Introduction to Microsoft SQL Server Big Data Clusters

Buck Woody – Applied Data Scientist
Microsoft Data Platform Team

June 2019

Microsoft

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

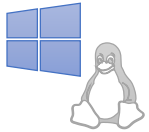
Every Industry classification benefits from Big Data, Retail and Finance leads the way

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

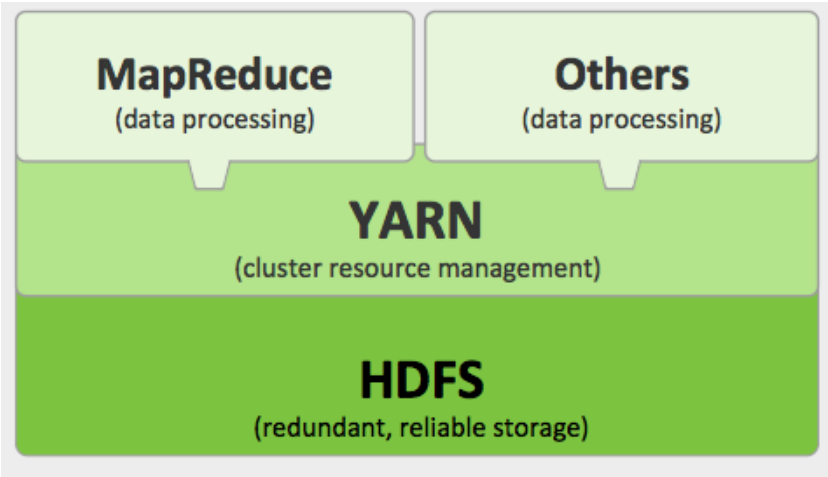
Scale-Out Processing

