Banking Dataset: Predictive Marketing Insights

University of Maryland, Baltimore County

Department of Data Science

Date: 05/14/2024

Project Advisor: Mesfin Abate

Group Members:

Lokesh Reddy Venna

Upendra Madaraboena

Charan Pathakamuri

Sashank Gandikota

Contents

1.Business Process and Requirements

1.1. Database System Overview:

The database system for this project is designed to support the direct marketing campaigns of a Portuguese banking institution focused on selling term deposits. The primary goal is to predict whether a client will subscribe to a term deposit based on various customer attributes and interaction history.

1.1.1. Customer Segmentation and Benefit:

The database system categorizes customers based on demographics, financial history, and previous interactions with the bank. By analyzing this data, the system aims to identify potential customers who are most likely to subscribe to a term deposit. This segmentation allows the bank to tailor its marketing efforts more effectively, resulting in higher conversion rates and increased revenue.

1.1.2. Key Components of the Database System:

- Customer Information: Includes demographic details such as age, education, marital status, and job type. Financial information such as income level, and loans(housing and personal).

- Interaction History: Tracks previous interactions between the bank and customers, including past term deposit subscriptions, responses to marketing campaigns, number of campaigns per individual.

- Predictive Analytics Models: Utilizes the Relational database management system to analyze historical data and predict the likelihood of a customer subscribing to a term deposit. This system trained on features such as customer demographics, financial behavior, and response patterns.

1.1.3. Benefits to Customers:

- Personalized Offers: Customers receive targeted offers based on their financial needs and behavior, enhancing their banking experience.

- Timely Communication: Customers are contacted through preferred channels at optimal times, reducing unnecessary marketing outreach and improving engagement.

- ❖ Financial Planning: By considering term deposits, customers can make informed decisions about their savings and investments, potentially increasing their wealth over time.

1.2. Data Storage and Capture Requirements:

1.2.1. Data to be Stored:

- ❖ Bank client data including age (numeric), job type (categorical), marital status (categorical), education level (categorical), credit default status (binary), average yearly balance (numeric), housing loan status (binary), personal loan status (binary).

- ❖ Last contact details including communication type (categorical), last contact day of the month (numeric), last contact month of the year (categorical), last contact duration in seconds (numeric).

- ❖ Campaign-related attributes such as the number of contacts performed during this campaign and for this client (numeric), days since the client was last contacted from a previous campaign (numeric), number of contacts performed before this campaign and for this client (numeric), outcome of the previous marketing campaign (categorical).

- ❖ Output variable: whether the client subscribed to a term deposit (binary).

1.2.2. Data Types:

Age: Numeric (Integer)

Job: Categorical (String)

Marital Status: Categorical (String)

Education: Categorical (String)

Default: Binary (String: "yes"/"no")

Balance: Numeric (Float)

Housing: Binary (String: "yes"/"no")

Loan: Binary (String: "yes"/"no")

Contact: Categorical (String)

Day: Numeric (Integer)

Month: Categorical (String)

Duration: Numeric (Float)

Campaign: Numeric (Integer)

Pdays: Numeric (Integer)

Previous: Numeric (Integer)

Poutcome: Categorical (String)

Y (Output): Binary (String: "yes"/"no")

1.2.3. Data Entry/Capture Methods:

- Automated Data Capture: Utilize automated systems to capture data from forms, CRM systems, or other digital sources. This ensures accuracy and efficiency in data entry.

- Manual Data Entry: In cases where automated capture is not possible, trained personnel should manually enter data into the system. Double-checking and validation protocols should be in place to minimize errors.

- Real-time Updates: Ensure that data is updated in real-time or at regular intervals to reflect the latest customer interactions and campaign outcomes.

1.2.4. Reports Needed:

1. Which contact method has highest number of outcomes?
2. Which age group has taken highest number of housing loans?
3. Which age group has the highest number of personal loans?
4. Which educational backgrounds have the highest average bank balance?
5. Which age group has the more credit default percentage?
6. Which job-category is more present in the dataset?
7. Does the campaign has more success or failure?
8. Which loan holders are in credit default?
9. Does Average and total count of educational backgrounds are same?
10. Which campaign has more Success(previous or present)?

1.3. Project Risks:

Data Privacy and Security Risks:

Risk: Unauthorized access or data breaches could compromise sensitive customer information, leading to legal and reputational consequences.

Mitigation: Implement robust data encryption, access controls, and regular security audits to safeguard customer data. Adhere to data protection regulations such as GDPR or CCPA.

Data Quality Risks:

Risk: Inaccurate or incomplete data could lead to biased predictions and unreliable insights, impacting the effectiveness of marketing campaigns.

Mitigation: Establish data validation and cleansing processes to ensure data accuracy and completeness. Regularly monitor data quality metrics and address any issues promptly.

Resource Constraints:

Risk: Insufficient resources, such as skilled personnel, technology infrastructure, or budgetary constraints, may hinder project execution and effectiveness.

Mitigation: Allocate adequate resources and invest in training for personnel involved in data analysis, marketing, and compliance. Prioritize key initiatives based on available resources and strategic goals.

Technology Risks:

Risk: Technical failures, system downtime, or compatibility issues with data management tools and platforms could disrupt project operations.

