Kamil Nowinski
Principal Microsoft Consultant

Microsoft® Most Valuable Professional

altius

PASS SQLSATURDAY SLOVENIA | 14 DEC 2019

Move part of your body to Azure SQL Data Warehouse

Dragon source: https://www.flickr.com/photos/wili/2628869994/in/gallery-41926029@N05-72157622307278981/

Thank you to our AWESOME sponsors!

gen-i

COMTRADE GAMING

Microsoft

Xnet
vaš zanesljiv IT partner

INSIGHT
KNOW YOUR NUMBERS

redgate
ingeniously simple

B2

softwareONE

MANNING

Kamil Nowiński

Microsoft Data Platform **MVP**
Speaker, blogger, data enthusiast

Principal Microsoft Consultant at Altius (www.altiusdata.com)
15+ yrs experience as DEV/BI/(DBA)

Member of the Data Community PL

Project member of „SCD Merge Wizard”

Founder of blog SQLPlayer (www.SQLplayer.net)

SQL Server Certificates:
MCITP, MCP, MCTS, MCSA, MCSE Data Platform,

MCSE Data Management & Analytics

Moreover: Bicycle, Running, Digital photography
@NowinskiK, @SQLPlayer

Blog

SQL Player
Play with data & have fun!

- Technical posts
- Various skill level
- Cheet sheets
- Recommended books
- Many useful other links
- Interviews (Podcast)

www.SQLPlayer.net

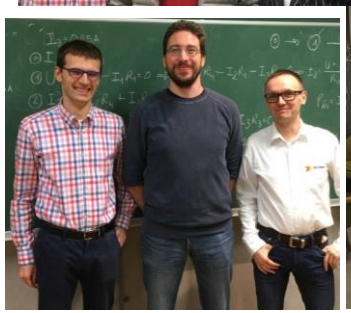
altius

@NowinskiK

PODCAST – interviews with...

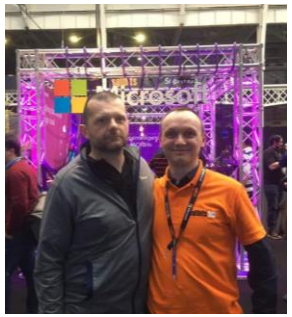
















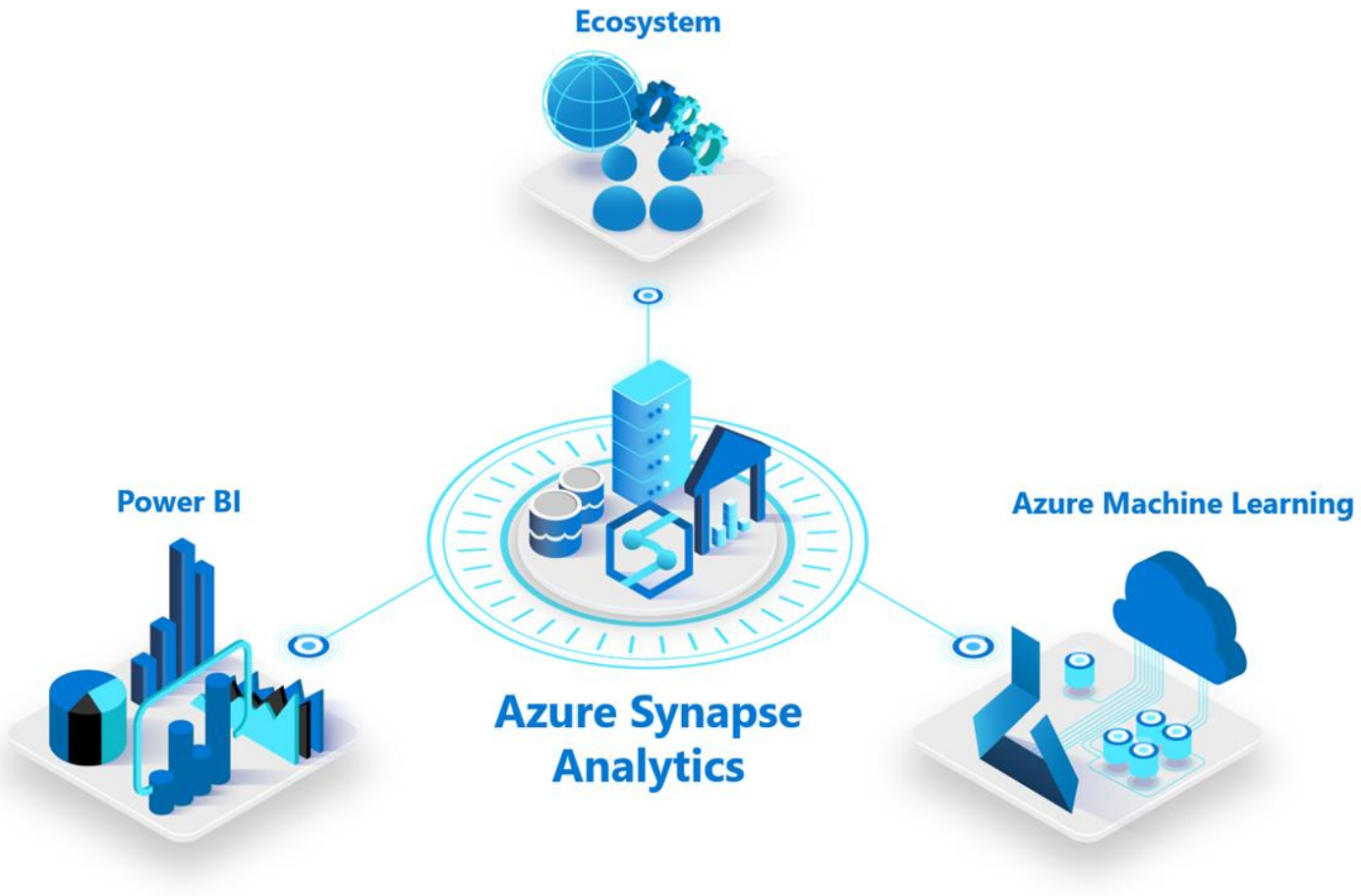
Scan me

SQL Player
Play with data & have fun!

