# PROGRAMMING HIVE

CKME 134 – BIG DATA ANALYTICS TOOLS

RYERSON UNIVERSITY

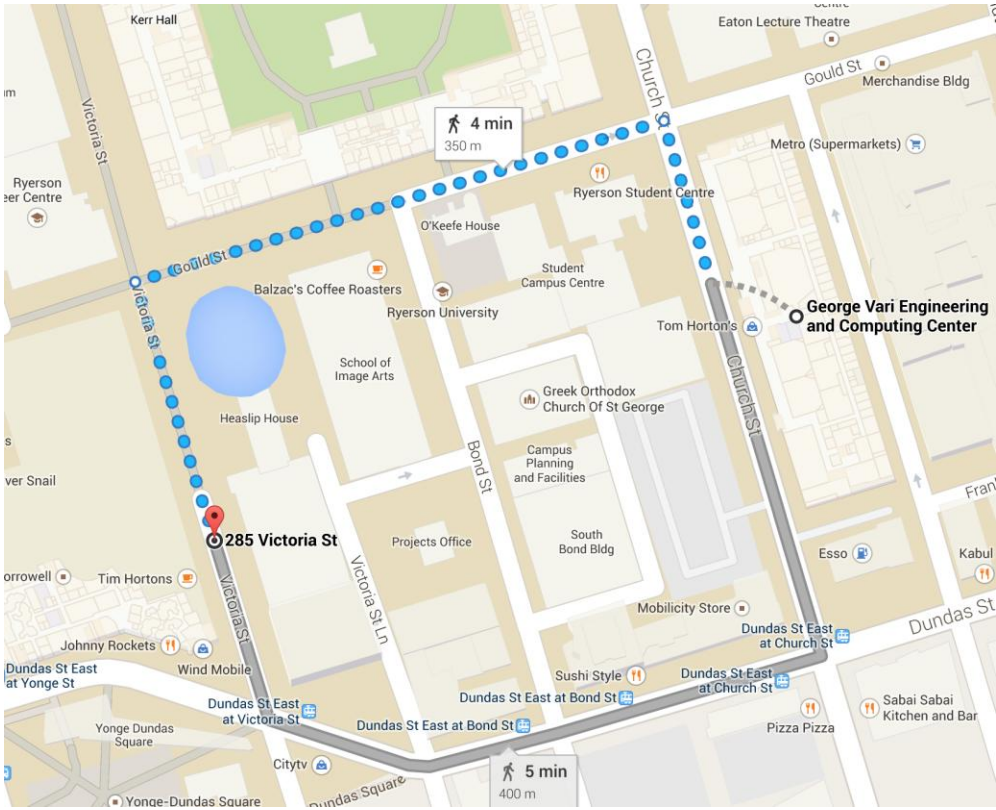
SPRING 2015

Instructor: Shaohua Zhang

# General Course Information



- Instructor
  - Shaohua Zhang
  - Ryerson shaohua.zhang@ryerson.ca
  - Personal shaohua.zhang@live.com
- GA
  - Behjat Soltanifar
  - behjat.soltanifar@ryerson.ca
- Lectures
  - 6:30~8:00
  - ENGLG06
- Lab
  - 8:00~9:30
  - 285 Victoria St (403/404)
    - Take the elevator to 4FL



# Course Outline (*subject to change*)



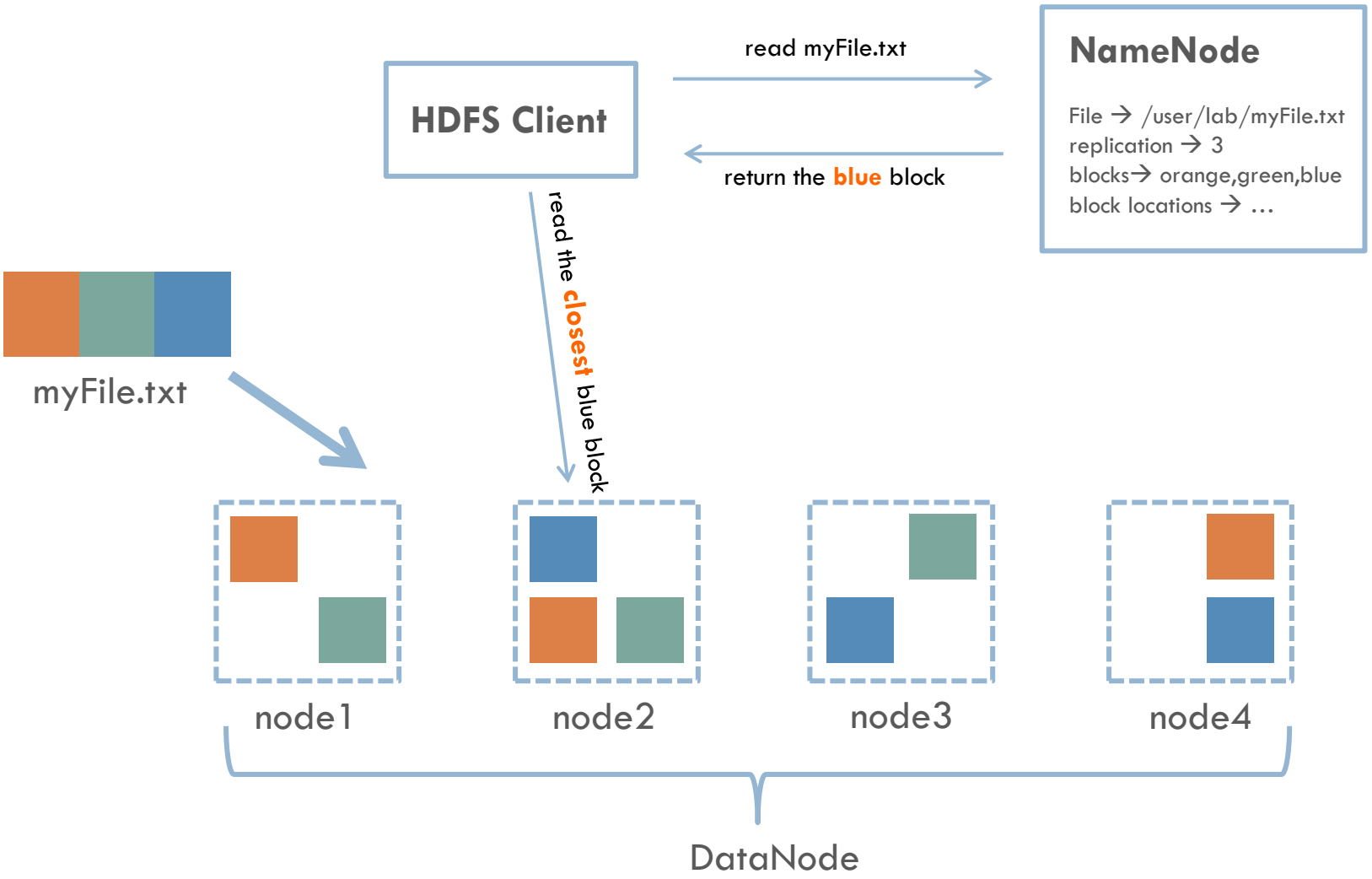
1. Intro to Big Data
2. Distributed Computing and MapReduce
3. Hadoop Ecosystem
4. Programming Hive
5. Advanced Hive
6. Programming Pig
7. Advanced Pig
8. Big Data Use Cases: Location Intelligence and Marketing Analytics
9. Big Data Use Cases: Recommendation Engine and Computational Advertising
10. Hadoop In Action: Building Data Pipelines
11. Beyond Hadoop: Spark
12. Beyond Hadoop: Real-Time Analytics

# Lecture 3 Recap

Catch-up Session
HDFS Architecture
MapReduce Architecture
RHadoop
Importance of Data Preparation

