Ben Weissman

@bweissman

PASS SQLSATURDAY VIENNA

Big Data Clusters

Make SQL Server your Data Hub for everything

Who am I?

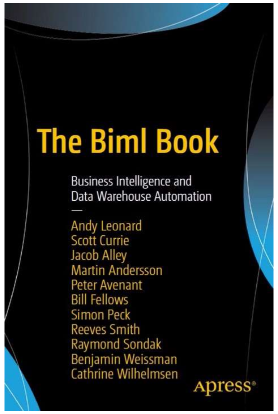
Ben Weissman, Solisyon, Nuernberg/Germany

@bweissman

b.weissman@solisyon.de

SQL Server since 6.5

Data Passionist



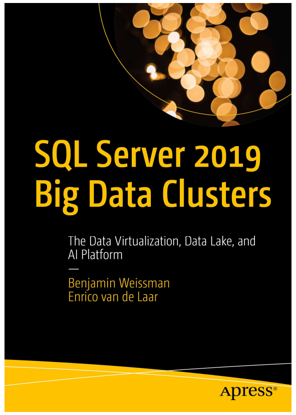
The Biml Book

Business Intelligence and Data Warehouse Automation

—

Andy Leonard
Scott Currie
Jacob Alley
Martin Andersson
Peter Avenant
Bill Fellows
Simon Peck
Reeves Smith
Raymond Sondak
Benjamin Weissman
Cathrine Wilhelmsen

Apress®



SQL Server 2019 Big Data Clusters

The Data Virtualization, Data Lake, and AI Platform

—

Benjamin Weissman
Enrico van de Laar

Apress®

Automating Custom SSIS Tasks in Biml

© 2019 Ben Weissman / Tillmann Eitelberg

Please Talk Data To Me

BimlHero CERTIFIED EXPERT

Microsoft CERTIFIED Solutions Associate

Machine Learning

Microsoft CERTIFIED Solutions Expert

Data Management and Analytics

MVP Microsoft® Most Valuable Professional

MICROSOFT PROFESSIONAL PROGRAM

Data Science

Big Data

Artificial Intelligence

Data Analysis

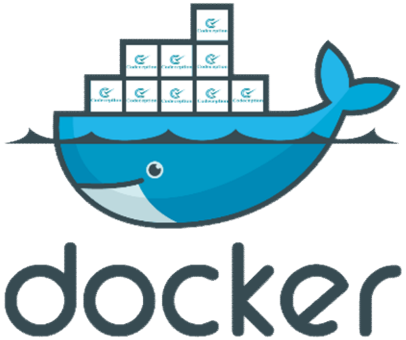
FRIEND OF redgate 2019

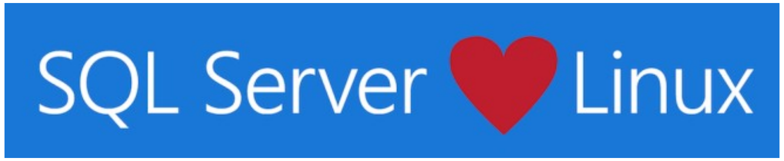
GENESEE ACADEMY World-Class Training

Certified Data Vault Modeler

What to look at before getting started…

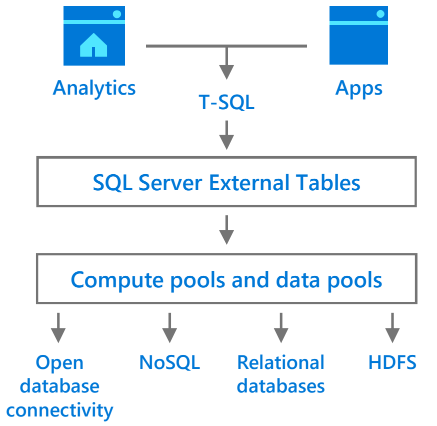
kubernetes





So, what is a SQL 2019 Big Data Cluster?

