Observability - Phase I

Client Success Story

Industry

Life Science

Client

Selection of Fortune 500

Challenge

- **Complex data platform**: The customer deployed a modern, cloud-based data platform using Azure Stack, Data Lake (ADLS), SQL for data storage, Azure Data Factory, Databricks, and Azure SQL for data processing, and PowerBI for reporting.
- **Rapid deployment by multiple teams**: Multiple data products were rapidly deployed on the platform by different cross-functional teams, resulting in a lack of unified visibility.
- **Lack of visibility**: The data platform owner struggled with a lack of visibility into platform usage and operations to ensure proper tool usage, optimal resource allocation, and increased platform adoption.

Solution

- **Comprehensive observability framework**: Architected and deployed an observability solution to monitor all major data platform components:
  - Data Processing Engines: Databricks, SQL, ADF (ETL).
  - Data Storage: Data Lake, SQL databases, file stores.
  - User activities/queries.
- **Use of open standards and API**: The solution uses open standards and API - Open Telemetry, allowing extensibility to other engines.
- **Observability dashboard**: Developed a dashboard that shows platform usage, alerts and trends, including:
  - Data object volumes and usage.
  - Processing job details with trends such as errors, processing times, and resource usage.
  - User activity trends, including query execution times.

Benefits (after first phase)

- **Improve data quality**: Monitor and detect anomalies in data pipelines to reduce errors.
- **Rapid problem resolution**: Real-time monitoring of data pipelines to quickly identify and resolve problems, minimize downtime, and ensure data availability within established SLAs.
- **Improve operational efficiency**: Optimize and streamline data operations to reduce costs and improve resource utilization.
- **Optimized resource allocation**: Better optimize resources, including personnel and infrastructure, resulting in cost savings.
- **Scalability and growth support**: Maintain performance goals by optimizing the handling of growing data volumes.

Observability - Phase II

Client Success Story

Industry

Life Science

Client

Selection of Fortune 500

Challenge

- **Complex data platform**: The customer deployed a modern, cloud-based data platform using Azure Stack, Data Lake (ADLS), SQL for data storage, Azure Data Factory, Databricks, and Azure SQL for data processing, and PowerBI for reporting.
- **Rapid deployment by multiple teams**: Multiple data products were rapidly deployed on the platform by different cross-functional teams, resulting in a lack of unified visibility.
- **Lack of visibility**: The data platform owner struggled with visibility into platform usage and operations to ensure proper tool utilization, optimal resource allocation, and increased platform adoption.
- **Data-driven business goals**: The customer wanted to develop a data-driven organization to support all key business decisions and implement KPI monitoring to set strategic direction.

Solution

- **Data catalog integration**: Implement a data catalog integrated with the observability dashboard, including:
  - Descriptions for all objects (tables, columns, pipelines, etc.)
  - Setting of the Tier 1
  - Assignment of owners to objects
  - Add tags and glossary terms
  - Defining data domains and products
  - Setting KPIs for OKRs
  - Adding DQ checks for tables and columns
- **Extended observability**: Extend observability to mission-critical data for fast and reliable decision making.

Benefits (after second phase)

- **Business risk mitigation**: Identify and resolve data issues before they impact critical business processes, reducing the risk of decisions based on inaccurate or incomplete information.
- **Increased data trust**: Ensure data reliability and integrity, building greater trust in the data.
- **Accelerated insight time**: Reduce the time to generate insights by quickly detecting and resolving data issues. Optimizing data processing prioritizes delivering the most relevant data.
- **Improved customer experience**: Reliable, high-quality data helps organizations understand customer behavior and preferences better.