Mitigation: Implement robust IT infrastructure, backup systems, and contingency plans to mitigate technology-related risks. Regularly update and maintain software systems to ensure reliability and performance.

By identifying and addressing these risks proactively, the project can enhance its resilience, optimize outcomes, and achieve its objectives effectively.

2. Design Specifications:

2.1. Relational concepts:

Customer Entity: Contains information about customers.

Attributes: age, job, marital, education, default, balance, housing, loan

Contact Entity: Information about marketing contacts with customers.

Attributes: contact, day, month, duration, campaign, pdays, previous

Outcome Entity: The outcome of marketing contacts.

Attributes: poutcome, y

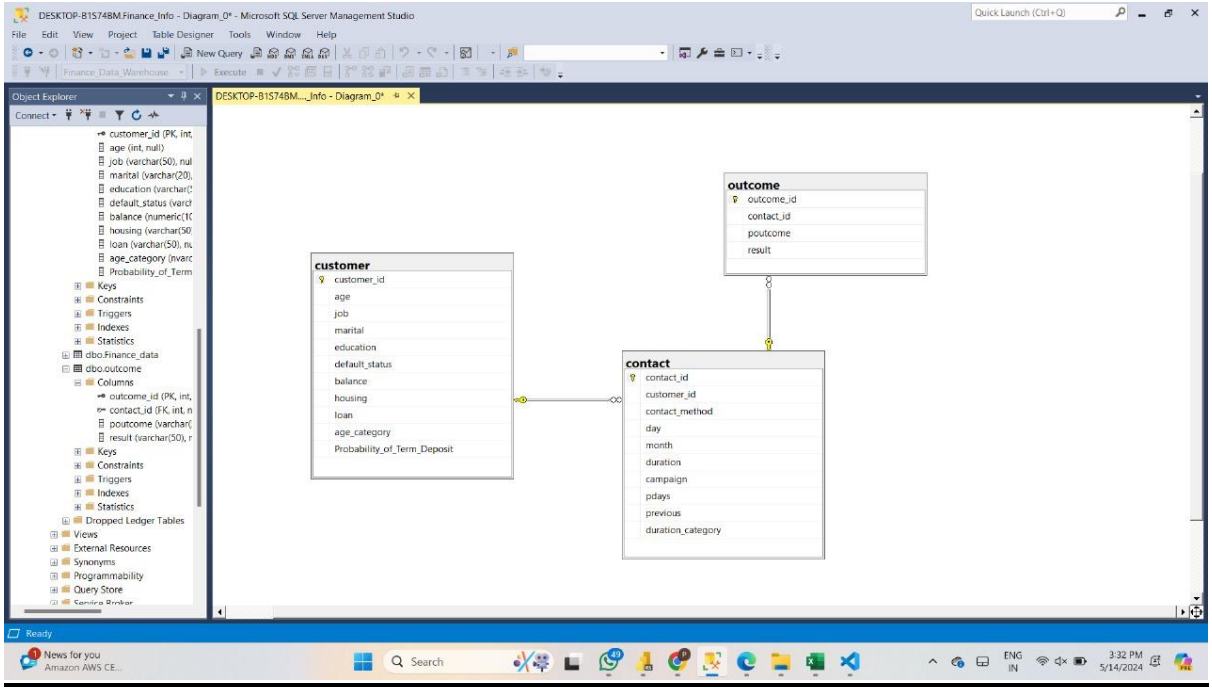
Relationships:

Customer - Contact: A customer can have multiple contacts, but each contact is associated with one customer (One-to-Many relationship).

Contact - Outcome: Each contact can lead to an outcome (One-to-One or One-to-Many relationship).

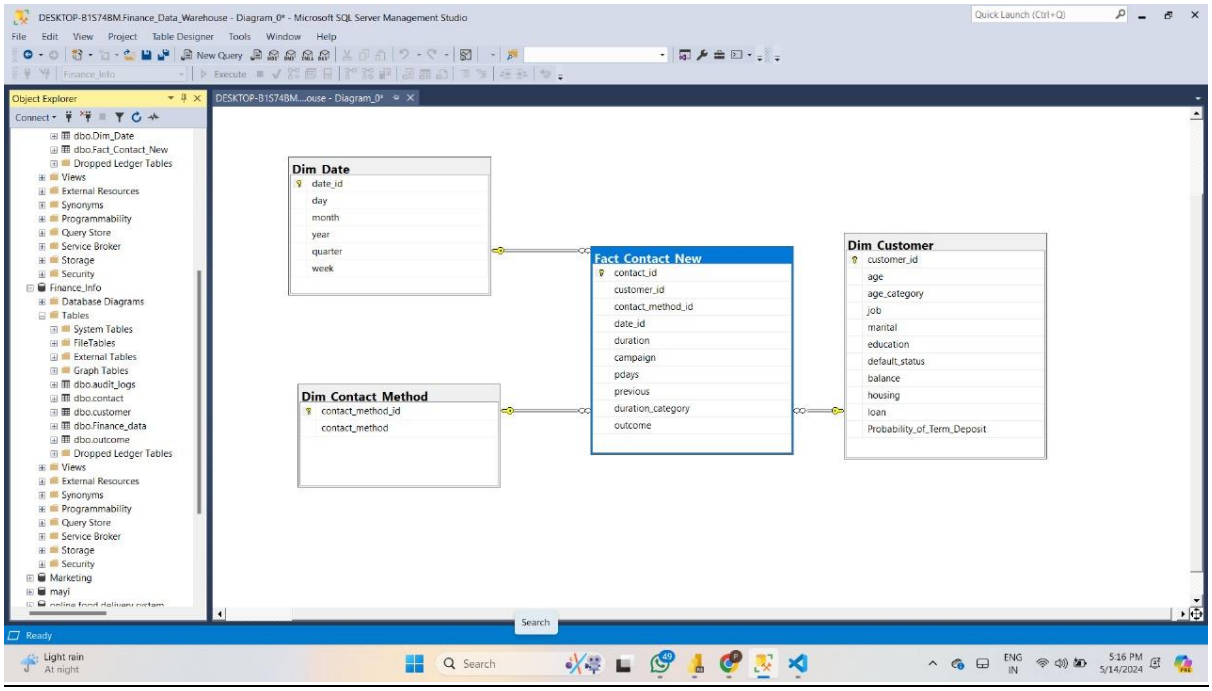
2.1.1. Technology to be used and ER Diagram:

