

APACHE SQOOP



Introduction



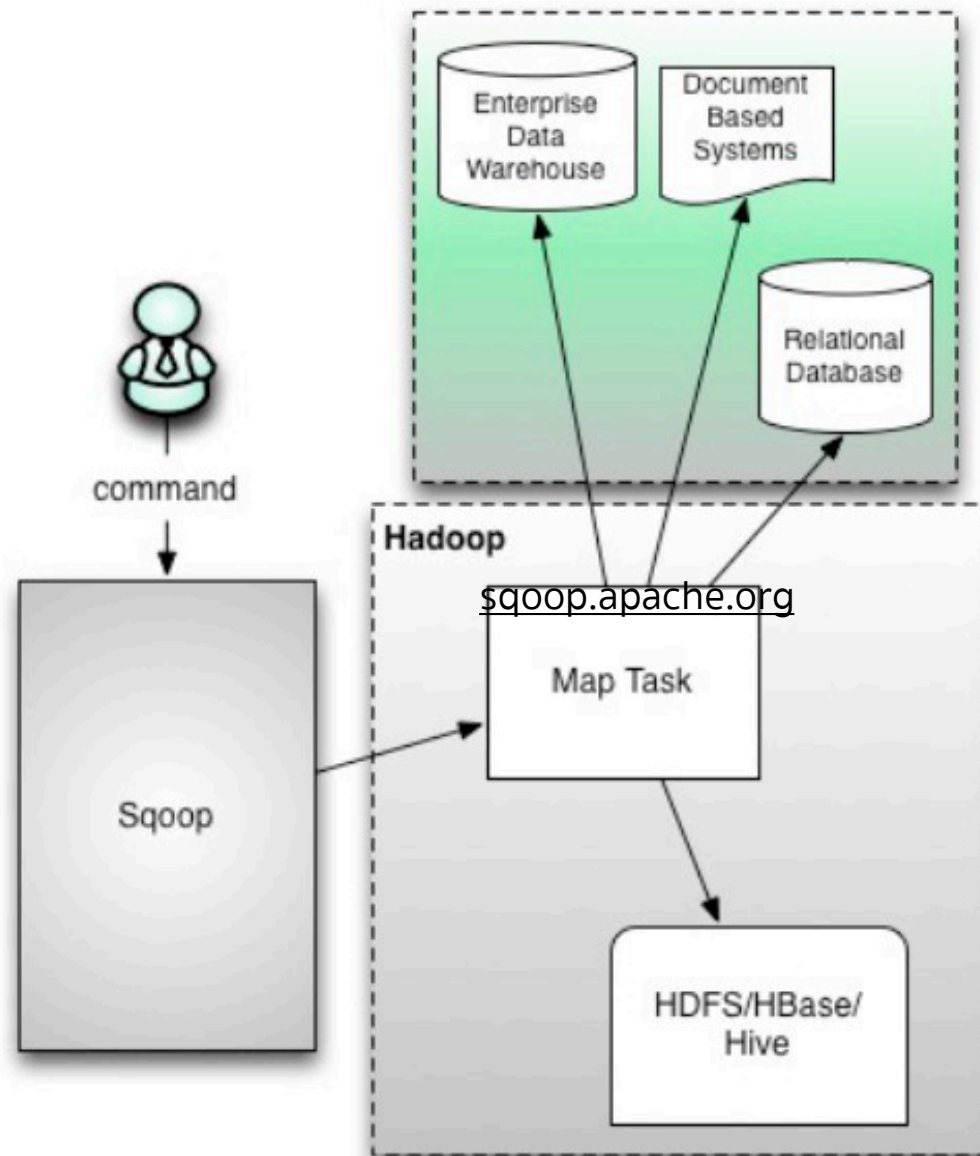
Scoop (“SQL-to-Hadoop”) is a straightforward command-line tool with the following capabilities:

Imports individual tables or entire databases to files in HDFS

Generates Java classes to allow you to interact with your imported data

Provides the ability to import from SQL databases straight into your Hive data warehouse