Scaled Processing and Scaled Storage

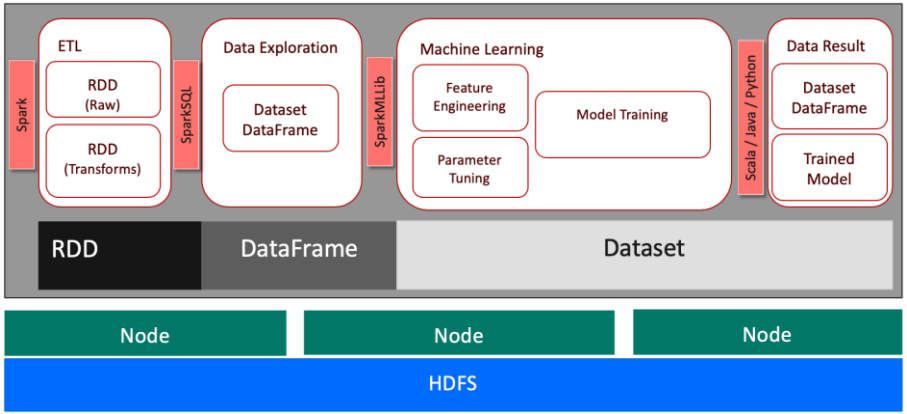
The foundations of scale



Hadoop



Spark

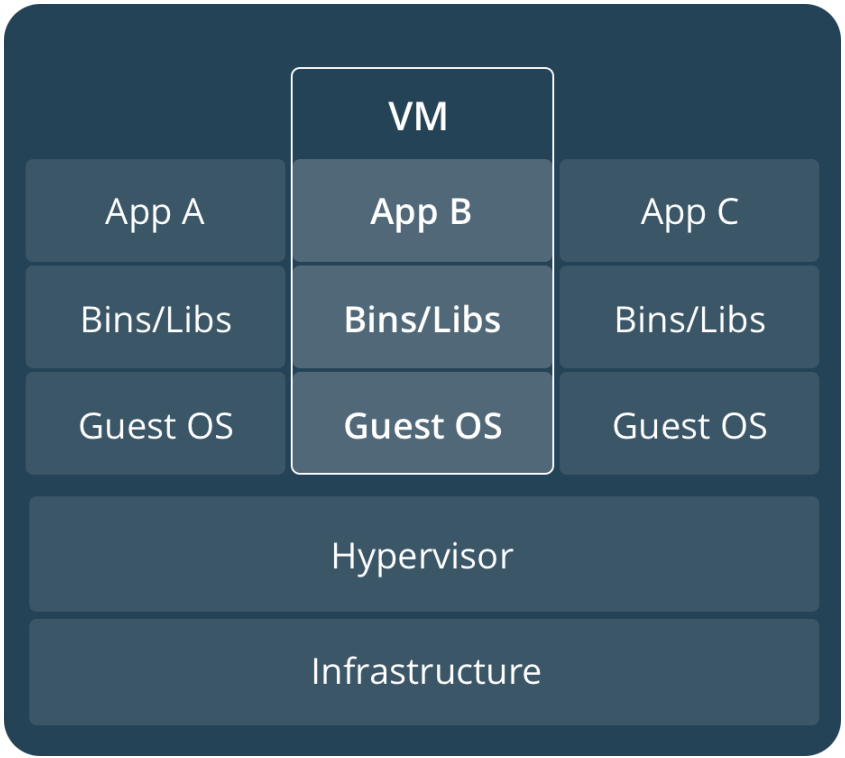


Virtualization

Hardware Abstraction

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



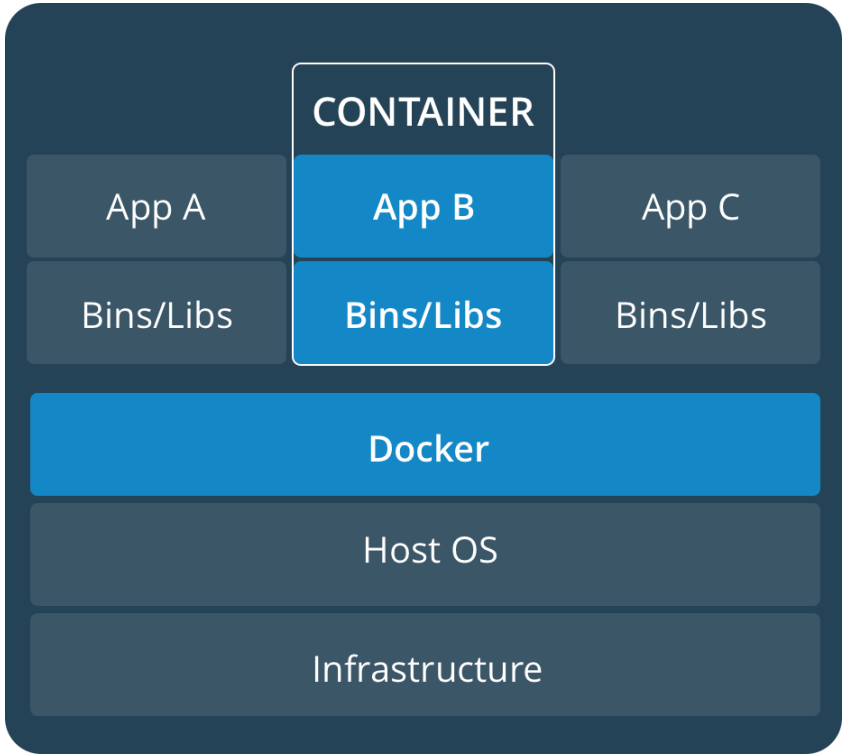
Containers

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

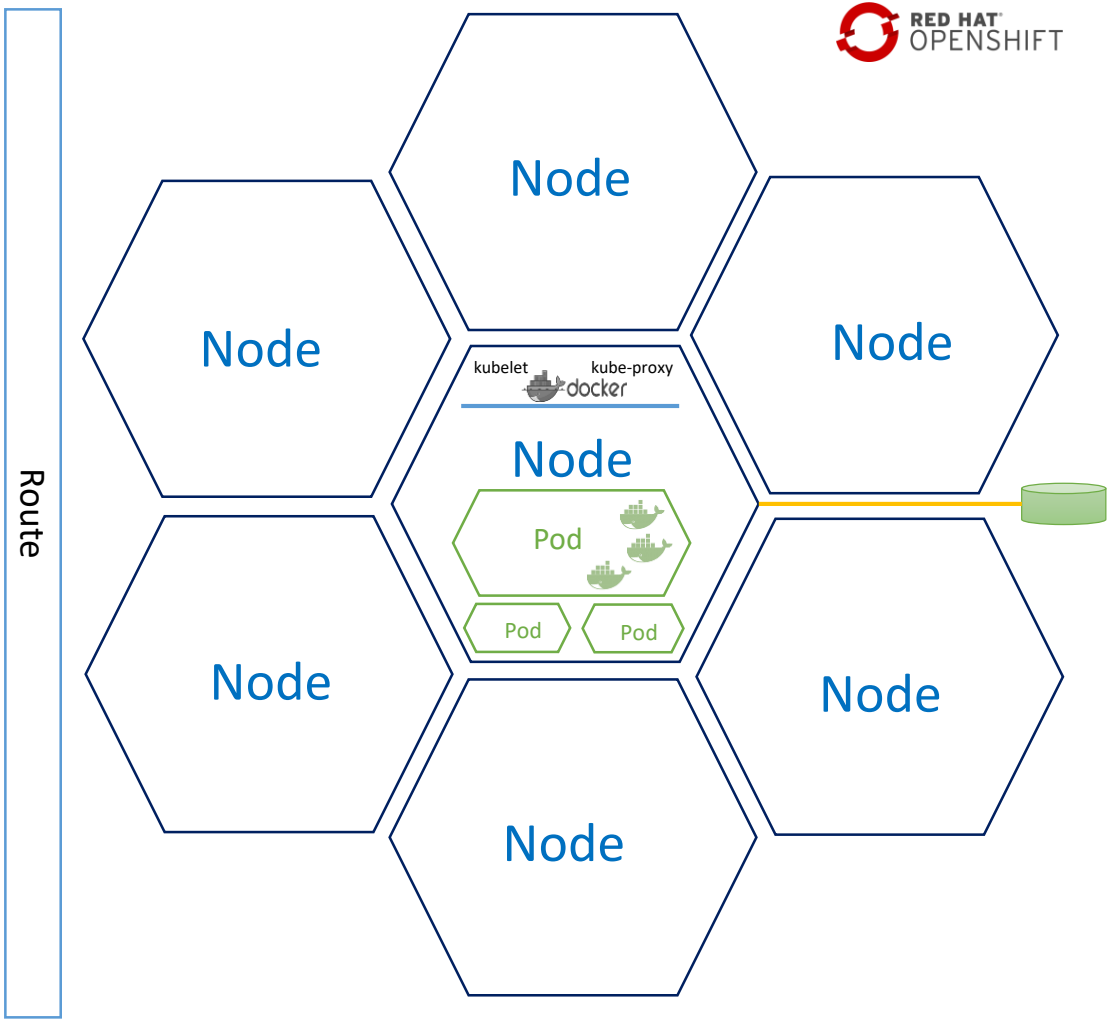
Scale is achieved through multiple containers



