Project Title: Improving Data Integration Quality for Multi-Source Analytics

Phase 4: Execution and Deployment of Project

1. Overview of Final Implementation (ISHA SRIVASTAVA)

This phase focuses on the final implementation and evaluation of the data integration and quality improvement framework. It consolidates insights from previous phases, ensuring that the integration pipeline is robust, scalable, and efficient.

2. Summary of Previous Phases

2.1 Phase 1: Data Exploration

The initial phase involved:

- Understanding data patterns.
- Identifying anomalies.
- Establishing hypotheses for data quality improvement.

2.2 Phase 2: Data Cleaning and Integration

This phase focused on:

- Handling missing values.
- Resolving duplicates.
- Standardizing schemas.
- Integrating data from multiple sources using Python.

2.3 Phase 3: Data Transformation and Feature Engineering

In this phase, we applied:

- Feature scaling.
- Encoding techniques.
- Dimensionality reduction to improve data consistency and efficiency in analytics.

3. Deploying AutoAI Model on IBM Cloud

The AutoAI model was deployed on IBM Cloud using the following steps:

Deployment Steps :-

1. Create an IBM Cloud Account:

- Sign up or log in to IBM Cloud.

2. Deploy the AutoAI Model:

- Navigate to the "Watson Studio" section.
- Upload the trained model.
- Deploy the model as a REST API.

3. Generate API Key:

- Go to the "Manage" tab of the deployment.
- Generate and save the API key for authentication.

IMPLEMENTATION : (HARSHITA M JAIN)