Data Virtualization



Analytics

Apps

T-SQL

SQL Server External Tables

Compute pools and data pools

Open database connectivity

NoSQL

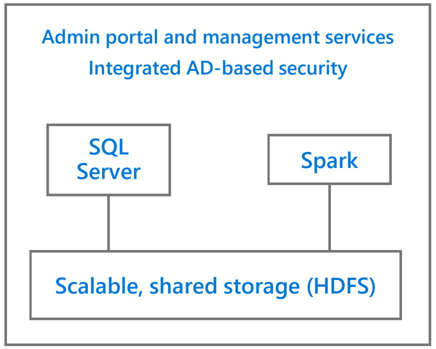
Relational databases

HDFS

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



Admin portal and management services

Integrated AD-based security

SQL Server

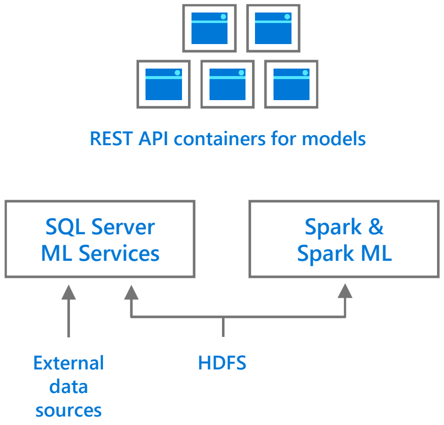
Spark

Scalable, shared storage (HDFS)

