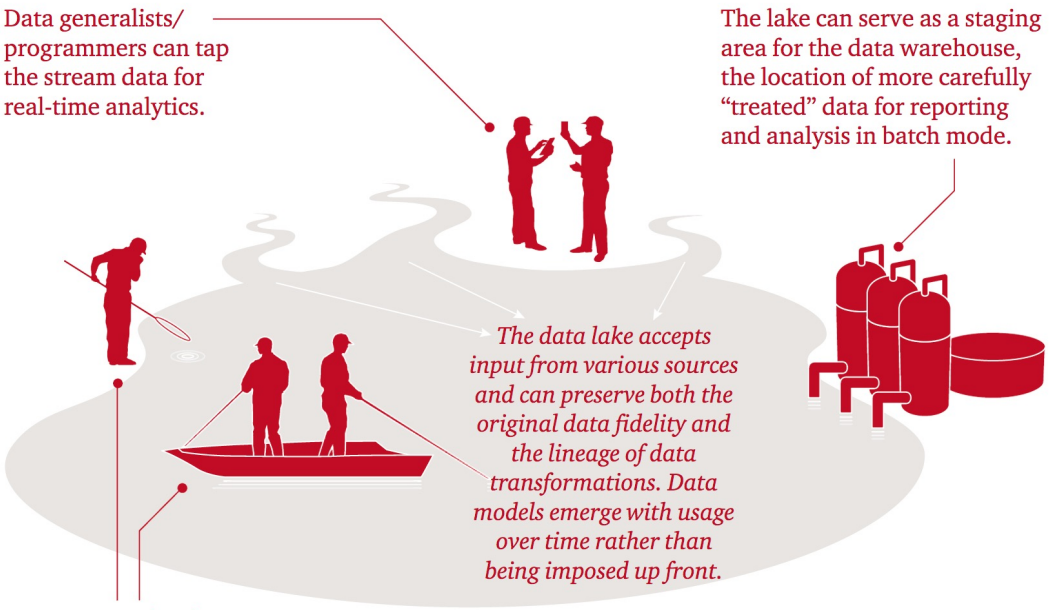
Data Lake

Enterprise Architectures for Big Data

What is a Data Lake?

A repository for large quantities and varieties of data, both structured and unstructured.



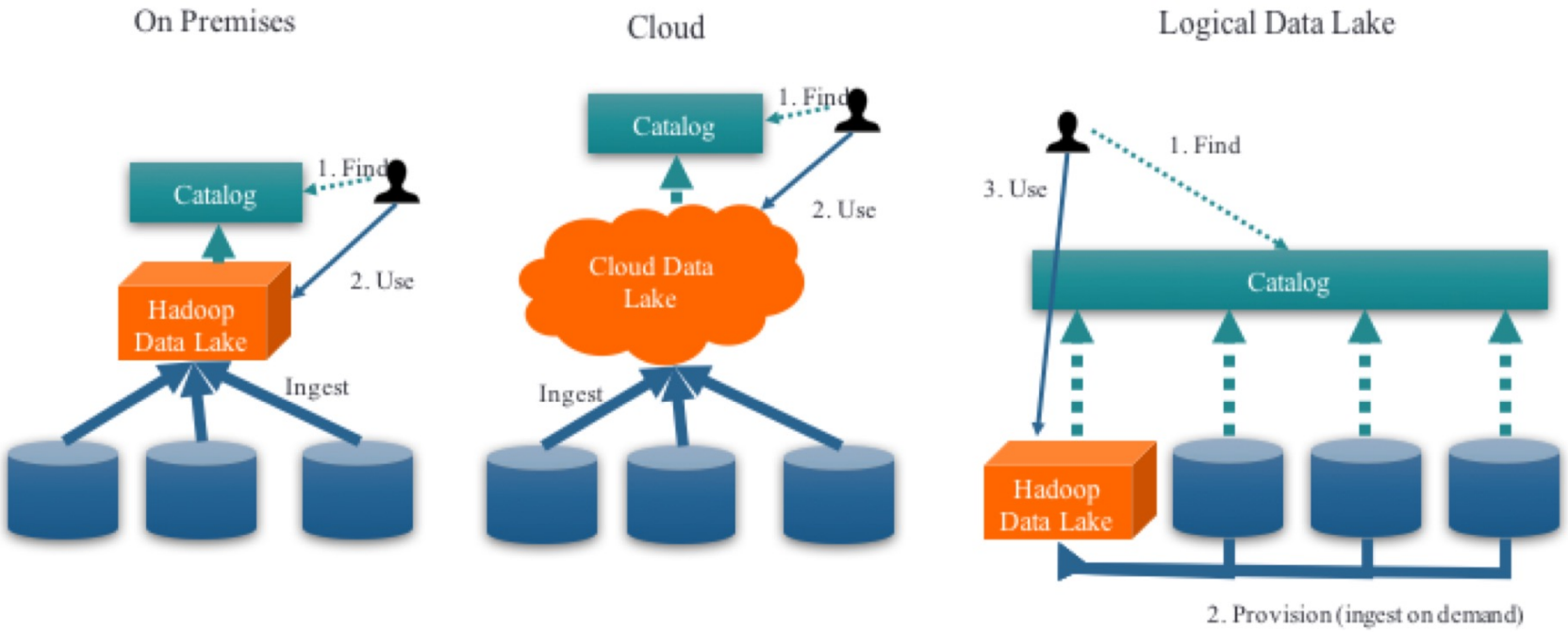
Data lakes take advantage of commodity cluster computing techniques for massively scalable, low-cost storage of data files in any format.