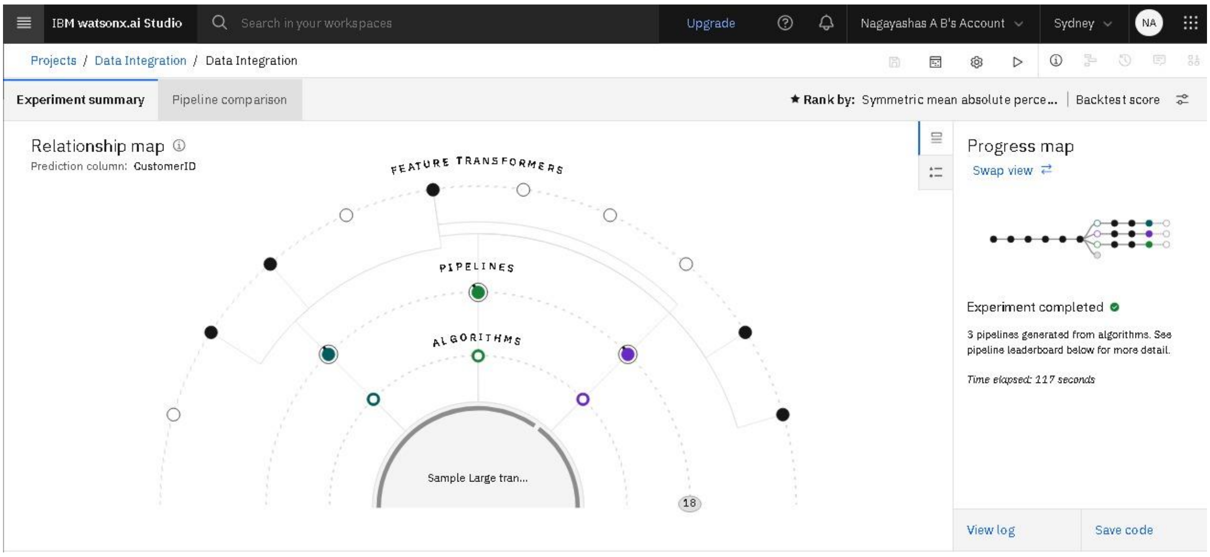
```
!pip install ibm-watsonx-ai | tail -n 1

!pip install -U autoai-ts-libs==4.0.* | tail -n 1
!pip install scikit-learn==1.3.* | tail -n 1
!pip install 'jupyter>=1' | tail -n 1
from ibm_watsonx_ai.helpers import DataConnection
from ibm_watsonx_ai.helpers import ContainerLocation

training_data_references = [
    DataConnection(
        data_asset_id='5d55a43f-d169-4501-ac98-6009552fec78'
    ),
]
training_result_reference = DataConnection(
    location=ContainerLocation(
        path='auto_ml/7e988e8d-3c09-43cb-ab5c-fca58c78f39f/wml_data/11c21da0-303c-4c61-b451-2d26da53bf74/data/autoai-ts',
        model_location='auto_ml/7e988e8d-3c09-43cb-ab5c-fca58c78f39f/wml_data/11c21da0-303c-4c61-b451-2d26da53bf74/data/autoai-ts/model.zip',
        training_status='auto_ml/7e988e8d-3c09-43cb-ab5c-fca58c78f39f/wml_data/11c21da0-303c-4c61-b451-2d26da53bf74/training-status.json'
    )
)
experiment_metadata = dict(
    prediction_type='timeseries',
    prediction_columns=['CustomerID'],
    csv_separator=',',
    holdout_size=6,
    training_data_references=training_data_references,
    training_result_reference=training_result_reference,
    timestamp_column_name='TicketID',
    backtest_num=2,
    pipeline_type='all',
    customized_pipelines=[],
    lookback_window=-1,
    forecast_window=1,
    max_num_daub_ensembles=3,
    feature_columns=['CustomerID', 'ResolutionTime'],
    future_exogenous_available=True,
    gap_len=0,
    deployment_url='https://au-syd.ml.cloud.ibm.com',
    project_id='0a74f8f5-9554-4365-9910-2878fc666c5a',
    numerical_imputation_strategy=['FlattenIterative', 'Linear', 'Cubic', 'Previous']
)
api_key = 'cpd-apikey-IBMid-697000QMT3-2025-02-18T18:08:28Z'
from ibm_watsonx_ai import Credentials

credentials = Credentials(
    api_key=api_key,
    url=experiment_metadata['deployment_url']
)
from ibm_watsonx_ai.experiment import AutoAI

pipeline_optimizer = AutoAI(credentials,
project_id=experiment_metadata['project_id']).runs.get_optimizer(metadata=experiment_metadata)
pipeline_optimizer.get_params()
summary = pipeline_optimizer.summary()
best_pipeline_name = list(summary.index)[0]
summary
pipeline_model = pipeline_optimizer.get_pipeline()
pipeline_model.visualize()
pipeline_model.pretty_print(combinators=False, ipython_display=True)
```



