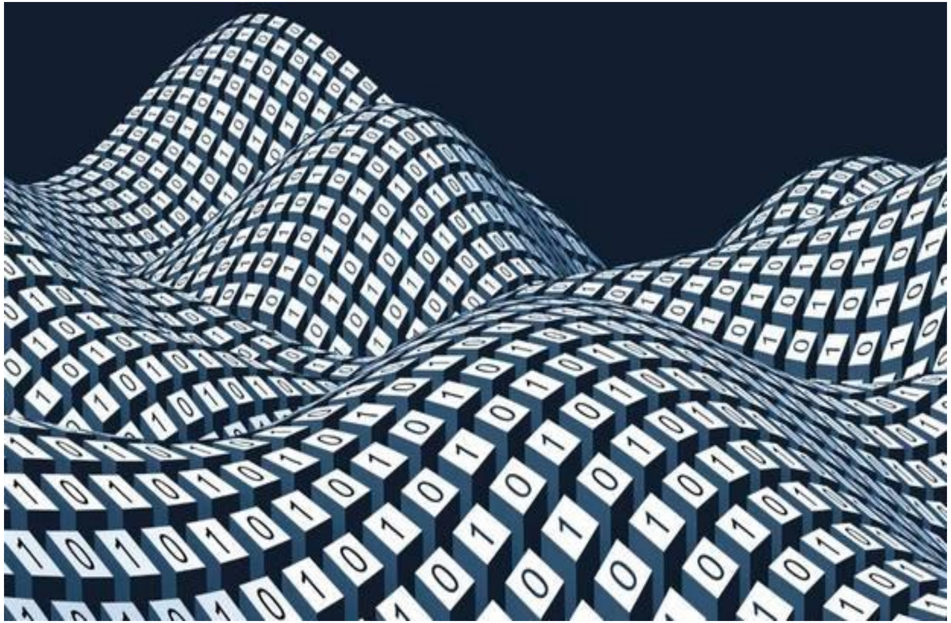
Visual Studio Microsoft Azure

Azure Global Bootcamp

Climbing Mountains of Data with Azure SQL DW



Bhavik Merchant
Data Platform Solution Architect
Microsoft Australia

Microsoft

Session Goals

- Gain high level understanding of Azure DW
- Know when to use it
- Explore new paradigms
- Highlight current limitations
- Compare with other Microsoft data offerings

Agenda

- Service overview
- Internal architecture, elasticity
- Provisioning, data distribution
- Visualising with Power BI
- Migration Steps
- Gotchas

Speaker Intro

- Worked with MS Data stack since SQL2000
- Was a consultant and practice lead
    - Business Intelligence (deepest/broadest knowledge)
    - SharePoint/Office 365 (enough to be dangerous!)
- Joined MSFT in January
    - Work with Enterprise clients across on-prem and cloud workloads
- Not a big data expert ☺

