Performance Evaluation of Apache Tajo

Jihoon Son / Gruter Inc.

Goals

- Performance comparison with other systems
- Scalability test of Tajo

Evaluation on Cloud Environment

- Google Cloud Platform
  - Instance type: n1-standard-8
    - 8 core, 30GB RAM

TPC-DS

- Data
  - 24 tables
    - Plain text format
    - Stored on Google Cloud Storage
- Query
  - Which can be executed on every system without modifications
    - For Hive, 0.12 doesn't support implicit join, so every query had to be changed

Performance Comparison with Other Systems

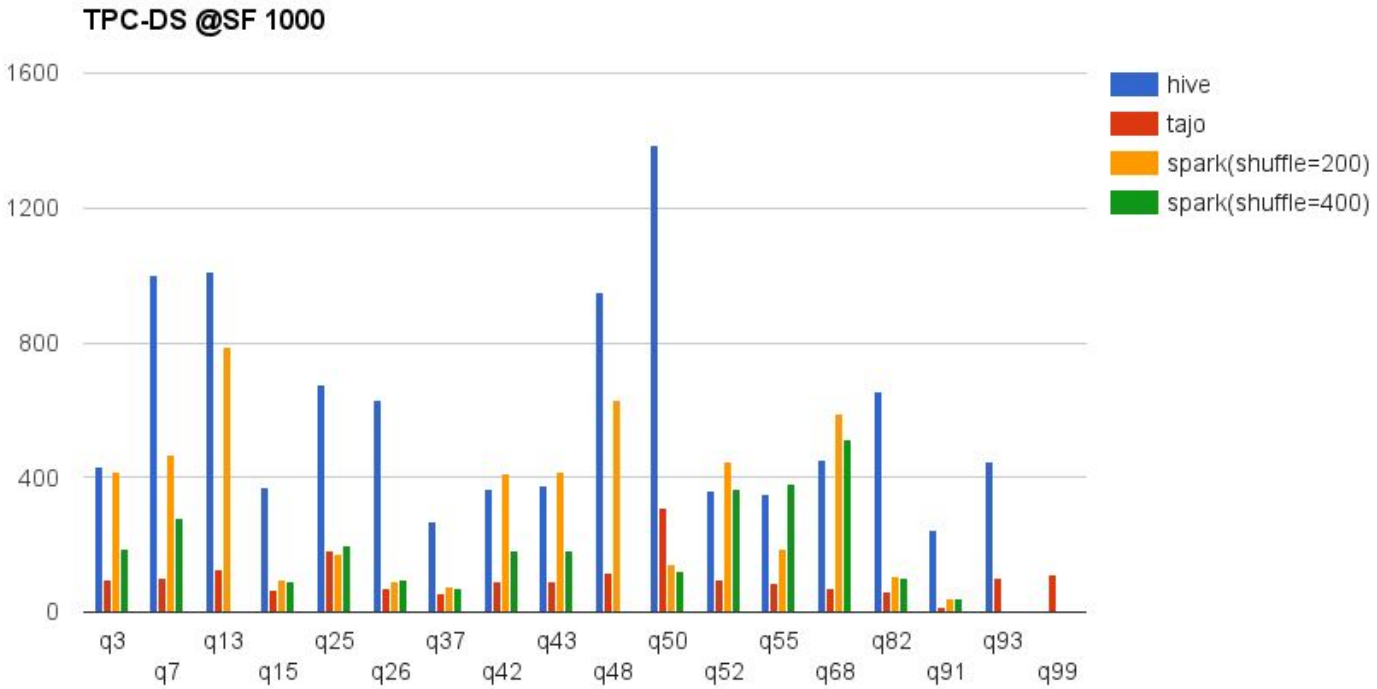
Target Systems

- Tajo (0.11.0)
  - Default configuration provided by GCP
    - Use the whole cpu and memory
- Hive (0.12)
  - Baseline performance
  - Default configuration provided by GCP
    - Use the whole cpu and memory

