**Problem Statement**

Retail companies struggle with integrating **in-store POS transactions, online orders, and inventory systems** into a single analytics-ready platform.  
They need:

* **Batch processing** for daily sales reconciliation.
* **Real-time ingestion** for fraud detection & stock alerts.
* **Cloud-native scalability** to support peak season sales.
* **Unified warehouse (Snowflake / Synapse)** for advanced reporting
* **Monitoring & error handling** for operational resilience.
* **Data security** for customer PII.
* **Cost transparency** for infrastructure & query optimization.

**Skill Towers Developed**

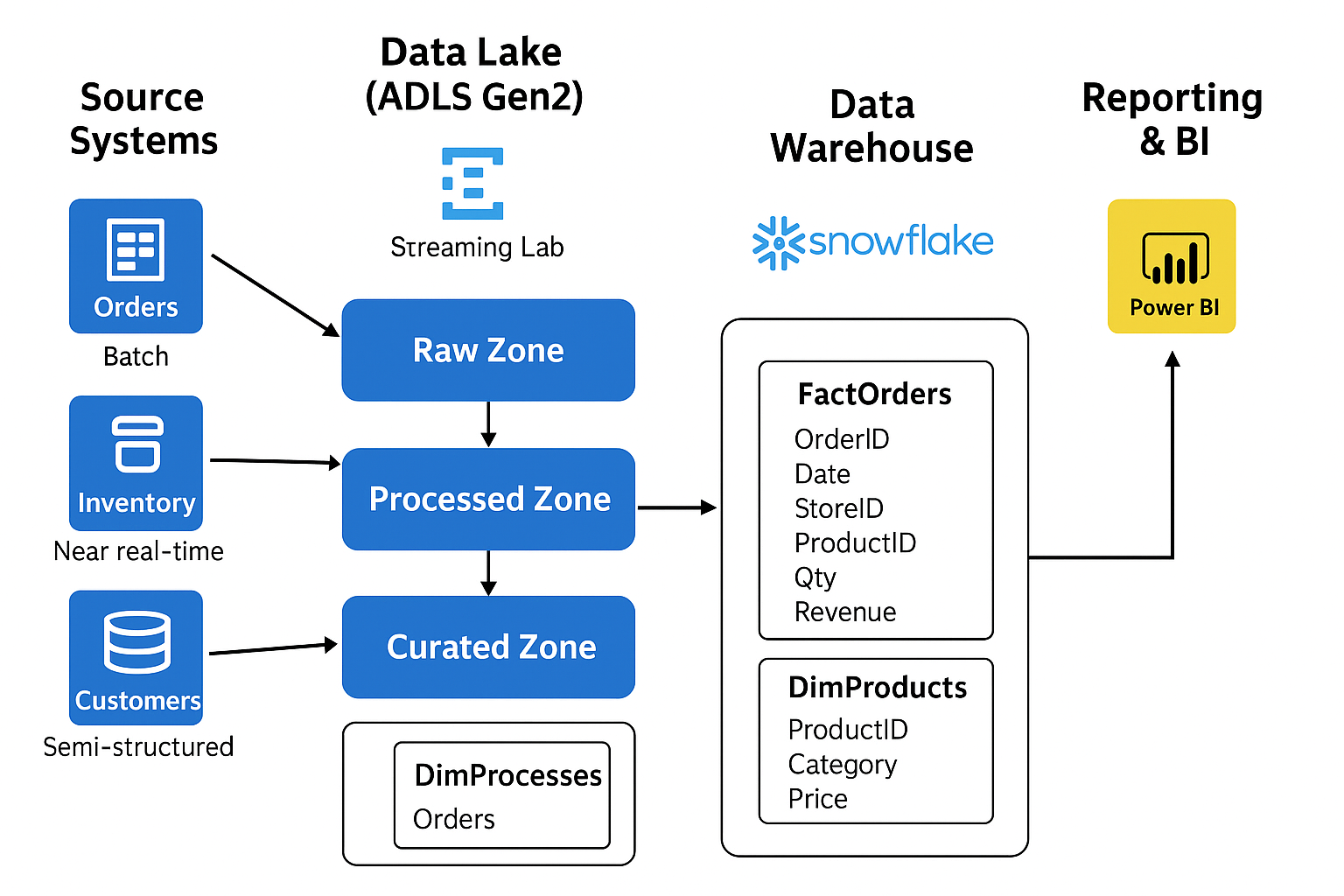
* **Data Engineering Fundamentals**: Batch & streaming design.
* **Azure Storage & ADF**: Ingestion pipelines for orders, customers, products.
* **Databricks & Delta Lake**: Cleansing, transformation, schema evolution.
* **Snowflake & Synapse**: Data warehousing & BI reporting.
* **Monitoring & Optimization**: Cost management, troubleshooting, scaling.
* **Advanced Topics**:
  + Snowpark ML for demand forecasting.
  + Synapse for interactive dashboards.
  + Azure Monitor + Log Analytics for pipeline & cluster monitoring.
  + Data security with PII masking.
  + Cost estimation for Snowflake compute & Azure infra

