Loading Azure SQL Data Warehouse

Warner Chaves
SQL MCM / MS DATA PLATFORM MVP

@warchav sqlturbo.com

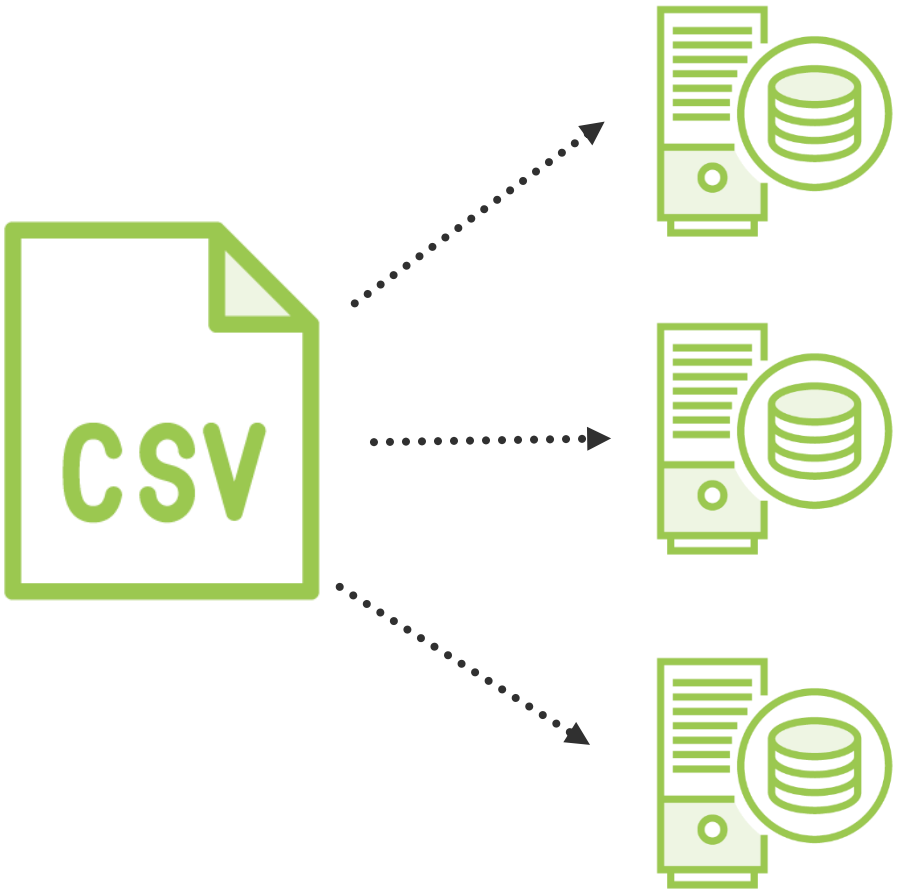
What's in this module?

Best Practices for loading data

Methods available for loading data

Azure Data Warehouse Migration utility

Loading an MPP System



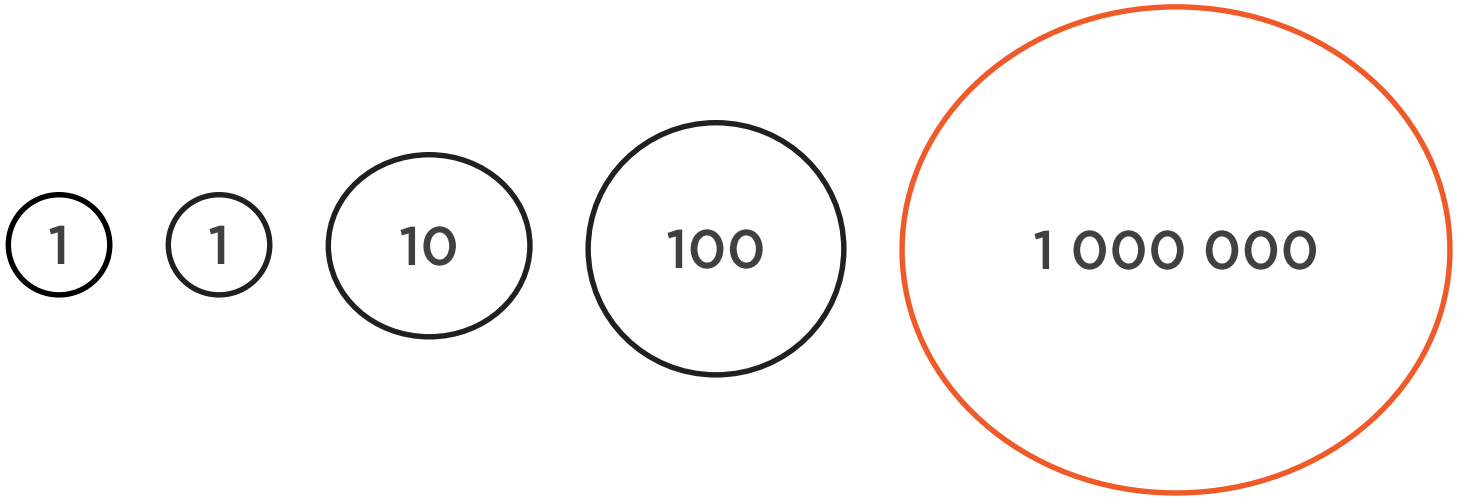
The main principle of loading data into Azure SQL Data Warehouse is to do as much work in parallel as possible.

