GOLD SPONSORS

elitmind
think bright

TECHNOLOGY
INNOVATION
DATA
KNOWLEDGE
tidk

VOLVO
Volvo Group IT

SILVER SPONSORS

dbWatch
DATABASE CONTROL

it KONTRAKT®

UBS

veeAM

BRONZE SPONSOR

KSIĄŻKI IT Z CAŁEGO ŚWIATA
NOVATECH
www.novatech.com.pl

STRATEGIC PARTNER

Microsoft

Move part of your body to Azure Data Warehouse

Kamil Nowiński

About me

Kamil Nowinski

Data Engineer at ASOS (www.asos.com)
13+ yrs experience as DEV/DBA

The Chairman of the Audit Committee of 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

BRAND NEW BLOG

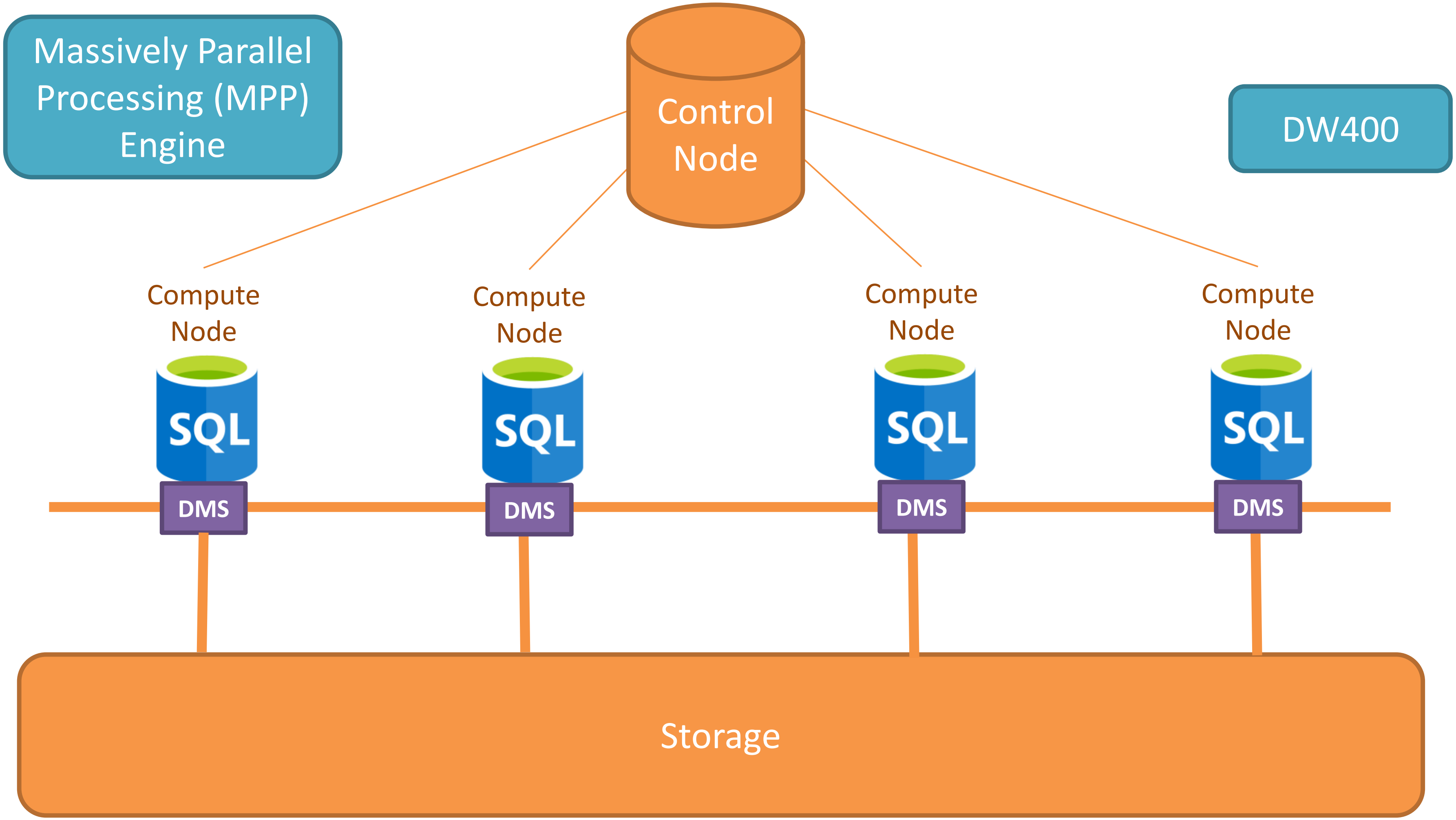
INTERVIEWS

SQLPlayer

ay with data & have fun!

