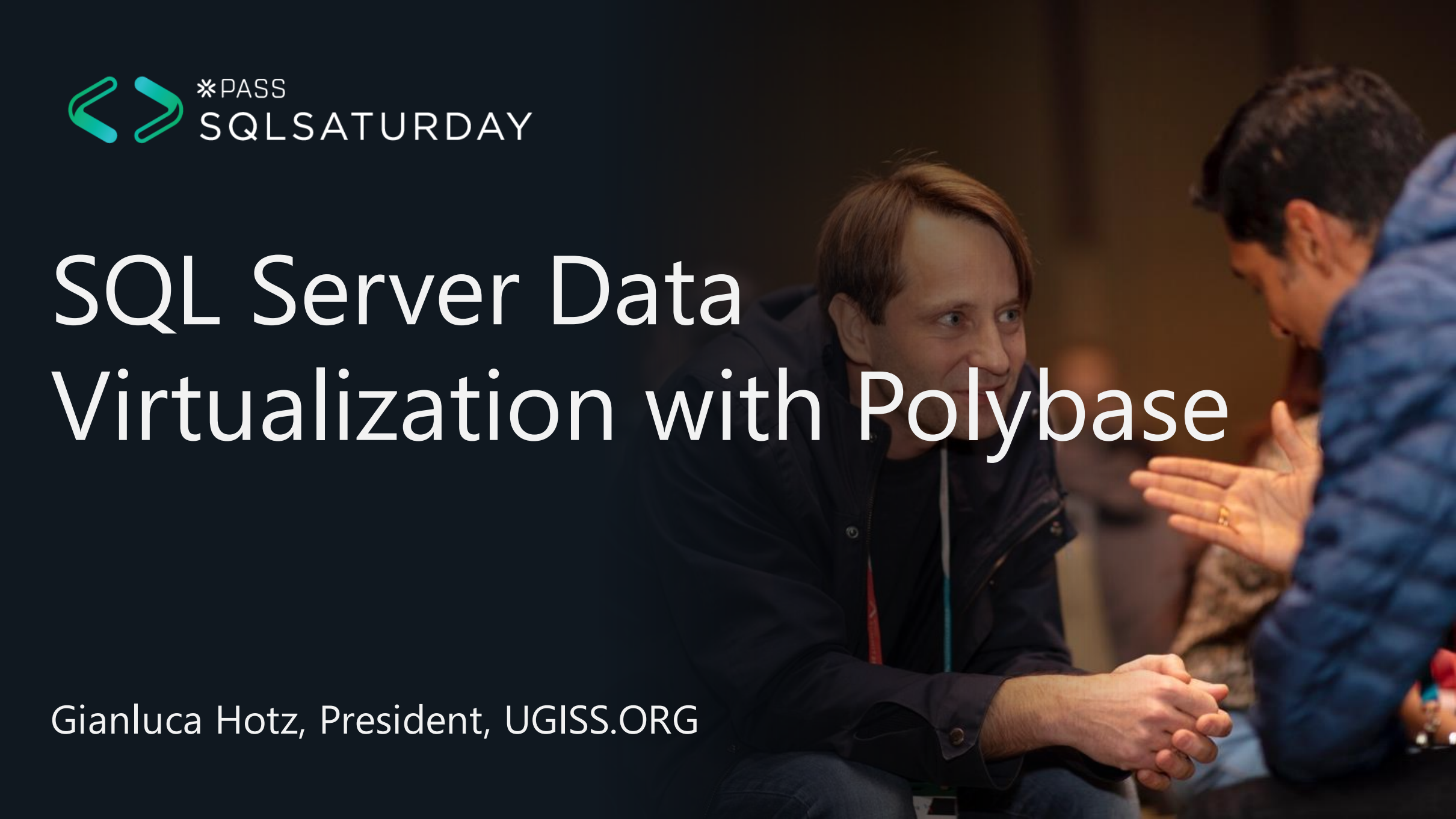


PASS SQLSATURDAY

SQL Server Data Virtualization with Polybase

Gianluca Hotz, President, UGISS.ORG

Own your career with interactive learning built by community and guided by data experts. Get involved. Get ahead.

Explore your PASS community

PASS MARATHON

Free online webinar events

PASS LOCAL GROUPS

Local user groups around the world

PASS SUMMIT

Connect with the global data community

PASS VIRTUAL GROUPS

Online special interest user groups

PASS.org

PASS CONNECTOR INSIGHTS

Learning on-demand and delivered to you

PASS VOLUNTEERS

Get involved

VIRTUAL

PASS SQLSATURDAY PORDENONE | 30 MAY 2020

Missed PASS Summit 2019?

