Aas Trailblazers

Unleashing insights

ETL vs ELT

SMP
(Symmetric
Multiprocessing)

VS.

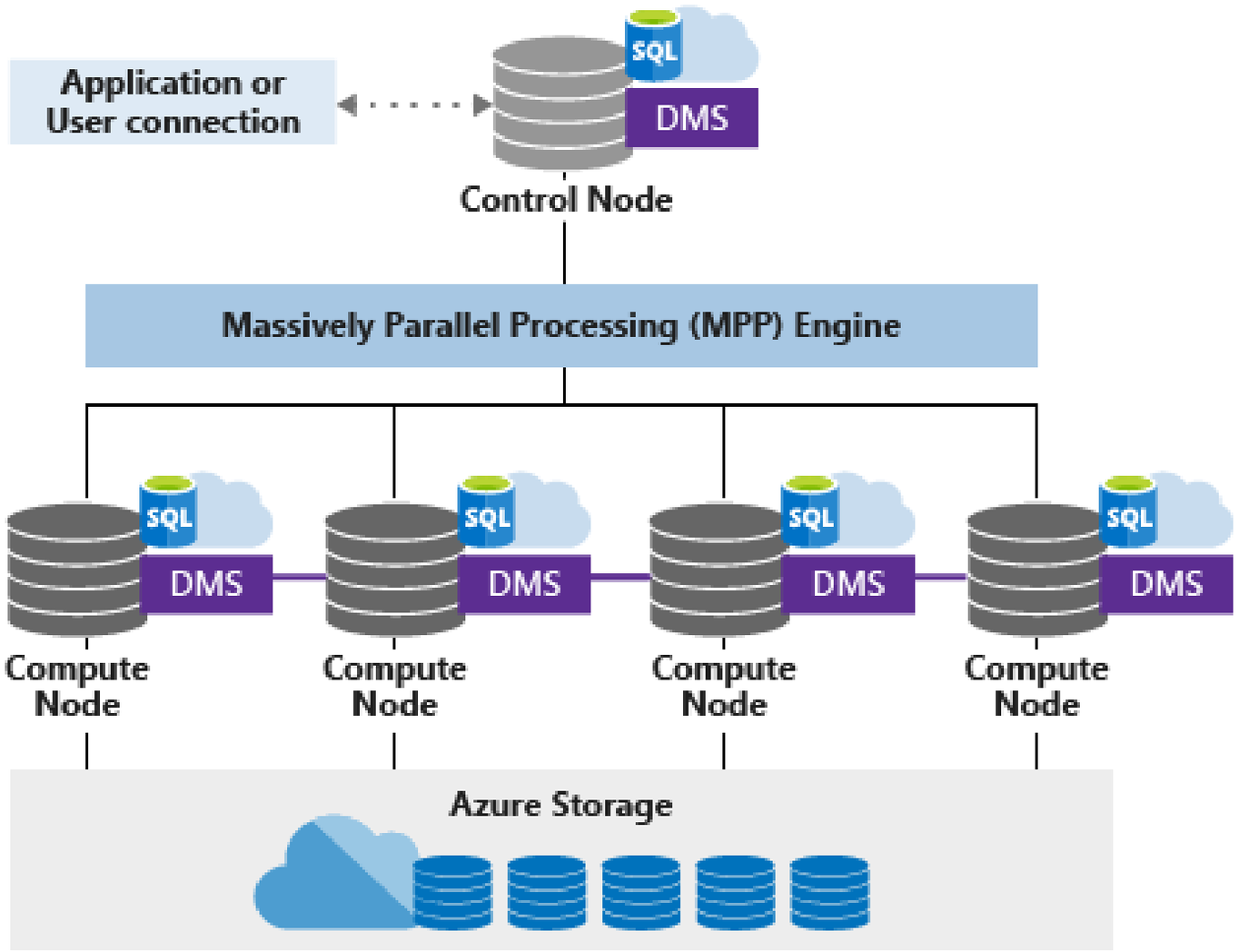
MPP
(Massively Parallel
Processing)