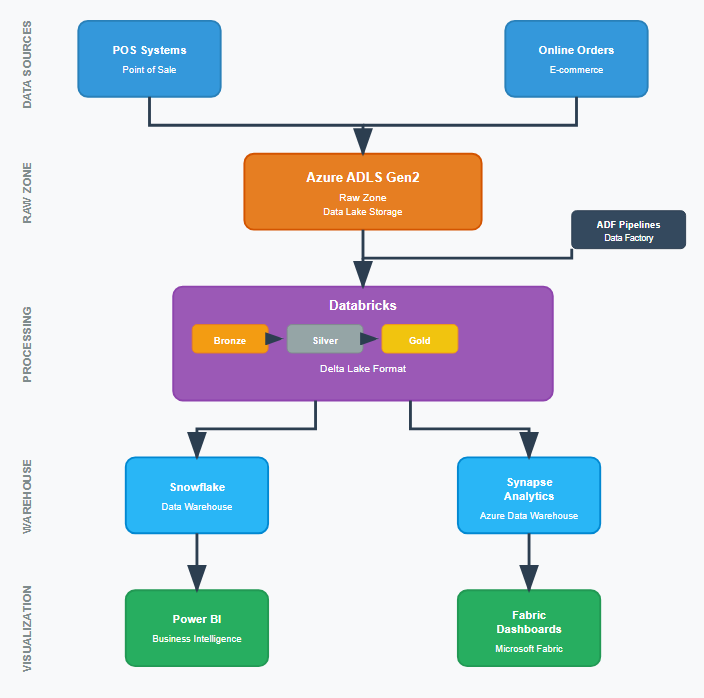
**Use Case / Architecture Diagram**

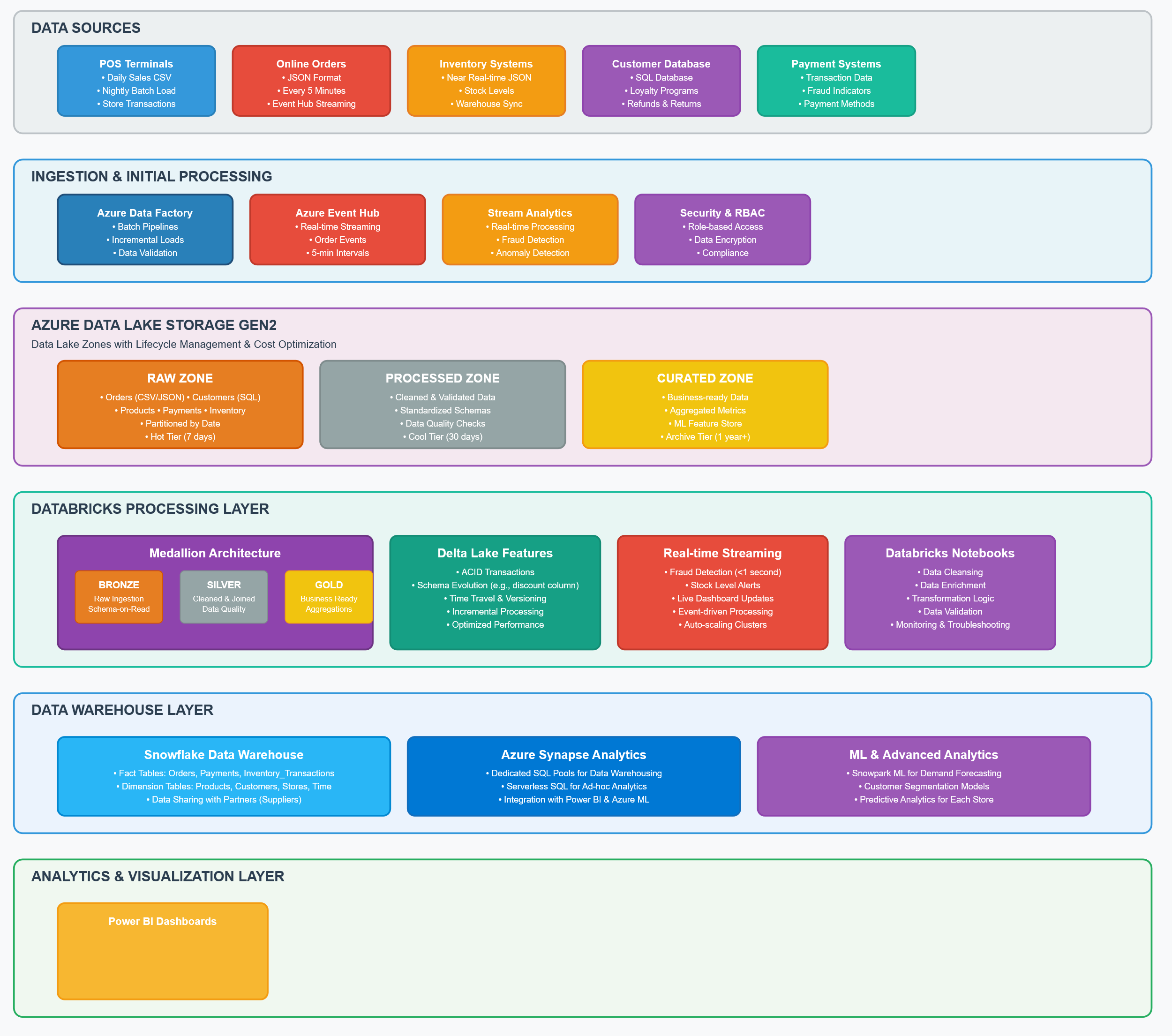
**Use Cases:**

1. **Order Ingestion** – Capture transactions from POS terminals (batch + streaming).
2. **Inventory Updates** – Sync stock in near-real-time across stores & warehouses.
3. **Customer Management** – Loyalty programs, refunds, personalization.
4. **Fraud Detection** – Flag abnormal orders instantly.
5. **Reporting & Analytics** – Sales by region, product, store, promotions.
6. **Error Handling & Monitoring** – Track failed ADF pipelines, Databricks job alerts, and log aggregation in Azure Monitor.
7. **Data Security** – Mask PII fields like customer email/phone in curated zones.
8. **Cost Estimation** – Estimate costs of ADF runs, Databricks clusters, and Snowflake query usage.

**Architecture:**







**User Stories**

* As a **store manager**, I want a **daily sales dashboard** so I can track revenue across stores.
* As an **inventory manager**, I want **stock alerts** when product levels are low.
* As a **fraud analyst**, I want to detect **unusually high transactions in real-time**.
* As a **data engineer**, I want a **scalable pipeline** that can process both batch & streaming data.
* As a **data scientist**, I want to use **Snowpark ML** to predict demand for each store.
* As a **cloud admin**, I want to monitor failed ADF/Databricks jobs and get alerts in Teams/Email.
