Data Ingestion Pipeline for Cloud-Based Storage: Automating Big Data with Microsoft Azure

Introduction

The industries are being transformed by way of “Cloud Computing” and “Big Data Analytics.” These two provide them with scalable platforms for processing the huge sets of data. This paper will dwell much on choosing some relevant datasets, establishing a pipeline on cloud using Microsoft Azure and automation processes towards achieving data consistency and integrity.

Dataset identification and selection

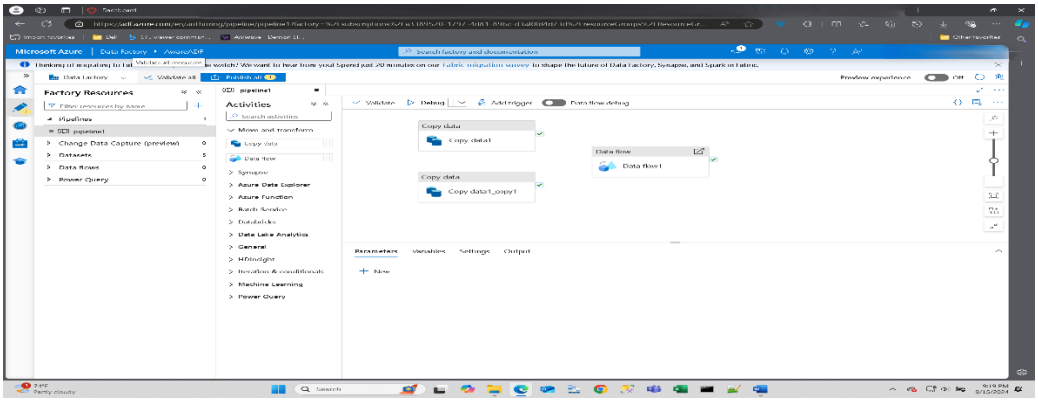
Project objective relevance: This project intends to find datasets related to the consumer electronics market (CE) relying mostly on companies like Samsung’s example. In achieving this goal Samsung employs different data sources such as customer reviews, product usage statistics among others, and supply chain logs just to mention a few to enhance its products and improve operational capacity. This scenario will be simulated using public e-commerce data sourced from sources like Kaggle giving insights into customer preferences, product performance as well as buying trends.

Format and Size Considerations: Data sets will be in cloud storage-compatible formats which include CSVs for structured data, JSON for transactions and other semi-structured data and Parquet file formats for supporting big data warehousing. Given all these, and considering Samsung operates globally, dataset sizes can be in the range of Gigabytes to Tera bytes. Microsoft Azure Blob Storage is a very good option when it comes to these large amounts of data because it is more affordable than other options in the market as well as highly scalable.

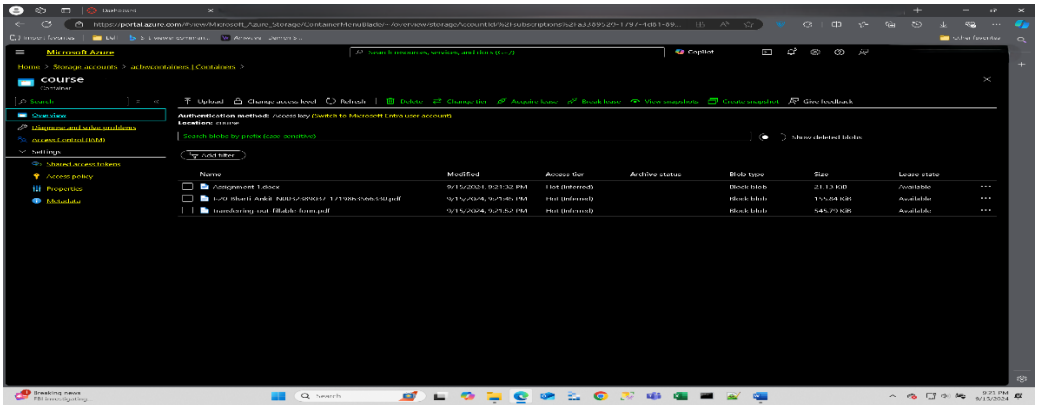
Designing the Data Ingestion Pipeline

1. Data Sources: Data is collected from APIs (for customer reviews), IoT sensors (for product metrics), or databases (for transaction logs). These sources provide real-time information on product usage and customer feedback.

