The Data Lake and OLAP

Rahul Dave(@rahuldave), Univ.Ai

Data Engineering Lifecycle

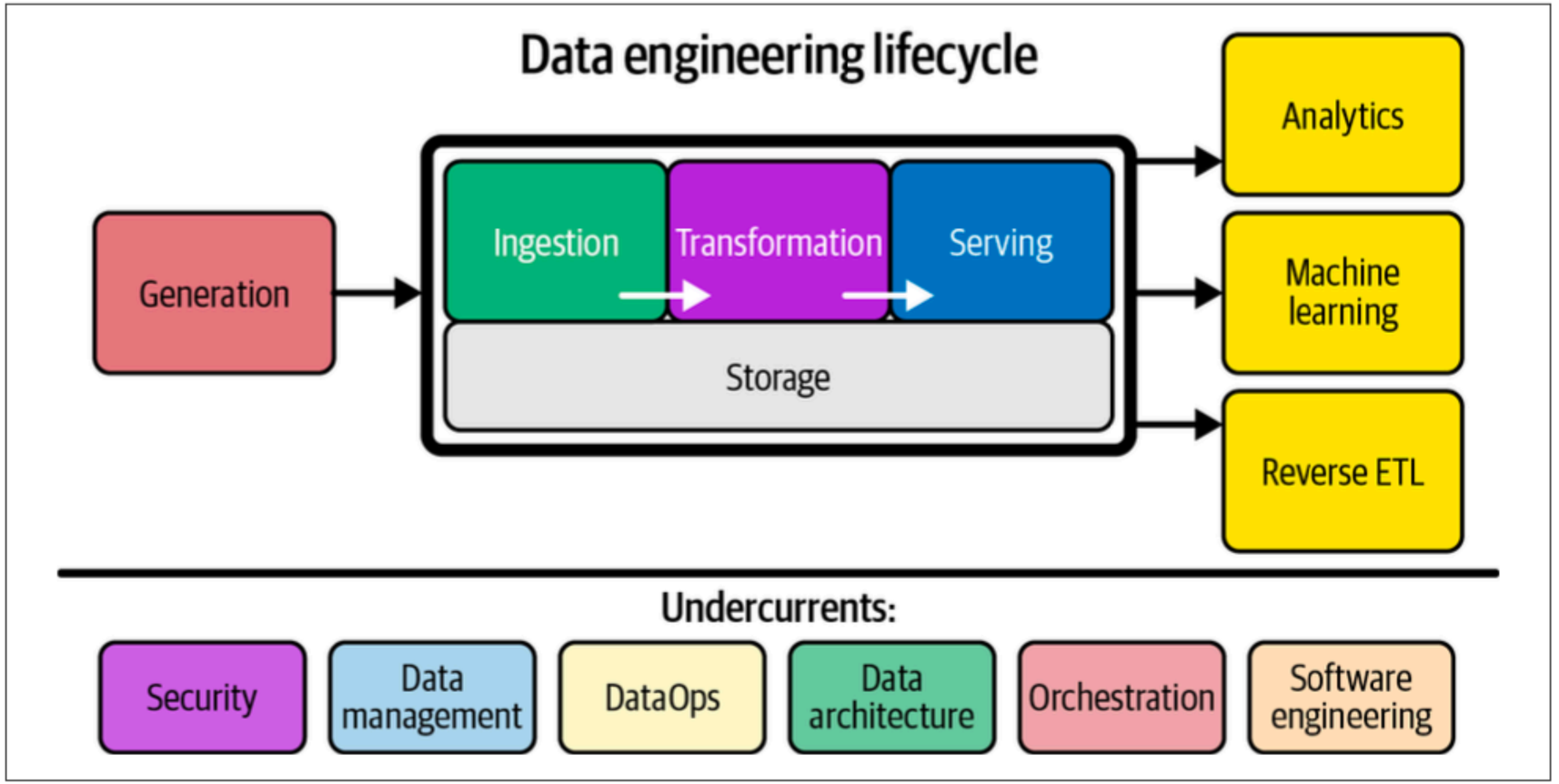
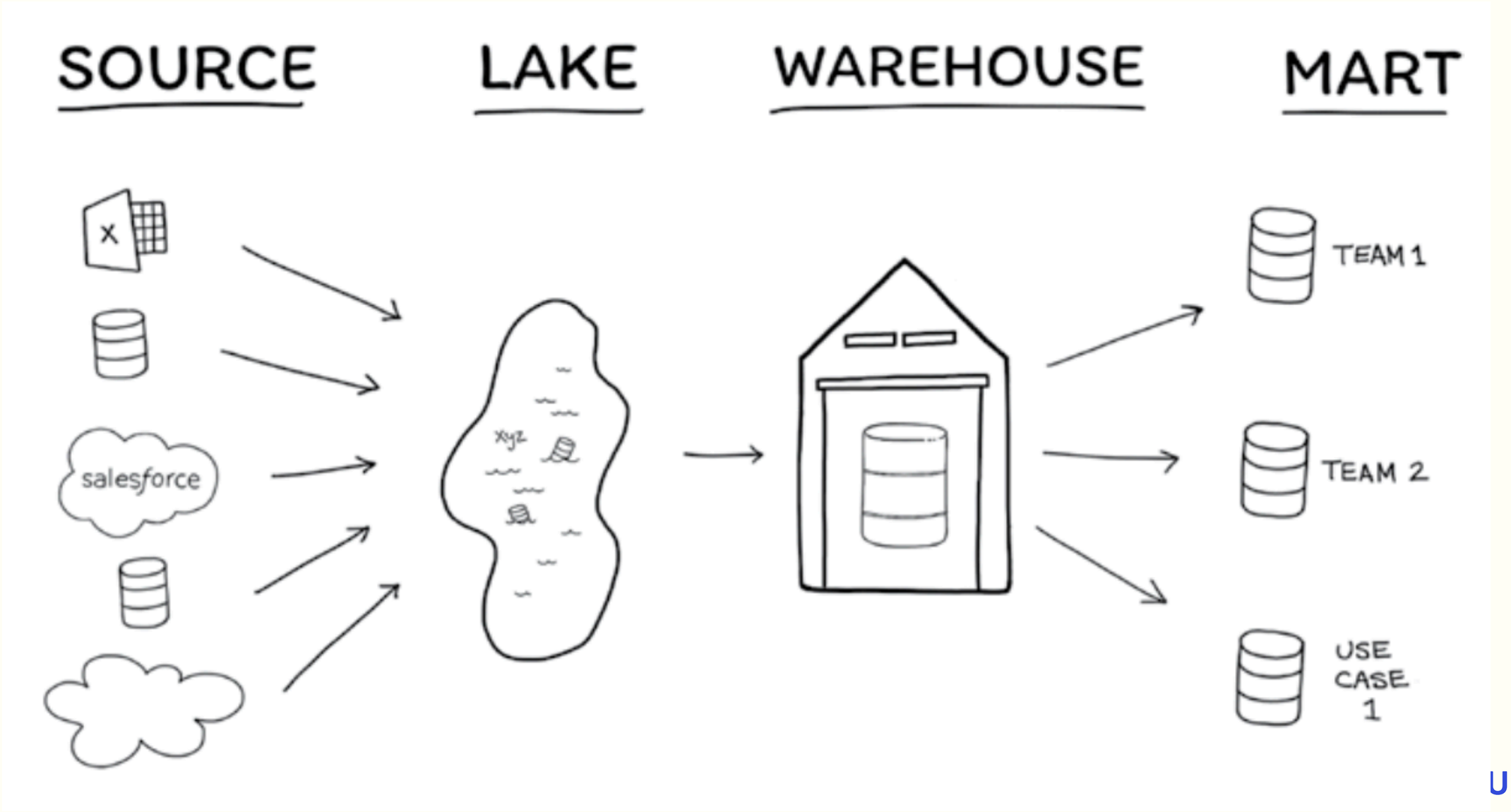


