SQL Server 2019
Big Data Clusters

Ben Weissman
@bweissman

Who am I?

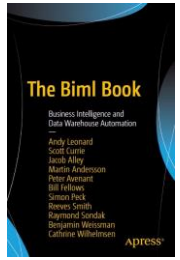
Ben Weissman
@bweissman
b.weissman@solisyon.de
http://biml-blog.de/

I ♥ DATA

DataGrillen
Data
Bratwurst
Beer

Please Talk Data To Me

SOLISYON



BimlHero
CERTIFIED EXPERT

MVP
Microsoft
Most Valuable
Professional

Microsoft
CERTIFIED
Solutions Associate
Machine Learning

Microsoft
CERTIFIED
Solutions Expert
Data Management and
Analytics

MICROSOFT
PROFESSIONAL
PROGRAM

Data Science
Big Data
Artificial Intelligence
Data Analysis

FRIEND OF
redgate
2019

GENESEE ACADEMY
World-Class Training

Certified Data Vault Modeler

Der Data Platform Podcast mit Biml Ben, Mr. T und Angry Frank

Please Talk Data To Me

Adaptive Query Processing ADF Azure Azure Data Studio **Azure Notebooks** Azure Stack Big Data Clusters **Biml** Black Panther Business Application Summit 2018 Data Platform **Data Platform Summit** dbatools **Docker** Flensburger Radler Alkoholfrei GDPR **Git Hub** Ignite Jupyter Notebooks Kubernetes **Las Vegas** Lissabon Microsoft Professional Program MPP **PASS Camp** PASS Deutschland e.V. PASS Essentials PASS Summit **Power BI** PowerShell Query Folding Regionalgruppen Solo SQL Management Studio 18 - Preview SQL Operations Studio **SQL Saturday** SQL Server 2019 Tabular Tomb Raider **tSQLt** TugaIT Visual Studio Code **WDC**

Ben Weissman
Biml Ben

Tillmann Eitelberg
Mr. T

Frank Geisler
Angry Frank

11	2396	1147	12
Episoden	Downloads	Sendeminuten	Gäste

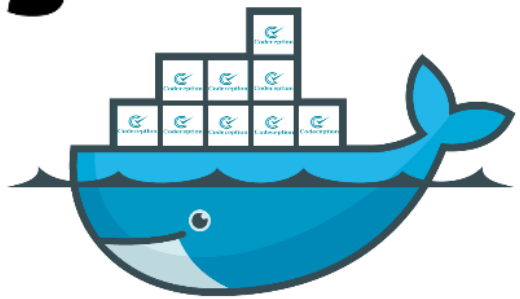
https://www.pleasetalkdatatome.de

Fun Facts on Big Data Clusters

- Some parts only run on Linux
- It's a „box product first" feature set
- It's actually not ONE feature but a huge feature set
- It's name is a bit misleading – not all of it is a cluster
- Some parts are currently in semi-private preview

What to look at before getting started…

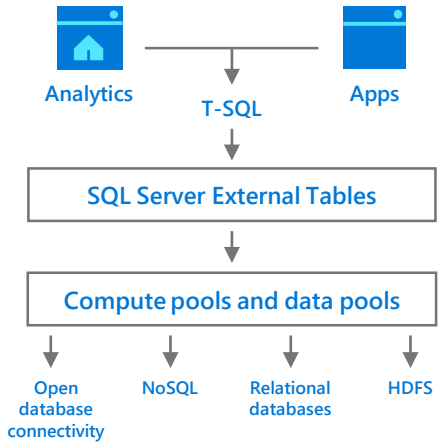
kubernetes



docker

So what is a Big Data Cluster in SQL 2019?!

