POC- AZURE DATA TOOLS (ETL Pipeline)

Table of Contents

Parth Salke

I. Problem Statement

Despite the growing availability of big data, effectively managing and processing data remains a challenge for many organizations. With the implementation of an ETL (Extract, Transform, Load) data pipeline utilizing Azure Data Factory, Dataflows, Databricks, and Azure Synapse Analytics, the aim is to streamline the data preprocessing process while ensuring custom validation. The challenge lies in efficiently handling diverse data sources, integrating custom validation rules, and executing seamless data transformation, ultimately enabling the organization to derive meaningful insights and make data-driven decisions. The need for a robust, scalable, and efficient data pipeline solution has become critical to enhancing data quality, reducing processing time, and improving overall data management for the organization's success and growth.

- **Key Challenges linked to manual Data Preprocessing and Validation**

1. Over several years, data management has been performed sufficiently but needs to expand to include full end-to-end data governance and data cataloging.
2. Limited Scalability: As data volumes grow, manual processes become increasingly impractical. The manual approach may not be able to scale to handle the volume and complexity of data efficiently.
3. Time Consuming: Manual data preprocessing and validation can be time-consuming, especially when dealing with large datasets.
4. Inconsistent Data Quality: Manual data preprocessing and validation may result in inconsistent data quality, leading to data discrepancies and unreliable analytical outputs.
5. Data Security Risks: Manual handling of data for preprocessing and validation can pose significant security risks, especially when dealing with sensitive or confidential information. Data security breaches, unauthorized access, or data leaks can occur.

II. Goals & Objectives

Data Governance- Using Managed Identities, Azure Active Directory (Entra ID) for access control and authentication, Azure Monitor- monitoring and alerting solutions

Architecture- Cloud-centric

Data Sources- Azure storage account (Blob- Binary format, CSV, Excel, JSON, Parquet, XML), Data Lake Gen1/ Gen2, SQL Database

Platform Capabilities- Data Quality Monitoring, Data Lineage

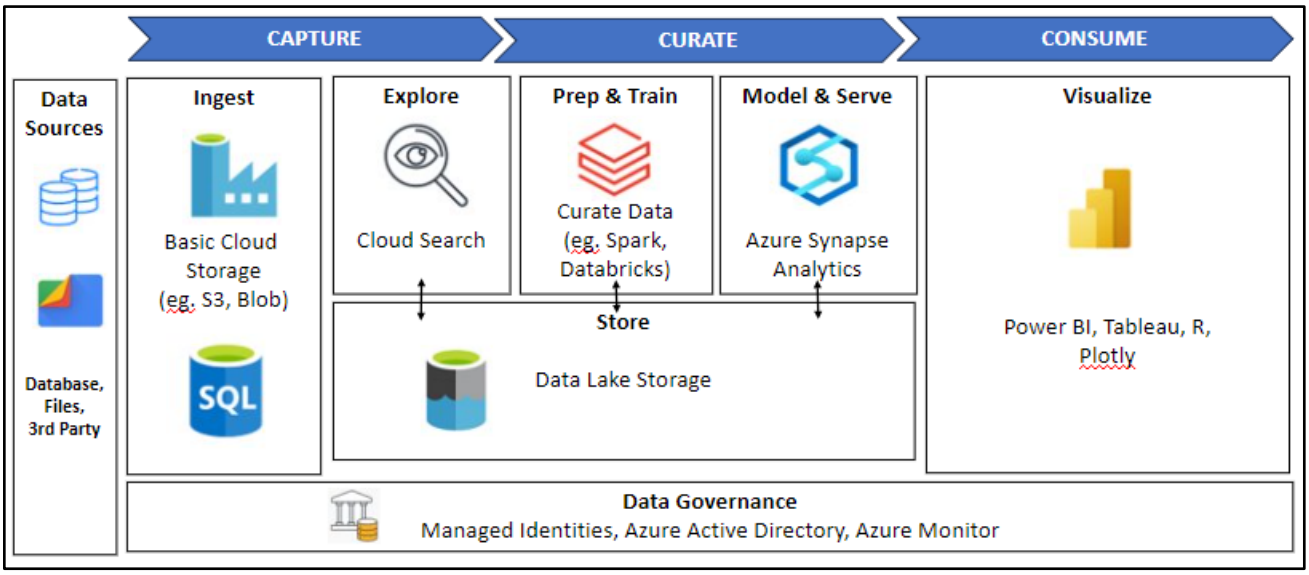
Data Formats- Structured, Semi-structured data

