CRM Data Management and Analytics Project Documentation

1. Project Objectives

- Normalize CRM data from raw CSV files to a well-structured relational database.
- Build a data warehouse (DW) with dimensional and fact tables to support advanced analytics.
- Create an efficient ETL pipeline for transforming raw data into meaningful insights.
- Provide interactive reports and dashboards using Power BI for business decision-making.
- Develop a machine learning (ML) model to predict customer behavior with high accuracy.

2. Architecture Overview

Our data architecture followed a multi-layered approach with three key components:

1. **Bronze Layer**: Raw data ingestion from CSVs to Azure Data Lake.
2. **Silver Layer**: Data cleansing and refinement using PySpark in Azure Synapse.
3. **Gold Layer**: Creation of dimensional (Dims) and fact tables for the data warehouse.

Additional components:

- Azure Database for hosting the normalized relational database.
- Azure Synapse Analytics for data processing and transformation.
- Power BI for visualization and decision-making insights.
- Machine Learning Model for predictive analytics.

3. Data Processing Workflow

1. **Data Normalization**:
    - Applied normalization based on the normal forms (1NF, 2NF, 3NF) to remove redundancy.
    - Uploaded the structured data to Azure Database.

2. **Data Warehouse Design**:
    - Built dimensional and fact tables using star schema.

4. ETL Process

1. **Bronze Layer (Raw Data Ingestion)**:
    - Azure Data Factory pipeline ingested data into Azure Data Lake as Parquet files.

2. **Silver Layer (Data Cleansing)**:
    - PySpark notebooks cleaned and transformed the data in Azure Synapse.

3. **Gold Layer (Data Warehouse)**:
    - Dimensional and fact tables were created, and a Lake Database was built for querying.

5. Power BI for Data Visualization

We used Power BI to create dashboards for KPIs such as customer trends, retention rates, and forecasts.

6. Machine Learning Model

- **Objective**: Predict customer churn and purchases.
- **Achieved Accuracy**: ~99%.
- **Tools**: Python, scikit-learn.

7. Technologies Used

- Azure Database
- Azure Data Factory
- Azure Synapse Analytics
- Azure Data Lake
- Power BI
- Python, PySpark, scikit-learn

8. Challenges and Learnings

- Normalization of CRM data was complex due to redundancy.
- Synchronizing Data Factory and Synapse pipelines required careful orchestration.
- Writing optimized Parquet files was crucial for performance.

9. Future Improvements

- Automate pipelines for real-time ingestion.
- Explore advanced ML models.
- Implement data governance practices.

Conclusion

This CRM project showcases the end-to-end transformation of raw customer data into actionable insights. The infrastructure built using Azure supports analytics, visualization, and predictive modeling for business value.