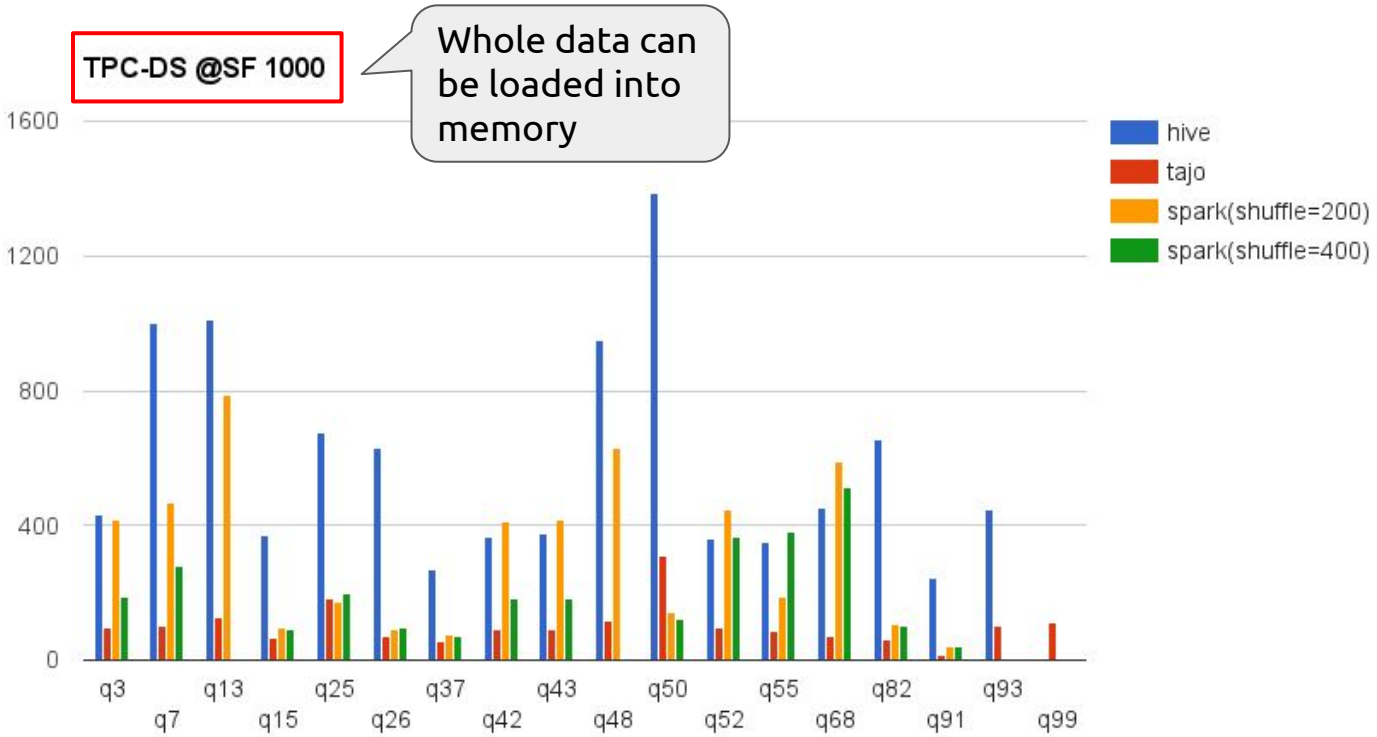
Target Systems

- Spark-SQL (1.5.0)
  - Default configuration provided by GCP
    - Use the whole cpu and memory
    - Tungsten enabled by default
  - `spark.sql.shuffle.partitions` is adjusted for better performance