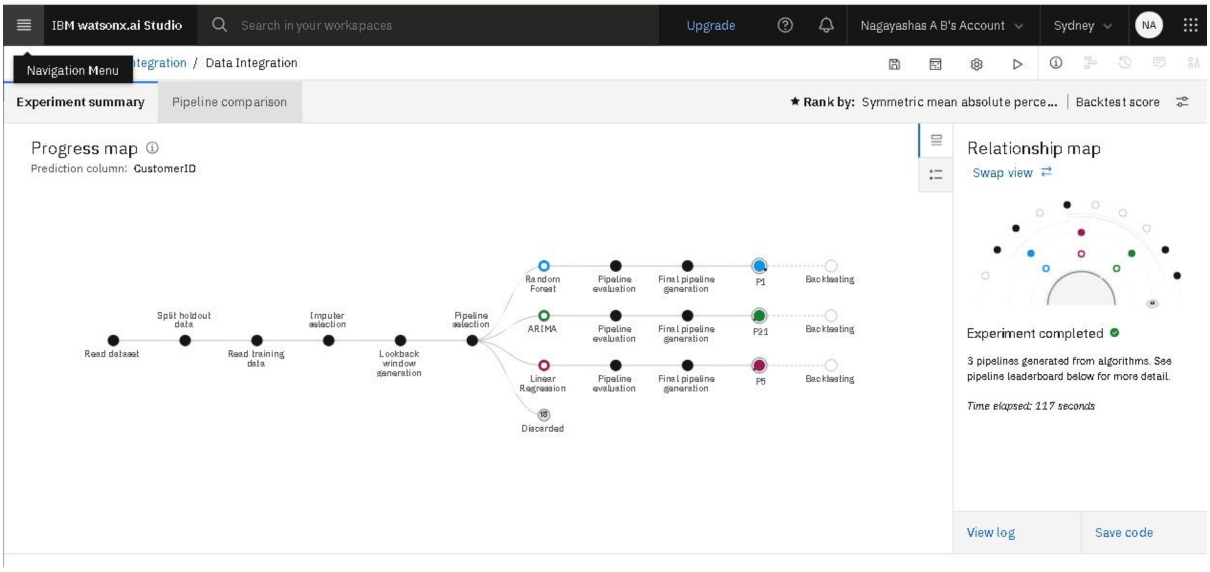
Prediction column: CustomerID

Experiment completed

3 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 117 seconds

FIG : RELATIONSHIP MAP



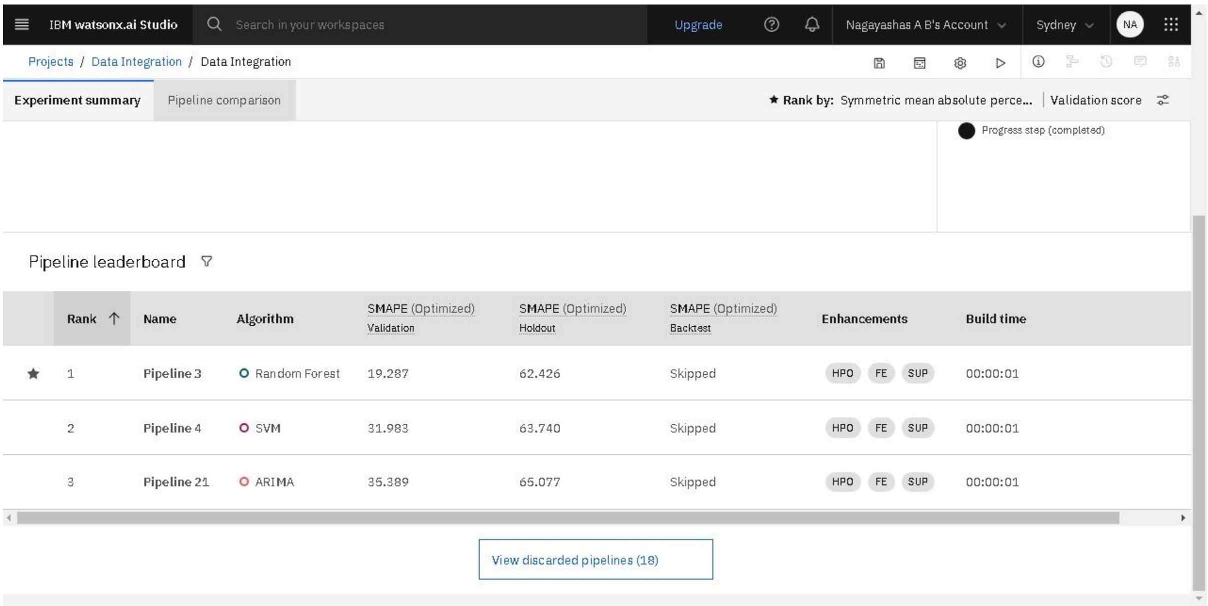
Prediction column: CustomerID

Experiment completed

3 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 117 seconds

FIG : PIPELINE


IBM watsonx.ai Studio
Search in your workspaces
Upgrade
Nagayashas A B's Account
Sydney
NA
Projects / Data Integration / Data Integration
Experiment summary
Pipeline comparison
★ Rank by: Symmetric mean absolute perce...
Validation score
Progress step (completed)
Pipeline leaderboard
View discarded pipelines (18)