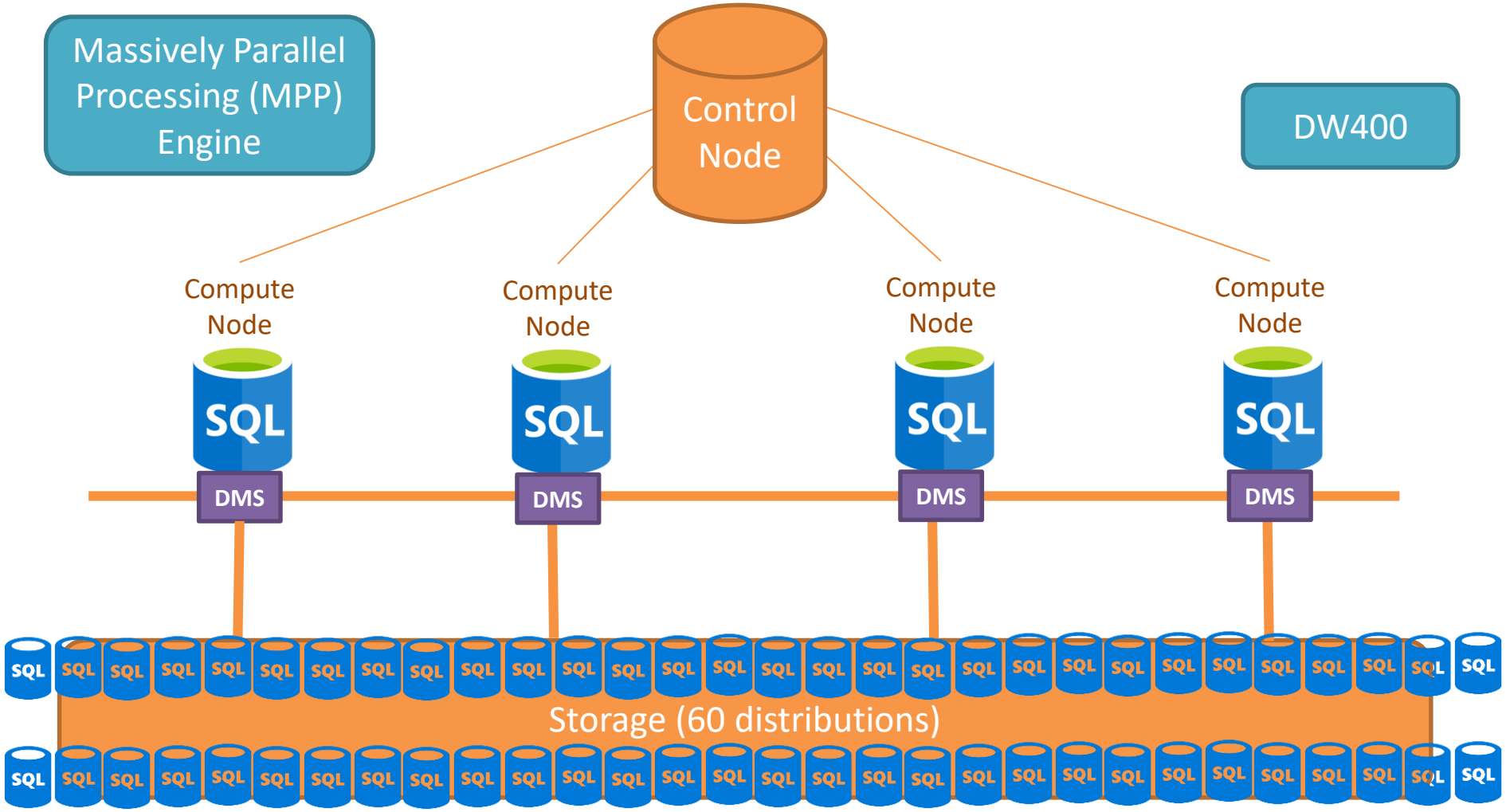
altius

@NowinskiK

Azure Synapse is Azure SQL Data Warehouse evolved



Azure SQL Data Warehouse Architecture



STORAGE

Table Distribution Options

- ROUND ROBIN
- HASH
- REPLICATED

Table Distribution Options: ROUND ROBIN

PROS:

- Default distribution
- Data distributed evenly across nodes
- East to start

CONS:

- Will incur more data movement at query time

Table Distribution Options: ROUND ROBIN

1	Poland
2	Germany
8	UK
...	
66	Switzerland
70	Ireland

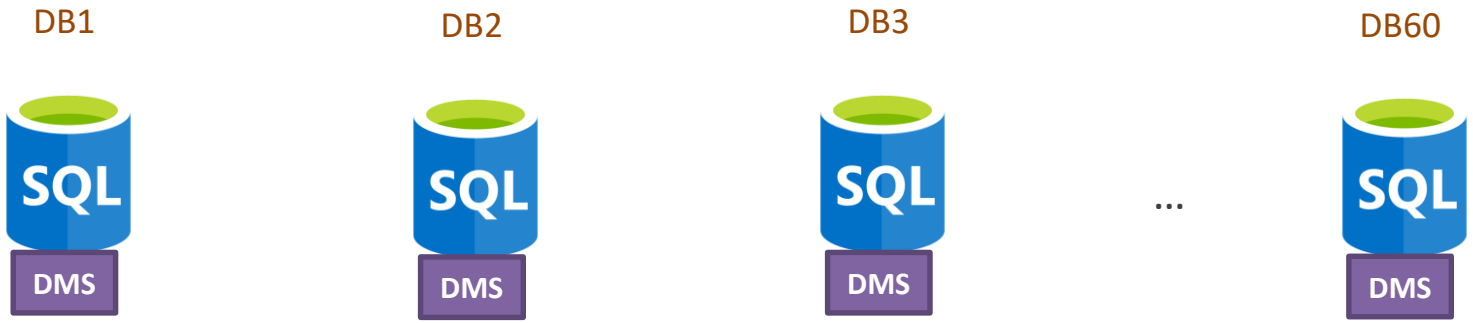


Table Distribution Options: HASH

PROS:

- Data divided across nodes based on hashing algorithm
- Same value produces the same hash value
- Single column only

CONS:

- Check for Data Skew, NULLs, -1, etc.

