DATA VIRTUALITY MASTERCLASS

Topic: Analytical Storage considerations

Welcome to the DV Masterclass! Agenda

Day 1

2:00pm - 3:00pm: Analytical storage considerations

3:00pm - 3:15pm: Break

3:15pm - 4:00pm: Support stories

4:00pm - 4:15pm: Break

4:15pm - 5:00pm: Logging with log4j

Day 2

2:00pm - 3:00pm: Sending formatted reports from Data Virtuality

3:00pm - 3:15pm: Break

3:15pm - 4:00pm: Updating Data Virtuality Server

4:00pm - 4:15pm: Break

4:15pm - 4:45pm: Receiving webhooks

What to expect from this session?

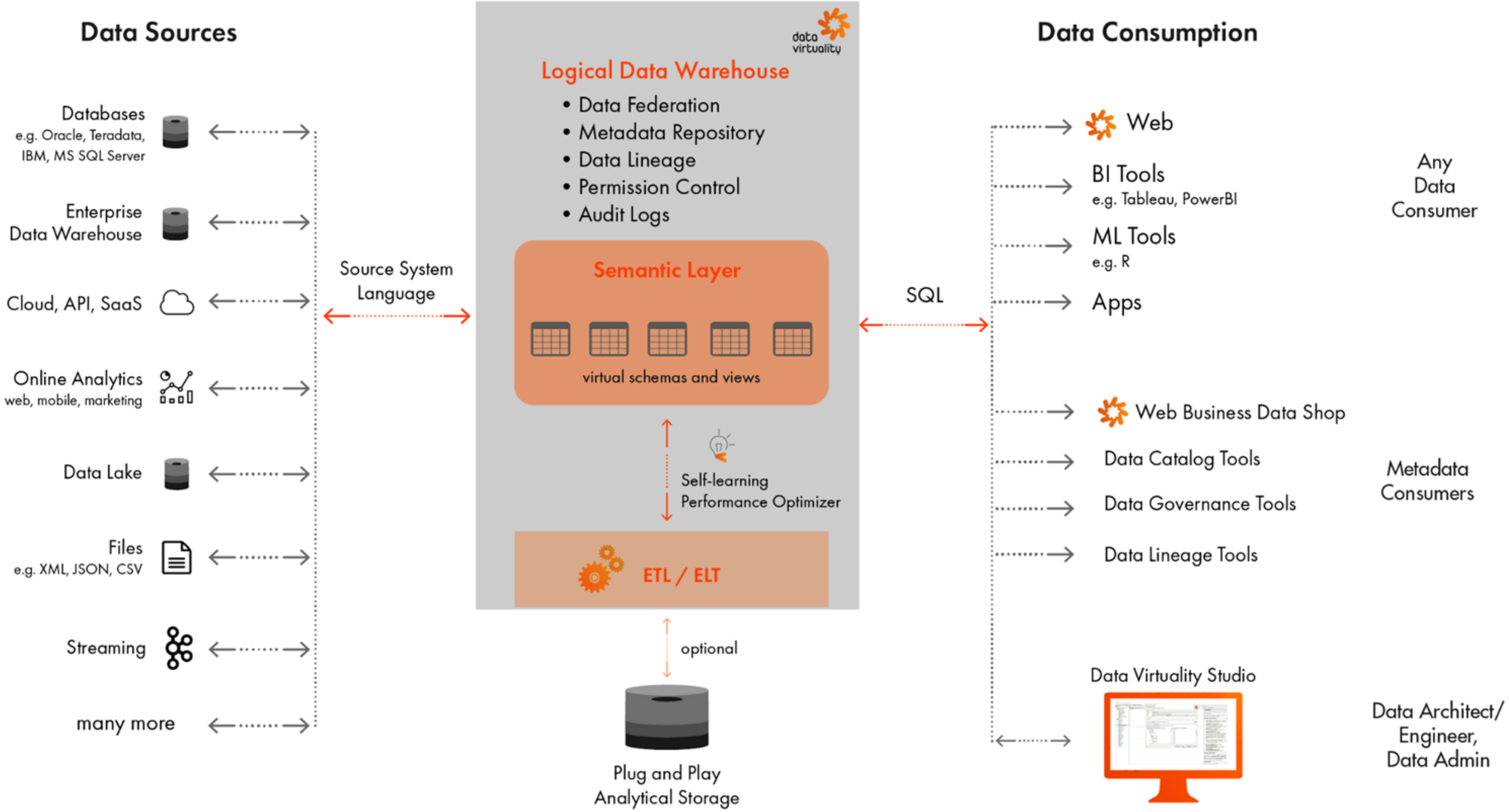
In this track, we will discuss the options of analytical storages for Data Virtuality.

