Deep-Dive into Polybase

Big Data for SQL Server 2016

Gerhard Brueckl

Our Sponsors

About Me

Gerhard Brückl

From Austria
Working with Microsoft Data Platform since 2006
Mainly focused on Analytics and Reporting
Big Data / IoT
Microsoft Azure

www.pmone.com

@gbrueckl blog.gbrueckl.at gerhard@gbrueckl.at

Polybase
Big Data for SQL Server

Big Data

- What is it?
- Why is it relevant for you?

Polybase

- Introduction
- Setup
- Using it

Scenarios

- Staging / Archive
- Import / Export

Performance

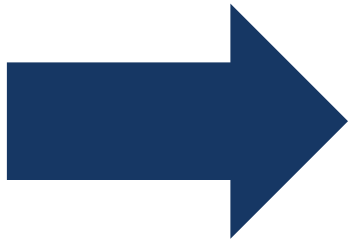
- Predicate Pushdown
- Query Distribution
- Azure SQL DW

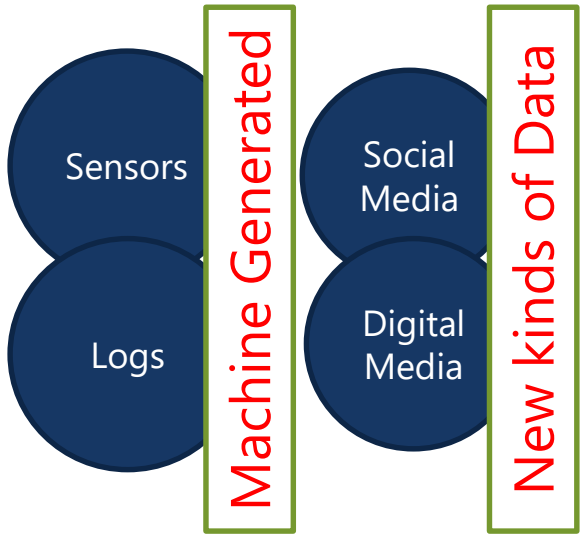
Big Data

Big Data

The way we generate data is changing

Cu	Product	Date	Amount
Cust	ProdA	Oct-06	100€
Cust2	ProdB	Oct-01	50€
CustN	ProdZ	Oct-01	… €



