

Revisiting the next generation of StockPrice infrastructure

August 1, 2021
Categories: MeRC
Tags: Research Stories, Using the Cloud

For many facets of our lives, long-term public good relies on a healthy tension between competition and stability. The age of digital disruption has profoundly changed the nature of competition in financial markets, to the extent that regulation has not always been adequate to ensure stability. Associate Professor Paul Lajbcygier and his colleague Rohan Fletcher from the Monash Business School are custodians of a longitudinal study seeking to understand when stability has been superseded by innovations. To their surprise, the recent Nectar Research Cloud refresh has caused a digital disruption to their own research, reducing analysis time from many months to one week, and in turn changing the focus of research.

How deep does the digital disruption rabbit hole go? We asked Paul and Rohan to tell us about it...

"With the advent of the IT revolution, financial exchanges have changed beyond recognition. With the Centre's help, we have focused our research on how the digital disruption has affected financial markets, considered welfare implications, and potential regulatory changes, with ramifications for regulators, traders, superannuants and all equity market stakeholders."

Associate Professor Paul Lajbcygier

Recently, the Monash eResearch centre has supported Paul's team by upgrading new essential hardware and infrastructure necessary to interrogate the vast data generated from the Australian equity markets.

"Without the Centre's help, our research would be impossible".

Associate Professor Paul Lajbcygier

To get an understanding how the refreshed Research Cloud would affect the team, they benchmarked the new hardware against their data. They rerun database MySQL code which searches vast amounts of ASX stock data in order to understand the costs of stock trading using new, innovative price impact models. That prior work led to an A* publication in 2020 in the Journal of Economics, Dynamics and Control ¹

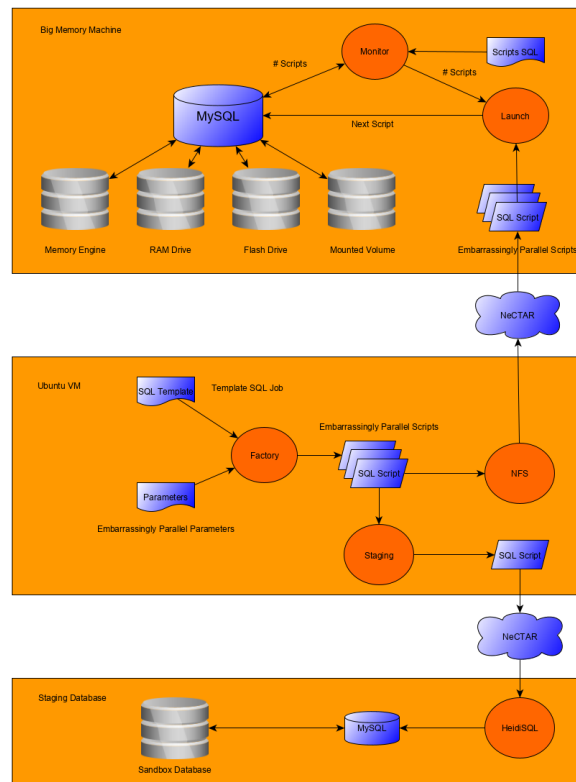
(<https://rcblog.erc.monash.edu.au/blog/2021/08/revisiting-the-next-generation-of-stockprice-infrastructure/#easy-footnote-bottom-1-2609>).

This analysis interrogates over 1300 stock's and their related trades and orders from 2007 to 2013, representing three terabytes of data.

"In order to implement this huge processing task, we have automated the breakdown of these MySQL queries by stock, year and month. This generates over 80,000 SQL scripts, which is a total of 3 gigabytes of SQL analysis code alone."

"With the latest hardware provided by Monash eResearch centre, this query took around one week, in contrast to the many months of required running time prior to the provision of the latest hardware."

Associate Professor Paul Lajbcygier



The architecture of the embarrassingly parallel Stock Price infrastructure hosted on the Nectar Research Cloud consists of a large memory machine running Ubuntu LTS, and a number of smaller staging, analysis machines based also on Ubuntu LTS and Windows.