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

AI/ML Platform



REST API containers for models

SQL Server ML Services

Spark & Spark ML

External data sources

HDFS

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

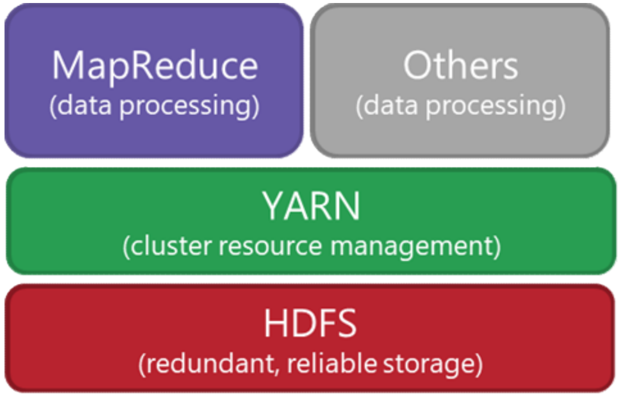
This slide: © by Microsoft

A little primer…

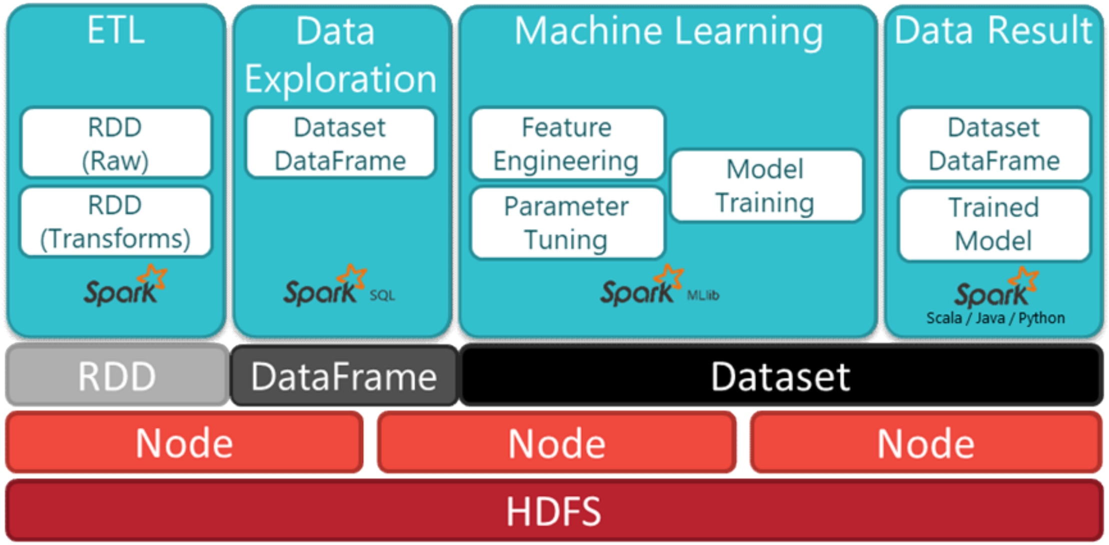
Scaled Processing and Scaled Storage

The foundations of scale

Hadoop



Spark



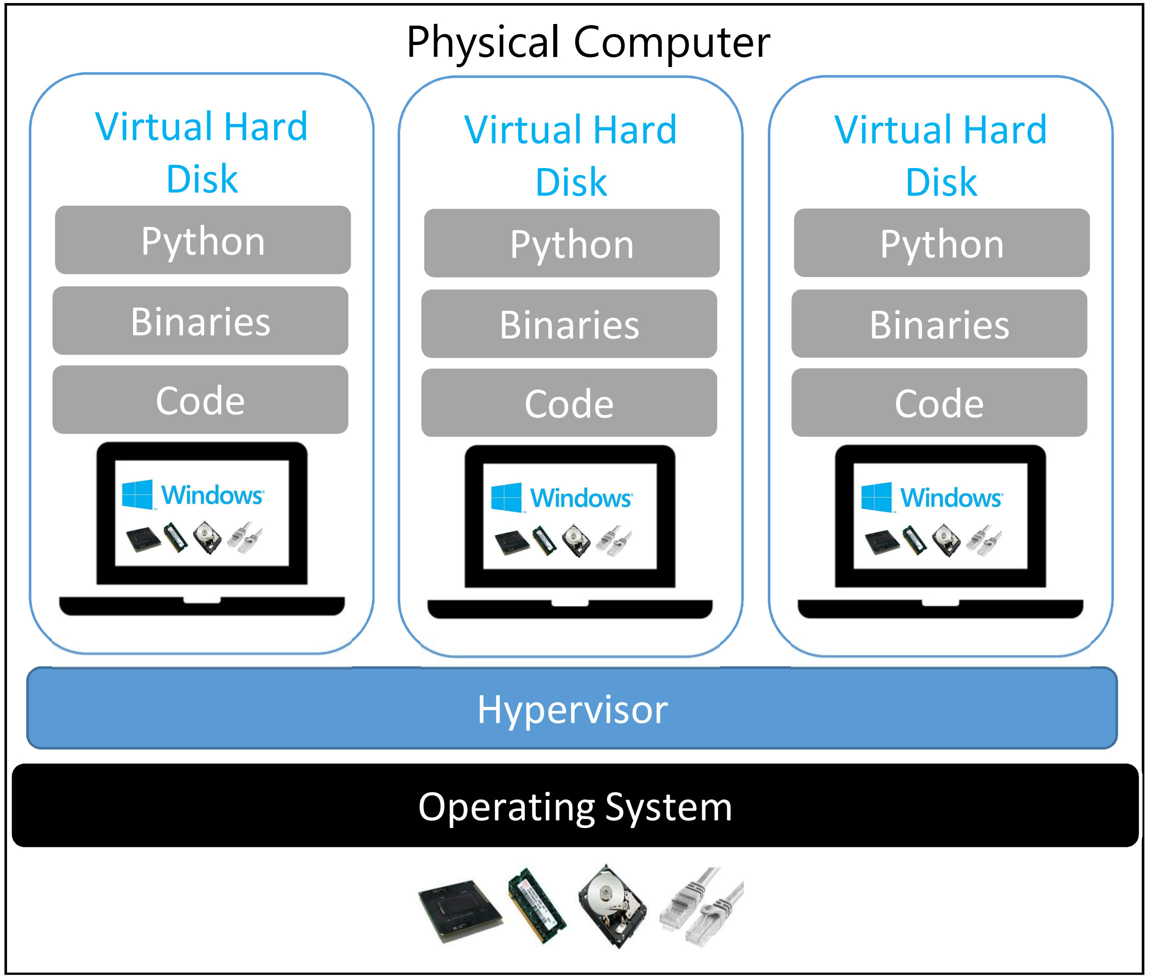
Virtualization

