Hive Queries Optimization Based on Acemap

Shiyuan Zhan

Database Group

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

1 Introduction

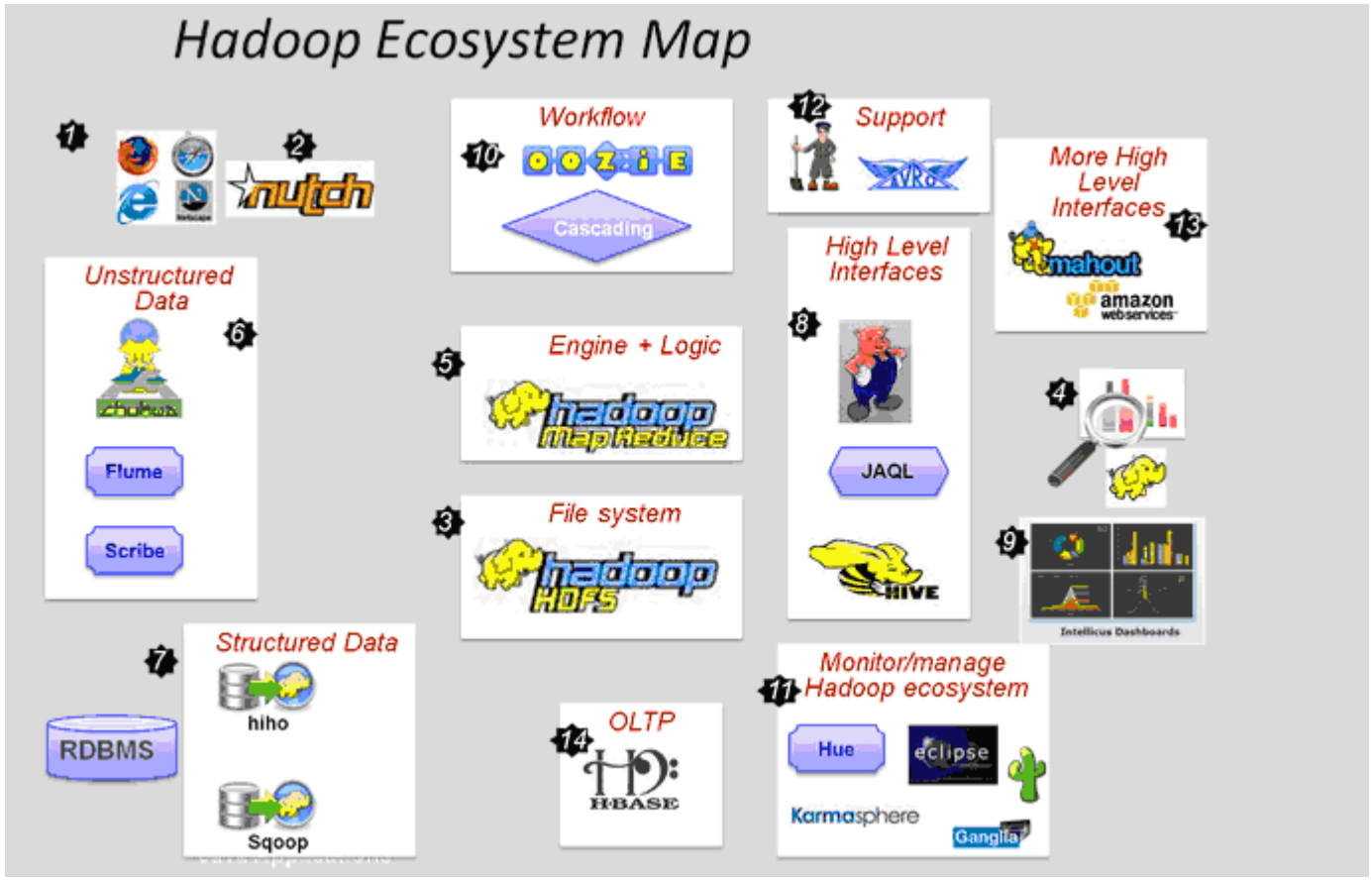
2 Environment

3 HIVE

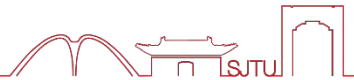
4 Result

5 Conclusion

1.1 Background



1.2 Goals



1. familiar with Hadoop cluster.
2. Optimize queries with Hive.
3. Compare the queries with Mysql.
4. Do some extra work about matching learning.