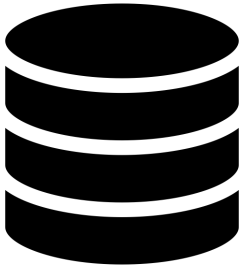
Source: http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf

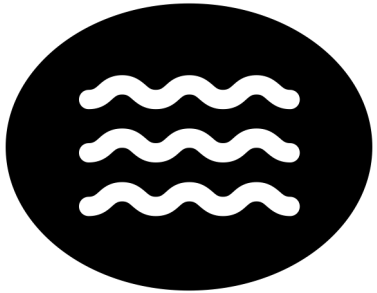
Data Lake

- Repository for raw data
- Includes structed, semi-structured and unstructured data
- Often Hadoop / HDFS based
- No upfront schema (schema at read vs. schema at write)
- Users often Data Scientists
- Could be a source for downstream systems like data warehouses
- Flexible usage and access

Different Data Lake Architectures

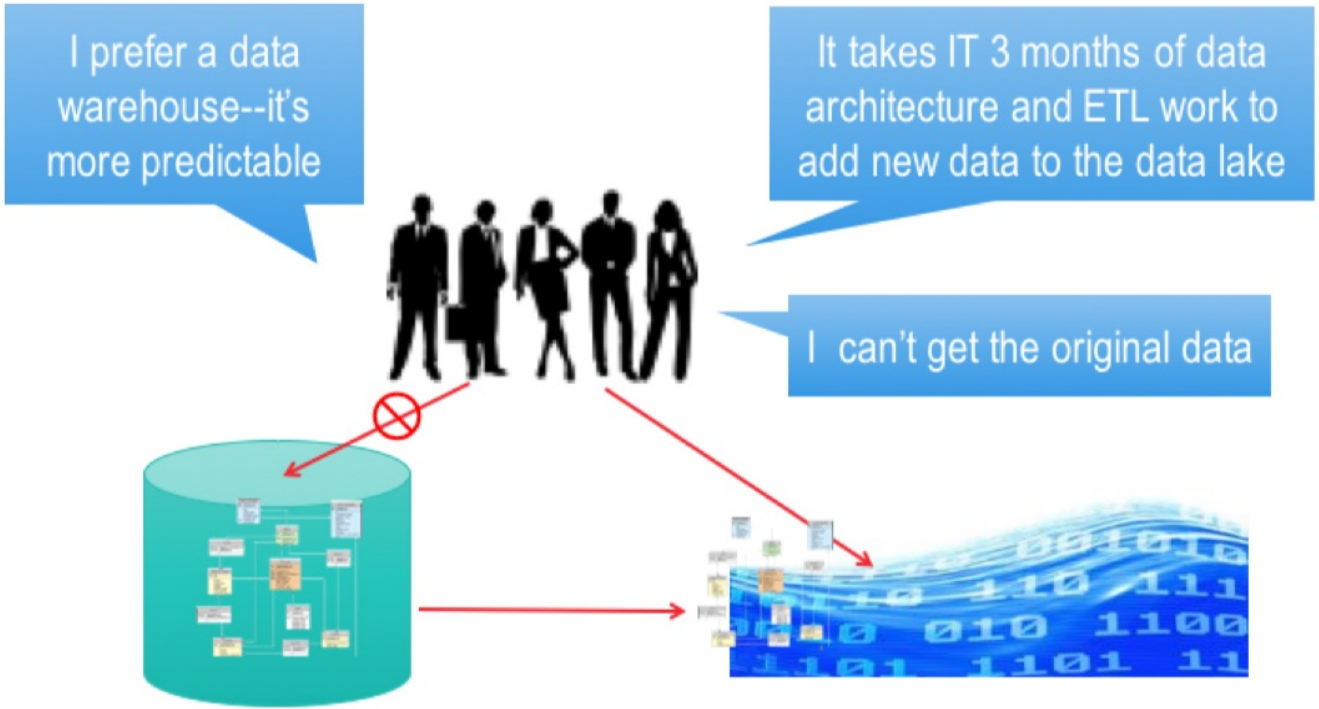






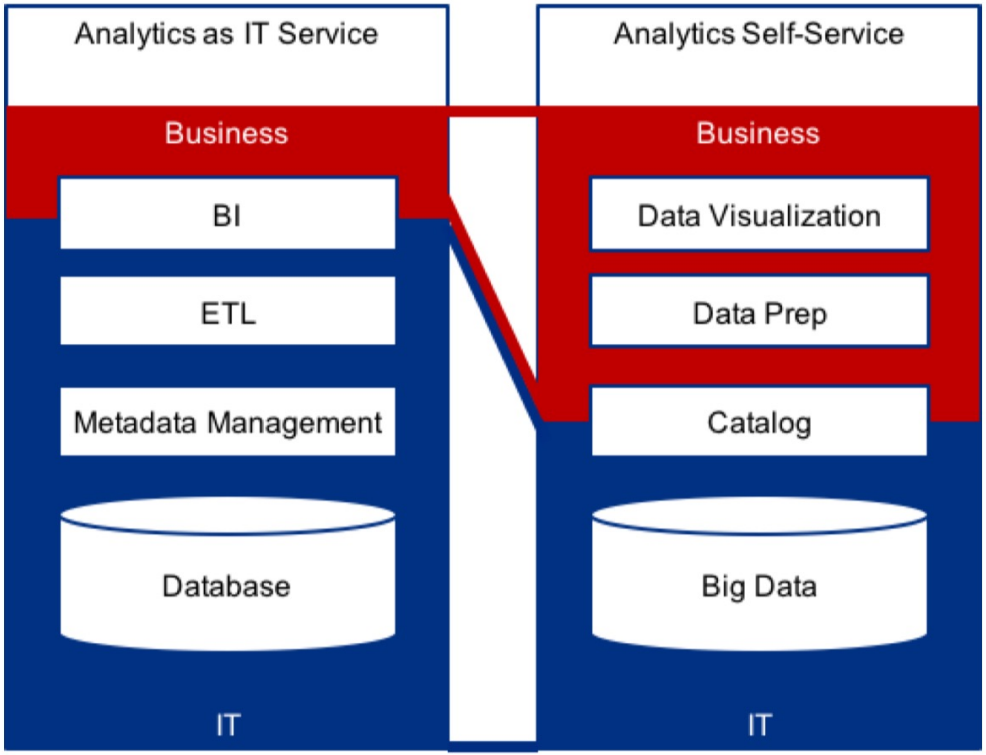
Data Warehouse vs. Data Lake

Drawbacks of data warehouse offloading



Source: Gorelik (2019)

Enabling analysts and reducing the load on IT with self-service analytics



Source: Gorelik (2019)