Database Management System (DBMS): we used a relational DBMS, SQL Server management studio to create an ER diagram.



2.2. Big data concepts:

2.2.1. Data-Warehouse Structure:

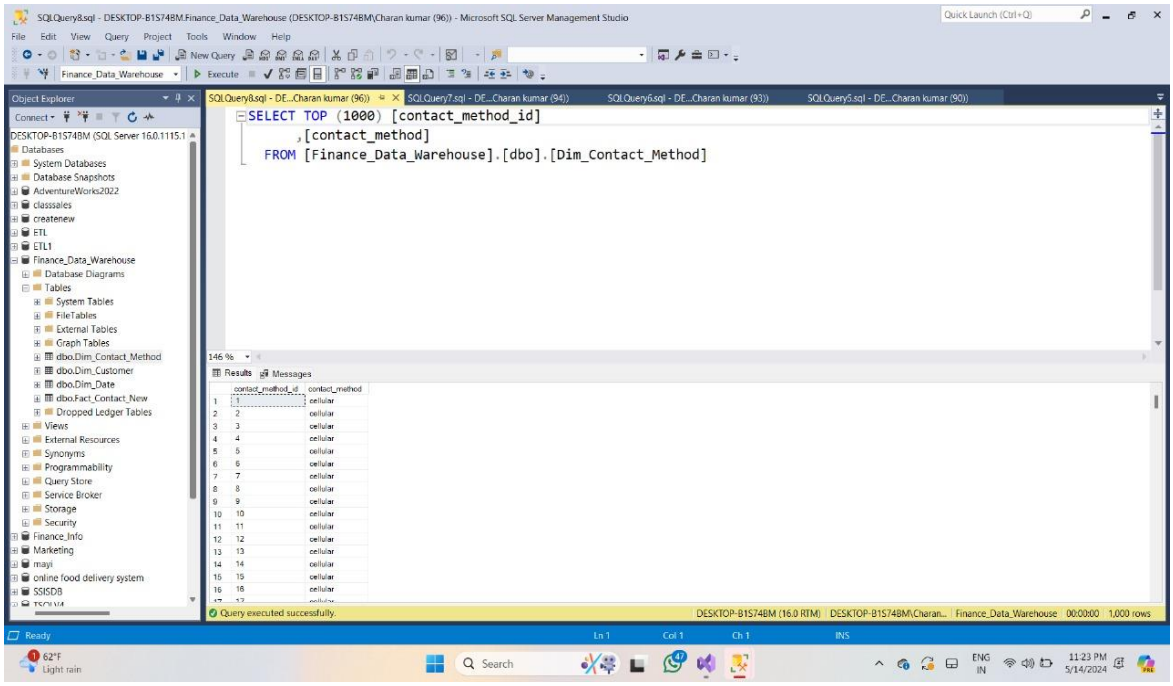


There are 3 dimensions tables and 1 Fact table.