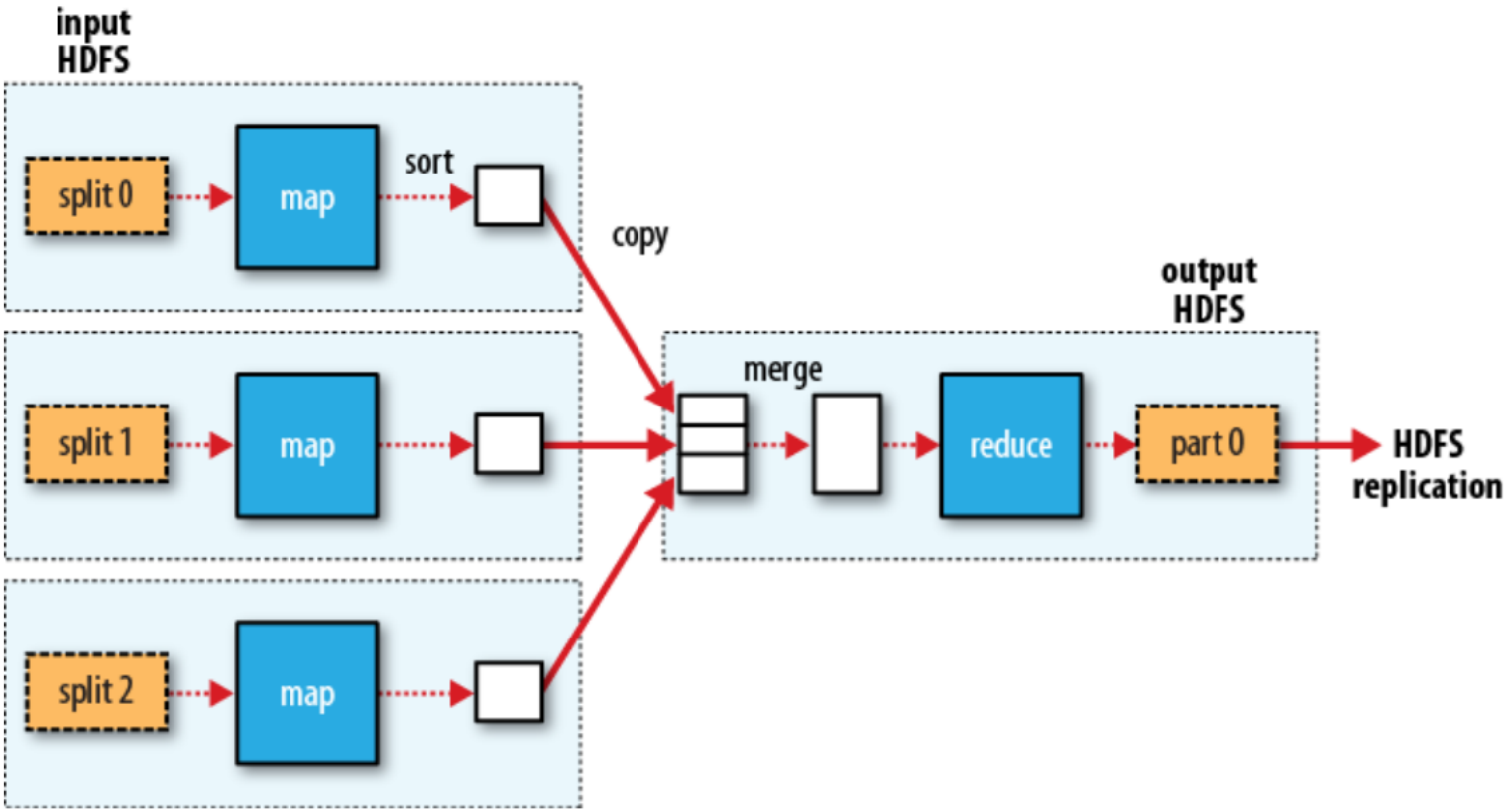
# Anatomy of a File Read





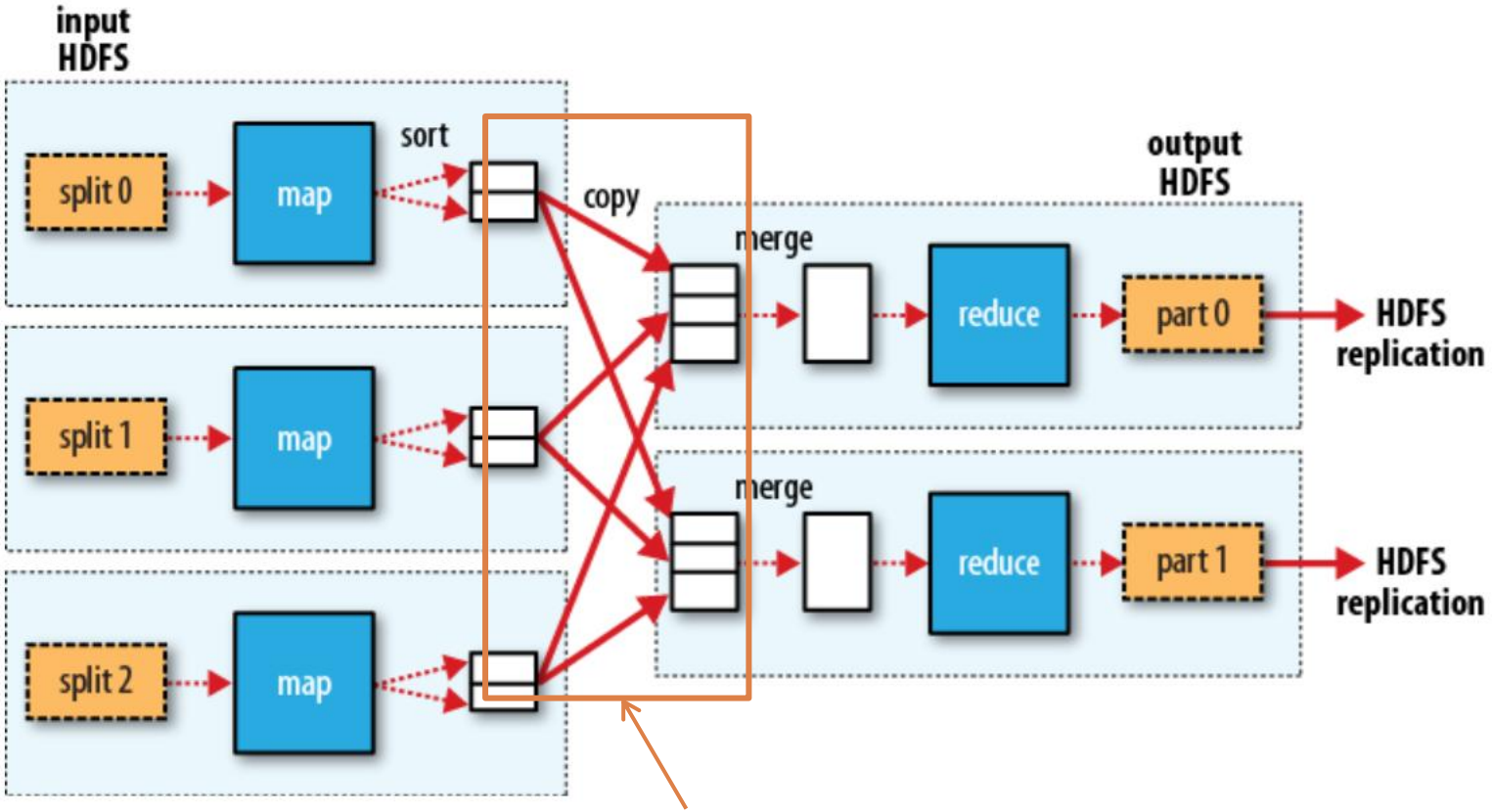
# MapReduce – 3 mappers, 1 reducers





# MapReduce – 3 mappers, 2 reducers

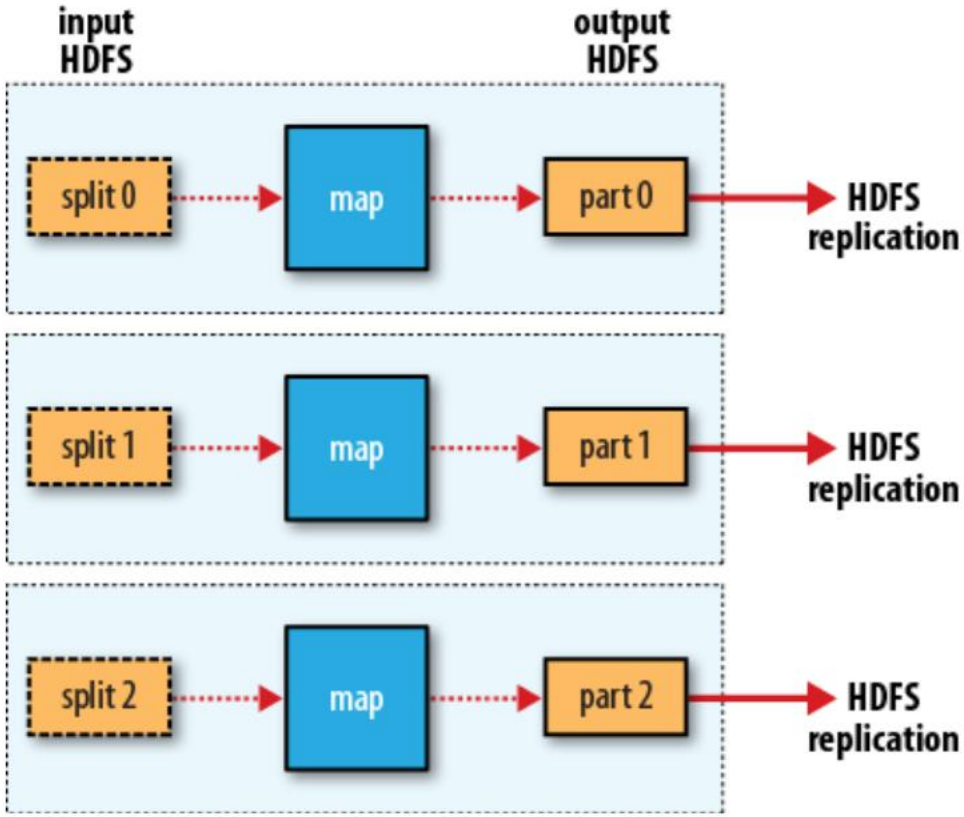





Reduce tasks are expensive due to data movements across network

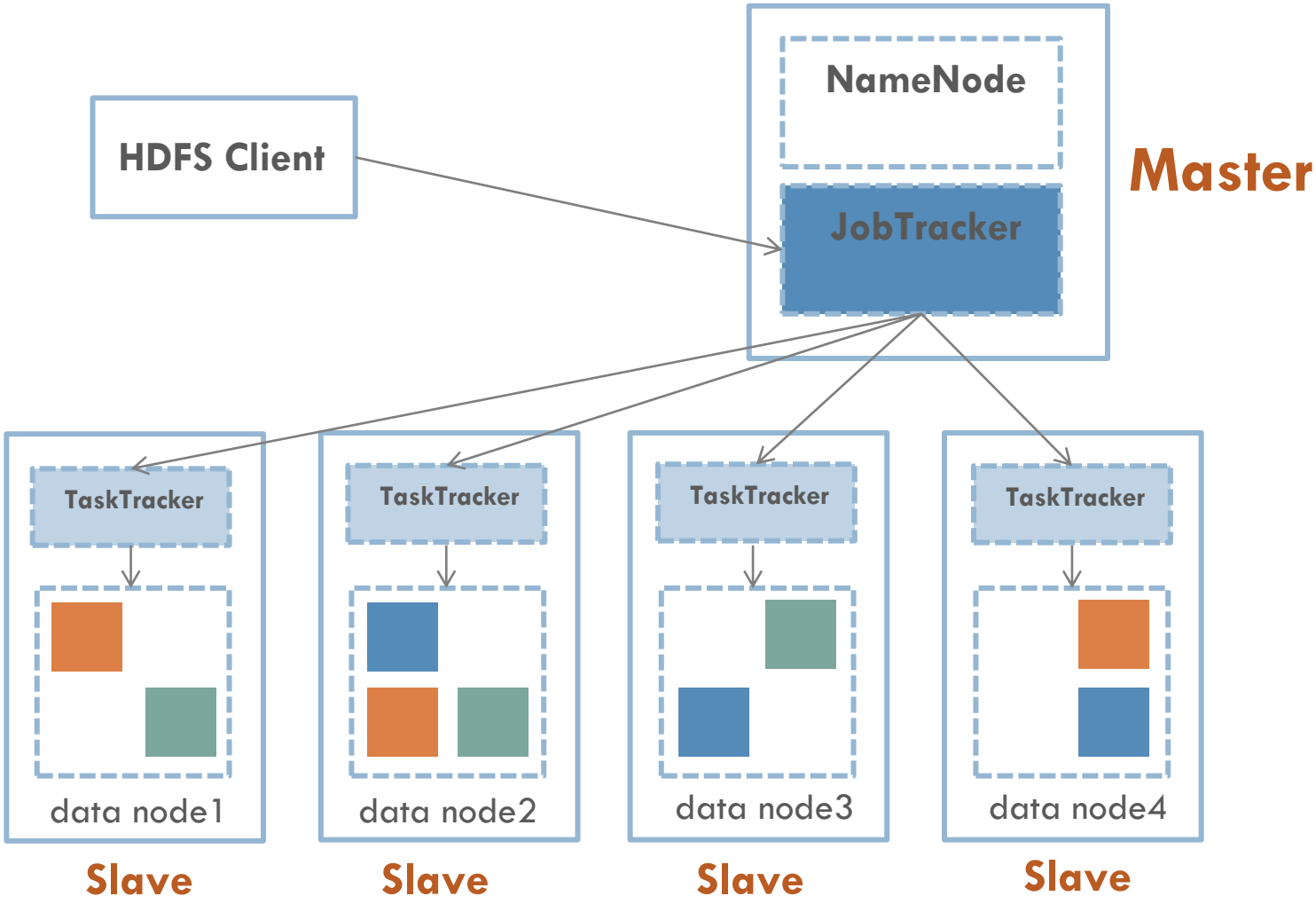
# MapReduce – Map Only Job





# MapReduce Architecture





# R & Hadoop



- RHIPE vs. RHadoop
- ORCH – Oracle R Connector for Hadoop
- RHadoop Wiki
  - https://github.com/RevolutionAnalytics/RHadoop/wiki
- Using RHadoop to predict website visitors
  - http://hortonworks.com/hadoop-tutorial/using-rhadoop-to-predict-visitors-amount/