Architecture Overview



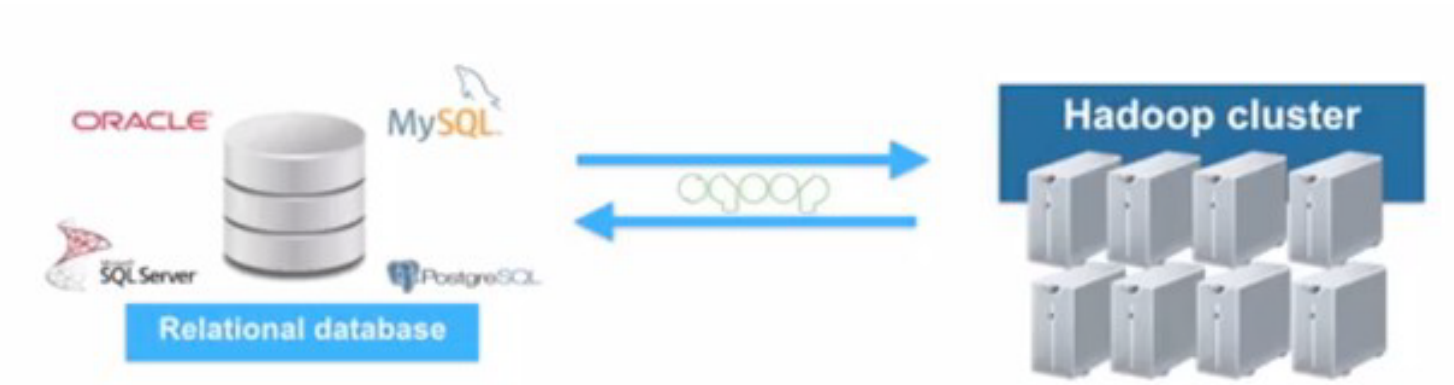
Sqoop Benefit

- **Leverages RDBMS metadata to get the column data types**
- **It is simple to script and uses SQL**
- **It can be used to handle change data capture by importing daily transactional data to Hadoop**
- **It uses MapReduce for export and import that enables parallel and efficient data movement**

Sqoop Mode

Sqoop import: Data moves from RDBMS to Hadoop

Sqoop export: Data moves from Hadoop to RDBMS



Use Case : Data Consolidation

- Integrate data from various organizational “data stores” to Hadoop for various data processing requirements



Import Commands

Parameters	Description
<code>--connect <jdbc-uri></code>	Specifies the server or database to connect to. It also specifies the port. For example: <code>--connect jdbc:mysql://host:port/databaseName</code>
<code>--connection-manager <class-name></code>	Specifies the connection manager class name.
<code>--driver <class-name></code>	Specifies the fully qualified name of the JDBC driver class.
<code>--hadoop-home <dir></code>	This parameter is used to override the <code>\$HADOOP_HOME</code> environment variable.
<code>-P</code>	If a user doesn't want to specify the database password along with the command, we can use the <code>-P</code> option to read the password from the console.
<code>--password <password></code>	Sets the authentication password required to connect to the input source.
<code>--username <username></code>	Sets the authentication username.
<code>--connection-param-file <properties-file></code>	Specifies the connection parameter's file.
<code>--help</code>	This option will provide the usage instructions.
<code>--verbose</code>	Prints more information during a query execution.

Export Commands

Parameters	Description
<code>--direct</code>	Use the direct mode to perform the export quickly. Note that it is only supported for MySQL.
<code>--export-dir<dir></code>	The location of input files in HDFS.
<code>--table <table-name></code>	Name of the output table (the RDBMS table).
<code>-m,--num-mappers <n></code>	Refers to the number of map tasks.
<code>--update-mode <mode></code>	Specifies how updates are performed when new rows are found with non-matching keys in the database. Legal values for the mode include <code>updateonly</code> (default) and <code>allowinsert</code> .
<code>--update-key <col-name></code>	The value of this column is used to identify the records that a user wants to update during the update mode. Use a comma-separated list of columns if there is more than one column.
<code>--staging-table <staging-table-name></code>	Specifies the name of the staging table. The staging table is used to stage the data before inserting it into the destination table.
<code>--clear-staging-table</code>	This argument is used to clean the data from the staging table.

