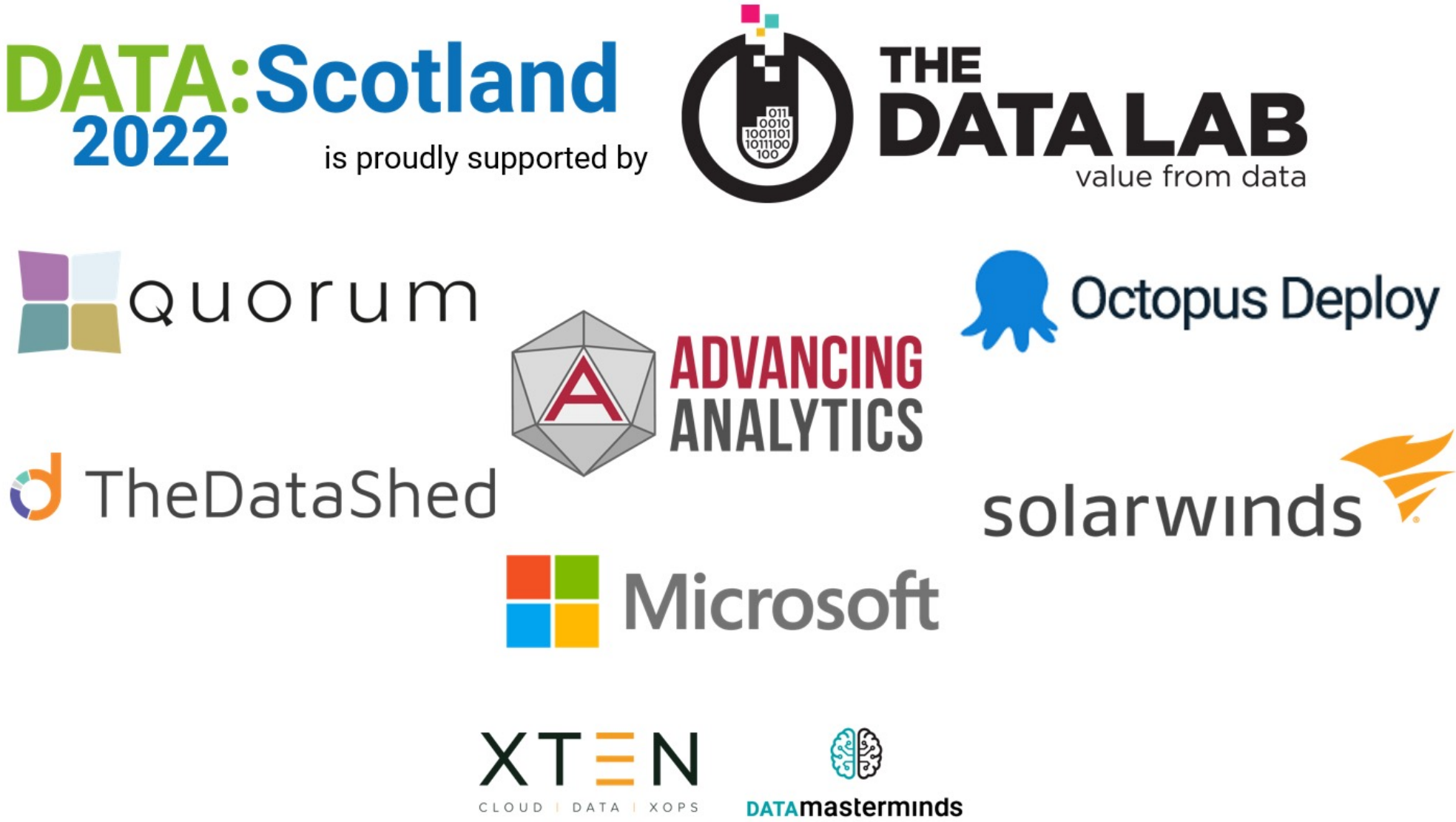
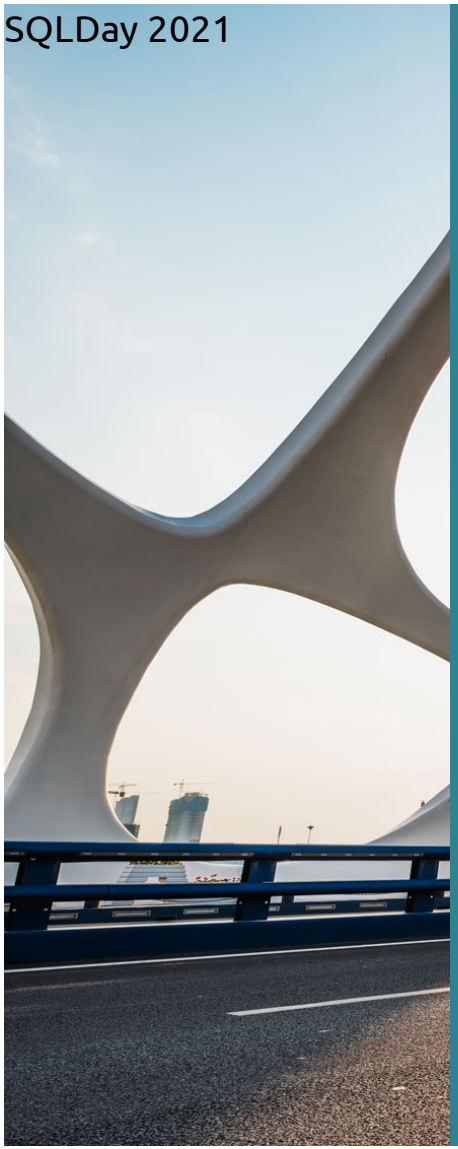

DATA:Scotland
2022
is proudly supported by
THE DATA LAB
value from data
quorum
Octopus Deploy
ADVANCING ANALYTICS
TheDataShed
solarwinds
Microsoft
XTEN
CLOUD | DATA | XOPS
DATAmasterminds


Capgemini
BATTLE OF THE CLOUD DATA WAREHOUSES
Johan Ludvig Brattås, Principal Solution Architect
Insights & Data


Capgemini
BATTLE OF THE CLOUD DATA PLATFORMS
Johan Ludvig Brattås, Principal Solution Architect
Insights & Data



AGENDA

A short historic introduction

Setting the stage

The opponents:

- **Redshift**
- **Big Query**
- **Databricks**
- **Snowflake**
- **Azure Synapse Analytics**

The outcome of the battle

THE ENTERPRISE DATA WAREHOUSE

Been around since 1988 (the Business Data Warehouse...)
Evolved over time
contains structured, schema on-write data
relational data
stores historical data that needs to be transformed
typically transform before storing
good for business decisions

THE PRINCIPLES OF THE ENTERPRISE DATA WAREHOUSE

high data quality and focus on exact numbers

«long» development life cycles

DATA LAKE

gained foothold around 2011

answer to pains in EDW on the 3 Vs of big data (volume, variety, velocity)
or 4 or 5 Vs...

typical varieties are:

the raw data lake

the business data lake or layered data lake

DATA LAKE

transformation only from raw layer to curated layers.

various use cases and users on various layers of the data lake

great for trends and ad-hoc analysis

short time to market development cycles

enables data science and self-service

danger of data swamp

DATA LAKE

Transformation happens in the upper layers

No normalization between source systems

Layered data lake gives different user groups different access in layers

Enables data science and self-service

Danger of data swamp (on-prem)

Strenght:

Great for trends and ad-hoc analysis

Short time to market development cycles

THE CLOUD DATA WAREHOUSE

Initially a response on challenges faced by traditional RDBMS

Massivelly Parallell Processing (MPP)

