

# Cloudera Impala

# Impala

- Cloudera's Massively Parallel Processing (MPP) SQL query engine.
- Impala brings scalable parallel database technology to Hadoop, enabling **low latency and high concurrency** SQL queries.
  - This is something not possible with Hive, HBase, etc
- Impala is promoted for data scientists to perform analytics on data stored in Hadoop via SQL or other BI tools.
- You can get interactive queries in (near) real-time.
- It consists of different **daemon processes** that run on **specific hosts within your cluster**.

# Impala

- Impala implements a distributed architecture based on **daemon processes** that are responsible for all aspects of query execution and that run on the same machines as the rest of the Hadoop infrastructure
- Impala is the **highest performing SQL-on-Hadoop system**, especially under multi-user workloads [see attached paper for details]

# Impala Daemon

- The core Impala component is a daemon process that runs on each node of the cluster, physically represented by the `impalad` process
- **Functions:**
  - It reads and writes to data files;
  - Accepts queries transmitted from the `impala-shell` command, Hue, JDBC, or ODBC;
  - Parallelizes the queries and distributes work to other nodes in the Impala cluster;

# Impala Queries

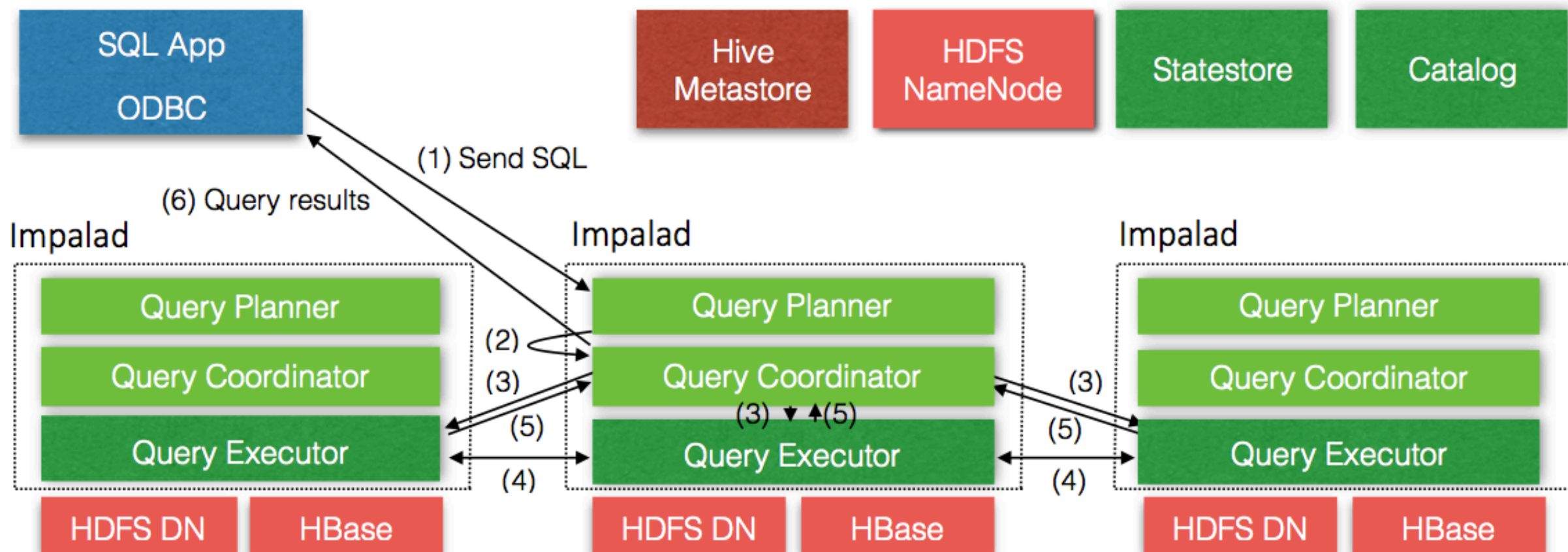
- You can submit a query to the Impala daemon running on any node, and **that node serves as the coordinator node for that query**
- The other nodes transmit partial results back to the coordinator, which constructs the final result set for a query
- The Impala daemons are in constant communication with the **statestore**, to confirm which nodes are healthy and can accept new work.