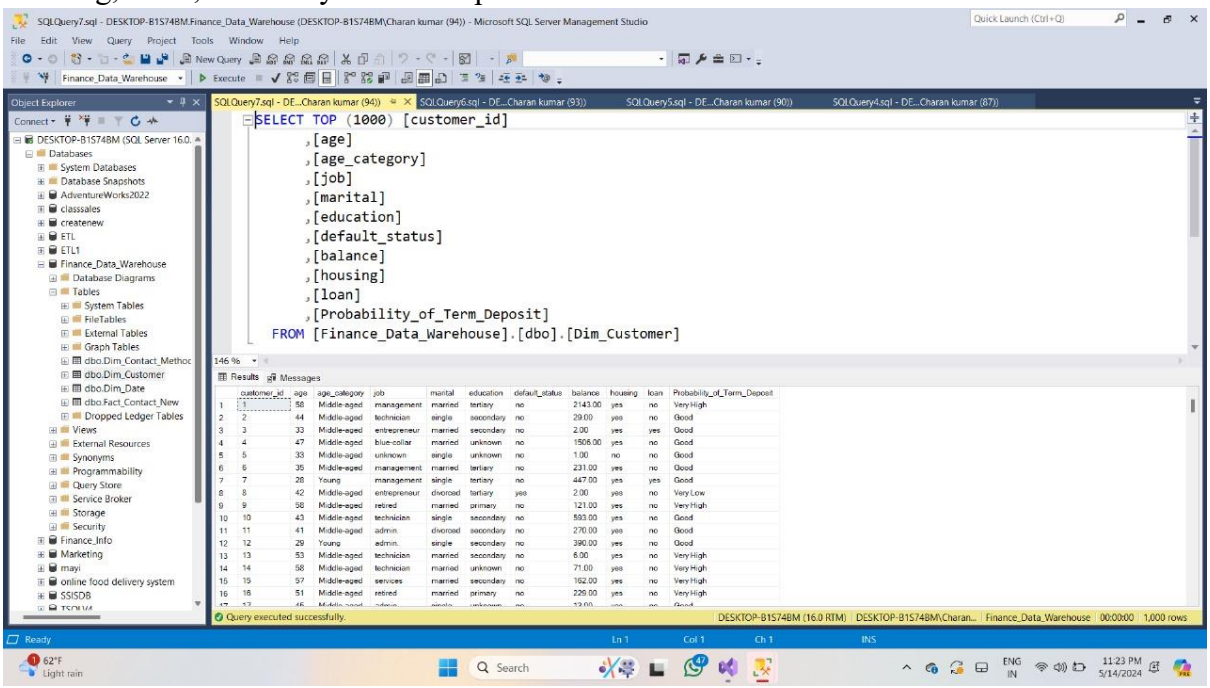
Dimension Tables:

1. Dim_Date:
   Primary Key: Date_id
   Other attributes: Day, Month, year, quarter, Week

2. Dim_Contact_method:
   Primary Key: Contact_method_id
   Other attributes: Contact_method



3. Dim_Customer:
   Primary Key: Customer_id
   Other atttributes: age, age_category, job, martial, education, Default Status, Balance, Housing, Loan, Proability of term deposit



Fact table:

Comntact_id, Customer_id, Contact_method_Id, date_id, duration, campaign, pdays, previous, Duration_Category, Outcome.

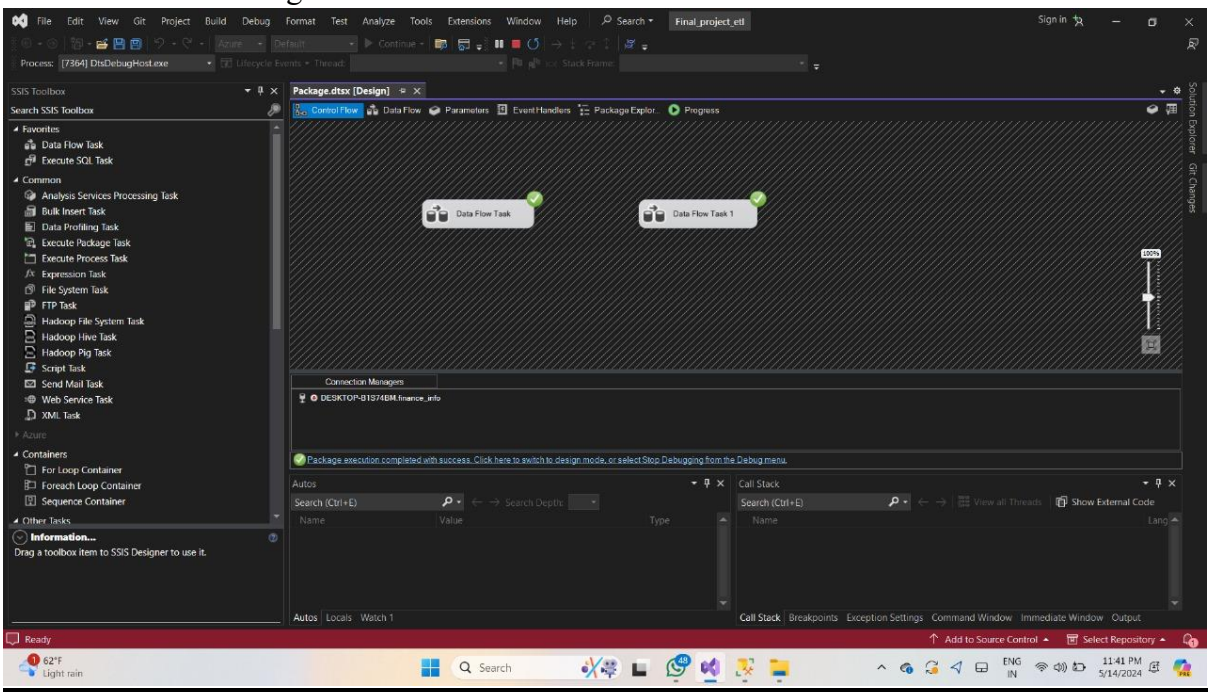
2.2.2 ETL Development:

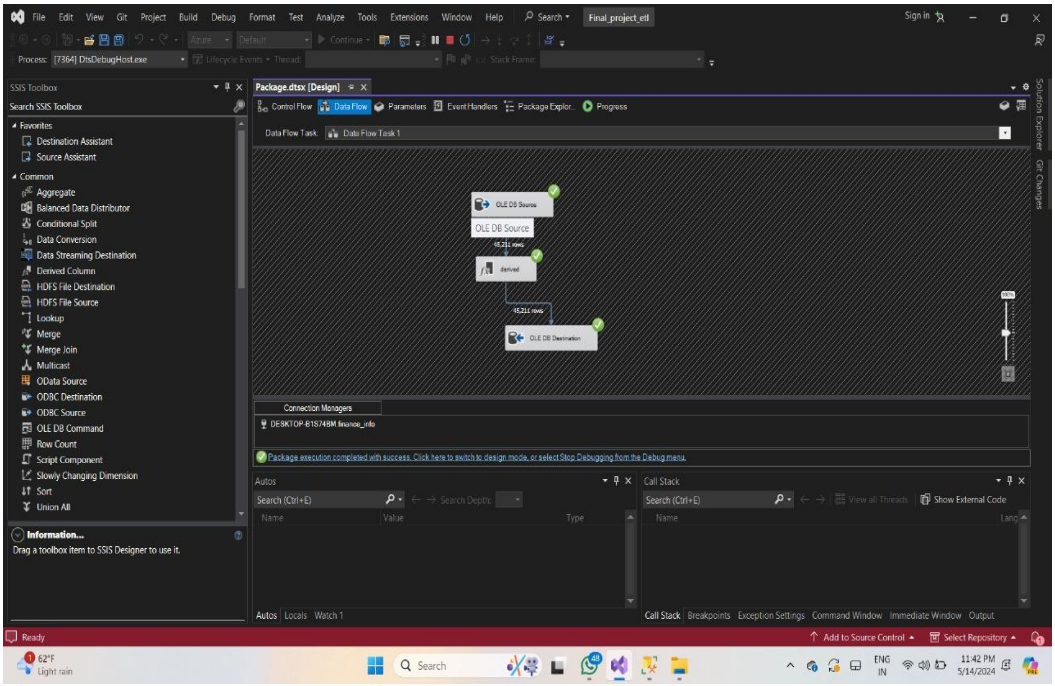
In ETL, we have derived three columns from already existing columns.

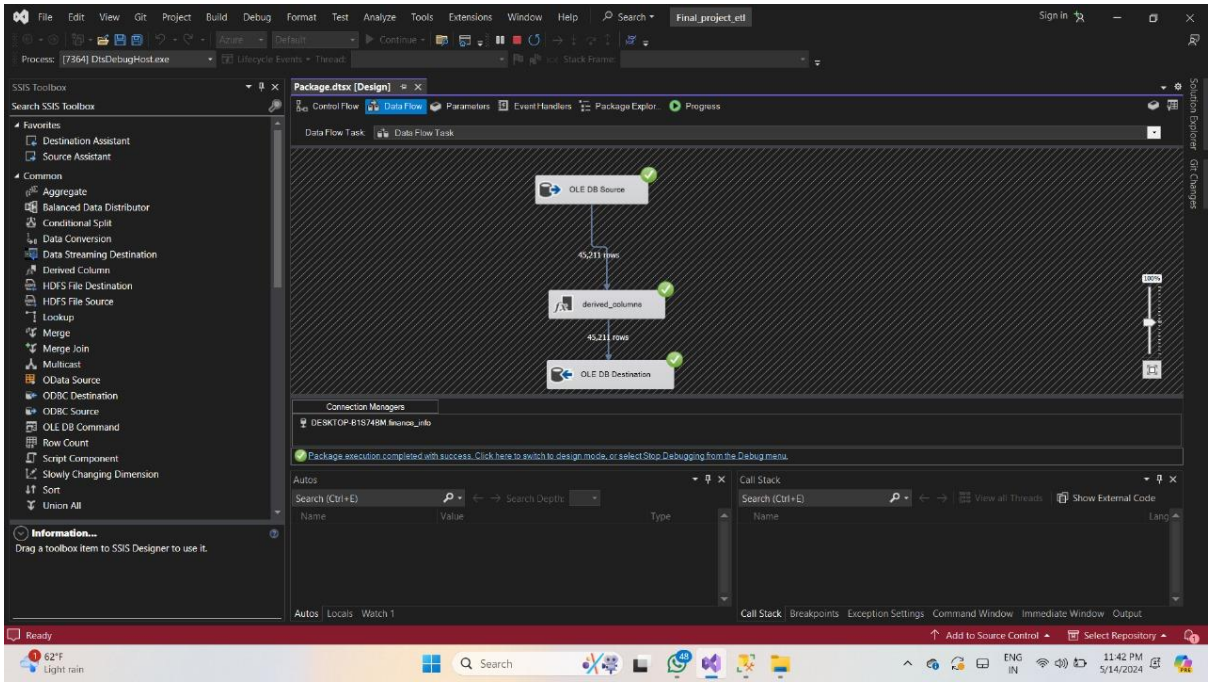
i) Age Category:
From age column, age category based on following criteria:
Below 30: young Age
Above 30 and below 60: Middle-Age
Above 60: Elderly

ii) Probability of Term Deposit:
We have calculated this value based on age and credit default.