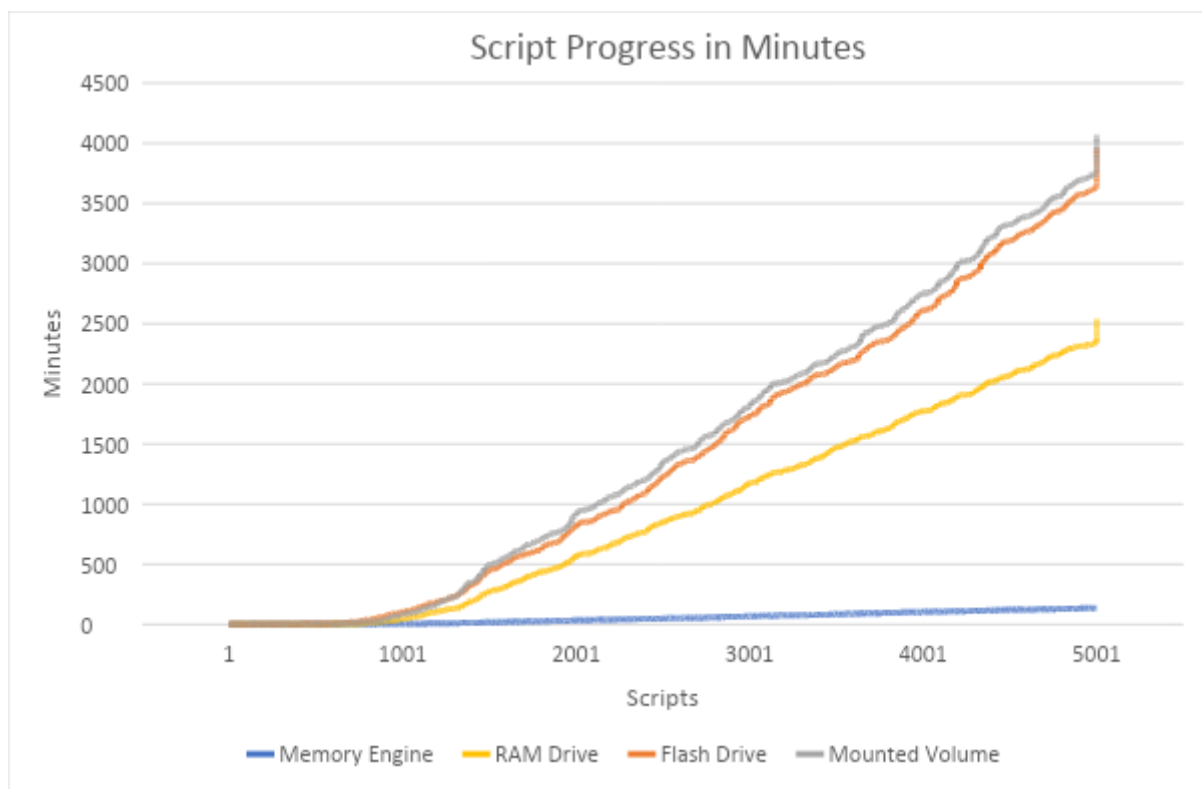
The project uses MySQL and bash scripts to facilitate the templating of SQL jobs. Scripts are generated on the small VM and then moved to the big memory machine for execution. The number of scripts on the big memory machine is monitored and is held at a pre-set maximum. On completion of a script on the MySQL database the next available script is launched automatically.

Testing of a single script may be performed on the staging database after some automatic modifications have been made to make it compatible for individual execution.

Below is an example chart comparing the execution time for an SQL script when utilising each of the four underlying database storage technologies now available on a big memory machine on the Monash node of the Research Cloud. These being: MySQL's Memory Storage Engine, utilising RAM Drive, utilising Flash Drive and utilising a mounted volume (a separate Ceph cluster via RoCE). Clearly, the use of the memory engine (blue line) provides the best performance.

"Repeating our published benchmark, the MySQL memory engine is approximately forty times better than the flash drive and mounted volume. This outstanding Memory engine performance occurs because the memory engine is internal to MySQL, thereby avoiding input/output lags required of the file system."