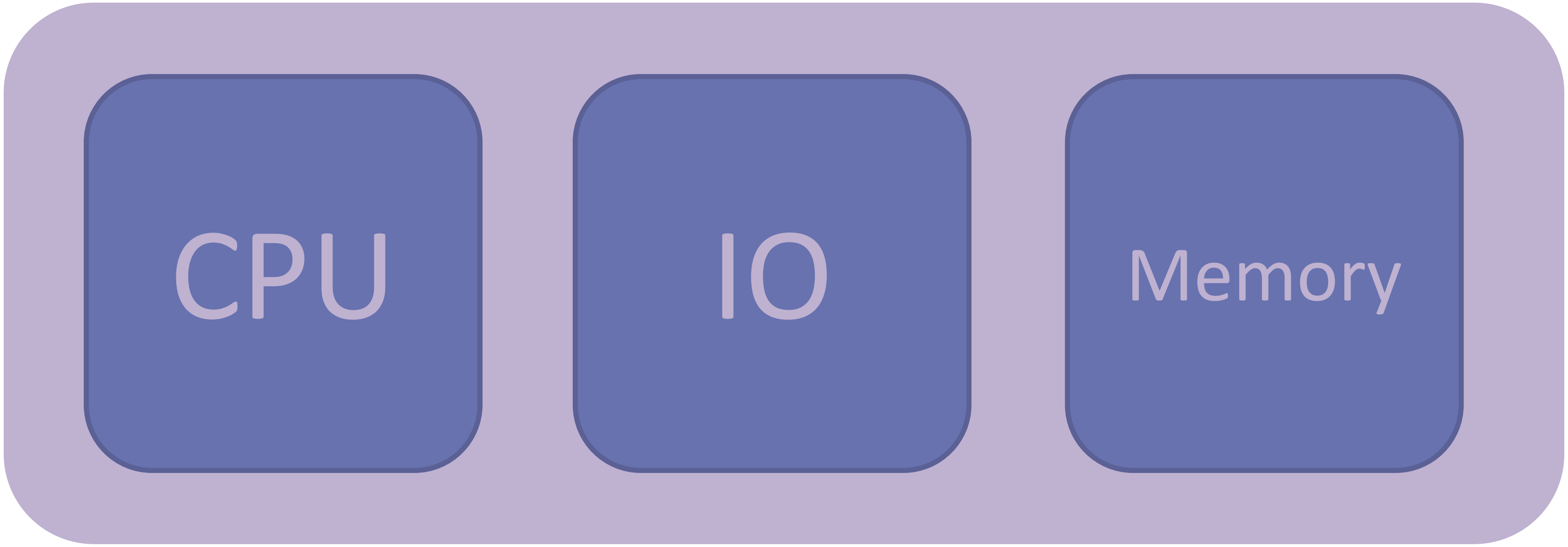
www.SQLPlayer.net

Azure SQL Data Warehouse Architecture



DWUs & cDWUs

- **DWU** – Data Warehouse Units
- **cDWU** – compute Data Warehouse Units
- Normalized amount of compute
- Converts to billing units i.e. what you pay



CPU

IO

Memory

DWUs & cDWUs

	DWU (Gen1)	cDWU (Gen2)
The optimized for	**Elasticity** performance tier	**Compute** performance tier
Support scaling compute up/down	YES	YES
Disk-based cache	NO	YES

Table Distribution Options: ROUND ROBIN

1	Poland
2	Germany
8	UK
…	
66	Switzerland
70	Ireland

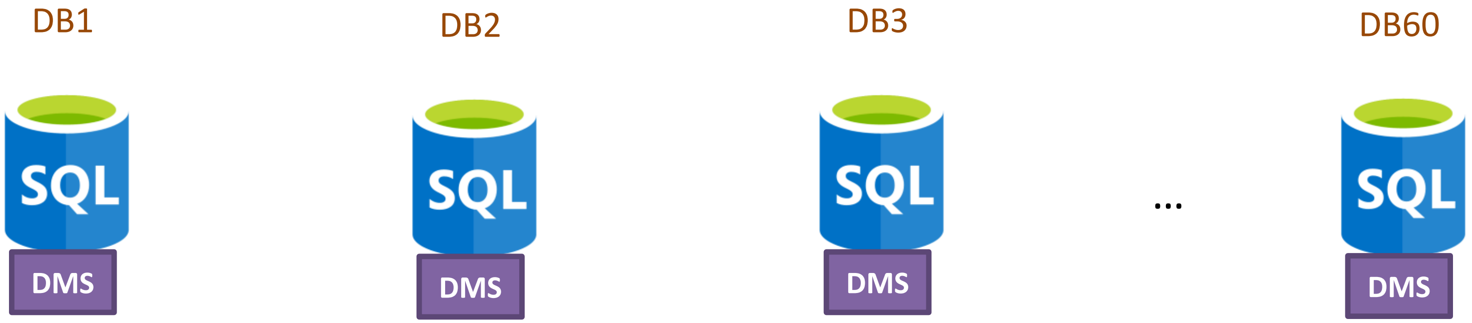


Table Distribution Options: ROUND ROBIN

PROS:

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

CONS:

- Will incur more data movement at query time

Table Distribution Options: HASH

1	Poland
2	Germany
8	UK
…	
66	Switzerland
70	Ireland

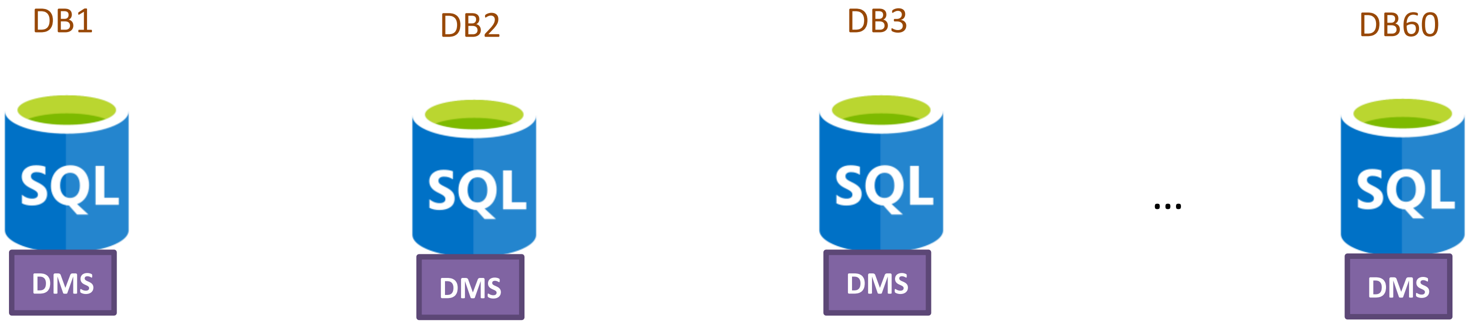


Table Distribution Options: HASH

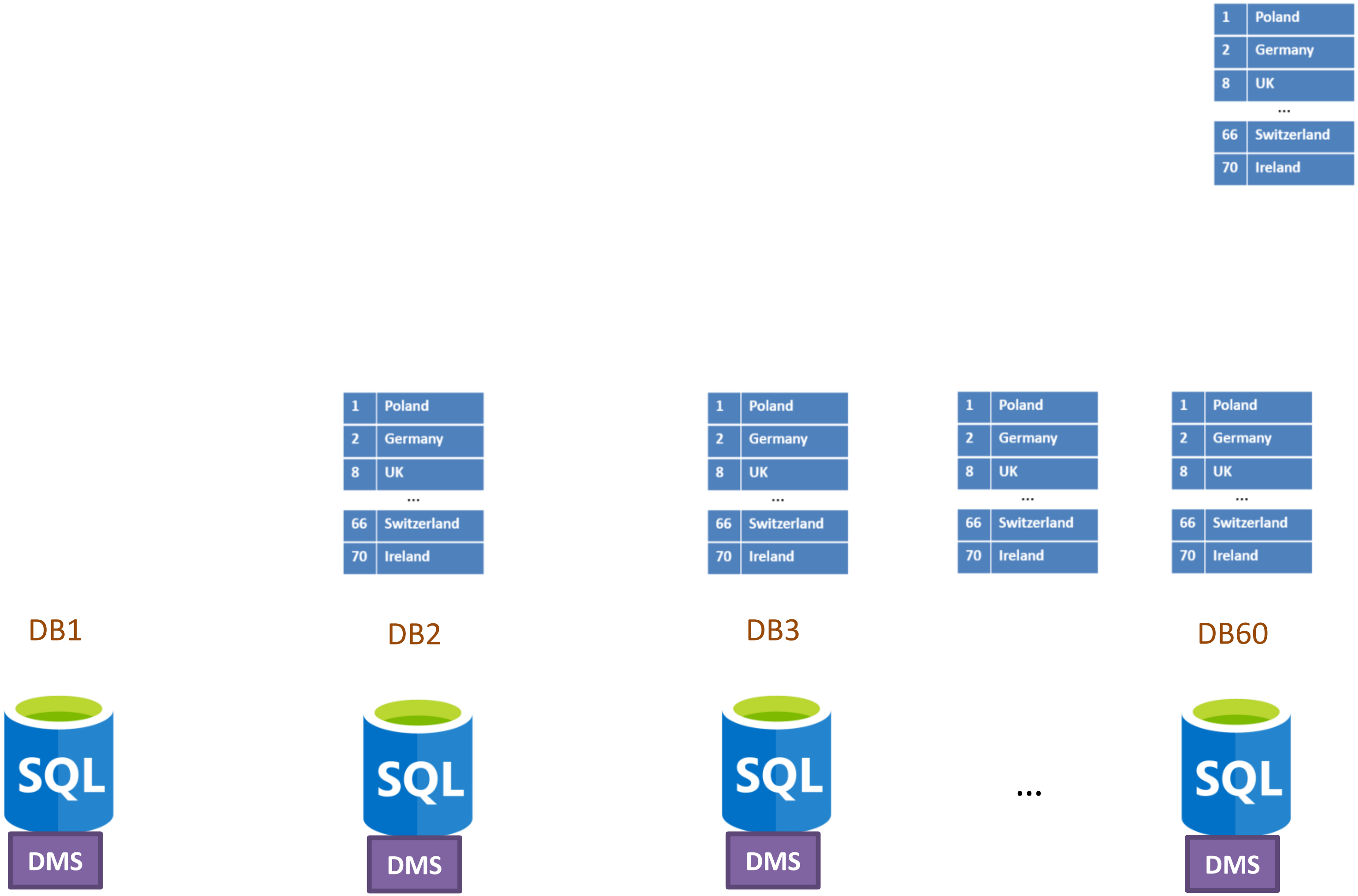
PROS:

- 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: REPLICATED



1	Poland
2	Germany
8	UK
	…
66	Switzerland
70	Ireland

1	Poland
2	Germany
8	UK
	…
66	Switzerland
70	Ireland

1	Poland
2	Germany
8	UK
	…
66	Switzerland
70	Ireland

1	Poland
2	Germany
8	UK
	…
66	Switzerland
70	Ireland

1	Poland
2	Germany
8	UK
	…
66	Switzerland
70	Ireland

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

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

- 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

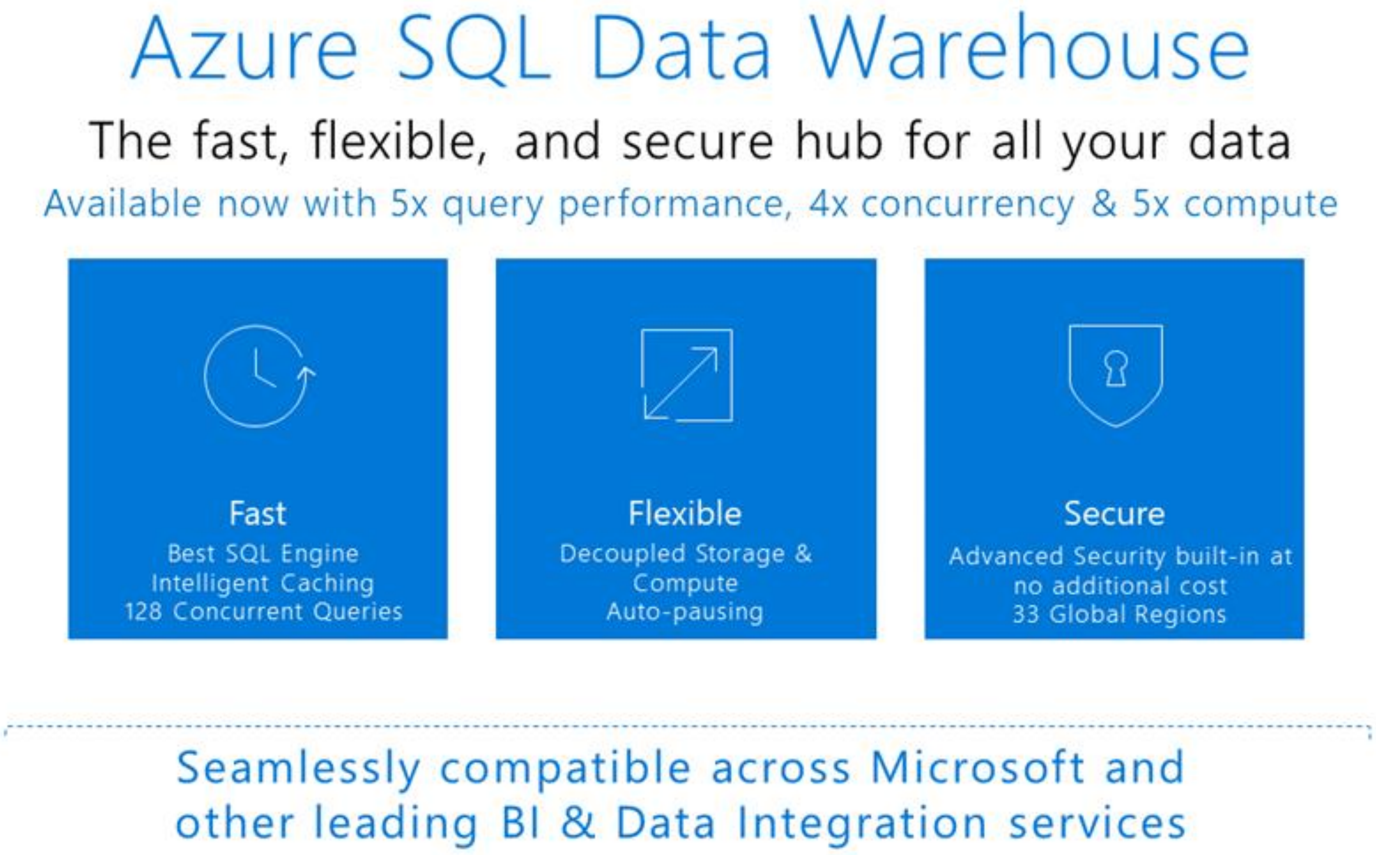
Data Distribution

DEMO

SQL Azure Data Warehouse

GEN 2

Fast, flexible, and secure cloud data warehouse



Azure SQL Data Warehouse

The fast, flexible, and secure hub for all your data

Available now with 5x query performance, 4x concurrency & 5x compute

Fast

Best SQL Engine
Intelligent Caching
128 Concurrent Queries

Flexible

Decoupled Storage & Compute
Auto-pausing

Secure

Advanced Security built-in at no additional cost
33 Global Regions

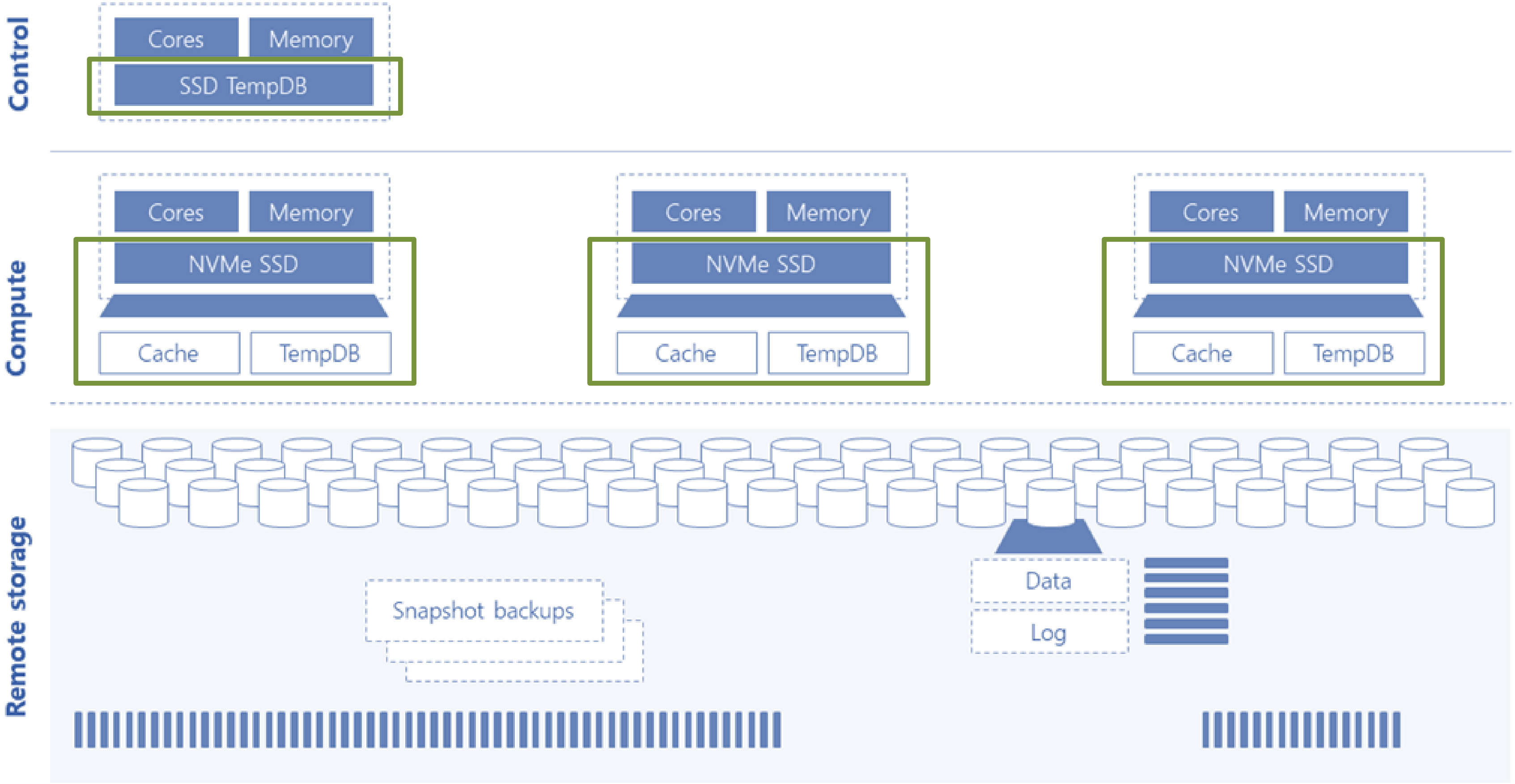
Seamlessly compatible across Microsoft and other leading BI & Data Integration services

Compute Optimized Gen2 Tier

- **Generally available** since 30/04/2018
- Expanded to **33** Azure regions
- Hardware innovations behind the scenes
- NVM Express (**NVMe**) solid-state drive (SSD)
- Generally offers up to **2GB/sec** of local I/O bandwidth
- **Adaptive caching** of recently used data on NVMe

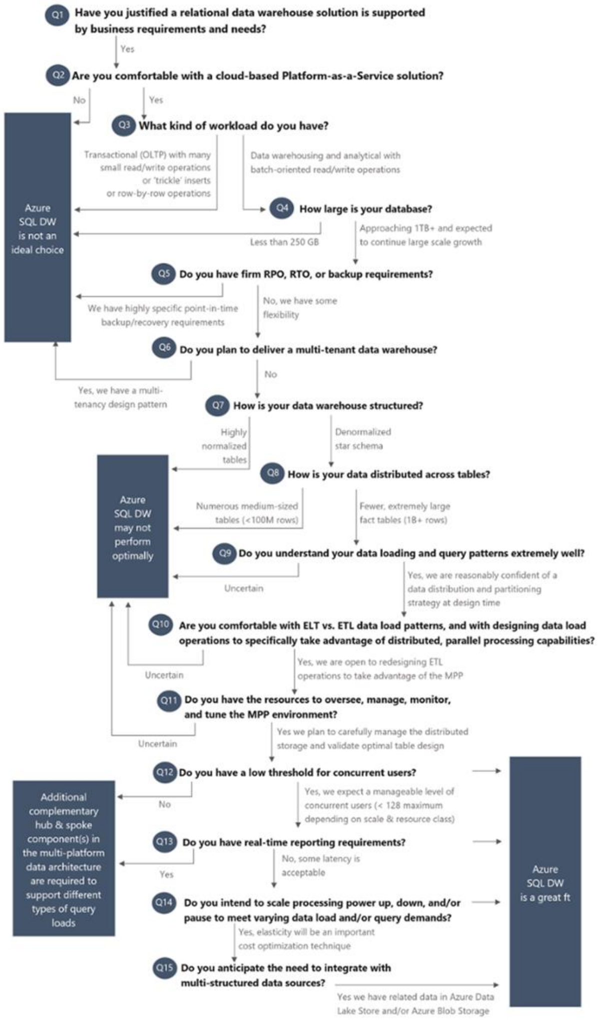
Compute Optimized Gen2 Tier

SQL DW Gen2 redefines performance with intelligent caching



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**



Q1 Have you justified a relational data warehouse solution is supported by business requirements and needs?