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

Hyper-V



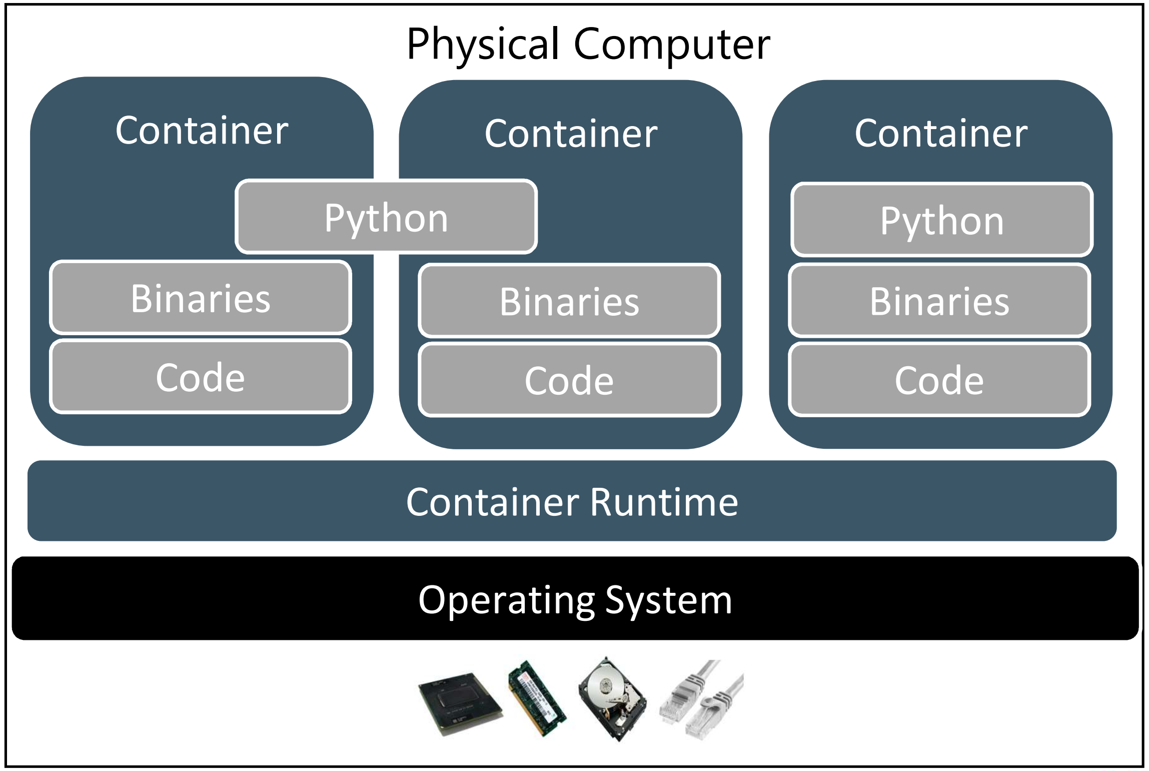
Containers

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

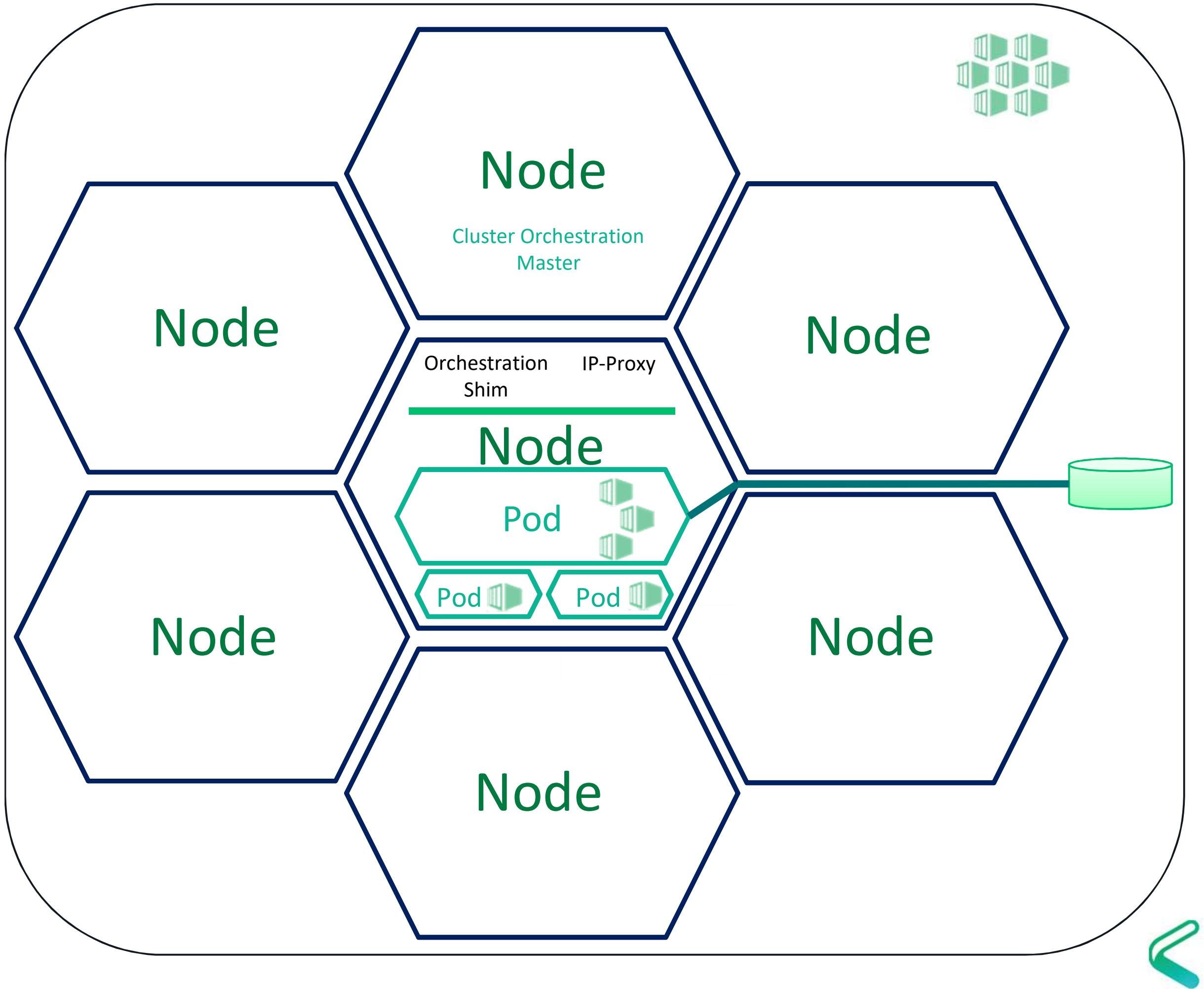
Scale is achieved through multiple containers