# Lecture 4 - Outline



- Hive CLI
- Hive Configurations
- Working with Hive Databases
- Creating Hive Table and Load Data into Tables
- Hive Data Type
- Hive Functions (Date, String)
- Hive Queries
  - where statement → filtering data
  - group by → aggregation
  - order → sorting
  - distinct → dedup
  - union
  - join
- Lab & Assignment 2

# Some Data Science Philosophy



- Most of the business questions can be answered with simple data exploration queries
- Always start with simple data exploration analysis and "ship faster"
  - Interaction with business/product managers are important
  - Don't create data science silos
- Your models are fragile
  - Product changes will break your models and algorithms
- Machine learning models are high-maintenance animals
  - Technical debt!
- Data preparation - 80% effort
  - You spend more time on Hive and Pig and even excel ☺
- Algorithms are the easy part – out of box solutions
  - You rarely have to invent new algorithms
  - Coming up with the right question, curating/collecting the right data and preparing the data are critical

Machine Learning: The High Interest Credit Card of Technical Debt: http://research.google.com/pubs/pub43146.html

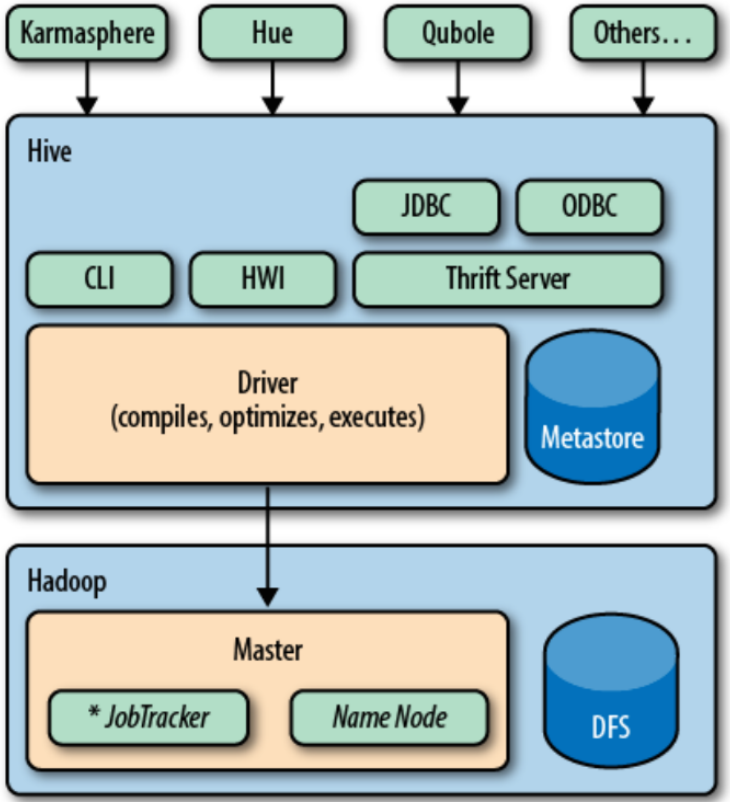
# Demo Data – Geotagged Tweets

| ID | DateTime | Latitude | Longitude | Tweet |
|---|---|---|---|---|
| USER_8d0e8566 | 2010-03-02T23:00:44 | 30.387524 | -91.109663 | Pre-workout prep has begun. |
| USER_8d0e8566 | 2010-04-02T23:04:20 | 30.387524 | -91.109663 | I really don't like that a college FB player's ON-FIELD production can be negated by a single workout. So proof's NOT in the pudding? |
| USER_87b48222 | 2010-03-02T23:23:29 | 37.530819 | -77.475577 | @USER_9bb099c2 15 pages??? fuck u mean!!?? damn. |
| USER_87b48222 | 2010-07-02T23:43:57 | 37.530819 | -77.475577 | @USER_e97d1292 lol do u know that song? |
| USER_01b8a291 | 2010-03-03T00:56:16 | 41.51179 | -95.893286 | HAHAHA OMG! I just found a baggie of weed that I hid from like four/five years ago!! Hahahaha!!! |
| USER_2e5f8774 | 2010-03-03T02:06:15 | 39.669307 | -79.85002 | @USER_2b2bd61b light skin free way and shit...lol Look like you sell bean pies |
| USER_942c68df | 2010-03-03T02:21:36 | 41.220425 | -85.861873 | These judges are being hard this year. |
| USER_8d0e8566 | 2010-04-03T02:28:12 | 30.387524 | -91.109663 | @USER_b7cdabe3 People don't dance like that to get the burn anymore. Its frowned upon..LOL. |
| USER_8d0e8566 | 2010-03-03T02:29:39 | 30.399934 | -91.121502 | RT @USER_9c9e75e2: Officially getting rid of my iPhone with its dysfunctional button this weekend\|Get a 9700 #BlackertheBerrytheSweetertheUse |
| USER_2e5f8774 | 2010-04-03T02:42:44 | 39.669307 | -79.85002 | @USER_7ac8dee6 Hey Cuz...Where u been at? |
| USER_8d0e8566 | 2010-05-03T02:43:01 | 30.393485 | -91.110458 | RT @USER_9c9e75e2: @USER_8d0e8566 I think that's the move!\|Make it happen and we can play Word Mole against each other. |
| USER_8d0e8566 | 2010-03-03T02:53:19 | 30.393485 | -91.110458 | @USER_b7cdabe3 Oh, okay! |
| USER_942c68df | 2010-07-03T02:55:36 | 41.234181 | -85.812994 | @USER_20c15b69 Me too. |
| USER_8d0e8566 | 2010-06-03T03:00:37 | 30.387524 | -91.109663 | The next 2hrs of tweets are @USER_fe579e73 for gibing me the idea with his #theory tweet |
| USER_942c68df | 2010-03-03T03:14:53 | 41.234181 | -85.812994 | @USER_21fe08ea Aww that sucks. If ya dont mind me asking, whats ruining your relationship? |
| USER_8d0e8566 | 2010-05-03T03:26:46 | 30.387524 | -91.109663 | @USER_fe579e73 did u change ur settings to use twitlonger? |
| USER_8d0e8566 | 2010-03-03T03:29:41 | 30.387524 | -91.109663 | RT @USER_de057bc2: Twitter is jacked up tonight\|Just on iPhones. #BlackertheBerrytheSweetertheUse |
| USER_8d0e8566 | 2010-03-03T03:33:47 | 30.387524 | -91.109663 | RT @USER_de057bc2: @USER_8d0e8566 EFF YO Blackberry\|Sore Loser |
| USER_8d0e8566 | 2010-06-03T03:47:43 | 30.387524 | -91.109663 | @USER_45b5c066 @USER_2b5b12ff The body nice but that had to be a contest at a Bukket Nekked. |
| USER_8d0e8566 | 2010-06-03T03:57:23 | 30.387524 | -91.109663 | #PeterWisdom "If u wake up and ur gal or the gal ur in bed with is staring at u,take solace in knowing she'll be sleep when u escape." LOL |
| USER_87b48222 | 2010-03-03T03:59:01 | 37.530819 | -77.475577 | Where do you those rip away jeans?!! @USER_af454d84 and where can I get some?! |
| USER_8d0e8566 | 2010-03-03T04:17:29 | 30.387524 | -91.109663 | @USER_b7cdabe3 LOL |
| USER_8d0e8566 | 2010-03-03T04:37:07 | 30.387524 | -91.109663 | RT @USER_45b5c066: #FamilyGuy Meg and Brian make out. Meg stalks him like Misery << did u just use a shag blog term?? #CLASSIC\|Did I? |

# Hive Environment



- CLI – Command Line Interface
- HWI – Hive Web Interface
- JDBC, ODBC, Thrift Server – Hive programmatic access
- Driver
  - Query interpretation
  - Query optimization
  - Execution via MapReduce
- Metastore
  - Keeps table schemas and other metadata
  - Usually MySQL database



DEMO

# Hive CLI



| Hive | Example |
|---|---|
| Start Hive CLI | $ hive |
| Executive one-off Hive query | $ hive –e "select count(*) from table;" |
| Execute Hive queries saved in a file | $ hive –f /home/lab/count_stats.hql |
| Run shell commands from Hive CLI | **hive**> ! ls -alF /home/lab/ ; |
| Run Hadoop shell commands from Hive CLI | **hive**> dfs -ls /user/lab/shakespeare ; |
| Execute a script file inside CLI | **hive**> source /home/lab/test.hql |
| Exit Hive CLI | **hive**> quit; |
| Force quit | Ctrl + Z |

DEMO

# Hive Environment Configuration



| Hive | Example |
| --- | --- |
| List hive settings | $ hive<br>hive> set; |
| Find hive settings | $ hive –e "set;" \| grep 'hive.cli.print.header' |
| Change/set environment parameter | hive> set hive.cli.print.header=true |
| Run initialization script | $ hive –i '/home/lab/hive-initialization-script.hql' |

DEMO

# Hive Environment Configuration



- Check your hive warehouse location

```
[root@sandbox lab]# hive -e "set -v;" | grep hive.metastore.warehouse.dir
15/01/29 21:38:48 WARN conf.HiveConf: HiveConf of name hive.optimize.mapjoin.mapreduce does not exist
15/01/29 21:38:48 WARN conf.HiveConf: HiveConf of name hive.heapsize does not exist
15/01/29 21:38:48 WARN conf.HiveConf: HiveConf of name hive.server2.enable.impersonation does not exist
15/01/29 21:38:48 WARN conf.HiveConf: HiveConf of name hive.auto.convert.sortmerge.join.noconditionaltask

Logging initialized using configuration in file:/etc/hive/conf/hive-log4j.properties
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hadoop/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hive/lib/hive-jdbc-0.14.0.2.2.0.0-2041-standalone.
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hive.metastore.warehouse.dir=/apps/hive/warehouse
[root@sandbox lab]#
```

# *Working with Hive Databases*

# Working with Hive Databases



| Hive | Example |
| --- | --- |
| List existing hive databases | hive> SHOW DATABASES; |
| Create a new database | hive> CREATE DATABASE twitter; |
| Create a new database | hive> CREATE DATABASE IF NOT EXISTS twitter; |
| Enter a default database | hive> USE twitter; |
| Show metadata of a database | hive> DESCRIBE DATABASE twitter; |
| Show extended details of a database | hive> DESCRIBE DATABASE EXTENDED twitter; |