Still a take on EDW

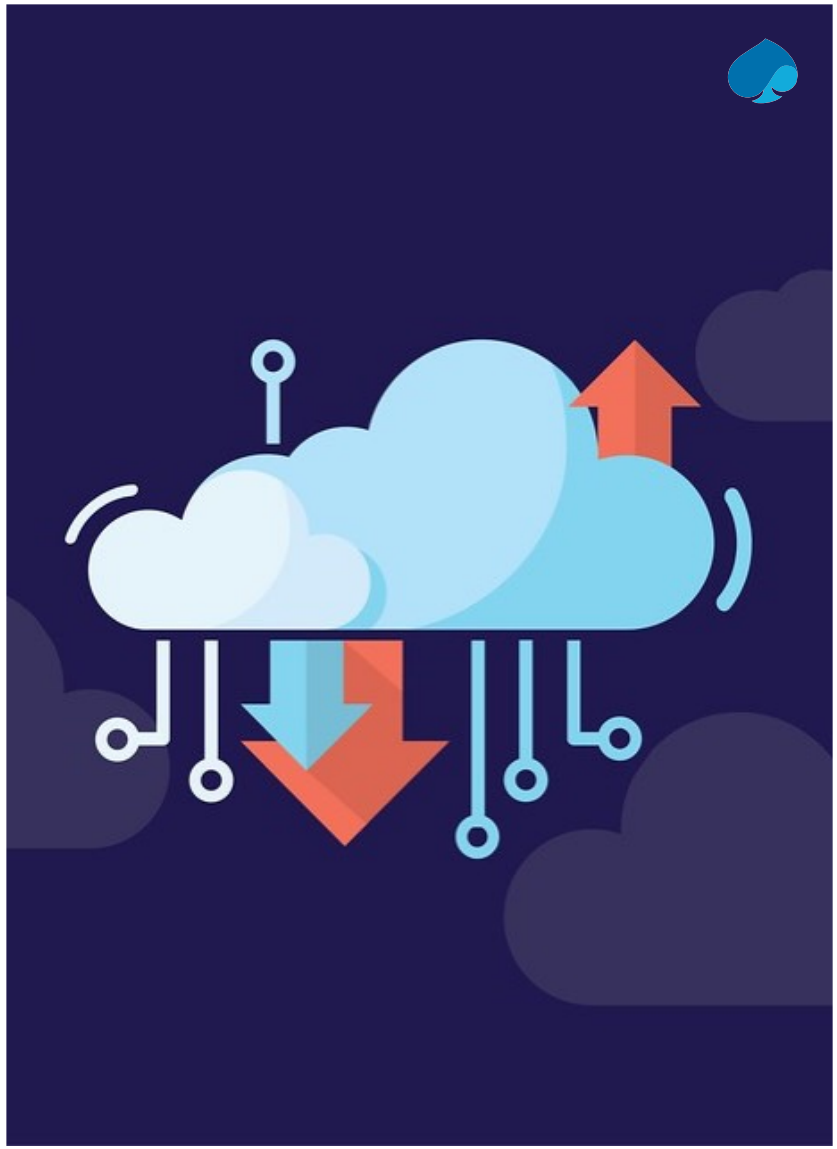
THE CLOUD DATA PLATFORM

Can data lake functionality and EDW merge somehow?

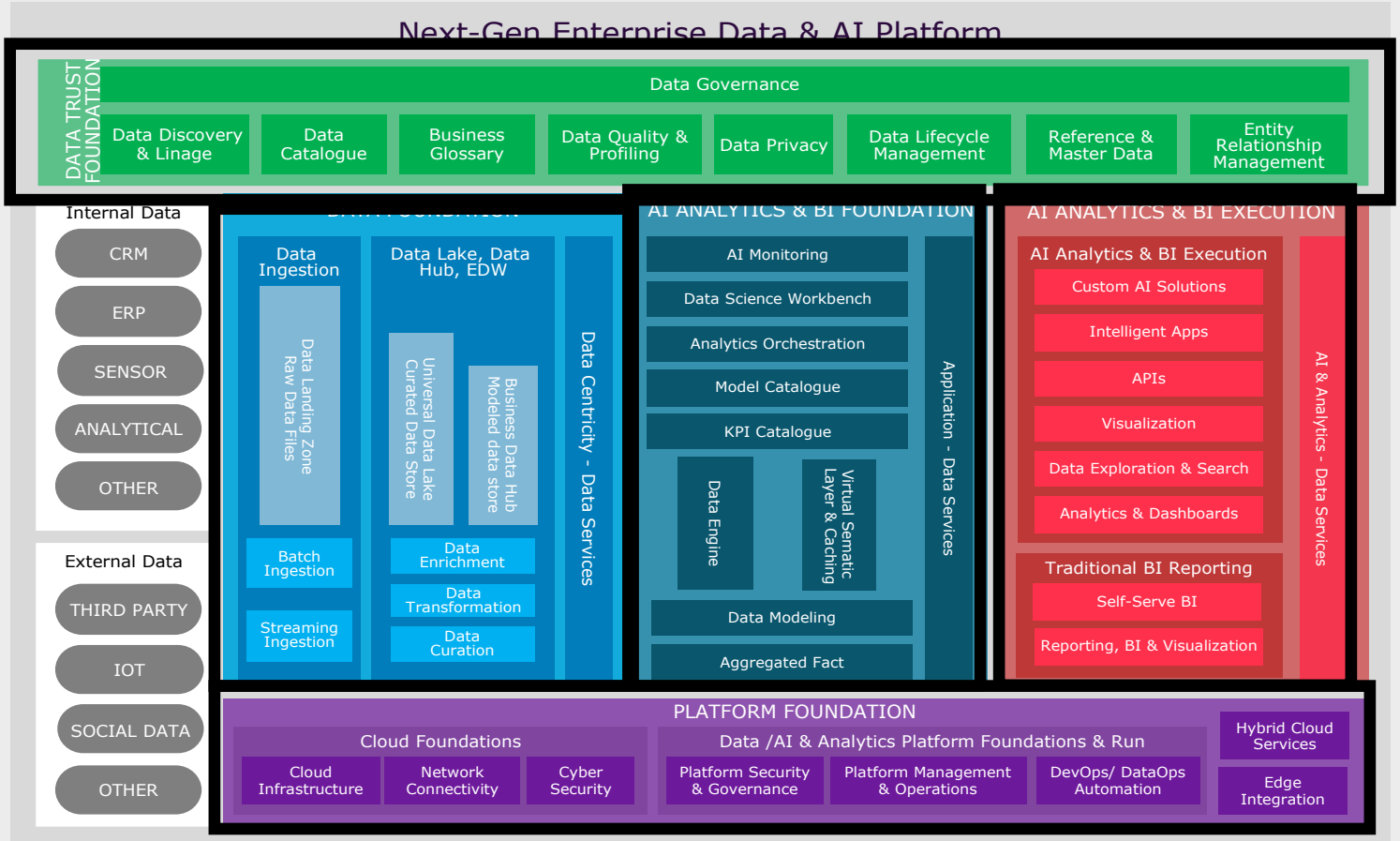
Suggestions for solving the issues:

- Logical data warehouse
- Cloud data warehouse
- Virtualization

Enter the new cloud data platforms



COMPONENTS OF A DATA PLATFORM



THE OPPONENTS

aws

Google Cloud

REDSHIFT

Redshift is a MPP RDBMS

Based on PostgreSQL

Columnar

Distributed databases

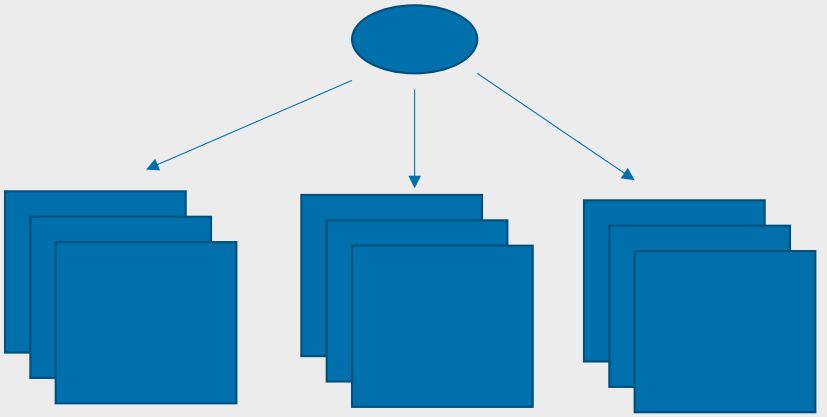
REDSHIFT

What is a distributed database?