Container Orchestration

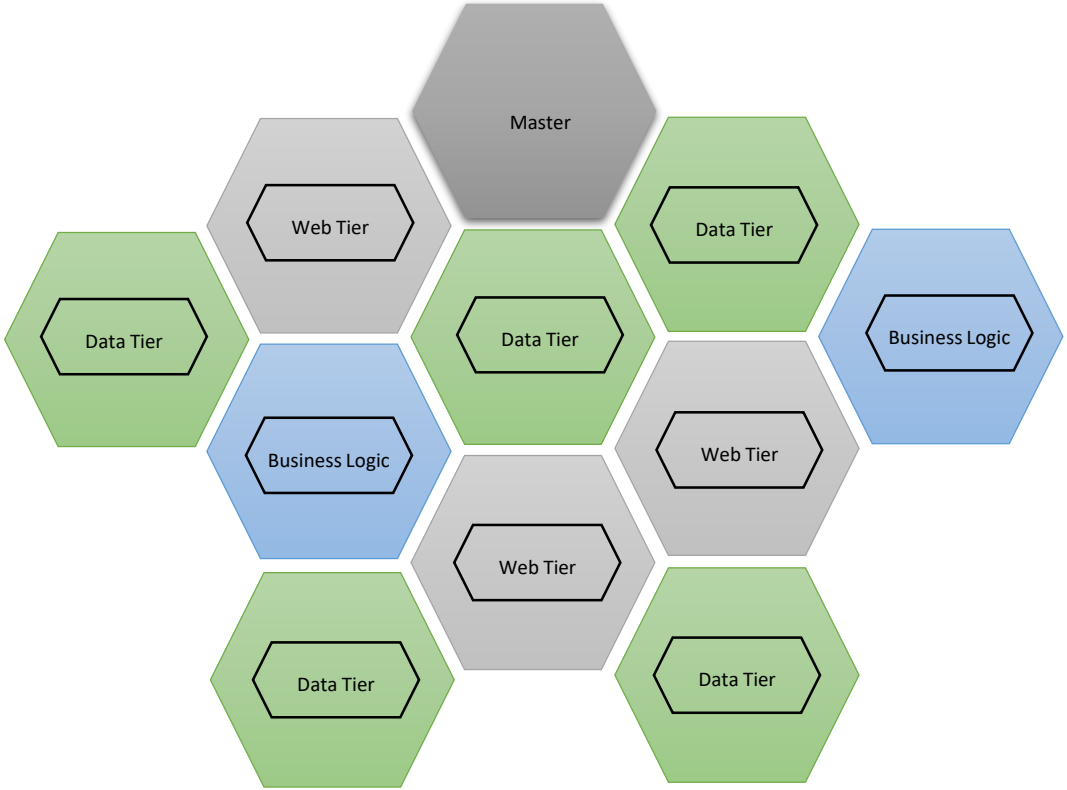
Containers at Scale

- 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

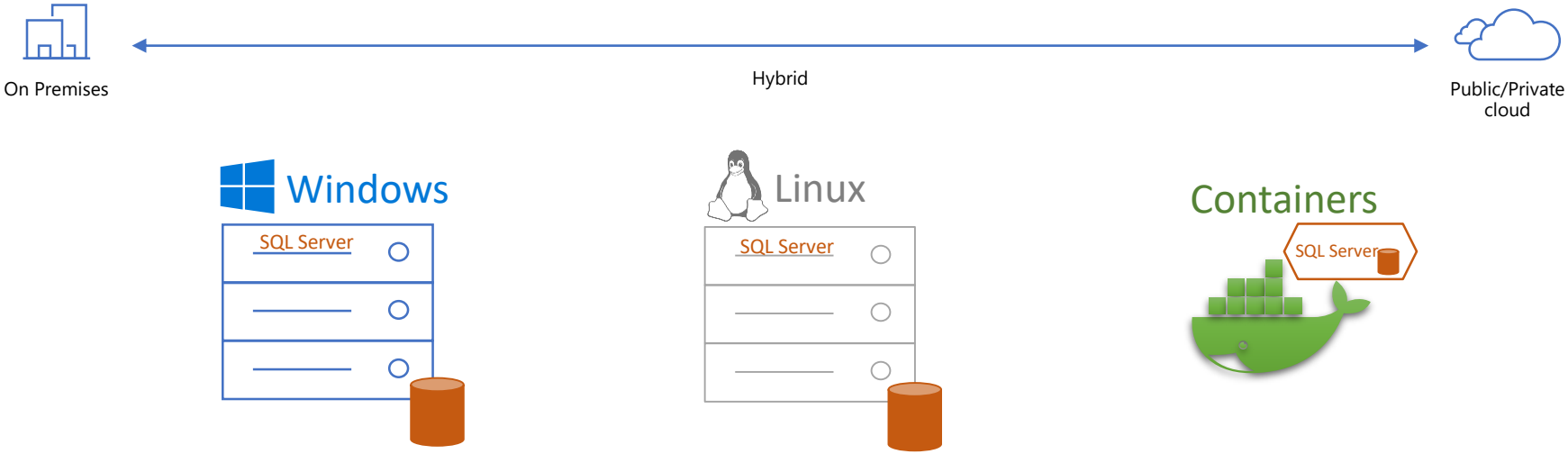
Scale by Purpose



SQL Server

Platform Evolution

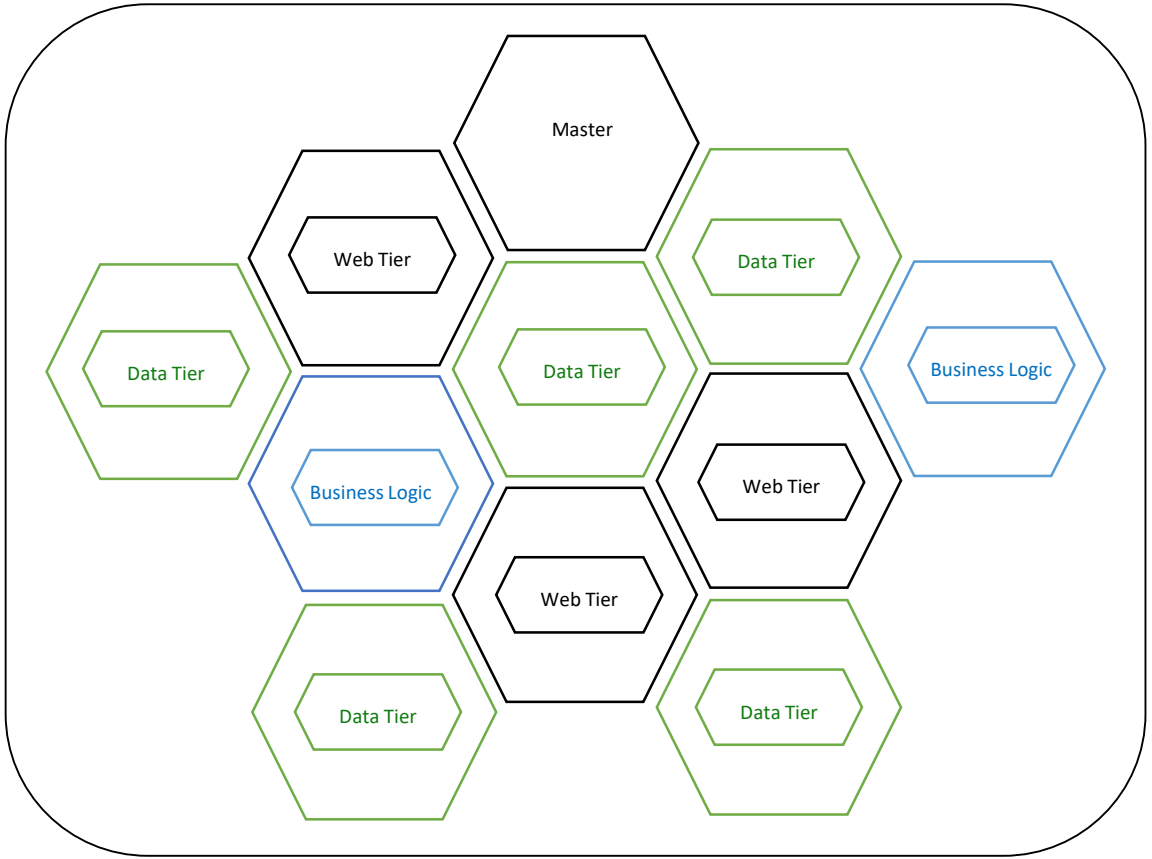
Microsoft SQL Server



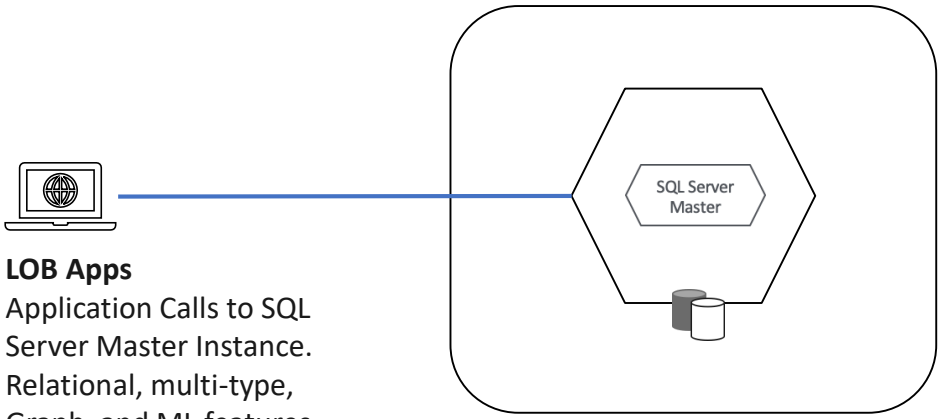