Data Warehouse Solves two main problems

1. Integrated Data Schema
2. Data Locality

Data Warehouses vs. Data Lakes

Data Warehouses

- Schema on Write
- ETL > Analytics/OLAP
- Queries mostly Interactive
- Structures Data
- Often RDBM based

Data Lakes

- Schema on Read
- EL > T > Analytics/OLAP
- Queries mostly Batch
- Structured, Semi-Structured and Unstructured
- Often Hadoop Based

Schema on-write vs. Schema on-read

Attibute	Schema On-Write (Early Binding) Data Warehousing	Schema on-read (Late Binding) Data Lake
Data Provider	• Evaluate Data • Define Data Structure • Collect Data & Ingest • Apply Structure	• Collect Data & Ingest
Data Consumer	• Answer Questions	• Evaluate Data • Define Data Structure • Apply Structure • Answer Questions
Ideal for	• Reused & Known Data • Consistent Results • The Masses	• Unfamiliar Data • Infrequent usage • Unstable source schema

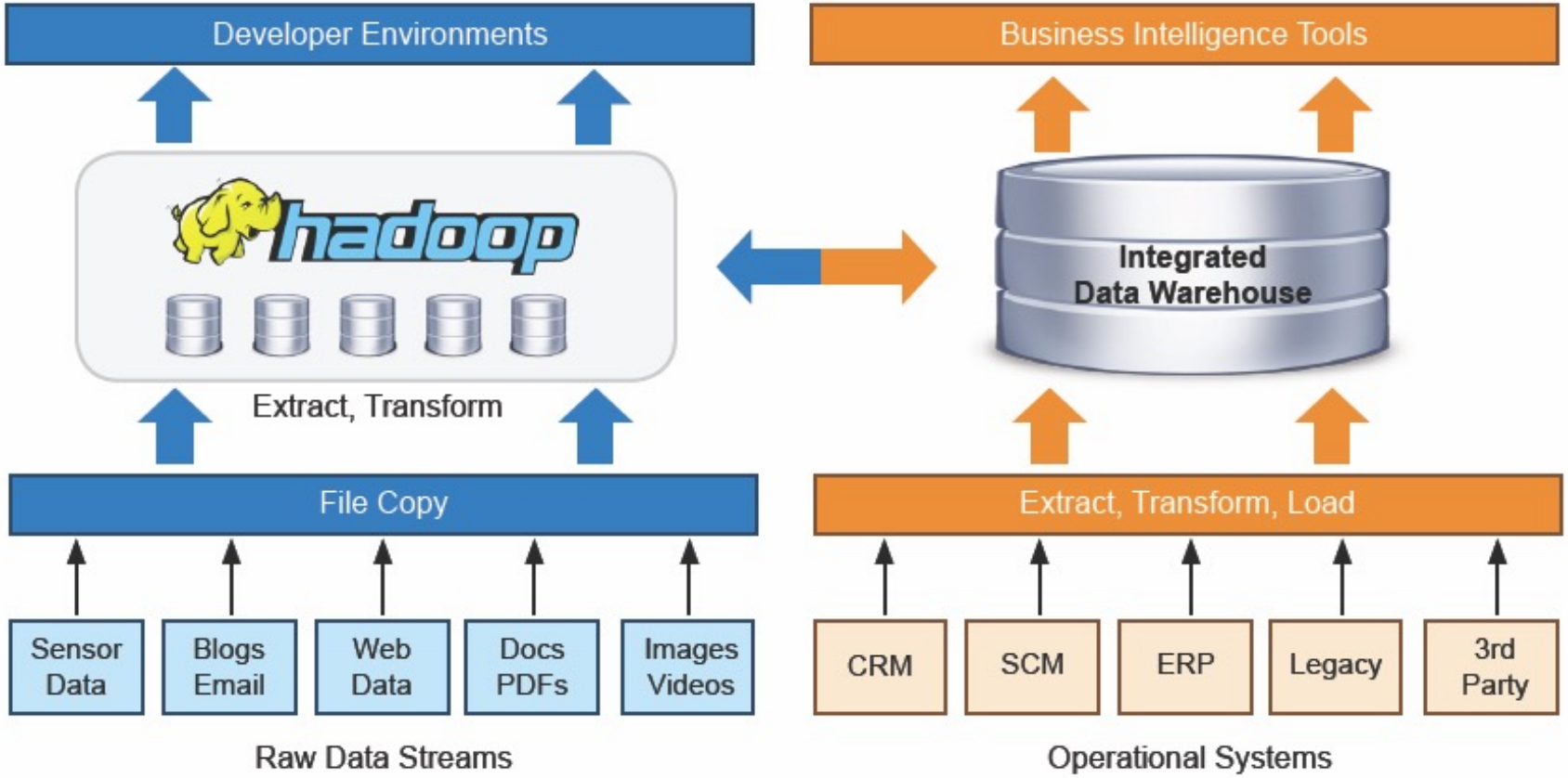
Source: Fang (2015)

