# BDAS: Comparison of Big Data Analytics Systems: Redshift & Hive

Prashant Ratnaparkhi, Insight Data Engineering Fellowship, New York

# Hive setup: 2 AWS clusters each with:


- 1 Namenode & 5 Datanodes
- Each node a r3.Large Instance
- Each instance with 16GB memory & either 100 (500) GB SSD storage

Note: First cluster with Hive system (HDFS+local metastore) with 50GB data & Second Hive cluster with 500GB data

# Redshift setup: One Amazon Redshift cluster with

- 1 Leader node & 5 worker nodes
- Each node dc1.large instance
- Each instance with 15GiB memory & 160GB SSD storage

Note: Two Redshift databases are created on this cluster. First database is populated with cluster with 50GB data & with 500GB data. The dataset is identical with the respective Hive clusters.

# Schema & table size summary

- Schema & Queries are based on TPC Big data Benchmark (TPC-BB)
- Schema consists of 23 tables representing a data warehouse of a large retailer

| Table | # Records [Total db size 50 GB] | # Records [Total db size 500 GB] |
|---|---|---|
| customer | 700,036 | 2,213,708 |
| item | 126,007 | 398,468 |
| web_sales | 50,606,188 | 595,629,807 |
| store_sales | 50,601,586 | 595,615,096 |
| web_clickstreams | 512,905,355 | 6,035,484,722 |
| … | | |

# Two Sample Queries

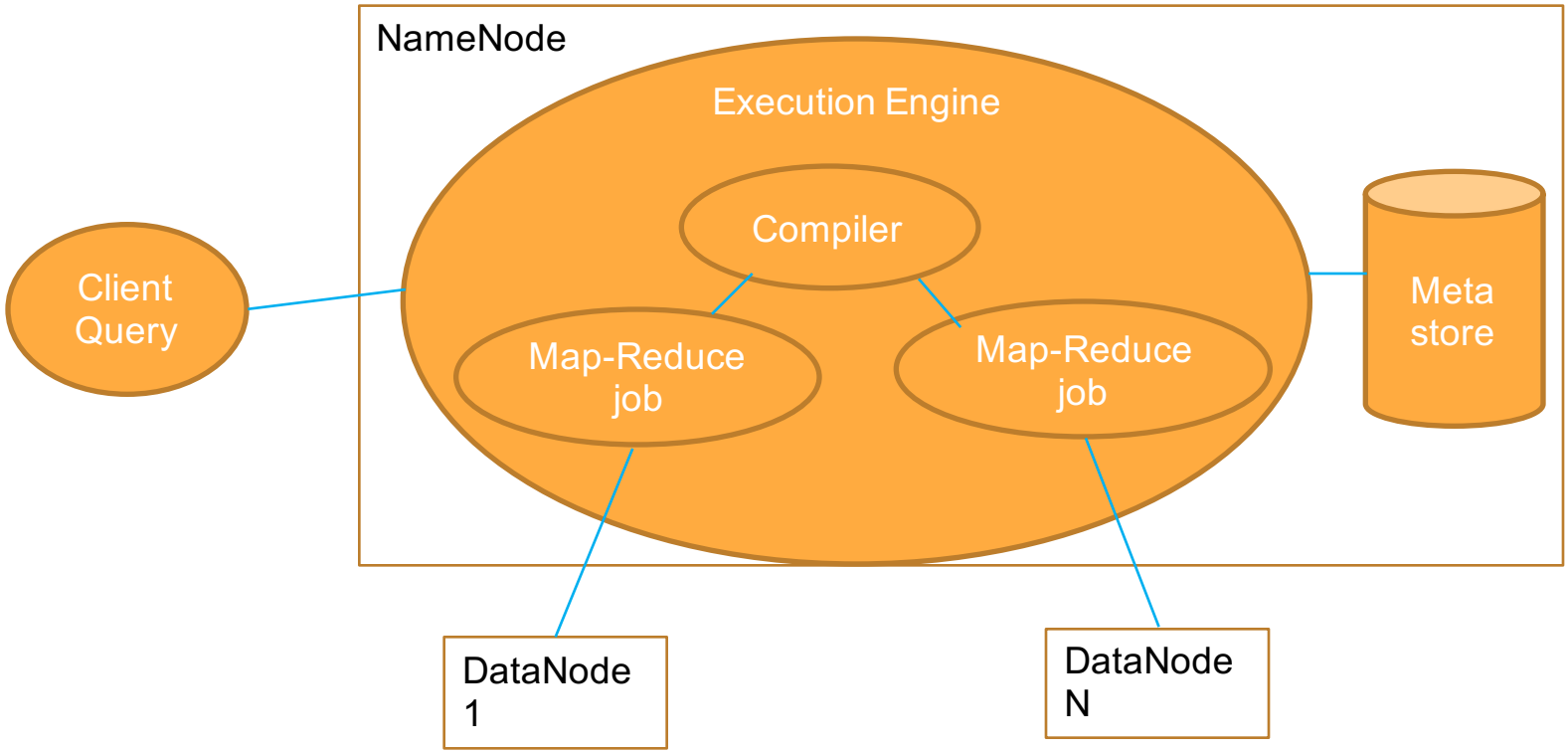
| # | Query | Operations |
|---|---|---|
| 1 | Customer-purchase behavior: List customers, who view online, followed by an in-store purchase | Multiple joins (web_clickstreams, items and store, items), order by and distinct |
| 2 | Market basket analysis: List the items sold frequently together, in certain stores, categories | Multiple joins (including one self join on store_sales and count, group by, having and order by clauses |