Table Distribution Options: HASH

1	Poland
2	Germany
8	UK
...	
66	Switzerland
70	Ireland

DB1

DB2

DB3

...

DB60

Table Distribution Options: REPLICATED

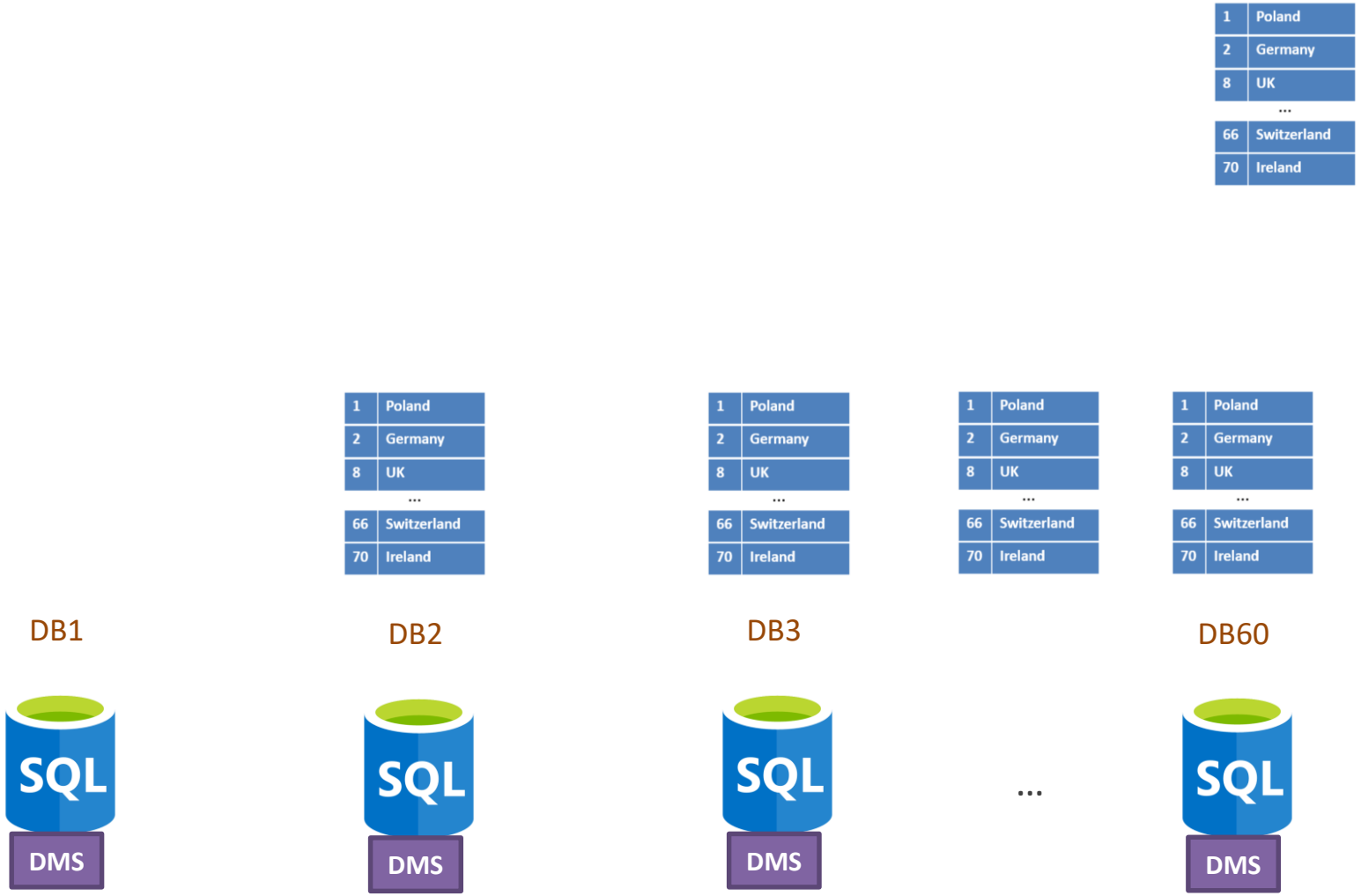
PROS:

- Data repeated on every node
- Simplifies many query plans and reduces data movement
- Best with joining hash table

CONS:

- Consume more space
- Joining two replicated tables runs on one node

Table Distribution Options: REPLICATED



Execution Plan – DMS Operations

DMS Operation	Description
ShuffleMoveOperation	Distribution → Hash algorithm → New distribution Changing the distribution column in preparation for join.
PartitionMoveOperation	Distribution → Control Node Aggregations - count(*) is count on nodes, sum of count
BroadcastMoveOperation	Distribution → Copy to all distributions Changes distributed table to replicated table for join.
TrimMoveOperation	Replicated table → Hash algorithm → Distribution When a replicated table needs to become distributed. Needed for outer joins.
MoveOperation	Control Node → Copy to all distributions Data moved from Control Node back to Compute Nodes resulting in a replicated table for further processing.
RoundRobinMoveOperation **HadoopRoundRobinMoveOperation**	Source → Round robin algorithm → Distribution Redistributes data to Round Robin Table.