Associate Professor Paul Lajbcygier

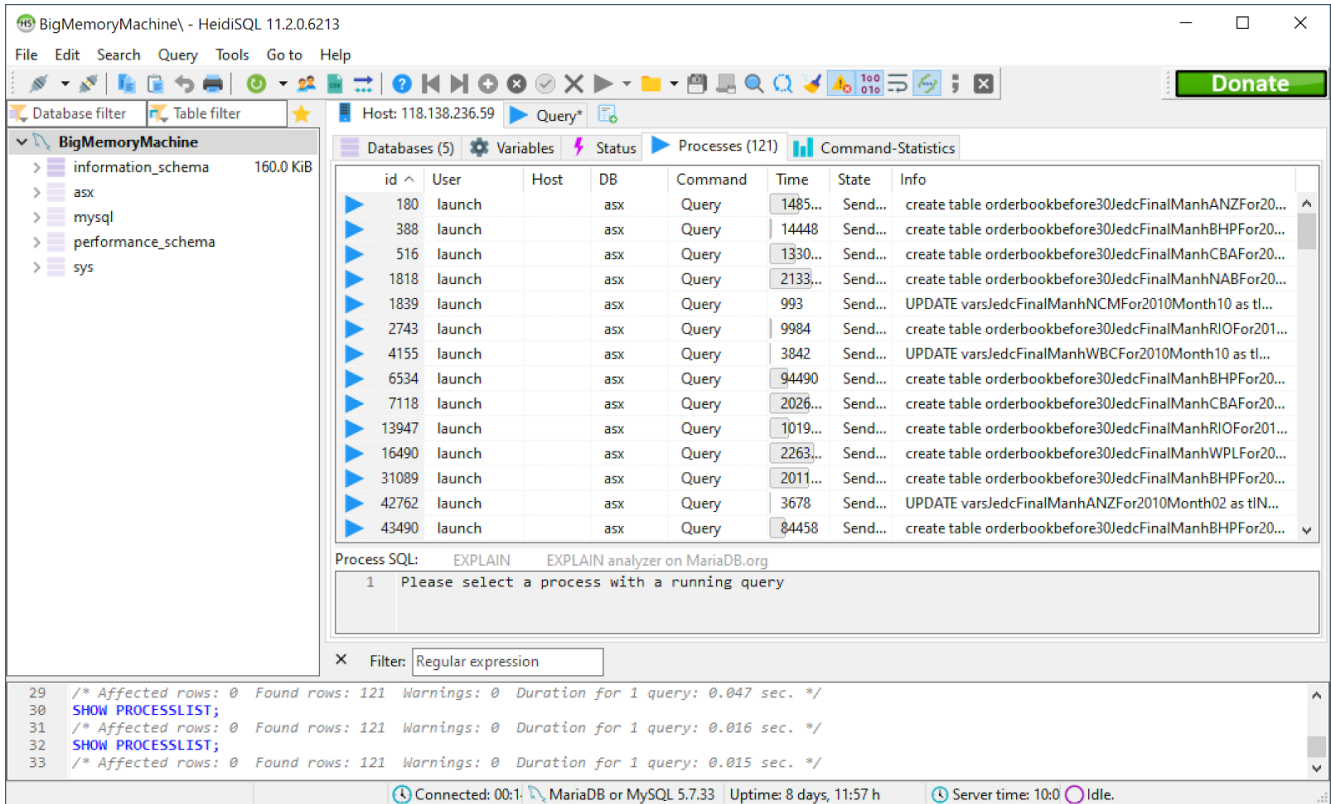


Stock Price search performance based on storage backends.

The team then sought observables to explain why the memory approach performed so well. Below is an example of system load as recorded using Ganglia running the same SQL query. To the left is the Memory Engine CPU load usage, followed by examples of the Flash Drive, Mounted Volume and RAM Drive CPU load usage.

"It is possible to see that the Memory engine utilises all 120 CPU processes consistently, in contrast to the right hand graph which shows other memory methods which do not efficiently utilise the new hardware and incur overheads due to the requirement that they must use the file system."

In addition to fine tuning MySQL to the specialist hardware (what they named the “Big Memory Machine”), the benchmarking necessitated the integration of a bespoke Microsoft Windows ecosystem of tools. They used the open source tool HeidiSQL, to both visualise and automate the decomposition of the analysis problem to 120 parallel executing SQL scripts.



Parallel Stock Price SQL executions.

To summarise, We've asked Paul what the overall impact of the revamped Stock Price infrastructure in answering their research questions.

“We're able to utilise data analysis on a more comprehensive data set including the ASX and the US NASDAQ, perform rapid prototyping with quick feedback; and complete analyses that would be intractable using the previous infrastructure.”