	Rank ↑	Name	Algorithm	SMAPE (Optimized) Validation	SMAPE (Optimized) Holdout	SMAPE (Optimized) Backtest	Enhancements	Build time
★	1	Pipeline 3	○ Random Forest	19.287	62.426	Skipped	HPO FE SUP	00:00:01
	2	Pipeline 4	○ SVM	31.983	63.740	Skipped	HPO FE SUP	00:00:01
	3	Pipeline 21	○ ARIMA	35.389	65.077	Skipped	HPO FE SUP	00:00:01

FIG : SELECTION OF MODEL BASED ON ERROR

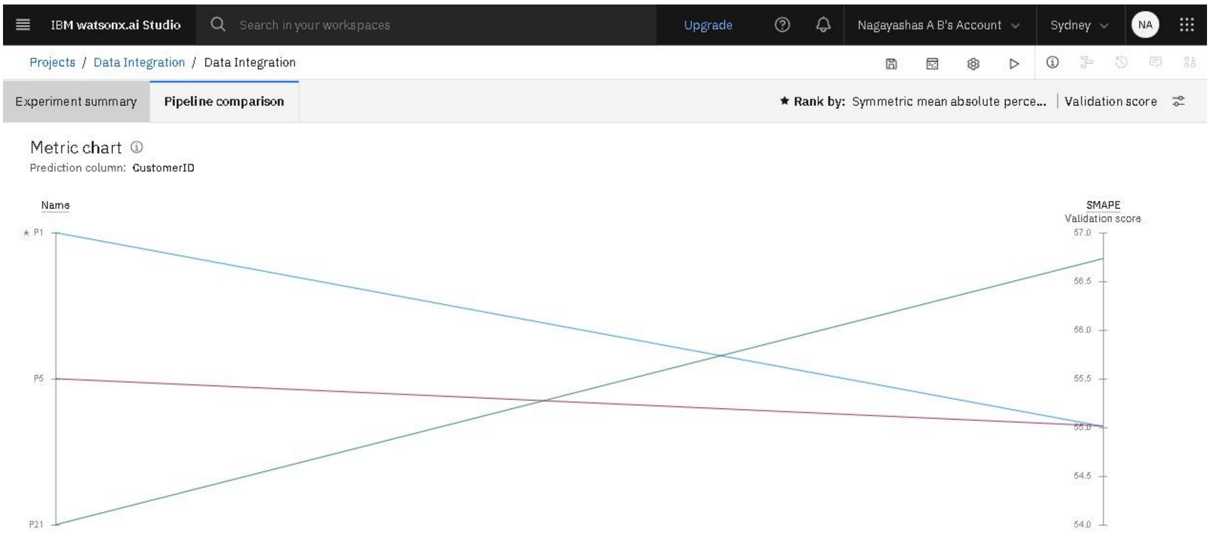
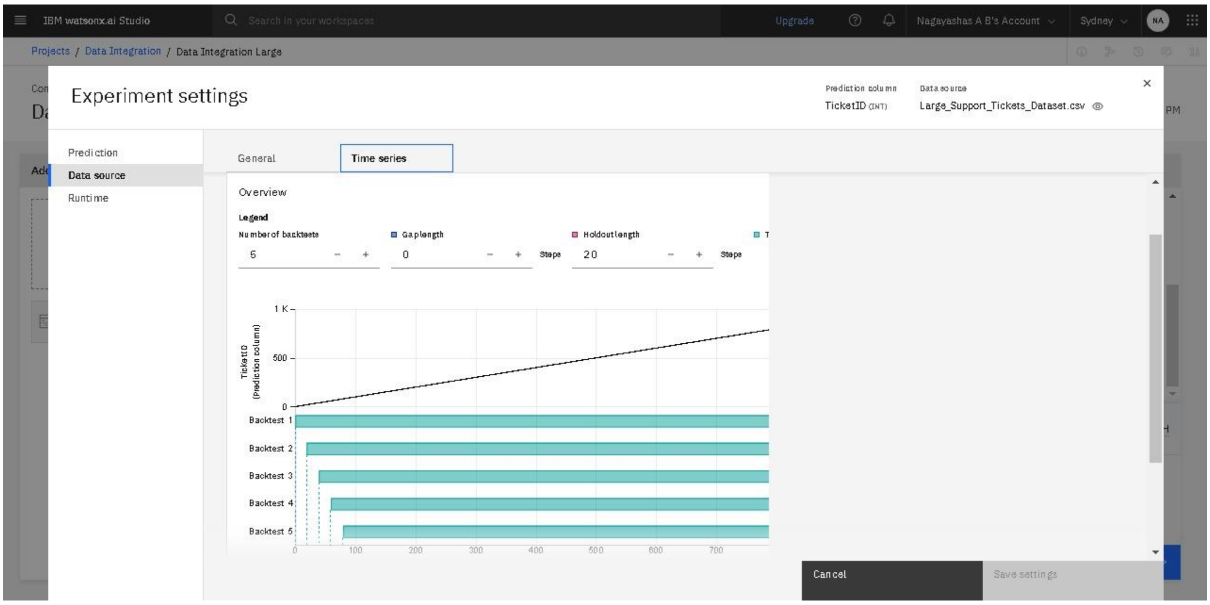