Platform Availability- Azure platform is available for retrieval 24x7

III. Key Assumptions

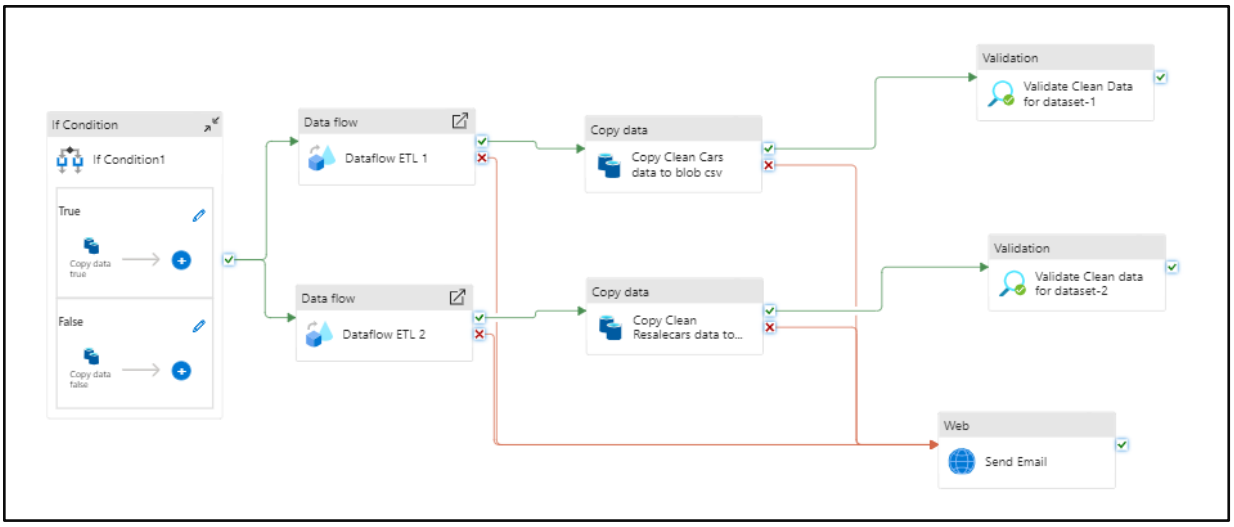
1. Data Availability: It is assumed that data will be available from the specified data sources, including Azure storage accounts (in various formats), Data Lake Gen1/Gen2, SQL databases, S3 buckets, etc.
2. Azure Services Performance: Any service degradation is expected to be addressed by Azure's support and to smoothly transition a pipeline through various environments, addressing linked service-related issues is essential.
3. Security and Compliance: The project assumes that Azure's security features, managed identities, and Azure Active Directory (Azure AD) will effectively safeguard data and ensure compliance with data security and privacy regulations.