- Purpose of the Analytical Storage in DV - Replication and Materialization
- Gathering your requirements
- Available Options for the Analytical Storage
- Comparing different Database Architectures
- Connection settings
- Switching your Analytical Storage

Purpose of the Analytical Storage in DV

Architecture

- Using the Analytical Storage as a repository
- Replication can go anywhere
- Materialization can only go to the Analytical Storage



Materialization in detail

- Materialized views / tables are written to the Analytical storage
- for each full materialization a new stage is created
- requests will access the old stage until the new stage has completed materializing
- Cleanup job will remove it
- for incremental materialization it will only be one stage, do not archive data there!
- The more you materialize, the more it becomes an operation on the Analytical Storage, using its capabilities
- This is why it is important to choose the right storage

Demo: effects of materialization of views

data virtuality

Gathering your requirements

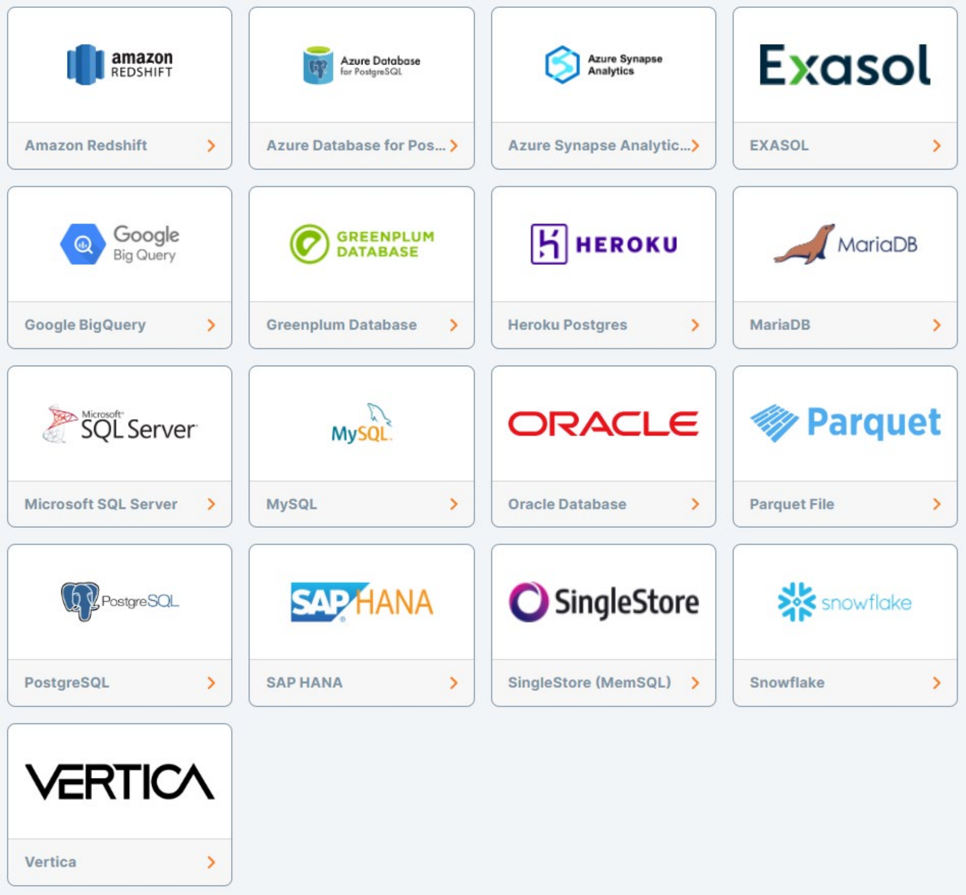
Question Catalog

- Where is my DV server located
  - Cloud or on prem?
- What amount of data do I expect
  - To be replicated and materialized
  - To be read in queries
- What are my access patterns
  - Join to other data
  - Pure analytical queries
- What is my desired performance
- How much do I want to pay
  - per hour, calculation or fixed