ETL – Extract, Transform and Load

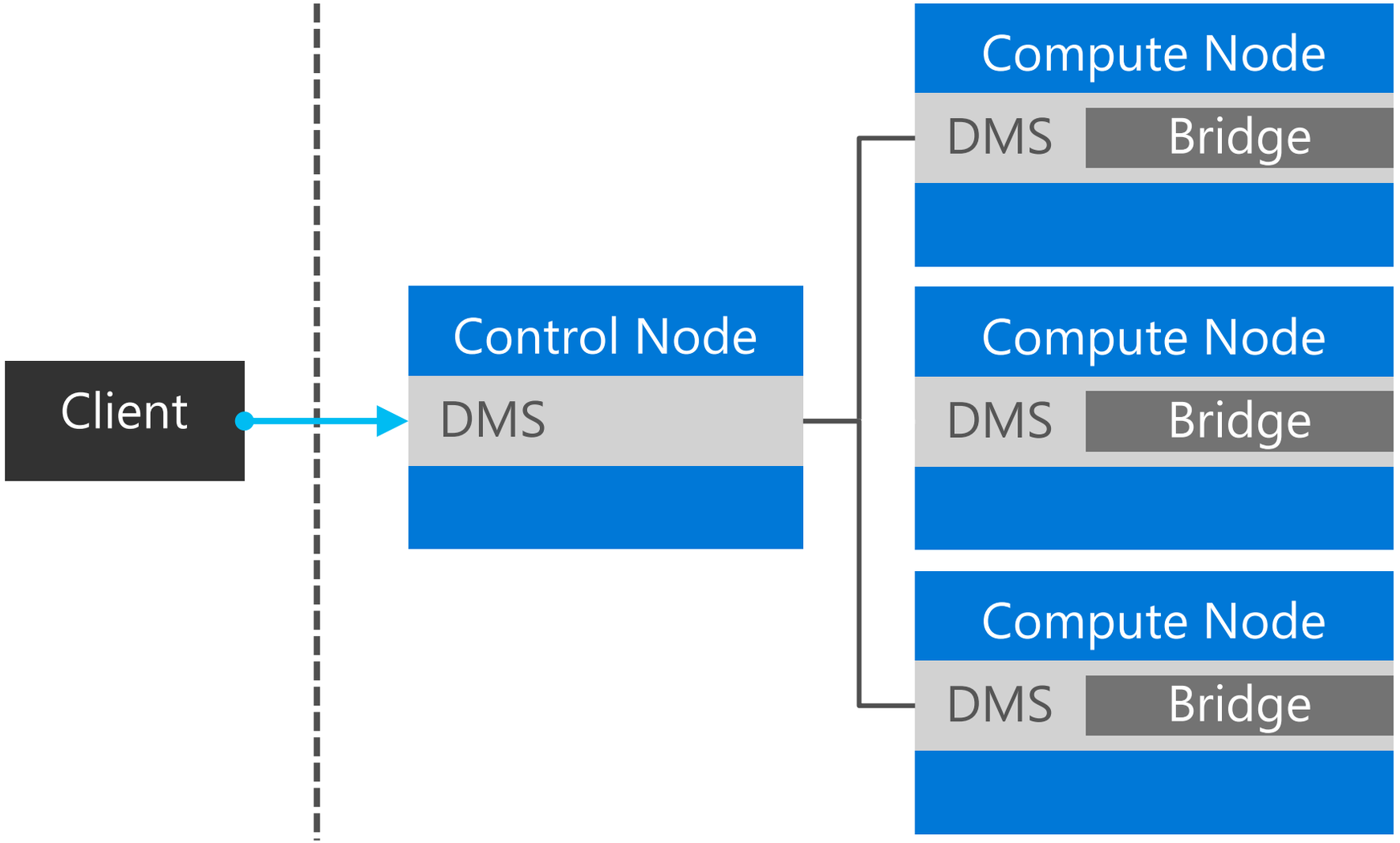
ELT – Extract, Load and Transform

To learn more about SMP vs MPP, please refer my earlier video, “Azure Synapse Analytics | Introduction and Getting Started”