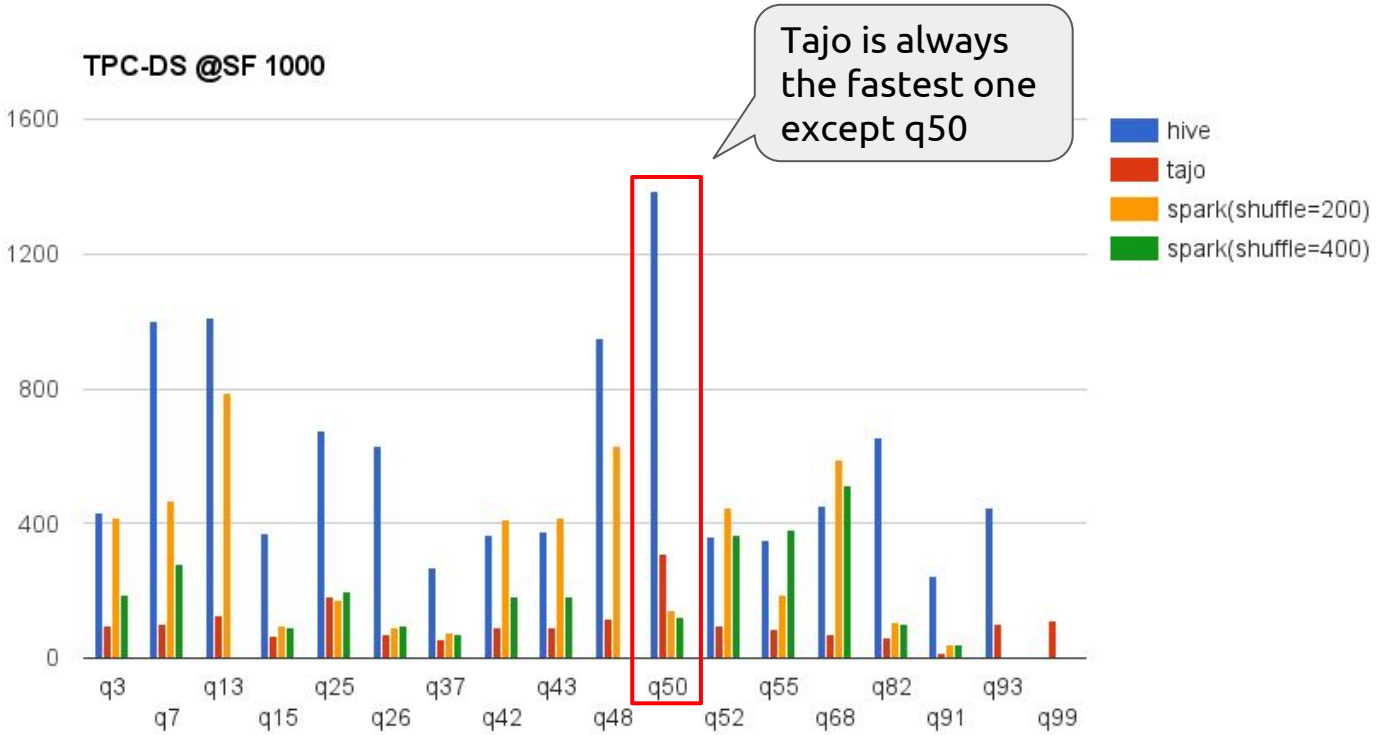
SF 1000, 50 instances



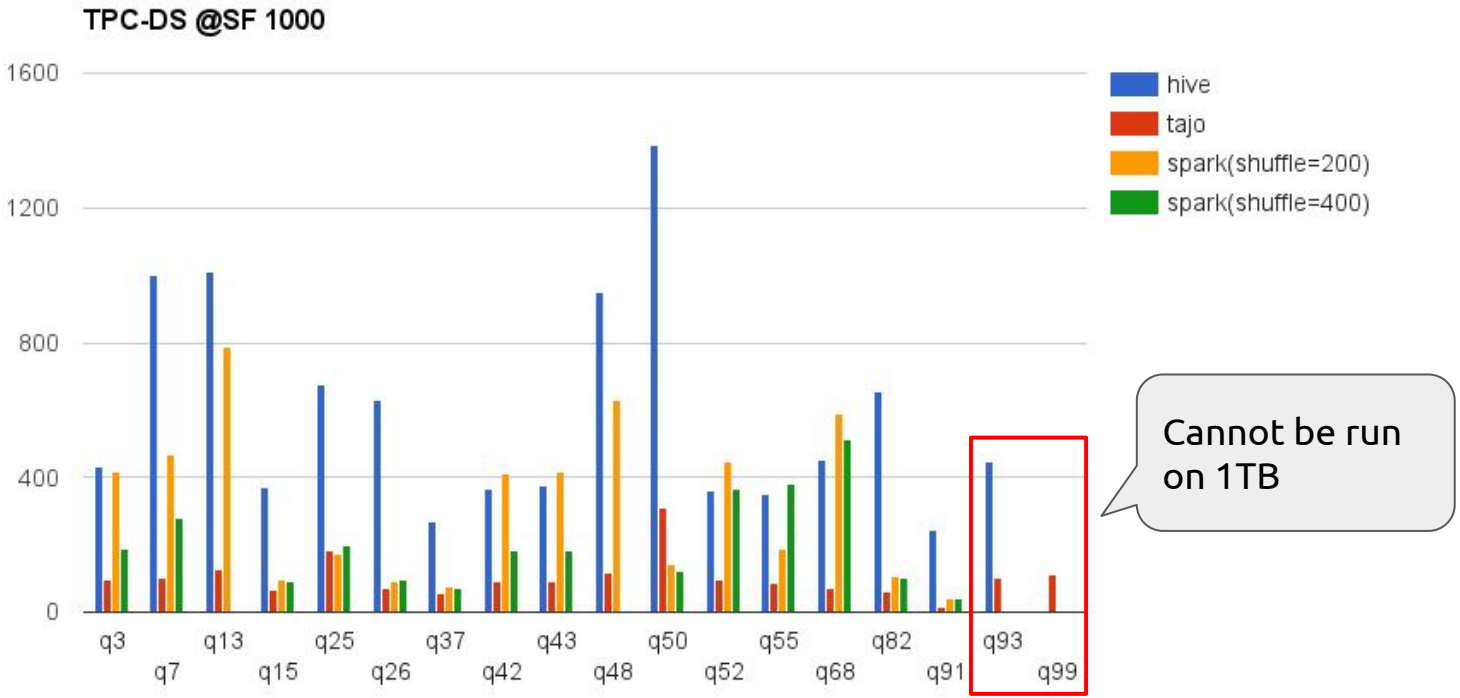
SF 1000, 50 instances



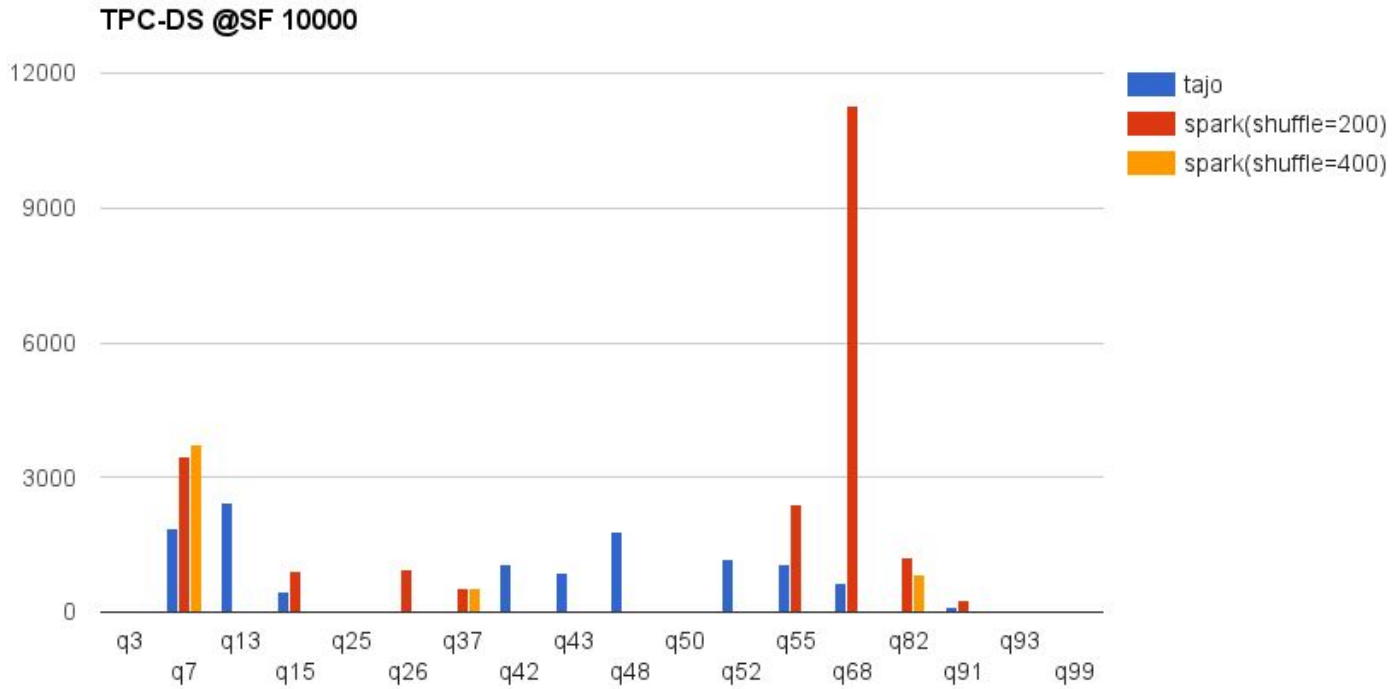
SF 1000, 50 instances



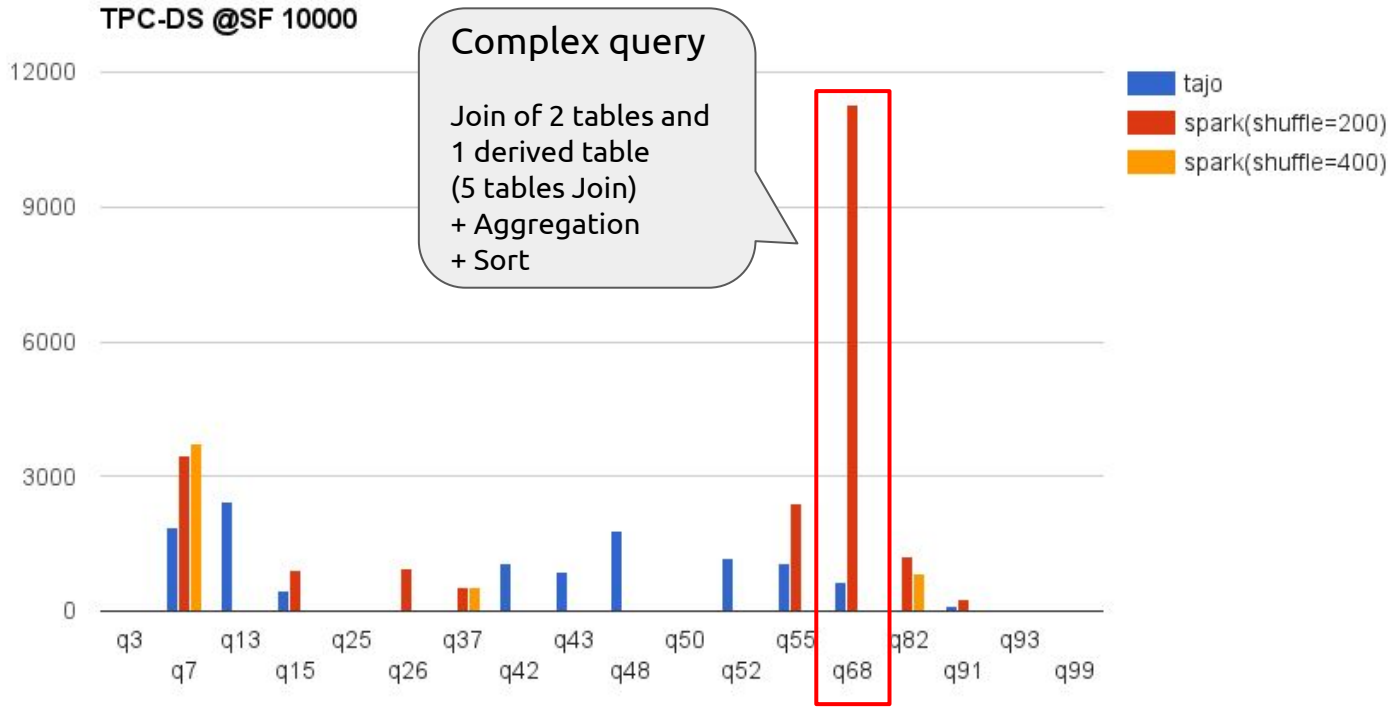