SQL Server 2019 Big Data Cluster – Complete Architecture

SQL Server 2019 and Big Data

Virtualization, Data Lake, Data Marts, and Spark



High Value, OLTP Data

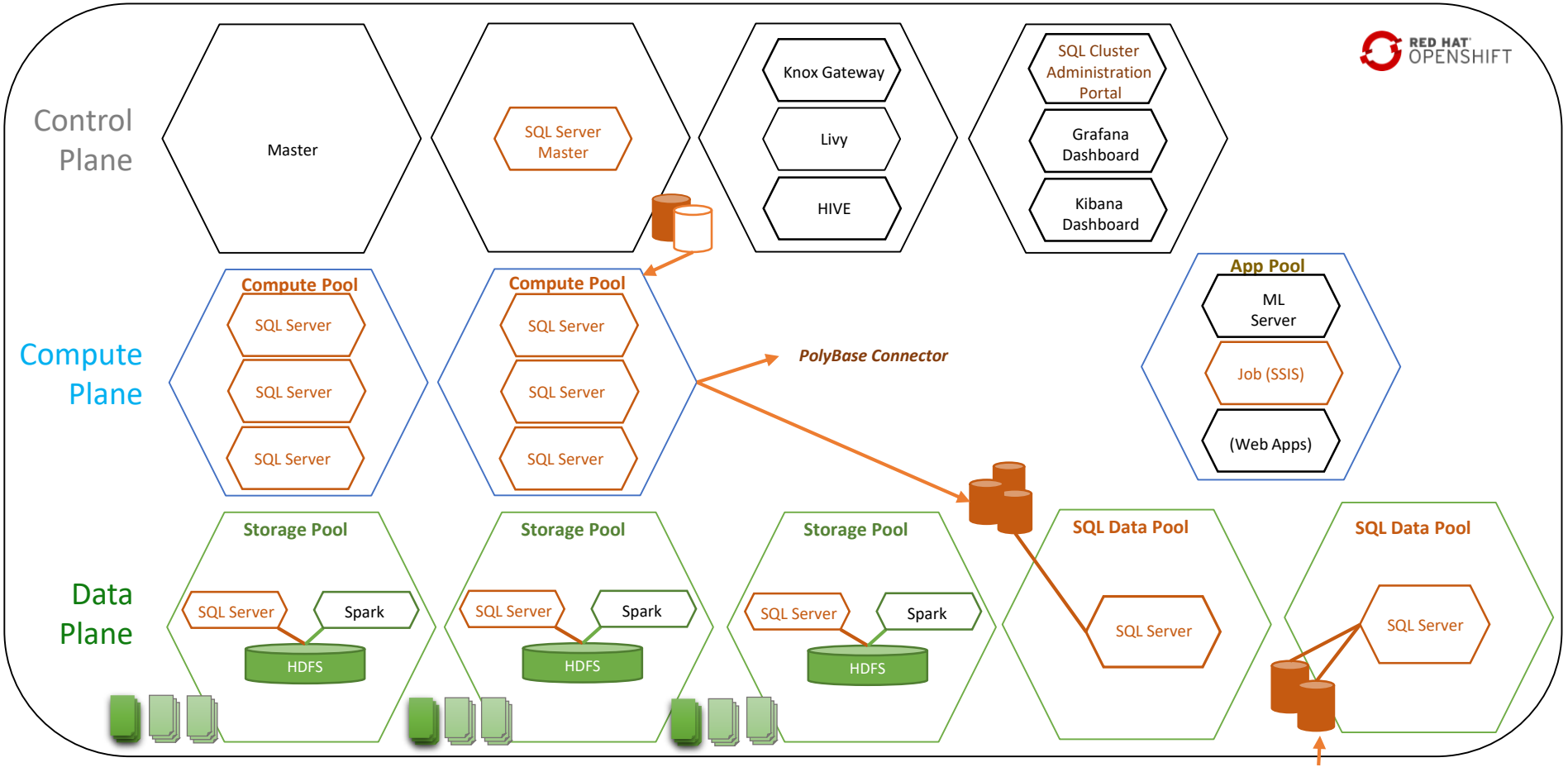


LOB Apps
Application Calls to SQL Server Master Instance. Relational, multi-type, Graph, and ML features supported. No code change.