Statistics

Statistics

- One or more columns of a table
- Indexed view
- External table

- Cost based Query Optimizer
- Candidate columns when used in:
  - JOIN
  - GROUP BY
  - WHERE
- Update statistics after incremental load
- Use multi-column statistics if needed

Important things

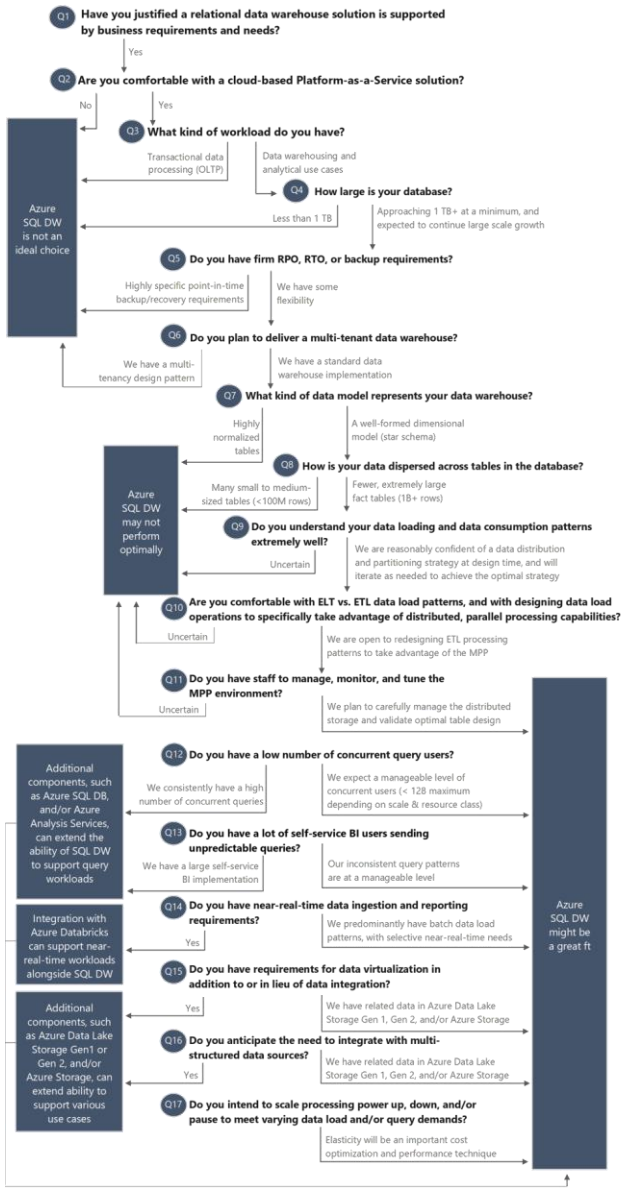
- SQL DW is based on an MPP architecture (not SMP)
  - The same engine under hood, but scale and concurrency are vary
- SIZE does really matter
- Individual table size and rowcount are important
- OLTP reporting type workloads are usually poor candidates
- Proper schema design – **important** in SQL Server
- Right schema desing – **CRITICAL** in SQL DW

Data Distribution

DEMO

Is Azure SQL Data Warehouse a good fit?

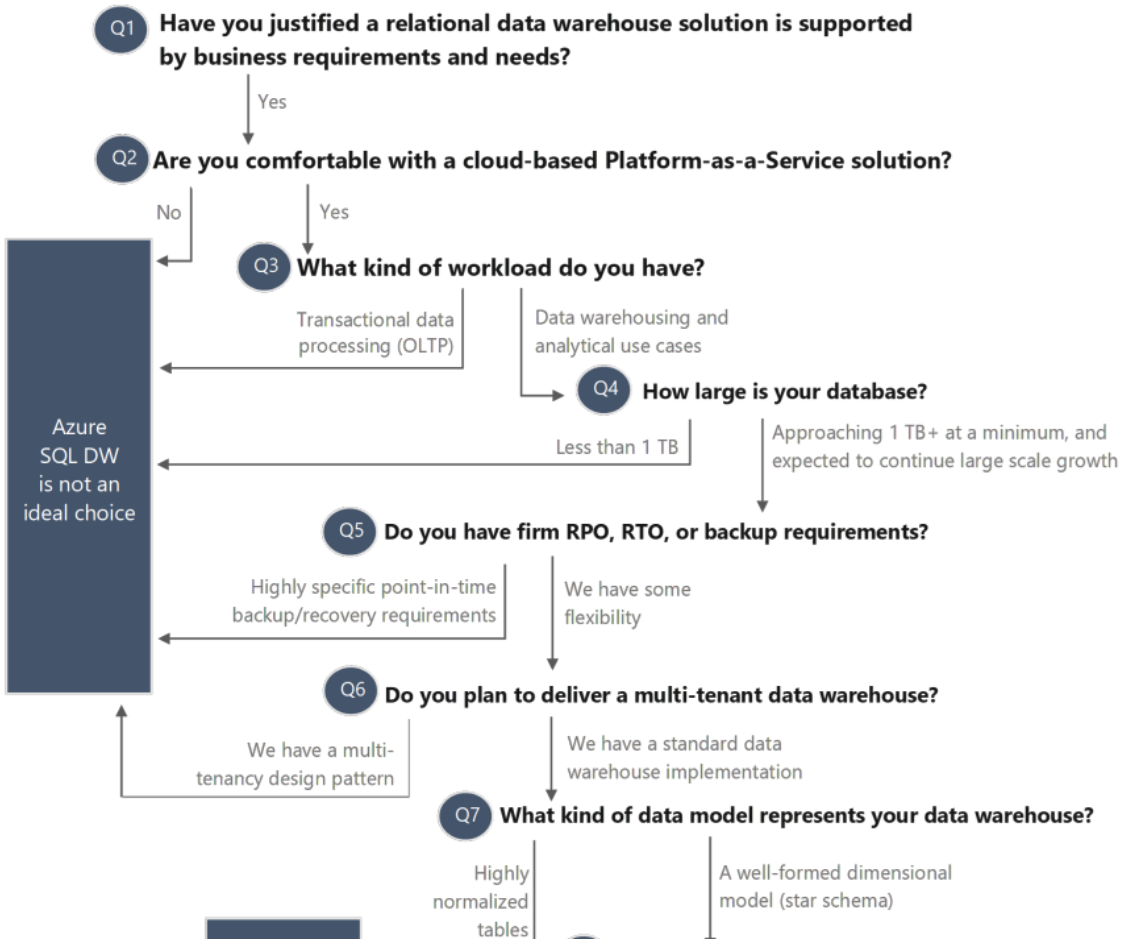
- Verify your source in many aspects
- Do answer for many questions
- Use form from more experienced
- Questions' diagram
- Ask **Melissa Coates**