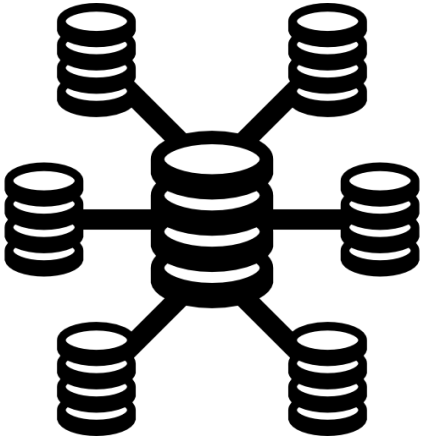


Big Data

The way we store data is changing

HDD/SSD
Raid

HDD/SSD Array
SAN / NAS

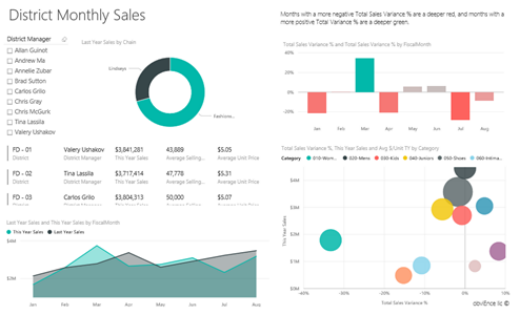


Multi-Node Cluster

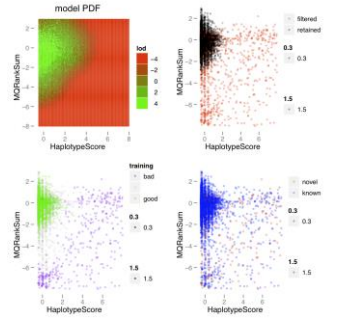
Big Data

The way we analyze data is changing

Static List Reports

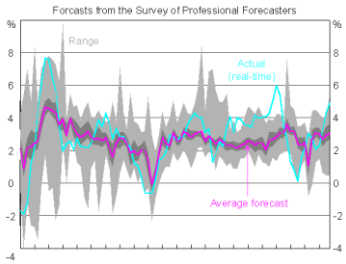


Interactive / AdHoc



Machine Learning

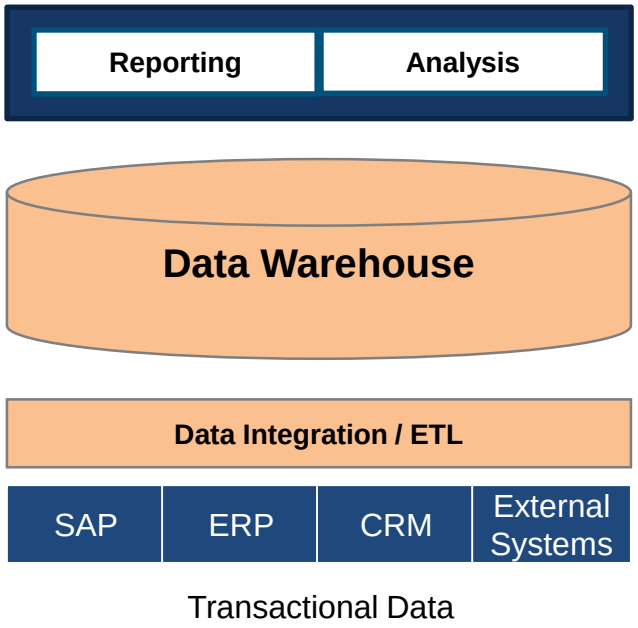
Advanced Analytics



Forecasting

Big Data and the classic DWH

Combining the “old” data with the “new” data



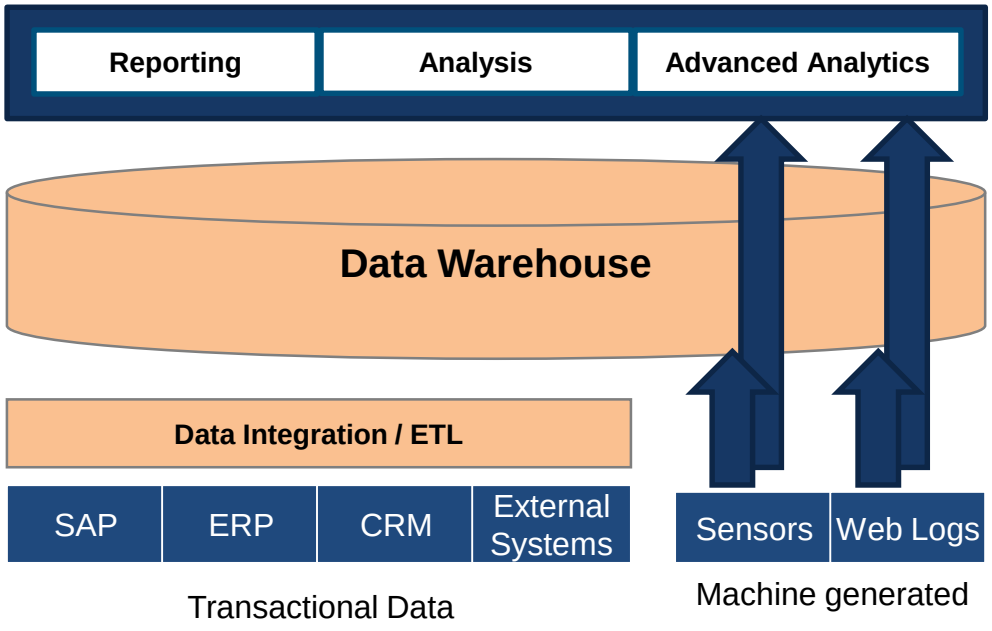
SQL Server Reporting Services

SQL Server RDBMS

SQL Server Integration Services

Big Data and the classic DWH

Combining the “old” data with the “new” data



SSRS / Power BI

SQL Server 2016

SSIS / Polybase

* Requires further processing

Polybase

What is Polybase

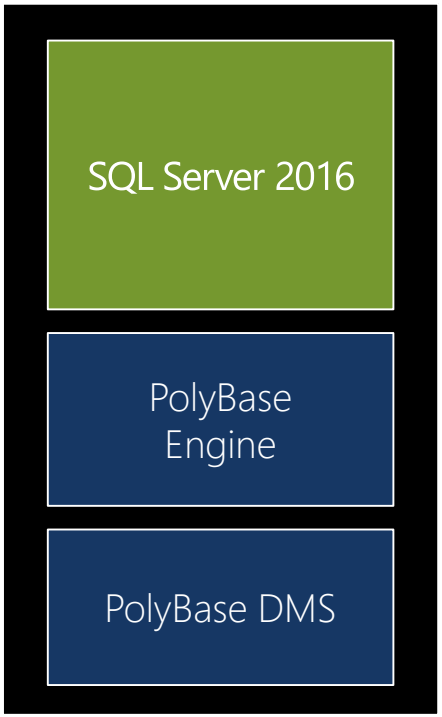
"Polybase is a Technology which allows us to access data which resides in a distributed file system (like HDFS) in a traditional way using regular SQL commands."

Analytical Platform System (APS)

Separate Storage and Compute

Scalability

SQL Server 2016

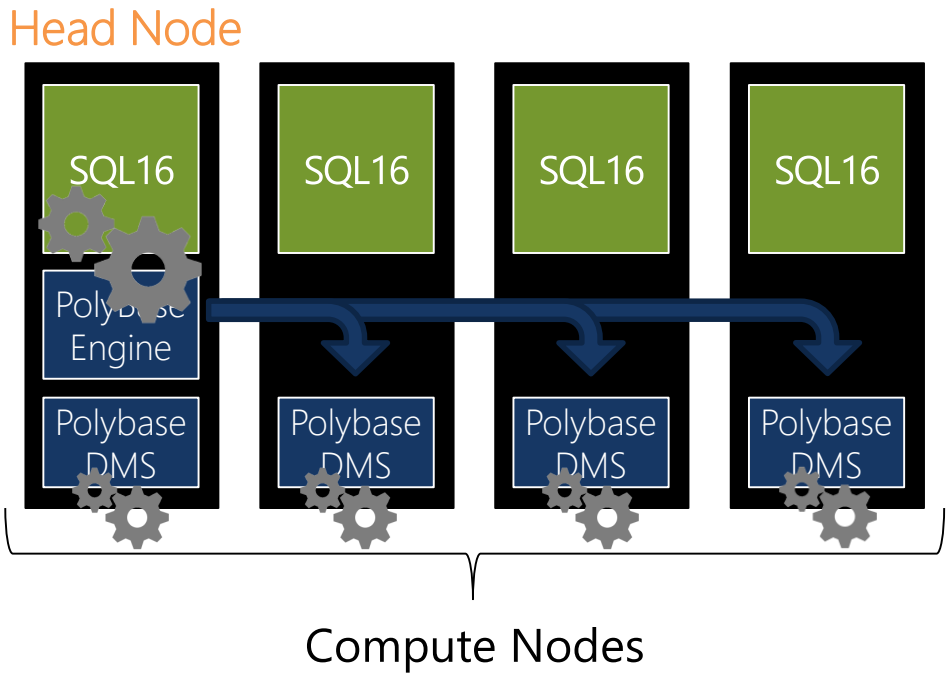


SQL Server 2016 Enterprise Edition

Windows Service responsible for distributed query processing.

