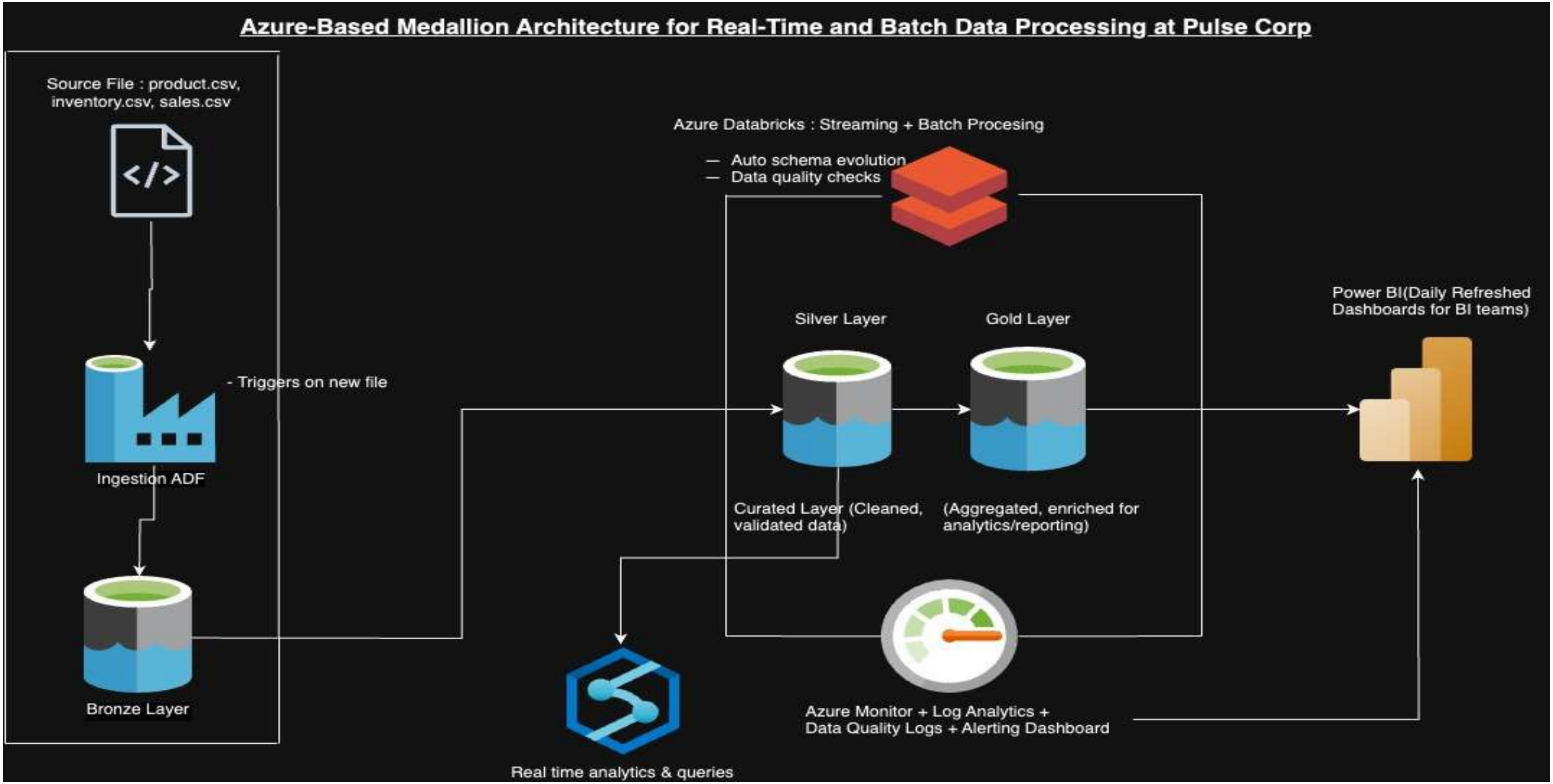
Pulse Corp -
Data Engineer Case Study

Nitesh Ranjan Singh
Data Engineer
+91 9113129159
niteshranjansingh85389@gmail.com

