Introduction to Microsoft SQL Server Big Data Clusters

Buck Woody – Applied Data Scientist
Microsoft Data Platform Team

June 2019
Data Growth
Computing and Storage advances impact data collection abilities

Computing and Storage technologies allow greater data collection points

They also allow longer historical data storage, and as time goes on become part of that storage lineage

Walmart is a classic example of data proliferation and leverage
## Use-Cases

Every Industry classification benefits from Big Data, Retail and Finance leads the way

<table>
<thead>
<tr>
<th>Industry Sector</th>
<th>Primary Use-Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retail</td>
<td>Demand prediction</td>
</tr>
<tr>
<td></td>
<td>In-store analytics</td>
</tr>
<tr>
<td></td>
<td>Supply chain optimization</td>
</tr>
<tr>
<td></td>
<td>Customer retention</td>
</tr>
<tr>
<td></td>
<td>Cost/Revenue analytics</td>
</tr>
<tr>
<td></td>
<td>HR analytics</td>
</tr>
<tr>
<td></td>
<td>Inventory control</td>
</tr>
<tr>
<td>Finance</td>
<td>Cyberattack Prevention</td>
</tr>
<tr>
<td></td>
<td>Fraud detection</td>
</tr>
<tr>
<td></td>
<td>Customer segmentation</td>
</tr>
<tr>
<td></td>
<td>Market analysis</td>
</tr>
<tr>
<td></td>
<td>Risk analysis</td>
</tr>
<tr>
<td></td>
<td>Blockchain</td>
</tr>
<tr>
<td></td>
<td>Customer retention</td>
</tr>
<tr>
<td>Healthcare</td>
<td>Fiscal control analytics</td>
</tr>
<tr>
<td></td>
<td>Disease Prevention prediction and classification</td>
</tr>
<tr>
<td></td>
<td>Clinical Trials optimization</td>
</tr>
<tr>
<td></td>
<td>Patient load analysis</td>
</tr>
<tr>
<td></td>
<td>Episode analytics</td>
</tr>
<tr>
<td>Public Sector</td>
<td>Revenue prediction</td>
</tr>
<tr>
<td></td>
<td>Education effectiveness analysis</td>
</tr>
<tr>
<td></td>
<td>Transportation analysis and prediction</td>
</tr>
<tr>
<td></td>
<td>Energy demand and supply prediction and control</td>
</tr>
<tr>
<td></td>
<td>Defense readiness predictions and threat analysis</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>Predictive Maintenance (PdM)</td>
</tr>
<tr>
<td></td>
<td>Anomaly Detection</td>
</tr>
<tr>
<td></td>
<td>Pattern analysis</td>
</tr>
<tr>
<td>Agriculture</td>
<td>Food Safety analysis</td>
</tr>
<tr>
<td></td>
<td>Crop forecasting</td>
</tr>
<tr>
<td></td>
<td>Market forecasting</td>
</tr>
<tr>
<td></td>
<td>Pipeline Optimization</td>
</tr>
</tbody>
</table>
Scale-Out Processing
Scaled Processing and Scaled Storage
The foundations of scale

Hadoop

- **MapReduce** (data processing)
- **Others** (data processing)
- **YARN** (cluster resource management)
- **HDFS** (redundant, reliable storage)

Spark

- **RDD** (Row)
- **RDD** (Transforms)
- **Dataset**
- **DataFrame**
- **Dataset**
- **Trained Model**
- **Model Training**
- **Feature Engineering**
- **Parameter Tuning**
- **SparkSQL**
- **SparkUI**
- **Scala/Java/Python**
- **Data Result**
- **Data Frame**
Virtualization

Building on hardware, you can create a complete “PC” on top of a Hypervisor layer, which abstracts out the hardware. You still own the Operating System and up.

This allows for scale by ring-fencing OS-level dependencies.
Containers
Abstracting the OS, Allowing complete portability

Containers go one level further than the Hypervisor, and focusing on binaries and applications

Storage and networking are a consideration

Scale is achieved through multiple containers
Container Orchestration
Containers at Scale

- **Container(s)** live in **Pods**
- **Pod(s)** are abstractions within **Nodes**
- **Node(s)** are PC’s or VM’s
- **Cluster(s)** are groups of **Nodes**
- Storage is by means of **Volume(s)** mounted through a **Claim**
- **Routing** provides external host name mapping and load balancing
Generic Cluster
Scale by Purpose
SQL Server
Platform Evolution