DEMO

```
[root@sandbox lab]# pwd
/home/lab
[root@sandbox lab]# hadoop fs -mkdir /user/lab/twitter
[root@sandbox lab]# hadoop fs -put full_text.txt /user/lab/twitter/
[root@sandbox lab]# hive
15/01/29 20:39:50 WARN conf.HiveConf: HiveConf of name hive.optimize.mapjoin.mapreduce does not exist
15/01/29 20:39:50 WARN conf.HiveConf: HiveConf of name hive.heapsize does not exist
15/01/29 20:39:50 WARN conf.HiveConf: HiveConf of name hive.server2.enable.impersonation does not exist
15/01/29 20:39:50 WARN conf.HiveConf: HiveConf of name hive.auto.convert.sortmerge.join.noconditionaltask

Logging initialized using configuration in file:/etc/hive/conf/hive-log4j.properties
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hadoop/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/imp
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hive/lib/hive-jdbc-0.14.0.2.2.0.0-2041-standalone.
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hive> show databases;
OK
default
foursquare
sz
xademo
Time taken: 3.094 seconds, Fetched: 4 row(s)
hive> create database twitter;
OK
Time taken: 2.049 seconds
hive> use twitter;
hive> show databases;
OK
default
foursquare
sz
twitter
xademo
Time taken: 0.042 seconds, Fetched: 5 row(s)
hive> describe database extended twitter;
OK
twitter         hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db       root    USER
Time taken: 0.078 seconds, Fetched: 1 row(s)
```

# *Working with Hive Tables*

# CREATE TABLE & LOAD DATA



| HiveQL Description | Script |
|---|---|
| **Create tab-delimited table** | *create table full_text (*<br>*id string,*<br>*ts string,*<br>*lat_lon string,*<br>*lat string,*<br>*lon string,*<br>*tweet string)*<br>*row format delimited*<br>*fields terminated by '\t' ;* |
| **Load data in HDFS into hive table** | *load data inpath '/user/lab/twitter/'*<br>*overwrite into table twitter.full_text;* |
| **Create table and load table in one query** | *create table full_text_2 as*<br>*select **<br>*from twitter.full_text;* |

DEMO

```
hive> use twitter;
OK
Time taken: 1.455 seconds
hive> show tables;
OK
Time taken: 0.257 seconds
hive> create table full_text (
    >           id string,
    >           ts string,
    >           lat_lon string,
    >           lat string,
    >           lon string,
    >           tweet string)
    > row format delimited
    > fields terminated by '\t' ;
OK
Time taken: 0.427 seconds
hive> describe full_text;
OK
id                      string
ts                      string
lat_lon                 string
lat                     string
lon                     string
tweet                   string
Time taken: 0.588 seconds, Fetched: 6 row(s)
hive> describe extended full_text;
OK
id                      string
ts                      string
lat_lon                 string
lat                     string
lon                     string
tweet                   string

Detailed Table Information      Table(tableName:full_text, dbName:twitter, owner:root, createTime:1422567700, lastAccessTime:0, retention
comment:null), FieldSchema(name:ts, type:string, comment:null), FieldSchema(name:lat_lon, type:string, comment:null), FieldSchema(name:la
 comment:null), FieldSchema(name:tweet, type:string, comment:null)], location:hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twi
Format, outputFormat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compressed:false, numBuckets:-1, serdeInfo:SerDeInfo(nam
impleSerDe, parameters:{serialization.format=    , field.delim=
Time taken: 0.572 seconds, Fetched: 8 row(s)
hive> dfs -ls /apps/hive/warehouse
    > ;
Found 8 items
drwxr-xr-x   - root hdfs          0 2015-01-15 06:10 /apps/hive/warehouse/foursquare.db
drwxr-xr-x   - hue  hdfs          0 2014-12-16 19:27 /apps/hive/warehouse/sample_07
drwxr-xr-x   - hue  hdfs          0 2014-12-16 19:27 /apps/hive/warehouse/sample_08
drwxr-xr-x   - hive hdfs          0 2015-01-09 22:25 /apps/hive/warehouse/sz.db
drwxr-xr-x   - root hdfs          0 2015-01-15 05:09 /apps/hive/warehouse/tweets1
drwxr-xr-x   - root hdfs          0 2015-01-29 21:41 /apps/hive/warehouse/twitter.db
drwxr-xr-x   - root hdfs          0 2015-01-24 05:39 /apps/hive/warehouse/wordcount
drwxr-xr-x   - hive hdfs          0 2014-12-16 19:43 /apps/hive/warehouse/xademo.db
hive> dfs -ls /apps/hive/warehouse/twitter.db;
Found 1 items
drwxr-xr-x   - root hdfs          0 2015-01-29 21:41 /apps/hive/warehouse/twitter.db/full_text
```

# DEMO

```
hive> dfs -ls /user/lab/twitter/;
Found 1 items
-rw-r--r--   1 root hdfs   57139942 2015-01-29 21:45 /user/lab/twitter/full_text.txt
hive> load data inpath '/user/lab/twitter/'
    > overwrite into table twitter.full_text;
Loading data to table twitter.full_text
Table twitter.full_text stats: [numFiles=1, numRows=0, totalSize=57139942, rawDataSize=0]
OK
Time taken: 4.424 seconds
hive> select * from twitter.full_text limit 5;
OK
USER_79321756   2010-03-03T04:15:26     ÜT: 47.528139,-122.197916       47.528139       -122.197916     RT @USER_2ff4faca: IF SHE DO IT 1 MO
&gt;haha. #cutthatout
USER_79321756   2010-03-03T04:55:32     ÜT: 47.528139,-122.197916       47.528139       -122.197916     @USER_77a4822d @USER_2ff4faca okay:)
USER_79321756   2010-03-03T05:13:34     ÜT: 47.528139,-122.197916       47.528139       -122.197916     RT @USER_5d4d777a: YOURE A FAG FOR G
OU ? A FUCKING NOBODY !!!!&gt;&gt;Lol! Dayum! Aye!
USER_79321756   2010-03-03T05:28:02     ÜT: 47.528139,-122.197916       47.528139       -122.197916     @USER_77a4822d yea ok..well answer t
USER_79321756   2010-03-03T05:56:13     ÜT: 47.528139,-122.197916       47.528139       -122.197916     A sprite can disappear in her mouth
Time taken: 1.179 seconds, Fetched: 5 row(s)
hive> describe twitter.full_text;
OK
id                      string
ts                      string
lat_lon                 string
lat                     string
lon                     string
tweet                   string
Time taken: 0.645 seconds, Fetched: 6 row(s)
hive> select id, tweet from twitter.full_text limit 5;
OK
USER_79321756   RT @USER_2ff4faca: IF SHE DO IT 1 MORE TIME.....IMA KNOCK HER DAMN KOOFIE OFF.....ON MY MOMMA&gt;&gt;haha. #cutthatout
USER_79321756   @USER_77a4822d @USER_2ff4faca okay:) lol. Saying ok to both of yall about to different things!:*
USER_79321756   RT @USER_5d4d777a: YOURE A FAG FOR GETTING IN THE MIDDLE OF THIS @USER_ab059bdc WHO THE FUCK ARE YOU ? A FUCKING NOBODY !!!!
USER_79321756   @USER_77a4822d yea ok..well answer that cheap as Sweden phone you came up on when I call.
USER_79321756   A sprite can disappear in her mouth - lil kim hmmmmm the can not the bottle right?
Time taken: 0.227 seconds, Fetched: 5 row(s)
```

DEMO

```
hive> use twitter;
OK
Time taken: 0.316 seconds
hive> show tables;
OK
full_text
Time taken: 0.173 seconds, Fetched: 1 row(s)
hive> create table full_text_2 as
    > select *
    > from full_text;
Query ID = root_20150129221111_251202e9-8d5c-4486-8be4-234526f5e3d6
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1422563374808_0001, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application_1
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2015-01-29 22:12:12,846 Stage-1 map = 0%,  reduce = 0%
2015-01-29 22:12:27,378 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.94 sec
MapReduce Total cumulative CPU time: 7 seconds 940 msec
Ended Job = job_1422563374808_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://sandbox.hortonworks.com:8020/tmp/hive/root/0ea5beb3-d499-458c-a255-85a90b63a799/hive_2
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db/full_text_2
Table twitter.full_text_2 stats: [numFiles=1, numRows=377616, totalSize=57139942, rawDataSize=56762326]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 7.94 sec   HDFS Read: 57140186 HDFS Write: 57140027 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 940 msec
OK
Time taken: 47.771 seconds
hive> dfs -ls /apps/hive/warehouse/twitter.db
    > ;
Found 2 items
drwxr-xr-x   - root hdfs          0 2015-01-29 21:45 /apps/hive/warehouse/twitter.db/full_text
drwxr-xr-x   - root hdfs          0 2015-01-29 22:12 /apps/hive/warehouse/twitter.db/full_text_2
```

# *Hive Data Types*

Hive Language Manual - Hive Data Types
https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types

# Hive Primitive Data Types



| Type | Size | Literal syntax examples |
|---|---|---|
| TINYINT | 1 byte signed integer. | 20 |
| SMALLINT | 2 byte signed integer. | 20 |
| INT | 4 byte signed integer. | 20 |
| BIGINT | 8 byte signed integer. | 20 |
| BOOLEAN | Boolean true or false. | TRUE |
| FLOAT | Single precision floating point. | 3.14159 |
| DOUBLE | Double precision floating point. | 3.14159 |
| STRING | Sequence of characters. The character set can be specified. Single or double quotes can be used. | 'Now is the time',"for all good men" |
| TIMESTAMP (v0.8.0+) | Integer, float, or string. | 1327882394 (Unix epoch seconds), 1327882394.123456789 (Unix epoch seconds plus nanoseconds), and '2012-02-03 12:34:56.123456789' (JDBC-compliant java.sql.Timestamp format) |
| BINARY (v0.8.0+) | Array of bytes. | See discussion below |

# *Hive Functions*

Hive Language Manual – Hive Functions
https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF#LanguageManualUDF-DateFunctions

# Hive Functions

DEMO



- String functions
- Date functions
- Condition functions
- Math functions

```
[root@sandbox lab]# hive -e "show functions" | tail -15
OK
Time taken: 4.691 seconds, Fetched: 196 row(s)
weekofyear
when
windowingtablefunction
xpath
xpath_boolean
xpath_double
xpath_float
xpath_int
xpath_long
xpath_number
xpath_short
xpath_string
year
|
~
[root@sandbox lab]# hive -e "describe function weekofyear;"
OK
weekofyear(date) - Returns the week of the year of the given
 date. A week is considered to start on a Monday and week 1
is the first week with >3 days.
Time taken: 4.551 seconds, Fetched: 1 row(s)
```