https://www.blue-granite.com/blog/is-azure-sql-data-warehouse-a-good-fit-updated

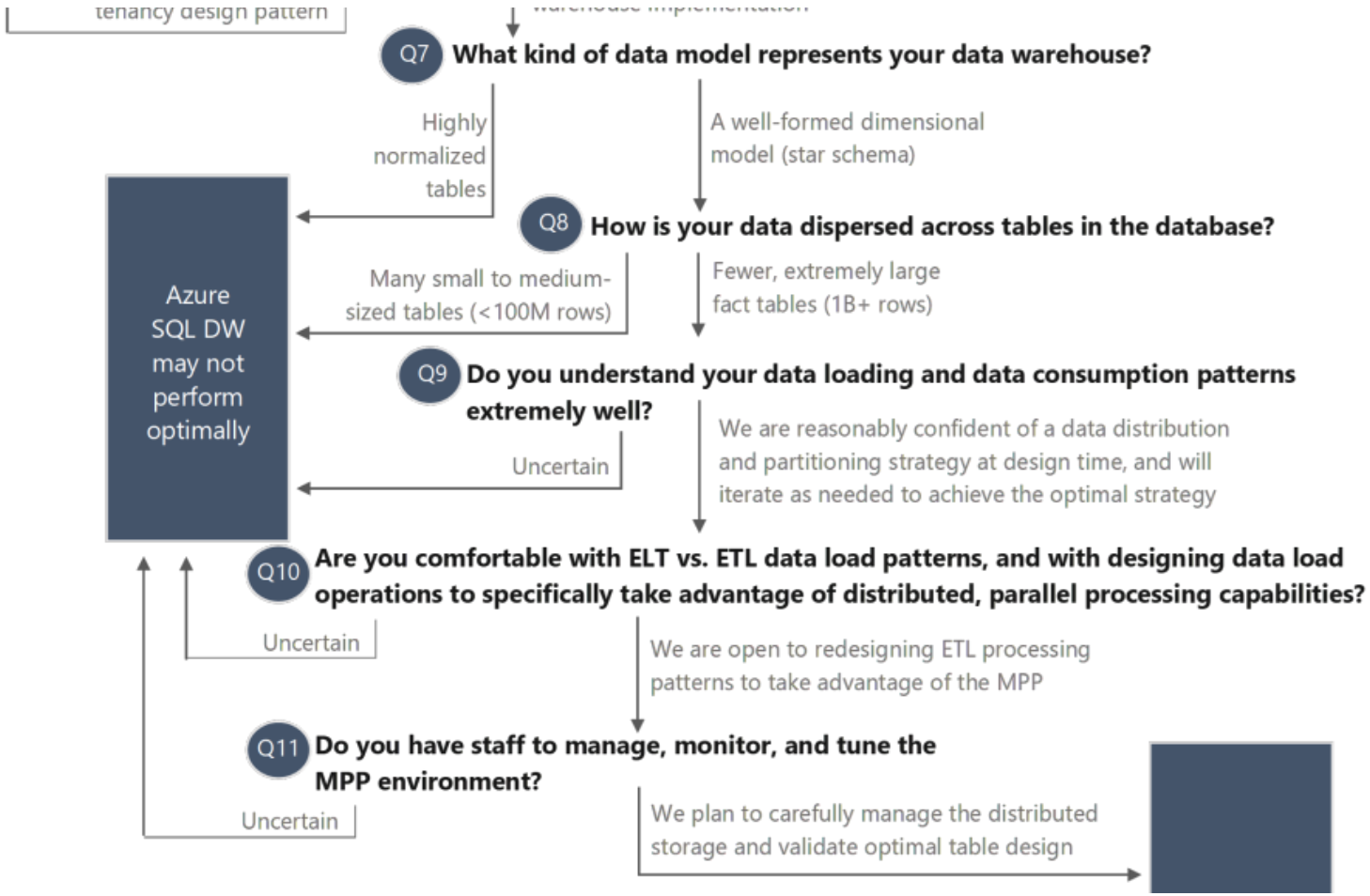
Is Azure SQL Data Warehouse a good fit? technology choice for your implementation?