Windows Service responsible for reading/writing data from/to HDFS

SQL Server Scale-Out Group



- Data Movement is still the most critical operation
- Create parallel reads across multiple compute nodes for faster throughput
- Each node reads only a subset of the data and returns results to the head node

Supported Sources

File Systems

- Hadoop Distributed File System (HDFS)
  - Hortonworks Data Platform (HDP)
  - Cloudera (CDH)
- Azure Blog Store

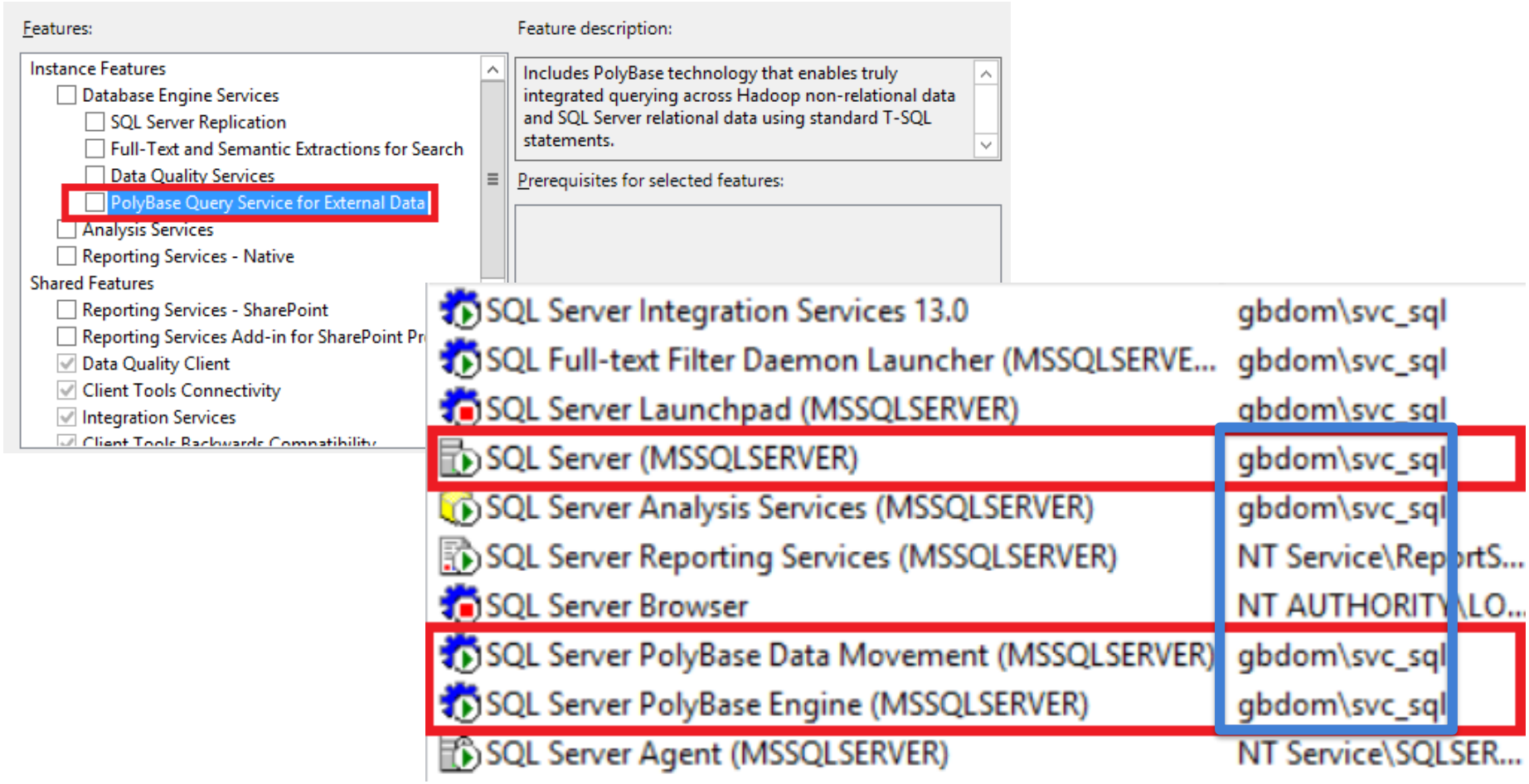
File Formats

- Delimited Text (CSV, TSV, …)
- ORC
- RC
- Parquet

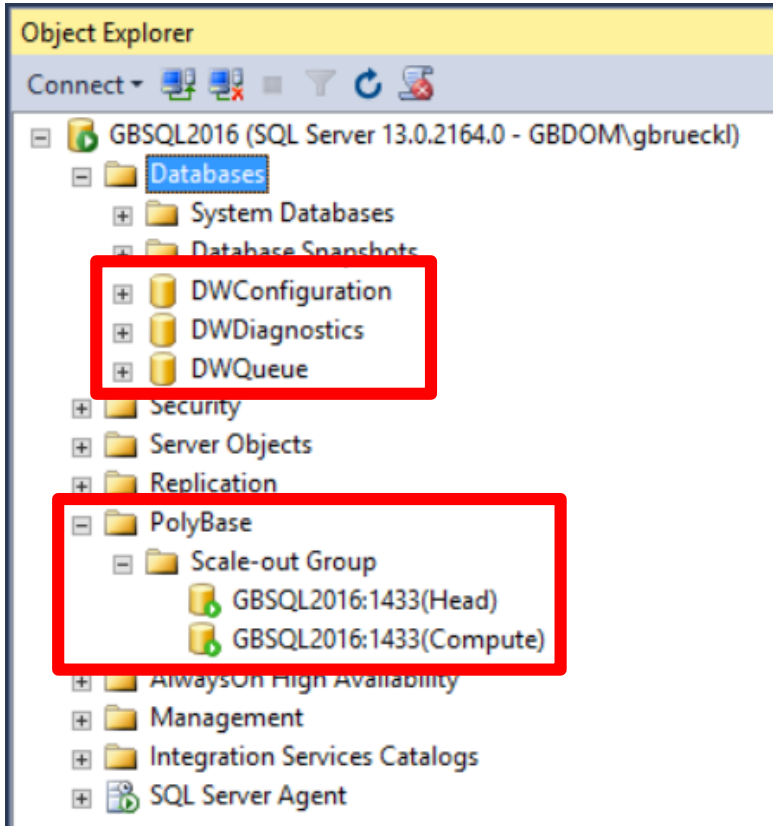
MORE TO COME!

Setup

Setting Up Polybase



PolyBase Internal Databases



DWConfiguration

Configuration inform for the PolyBase engine a MS services

DWDiagnost

Diagnostic distributed query system

DW eue

mation for rollback.

DO NOT ALTER THESE!