# Date Functions

DEMO

## Convert Datetime String to Timestamp

```
hive> create table full_text_ts as
    > select id, cast(concat(substr(ts,1,10), ' ', substr(ts,12,8)) as timestamp) as ts, lat, lon, tweet
    > from full_text;
Query ID = root_20150129224949_1a13429f-c16b-440b-be25-eb87d2ac3b3b
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1422563374808_0004, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application_
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2015-01-29 22:49:38,597 Stage-1 map = 0%,  reduce = 0%
2015-01-29 22:49:58,996 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.47 sec
MapReduce Total cumulative CPU time: 13 seconds 470 msec
Ended Job = job_1422563374808_0004
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://sandbox.hortonworks.com:8020/tmp/hive/root/0ea5beb3-d499-458c-a255-85a90b63a799/hive_
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db/full_text_ts
Table twitter.full_text_ts stats: [numFiles=1, numRows=377616, totalSize=47273124, rawDataSize=46895508]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 13.47 sec   HDFS Read: 57140186 HDFS Write: 47273210 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 470 msec
OK
Time taken: 36.014 seconds
hive> describe full_text_ts;
OK
id                      string
ts                      timestamp
lat                     string
lon                     string
tweet                   string
Time taken: 0.651 seconds, Fetched: 5 row(s)
```

# Date Functions

DEMO

## *Extract year, month and day from timestamp*



```
hive> select ts, unix_timestamp(ts) as unix_timestamp, to_date(ts) as date, year(ts) as year, month(ts) as month, day(ts) as day
    > from full_text_ts
    > limit 10;
OK
2010-03-03 04:15:26     1267589726      2010-03-03      2010    3       3
2010-03-03 04:55:32     1267592132      2010-03-03      2010    3       3
2010-03-03 05:13:34     1267593214      2010-03-03      2010    3       3
2010-03-03 05:28:02     1267594082      2010-03-03      2010    3       3
2010-03-03 05:56:13     1267595773      2010-03-03      2010    3       3
2010-03-03 16:52:44     1267635164      2010-03-03      2010    3       3
2010-03-03 16:57:24     1267635444      2010-03-03      2010    3       3
2010-03-03 20:20:40     1267647640      2010-03-03      2010    3       3
2010-03-03 23:23:33     1267658613      2010-03-03      2010    3       3
2010-03-03 23:37:36     1267659456      2010-03-03      2010    3       3
Time taken: 0.19 seconds, Fetched: 10 row(s)
```

DEMO

# String Functions

```
hive> select id, ts, trim(lower(tweet)) as tweet
    > from full_text_ts
    > limit 5;
OK
USER_79321756   2010-03-03 04:15:26     rt @user_2ff4faca: if she do it 1 more time......ima knock her damn koofie off.....on my momma&g
USER_79321756   2010-03-03 04:55:32     @user_77a4822d @user_2ff4faca okay:) lol. saying ok to both of yall about to different things!:*
USER_79321756   2010-03-03 05:13:34     rt @user_5d4d777a: youre a fag for getting in the middle of this @user_ab059bdc who the fuck are
USER_79321756   2010-03-03 05:28:02     @user_77a4822d yea ok..well answer that cheap as sweden phone you came up on when i call.
USER_79321756   2010-03-03 05:56:13     a sprite can disappear in her mouth - lil kim hmmmmm the can not the bottle right?
Time taken: 0.206 seconds, Fetched: 5 row(s)
hive> select id, ts, trim(upper(tweet)) as tweet
    > from full_text_ts
    > limit 5;
OK
USER_79321756   2010-03-03 04:15:26     RT @USER_2FF4FACA: IF SHE DO IT 1 MORE TIME......IMA KNOCK HER DAMN KOOFIE OFF.....ON MY MOMMA&G
USER_79321756   2010-03-03 04:55:32     @USER_77A4822D @USER_2FF4FACA OKAY:) LOL. SAYING OK TO BOTH OF YALL ABOUT TO DIFFERENT THINGS!:*
USER_79321756   2010-03-03 05:13:34     RT @USER_5D4D777A: YOURE A FAG FOR GETTING IN THE MIDDLE OF THIS @USER_AB059BDC WHO THE FUCK ARE
USER_79321756   2010-03-03 05:28:02     @USER_77A4822D YEA OK..WELL ANSWER THAT CHEAP AS SWEDEN PHONE YOU CAME UP ON WHEN I CALL.
USER_79321756   2010-03-03 05:56:13     A SPRITE CAN DISAPPEAR IN HER MOUTH - LIL KIM HMMMMM THE CAN NOT THE BOTTLE RIGHT?
Time taken: 0.205 seconds, Fetched: 5 row(s)
hive> select id, ts, length(tweet) as length
    > from full_text_ts
    > limit 5;
OK
USER_79321756   2010-03-03 04:15:26     119
USER_79321756   2010-03-03 04:55:32     96
USER_79321756   2010-03-03 05:13:34     148
USER_79321756   2010-03-03 05:28:02     89
USER_79321756   2010-03-03 05:56:13     82
Time taken: 0.198 seconds, Fetched: 5 row(s)
hive> select id, ts, sentences(tweet) as tokens
    > from full_text_ts
    > limit 5;
OK
USER_79321756   2010-03-03 04:15:26     [["RT","USER","2ff4faca","IF","SHE","DO","IT","1","MORE","TIME","IMA","KNOCK","HER","DAMN","KOOF
USER_79321756   2010-03-03 04:55:32     [["USER","77a4822d","USER","2ff4faca","okay","lol"],["Saying","ok","to","both","of","yall","abou
USER_79321756   2010-03-03 05:13:34     [["RT","USER","5d4d777a","YOURE","A","FAG","FOR","GETTING","IN","THE","MIDDLE","OF","THIS","USER
],["gt","gt","Lol"],["Dayum"],["Aye"]]
USER_79321756   2010-03-03 05:28:02     [["USER","77a4822d","yea","ok","well","answer","that","cheap","as","Sweden","phone","you","came"
USER_79321756   2010-03-03 05:56:13     [["A","sprite","can","disappear","in","her","mouth","lil","kim","hmmmmm","the","can","not","the"
Time taken: 0.203 seconds, Fetched: 5 row(s)
```

# String Functions

DEMO

## *Find twitter handles mentioned in a tweet*



```
hive> select id, ts, regexp_extract(lower(tweet), '(.*)@user_(\\S{8})([:| ])(.*)', 2) as patterns
    > from full_text_ts
    > limit 5;
OK
USER_79321756   2010-03-03 04:15:26     2ff4faca
USER_79321756   2010-03-03 04:55:32     2ff4faca
USER_79321756   2010-03-03 05:13:34     ab059bdc
USER_79321756   2010-03-03 05:28:02     77a4822d
USER_79321756   2010-03-03 05:56:13
Time taken: 0.197 seconds, Fetched: 5 row(s)
```

# String Functions

DEMO

## Find users who like to tweet long sentences



```
hive> select t.id, t.len, t.tweet
    > from (select id, tweet, length(tweet) as len from full_text_ts) t
    > order by len desc
    > limit 10;
Query ID = root_20150130024040_ceb702a7-b210-4d34-a361-333349a491d0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1422563374808_0010, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application_1422563374808_0010/
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-01-30 02:40:16,795 Stage-1 map = 0%,  reduce = 0%
2015-01-30 02:40:29,180 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.3 sec
2015-01-30 02:40:40,229 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.66 sec
MapReduce Total cumulative CPU time: 8 seconds 660 msec
Ended Job = job_1422563374808_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.66 sec   HDFS Read: 47273366 HDFS Write: 2511 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 660 msec
OK
USER_f02c28a6   288     #FF @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014
14dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9
USER_79daf9ff   252     #FF the wife @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @US
@USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0
USER_6113e677   247     Follow: @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_61
_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677
USER_a0ec1a93   232     RT @USER_6172fd58: #FF @USER_1545c105 @USER_1545c105 @USER_1545c105 @USER_1545c105 @USER_1545c105 @USER_1545c105 @USER_1545c105 @USER_15
_1545c105 @USER_1545c105 @USER_1545c105 @USER_1545c105 @USER_1545c105 @USER_1545c105
USER_dbf87c8d   228     #FF @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958
958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7 @USER_65d958f7
USER_54d59141   224     One of my fav tweepz @USER_022b3e2b FOLLOW NOW ----> @USER_022b3e2b @USER_022b3e2b @USER_022b3e2b @USER_022b3e2b @USER_022b3e2b @USER_02
t @USER_022b3e2b @USER_022b3e2b @USER_022b3e2b @USER_022b3e2b @USER_022b3e2b @USER_022b3e2b
USER_82d93fd1   217     #FF DJs @USER_5fb6ea48 @USER_c2e6fa2d @USER_1c2fa926 @USER_f0768d3a @USER_754fc8d7 @USER_3e0a952b @USER_1bf1a9cc @USER_64938b2e @USER_29
_33a22d08 @USER_139b2a1c @USER_e43d3399
USER_e970a234   213     #FF @USER_23a84292 @USER_c93efdf6 @USER_152054b5 @USER_bdd53ac4 @USER_28c944d6 @USER_2075c9ca @USER_255a54be @USER_a33730a0 @USER_ff04d2
7473f @USER_7331f119 @USER_0e6656e0
USER_915e3f8c   209     @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @
6 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966 @USER_e7961966
USER_cd1c16ea   209     #FF&gt;&gt;&gt;&gt; my niggas @USER_b05bd56f @USER_9ad7c4d9 @USER_46a17eb2 @USER_32e22fb4 @USER_b8878efc @USER_b1d0a7af @USER_bc270513 @
d @USER_46a17eb2 @USER_a657080e
Time taken: 38.763 seconds, Fetched: 10 row(s)
```

Hmmmmm… need better filters. Twitter has a 140 char limit

# String Functions

DEMO

## *Find users who like to tweet long sentences*

```
hive> select t.id, t.len, t.trimmed_tweet
    > from (select id, regexp_replace(tweet, "@USER_\\w{8}", " ") as trimmed_tweet, length(regexp_replace(tweet, "@USER_\\w{8}", " "))
 as len from full_text_ts) t
    > order by len desc
    > limit 10;
Query ID = root_20150130024444_daa82803-fbd5-4956-ab4a-627aaf962cb6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1422563374808_0011, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application_1422563374808_0011/
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-01-30 02:45:05,881 Stage-1 map = 0%,  reduce = 0%
2015-01-30 02:45:21,574 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.7 sec
2015-01-30 02:45:31,758 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.07 sec
MapReduce Total cumulative CPU time: 11 seconds 70 msec
Ended Job = job_1422563374808_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.07 sec   HDFS Read: 47273366 HDFS Write: 1677 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 70 msec
OK
USER_2f88c77e   158     RT@MZ_BURSON_420 RT@c_dubble_u: FL . . Cali . . Texas? Help me out yall where should I go to live?&lt;&lt;&lt;
&lt;&lt; CALI FOR SURE whre u at now &gt;&gt; VA
USER_3ca302f1   151     Shout out to all my farmers buly, linden, northside niggas, & all my rosedale niggas &lt;&lt; I love yal & mis
s yal niggas &gt;&gt; NO HOMO! B Bak soon
USER_2c8d1305   150     #ff  &lt;&lt;&lt;&lt;&lt;- this girl has been by my side for the past 9yrs!!! She's an awesome friend!! I woul
dn't trade her for NOTHING in the world!
USER_3ca302f1   150     MY MOTTO &gt;&gt; GET THE MONEY, THE CAKE, THE GUAP, THE FEDDI, THE SCRATCH. THE DOE, THE MULA, THE BREAD, THE
 BEJAMINS &lt;&lt; wateva u wana call it
```

Better!