Yes

Q2 Are you comfortable with a cloud-based Platform-as-a-Service solution?

No

Yes

Azure SQL DW is not an ideal choice

Q3 What kind of workload do you have?

Transactional (OLTP) with many small read/write operations or 'trickle' inserts or row-by-row operations

Data warehousing and analytical with batch-oriented read/write operations

Q4 How large is your database?

Less than 250 GB

Approaching 1TB+ and expected to continue large scale growth

Q5 Do you have firm RPO, RTO, or backup requirements?

We have highly specific point-in-time backup/recovery requirements

No, we have some flexibility

Q6 Do you plan to deliver a multi-tenant data warehouse?

Yes, we have a multi-tenancy design pattern

No

Q7 How is your data warehouse structured?

Highly normalized tables

Denormalized star schema

Azure SQL DW may not perform optimally

Q8 How is your data distributed across tables?

Numerous medium-sized tables (<100M rows)

Fewer, extremely large fact tables (1B+ rows)

Q9 Do you understand your data loading and query patterns extremely well?

Uncertain

Yes, we are reasonably confident of a data distribution and partitioning strategy at design time

Q10 Are you comfortable with ELT vs. ETL data load patterns, and with designing data load operations to specifically take advantage of distributed, parallel processing capabilities?

Uncertain

Yes, we are open to redesigning ETL operations to take advantage of the MPP

Q11 Do you have the resources to oversee, manage, monitor, and tune the MPP environment?

Uncertain

Yes we plan to carefully manage the distributed storage and validate optimal table design

Q12 Do you have a low threshold for concurrent users?

No

Yes, we expect a manageable level of concurrent users (< 128 maximum depending on scale & resource class)

Additional complementary hub & spoke component(s) in the multi-platform data architecture are required to support different types of query loads

Q13 Do you have real-time reporting requirements?

Yes

No, some latency is acceptable

Q14 Do you intend to scale processing power up, down, and/or pause to meet varying data load and/or query demands?

Yes, elasticity will be an important cost optimization technique

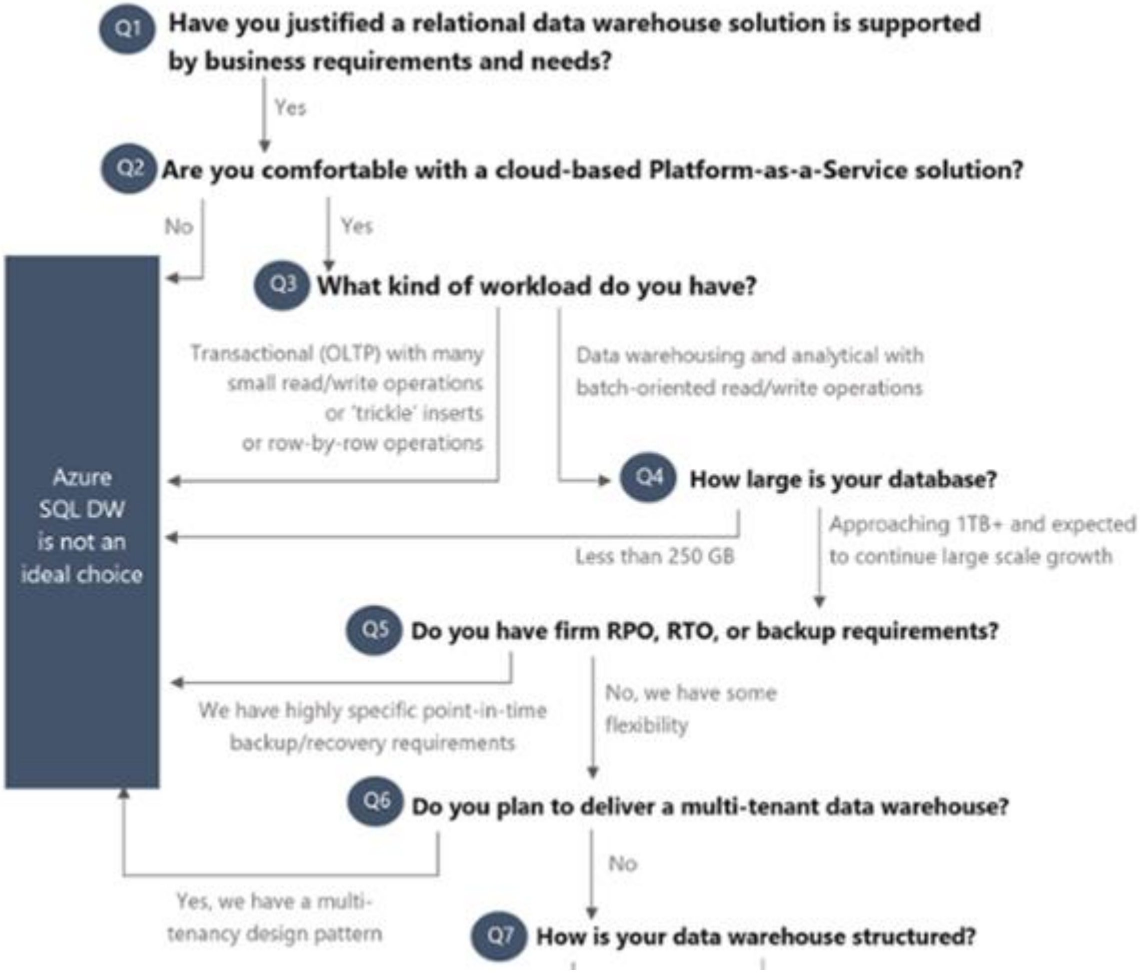
Q15 Do you anticipate the need to integrate with multi-structured data sources?

Yes we have related data in Azure Data Lake Store and/or Azure Blob Storage

Azure SQL DW is a great ft

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

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



Q1 Have you justified a relational data warehouse solution is supported by business requirements and needs?

Yes

Q2 Are you comfortable with a cloud-based Platform-as-a-Service solution?

No

Yes

Q3 What kind of workload do you have?

Transactional (OLTP) with many small read/write operations or 'trickle' inserts or row-by-row operations

Data warehousing and analytical with batch-oriented read/write operations

Azure SQL DW is not an ideal choice

Q4 How large is your database?

Less than 250 GB

Approaching 1TB+ and expected to continue large scale growth

Q5 Do you have firm RPO, RTO, or backup requirements?

