Data Migration – Azure Synapse Service

Eshant Garg
Data Engineer, Architect, Advisor
eshant.garg@gmail.com

Introduction

Best practices for Data Load in MPP Architecture

Different loading methods

Why PolyBase is preferred

SSIS vs PolyBase loading methods

PolyBase process in detail

Demo – Loading with different techniques

Best Practices for Data Laod

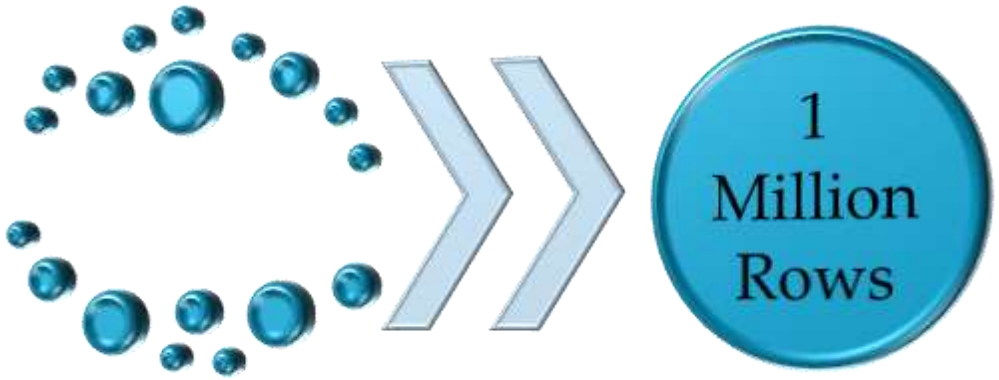
Data Warehouse Readers

Your DWUs have a direct impact on how fast you can load data in parallel

No of DWU	**100**	**200**	**300**	**400**	**500**	**600**	**1000**	**1500**	**2000**
Readers	8	16	24	32	40	48	60	60	60
Writers	60	60	60	60	60	60	60	60	60

Optimize Insert Batch Size

Avoid trickle insert pattern. Ideal batch size is 1 million or more direct or in a file