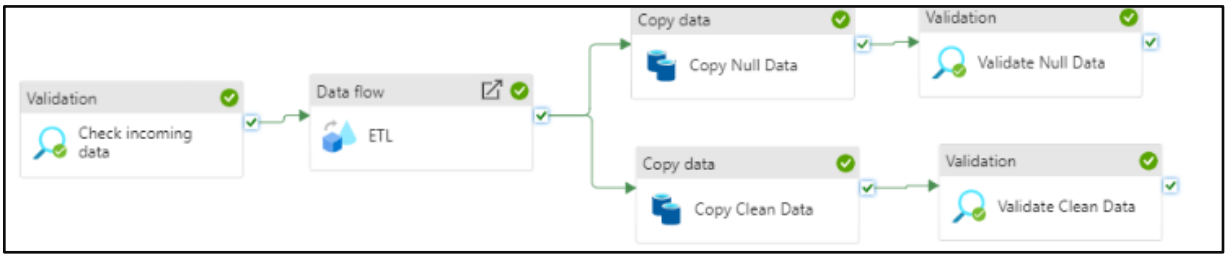
IV. Target Architecture

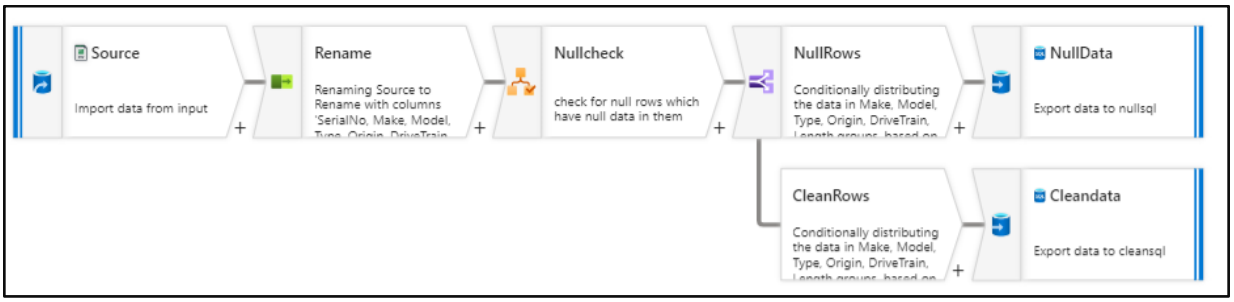


V. Architecture Diagram

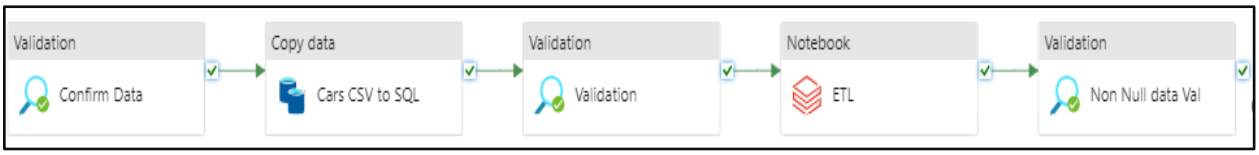
a. Using Dataflow







b. Using Databricks



VI. Target State

Parth Salke

Target Goals for end-to-end ETL Pipeline	
Diverse Data Handling	The solution will be capable of processing various data formats, including structured and semi-structured data
Automated Data Processing	The project aims to automate data preprocessing, transformation, and validation processes, reducing the reliance on manual efforts.
Comprehensive Data Governance	The organization will have a well-defined and implemented data governance framework in place, utilizing managed identities, Azure Active Directory (Azure AD), and Entra ID for access control and authentication.
Scalability	The organization will have a well-defined and implemented data governance framework in place, utilizing managed identities, Azure Active Directory (Azure AD), and Entra ID for access control and authentication.
Monitor & Alerts	The project aims to implement a fallback mechanism using Logic Apps to send custom email alerts in the event of pipeline run failures. This feature ensures timely notification and proactive issue resolution and monitor runs and triggers using Azure Monitor.

VII. Conclusion

In summary, the Azure-based end-to-end data pipeline project is poised to address the challenges of manual data preprocessing and validation and is a technical and data governance upgrade. By automating processes, enhancing data quality, and ensuring security within the Azure platform, we aim to streamline data management, reduce costs, and drive data-driven decision-making. The current ETL pipeline has reduced the processing and transformation time by 20% than the previous one.

The target state, as described, represents our vision for a more efficient and secure data ecosystem. It is our strategic response to the growing demands of managing diverse and expanding datasets, and it sets the stage for our organization's continued growth and competitiveness in the data-driven landscape.