Schema on-write vs. Schema on-read

Attibute	Data Warehousing	Data Lake
Workload	• Hundreds to thousands of concurrent users • Performing online (interactive) analytics • Advanced workload management capabilities • Batch processing	• Batch processing of data at scale • Currently improving its capabilities to support more interactive users
Schema	• Typically schema is defined before data is stored • Requires work at the beginning of the process, but offers performance, security and integration	• Typically schema is defined after data is stored • Offers extreme agility and ease of data capture, but requires work at the end of the process. • Works well for data types where data value is not known
Scale	• Large data volumes at moderate cost	• Extreme data volumes at low cost

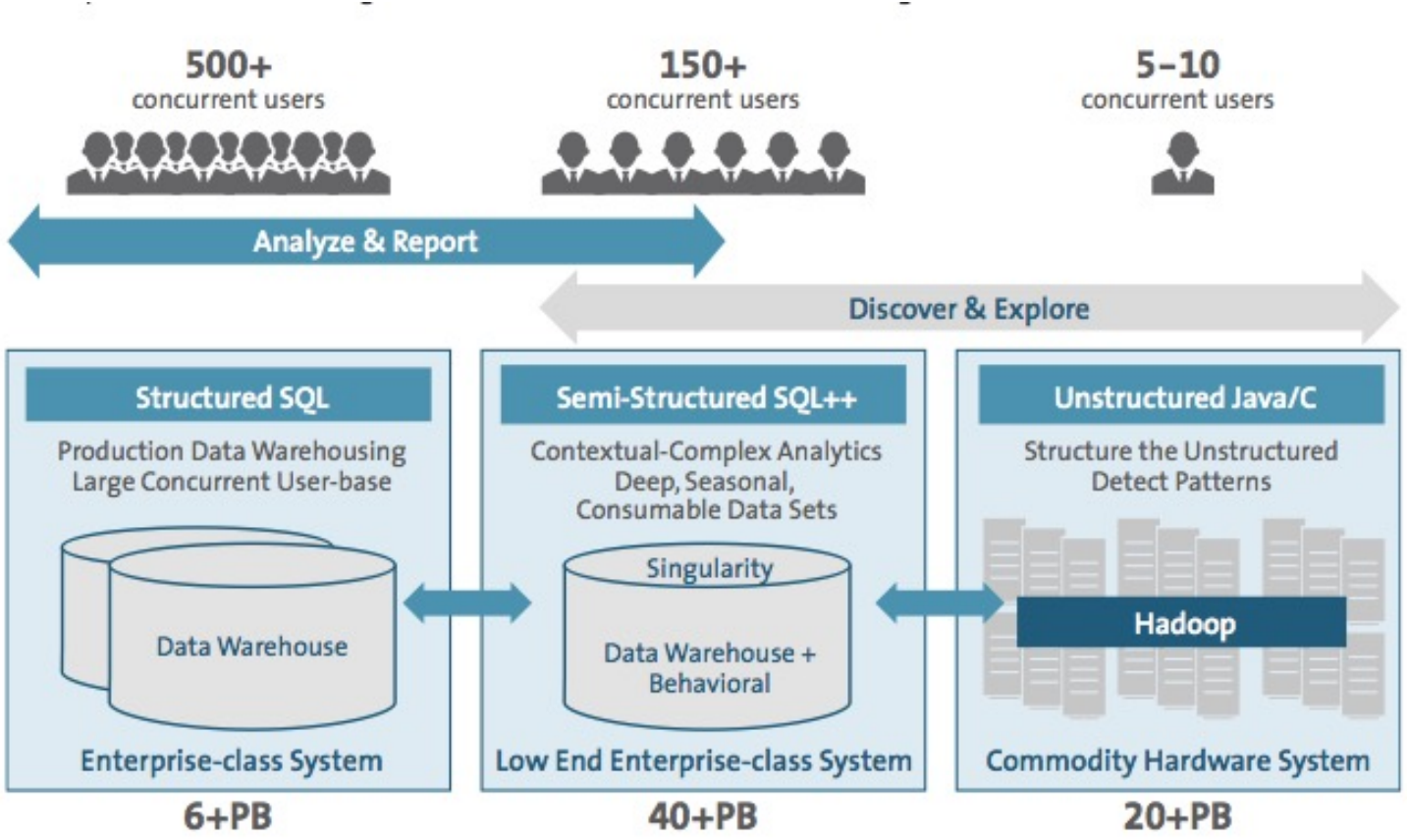
Source: Fang (2015)

