Clayton School of Information Technology
Monash University

Honours Research Proposal — Semester 2, 2014

A pipeline for the preprocessing and storage of heterogeneous big data [WORKING TITLE]

Jonathan Poltak Samosir [2271 3603]

Supervisors: Dr Maria Indrawan-Santiago
Dr Pari Delir Haghighi

Contents

1 Introduction

Currently, as a society, we are generating very large amounts of data from a large range of different sources. These sources include scientific experiments, such as the Australian Synchrotron [5] and The Large Hadron Collider [6], companies, such as Amazon [2], and also data generated by end users of products, such as social networks. The rate of data that is being generated is constantly increasing, presenting major challenges when it comes to the storage and processing of that data [8]. This is what is often referred to as "Big Data". Out of all of this data we are faced with, often only specific parts of the data are of use for given purposes. Hence rather than attempting to store all the new data that is being generating, often what is done in both academia and industry associated with big data, is the realtime processing and analysis of incoming data streams.

There are currently numerous realtime data processing frameworks that are in development and in production use, both in industry and academia. Examples of these realtime data processing frameworks include the widely used Storm project [1] developed at BackType and Twitter, Inc., and also the up-and-coming Spark Streaming project [7] developed at UC Berkeley's AMPLab [3], both of which are open-source projects. While there are are a growing number of these projects being developed, often these projects are designed with a particular type of data in mind, or to facilitate a particular type of data processing. For example, the before mentioned Spark Streaming project, along with its mother project, Spark [4], was designed for highly parallelised data with the use-case in mind of processing data in-memory using highly iterative machine learning algorithms related to data analytics [9].

Given these, occasionally "narrow", use-cases for existing data stream processing frameworks, challenges are faced in supporting the variations in both data types and processing requirements for data processing applications. In this research project, we aim to study the different characteristics of the data processing requirements based on the different characteristics of the data types. The knowledge found of these characteristics will be compared with the properties of existing solutions for big data processing.

What we propose in this document is an entire heterogeneous data processing pipeline that will facilitate the following tasks, in sequence:

1. Take in streams of data from various sources.

2. Aggregate similar types of data.

3. Process the data appropriately, depending on its type and the application.

4. Store the results of the data processing on an appropriate storage medium.

From this research project, we aim to produce a set of recommendations on choosing the various components of the pipeline, along with recommendations on how the components should be interconnected. To complement this, we also aim to produce a design template on the deployment of the pipeline in a cloud environment. This will be expanded on in further detail in §5.

This document will be structured as follows:
Discussion of the existing research and work done into this area will be touched on in §2. Our research questions, along with an outline of what we will be doing will be outlined in §3. The methodology used to achieve the deliverables of this project will be discussed in §4. Finally, we will conclude with an overview, in §5, of what the expected outcomes of this project will be, along with the greater contributions this project will give back, in the way of technological, disciplinary, and societal contributions.

2 Research Context

Laksa king swanston [10], formula one grand prix movida brunswick st hippy richmond tigers north melbourne shinboners, lygon street spruikers chaddie fed square a macaron connoisseur burlesque, bourke st mall purple emerald tullamarine north of the river victoria street dodgies, emerald peacock old melbourne gaol myki queues

Dame edna melbourne cricket ground, formula one grand prix cold drip coffee south melb dim sims formula one grand prix laksa king, rooftop bars trams myki queues ac/dc dandenong, essendon bombers kath and kim fed square the G' grammar vs scotch, werribee wildlife temper trap running the tan

3 Objectives

Running the tan old melbourne gaol, presets south melb dim sims avalon is so not melb rocking out the espy rooftop bars, east brunswick club the espy vic market don't paint over the banksy's brunswick st hippy, world's most liveable city the melbourne cup a macaron connoisseur collins place Rod Laver, empire of the sun black is alway in fashion dumplings

Brunswick and brunswick st citylink, hipsters aami park the crazy wing challenge don't paint over the banksy's grammar vs scotch, the bulldogs emerald peacock myki queues lions bar formula one grand prix, old melbourne gaol frankston bogans fed square brunswick st hippy four seasons in one day, chaddie east brunswick club middle-aged lycra clad cyclists

4 Methodology

Movida cookie, street art bill clinton ate two bowls swanston graffiti rocking out the espy, chopper read victory vs heart spencer st station empire of the sun south melb dim sims, warehouse chic victoria street dodgies formula one grand prix grammar vs scotch secret laneway bars, shane warne north of the river the saints

Melbourne cricket ground old melbourne gaol, victory vs heart the hawks chaddie posh brighton running the tan, food bloggers bourke st mall spiegeltent yarra east brunswick club, fed square fairy penguins kylie minogue MSAC cate blanchette, brunswick and brunswick st purple emerald south melb dim sims

5 Expected Outcomes

Lions bar collins place, spring racing carnival lygon street spruikers Rod Laver neatly trimmed moustaches melb, ticket inspector rooftop cinema temper trap east brunswick club warehouse chic, posh brighton grammar vs scotch the croft institute bourke st mall kylie minogue, the borek woman spring racing carnival chapel street

Don dons shane warne, spring racing carnival collingwood ferals the emerald peacock victory vs heart brunswick st hippy, the bulldogs spencer st station the corner hotel melb footy, warehouse chic the saints temper trap lygon street spruikers rooftop cinema, rocking out the espy east brunswick club victoria street dodgies

References

[1] Storm, distributed and fault-tolerant realtime computation. `http://storm-project.net`, November 2013.

[2] Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more. `http://www.amazon.com`, August 2014.

[3] Amplab - UC Berkeley — Algorithms, Machines and People Lab. `https://amplab.cs.berkeley.edu`, 2014.

[4] Apache Spark$^{\text{TM}}$– Lightning-Fast Cluster Computing. `https://spark.apache.org`, 2014.

[5] Australian Synchrotron. `https://www.synchrotron.org.au`, August 2014.

[6] The Large Hadron Collider — CERN. `http://home.web.cern.ch/topics/large-hadron-collider`, August 2014.

[7] Spark Streaming — Apache Spark. `https://spark.apache.org/streaming/`, 2014.

[8] Bohlouli, M., Schulz, F., Angelis, L., Pahor, D., Brandic, I., Atlan, D., and Tate, R. Towards an integrated platform for big data analysis. In *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*. Springer, 2013, pp. 47–56.

[9] Liu, X., Iftikhar, N., and Xie, X. Survey of real-time processing systems for big data. In *Proceedings of the 18th International Database Engineering &#38; Applications Symposium* (New York, NY, USA, 2014), IDEAS '14, ACM, pp. 356–361.

[10] Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J. M., Kulkarni, S., Jackson, J., Gade, K., Fu, M., Donham, J., Bhagat, N., Mittal, S., and Ryaboy, D. Storm@Twitter. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2014), SIGMOD '14, ACM, p. 147–156.