Container Orchestration

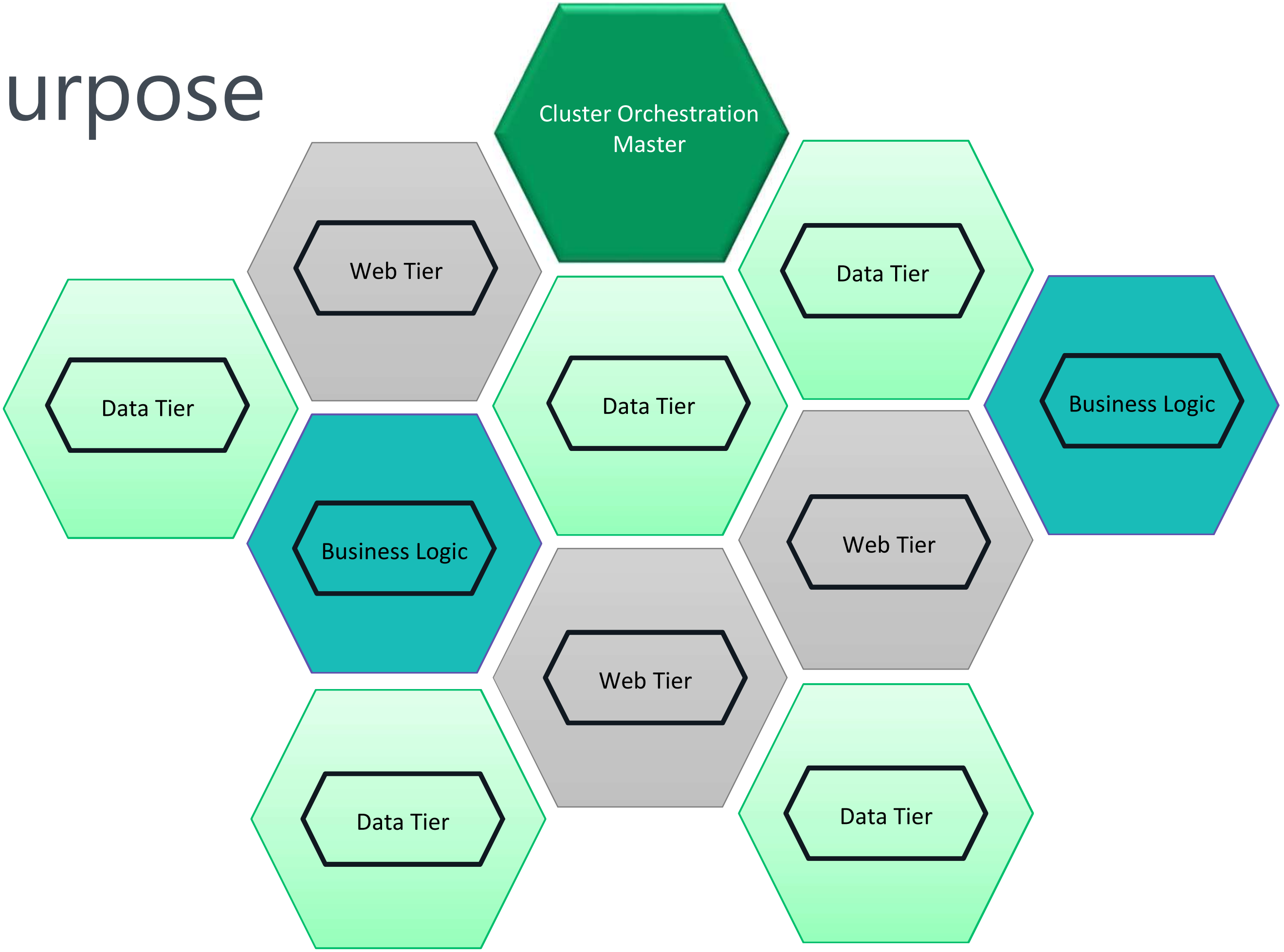
Containers at Scale

- **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*



Generic Cluster

Scale by Purpose



Want to learn more…

…without all that tech stuff?