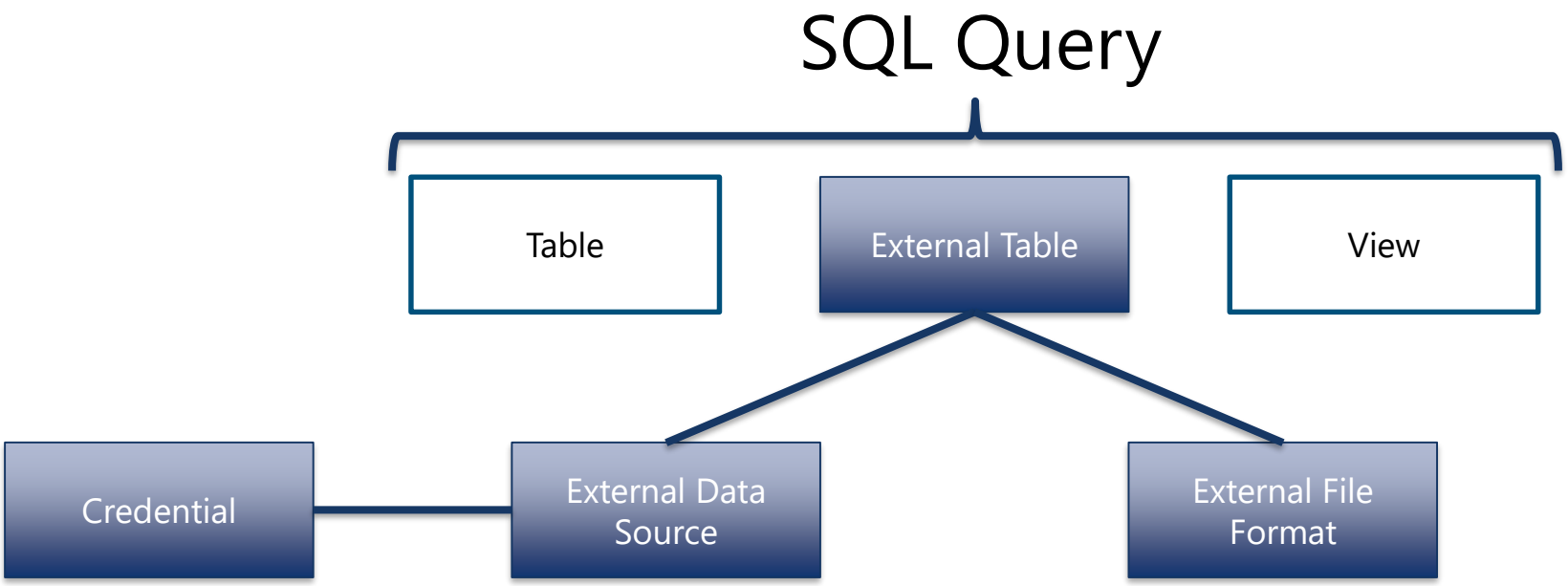
Configure Hadoop Connectivity

Different Sources require different settings

```
--Configure Hadoop connectivity
sp_configure
     @configname = 'hadoop connectivity',
     @configvalue = { 0 - 7 }
[;]
```

Value	Description
0	Disable Hadoop connectivity
1	HDP 1.3 on Windows
	Azure Blob Store
2	HDP 1.3 on Linux
3	CDH 4.3 on Linux
4	HDP 2.0 on Windows
	Azure Blob Store
5	HDP 2.0 on Linux
6	CHD 5.1+ on Linux
7 (default)	HDP 2.1+ on Linux
	HDP 2.1+ on Windows
	Azure Blob Store

External Objects



MasterKey and Credential

MasterKey

- Encrypt sensitive Information

```
-- Create a Master Key using my own password.
CREATE MASTER KEY ENCRYPTION BY
PASSWORD='Pass@word1!';

-- Create database Credential
CREATE DATABASE SCOPED CREDENTIAL CRED_AzureStorage
WITH
IDENTITY = 'gbdomaindata', --> Storage Account
SECRET='<AccessKey>'; --> AccessKey
```