We have highly specific point-in-time backup/recovery requirements

No, we have some flexibility

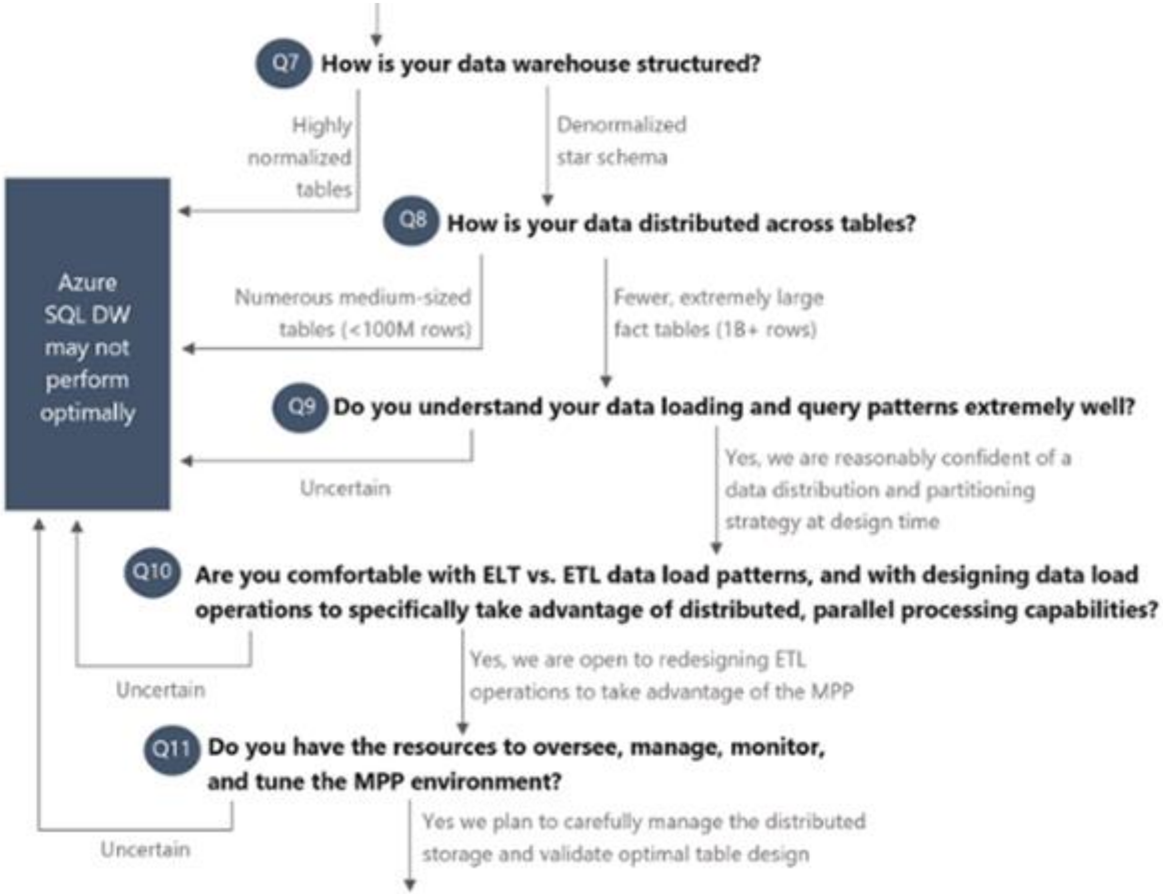
Q6 Do you plan to deliver a multi-tenant data warehouse?

Yes, we have a multi-tenancy design pattern

No

Q7 How is your data warehouse structured?

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



Q7 How is your data warehouse structured?

Highly normalized tables

Denormalized star schema

Azure SQL DW may not perform optimally

Q8 How is your data distributed across tables?

Numerous medium-sized tables (<100M rows)

Fewer, extremely large fact tables (1B+ rows)

Q9 Do you understand your data loading and query patterns extremely well?

Uncertain

Yes, we are reasonably confident of a data distribution and partitioning strategy at design time

Q10 Are you comfortable with ELT vs. ETL data load patterns, and with designing data load operations to specifically take advantage of distributed, parallel processing capabilities?

Uncertain

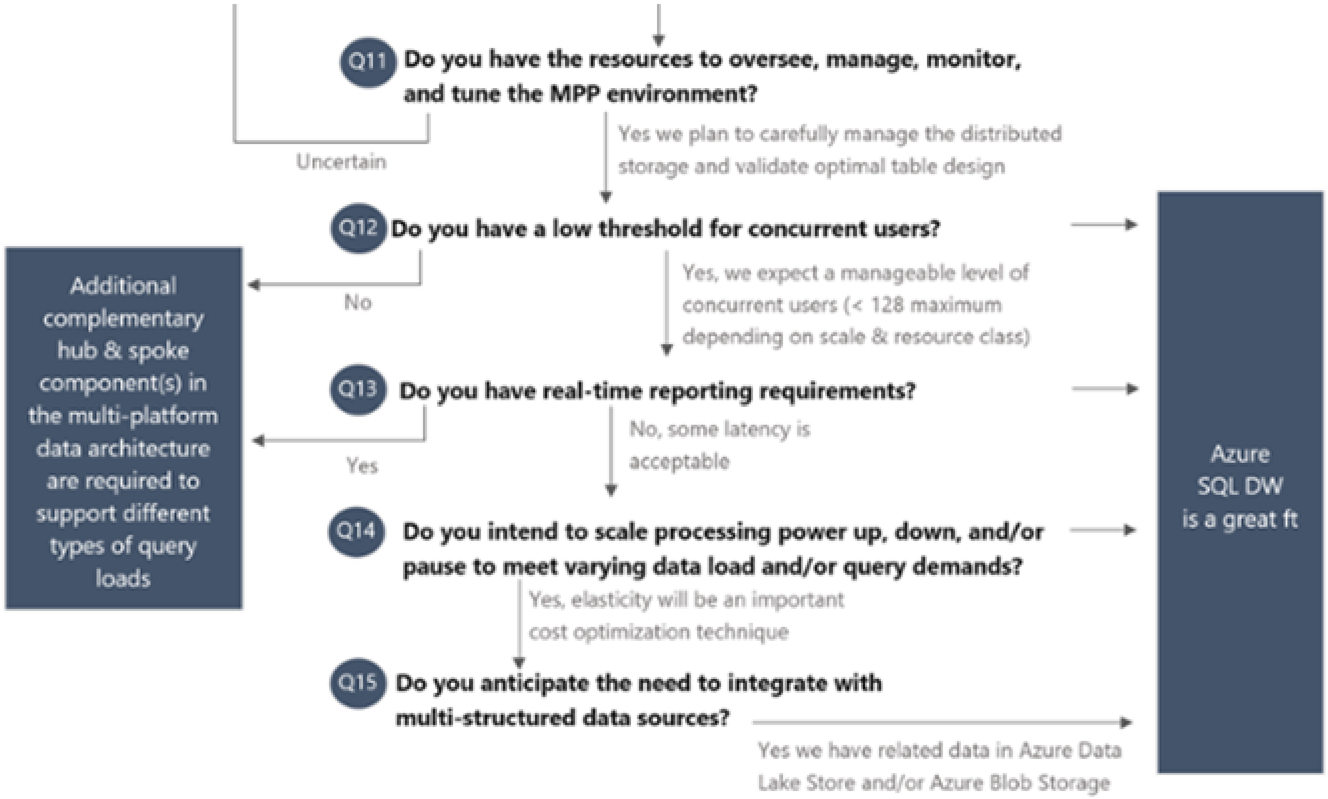
Yes, we are open to redesigning ETL operations to take advantage of the MPP

