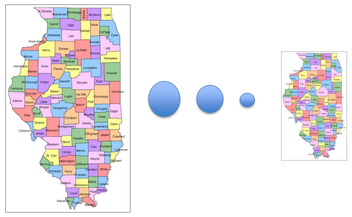
MapReduce Functions on GasDay™ Data Using Hadoop

By: Darla Ahlert and Dr. George Corliss

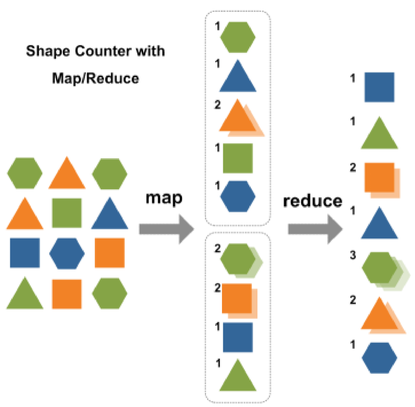
Introduction

- GasDay™ forecasts natural gas usage for many utilities around the United States
- Over 600,000 days of GasDay data; each having 145 different pieces of data
- Need to process all of this data in an effective, time-efficient manner
- Found MapReduce and Hadoop frameworks

MapReduce

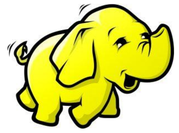


- Software framework allowing processing of large datasets
- Set of two functions, written by the user, that work together and are easily modified to fit many different situations
- Map( ) : Takes input data, processes it, and produces a set of intermediate key/value pairs
- Key/Value pairs get sorted
- Reduce( ) : Takes sorted intermediate pairs, merges similar keys, and forms a smaller list of key/value pairs



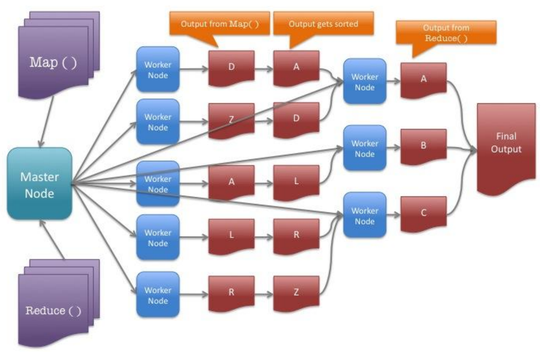
Hadoop Software Framework

- Allows users to distribute mass amounts of data across multiple Linux machines (cluster)
- Hadoop Distributed File System (HDFS) allows MapReduce framework to run across the cluster
- One machine set as “master”; the rest set as “slaves” or “workers”



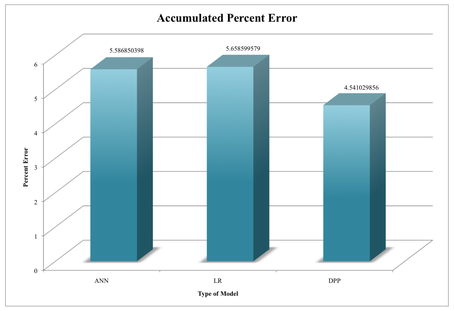
Working Together

- Master: keeps track of important information; sends Map( ) and Reduce( ) tasks to worker nodes
- Workers: complete Map( ) and Reduce( ) tasks
- Map( ) tasks done in parallel with one another
- Reduce( ) tasks also done in parallel with one another
- Parallelization allows simple programs to run over large datasets in a very short amount of time

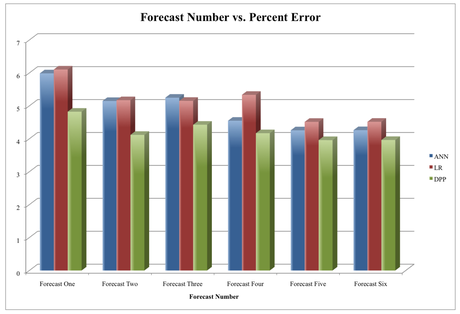


Results

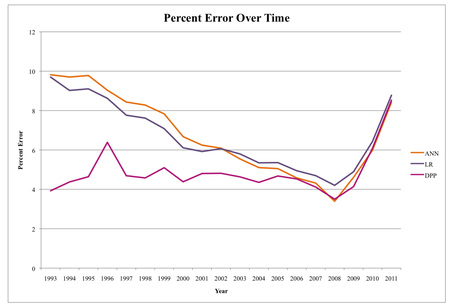
- Frameworks enable us to gain insight into GasDay’s vast amount of data



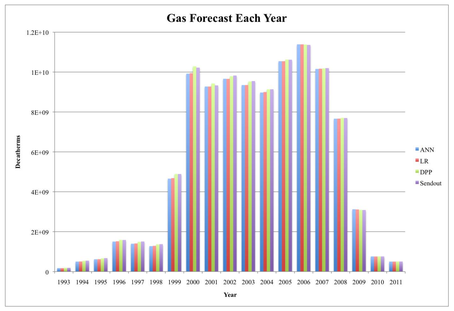
Most accurate type of model



Most accurate forecast number



Forecast accuracy over time



Total amount of gas forecasted each year

Future Work

- Improve on work done throughout my time with GasDay
- Continue to gain more insight into the GasDay data
- Different questions to ask data
- Find every data file from GasDay forecasts

References

[1] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, January 2008.
[2] R. C. Taylor, “An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics,” BMC Bioinformatics, vol. 11, pp. 1–6, July 2010