2. Environment

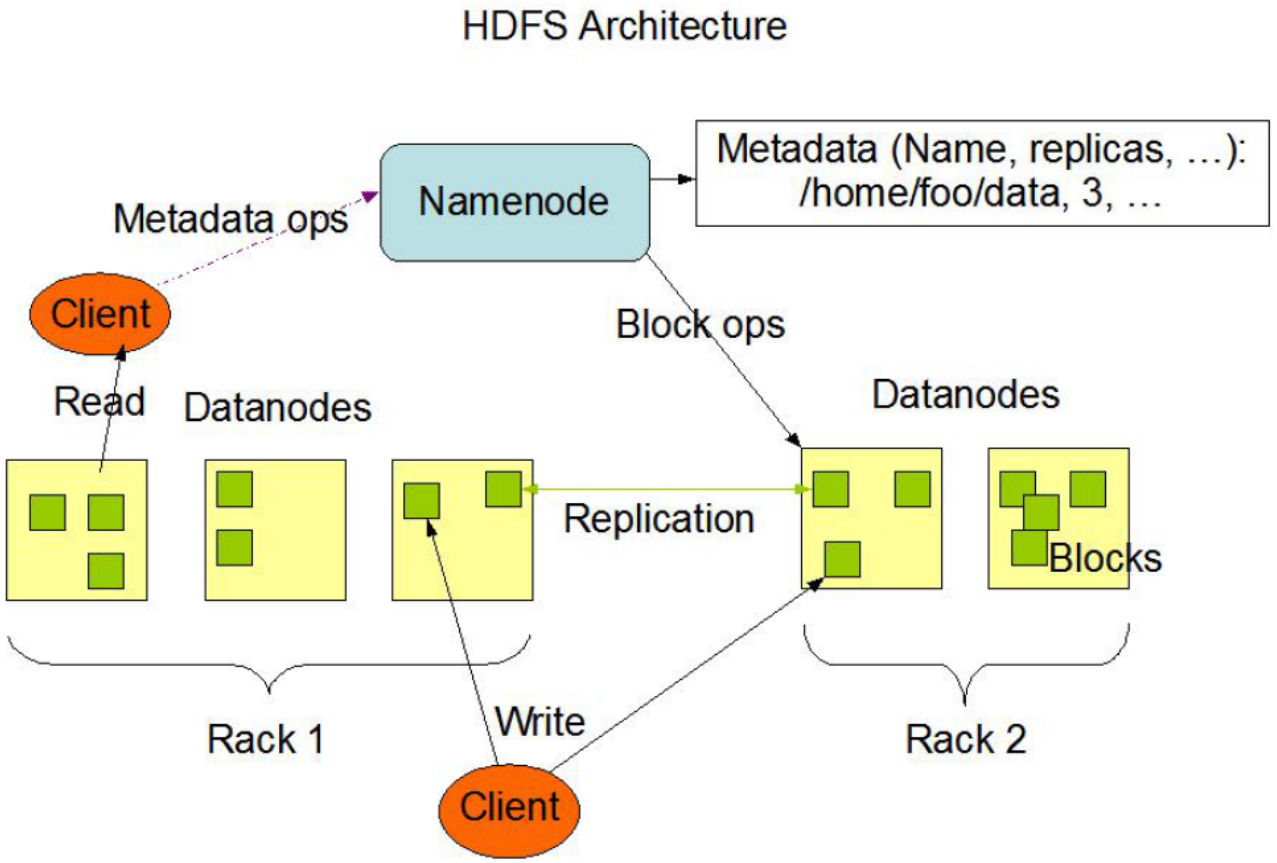
Started	Tue Apr 18 17:29:30 CST 2017
Version	2.7.3
Configured Capacity	15.06 TB
Live Nodes	3

