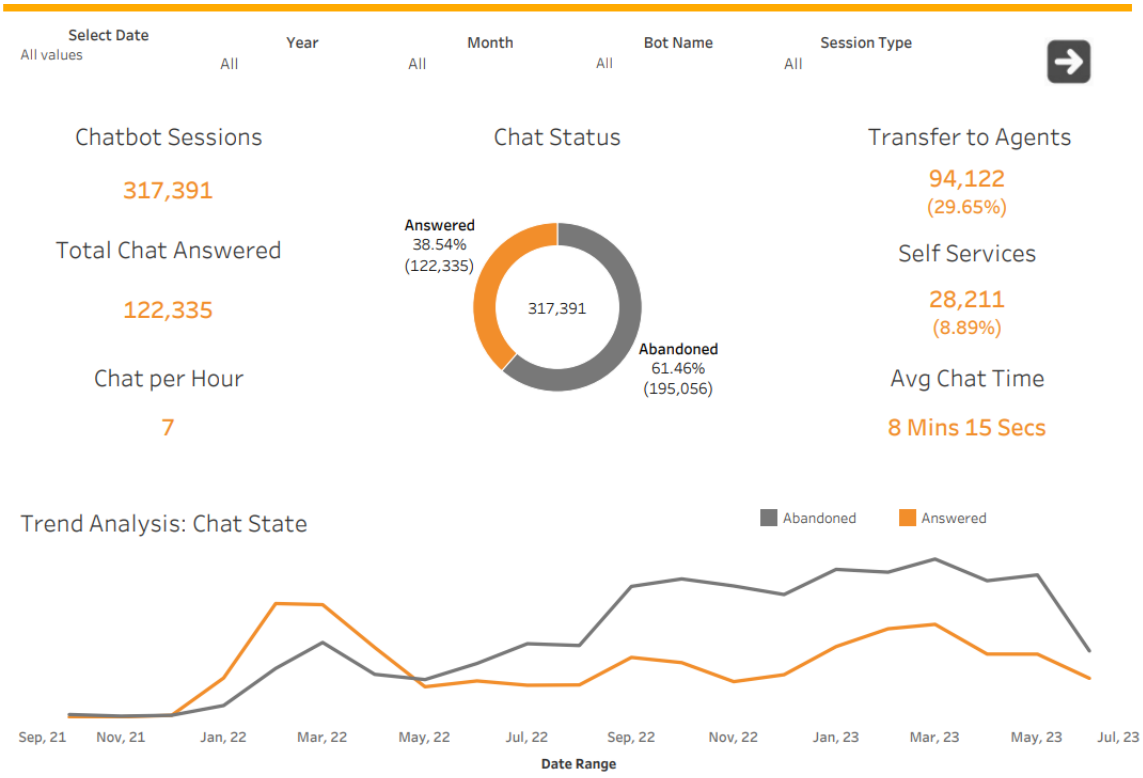


2.1.4 Data automation DAG

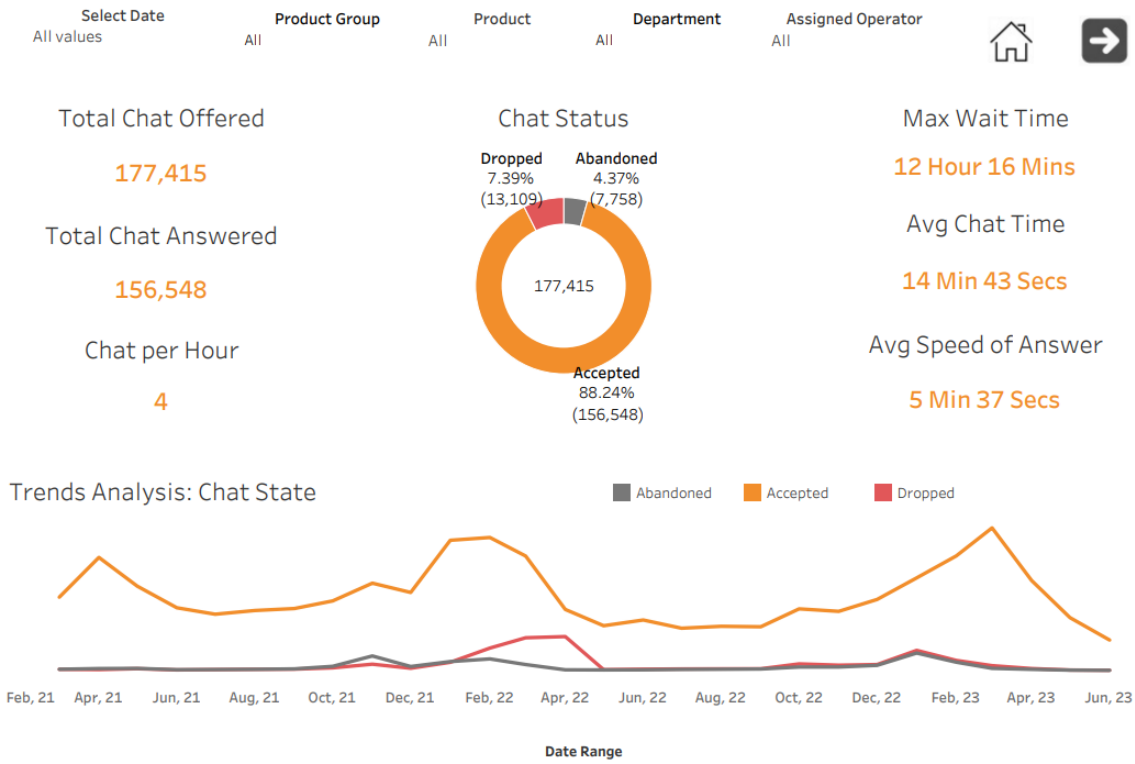


2.1.5 Some useful metrics/insights

Chatbot analysis (Number of chats peak in March 2022 and 2023):



Chat engine analysis (Number of chats peak in March 2022 and 2023):