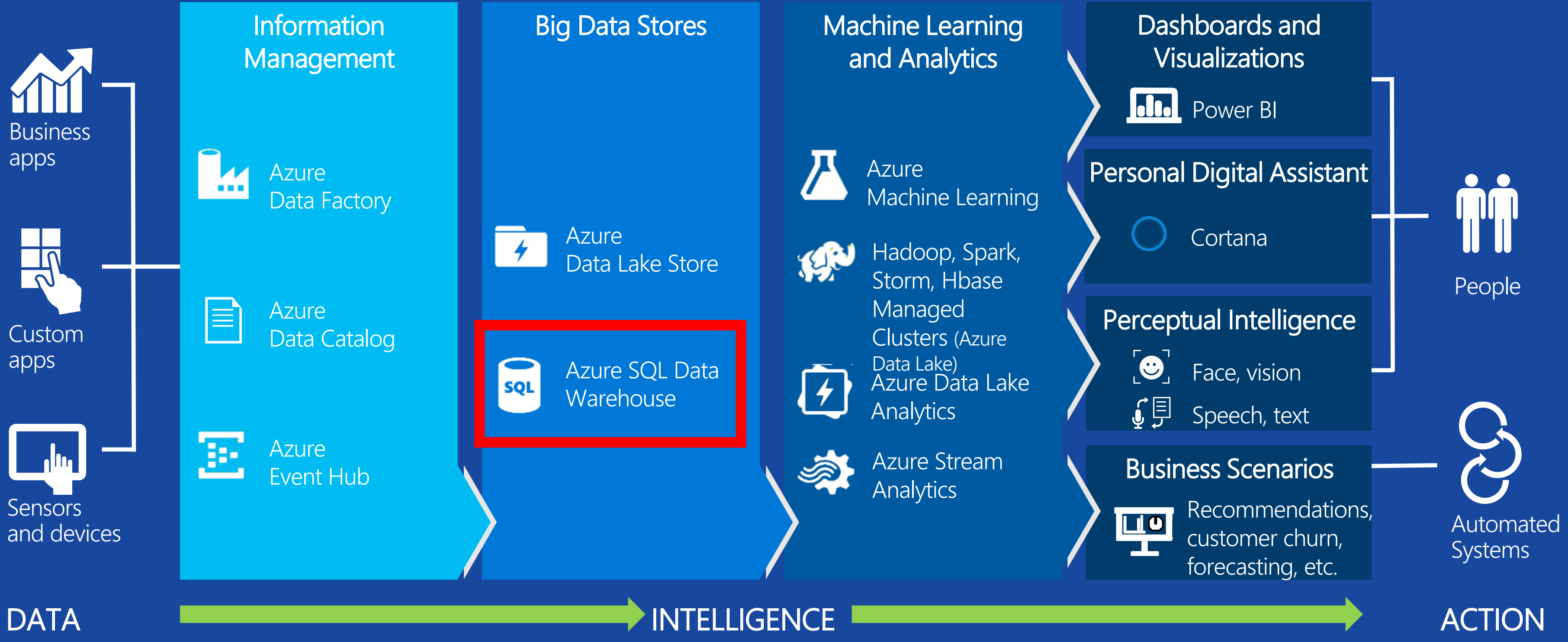
Overview

What is Azure SQL DW?

- Cloud based, scale out DB
- Designed for **massive** data volumes (60Tb max)
  - 5x compression = approx. 300Tb user data
- Built on MPP architecture
- Separates compute and storage
- Flexible: Scale out, scale back, pause/resume
- Allows seamless querying over Hadoop
- Supports hybrid architecture with APS

Cortana Intelligence

The Azure Data Platform at a Glance



Business apps

Custom apps

Sensors and devices

Information Management

Azure Data Factory

Azure Data Catalog

Azure Event Hub

Big Data Stores

Azure Data Lake Store

Azure SQL Data Warehouse

Machine Learning and Analytics

Azure Machine Learning

Hadoop, Spark, Storm, Hbase Managed Clusters (Azure Data Lake)

Azure Data Lake Analytics

Azure Stream Analytics

Dashboards and Visualizations

Power BI

Personal Digital Assistant

Cortana

Perceptual Intelligence

Face, vision

Speech, text

Business Scenarios

Recommendations, customer churn, forecasting, etc.

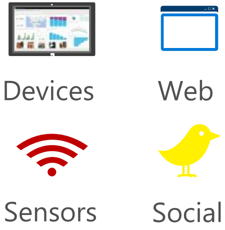
People

Automated Systems

DATA → INTELLIGENCE → ACTION

Why is Azure SQL DW Relevant?

- Increased data types and volumes
- Varied data sources
- Added complexity and cost
- Typical Big Data offerings have steep learning curve



What are the Benefits?

- Deploy within 10 minutes
  - No architecting, configuring, tuning*
- Use familiar paradigms
  - SQL tables, stored procs, indexes, partitions
  - T-SQL
- Cost control via elasticity – think peak/off-peak
- Hybrid architectures possible (sensitive vs non)
- Familiar analytical tools – Power BI, Excel, SSRS

Internals

SMP vs MPP

SMP/NUMA

Multiple CPUs are used to complete individual processes simultaneously

All CPUs share the same memory (SMP) OR different Groups of CPUs use different sets of memory on the same machine (NUMA)

All SQL Server implementations up until now have been SMP/NUMA

Massively parallel processing (MPP)