Data Warehouse Readers

Number of:	DWU									
	100	200	300	400	500	600	1,000	1,200	1,500	2,000
Readers	8	16	24	32	40	48	60	60	60	60
Writers	60	60	60	60	60	60	60	60	60	60

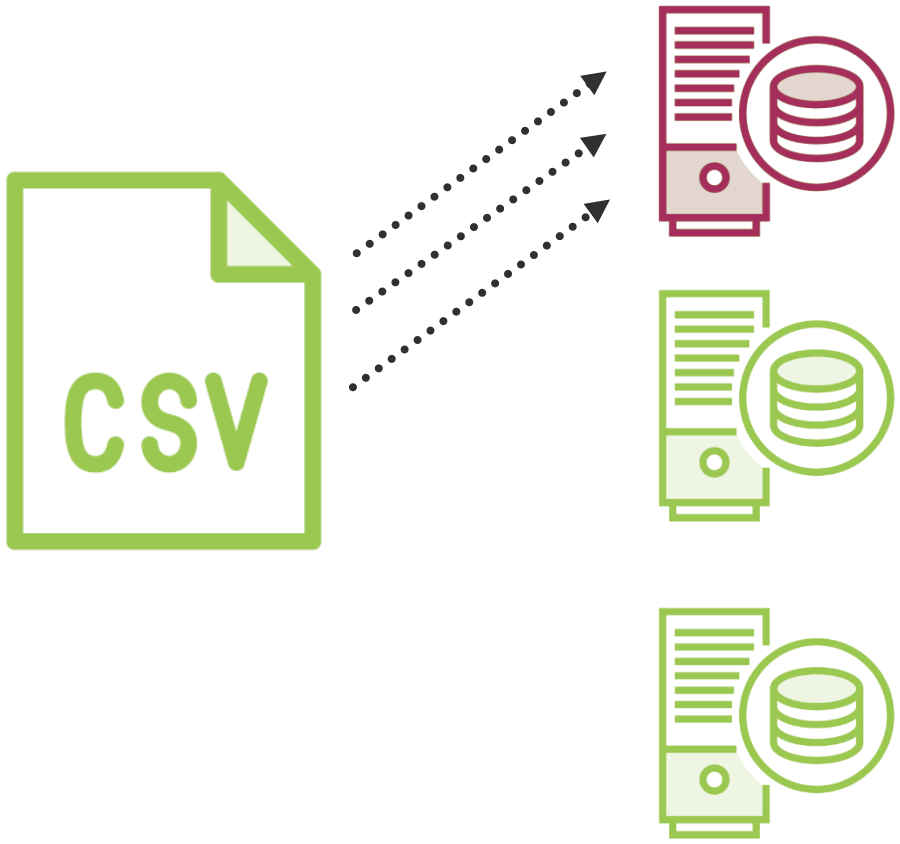
Your DWUs have a direct impact on how fast you can load data in parallel.