Get the Recordings

Download all PASS Summit sessions on Data Management, Analytics, or Architecture for only $399 USD

More options available at
PASSstuff.com

We are covering all bases to ensure our community can continue reaching new and exciting heights. Plans are underway for the in-person event you all know and love along with a new venture, a new opportunity: a PASS Summit 2020 Virtual Event.

Find out more at PASS.org/summit

PASS SQLSATURDAY

Thank you to our Global Sponsors and Supporters

Microsoft Azure

DELL Technologies

Google Cloud

IDERA

Quest

SentryOne

vmware

VIRTUAL PASS SQLSATURDAY PORDENONE | 30 MAY 2020

PASS SQLSATURDAY

Thank you to our Local Sponsors and Supporters

Microsoft

This event was sponsored by Microsoft

Learn more about SQL Server 2019 today:

-Get free training: aka.ms/sqlworkshops

-Download the SQL19 eBook: aka.ms/sql19_ebook

Chi sono?

Gianluca Hotz | @glhotz | ghotz@ugiss.org

Consulente indipendente

- 20+ anni su SQL Server (dalla 4.21 nel 1996)
- Modellazione e sviluppo database, dimensionamento e amministrazione database server, aggiornamenti e migrazioni, performance tuning

Community

- 20+ anni Microsoft MVP SQL Server/Data Platform (dal 1998)
- VMware Experts SQL Server
- Fondatore e presidente UGISS (PASS Chapter)
- Co-organizzatore DAMAG Meetup Community

Problema

Sistemi DSS/OLAP/DWH necessitano di processi

- ETL (Extract, Transform, Load)
- ELT (Extract, Load, Transform)

Causano «Data Movement»!

Mitigato da virtualizzazione accesso ai dati

Data Movement

- Costi
  - Costi storage duplicati
  - Sforzo per costruire/mantenere pipeline dati
- Velocità
  - Tempo integrazione nuovi dati -> sfuggono opportunità!
  - Latenza disponibilità dati -> sfuggono opportunità
- Sicurezza
  - Superfice di attacco più ampia
  - Potenziale inconsistenza politiche di sicurezza
- Qualità
  - Pipeline dati possono introdurre problemi di qualità…
- Compliance
  - Maggiori problemi di governance

Virtualizzazione

https://en.wikipedia.org/wiki/Data_virtualization

"Unlike the traditional extract, transform, load (ETL) process, the data remains in place, and real-time access is given to the source system for the data. This reduces the risk of data errors, of the workload moving data around that may never be used, and it does not attempt to impose a single data model on the data"

- Dati rimangono dove sono
- Viene dato accesso in tempo reale al dato
- Non si cerca di imporre un solo modello…

Virtualizzazione e SQL Server

Open Data Service Gateway (< SQL Server 7.0!!)

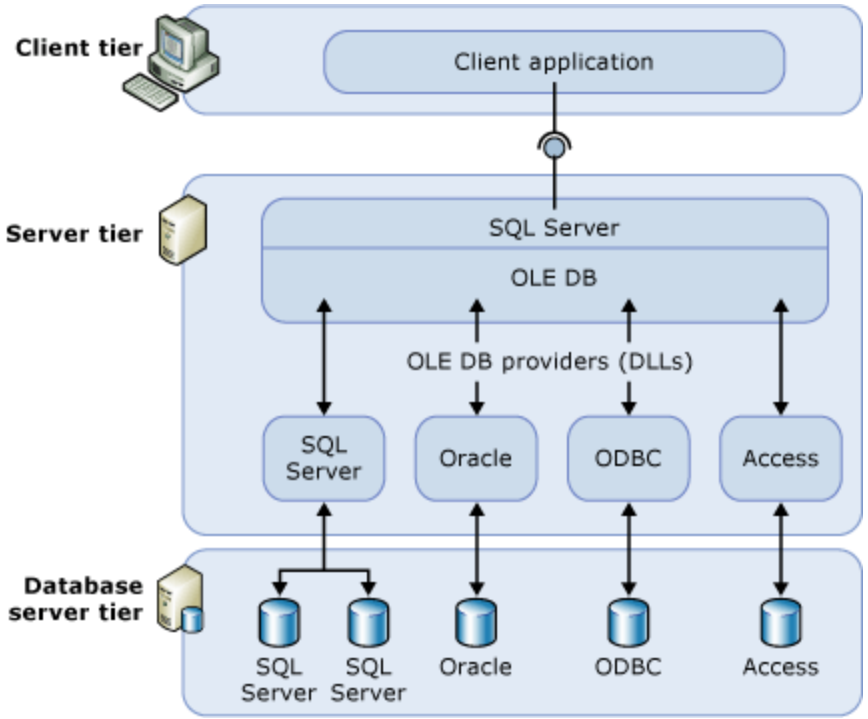
Linked Servers

PolyBase

Linked Servers

- Componenti
  - OLE DB provider
  - OLE DB data source
- Utilizzo
  - Accesso a fonti dati esterne eterogenee
  - Query e transazioni distribuite
- Supporto
  - SQL Server
  - Azure SQL Database Managed Instance
    - limitato SQL Database/SQL Database M.I./SQL Server

