DADS 6002 / CI 7301
Big Data Analytics
Hive Lab

Hive Lab

- Start Hive

```
#  hive
hive >
```

- Exit from Hive

```
hive > exit;
```

- Create a database

```
hive> create database my_db;
hive> show databases;
hive> use my_db;
```

Hive Lab

- Create and alter a table

hive > create table test (id int, name string) row format delimited fields terminated by ‘,’ stored as textfile;

hive > show tables;

hive > describe test;

hive > alter table test add columns (address string);

hive > describe test;

hive > drop table test;

Hive Lab

- Download a zip file ml-100k.zip from MS Teams into the shared folder and copy to the working directory

  (or download it from the web as follow.

  # wget http://files.grouplens.org/datasets/movielens/ml-100k.zip)

unzip ml-100k.zip

more ml-100k/u.user

Hive Lab

```
# cd ml-100k
# hadoop fs –mkdir /user/cloudera/movielens
# hadoop fs –put u.user /user/cloudera/movielens
# hadoop fs –ls /user/cloudera/movielens
# hive
hive > create table users ( userid int, age int,
    gender string, occupation string, zipcode string )
    row format delimited fields terminated by ‘|’
    stored as textfile;
```

Hive Lab

hive > load data inpath
‘/user/cloudera/movielens/u.user’ overwrite into
table users;

hive > select * from users;

hive > select zipcode, count(1) as count, avg(age)
as age from users

group by zipcode order by count desc;

hive > quit;

Hive Lab

- Download a file wlog from MS Teams into the shared folder and copy it to the working directory
- Execute the following commands.

hadoop fs –mkdir /user/cloudera/weblog

hadoop fs –put wlog /user/cloudera/weblog

hadoop fs –ls /user/cloudera/weblog

Hive Lab

```
hive> create table weblogtest ( host string, object string,
    time string )
ROW FORMAT SERDE
    'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
    WITH SERDEPROPERTIES
("input.regex" = "([^ ]+) \"([^ ]+)\" ([0-9]+)" ) stored as
    textfile;
```

Hive Lab

hive > load data inpath ‘/user/cloudera/weblog/wlog’
overwrite into table weblogtest;

hive > select count(1) from weblogtest;

hive > select object, count(1) as count from weblogtest
group by object order by count desc;

Hive Lab

```
hive> create external table weblog (
            host string,
            object string,
            time string )
ROW FORMAT SERDE
    'org.apache.hadoop.hive.contrib.serde2.RegexSerDe' WITH
    SERDEPROPERTIES
("input.regex" = "([^ ]+) \"([^ ]+)\" ([0-9]+)" ) stored as textfile location
    “/user/cloudera/weblog”;
hive > select object, count(1) as count from weblog group by object
    order by count desc;
```