David P. Stonehouse

University of Maryland University College

DATA 650 – Big Data Analytics

Fall 2016, Section 9040, Professor Gortcheva

Assignment 3: BigSQL with Yelp Data

## Lessons Learned from the Walk-through exercise

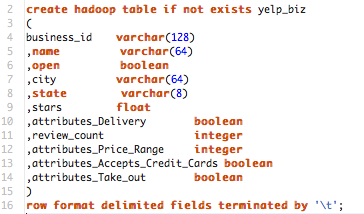
The BigSQL walk-through demonstrated the power of using standard SQL commands to interact with a Hadoop repository.

.

## Statement of Work

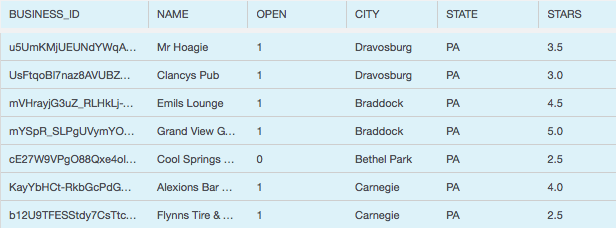
Yelp is a web service that allows users to share and find opinions on businesses, especially restaurants. Yelp has provided data for academic data scientists to study and mine for useful patterns. The data includes information about reviews, reviewers (users) and businesses. By studying this data with SQL commands in BigSQL, users can gain insights about samples of reviews. Those insights include how the review compare with other reviews by the same reviewer and other reviews for the same business. It will also provide a nice way to filter and search the reviews according to information about the reviewers and businesses.

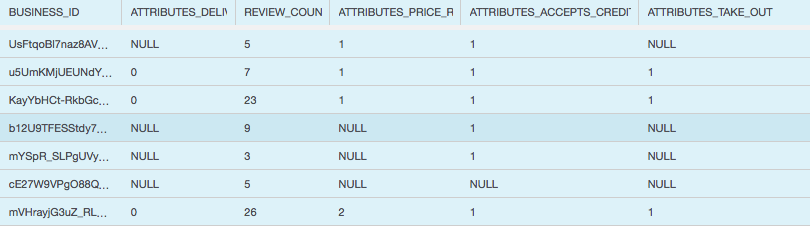
## Database Implementation

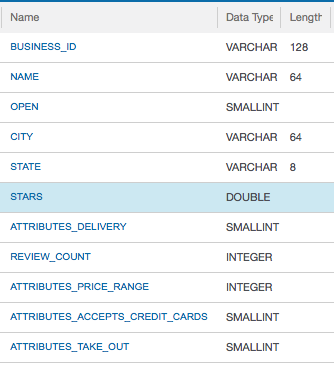




../../Desktop/Screen%20Shot%202016-10-27%20at%205.55.08%20PM.png







Challenges

* JSON to CSV
* Variable names – find/replace periods and spaces with underscores
* Character values – surround with quotes
* Embedded commas – for now, skip
* embedded return characters
* Empty values -> NULL
* Dates
* Too many columns in biz – opened in excel, removed columns not of interest, save as tab delimited
* Convert line endings
* Remove commas, single quotes, double quotes
* Boolean – BigSQL – have to use 0/1 instead of TRUE/FALSE
* Accidentally added a row twice, Hadoop table doesn’t allow selective delete, had to do drop table
* Name collisions – “date”

## Using the database to retrieve data

appendix has a listing of words that correlate with “trailwood” and “birch”, two of the top agenda item words.

## Summary and Recommendations

The data BigSQL offers ways multiple ways to load the data in. For smaller scale loading, INSERT SQL commands can be used. There is also a LOAD command that can be used to pull data from another database system. IBM claims some significant improvements over MapReduce in BigSQL compared with a competing implementation built with MapReduce. It also provides some key capabilities related to row and column level access.

## Conclusions, Limitations and Future Research

## References

IBM, “IBM Knowledge Center - Data types that are supported by Big SQL”. Retrieved from <http://www.ibm.com/support/knowledgecenter/SSPT3X_4.0.0/com.ibm.swg.im.infosphere.biginsights.dev.doc/doc/biga_numbers.html>

http://www.ibm.com/developerworks/library/bd-bigsql/

## Appendix: Supporting Information

**Selected logs from R code:**

**R source code:**

Now try with tf-idf

kfit <- DoKMeans(m.tf.idf.transpose2,8)