Business Problems Report

CloudNData
Azure Dataflow Optimization

1. Business Problem Statement

The goal of this project is to address the challenges faced by CloudNData in effectively processing, analyzing, and visualizing sales data to make data-driven decisions. Currently, the company faces inefficiencies in extracting meaningful insights from large volumes of data due to the complexity of existing systems and tools.

Key issues include:

- **Manual Data Handling**: Manual processes are currently in place for transforming and preparing data, leading to errors and time delays.
- **Inefficiency in Data Transformation**: Current data pipelines are inefficient, requiring manual intervention and resulting in high ETL times.
- **Lack of Real-Time Insights**: Business decision-makers lack access to real-time analytics, affecting their ability to make informed decisions quickly.
- **High Costs in Data Analytics**: Existing data querying processes are costly and inefficient, with no clear way to optimize resources and reduce unnecessary overhead.

2. Business Objectives

To address these challenges, the project will aim to:

- Build a scalable, automated, and dynamic ETL pipeline for processing AdventureWorks sales data.
- Optimize the performance and cost-efficiency of data storage and analytics operations.
- Provide business stakeholders with easy access to data and insights through visualization and reporting tools.

3. Project Overview

The project implements an **end-to-end data pipeline** on Azure Cloud to solve these business problems. This pipeline integrates several Azure services, including Data Lake, Data Factory, Databricks, Synapse Analytics, and Power BI. By adopting a **Medallion Architecture** (Bronze, Silver, and Gold layers), this solution automates the data ingestion, transformation, and analytics process.

Key features include:

- **Dynamic ETL Pipeline**: Automated and optimized for reducing processing time by 70%.
- **Medallion Architecture**: Structured storage layers (Bronze for raw data, Silver for transformed data, and Gold for serving business insights).
- **Cost Efficiency**: Serverless SQL pool integration to reduce the costs associated with large-scale data querying.
- **Advanced Analytics & Visualization**: Power BI integration for accessible reporting and visualization of sales performance.

4. Guidelines for the Project Team

1. **Data Handling & Transformation**:
    - Ensure that data quality is maintained at every stage of the pipeline. Pay close attention to ensuring no data is lost during the transformation steps.
    - Leverage the **medallion architecture** to allow easy traceability and processing of data through different stages (raw, transformed, and final).
2. **Optimization**:
    - The goal is to reduce ETL time by **70%**. This can be achieved through optimizations such as batch processing, parallelism in Databricks, and using **serverless SQL pools** for cost-effective querying.
    - Use **Parquet file format** for efficient storage and retrieval, leveraging **Snappy compression** for optimal performance.
3. **Security & Access Control**:
    - Implement robust security measures, such as **Azure Entra ID integration** and **Role-Based Access Control (RBAC)** to ensure that sensitive data is protected.
    - Make use of **managed identities** for authentication across Azure services, reducing the risks associated with manually managed secrets.
4. **Data Visualization**:
    - Power BI will be the tool for visualizing the final processed data. Ensure that business users have access to dynamic dashboards that provide insights into key performance metrics.
    - Provide interactive reporting capabilities so that stakeholders can drill into the data for more detailed analysis.
5. **Collaboration & Communication**:
    - Keep the project documentation up-to-date and ensure team members have access to the necessary resources.
    - Regularly update stakeholders on progress, especially when key milestones are achieved.

5. Expected Deliverables

- **End-to-End Data Pipeline**: Fully functional ETL pipeline with automated data ingestion, transformation, and analytics.

- **Optimized Data Storage and Querying**: Implementation of **serverless SQL pools** and use of **Parquet format** for storage.
- **Power BI Dashboards**: Interactive reports showcasing key business insights, including sales data trends, customer segments, and performance metrics.
- **Documentation**: Clear and concise technical documentation outlining the architecture, setup, and usage of the system.

6. Timeline and Milestones

- **Week 1-2**: Setup of Azure resources (Data Lake, Synapse Analytics, Data Factory, Databricks).
- **Week 3-4**: Development of dynamic pipeline and data transformation processes.
- **Week 5**: Implementation of Power BI integration for reporting and visualization.
- **Week 6**: Final testing and optimization for performance and cost-efficiency.
- **Week 7**: Project handover and documentation.

7. Conclusion

This project will significantly enhance the ability of CloudNData to make data-driven decisions, streamline operations, and optimize costs. By addressing the existing challenges with an automated, scalable solution, the company will be better equipped to stay competitive in the data-driven marketplace.

Document Notes

- Prepared by: Md Tarif
- Reviewed by: Abhranil Chatterjee
- Contact: mdtarif.chat@gmail.com