SQL Server 2019 and Big Data

Cluster Architecture: Data Lake, Data Mart, and Spark

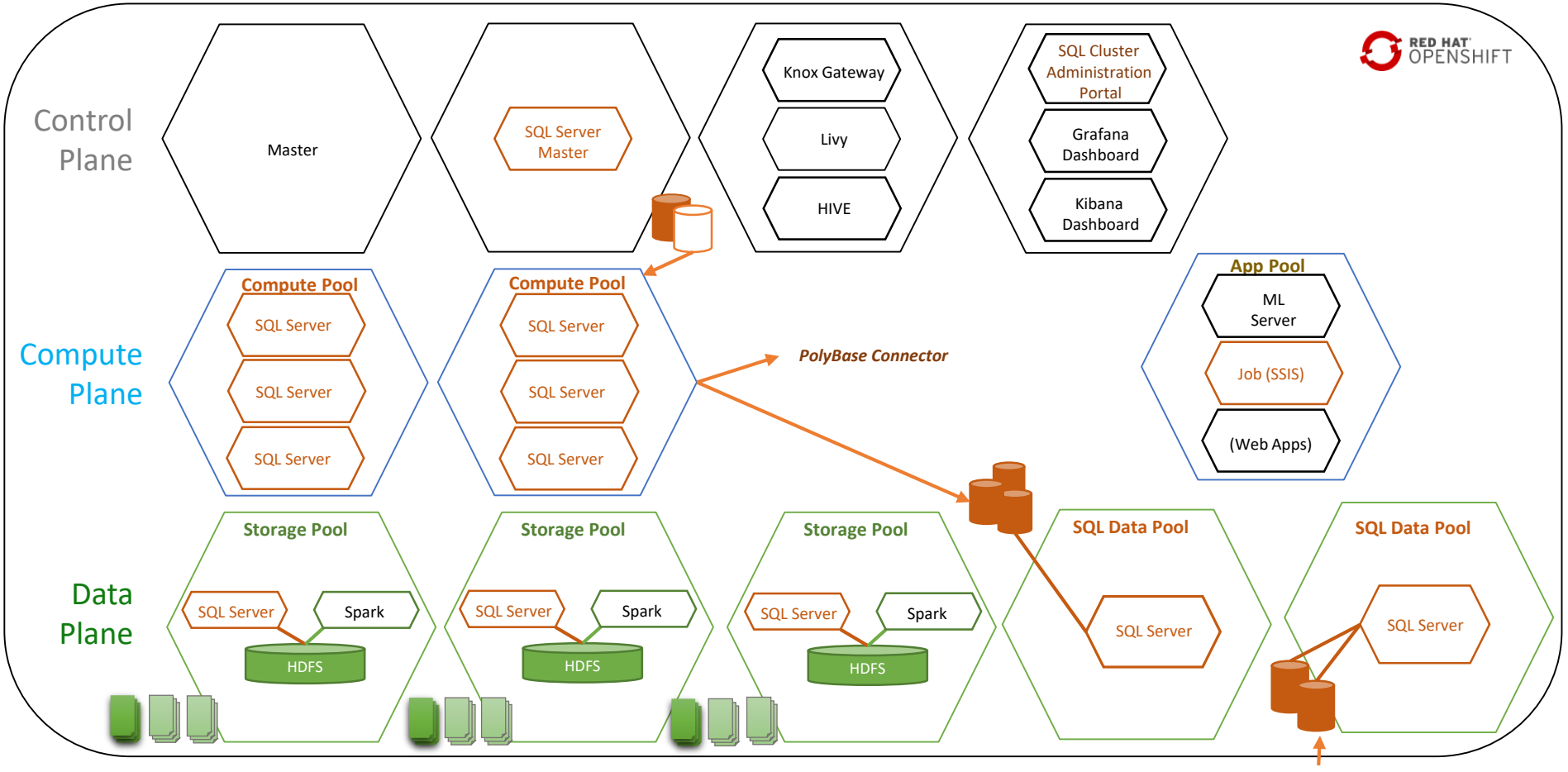
Microsoft SQL Server



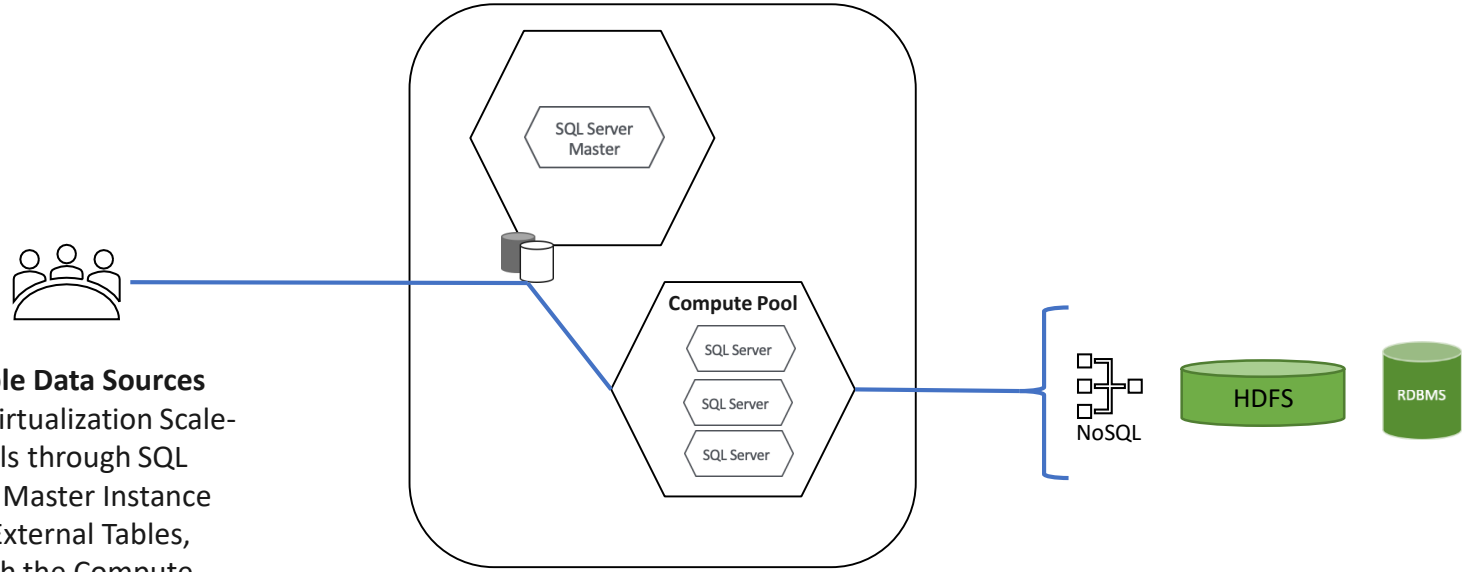
SQL Server 2019 Big Data – Data Virtualization