Credential

- Access to DataSource

MSDN

External Data Sources

Reference to external Storage Service

- Type
  - Azure Blob Store
  - Hadoop / HDFS
    - Hortonworks
    - Cloudera
- Location

```
-- Create the External Data Source using Credential
CREATE EXTERNAL DATA SOURCE AzureStorage
WITH (
    TYPE = HADOOP,
    LOCATION =
'wasbs://<container>@<name>.blob.core.windows.net',
     CREDENTIAL = CRED_AzureStorage
);
```

MSDN

External File Format

Definition of File Format

Format Type

- Delimited Text
- ORC
- RCFile
- Parquet

Data Compression

Format Options

SerDe

```
-- Create External File Format
CREATE EXTERNAL FILE FORMAT CSV_Quoted
WITH (
    FORMAT_TYPE = DELIMITEDTEXT,
    FORMAT_OPTIONS (
        FIELD_TERMINATOR = ',',
        STRING_DELIMITER = '"',
        USE_TYPE_DEFAULT = FALSE
    )
);
```

MSDN

External Tables

Definition of external data

- Columns
- Data Type
- Reject Settings

External Data Source

- Location

External File Format

```
-- Create the External Table
CREATE EXTERNAL TABLE [azure].[DimCurrency] (
CurrencyKey int NOT NULL,
CurrencyAlternateKey nchar(3) NOT NULL,
CurrencyName nvarchar(50) NOT NULL
)
WITH
(LOCATION='/AdventureWorksDW2012/DimCurrency',
      DATA_SOURCE = AzureStorage,
      FILE_FORMAT = CSV_Quoted,
      REJECT_TYPE = VALUE,
      REJECT_VALUE = 0
);
```

Working with External Tables

SELECT

- Export: INSERT / CETAS

Full SQL Syntax

- Joins
- Group By
- Aggregation

```
SELECT
    dim.[ProductGroup],
    SUM(ext.[Revenue]) AS TotalRevenue
FROM [dwh].[DimProduct] dim
INNER JOIN [ext].[Sales] ext
    ON dim.[ProductKey]
        = ext.[ProductKey]
WHERE ext.[Quantity] > 10
GROUP BY dim.[ProductGroup]
HAVING SUM(ext.[Revenue] > 100
```