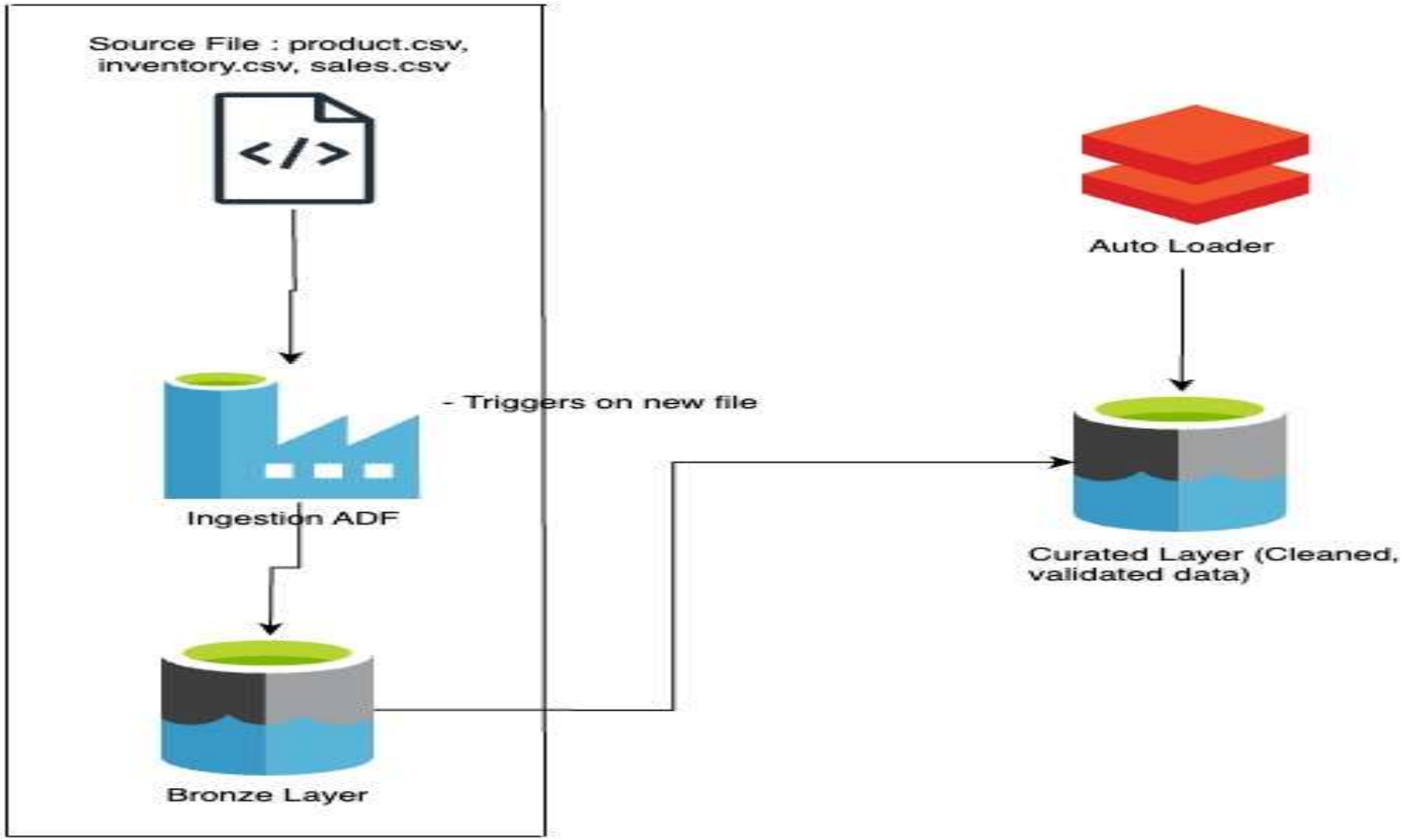
Azure-Based Medallion Architecture for Real-Time and Batch Data Processing at Pulse Corp



How the Solution Supports Pulse Corp’s Analytical and Reporting Needs ?

- **Real-Time Analytics**
  → Streaming via **Azure Databricks** provides up-to-date access for operational dashboards and ad-hoc insights.
- **Daily Business Reporting**
  → Curated data in **Gold Layer** powers **Power BI** dashboards refreshed daily—meeting reporting SLAs.
- **High Data Quality**
  → **Silver Layer** ensures cleaned, validated data. Built-in quality checks increase trust in insights.
- **Zero Manual Intervention**
  → Auto-triggered pipelines + **Azure Monitor** ensure 0-click operation with full visibility and alerts.
- **Schema Flexibility**
  → **Auto schema evolution** handles source file changes without breaking the pipeline or needing
- **Scalable & Modular Design**
  → Built on **Medallion Architecture**, the system is scalable, modular, and ready for future data expansion