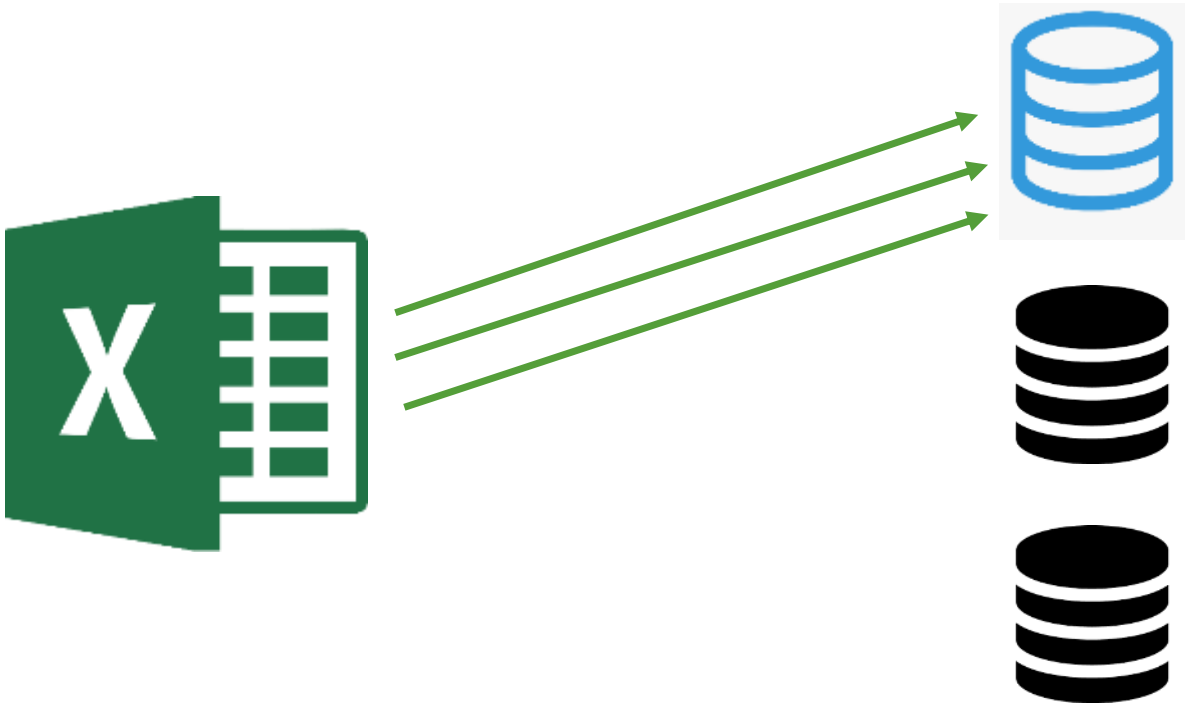


Combine rows to make it a batch of 1 million

Ideal batch size

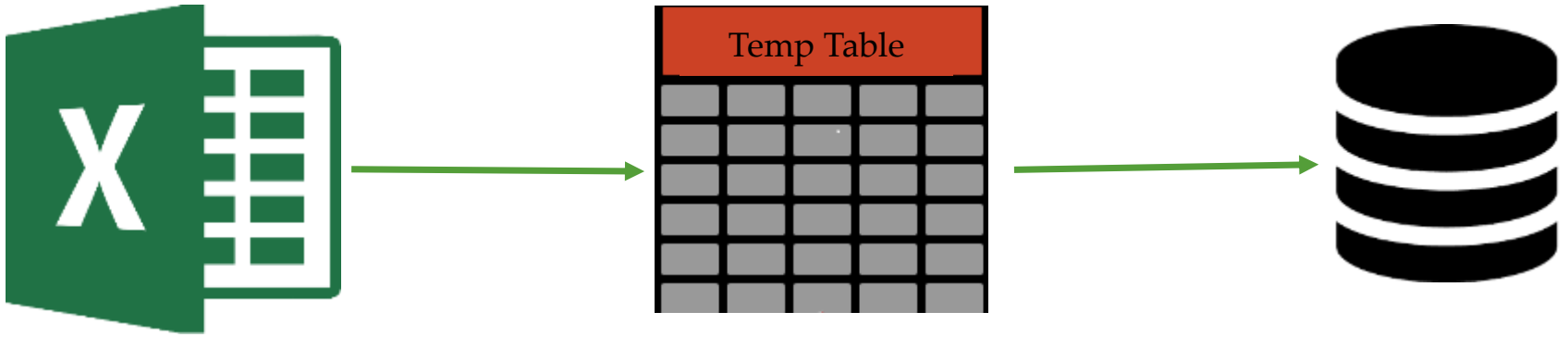
Avoid ordered data

Data ordered by distribution key can introduce hot spots that slow down the load operation



Using temporary tables

Stage and transform on a Temp Heap table before moving to permanent storage.



CREATE TABLE AS

```
CREATE TABLE #tmp_fct
WITH
(
DISTRIBUTION = ROUND_ROBIN
)
AS
SELECT *
FROM
[dbo].[FactInternetSales];
```

- Fully Parallel operation
- It is minimally logged
- It can change: distribution, table type, partitioning