2.1 HDFS

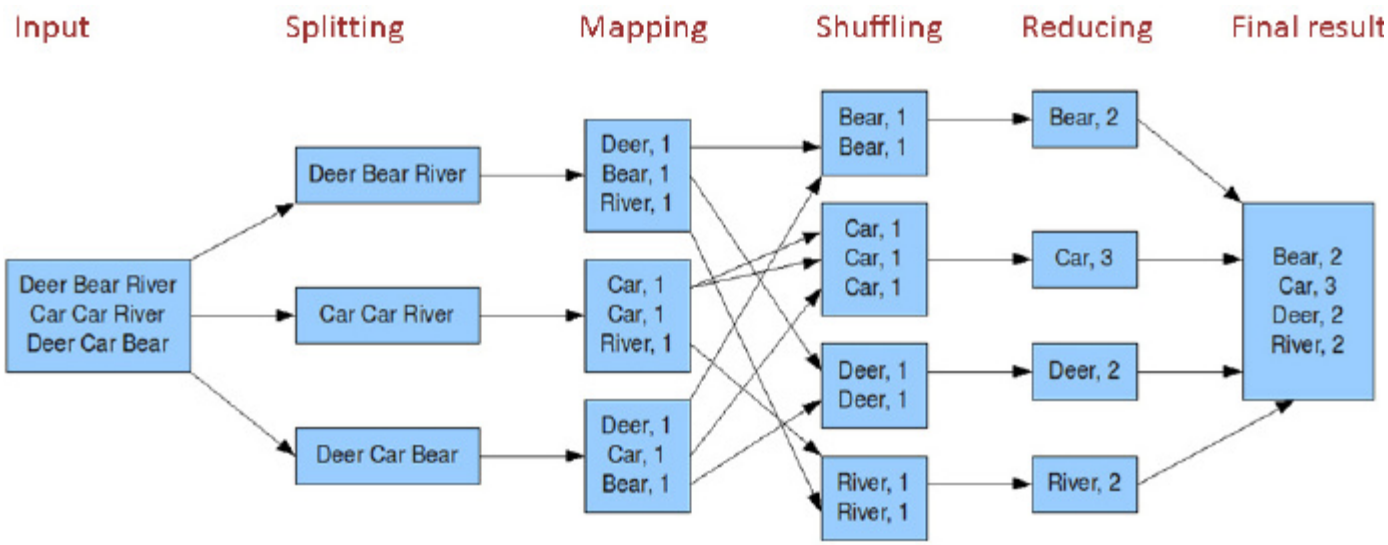
Some features of HDFS:

1. Streaming Data Access
2. Simple Coherency Model
3. Large Data Sets
4. “Moving Computation is Cheaper than Moving Data
5. Portability Across Heterogeneous Hardware and Software Platforms