Distributing data across databases within cluster

AWS distribution types:

Key
All
Even



REDSHIFT

Redshift is a MPP RDBMS

Based on PostgreSQL

Columnar

Distributed databases

3 versions:
Dense Storage - DS2
Dense Compute – DC2
Aqua Advanced query Accelerator – RA3

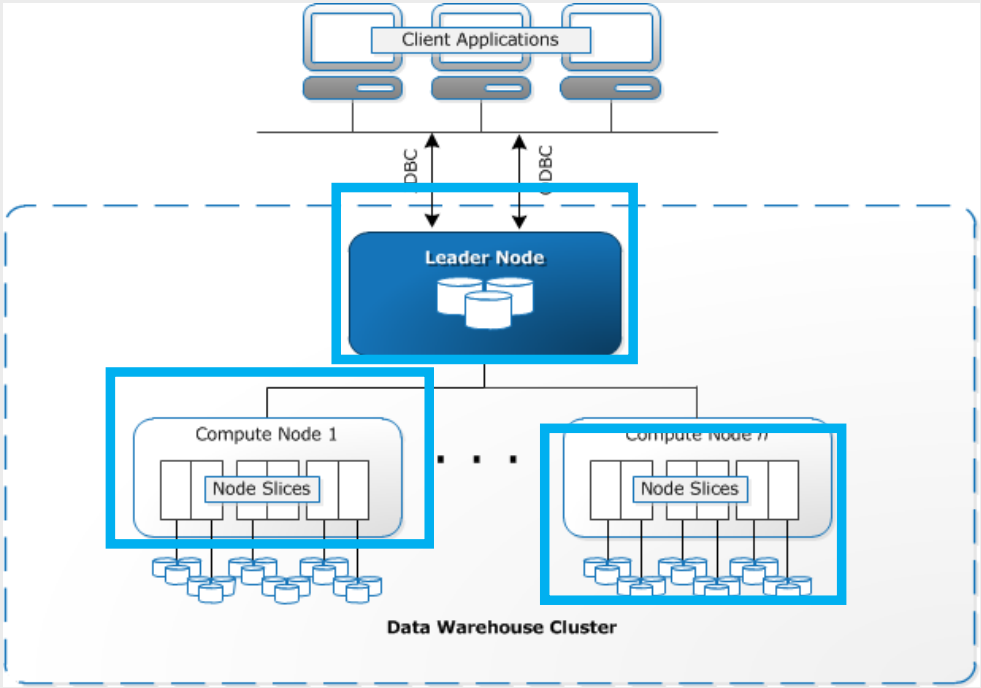
REDSHIFT

Architecture of DS & DC versions

Leader node

Compute nodes

Compute slices and Storage combined



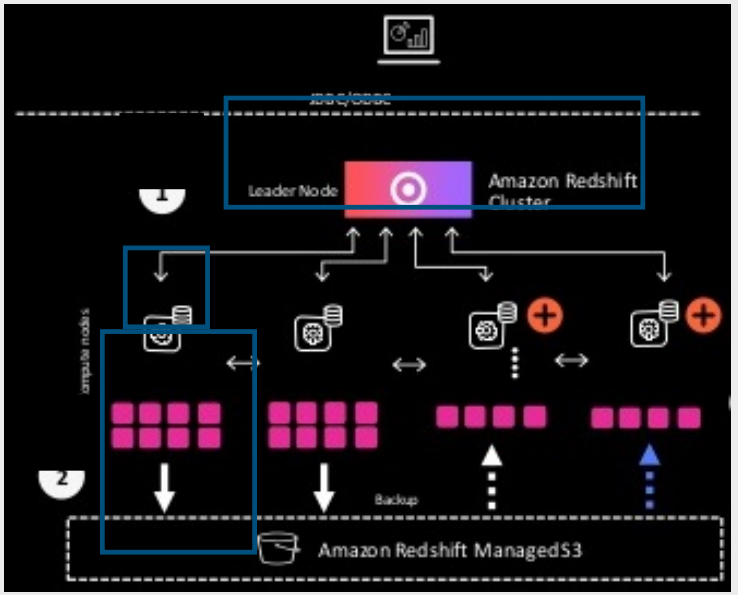
REDSHIFT

Architecture of RA3 versions

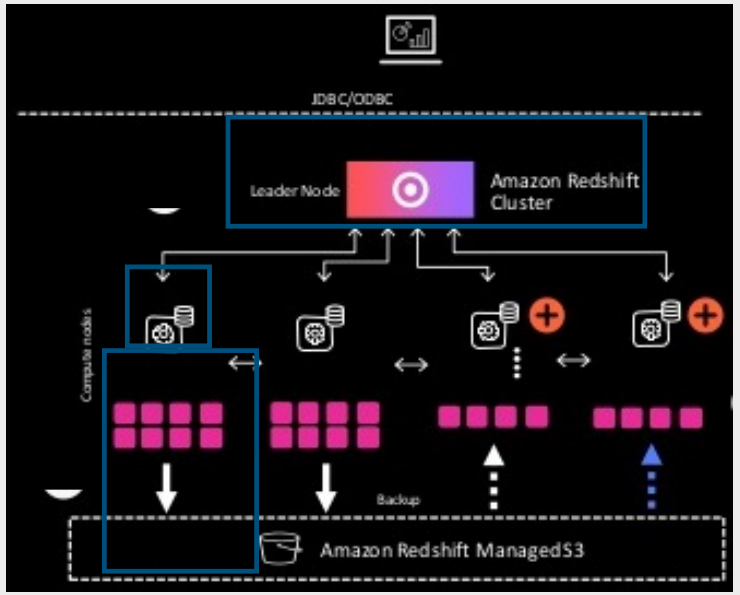
Leader node

Compute nodes

Managed Storage – SSD + S3



REDSHIFT