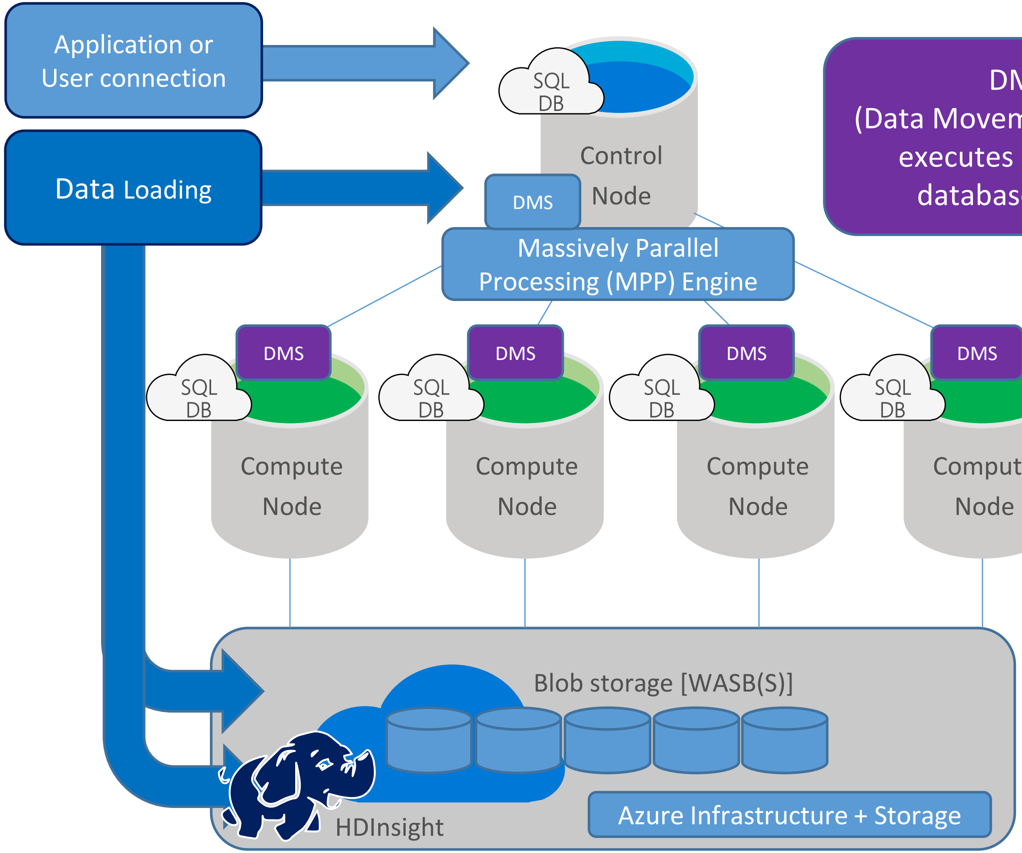
Multiple Nodes (computers) get utilized to process a single task

Many separate CPUs running in parallel across multiple nodes to execute a single task

Each set of CPUs has its own memory

Applications must be segmented, using high-speed communications between nodes

What’s in the Box?



Storage and Compute are de-coupled, enabling a true elastic service and separate charging for both compute and storage

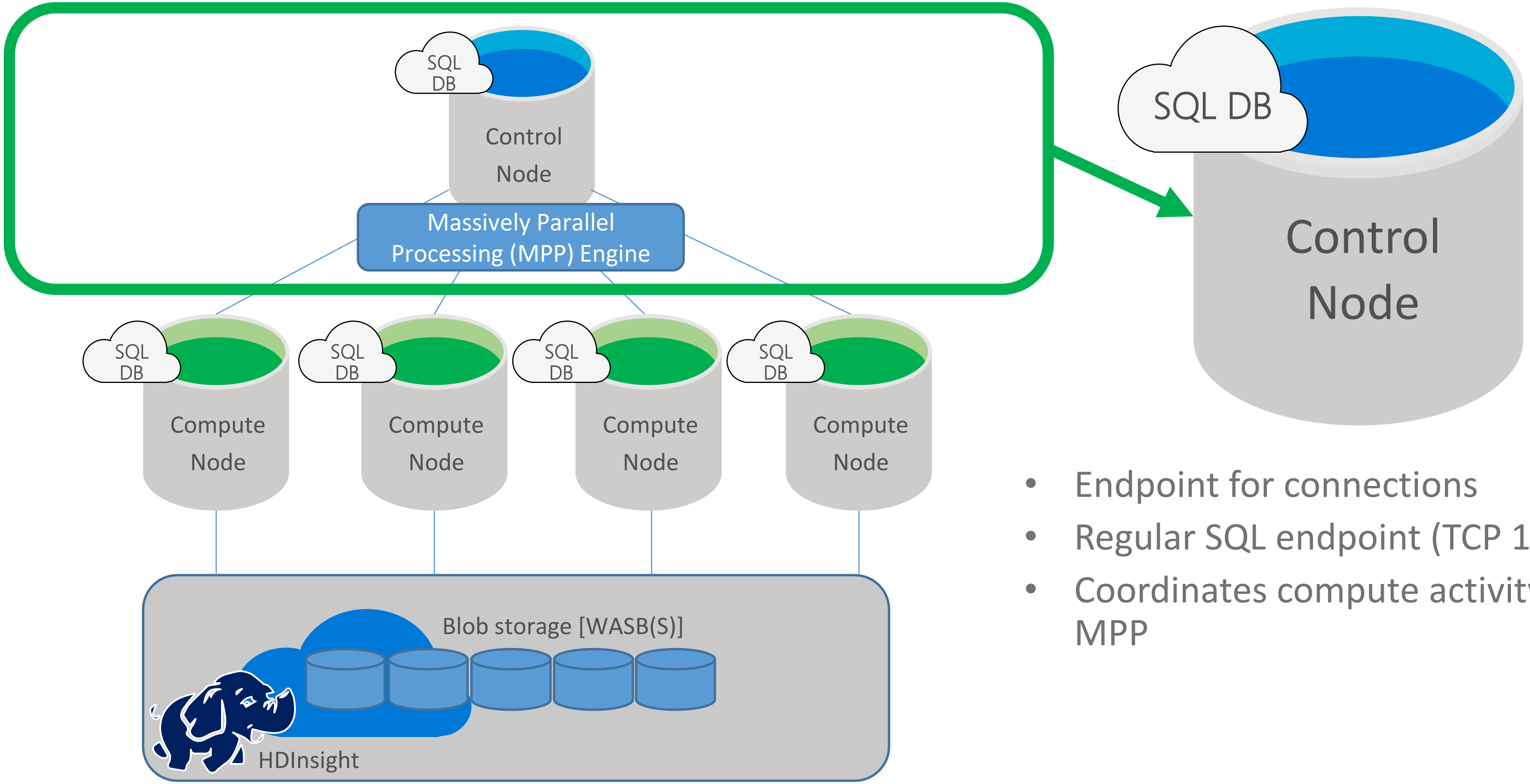
DMS
(Data Movement Service) executes across all database nodes

