



Best Practices and Insights when migrating to Apache Iceberg for Data Engineers

Amit Gilad
Data engineer

Agenda

- Introduction
- Ingestion
- Compaction
- Maintenance
- Monitoring
- Benchmarks

About me





About Cloudinary

- 10,000 customers
- 2,000,000 Developers
- 60 Billion assets
- 30 PB monthly bandwidth



- 10-20 TB of logs daily
- 50 Billion records every day
- ~7-14 GB every minute



Why ?



Cost



Time travel



Features

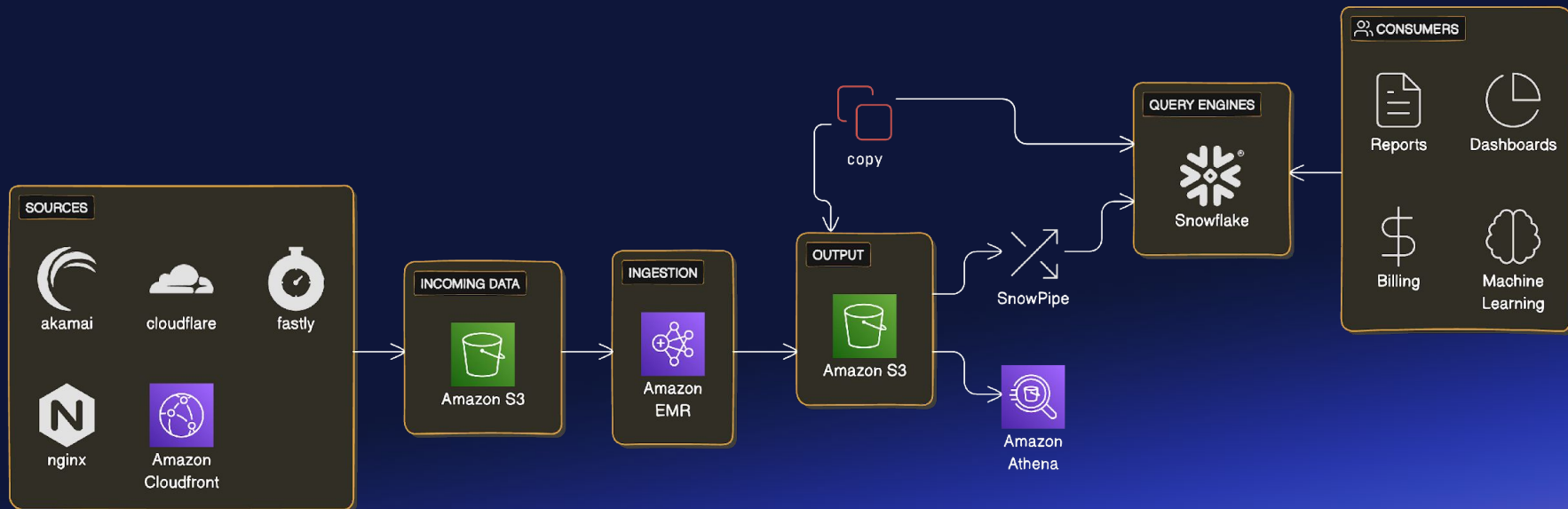


Multiple engines



Data retention

Previous Architecture





End Result

- Reduce Storage cost by 25%
- Data retention 6X
- Query cost reduced by 25-40%
- Single copy
- Reduce query execution time by 30% to 50%



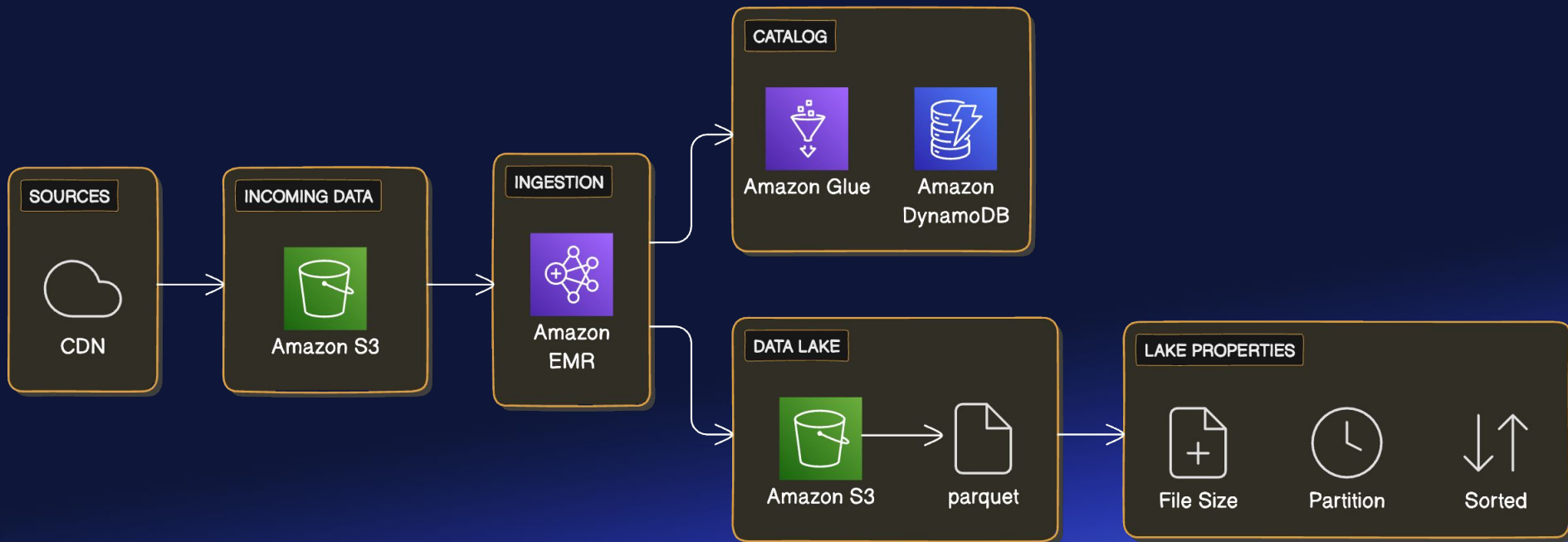
Getting ready

- Select tables for migration
- Mission critical queries
- Cost
- Execution time
- Identify users & what tools they use

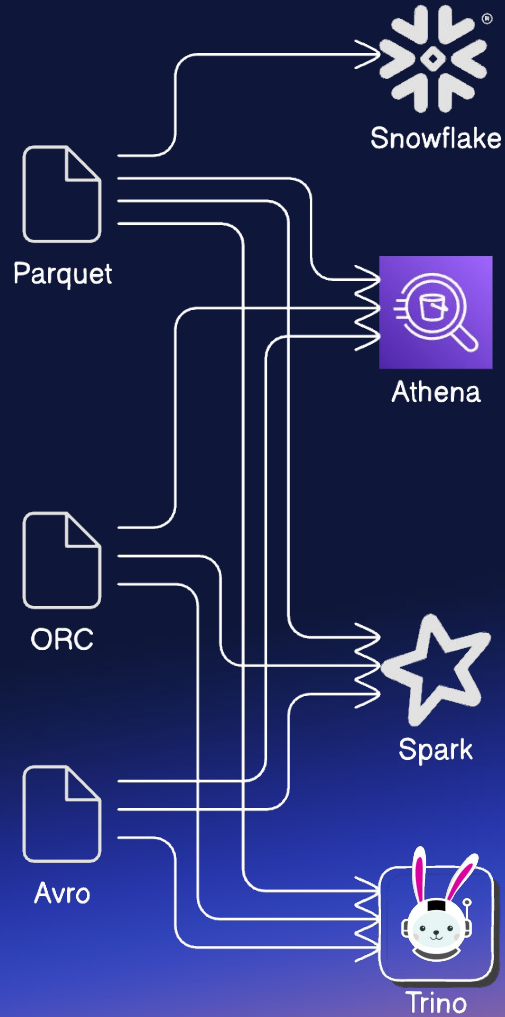


The Journey begins

Ingestion



File Format



Compression ?



Table configuration

- `write.target-file-size-bytes = 1073741824`
- `write.distribution-mode = hash`
- `write.parquet.compression-codec = zstd`
- `write.metadata.delete-after-commit.enabled = true`
- `write.metadata.previous-versions-max = 500`

Adaptive Query Execution = Spark > 3.0.0

- `spark.sql.adaptive.enabled`
 - `spark.sql.adaptive.coalescePartitions.enabled`
- } SMALL FILES

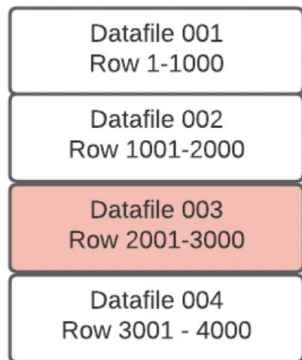
- `spark.sql.adaptive.skewJoin.enabled`
- } BIG FILES

CoW vs MoR

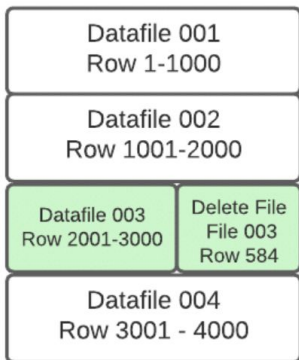
Merge-on-Read

DELETE FROM table where id = 2585

Before Update



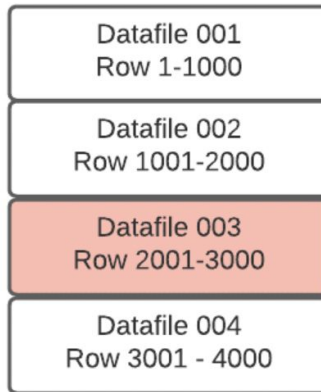
After Update



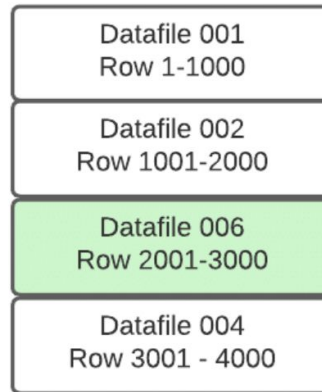
Copy-on-Write

DELETE FROM table where id = 2585

Before Update



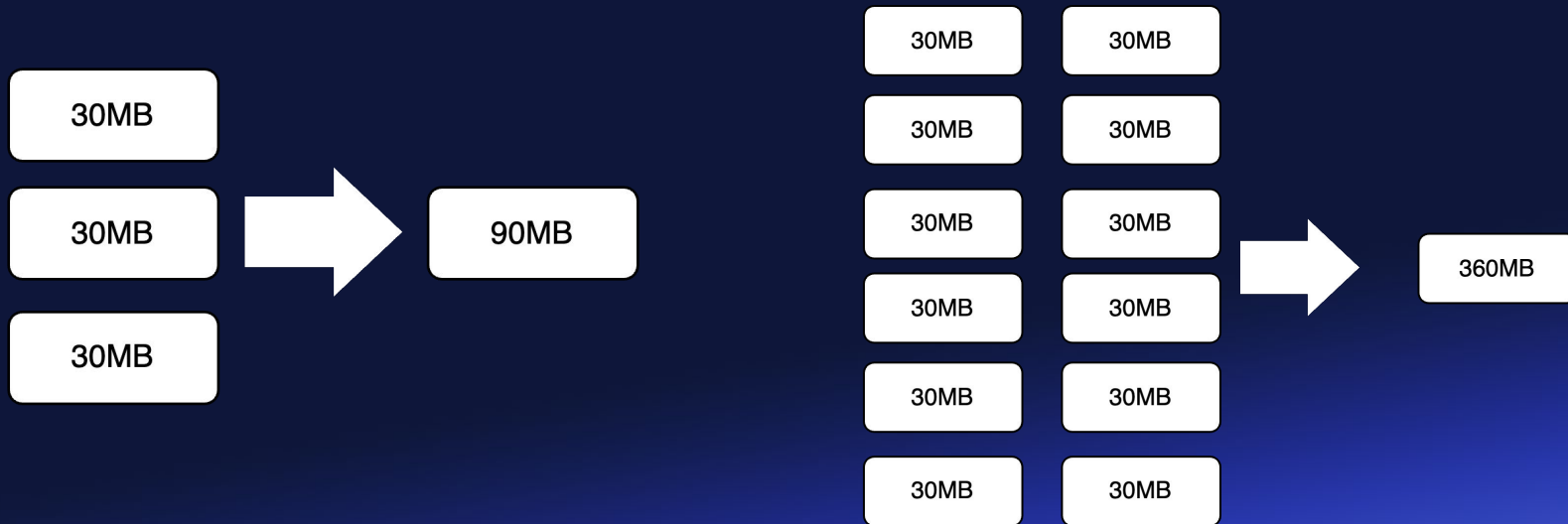
After Update





Rewrite data files (Compaction)

Compact Small files



Rewrite deletes

delete-file-threshold - 2147483647 b



Rewrite data file (compaction) strategy

BinPack

Simple merge or split of files in targeted partitions

- Light operation
- No shuffles

Sort

Shuffle the data in targeted partitions based on hierarchical sort key(s)

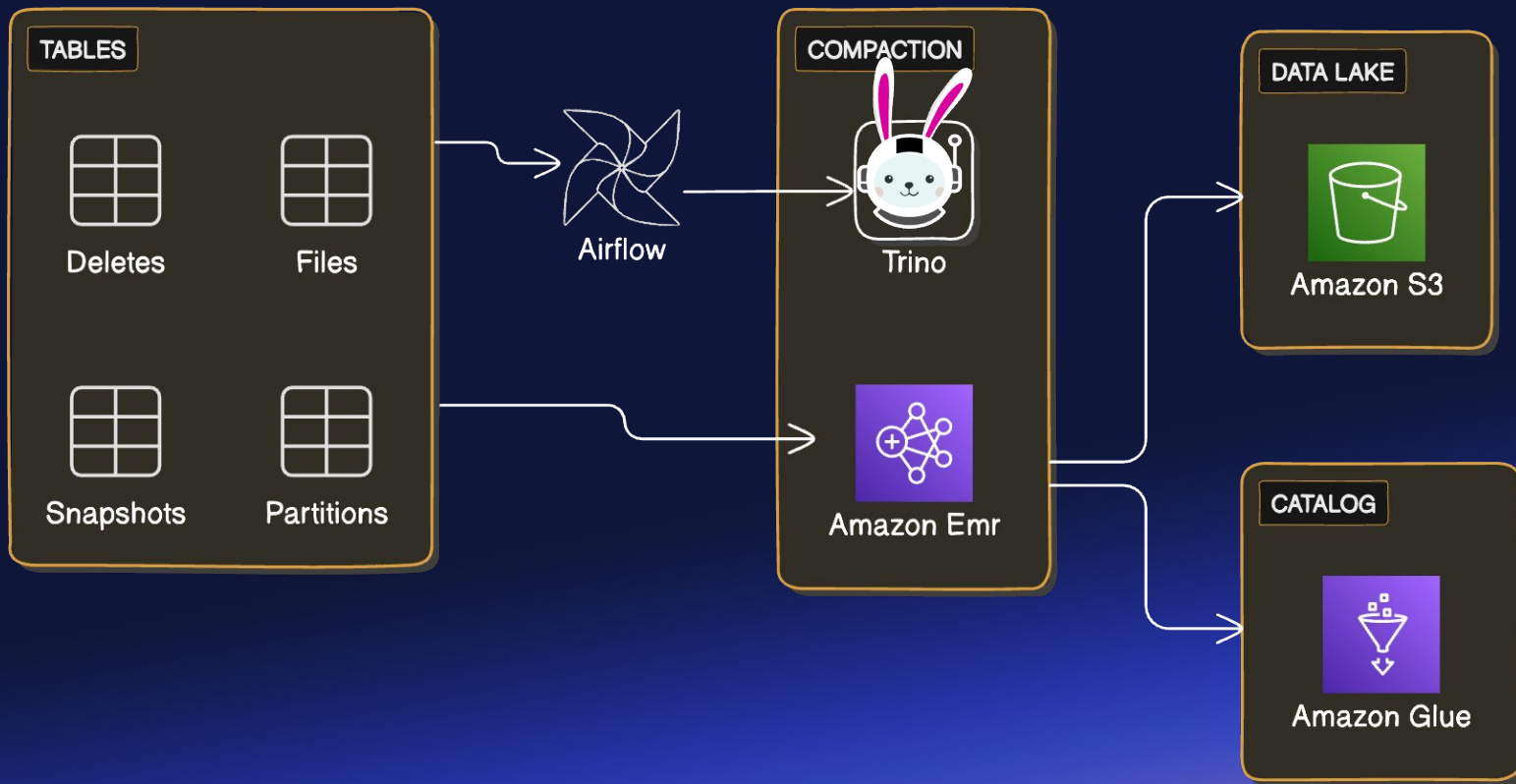
- Medium operation
- Range based shuffle
- Efficient read against sorted columns

Z-order

Cluster the data in targeted partitions based on **multiple columns**

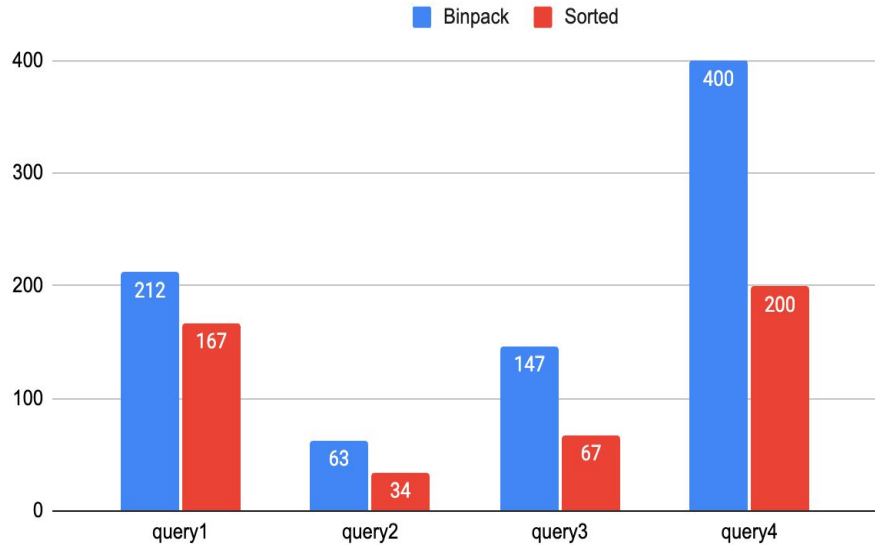
- Expensive operation
- Consider Z-order when using filters on **multiple dimensions**.
- Columns with high cardinality are best suited for Z-ordering.

Compaction Architecture

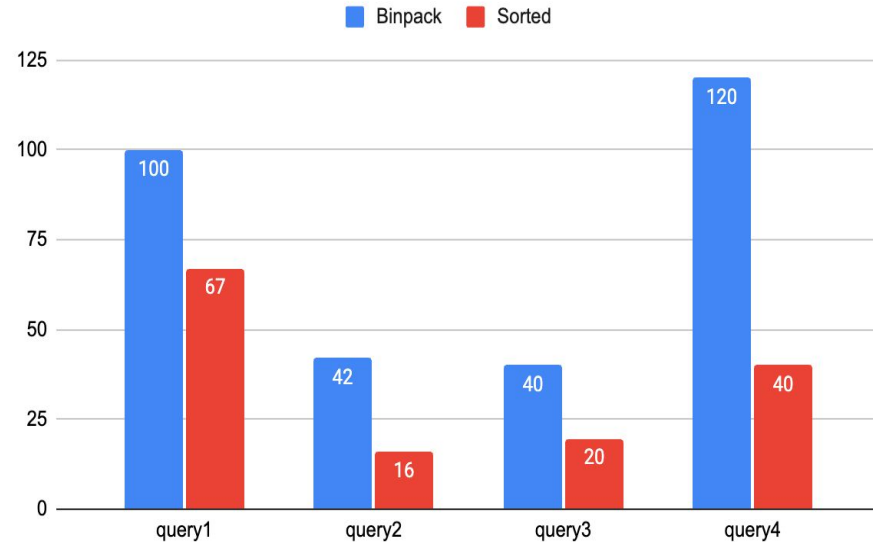


Binpack vs Sorting

Data scanned(GB)



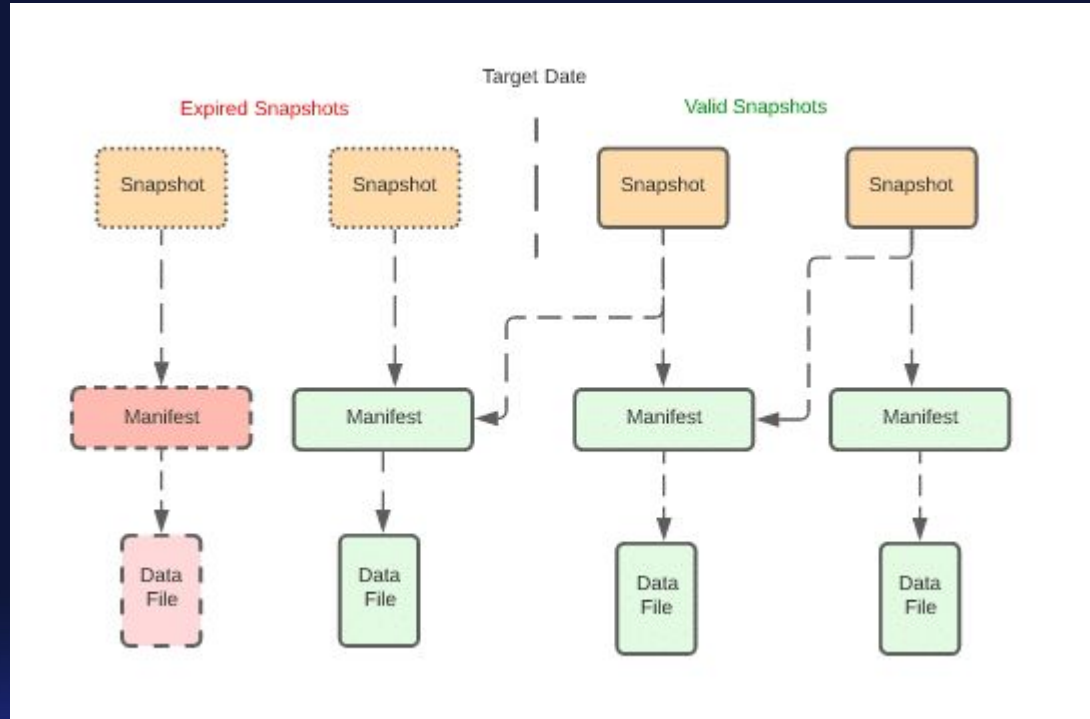
Query Time





Maintenance

Expire snapshots

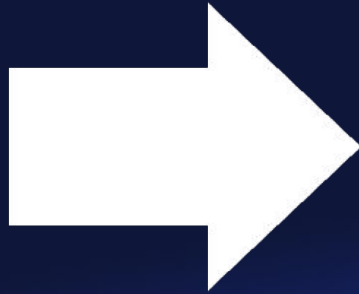


Rewrite manifests

256KB

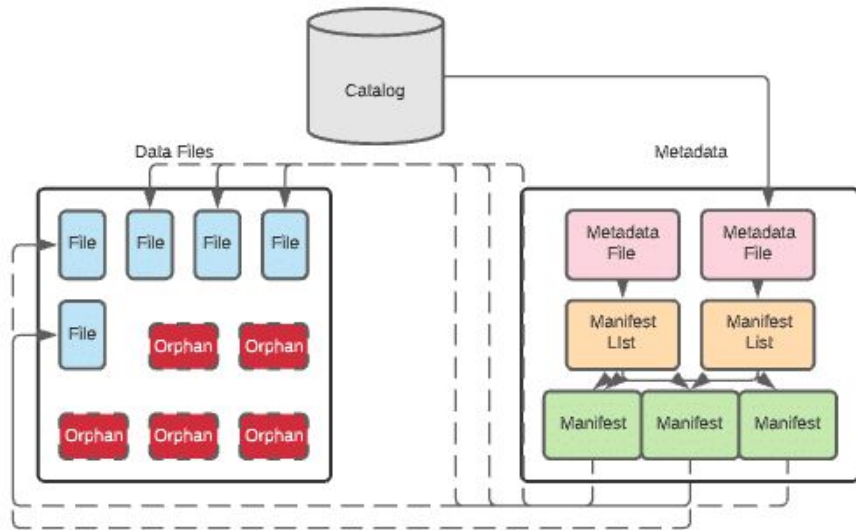
1MB

2MB

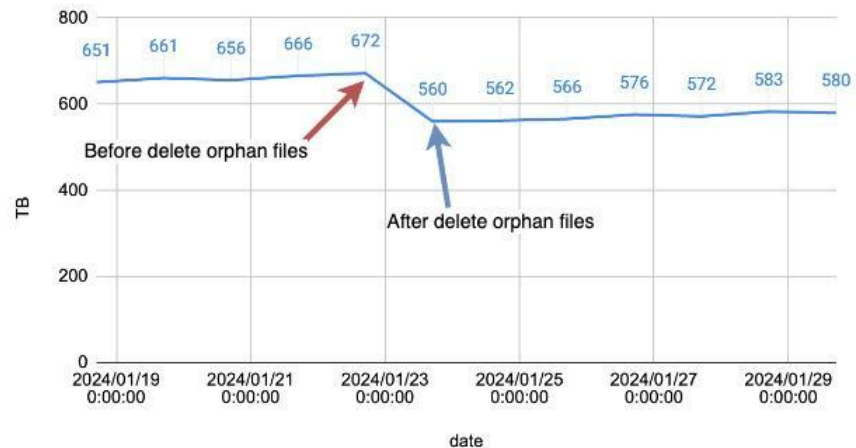


3MB

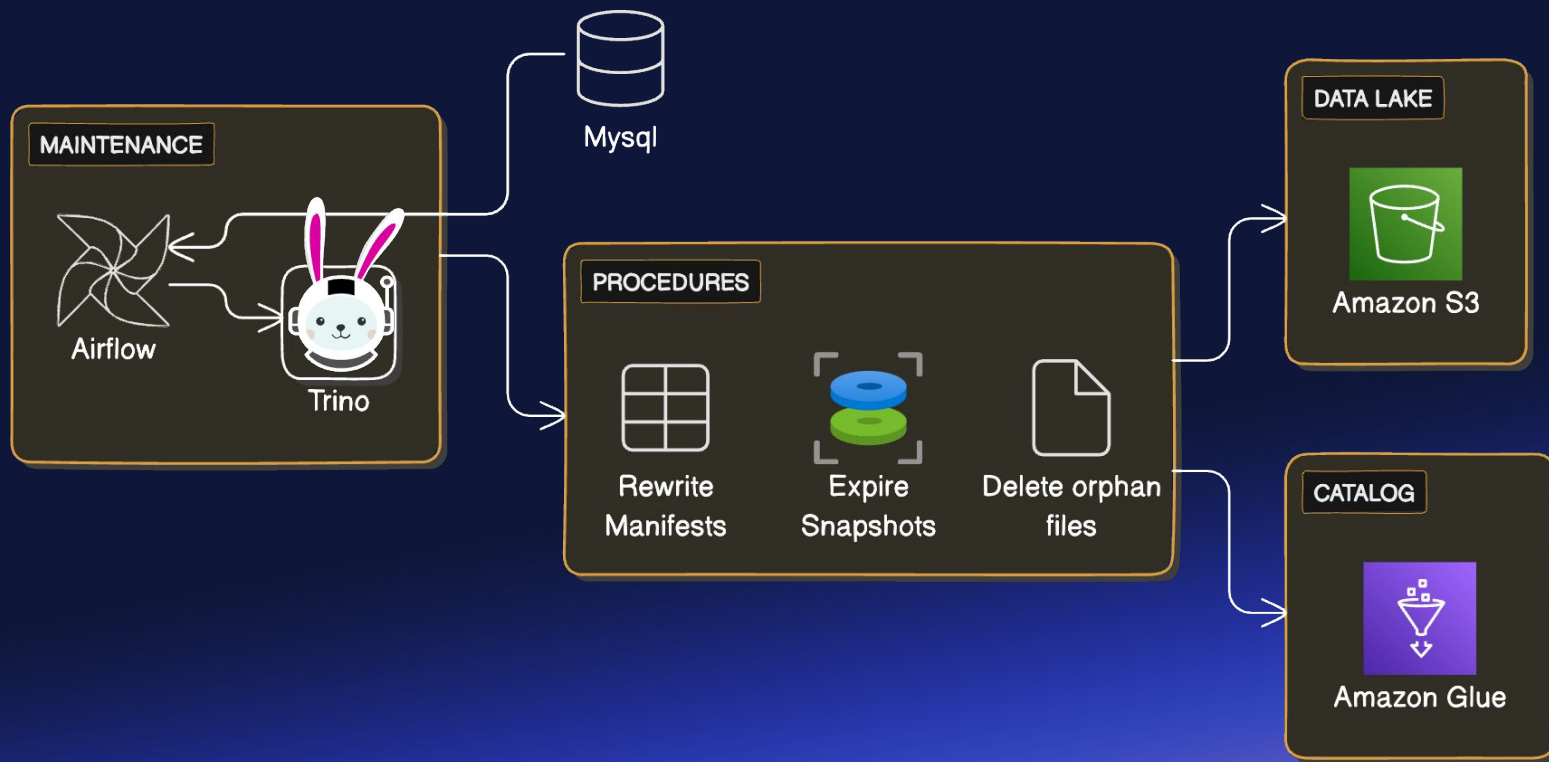
Delete orphan files



Total table size



Maintenance Architecture





Monitoring

Metadata tables- (snapshots)

committed_at_hour	operation	added_records	added_data_files	deleted_data_files	size_gb	avg_file_size_mb
2024-03-10 21:00:00.000 UTC	replace	81026952	40	679	19	505
2024-03-10 21:00:00.000 UTC	append	94690651	781		21	29
2024-03-10 20:00:00.000 UTC	replace	704952135	333	5742	165	510
2024-03-10 20:00:00.000 UTC	append	700412961	5824		161	29
2024-03-10 19:00:00.000 UTC	append	894800167	7485		192	27
2024-03-10 19:00:00.000 UTC	replace	1024656767	458	8451	229	513
2024-03-10 18:00:00.000 UTC	replace	809592621	361	6868	180	513

Metrics reporter api

Spark Iceberg Metrics ▾

Share Show Overlays Configure Add Widgets

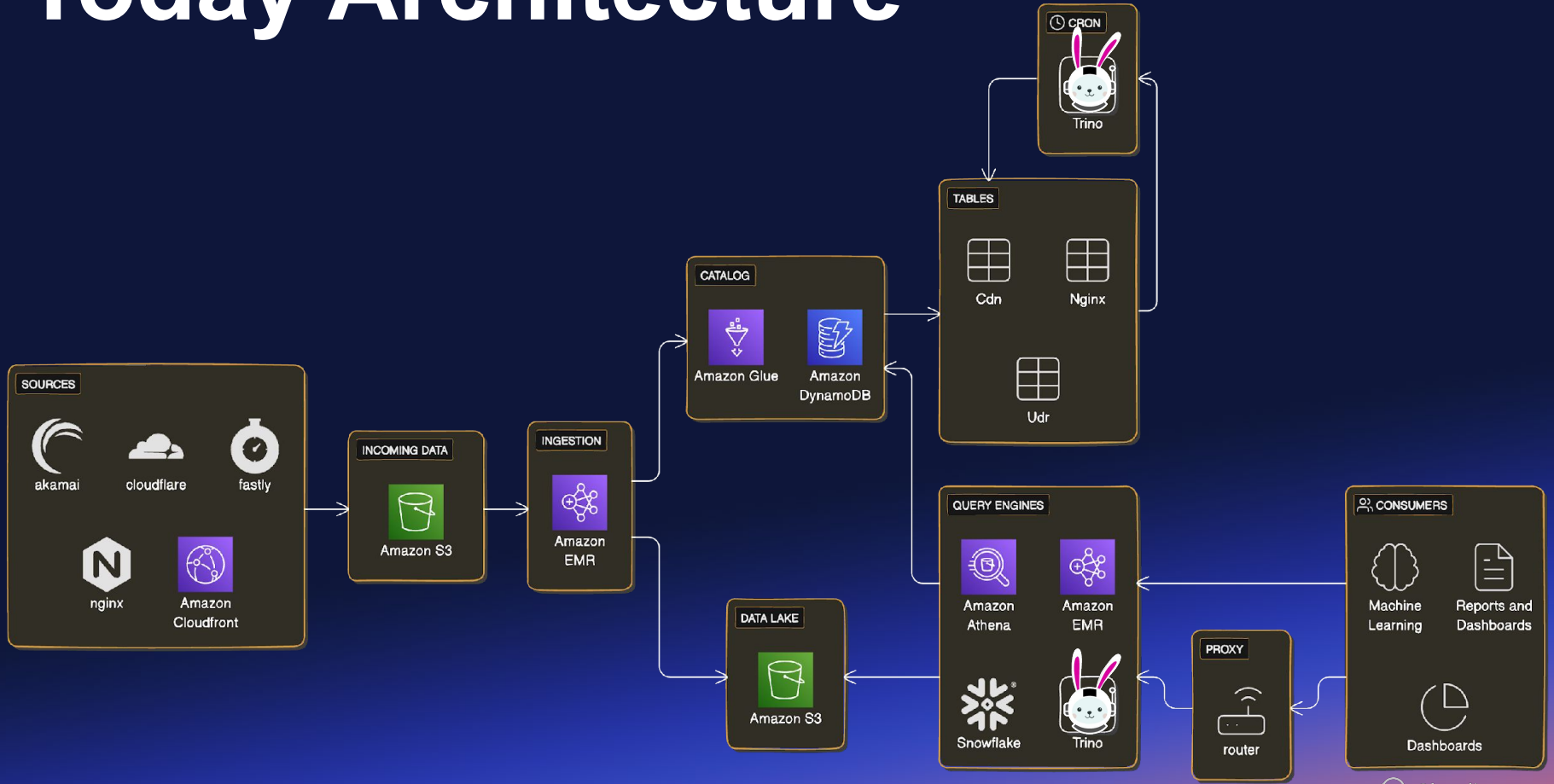
Saved Views cluster spark_production_writer_cdnline

1mo Past Month

The dashboard displays the following metrics:

- SQS - cdnline:** A line chart showing a sharp spike to over 10k around May 15, followed by a low, stable baseline.
- Total Table Size:** A large text display showing 477T, with a shaded area indicating a range from 4.5T to 476.6T.
- Added File Size:** A line chart showing a significant spike to 4T around May 22, followed by a fluctuating baseline.
- Processed Rawjson Files:** A line chart showing a steady increase from near zero to approximately 200k files by May 19, with subsequent fluctuations.
- Total Data Files:** A line chart showing a sharp increase to about 0.8M files around May 15, followed by a dip and then a steady rise to nearly 1M files.
- Total Table Size (600T):** A line chart showing a sharp increase to 400T around May 15, followed by a dip and then a steady rise to over 600T.
- Added Positional Deletes:** A line chart showing a cluster of spikes between May 20 and May 25, peaking at approximately 50k.
- Total Delete Files:** A line chart showing a single sharp spike to 200 files around May 15, followed by a very low baseline.
- Total Records:** A line chart showing a sharp increase to about 1.8T records around May 15, followed by a dip and then a steady rise to over 2T records.

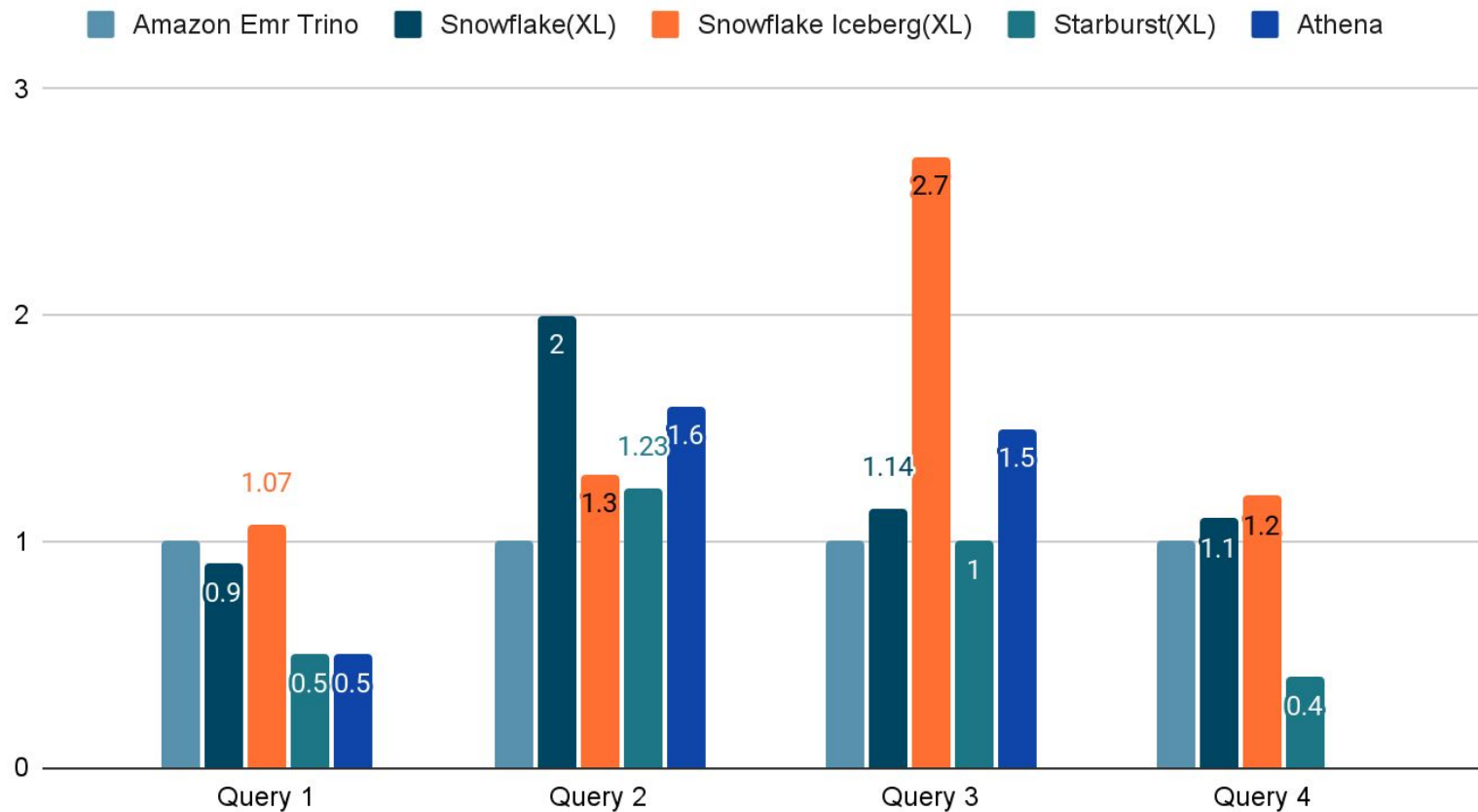
Today Architecture



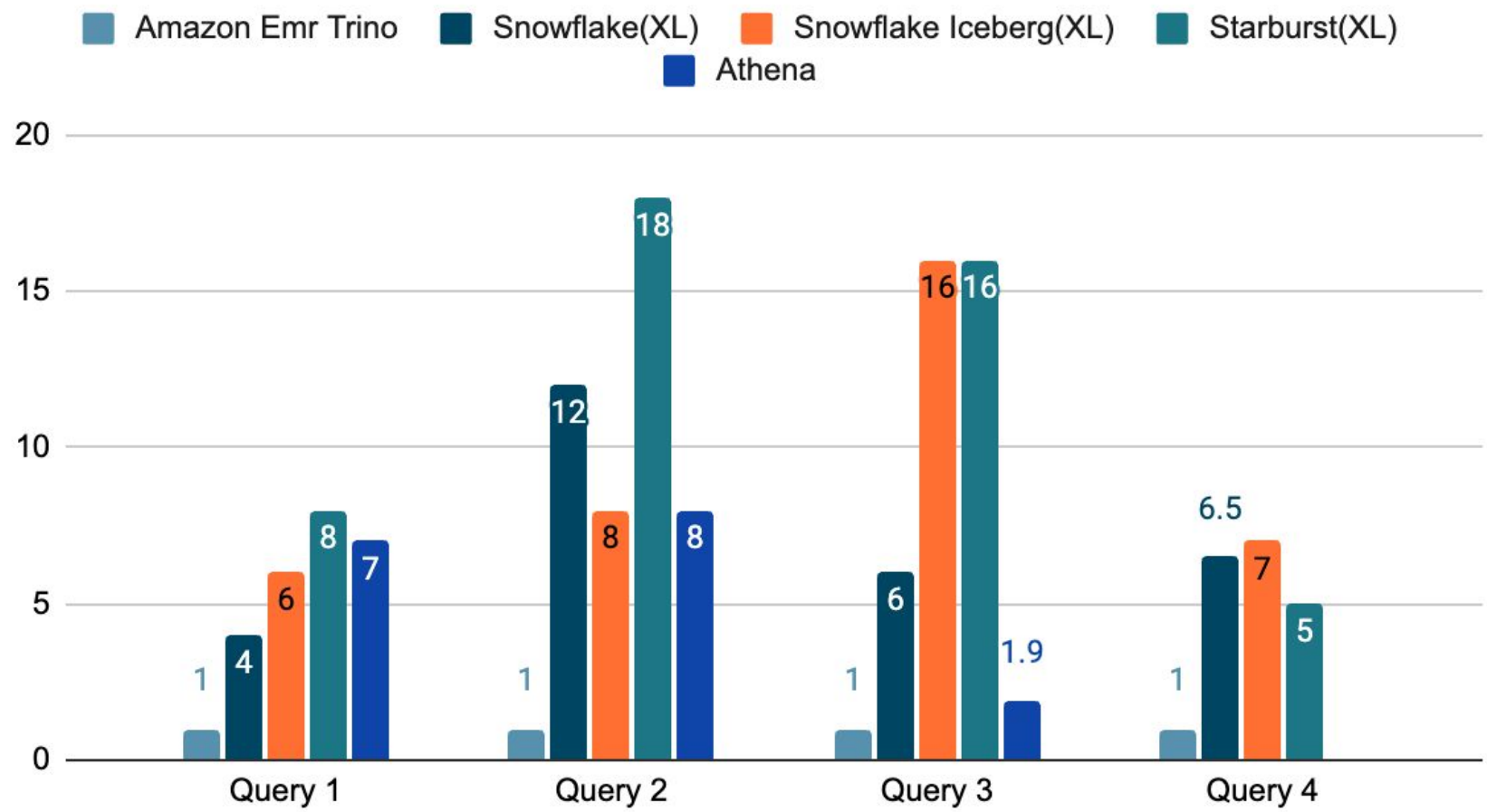


Benchmarks

Query duration normalized



Query cost normalized



Amit Gilad

Questions?
Thanks

