INDIAN INSTITUTE OF MANAGEMENT CALCUTTA

WORKING PAPER SERIES

WPS No. 718/ December 2012

The IIM Calcutta Data Ce rospective, Programming Tutorial and Futur irections

by

P.Srikant
Doctoral Student IIM C H. Road, Joka P.O., Kolkata 700 104 India

&

Chakrabarti
Professor, IIM Calcutta, I rbour Road, Joka P.O., Kolkata 700

The IIM Calcutta Data Centre : A Retrospective, Programming Tutorial and Future Directions

P. Srikant
IIM Calcutta

I describe the development from scratch of a high-frequency financial research data centre, using freely available tools. The design provides an acceptable level of performance at a low cost. I provide examples of usage, and suggest possible paths for the further development of the data centre.

Address for Correspondence:

P. Srikant
Fellow Programmes and Research, IIM Calcutta
Diamond Harbour Road, Joka, Kolkata - 700 104, India
phone: +91 91636 56187
email: p.srikant@gmail.com

(Questions and comments are welcome at the above address)

1 Introduction

The Data Centre was at the Finance Research and Trading Laboratory, IIM Calcutta, was initiated in April 2010. This data centre is now a unique repository of Indian tick data, from which a stream of academic research in asset pricing and microstructure has started emerging. The data is also used by students inclined towards trading careers to devise strategies in partnership with financial firms. The time now seems appropriate to take stock of the data centre's achievements over the past t.0f..30 Twrd em

ca eseae at the timre.Alswrd em-1303 0 TD.0004 Tc-.0004 Tw(o,t t is r tible providesshorts)

ommes(ugge)-e

suspect and the data is exceedingly slow to access. Also, intraday data is typically not available in commercial databases.

The development of databases is an important precursor to financial research. There have been several instances where academics have taken the lead in establishing such databases. To cite a few examples, the Center for Research in Security Prices at Chicago was created when a professor was asked by the industry to determine historical stock market return for which they had to create a portfolio, which later on became the CRSP equally and value weighted indexes and the basis for almost all event studies in the US. Empirical market microstructure was born when Bob Wood at the Univ. of Memphis created the ISSM database from intraday trading data. Ghon Rhee created the PACCAP database at URI which provided a unique competitive advantage for their Ph.D. students.[1]

compelled perforce to compress the files. I considered various possible compression formats, including writing my own, before selecting gzip. Gzip [Nelson and Gailly, 1995] is a widely used file compression program, which reduced the file sizes to about 1.2 GB each (The average size of compressed files now is 2.5 GB). This bought us some critical breathing time - I could now store a couple of months of data on the server. Later, the institute procured a 1 terabyte drive for some $150, which could hold two years of data. This was the only financial cost incurred by the project at the time.

4 Design Choices

The design of a large database requires various design choices. At 15 GB a day, a year's data would occupy some 4 TB of space. Even compressed, the data would occupy approximately half a terabyte.

[Jacob and Shasha, 1999] list various models for time series systems. A tick database for financial instruments is typically characterized by irregular time series of non-periodic data. The level of activity across stocks varies widely - some stocks have thousands of trades each hour, while others may only have a few dozen. The design of an appropriate architecture needs to consider the nature of data, the frequency of updates, and the use cases for such data.

I therefore planned to always retain the compressed daily files as a backup, while investigating other storage formats. Since the archival process was organized around a daily recording routine, I decided that this would also be the unit of storage.

Some specific design choices I considered are listed below:

1. *All information or select fields:* Different exchanges publish different kinds of information. For example, the NSE publishes only one level of bid and ask, while the BSE publishes 5 levels. While I received suggestions to standardize and store only a few fields, which would have led to smaller files, I decided not to throw away any information - while unnecessary information could always be ignored, it would be difficult if not impossible to recreate data fields which we dropped.
2. *One file or many files:* Live data is received in the form of messages, ordered by time. Messages from all exchanges and all instruments are broadcast on the same network. In general, the communication network does not guarantee that messages will be sent in the order they were received. Our timestamp on messages has a minimum resolution of one second. If all messages are stored in the same file, potentially valuable sequence information is retained. This sequence information - the ordering of events within a second - would be irretrievably lost if each ticker was stored in a separate file.
   It is worth mentioning here that I had considered other attempts to deal with this issue. For example, NEEDS provided a sequence count for its data on the TSE - this is timestamped to the nearest minute. The NYSE TAQ dataset sorts all trades in a day by ticker, and uses an indexed sequential file to access a desired ticker. (I understand a project was subsequently undertaken to split files by ticker, which seems an inelegant and lossy solution.)
   Financial research is rarely done on one stock, so a frequent case would be to run a