Synapse SQL – MPP Architecture

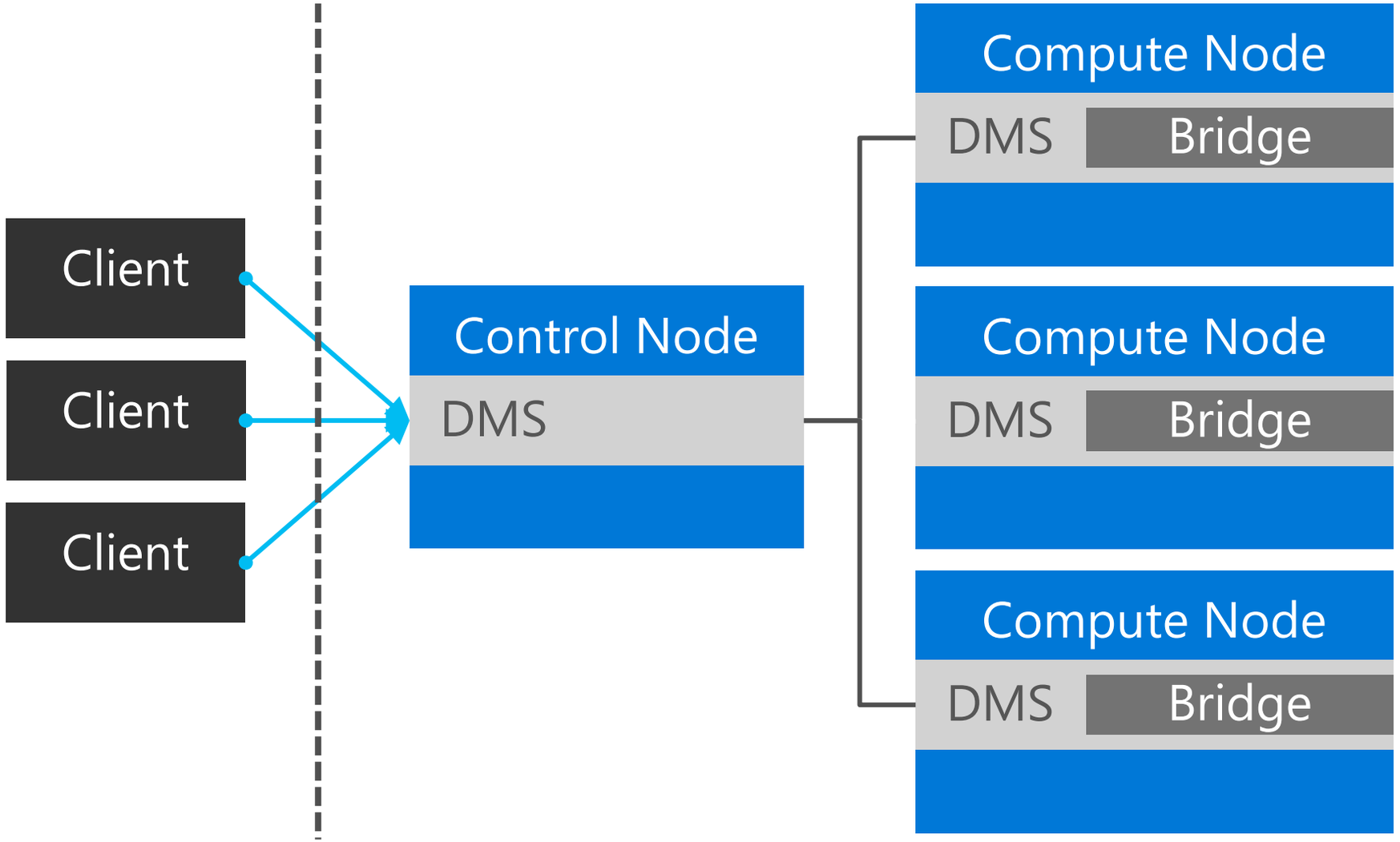


Single Gated – Data Load



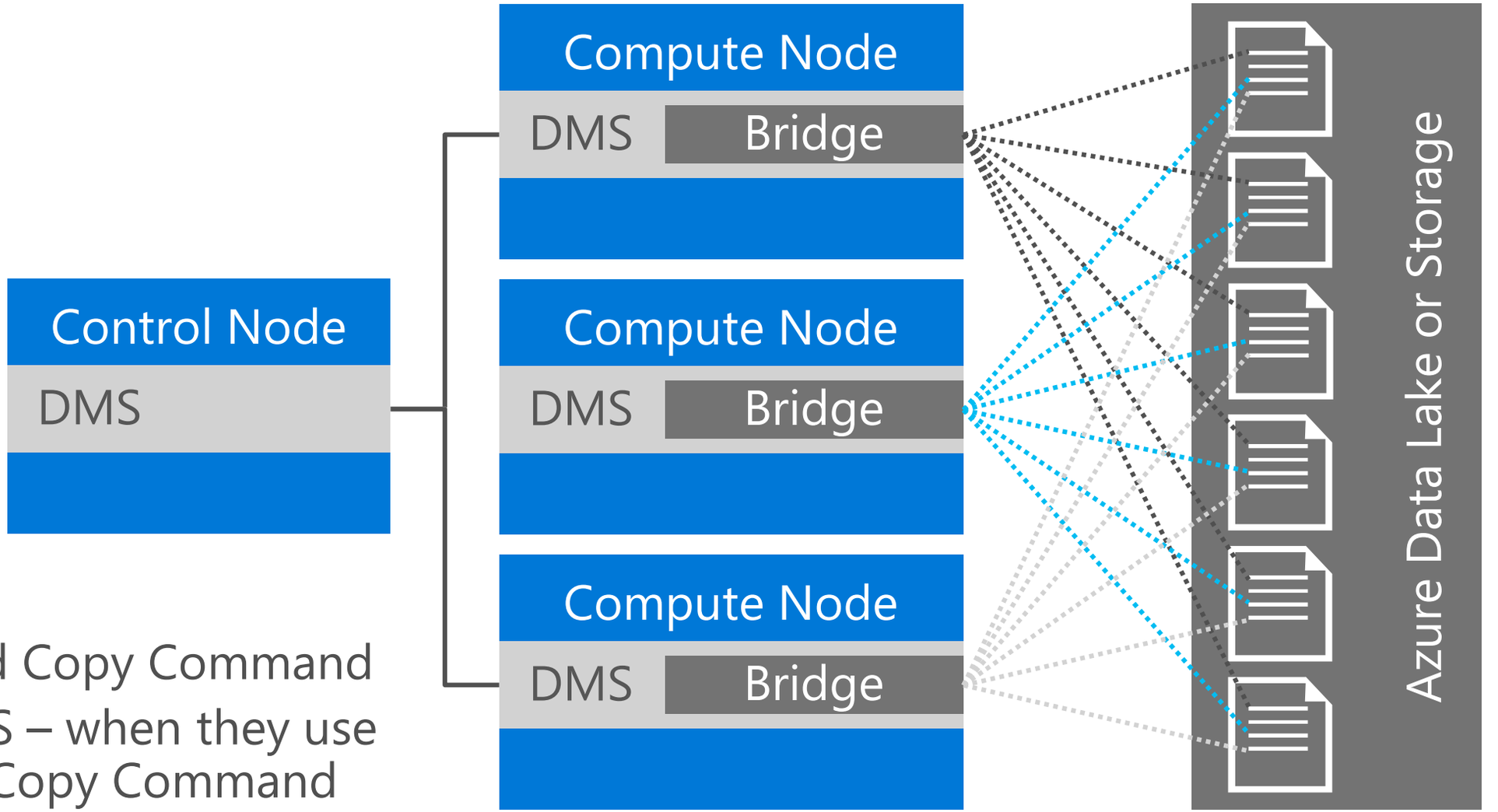
Applies to bcp or SqlBulkCopy API https://docs.microsoft.com/en-us/sql/tools/bcp-utility?view=sql-server-ver15