Compute
Scale compute up or down when required

Pause, Restart, Stop, Start.

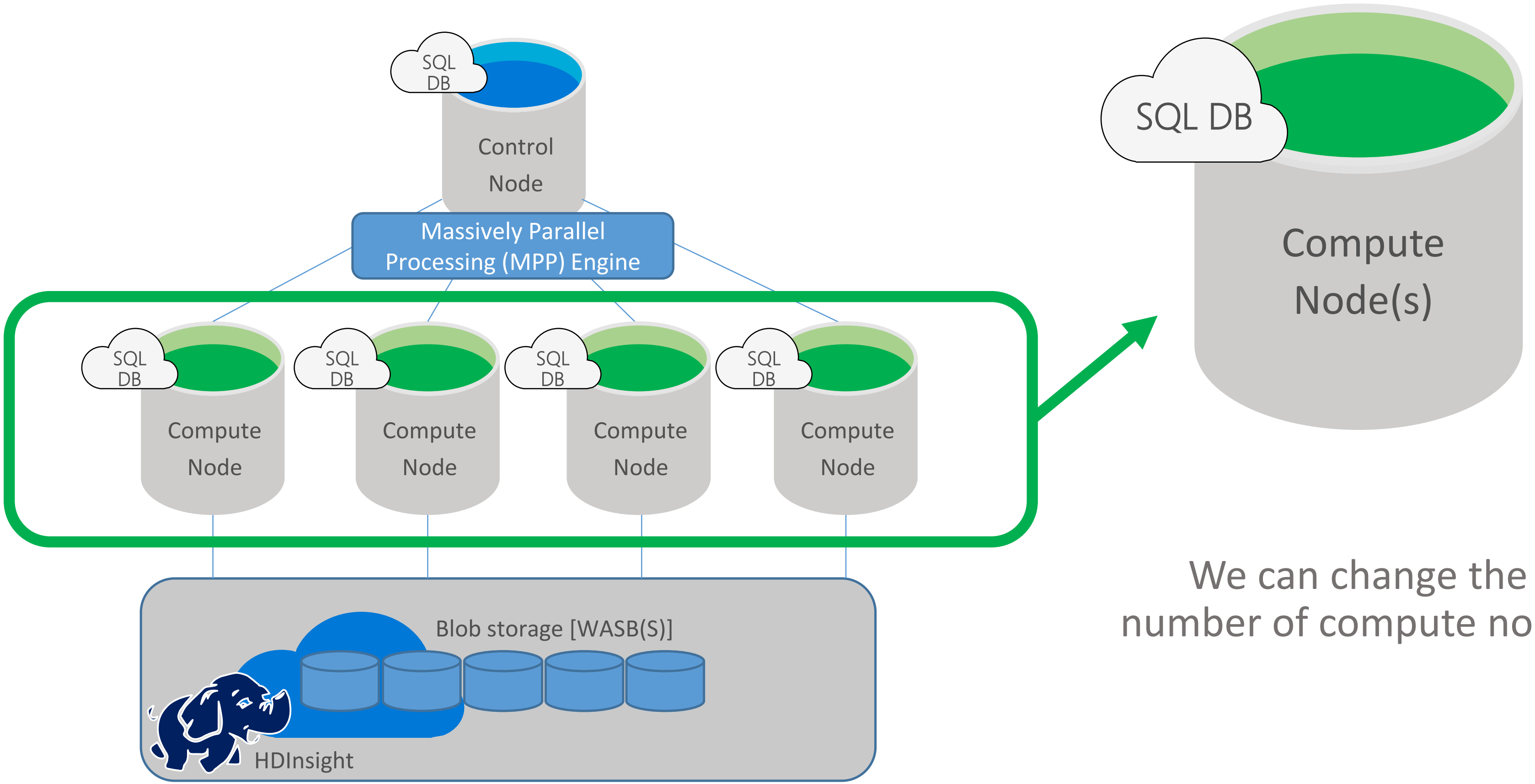
Storage
Add/Load data to WASB(S) without incurring compute costs

