**Day 25 Notes**

**Name: Podutur Lahari - DE126**

**Date:16-12-2024**

**Integrating Azure Databricks with Azure Synapse Analytics**

Azure Databricks and Azure Synapse Analytics work together to create a modern data architecture. Below are concise notes on how to integrate them and best practices for administration, security, tools, runtime management, high availability (HA), disaster recovery (DR), and cluster optimization.

**Integration Steps**

1. **Data Flow**:
   * Use Azure Data Lake as a storage layer.
   * Use Azure Databricks for data processing and transformation.
   * Use Azure Synapse Analytics for data warehousing and reporting.
2. **Connectivity**:
   * Link Databricks to Synapse using a JDBC or ODBC driver for direct queries.
   * Use Azure Synapse Connector for seamless data movement between the two.
3. **Orchestration**:
   * Leverage Azure Data Factory to orchestrate data pipelines between Databricks and Synapse.

**Best Practices**

**1. Workspace Administration**

* Define roles and permissions in Azure Databricks to control access.
* Use resource tagging to organize and manage Databricks resources.
* Regularly monitor workspace usage and cost using Azure Cost Management.

**2. Security**

* **Authentication**: Use Azure Active Directory (AAD) for single sign-on and access control.
* **Data Encryption**: Enable encryption for both in-transit (TLS) and at-rest data.
* **Network Isolation**: Use private endpoints for Databricks workspaces to restrict access.
* **Secrets Management**: Store database credentials securely in Azure Key Vault.

**3. Tools Integration**

* Use **Azure Synapse Studio** for analytics and visualization.
* Leverage **Azure Monitor** for logging and monitoring Databricks performance.
* Integrate **Power BI** for data visualization and reporting.

**4. Databricks Runtime**

* Use the latest Databricks Runtime version for performance and security enhancements.
* For machine learning workloads, use the ML runtime optimized for TensorFlow and PyTorch.
* Use the Delta Lake runtime for ACID transactions and time-travel queries.

**5. High Availability (HA) & Disaster Recovery (DR)**

* **HA**: Use multi-zone clusters for high availability.
* **DR**:
  + Regularly back up data to Azure Data Lake.
  + Enable Delta Lake’s versioning for easy data recovery.
  + Document and automate cluster configurations for quick restoration.

**6. Cluster Management**

* **Auto-scaling**: Enable auto-scaling to optimize resource usage and costs.
* **Cluster Pools**: Use cluster pools for faster startup times.
* **Instance Types**: Choose compute-optimized VMs for high processing workloads.
* **Lifecycle Management**: Schedule cluster termination for idle periods to reduce costs.