* As a **compliance officer**, I want to ensure that customer PII data is masked before being shared with BI dashboards.
* As a **finance manager**, I want to estimate and optimize query and infra costs across Azure and Snowflake.

**Expected Deliverables**

* Data Lake zones: Raw / Processed / Curated for **orders, customers, products, payments**.
* ADF pipelines for batch ingestion of POS data.
* Databricks notebooks for cleansing, enrichment, and Delta Lake transformations.
* Incremental data ingestion (daily orders).
* Snowflake warehouse with **fact (Orders)** and **dimension tables (Products, Customers, Stores)**.
* Power BI / Synapse dashboards showing:
  + Daily/Monthly Sales
  + Top-selling products
  + Inventory shortages
  + Suspicious transactions
* Error Handling & Monitoring:
  + ADF failure alerts → Azure Monitor + Action Groups
  + Databricks logs → Log Analytics + custom metrics
  + Health dashboards for pipelines & jobs
* Data Security:
  + Mask customer email addresses & mobile numbers in curated zone and BI layer.
  + Example: customer\_email → XXXX@domain.com
* Cost Estimation:
  + Snowflake: Warehouse credit usage, storage, caching benefits.
  + Azure: ADF pipeline activity cost, Databricks cluster cost per job, storage lifecycle policies.
* Documentation: Architecture diagram + Troubleshooting Playbook.