Control Node



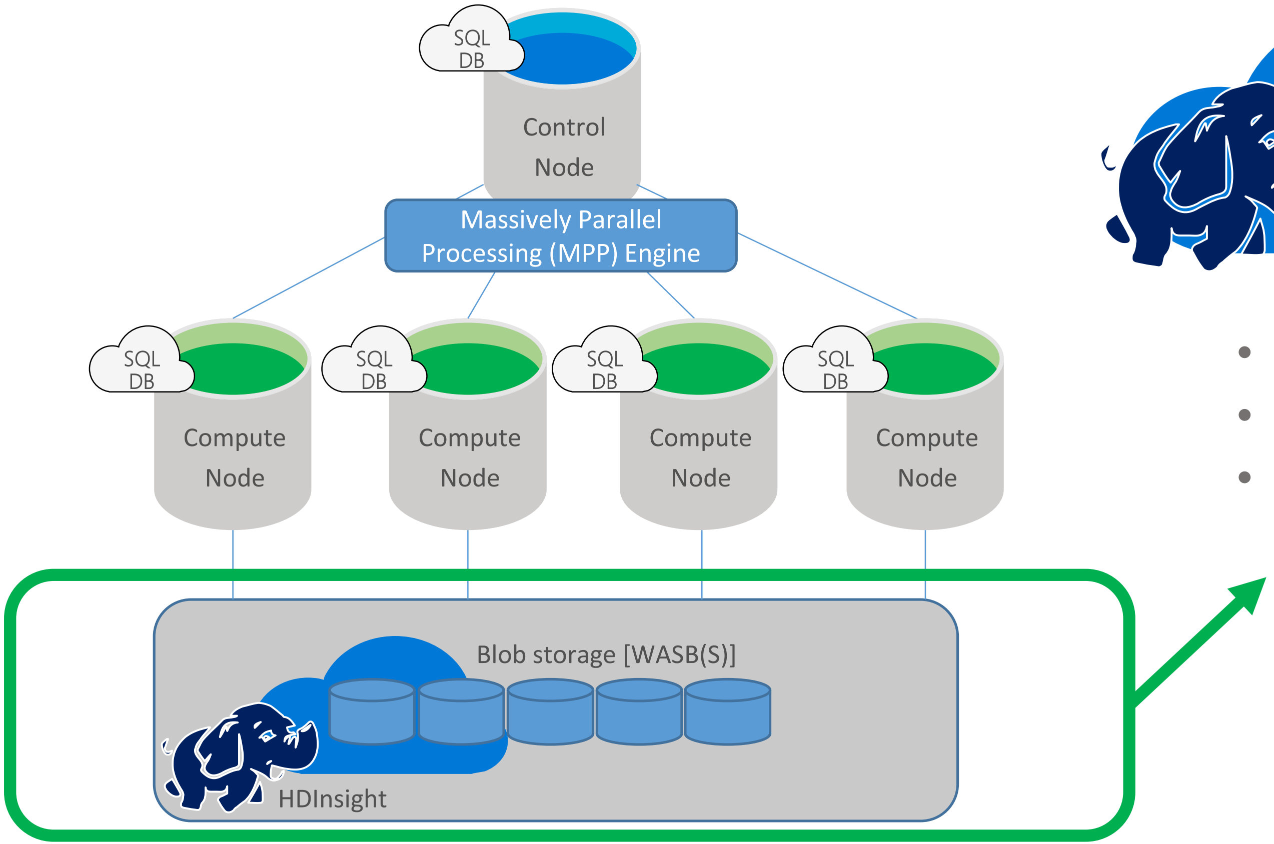
- Endpoint for connections
- Regular SQL endpoint (TCP 1433)
- Coordinates compute activity using MPP

