Data Council
Singapore 2019

presto

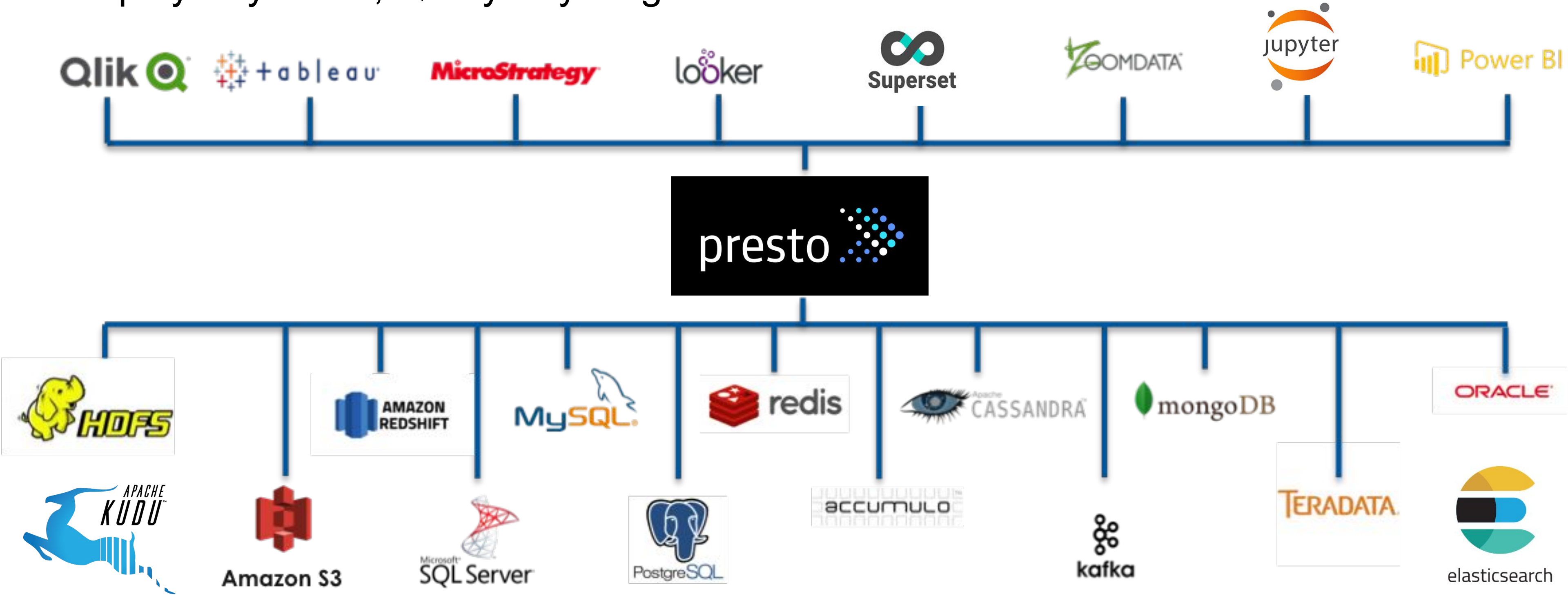
Optimizing Performance of SQL-on-Anything Engine

@prestosql @starburstdata

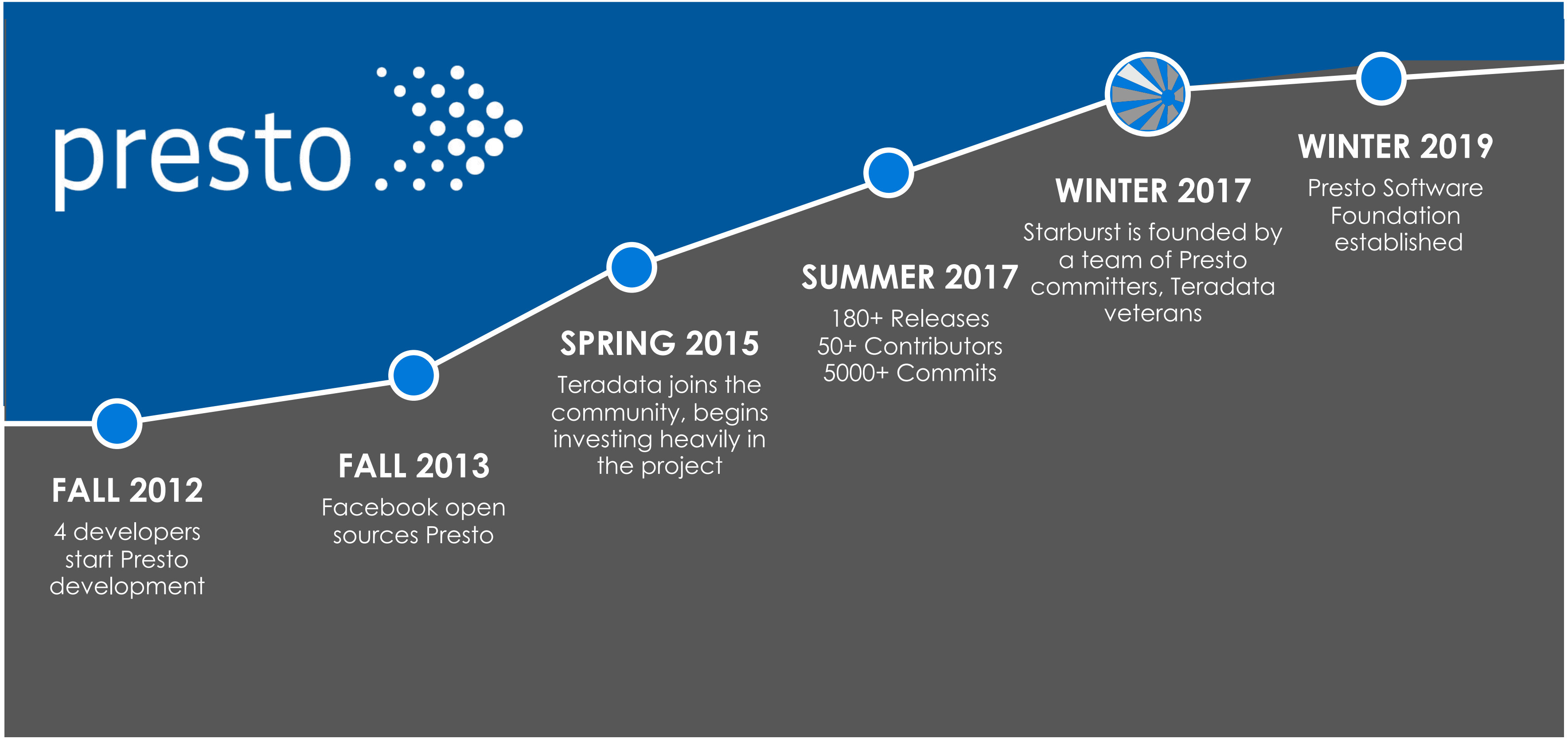
Kamil Bajda-Pawlikowski, CTO Starburst

Presto: SQL-on-Anything

Deploy Anywhere, Query Anything



Project History



presto

FALL 2012

4 developers
start Presto
development

FALL 2013

Facebook open
sources Presto

SPRING 2015

Teradata joins the
community, begins
investing heavily in
the project

SUMMER 2017

180+ Releases
50+ Contributors
5000+ Commits

WINTER 2017

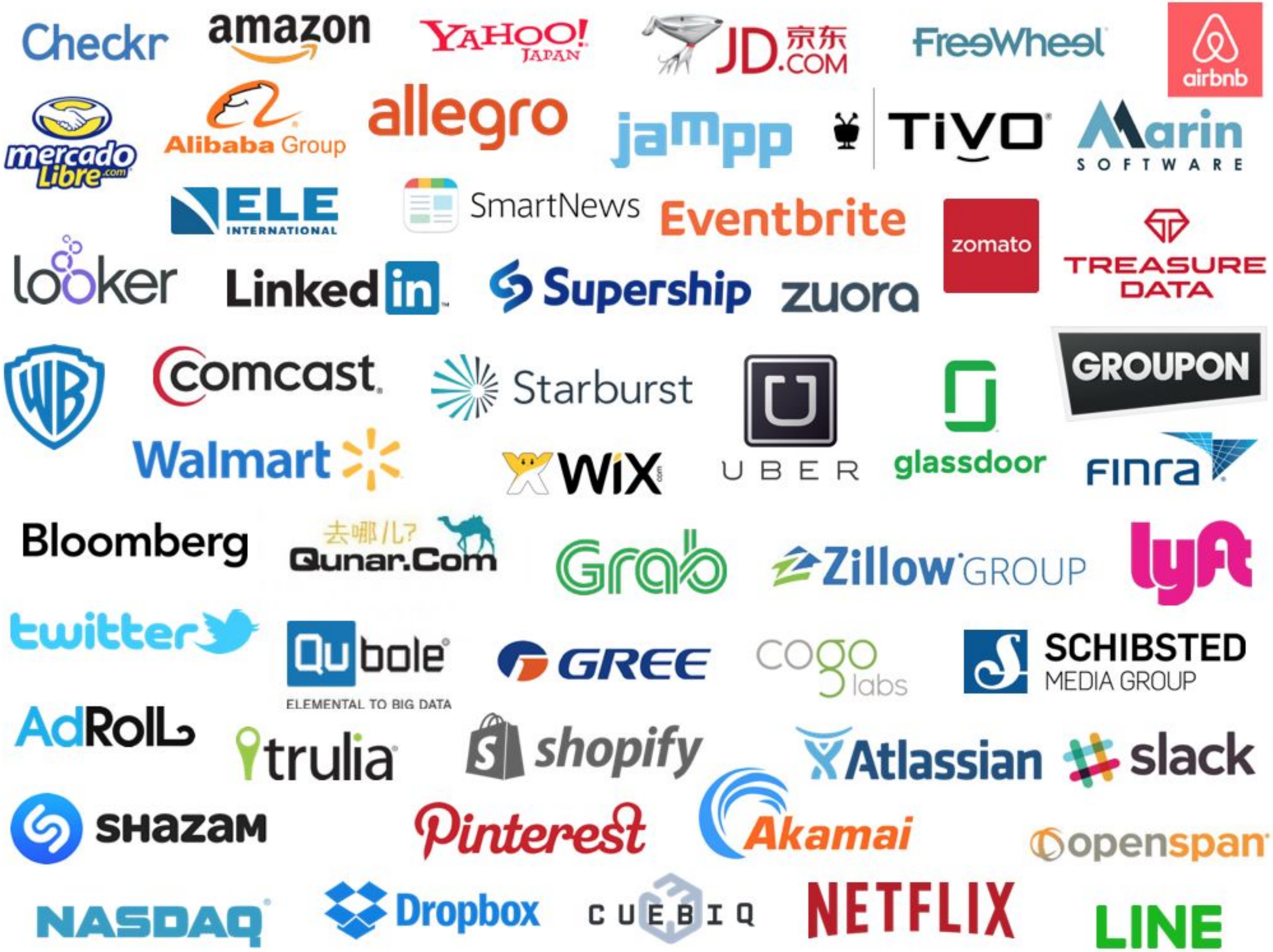
Starburst is founded by
a team of Presto
committers, Teradata
veterans

WINTER 2019

Presto Software
Foundation
established

Community

See more at our Wiki



Presto in Production

Facebook: 10,000+ of nodes, HDFS (ORC, RCFile), sharded MySQL, 1000s of users

Uber: 2,000+ nodes (several clusters on premises) with 160K+ queries daily over HDFS (Parquet/ORC)

Twitter: 2,000+ nodes (several clusters on premises and GCP), 20K+ queries daily (Parquet)

LinkedIn: 500+ nodes, 200K+ queries daily over HDFS (ORC), and ~1000 users

Lyft: 400+ nodes in AWS, 100K+ queries daily, 20+ PBs in S3 (Parquet)

Netflix: 300+ nodes in AWS, 100+ PB in S3 (Parquet)

Yahoo! Japan: 200+ nodes for HDFS (ORC), and ObjectStore

FINRA: 120+ nodes in AWS, 4PB in S3 (ORC), 200+ users

Why Presto?

Community-driven open source project

High performance ANSI SQL engine

- New Cost-Based Query Optimizer
- Proven scalability
- High concurrency

Separation of compute and storage

- Scale storage and compute independently
- No ETL or data integration necessary to get to insights
- SQL-on-anything

No vendor lock-in

- No Hadoop distro vendor lock-in
- No storage engine vendor lock-in
- No cloud vendor lock-in

Beyond ANSI SQL

Presto offers a wide variety of built-in functions including:

- regular expression functions
- lambda expressions and functions
- geospatial functions

Complex data types:

- JSON
- ARRAY
- MAP
- ROW / STRUCT

```
SELECT regexp_extract_all('1a 2b 14m', '\d+');  -- [1, 2, 14]

SELECT filter(ARRAY [5, -6, NULL, 7], x -> x > 0); -- [5, 7]

SELECT transform(ARRAY [5, 6], x -> x + 1); -- [6, 7]


SELECT c.city_id, count(*) as trip_count
FROM trips_table as t
JOIN city_table as c
ON st_contains(c.geo_shape,
     st_point(t.dest_lng, t.dest_lat))
WHERE t.trip_date = ‘2018-05-01’
GROUP BY 1;
```

Tools, bindings, extensibility

redash
Superset
jupyter

JDBC / ODBC drivers for BI/SQL tools

C/C++, Go, Java, Node.js, Python, PHP, R and Ruby on Rails

UDFs, UDAFs, Connector SPI

More connectors

https://www.starburstdata.com/technical-blog/starburst-presto-databricks-delta-lake-support/

https://streaml.io/blog/querying-data-streams-with-apache-pulsar-sql

http://iceberg.incubator.apache.org/

https://eng.uber.com/apache-hudi/

https://tiledb.io/press/tiledb-presto

https://engineering.grab.com/big-data-real-time-presto-talariadb

https://blog.yugabyte.com/presto-on-yugabyte-db-interactive-olap-sql-queries-made-easy-facebook/

Enterprise edition

Founded by Presto committers:

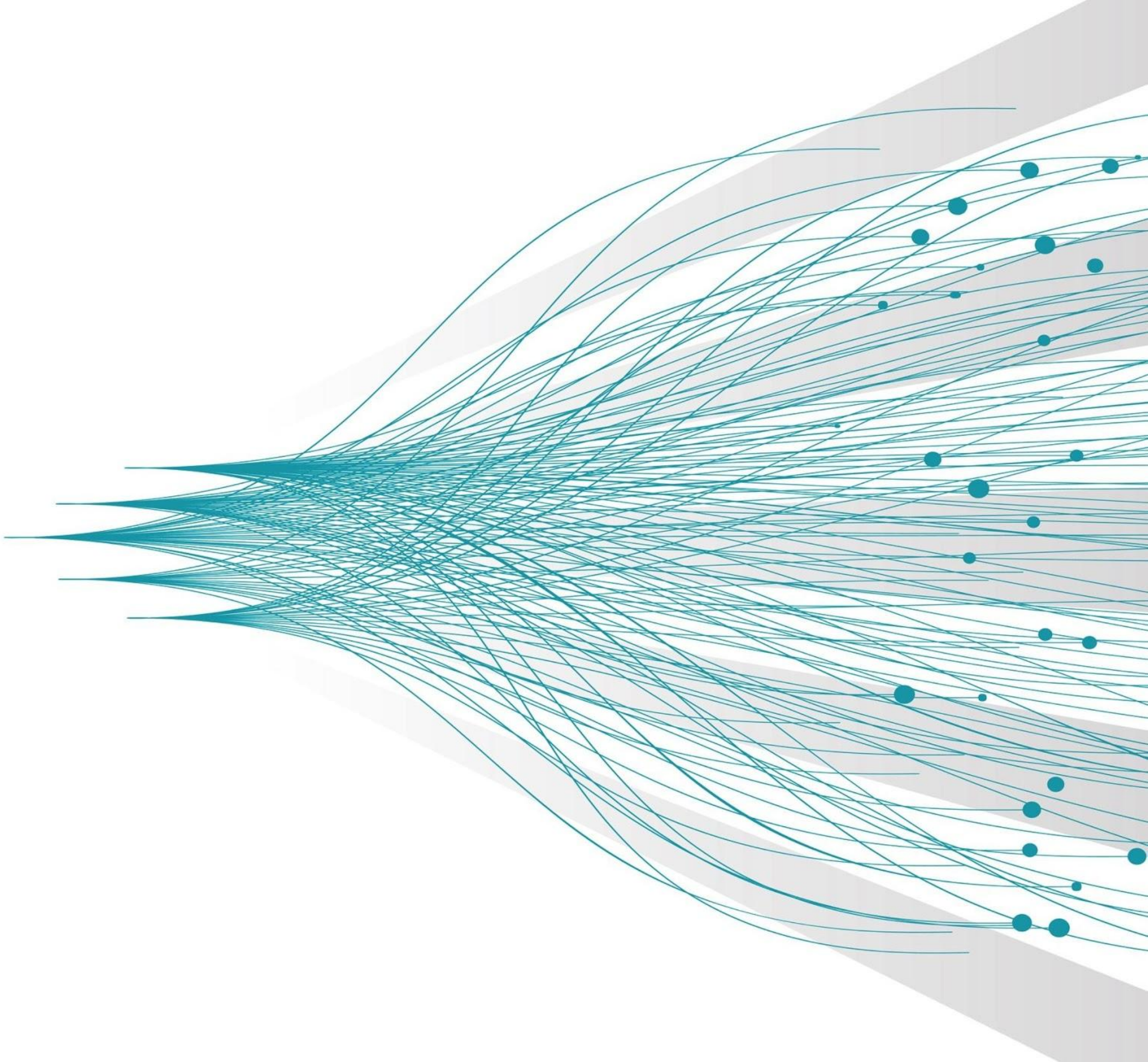
- Over 4 years of contributions to Presto
- Presto distro for on-prem and cloud env
- Supporting large customers in production
- Enterprise subscription add-ons (ODBC, Ranger, Sentry, Oracle, Teradata, K8S)

Notable features contributed:

- ANSI SQL syntax enhancements
- Execution engine improvements
- Security integrations
- Spill to disk
- Cost-Based Optimizer

https://www.starburstdata.com/presto-enterprise/

Performance



Built for Performance

Query Execution Engine:

- MPP-style **pipelined** in-memory execution
- **Columnar** and **vectorized** data processing
- Runtime query **bytecode compilation**
- Memory **efficient** data structures
- Multi-threaded multi-core execution
- Optimized readers for **columnar formats** (ORC and Parquet)
- Predicate and column projection **pushdown**
- Now also **Cost-Based Optimizer**

CBO in a nutshell

Presto Cost-Based Optimizer includes:

- support for **statistics** stored in Hive Metastore
- **join reordering** based on selectivity estimates and cost
- automatic **join type** selection (repartitioned vs broadcast)
- automatic left/right side selection for joined tables

https://www.starburstdata.com/technical-blog/

Statistics & Cost

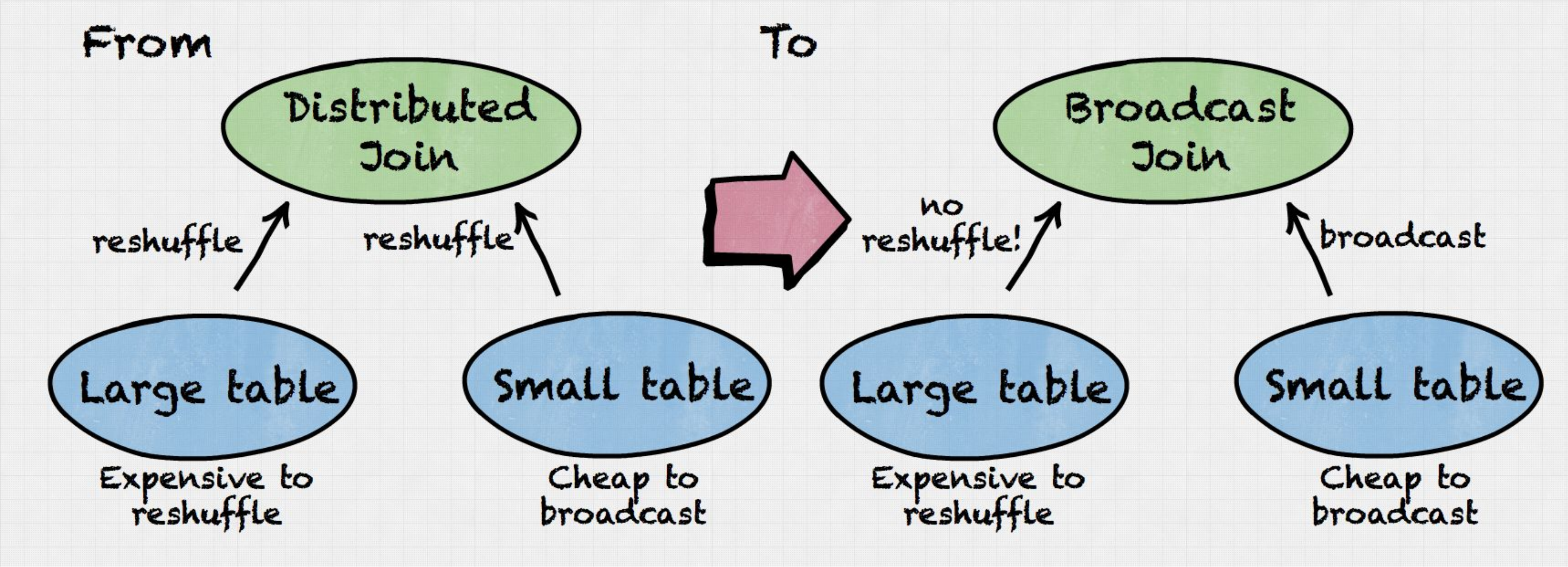
Hive Metastore statistics:

- number of rows in a table
- number of distinct values in a column
- fraction of NULL values in a column
- minimum/maximum value in a column
- average data size for a column

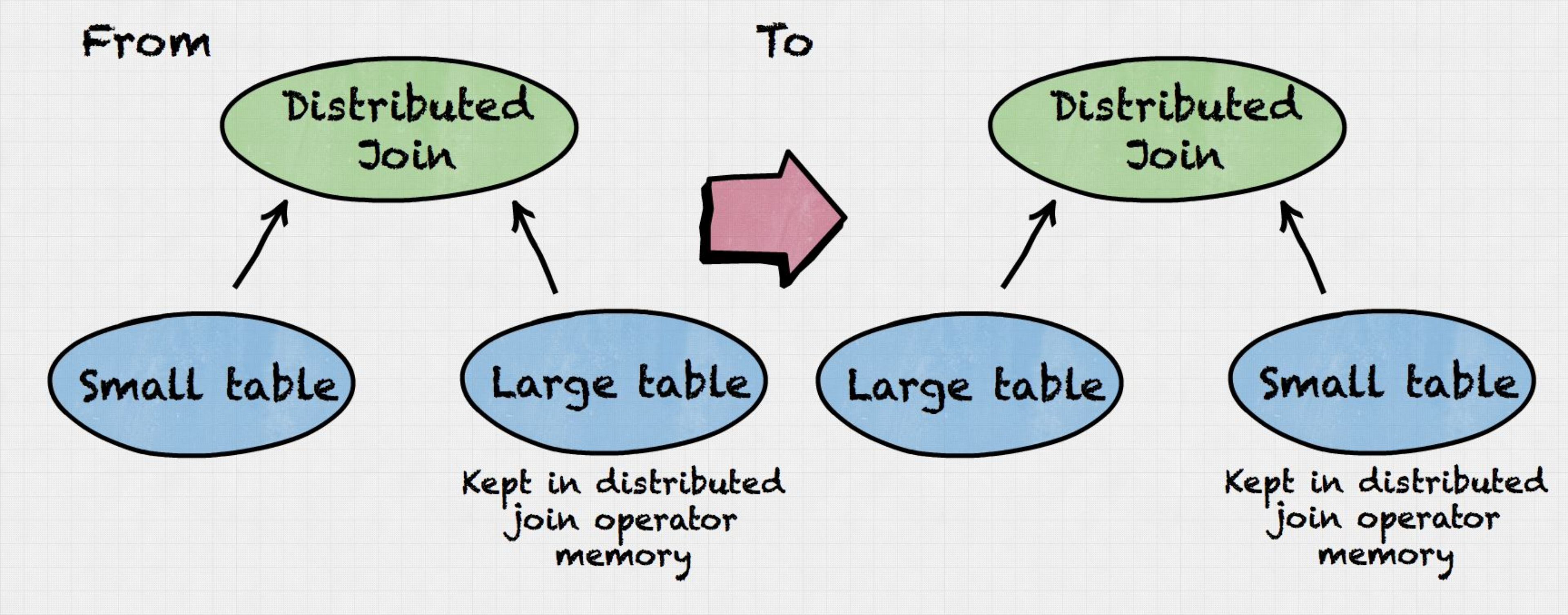
Cost calculation includes:

- CPU
- Memory
- Network I/O

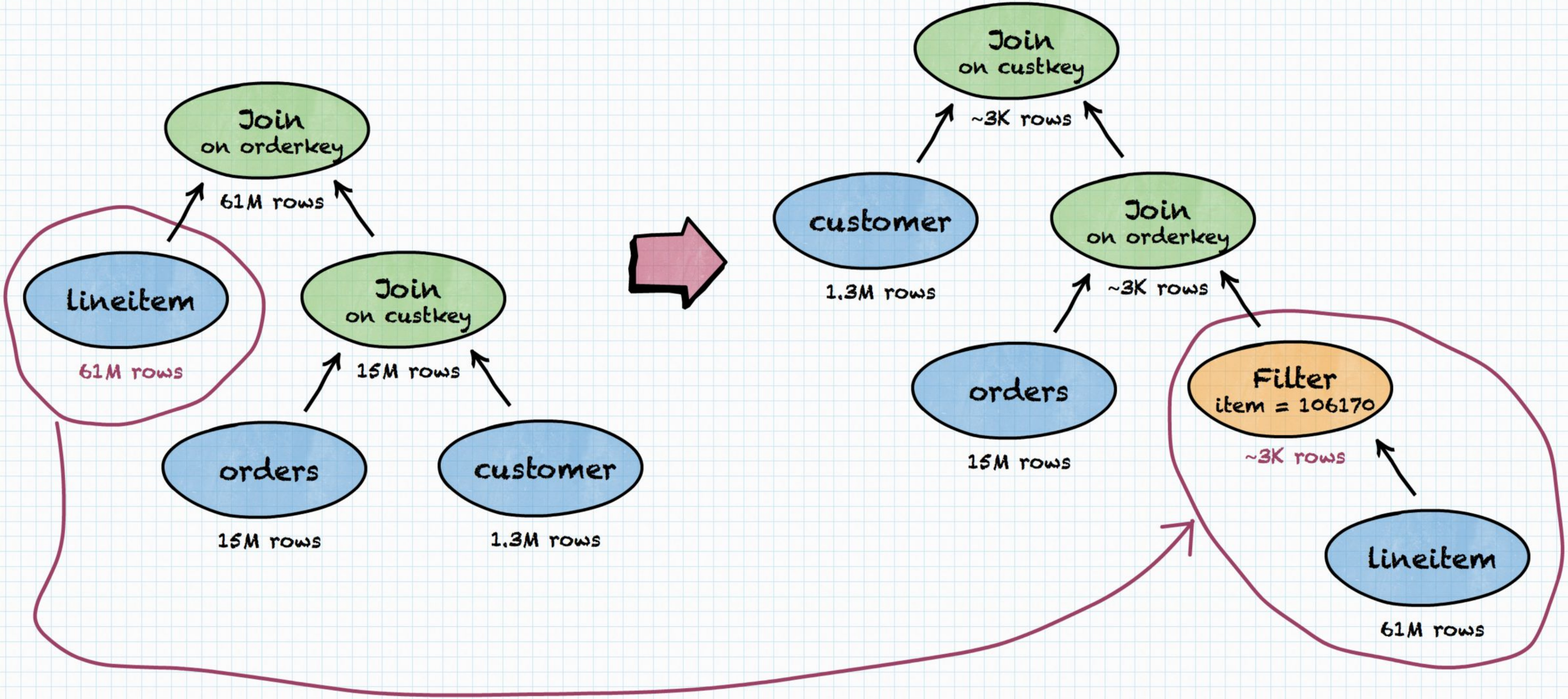
Join type selection



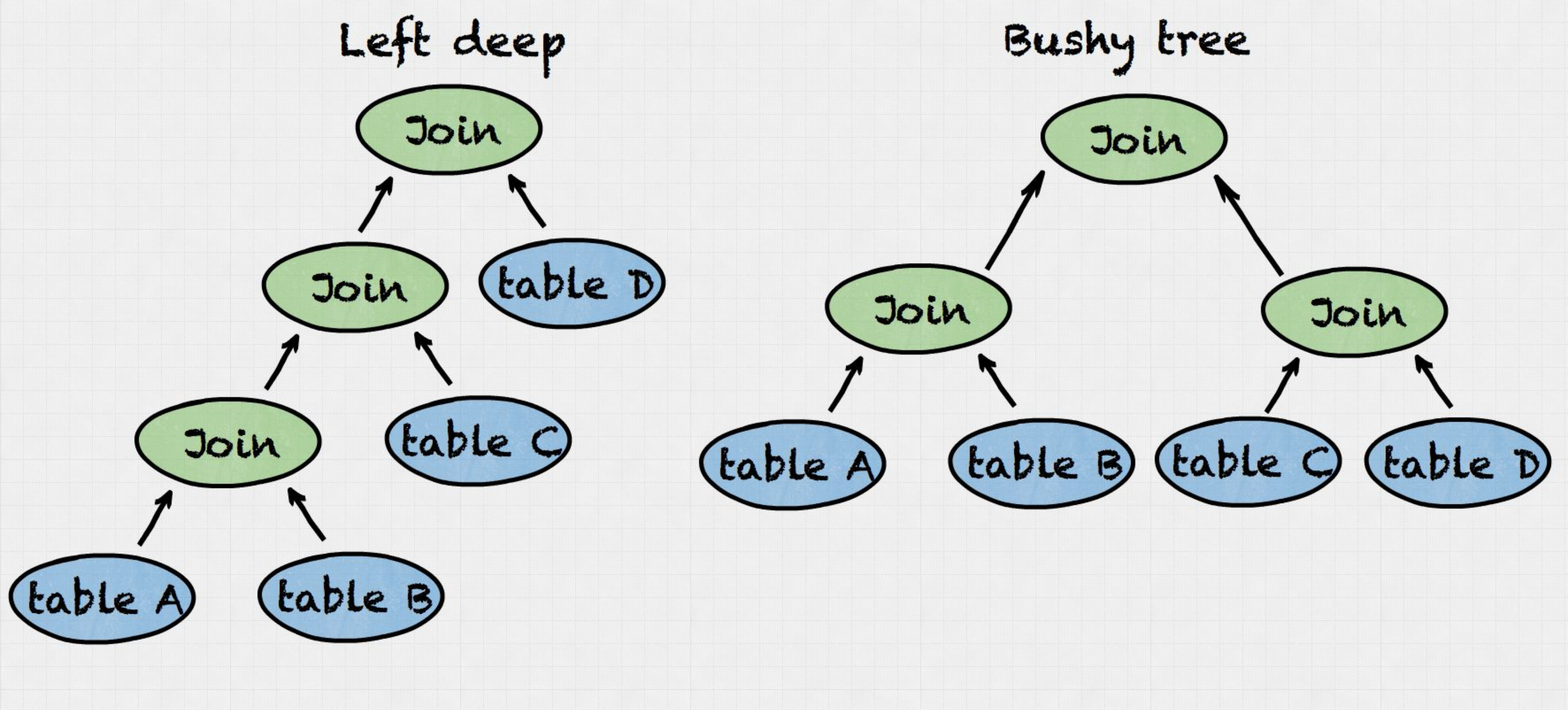
Join left/right side decision



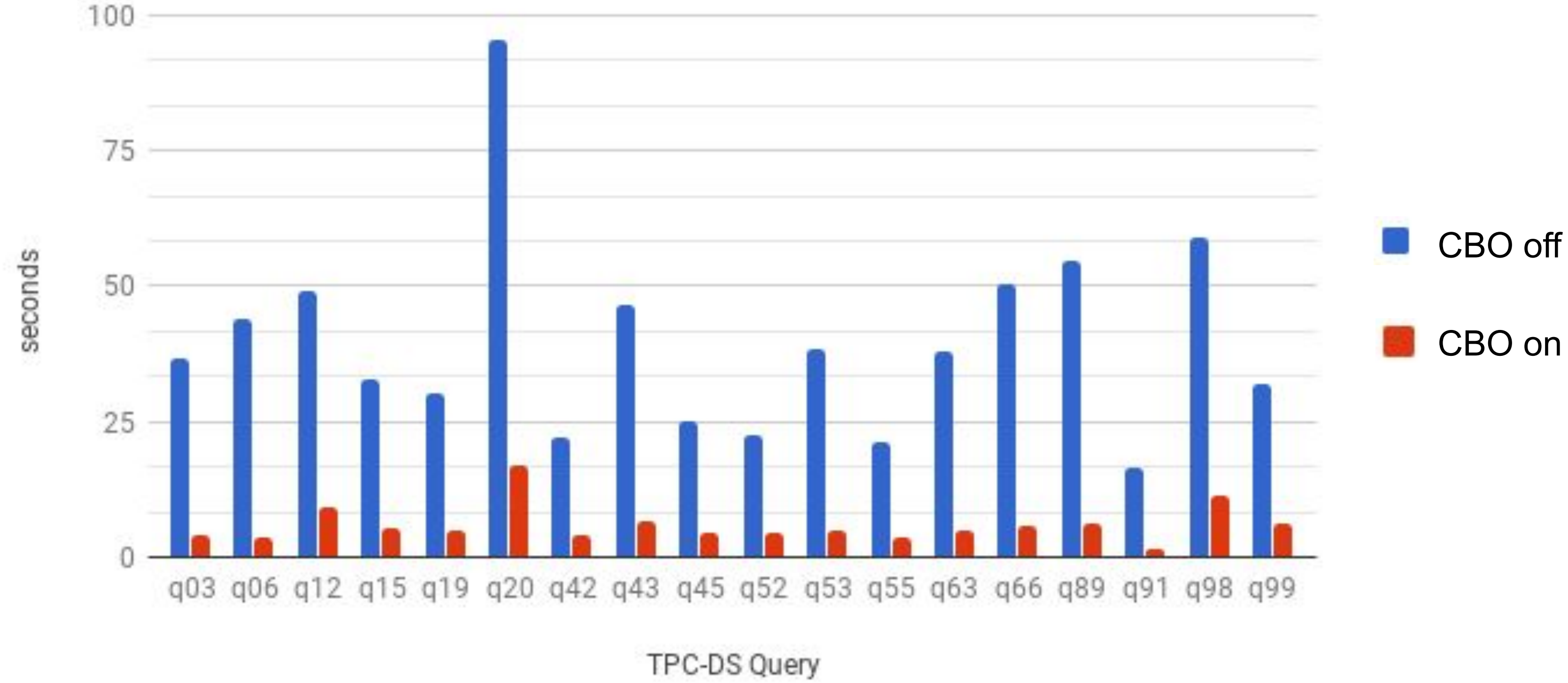
Join reordering with filter



Join tree shapes



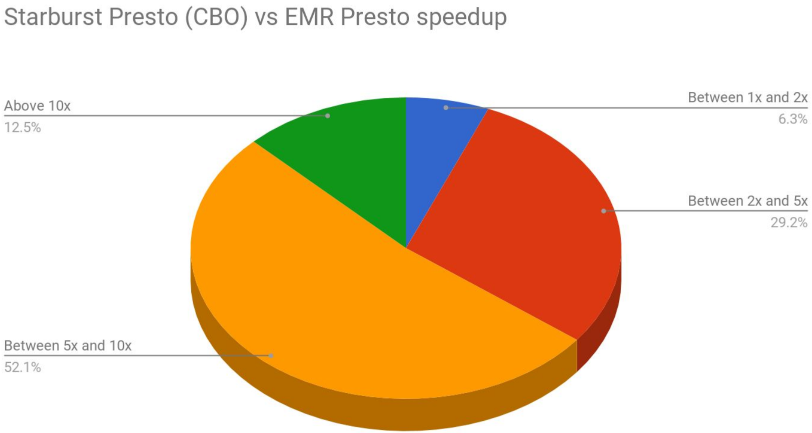
Benchmark results

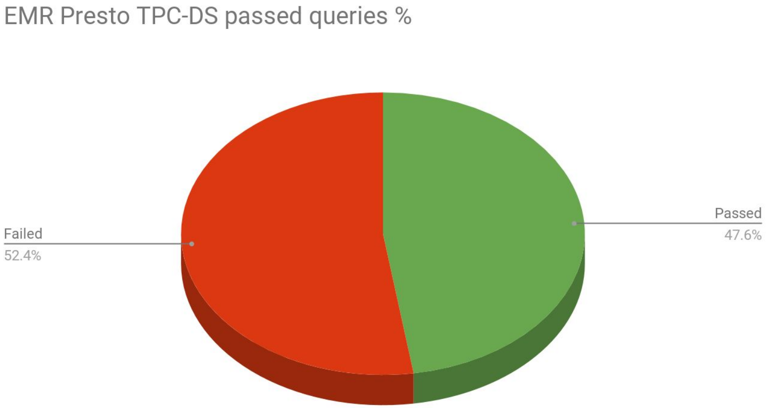


https://www.starburstdata.com/presto-benchmarks/

Benchmark results

- on average 7x improvement vs EMR Presto
- EMR Presto cannot execute many TPC-DS queries
- All TPC-DS queries pass on Starburst Presto





https://www.starburstdata.com/presto-aws/

Recent CBO enhancements

- Deciding on semi-join distribution type based on cost
- Capping a broadcasted table size
- Various minor fixes in cardinality estimation
- ANALYZE table (native in Presto)
- Stats for AWS Glue Catalog
- Enabling DBMS federation use cases

What’s next for Optimizer

- Enhanced stats support
    - Improved stats for Hive
    - Stats for more DBMS and NoSQL connectors
    - Tolerate missing / incomplete stats
- Core CBO improvements
    - Cost more operators
    - Adjust cost model weights based on the hardware
    - Adaptive optimizations
    - Introduce Traits
- Involve connectors in optimizations

Further reading

https://www.prestosql.io

https://www.starburstdata.com

https://fivetran.com/blog/warehouse-benchmark

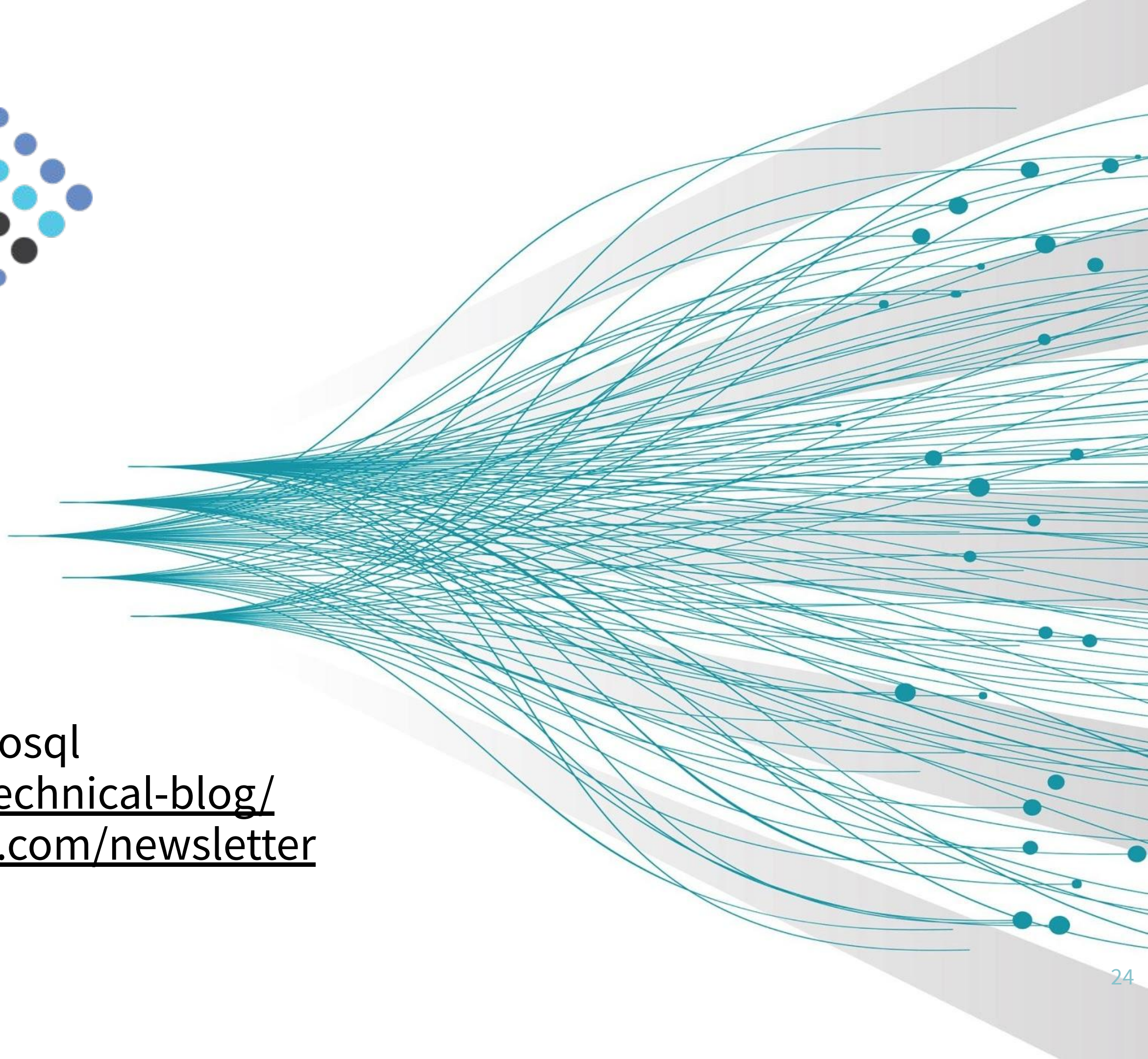
https://www.concurrencylabs.com/blog/starburst-presto-vs-aws-emr-sql/

http://bytes.schibsted.com/bigdata-sql-query-engine-benchmark/

https://virtuslab.com/blog/benchmarking-spark-sql-presto-hive-bi-processing-googles-cloud-dataproc/

presto

Thank You!



Twitter: @starburstdata @prestosql
Blog: www.starburstdata.com/technical-blog/
Newsletter: www.starburstdata.com/newsletter