**Pyspark-snowflake-integration**

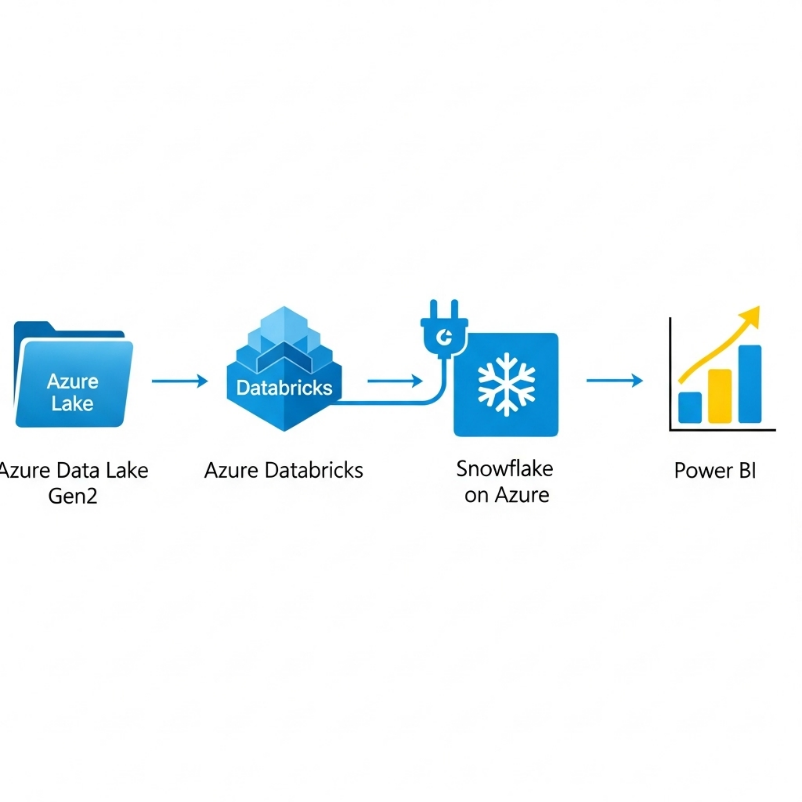
This project demonstrates the integration of **Azure Databricks** (Apache Spark) with **Snowflake Data Cloud** on the Microsoft Azure platform to create a robust, scalable, and high-performance data processing pipeline. The architecture leverages Azure-native services for security, storage, and orchestration while utilizing Snowflake's cloud data platform for enterprise-scale data warehousing.

**2. Objectives**

* **Azure-Native Integration:** Establish secure, optimized connectivity between Azure Databricks and Snowflake using Azure-specific configurations and services.
* **Unified Data Platform:** Create a modern data architecture that combines Azure's data processing capabilities with Snowflake's data warehousing strengths.
* **Enterprise Security:** Implement Azure Active Directory integration and private endpoint connectivity for enhanced security.
* **Cost-Optimized Performance:** Leverage Azure Spot Instances for Databricks clusters and Snowflake's auto-scaling capabilities.
* **End-to-End Pipeline:** Build a complete, production-ready ETL/ELT pipeline with monitoring and error handling.

**3. Azure System Architecture**

The solution integrates multiple Azure services with Snowflake to create a comprehensive data platform.



**3.1. Azure Components**

1. **Azure Data Lake Storage Gen2 (ADLS Gen2):**
   * **Role:** Primary data lake for raw and processed data
   * **Format:** Parquet files organized in medallion architecture (Bronze, Silver, Gold layers)
   * **Integration:** Native integration with Azure Databricks using ABFS driver
2. **Azure Databricks:**
   * **Role:** Distributed data processing and transformation engine
   * **Configuration:**
     + Runtime: 13.3 LTS (Spark 3.4.1, Scala 2.12)
     + Worker nodes: Standard\_F4s\_v2 (Spot instances for cost optimization)
     + Photon Acceleration: Enabled for improved performance
3. **Azure Virtual Network (VNet):**
   * **Role:** Secure network isolation
   * **Components:**
     + Databricks deployed in VNet injection mode
     + Private Endpoints for Snowflake connectivity
     + Network Security Groups (NSGs) for traffic control
4. **Azure Key Vault:**
   * **Role:** Centralized secrets management
   * **Storage:** Snowflake credentials, service principal keys, connection strings
5. **Azure Data Factory:**
   * **Role:** Pipeline orchestration and scheduling
   * **Integration:** Triggers Databricks notebooks and monitors pipeline execution

**3.2. Snowflake Configuration on Azure**

* **Cloud Region:** East US 2 (Azure)
* **Storage:** Azure Blob Storage (internal)
* **Connectivity:** Private Link integration with Azure VNet
* **Authentication:** Azure AD integration for single sign-on