SQL Server 2019 big data cluster

Microsoft SQL Server



Data Virtualization

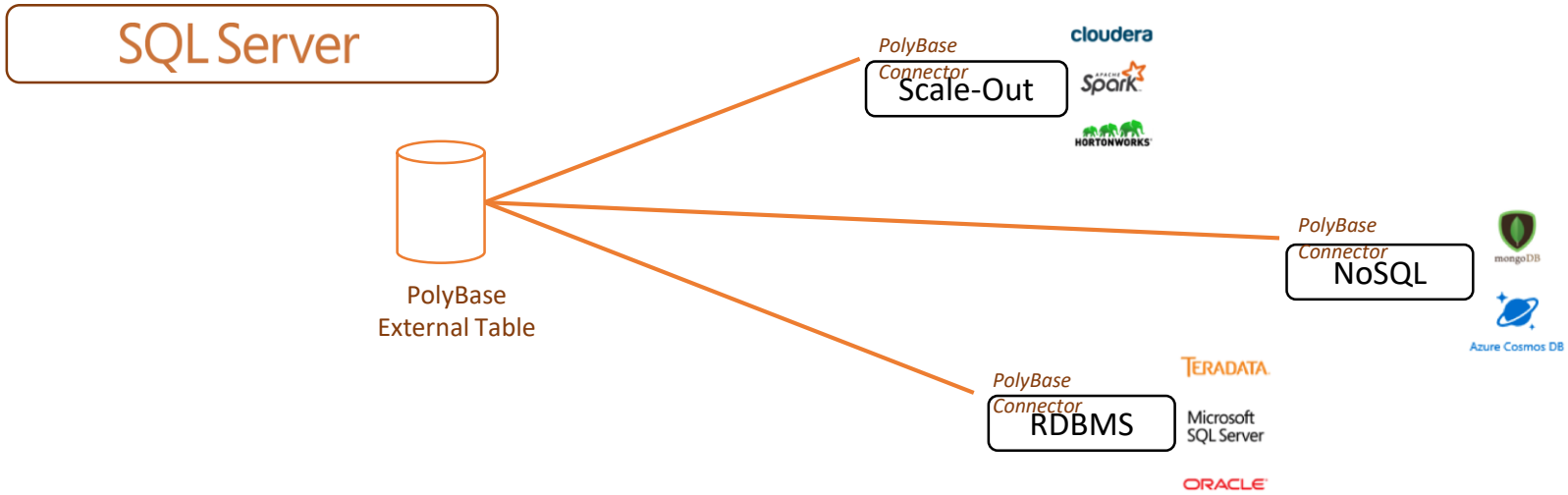


Multiple Data Sources
Data Virtualization Scale-out calls through SQL Server Master Instance using External Tables, through the Compute Pool using PolyBase Connectors at the Source