2.1 HDFS


HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

2.2 MapReduce

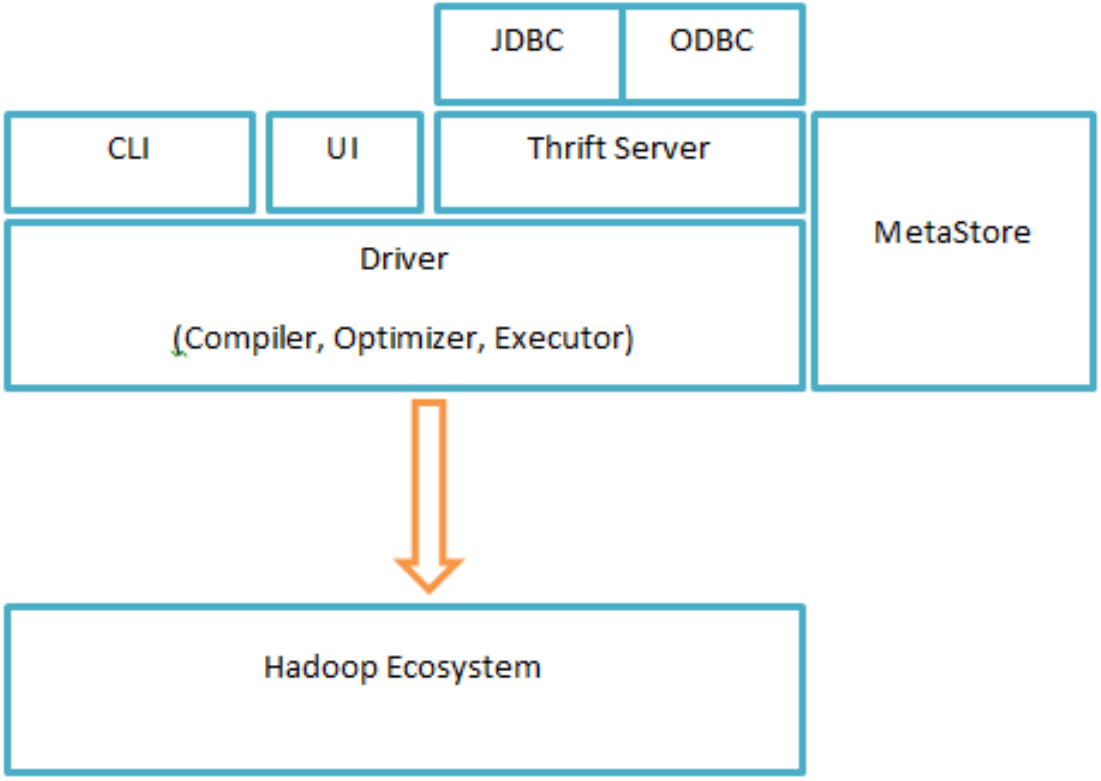


3. HIVE

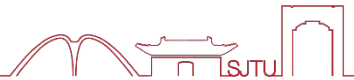
Some features of HDFS:

1. Tools to enable easy access to data via SQL, thus enabling data warehousing tasks such as extract/transform/load (ETL), reporting, and data analysis.
2. A mechanism to impose structure on a variety of data
3. Access to files stored either directly in Apache HDFSTM or in other data storage systems such as Apache HBaseTM
4. Query execution via Apache TezTM, Apache SparkTM, or MapReduce
5. Procedural language with HPL-SQL
6. Sub-second query retrieval via Hive LLAP, Apache YARN and Apache Slider.

3. HIVE


JDBC
ODBC
CLI
UI
Thrift Server
MetaStore
Driver
(Compiler, Optimizer, Executor)
Hadoop Ecosystem

