Summer Internship Report

Submitted in partial fulfillment of the requirements

Of

POST GRADUATE DIPLOMA

IN

BIG DATA ANALYTICS

Internal Obligation Logic Revamp

Mr. Akshada Rane

Roll Number: B2022009

GOA INSTITUTE OF MANAGEMENT

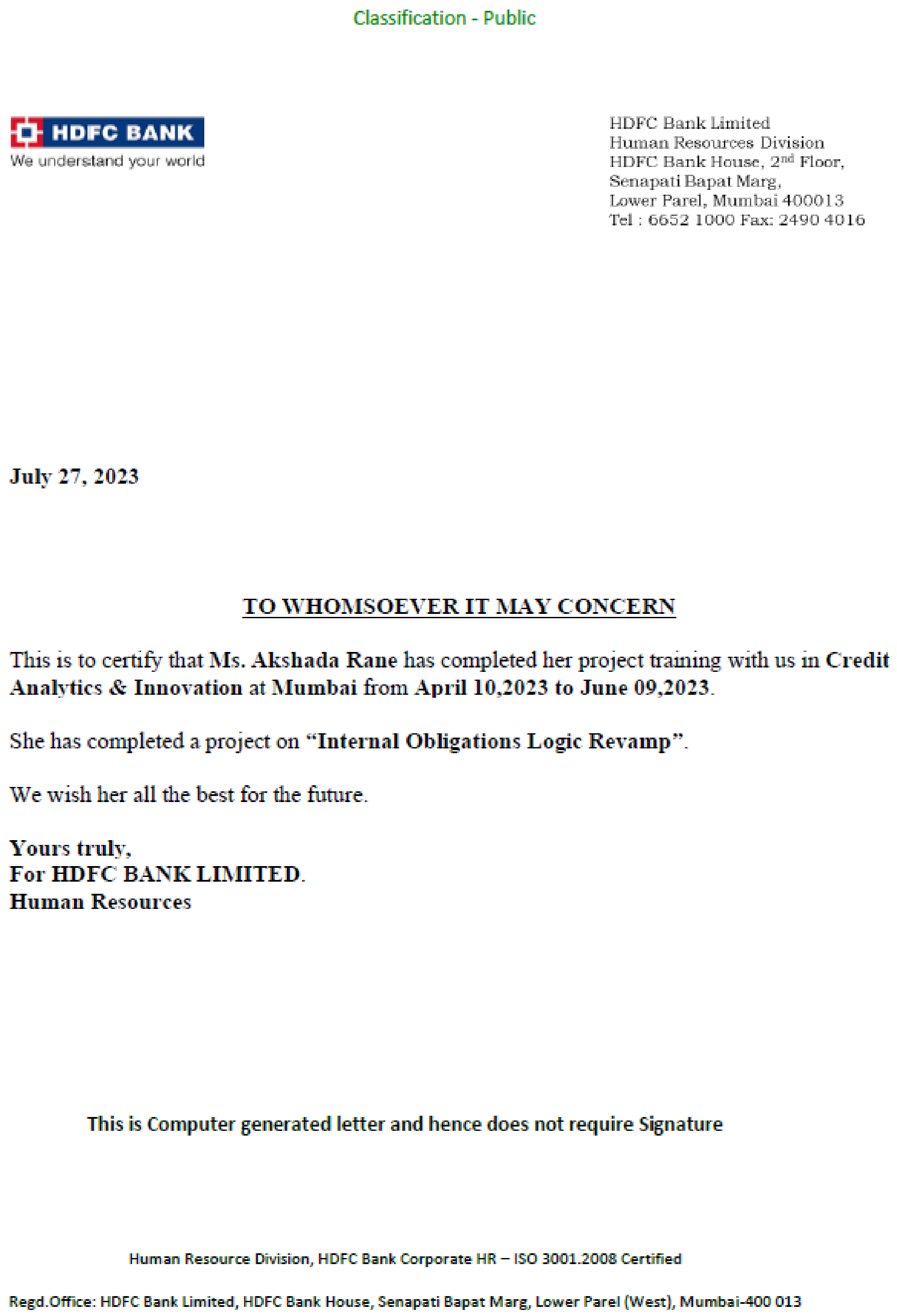
SANQUELIM - GOA

Batch: **2022-2024**

CERTIFICATE

This is to certify that this project report entitled **"Internal Obligation Logic Revamp"** has been submitted in partial fulfillment of the requirements for Post Graduate Diploma in Big Data Analytics at Goa Institute of Management is a bonafide work of Mr. Akshada Rane.

Mentor: Prof. Dr. Alekh Gaur ______________________



Classification - Public

HDFC Bank Limited
Human Resources Division
HDFC Bank House, 2nd Floor,
Senapati Bapat Marg,
Lower Parel, Mumbai 400013
Tel : 6652 1000 Fax: 2490 4016

July 27, 2023

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Ms. Akshada Rane** has completed her project training with us in **Credit Analytics & Innovation** at **Mumbai** from **April 10,2023 to June 09,2023**.

She has completed a project on **"Internal Obligations Logic Revamp"**.

We wish her all the best for the future.

Yours truly,
For HDFC BANK LIMITED.
Human Resources

This is Computer generated letter and hence does not require Signature

Human Resource Division, HDFC Bank Corporate HR – ISO 3001.2008 Certified

Regd.Office: HDFC Bank Limited, HDFC Bank House, Senapati Bapat Marg, Lower Parel (West), Mumbai-400 013

ACKNOWLEDGEMENT

I am grateful to **M/s. HDFC Bank LTD** for accepting me as an intern and providing me exponential learning opportunity in the areas I studied at Goa Institute of Management.

First and foremost, I am deeply thankful to my college faculty Prof. **Dr. Alekh Gaur** for your support and guidance throughout this journey have been instrumental in shaping my understanding of the business world.

I would also like to extend my appreciation to Corporate Relations and Placement Cell, PGDM-BDA for providing me with the opportunity to undertake this internship, which has been an invaluable learning experience.

Lastly, I would like to thank **Ms. Rinkesh Arondekar** who was my mentor in the internship for providing a guidance in resolving all the problems.

Table of Contents

1. Introduction to the company – HDFC Bank LTD.

1.1 Overview

HDFC Ltd, an early pioneer in the Indian financial sector,

received preliminary approval from the Reserve Bank of India (RBI) in 1994 to establish a private sector bank, marking a significant milestone in the liberalization of India's banking industry. HDFC Bank, officially incorporated in August 1994 with its headquarters in Mumbai, commenced operations as a Scheduled Commercial Bank in January 1995. Over the years, HDFC Ltd had built a strong reputation as one of the leading housing finance companies in India, offering a diverse range of products. HDFC Bank, on the other hand, expanded its services to encompass urban, semi-urban, and rural India, seamlessly providing home loans among its extensive product suite. This journey culminated on April 4, 2022, when the merger of HDFC Limited, India's largest housing finance company, and HDFC Bank, the country's largest private sector bank, was announced.

As of July 31, 2023, HDFC Bank boasted a widespread distribution network, comprising 7,895 branches and 20,462 ATMs/Cash Recycler Machines in 3,825 cities and towns across India. The merger integrated HDFC Ltd.'s distribution network, including 737 outlets and 214 offices of HDFC Sales Private Limited, into the bank's existing network. Furthermore, HDFC Bank extended its reach internationally, with branches in four countries and three representative offices in Dubai, London, and Singapore. These international branches and offices catered to Non-Resident Indians and Persons of Indian Origin, offering them a range of home loan products. This merger marked a significant step in the evolution of India's financial landscape, combining the strengths of two key players to provide enhanced financial services to a broader customer base.

The bank is committed to maintaining the highest level of ethical standards, professional integrity, corporate governance and regulatory compliance. HDFC Bank’s business philosophy is based on five core values as mentioned below:

Excellence

Customer Focus

Product Leadership

People

Sustainability

1.2 Departments

1. Product Team: This team is primarily responsible for the development and enhancement of the bank's products, with a focus on loans. They work on creating and improving the loan products offered by HDFC Bank.

2. Policy Team: The Policy Team plays a crucial role in innovating and adapting the products developed by the Product Team to ensure they comply with regulatory rules and reduce the risk of defaulters. They work to make the bank's loan products align with policies and regulations.

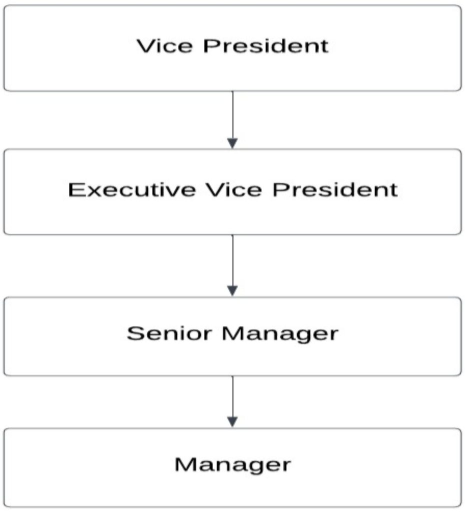
3. Campaign Analytics Team: This team is tasked with analysing data related to Existing to Bank (ETB) customers, who already have a relationship with the bank. They develop qualification criteria or logic to determine the creditworthiness of these customers and to identify those eligible for specific loan offers or products. Their goal is to increase the target customer base without increasing the defaulter rate.

4. Campaign Operation Team: Once the Campaign Analytics Team has developed the qualification criteria, the Campaign Operation Team is responsible for the actual rollout of offers and products to the relevant customers. They execute the distribution of messages, emails, and offers to creditworthy customers based on the criteria.

5. Data Scientist: Data Scientists in the organization are responsible for developing statistical models. They work closely with teams like Campaign Analytics to create data-driven insights and models that aid in decision-making. They serve as internal clients (stakeholders) for teams like Campaign Analytics.

In summary, the Product Team and Policy Team collaborate creating and adapting loan products, Campaign Analytics Team assesses the creditworthiness of ETB customers. The Campaign Operation Team executes the distribution of offers to eligible customers. Data Scientists develop statistical models, and Data Engineers maintain the data infrastructure and pipelines to support these processes. These teams work together to manage and mitigate the risk of defaulters in credit management while promoting HDFC Bank's products.

1.3 Hierarchy



2. Assignments

Internal Obligation Logic Revamp - HDFC Bank LTD

3. Project 1

3.1 Objectives

Correcting the Existing Logic of Customer’s Internal fixed Obligation Calculation through their Transaction Details.