SQL Server 2019 and Big Data

Data Virtualization

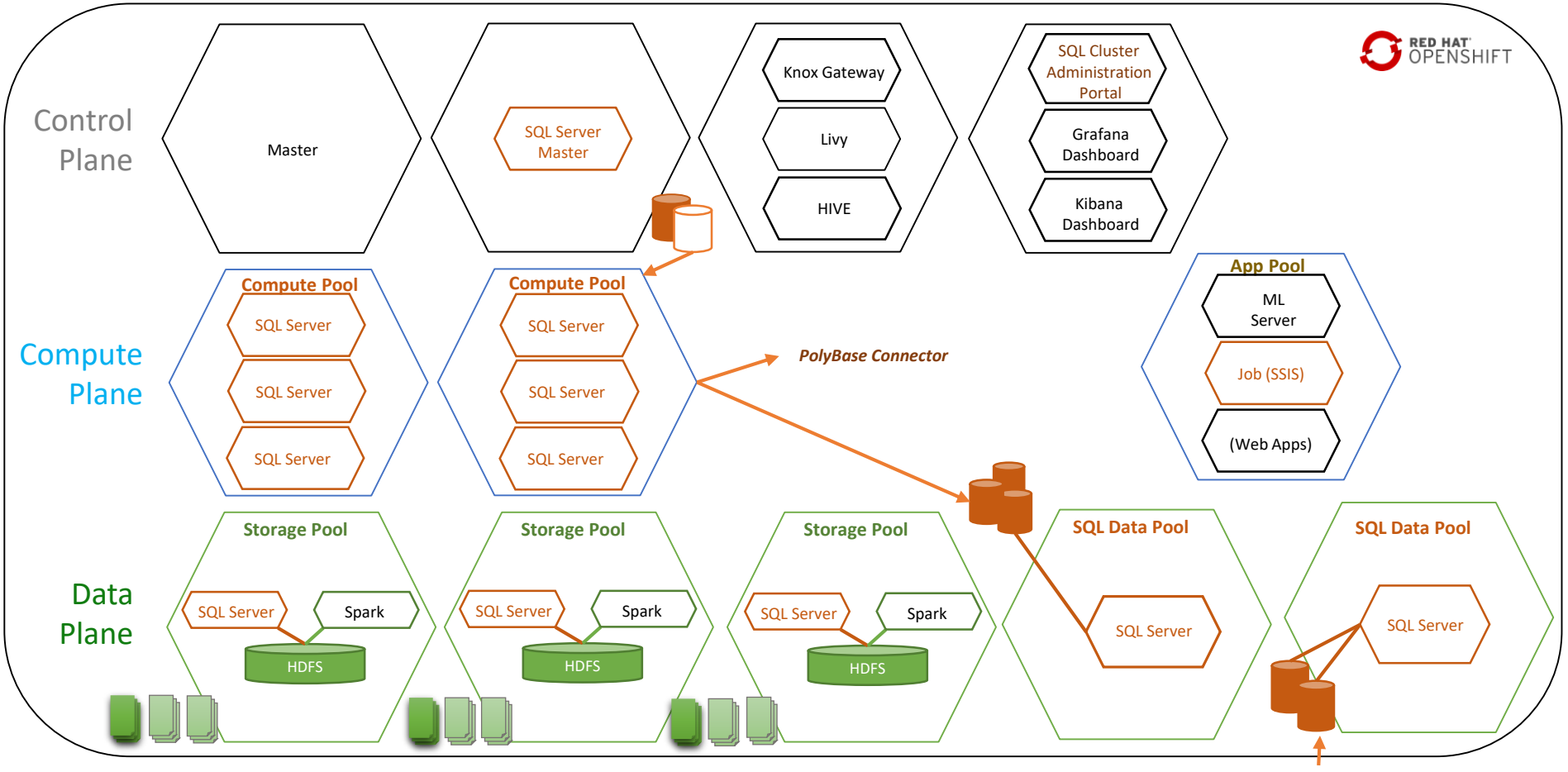
Microsoft SQL Server



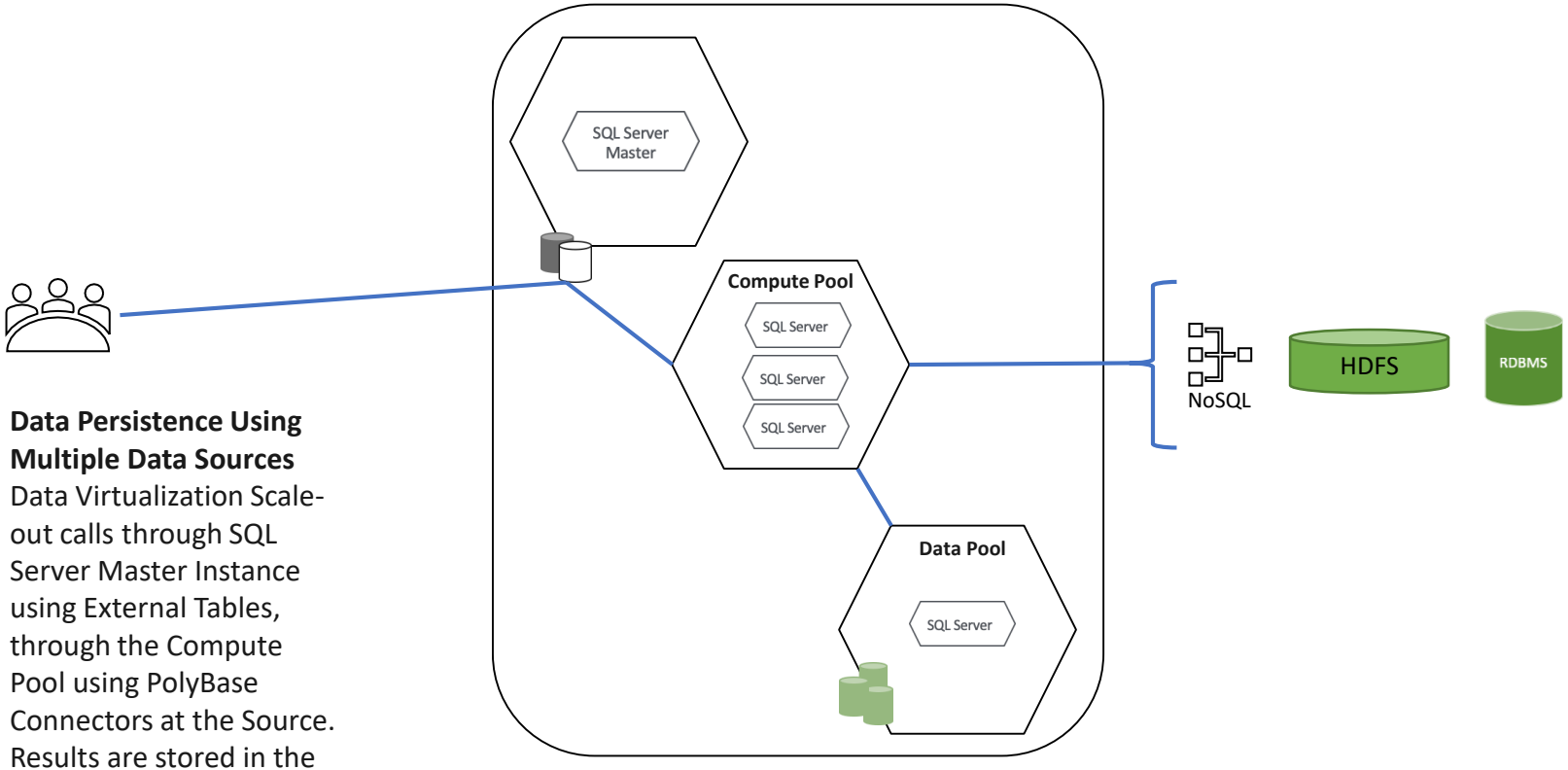
SQL Server 2019 Big Data Cluster – Data Mart

SQL Server 2019 big data cluster

Microsoft SQL Server



Data Mart

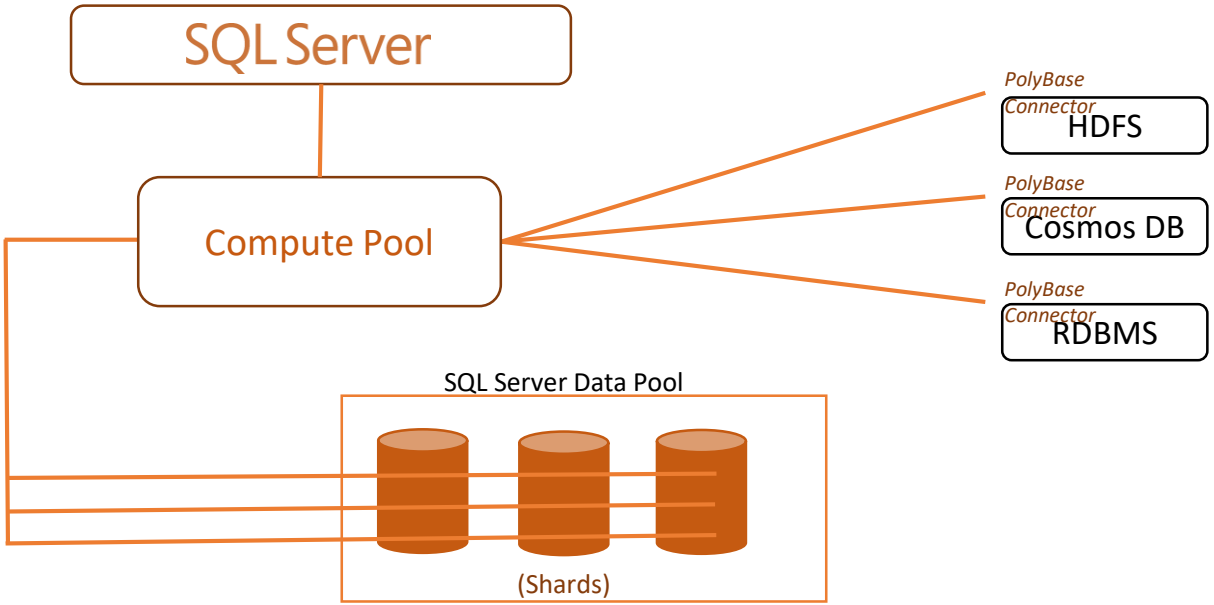


Data Persistence Using Multiple Data Sources
Data Virtualization Scale-out calls through SQL Server Master Instance using External Tables, through the Compute Pool using PolyBase Connectors at the Source. Results are stored in the Shards of the Data Pool.