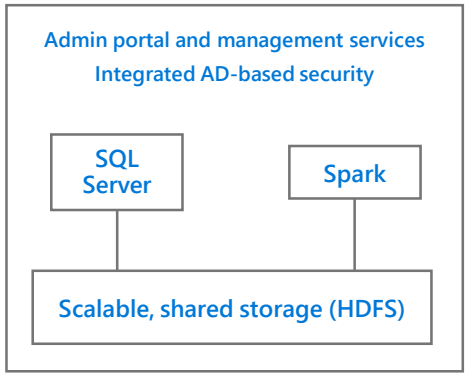
Data virtualization



Combine data from many sources without moving or replicating it

Scale out compute and caching to boost performance

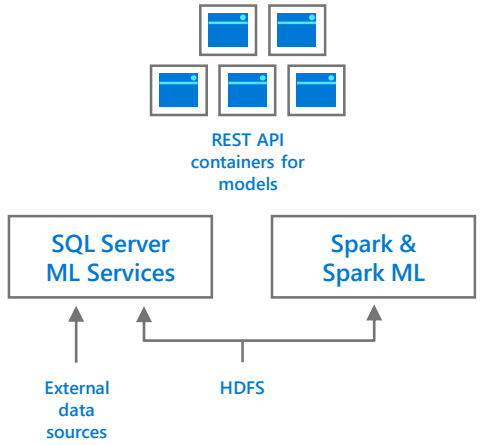
Managed SQL Server, Spark, and data lake



Store high volume data in a data lake and access it easily using either SQL or Spark

Management services, admin portal, and integrated security make it all easy to manage

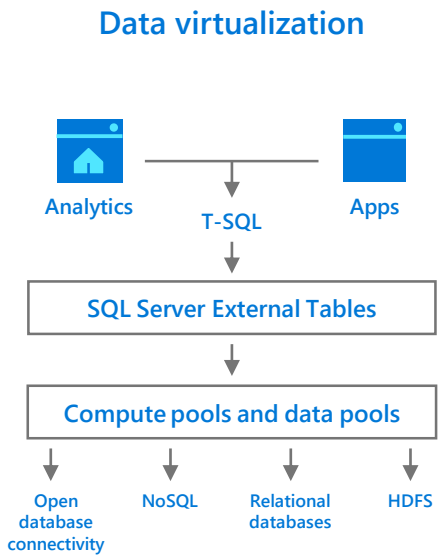
Complete AI platform



Easily feed integrated data from many sources to your model training

Ingest and prep data and then train, store, and operationalize your models all in one system

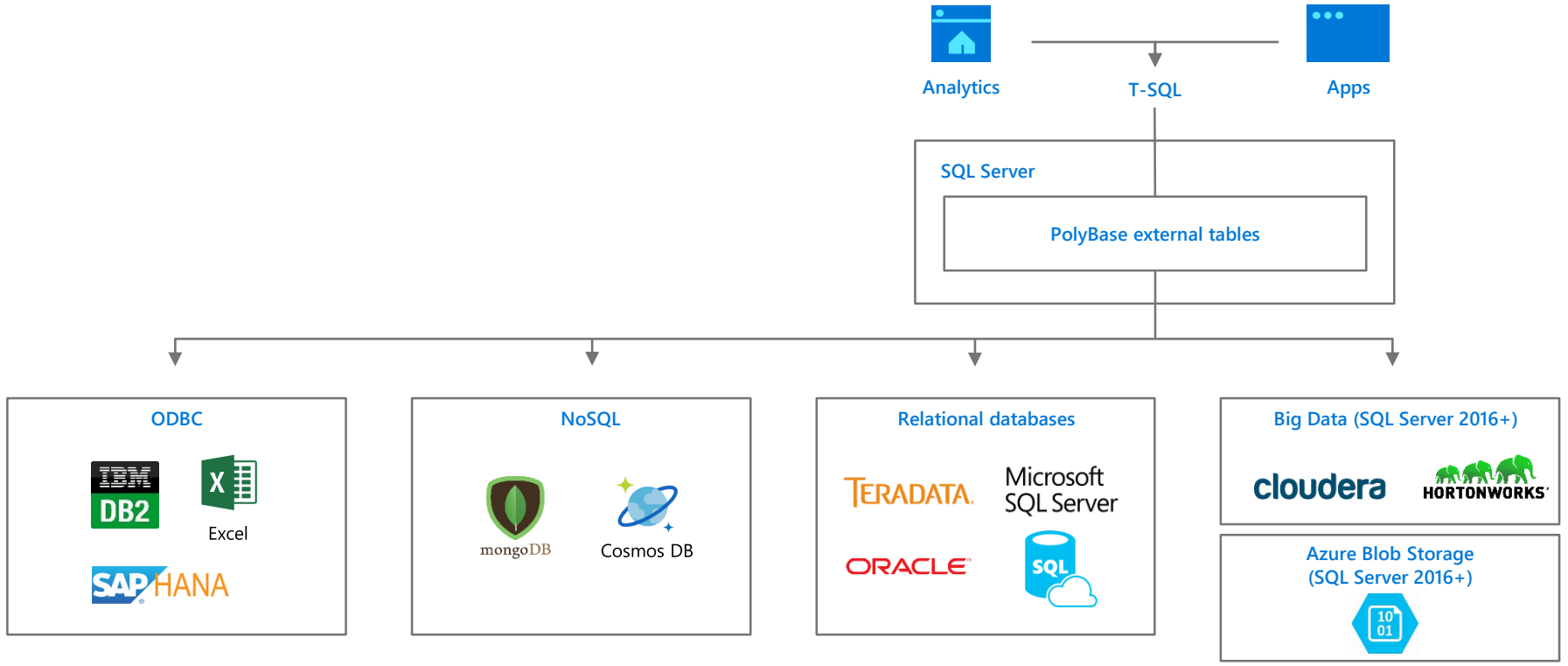
Data Virtualization – Is this the END of SSIS?!