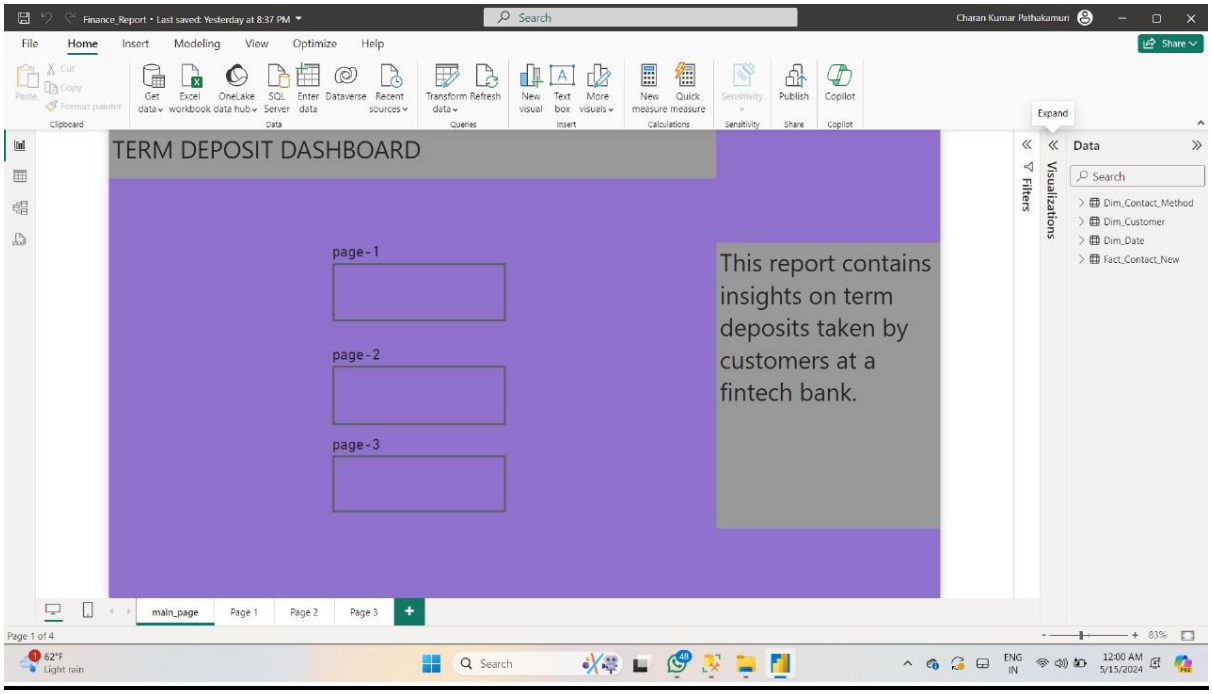
iii) Duration Category:
This has been categorized based on duration column.
If the call duration is less than 180: small.
If the call higher than 180 and lesser than 360: Medium
Greater than 360: Long.