3.2 Domain Understanding

Below are the Basic Concepts & problems & issues that were faced

Basic Concepts:

Fixed Obligation to Income Ratio (FOIR): Total Fixed Obligation/Monthly Salary

FOIR represents the ratio of a borrower's total fixed financial obligations to their monthly income. Fixed obligations include mandatory EMIs (Equated Monthly Instalments) that cannot be cancelled, such as loan repayments. We exclude certain expenses like mutual funds, leisure EMIs (e.g., OTT subscriptions), and periodic payments that can be cancelled at any time.

A higher FOIR indicates a higher risk associated with granting a loan.

Calculating Fixed Obligations

Fixed obligations can be categorized into two types: external and internal.

1. External Obligation: Calculated using credit bureau data.

2. Internal Obligation: Calculated using transactional data of HDFC Bank customers who use services like current and savings accounts. Various modules contribute to internal obligations:

- ACH (Automated Clearing House): Involves electronic fund transfers between banks, which customers are obligated to pay. These payments typically occur daily.

- ECS (Electronic Clearance Service): Similar to ACH, ECS helps customers pay EMIs by automatically debiting the specified bank account on a scheduled basis.

- Standing Instruction (SI): Transactions are automatically processed based on a specific cycle.

Internal Loans: These are loans provided by HDFC Bank.

PDC (Post Dated Cheques): Cheques with future dates specified for encashment.

- Liability: Savings and current accounts.

- Assets: Only loan accounts.

- Credit Card: Credit card information.

Campaign Analytics Team: This team is responsible for developing qualification criteria for offering products and services to customers. These criteria aim to cover the maximum number of customers while minimizing risk.

Loan Approval Processes:

- Straight Through Process (STP): Loans are approved in just 10 seconds without manual underwriting. The most eligible candidates, often those with salary accounts in HDFC Bank, receive loans through this process.

- GC/PQ (Green Channel/Pre-Qualification): In this process, an underwriter is involved, who may verify documents like salary slips for customers without HDFC salary accounts. The rules here are less strict compared to STP loans due to the presence of an underwriter.

The above analysis is conducted using transaction details from the existing customer database to classify and provide them with suitable offers.

Problems/Issues Faced

Correcting the Existing Logic of Customer’s Internal fixed Obligation Calculation through their Transaction Details.

The existing logic analyses transaction details of clients, specifically the narration field, to determine if a transaction is a fixed obligation or not.

Issues:

However, there are several inaccuracies in this process, resulting in incorrect calculations of fixed obligations which are as follows:

Post Analysing 10 Samples & reviewing the existing logic code, Issues observed were payments done to insurance, donations & stocks were incorrectly calculated as fixed obligations.

In recurring EMI payment, even slight variation in monthly amount is excluded.

Sample Cases:

Issues in Existing Logic	Sample of Narration field
Payments done to stock investments getting included	ACH D-BD-BSE Limited-TXVN31573853
Insurance payments getting included while calculation of fixed obligations	ACH D-TP ACH MaxLifeInsu-1108728444 ACH D-LIC OF INDIA-3173577230323
Payments done for donations getting included	ACH D-BD-UNICEF-TPJA2495657
In recurring EMI payments, even slight variation in monthly amount is getting excluded	

Root Cause: The absence of an extensive list of keywords and the infrequent updating of keywords are the root causes of this problem.

Only words like "mf" and "mutual" are currently excluded from analysis of ACH transactions if they appear in the transaction's narrative field; otherwise, all transactions are treated as fixed obligations. Therefore, Non-fixed obligations are also therefore included in the calculation of fixed obligations.

Business impacting bank:

The above problems impact bank's business such as

(1) The bank may give a lower loan amount to eligible customers, reducing profitability,

(2) The bank may face credit risk by providing higher loan amounts to customers who actually have lower fixed obligations.

(3) Can also loose on potential customers, also reducing profitability.

3.3. Data Understanding

To identify customers' fixed obligations, we conducted a keyword-driven analysis of their transaction details. Our approach began with a comprehensive compilation of keywords, aiming to capture a wide range of financial commitments. Initially, we analyzed transaction data from 10 sample customers to establish a foundational keyword list. Subsequently, we expanded our analysis to encompass a substantial dataset, including 1 million ACH and 1 million ECS transactions, further refining our keyword list. Employing data preprocessing techniques and automation via fuzzy logic, we achieved both precision and efficiency in recognizing fixed obligations from transaction narratives, crucial for calculating the Fixed Obligation to Income Ratio (FOIR).

3.4. Methodology: Tools, Techniques and Framework

3.4.1 Tools

Azure ML notebook

Excel

3.4.2 Techniques

Data Preprocessing

Text Analysis

Keyword Identification

Fuzzy Logic

3.4.3 Framework

After determining the issues in existing logic, payments made for insurance, stocks, and donations were to be excluded in the calculation of fixed obligations.

To solve the recurring EMI payment issue:

- Lender specific and some percentage deviation in EMI payment can be included.
- Only 1 or 2 months of transaction amount consistency to be checked instead of 3 months.
- To exclude payments done to insurance, stocks, donations list of keywords to be made exhaustive & should be periodically updated to keep the list of keywords up to date to give accurate results.

To improve the list of keywords in existing logic:

- 10 Samples of customer's transactions details of users were analysed.
- TTD logic was analysed especially EMI-master which contained various inclusion & exclusion keywords.

- 10L ACH & 10L ECS transactions were analysed.

Existing logic works mainly on the transaction types i.e., certain transaction types is assigned to each transactions & based on the keywords used in these transaction's narration field it is categorized as whether transaction is a fixed obligation or not & also mainly works on exclusion logic i.e., if narration field contains certain keywords, then exclude from calculation of fixed obligation else include else transaction amount to be included while calculation of fixed obligations.

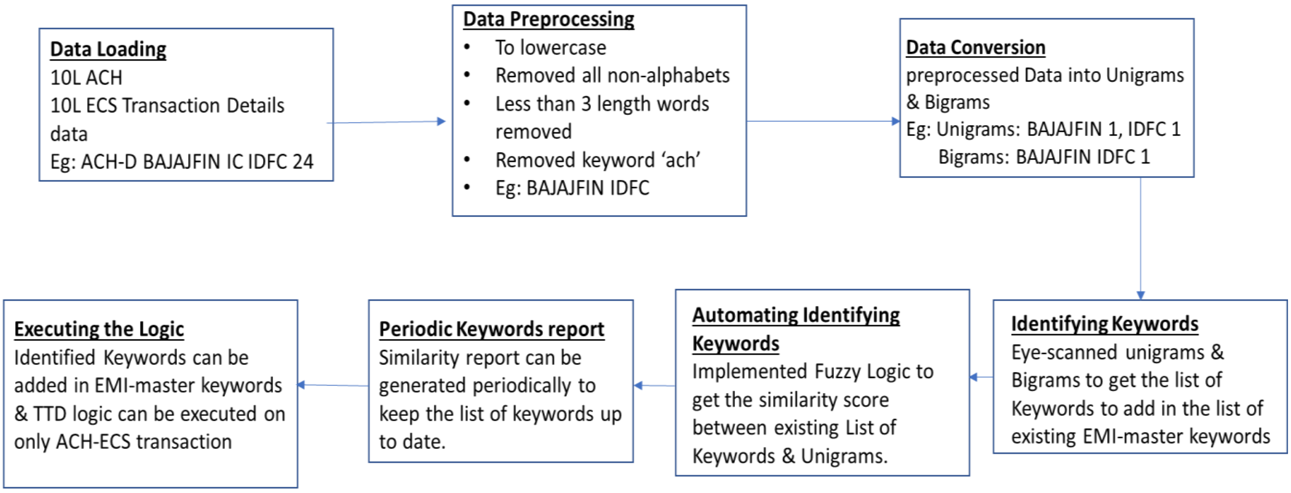
The Through the Door (TTD) logic incorporates the inclusion and exclusion of keywords to improve the accuracy of fixed obligation calculations. It utilizes an EMI_master Excel file that contains a list of keywords, further classified as Lender_Specific, Company_Specific, and Exclusion_List. Transactions with keywords from the exclusion list are not considered as fixed obligations.

However, there are limitations in the TTD logic related to the consideration of UPI, IMPS, and NEFT transactions, which may not reliably indicate fixed obligations. Additionally, the presence of the keyword "EMI" in the description field is currently considered as a fixed obligation, even though it may not always be reliable.

To overcome these limitations, the TTD logic can be executed on only ECS and ACH transactions, as they provide more accurate results in identifying fixed obligations. These transaction types are considered more reliable indicators of fixed obligations.

TTD EMI-master keywords can be compared and included in existing logic to make the list of keywords more exhaustive.

3.5 Work done



Data Loading
10L ACH
10L ECS Transaction Details data
Eg: ACH-D BAJAJFIN IC IDFC 24

Data Preprocessing
- To lowercase
- Removed all non-alphabets
- Less than 3 length words removed
- Removed keyword 'ach'
- Eg: BAJAJFIN IDFC

Data Conversion
preprocessed Data into Unigrams & Bigrams
Eg: Unigrams: BAJAJFIN 1, IDFC 1
Bigrams: BAJAJFIN IDFC 1

Identifying Keywords
Eye-scanned unigrams & Bigrams to get the list of Keywords to add in the list of existing EMI-master keywords

Automating Identifying Keywords
Implemented Fuzzy Logic to get the similarity score between existing List of Keywords & Unigrams.

Periodic Keywords report
Similarity report can be generated periodically to keep the list of keywords up to date.

Executing the Logic
Identified Keywords can be added in EMI-master keywords & TTD logic can be executed on only ACH-ECS transaction

Primary issue was to make the list of keywords as exhaustive as possible, therefore to identify these lists of keywords, we went through 10 Sample Customer’s transaction details & analysed their narration field to identify some of the keywords.

To make the list of keywords more exhaustive, 10L ACH & 10L ECS transactions were also analysed to come up with more relatable keywords, process followed to analyse these keywords is as followed:

First the 20L of transaction data was pre-processed:

1. Removal of non-alphabetic characters: Any non-alphabetic characters present in the narration field were eliminated. This step helps to ensure that only relevant text is considered for keyword analysis.

2. Conversion to lowercase: The narration field was converted to lowercase to standardize the text and avoid discrepancies due to capitalization.