**3.3. Data Flow**

1. **Ingest:** Data lands in ADLS Gen2 Bronze container from various sources (Azure SQL, Cosmos DB, Event Hubs)
2. **Process:** Azure Databricks reads from Bronze, applies transformations using PySpark, writes to Silver/Gold layers
3. **Load:** Transformed data from Gold layer is loaded to Snowflake using Spark-Snowflake connector
4. **Serve:** Snowflake serves as the presentation layer for BI tools (Power BI, Tableau)

**4. Azure Technology Stack**

| Component | Technology / Service | Specification | Purpose |
| --- | --- | --- | --- |
| **Data Processing** | Azure Databricks | Runtime 13.3 LTS, Photon Enabled | Distributed data processing |
| **Data Warehouse** | Snowflake on Azure | Business Critical Edition | Cloud data warehousing |
| **Cloud Storage** | Azure Data Lake Gen2 | Hot Tier, Hierarchical Namespace | Data lake storage |
| **Orchestration** | Azure Data Factory | v2 | Pipeline scheduling & monitoring |
| **Secrets Management** | Azure Key Vault | Standard Tier | Secure credentials storage |
| **Networking** | Azure Virtual Network | Private Endpoints, NSGs | Secure network isolation |
| **Identity** | Azure Active Directory | Enterprise Edition | Unified authentication |
| **Monitoring** | Azure Monitor | Log Analytics Workspace | Centralized logging & metrics |

**5. Implementation on Azure**

**5.1. Prerequisites and Azure Configuration**

**A. Azure Infrastructure Setup:**

bash

**

**B. Snowflake Azure Configuration:**

sql

**

**5.2. Azure Data Factory Pipeline**



**6. Azure-Specific Results and Validation**

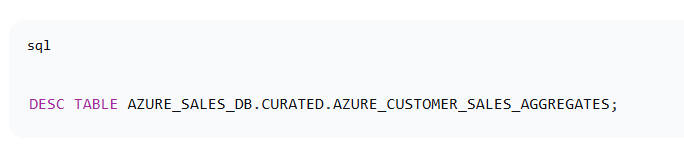
**6.1. Performance Metrics on Azure**

The pipeline was tested with a **~500 GB dataset** in Azure Data Lake Gen2:

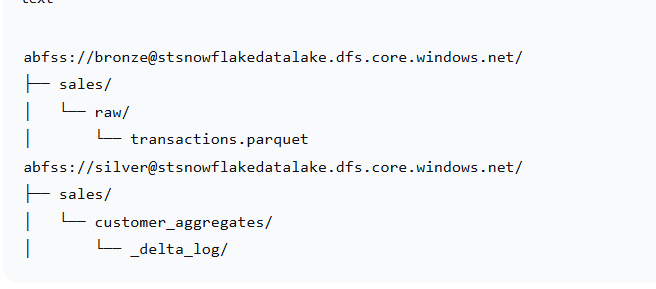
* **Data Read Time (ADLS Gen2 to Databricks):** ~8 minutes
* **Transformation Time (Databricks, 8-node cluster):** ~6 minutes
* **Data Write Time (Databricks to Snowflake):** ~10 minutes
* **Total Pipeline Duration:** **~24 minutes**
* **Cost:** ~$45 per execution (Databricks + Snowflake compute)

**6.2. Output Validation**

**Snowflake Table Schema:**

**Azure Data Lake Structure:**

text

**6.3. Azure Monitor Dashboard**



Key metrics monitored:

* Databricks job duration
* Snowflake warehouse credit usage
* Data volume processed
* Success/failure rates

**7. Azure-Specific Challenges and Solutions**

| Challenge | Azure-Specific Solution |
| --- | --- |
| **Network Security** | Implemented VNet injection for Databricks, Private Endpoints for Snowflake connectivity, and Network Security Groups |
| **Credential Management** | Used Azure Key Vault integrated with Databricks secret scopes for secure credential storage |
| **Cost Optimization** | Configured Databricks with Spot instances and Snowflake with auto-suspend warehouse settings |
| **Data Governance** | Implemented Azure Purview for data cataloging and lineage tracking |
| **Performance Tuning** | Used Photon engine in Databricks and optimized partitioning strategies for Azure storage |

**8. Security Implementation**

**8.1. Azure Security Features**

1. **Managed Identity:** Databricks uses managed identity to access Azure resources
2. **Private Endpoints:** All traffic between Azure and Snowflake uses private networking
3. **Encryption:** Data encrypted at rest (Azure Storage Service Encryption) and in transit (TLS 1.2)
4. **Azure AD Integration:** Single sign-on for both Azure services and Snowflake
5. **Role-Based Access Control (RBAC):** Fine-grained permissions for Azure resources