

Hive Activity Sheet

Files

Commands:

To enter the hive prompt/CLI
\$ hive

To list all the existing databases:
> show databases;

To create a new database:
> CREATE DATABASE IF NOT EXISTS testing;

To change/use to the current database:
> use testing;

To create a new table in the above database:
> CREATE TABLE posts (user STRING, post STRING, time BIGINT) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

To view/list all the tables in the current database:
> show tables;

To view the schema/structure of the table:
> describe posts;

To load data (file) from local file system to Hive:
> LOAD DATA LOCAL INPATH '/home/insofe/user-posts.txt' OVERWRITE
INTO TABLE posts;

To view the contents of the above table:
> select count (1) from posts;
> select * from posts where user="user2";
> select * from posts where time<=1343182133839 limit 2;

To drop (remove the data and schema) the table:
> DROP TABLE posts;

To Load data from HDFS:
> LOAD DATA INPATH '/home/insofe/user-posts.txt' OVERWRITE INTO
TABLE posts;

```
> CREATE EXTERNAL TABLE postsloc (user STRING, post STRING, time BIGINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 'user/insofe/';
```

```
> CREATE EXTERNAL TABLE posts (user STRING, post STRING, time BIGINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
```

Schema Violations:

null is set for any value that violates pre-defined schema.

```
> LOAD DATA LOCAL INPATH '/home/insofe/user-posts-inconsistentFormat.txt' OVERWRITE INTO TABLE posts;
```

```
> drop table posts;
```

Partitions:

```
> CREATE TABLE posts (user STRING, post STRING, time BIGINT)
  PARTITIONED BY(country STRING)
  ROW FORMAT DELIMITED
  FIELDS TERMINATED BY ','
  STORED AS TEXTFILE;
```

Below command shows an error

```
> LOAD DATA LOCAL INPATH '/home/insofe/user-posts.txt' OVERWRITE
  INTO TABLE posts;
```

Partitions:

```
> LOAD DATA LOCAL INPATH '/home/insofe/user-posts-US.txt'
  OVERWRITE INTO TABLE posts PARTITION(country='US');
```

```
> LOAD DATA LOCAL INPATH '/home/insofe/user-posts-AUSTRALIA.txt'
  OVERWRITE INTO TABLE posts PARTITION(country='AUSTRALIA');
```

Show partitions posts;

```
> select * from posts where country='US' limit 10;
```

Buckets:

```
> CREATE TABLE post_count (user STRING, count INT) CLUSTERED BY
  (user) INTO 5 BUCKETS;
```

```
> set hive.enforce.bucketing = true;

> insert overwrite table post_count select user, count(post)
from posts group by user;

> select * from post_count TABLESAMPLE(BUCKET 1 OUT OF 2);
```

Joins:

Default Join is Inner Join.

Rows are joined where the keys match.

Rows that do not have matches are not included in the result.

```
> drop table posts;

> CREATE TABLE posts (user STRING, post STRING, time BIGINT) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

> LOAD DATA LOCAL INPATH '/home/insofe/user-posts.txt' OVERWRITE
INTO TABLE posts;

> CREATE TABLE likes (user STRING, post STRING, time BIGINT) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

> LOAD DATA LOCAL INPATH '/home/insofe/user-likes.txt' OVERWRITE
INTO TABLE likes;

> SELECT p.*, l.* FROM posts p JOIN likes l ON (p.user = l.user)
limit 10;
```

Outer Join - Rows, which will not join with the other table, are still included in the result

Left Outer Join - Row from the first table are included whether they have a match or not. Columns from the unmatched (second) table are set to null.

```
> SELECT p.*, l.* FROM posts p LEFT OUTER JOIN likes l ON
(p.user = l.user) limit 10;
```

Right Outer Join - The opposite of Left Outer Join: Rows from the second table are included no matter what. Columns from the unmatched (first) table are set to null.

```
> SELECT p.*, l.* FROM posts p RIGHT OUTER JOIN likes l ON  
(p.user = l.user) limit 10;
```

Full Outer Join - Rows from both sides are included. For unmatched rows the columns from the other table are set to null.

```
> SELECT p.*, l.* FROM posts p FULL OUTER JOIN likes l ON  
(p.user = l.user) limit 10;
```