Database Administration

José Orlando Pereira

Departamento de Informática
Universidade do Minho

What next?

- Data storage and processing is changing:
  - Scale and new architectures (cloud)
  - New tasks and purposes (AI)

- Warning: The next slides include only pointers to key concepts! Check references for details.

Roadmap

- Analytical processing
- Transaction processing

Hadoop Distributed File System

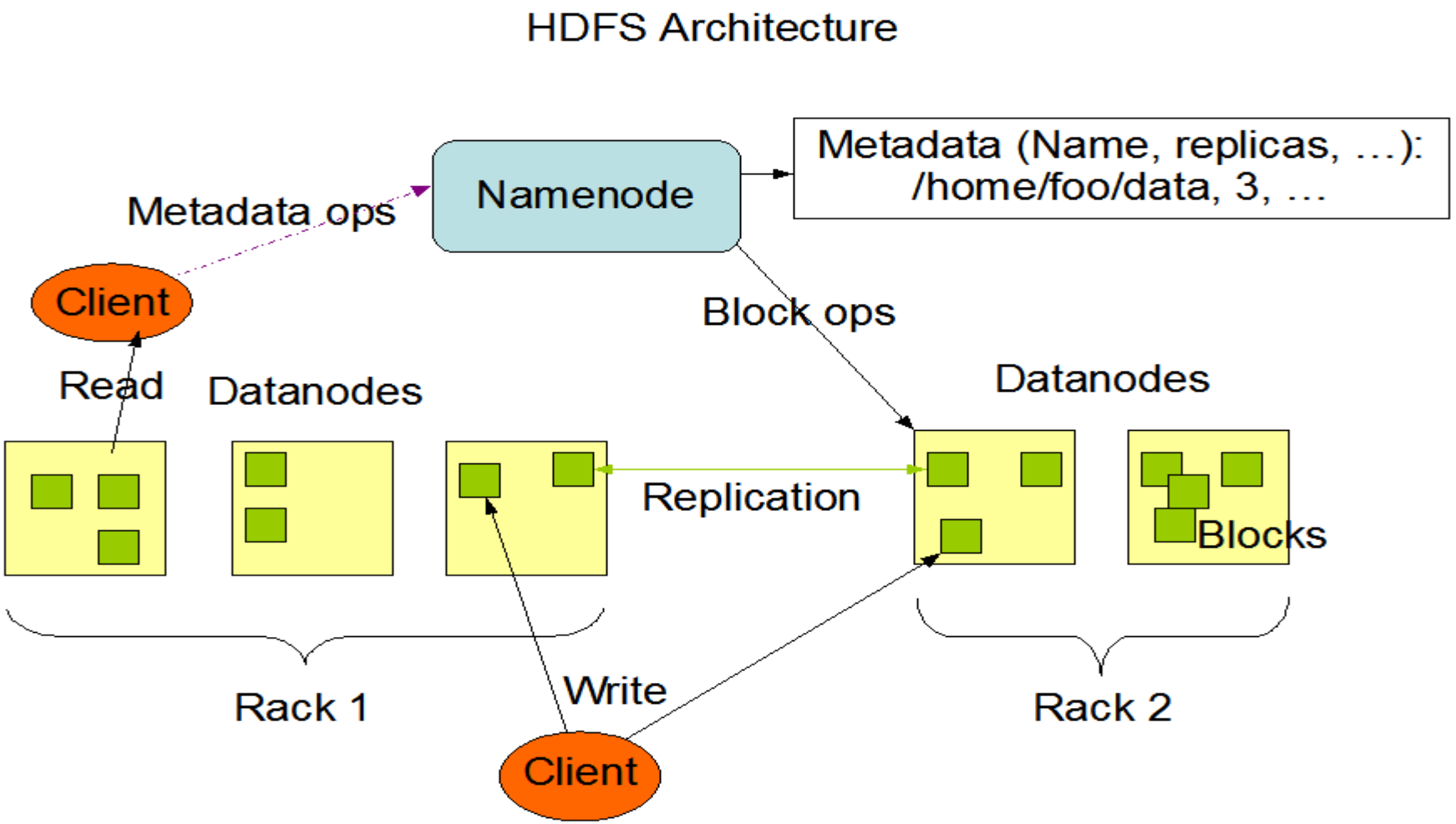


Image from: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

Amazon Dynamo

- Planet-scale key-value data store
- Showcase of distributed systems techniques!

Table 1: Summary of techniques used in *Dynamo* and their advantages.

Problem	Technique	Advantage
Partitioning	Consistent Hashing	Incremental Scalability
High Availability for writes	Vector clocks with reconciliation during reads	Version size is decoupled from update rates.
Handling temporary failures	Sloppy Quorum and hinted handoff	Provides high availability and durability guarantee when some of the replicas are not available.
Recovering from permanent failures	Anti-entropy using Merkle trees	Synchronizes divergent replicas in the background.
Membership and failure detection	Gossip-based membership protocol and failure detection.	Preserves symmetry and avoids having a centralized registry for storing membership and node liveness information.

Source: G. DeCandia et al., "Dynamo: Amazon's highly available key-value store," Oper. Syst. Rev., vol. 41, no. 6, pp. 205–220, Oct. 2007
Available: https://doi.org/10.1145/1323293.1294281

Cloud object storage

Capacity and throughput	Amazon S3 holds more than **280 trillion objects** and averages over **100 million requests per second**
Events	Every day, Amazon S3 sends over **125 billion event notifications** to serverless applications
Replication	Customers use Amazon S3 Replication to **move more than 100 PB** of data per week
Cold Storage Retrieval	Every day, customers **restore more than 1PB** from the S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive storage classes
Data Integrity Checks	Amazon S3 performs over **4 billion checksum computations per second**
Cost Optimization	On average, customers using Amazon S3 Storage Lens advanced metrics and recommendations have obtained **cost savings 6x greater** than the Storage Lens cost in the first six months of using it.
Flexibility	**Hundreds of thousands of data lakes** are built on Amazon S3

Source: "Building and operating a pretty big storage system called S3," All Things Distributed, July 27, 2023.
https://www.allthingsdistributed.com/2023/07/building-and-operating-a-pretty-big-storage-system.html

Datalake

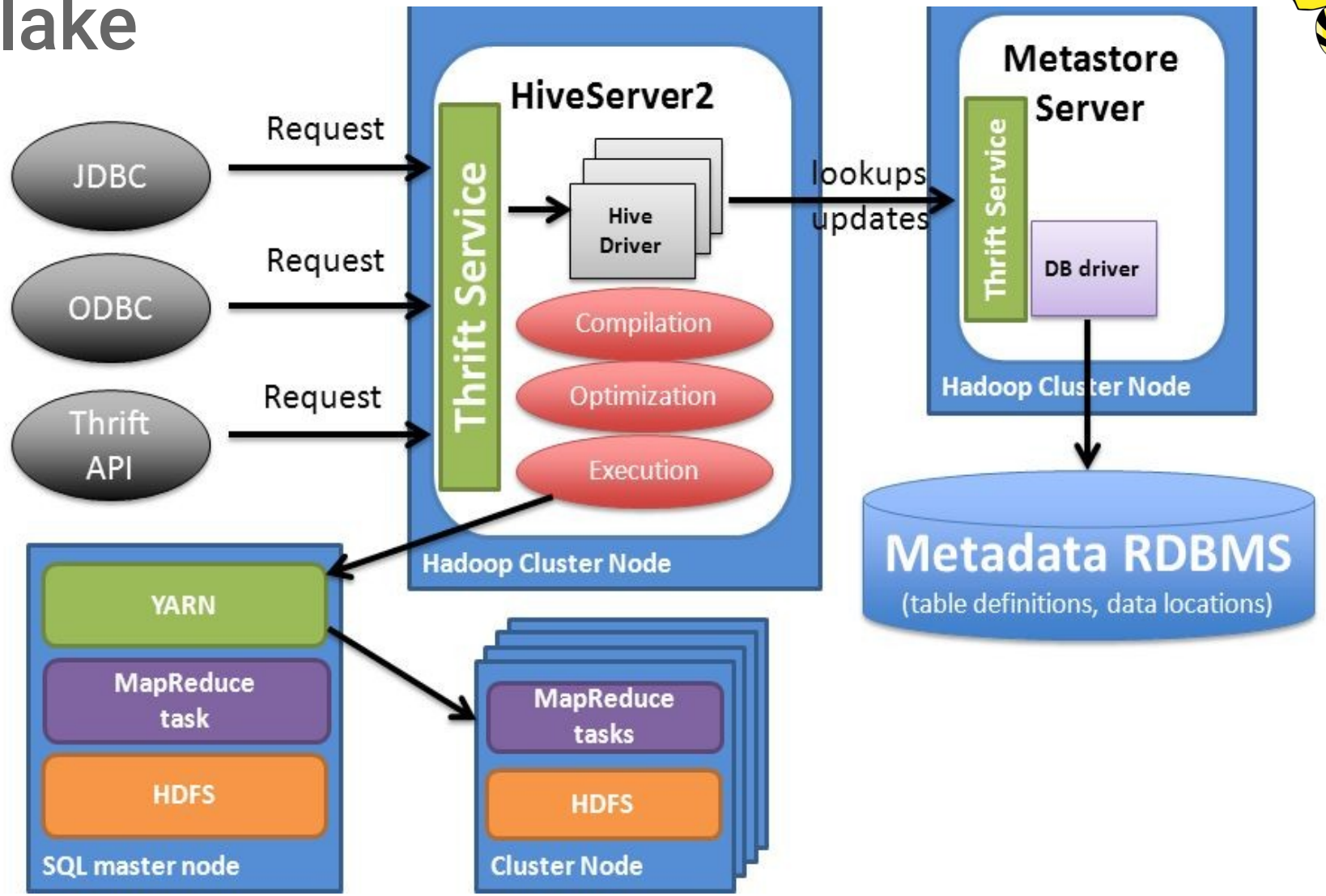


Image from: https://andr83.io/1123/

Datalake

- Back to SQL!
- Using many file formats:
  - CSV, JSON, Parquet, ...
- Many open source options for processing layers:
  - Spark
  - DataFusion
  - ...
  - commercial cloud services...

DataOps and DAGs

SQLMesh

- Complex tasks, composed by resource-intensive processing steps
  - Automatically recompute when data is updated
  - Explore alternatives efficiently and without endangering current production data

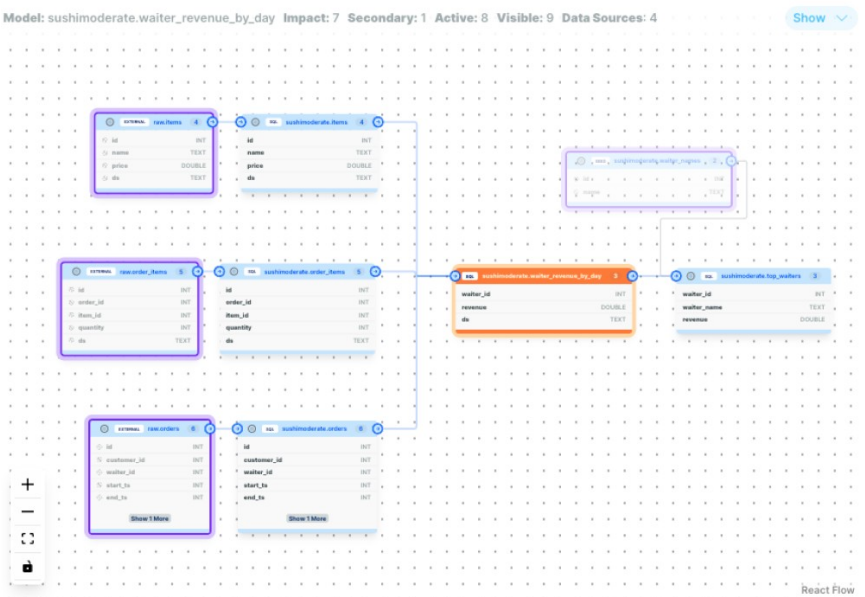


Image from: https://www.tobikodata.com/blog/sqlmesh-ui

DataOps and DAGs

- Complex workflows with:
  - Data processing tasks over multiple repositores / tools
  - Python processing steps
  - Arbitrary tools

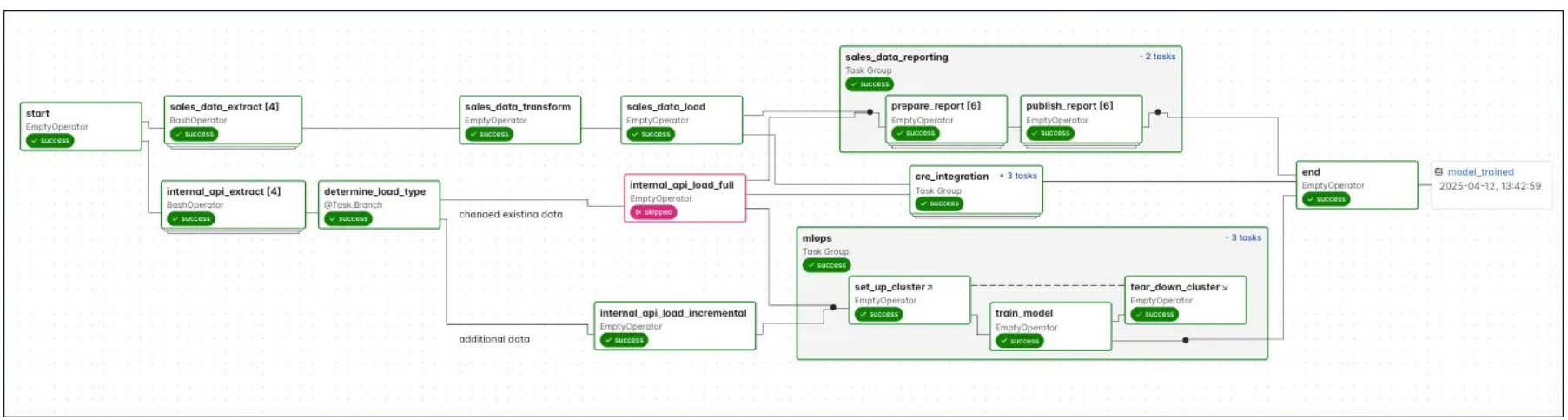


Image from: https://www.astronomer.io/docs/learn/dags

Roadmap

- Analytical processing
- Transaction processing

Lakehouse

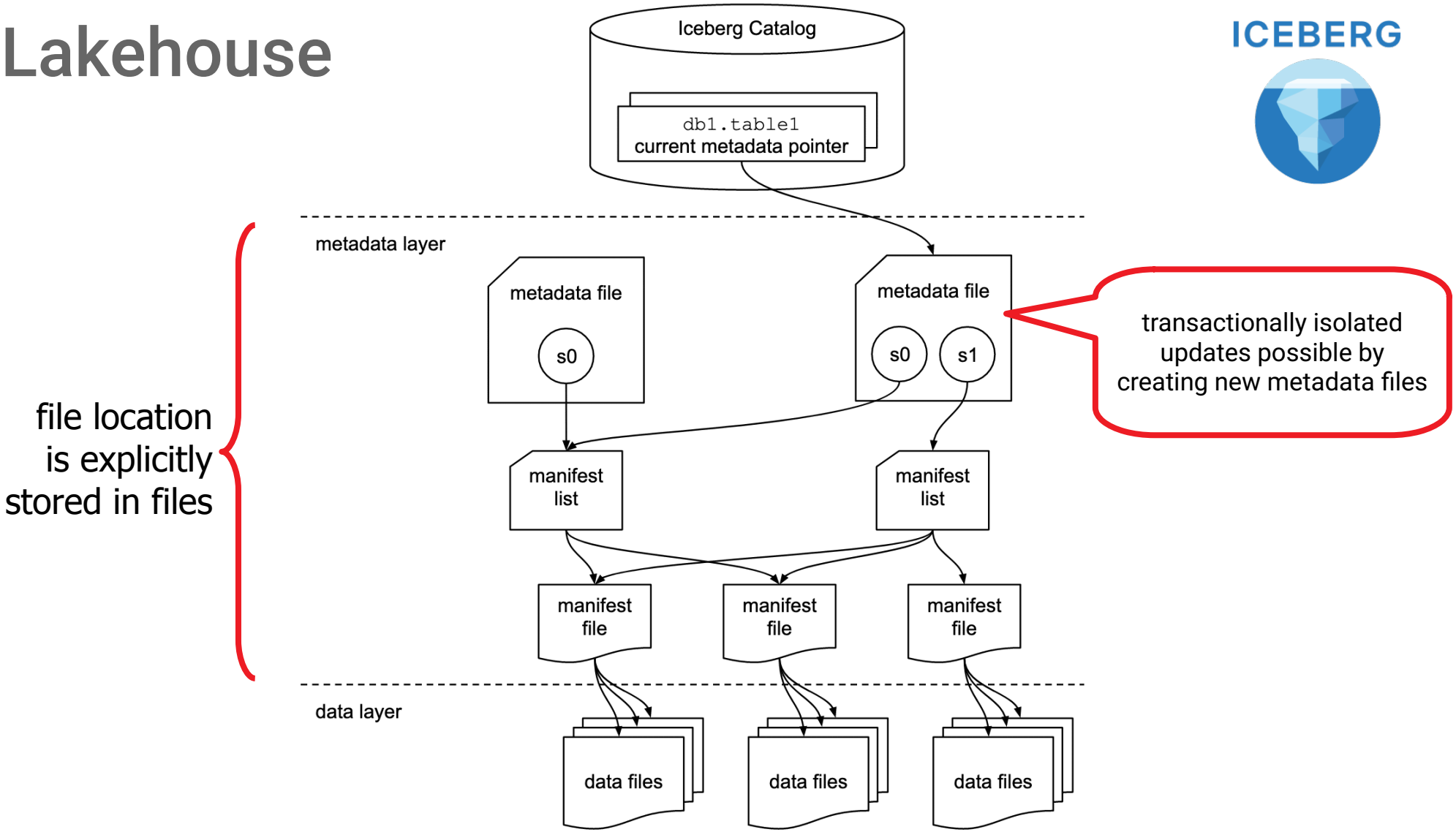


Image from: https://www.dremio.com/resources/guides/apache-iceberg-an-architectural-look-under-the-covers/

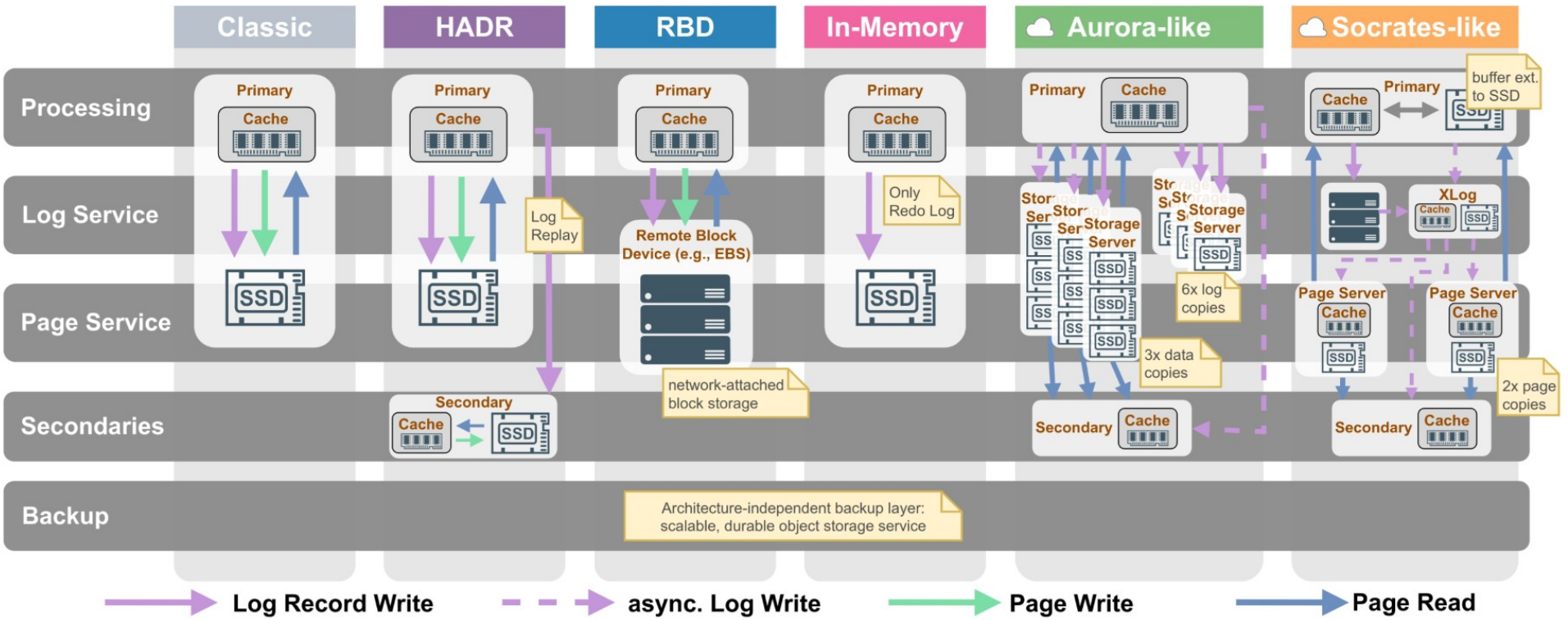
Iceberg

- Metadata pointer identifies the file that contains the most current information
  - Might itself be stored in a file
- A “metadata file” contains a list of snapshots, that described the file at different times
- A “manifest list” contains a list of fragments that exist at the same time
- A “manifest file” decribes an actual physical file
  - Contains statistics
- All information, except the pointer is immutable
  - No inconsistency
  - Can be stored in cloud object stores (e.g. S3)

Transactions on Iceberg

- New file fragments of new versions of existing fragments can be created by I/U/D
  - Corresponding Manifest, Manifest list, and Metadata files are created describing the new version
  - The new version is registered as current, but keeps history of previous snapshots
- Enables data modification
  - “Lakehouse”

Page-based cloud native



Source M. Haubenschild and V. Leis, "OLTP in the cloud: architectures, tradeoffs, and cost," VLDB J., May 2025. Available: https://www.cs.cit.tum.de/fileadmin/w00cfj/dis/papers/CloudOLTP.pdf

Log-structured merge-tree (LSM-tree)

- Write to the C_0 (in-memory) tree
- Read by traversing trees C_0, C_1, …
  - Use indexes (Bloom filters) to skip levels

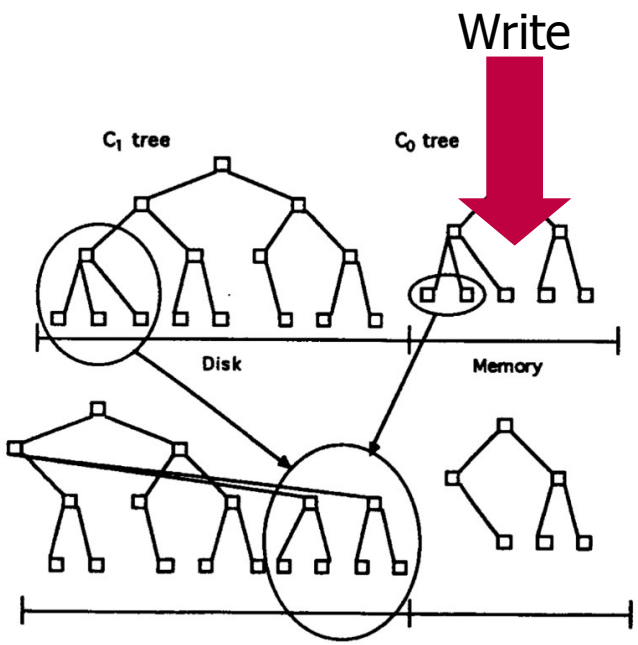


Fig. 2. Conceptual picture of rolling merge steps, with result written back to disk

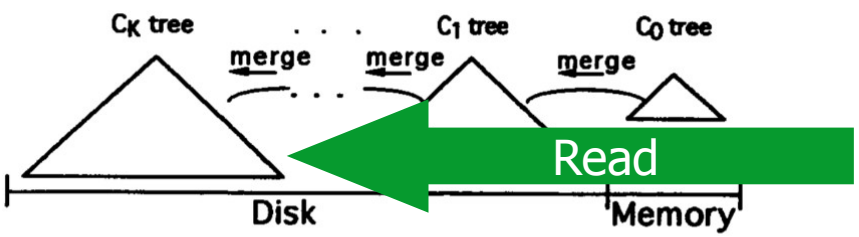
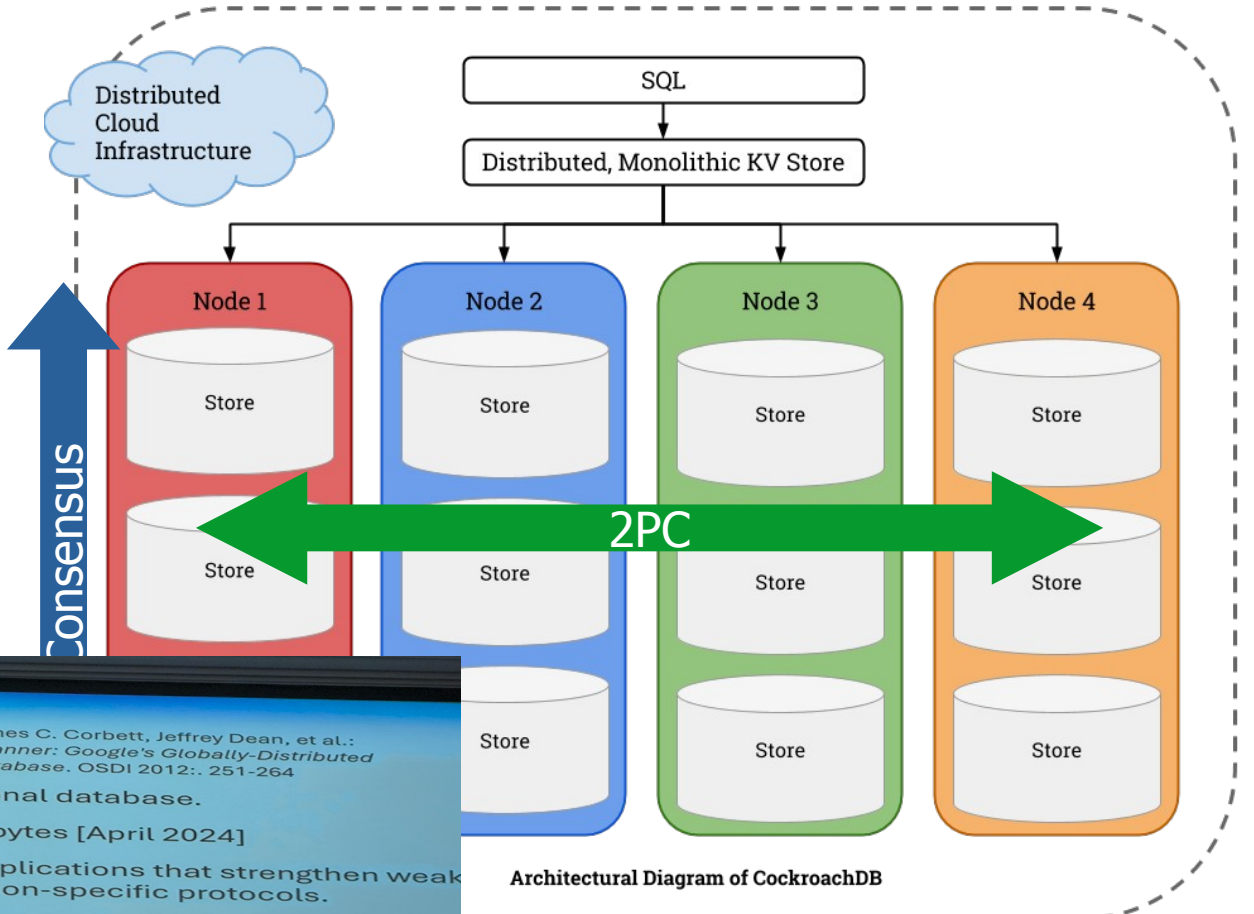


Fig. 4. An LSM-tree of $K + 1$ components

Source: P. O'Neil, E. Cheng, D. Gawlick, and E. O'Neil, "The log-structured merge-tree (LSM-tree)," Acta Inform., vol. 33, no. 4, pp. 351–385, June 1996, Available: https://doi.org/10.1007/s002360050048

Layered architecture

- SQL processing on top of a distributed transactional key-value store
  - Spanner
  - CockRoachDB
  - …



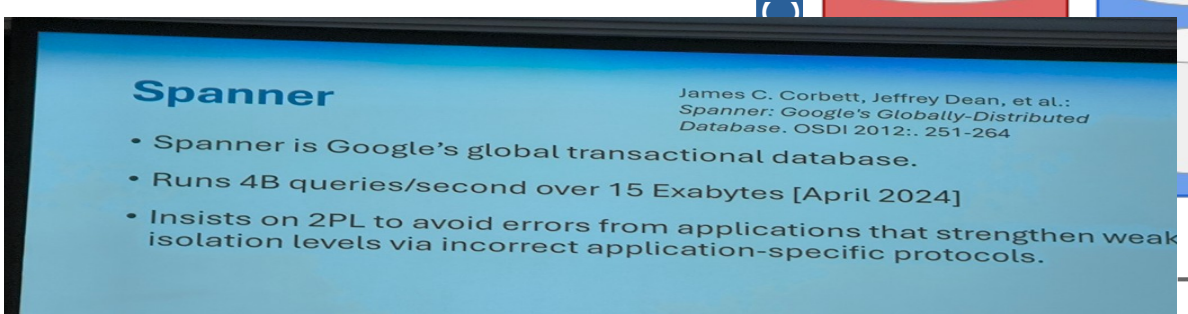


Image from: https://www.cockroachlabs.com/blog/the-new-stack-meet-cockroachdb-the-resilient-sql-database/