MapReduce Low Latency/High Frequency Financial Data Processing

Parth Gandhi (parth@perceptiveit.com.au)
Prof. Sanjay Chawla
School of Information Technologies

SUMMARY

We look at the effectiveness of using Apache Pig for Financial Analysis of real Equities and other securities. By using a cluster of multiple notes, large data sets of financial data can be analyzed with scalable processing times

MOTIVATION

- Financial analysis is the basis of Algorithmic trading
- Using larger datasets can improve one's ability to trade in the market
- More accurate predictions and prediction models can be built with larger datasets
- Currently the market uses R, R Parallel and Excel to perform data analysis. Only a very small subset of systems are Hadoop based
- Big Data is the next generation of Algorithmic Trading (Qin 2012)

BACKGROUND

- Financial Analysis includes calculations such as Correlation, Average Return, Sharpe Ratio, Regression and many more
- Traditionally such calculations are performed using Excel, R, Python but are restricted to one machine and the resources it can provide.
- Distributions such as R Parallel allow for parallel processing of data within one node but data is still constrained by the system's resources

APACHE PIG

- Ideal for Research and Machine Learning
- Uses high level language "Pig Latin" (fig 1. shows example of Pig Latin)
- Allows User Defined Functions (UDF)
- Compiler produces MapReduce jobs
- Can be used with Amazon Elastic Map Reduce
- Built for processing

```
>> A = LOAD 'myData';

>> Dump A;
```

Figure 1: An example of Pig Latin

DATA WORKFLOW FOR ANALYSIS

- Firstly data needs to be collected. In this case we use the Reuters Database provided by Sirca to gain access to data
- We only use the CSV export tool because the API requires special access permissions.
- Collected data is then pre-processed using a series of Python Scripts. This is recommended so queries can be faster when running through Pig.
- Pre-processed data is then processed in Pig. Initially we test the data by generating Correlation between two data streams.
- We output a series of text files that contain processed output
- We then verify the output using R

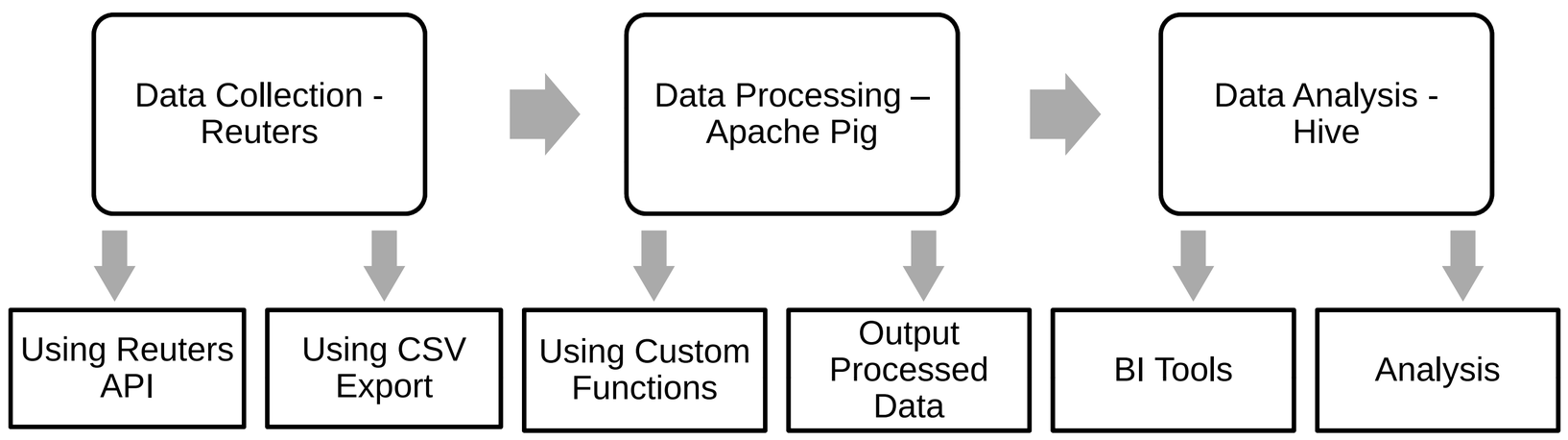

Figure 2: Proposed workflow for Collection, Processing and Analysis for Financial Data (Chauhan 2012)

WORKING WITH BIG DATA

- Data files generated from Reuters vary largely in size.
- A one year data stream of the Australian Dollar, including the time stamp, open, close, bid and ask prices per tick generates a 4.1GB CSV file
- A file of this size cannot be opened in Excel – Tested on a 6 core machine with 32GB RAM
- A file of this size can only be pre-processed in python if read line by line. This file would need to be pre-processed to be used in R unless the machine has more RAM than required for the entire file, the OS, R and any other services
- This file can be easily processed using Pig

PREPROCESSING DATA

- Involves the processing of missing data
- Any ticks missing data are omitted from calculation
- Often ticks have missing data when two data streams are combined due to time-zone conflicts
- Files need to be read line by line so there are no RAM issues when working with large CSV files

PIG - USER DEFINED FUNCTIONS

- Pig has built in mathematical functions required to perform operations
- Python/Jython math libraries can be used to perform most operations required to write functions
- Correlation of time series has already been tested
- Linear regression is currently being implemented
- Pig generally sees a 2x speedup on a 2 node cluster (fig 3)

Operation	Program	Run-time
Correlation	Excel	Couldn't open file
Correlation	R	4-6 Hours
Correlation	Pig	2-3 Hours

Figure 3: Testing Correlation on a 4.1Gb CSV

INDUSTRY APPLICATIONS

- Allows for 'Big Data' to be processed
- Can be used in real-time with live data
- Build a Hive Data Warehouse and pipe data into it for analysis
- Work with high frequency/low latency Nano-second data

FUTURE WORK

- Develop Hive Data Warehouse to store pre-processed data and use for further analysis and reporting
- Use Reuters API for data
- Develop preprocessing scripts to process data from Reuters API
- Use Hive Analysis to generate trading strategies for next generation Algorithmic Trading
- Implement into existing Algorithmic Trading workflow for faster and larger data processing

CONCLUSION

This is the first time that MapReduce data processing has been applied to financial data. The evolution of Apache Pig as a scripting tool as well as its implementation on Hadoop makes this the ideal platform for financial data processing. This allows for scalable and cheap processing using Amazon Web Services and is an industry changer.

Acknowledgements

This work was supported by Sirca, who provided the data required for the testing of the functions written in Pig.

References

- Chauhan A, 2012, 'Hive & Pig', presented at 'Extremely Large Databases Conference, Stanford, 10-13 September
- Qin, X, 2012. Making Use of the Big Data: Next Generation of Algorithm Trading. Artificial Intelligence and Computational Intelligence, [Online]. 7530, 34-41. Available at: http://dx.doi.org/10.1007/978-3-642-33478-8_5 [Accessed 22 October 2012].
- Goodhart, C. A. E. & O'Hara, M. 1997. High frequency data in financial markets: Issues and applications. Journal of Empirical Finance, 4, 73-114.