Data Lake

DATASIENS

A central repository for all data types, structured, semi-structured and unstructured.

Technology

Cloud Storage

- cloud native technology such as Amazon S3, Azure Data Lake Storage (ADLS) and Google Cloud Storage (GCS)

HDFS

- Hadoop distributed file system

Object Storage

- On-premise equivalent of cloud storage from vendors like Dell, Hitachi, HP, IBM etc.
- Open source variants like Minio and CEPH

Principles

Schema on read

- data is not checked for structure or consistency during writes.
- the onus of verifying the data and its structure lies on the reader.

In-place analytics

- Instead of moving data from one database table to another, schema on read makes it possible to read the same data file in different ways, thus minimizing storage utilization

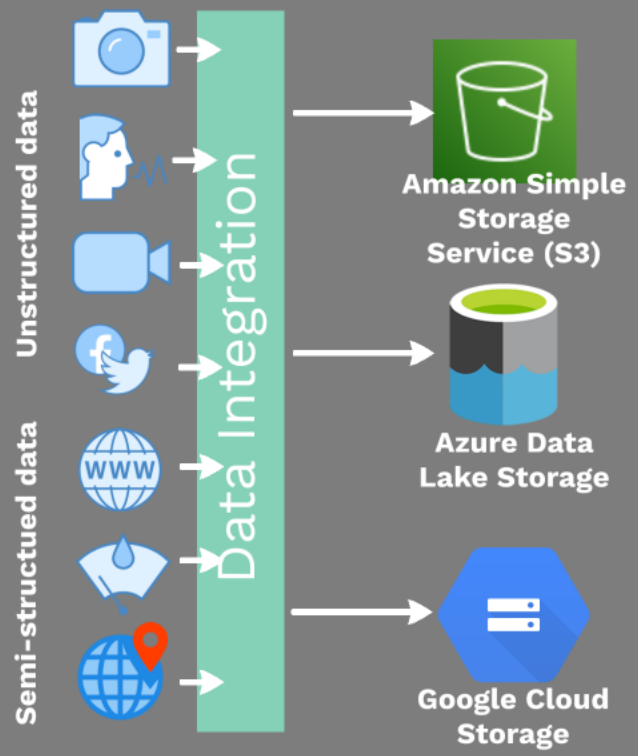
ELT versus ETL

- Data is Extracted, Loaded and then Transformed in a data lake.
- Data is Extracted, Transformed and then Loaded into a data warehouse. A data warehouse applies the principle of schema on write.

Cloud Storage

DATASIENS

Virtually unlimited scale of data storage hosted by a cloud provider and accessible via the network.



Concepts

Bucket or Container

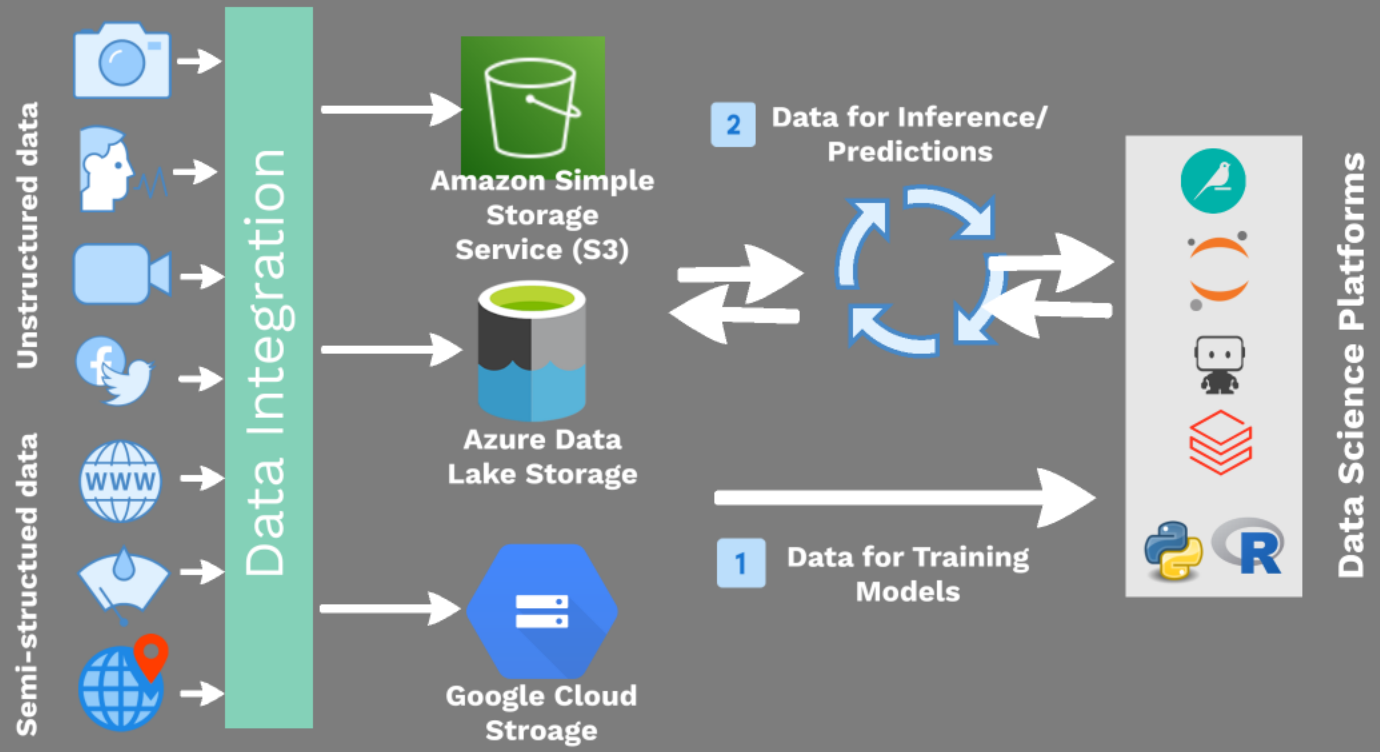
- As the name suggests they are basic containers which are used to organize data.
- Important thing to note is that buckets cannot be nested.
- Bucket names have to be unique globally.
- You cannot rename a bucket.