# Conditional Function

DEMO

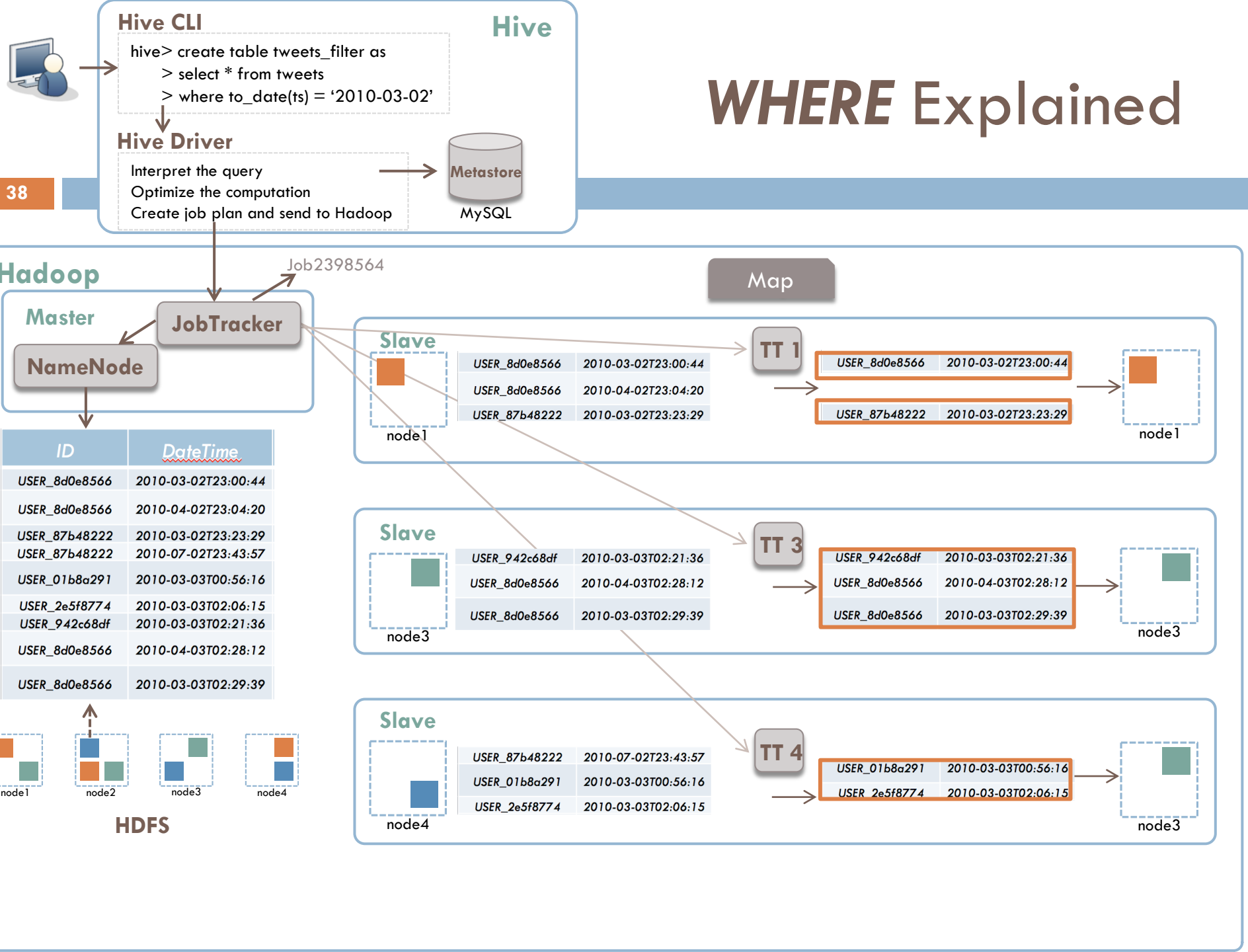
## *Find users who like to tw-eating*



```
hive> select * from
    >     (select id, ts, case when hour(ts) = 7 then 'breakfast'
    >                          when hour(ts) = 12 then 'lunch'
    >                          when hour(ts) = 19 then 'dinner'
    >                     end as tw_eating,
    >             lat, lon
    >     from full_text_ts) t
    > where t.tw_eating in ('breakfast','lunch','dinner')
    > limit 10;
OK
USER_79321756   2010-03-04 07:12:13     breakfast       47.528139       -122.197916
USER_79321756   2010-03-04 07:12:22     breakfast       47.528139       -122.197916
USER_79321756   2010-03-05 07:10:00     breakfast       47.528139       -122.197916
USER_79321756   2010-03-05 07:33:45     breakfast       47.528139       -122.197916
USER_79321756   2010-03-05 07:57:25     breakfast       47.528139       -122.197916
USER_79321756   2010-03-05 07:58:56     breakfast       47.528139       -122.197916
USER_79321756   2010-03-05 07:59:39     breakfast       47.528139       -122.197916
USER_79321756   2010-03-05 19:37:52     dinner  47.528139       -122.197916
USER_79321756   2010-03-07 19:27:02     dinner  47.528139       -122.197916
USER_6197f95d   2010-03-03 12:01:41     lunch   40.221968       -74.734795
Time taken: 0.201 seconds, Fetched: 10 row(s)
```

# *Hive Queries*

# *WHERE* Explained



# *WHERE Clause*– Filtering Data

DEMO

## *Find all tweets by a user*



```
hive> select id, ts, lat, lon, tweet
    > from full_text_ts
    > where id='USER_ae406f1d';
OK
USER_ae406f1d   2010-03-03 00:29:19     40.729685       -74.006611      @USER_544d2831 SMDH hahahaha!!! And it's still alive?! It
USER_ae406f1d   2010-03-03 00:36:31     40.729685       -74.006611      RT @USER_6d01ce88 @USER_ae406f1d We can push him off the
OOO #GoodOne!!!!!!!
USER_ae406f1d   2010-03-03 03:48:28     40.729685       -74.006611      RT @USER_0dbef2b0: Jeezy is taking NOOOOOO prisoners. I p
orture!! :-p
USER_ae406f1d   2010-03-03 04:42:44     40.729685       -74.006611      I am finally done sorting clothes NOT putting them away T
t friend!!
USER_ae406f1d   2010-03-03 05:21:41     40.729685       -74.006611      Grrr don't u hate when you have a song stuck in ur head b
!
USER_ae406f1d   2010-03-03 17:25:55     40.729685       -74.006611      #nowplaying After 7 - Ready Or Not how did i forget about
 know it baby"!!
USER_ae406f1d   2010-03-03 17:41:44     40.729685       -74.006611      @USER_31ba4341 LMAOOOOOOOOO I cant!! i cant!! hahahaha
USER_ae406f1d   2010-03-03 18:04:29     40.729685       -74.006611      Hands Down!!! The Tony Toni Tone station on Pandora is TH
USER_ae406f1d   2010-03-03 18:30:36     40.729685       -74.006611      @USER_b3f24256 cant sprint transfer all that for you? if
USER_ae406f1d   2010-03-03 19:52:08     40.729685       -74.006611      i hate.... slow work days :(
USER_ae406f1d   2010-03-04 06:18:50     40.729685       -74.006611      RT @USER_a03c275d: @USER_ae406f1d @USER_6d01ce88 I threw
```

# *WHERE Clause*– Filtering Data

## *Calculate # of tweets on a specific date*

DEMO



```
hive> select count(*)
    > from full_text_ts
    > where to_date(ts) = '2010-03-07'
    > ;
Query ID = root_20150130030505_a0a794c2-95ba-42a0-a663-b5d1f035eebf
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1422563374808_0012, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application_1
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-01-30 03:05:22,169 Stage-1 map = 0%,  reduce = 0%
2015-01-30 03:05:36,416 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.11 sec
2015-01-30 03:05:47,665 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.59 sec
MapReduce Total cumulative CPU time: 9 seconds 590 msec
Ended Job = job_1422563374808_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.59 sec   HDFS Read: 47273366 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 590 msec
OK
65166
Time taken: 38.884 seconds, Fetched: 1 row(s)
```

# *WHERE Clause*– Filtering Data

## *Find all tweets tweeted from NYC vicinity*

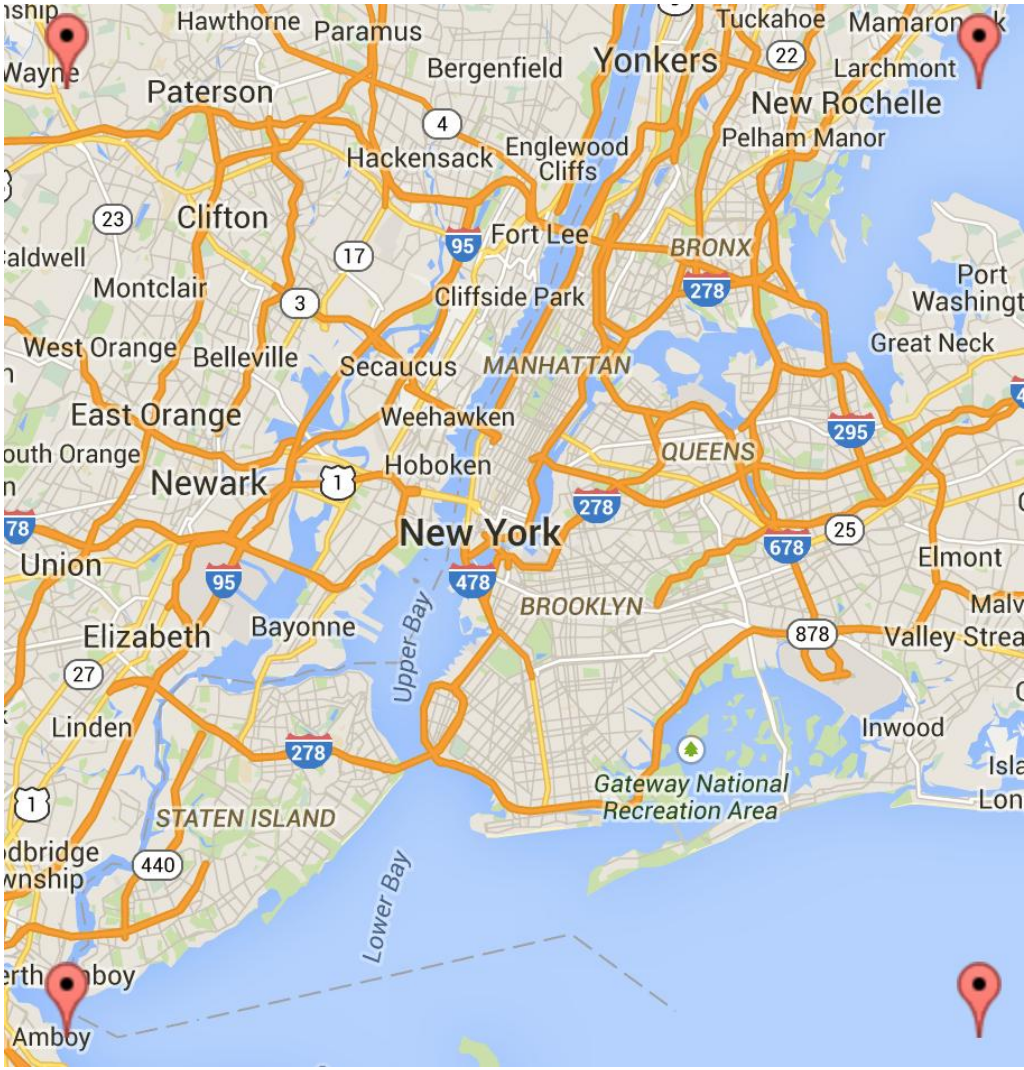


- Find NYC bounding box
  - https://www.flickr.com/places/info/2459115

**Details**

| | |
|---|---|
| Name: | New York |
| WOEID: | 2459115 |
| PlaceType ID | 7 (*Town*) |
| Bounding Box: | -74.2589, 40.4774, -73.7004, 40.9176 |
| Centroid: | -74.0071, 40.7146 |
| Locality: | New York |
| Region: | New York |
| Country: | United States |
| Timezone: | America/New_York |