# Impala Statestore

- The Impala statestore checks on the health of Impala daemons on all the nodes in a cluster, and continuously relays its findings to each of those daemons

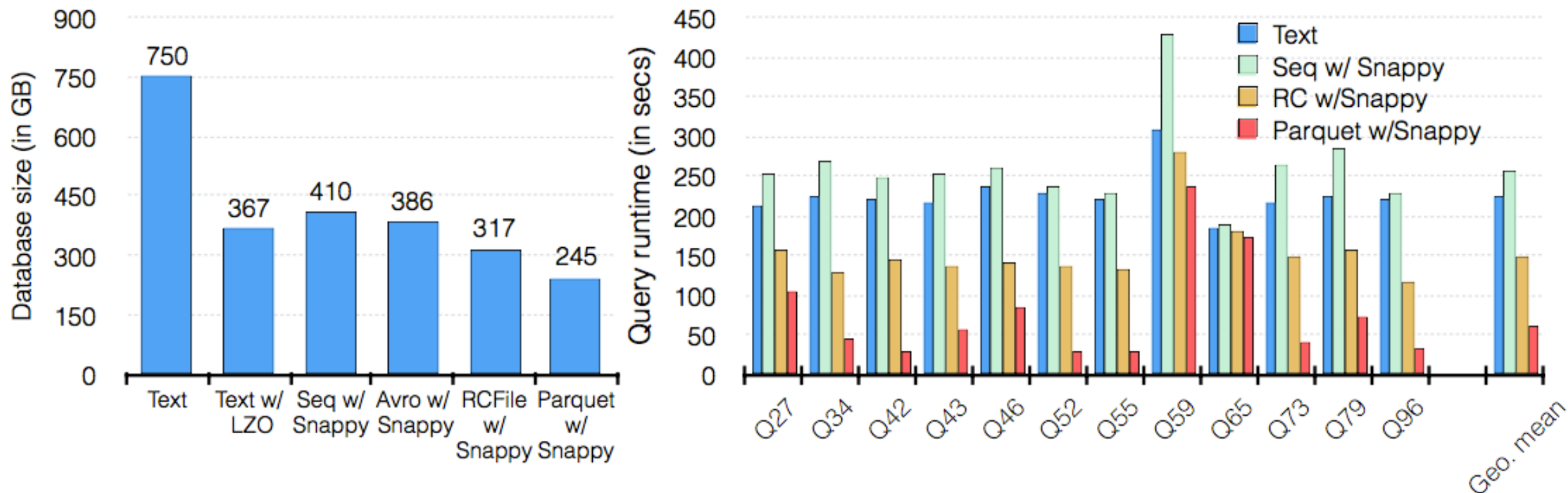
# Impala Catalog Service

- Catalog service relays the metadata changes from Impala SQL statements to all the nodes in a cluster. It is physically represented by a daemon process named catalogd; you only need such a process on one node in the cluster.