Architecture of RA3 versions

Leader node

Compute nodes

Managed Storage – SSD + S3

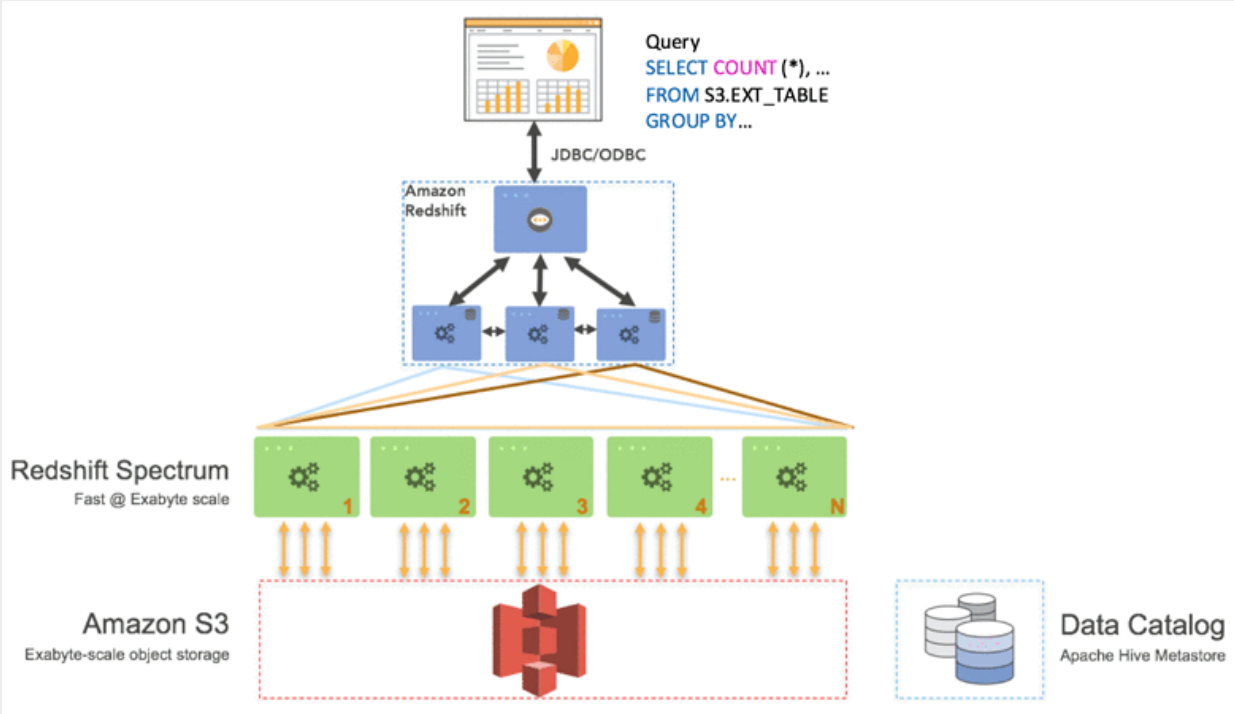
REDSHIFT

Redshift Spectrum

Query your data lake with SQL

External tables

Materialized views



Snowflake is a «cloud data platform»

Based on ANSI SQL

Columnar

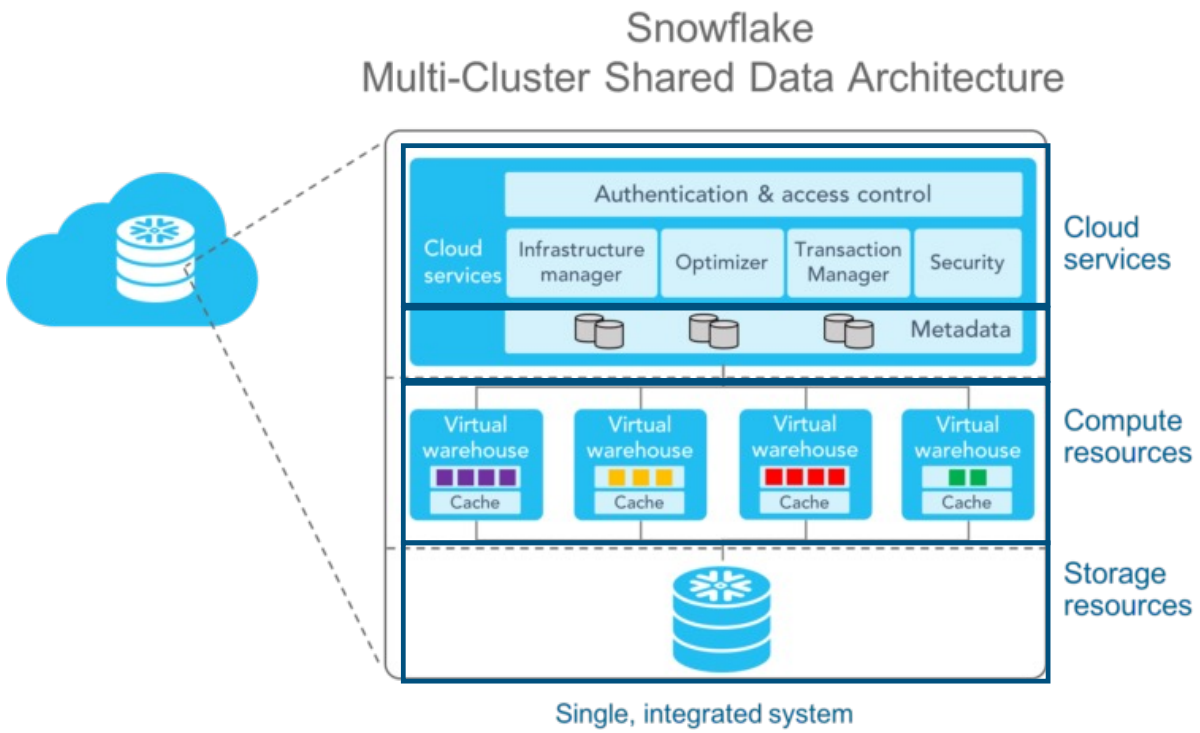
Separate compute – called warehouses

Storage is separate from compute, but shared among the warehouses

«Maintenance free»

snowflake®

- Snowflake architecture
- Common Services
- Separate metadata repository
- Compute resources with caching
- Shared storage
- Connectors to main clouds