Combine data from many sources without moving or replicating it

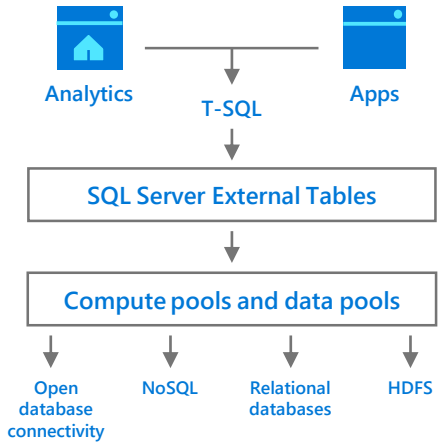
Scale out compute and caching to boost performance

Easily combine across relational and non-relational data stores



Data Virtualization – So it‘s a linked server?

Data virtualization



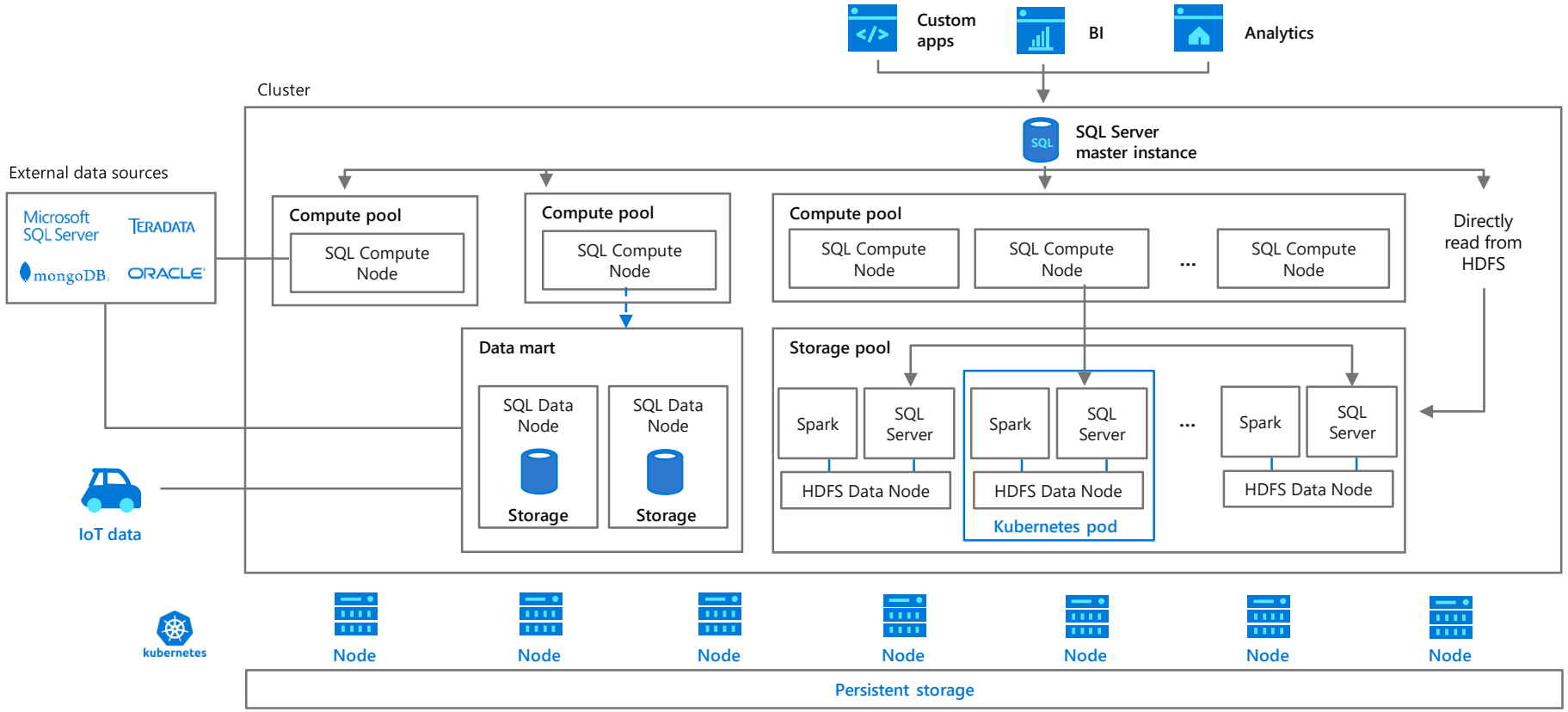
Combine data from many sources without moving or replicating it