Loading Data from RDBMS to Hadoop

Configuring MySQL On Cloudera.Quickstart

```
$ sudo /usr/bin/mysql_secure_installation
```

Enter current password for root (enter for none):

OK, successfully used password, moving on...

Set root password? [Y/n] **Y**

New password:**cloudera**

Re-enter new password:**cloudera**

Remove anonymous users? [Y/n] **Y**

Disallow root login remotely? [Y/n] **N**

Remove test database and access to it [Y/n] **Y**

Reload privilege tables now? [Y/n] **Y**

All done!

Running MySQL

\$ mysql -uroot -p"cloudera"

```
[cloudera@quickstart ~]$ mysql -uroot -p"cloudera"
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 389
Server version: 5.1.73 Source distribution
```

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

```
mysql> █
```

mysql> show databases;

```
mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| cm |
| firehose |
| hue |
| metastore |
| mysql |
| nav |
| navms |
| oozie |
| retail_db |
| rman |
| sentry |
+-----+
12 rows in set (0.01 sec)
```

Prepare a test database table

```
mysql> CREATE DATABASE test_mysql_db;
mysql> USE test_mysql_db;
mysql> CREATE TABLE country_tbl(id INT NOT NULL, country
VARCHAR(50), PRIMARY KEY (id));
mysql> INSERT INTO country_tbl VALUES(1, 'USA');
mysql> INSERT INTO country_tbl VALUES(2, 'CANADA');
mysql> INSERT INTO country_tbl VALUES(3, 'Mexico');
mysql> INSERT INTO country_tbl VALUES(4, 'Brazil');
mysql> INSERT INTO country_tbl VALUES(61, 'Japan');
mysql> INSERT INTO country_tbl VALUES(65, 'Singapore');
mysql> INSERT INTO country_tbl VALUES(66, 'Thailand');
```

View data in the table

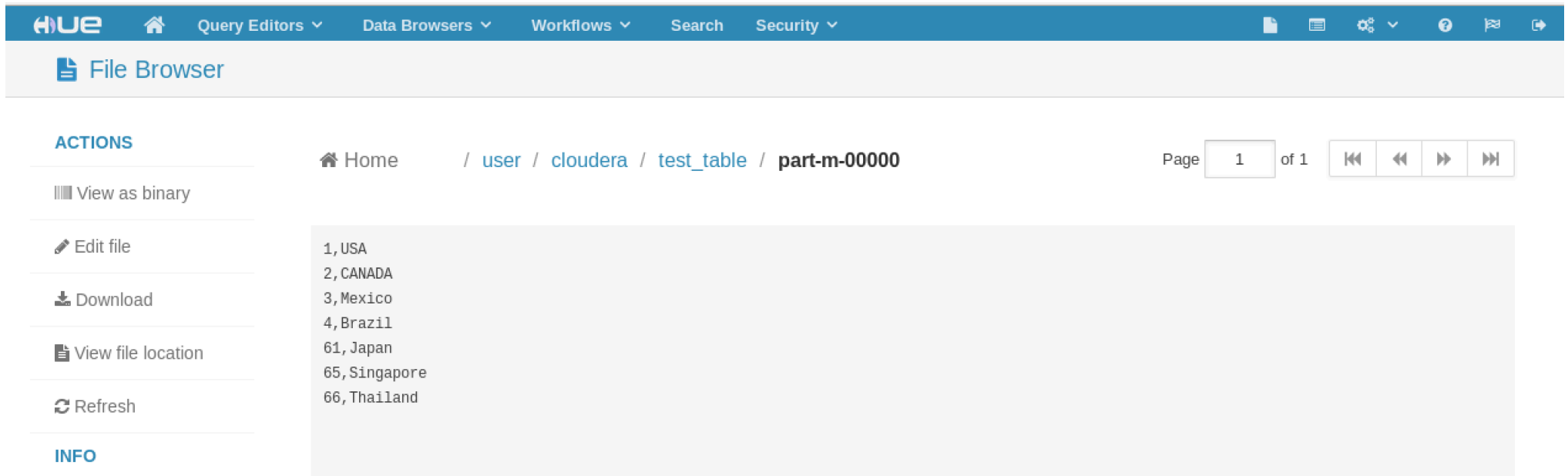
```
mysql> SELECT * FROM country_tbl;
```

```
+-----+-----+
| id | country |
+-----+-----+
| 1  | USA     |
| 2  | CANADA  |
| 3  | Mexico  |
| 4  | Brazil  |
| 61 | Japan   |
| 65 | Singapore |
| 66 | Thailand |
+-----+-----+
7 rows in set (0.00 sec)
```

```
mysql> exit;
```