- Lat-lon Plotter
  - http://www.darrinward.com/lat-long/?id=432657



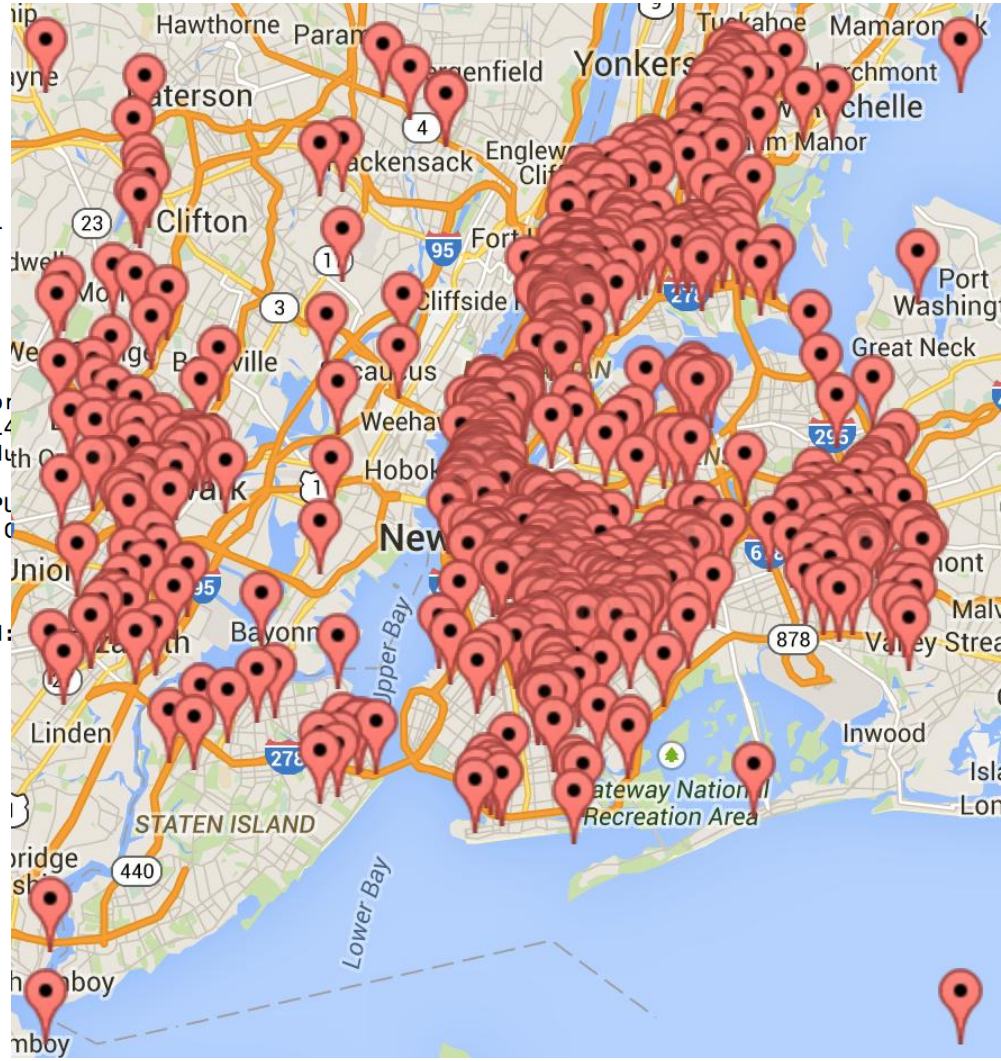
# *WHERE Clause*– Filtering Data

DEMO

## *Find all tweets tweeted from NYC vicinity*



```
hive> select distinct id, lat, lon
    > from full_text_ts
    > where lat > 40.4774 and lat < 40.9176 and
    >       lon > -74.2589 and lon < -73.7004
    > limit 20;
Query ID = root_20150130033232_d754b385-1e46-44b1-b679-28644b098305
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1422563374808_0015, Tracking URL = http://sandbox.hor
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_14
Hadoop job information for Stage-1: number of mappers: 1; number of redu
2015-01-30 03:32:46,534 Stage-1 map = 0%,  reduce = 0%
2015-01-30 03:32:57,671 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU
2015-01-30 03:33:08,659 Stage-1 map = 100%,  reduce = 100%, Cumulative C
MapReduce Total cumulative CPU time: 8 seconds 40 msec
Ended Job = job_1422563374808_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.04 sec   HDFS Read:
Total MapReduce CPU Time Spent: 8 seconds 40 msec
OK
USER_002ac912   40.702701       -73.744705
USER_004aa25a   40.671336       -73.859399
USER_004aa25a   40.706899       -73.99104
USER_004aa25a   40.714269       -74.005972
USER_00839c9b   40.70301        -74.094769
USER_0089c6ed   40.81632        -73.922097
USER_0089c6ed   40.827429       -73.925756
USER_0089c6ed   40.833119       -73.921557
USER_0089c6ed   40.833397       -73.916811
USER_0089c6ed   40.835406       -73.920573
USER_0089c6ed   40.843846       -73.841466
USER_00afb090   40.653857       -73.916457
USER_00bc828f   40.825037       -73.86694
USER_00c2273d   40.675074       -73.910702
```



# *GROUP BY* Explained

| ID | DateTime | Latitude | Longitude | Tweet |
|---|---|---|---|---|
| USER_8d0e8566 | 2010-03-02T23:00:44 | 30.387524 | -91.109663 | Pre-workout prep has begun. |
| USER_8d0e8566 | 2010-04-02T23:04:20 | 30.387524 | -91.109663 | I really don't like that a college FB player's ON-FIELD production can be negated by a single workout. So proof's NOT in the pudding? |
| USER_87b48222 | 2010-03-02T23:23:29 | 37.530819 | -77.475577 | @USER_9bb099c2 15 pages??? fuck u mean!!?? damn. |
| USER_87b48222 | 2010-07-02T23:43:57 | 37.530819 | -77.475577 | @USER_e97d1292 lol do u know that song? |
| USER_01b8a291 | 2010-03-03T00:56:16 | 41.51179 | -95.893286 | HAHAHA OMG! I just found a baggie of weed that I hid from like four/five years ago!! Hahahaha!!! |
| USER_2e5f8774 | 2010-03-03T02:06:15 | 39.669307 | -79.85002 | @USER_2b2bd61b light skin free way and shit...lol Look like you sell bean pies |
| USER_942c68df | 2010-03-03T02:21:36 | 41.220425 | -85.861873 | These judges are being hard this year. |
| USER_8d0e8566 | 2010-04-03T02:28:12 | 30.387524 | -91.109663 | @USER_b7cdabe3 People don't dance like that to get the burn anymore. Its frowned upon..LOL. |
| USER_8d0e8566 | 2010-03-03T02:29:39 | 30.399934 | -91.121502 | RT @USER_9c9e75e2: Officially getting rid of my iPhone with its dysfunctional button this weekend\|Get a 9700 #BlackertheBerrytheSweetertheUse |

```
hive> select id, count(*) as cnt
    > from tweets
    > group by id;
```

→

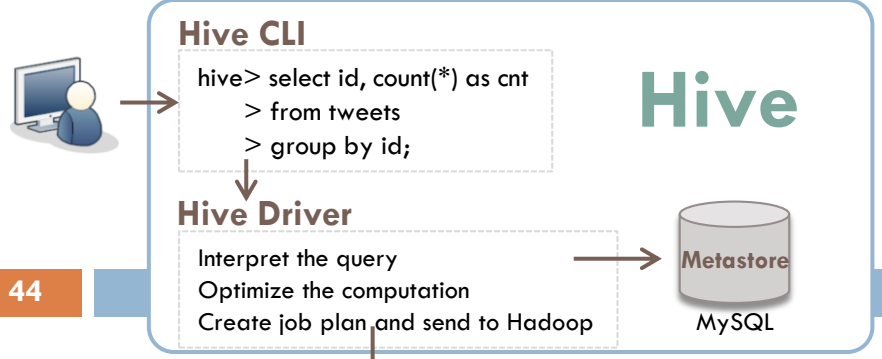
| ID | CNT |
|---|---|
| USER_8d0e8566 | 4 |
| USER_87b48222 | 2 |
| USER_01b8a291 | 1 |
| USER_2e5f8774 | 1 |
| USER_942c68df | 1 |