DEMO:
Linked Servers



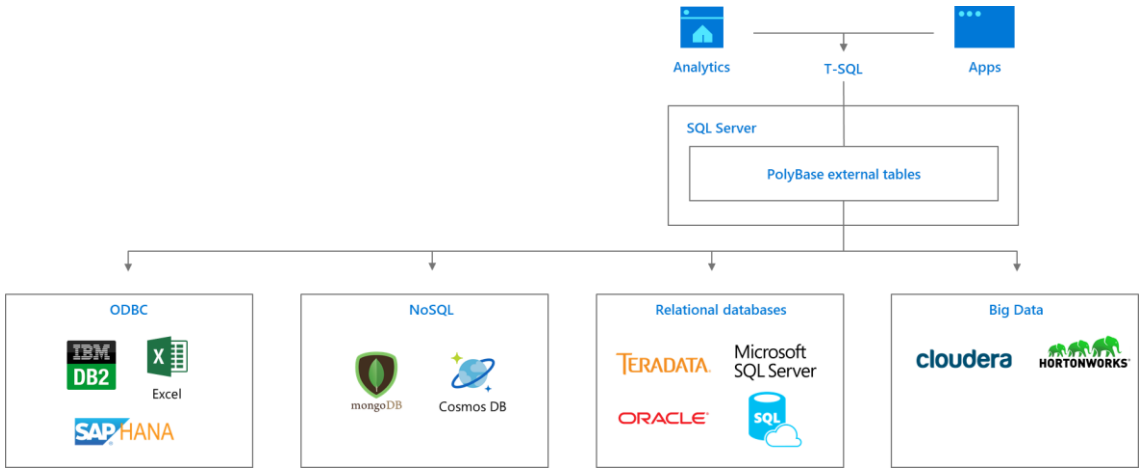
PolyBase

Distributed Compute Engine

- Integrato con SQL Server
- Combina fonti eterogenee tramite query distribuite

Componenti Virtualizzazione

- External Data Source
- External Tables



Supporto PolyBase

SQL Server 2016+

- Azure Blob Storage, Hadoop (Cloudera or Hortonworks)

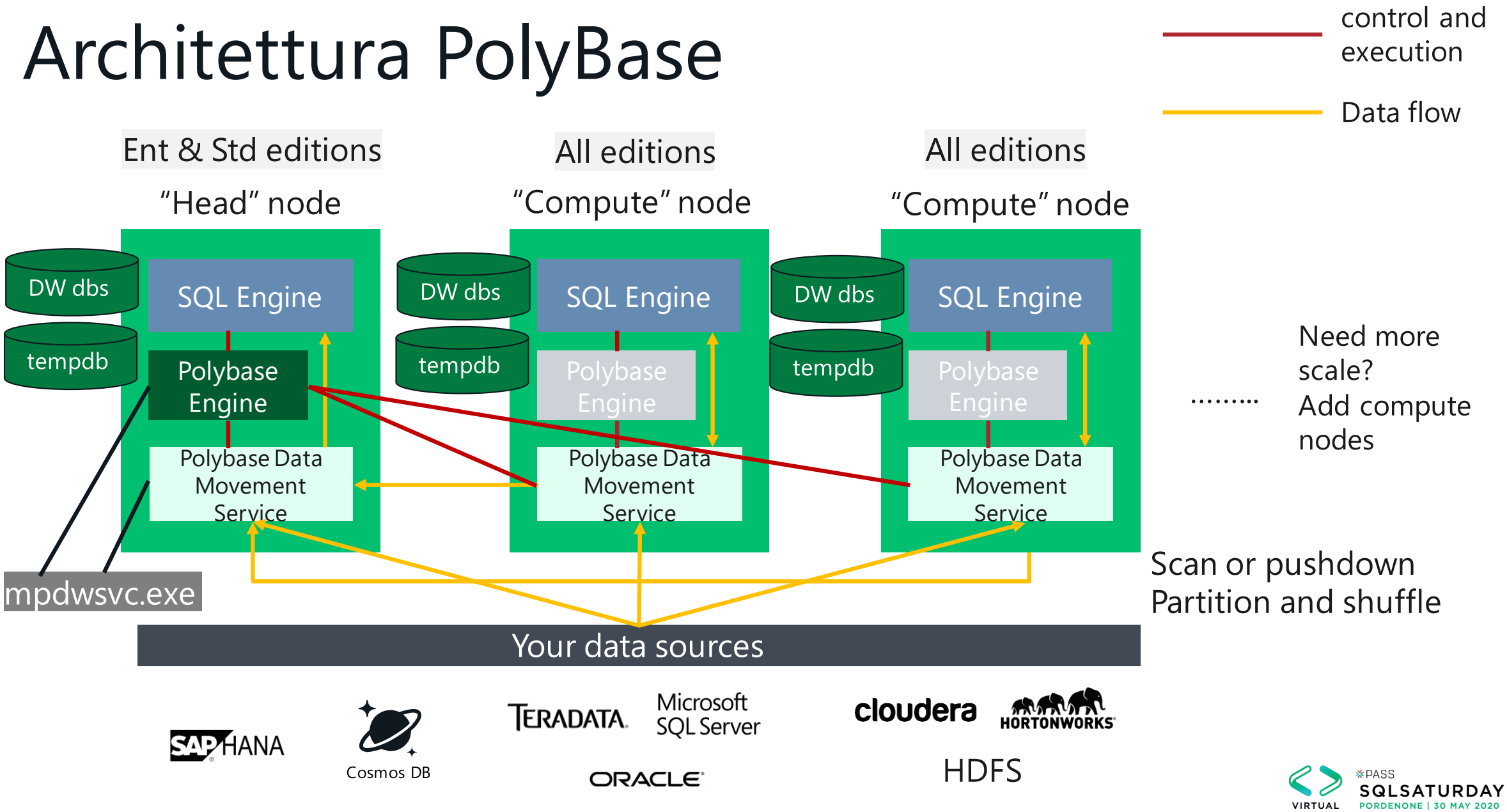
(SQL Server 2017+)

