😊 Loading data into SQL Pool using Polybase

PolyBase in **Azure Synapse Analytics** is a powerful data virtualization technology that allows you to query and integrate external data sources seamlessly. It provides the ability to access and query data stored outside the database—such as in **Azure Data Lake Storage**, **Azure Blob Storage**, or other relational and non-relational data stores—using standard **T-SQL** syntax.

Key Features of PolyBase

1. **Data Virtualization:**
    - Query external data without moving or copying it into the database.
    - Enables seamless integration of structured and unstructured data from diverse sources.
2. **High Performance:**
    - Leverages **Massively Parallel Processing (MPP)** architecture for fast query execution.
    - Suitable for large-scale analytical queries involving both local and external data.
3. **Standard T-SQL Syntax:**
    - No need to learn new query languages; you can use familiar SQL syntax to interact with external data.
4. **Wide Range of Supported Data Sources:**
    - **Azure Data Lake Storage**
    - **Azure Blob Storage**
    - **SQL Server** or **Azure SQL Database**
    - **Oracle**, **Teradata**, **MongoDB**, and more.
5. **Integration with Synapse Analytics:**
    - PolyBase plays a key role in loading data into **Dedicated SQL Pools** and in querying external data sources directly from **Serverless SQL Pools**.

How PolyBase Works

1. **External Table Creation:**
    - Define an **external table** that maps to the external data source.
    - The external table acts as a schema-on-read layer, enabling you to query external data as if it were local.
2. **Data Retrieval:**
    - When a query is executed, PolyBase retrieves the necessary data from the external source.

    - It optimizes data access by pushing down predicates (e.g., filters) to the external source when possible.
3. **Query Execution:**
    - Query execution combines data from local and external sources, allowing for complex joins and aggregations.

Use Cases

1. **Big Data Analytics:**
    - Query large datasets stored in Azure Data Lake without copying data into Synapse.
2. **Data Integration:**
    - Combine on-premises or cloud-based data sources with data stored in Synapse.
3. **ETL Workflows:**
    - Use PolyBase to ingest and transform data efficiently during extract, transform, and load (ETL) processes.
4. **Cost Optimization:**
    - Avoid unnecessary data duplication by querying data in its original location.

Benefits of Using PolyBase

- **Scalability:** Handles large datasets with ease, leveraging the MPP architecture of Synapse.
- **Simplicity:** Reduces the need for complex data movement pipelines by enabling direct querying.
- **Cost Efficiency:** Minimizes storage costs by eliminating the need to store duplicate copies of data.

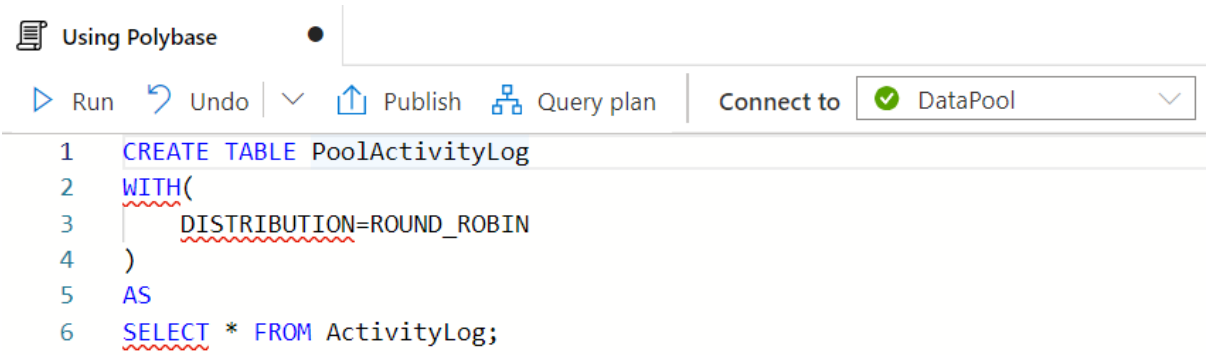
Considerations and Limitations

- **Performance:** PolyBase works best with large, read-intensive workloads; smaller queries may incur latency due to network overhead.
- **Data Format Support:** Works with common formats like **Parquet**, **CSV**, **ORC**, but format support varies by source.
- **Access Permissions:** Requires appropriate credentials and permissions to access external data.

