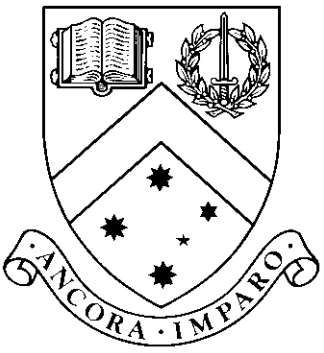
Clayton School of Information Technology
Monash University



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 [7], 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 [9]. 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 [8] developed at UC Berkeley's AMPLab [3], both of which are open-source projects. While there 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 [11].

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

3 Objectives

3.1 Research questions

The following research questions will be the main focus of our project:

1. What classification methods can we best use to classify arbitrary data?
2. How the data arriving in the pipeline be preprocessed and structured to adhere to the interface of the appropriate storage system?
3. How can existing data stream processing solutions be used or be altered for use within this pipeline?

3.2 Research aims

The main aim or goal of this research project is to develop a set of recommendations that can lead to the creation of this data stream processing pipeline. While we do not aim to have a fully usable, production-ready piece of software as a deliverable, we want to at least have a proof-of-concept working in the National eResearch Collaboration Tools and Resources (NeCTAR) cloud [6].

4 Research Design

4.1 Methodology

4.2 Proposed thesis chapter headings

The proposed structure of the thesis is as follows:

1. Introduction
    1.1. Overview
    1.2. Background
    1.3. Research problem
    1.4. Research questions
    1.5. Research scope
    1.6. Conclusion and thesis structure
2. Literature Review
    2.1. Introduction
    2.2. Definition of terms
    2.3. Big data in industry and academia
    2.4. Batch data processing
    2.5. Overview of batch data processing technologies
    2.6. Realtime data processing
    2.7. Overview of realtime data processing technologies
    2.8. Disruptive research in big data
    2.9. Conclusion

3. Streaming Data Preprocessing Pipeline Model
    3.1. Introduction
    3.2. Need for preprocessing pipeline
    3.3. Overview of pipeline
    3.4. Usage of pipeline
    3.5. Conclusion
4. Research Method and Implementation
    4.1. Introduction
    4.2. Chosen research method
    4.3. Data classification method
    4.4. Data processing technologies for pipeline
    4.5. Implementation of pipeline in NeCTA cloud
    4.6. Formulation of pipeline recommendations
    4.7. Conclusion
5. Discussion and Evaluation
    5.1. Introduction
    5.2. Further methods for data classification
    5.3. Future of big data research
    5.4. Evaluation of project
    5.5. Conclusion
6. Conclusion
    6.1. Overview
    6.2. Research contributions
    6.3. Research limitations
    6.4. Future research
7. Reference List
8. Appendices

4.3 Timetable

4.4 Potential difficulties

While we believe that most components of this project are very much feasible given the time and resources we have allocated so far, we have identified a small number of possible difficulties that may be encountered as the project progresses. The most obvious difficulty so far that we have identified is the possibility of acquiring a substantial amount of data that we can use for both testing and during the evaluation stages of the project. As this data will be used to test our data classification methods and evaluate our pipeline deployed in the cloud, it will need to be diverse. By diverse, what we mean is it must display heterogeneity in terms of its type and origin; data from many different sources would be ideal.

Currently we have no concrete leads on the acquisition of this data, although we will look into collaboration with other data-based research projects ongoing at Monash. We also are yet to explore freely available data sets, such as the Enron corpus email dataset [10], although these may definitely be taken into consideration at a later stage in the project in the case that data acquisition proves infeasible.

4.5 Special facilities required

As we are aiming to deploy a proof-of-concept of this pipeline, the main special facility needed access to is a cloud solution that enables us to install and test our pipeline. For this, we have already requested access for what we are planning to do in this project on the National eResearch Collaboration Tools and Resources (NeCTAR) cloud [6]. This cloud is funded by the Australian Government and available to Australian researchers in many different disciplines.

With access to this cloud for the duration of this project, we will be able to install and perform qualitative comparisons between the numerous realtime data processing solutions available as of now. This will assist us in making our recommendations for the pipeline based on the classification of particular set of data.

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^{TM}– Lightning-Fast Cluster Computing. https://spark.apache.org, 2014.

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

[6] home — NeCTAR. http://www.nectar.org.au, August 2014.

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

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

[9] 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.

[10] Klimt, B., and Yang, Y. Introducing the enron corpus. In *CEAS* (2004).

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