Q11 Do you have the resources to oversee, manage, monitor, and tune the MPP environment?

Uncertain

Yes we plan to carefully manage the distributed storage and validate optimal table design

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



Q11 Do you have the resources to oversee, manage, monitor, and tune the MPP environment?

Uncertain

Yes we plan to carefully manage the distributed storage and validate optimal table design

Q12 Do you have a low threshold for concurrent users?

No

Yes, we expect a manageable level of concurrent users (< 128 maximum depending on scale & resource class)

Q13 Do you have real-time reporting requirements?

Yes

No, some latency is acceptable

Q14 Do you intend to scale processing power up, down, and/or pause to meet varying data load and/or query demands?

Yes, elasticity will be an important cost optimization technique

Q15 Do you anticipate the need to integrate with multi-structured data sources?

Yes we have related data in Azure Data Lake Store and/or Azure Blob Storage

Additional complementary hub & spoke component(s) in the multi-platform data architecture are required to support different types of query loads

Azure SQL DW is a great ft

Data Preparation

- 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

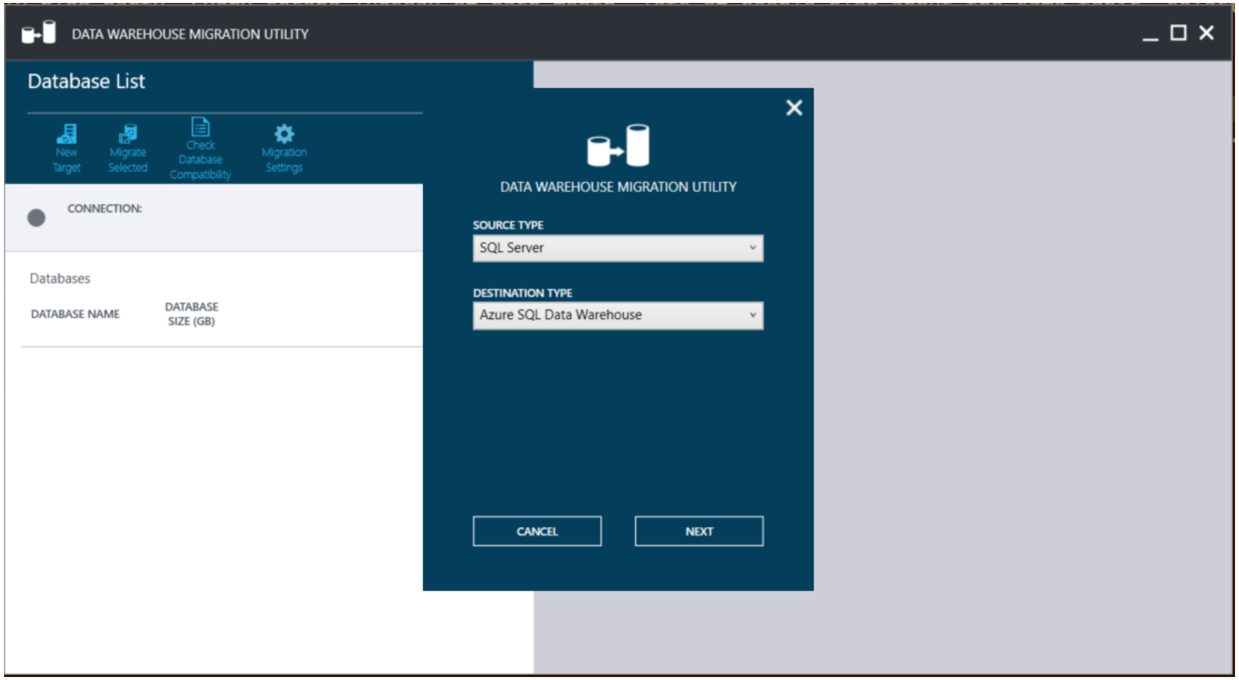
- Use Migration Tool
- Understand current T-SQL surface area and workarounds
- Avoid Singelton DML operations (INSERT, UPDATE, DELETE)
  - Batch DML if possible
  - If unavoidable, wrap in transaction (BEGIN TRAN ... COMMIT)
- Use heap table OR temp table for staging data
- Avoid large fully logged operations
  - Considers CTAS as this is minimal logged operation
  - Use LOJ as alternative DELETE
  - 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
- Explot 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



