

Warehouse, lake, or lakehouse?

D8A.ACADEMY/RESOURCES/WAREHOUSE-LAKE-LAKEHOUSE

Clean, structured data. BI dashboards every morning.



Warehouse

Snowflake, BigQuery, Redshift

Raw logs, events, images. Data scientists want everything.



Data lake

S3 or ADLS + Parquet

BI and ML on one platform, without copying data twice.



Lakehouse

Delta Lake, Iceberg, Hudi

A few dozen GB and a team of five.

Postgres

Seriously. It will hold for years.

WORTH MEMORIZING

01 The real differences

Schema timing is the whole story. Everything else follows from it.

	WAREHOUSE	LAKE	LAKEHOUSE
schema	on write	on read	on read + enforced
data	structured	anything	anything
queries	fast SQL	slow, varied	fast SQL
storage cost	high	low	low
main user	analysts	data scientists	both

02 Trap: the data swamp

A lake without ownership is a folder called `final_v2` the size of a company.

the slide lake with no catalog, no owners, no contracts

the symptom nobody trusts any number that comes out of it

the fix catalog + schema checks at the door, from day one

→ The swamp is not a storage failure. It is a governance failure that storage made cheap.