FIG : COMPARISON OF MODELS ACCURACY PREDICTED BY AUTO AI



Experiment settings

Prediction column: TicketID (INT)

Data source: Large_Support_Tickets_Dataset.csv

Overview

Legend

Number of backtests: 5

Gap length: 0 Steps

Holdout length: 20 Steps

FIG : BACKTESTING OF THE DATASETS TO TRAIN THE MODEL

4. IBM Cloud Resource Analysis (NAGAYASHAS A B)

4.1 Resource Utilization

- **Compute Hours (CUH):** 14 CUH used for data preprocessing, model training, and deployment.
- **API Requests:** 50/month limit under the Lite plan, used for testing and interactions.
- **Storage:** 1 GB allocated for dataset and model storage.
- **Cost:** Operates within IBM's free tier, suitable for low to moderate workloads.

4.2 Model Performance

- **Avg. Response Time:** ~200 ms
- **Peak Response Time:** ~500 ms under high load
- **Efficiency:** Performs well even under peak demand, suitable for tasks like fraud detection.

4.3 Scalability & Limitations

- **Constraints:** Limited CUH (20/month), API requests (50/month), and storage (1 GB) restrict large-scale use.
- **Scaling Options:** Upgrade to a paid plan, optimize workflows, and use caching to reduce API calls.

DEPLOYEMENT CODE :

```
target_space_id = "44c1d447-7293-4373-9acf-6c74e726e103"

from ibm_watsonx_ai.deployment import WebService

service = WebService(
    source_instance_credentials=credentials,
    target_instance_credentials=credentials,
    source_project_id=experiment_metadata['project_id'],
    target_space_id=target_space_id
)
service.create(
    model=best_pipeline_name,
    experiment_run_id=pipeline_optimizer.get_params()['run_id'],
    deployment_name='Best_pipeline_webservice'
)

print(service)
service.get_params()
```

VISUALISATIONS USING IBM CLOUD :