3. Removal of short words: Words with a length of less than 2 characters were removed. This helps to filter out insignificant words that may not contribute to the identification of keywords.

4. Exclusion of "Ach" keyword: The keyword "Ach" was specifically removed from the narration field. This step likely eliminates any instances where the keyword itself may lead to false identification of fixed obligations.

After the pre-processing steps, the narration field was further analysed using a combination of unigrams (single words) and bigrams (pairs of words) to identify repeating keywords and patterns.

To identify additional keywords, the unigrams and bigrams were manually eye scanned. However, this process was found to be tedious and potentially less accurate due to the subjective nature of human analysis.

To automate the keyword identification process and improve accuracy, a fuzzy logic approach was implemented. This involved calculating similarity scores between the existing sample keywords and the unigrams and bigrams. Keywords with lower similarity scores were given more focus for further analysis.

This automated approach can be performed periodically to capture any evolving patterns and keywords in the transaction data & give more accurate results.

These newly list of keywords was added to EMI-master & TTD logic was executed on only ECS & ACH transaction details database by putting the filter on transaction types 9260-ECS, 9990-ACH etc.

3.6 Analytical Insights (Optional)

After performing the above implementation, the comparative results with analysis of existing logic & modified TTD logic are mentioned below:

Sr No	APP_ID_C	XNSYearMonth	Existing Logic	New(TTD Logic)	New (Rev_TTD_logic)	Remarks
1	47188437	2023-03	5000	0	0	ACH D- TP ACH INDIAfOLINE-1094462251 -> Considered in existing
2	57125253	2023-03	15000	0	0	ACH D- BD-BSE LIMITED-TXVN31697227 -> Considered in existing
3	57125259	2023-03	47079	0	41629	ACH D- RETAIL ASSETS CPC BE-NCACDOSYM230 -> Can consider as a loan
4	73461559	2023-04	12000	0	0	ACH D- INDIAN CLEARING CORP-D9137051X034 -

						>Can consider as loan in existing
5	66776496	2023-04	14312	4467	18279	ACH D-015720910000003-0BA3771A1A4DFE942 -> Not getting considered in TTD
6	57125712	2023-03	23499	40877	40877	existing -> consistency issue ABL not included, Justdial, LIC wrongly included
7	43976522	2023-03	5000	65936	65936	ACH D-HDFCLTD-340595230-> consistency issue in existing, Etmoney-> included in existing
8	57125781	2023-03	88978	85436	85436	CTAAIL -> included in existing
9	8814317	2023-03	117029	117029	117029	ACH D-HDFCLTD-339284260 -> included in existing
10	5877290	2023-03	135586	119677	119677	ACH D-TP ACH HDFCSTDLIF -1093283860 -> wrongly included in existing

			463483	433422	488863	

As verified out of 10 Samples 8 Samples gives accurate results.

Limitations:

ACH D-015720910000003-0BA3771A1A4DFE942: If narration field does not contain inclusion & exclusion keyword then that transactions will not be included.

Next Steps:

Periodic updation of EMI-master keywords.

Implement the fuzzy logic algorithm to automate the identification of new keywords from transaction data

By deploying the solution, we can enhance the accuracy of fixed obligation calculations, leading to improved loan assessments, reduced credit risks & increased profitability

4. Contributions and learning

Major contribution in the internship period to the team and the organization is:

1. Keyword Exhaustiveness: The primary contribution of our approach was the creation of an exhaustive list of keywords. This was achieved through the meticulous analysis of transaction details from 10 sample customers, followed by an extensive examination of 10 lakh (1 million) ACH and 10 lakh ECS transactions. This comprehensive keyword list forms the foundation for accurately identifying fixed obligations.

2. Data Preprocessing for Precision: We contributed to the process by implementing robust data preprocessing techniques. By removing non-alphabetic characters, converting text to lowercase, and eliminating short words, we ensured that the keyword analysis focused on relevant information, reducing noise in the data.

3. Fuzzy Logic Automation: One of the key innovations was the implementation of fuzzy logic for keyword identification. This automated approach calculated similarity scores between existing keywords and transaction data, prioritizing keywords with lower similarity scores for further analysis. This automation not only saved time but also improved the accuracy of fixed obligation identification.

4. Scalability and Periodicity: We demonstrated the scalability of our approach by analysing a substantial dataset of transaction details. Additionally, our method is designed to be performed periodically, capturing evolving patterns and keywords in transaction data over time. This adaptability ensures that our results remain accurate and up-to-date.

Major learnings during the internship period are:

1. Data Preprocessing Importance: We learned that data preprocessing plays a critical role in ensuring the accuracy of keyword analysis. Techniques like removing non-alphabetic characters and converting text to lowercase are essential for cleaning the data and reducing potential sources of error.

2. Automation Enhances Efficiency: Our experience highlighted the efficiency gains achieved through automation, particularly when dealing with large datasets. The use of fuzzy logic for keyword identification not only expedited the process but also reduced the subjectivity associated with manual analysis.

3. Keyword Evolution: We recognized that transaction data and financial landscapes evolve over time. By designing our approach to be performed periodically, we acknowledged the importance of staying up-to-date with emerging patterns and keywords to maintain the accuracy of fixed obligation identification.

4. Collaborative Approach: Collaboration between data analysts, domain experts, and automation specialists was instrumental in the success of our method. Combining domain knowledge with automation technologies led to a more effective and accurate keyword identification process.

In summary, our contributions encompassed the creation of an exhaustive keyword list, the implementation of precise data preprocessing techniques, the automation of keyword identification using fuzzy logic, and the recognition of scalability and periodicity as crucial factors. Our learnings emphasized the significance of data cleaning, automation, staying updated with evolving patterns, and collaborative efforts in data analysis processes.

5. Area of Improvement

Keyword List Enrichment: While efforts were made to compile an exhaustive list of keywords, continuous enrichment is essential. Regularly updating the keyword list based on evolving transaction narratives and financial products can further improve accuracy.

User Feedback Integration: Solicit feedback from users or domain experts who interact with the results of keyword identification. Their input can help refine the keyword list and improve the relevance of identified fixed obligations.

Feedback Loop with Stakeholders: Establish a feedback loop with stakeholders, such as loan officers or policy teams, to understand how the identified fixed obligations impact decision-making. This feedback can inform improvements in the keyword identification process.

Documentation and Knowledge Sharing: Document the keyword identification process thoroughly and share knowledge within the team. Clear documentation ensures continuity even if team members change.

6. Conclusion

In conclusion, our meticulous approach to identifying fixed obligations through keyword analysis in customer transaction details has yielded valuable insights and efficiencies. The creation of an extensive keyword list, coupled with data preprocessing techniques, has substantially improved the accuracy of our analysis. We have learned the importance of maintaining the exhaustiveness of our keyword database, as financial transactions evolve over time.

Furthermore, the implementation of automated fuzzy logic for keyword identification has significantly streamlined our process, reduced subjectivity and enhancing both efficiency and accuracy. This approach has paved the way for periodic automation, ensuring that our analyses remain relevant and up-to-date in a dynamic financial landscape. By integrating these newly identified keywords into our systems, we have effectively assessed fixed obligations and contributed to the calculation of the Fixed Obligation to Income Ratio (FOIR). Overall, these learnings and methodologies position us well for continued success in similar financial assessments in the future.

7. Notes and References

https://learn.microsoft.com/en-us/certifications/exams/dp-900/

https://www.youtube.com/watch?v=HXV3zeQKqGY&t=1390s

https://cred.club/check-your-credit-score/articles/difference-between-cibil-report-and-cibil-credit-score

Summer Internship Report

Submitted in partial fulfillment of the requirements

Of

POST GRADUATE DIPLOMA

IN

BIG DATA ANALYTICS

Data Processing Pipeline in AWS

Mr. Akshada Rane

Roll Number: B2022009

GOA INSTITUTE OF MANAGEMENT

SANQUELIM - GOA

Batch: **2022-2024**

CERTIFICATE

This is to certify that this project report entitled **“Data Pipeline in AWS”** has been submitted in partial fulfillment of the requirements for Post Graduate Diploma in Big Data Analytics at Goa Institute of Management is a bonafide work of Mr. Akshada Rane.

Mentor: Prof. Dr. Alekh Gaur _______________________

Kellogg’s.

TO WHOMSOEVER IT MAY CONCERN

ACKNOWLEDGEMENT

I am grateful to **M/s. Kellogg's** for accepting me as an Data Engineer intern and providing me exponential learning opportunity in the areas I studied at Goa Institute of Management.

First and foremost, I am deeply thankful to my college faculty **Prof. Dr. Alekh Gaur** for your support and guidance throughout this journey have been instrumental in shaping my understanding of the business world.

I would also like to extend my appreciation to Corporate Relations and Placement Cell, PGDM-BDA for providing me with the opportunity to undertake this internship, which has been an invaluable learning experience.

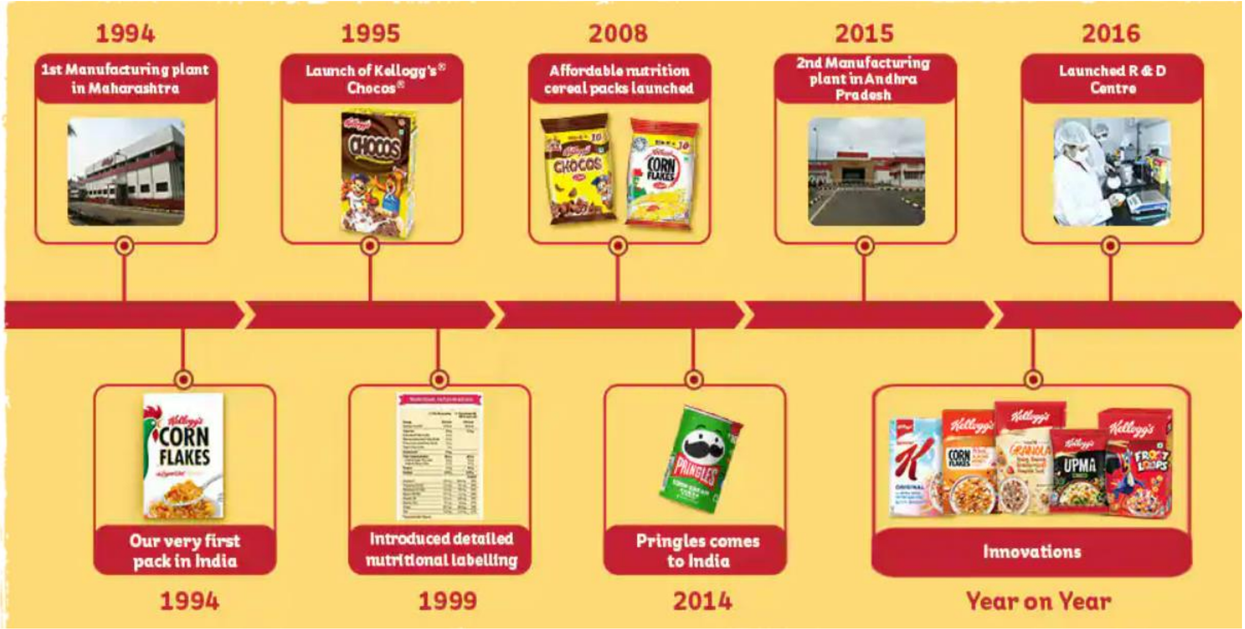
Lastly, I would like to thank **Ms. Rahul** who was my mentor in the internship for providing a guidance in resolving all the problems.