Scale out compute and caching to boost performance

Linked Servers

PolyBase External tables

So why do they call it a cluster?!



This slide: © by Microsoft

How can I get it installed? – Polybase only

- Install Java JRE
- Get the latest CTP from http://microsoft.com/sql
- Install SQL Server on Windows or Linux including Polybase
- Use EVALUATION edition!
- Enable Polybase after installation:

```
exec sp_configure @configname = 'polybase enabled', @configvalue = 1;
RECONFIGURE
```

- Restart SQL Server
- Install Azure Data Studio
- Install the vNext Extension for Azure Data Studio

How can I get it installed? – The full package

- Sign up for the preview program: https://aka.ms/eapsignup
- Install Kubernetes-CLI, MSSQLCTL, Python, azure-cli, curl*
- Install Azure Data Studio
  - Add vNext Extension
- Decide on a Kubernetes environment
  - Docker or Minikube
  - AKS
  - Something completely different ☺
  - (many but not all are supported)
- Set environment variables**
- Deploy the cluster using

  mssqlctl create cluster …
- When using AKS, consider this script:

  https://github.com/Microsoft/sql-server-samples/tree/master/samples/features/sql-big-data-cluster/deployment

Picture: © Klaus Aschenbrenner

* Prerequisits

```
Set-ExecutionPolicy Bypass -Scope Process -Force; iex ((New-Object System.Net.WebClient).DownloadString('https://chocolatey.org/install.ps1'))
choco install azure-cli -y
choco install azure-data-studio -y
choco install python3 -y
choco install notepadplusplus -y
$env:Path = [System.Environment]::GetEnvironmentVariable("Path","Machine") + ";" + [System.Environment]::GetEnvironmentVariable("Path","User")
python -m pip install --upgrade pip
python -m pip install requests
python -m pip install requests --upgrade
choco install curl -y
choco install 7zip -y
choco install kubernetes-cli -y
pip3 install kubernetes
pip3 install  -r https://private-repo.microsoft.com/python/ctp3.0/mssqlctl/requirements.txt
```

** Environment Variables

SET CONTROLLER_USERNAME=<controller_admin_name - can be anything>
SET CONTROLLER_PASSWORD=<controller_admin_password - can be anything, password complexity compliant>
SET KNOX_PASSWORD=<knox_password - can be anything, password complexity compliant>
SET MSSQL_SA_PASSWORD=<sa_password_of_master_sql_instance - can be anything, password complexity compliant>