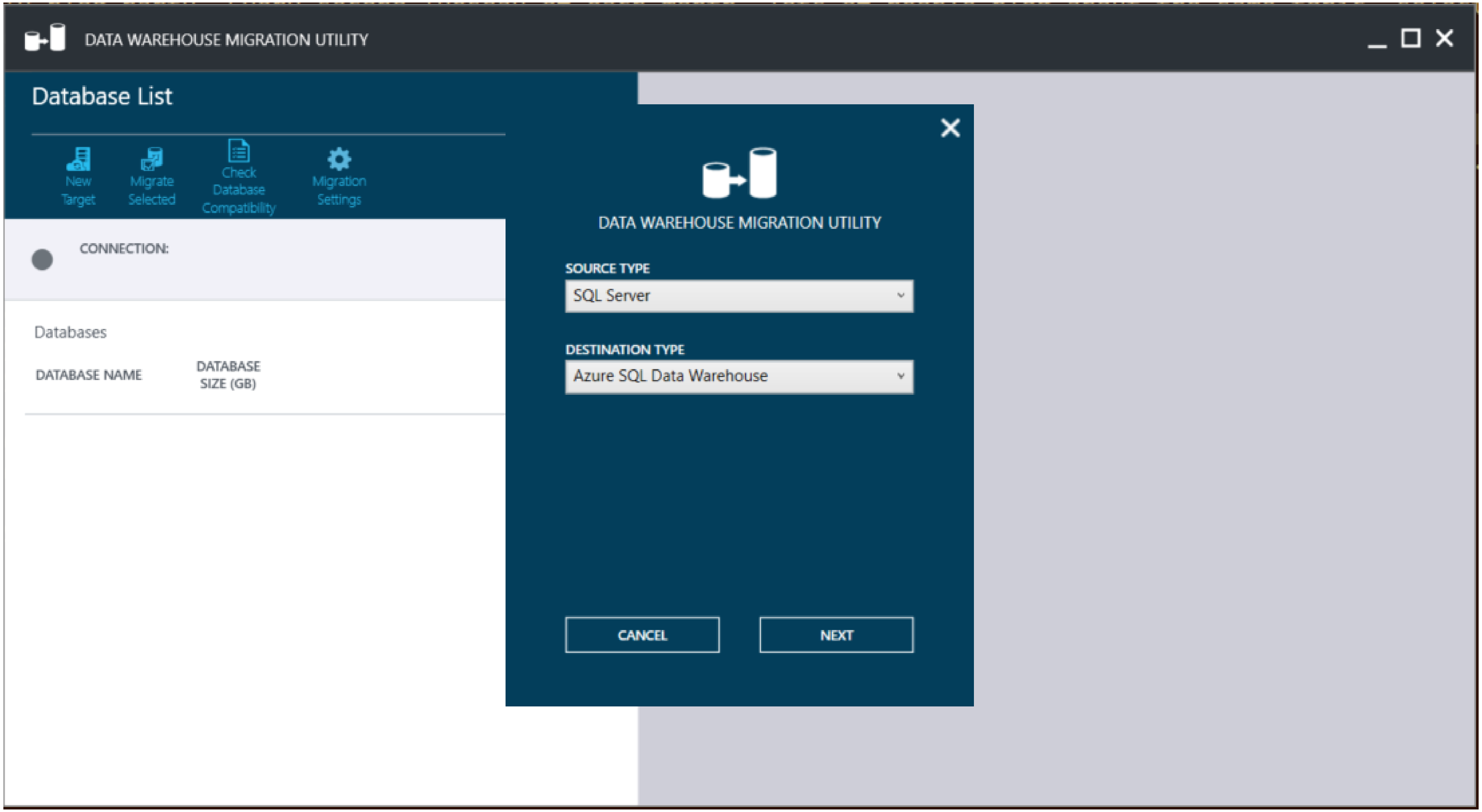
SOURCE TYPE: SQL Server

DESTINATION TYPE: Azure SQL Data Warehouse

Data Migration (WWI)

DEMO

Data Warehouse Migration Utility (Preview)



DATA WAREHOUSE MIGRATION UTILITY

Database List

CONNECTION:

Databases

DATABASE NAME | DATABASE SIZE (GB)

DATA WAREHOUSE MIGRATION UTILITY

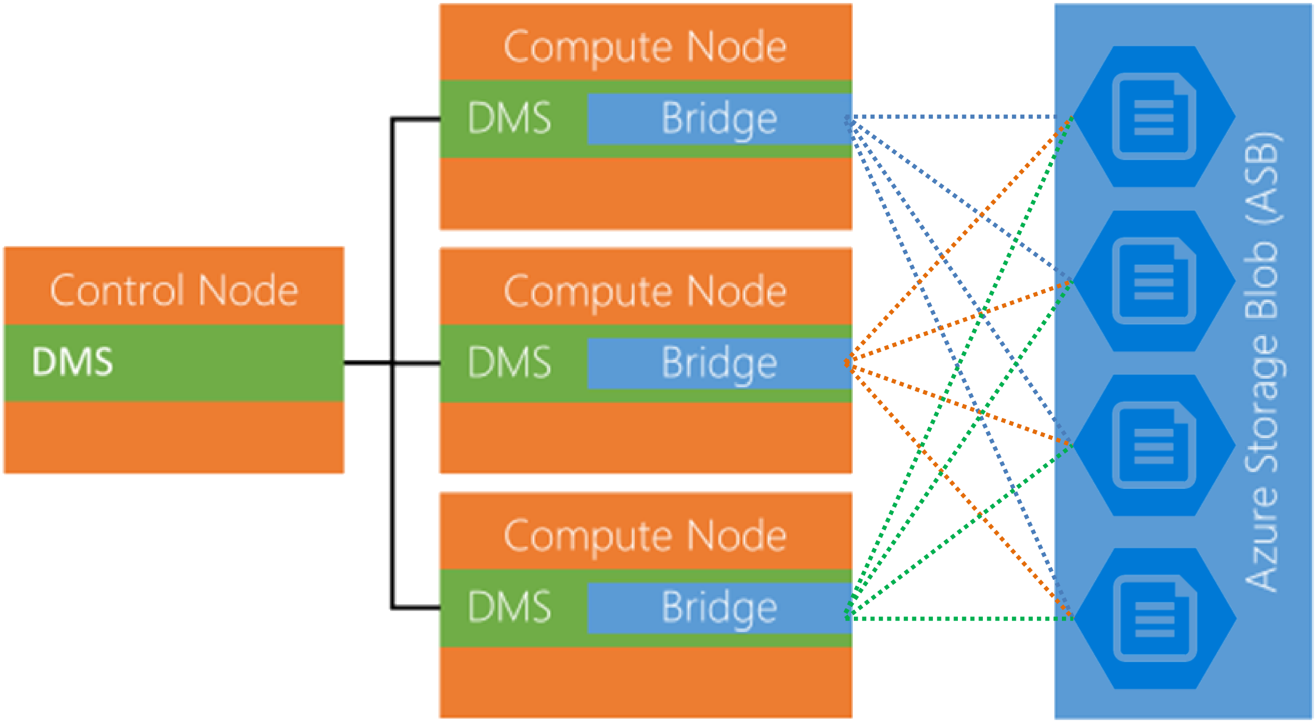
SOURCE TYPE: SQL Server

DESTINATION TYPE: Azure SQL Data Warehouse

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
  - No reject files & reason for rejected rows

Parallel Loading with PolyBase



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

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)

Parallel Loading with PolyBase

DEMO

Thank you

kamil@nowinski.net

@NowinskiK

http://SQLPlayer.net

Kamil Nowinski | MCSE Data Platform

Data Community

SQL Day

GOLD SPONSORS

elitmind think bright

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

VOLVO

Volvo Group IT

SILVER SPONSORS

dbWatch DATABASE CONTROL

it KONTRAKT®

UBS

veeAM

BRONZE SPONSOR

KSIĄŻKI IT Z CAŁEGO ŚWIATA

NOVATECH

www.novatech.com.pl

STRATEGIC PARTNER

Microsoft