



Starburst

presto 

The Presto Cost-Based Optimizer for interactive SQL on anything

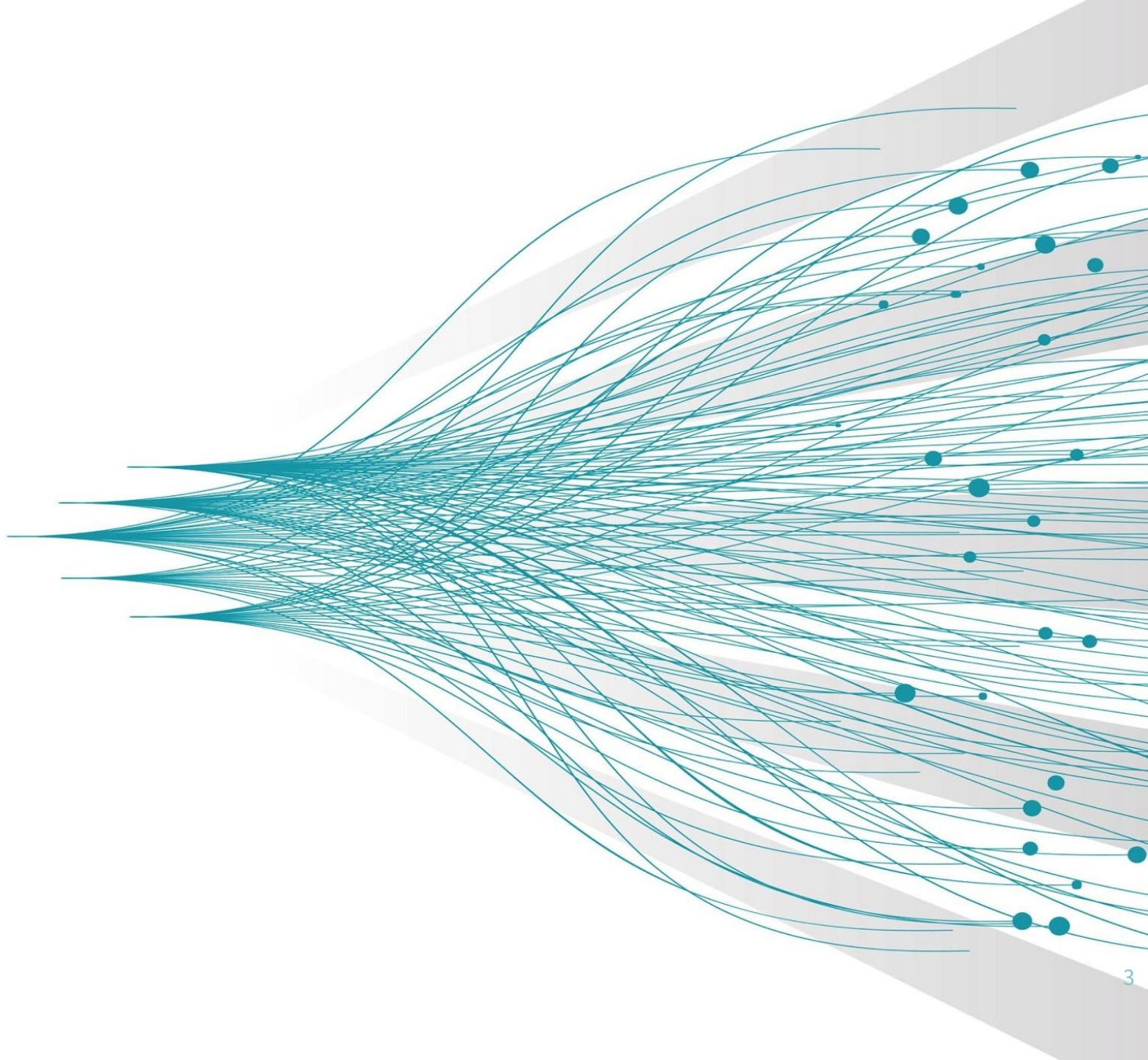
Strata Data Conference - London
May 1, 2019

Wojciech.Biela@starburstdata.com
Piotr.Findeisen@starburstdata.com

Agenda

- Who we are?
- Presto - quick recap
- Presto's CBO launch
- Recent CBO enhancements
- CBO Roadmap

Starburst



Starburst Data

The Presto^{⚙️} Experts.



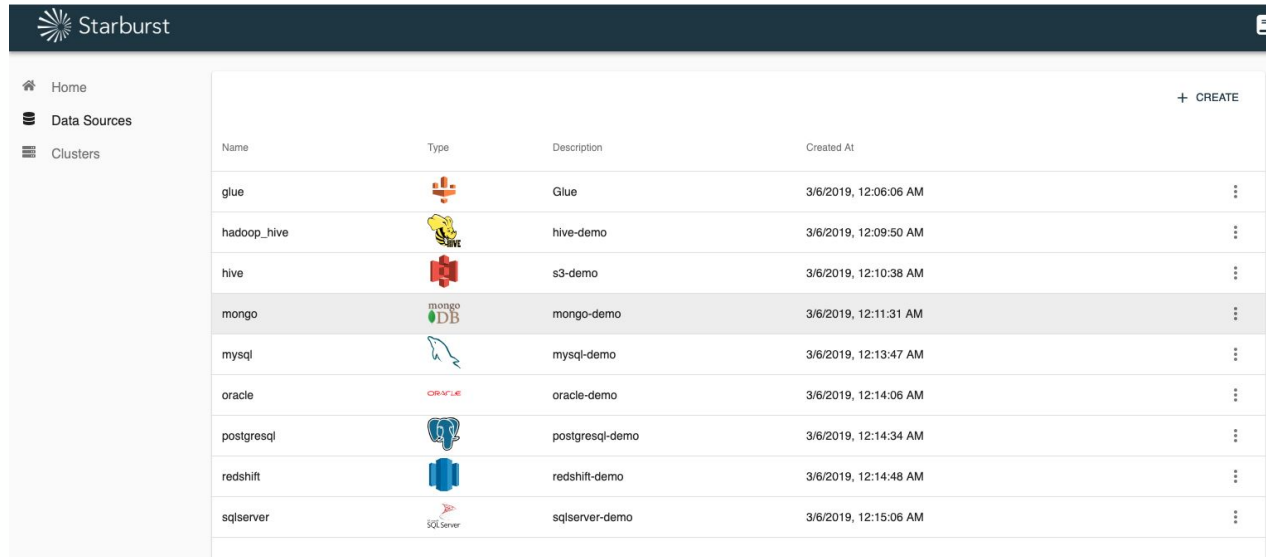
Founded by Presto committers:

- Over 4 years of contributions to Presto
- Presto distro for on-prem and cloud env
- Supporting large customers in production
- Enterprise subscription add-ons










Notable features contributed:

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

Starburst Presto & Cloud



The screenshot shows the Starburst web interface. On the left is a navigation menu with 'Home', 'Data Sources', and 'Clusters'. The main area displays a table of data sources with columns for Name, Type, Description, and Created At. A '+ CREATE' button is in the top right of the table area.

Name	Type	Description	Created At
glue		Glue	3/6/2019, 12:06:06 AM
hadoop_hive		hive-demo	3/6/2019, 12:09:50 AM
hive		s3-demo	3/6/2019, 12:10:38 AM
mongo		mongo-demo	3/6/2019, 12:11:31 AM
mysql		mysql-demo	3/6/2019, 12:13:47 AM
oracle		oracle-demo	3/6/2019, 12:14:06 AM
postgresql		postgresql-demo	3/6/2019, 12:14:34 AM
redshift		redshift-demo	3/6/2019, 12:14:48 AM
sqlserver		sqlserver-demo	3/6/2019, 12:15:06 AM

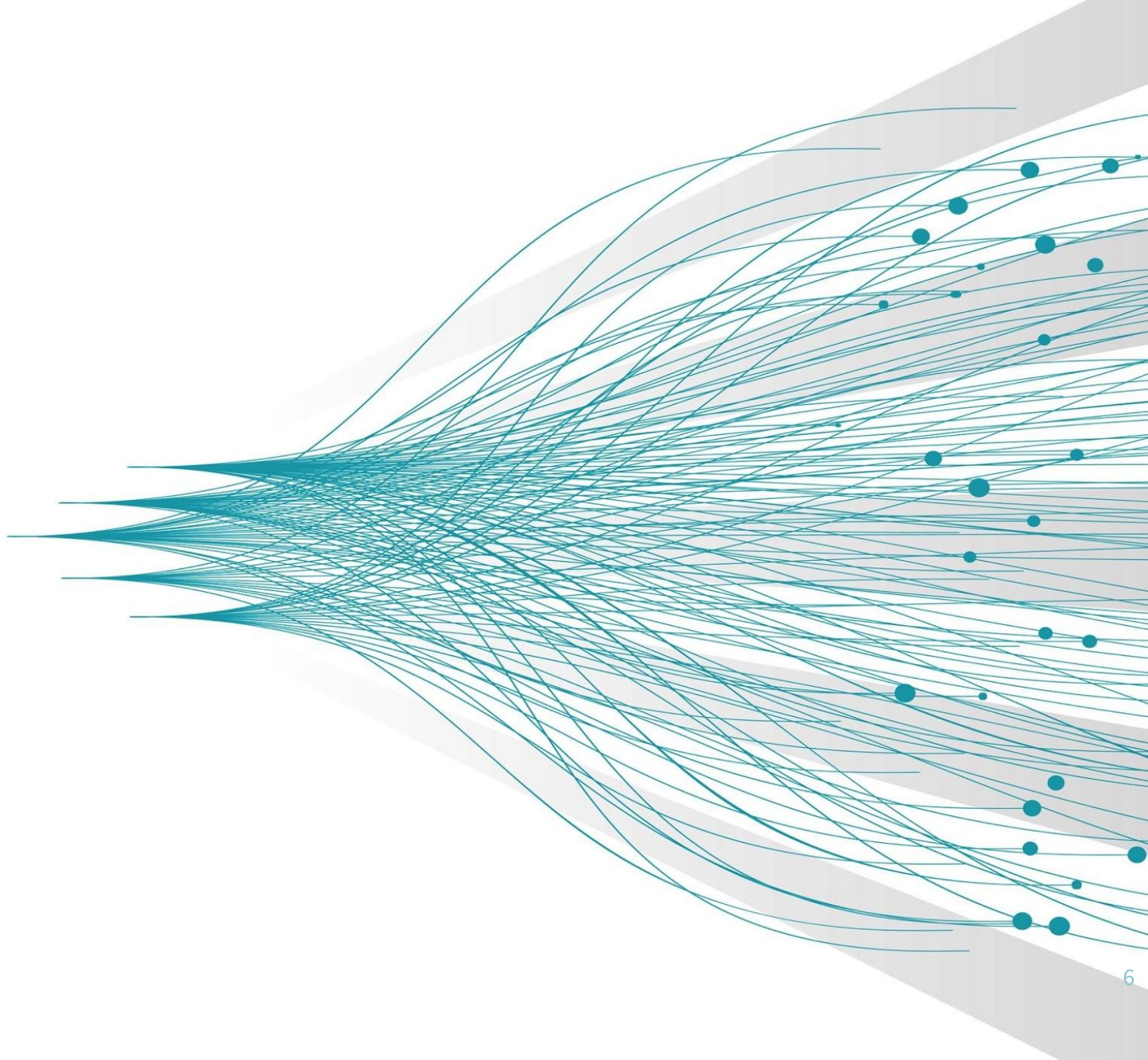


Azure HDInsight

Starburst Presto

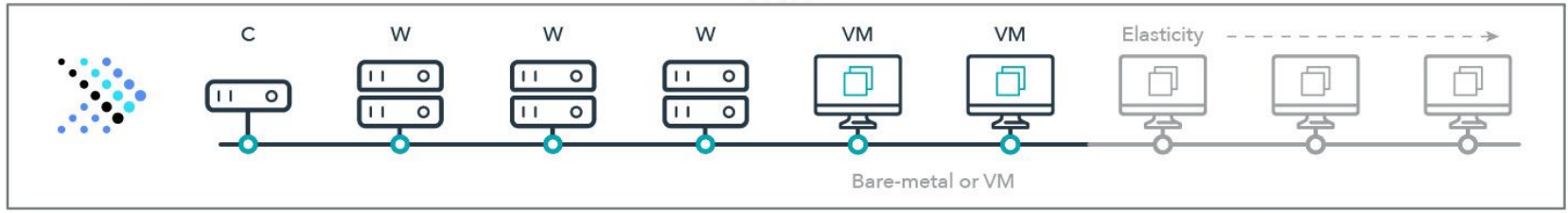
<https://www.starburstdata.com/technical-blog/announcing-starburst-enterprise-302e-with-mission-control/>

Presto





Presto Cluster



Object Storage



NoSQL Sources



RDBMS Storage



Hadoop



Project History



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

The Presto fan club



See more at <https://github.com/prestosql/presto/wiki/Presto-Users>

Presto in Production

- **Facebook:** 10,000+ of nodes, HDFS (ORC, RCFile), sharded MySQL, 1000s of users
- **Uber:** 2,000+ nodes (clusters on prem.) 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

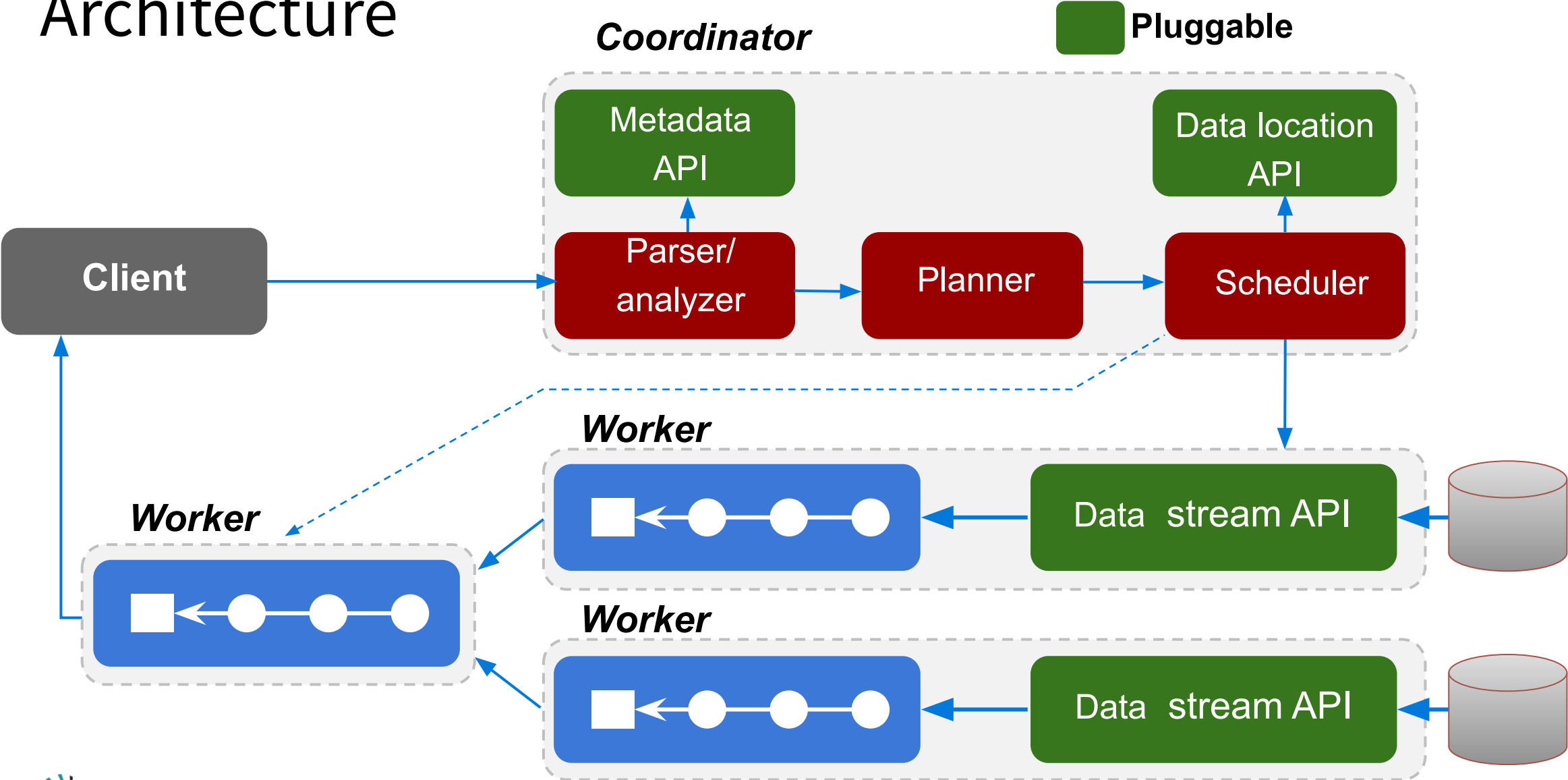
- No Hadoop distro vendor lock-in
- No storage engine vendor lock-in
- No cloud vendor lock-in

Built for Performance

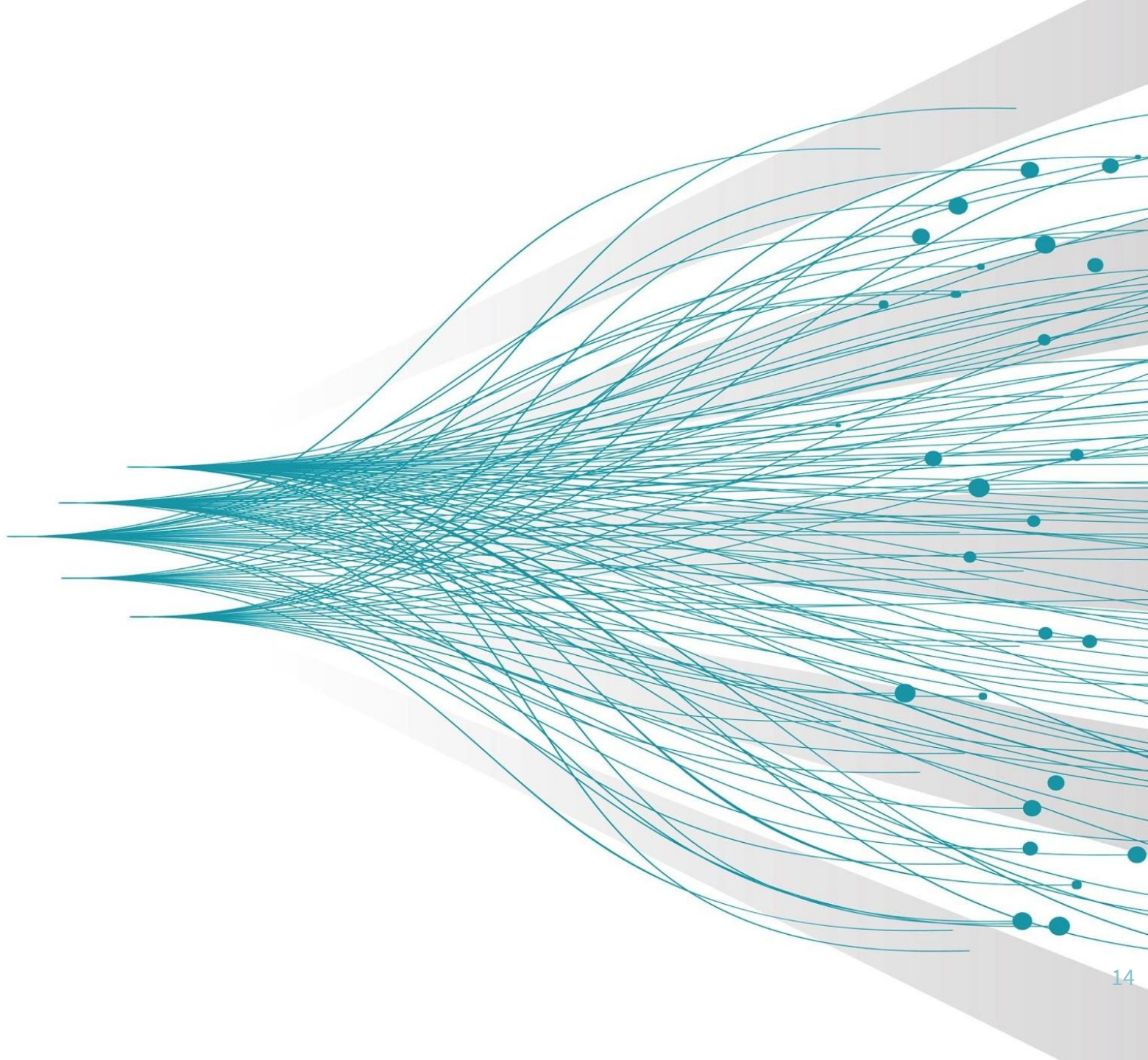
Query Execution Engine:

- MPP-style **pipelined** in-memory execution
- **Vectorized** data processing
- Runtime query **bytecode generation**
- 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**

Architecture



Presto CBO



Before we optimize ...

Join in Presto

- Hash Join
- Right table is in memory ("build table")
- Left table is streamed ("probe table")
- Can be broadcast or repartitioned

Before we optimize ...

Join in Presto

- Hash Join
- Right table is in memory ("build table")
- Left table is streamed ("probe table")
- Can be broadcast or repartitioned

- A join can be followed by a join, can be followed by a join...

CBO in a nutshell

Cost-Based Optimizer v1 includes:

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

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

Cost and Statistics

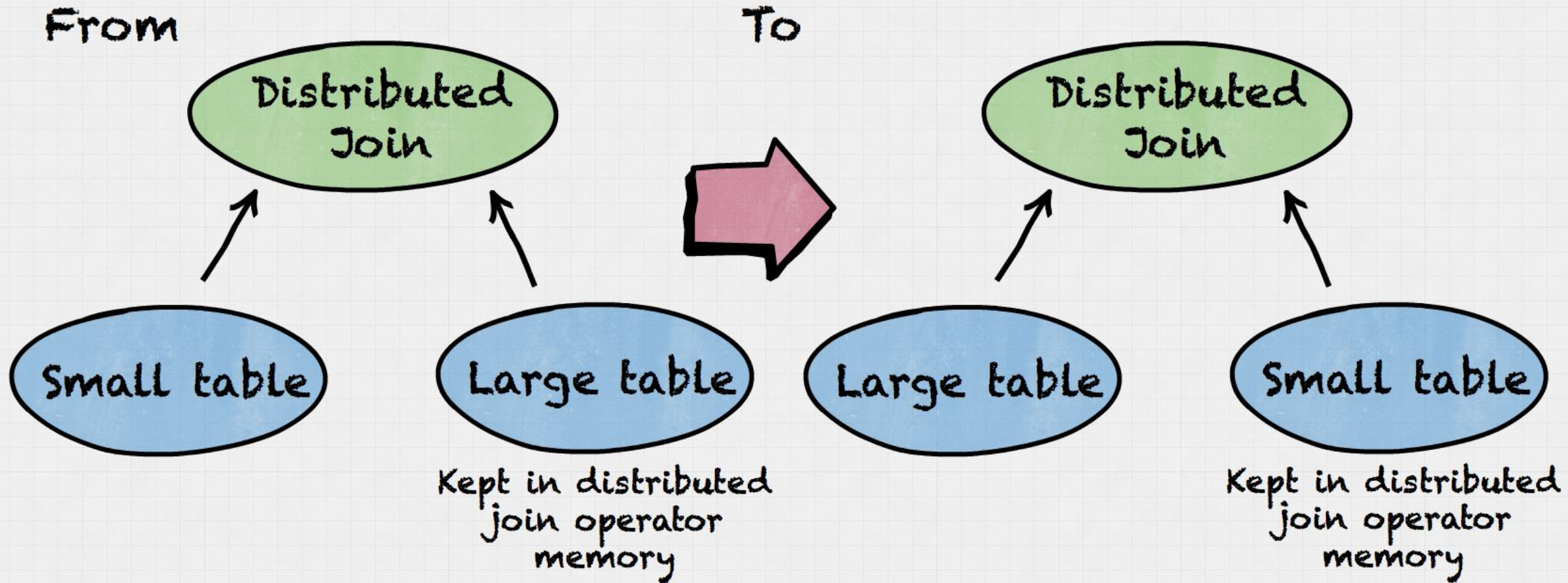
Cost calculation includes:

- CPU
- Memory usage
- Network I/O

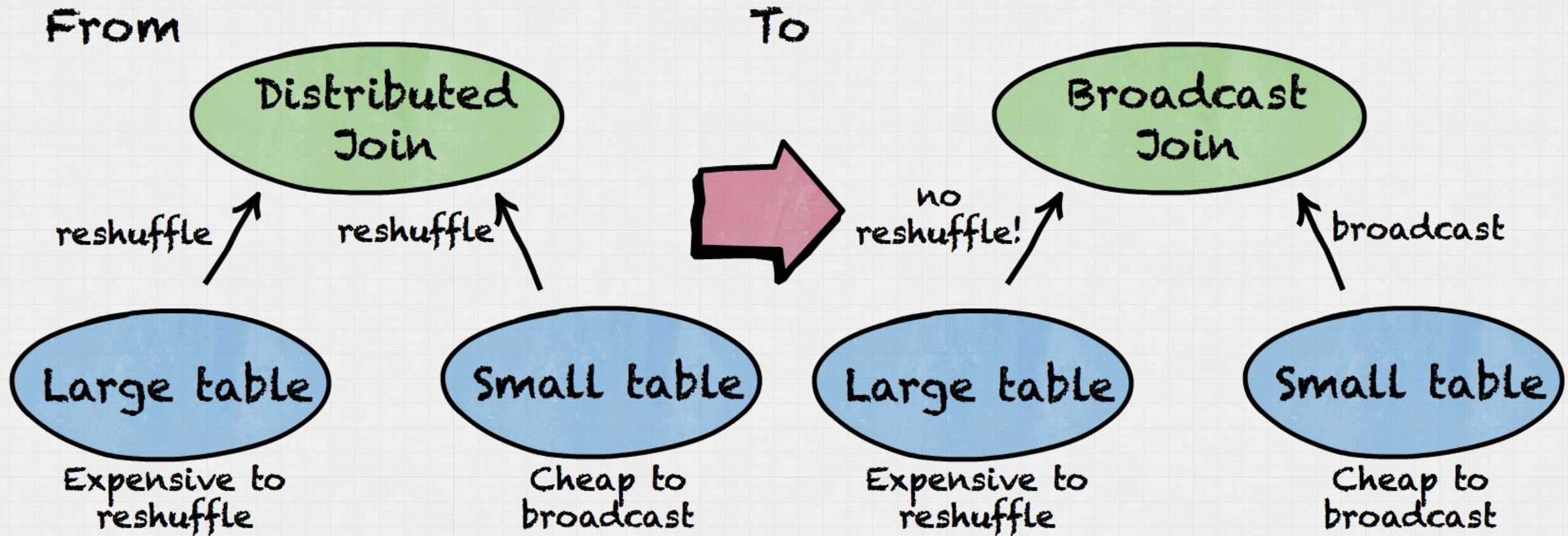
Hive Metastore statistics:

- 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

Join left/right side decision



Join type selection

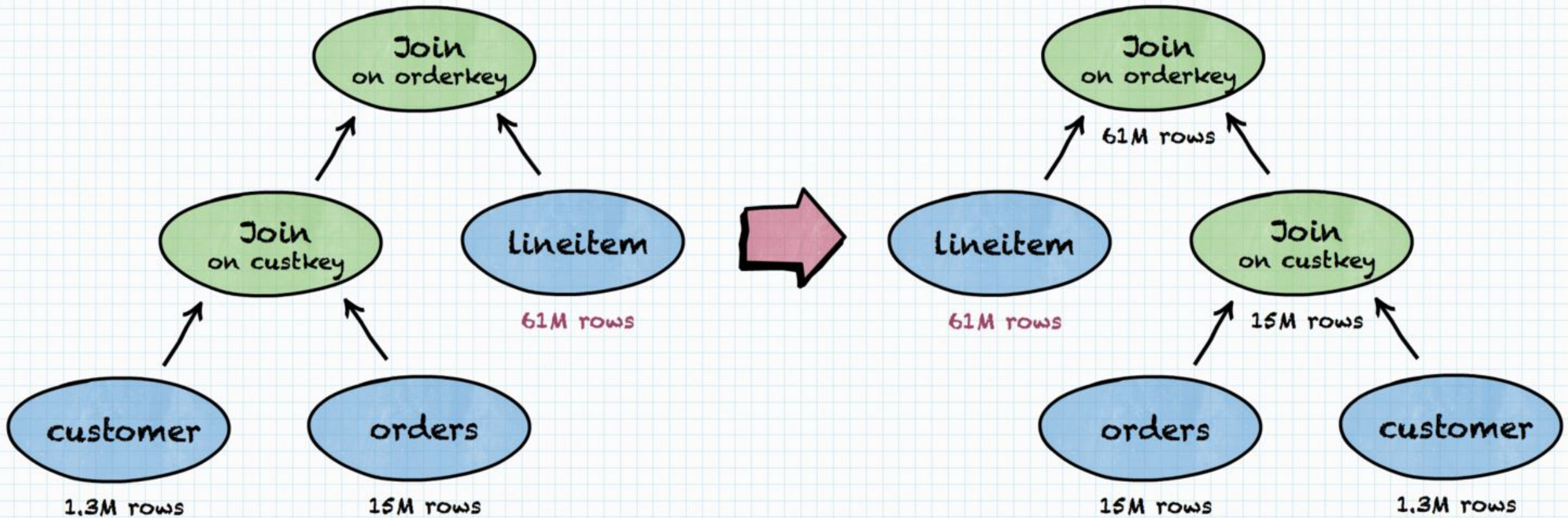


Join reordering

"Which customers are spending the most at our shop?"

```
SELECT c.custkey, sum(l.price)
FROM customer c
JOIN orders o ON c.custkey = o.custkey
JOIN lineitem l ON o.orderkey = l.orderkey
GROUP BY c.custkey
ORDER BY sum(l.price) DESC;
```

Join reordering

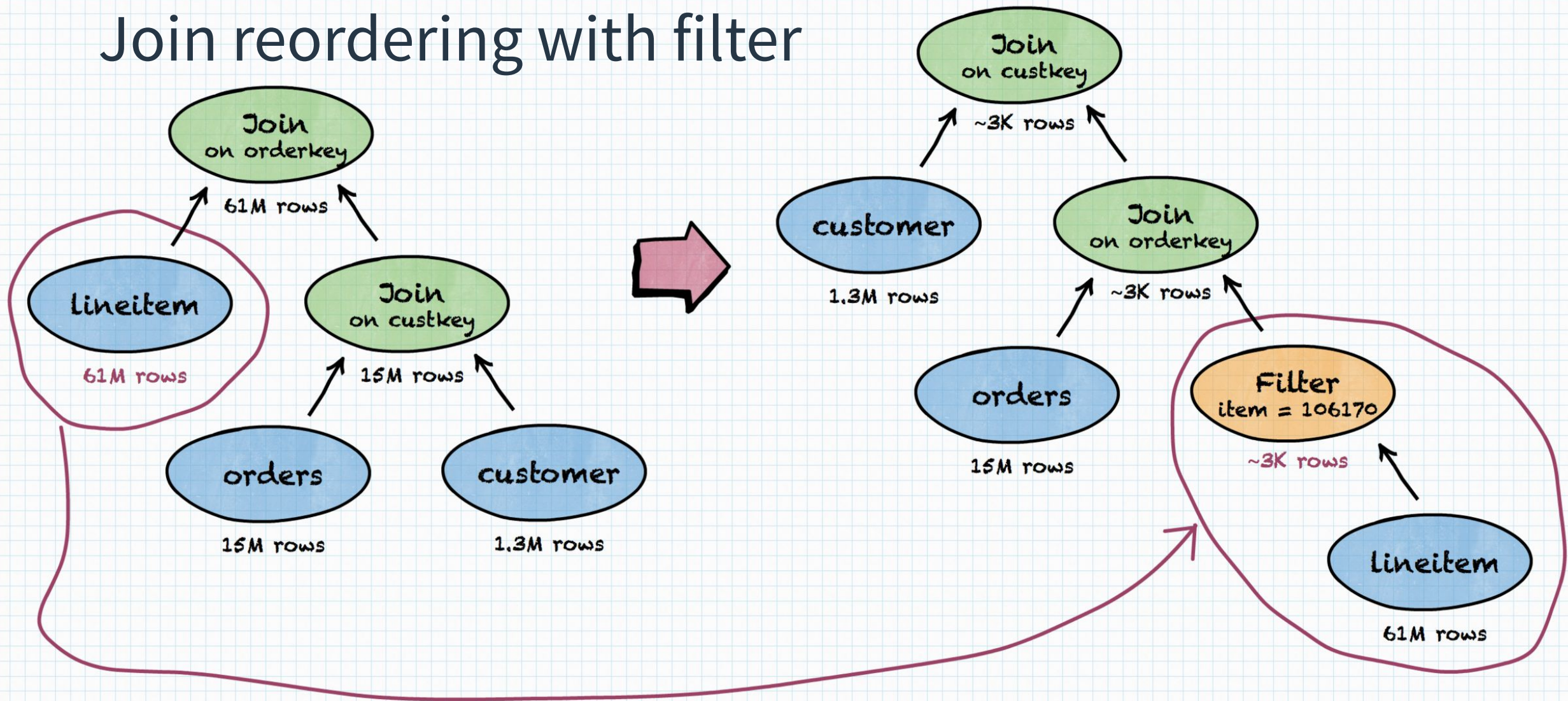


Join reordering with filter

"Which customers are spending the most on coffee?"

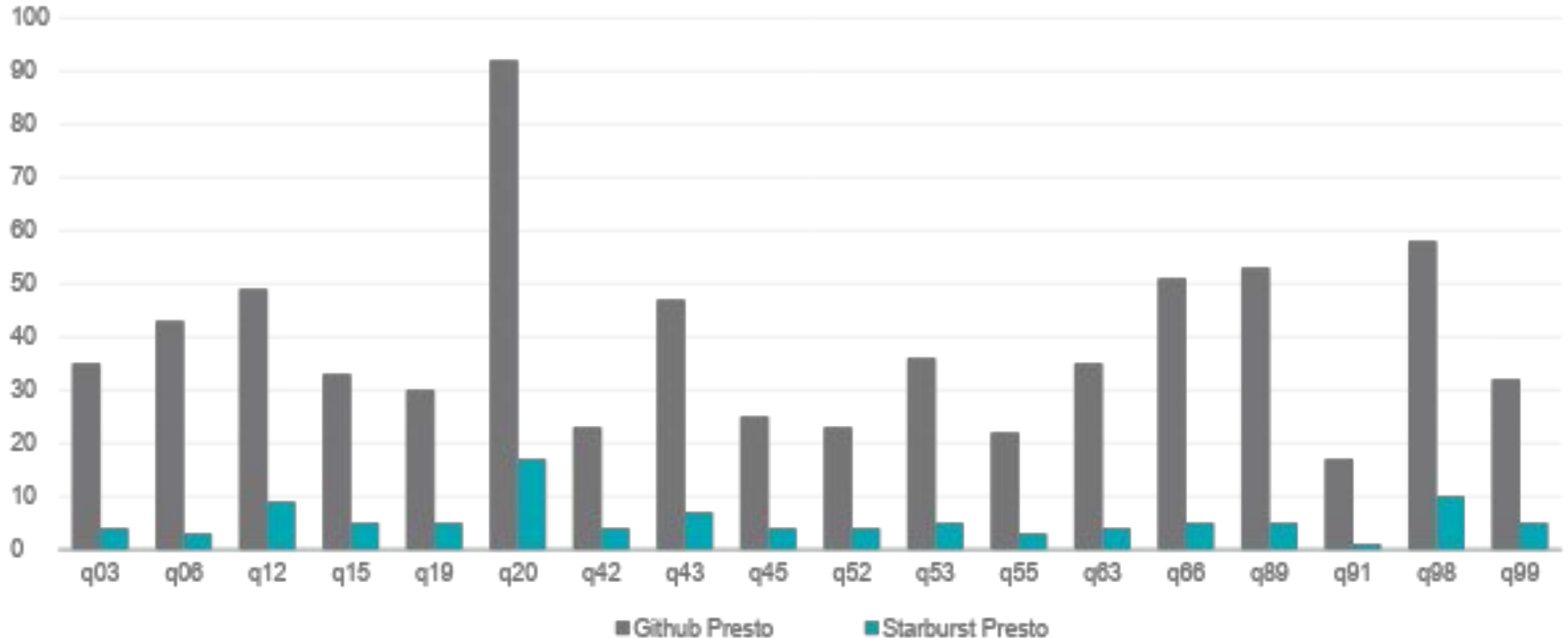
```
SELECT c.custkey, sum(l.price)
FROM customer c
JOIN orders o ON c.custkey = o.custkey
JOIN lineitem l ON o.orderkey = l.orderkey
WHERE l.item = 'coffee'
GROUP BY c.custkey
ORDER BY sum(l.price) DESC;
```

Join reordering with filter



Presto CBO Speedup

Duration of TPC-DS queries (lower is better)

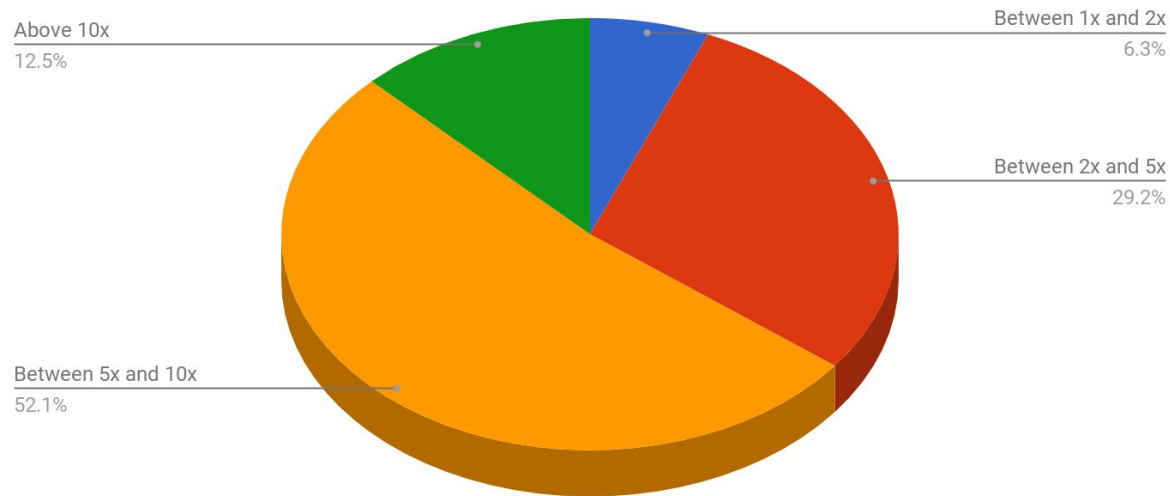


<https://www.starburstdata.com/presto-benchmarks/>

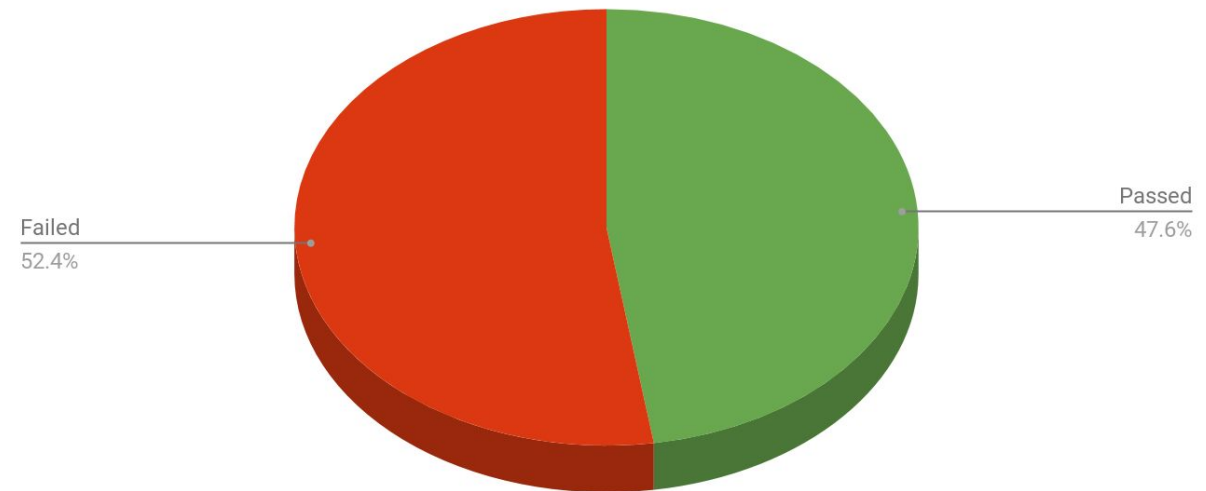
Cloud cost reduction

- on average 7x improvement vs EMR Presto (up to 18x faster!)
- EMR Presto cannot execute many TPC-DS queries (all pass on Starburst Presto)
- on average 4x improvement on simpler queries benchmark (TPC-H, 3rd party benchmark)

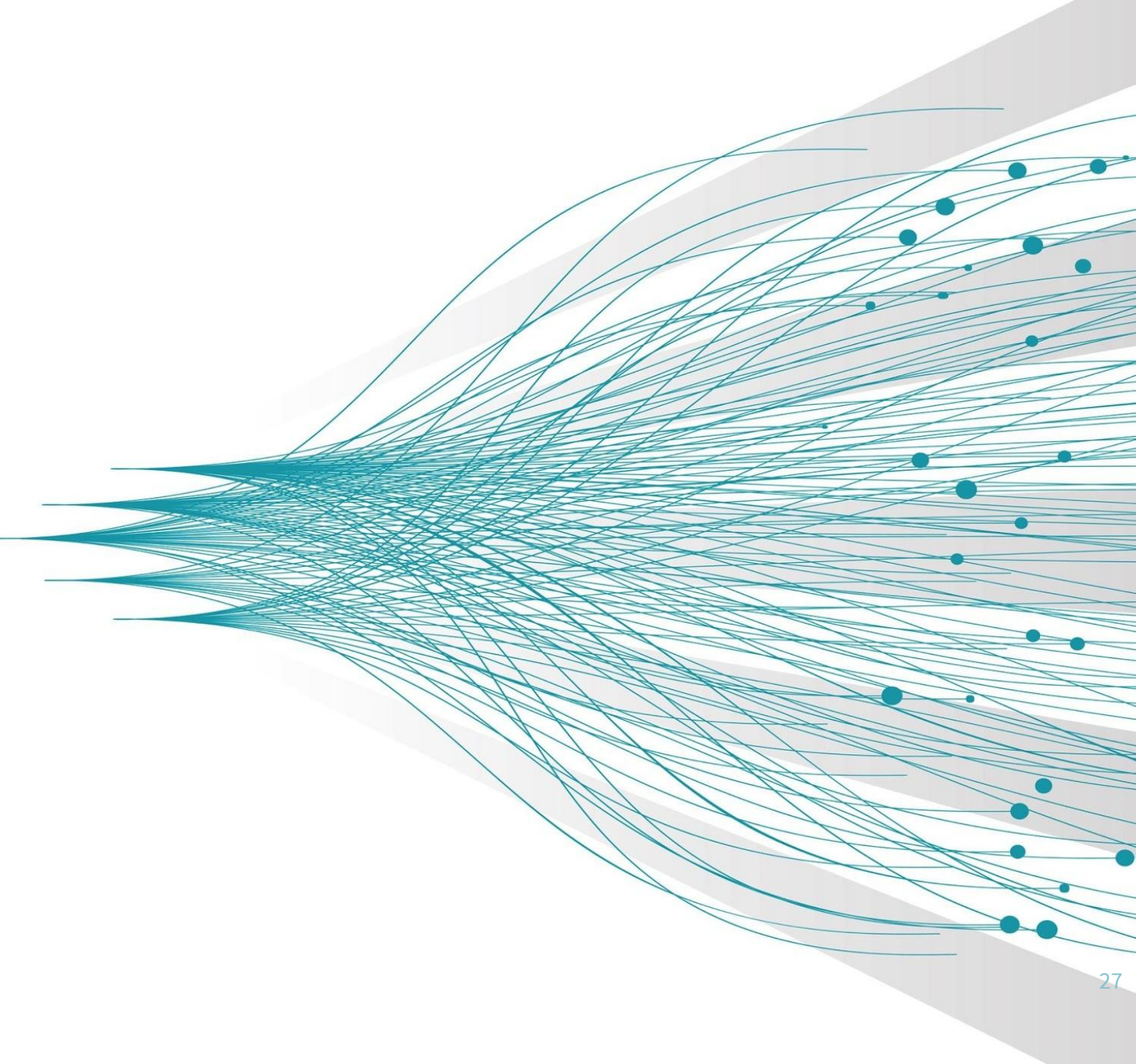
Starburst Presto (CBO) vs EMR Presto speedup



EMR Presto TPC-DS passed queries %



CBO Next Chapter



Enhancements to date

- Deciding on semi-join distribution type based on cost
 - "... WHERE x IN (SELECT y FROM ...)" queries
- Capping a broadcasted table size
- Various minor fixes in cardinality estimation
- Peak memory estimation (more meaningful memory cost metric)

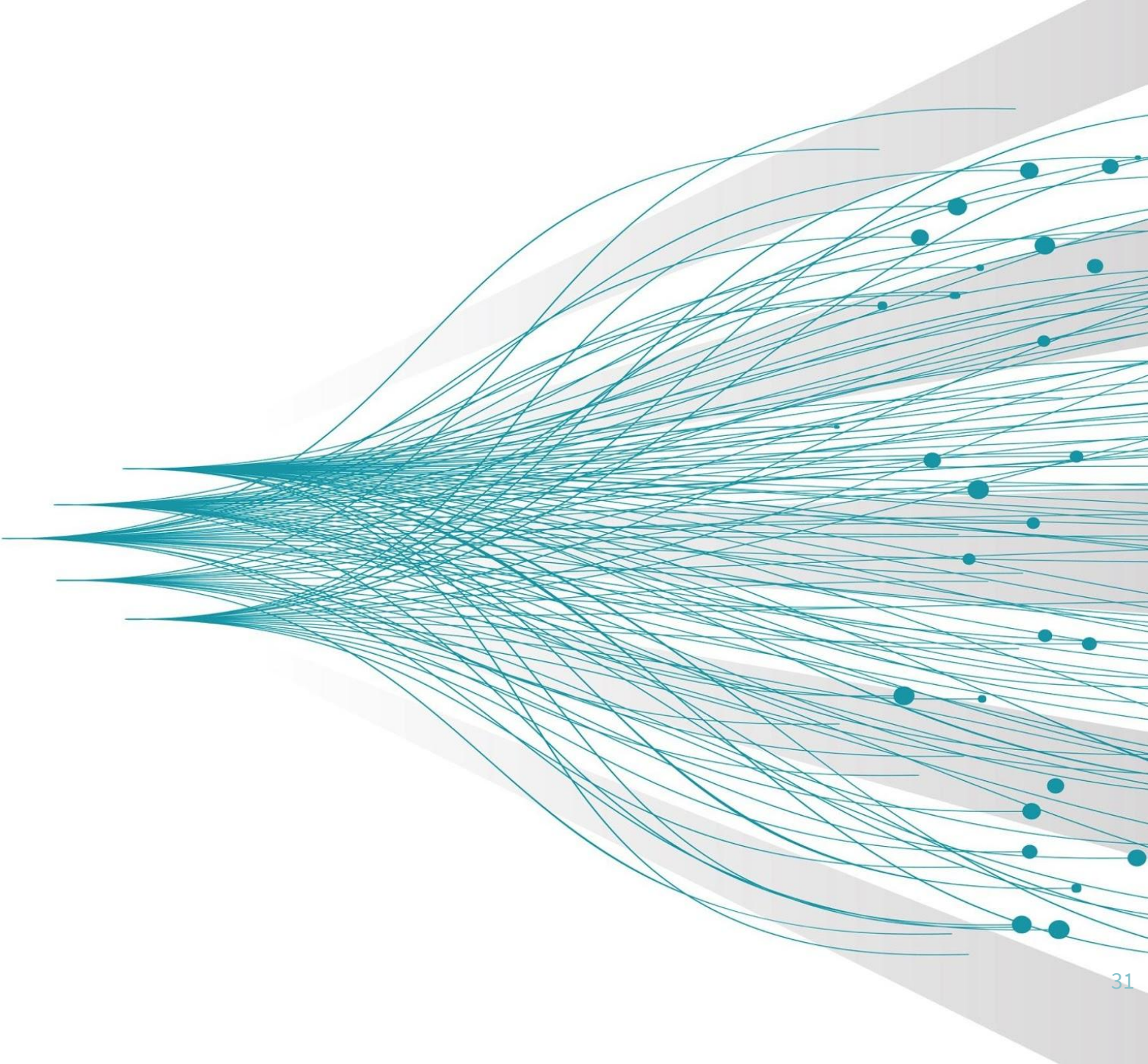
Collecting statistics

- ANALYZE table
 - native in Presto
 - Hive runtime no longer needed to collect stats
- Collecting table statistics on write
 - low overhead, the data is being processed in memory
- Stats for AWS Glue Catalog
 - Glue does not provide statistics support out of the box
 - exclusive from Starburst

Statistics in other connectors

- CBO initially supported with Hive connector only
- We added stats for RDBMS connectors (Starburst)
 - PostgreSQL
 - MySQL
 - SQL Server
 - Teradata
 - Oracle
- Enables CBO in federation use-cases

CBO Roadmap



What's next

- Stats support
 - Improved stats for Hive and Glue, e.g. NDV estimates across partitions
 - Stats for NoSQL connectors
- Core CBO and engine enhancements
 - Involve connectors in optimizations ("*-pushdown")
 - Peak memory-based decisions
 - Adjust cost model weights based on the hardware
 - Cost and stats for more operation types
 - Introduce Traits — consider alternative plans during optimizations
 - Late materialization
 - Adaptive optimizations

Further reading

<https://prestosql.io/>

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

<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-goo-gles-cloud-dataproc/>

presto



Thank You!

Twitter: @starburstdata @prestosql

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

Newsletter: www.starburstdata.com/newsletter

Rate today's session

Cyberconflict: A new era of war, sabotage, and fear See passes & pricing

David Sanger (The New York Times)
9:55am-10:10am Wednesday, March 27, 2019
Location: Ballroom
Secondary topics: Security and Privacy

Rate This Session 

We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the *New York Times* as well as a national security and political contributor for CNN and a frequent guest on *CBS This Morning*, *Face the Nation*, and many PBS shows.




Session page on conference website

✓ Attending Notes Remove

Cyberconflict: A new era of war, sabotage, and fear

🕒 9:55 AM - 10:10 AM, Wed, Mar 27, 2019

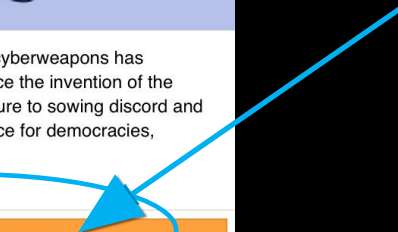
Speakers

 **David Sanger**
National Security Correspondent
The New York Times

📍 Ballroom

Keynotes

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

 **SESSION EVALUATION**

O'Reilly Events App