IMPALA

Thaveewat Khanan

2603615 Organization Data Management

open source massively parallel processing (MPP) SQL query engine

Cloudera Impala is a query engine that runs on Apache Hadoop. Impala brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache HBase without requiring data movement or transformation.

What is Impala ?

General--- purpose SQL engine
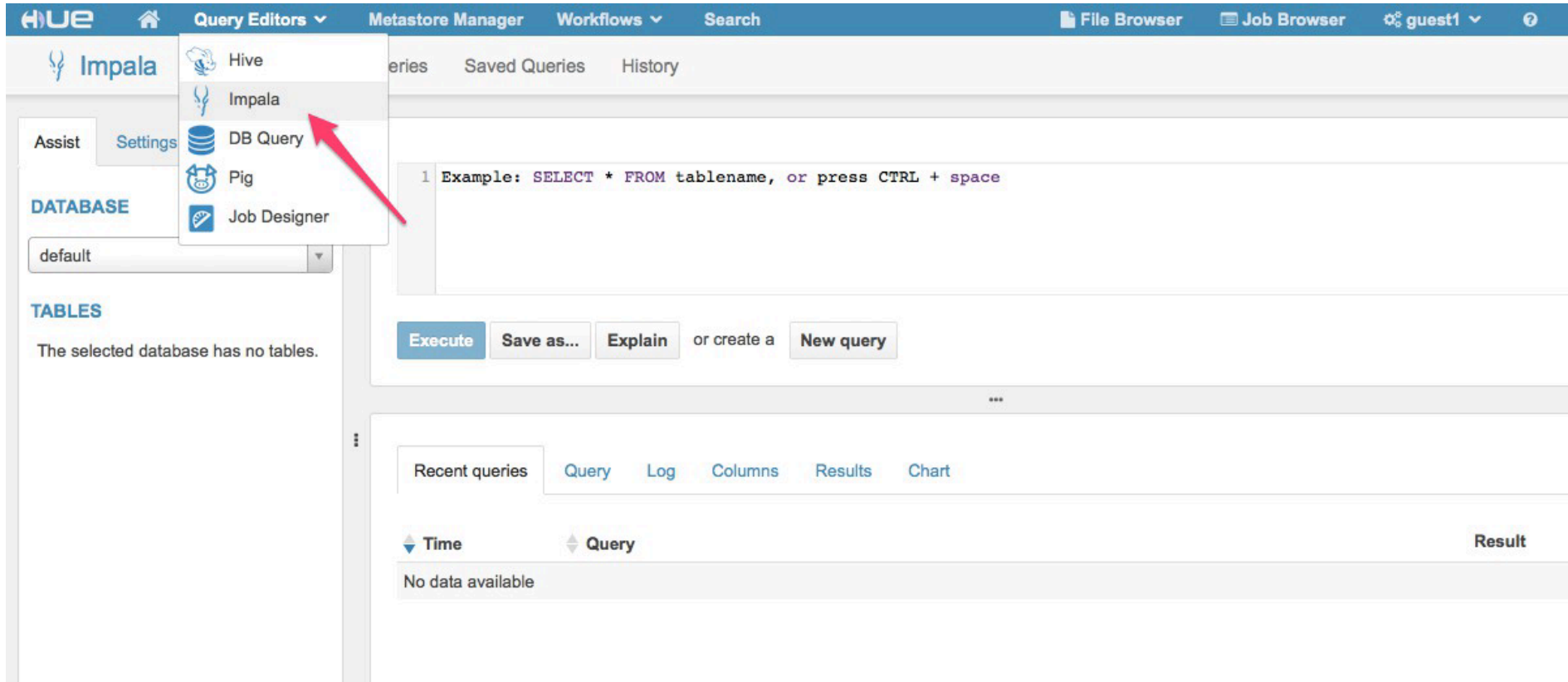
Real--time queries in Apache Hadoop