1. Introduction to the company – Kellogg’s

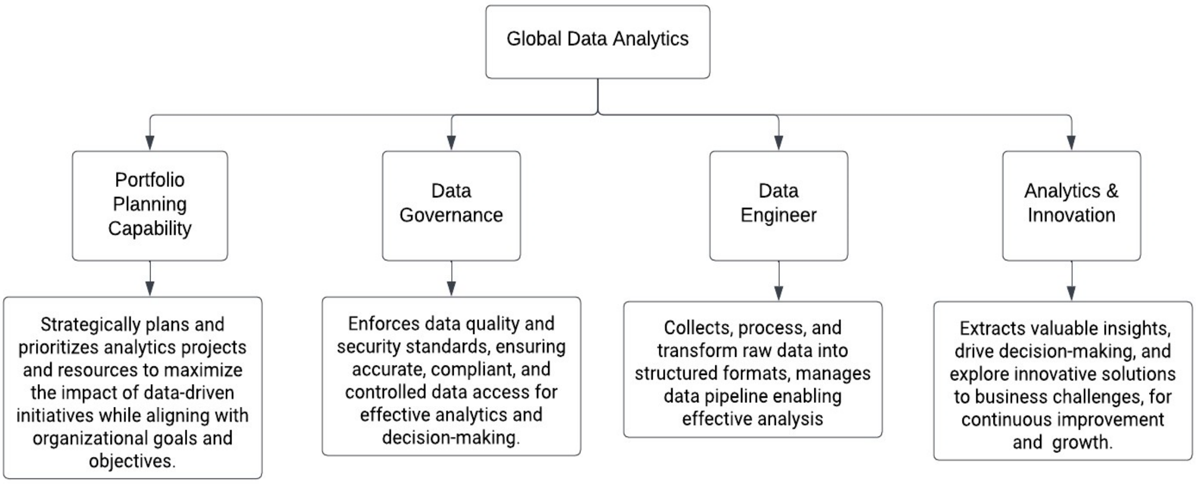
1.1 Overview

Kellogg's, a global leader, carries forward the enduring values instilled by W.K. Kellogg more than a century ago. Operating in 180 countries, including India and South Asia, Kellogg's commitment remains providing families with enhanced breakfast choices, a legacy stemming from W.K. Kellogg's pioneering introduction of breakfast cereal in 1898. The traditional corn-flaking method continues to be a hallmark, proven to deliver superior taste.

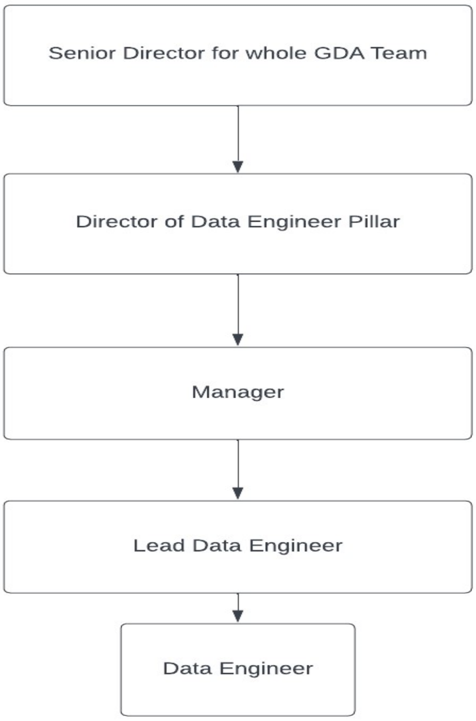
W.K. Kellogg's dedication to nutrition and well-being remains at the heart of Kellogg's mission. With active participation in India's "Make in India" initiative since 1994, Kellogg's offers a diverse range of cereals made from staple grains like wheat, corn, and rice, addressing the nutritional needs of homemakers, children, and adults alike. Kellogg's journey includes milestones such as Kellogg's Chocos in 1996, Project Red Alert's iron-fortified Kellogg's Corn Flakes in 1999, and the introduction of affordable INR 10/- SKUs in 2008. Expanding further, Kellogg's introduced Pringles chips in 2014 and ventured into the Indian breakfast market with Kellogg Upma in 2020. The year 2021 saw the introduction of Kellogg's Froot Loops, underscoring Kellogg's ongoing commitment to offering nutritious and delicious options to time-pressed consumers.


1994
1st Manufacturing plant in Maharashtra
1995
Launch of Kellogg's® Chocos®
2008
Affordable nutrition cereal packs launched
2015
2nd Manufacturing plant in Andhra Pradesh
2016
Launched R & D Centre
Our very first pack in India
1994
Introduced detailed nutritional labelling
1999
Pringles comes to India
2014
Innovations
Year on Year

1.2 Departments


Global Data Analytics
Portfolio Planning Capability
Data Governance
Data Engineer
Analytics & Innovation
Strategically plans and prioritizes analytics projects and resources to maximize the impact of data-driven initiatives while aligning with organizational goals and objectives.
Enforces data quality and security standards, ensuring accurate, compliant, and controlled data access for effective analytics and decision-making.
Collects, process, and transform raw data into structured formats, manages data pipeline enabling effective analysis
Extracts valuable insights, drive decision-making, and explore innovative solutions to business challenges, for continuous improvement and growth.

1.3 Hierarchy


Senior Director for whole GDA Team
Director of Data Engineer Pillar
Manager
Lead Data Engineer
Data Engineer

2. Assignments

1. Data Processing Pipeline in AWS
2. Conducted a POC of Apache Hudi
3. Implemented CICD (Continuous Integration & Continuous Deployment) pipeline

Project 1 -Phase 1

3.1 Objectives

The project implemented an AWS-based data processing and orchestration system with AWS S3, AWS Glue, Amazon Redshift, and Hudi. It uses event triggers, state machines, and Athena queries to efficiently handle the ingestion, transformation, and archiving of raw CSV files, ensuring data integrity and providing real-time monitoring. The primary objectives of the project were to automate data processing, implement SCD Type 2 transformations, maintain data quality, and enhance operational efficiency.

3.2 Domain Understanding

Data Warehouse:

Definition: A data warehouse is a centralized repository for storing, organizing, and managing data from various sources within an organization. It is designed to support business intelligence (BI) and data analytics activities.

Purpose: The primary purpose of a data warehouse is to facilitate data-driven decision-making by providing a structured and optimized environment for storing and analysing data.

Key Components: Data sources (where data originates), ETL (Extract, Transform, Load) processes (used to clean and transform data), a data repository (the warehouse itself), and tools for data querying and reporting are essential components of a data warehouse.

Key Challenges: Common challenges in data warehousing include managing the cost of infrastructure, dealing with the complexity of data integration, and ensuring data freshness or timeliness.

Data Lake:

Definition: A data lake is a flexible storage repository that can hold both structured and unstructured data at any scale. It is designed to store vast amounts of data in various formats.

Purpose: Data lakes provide a scalable and versatile repository for storing diverse data types from different sources. They enable organizations to retain raw, unprocessed data for future analysis.

Key Challenges: Challenges in data lakes include issues related to data reliability (lack of ACID transaction support), data quality (schema-on-read vs. schema-on-write approaches), and performance when dealing with large volumes of data.

Apache Hudi:

Lakehouse: A data lake house combines the best aspects of data warehouses and data lakes. It offers unified storage, processing, and analytics capabilities.

Apache Hudi: Apache Hudi is an open-source data management framework that serves as the foundation for a data lake house. It enhances efficiency, provides transactional capabilities, and ensures ACID compliance in data lakes.

Use Case: Apache Hudi is suitable for real-time data ingestion and managing historical data.

Parquet:

Definition: Parquet is an open-source, column-oriented data file format designed for efficient data storage and retrieval. It is highly optimized for analytics workloads.

Main Functionalities: Parquet offers columnar storage (which enhances query performance), has wide ecosystem support, and provides compression and encoding capabilities for efficient storage.

Glue:

Definition: AWS Glue is a serverless data integration service that offers fully managed ETL (Extract, Transform, Load) capabilities.

Key Features: Glue is serverless and scalable, provides a data catalog and metadata management, and seamlessly integrates with other AWS services for data processing and transformation.

Step Function:

Definition: AWS Step Functions is a serverless orchestration service that enables the coordination and management of workflows for applications.

Key Features: Step Functions offer workflow automation, state management, and a visual representation of workflow steps.

3.3. Data Understanding

In this section, we delve into a comprehensive understanding of the data management project we undertook. The project centered around orchestrating the efficient transformation of data

3.4. Methodology : Tools,Techniques and Framework

3.4.1 Tools

A fundamental component of our methodology was the utilization of several AWS (Amazon Web Services) tools and services.

AWS S3 (Amazon Simple Storage Service): A scalable storage service for data ingestion and retrieval.

AWS Glue: A serverless data integration service for ETL (Extract, Transform, Load) processes.

AWS Step Functions: A serverless orchestration service for workflow coordination.

Amazon Redshift: A fully managed data warehouse service for high-performance analytics.