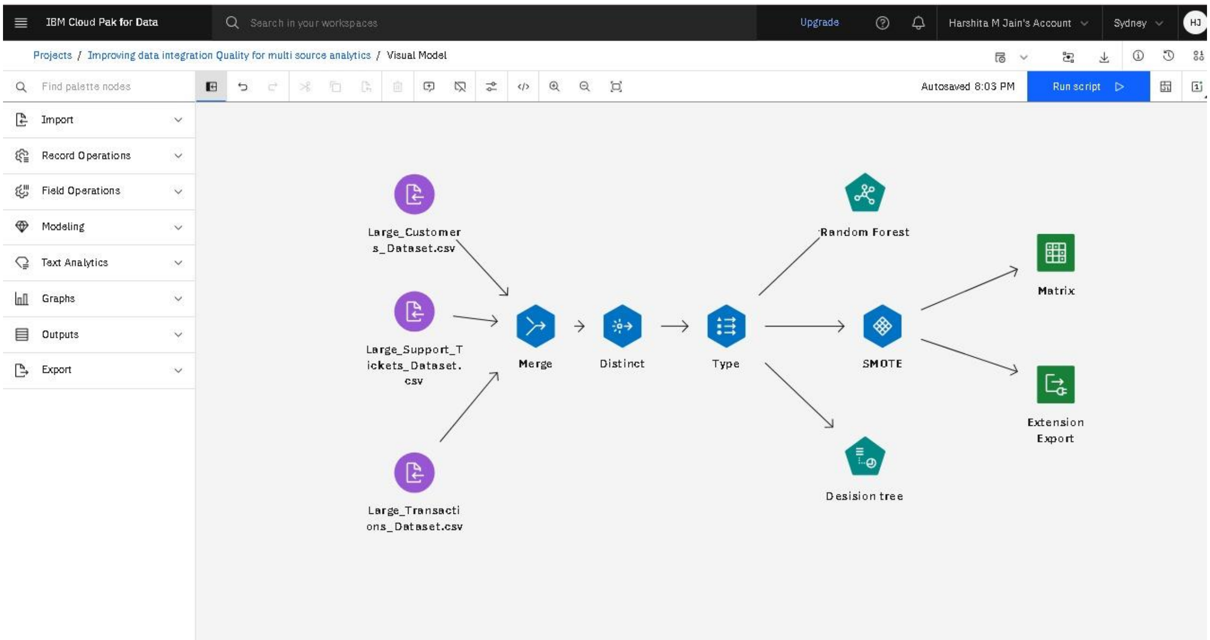
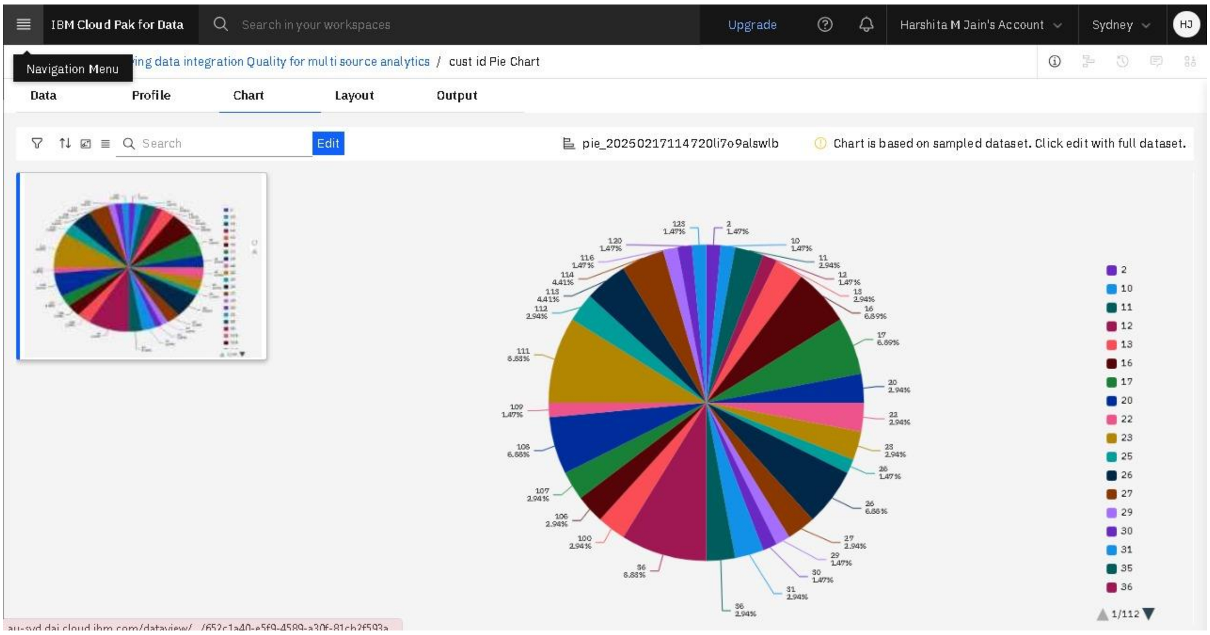