On Premises

Windows

Linux

Containers

SQL Server

Public/Private cloud

Hybrid
SQL Server 2019 Big Data Cluster – Complete Architecture
SQL Server 2019 and Big Data
Virtualization, Data Lake, Data Marts, and Spark
LOB Apps
Application Calls to SQL Server Master Instance. Relational, multi-type, Graph, and ML features supported. No code change.
SQL Server 2019 and Big Data
Cluster Architecture: Data Lake, Data Mart, and Spark

Control Plane
- Master

Compute Plane
- SQL Server
- SQL Server
- SQL Server
- SQL Server
- SQL Server

Data Plane
- Storage Pool
  - SQL Server
  - Spark
  - HDFS
- Storage Pool
  - SQL Server
  - Spark
  - HDFS
- Storage Pool
  - SQL Server
  - Spark
  - HDFS
- SQL Data Pool
  - SQL Server

App Pool
- ML Server
- Job (SSIS)
- (Web Apps)

Tools:
- Knox Gateway
- Livy
- HIVE
- SQL Cluster Administration Portal
- Grafana Dashboard
- Kibana Dashboard
- PolyBase Connector
SQL Server 2019 Big Data – Data Virtualization
SQL Server 2019 big data cluster

Control Plane

- Master

Compute Plane

- Compute Pool
  - SQL Server
  - SQL Server
  - SQL Server
  - SQL Server

- Compute Pool
  - SQL Server
  - SQL Server
  - SQL Server
  - SQL Server

Data Plane

- Storage Pool
  - SQL Server
  - Spark
  - HDFS

- Storage Pool
  - SQL Server
  - Spark
  - HDFS

- Storage Pool
  - SQL Server
  - Spark
  - HDFS

App Pool

- SQL Data Pool
  - SQL Server

- SQL Data Pool
  - SQL Server

- SQL Cluster Administration Portal
- Grafana Dashboard
- Kibana Dashboard
- Knox Gateway
- Livy
- HIVE
- SQL Server
- Spark
- PolyBase Connector
- ML Server
- Job (SSIS)
- (Web Apps)
Data Virtualization

**Multiple Data Sources**
Data Virtualization Scale-out calls through SQL Server Master Instance using External Tables, through the Compute Pool using PolyBase Connectors at the Source.
SQL Server 2019 and Big Data
Data Virtualization

PolyBase Connector
Scale-Out

PolyBase Connector
NoSQL

PolyBase Connector
RDBMS
SQL Server 2019 Big Data Cluster – Data Mart
Data Persistence Using Multiple Data Sources
Data Virtualization Scale-out calls through SQL Server Master Instance using External Tables, through the Compute Pool using PolyBase Connectors at the Source. Results are stored in the Shards of the Data Pool.
SQL Server 2019 and Big Data

Data Mart

SQL Server

Compute Pool

PolyBase Connector
- HDFS
- Cosmos DB
- RDBMS

SQL Server Data Pool
(Shards)
SQL Server 2019 Big Data Cluster – Data Lake, Machine Learning and Spark
Data Lake, Scale, ML and AI

**Multiple Data Sources**
Data Virtualization Scale-out calls through SQL Server Master Instance using External Tables through the Compute Pool to the Data Pool.

**Scaled Data Analysis**
Data Mart Scale-out calls through SQL Server Master Instance using External Tables into Data Pool. Direct calls to a Data Lake (HDFS) using the Storage Pool.

**Data Science**
Data Engineering and Pipelines for Models with big data using Notebooks and other tools through to Spark, ingesting and processing data using the Storage Pool.

**AI Enablement**
Prediction and Classification Scoring to AI apps using the App Pool.
SQL Server 2019 Big Data Cluster – Tools, Management and Monitoring
Takeaways

• SQL Server 2019 Big Data cluster includes SQL Server together with the HDFS and Spark Compute engine as one package for big data processing, Machine Learning and AI

• Spark is a distributed compute engine that provides a unified framework for E2E big data processing pipeline including Machine learning and AI

• You can use SQL Server 2019 to create a secure, hybrid, machine learning architecture starting with data preparation, training a machine learning model, operationalizing your Model and using it for scoring

• Go Do > Practice with installing, configuring, and operating SQL Server 2019
• Go Do > Download this deck and practice a demo on Big Data Clusters on SQL Server
• Go Do > Follow a complete workshop
Resources

- Official documentation – aka.ms/bdc
- In-depth training - aka.ms/sqlworkshops
THANK YOU

linkedin.com/company/Red-Hat
facebook.com/RedHatInc
youtube.com/user/RedHatVideos
twitter.com/RedHat
THANK YOU

linkedIn.com/company/Red-Hat
youtube.com/user/RedHatVideos
facebook.com/RedHatInc
twitter.com/RedHat