- Operazioni BULK INSERT/OPENROWSET Blob Storage

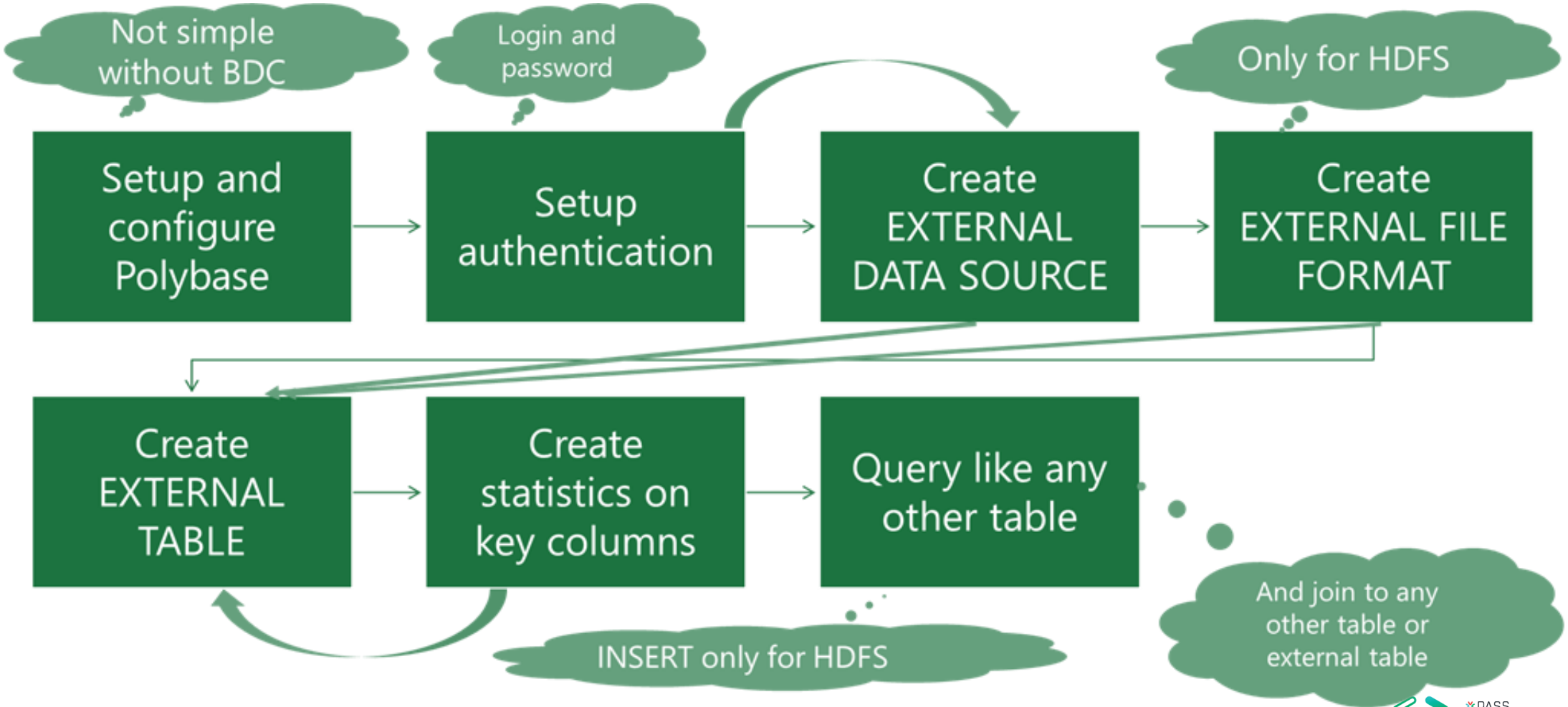
SQL Server 2019

- SQL Server (all flavors), Oracle, Teradata, MongoDB, CosmosDB (Mongo API)
- ODBC Generic Type (e.g. DB2, SAP Hana, Excel, ...)