**Milestones & Duration**

| Milestone | Duration | Details |
| --- | --- | --- |
| M1 – Environment Setup | 1 hr | Setup Azure resources, ADLS, Resource Groups |
| M2 – Storage & Ingestion | 1 hr | Create ADLS zones, ingest POS data (CSV, JSON) with ADF |
| M3 – Transformation in Databricks | 1 hr | Bronze → Silver → Gold Delta pipelines |
| M4 – Warehouse (Snowflake/Synapse) | 1 hr | Load curated POS data, create fact/dimension tables |
| M5 – Advanced Features | 1 hr | Incremental loads, real-time fraud detection, schema evolution |
| M6 – Error Handling & Monitoring | 1 hr | Configure pipeline monitoring, alerts, Databricks job logging |
| M7 – Data Security & PII Masking | 1 hr | Implement column masking in curated zone and BI |
| M8 – Cost Estimation | 1 hr | Estimate infra + Snowflake costs, document optimization strategies |
| M9 – Reporting & BI | 1 hr | Power BI dashboards & Synapse integration |
| M10 – Capstone Delivery | 1 hr | Final demo + documentation submission |

**Implementation Notes**

* **Data Sources:**
  + **Orders** – CSV, daily batch ingestion.
  + **Inventory** – JSON, near real-time ingestion.
  + **Customer details** – SQL DB, periodic sync.
* **Batch Example:**
  + Nightly load of all store sales via **ADF pipelines**.
  + Incremental load for daily new orders.
* **Streaming Example:**
  + Capture new orders every 5 minutes via Event Hub → Databricks Structured Streaming.
  + Feed real-time fraud detection alerts to monitoring dashboards.
* **Delta Lake:**
  + Bronze → Silver → Gold pipeline for ACID transactions.
  + Support schema evolution (e.g., adding discount column without breaking pipeline).
  + Enable time travel for troubleshooting and rollback.
* **Snowflake / Synapse:**
  + Curated data warehouse layer for reporting.
  + Fact tables: Orders, Payments.
  + Dimension tables: Products, Customers, Stores.
  + Use Snowpark for ML-driven demand forecasting.
  + Enable data sharing with suppliers and partners.
* **Synapse / Power BI:**
  + Build interactive dashboards for sales, inventory shortages, and fraud alerts.
  + Embed RLS (Row-Level Security) for role-based access.
* **Error Handling & Monitoring:**
  + ADF → Integration with Azure Monitor for pipeline failure alerts.
  + Databricks → Push logs & metrics to Log Analytics, enable job-level retries.
  + Action Groups → Email/Teams notifications for critical failures.
  + Build a Monitoring Dashboard for pipeline health & cluster costs.
* **Data Security:**
  + Apply masking on PII fields (emails, phone numbers) in curated zone.
    - Example: john.doe@email.com → \*\*\*\*@email.com
  + Use Azure Key Vault for credential and connection string management.
  + Ensure RBAC & ADLS ACLs for secure data access.
* **Cost Estimation & Optimization:**
  + ADF – Activity & pipeline execution cost monitoring.
  + Databricks – Track cost per cluster/job, use auto-termination.
  + Snowflake – Estimate query credit usage; leverage caching & warehouse sizing.
  + Storage Lifecycle – Tier older data in ADLS to optimize storage costs.
  + Document cost impact of peak sales season workloads.

**Evaluation Rubrics**

| **Criteria** | **Weightage** |
| --- | --- |
| Correct Azure setup (RG, ADLS, RBAC, Lifecycle) | 10% |
| ADF pipelines for batch & incremental ingestion | 15% |
| Databricks transformations & Delta Lake implementation | 15% |
| Warehouse schema design (Snowflake/Synapse) | 15% |
| BI Dashboard (sales, inventory, fraud detection) | 10% |
| **Error Handling & Monitoring setup** | 10% |
| **Data Security (PII masking)** | 10% |
| **Cost Estimation & Optimization** | 5% |
| Documentation (architecture, troubleshooting) | 5% |
| Bonus: Real-time fraud detection / ML demand prediction | 5% |