Scenarios

Scenarios

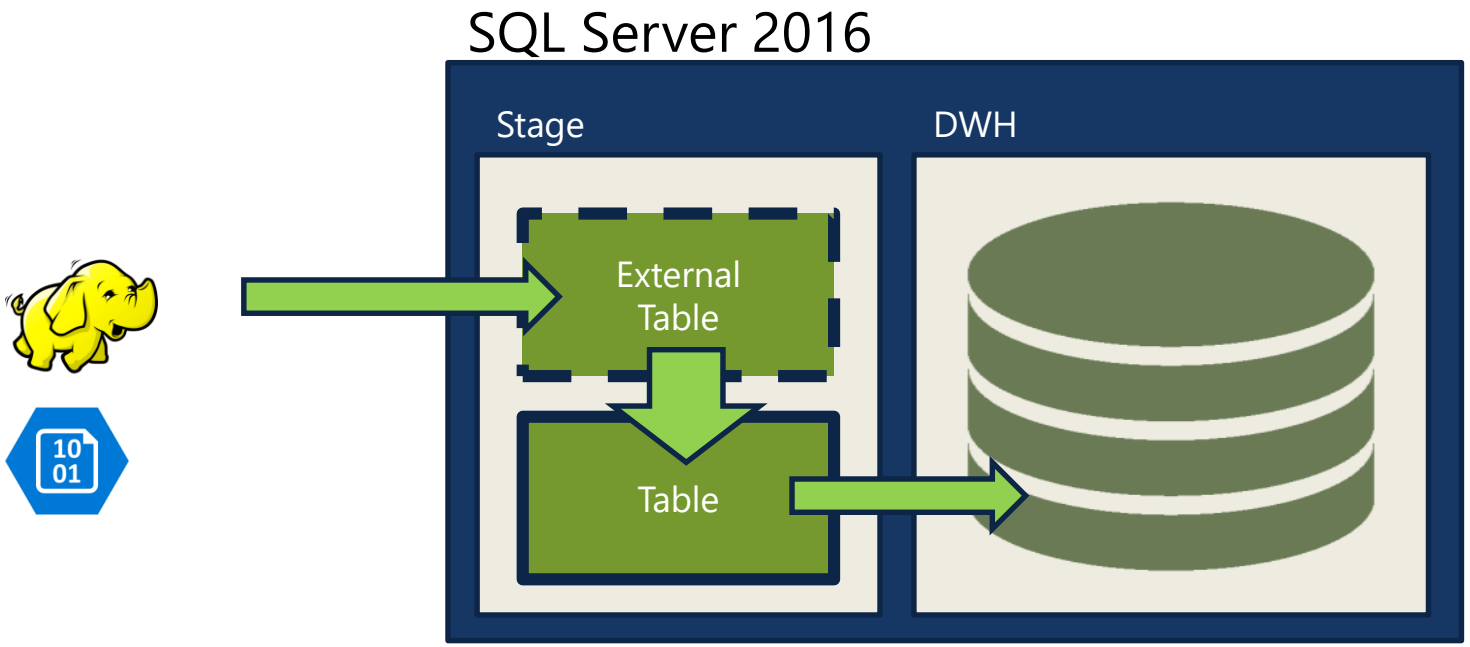
Staging

Archive

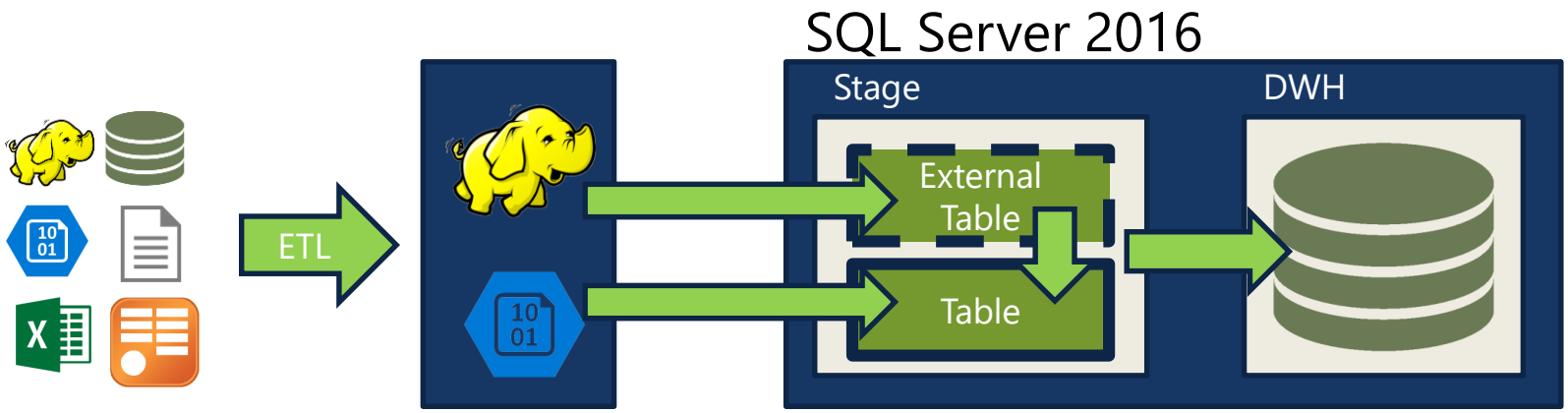
Import / Export

SQL Interface for HDFS

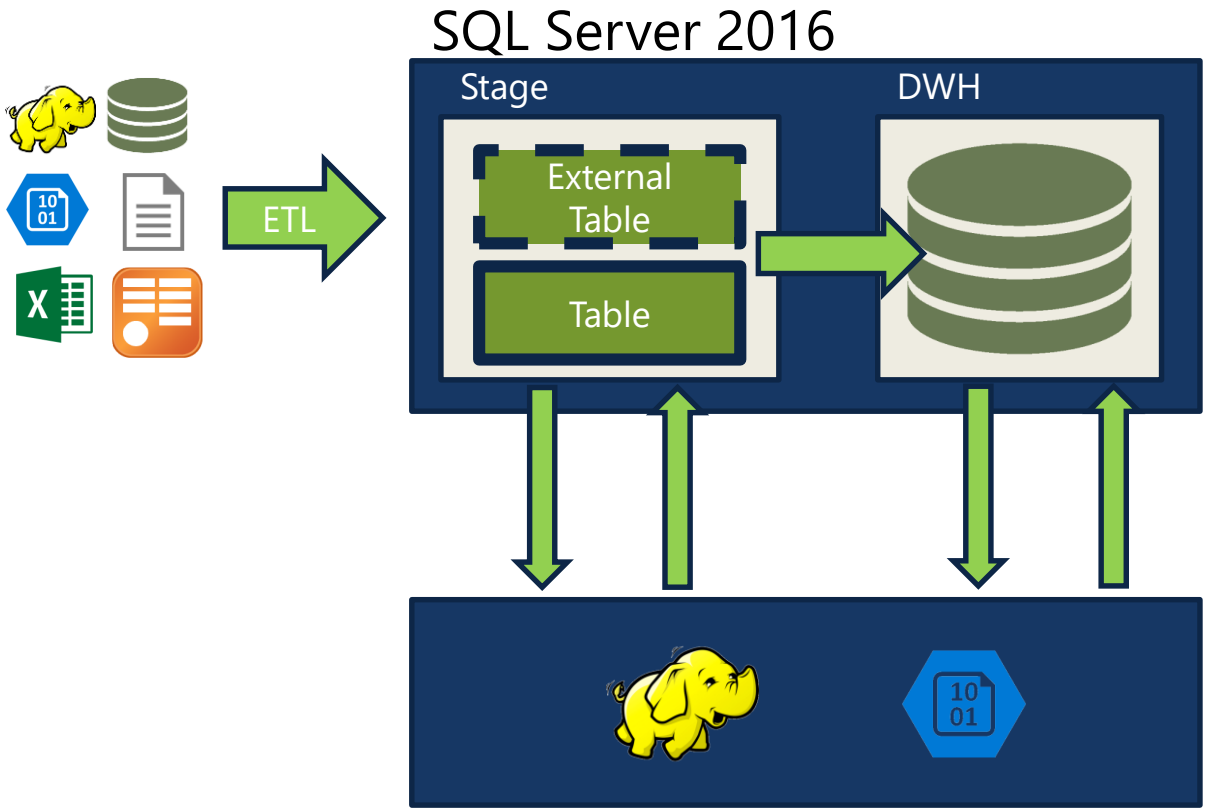
Polybase for Staging