Opensource under Apache License

Runs directly within Hadoop
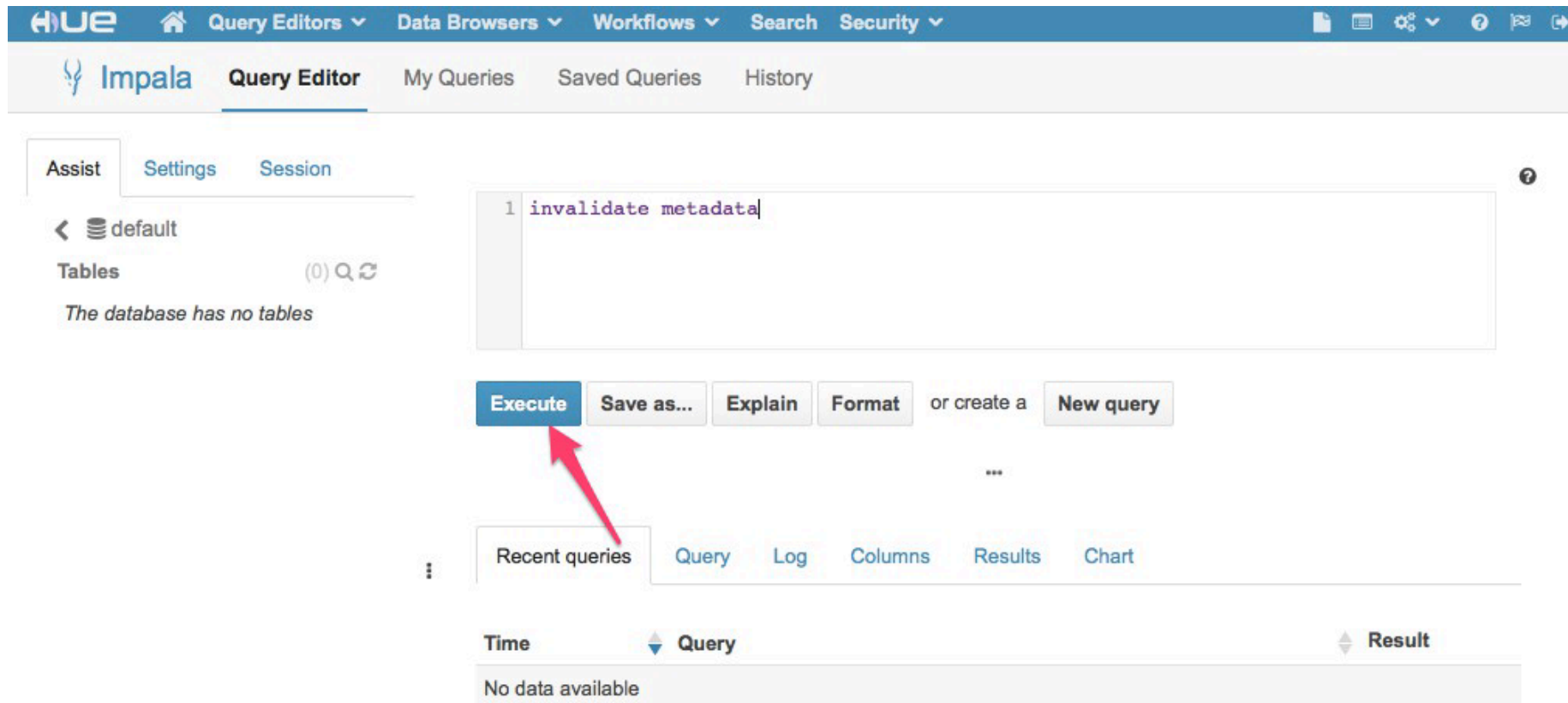
High performance

- – C++ instead of Java

- – Runtime code generator

- – Roughly 4-100 x Hive

Start Impala Query Editor

Update the list of tables/metadata by execute the command invalidate metadata

Run Impala On Terminal

Open Terminal

Type command > impala-shell

[quickstart.cloudera:21000] > select * from users;

```
[[quickstart.cloudera:21000] > select * from users;
Query: select * from users
+--------+-----+--------+---------------+---------+
| userid | age | gender | occupation    | zipcode |
+--------+-----+--------+---------------+---------+
| 1      | 24  | M      | technician    | 85711   |
| 2      | 53  | F      | other         | 94043   |
| 3      | 23  | M      | writer        | 32067   |
| 4      | 24  | M      | technician    | 43537   |
| 5      | 33  | F      | other         | 15213   |
| 6      | 42  | M      | executive     | 98101   |
| 7      | 57  | M      | administrator | 91344   |
| 8      | 36  | M      | administrator | 05201   |
| 9      | 29  | M      | student       | 01002   |
| 10     | 53  | M      | lawyer        | 90703   |
| 11     | 39  | F      | other         | 30329   |
| 12     | 28  | F      | other         | 06405   |
| 13     | 47  | M      | educator      | 29206   |
| 14     | 45  | M      | scientist     | 55106   |
| 15     | 49  | F      | educator      | 97301   |
| 16     | 21  | M      | entertainment | 10309   |
```