Data Lakehouse

→ Data Warehouse

Pros

- BI
- Analytics
- Structured & clean data
- Predefined Schemas

Cons

- No support for semi structured or unstructured data.
- Inflexible Schemas
- Struggled with volume and velocity upsticks
- Long processing time.

→ Data lake

Pros

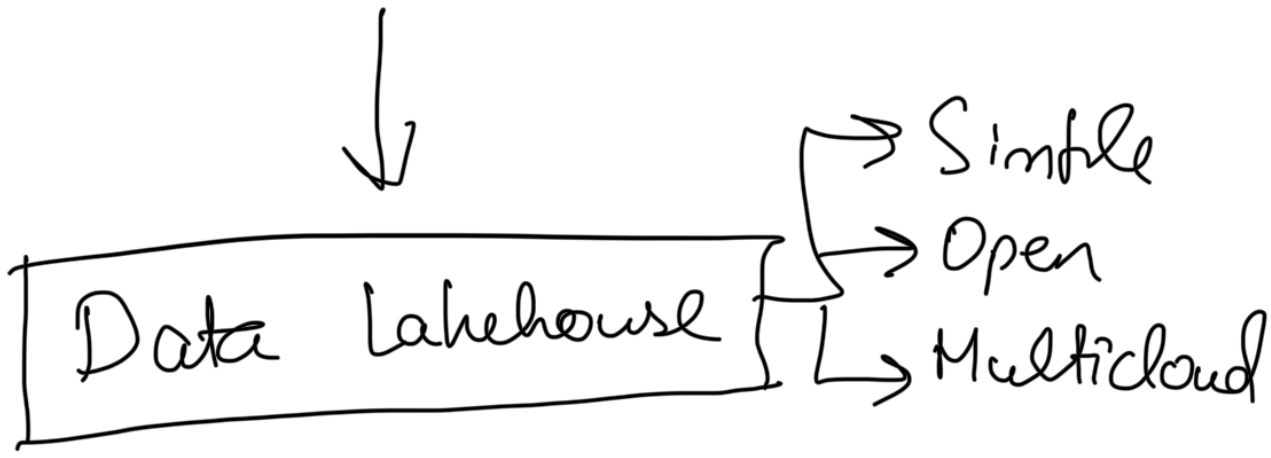
- Flexible data storage.
- Streaming support.
- Cost efficient in Cloud.
- Support AI & ML.

Cons

- No transactional support.
- Poor data reliability
- Slow analysis performance
- Data governance concerns.
- Data warehouse

still needed.

Data Lake + Data Warehouse



- All ML, SQL, BI and streaming use cases.
- One security and governance approach for all data assests and all cloud.
- An open and reliable data platform to efficiently handle all data types.

Key features:

- Transaction support.
- Schema enforcement and governance.
- Data governance.

- BI support.
- Decoupled storage from compute.
- Open storage formats.
- Support for diverse data types.
- Support for diverse workloads.
- End-to-end streaming.

Problem with Data lake

- Lack of ACID transaction support.
- Lack of schema enforcement.
- Lack of integration with a data catalog.
- Ineffective partitioning.
- Too many small files.

→ Delta lake

- ACID transaction guarantees.
- Scalable data and metadata handling.
- Audit history and time travel.
- Scheme enforcement and schema evolution.
- Support for deletes, updates and merges.
- Unified streaming and batch data processing.
- Compatible with Apache Spark.
- Uses delta tables.
- Has a transaction log.
- Open source.

→ Photon : Next gen query engine. It is compatible with spark APIs. It offers 2x the speed on the TPC

DS ITB benchmark.

→ Challenges to AI and Data governance

- Diversity of data and AI assets.
- Using two disparate and incompatible data platform.
- Rise of multi cloud adoption.
- Fragmented tool usage for data governance.

→ Unity Catalogue

→ Data Sharing with Delta Sharing

- Open cross platform sharing.
- Share live data without copying it.
- Centralized administration and governance.
- Marketplace for data products.

- Privacy – safe data clean rooms.

→ Divided Security Architecture

- Control plane
- Data plane

Compliance

- SOC 2 Type II
- ISO 27001
- ISO 27017
- ISO 27018
- FedRAMP High
- HITRUST
- HiPAA
- PCI

Also it is GDPR and CCPA ready.