Architettura PolyBase



DEMO: PolyBase



PolyBase con Azure SQL Database

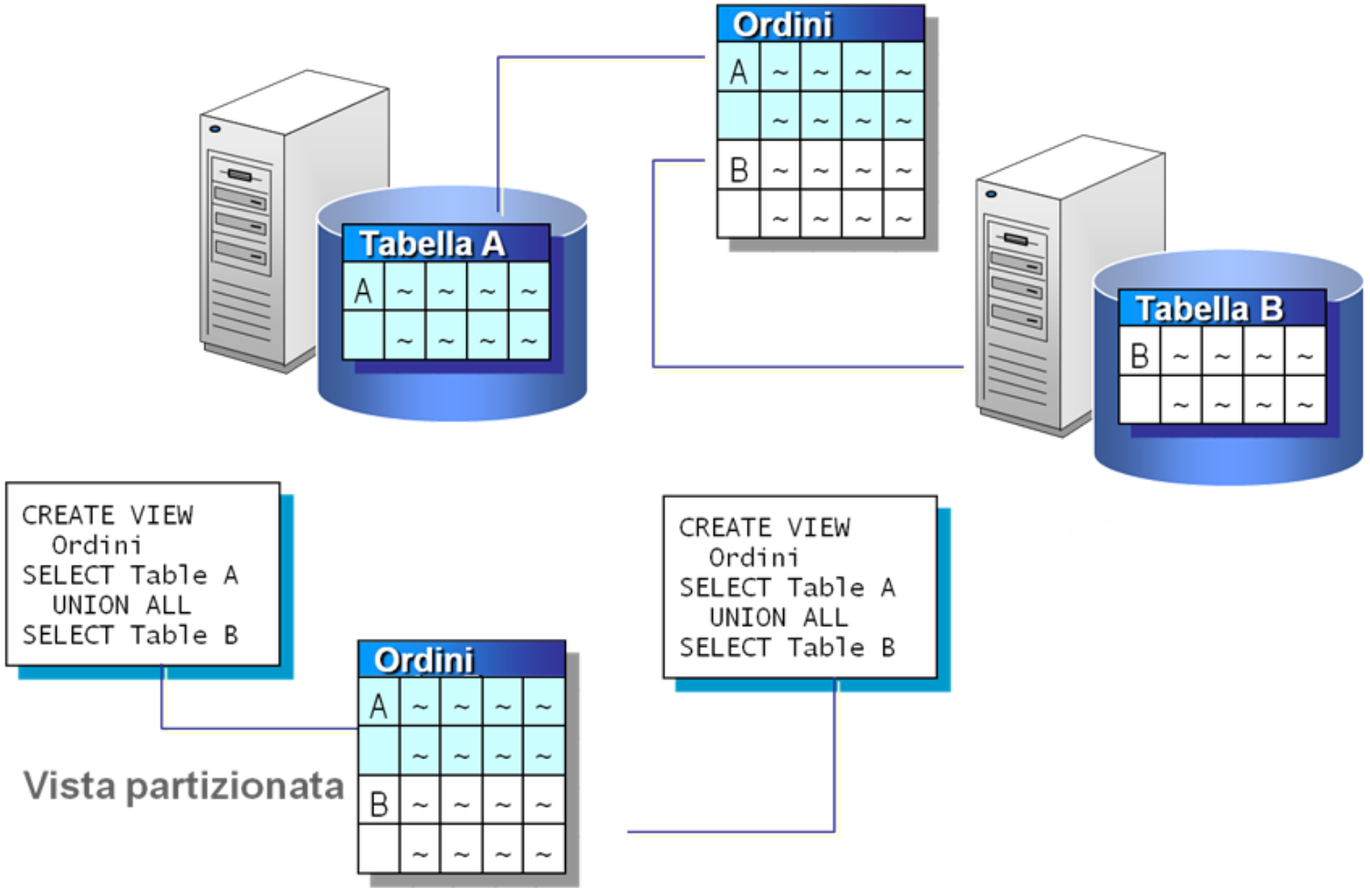
Bug nella versione 2019 RTM

- https://github.com/MicrosoftDocs/sql-docs/issues/3727
- Bisogna specificare database nel data source
- **CONNECTION_OPTIONS = 'Database=pbdemo’**
- Fix non documentato nella GDR
  - https://support.microsoft.com/en-us/help/4517790