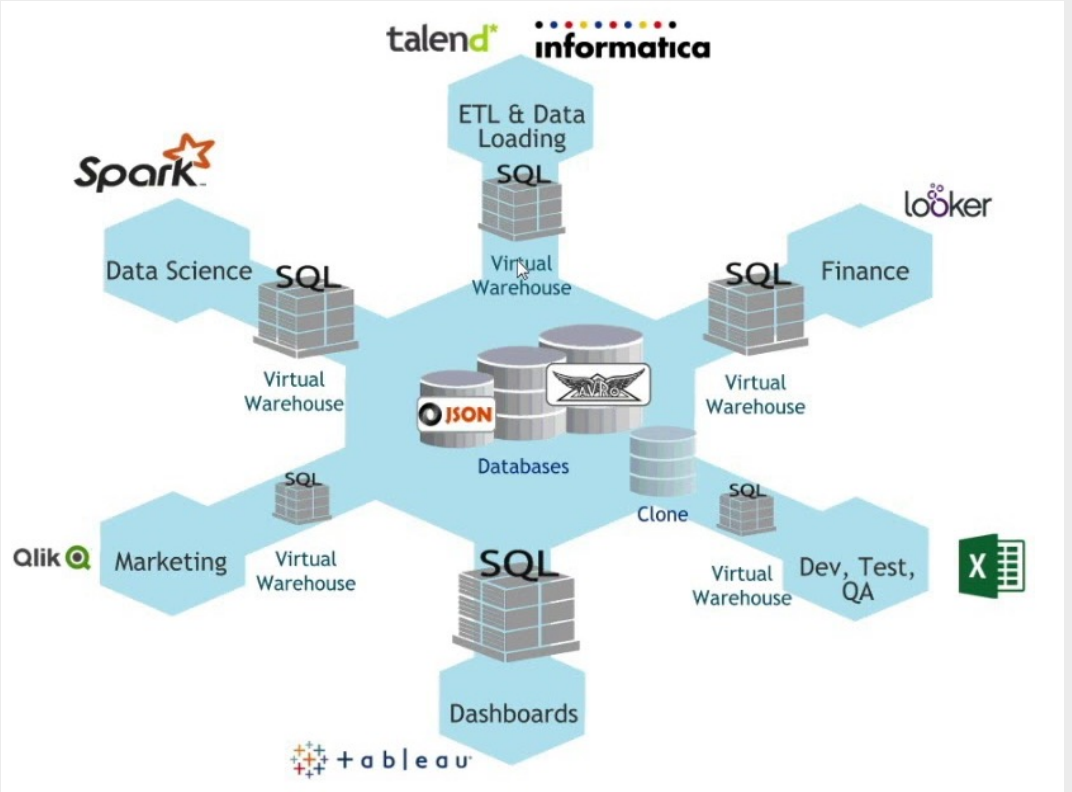
snowflake®

Dedicated warehouses per use case

Auto-scalable

Auto-pause option

Quick data cloning for dev/test



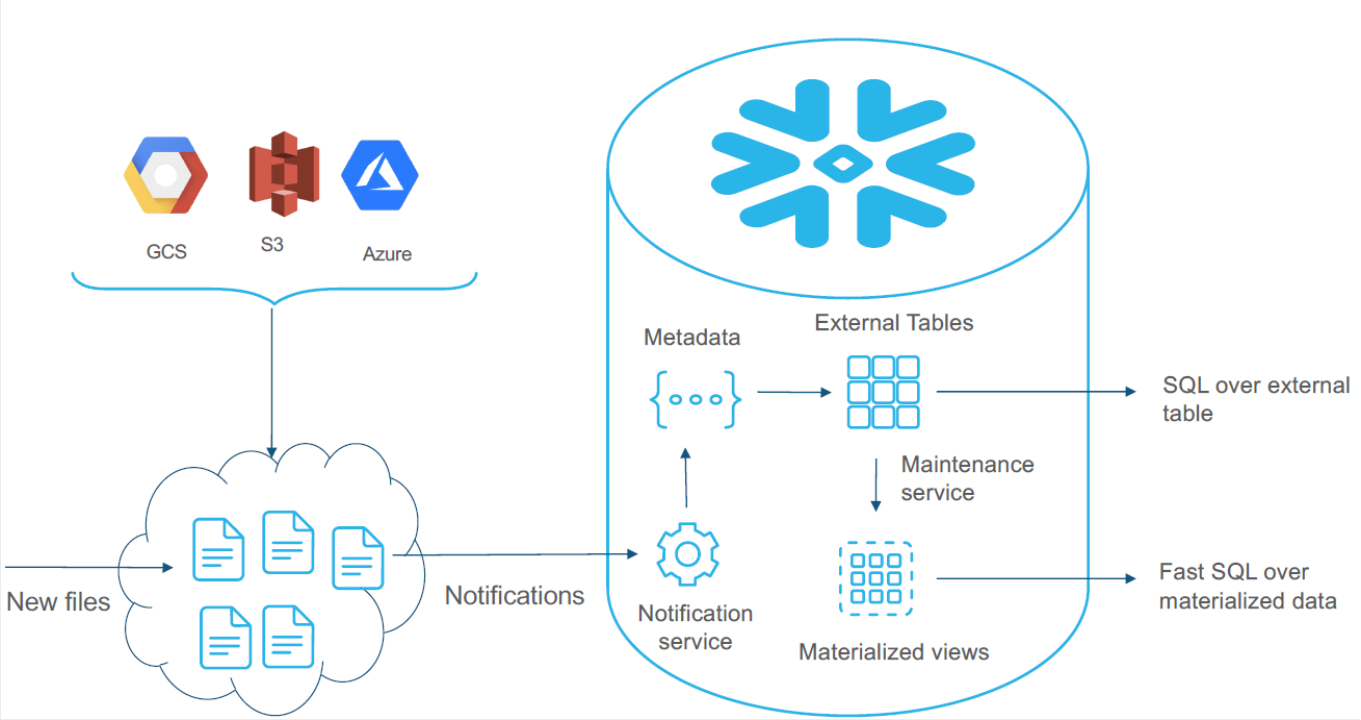
snowflake®

Query over data lakes

File share

External tables

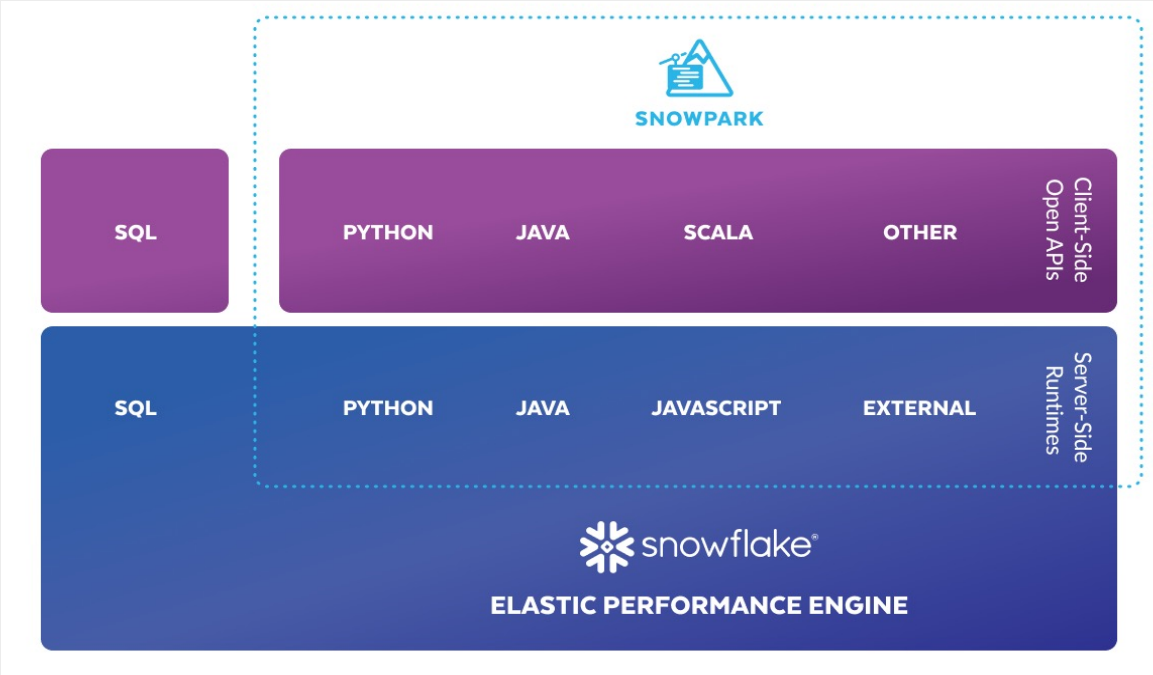
Materialized views



snowflake®

Why Snowflake is a «cloud data platform»

Now also with Snowpark...



Azure Synapse Analytics

Synapse dedicated SQL Pools – the service formerly known as SQL Data Warehouse

SQL Server MPP

T-SQL

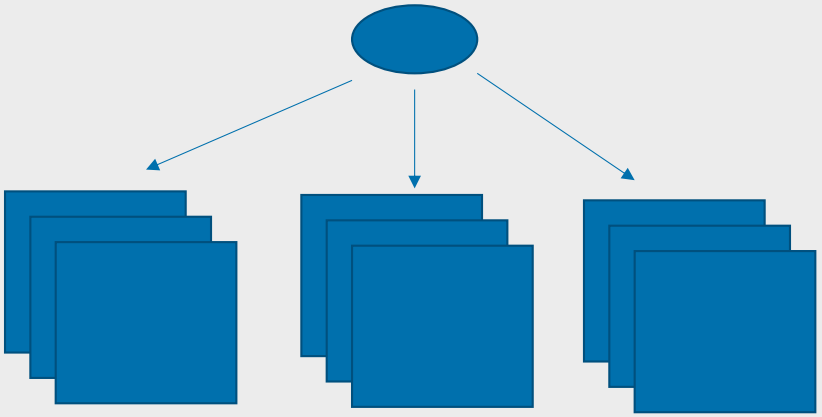
Columnar per default (rowbased optional)

Separate compute and storage

Distributed databases

Azure Synapse Analytics

What is a distributed database?