Optimize Insert Batch Size



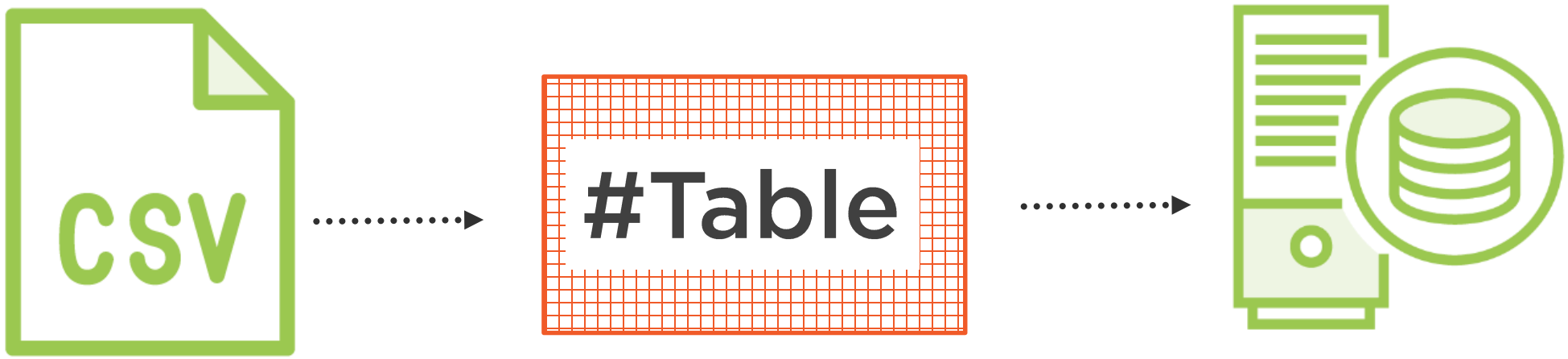
Avoid trickle insert pattern. Ideal batch size is 1 million or more direct or in a file.

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

CREATE TABLE AS

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

User Resource Class

User resource classes are database roles that govern how many resources are given to a query.

Four Resource Classes

Class	Smallrc	Mediumrc	Largerc	Xlargerc
Default	X			
Memory	100MB	100MB-1600MB	200MB-3200MB	400MB-6400MB

The lower range corresponds to DWU100, the upper range to DWU2000.

For fast and high quality loads, create a user just for loading which utilizes a medium or large resource class.

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

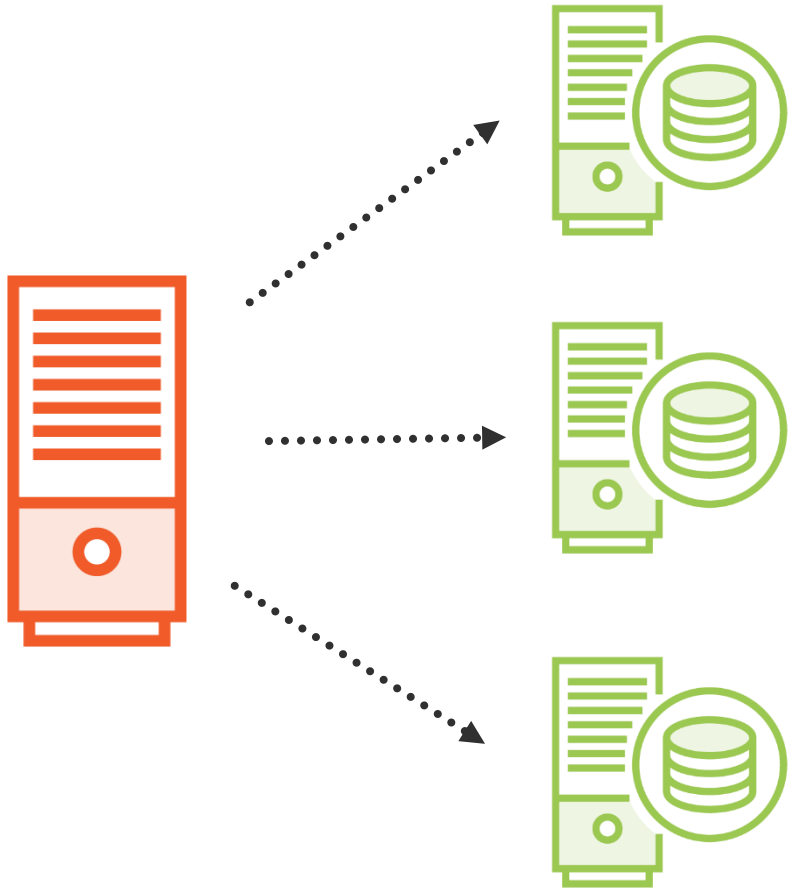
PolyBase

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.