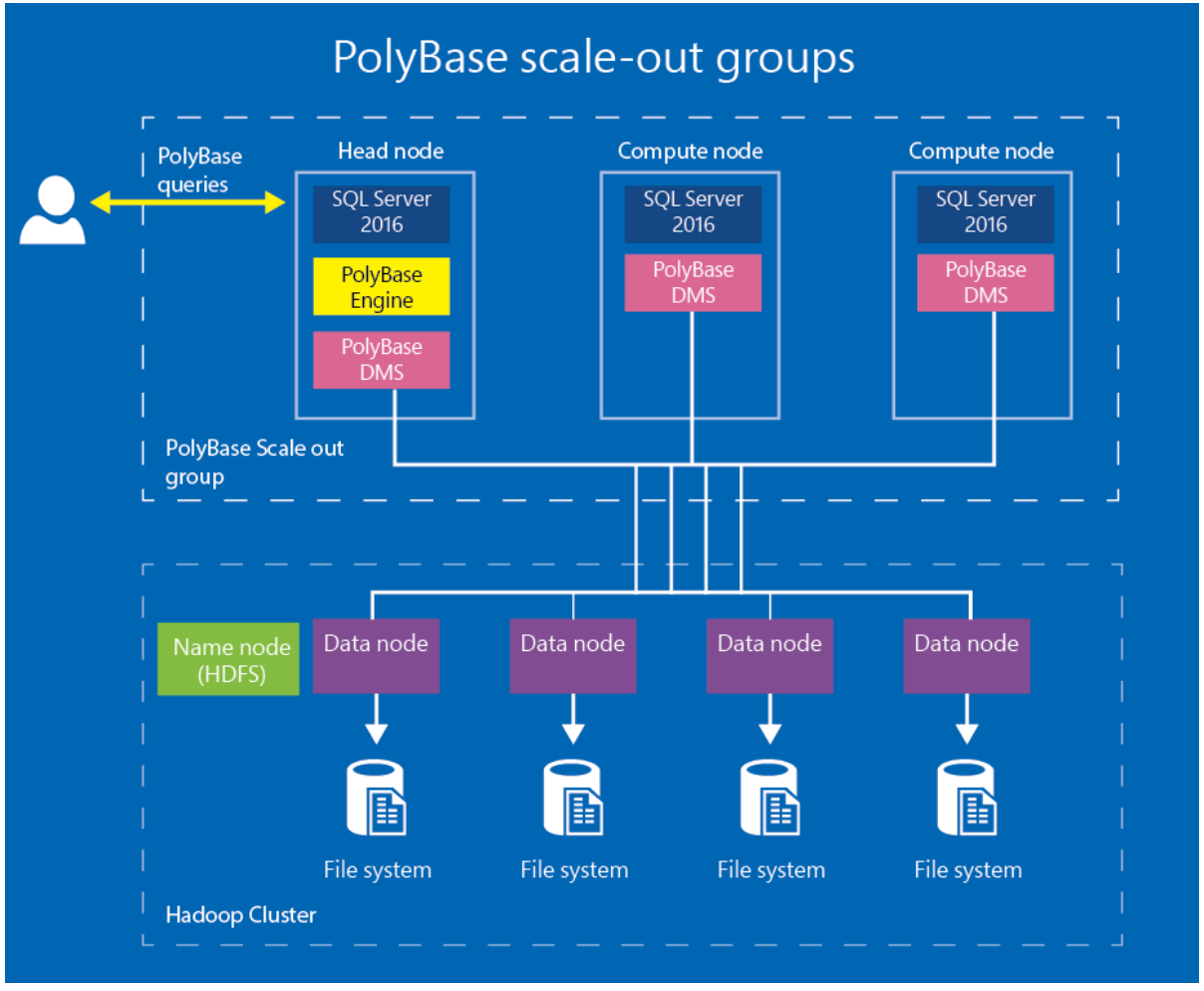
Scale-Out con Linked Server

- Distributed Partioned Views
  - Aka Federated Database Server in SQL Server 2000
- Partizionamento orizzontale tramite viste
  - Tabelle con partizioni dati distribuite su più istanze
  - Vincoli CHECK identificano la partizione
  - Viste su tutti i nodi con SELECT * [...] UNION ALL
- Query Processor riconosce lo scenario
  - Distribuisce il carico eseguendo query remote/locali
  - Predicato deve operare su colonna del vincolo CHECK
- Disponibili ancora solo con edizione Enterprise
- Funzionalità poco documentata...