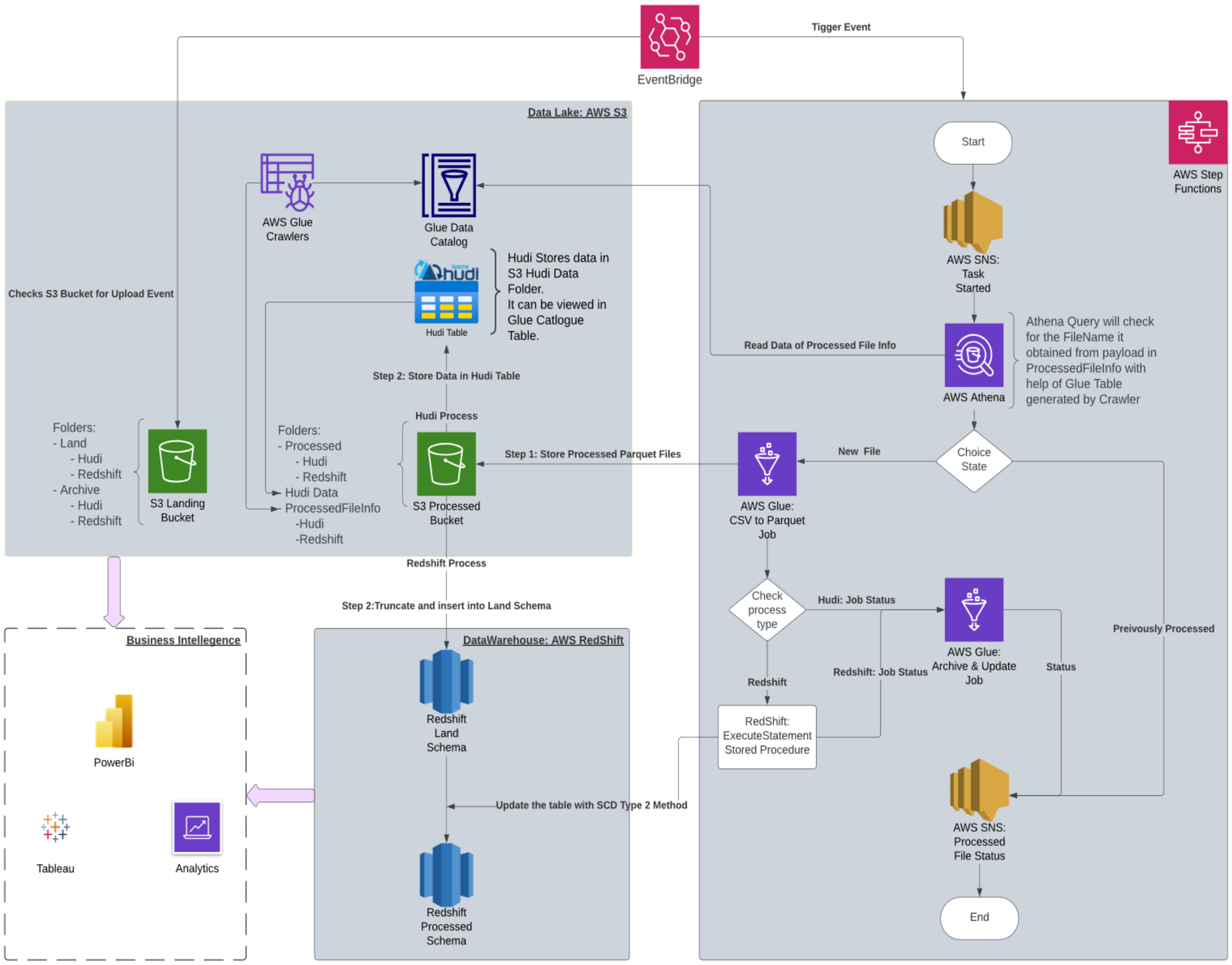
Apache Hudi: An open-source data management framework for efficient data storage, processing, and analytics.

SNS (Amazon Simple Notification Service): A notification service for real-time event communication.

EventBridge: An event-driven service for triggering workflows based on specific events and rules.

3.4.2 Framework

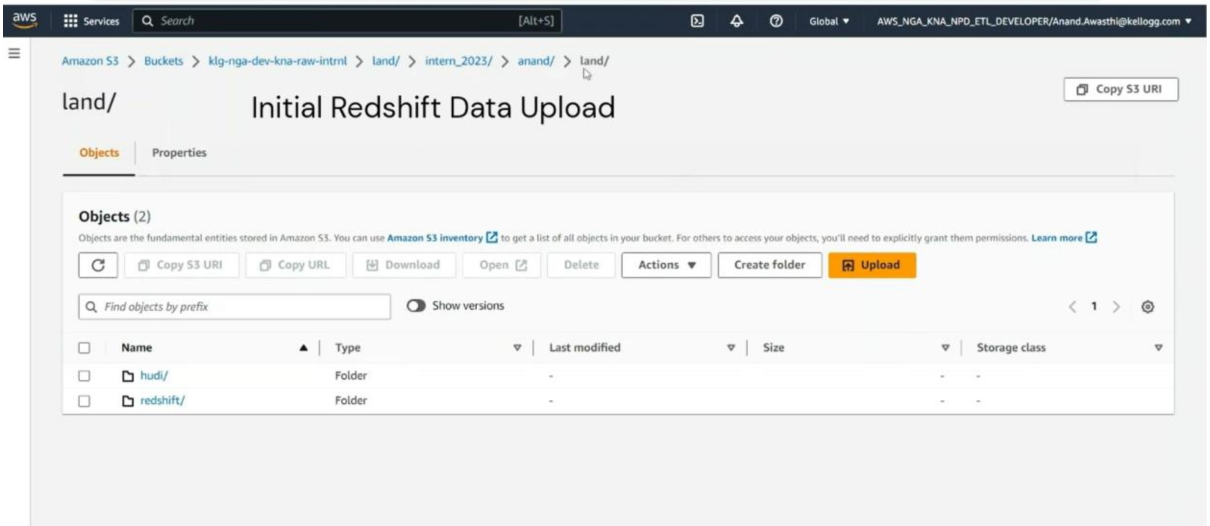
Below is the process diagram of data pipeline:


EventBridge
Tigger Event
Data Lake: AWS S3
AWS Glue Crawlers
Glue Data Catalog
Hudi Table
Hudi Stores data in S3 Hudi Data Folder.
It can be viewed in Glue Catlogue Table.
Checks S3 Bucket for Upload Event
Step 2: Store Data in Hudi Table
Hudi Process
Folders:
- Land
- Hudi
- Redshift
- Archive
- Hudi
- Redshift
S3 Landing Bucket
Folders:
- Processed
- Hudi
- Redshift
Hudi Data
ProcessedFileInfo
-Hudi
-Redshift
S3 Processed Bucket
Step 1: Store Processed Parquet Files
Redshift Process
Step 2:Truncate and insert into Land Schema
Business Intellegence
PowerBi
Tableau
Analytics
DataWarehouse: AWS RedShift
Redshift Land Schema
Update the table with SCD Type 2 Method
Redshift Processed Schema
AWS Step Functions
Start
AWS SNS: Task Started
Read Data of Processed File Info
AWS Athena
Athena Query will check for the FileName it obtained from payload in ProcessedFileInfo with help of Glue Table generated by Crawler
New File
Choice State
AWS Glue: CSV to Parquet Job
Check process type
Hudi: Job Status
AWS Glue: Archive & Update Job
Preivously Processed
Redshift
Redshift: Job Status
Status
RedShift: ExecuteStatement Stored Procedure
AWS SNS: Processed File Status
End

3.5 Work done

Our work encompassed the entire data processing and orchestration workflow. Raw CSV files were ingested into AWS S3 buckets, differentiating between files destined for Redshift and those for Hudi. A state machine orchestration was established through AWS EventBridge, incorporating Athena queries to ensure data freshness and prevent duplicate processing. Glue jobs were implemented to convert CSV files to Parquet, with resultant files stored in the Processed folder of S3. In the case of Redshift, data was loaded into landing tables and underwent SCD Type 2 transformations & were stored in Redshift processed table using Stored procedures. For Hudi, data remained in Hudi format in the S3 Processed bucket. Furthermore, we maintained an information CSV file for real-time monitoring and archival of processed data, ensuring data quality and governance. The completion of tasks triggered SNS notifications, and CSV files were archived from the Land table to the S3 Archive folder, concluding our comprehensive data processing workflow.

Find below the screenshots of the work done in the AWS:



Amazon S3 > Buckets > klg-nga-dev-kna-raw-intrnl > land/ > intern_2023/ > anand/ > land/

land/

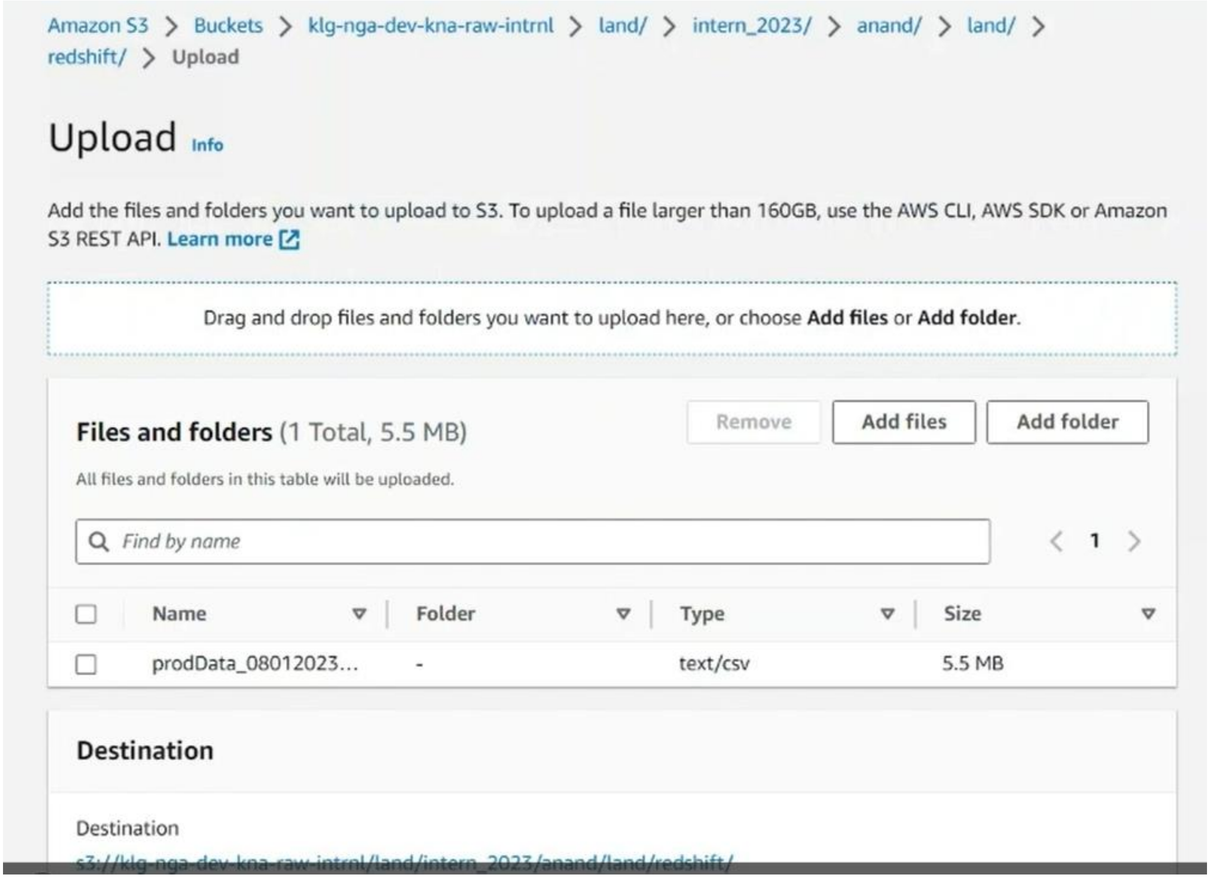
Initial Redshift Data Upload

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	hudi/	Folder	-	-	-
□	redshift/	Folder	-	-	-