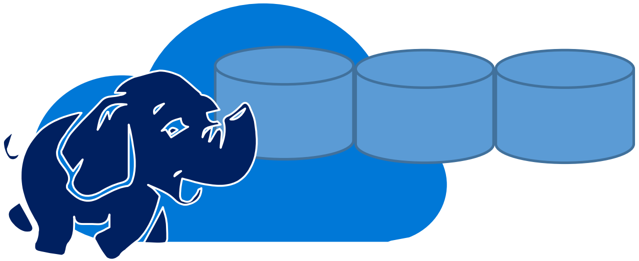
Compute Nodes



We can change the number of compute nodes

Blob storage





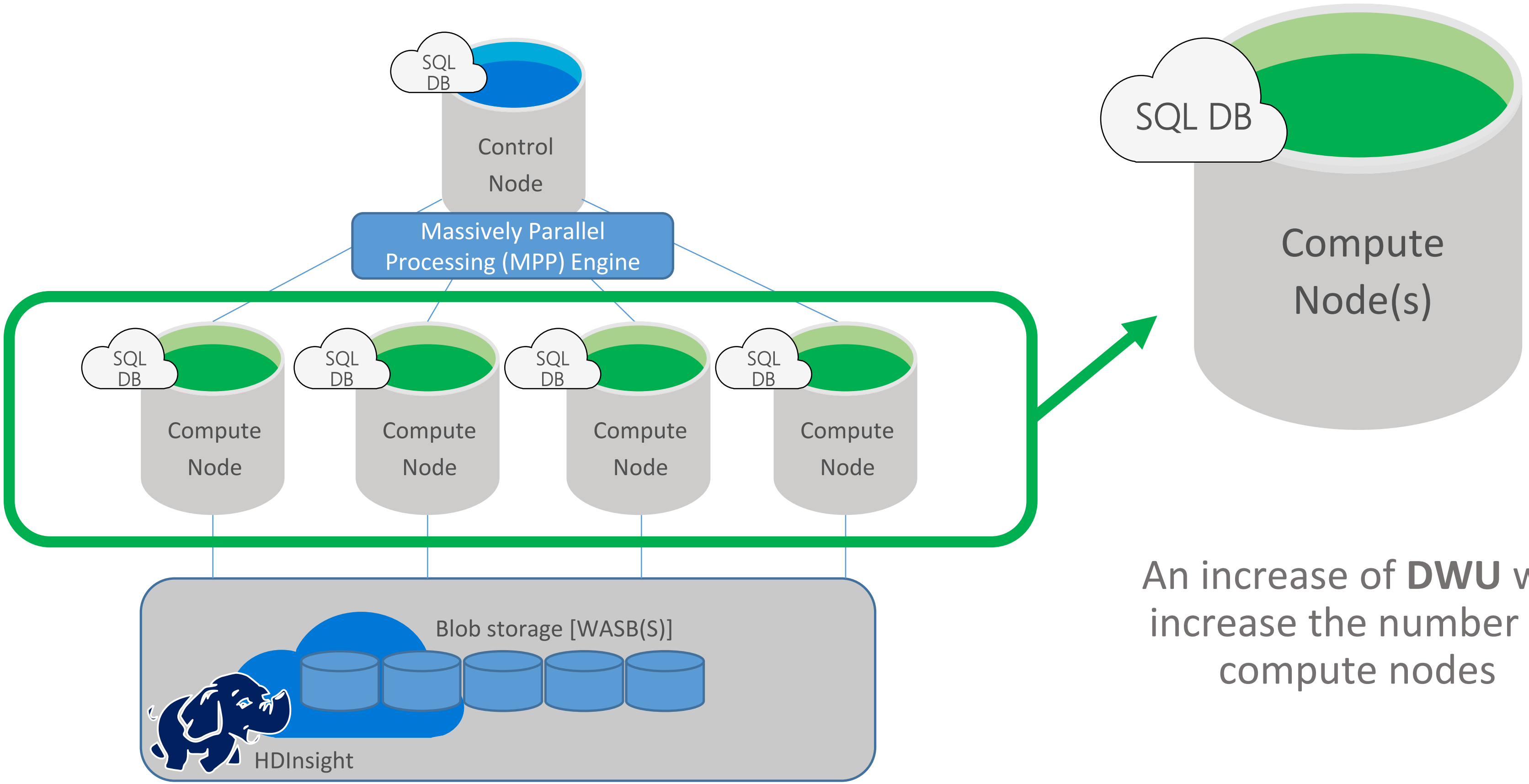
- RA-GRS storage
- +PB’s of storage
- Load data without incurring compute costs

Demo:

Provisioning & Accessing

Elasticity

Compute Nodes Again



An increase of **DWU** will increase the number of compute nodes

DW Who?