Importing data from MySQL to HDFS

```
$ sqoop import --connect jdbc:mysql://localhost/test_mysql_db --  
username root --password cloudera --table country_tbl --target-dir /  
user/cloudera/test_table -m 1
```

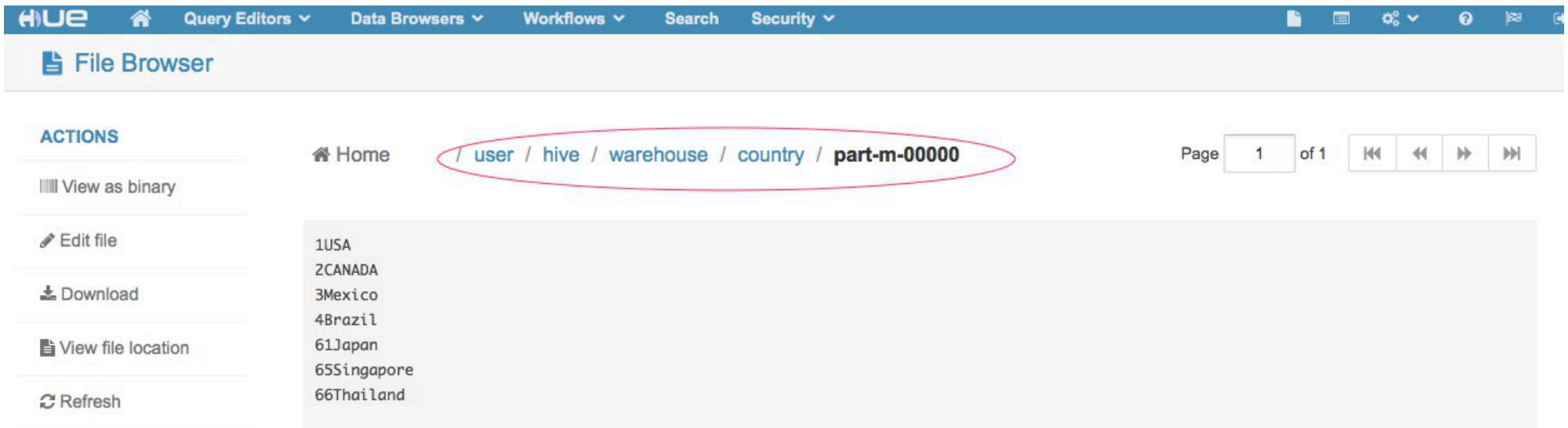


The screenshot shows the HUE web interface. The top navigation bar includes 'HUE', a home icon, and dropdown menus for 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. On the right of the bar are icons for file operations. Below the bar, the 'File Browser' section is active. The left sidebar contains an 'ACTIONS' menu with options: 'View as binary', 'Edit file', 'Download', 'View file location', and 'Refresh'. The main content area shows the breadcrumb path: 'Home / user / cloudera / test_table / part-m-00000'. To the right of the path is a pagination control showing 'Page 1 of 1' and navigation buttons. The file content is displayed as a list of lines:

```
1, USA  
2, CANADA  
3, Mexico  
4, Brazil  
61, Japan  
65, Singapore  
66, Thailand
```

