Steel Industry

-Jahnavi Sharma
https://www.linkedin.com/in/jahnavi-sh/

Mission Objectives

This report outlines the methodology and technological framework for enhancing sales strategies using an integrated, scalable, and efficient cloud-based architecture. The system uses modern data ingestion, storage, processing and visualization techniques to enable actionable insights that drive decision-making and improve operational efficiency.

OBJECTIVES AND GOALS:

1. Unified Data Integration

- Integrating customer feedback, inventory data, ERP records, and IoT sensor data into a centralized repository.
- Simplifying access and reducing silos between departments for a more holistic view of business operations.

2. Scalability and Flexibility

- Designing an architecture that grows with the business. Easily accommodates with the increased data volumes and new data sources without disruptions.

3. Real-time Data Processing

- Utilizing streaming data to monitor ongoing operations, inventory levels, and customer trends.
- Enabling agile decision-making, such as adjusting sales strategies based on real-time demand patterns.

4. Data Accuracy and Consistency

- Implementing validation and verification mechanisms to ensure reliable and consistent data across all systems.

5. Automated Insights Delivery

- Reducing dependency on manual data analysis by delivering pre-processed insights and predictive analytics.

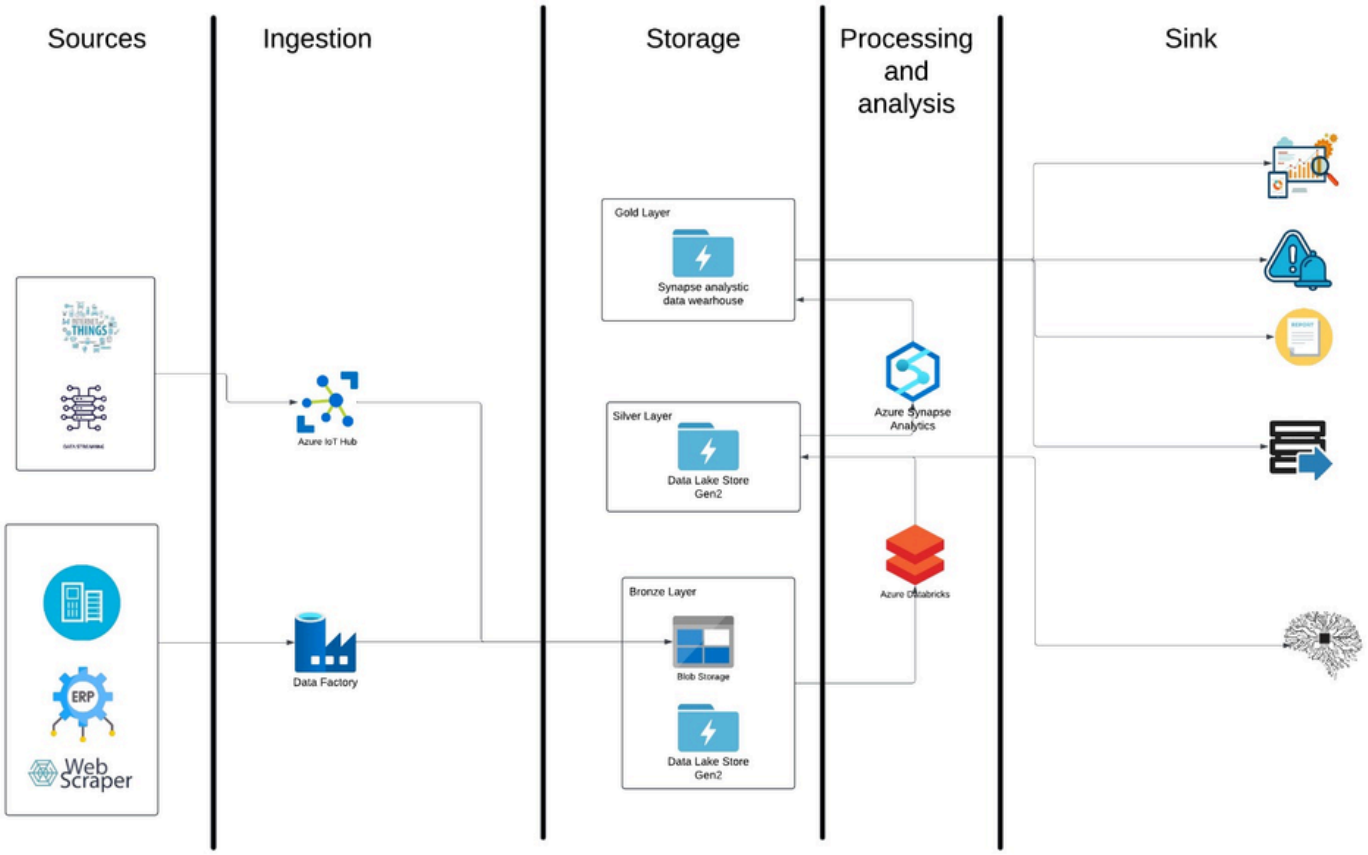
6. Interpretability

- Creating clear and user-friendly visualizations, dashboards, and reports.

7. Performance Monitoring

- Defining KPIs to assess data flow, system uptime, and operational efficiency.