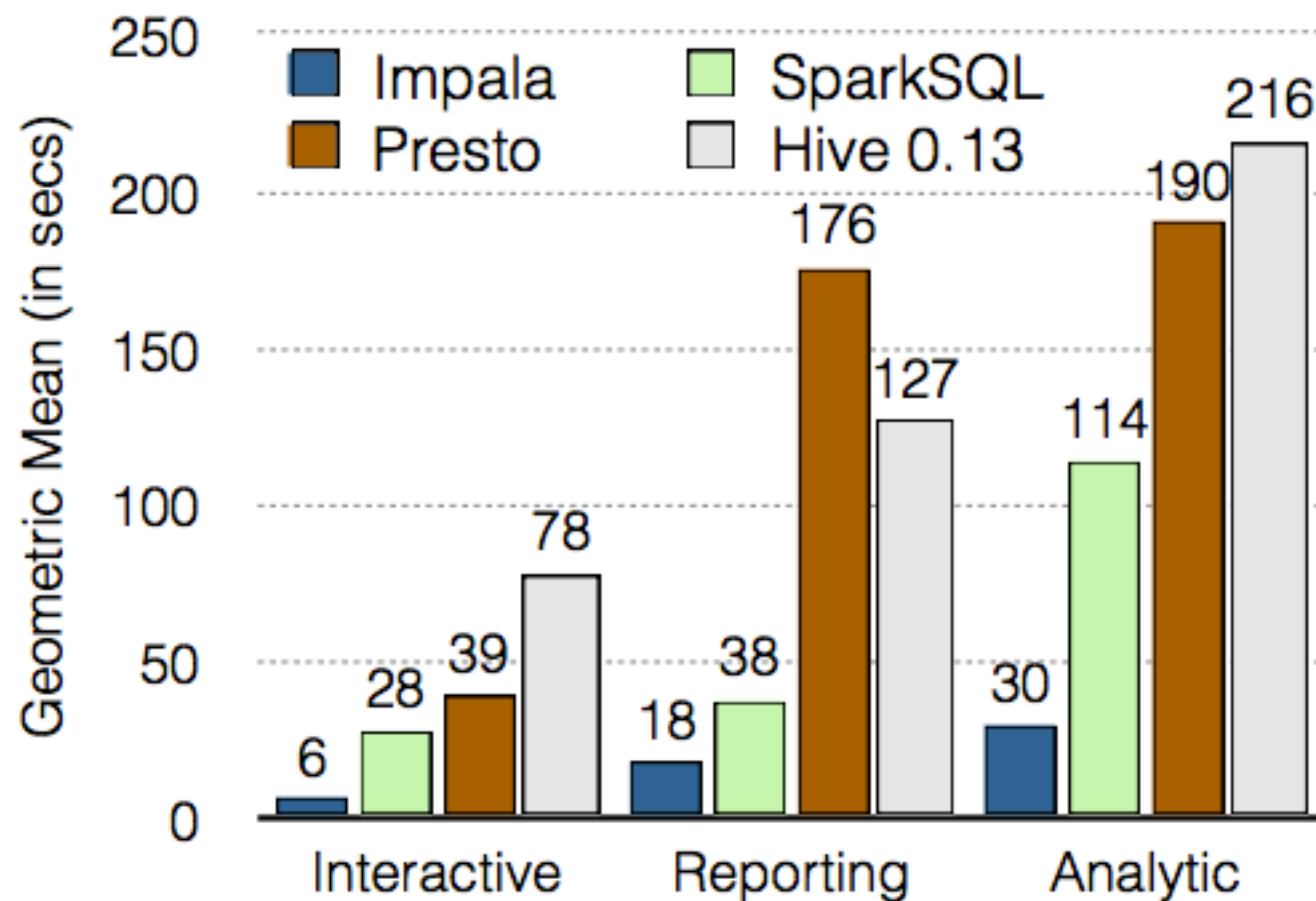


# Impala supports various file formats



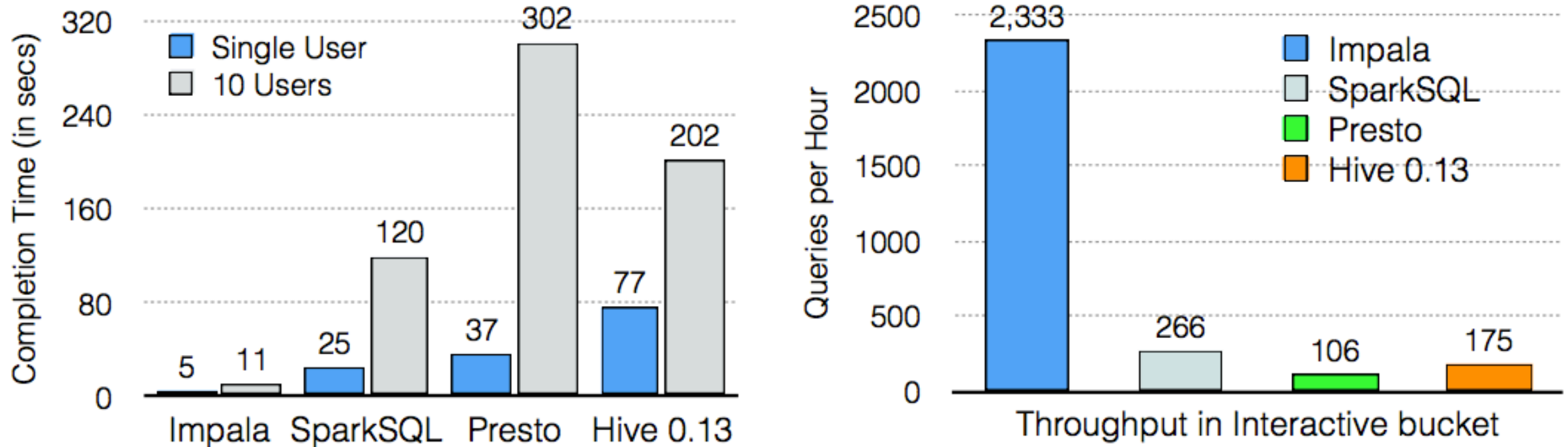
Apache Parquet is a high efficiency columnar storage format available to any project in the Hadoop ecosystem,