Available Options for the Analytical Storage in DV

Overview

- Available options: https://datavirtuality.com/en/connectors/#connectors-data-warehouses
- Limitations: Parquet, Google BQ



Amazon Redshift
Azure Database for Pos...
Azure Synapse Analytic...
EXASOL
Google BigQuery
Greenplum Database
Heroku Postgres
MariaDB
Microsoft SQL Server
MySQL
Oracle Database
Parquet File
PostgreSQL
SAP HANA
SingleStore (MemSQL)
Snowflake
Vertica

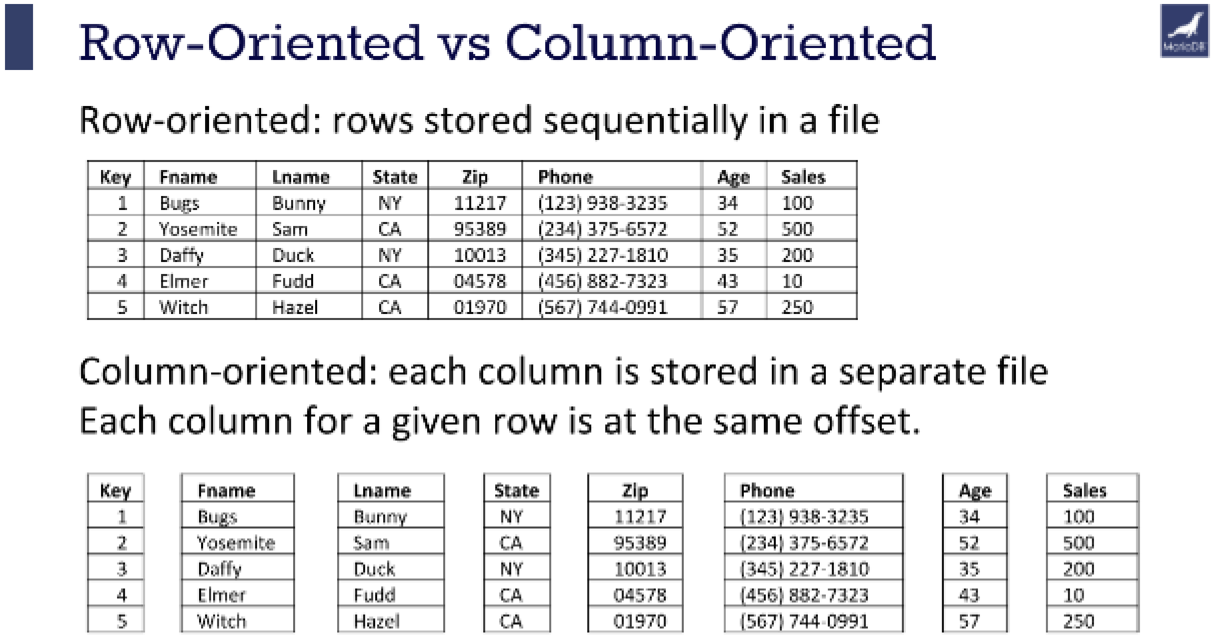
Comparing Database Architectures

Database Architectures

- In memory: frequently accessed / all data is kept in RAM for high query performance
- Columnar: each column’s data is stored in one file

Columnar Storage

- Efficient if aggregations are done over few columns
- Efficient if all values in a column are updated
- Less efficient when many columns are read
- Less efficient when writing a full row
- Compression is efficient, as one row has one data type



Row-Oriented vs Column-Oriented

Row-oriented: rows stored sequentially in a file

Key	Fname	Lname	State	Zip	Phone	Age	Sales
1	Bugs	Bunny	NY	11217	(123) 938-3235	34	100
2	Yosemite	Sam	CA	95389	(234) 375-6572	52	500
3	Daffy	Duck	NY	10013	(345) 227-1810	35	200
4	Elmer	Fudd	CA	04578	(456) 882-7323	43	10
5	Witch	Hazel	CA	01970	(567) 744-0991	57	250

Column-oriented: each column is stored in a separate file
Each column for a given row is at the same offset.