SET DOCKER_REGISTRY=private-repo.microsoft.com
SET DOCKER_REPOSITORY=mssql-private-preview
SET DOCKER_USERNAME=<your username, credentials provided by Microsoft>
SET DOCKER_PASSWORD=<your password, credentials provided by Microsoft>
SET DOCKER_EMAIL=<your Docker email, use the username provided by Microsoft>
SET DOCKER_PRIVATE_REGISTRY="1"

It's Demo Time: Polybase on prem

- Creating external tables from SQL Server Sources using Azure Data Studio
  - Master Key
  - Credentials
  - Data Source
  - Tables
- Automating external tables with Biml* ☺

*https://www.solisyon.de/biml-polybase-external-tables/

It‘s Demo Time: Kubernetes Cluster

- Deploying a cluster with some sample data*
- The Cluster Portal
- Play with it using T-SQL
  - Query HDFS Data
  - Write/Read Data from Data Pool
- Play with it using Notebooks
  - Read/Analyze Data with Spark

*https://github.com/Microsoft/sql-server-samples/tree/master/samples/features/sql-big-data-cluster/

Kubernetes Cluster – Install Cluster (incl. AKS)

```
SQL Server big data cluster connection endpoints:
SQL Server master instance:
IP              PORT
40.113.127.13   31433

HDFS/KNOX:
IP              PORT
13.94.244.250   30443

Cluster administration portal (https://<ip>:<port>):
IP             PORT
40.68.84.89    30777
```

*If you forget about these…
kubectl get service -n <clustername>

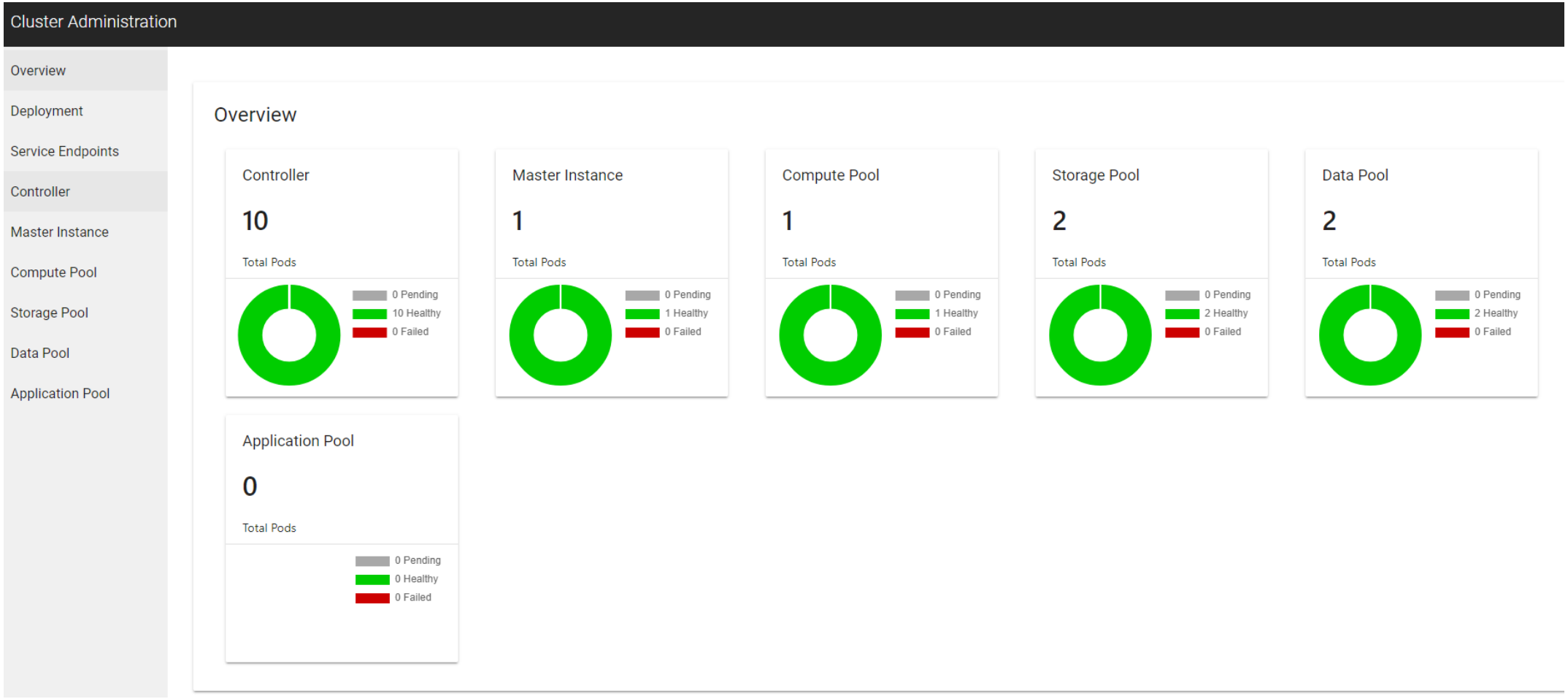
Kubernetes Cluster – Install Sample Data

