Analysis and Visualization California Transportation Data Using Chameleon Cloud

Lei Cao, Bin Tang, Mohsen Beheshti
lcao4@toromail.csudh.edu, btang, mbeheshti@csudh.edu
Computer Science Department, California State University Dominguez Hills

Abstract

The Highway Safety Information System (HSIS) (https://www.hsisinfo.org/) of U.S. Department of Transportation is a multistate database that contains crash, roadway inventory, and traffic volume data for a select group of States including California. However, none of the data at HSIS is currently being processed in a cloud environment. Chameleon Cloud is an NSF-funded cloud test bed which a configurable experimental environment for large-scale cloud research. We, therefore, use Chameleon Cloud to process the data to understand the tradeoffs between traffic safety and roadway designs, and ultimately explore how data clouds can benefit next generation intelligent transportation system.

Intent

We retrieved a California Traffic Accident Dataset 2001 - 2012 from HSIS and processed it in a Hadoop cluster in Chameleon Cloud. Using Chameleon Cloud, we'd like to find out the answer to this question: "How is the trend of California traffic accident related to social factors such as economy and employment rate?"

Materials and Methods

Figure 1 shows the entire process of the HSIS data process in Chameleon Cloud. Firstly, the data set is downloaded directly to the Virtual Machine server, and stored to Hadoop Distributed File System (HDFS). Then in Jupyter Notebook, an interactive coding platform, the dataset is read as Spark RDDs (Resilient Distributed Datasets). The data wrangling process comes next. For every year of the 12 years, there are five data files, respectively named Accident, Intersection, Occupant, Roadway and Vehicle.

Then a combined master file of the 12 Accident files is generated. After a proper data cleaning and analysis using Spark and Pandas libraries, we were able to find the annual accident number for 8 different incorporated population groups (Figure 2). Using Matplotlib python library, we were able to visualize the plot. There are two significant low ebbs, 2002 and 2007-2009, respectively conforming to the aftermath of 9/11/2001 terrorist attack and 2007-2009 global financial crisis.