2. Ingestion Tools: Azure Data Factory is a key tool for the data ingestion process. It can extract, transform, and load (ETL) data from various sources into cloud storage. Azure Data Factory also supports both real-time and batch data ingestion, making it suitable for continuous streams of data from IoT devices or daily updates from e-commerce platforms.



3. Storage Solution: Azure Blob Storage is chosen as the primary storage solution due to its scalability and support for various data formats. Blob Storage integrates easily with other Azure services, ensuring that data from multiple sources can be processed efficiently.



Using Microsoft Azure Services

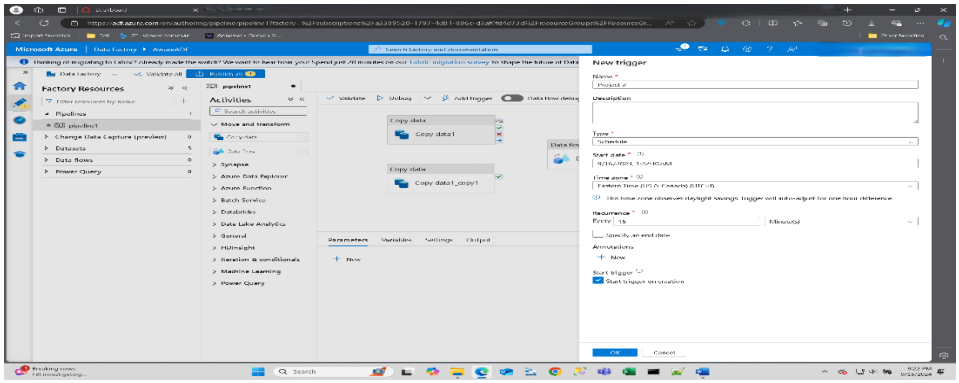
For this project, Azure Data Factory will be used to automate the data ingestion process. Azure Data Factory is ideal for hybrid data integration, allowing the ingestion of data from cloud, on-premises, and external sources into a unified platform. The data will be stored in Azure Blob Storage, which offers high scalability, security, and cost efficiency for storing large datasets. Blob Storage supports CSV, JSON, and Parquet formats, which are commonly used in analytics.

Automating the Data Ingestion Process

1. Automation with Azure Data Factory and Functions-

Automation is key to ensuring efficient data ingestion. ADF will automate the extraction of data from APIs or IoT systems and transform it into structured formats for analysis. For example, a

pipeline in Azure Data Factory can be scheduled to pull daily customer reviews and transform them into JSON or Parquet before storing them in Blob Storage.



Additionally, Azure Functions, a serverless compute service, can trigger automated tasks such as data validation or monitoring. For instance, when new data arrives in Blob Storage, Azure Functions can validate the schema and integrity of the incoming data.

2. Ensuring Data Integrity and Consistency-

It is imperative that the ingested data is valid and consistent for analysis purposes. Many times, Azure Data Factory contains built-in data validation, in which it checks if each dataset follows some set of schemas. For instance, with customer reviews, there are stipulations that require the customer to provide product id, review and rating. Schema enforcement in the pipeline eliminates the threat of data corruption of data that is yet to be processed or analyzed.

Conclusion

Choosing appropriate datasets and employing Microsoft Azure, Samsung can construct automated data ingestion systems that are efficient and expand with time. Several solutions provided by Microsoft help to collect information from different sources including ADF while data can be held in an Azure Blob Storage in a cheap and scalable manner. Employing automation in the ingestion process through Azure Functions and Azure Monitor assures the reliability of data and uniformity facilitating timely insights and improvements in operations. This project shows how useful tools of Azure can be used to develop the effective data pipeline that is needed for big data analytics.

References

Microsoft Azure. (n.d.). Data ingestion in cloud computing. Retrieved from https://azure.microsoft.com/en-us/resources/data-ingestion/

Microsoft Azure. (n.d.). Azure Data Factory. Retrieved from https://learn.microsoft.com/en-us/azure/data-factory/introduction