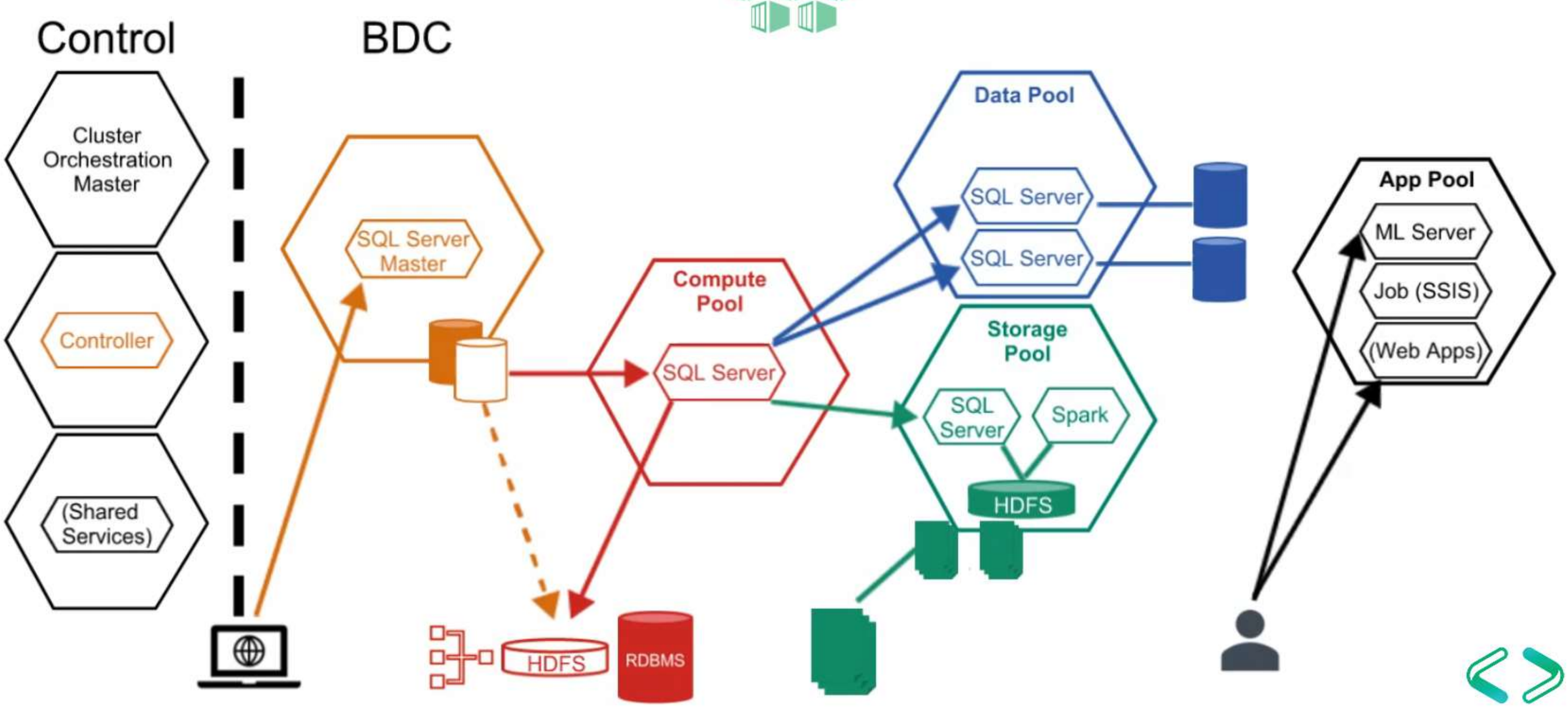
Or talk to this guy…

@nocentino

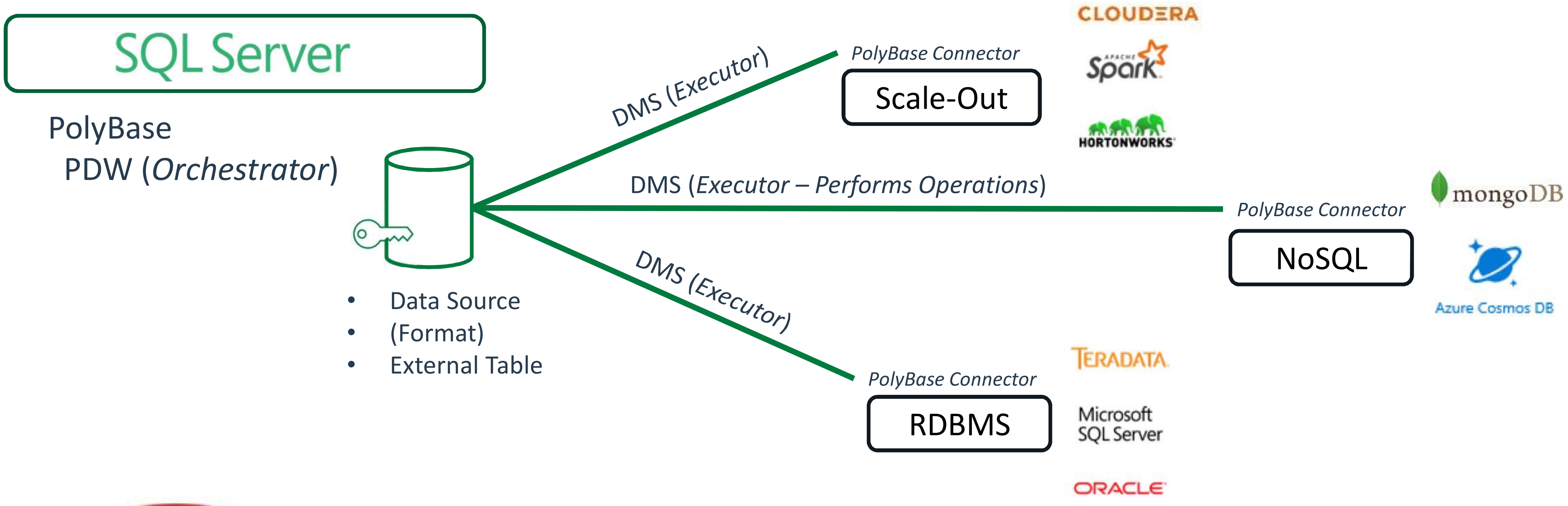
https://www.cncf.io/wp-content/uploads/2019/07/The-Illustrated-Childrens-Guide-to-Kubernetes.pdf

Complete Architecture

OLTP, Data Virtualization, Data Mart and Big Data



Data Virtualization



SQL Server

PolyBase

PDW (*Orchestrator*)

- Data Source
- (Format)
- External Table

DMS (*Executor*)

PolyBase Connector

Scale-Out

DMS (*Executor – Performs Operations*)

PolyBase Connector

NoSQL

DMS (*Executor*)