- Q3
  - OLTP?
  - DW / Analytical workload?
- Q4
  - <1 TB?
  - >1 TB
- Q6
  - Multitenant?
  - Standard implementation



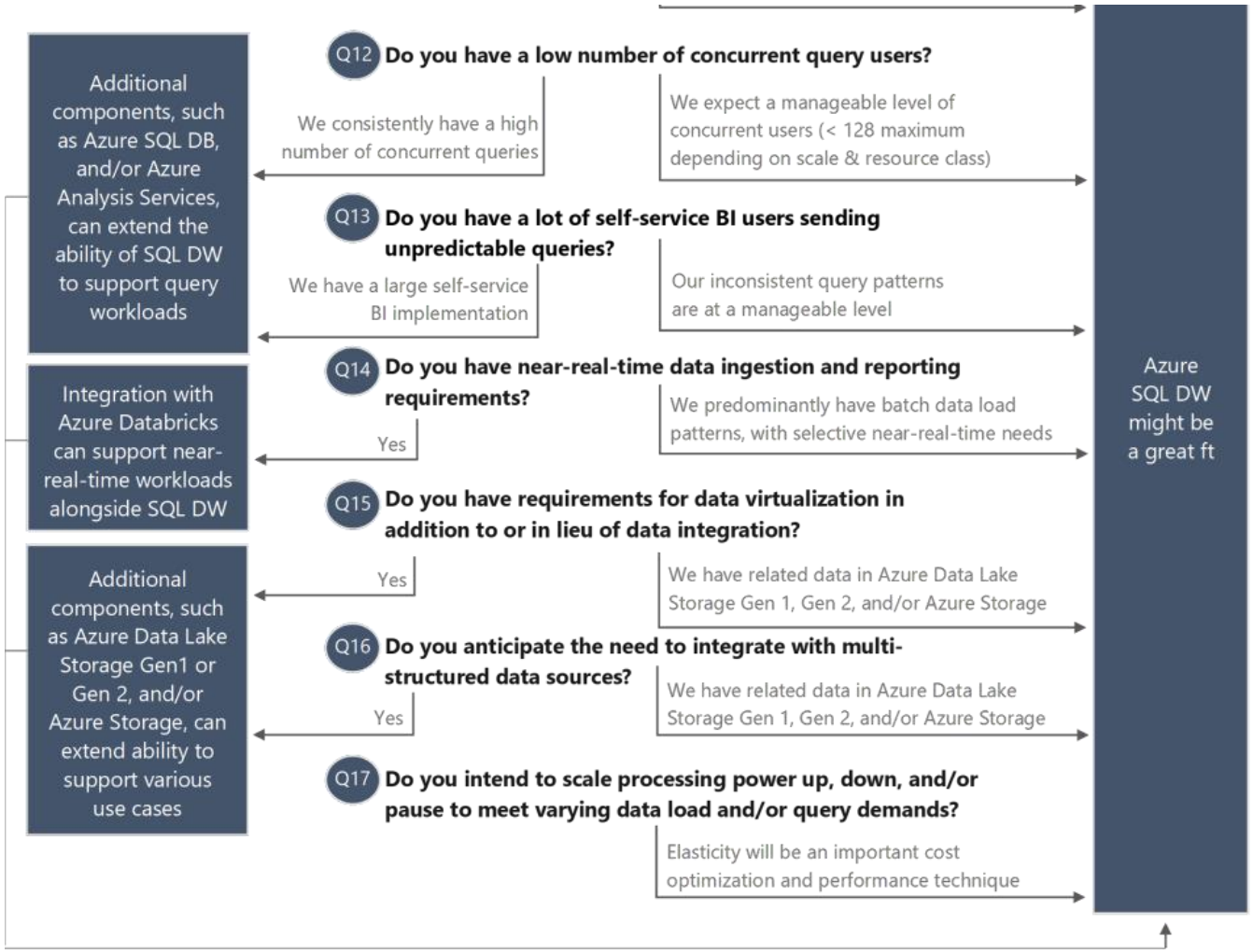
Is Azure SQL Data Warehouse the best technology choice for your implementation?

- Q7
  - Highly normalized tables?
  - Well-formed / star schema
- Q8: Number of tables & rows?
  - Many small/medium tables?
  - Fewer / large fact tables (1B+ rows)
- Q11: Skilled stuff



Is Azure SQL Data Warehouse the best technology choice for your implementation?

- Q12: Concurrent queries
  - High number
  - <128 active sessions
- Q14: Frequency of ingestion?
  - Near-real-time



PREPARATION & COPY

Data Preparation: files

- Filter essential objects to migrate
- Create performant local storage to receive exported data
- Establish standard or dedicated connectivity to cloud
- Choose region nearest to you with Azure SQL DW
- PolyBase: One folder per table in storage container

Data Migration Recommendations

- Use Migration Tool
- Understand current T-SQL surface area and workarounds
- Avoid Singelton DML operations (INSERT, UPDATE, DELETE)
  - Batch DML if possible
  - If unavoidable, wrap in transaction (BEGIN TRAN ... COMMIT)
- Use heap table OR temp table for staging data
- Avoid large fully logged operations
  - Considers CTAS as this is minimal logged operation
  - Process by partition to leverage parallelism and partition switching
- Design retry logic to address service disruption