- Data Warehouse Unit (DWU)
- Derived from tests, gives competitive performance
- Blocks of 100, from 100 to 2000. Not all multiples.
- How many do I need?
  - Start small, test performance
  - Increase till you reach required performance levels
  - Scales linearly
  - Results are inaccurate with volumes < 1Tb
- Changes applied within 5 mins

How to Scale?

- Adjust slider in Azure Portal

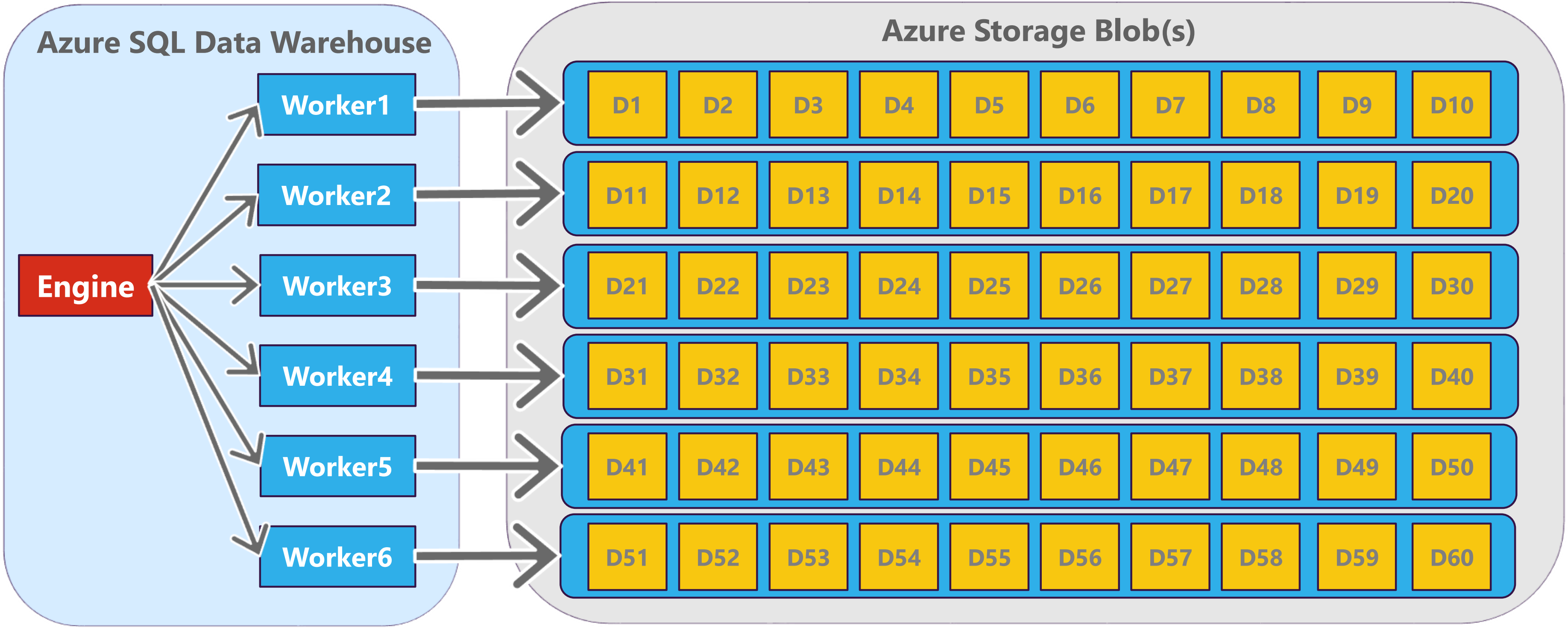
- Use T-SQL

  -
    ```
    ALTER DATABASE MyDB
    MODIFY (SERVICE_OBJECTIVE = 'DW1000');
    ```

- Use PowerShell

  -
    ```
    Set-AzureRmSqlDatabase -DatabaseName "MyDB" -ServerName
    "MyServer.database.windows.net" -RequestedServiceObjectiveName
    "DW1000"
    ```

Architecture of SQL DW for DWU600



Azure SQL Data Warehouse

Engine

Worker1

Worker2

Worker3

Worker4

Worker5

Worker6

Azure Storage Blob(s)