SQL Server 2019 and Big Data

Data Mart

Microsoft SQL Server



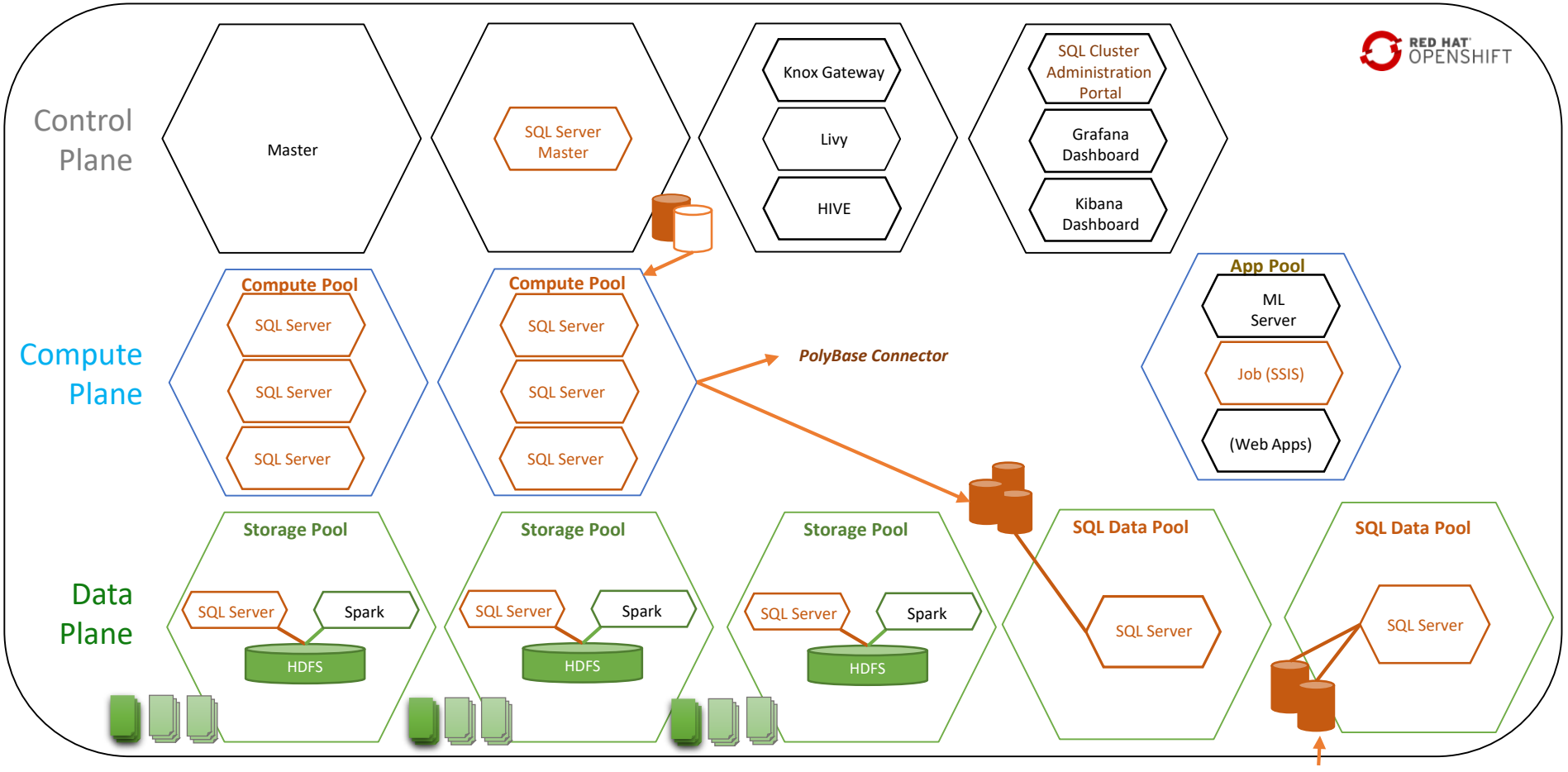
SQL Server Big Data Cluster – Data Mart

Example

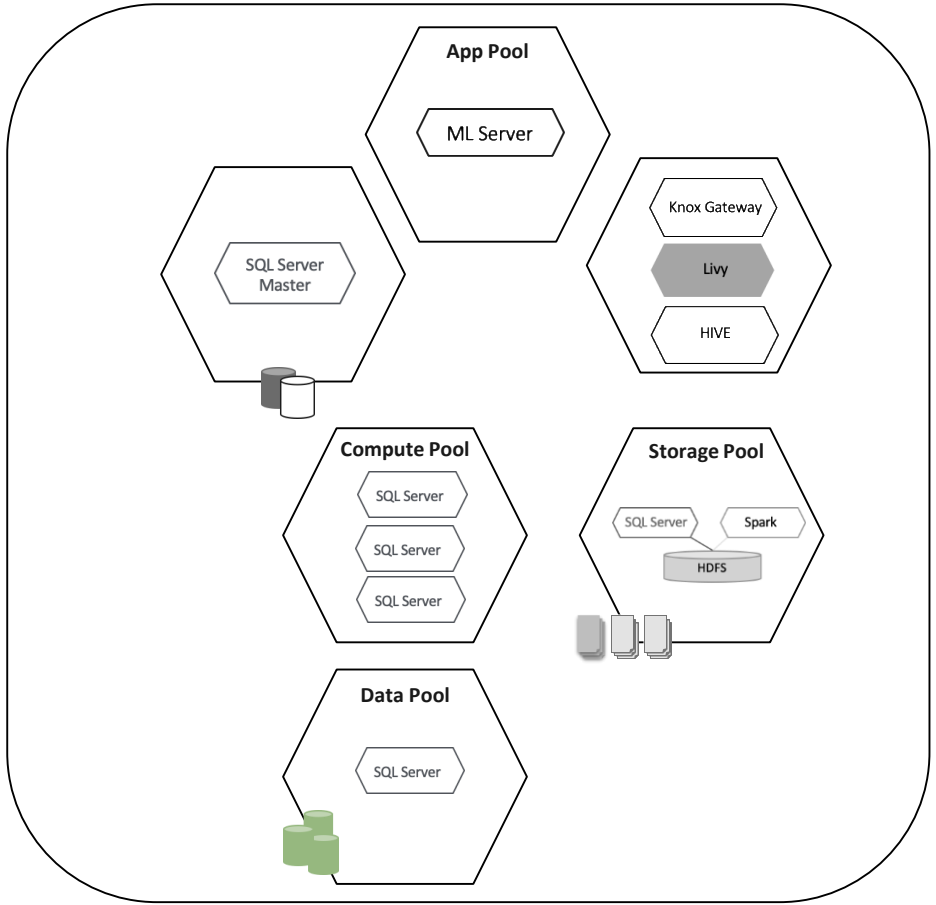
SQL Server 2019 Big Data Cluster – Data Lake, Machine Learning and Spark

SQL Server 2019 big data cluster

Microsoft SQL Server



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.

AI Enablement
Prediction and Classification Scoring to AI apps using the App 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

Spark Query Notebook

Example

SQL Server 2019 Big Data Cluster – Tools, Management and Monitoring

SQL Server Big Data Cluster – Management and Monitoring

Example

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

RED HAT SUMMIT

THANK YOU


linkedin.com/company/Red-Hat


facebook.com/RedHatinc

youtube.com/user/RedHatVideos


twitter.com/RedHat

THANK YOU

linkedin.com/company/Red-Hat

facebook.com/RedHatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat