What Can Iceberg Do for You?

dremio Capitalize ANALYTICS

Agenda

- What is Iceberg?
- Why Iceberg?
- Iceberg for Analytics
- The Art of the Possible – What Will Iceberg Unlock?
- About Capitalize
- Q&A

Capitalize ANALYTICS

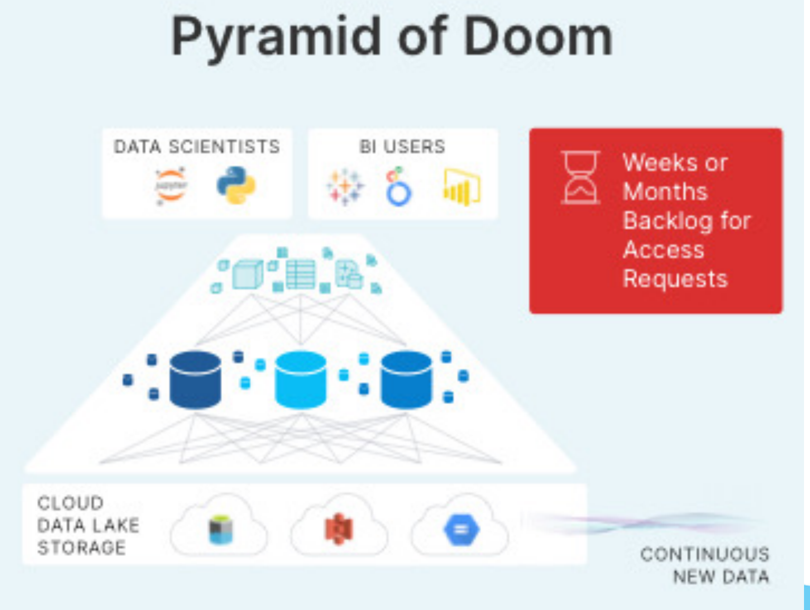
How did we get here?

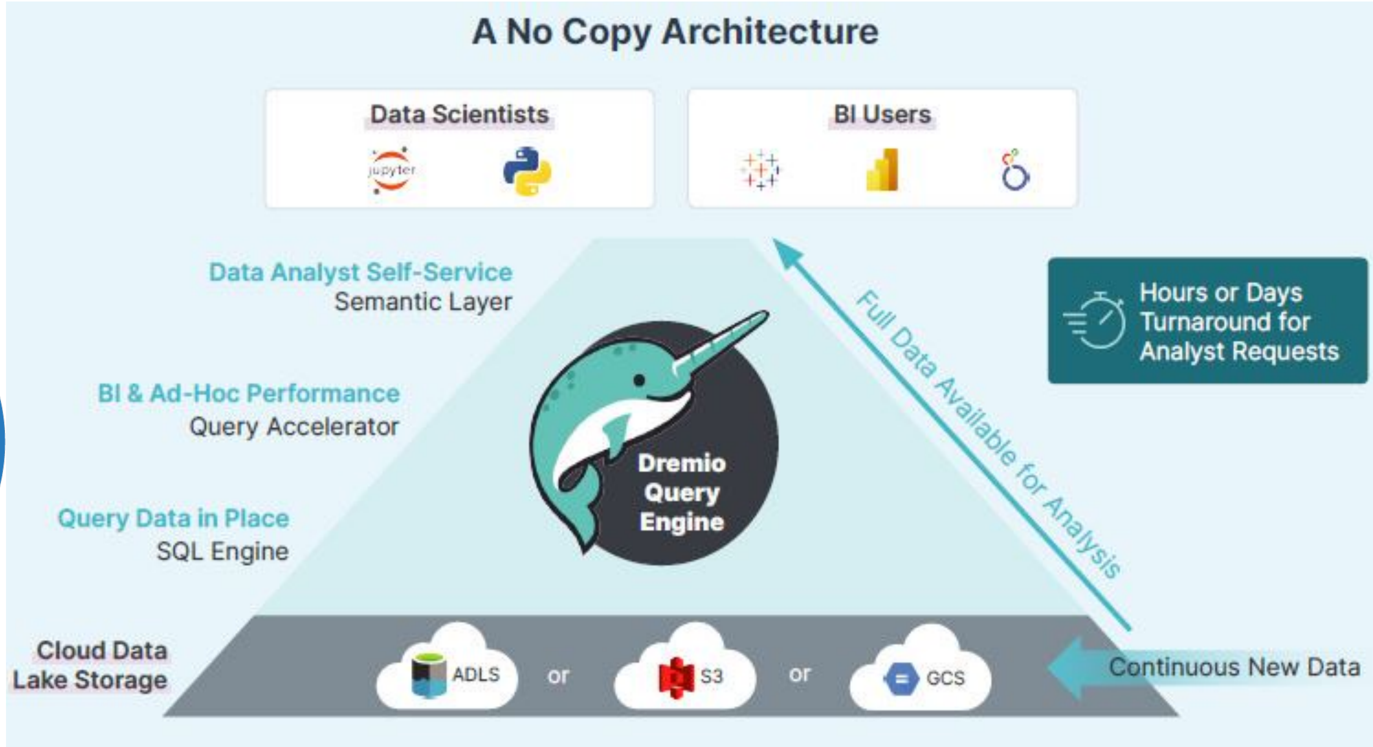
- Shift from Enterprise Data Warehouse to Data Lake
  - Cloud infrastructure and data storage cost declining
  - Less implementation and maintenance
  - *New Standards*: Hive, Spark, Presto, Trino

- ***Problem***: Data volumes increased exponentially over time
  - Folder level architecture lacks scalability
  - Band aid solutions cripple performance and increase overhead

Iceberg to save the day!

- ***Solution***: Data Lakehouse
  - Improves the “new” standards
  - Best practices from EDW with Data Lake efficiency
  - Focus on performance and scalability





Where did Iceberg come from?

- Originally developed at Netflix
  - Outgrew existing standard (Hive)
  - Built from the ground up to handle size and scale in modern cloud infrastructure
- The “key” was shifting focusing from folder level to file level
- Open-sourced in 2018, Productionalized in 2020

What is Iceberg?

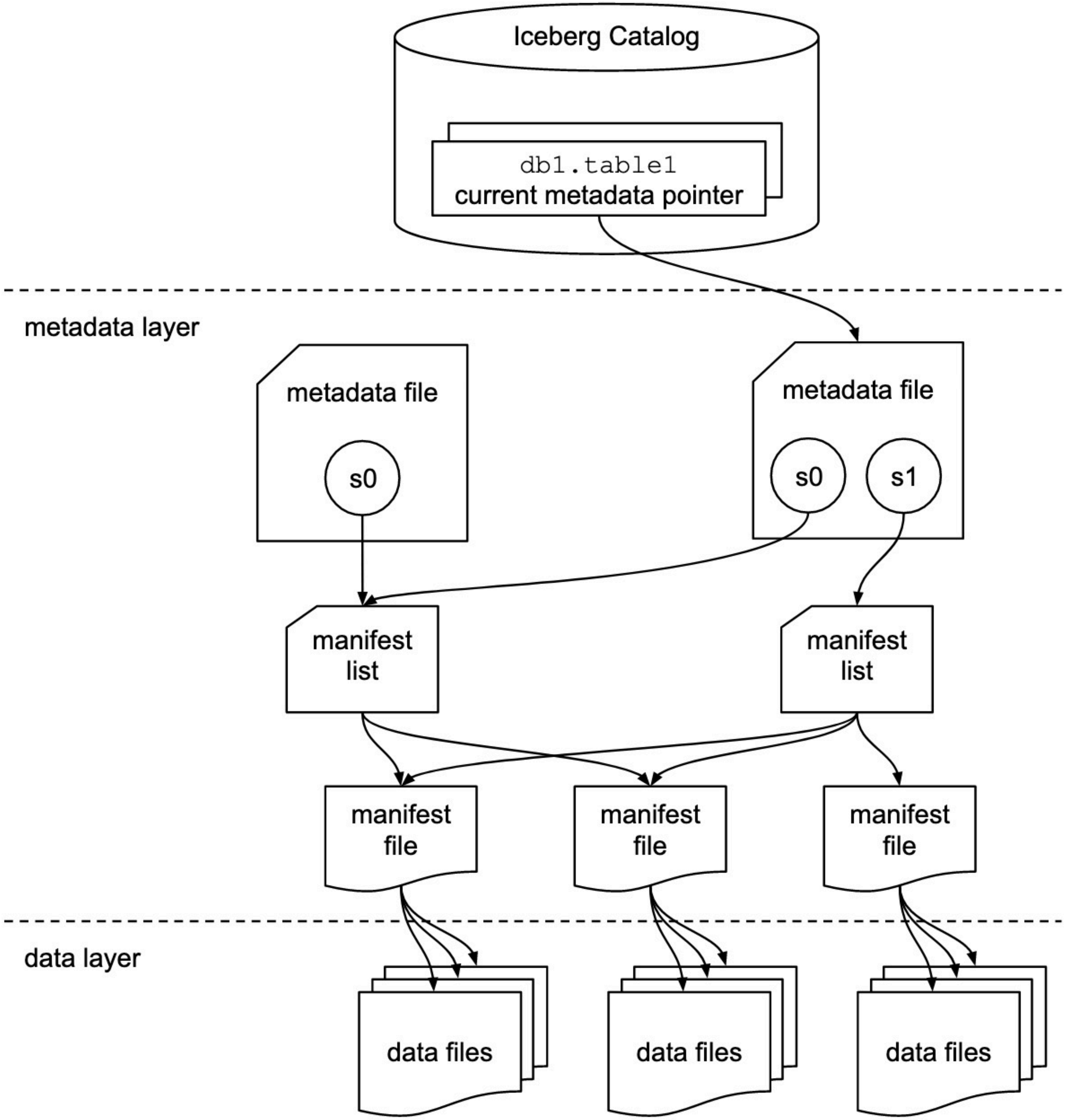
- Open table format designed for very large tables
- Provides standardization to managing files within a table
- Operates as an abstraction layer between physical files and table structure

Easy as 1-2-3

1. The Iceberg catalog

2. The metadata layer

3. The data layer



Three Layers of Architecture

1. **The Iceberg catalog**
   - Provides the location of the current metadata pointer
   - Pinpoints where to read or write data for a given table

2. **The metadata layer**
   a. *Metadata file*: schema, partitions, snapshots
   b. *Manifest list*: list of files and mappings to snapshots
   c. *Manifest files*: tracks data files and statistics

3. **The data layer**
   - Partition membership, record count, lower- & upper-bounds of columns.

Why Iceberg?

- Schema evolution
  - add, drop, update, or rename column commands with no side effects or inconsistency.
- Partition evolution
  - Facilitates the modification of partition layouts without needing to rewrite the entire table.
- Version rollback
  - Corrects problems quickly by resetting tables to a known good state.
- Transactional consistency
  - Avoids partial or uncommitted changes by tracking atomic transactions with ACID properties.

Why Iceberg? (cont.)

- Optimized processes
  - Utilizes advanced filtering to limit user mistakes causing slow queries.

- Time travel
  - Provides previous versions of the table for comparison and reproduction of queries

- Increased performance
  - Files are intelligently filtered via advanced partition pruning and column-level statistics.

Iceberg for Analytics

- Correct and consistent view of a table
- Faster query execution
- Better and safer table evolution
- Scalability at data, user, and applications levels
- End User Ease of Use
- Data Democratization

What Will Iceberg Unlock?

- More data, less problems
- Highly performative and interactive Dashboards
- Increased throughput for data pipelines
- Larger, more complete data sets for modeling
- Real-Real Time & Streaming capabilities

About Capitalize

✓ Founded in 2005 (Analytics practice in 2012)

✓ Professionals Across the US (18 states)

✓ Over 750 clients in North America

✓ Functional Area and Vertical expertise

✓ Fully integrated Analytics Services Firm

Vendors:	**Verticals**:	**Functions:**
Alteryx	Energy (O&G)	Accounting
Automation Anywhere	Financial Services	Audit
DataRobot	Health Care	Engineering
Dremio	Hospitality	Finance (FP&A)
IBM (Cognos)	Insurance	Human Resources
Power BI	K12/Higher Education	Logistics
SnapLogic	Logistics	Marketing
Snowflake	Distribution	Quality/Maintenance
Tableau	Manufacturing	Sales
UiPath	Public Sector	SalesOps
Vertica	Retail	Supply Chain
Workday Adaptive	Services	Tax

Capitalize
ANALYTICS

We Help Firms…

- Focus on overcoming data challenges
- Achieve analytics at Google-like speed
- Accelerate reporting and statutory filings
- Gain better, faster insight to data
- Be a data driven organization
- Achieve efficiencies via automation to enable all the above!

Capitalize
ANALYTICS

Technology Partners

alteryx

IBM COGNOS

UiPath™

snowflake

snapLogic®

workday + Adaptive Insights

Power BI

dremio

VERTICA

ABBYY®

tableau®

ThoughtSpot

DataRobot

collibra™

AUTOMATION ANYWHERE

Capitalize ANALYTICS

Wrap Up

- Q&A

Capitalize
ANALYTICS

References

- https://iceberg.apache.org/
- https://medium.com/expedia-group-tech/a-short-introduction-to-apache-iceberg-d34f628b6799
- https://www.dremio.com/resources/guides/apache-iceberg-an-architectural-look-under-the-covers/
- http://docs.dremio.com/data-formats/apache-iceberg/#:~:text=Additionally%2C%20Iceberg%20intelligently%20organizes%20snapshot,working%20at%20data%20lake%20scale
- https://thenewstack.io/apache-iceberg-a-different-table-design-for-big-data/

Capitalize
ANALYTICS