D1 D2 D3 D4 D5 D6 D7 D8 D9 D10

D11 D12 D13 D14 D15 D16 D17 D18 D19 D20

D21 D22 D23 D24 D25 D26 D27 D28 D29 D30

D31 D32 D33 D34 D35 D36 D37 D38 D39 D40

D41 D42 D43 D44 D45 D46 D47 D48 D49 D50

D51 D52 D53 D54 D55 D56 D57 D58 D59 D60

Data Distribution

Distribution Principles

- All tables in Azure DW are distributed
- Two choices, decided at table level
  - Round-Robin: Evenly but randomly (default)
  - Hash: Based on hashing values from a single column
- Hash is preferred where clusters of tables share common join/aggregation criteria
  - Prevents DMS having to shunt data from one compute node to another
- Currently always 60 SQL databases behind the scenes

Partitioning Tables

- Same as traditional SQL Server
- Partition swapping for “hot” data
- Allows query optimiser to exclude large row sets
- Minimum 100k rows per distribution for best performance and compression
  - Compression only applied after this threshold
- Partitions are also distributed

Distribution/Partitioning Syntax

```
CREATE TABLE [dbo].[myTable]
(
          [ProductKey] int NOT NULL
        , [OrderDateKey] int NOT NULL
        , [CustomerKey] int NOT NULL
        , [SalesOrderNumber] nvarchar(20) NOT NULL
        , [OrderQuantity] smallint NOT NULL
        , [SalesAmount] money NOT NULL
)

WITH
(
        CLUSTERED COLUMNSTORE INDEX
        , DISTRIBUTION = HASH([ProductKey])
        , PARTITION (
                [OrderDateKey]
                RANGE RIGHT FOR VALUES (20000101, 20010101, 20020101, 20030101, 20040101, 20050101)
        )
) ;
```

Changing Distribution Method

- Cannot use ALTER TABLE
- Must copy into new table
  - Use CREATE TABLE AS SELECT (CTAS)
- Same as previous example, add AS SELECT after

Demo:

Elasticity & Distribution

Loading Data

Various Data Loading options

- Directly into Azure DW
  - SSIS or 3rd Party Integration tools (OLE, ADO, ODBC)
- Via PolyBase (few GB)
  1. Create Azure storage account and container
  2. Use AZCopy to upload files
  3. Create External Data Source and Table
  4. Use CREATE TABLE AS SELECT syntax to load in to DW
- Import/Export service for larger volumes
  - Use Azure Import/Export Tool (supports up to 8TB per disk)
  - It will encrypt drive
  - Create Job in portal and send via courier

Visualising with Power BI

- Connect directly from Azure Portal
  - Opens in PowerBI.com
  - Works for single tables/views
- Launch Power BI Desktop, connect to Azure DW
  - Must specify relationships

Demo:

Visualising in Power BI

Security

Authentication

- Currently only SQL Server auth
- “Server admin” login is god, any DB
- Create login and database users as normal
  - CREATE LOGIN (on master)
  - CREATE USER (on database)

Authorisation

- Use standard database roles/permission
  - EXEC sp_addrolemember 'db_datareader', <user>';
  - EXEC sp_addrolemember 'db_datawriter', <user>';
- Other standard permissions also available

Encryption

- Uses Transparent Data Encryption (TDE)
  - ALTER DATABASE [myDB] SET ENCRYPTION ON;
- Can also enable via Azure portal

Caveats

- No PK/FK relationships
  - A result of MPP architecture. Workaround – views
- Other unsupported constructs
  - E.g. Constraints, triggers
- Must manually create stats on columns
  - Don’t skip! Choose join/group columns. Regularly update.
- Cant use fully SSMS/SSDT fully at present
- Not all data types supported
  - Guide available online for conversion
- 32k buffer size for DMS

Understand SQL Data Warehouse

Where does SQL Data Warehouse fit?

SQL Server VM (Iaas)	Azure SQL Database	Azure SQL Data Warehouse	Azure Data Lake
OLTP / DW workloads	OLTP/ DW workloads	DW workloads only	Non-relational
Lift and Shift	Net new development	Fully managed	Cheap, flexible Access
Customer managed	Fully managed service	Dynamic Pause/Scale	Processing raw data
1TB+	1-500 GB	250GB – PB+	1 TB+

Final Takeaways

- Probably the best price/performance out there
  - $1.31 per hour per DWU
- Cost effective
  - Only MPP DW that separate compute and storage
  - Can scale on demand
  - Can pause/resume
- Excellent performance
  - Scaled AW sample gave 30s query time for aggregation/join over 5 billion rows
- Polybase allows hybrid architecture
- Integrated with Azure ML, Data Factory, Power BI