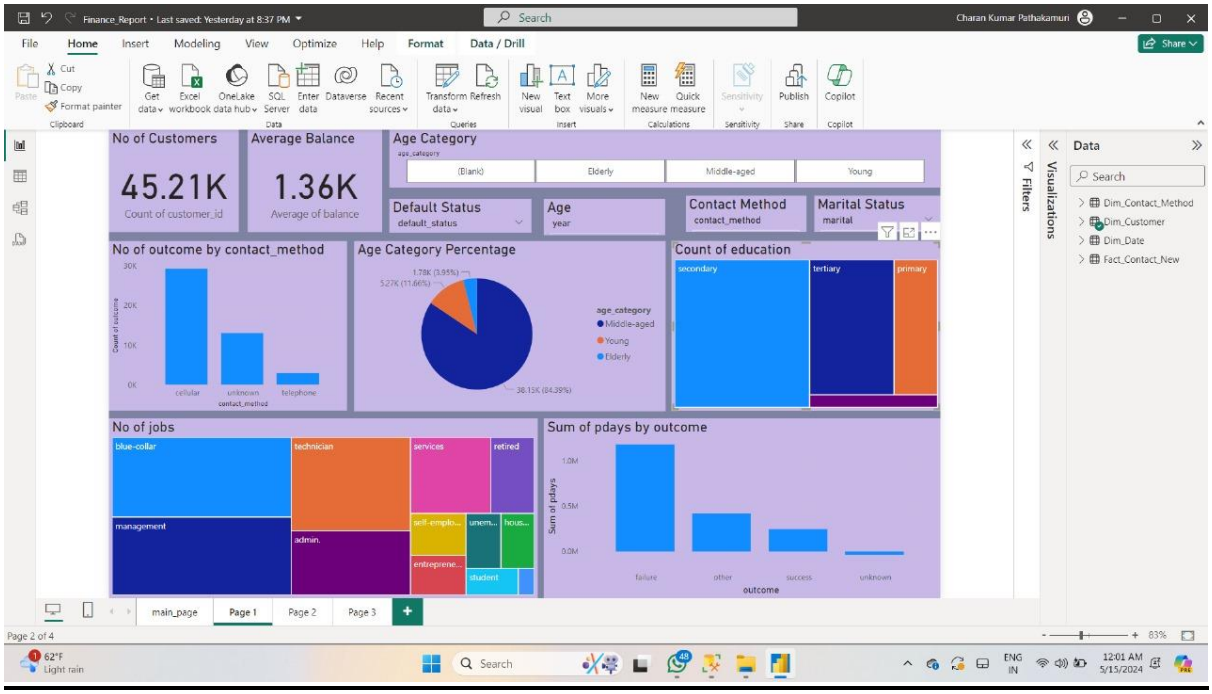


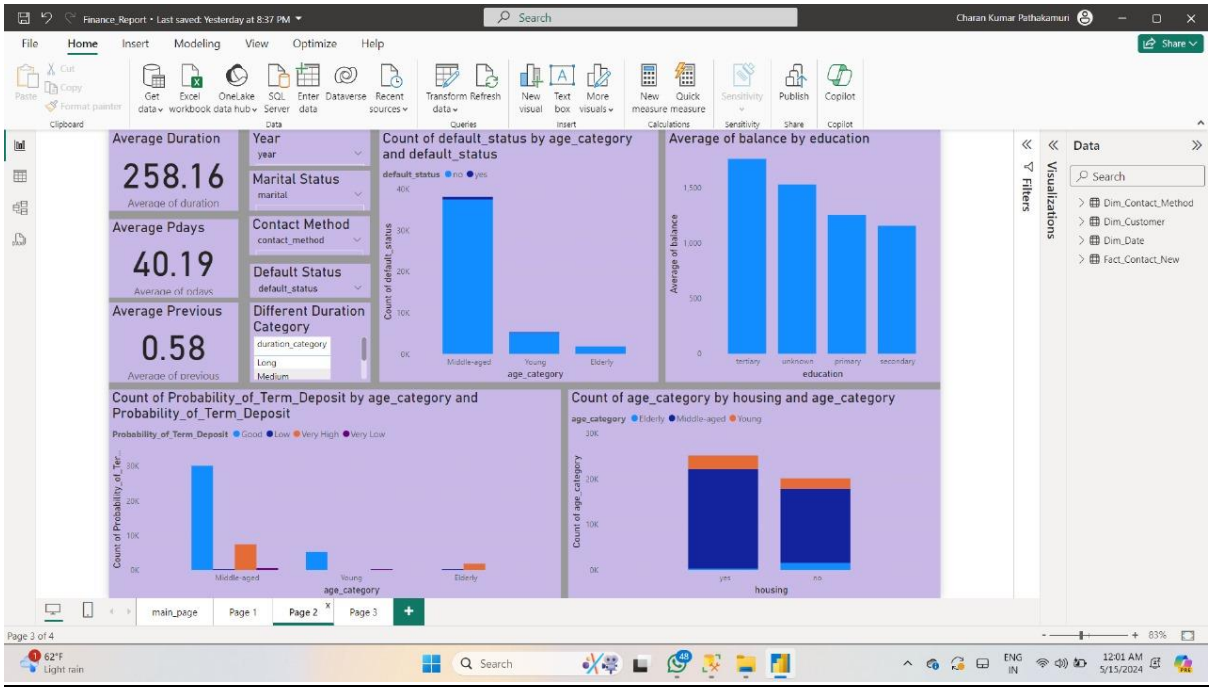

File Edit View Git Project Build Debug Format Test Analyze Tools Extensions Window Help
Final_project_etl
Sign in
SSIS Toolbox
Search SSIS Toolbox
Favorites
Destination Assistant
Source Assistant
Common
Aggregate
Balanced Data Distributor
Conditional Split
Data Conversion
Data Streaming Destination
Derived Column
HDFS File Destination
HDFS File Source
Lookup
Merge
Merge Join
Multicast
OData Source
ODBC Destination
ODBC Source
OLE DB Command
Row Count
Script Component
Slowly Changing Dimension
Sort
Union All
Information...
Drag a toolbox item to SSIS Designer to use it.
Package.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explor...
Progress
Data Flow Task
Data Flow Task 1
OLE DB Source
45,211 rows
derived
45,211 rows
OLE DB Destination
Connection Managers
DESKTOP-B1S748M.finance_info
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.
Autos
Search (Ctrl+E)
Name
Value
Type
Call Stack
Search (Ctrl+E)
Show External Code
Name
Autos Locals Watch 1
Call Stack Breakpoints Exception Settings Command Window Immediate Window Output
Ready
Add to Source Control
Select Repository
62°F
Light rain
Search
ENG
IN
11:42 PM
5/14/2024