# How does Impala compare



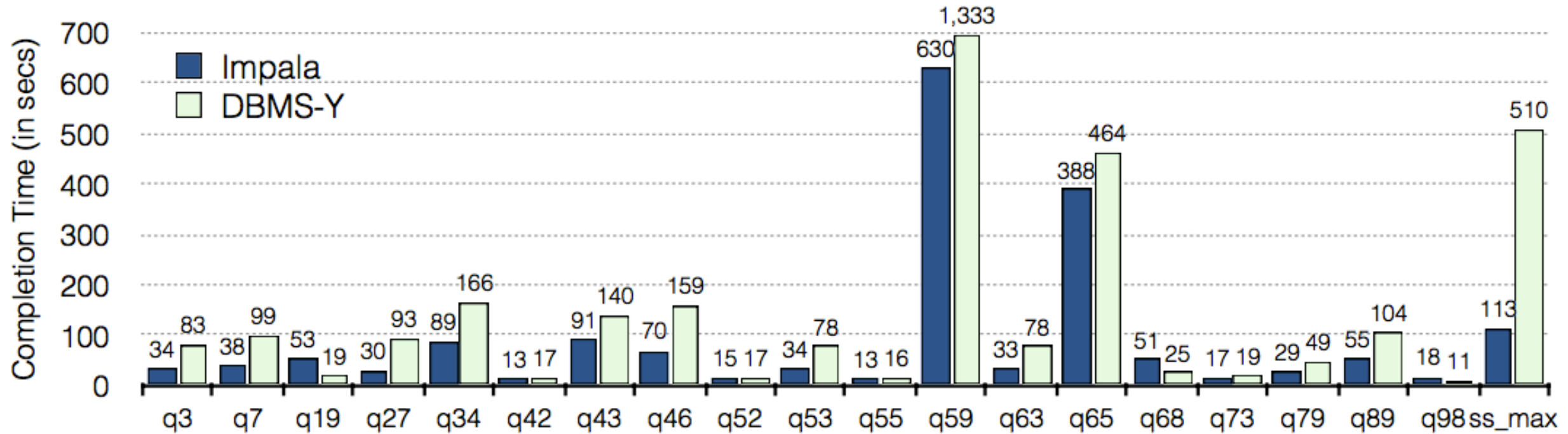
**Figure 6: Comparison of query response times on single-user runs.**

# How does Impala compare



**Figure 7: Comparison of query response times and throughput on multi-user runs.**

# How does Impala compare



**Figure 8: Comparison of the performance of Impala and a commercial analytic RDBMS.**

# What makes Impala popular

- Distributed query processing
- No single point of failure - each datanode can act as coordinator for any query.
- Low latency, fast queries
- Supports a variety of file formats (including compressed)
- It's the fastest SQL query engine – even when compared to Spark SQL

# How to get access to Impala

- Cloudera distribution (5.5)
- Docker container:  
<https://hub.docker.com/r/cloudera/impala-dev/>
- More details:  
<http://impala.io/index.html>