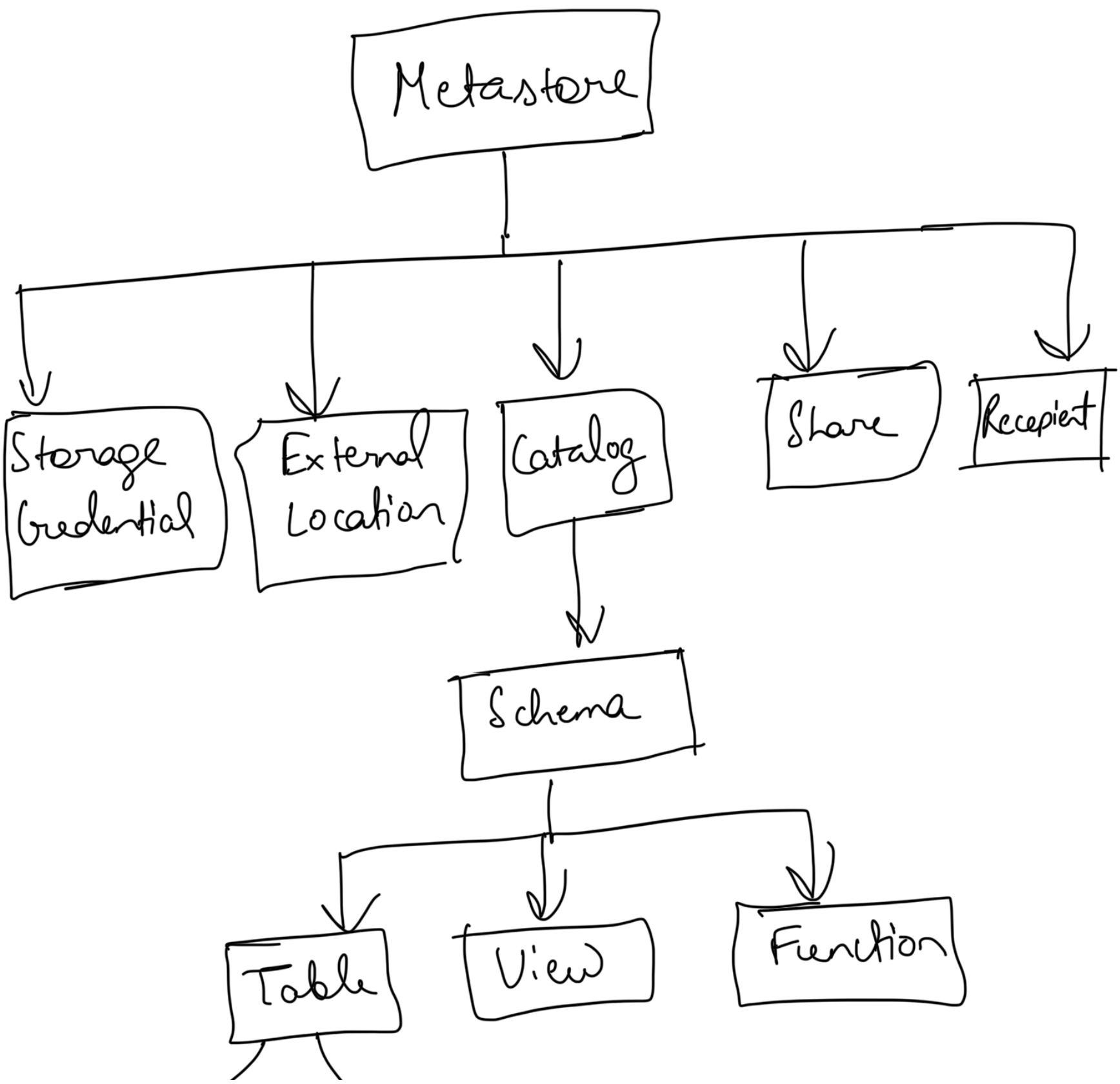
Compute resource challenges

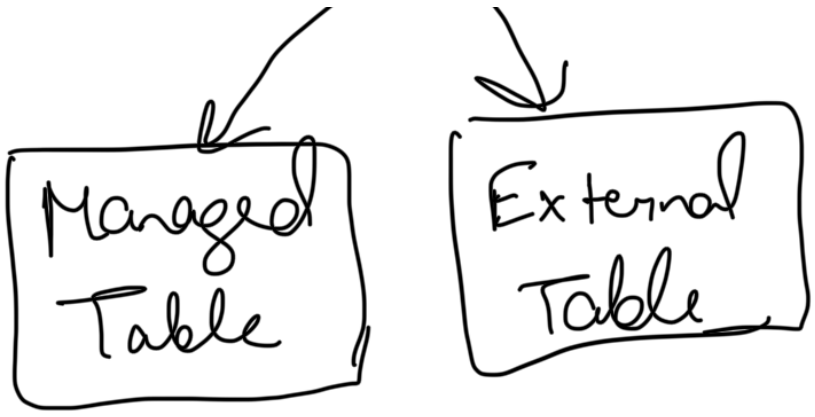
- Cluster creation is complicated.
- Environment setup is slow.
- Business cloud account limitations and resource options.

- Long running clusters.
- Over provisioning of resources.
- Higher resource cost.
- High admin overhead.

- Unity catalogue uses three level namespace.

Eg: Select * From catalog.schema.table





Key benefit of Data Warehousing with Databricks Lakehouse Platform

- Best price & performance.
- Built-in governance.
- A rich ecosystem.
- Break down silos.

Data is valuable business asset.

Challenge for data engineering

- Complex data ingestion methods.
- Support for data engineering principles
- Third party orchestration tool.
- Pipeline and architecture performance tuning.
- Inconsistencies between data warehouse and data lake provides.

Database Workflows

It is first fully managed orchestration service embedded in Databricks lakehouse platform.