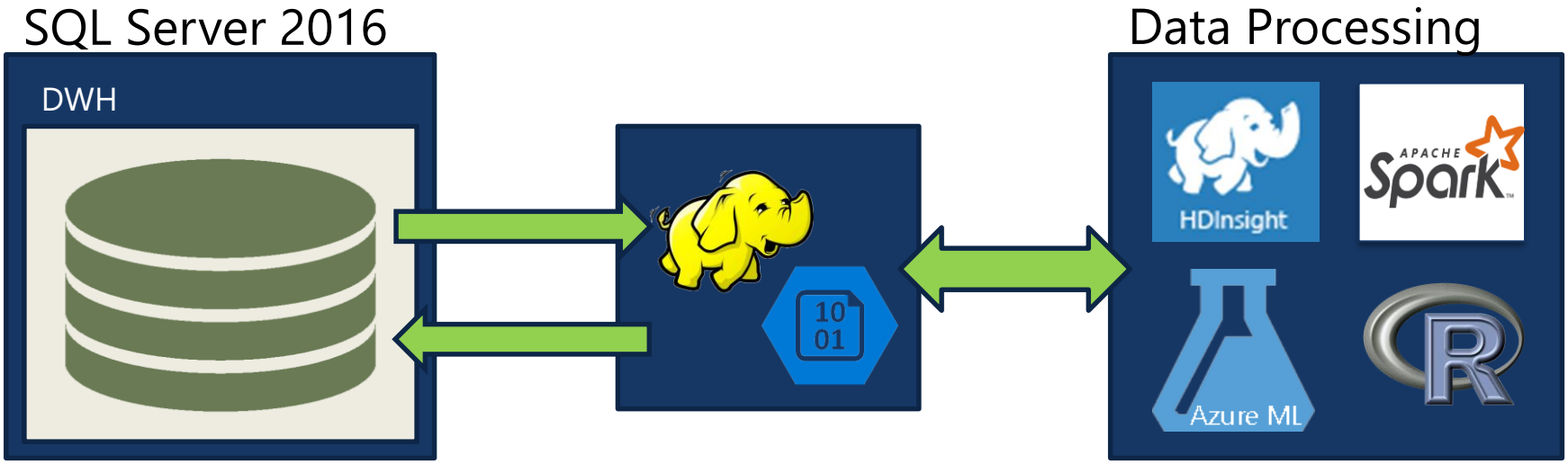
Polybase for Staging Archive



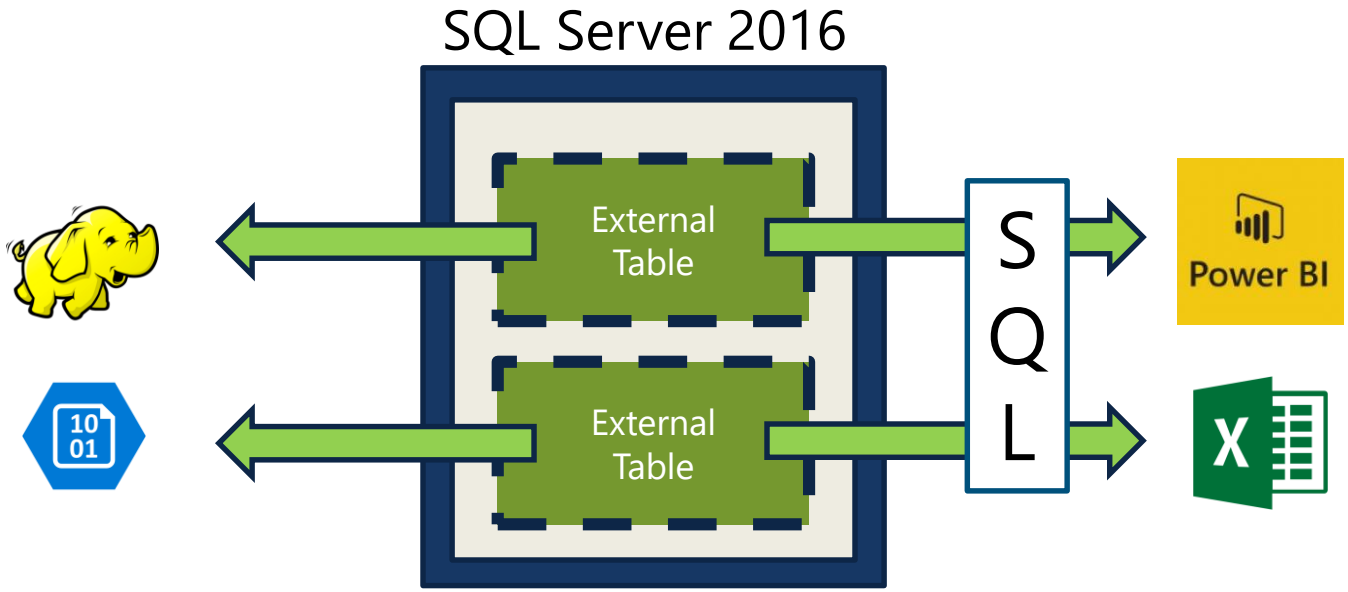
Polybase for Archiving (1)



Polybase for Import / Export



Polybase as SQL Interface



Performance

Predicates – HDFS only!