DEMO: Viste partizionate distribuite



DEMO: PolyBase Scale-Out Groups



A confronto

External Tables

- Database Scoped
- ODBC Drivers
- Read-only*
- Scale out queries with push-down
- Failover with AG
- Basic authentication
- Distributed Transactions not supported
- Better for OLAP workloads

* Insert into HDFS allowed

Linked Servers

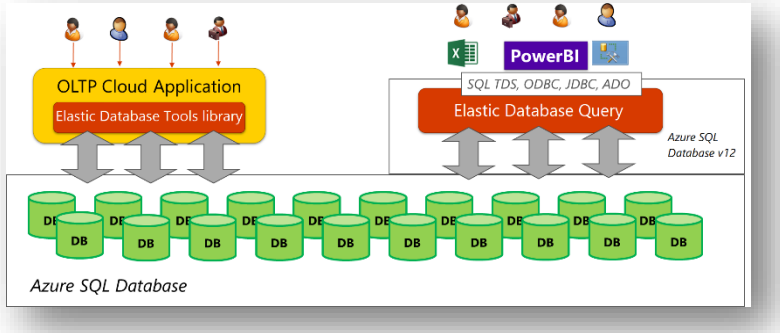
- Instance Scoped
- OLEDB Providers
- Read/write
- Single threaded queries with (limited) push-down
- Requires separate config from AG
- Basic and integrated authentication
- Distributed Transactions supported
- Better for OLTP workloads

Scale-Out OLTP Azure SQL Database

Elastic Database Queries

- Supporto T-SQL per query distribuite
  - CREATE EXTERNAL DATA SOURCE
  - CREATE EXTERNAL TABLE
- Partizionamento verticale: «cross-database queries»
  - TYPE = RDBMS
- Partizionamento orizzontale: «Sharding»
  - TYPE = SHARD_MAP_MANAGER

https://azure.microsoft.com/en-us/documentation/articles/sql-database-elastic-query-overview



Scale-Out OLAP

Azure Synapse Analytics

- aka Azure SQL Data Warehouse
- aka APS/PDW in Azure
- PolyBase è nato qui, principalmente per integrazione HDFS

Altre soluzioni virtualizzazione OLAP

Power BI

- Limite memoria portato da 10GB a 400GB
- 20-30% circa per esecuzione query
- https://docs.microsoft.com/en-us/power-bi/power-bi-data-sources