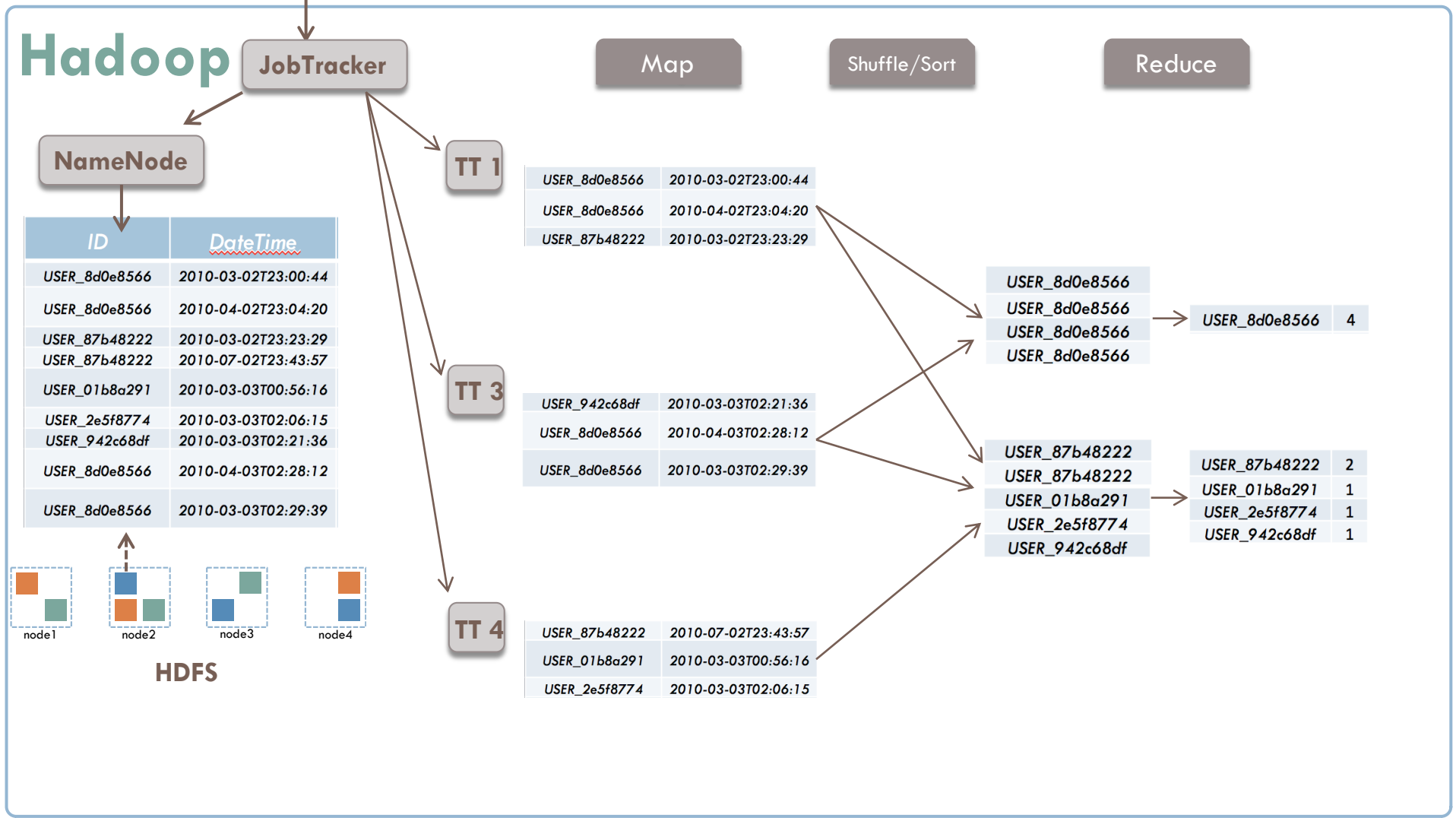
```
hive> select id, avg(latitude) as lat_center, avg(longitude) as lng_center
    > from tweets
    > group by id;
```

→

| ID | lat_center | lng_center |
|---|---|---|
| USER_8d0e8566 | 30.397321 | -91.229643 |
| USER_87b48222 | 37.470001 | -77.492234 |
| USER_01b8a291 | 41.51179 | -95.893286 |
| USER_2e5f8774 | 39.669307 | -79.85002 |
| USER_942c68df | 41.220425 | -85.861873 |

# GROUP BY Explained



# *GROUP BY* - Aggregation

DEMO

## *Calculate # of tweets per user*



```
hive>
    > create table twitter.tweets_per_user as
    > select id, COUNT(*) as cnt
    > from full_text_ts
    > group by id;
Query ID = root_20150130034848_4573d6e3-a2c9-4236-9637-9e62cdc17f10
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1422563374808_0016, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/a
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-01-30 03:48:15,559 Stage-1 map = 0%,  reduce = 0%
2015-01-30 03:48:27,633 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.88 sec
2015-01-30 03:48:39,852 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.41 sec
MapReduce Total cumulative CPU time: 9 seconds 410 msec
Ended Job = job_1422563374808_0016
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db/tweets_per_use
Table twitter.tweets_per_user stats: [numFiles=1, numRows=9475, totalSize=161391, rawDataSize=151
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.41 sec   HDFS Read: 47273366 HDFS Write: 161
Total MapReduce CPU Time Spent: 9 seconds 410 msec
OK
Time taken: 40.962 seconds
hive> select * from tweets_per_user limit 5;
OK
USER_00024ea8   64
USER_001152b0   20
USER_001757c2   33
USER_001ac628   56
USER_002ac912   59
Time taken: 0.187 seconds, Fetched: 5 row(s)
```

# ORDER BY

DEMO

## *Find top 10 tweeters in New York City*



```
hive> select id, count(*) as cnt
    > from full_text_ts
    > where lat > 40.4774 and lat < 40.9176 and
    >       lon > -74.2589 and lon < -73.7004
    > group by id
    > order by cnt desc
    > limit 15;
```

```
OK
USER_f35e4685   259
USER_c913f269   252
USER_2e157dc3   243
USER_c8613ca2   228
USER_75d22fa8   208
USER_c6710d1e   201
USER_18c466a9   194
USER_251d06ba   180
USER_1cd92470   171
USER_5437bd11   169
USER_4a62e18c   161
USER_d9724011   159
USER_2105c63f   158
USER_558acca4   157
USER_36ee45f9   147
Time taken: 73.051 seconds, Fetched: 15 row(s)
```

# DISTINCT

DEMO

*Find # of days this dataset spans*



```
hive> select count(distinct to_date(ts))
    > from full_text_ts;
Query ID = root_20150130042121_53e0176a-40e3-4623-a2d1-f8dc44ae9019
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1422563374808_0021, Tracking URL = http://sandbox.hortonworks.com:8088/
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-01-30 04:21:59,905 Stage-1 map = 0%,  reduce = 0%
2015-01-30 04:22:13,110 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.14 sec
2015-01-30 04:22:23,247 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.56 sec
MapReduce Total cumulative CPU time: 8 seconds 560 msec
Ended Job = job_1422563374808_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.56 sec   HDFS Read: 47273366 HDFS Wr
Total MapReduce CPU Time Spent: 8 seconds 560 msec
OK
6
Time taken: 37.928 seconds, Fetched: 1 row(s)
```

# *JOIN* – Type of joins



- Inner Join
- Left/Right/Full Outer Join
- Semi-Join
- Map-Only Join
- Cross-Join

# JOIN

DEMO

## Find Weekend Tweets (prepare lookup table)



```
[root@sandbox lab]# hadoop fs -put dayofweek.txt /user/lab/
[root@sandbox lab]# hadoop fs -ls /user/lab
Found 1 items
-rw-r--r--   1 root hdfs        115 2015-01-30 05:06 /user/lab/dayofweek.txt
[root@sandbox lab]# hive
hive> use twitter;
OK
Time taken: 2.942 seconds
hive> show tables;
OK
full_text
full_text_2
full_text_ts
tweets_per_user
Time taken: 0.409 seconds, Fetched: 4 row(s)
hive> create table dayofweek (date string, dayofweek string)
    > row format delimited
    > fields terminated by '\t';
OK
Time taken: 0.787 seconds
hive> load data inpath '/user/lab/dayofweek.txt'
    > overwrite into table twitter.dayofweek;
Loading data to table twitter.dayofweek
Table twitter.dayofweek stats: [numFiles=1, numRows=0, totalSize=115, rawDataSize=0]
OK
Time taken: 1.399 seconds
hive> select * from dayofweek;
OK
2010-03-02      Tuesday
2010-03-03      Wednesday
2010-03-04      Thursday
2010-03-05      Friday
2010-03-06      Saturday
2010-03-07      Sunday
Time taken: 0.865 seconds, Fetched: 6 row(s)
```

# JOIN – INNER JOIN

DEMO

## Find Weekend Tweets

```
hive> create table weekend_tweets as
    > select a.id, a.ts, b.dayofweek, a.lat, a.lon, a.tweet
    > from full_text_ts a JOIN dayofweek b
    >      ON to_date(a.ts) = b.date AND b.dayofweek IN ('Saturday','Sunday');
Query ID = root_20150130052020_c1757f46-63df-4b16-9754-dd529d9d366a
Total jobs = 1
Execution log at: /tmp/root/root_20150130052020_c1757f46-63df-4b16-9754-dd529d9d366a.log
2015-01-30 05:20:41     Starting to launch local task to process map join;      maximum memory = 260177920
2015-01-30 05:20:45     Dump the side-table for tag: 1 with group count: 2 into file: file:/tmp/root/9c4ae27d-ef3d-469b-a0d8-f3a936781
ed9/hive_2015-01-30_05-20-31_545_2117097726360411263-1/-local-10003/HashTable-Stage-4/MapJoin-mapfile01--.hashtable
2015-01-30 05:20:45     Uploaded 1 File to: file:/tmp/root/9c4ae27d-ef3d-469b-a0d8-f3a936781ed9/hive_2015-01-30_05-20-31_545_211709772
6360411263-1/-local-10003/HashTable-Stage-4/MapJoin-mapfile01--.hashtable (334 bytes)
2015-01-30 05:20:45     End of local task; Time Taken: 3.489 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1422563374808_0022, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application_1422563374808_0022/
Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job  -kill job_1422563374808_0022
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2015-01-30 05:21:00,990 Stage-4 map = 0%,  reduce = 0%
2015-01-30 05:21:18,186 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 10.64 sec
MapReduce Total cumulative CPU time: 10 seconds 640 msec
Ended Job = job_1422563374808_0022
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db/weekend_tweets
Table twitter.weekend_tweets stats: [numFiles=1, numRows=130017, totalSize=17069733, rawDataSize=16939716]
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1   Cumulative CPU: 10.64 sec   HDFS Read: 47273366 HDFS Write: 17069821 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 640 msec
OK
Time taken: 48.238 seconds
hive> select * from weekend_tweets limit 5;
OK
USER_79321756   2010-03-06 00:57:49     Saturday        47.528139       -122.197916     Haha. We had to stop by my grandmas cause my l
il bros mouth started bleeding hella! Ugh! She pulled that tooth out! Gotta love her!
USER_79321756   2010-03-06 03:40:59     Saturday        47.528139       -122.197916     Apple bees for dinner with the fam and my bf @
USER_60939380 !:) ima start my patron feast there! @USER_6841a093 how ya feeeeeel?
USER_79321756   2010-03-06 04:07:41     Saturday        47.528139       -122.197916     Allllllll da way turnnnnnd up! You can catch
me out gsc bouncin tonight!!
```

# *Lab & Assignment*

# Lab & Assigments



- Lab Computer
  - Username: datastudent
  - Password: datastudent
- Lab 4
  - Go through all the demos
  - Script will be posted after tonight's class
- Assignment 2

# Assignment 2



- *Find hour of the day that generated most number of tweets on March 6, 2010*
- *Find the most mobile tweeter*
  - *User who tweeted from most distinct location (lat-lon)*
- *Find 3 most popular topics (hashtags)*
- *Find most frequently mentioned twitter handle @*