Push-able

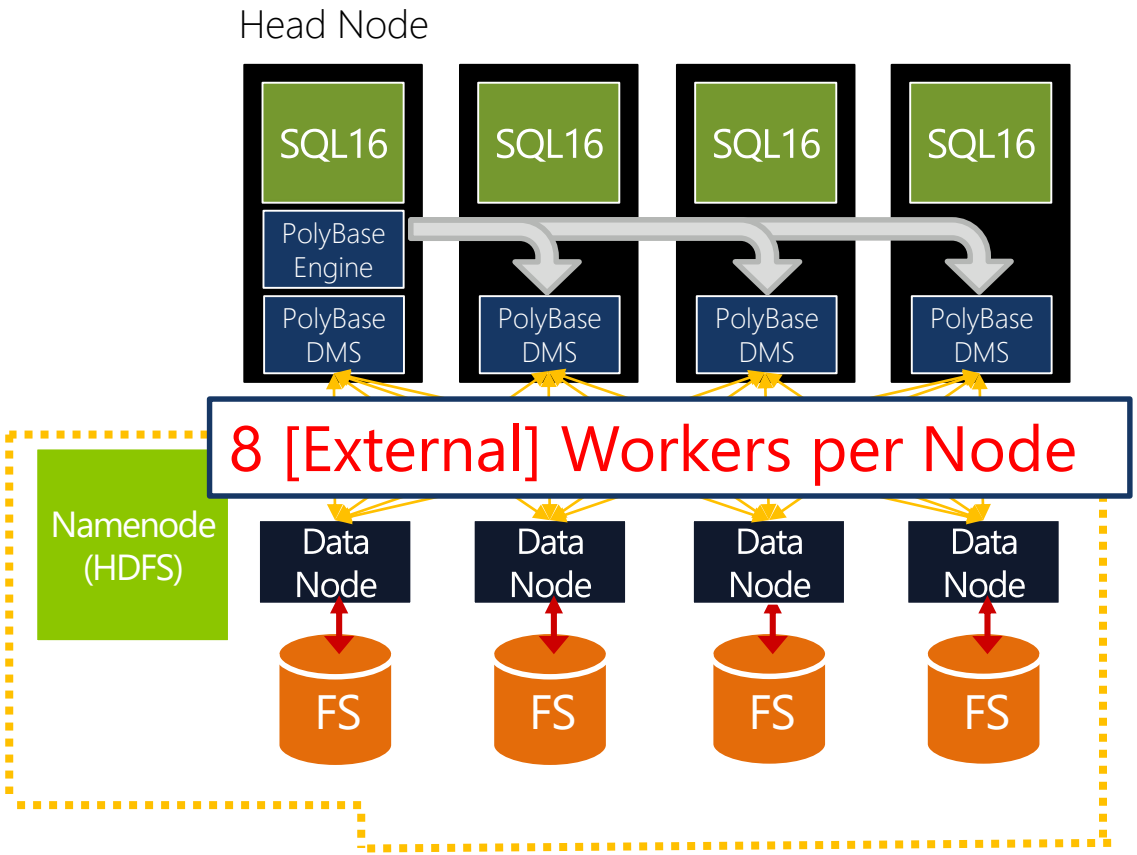
- Select subset of columns
- Arithmetic Operators
- Comparison Operators
- Logical Operators
- Unary Operators

Non-push-able

- Joins
- Complex Operators
- Group By *
- Aggregations *

* Might come in the future

Query Distribution



File Layout

One Big File vs. Many Small Files

Compressed vs. Uncompressed Files

File Format

Align with Number of Readers !!!

No Partitioning yet !!!

File Format – CSV

UTF-8 only!

Remove Header Rows

No Linebreaks in Text

Proper Date-Format

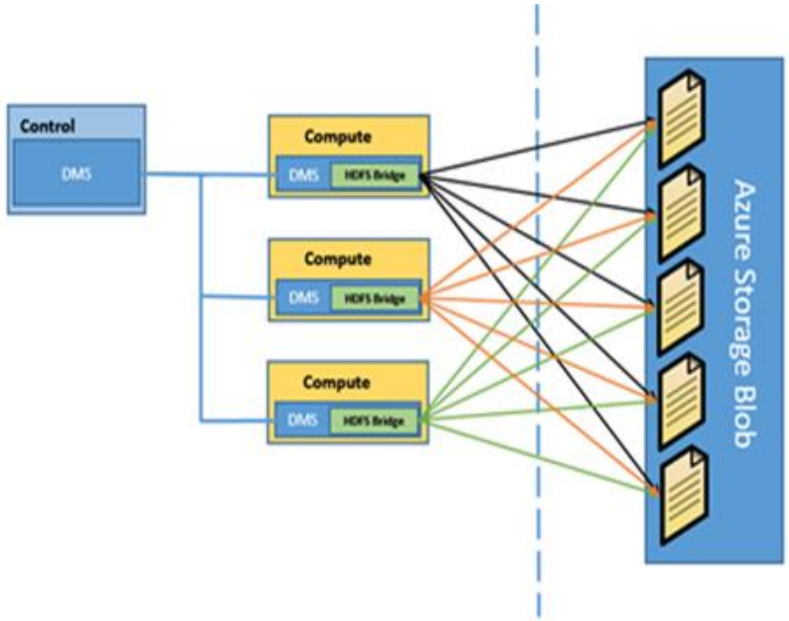
- [Date] = yyyy-MM-dd
- [Time] = hh:mm:ss

Azure SQL Data Warehouse

APS/PDW in the cloud

- Pause/Stop Compute
- Persisted Storage

- Always 60 Distributions



Number of:	DWU									
	100	200	300	400	500	600	1,000	1,200	1,500	2,000
Compute Nodes	1	2	3	4	5	6	10	12	15	20
Readers	8	16	24	32	40	48	80	96	120	160
Writers	60	60	60	60	60	60	80	96	120	160

Save the Dates!

PASS Camp 2016 - 06. to 09. December 2016

SQL Konferenz 2017 - 14. to 16. February 2017

SQL Saturday Vienna Freitag, 20. Januar 2017

How did you like it?

Please give feedback

- to the event:
- http://goo.gl/5tStsd

- to me as a speaker:
- http://goo.gl/TgViTX

Questions?

Gerhard Brückl

From Austria
Working with Microsoft Data Platform since 2006
Mainly focused on Analytics and Reporting
Big Data / IoT
Microsoft Azure

www.pmone.com

@gbrueckl blog.gbrueckl.at gerhard@gbrueckl.at