SF 1000, 50 instances



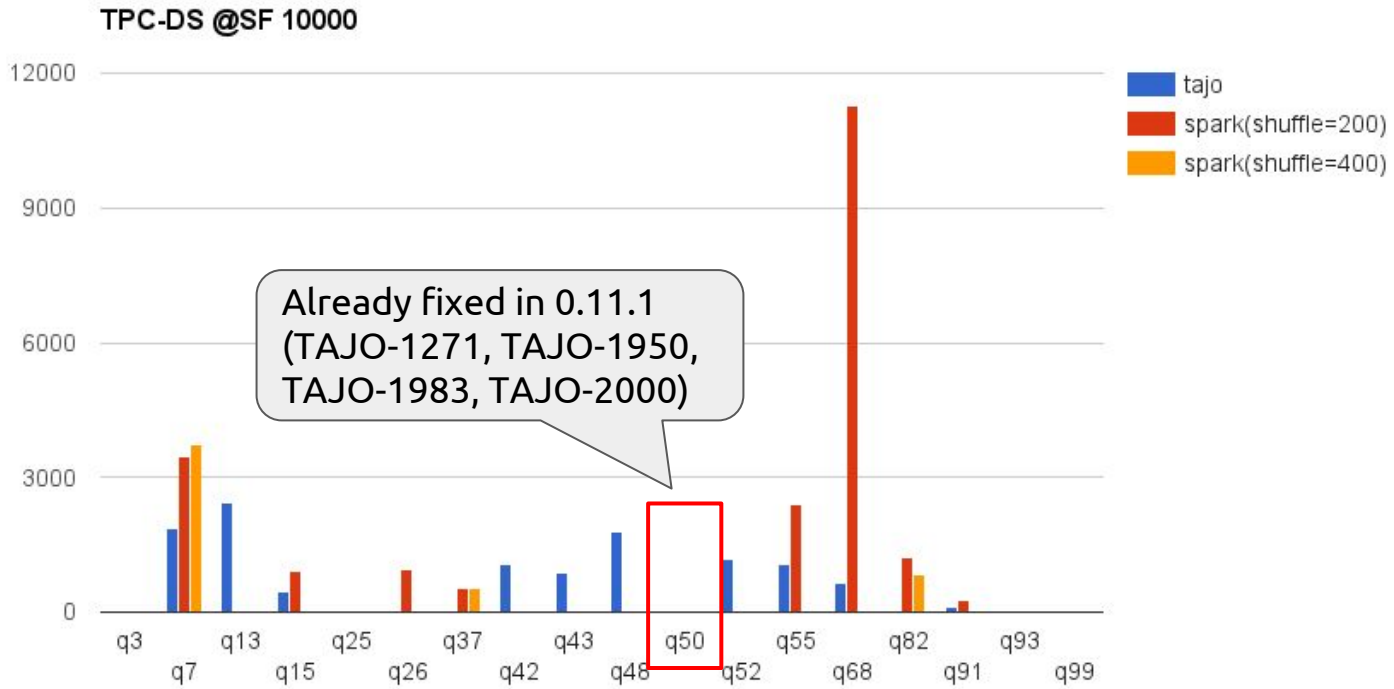
SF 10000, 50 instances



SF 10000, 50 instances

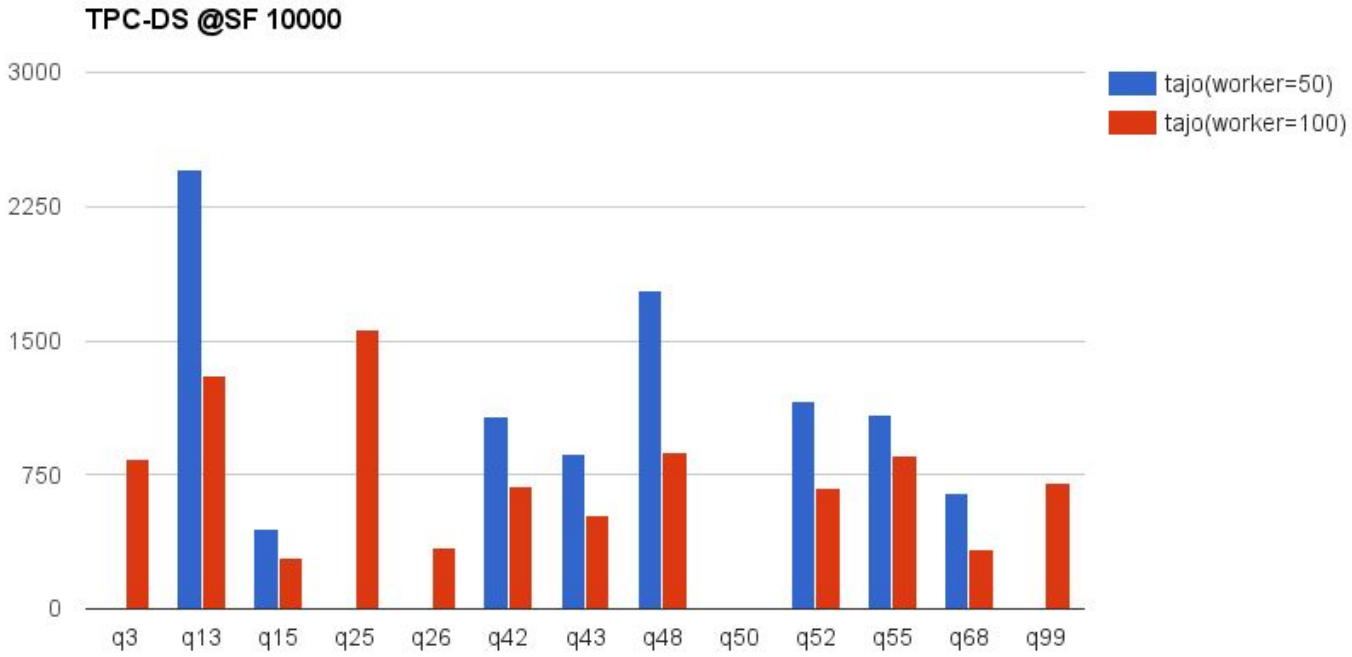


SF 10000, 50 instances

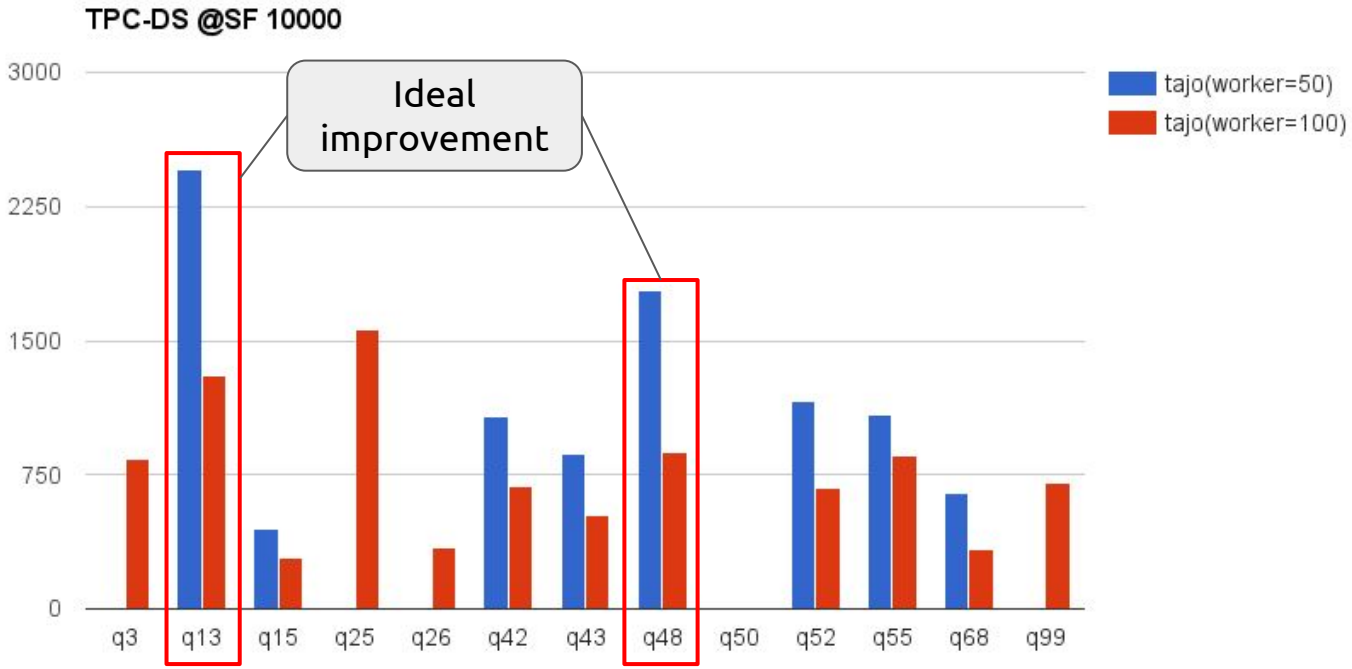