PolyBase Connector

RDBMS

DEMO

PolyBase in SQL 2019

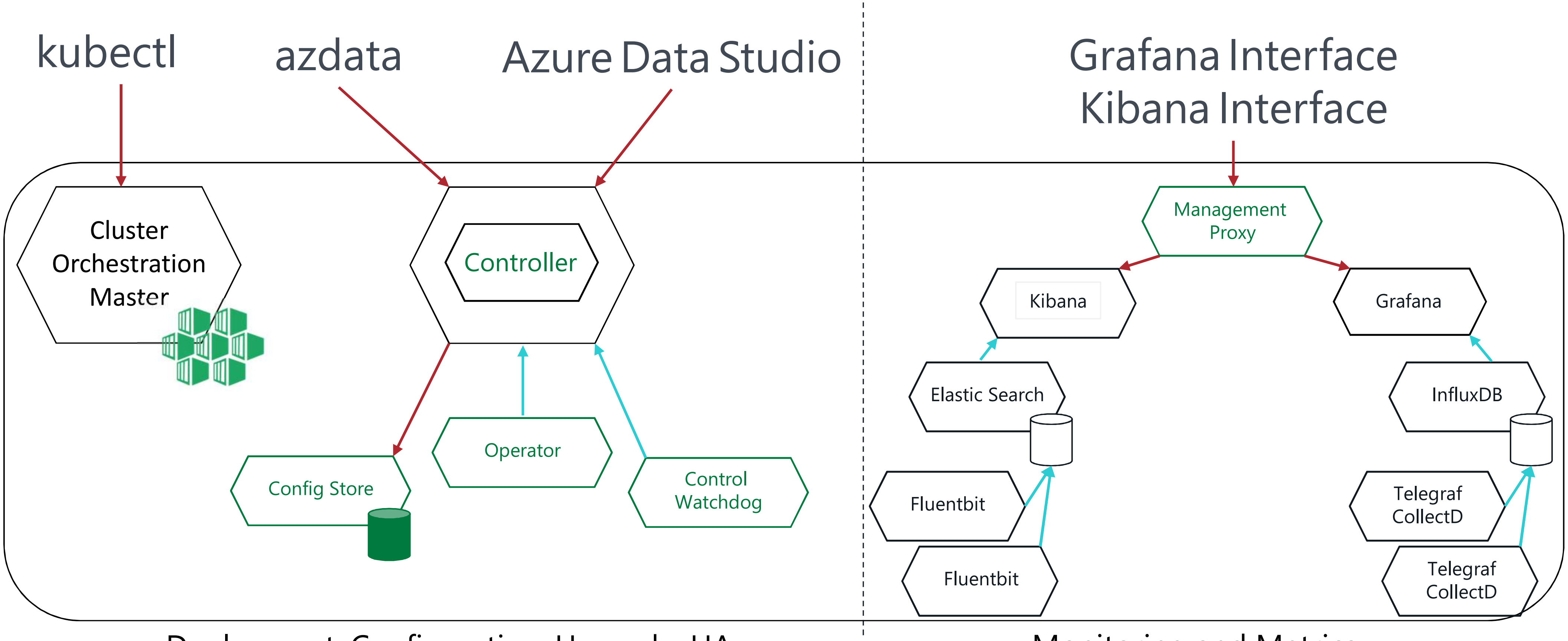
Storage Pool

Data Pool

Spark Queries

Tools, Management and Monitoring

Managing the Big Data Cluster



Deployment, Configuration, Upgrade, HA

Monitoring and Metrics

Deployment

How can I get it installed?

PolyBase only

- Get SQL 2019 from http://microsoft.com/sql
- Install SQL Server on Windows or Linux including PolyBase
- Enable PolyBase after installation:

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

- Restart SQL Server
- Install Azure Data Studio and Data Virtualization Extension

How can I get it installed?

The full package

- Decide on a Kubernetes environment (AKS, kubeadm, …)
- Install Azure Data Studio and Data Virtualization Extension
- Install Prerequisites*
- Deploy the cluster using azdata/Azure Data Studio

* Prerequisites

```
Set-ExecutionPolicy Bypass -Scope Process -Force; iex ((New-Object System.Net.WebClient).DownloadString('https://chocolatey.org/install.ps1‘))
choco install notepadplusplus -y
choco install 7zip -y
choco install curl -y
choco install sqlserver-cmdlineutils -y
choco install azure-cli -y
choco install azure-data-studio -y
choco install python3 -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 kubernetes-cli -y
pip3 install Kubernetes
choco install visualcpp-build-tools -y
pip3 install -r https://aka.ms/azdata
```

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

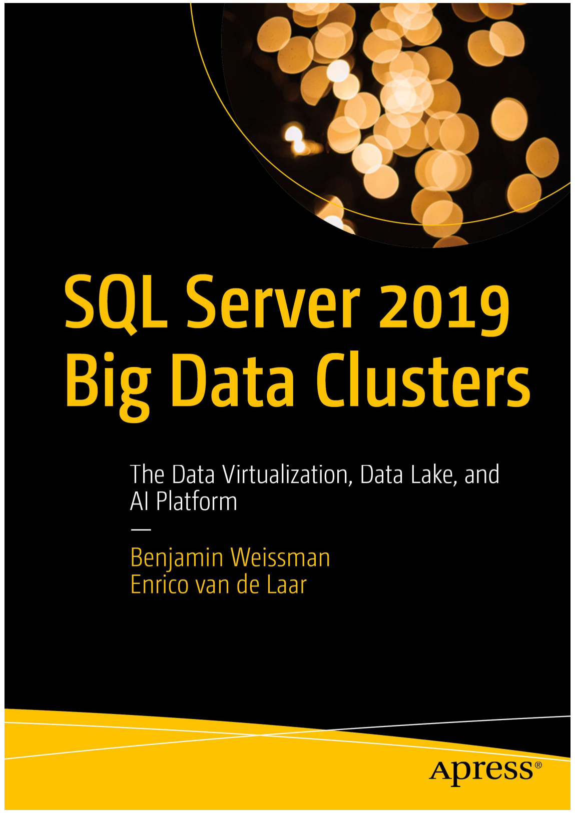
Questions?

Ben Weissman

🐦 @bweissman

linkedin.com/in/weissmanben/

Thank you for your time!



SQL Server 2019 Big Data Clusters

The Data Virtualization, Data Lake, and AI Platform

—

Benjamin Weissman
Enrico van de Laar

Apress®