Figure 1: File structure in S3 land bucket



Amazon S3 > Buckets > klg-nga-dev-kna-raw-intrnl > land/ > intern_2023/ > anand/ > land/ > redshift/ > Upload

Upload

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (1 Total, 5.5 MB)

All files and folders in this table will be uploaded.

□	Name	Folder	Type	Size
□	prodData_08012023...	-	text/csv	5.5 MB

Destination

Destination

s3://klg-nga-dev-kna-raw-intrnl/land/intern_2023/anand/land/redshift/

Figure 2: File upload in redshift folder

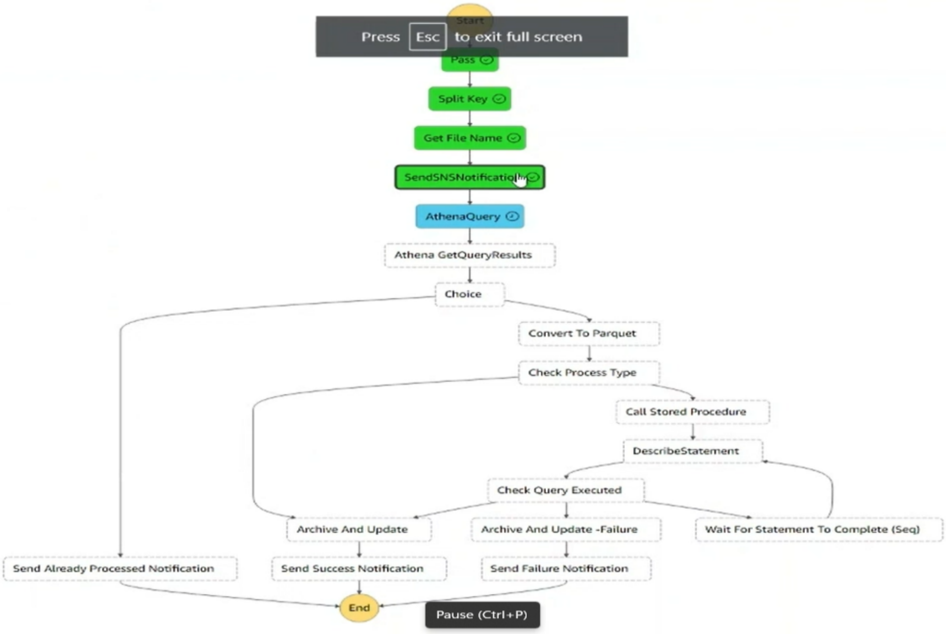
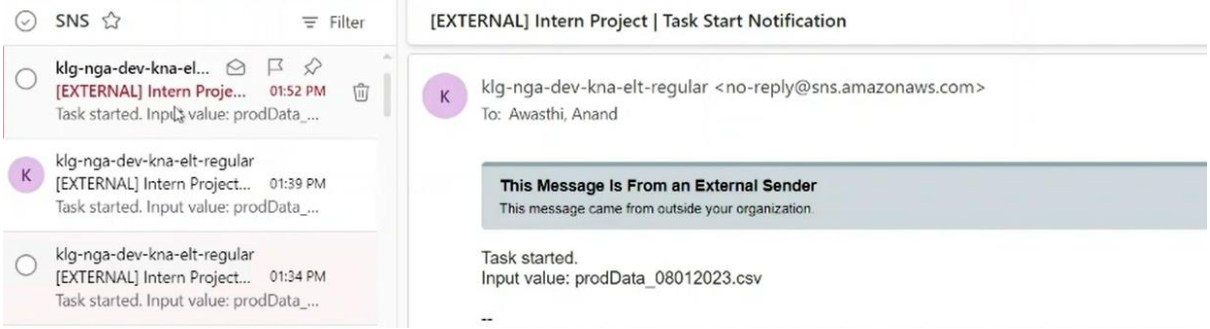


Figure 3: Step Function triggered



klg-nga-dev-kna-el...
[EXTERNAL] Intern Proje... 01:52 PM
Task started. Input value: prodData_...

klg-nga-dev-kna-elt-regular
[EXTERNAL] Intern Project... 01:39 PM
Task started. Input value: prodData_...

klg-nga-dev-kna-elt-regular
[EXTERNAL] Intern Project... 01:34 PM
Task started. Input value: prodData_...

[EXTERNAL] Intern Project | Task Start Notification

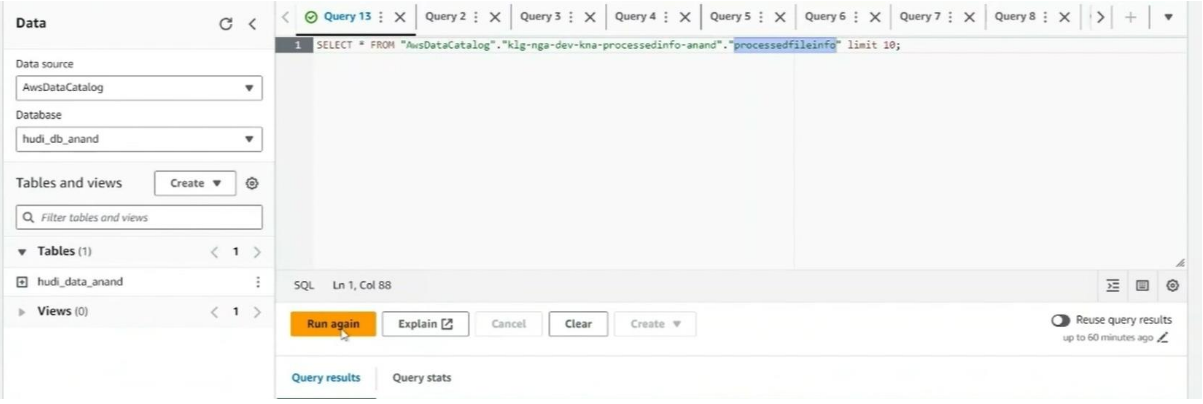
klg-nga-dev-kna-elt-regular <no-reply@sns.amazonaws.com>
To: Awasthi, Anand

This Message Is From an External Sender
This message came from outside your organization.

Task started.
Input value: prodData_08012023.csv

--

Figure 4: Received the SNS notification stating that the task has started

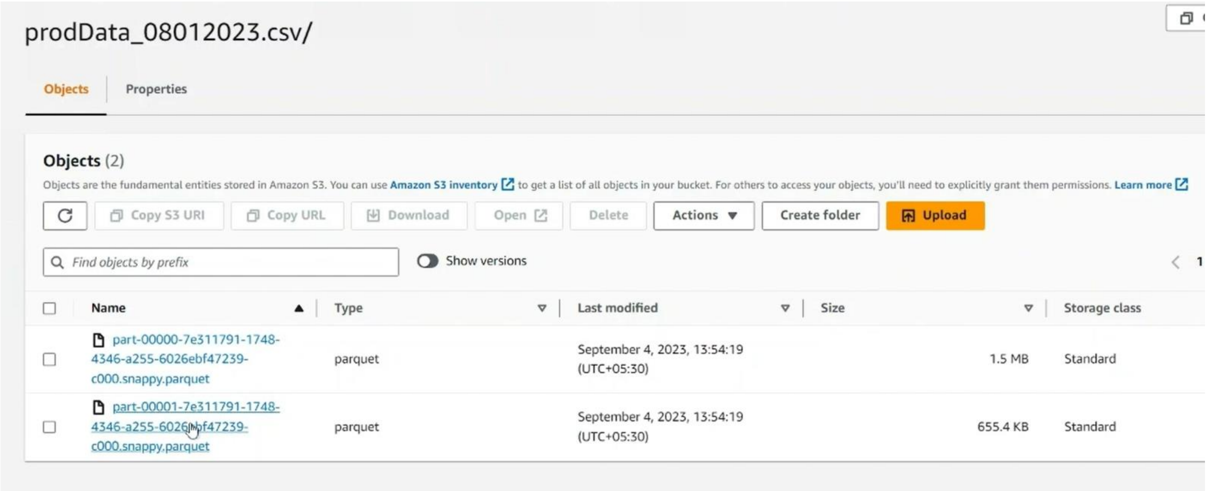


Data source: AwsDataCatalog

Database: hudi_db_anand

```
SELECT * FROM "AwsDataCatalog"."klg-nga-dev-kna-processedinfo-anand"."processedfileinfo" limit 10;
```

