Optimising IungoRetail Ltd data pipeline

Link to Pipeline pdf: https://pdf.ac/2D8107

Introduction

IungoRetail Ltd is expanding its retail operations across both online and physical stores. To support business growth, the company needs a scalable, efficient, and cost-effective data pipeline. The existing data pipeline has issues with bottlenecks, data silos, scalability issues, and high costs. This report explains a data pipeline that uses Microsoft Azure services to make data ingestion, transformation, and analysis smoother with a focus on security, monitoring, and automation.

Data ingestion

The pipeline begins with two data sources:

- Online Store Data: Includes website logs, e-commerce transactions, and marketing data.
- In-Store Data: Includes POS transactions, loyalty program details, and sensor data.

Each data source is ingested into Azure Data Factory, which is used as the main data movement and orchestration tool.

From Azure Data Factory, the data is categorised:

- SQL databases or Event Hub (for structured e-commerce and POS data)
- IOT (for real-time and customer interaction data)

Event Hub and IOT data merge before entering Azure Data Factory again for further processing. Both streams are then combined and stored centrally in Azure Data Lake Storage.

Data storage

The Azure Data Lake Storage is the central storage facility, where all data from online and in-store operations is collected. This storage is constantly integrated with Azure Synapse Analytics, enabling advanced data queries and analysis.

The stored data moves into the transformation section, which processes the raw data into structured formats. This layer includes:

- Cleaning
- Aggregation
- Data Reduction
- Deduplication

Each transformation step is executed using Azure Data Factory and Azure Databricks, ensuring high-speed processing.

Azure DevOps is integrated, facilitating control and smooth updates for continuous automation and efficiency which IungoRetail ltd have had issues with.

Data analysis

Once the data is transformed, it is passed to Azure Data Factory again before being fed into the Analysis section:

- Power BI generating dashboards and interactive reports.
- Dashboards and Reports enable business intelligence insights (another need for IungoRetail ltd).

These insights are then shared with the Data Analysis team, allowing them to look at how their customers are interacting with them and seeing how to improve processes.

Security

To make sure all data is secure and monitored, I've created a Security & Monitoring section which is integrated which includes:

- Azure Security Center - Identifies security vulnerabilities in databases, storage, and applications.
- Azure Monitor - Captures metrics, logs, and traces from data pipelines and storage.

- Azure Key Vault - Ensures data security and secrets management for DataOps pipelines.

This section connects both online and in-store data sources, ensuring continuous security checks and monitoring from the moment data is ingested. This is used throughout the ETL process.

Benefits

Improved performance

- Azure Databricks provide fast data processing, eliminating ETL bottlenecks and uses Apache Sparks library.
- Processing only new data instead of reloading everything makes data updates faster and more efficient.
- Azure Synapse Analytics allows running multiple queries at the same time for large datasets.

Elimination of data silos

- Azure Data Lake Storage stores data from all sources, providing a 360-degree customer view. This is another issue that IungoRetail ltd had.
- Azure Data Factory automates data movement and combines structured and unstructured data easily.

Cost efficiency

- Storing data in the cloud lowers infrastructure expenses.
- Better ETL processes use computing power more efficiently, cutting cloud costs.
- Keeping track of system activity helps catch issues early and prevent expensive downtime.

Automation

- Azure DevOps handles deployment automatically, cutting down on manual work and mistakes.
- Azure Security Center and Monitor provide real-time tracking of system performance and security threats.

Conclusion

The new Azure-based pipeline greatly improves IungoRetail's data system by addressing issues with speed, scalability, and data integration. Using Azure Data Factory, Databricks, Synapse Analytics, and Power BI, the updated setup allows for real-time insights, smoother operations, and lower costs. Also, automated monitoring and security features help protect data and ensure compliance.

With this improved system, IungoRetail can make smarter business decisions, enhance customer experiences, and support long-term growth across both online and physical stores.