|  |
| --- |
| CREATE EXTERNAL TABLE tweets (     id BIGINT,     created\_at STRING,     source STRING,     favorited BOOLEAN,     retweet\_count INT,     retweeted\_status STRUCT<        text:STRING,        user:STRUCT<screen\_name:STRING,name:STRING>>,     entities STRUCT<        urls:ARRAY<STRUCT<expanded\_url:STRING>>,        user\_mentions:ARRAY<STRUCT<screen\_name:STRING,name:STRING>>,        hashtags:ARRAY<STRUCT<text:STRING>>>,     text STRING,     user STRUCT<        screen\_name:STRING,        name:STRING,        friends\_count:INT,        followers\_count:INT,        statuses\_count:INT,        verified:BOOLEAN,        utc\_offset:INT,        time\_zone:STRING>,     in\_reply\_to\_screen\_name STRING  )  ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'  LOCATION '/user/flume/tweets'; |

Now that we have the data in HDFS and the table created in Hive, lets run some queries in Hive.  
  
One of the way to determine who is the most influential person in a particular field is to to figure out whose tweets are re-tweeted the most. Give enough time for Flume to collect Tweets from Twitter to HDFS and then run the below query in Hive to determine the most influential person.

[?](http://www.thecloudavenue.com/2013/03/analyse-tweets-using-flume-hadoop-and.html)

|  |  |
| --- | --- |
| 1 | SELECT t.retweeted\_screen\_name, sum(retweets) AS total\_retweets, count(\*) AS tweet\_count FROM (SELECT retweeted\_status.user.screen\_name as retweeted\_screen\_name, retweeted\_status.text, max(retweet\_count) as retweets FROM tweets GROUP BY retweeted\_status.user.screen\_name, retweeted\_status.text) t GROUP BY t.retweeted\_screen\_name ORDER BY total\_retweets DESC LIMIT 10; |

Similarly to know which user has the most number of followers, the below query helps.

[?](http://www.thecloudavenue.com/2013/03/analyse-tweets-using-flume-hadoop-and.html)

|  |  |
| --- | --- |
| 1 | select user.screen\_name, user.followers\_count c from tweets order |