Loading Methods

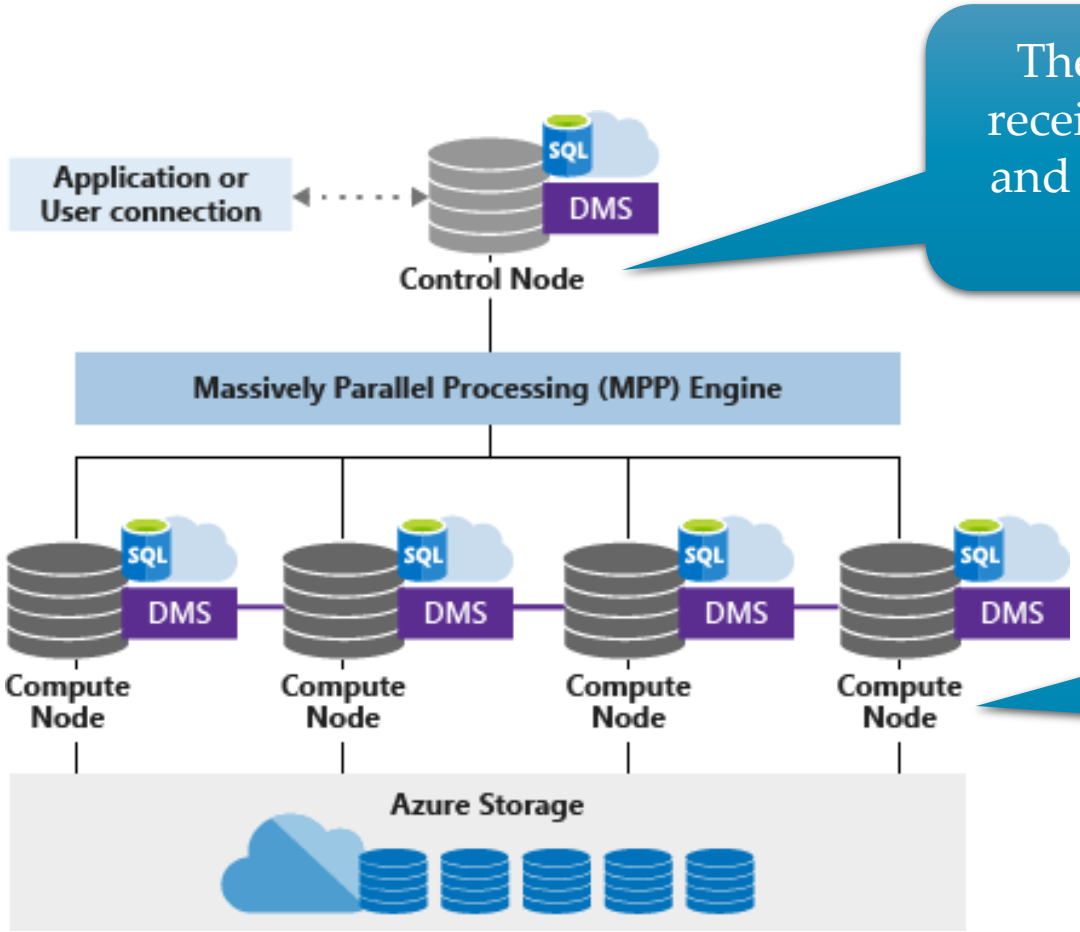
Single client loading methods

- SSIS
- Azure Data Factory
- BCP
- Can add some parallel capabilities but are bottlenecked at the control node

Parallel readers loading methods

- PolyBase
- Reads from Azure blob Storage and loads the contents into Azure SQL DW
- Bypasses the Control Node and loads directly into the Compute Nodes