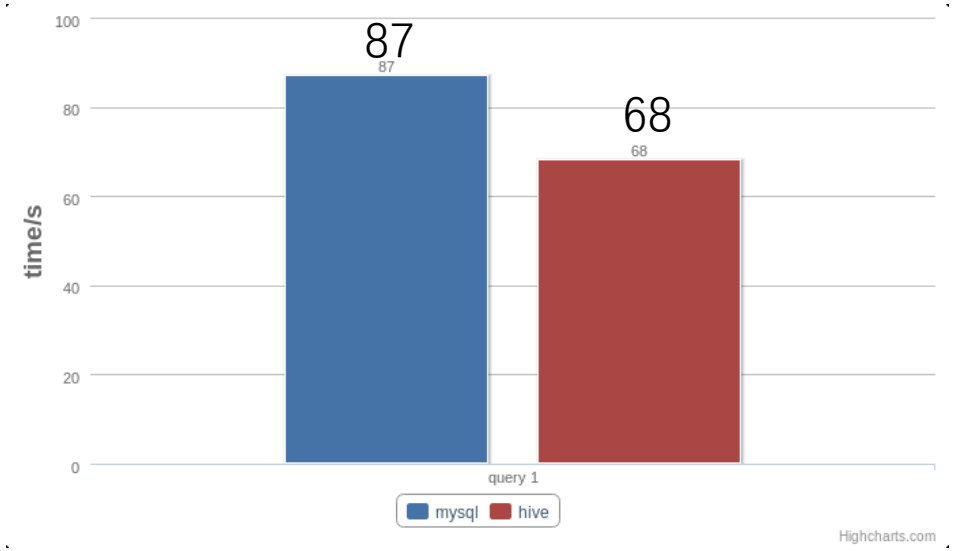
4.1 Queries




```
1 select PaperReferences.PaperID,count(*)
2     from PaperReferences inner join
3     (select * from PaperAuthorAffiliations
4     where AuthorID='0C7733DB') as TB
5     on PaperReferences.PaperReferenceID = TB.PaperID
6     group by PaperReferences.PaperID
```

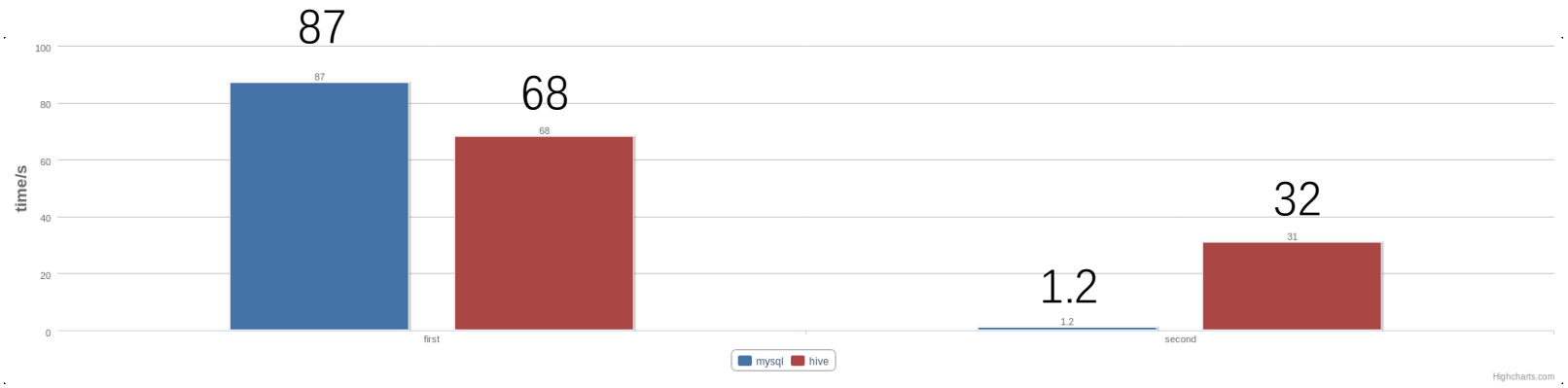
```
1 SELECT  count(*),SUM(SCICitation) as sum from
2     PaperSciReferencesCount CROSS JOIN
3     (select PaperID
4     from PaperAuthorAffiliations
5     where AuthorID = '0000194E' ) AS TB1 on
6     PaperSciReferencesCount.PaperReferenceID = TB1.PaperID
```

4.1 Queries


100
80
60
40
20
0
time/s
87
87
68
68
query 1
mysql
hive
Highcharts.com


3.7
3.4
time/s

4.1 Queries


87
68
32
1.2
time/s
100
80
60
40
20
0
87
68
31
1.2
first
second
mysql
hive
Highcharts.com

5. Conclusion

Hive is an great database for its speed to query. Also Hive is not adapt well with our Acemap system because the query in our website is not so complicated to use Hive. We can use Hive to do some data mining work.

Thanks!