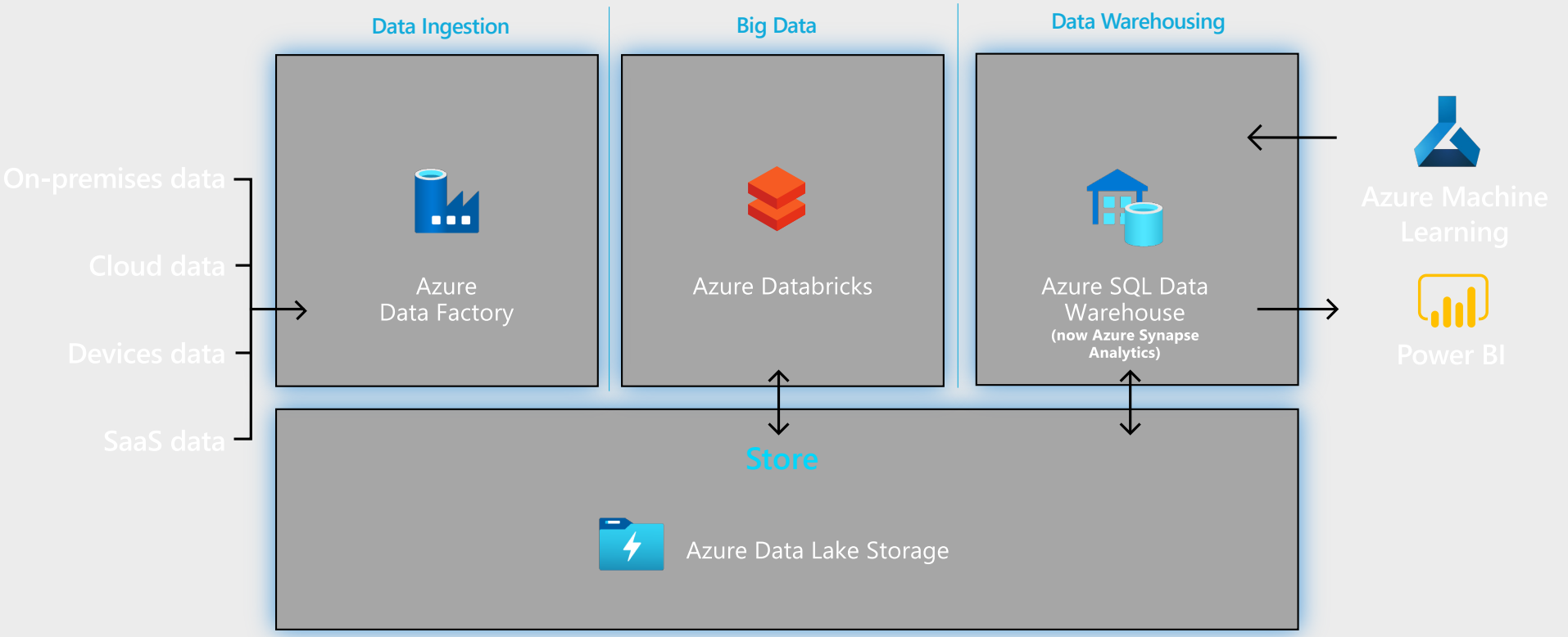


Distributing data across databases within cluster

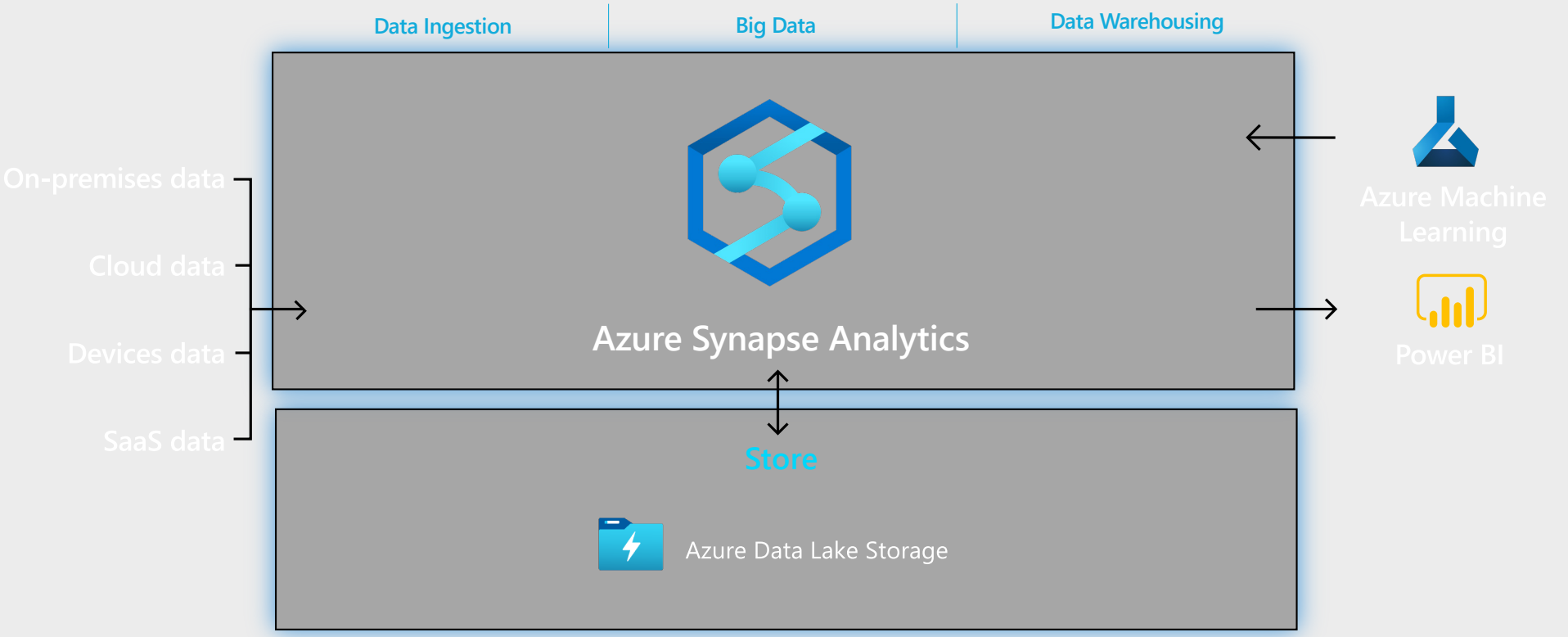
SQL DWH distribution types:

Partition key
All
Round Robin

Azure Synapse Analytics



Azure Synapse Analytics



Azure Synapse Analytics

Synapse – more than just MPP RDBMS

Serverless SQL

Polybase 2.0 for queries on external data sources

Azure Spark

Azure Data Factory

Power BI

Single workspace

Azure Data Explorer

Purview integration

BIGQUERY

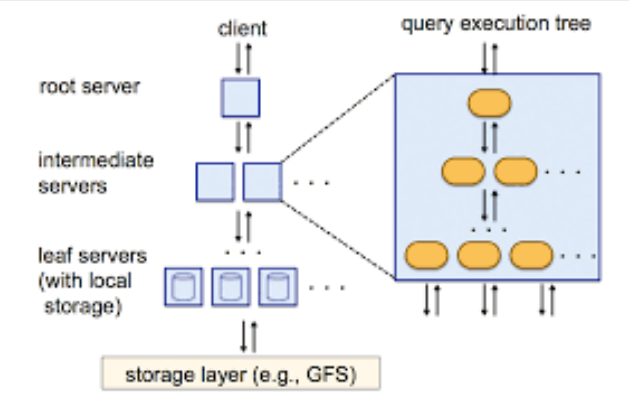
Google Cloud

«Serverless»