Data Pipeline Overview



The architecture is designed to manage data through a streamlined pipeline, as shown in the cloud architecture diagram. The pipeline is divided into four key components: Sources, Ingestion, Storage and Processing and Analysis, with final outputs directed to Sink systems for visualizations, reporting, and integration.

SOURCES

The data is collected from various sources, including:

1. IoT devices for real-time metrics and telemetry.
2. Enterprise Resource Planning (ERP) systems for internal operations data.
3. Web scraping tools for competitive analysis and market trends.

Data Pipeline Overview

INGESTION

The ingestion layer integrates data from multiple sources using:

1. Azure IoT Hub: For real-time data streaming from IoT devices
2. Azure Data Factory: For orchestrating batch and pipeline data integration from ERP systems and web scrapers.

This ensures a seamless and consistent flow of structured and unstructured data into the storage system.

STORAGE

The storage layer follows a multi-layered architecture to enable efficient data organization and transformation:

1. Bronze layer: Raw data is stored in Azure Blob Storage and Data Lake Store Gen2. This is done to preserve its original format for traceability.
2. Silver layer: Refined and cleansed data is organized within Data Lake Store Gen2.
3. Gold layer: Finalized, analytics-ready datasets are stored in Azure Synapse Analytics Data Warehouse.

PROCESSING AND ANALYSIS

Advanced data processing and analysis are carried out using:

1. Azure Synapse Analytics: for querying, analytics and integrating data from the gold layer.
2. Azure Databricks: for big data processing, machine learning model training and exploratory analysis.

These tools work together to transform raw data into actionable insights.

SINK

The final outputs are directed to the following:

1. Interactive dashboard: Visualizations to track performance metrics and KPIs.
2. Reports: Automated periodic reports
3. Alerts: Notifications for anomalies, trends, or critical decision points.
4. Integration systems: Interface with CRM, ERP, and other downstream systems for seamless operations.

Pipeline Strategy

The pipeline strategy outlines the methodologies and practices to ensure the smooth operations of the data pipeline. It focuses on maintaining the reliability, efficiency, and scalability of the data flow through all stages of the cloud architecture.

Pipeline Strategy Framework

1. **Automation and Orchestration**

- Tools: Azure Data Factory orchestrates the batch and pipeline data integration, ensuring seamless flow between various components.
- Methodology: Automation scripts handle routine data ingestion, transformation and transfers. This helps reduce manual intervention.

2. **Modularity**

- Each pipeline component i.e. Ingestion, Storage, Processing, and Sink, is modularized. This allows for independent upgrades, maintenance, and testing without disrupting the entire system.

3. Data Validation

- Real-time Validation: Azure IoT Hub includes real-time checks for incoming IoT data to ensure completeness and accuracy of data.
- Batch Validation: Azure Data Factory includes validation rules for structured ERP data and web-scraped datasets before ingestion.

4. Scalability

- Azure Synapse Analytics and Databricks allow dynamic scaling to handle fluctuating data volume and computational requirements.

5. Monitoring and Logging

- Azure Monitor does real-time monitoring. It tracks pipeline health and generates alerts about delays, anomalies or performance bottlenecks.
- Azure Log Analytics stores logs, enabling detailed troubleshooting.

Pipeline Failure Strategy

The pipeline failure strategy defines the procedure to handle and recover from failures, ensuring minimal disruption and quick resolution.

Failure Identification

1. Ingestion Failures

- Common causes: Data source downtime, malformed data, network interruptions.
- Mitigation: Redundant ingestion paths with retry mechanisms are configured in Azure IoT Hub and Data Factory.

2. Storage Failure

- Common causes: Insufficient storage, misconfiguration in storage tiers, or file corruption.
- Mitigation: Automated scaling in Azure Blob storage and Data Lake ensures storage capacity, and backup policies safeguard against data loss.

3. Processing Failure

- Common causes: Insufficient compute resources, incompatible data formats, or errors in ETL processes.
- Mitigation: Job retries and resources scaling in Azure Databricks and Synapse Analytics. Alerts trigger corrective actions automatically.

4. Sink Failure

- Common cause: Downtime in dashboard systems, reporting tools, or integration endpoints.
- Mitigation: Integration systems have backup servers to ensure continued operation.

RECOVERY FRAMEWORK

1. Automatics failover: Azure provides automatics failover to secondary regions in case of regional outages, ensuring high availability.
2. Retry logic: ADF retries failed ingestion jobs three times before escalating the issue to the monitoring team.
3. Alerting mechanism: Azure alerts notify immediately upon detecting pipeline failure, providing detailed logs for rapid diagnosis.
4. Rollback mechanisms: In the event of critical errors, pipelines can revert to the last stable state.
5. Disaster recovery: It includes daily backups of all data layers and metadata, stored in Azure Backup and Recovery Vault.

Conclusion

This initiative establishes a scalable, robust data architecture, integrating diverse sources and enabling real-time processing. It ensures data accuracy, automated insights delivery, and seamless interoperability. By empowering decision-making through dynamic dashboards, reports, and alerts, the system enhances operational excellence, fosters collaboration, and supports continuous improvement across critical business functions