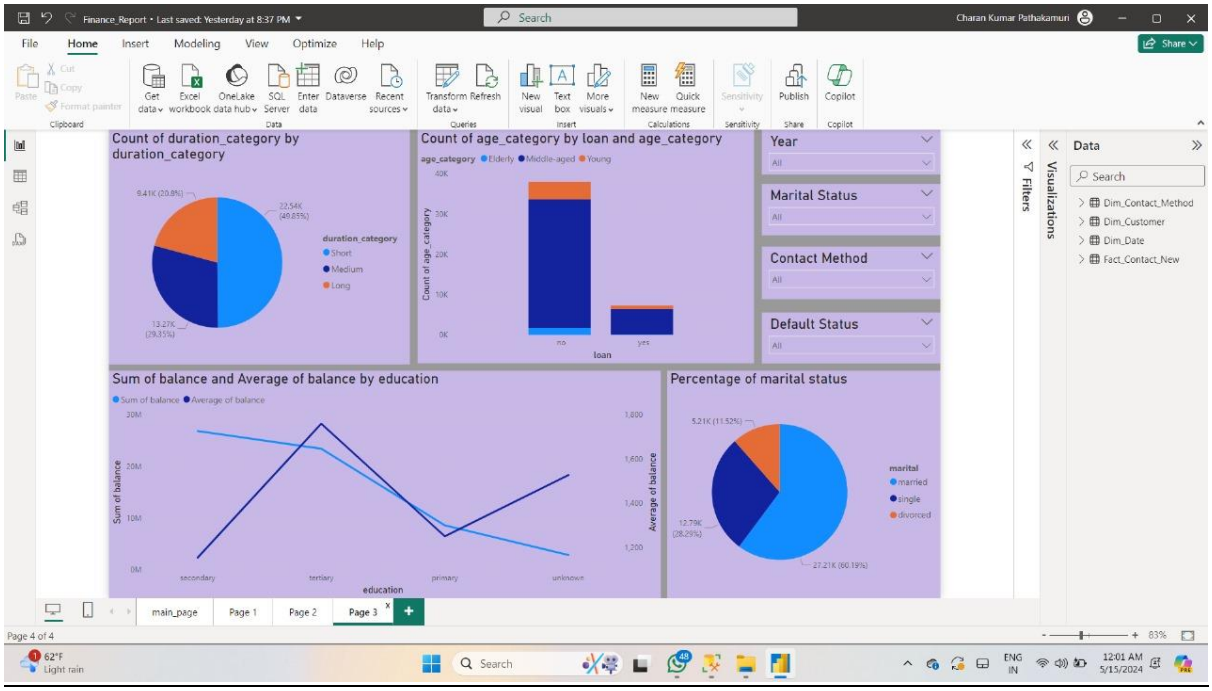

File Edit View Git Project Build Debug Format Test Analyze Tools Extensions Window Help
Final_project_etl
Sign in
SSIS Toolbox
Search SSIS Toolbox
Favorites
Destination Assistant
Source Assistant
Common
Aggregate
Balanced Data Distributor
Conditional Split
Data Conversion
Data Streaming Destination
Derived Column
HDFS File Destination
HDFS File Source
Lookup
Merge
Merge Join
Multicast
OData Source
ODBC Destination
ODBC Source
OLE DB Command
Row Count
Script Component
Slowly Changing Dimension
Sort
Union All
Information...
Drag a toolbox item to SSIS Designer to use it.
Package.dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explor...
Progress
Data Flow Task
Data Flow Task
OLE DB Source
45,211 rows
derived_columns
45,211 rows
OLE DB Destination
Connection Managers
DESKTOP-B1S748M.finance_info
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.
Autos
Search (Ctrl+E)
Name
Value
Type
Call Stack
Search (Ctrl+E)
Show External Code
Name
Autos Locals Watch 1
Call Stack Breakpoints Exception Settings Command Window Immediate Window Output
Ready
Add to Source Control
Select Repository
62°F
Light rain
Search
ENG
IN
11:42 PM
5/14/2024

2.2.3 Power BI:


TERM DEPOSIT DASHBOARD
page-1
page-2
page-3
This report contains insights on term deposits taken by customers at a fintech bank.
main_page
Page 1
Page 2
Page 3


No of Customers
45.21K
Count of customer_id
Average Balance
1.36K
Average of balance
Age Category
Default Status
Age
Contact Method
Marital Status
No of outcome by contact_method
Age Category Percentage
Count of education
No of jobs
Sum of pdays by outcome
main_page
Page 1
Page 2
Page 3





2.3 Cost Estimations:

2.3.1 Software Costs:

Microsoft SQL Server:

a) Licensing: SQL Server enterprise edition will cost $15,123

b) ODBC Driver: Which is free

Python Libraries:

a) Pandas: Free and Open source

b) SQL Alcheny: free and open source.

c) pyodbc: free and open-source.

d) Openpyxl: free and open source.

Visual Studio 2022: Enterprise Edition : $1,700

SSIS Package: $3,400

Microsoft Fabric Power BI : $5000

Visual Studio Code : Free and Open Source

Total software Costs: $25,223

2.3.2 Hardware Costs:

On-premises:

Server Chassis: $1,000

CPU : $2000

Ram: 200

Storage: $200

Network Interface: $500

Cooling Systems: $250

RAID Controllers: $200

Backup Solutions: $500

Total: 4,850

2.3.3 Staff Resources:

Developer: 10,000

DBA: 15,000

Data Engineer: $5000

Power BI developer: $3000

Total: $33,000

Miscellaneous:

Training and Documentation: $1000

Total cost: $63,850

2.4 Technologies Used:

- Python
- Pandas.
- SQL Alchemy Engine
- Openpyxl
- SQL Server
- Pyodbc for data Migration
- SSIS used for ETL.
- Power Bi
- Visual Studio 2022

2.5.Teamates Contributions:

Lokesh Reddy: Data Base Designing, Modelling and normalizing

Upendra: ETL, Data Quality

Charan: Power BI, Data Exploring

Sashank: Data Cleaning, Loading