Coexistence of Hadoop and DW

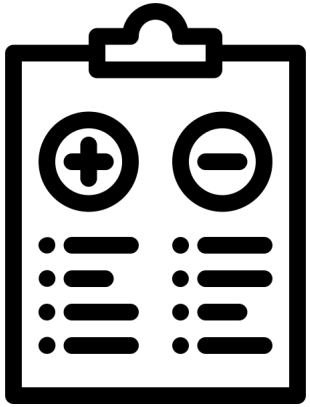


Source: Teradata

Big Data & Hadoop at eBay



Source: Hortonworks



Pro and Cons

Advantages & Disadvantages of Data Lakes

Advantages

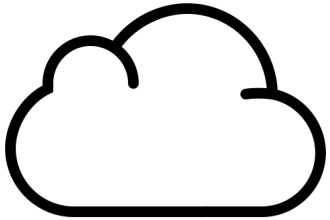
- Size
- Low cost
- Flexibility
- Easy accessibility
- No Data Silos
- “let’s store everything” strategy
- Self-Service Analytics
- Data Catalog of all data

Disadvantages

Often Problems with

- Data Quality
- Data Lineage (origin of the data)
- Missing Metadata
- Different Schemas
- Not Integrated
- Risk of a “data graveyard” / “Data Swamp”

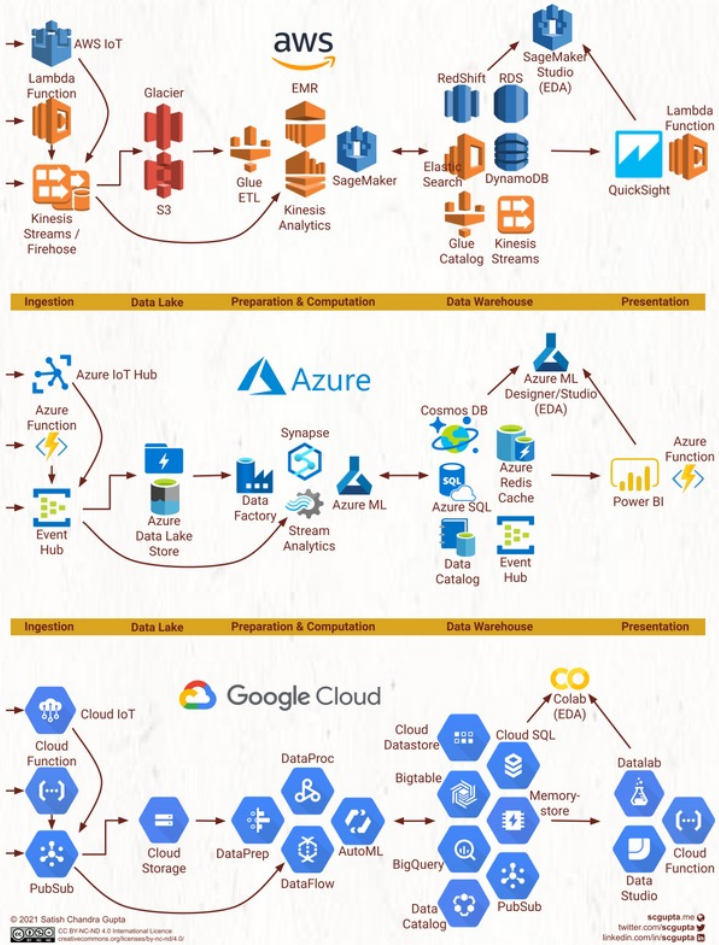
(However, Data Catalog and ELT can help with lineage, meta data and quality)