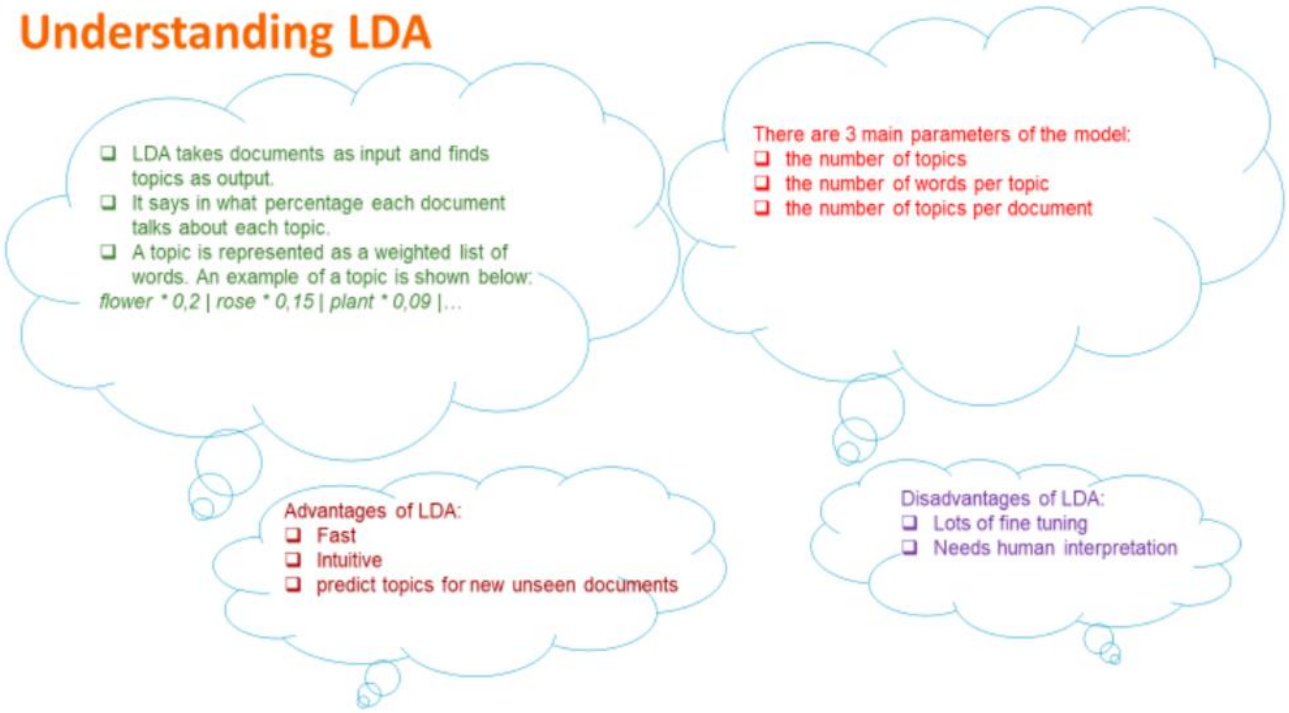


2.2 Text analytics of chatbot data/service tickets

After extracting data from the above process, the next task was to work on the text analytics of customers' interactions with the chat interfaces. The NLP model used is LDA (Latent Dirichlet Allocation).

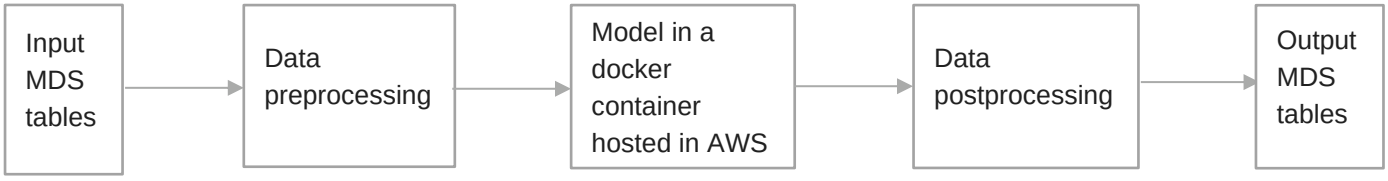
2.2.1 About LDA

(Unsupervised learning)



2.2.2 Deployment of the model

(Automated by scheduling daily/monthly in AWS)



2.2.3 Visualizing the model output

The output of the topic model is visualized through a Tableau dashboard (Input data for the dashboard is output data of the deployed model) where the conversations corresponding to a particular topic (a particular issue of customer – such as their electronic files are rejected) are displayed and a possible way to resolve the issue is presented (in the "Areas of opportunity box").

2.3 Finance metrics optimization of call center resources

The objective of this project is to allocate an optimal number of employees in the call center such that they can handle maximum number of incoming calls and minimal cost is incurred while allocating the call center employees.

2.3.1 Proof of concept

(On dummy data, can be later applied on real data)

Mathematical model:

Inputs:

Average cost per FTE in High-cost location per month: USD 2000

Average cost per FTE in Low-cost location per month: USD 1000

Average handle time per call for FTEs in High-cost location per month: 10 minutes

Average handle time per call for FTEs in Low-cost location per month: 20 minutes

X number of FTE to be employed in High-cost location

Y number of FTE to be employed in Low-cost location

Forecast of maximum number of calls for an entire product division in a day in 2023 = 10000 calls

Time available from each FTE per day= 8hrs

Considering 60% utilization, time available from each FTE per day=4.8hours=288 minutes

Maximum Number of calls that can be handled by FTE in high-cost location= (X*288)/10

Maximum Number of calls that can be handled by FTE in Low-cost location= (Y*288)/20

Constraint:

{(X*288)/10} + {(Y*288)/20}>=10000

X+Y>=242

Objective function to be minimized:

(X*2000) + (Y*1000)

3. Conclusion

From the internship, the major takeaways have been the practical experience of various Machine Learning and Data Engineering frameworks. Also, suggestions can be made to the company to use tools like generative AI to reduce dependency on chatbots and improve the capabilities of the “self-serve” platform (which is the first platform customers approach with their issues).