Key
1
2
3
4
5

Fname
Bugs
Yosemite
Daffy
Elmer
Witch

Lname
Bunny
Sam
Duck
Fudd
Hazel

State
NY
CA
NY
CA
CA

Zip
11217
95389
10013
04578
01970

Phone
(123) 938-3235
(234) 375-6572
(345) 227-1810
(456) 882-7323
(567) 744-0991

Age
34
52
35
43
57

Sales
100
500
200
10
250

Image: mariadb.com

Database	Storage	Native Columnar	Cloud Only	Pricing	Remarks
Amazon Redshift	On compute nodes	Yes	Yes, AWS	Hourly per node	In use by many DV customers, VACUUM!
PostgreSQL / Heroku / Azure / RDS	Disk based	No, but possible via cstore_fdw extension	No / Yes	Free / hourly in cloud	DV’s favourite, good and stable starting point
Azure Synapse Analytics	Data Lake Storage	Yes	Yes	Hourly per node	
Exasol	Active data in memory, rest on disk	Yes	No	Buy / Hourly	
Greenplum / Vmware Tanzu GP	Disk based	Columnar, row or both	No	Community / $995 per CPU (Basic)	MPP version of PostgreSQL
MariaDB	Disk based	No, but possible via MariaDB ColumnStore extension	No	Free	Claims to be faster than MySQL and better with larger data
MS SQL Server	Disk based (partial in memory with In-Memory-OLTP)	No, but possible via Columnstore Index	No	Buy	Limits apply
MySQL	Disk based	No, but MEMORY Storage Engine available	No	Free, Buy	Limits apply
Oracle	Disk based, (partial in memory with Oracle Database In-Memory (12c+)	No, but possible via In-Memory Column Store Architecture	No	Buy	One of the most stable and mature databases on the market
SAP HANA	In Memory	Columnar or row-based	No	$$$	
SingleStore / former memsql	In Memory	Columnar or row-based	No	Limited Standard ed. free, Hourly	
Snowflake	Data Lake Storage	Yes	Yes, AWS, Azure, GCP	$2-4 compute / h, pre purchase discount	Scales up automatically, cost prediction difficult
HP Vertica	Disk Based	Yes	No	1TB free, per hour and TB pricing available	

Connection Settings

Connection settings

- Pushdown is affecting performance A LOT: https://documentation.datavirtuality.com/24/performance-optimization-guide
- Check if new translator options are available after an update: https://documentation.datavirtuality.com/24/reference-guide/connecting-data-sources/translators
- Check if JDBC options make sense: https://documentation.datavirtuality.com/24/reference-guide/connecting-data-sources/jdbc-connectors

data virtuality

Demo: Performance of disk based versus in memory

Switching your Analytical Storage

Switching your Analytical Storage

- Benchmark it first!
  - Write performance, copyover simulates full materialization
  - Reading performance
  - There is no difference between a “connection” and the Analytical Storage, except materialization
- If there is a native migration tool, it might be faster
- Or use the following script:

```
BEGIN
    LOOP ON (SELECT * FROM "SYS.Tables" WHERE SchemaName = 'dwh') AS dwh_table
    BEGIN
        EXECUTE IMMEDIATE 'SELECT * INTO dwh_new."' || dwh_table."Name" || '" FROM dwh."' || dwh_table."Name" ||
'"' ;
    END
END;;
```

- Challenge: data types for existing jobs
- Let it run as a job

Summary

- Choosing the right Analytical Storage will certainly affect performance
- It is easy to switch the AS in DV
- Consider pricing
- Benchmark first
- -> World domination

Any feedback / questions?

Thank you!

Please feel free to contact us at:
presales@datavirtuality.com

or

visit us at:
datavirtuality.com