**Day 26 Notes**

**Name: Podutur Lahari - DE126**

**Date:19-12-2024**

**Setting Up Azure Synapse Analytics and Integration with Azure Databricks**

**1. Setting Up Azure Analytics**

**Steps to Set Up**

1. **Create a Synapse Workspace**:
   * Navigate to the Azure portal.
   * Search for **Azure Synapse Analytics** and create a workspace.
   * Configure resource details (resource group, region, and workspace name).
2. **Enable Managed VNet (Optional)**:
   * Use a managed Virtual Network (VNet) for secure network traffic.
3. **Add a Dedicated SQL Pool (Optional)**:
   * Create a dedicated SQL pool for scalable, high-performance queries.
4. **Connect to Data Sources**:
   * Link your Synapse workspace to Azure Data Lake, Blob Storage, or other sources.
5. **Synapse Studio**:
   * Use Synapse Studio for managing pipelines, queries, and datasets.

**2. Integrating Azure Synapse Analytics with Azure Databricks**

**Steps to Integrate**

1. **Set Up Azure Databricks**:
   * Create a Databricks workspace in the Azure portal.
   * Configure it with the required permissions and virtual networks.
2. **Install Synapse Connector**:
   * In Databricks, install the **Azure Synapse Connector** from Maven:
   * com.microsoft.azure:spark-mssql-connector:<version>
3. **Establish Connection**:
   * Use JDBC or ODBC for connecting Databricks to Synapse Analytics:
   * synapse\_jdbc\_url = "jdbc:sqlserver://<synapse-server>.database.windows.net:1433;database=<database-name>;encrypt=true;trustServerCertificate=false"
4. **Authenticate**:
   * Use a service principal or Azure Managed Identity for authentication.
5. **Data Flow**:
   * Transform data in Databricks using Spark and load it into Synapse Analytics.
6. **Orchestrate Pipelines**:
   * Use Azure Data Factory or Synapse Pipelines to schedule and manage workflows.

**Best Practices for Integration**

* **Performance**: Use partitioning and caching in Databricks for efficient data processing.
* **Security**: Secure data transfer with private endpoints and Azure Key Vault.
* **Delta Lake**: Leverage Delta Lake for seamless data transfer with ACID compliance.
* **Data Formats**: Use Parquet or ORC for optimized data storage and querying.