Single Gated – Data Load – Parallelized



Applies to bcp or SqlBulkCopy API

Parallel – Data Load

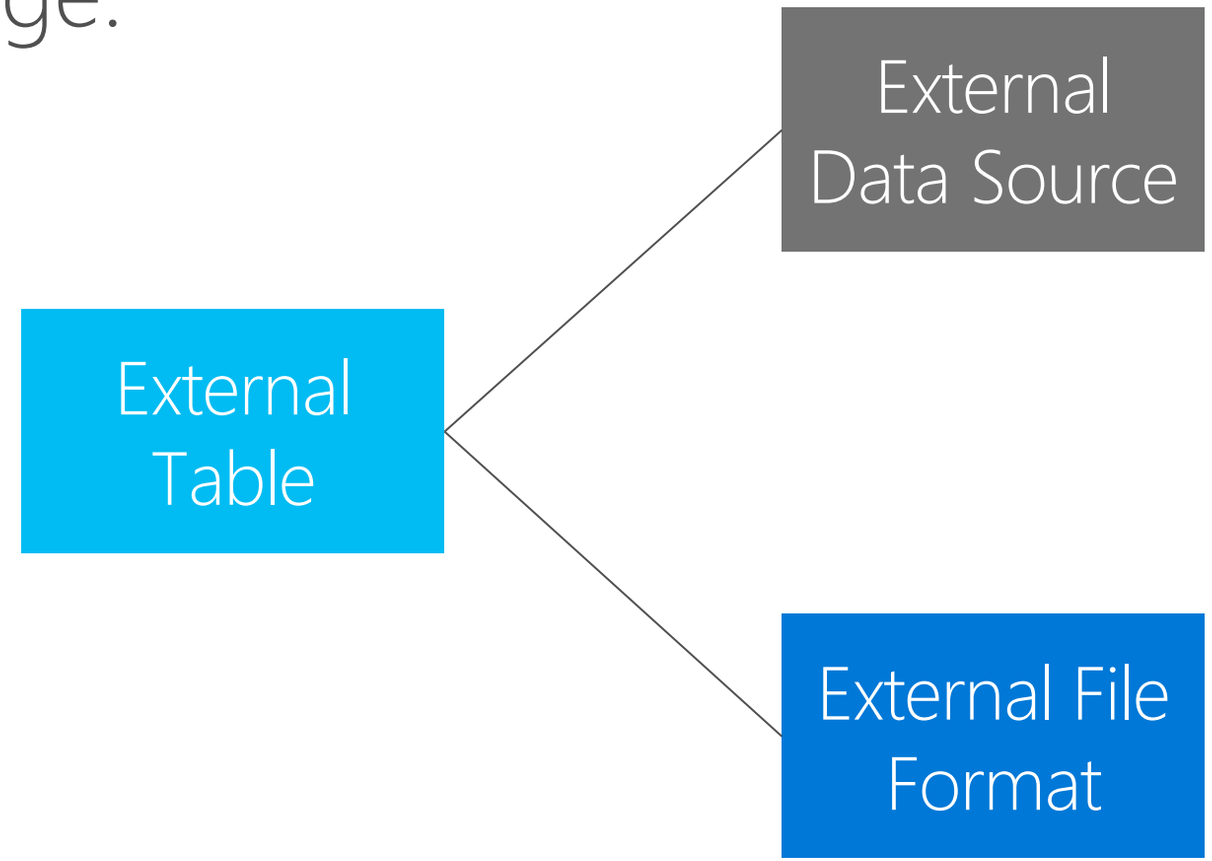


Applies to

- Polybase and Copy Command
- ADF and SSIS – when they use Polybase or Copy Command
- Spark Connector – uses Polybase

Polybase

- PolyBase is a feature that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.



sys.tables
sys.external_tables

Polybase – External Table Types

Depending on the type of the external data source, you can use two types of external tables:

- Hadoop external tables that you can use to read and export data in various data formats such as CSV, Parquet, and ORC. Hadoop external tables are available in dedicated SQL pools, but they aren't available in serverless SQL pools.
- Native external tables that you can use to read and export data in various data formats such as CSV and Parquet. Native external tables are available in serverless SQL pools, and they are in preview in dedicated Synapse SQL pools.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop

Polybase – External Table Types

External table type	Hadoop	Native
Dedicated SQL pool	Available	Parquet tables are available in **gated preview** - contact your Microsoft Technical Account Manager or Cloud Solution Architect to check if you can add your dedicated SQL pool to the gated preview.
Serverless SQL pool	Not available	Available
Supported formats	Delimited/CSV, Parquet, ORC, Hive RC, and RC	Serverless SQL pool: Delimited/CSV, Parquet, and Delta Lake(preview) Dedicated SQL pool: Parquet
Folder partition elimination	No	Only for partitioned tables synchronized from Apache Spark pools in Synapse workspace to serverless SQL pools
Custom format for location	Yes	Yes, using wildcards like `/year=*/month=*/day=*`
Recursive folder scan	No	Only in serverless SQL pools when specified `/**` at the end of the location path
Storage filter pushdown	No	Yes in serverless SQL pool. For the string pushdown, you need to use `Latin1_General_100_BIN2_UTF8` collation on the `VARCHAR` columns.
Storage authentication	Storage Access Key(SAK), AAD passthrough, Managed identity, Custom application Azure AD identity	Shared Access Signature(SAS), AAD passthrough, Managed identity

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop

Data Import – with CREATE TABLE AS (CTAS)

- Use CTAS to load data into staging table
- Fully parallelized operation
- Fastest way to import data
- CTAS provides flexibility to:
  - Re-create a table with a different distribution type or do hash distribution on different column
  - Change index type for the new table

```
CREATE TABLE [stg].[DimProduct]
WITH (HEAP, DISTRIBUTION =
ROUND_ROBIN)
AS
SELECT * FROM [asb].[DimProduct]
OPTION (LABEL = 'CTAS : Load
[stg].[DimProduct]');
GO
```

Data Export – with CREATE EXTERNAL TABLE AS (CETAS)

- Use CETAS to export data to Azure Storage or Data Lake
- Fully parallelized operation
- Fastest way to export the result of a T-SQL query

```
CREATE EXTERNAL TABLE [asb].[DimProduct_2]
WITH (
        LOCATION = '/DimProduct_2/',
        DATA_SOURCE = AzureDataLake_secured,
        FILE_FORMAT = TextFileFormatContoso
)
AS
SELECT * FROM [stg].[DimProduct]
GO
```

```
CREATE EXTERNAL TABLE asb.DimActiveCustomers
WITH
    (
        LOCATION = '/ActiveCustomers/',
        DATA_SOURCE = AzureDataLake_secured,
        FILE_FORMAT = TextFileFormatContoso
    )
AS
SELECT T2.* FROM cso.FactInternetSales T1 JOIN cso.DimCustomer T2
ON ( T1.CustomerKey = T2.CustomerKey)
GO
```

Copy Command

Overview

Copies data from source to destination

Benefits

Retrieves data from all files from the folder and all its subfolders.

Supports Azure Data Lake Storage (ADLS) Gen 2 and Azure Blob Storage.

Supports CSV, PARQUET, ORC file formats

Supports multiple locations from the same storage account, separated by comma

```
COPY INTO test_1
FROM 'https://XXX.blob.core.windows.net/customerdatasets/test_1.txt'
WITH (
   FILE_TYPE = 'CSV',
   CREDENTIAL=(IDENTITY= 'Shared Access Signature',
SECRET='<Your_SAS_Token>'),
   FIELDQUOTE = '"',
   FIELDTERMINATOR=';',
   ROWTERMINATOR='0X0A',
   ENCODING = 'UTF8',
   DATEFORMAT = 'ymd',
   MAXERRORS = 10,
   ERRORFILE = '/errorsfolder/'--path starting from the storage container,
   IDENTITY_INSERT
   )
```

```
COPY INTO test_parquet
FROM 'https://XXX.blob.core.windows.net/customerdatasets/test.parquet'
WITH (
   FILE_FORMAT = myFileFormat
   CREDENTIAL=(IDENTITY= 'Shared Access Signature',
SECRET='<Your_SAS_Token>')
)
```

https://docs.microsoft.com/en-us/sql/t-sql/statements/copy-into-transact-sql?view=azure-sqldw-latest

Demo