ANSI SQL

Separate compute and storage

Columnar



BigQuery

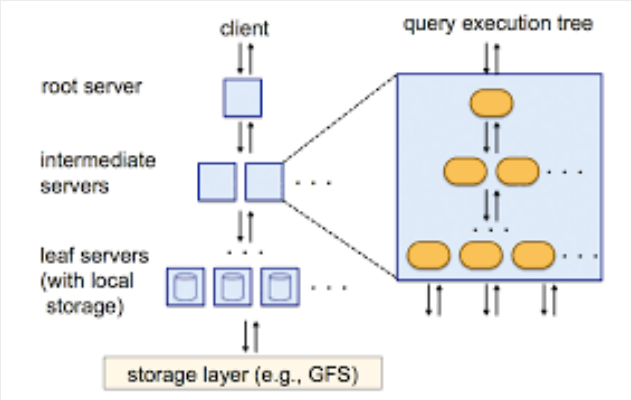
Google Cloud

Pay per query – not capacity

Pay for storage

Optimized for wide tables

Supports streaming data as well as batch



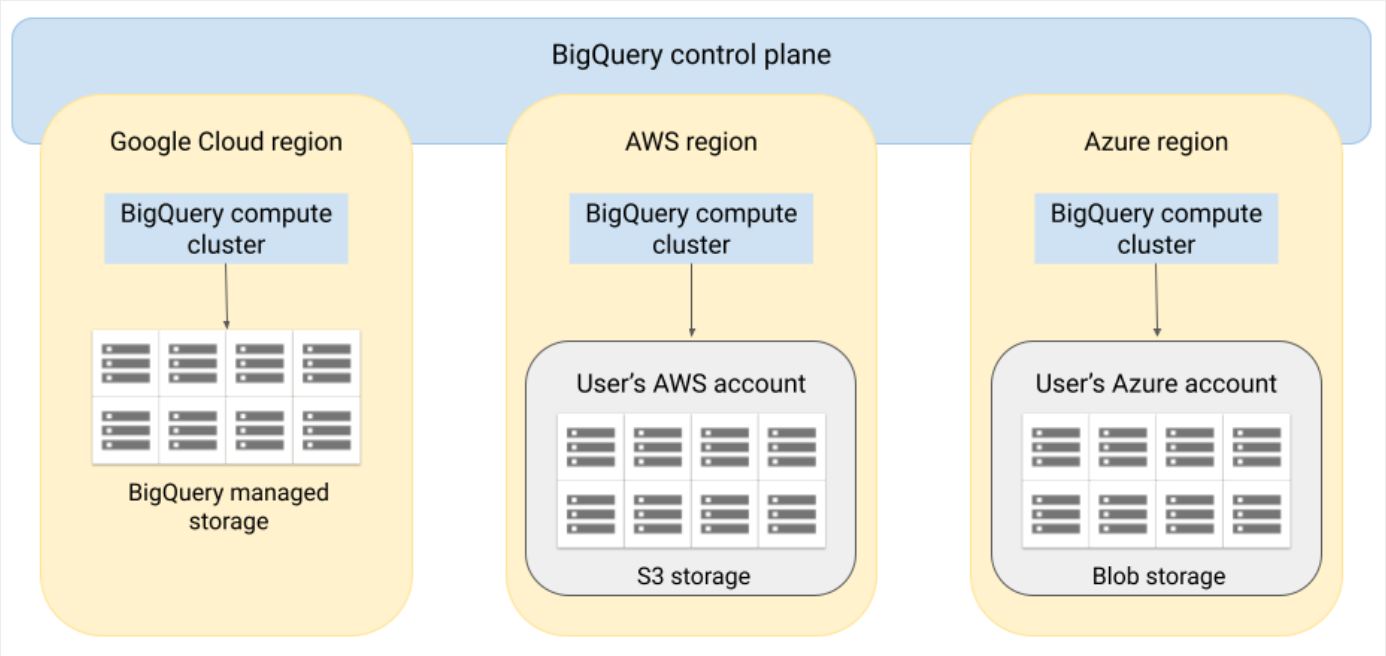
BigQuery

Google Cloud

BigQuery Omni

Multi-cloud functionality

Enabled by Anthos



Databricks – started out as containerized Spark.
Available on all hyperscalers!

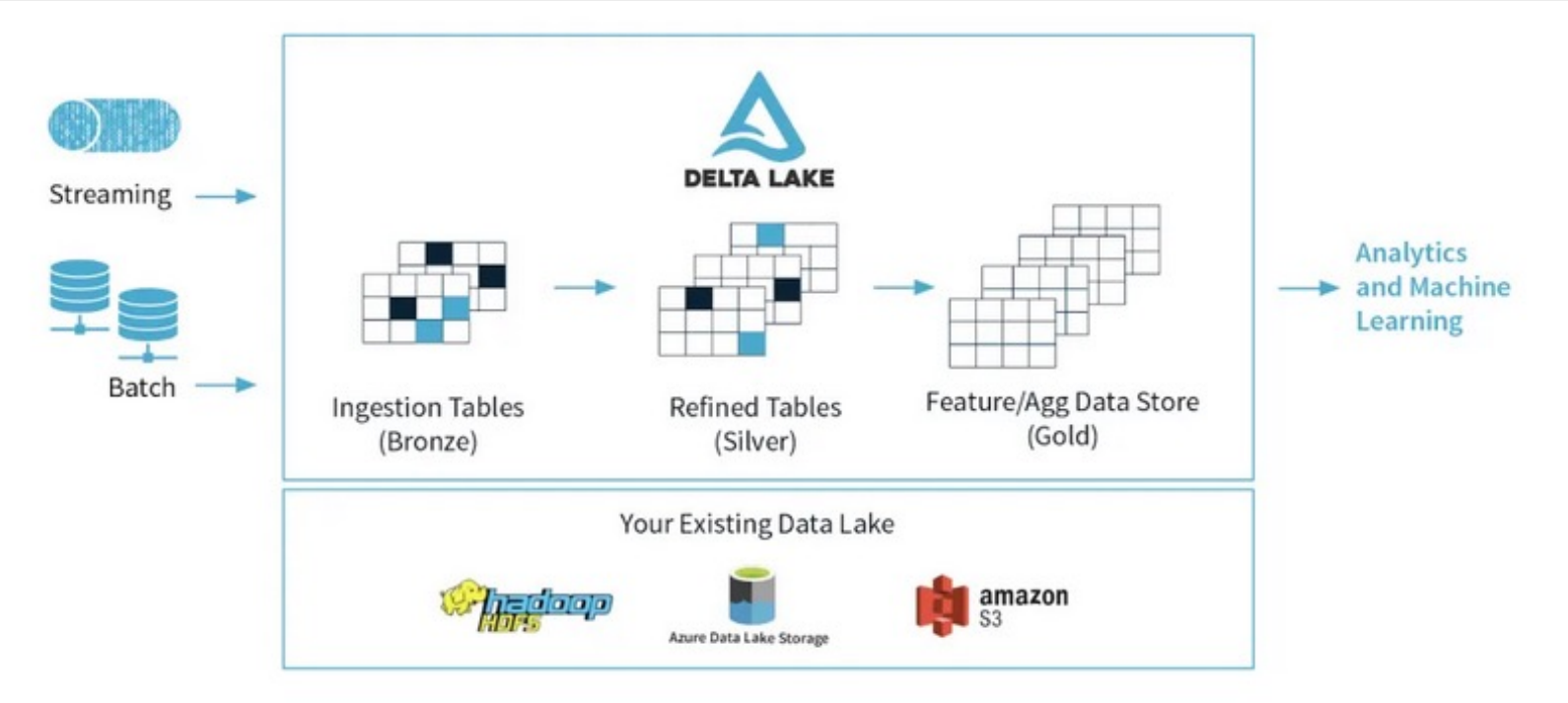
- Distributed In-memory processing
- Multi-language support – Scala, Python, Spark SQL...
- Common development environment with workspaces and notebooks
- Not a relational database

However: Enter Delta lake

- Filebased (parquet files)
- Brings ACID principles to data lake (supports S3, ADLS, GCS, and HDFS)
- Unified batch and streaming source and sink
- Schema enforcement
- Schema evolution
- Transaction logs
- Time travel
- Updates and deletes
- Audit history
- Open source

databricks

Databricks Data Lakehouse



Databricks is now a cloud data platform

- SQL Analytics (Photon engine)
- SQL Serverless (preview)
- Unity Catalog
- Delta Live Tables

ANALYSIS

Functionality:

All opponents superficially resemble each other:
- Query using some SQL dialect – but also support for Python/R

Yes – easy to start using.

