Assignment 2

Discussion Questions (5 points)

1. What key considerations should one consider before migrating from a RDBMS to a NoSQL data store?
2. Given that HBase stores rowkeys in lexicographical order, what possible performance impacts should you keep in mind when designing a rowkey in HBase? Why might prefixing a rowkey with a timestamp be problematic? How could you design your rowkeys to minimize such impacts?
3. Many big data companies including Facebook, Trend Micro, and StumbleUpon integrate Hive and HBase to enable their data analytics platform (http://lanyrd.com/2011/oscon-data/sghmh/). Describe the motivations for combining Hive with HBase and provide example use cases for each in the context of a large, real-time distributed analytics system.

Assignment (Total 15 points)

The Resources section contains a set of CSV data containing one hour of aggregated traffic statistics from Wikipedia[1]: This record in the file contains a count of page views for a specific page on Wikipedia. The first column is the language code, second is the page name, third is the number of page views, and fourth is the size of the page in bytes.

(5 points for successful loading from MySQL to HBase)

Import the pagecounts CSV into a MySQL table using the following schema:

```
CREATE TABLE wikistats (
  language varchar(2) NOT NULL,
  pagename text NOT NULL,
  count int(11) DEFAULT NULL,
```

[1] Wikistats: http://www.mediawiki.org/wiki/Analytics/Wikistats

```
        size int(11) DEFAULT NULL);
```

After loading the data you should have a count of 1,461,294 total rows in the table. Use Sqoop's import tool and the –where argument: http://sqoop.apache.org/docs/1.4.4/SqoopUserGuide.html#_selecting_the_data_to_import to import the rows with language code "en" into an HBase table. If you did this correctly, your import should have retrieved 858,131 records.

(5 points for schema definition)

In your submission, indicate the following attributes of the HBase table (or copy and paste your import command):

1. rowkey
2. column families
    a. columns and any counters

(5 points for HBase commands)

Perform the following HBase commands:

1. Get the HBase row data for the row where pagename = "Hadoop" and provide results
2. Put a new column family, "meta", and column, "url", for the "Hadoop" row and set the value to http://en.wikipedia.org/wiki/Hadoop
3. Add a new row with pagename = "Doge (meme)", language = "en", size = 111, and increment or set the count to 50.

Submit the output of 'describe' on your HBase table and the above commands and result output.