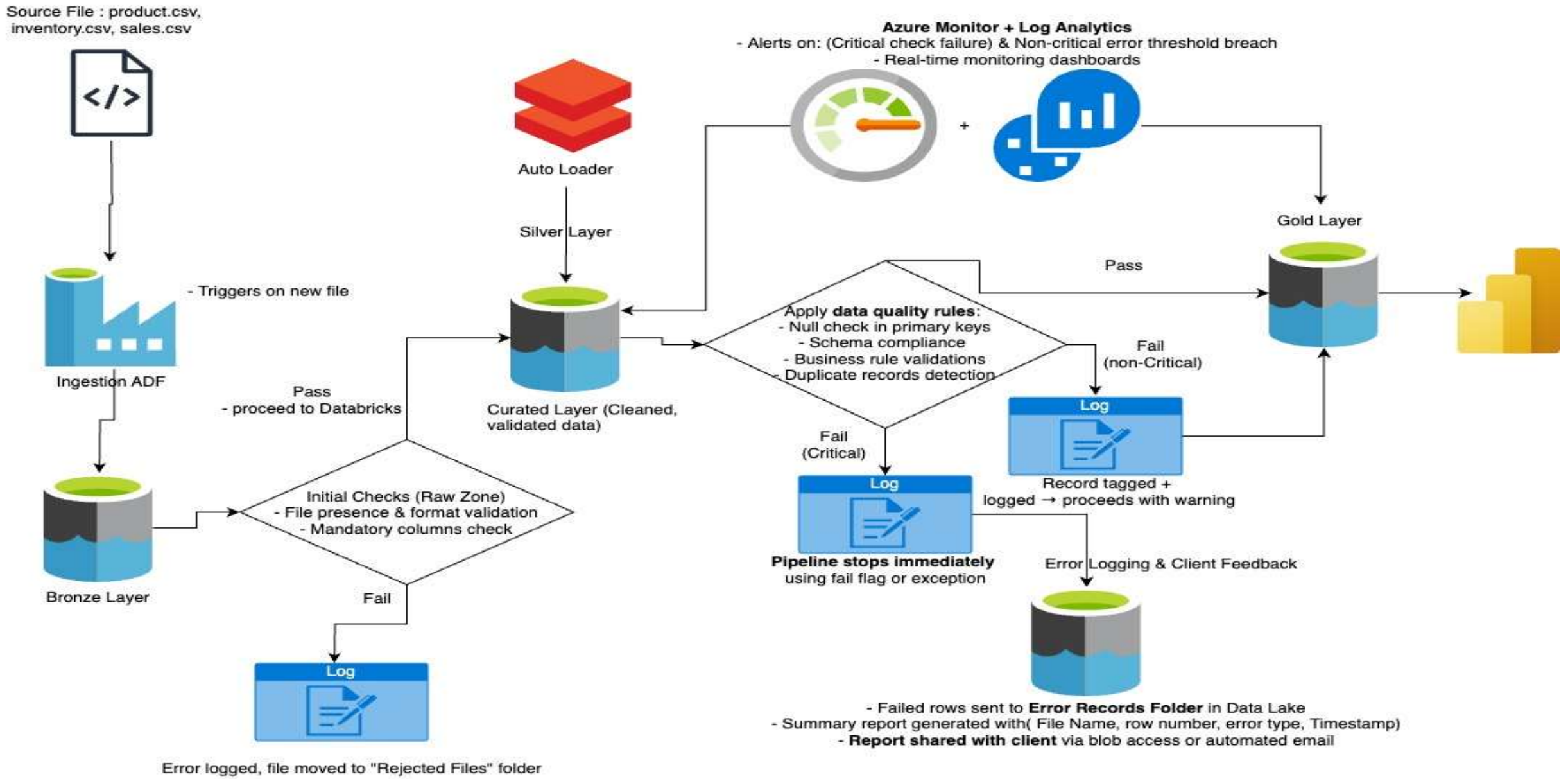
Data Ingestion Framework



Data Ingestion Framework – Schema Flexible Design

- **Automated File Ingestion via ADF**
  → Azure Data Factory (ADF) detects and ingests new files (product.csv, inventory.csv, sales.csv) into the **Bronze Layer** (raw zone).
- **Schema-On-Read Raw Storage**
  → Files are stored in their original structure, supporting layout changes without breaking pipelines
- **Databricks Auto Loader for Evolved Schemas**
  → Auto Loader reads from the Bronze Layer, supports **auto schema inference** and **mergeSchema=true** for layout flexibility.
- **Dynamic Schema Mapping to Curated Layer**
  → Cleaned and validated data is written to the **Curated (Silver) Layer**, applying transformation logic even with evolving columns.
- **No Pipeline Modifications Needed**
  → The framework handles added columns or reordered fields **without needing upstream changes**—ensuring stability.
- **Scalable and Future-Proof**
  → Supports both batch and streaming, with optional schema registry and logging for **schema version tracking and auditing**.

Data Quality Management Framework – End-to-End Automated Validation & Error Handling (Azure-based)



Data Quality Management Framework – Automated & Scalable Design

- **DQ Checks Embedded in Silver Layer (Databricks)**
  – Automated rules validate schema, nulls, duplicates, and business logic using notebooks or DLT pipelines.
- **Critical vs. Non-Critical Checks**
  – **Critical** check failures (e.g. missing keys) immediately stop the pipeline.
  – **Non-Critical** violations are logged, and flagged rows continue for further review.
- **Real-Time Error Logging & Client Feedback**
  – Failed records are written to a structured **Error Zone** in Data Lake.
  – A summary error report is auto-generated and shared with the client.
- **Monitoring with Azure Monitor + Log Analytics**
  – Alerts trigger for DQ failures, with dashboards to track rule performance and trends.
- **Data Propagation Control**
  – Only validated records from Silver move to the **Gold Layer**, ensuring clean data for analytics and BI.
- **Fully Automated & Scalable**
  – No manual intervention needed; modular DQ design supports evolving business rules and schema changes.