Control Node

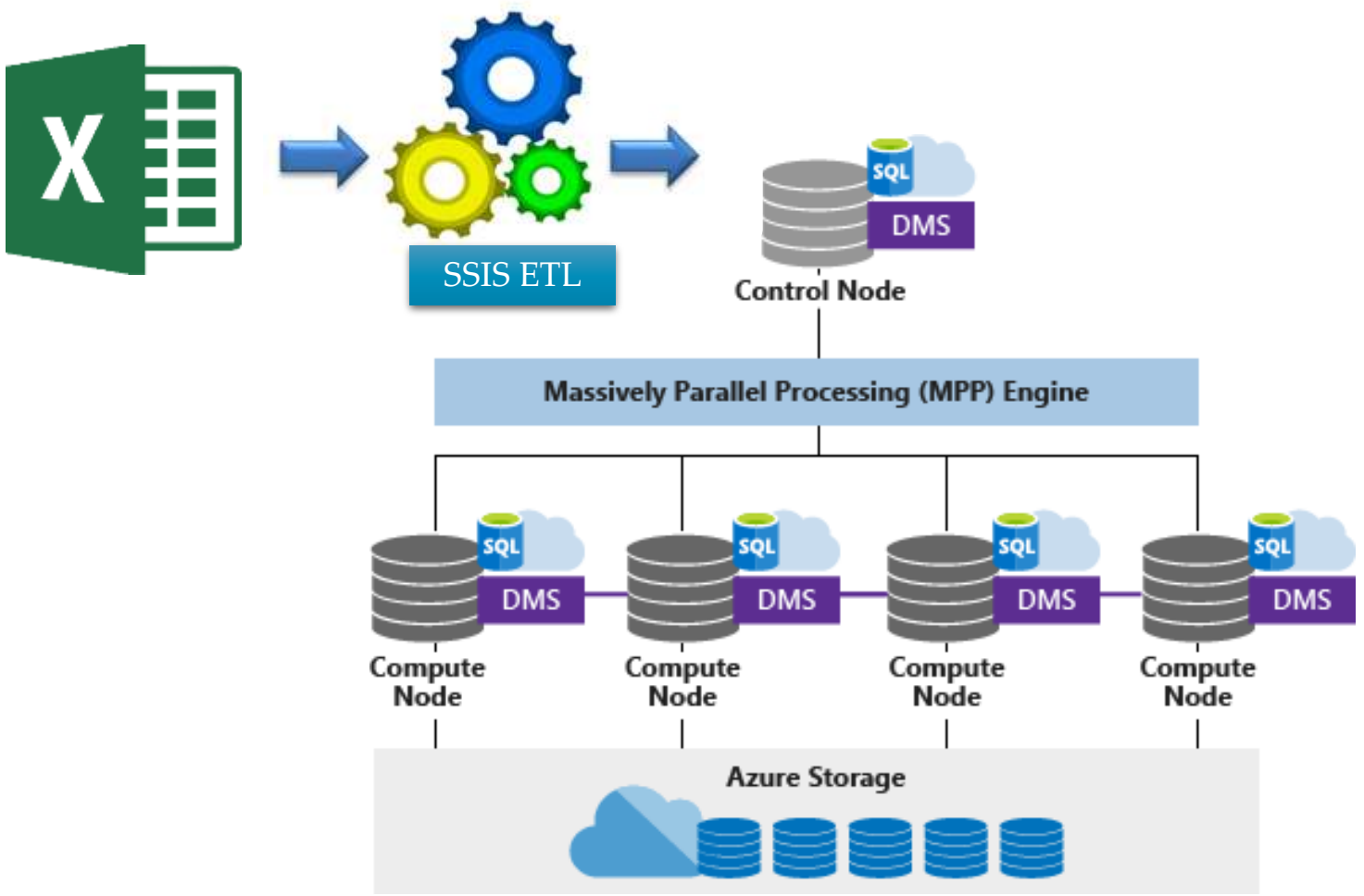


The Control node receives connections and orchestrates the queries

The Compute nodes do processing on the data and scale with the DWUs.