.\bootstrap-sample-db.cmd

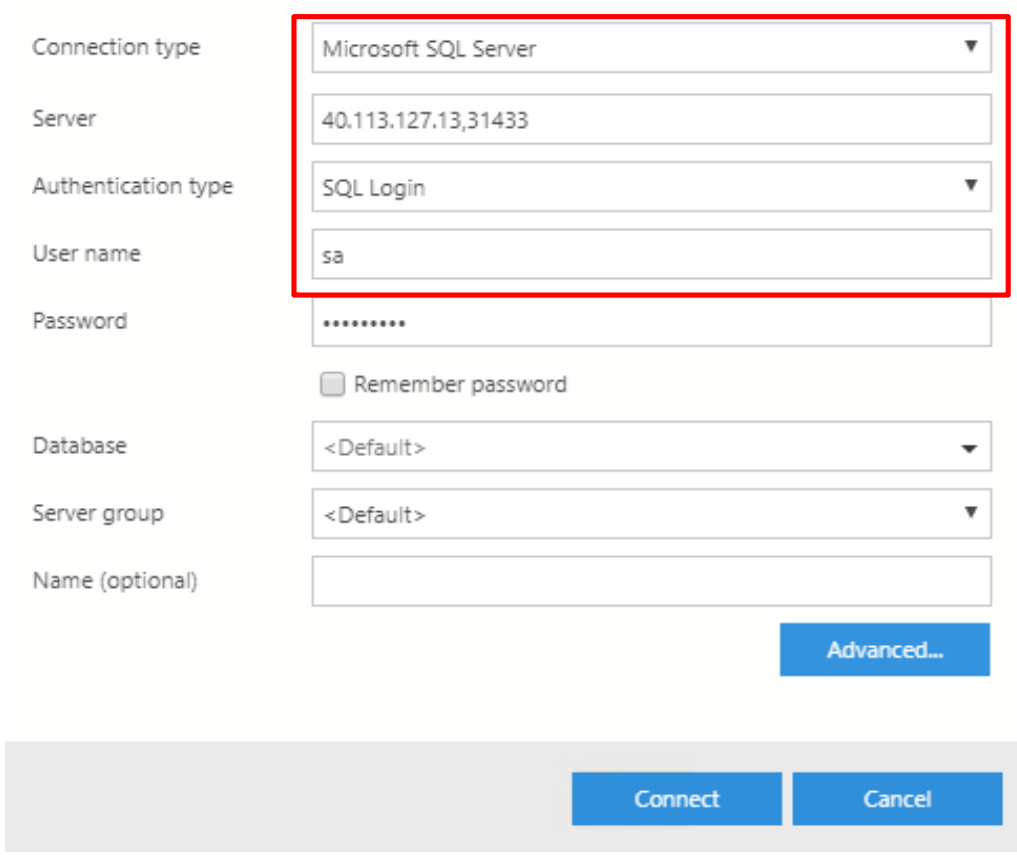
USAGE: .\bootstrap-sample-db.cmd <CLUSTER_NAMESPACE> <SQL_MASTER_IP> <SQL_MASTER_SA_PASSWORD> <BACKUP_FILE_PATH> <KNOX_IP> [<KNOX_PASSWORD>]
Default ports are assumed for SQL Master instance & Knox gateway.