Figure 5: Checking in Athena if the file is already processed



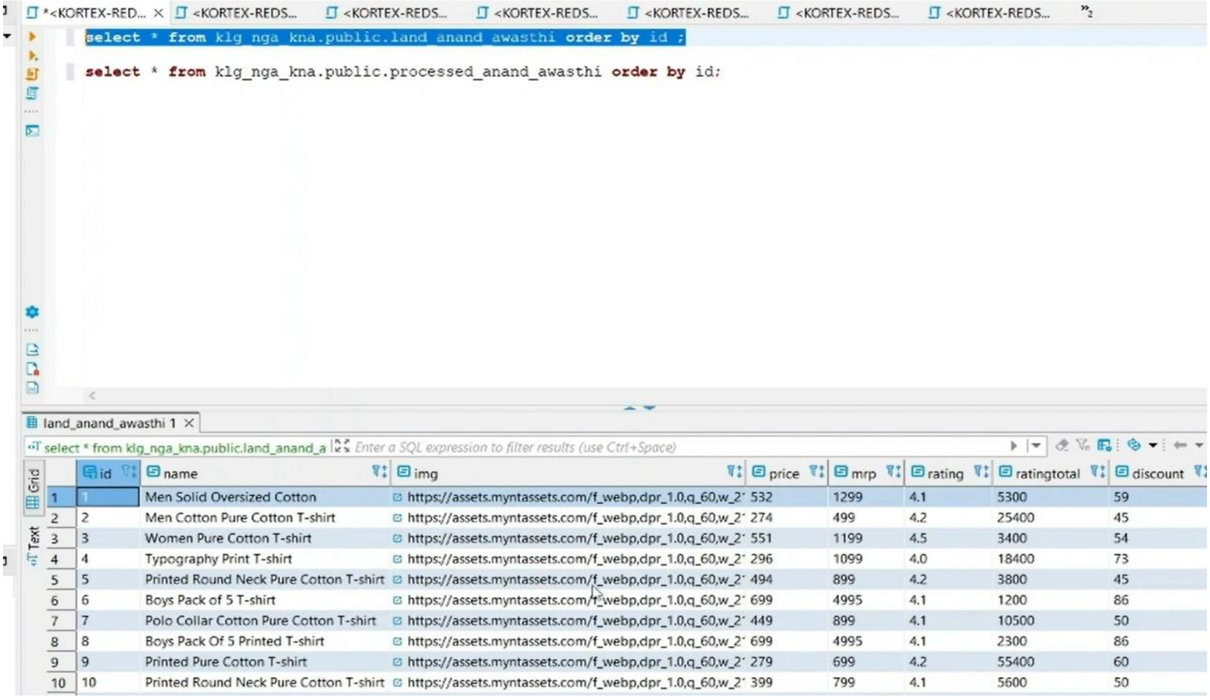
prodData_08012023.csv/

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	part-00000-7e311791-1748-4346-a255-6026ebf47239-c000.snappy.parquet	parquet	September 4, 2023, 13:54:19 (UTC+05:30)	1.5 MB	Standard
□	part-00001-7e311791-1748-4346-a255-6026ebf47239-c000.snappy.parquet	parquet	September 4, 2023, 13:54:19 (UTC+05:30)	655.4 KB	Standard

Figure 6: Once Glue Job is completed, transformed parquet file is stored in processed bucket of S3

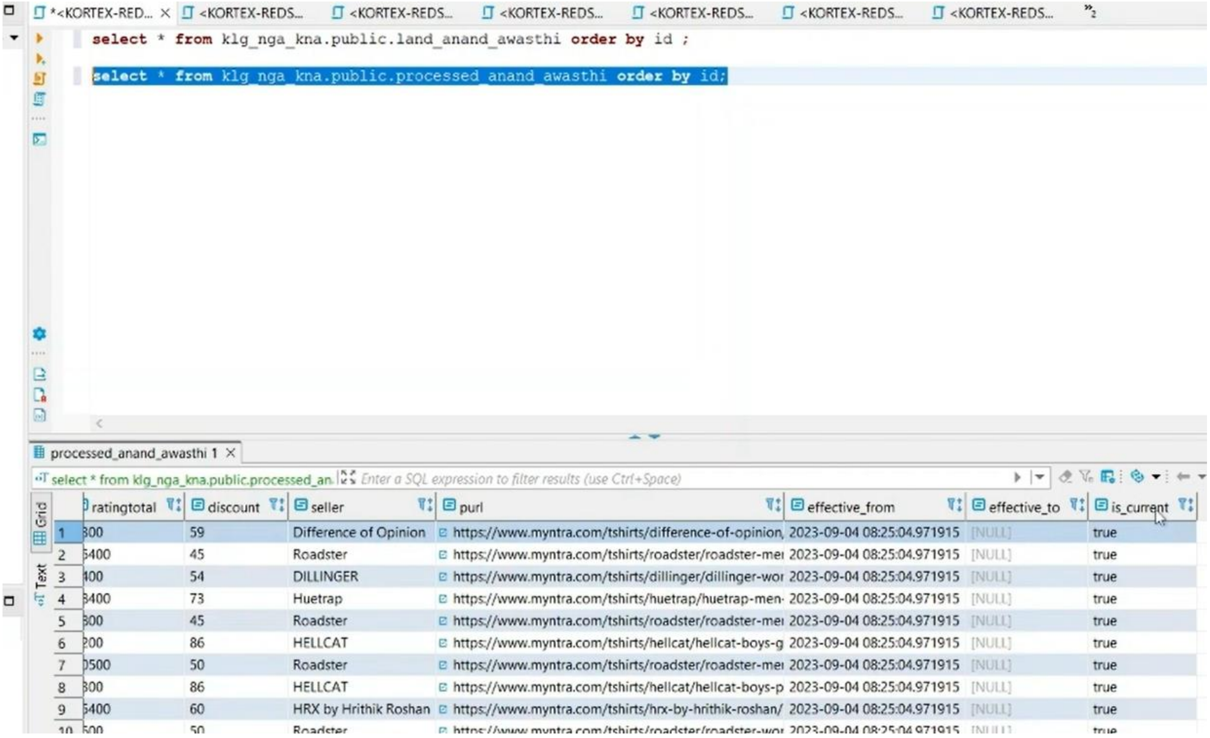


```
select * from klg_nga_kna.public.land_anand_awasthi order by id ;

select * from klg_nga_kna.public.processed_anand_awasthi order by id;
```

select * from klg_nga_kna.public.land_anand_a

	id	name	img	price	mrp	rating	ratingtotal	discount
1	1	Men Solid Oversized Cotton	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	532	1299	4.1	5300	59
2	2	Men Cotton Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	274	499	4.2	25400	45
3	3	Women Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	551	1199	4.5	3400	54
4	4	Typography Print T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	296	1099	4.0	18400	73
5	5	Printed Round Neck Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	494	899	4.2	3800	45
6	6	Boys Pack of 5 T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	699	4995	4.1	1200	86
7	7	Polo Collar Cotton Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	449	899	4.1	10500	50
8	8	Boys Pack Of 5 Printed T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	699	4995	4.1	2300	86
9	9	Printed Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	279	699	4.2	55400	60
10	10	Printed Round Neck Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	399	799	4.1	5600	50

Figure 7: Redshift landing table records

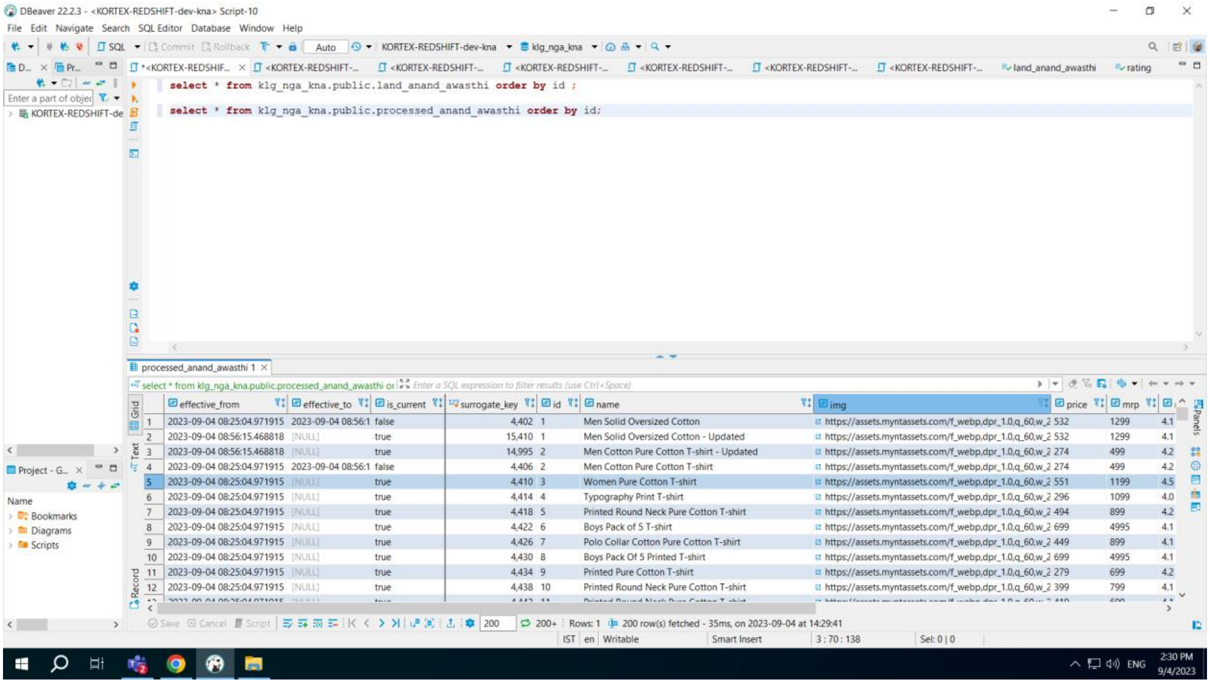


```
select * from klg_nga_kna.public.land_anand_awasthi order by id ;

select * from klg_nga_kna.public.processed_anand_awasthi order by id;
```

processed_anand_awasthi 1

	ratingtotal	discount	seller	purl	effective_from	effective_to	is_current
1	300	59	Difference of Opinion	https://www.myntra.com/tshirts/difference-of-opinion,	2023-09-04 08:25:04.971915	[NULL]	true
2	5400	45	Roadster	https://www.myntra.com/tshirts/roadster/roadster-mei	2023-09-04 08:25:04.971915	[NULL]	true
3	400	54	DILLINGER	https://www.myntra.com/tshirts/dillinger/dillinger-wor	2023-09-04 08:25:04.971915	[NULL]	true
4	3400	73	Huetrap	https://www.myntra.com/tshirts/huetrap/huetrap-men-	2023-09-04 08:25:04.971915	[NULL]	true
5	300	45	Roadster	https://www.myntra.com/tshirts/roadster/roadster-mei	2023-09-04 08:25:04.971915	[NULL]	true
6	200	86	HELLCAT	https://www.myntra.com/tshirts/hellcat/hellcat-boys-g	2023-09-04 08:25:04.971915	[NULL]	true
7	0500	50	Roadster	https://www.myntra.com/tshirts/roadster/roadster-mei	2023-09-04 08:25:04.971915	[NULL]	true
8	300	86	HELLCAT	https://www.myntra.com/tshirts/hellcat/hellcat-boys-p	2023-09-04 08:25:04.971915	[NULL]	true
9	5400	60	HRX by Hrithik Roshan	https://www.myntra.com/tshirts/hrx-by-hrithik-roshan/	2023-09-04 08:25:04.971915	[NULL]	true
10	500	50	Roadster	https://www.myntra.com/tshirts/roadster/roadster-wor	2023-09-04 08:25:04.971915	[NULL]	true