Importing data from MySQL to Hive Table

```
$ sqoop import--connect jdbc:mysql://localhost/test_mysql_db --  
username root --password cloudera --table country_tbl --hive-import --  
hive-table country -m 1
```



The screenshot shows the Hue File Browser interface. The top navigation bar includes 'HUE', 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. The main header is 'File Browser'. On the left, there is an 'ACTIONS' sidebar with options: 'View as binary', 'Edit file', 'Download', 'View file location', and 'Refresh'. The main content area shows a breadcrumb path: 'Home / user / hive / warehouse / country / part-m-00000'. The path 'user / hive / warehouse / country / part-m-00000' is circled in red. Below the path, a list of country codes is displayed: 1USA, 2CANADA, 3Mexico, 4Brazil, 61Japan, 65Singapore, and 66Thailand. On the right, there is a pagination control showing 'Page 1 of 1' and navigation buttons.

Reviewing data from Hive Table

```
[cloudera@quickstart ~]$ hive
```

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
```

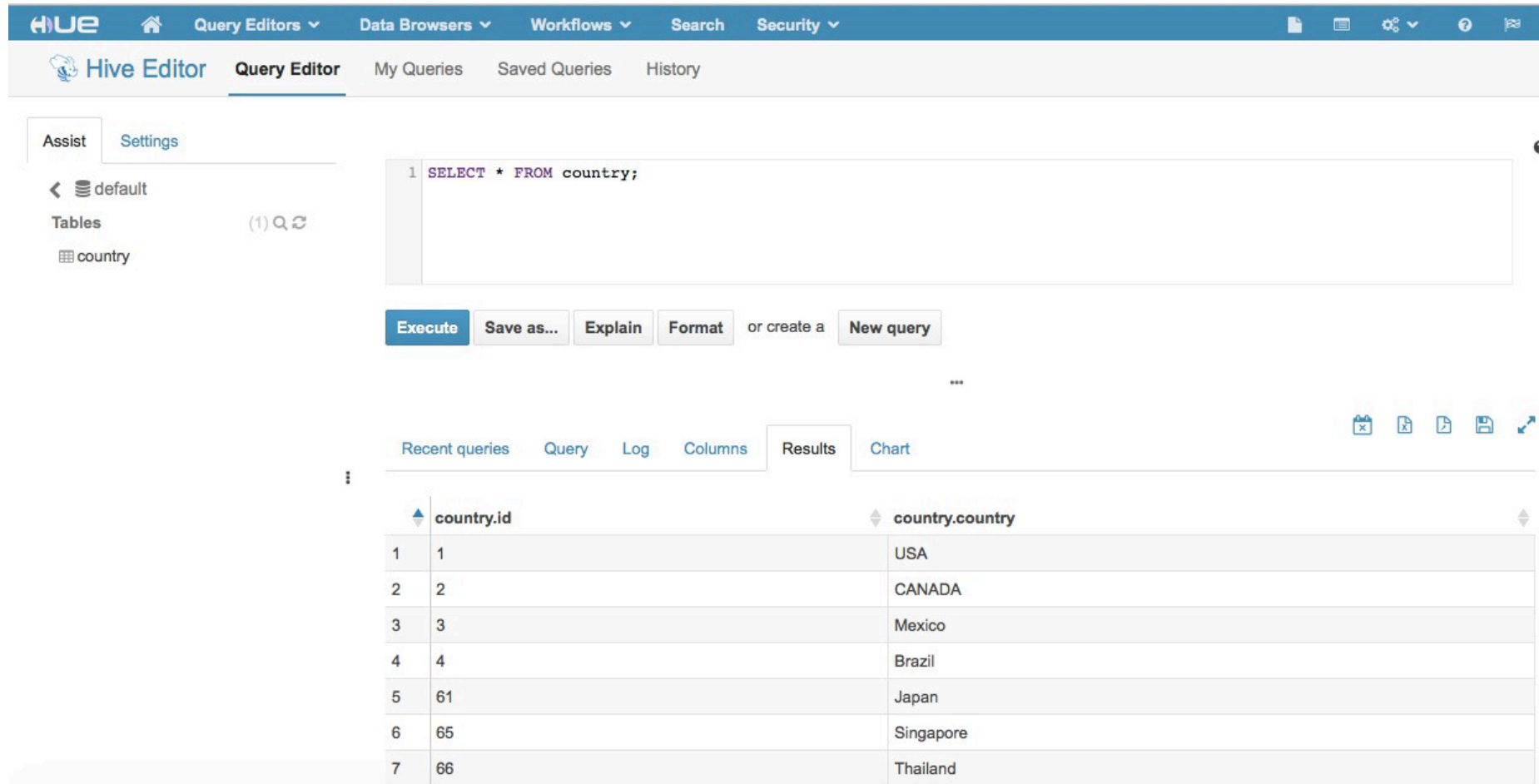
```
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
```