Blobs or objects

- The contents inside the bucket are called objects or blobs.
- Objects are immutable, if you try to overwriteone of them, a new object gets created and the old one gets deleted.
- Objects are just files of different types.
- Directories are also called as prefix, they are not real directories like on a Windows file system but just an emulation of the same concept.

Cloud Storage for Data Science

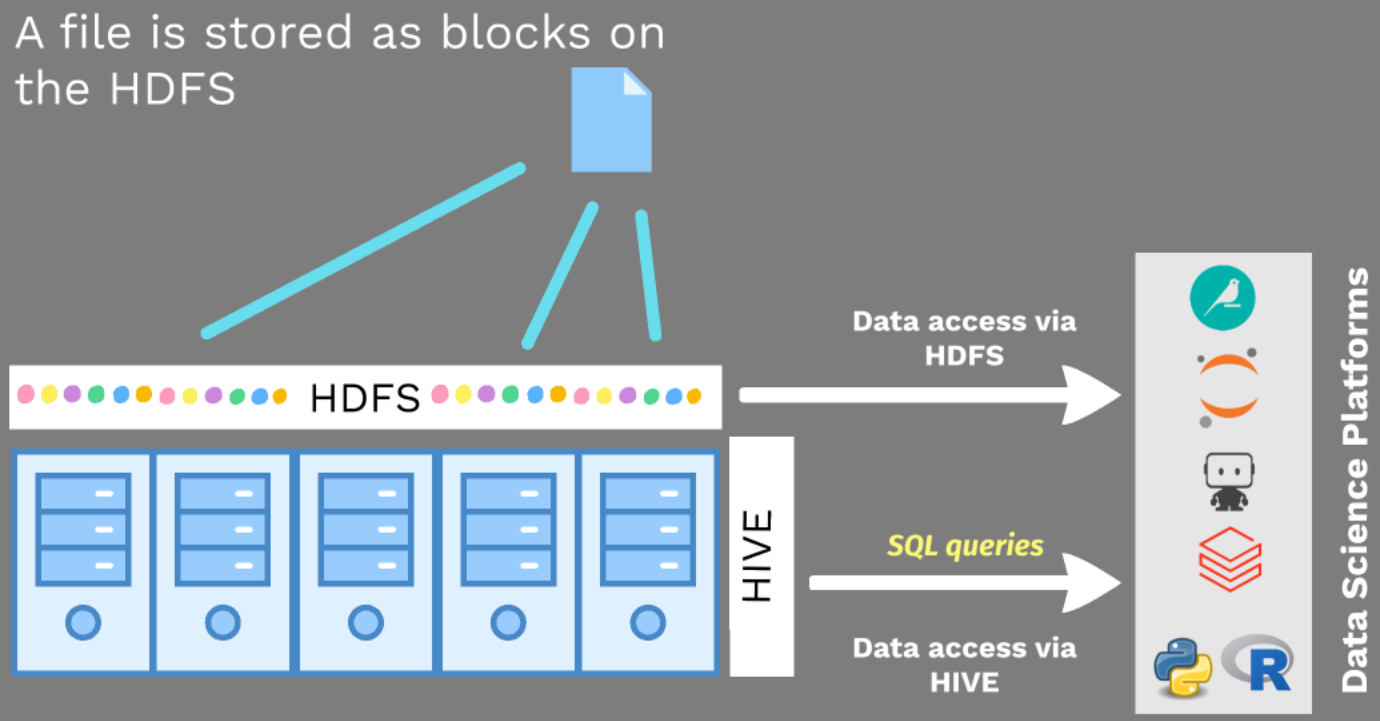
DATASIENS



HDFS

DATASIENS

Hadoop Distributed File System

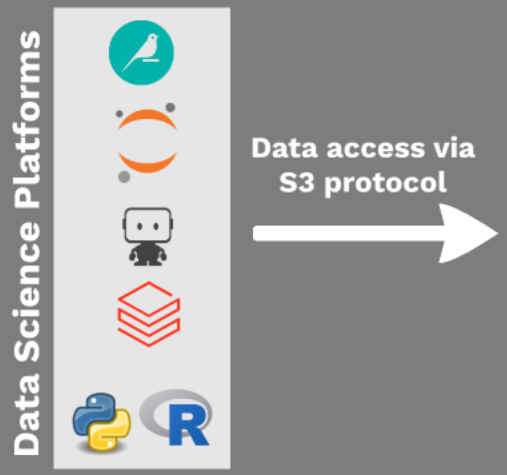


The blocks are distributed across several inexpensive computers/servers running Linux operating system, thus making it fault tolerant.

Object storage

DATASIENS

On-premise equivalent of the cloud storage.



HP Scality

Hitachi Content Platform (HCP)

Cohesity

Huawei Oceanstor

Netapp

Minio (open source)

CEPH (open source)