However, Databricks is the only solution that did not start out as RDBMS

ANALYSIS

Functionality:

All opponents superficially resemble each other:
- Query against external data

Yes, but only Synapse and Databricks offers connections to other RDBMS as part of the package – with BiqQuery also against CloudSQL

And Redshift Spectrum incurs extra costs per query

ANALYSIS

Functionality:

All opponents superficially resemble each other:
- Scalable

Yes, but Redshift only offers 2 different scales per version

BigQuery is considered serverless in this setting

Redshift, Synapse and Databricks also offers serverless features

ANALYSIS

Functionality:

Loading data in:

- All supports external ETL tools
- All services offer tools as part of package as well – though Redshift depends on AWS services such as Glue
- Redshift and Synapse SQL Pools need design decisions on distribution
- BigQuery need design decision to avoid joins
- Snowflake and Databricks don't have these constraints

Analysis

Functionality:

Querying against data:
- All offer browser based query tool
- All offer API’s
- Synapse and Databricks have the richest development workspaces

Analysis

Speed:

Everyone claim they are the fastest!
Which one wins?

Eh, it really depends…

Analysis

Availability:

Synapse available worldwide

Snowflake worldwide, but fewer data centers per region

Redshift worldwide

BigQuery worldwide

Databricks worldwide (and on all hyperscalers)

Analysis

Pricing:

Redshift has varying prices per version and then per tier. As well as separate pricing for Spectrum

Snowflake prices storage per TB, and compute per second and priced according to node size at compute time. Scalable

Synapse has auto scalability up and down on SQL DWH, serverless options, Spark cluster selections. Pricing varies for each part of the service

BigQuery prices for storage, streaming inserts and per Mb passed through query
– though there are various pricing alternatives

Databricks pricing for compute size, size of cluster – analytics engine has separate pricing.

Analysis

Other features:

Azure Synapse Link + Cosmos DB or Dataverse or SQL

Azure Synapse Pathway

Analysis

Why aren’t every player «real data platforms»?:

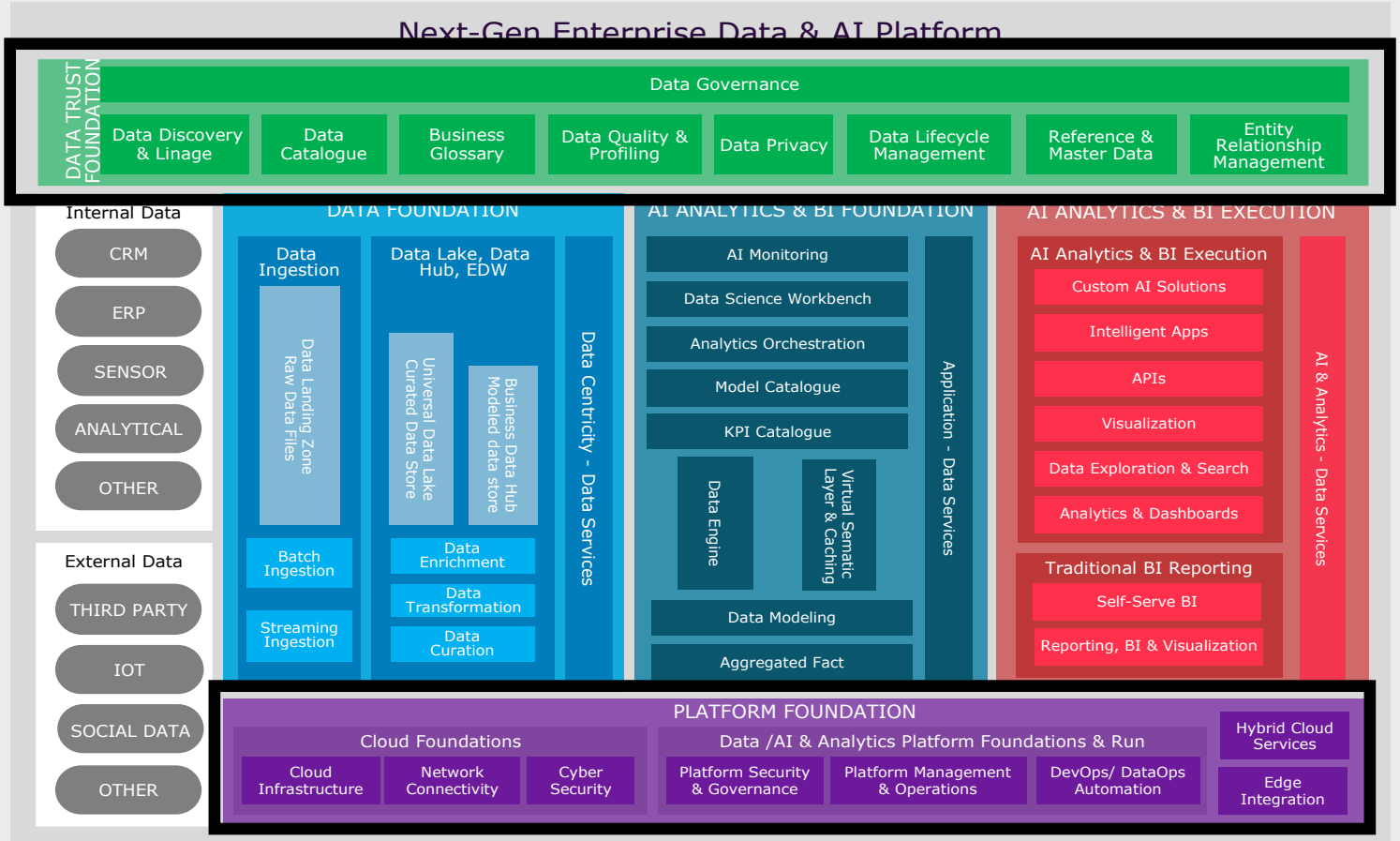
AWS,GCP and Snowflake lacking in the data governance and trust domains.

AWS has a catalog if you use Glue for ETL – but…

None have proper lineage, data quality and governance tools

There are 3rd party tools of course.

COMPONENTS OF A DATA PLATFORM



DO WE HAVE A WINNER?

Session Feedback

Event Feedback

About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of 270,000 team members in nearly 50 countries. With its strong 50 year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fuelled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2020 global revenues of 16 billion.

Get the Future You Want | www.capgemini.com

Name, Last Name

Title/Role
Capgemini Office (Optional)
Address Line 1
Address Line 2
Address Line 3

Name, Last Name

Title/Role
Capgemini Office (Optional)
Address Line 1
Address Line 2
Address Line 3

Name, Last Name

Title/Role
Capgemini Office (Optional)
Address Line 1
Address Line 2
Address Line 3

Name, Last Name

Title/Role
Capgemini Office (Optional)
Address Line 1
Address Line 2
Address Line 3