```
hive> show tables;
```

```
hive> select * from country;
```

```
...  
1      USA  
2      CANADA  
3      Mexico  
4      Brazil  
61     Japan  
65     Singapore  
66     Thailand  
Time taken: 0.587 seconds, Fetched: 7 row(s)
```

Running from Hue: Beewax



The screenshot shows the Hue Query Editor interface. The top navigation bar includes 'HUE', 'Home', 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. Below this, the 'Hive Editor' tab is active, showing 'Query Editor', 'My Queries', 'Saved Queries', and 'History'. On the left sidebar, 'Assist' and 'Settings' are visible, along with a 'Tables' section listing 'country'. The main query editor contains the SQL query: `1 SELECT * FROM country;`. Below the query editor are buttons for 'Execute', 'Save as...', 'Explain', 'Format', and 'New query'. The 'Results' tab is selected, displaying a table with two columns: 'country.id' and 'country.country'. The table contains 7 rows of data.

	country.id	country.country
1	1	USA
2	2	CANADA
3	3	Mexico
4	4	Brazil
5	61	Japan
6	65	Singapore
7	66	Thailand

Importing data from MySQL to HBase

```
$ sqoop import--connect jdbc:mysql://localhost/test_mysql_db --  
username root --password cloudera --table country_tbl --hbase-  
table country --column-family hbase_country_cf --hbase-row-key  
id --hbase-create-table -m 1
```

Start HBase

```
$ hbase shell
```

```
hbase(main):001:0> list
```

```
hbase(main):001:0> list
```

```
TABLE
```

```
country
```

```
employee
```

```
student
```

```
3 row(s) in 0.3720 seconds
```

```
=> ["country", "employee", "student"]
```

Viewing Hbase data

```
hbase(main):003:0> scan 'country'
ROW          COLUMN+CELL
 1          column=hbase_country_cf:country, timestamp=1468081466623, value=USA
 2          column=hbase_country_cf:country, timestamp=1468081466623, value=CANADA
 3          column=hbase_country_cf:country, timestamp=1468081466623, value=Mexico
 4          column=hbase_country_cf:country, timestamp=1468081466623, value=Brazil
61          column=hbase_country_cf:country, timestamp=1468081466623, value=Japan
65          column=hbase_country_cf:country, timestamp=1468081466623, value=Singapore
66          column=hbase_country_cf:country, timestamp=1468081466623, value=Thailand
7 row(s) in 0.1670 seconds
```

Viewing data from Hbase browser

HUE Home Query Editors Data Browsers Workflows Search Security

HBase Browser

Home - Cluster / country Switch Cluster

row_key, row_prefix* +scan_len [col1, family:col2, fam3:, col_prefix* Q hbase_country_cf: Filter Columns/Families ☒ All Sort By ASC

1	<div>hbase_country_cf: country</div> <div>USA</div>
2	<div>hbase_country_cf: country</div> <div>CANADA</div>
3	<div>hbase_country_cf: country</div> <div>Mexico</div>