

# Cheat Sheet

## Hive for SQL Users

### Contents

- 1 | Additional Resources
- 2 | Query, Metadata
- 3 | Current SQL Compatibility, Command Line, Hive Shell

If you're already a SQL user then working with Hadoop may be a little easier than you think, thanks to Apache Hive. Apache Hive is data warehouse infrastructure built on top of Apache™ Hadoop® for providing data summarization, ad hoc query, and analysis of large datasets. It provides a mechanism to project structure onto the data in Hadoop and to query that data using a SQL-like language called HiveQL (HQL).

Use this handy cheat sheet (based on this [original MySQL cheat sheet](#)) to get going with Hive and Hadoop.

### Additional Resources



Learn to become fluent in Apache Hive with the Hive Language Manual:

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual>



Get in the Hortonworks Sandbox and try out Hadoop with interactive tutorials:

<http://hortonworks.com/sandbox>



Try out Qubole Data Service (QDS) for free:

<http://www.qubole.com/try>

## Query

| Function   | MySQL  | HiveQL  |
|--|--|---|
| Retrieving information   | <code>SELECT from_columns FROM table WHERE conditions;</code>                      | <code>SELECT from_columns FROM table WHERE conditions;</code>                         |
| All values   | <code>SELECT * FROM table;</code>  | <code>SELECT * FROM table;</code>   |
| Some values  | <code>SELECT * FROM table WHERE rec_name = "value";</code>                         | <code>SELECT * FROM table WHERE rec_name = "value";</code>                            |
| Multiple criteria  | <code>SELECT * FROM table WHERE rec1="value1" AND rec2="value2";</code>            | <code>SELECT * FROM TABLE WHERE rec1 = "value1" AND rec2 = "value2";</code>           |
| Selecting specific columns   | <code>SELECT column_name FROM table;</code>  | <code>SELECT column_name FROM table;</code>   |
| Retrieving unique output records                                       | <code>SELECT DISTINCT column_name FROM table;</code>                               | <code>SELECT DISTINCT column_name FROM table;</code>                                  |
| Sorting  | <code>SELECT col1, col2 FROM table ORDER BY col2;</code>                           | <code>SELECT col1, col2 FROM table ORDER BY col2;</code>                              |
| Sorting backward   | <code>SELECT col1, col2 FROM table ORDER BY col2 DESC;</code>                      | <code>SELECT col1, col2 FROM table ORDER BY col2 DESC;</code>                         |
| Counting rows  | <code>SELECT COUNT(*) FROM table;</code>   | <code>SELECT COUNT(*) FROM table;</code>  |
| Grouping with counting   | <code>SELECT owner, COUNT(*) FROM table GROUP BY owner;</code>                     | <code>SELECT owner, COUNT(*) FROM table GROUP BY owner;</code>                        |
| Maximum value  | <code>SELECT MAX(col_name) AS label FROM table;</code>                             | <code>SELECT MAX(col_name) AS label FROM table;</code>                                |
| Selecting from multiple tables<br>(Join same table using alias w/"AS") | <code>SELECT pet.name, comment FROM pet, event WHERE pet.name = event.name;</code> | <code>SELECT pet.name, comment FROM pet JOIN event ON (pet.name = event.name);</code> |

## Metadata

| Function                         | MySQL                                 | HiveQL  |
|----------------------------------|---------------------------------------|---|
| Selecting a database             | <code>USE database;</code>            | <code>USE database;</code>                        |
| Listing databases                | <code>SHOW DATABASES;</code>          | <code>SHOW DATABASES;</code>                      |
| Listing tables in a database     | <code>SHOW TABLES;</code>             | <code>SHOW TABLES;</code>                         |
| Describing the format of a table | <code>DESCRIBE table;</code>          | <code>DESCRIBE (FORMATTED EXTENDED) table;</code> |
| Creating a database              | <code>CREATE DATABASE db_name;</code> | <code>CREATE DATABASE db_name;</code>             |
| Dropping a database              | <code>DROP DATABASE db_name;</code>   | <code>DROP DATABASE db_name (CASCADE);</code>     |

## Current SQL Compatibility

| Hive SQL Datatypes        | Hive SQL Semantics                                   |
|---------------------------|--|
| INT                       | SELECT, LOAD INSERT from query                       |
| TINYINT/SMALLINT/BIGINT   | Expressions in WHERE and HAVING                      |
| BOOLEAN                   | GROUP BY, ORDER BY, SORT BY                          |
| FLOAT                     | Sub-queries in FROM clause                           |
| DOUBLE                    | GROUP BY, ORDER BY                                   |
| STRING                    | CLUSTER BY, DISTRIBUTE BY                            |
| TIMESTAMP                 | ROLLUP and CUBE                                      |
| BINARY                    | UNION  |
| ARRAY, MAP, STRUCT, UNION | LEFT, RIGHT and FULL INNER/OUTER JOIN                |
| DECIMAL                   | CROSS JOIN, LEFT SEMI JOIN                           |
| CHAR                      | Windowing functions (OVER, RANK, etc)                |
| VARCHAR                   | INTERSECT, EXCEPT, UNION, DISTINCT                   |
| DATE                      | Sub-queries in WHERE (IN, NOT IN, EXISTS/NOT EXISTS) |
|                           | Sub-queries in HAVING                                |

| Color Key |
|-----------|
| Hive 0.10 |
| Hive 0.11 |
| FUTURE    |

## Command Line

| Function                   | Hive  |
|----------------------------|---|
| Run query                  | hive -e 'select a.col from tab1 a'  |
| Run query silent mode      | hive -S -e 'select a.col from tab1 a'                                       |
| Set hive config variables  | hive -e 'select a.col from tab1 a' -hiveconf hive.root.logger=DEBUG,console |
| Use initialization script  | hive -i initialize.sql  |
| Run non-interactive script | hive -f script.sql  |

## Hive Shell

| Function                              | Hive                       |
|---------------------------------------|----------------------------|
| Run script inside shell               | source file_name           |
| Run ls (dfs) commands                 | dfs -ls /user              |
| Run ls (bash command) from shell      | !ls                        |
| Set configuration variables           | set mapred.reduce.tasks=32 |
| TAB auto completion                   | set hive.<TAB>             |
| Show all variables starting with hive | set                        |
| Revert all variables                  | reset                      |
| Add jar to distributed cache          | add jar jar_path           |
| Show all jars in distributed cache    | list jars                  |
| Delete jar from distributed cache     | delete jar jar_name        |