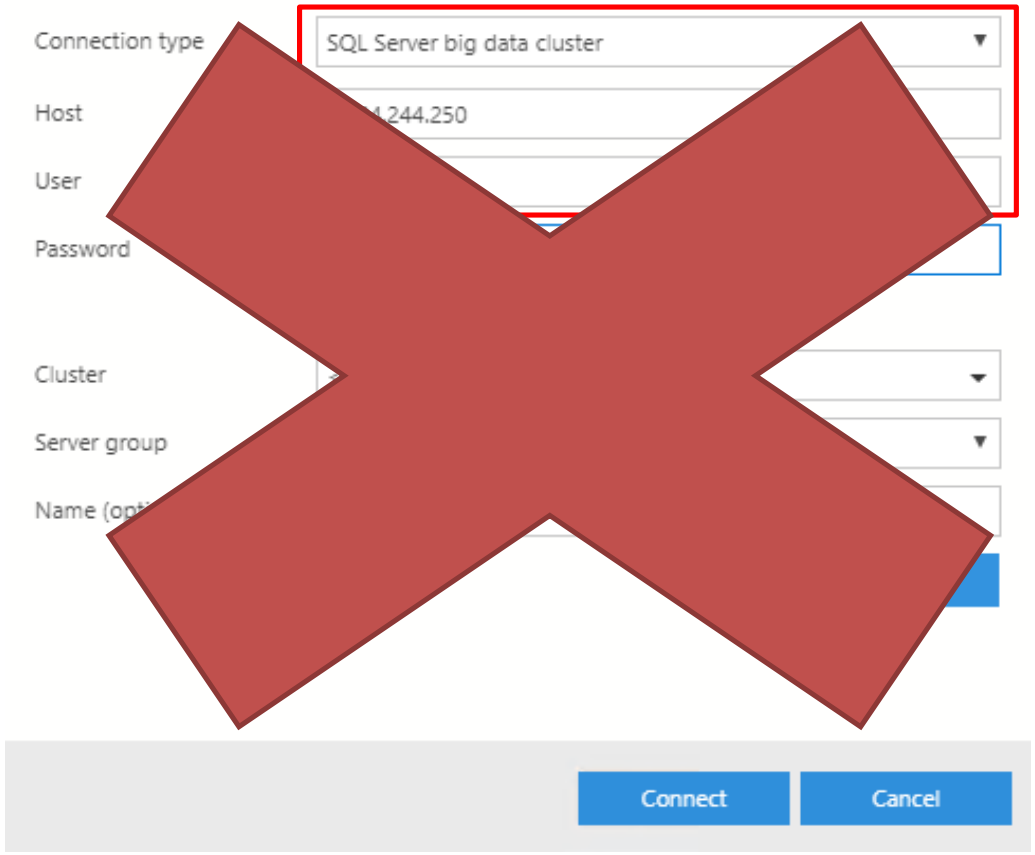
https://github.com/Microsoft/sql-server-samples/tree/master/samples/features/sql-big-data-cluster

Kubernetes Cluster – Portal



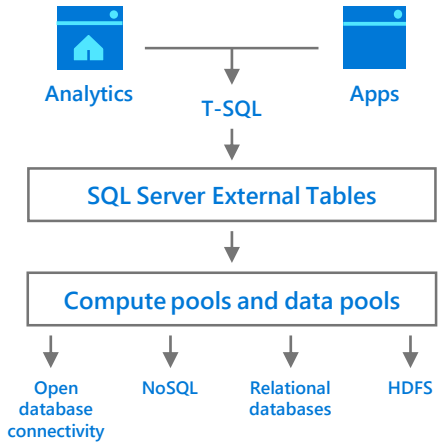
Kubernetes Cluster – Connect in ADS





So what is a Big Data Cluster in SQL 2019?!