# Summary of Results of selected queries

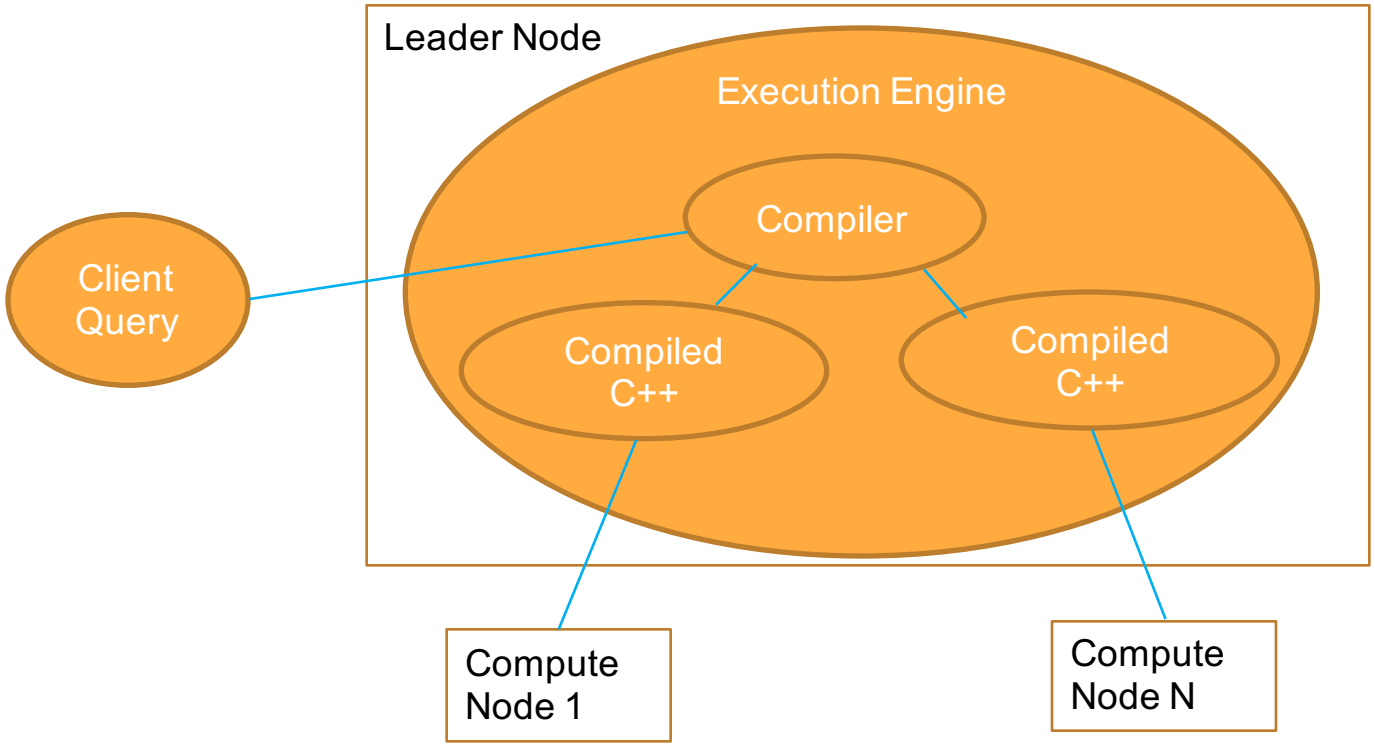
Task execution time is listed in the table below

| Task | Hive (seconds) | RedShift (seconds) | Dataset (GB) |
|---|---|---|---|
| Query1 | 288 | 2.00 | 50 GB |
| Query1 | 2392 | 19.18 | 500 GB |
| Query2 | 376 | 4.00 | 50 GB |
| Query2 | 1689 | 22.52 | 500 GB |

# Hive Query Processing



# Amazon Redshift Query Processing



# Cost of this benchmark execution

| # | Item | Duration in hours | Price node per hour in USD | Cost in USD |
|---|---|---|---|---|
| 1 | Redshift cluster: 6 dc1.large nodes (1 leader, 5 compute) | 408 | $0.250 | $102.00 |
| 2 | Hive Cluster: 12 m3.large nodes (2 master, 6 data) | 384 | $0.133 | $51.07 |
| 3 | Hive Cluster: 12 r3.large nodes (2 master, 6 data) | 96 | $0.166 | $15.94 |
| Total cost of Hive clusters | | | | $67.01 |
| Total cost of Redshift cluster | | | | $102.00 |
| Grand Total | | | | $169.01 |

# Prashant Ratnaparkhi: Technologist with experience in software product development

- Master (Systems & Control Engineering), IIT, Mumbai, India
- Data Scientist Certification from Johns Hopkins/Coursera
- Enjoy running, skiing, swimming biking & Ultimate Frisbee