Data Migration Recommendations

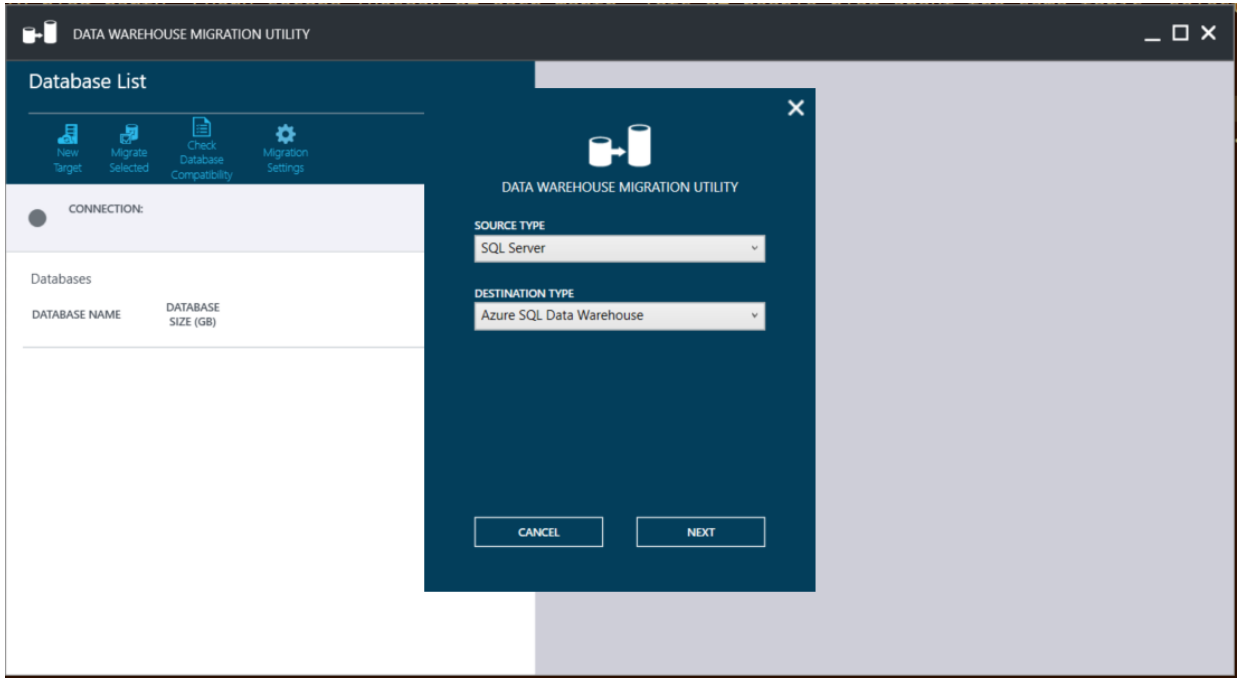
- Data Format Conversion
  - Data Format, Field delimiters, Escaping, Field order, encoding
- Compression
  - Use Gzip, ORC, parquet
- Export
  - BCP for fast export
  - Multiple files per large table, one folder per table
- Copy
  - AZCopy
  - Data Movement Library

Data Migration Tips

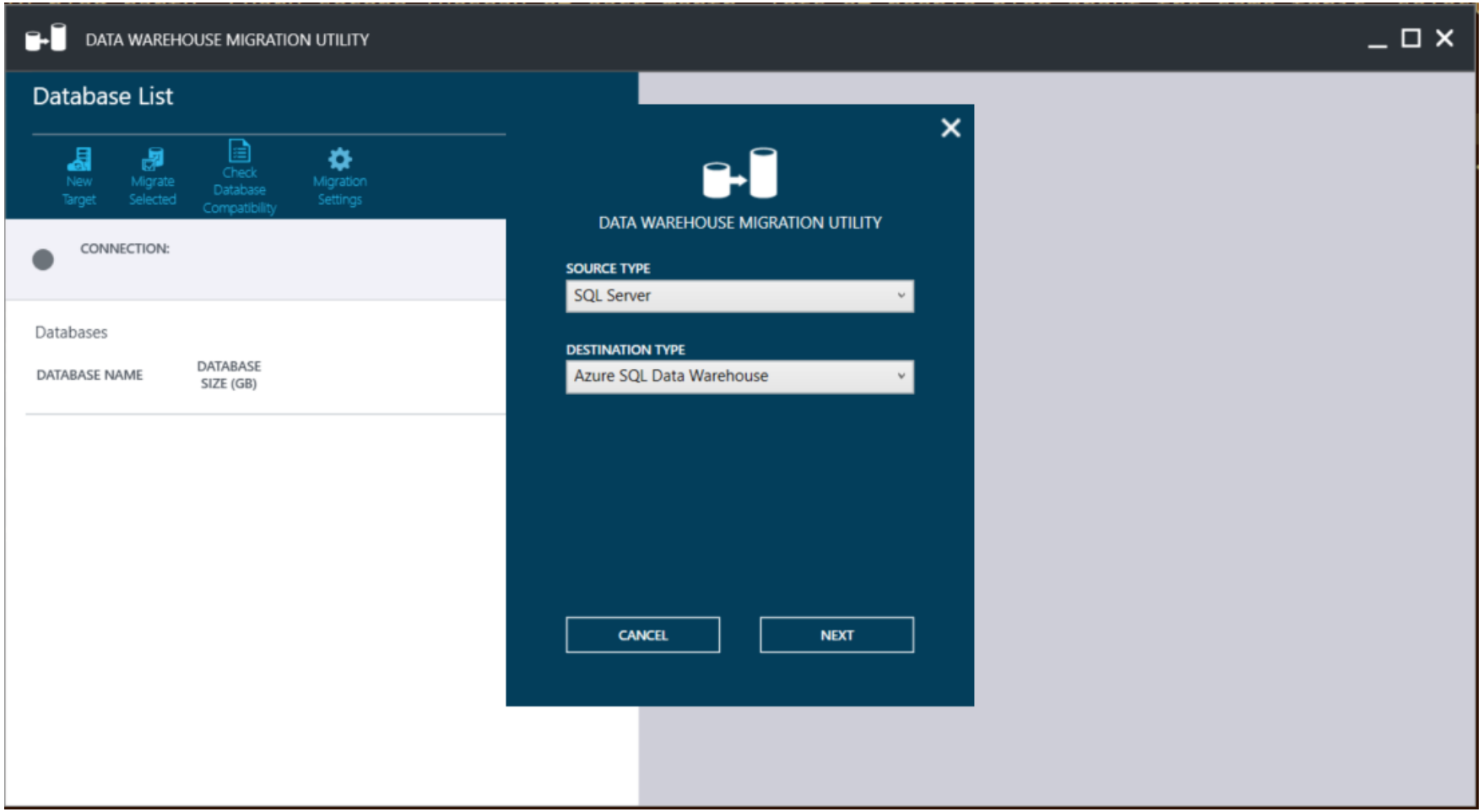
- Incorrect format means migration needs to be entirely repeated
- Exploit bcp options, hints, parralellism
- Multiple compressed files, split files
- Parallel import, reliable transfer
- Don't use multiple files in the same gzipped file
- Efficient Copy
  - Parallel, Async, Resumable
  - Limit concurrent copies if low bandwith
- Very large Data transfer
  - Express Route, Import/Export Service