Figure 8: SCD type-2 transformed data in redshift processed table



DBeaver 22.2.3 - <KORTEX-REDSHIFT-dev-kna> Script-10

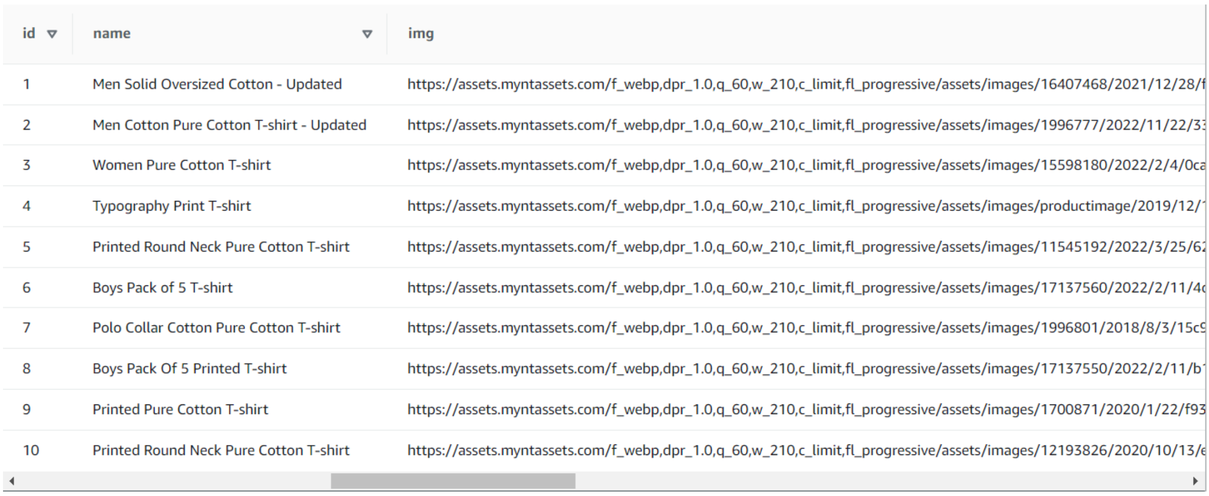
```
select * from klg_nga_kna.public.land_anand_awasthi order by id ;

select * from klg_nga_kna.public.processed_anand_awasthi order by id;
```

processed_anand_awasthi 1

	effective_from	effective_to	is_current	surrogate_key	id	name	img	price	mrp	
1	2023-09-04 08:25:04.971915	2023-09-04 08:56:1	false	4,402	1	Men Solid Oversized Cotton	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	532	1299	4.1
2	2023-09-04 08:56:15.468818	[NULL]	true	15,410	1	Men Solid Oversized Cotton - Updated	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	532	1299	4.1
3	2023-09-04 08:56:15.468818	[NULL]	true	14,995	2	Men Cotton Pure Cotton T-shirt - Updated	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	274	499	4.2
4	2023-09-04 08:25:04.971915	2023-09-04 08:56:1	false	4,406	2	Men Cotton Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	274	499	4.2
5	2023-09-04 08:25:04.971915	[NULL]	true	4,410	3	Women Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	551	1199	4.5
6	2023-09-04 08:25:04.971915	[NULL]	true	4,414	4	Typography Print T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	296	1099	4.0
7	2023-09-04 08:25:04.971915	[NULL]	true	4,418	5	Printed Round Neck Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	494	899	4.2
8	2023-09-04 08:25:04.971915	[NULL]	true	4,422	6	Boys Pack of 5 T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	699	4995	4.1
9	2023-09-04 08:25:04.971915	[NULL]	true	4,426	7	Polo Collar Cotton Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	449	899	4.1
10	2023-09-04 08:25:04.971915	[NULL]	true	4,430	8	Boys Pack Of 5 Printed T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	699	4995	4.1
11	2023-09-04 08:25:04.971915	[NULL]	true	4,434	9	Printed Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	279	699	4.2
12	2023-09-04 08:25:04.971915	[NULL]	true	4,438	10	Printed Round Neck Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_2	399	799	4.1

200 | 200+ | Rows: 1 | 200 row(s) fetched - 35ms, on 2023-09-04 at 14:29:41

Figure 9: Updated CSV uploaded in redshift landing table, shows the historical as well as latest data recorded in processed S3 bucket



id	name	img
1	Men Solid Oversized Cotton - Updated	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/16407468/2021/12/28/f
2	Men Cotton Pure Cotton T-shirt - Updated	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/1996777/2022/11/22/3
3	Women Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/15598180/2022/2/4/0ca
4	Typography Print T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/productimage/2019/12/
5	Printed Round Neck Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/11545192/2022/3/25/6
6	Boys Pack of 5 T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/17137560/2022/2/11/4
7	Polo Collar Cotton Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/1996801/2018/8/3/15c
8	Boys Pack Of 5 Printed T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/17137550/2022/2/11/b
9	Printed Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/1700871/2020/1/22/f93
10	Printed Round Neck Pure Cotton T-shirt	https://assets.myntassets.com/f_webp,dpr_1.0,q_60,w_210,c_limit,fl_progressive/assets/images/12193826/2020/10/13/e

Figure 10: Apache Hudi updated table

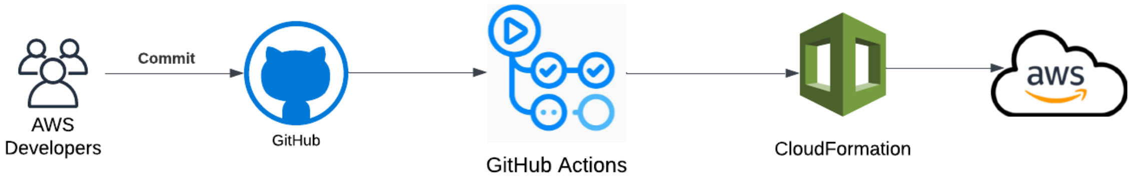
Phase 2:

The proof of concept for Apache Hudi aimed to demonstrate its capability to effectively roll back to a previous commit in cases where data upserts resulted in errors. Apache Hudi is a powerful data management framework that allows for efficient data storage and management within a data lakehouse architecture. In this context, the ability to roll back to a previous commit is crucial for data reliability and integrity. By conducting this proof of concept, the team sought to ensure that data inconsistencies or errors resulting from upsert operations could be swiftly and accurately corrected. This capability not only enhances data quality but also minimizes the potential impact of errors on downstream processes and analytics, making Apache Hudi a valuable asset in maintaining a robust and dependable data infrastructure.

Phase 3:

CICD Pipeline:

Streamlining Infrastructure & Deployment:



The provided sequence outlines a common workflow for AWS developers:

1. Developers Commit Code to GitHub: Developers write and commit their code changes to a GitHub repository. This serves as the central repository for source code.

2. GitHub Actions: GitHub Actions is used to set up workflows that automate various tasks, such as code testing, building, and deployment. These workflows are triggered automatically when code is committed or pushed to the repository.

3. CloudFormation: AWS CloudFormation is employed to define and provision the infrastructure required for the application or service being developed. Infrastructure as Code (IaC) templates are used to specify the AWS resources and their configurations.

4. AWS: Once the CloudFormation stack is defined, it is deployed on AWS. This includes creating and configuring resources such as EC2 instances, databases, and networking components based on the CloudFormation template.

In this workflow, the automation provided by GitHub Actions and the use of CloudFormation for infrastructure provisioning streamline the development and deployment process, enabling developers to focus on writing code and ensuring consistent and reliable deployments on AWS.

4. Contributions and learning

Contributions:

During internship at Kellogg's, had the opportunity to contribute significantly to the organization's data management and analytics initiatives. By implementing an efficient data processing workflow, played a vital role in enhancing data-driven decision-making capabilities. The automation of data ingestion, transformation, and storage in AWS S3 and Redshift streamlined the data handling process. The integration of Apache Hudi and the creation of a data lakehouse architecture further optimized data storage and analytics. Additionally, the implementation of real-time monitoring through AWS services, such as SNS notifications and EventBridge, improved operational efficiency and data governance.

Learning

My time at Kellogg's was a tremendous learning experience. I gained a deep understanding of data warehousing, data lakes, and data management best practices. Working with AWS tools and services, such as S3, Glue, and Step Functions, enhanced my technical skills in data integration and orchestration. The exposure to Apache Hudi broadened my knowledge of open-source data management frameworks. Moreover, I learned the importance of data quality, real-time monitoring, and the value of efficient ETL processes in a corporate environment

5. Area of Improvement

While I made significant contributions during my internship, I also identified areas for improvement. One such area is the further optimization of data processing pipelines to enhance scalability and efficiency, especially as data volumes continue to grow. Additionally, refining data quality checks and incorporating advanced analytics techniques could enhance the organization's decision-making capabilities. Finally, ongoing training and development in emerging technologies and data management practices would contribute to the organization's long-term success.

6. Conclusion

In conclusion, my internship at Kellogg's provided an invaluable opportunity to contribute to the organization's data management endeavours. Acquired a wealth of knowledge and skills, and I am grateful for the experiences and challenges that have contributed to my professional growth. While there are areas for improvement, I am confident that the foundation laid during my internship will serve as a stepping stone for future endeavours in data management and analytics. I am thankful for the support and guidance I received throughout my internship and look forward to applying what I've learned in my future career.

7. Notes and References

https://www.youtube.com/channel/UCzpHRBVnkzBfSsXostYuW1g

https://docs.aws.amazon.com/

https://hudi.apache.org/docs/use_cases/

https://medium.com/intermix-io/4-amazon-redshift-use-cases-collect-store-analyze-share-data-319053013e27