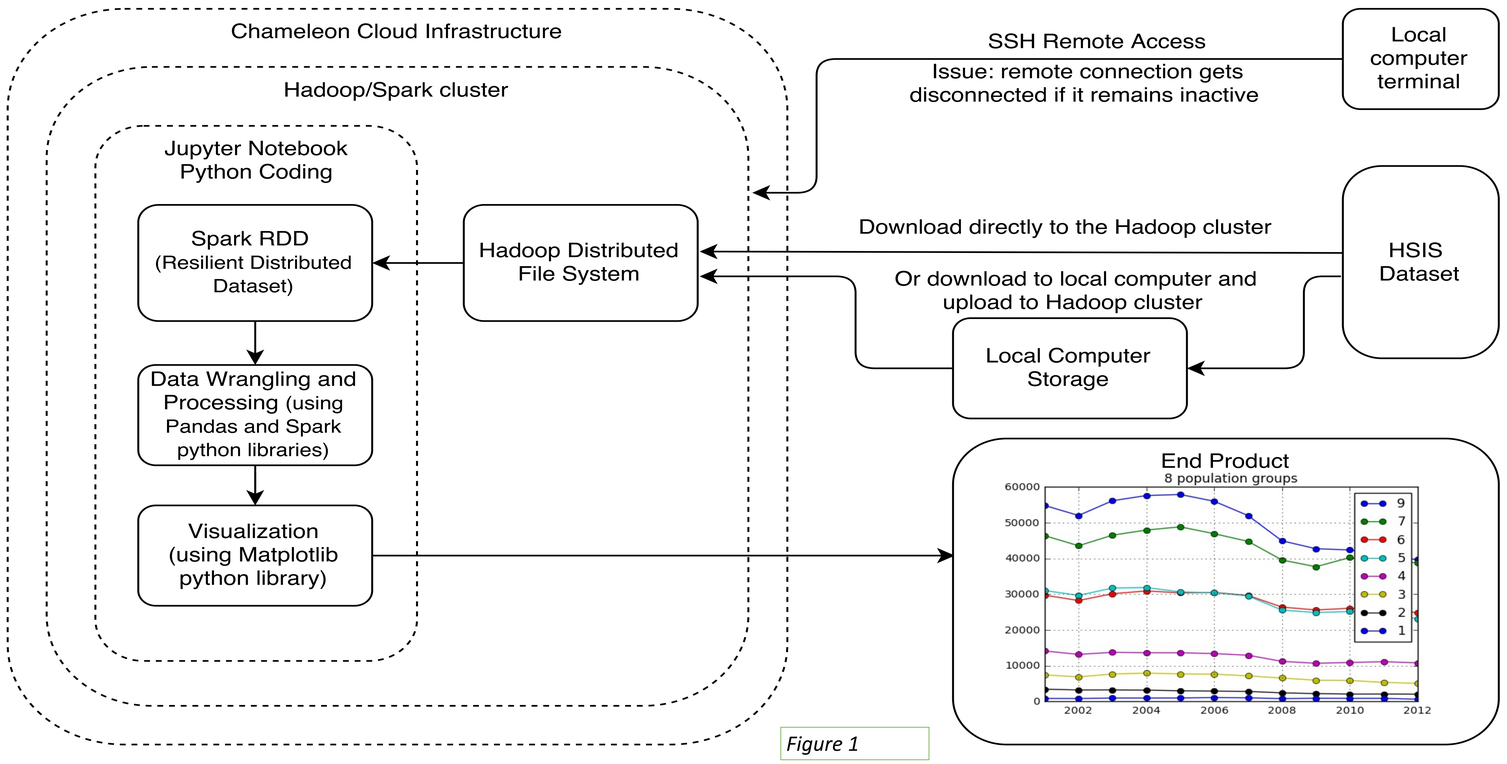

Figure 1

Conclusions

The result shows us that there's actually a connection between the trend of California traffic accident and social factors, which paves our way to explore more of the utilization of cloud computing in the attempt on the next generation intelligent transportation system.

Chameleon Cloud is ideal for our project because it's free to researchers, and it provides an easy OpenStack GUI, giving us a quite convenient configuration of virtual machine instances.

Acknowledgements

This is a collaborated project with Dr. Lu Sun, Professor of Civil Engineering Department from The Catholic University of America. We thank Chameleon Cloud staff for proving us an excellent support for our project.

This research is funded in part through the National Science Foundation (NSF) under Grant No.1649271 and CNS-1551221

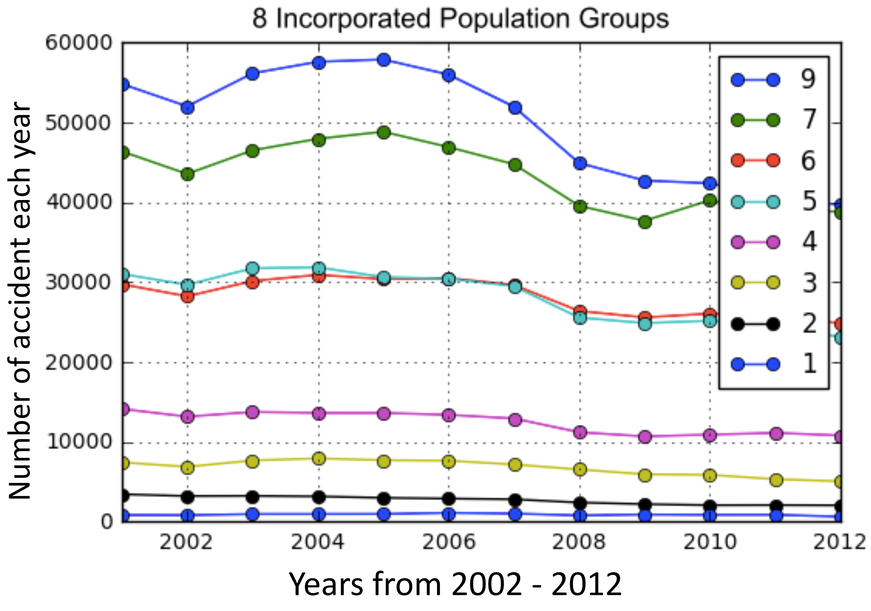

Figure 2