Data Migration (WWI)

DEMO



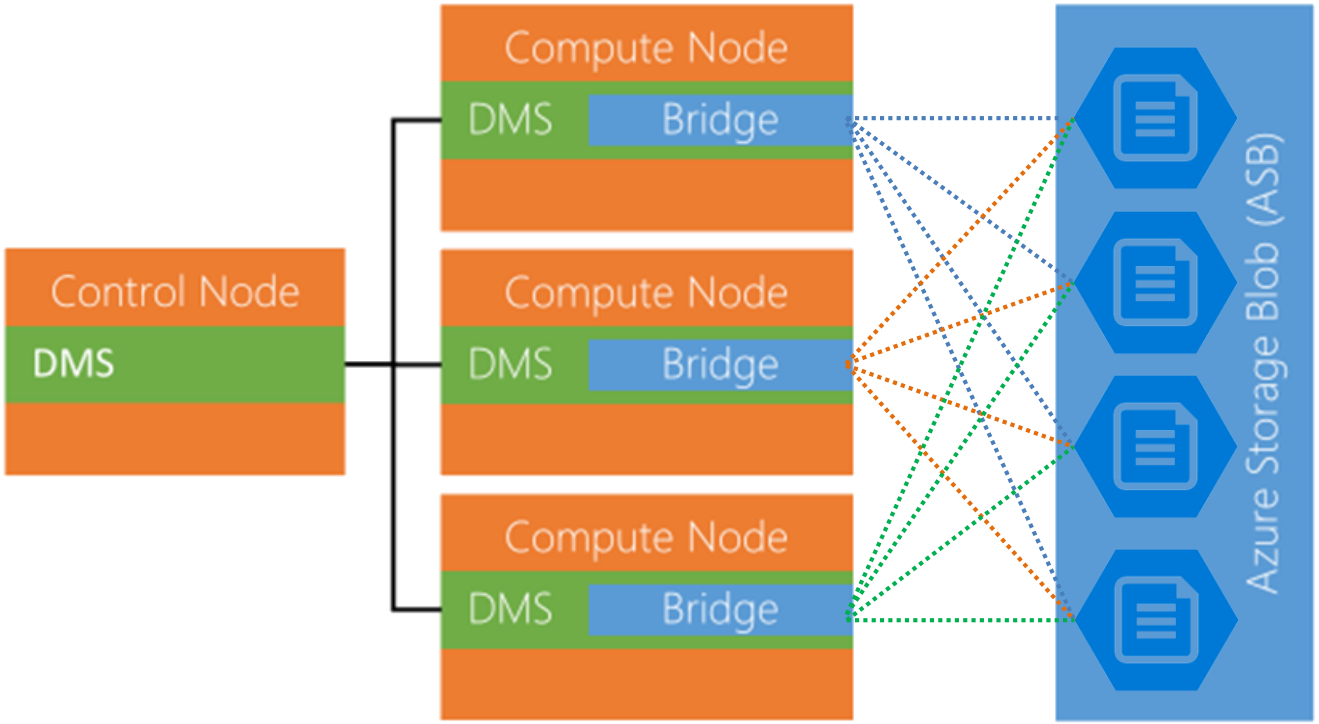
Data Warehouse Migration Utility (Preview)



Data Loading Recommendations

- PolyBase and SSIS (with 2017 Azure feature pack) the fastest method
  - Upload to BLOB via AZCOPY or PowerShell library
  - Historical load – use CTAS
  - Incremental – use INSERT...SELECT
  - UTF-8, UTF-16 also supports
- Use the highest resource class (without sacrificing concurrency)
- Increase DWU before load, decrease once done
- ADLS supported
- Doesn’t support:
  - Extended ASCII
  - Custom multi-date format

Parallel Loading with PolyBase



Data Loading Options

	PolyBase	SSIS *	ADF	BCP	SqlBulk Copy
Rate	Fastest				Slowest
Rate increase as DWU increases	Yes	Yes	Yes	No	No
Rate increases as you add concurrent load	No	No	No	Yes	Yes

* With SSMS Azure Feature Pack June 2017 (or newer)

PolyBase characteristics

- Single PolyBase load provides best performance for non-compressed files
- Load performance scales as you increase service level objective (SLO)
  - Number of files should be greater than of equal to the total number of readers of your service level objective (SLO)
- Automatically parallelizes data load process;
  - no need to manually break the input data into multiple files and issue concurrent loads
  - Each reader slice 512 MB block from data files
- Max throughput depends on number of readers available on the DWU level
- Multiple readers will not work against a compressed text file (gzip)
  - Only a single reader is used per compressed file since uncompressing the file in the buffer is single threaded
  - Alternatively, generate multiple compressed files

Parallel Loading with PolyBase

DEMO

Resources

- Azure SQL Data Warehouse -> Azure Synapse Analytics
- YouTube – sessions, webinars
- Seven Key Principles of Cloud Security and Privacy (white paper)
- And finally:
- SQLPlayer.net blog

Questions?

Thank you!

kamil@nowinski.net

@NowinskiK @SQLPlayer

SQLPlayer.net

https://github.com/NowinskiK/CommunityEvents

Kamil Nowinski
Microsoft Data Platform MVP
MCSE Data Platform & MCSE Data Management and Analytics