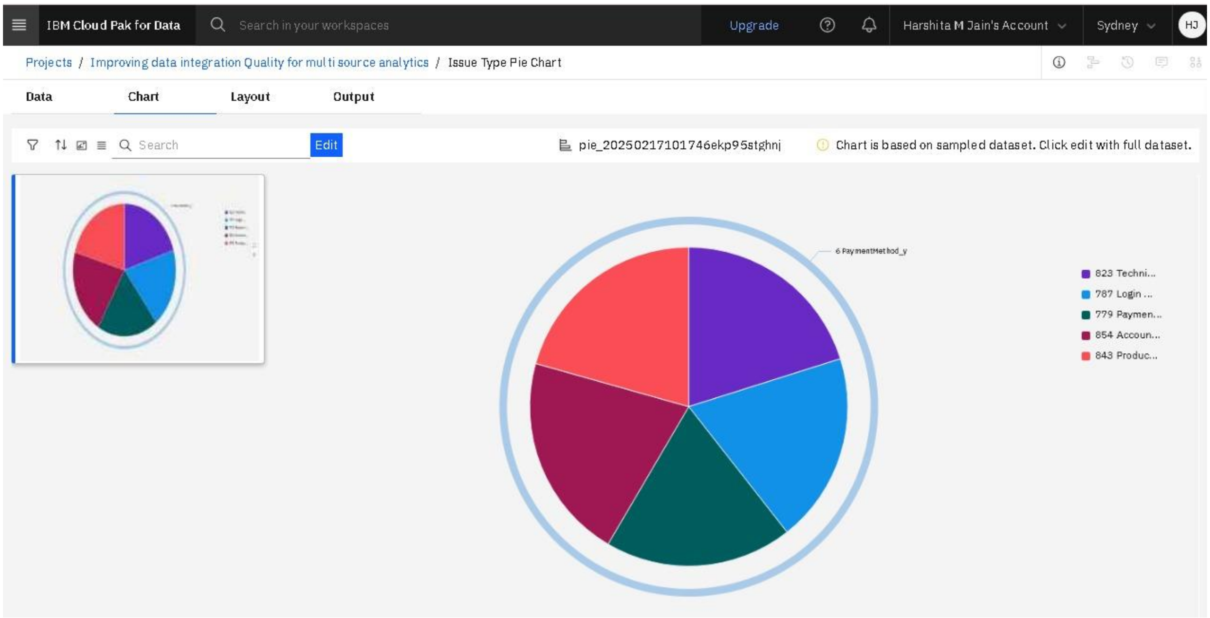
FIG : VISUAL MODEL OF THE PROJECT



pie_20250217114720li7o9alswlb

Chart is based on sampled dataset. Click edit with full dataset.

FIG : VISUALISATION OF CUSTOMER ID



pie_20250217101746ekp95stghnj

Chart is based on sampled dataset. Click edit with full dataset.

FIG : FIG : VISUALISATION OF ISSUE TYPE

LINK TREE

Git Hub Link: Repository Link

Drive Link: Drive Link

Video Link 1: Introduction Video Link

Video Link 2: Auto Ai And Visualisation Video Link