Data Lake in the Cloud

Big Data Pipelines on AWS, Microsoft Azure, and GCP

scgupta.link/big-data-pipeline



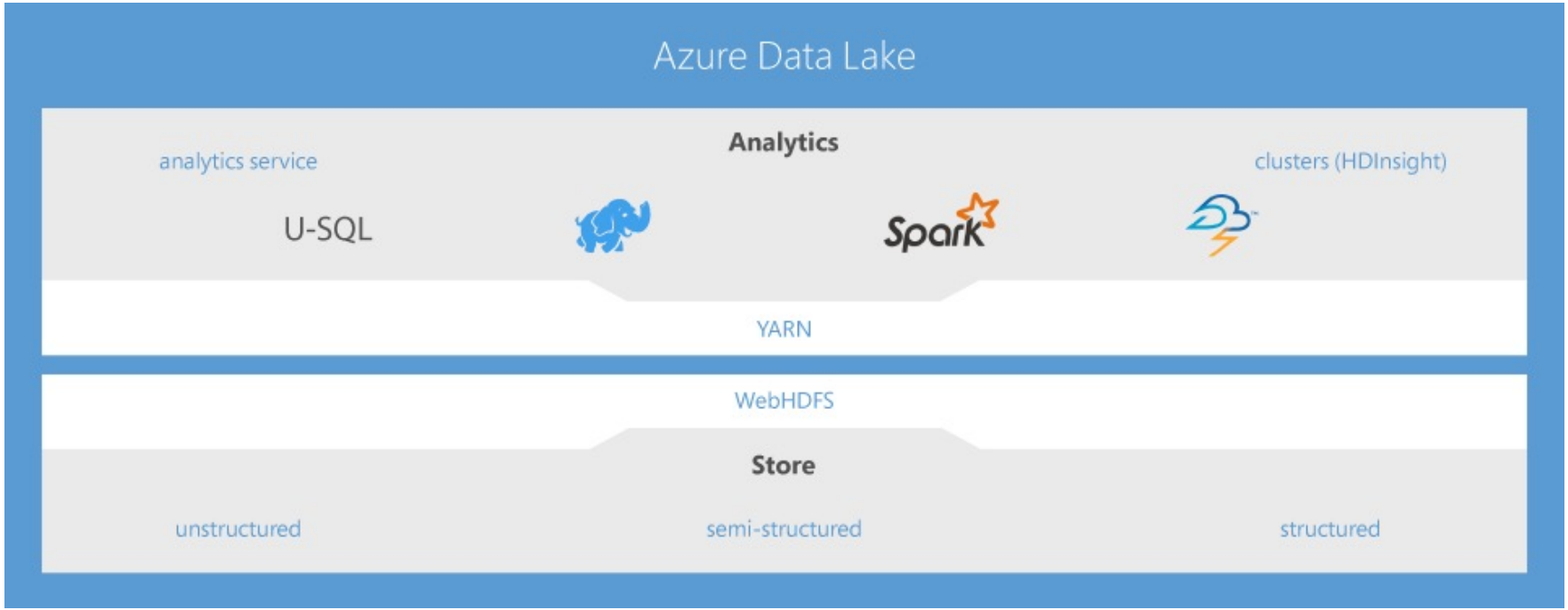
Amazon AWS Data Lake Offerings

EC2 – Elastic, On-Demand Cluster

S3 – Unlimited Storage

EMR – Scale-Out Computing

Microsoft Azure Data Lake



Source: Gorelik (2019)