Figure 1-1. The data engineering lifecycle

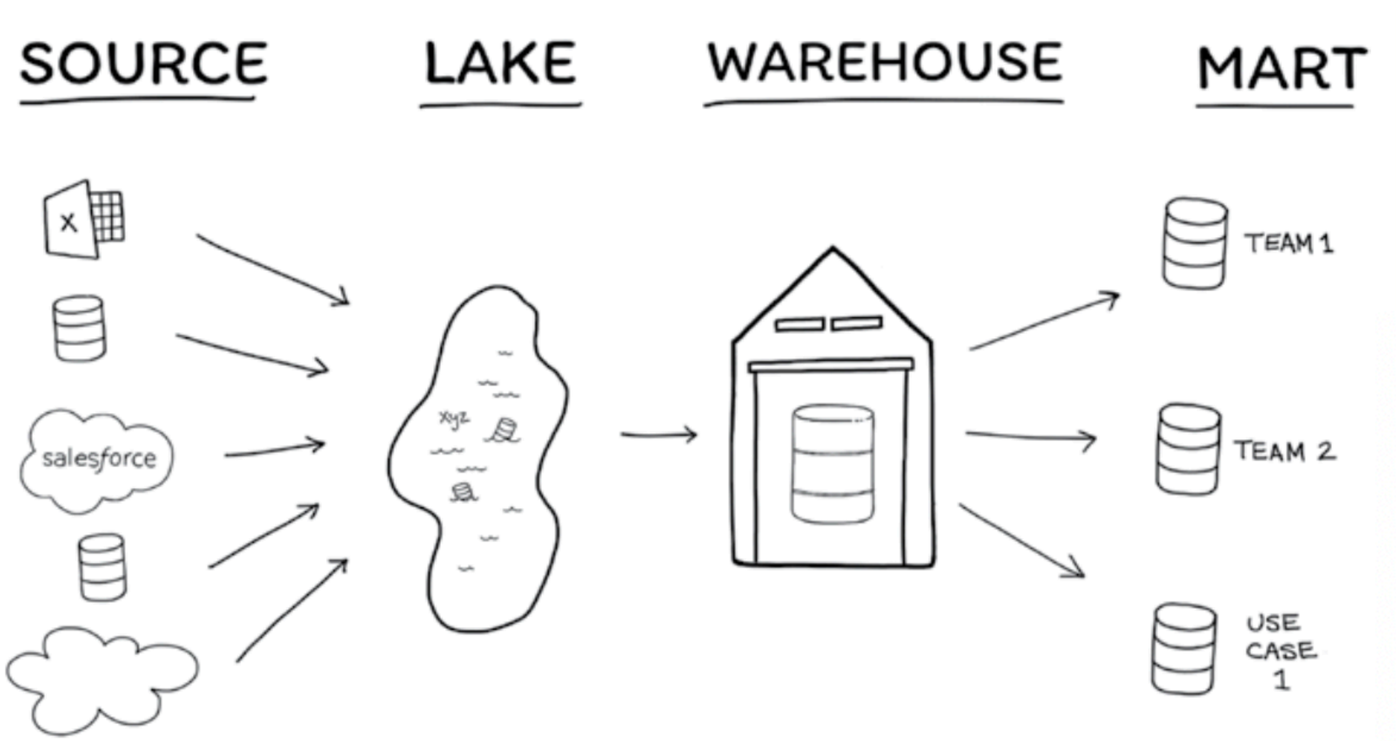
The Arrival of Hadoop

- In 2003, Google published the GFS paper, and in 2004 the Map-Reduce Paper
- Hadoop was contributed by Yahoo in 2006
- The big difference with Hadoop was that one brought the compute to the data and spread the storage out: this was the genesis of HDFS
- The disadvantage was that people now needed to maintain Hadoop/HDFS clusters and this was a lot of work in itself
- Until the cloud: now you can spin up on-demand HDFS clusters for a computation on AWS EMR with the original data on AWS S3.

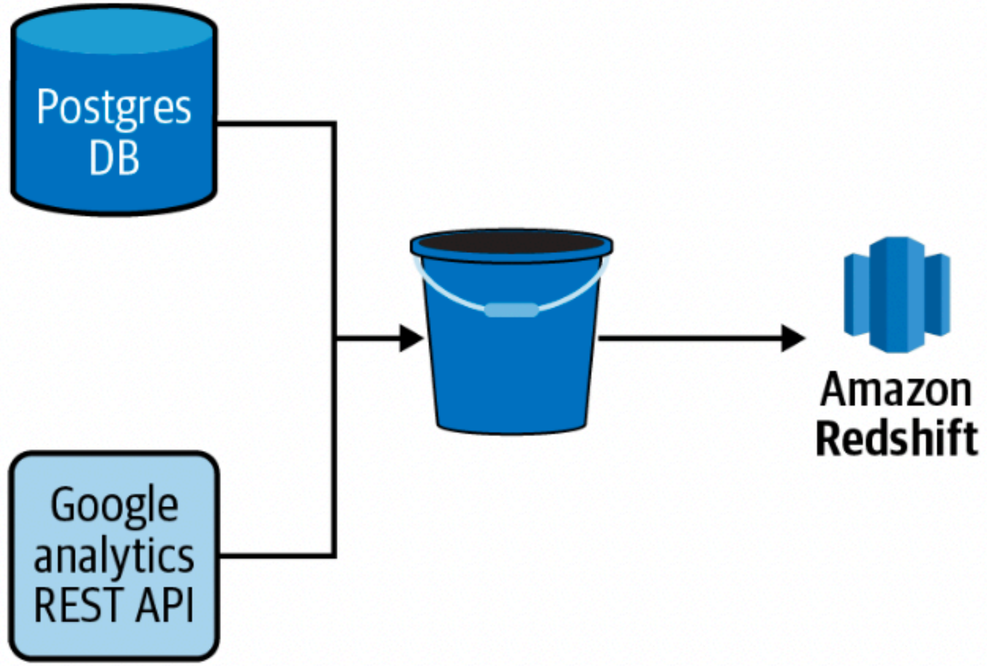
Introducing the Lake



The Lake is usually built on Object Storage.



A lake is often a useful intermediary. All kinds of data: tabular/image/text can be dumped there. It can also be used for post warehousing data outputs for downstream pursuits such as machine learning.



Analysis using non-SQL code such as using SPARK can also be carried out in the data lake.

SQL on the lake

- The trend was started by facebook with Hive, and then Presto. SparkSQL implemented SQL in the spark ecosystem
- Now you have many choices including duckdb, Presto, trino, sparksql, dremio. SQL capability is a MUST for any new engine.
- Meanwhile Redshift can now query parquet files in S3 and Snowflake is built on a micro-partioned architecture. Duckdb has its own database format but will query parquet and CSV, just like SPARK does.
- Thus there is a great convergence towards SQL everywhere

To SQL or Not to SQL(Spark).
Thats the question...

- With object storage and SQL-on-file as base layer it is not entirely clear what is in the database and what is not
- There are two kinds of ELT: into warehouse, and into lake and (possibly) then with some T into warehouse (a kind of ETL).
- Dont focus on ETL/ELT, focus on the needs of the data
- We should think about SQL workflows vs non-SQL workflows instead. What transformations would be hard to express with SQL? There you want to use Spark and PySpark.