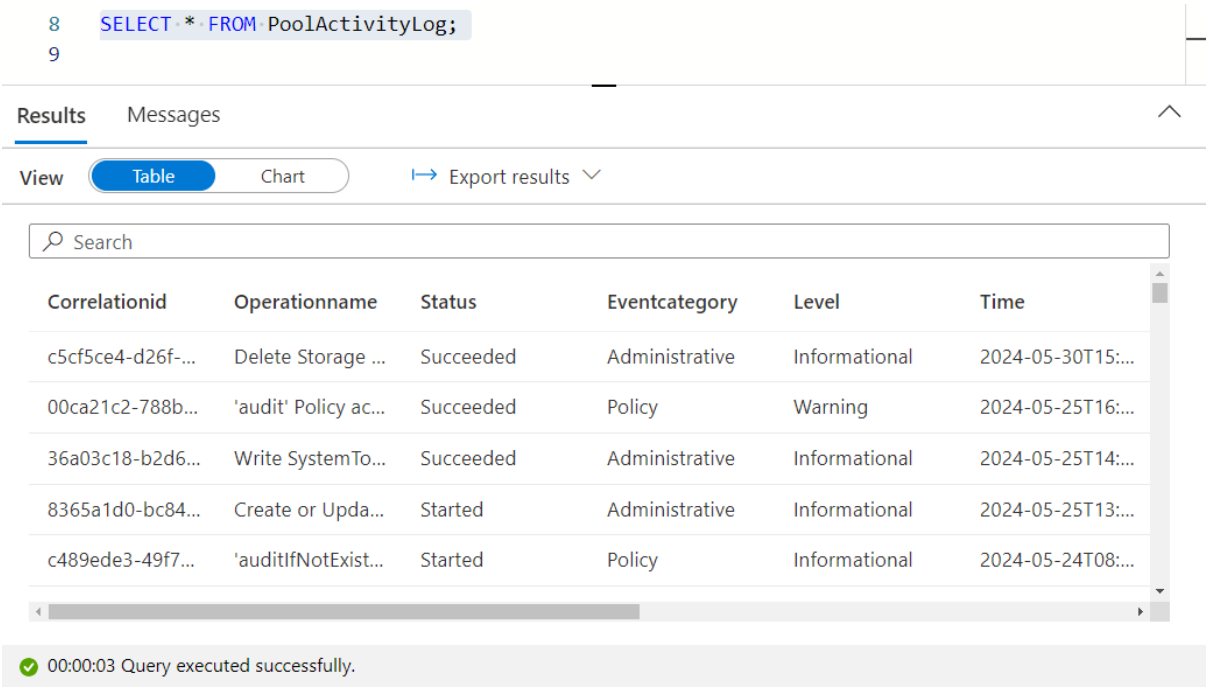
PolyBase is a versatile tool for integrating and analyzing data across diverse environments in Azure Synapse Analytics.

😄 To begin with the Lab

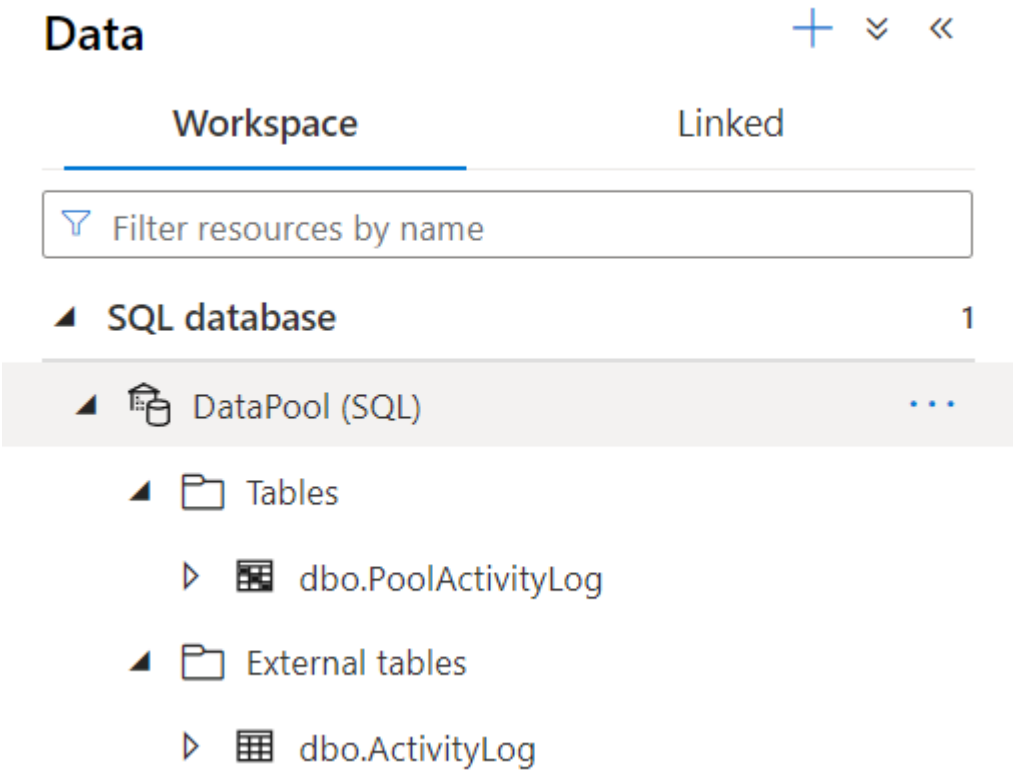
1. In this lab we are going to use Polybase to load data from our Activity Log External table to a new table called Pool Activity Log.



2. Below you can see that the data is transferred from our external table to the Pool Activity Log table.
3. This is also a persistent table, meaning it contains the data inside it. The data inside this table is not being fetched from the data lake. The data is stored inside this table.



4. Also, remember to publish the changes to save everything.
5. If you go to the data tab open the workspace section and expand the SQL Database, you will see two types of Tables: external and normal.
6. The normal table is the persisted table which we just created using the Polybase.


Data
Workspace
Linked
Filter resources by name
SQL database
1
DataPool (SQL)
Tables
dbo.PoolActivityLog
External tables
dbo.ActivityLog