Scalability Test

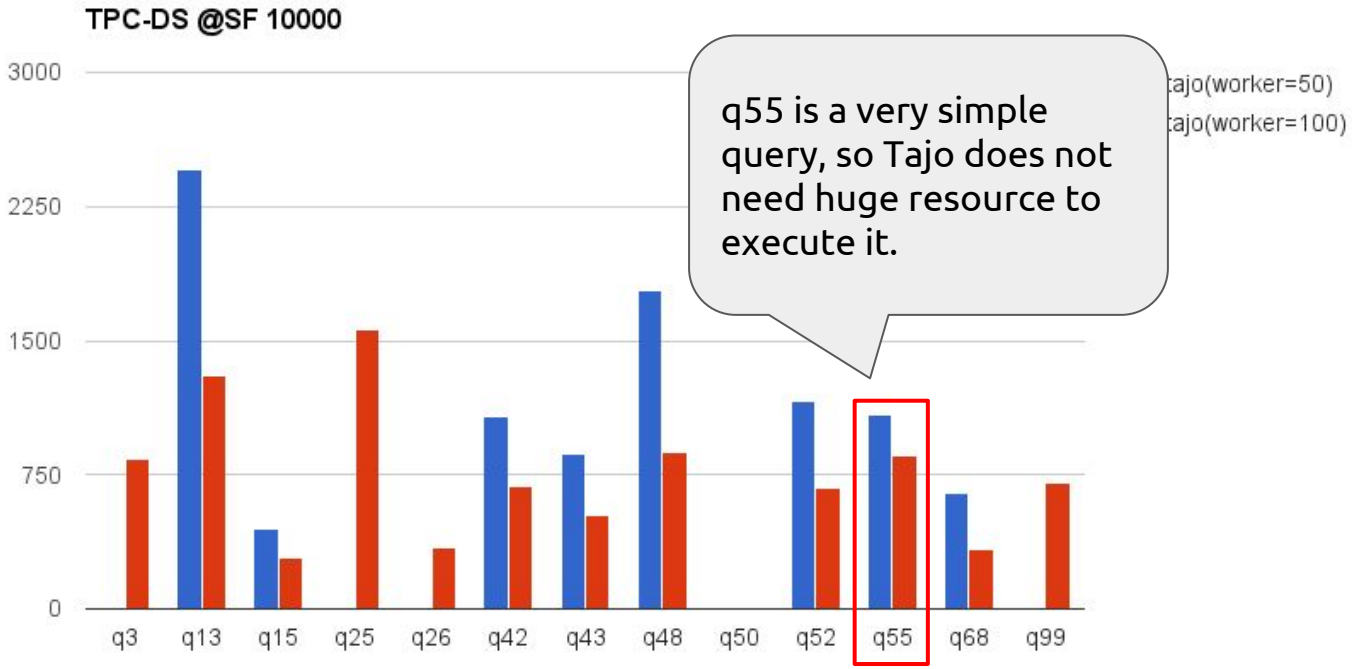
SF 10000


TPC-DS @SF 10000
3000
2250
1500
750
0
q3
q13
q15
q25
q26
q42
q43
q48
q50
q52
q55
q68
q99
tajo(worker=50)
tajo(worker=100)

SF 10000


TPC-DS @SF 10000
Ideal improvement
tajo(worker=50)
tajo(worker=100)
3000
2250
1500
750
0
q3
q13
q15
q25
q26
q42
q43
q48
q50
q52
q55
q68
q99

SF 10000


TPC-DS @SF 10000
3000
2250
1500
750
0
q3
q13
q15
q25
q26
q42
q43
q48
q50
q52
q55
q68
q99
ajo(worker=50)
ajo(worker=100)
q55 is a very simple query, so Tajo does not need huge resource to execute it.

Get Involved!

- We are recruiting contributors!
- General
  - http://tajo.apache.org/
- Getting Started
  - http://tajo.apache.org/docs/current/getting_started.html
- Downloads
  - http://tajo.apache.org/downloads.html
- Issue tracker
  - http://issues.apache.org/jira/browse/TAJO
- Join the mailing list
  - dev-subscribe@tajo.apache.org
  - issues-subscribe@tajo.apache.org

Q & A