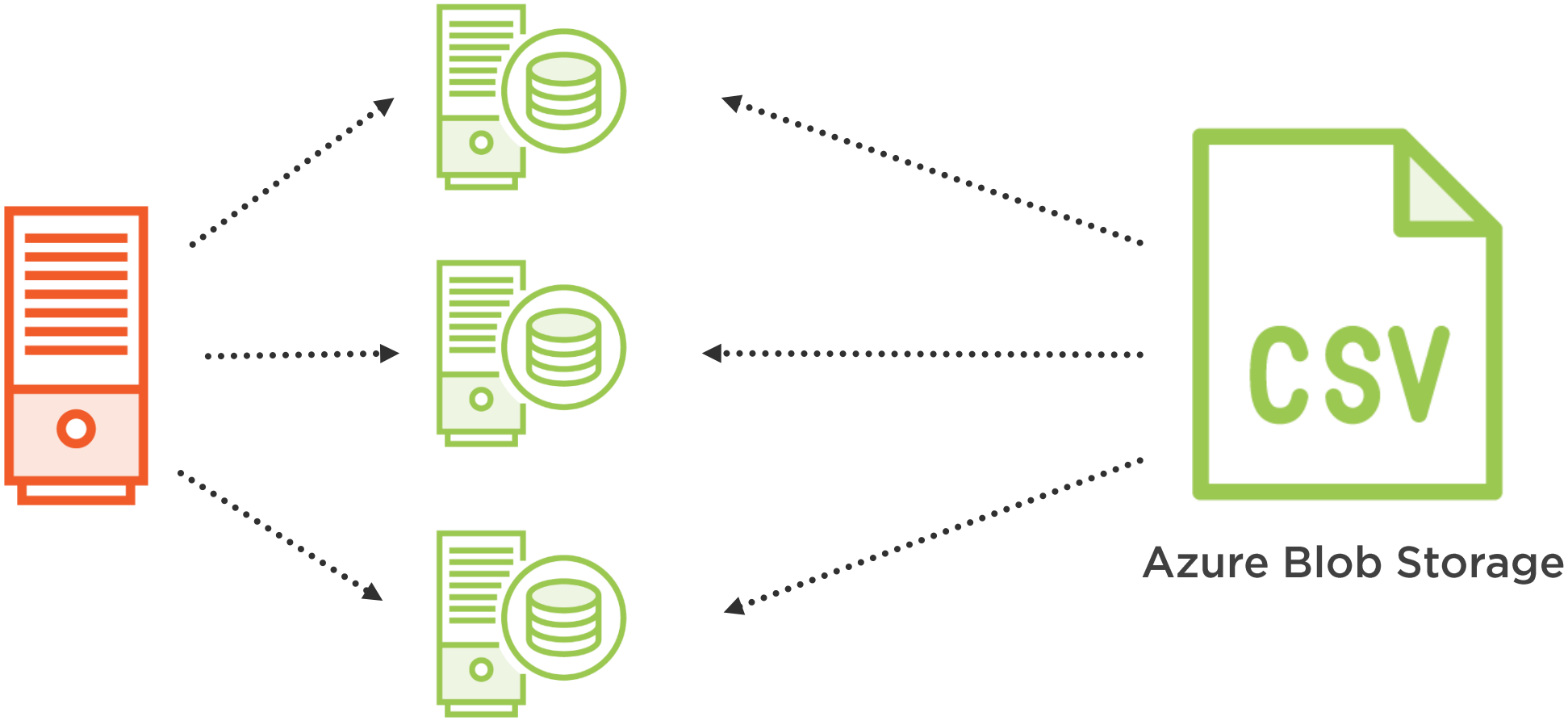
The Compute Nodes do processing on the data and scale with the DWUs.

Loading with SSIS

Demo

Loading data with SSIS

Loading with PolyBase



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

Also gzip, zlib and Snappy compressed files.

Multiple readers will not work against a compressed file.

PolyBase Setup

1. Create a master key
2. Create a database scoped credential with the storage key
3. Create an external data source
4. Create external file format
5. Create an external table
6. Load from the external table

Demo

Loading data with PolyBase

Migration Utility

1. Supports SQL Server 2012+ and Azure SQL Database
2. Provides a migration report pointing out possible issues
3. Assists with schema migration
4. Assists with data migration

Demo

Using the Azure SQL Data Warehouse migration utility

Summary

There are several Best Practices for loading data that can help you get the best loading times.

There are two node types: Control and Compute nodes.

There are 2 types of load methods: single-client and fully parallel with PolyBase.

The Azure Data Warehouse migration utility can be used to plan your migration.

Next Module:

Querying and Tuning Azure SQL Data Warehouse