program for all stocks, or a list of stocks, for a day at a time. Any user with the rudiments of programming knowledge could use elementary data structures like a hashtable for each stock to store intermediate computations of interest. In addition, one file most closely mimics a live data stream, so this allows for easy translation of strategies from a historical to a live environment.

3. *Choice of database:* While compressed files can be used for many research tasks, a database typically provides indexes that speed up queries for specific instruments. Unlike in an investment bank, I did not have an unlimited budget at my disposal. Commercial time-series databases typically cost over

```
   1 srikant    users      2436360072 Oct  6 06:30 20121005.txt.gz
bash-2.05b$ gzcat 20121005.txt.gz
```

```
my $gz = gzopen("20121005.txt.gz", "rb") ;
$gz->gzseek( 1700 * 1024 * 1024 );
```

```
table = h5file.root.mktdata.ticks

for row in table.where ( ’name
                        o = o + fname + "=" + fval + "|"
        t['other']= o

        t.append()

table.flush()
h5file.close()
f.close()
```

combination of Hadoop and HDF5 should be an efficient yet relatively low-cost solution for research data services. Hadoop also comes with its own database, HBase, which might also prove to be an efficient solution.

One of my goals in writing this article at this time is also to caution against following unfruitful lines of work. I believe it is not advisable to split the files by ticker, or to use a relational database to store this data. Another caution is that a continued effort is required to ensure that the recording process continues to function smoothly, and that the data centre’s recorded data is fully backed up.

I am fully conscious of the limitations of what I have developed. However, many of these design choices were dictated by resource limitations, some of which no longer hold. Also, some technologies have gained wide acceptance within the last two years. An exciting direction forward is the use of Apache Hadoop to parallelize computations so that files corresponding to multiple dates worth of data can be processed simulataneously. Deploying such a platform will need competent technology inputs. I understand cloud computing and technologies like Hadoop are actively being investigated by the MIS department, and we may be able to leverage on their expertise.

As future work, a library could be developed to provide programming interfaces which insulates the end user from the underlying storage format. A possible programming semantic to calculate vwap over an interval may look like this:

```
volume = value = 0;

sim = new Simulator(  );
sim.setStart(20121005:09:00:00);
sim.setEnd(20121005:09:30:00 );
sim.run();

while ( record = sim.getNextRecord() )
{
   if ( record.ticker == "ABB" and record.hasField( "lvol" ) )
   {
               volume += record.lvol;
               value += record.last * record.lvol;
    }
}

sim.stop();

print value / volume;
```

8 Acknowledgments

In early 2010, Prof. Dey visited IIM Calcutta and taught an course in Empirical Market Microstructure, which I attended. There was then a complete absence of high-frequency data in India, which made it impossible to perform research on intraday asset prices and market microstructure. Malay pointed out that the absence of a database was not just a constraint, but

also an opportunity to build one. I would like to thank Prof. Dey for suggesting the project, and for his support and guidance during the project.

Based on Malay's suggestion, I decided to investigate the technical feasibility of setting up a data centre. I am grateful to my faculty advisor, Prof. B. B. Chakrabarti, for agreeing that this project would also serve to meet my PhD summer research project requirement.[3] We believed at the time that creating such a centre would add to the institute's credentials as a premier finance school, and signal the institute's commitment to research.

I would also like to acknowledge the consistent and diligent work of Priyanka, the Assistant Manager of the finance laboratory, in ensuring that the recording process has continued to operate smoothly. In addition, interns at the lab at the time included Shishir Kumar, Rishiraj Diwakar, and Rahul Kumar, who helped developed a few applications using this data.

Finally, I would like to thank th

2. http://www.elitetrader.com/