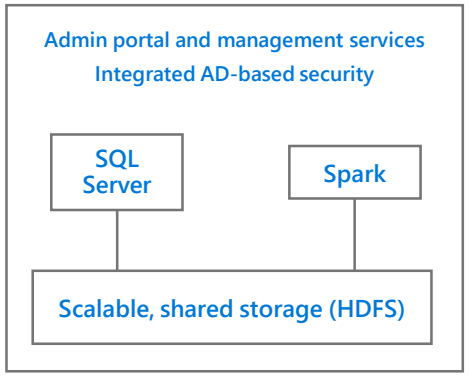
Data virtualization



- No Data redundancy
- Real time data

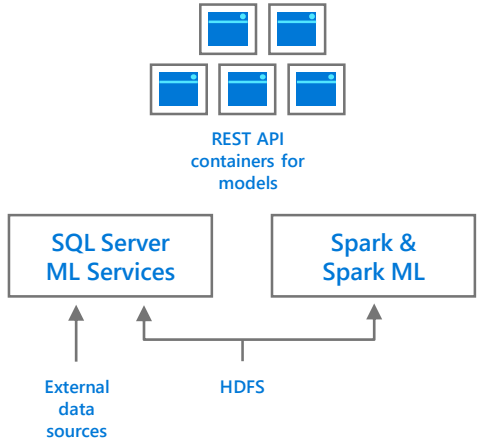
- No extra indexing
- Extra load on source
- Read only

Managed SQL Server, Spark, and data lake



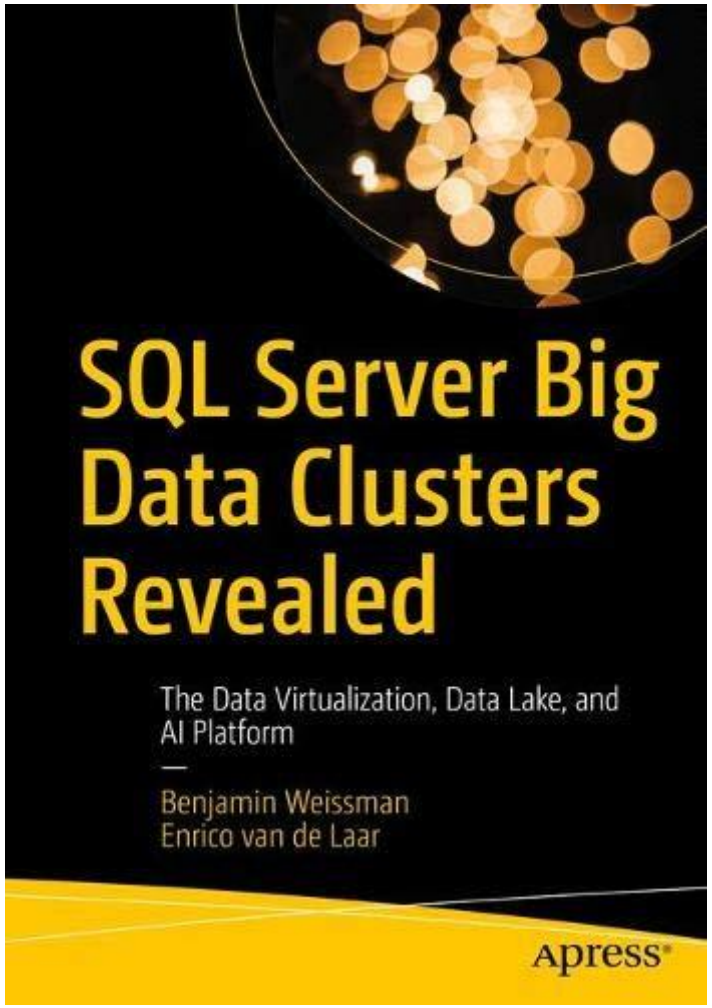
- Store high volume data in a data lake and access it easily using either SQL or Spark
- Management services, admin portal, and integrated security make it all easy to manage

Complete AI platform



- Easily feed integrated data from many sources to your model training
- Ingest and prep data and then train, store, and operationalize your models all in one system

Announcing


SQL Server Big Data Clusters Revealed
The Data Virtualization, Data Lake, and AI Platform
Benjamin Weissman
Enrico van de Laar
Apress

Any questions?

Ben Weissman
@bweissman