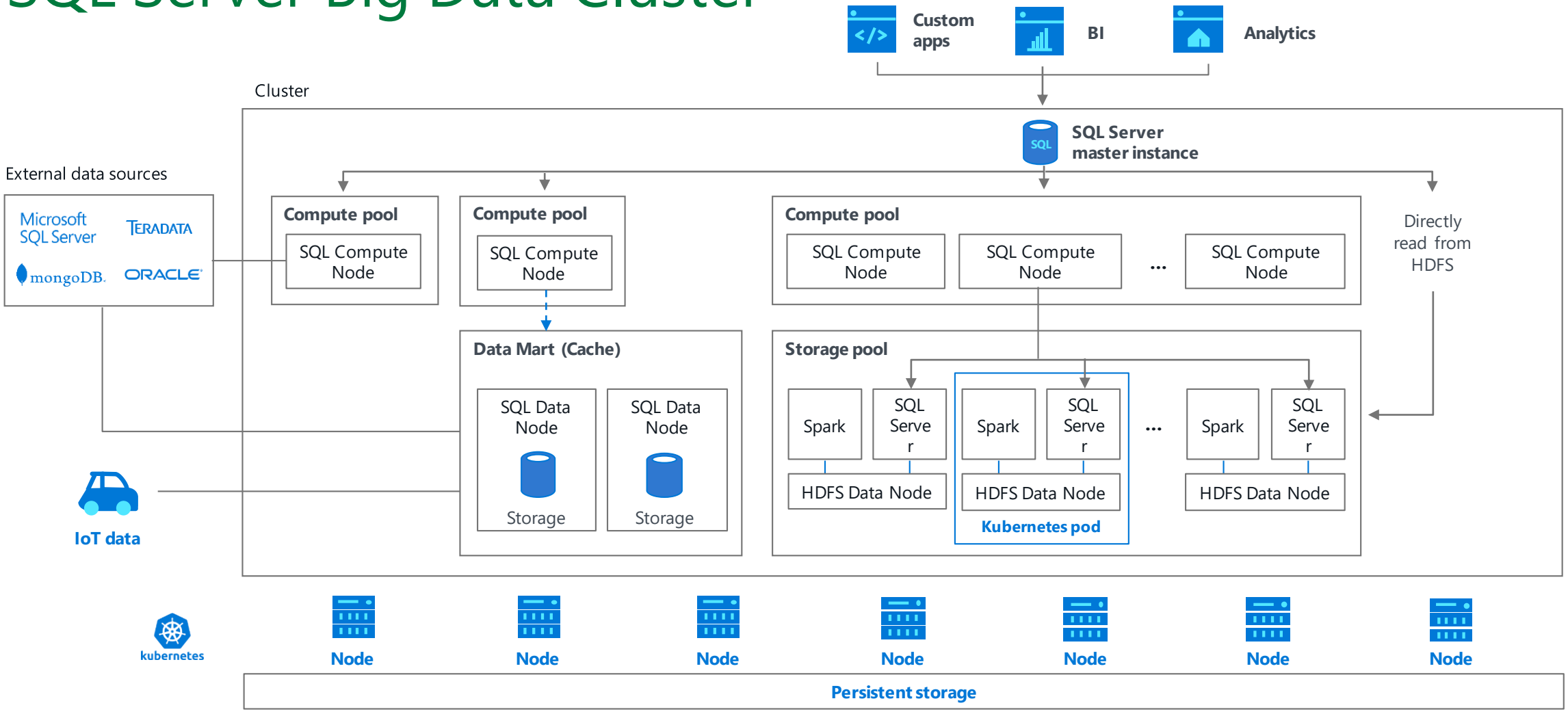
Azure Analysis Services

- https://docs.microsoft.com/en-us/azure/analysis-services/analysis-services-datasource

On-Premise Data Gateway

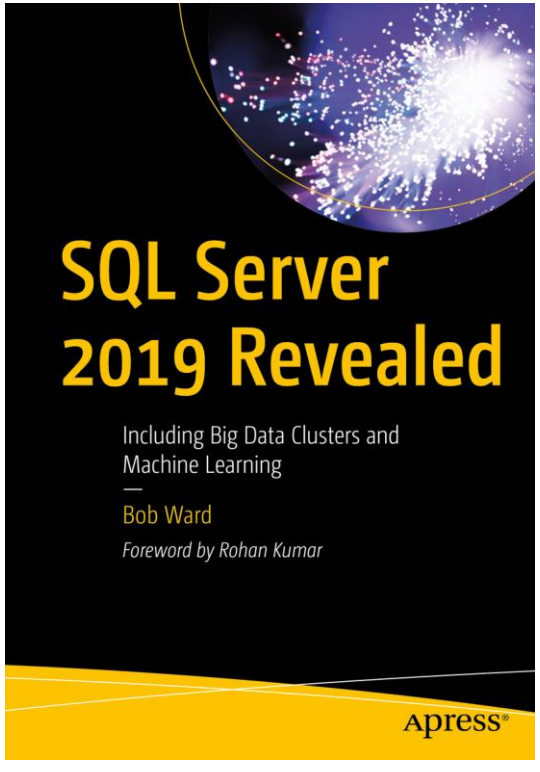
- Può essere usato da Power BI o Azure Analysis Service
- https://docs.microsoft.com/en-us/data-integration/gateway/service-gateway-onprem

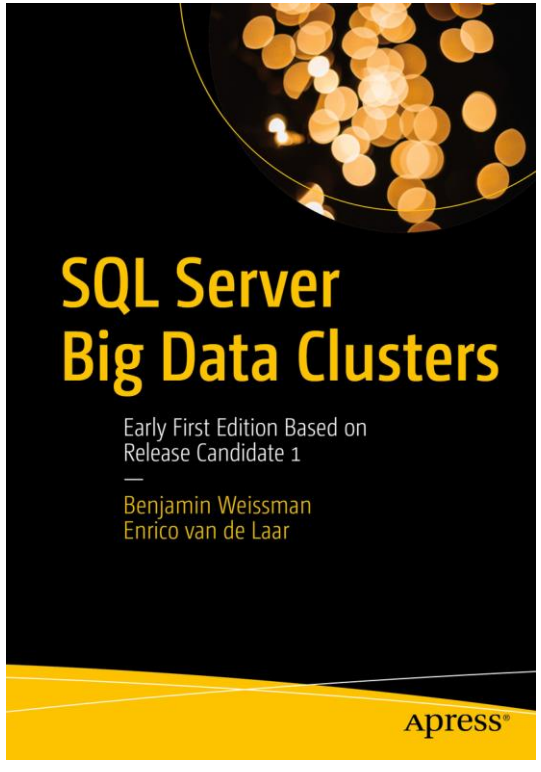
SQL Server Big Data Cluster



https://docs.microsoft.com/en-us/sql/big-data-cluster/big-data-cluster-overview

Risorse


SQL Server
2019 Revealed
Including Big Data Clusters and
Machine Learning
Bob Ward
Foreword by Rohan Kumar
Apress


SQL Server
Big Data Clusters
Early First Edition Based on
Release Candidate 1
Benjamin Weissman
Enrico van de Laar
Apress

PASS SQLSATURDAY

Ricordatevi di compilare il feedback form ☺

https://speakerscore.com/8P1Q

Thank you

#SqlSat921


PASS