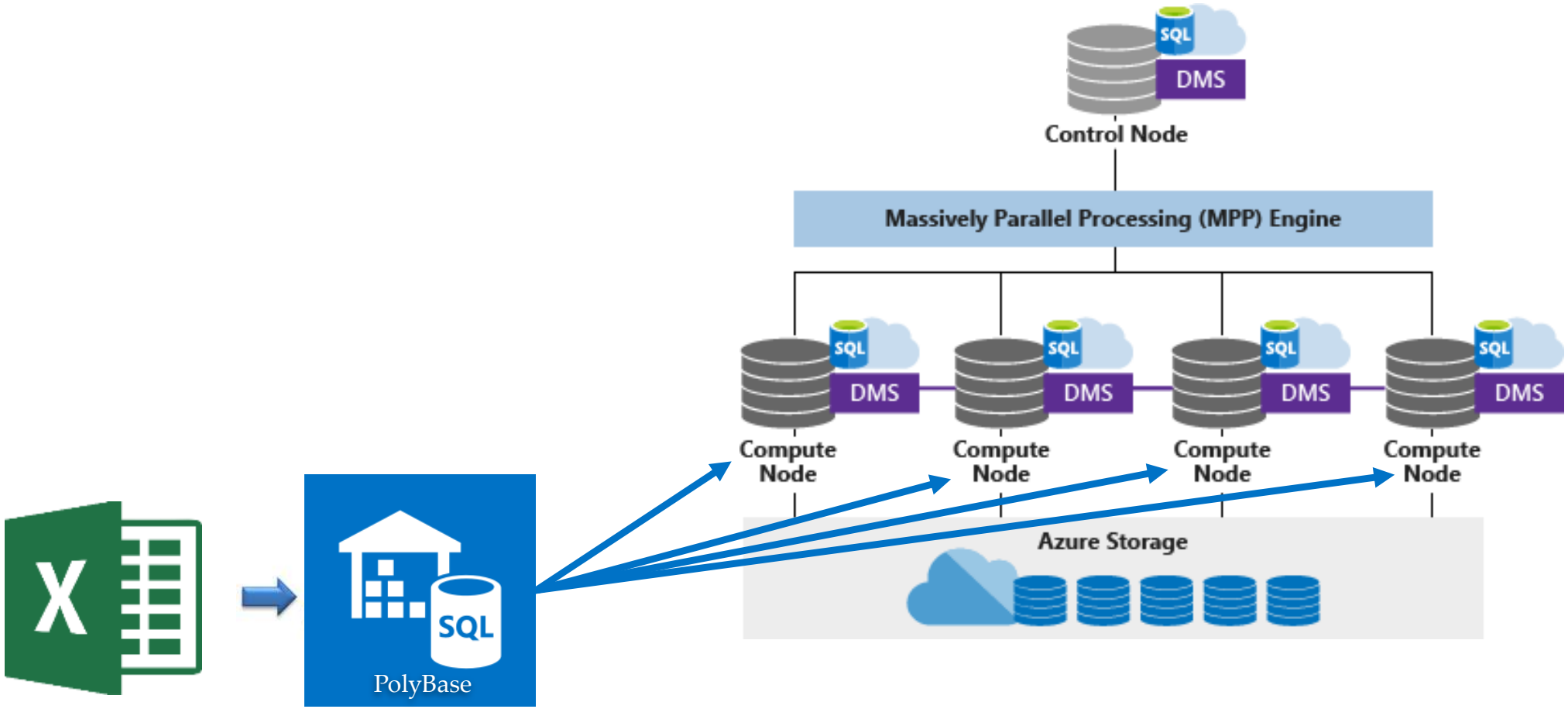
Source: Microsoft

Loading with SSIS



Source: Microsoft

Loading with PolyBase

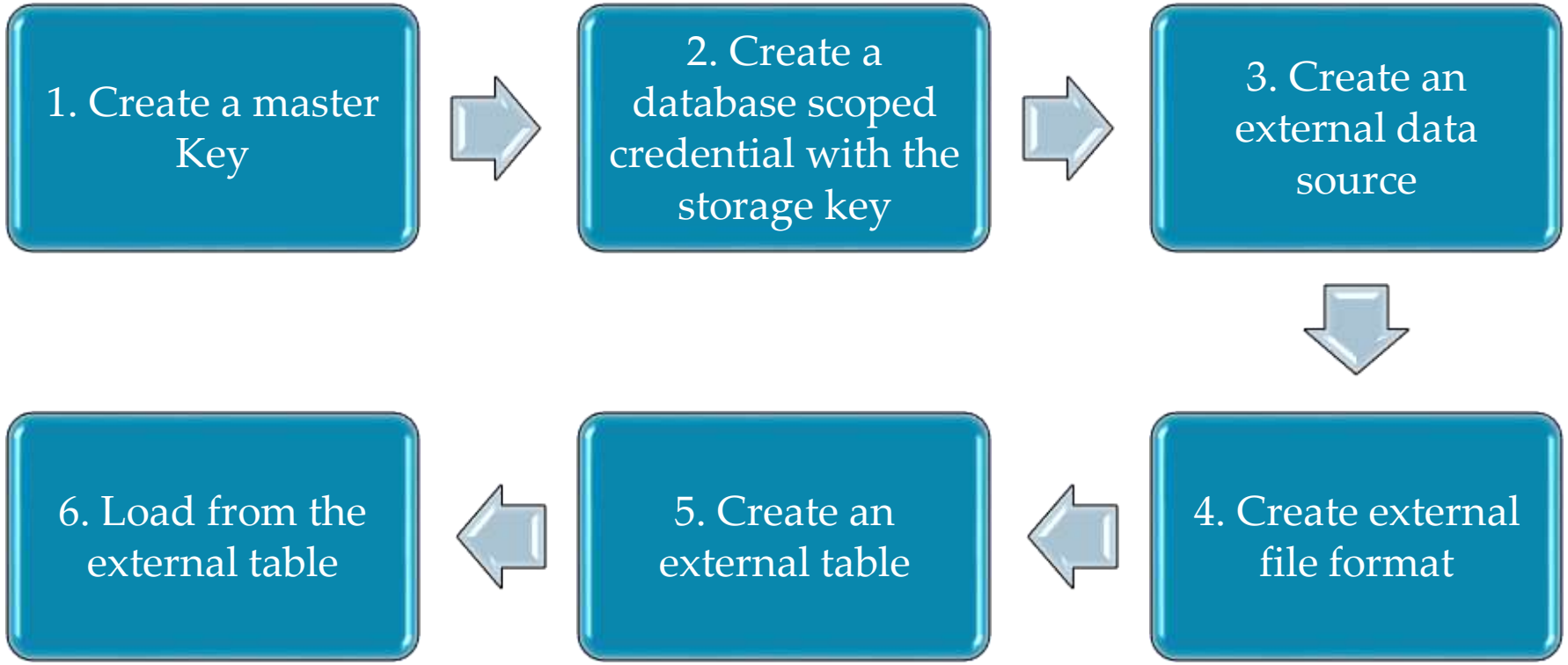


Source: Microsoft

PolyBase can load data from UTF-8 delimited text files and popular Hadoop file formats like RC File, ORC, and Parquet

Multiple readers will not work against a compressed file

PolyBase Setup



DEMO

Load Dimension table from on-premises to Azure Synapse Data warehouse using SSIS

Demo – PolyBase

1. Export table to flat file
2. Create blob storage account
3. Upload flat file to blob storage
4. Run PolyBase 6 steps process
5. Monitor and confirm successful migration
6. Confirm 60 distributions in destination table

Azure Synapse Analytics

Demo – Loading with Data Factory

1. Export table to flat file
2. Create blob storage account
3. Upload flat file to blob storage
4. Create Data Factory account
5. Create destination table
6. Create and run Data Factory pipeline to move data
7. Monitor and verify destination table

Demo – Loading with Data Factory

1. Export Fact table to flat file
2. Create Lake storage account
3. Upload flat file to Data Lake storage
4. Create Data Factory account
5. Create destination table
6. Create and run Data Factory pipeline to move data
7. Monitor and verify destination table

Demo

- Provision Azure SQL Data warehouse
- Pause/Resume compute
- Firewall rules
- Connect with SSMS and create login and user to load data
- Link to NYC taxi source data
- Create external tables
- Load and Examine data

Summary

Data Migration – Best Practices

Control node vs Compute nodes role in loading

Single client (SSIS) vs Fully Parallel (PolyBase)

Why PolyBase is preferred

PolyBase setup

Demo – using different scenarios

Azure
Synapse
Analytics