

Clayton School of Information Technology
Monash University



Honours Thesis — Semester 1, 2015

A study of data stream processing systems for use
with railway

Jonathan Poltak Samosir

[2271 3603]

Supervisors: Dr Maria Indrawan-Santiago
Dr Pari Delir Haghighi

Contents

1	Introduction	1
1.1	Research Context	1
1.1.1	Big Data	1
1.1.2	Batch data processing	2
1.1.3	Realtime data processing	2
1.1.4	Monash IRT Railway Project	3
1.2	Research Aim	3
1.3	Research Questions	4
1.4	Significance and Contribution	4
1.5	Thesis Structure	4
1.6	Summary	5
2	Literature Review	6
2.1	Big data processing background	6
2.2	Batch data processing	7
2.2.1	MapReduce and GFS	7
2.2.2	Hadoop MapReduce and HDFS	8
2.2.3	Pig and Hive	9
2.3	Realtime data processing	9
2.3.1	Hadoop YARN	10
2.3.2	Storm	11
2.3.3	Spark Streaming	12
2.3.4	Samza	13
2.3.5	S4	13
2.4	Discussion and analysis	14
2.5	Conclusion	17
3	Data Steam Processing System Technology Overview	18
3.1	Data Steam Processing System Technology Choices	18
3.1.1	Samza Programming	18
3.2	Storm Programming	20
3.3	Spark Streaming Programming	21
3.4	Evaluation Method & Approach	22
3.4.1	Quantitative methods	22
3.4.2	Qualitative methods	23
3.5	Overview of the data	24
3.6	Conclusion	25
4	Pipeline Implementation	26
4.1	Data Filtering Pipeline Design	26
4.2	Testing Environment Details	27
4.2.1	Details on Host System	27
4.2.2	Details on Data Stream Processing System Technologies	28
4.3	Implementation of Pipelines in Data Stream Processing System Technologies	28
4.3.1	Samza	28
4.3.2	Storm	28
4.3.3	Spark Streaming (JVM Spark)	28
4.3.4	Spark Streaming (PySpark)	28
4.4	Conclusion	28

5	Discussion and Evaluation	29
5.1	Testing environment	29
5.2	Quantitative Evaluations	29
5.2.1	Testing data	29
5.2.2	Tests to be performed	30
5.2.3	Outcomes of Tests Performed	31
5.3	Qualitative Evaluations	31
5.4	Tests to be performed	31
5.5	DSPS Technology Recommendations	31
6	Conclusions	32
6.1	Research Contribution	32
6.2	Future Work	32

Abstract

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 [7] and The Large Hadron Collider [10], companies, such as Amazon [4], 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 [23]. This is what is often referred to now as “Big Data”. A further big data project, that will be looked into as the basis of the project outlined in this thesis, is the automated monitoring of railway tracks and cars by the Institute of Railway Technology at Monash University (IRT) [9].

Out of all of these data that are faced in such projects, often only specific parts of the data are of particular use for given purposes. Hence, rather than attempting to store all the new data that is being generated, an increasingly popular method of dealing with such data, in both academia and industry associated with big data, is the processing and analysis of data in realtime as it is received. This is also known as online processing, or single pass processing, where data is processed in real time before being discarded. Further options also exist where data is stored after processing, depending on the specific use-case.

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 [2], developed at BackType and Twitter, Inc., and also the up-and-coming Spark Streaming project [12], developed at UC Berkeley’s AMPLab [5], 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 processing in mind. For example, the before mentioned Spark Streaming project, along with its mother project, Spark [6], was originally designed for highly parallelisable data with the use-case in mind of processing data in-memory using highly iterative machine learning algorithms related to data analytics [59].

What is proposed in this chapter is a realtime big data processing pipeline for the aforementioned railway monitoring system by the IRT at Monash University. This processing pipeline will allow the possibility that data is able to be streamed in realtime from the railway and processed according to the team’s given processing requirements.

This chapter will be structured as follows:

A brief outline of the domain on big data processing, both batch and realtime processing, will be given in §1.1. Furthermore, an overview of the Monash University Institute of Railway Technology’s project, upon which this project is based on, will be covered in the aforementioned chapter. In §1.2, the goals and aims of the project will be given, as well as more specific goals and aims relating to the subproject covered in this thesis. Research questions for this subproject will be given in §1.3. Finally, the structure of this overall thesis will be given in §1.5, before concluding this chapter in §1.6.

1.1 Research Context

1.1.1 Big Data

Big data, as explained previously, is becoming commonplace in both industry and academia. Everyday companies are finding that they are generating a large volume data and that their traditional relational database management system (RDMBS) solutions cannot scale

to the epic proportions needed to handle this data in an efficient and robust manner [60]. Hence, companies and academics alike have started looking at alternative solutions designed with the goal of handling these massive datasets.

The most popular solution for this problem, up until recently, has been the MapReduce model of programming along with some type of scalable distributed storage system [21]. The MapReduce model was started at Google, Inc. with their own proprietary implementation along with their proprietary distributed file system, known as the Google File System (GFS) [41]. Without going into the low-level details of MapReduce and GFS, the use of this solution at Google allowed the company to easily handle all the data that was coming into their servers, including that related to Google Search, and perform the necessary processing operations that was needed at the time [41] [35].

1.1.2 Batch data processing

From the success of MapReduce usage combined with GFS at Google, the open-source community responded swiftly with the development of the Apache Hadoop framework ¹. Hadoop originally offered an open-source implementation of MapReduce and their own open-source distributed file system known as the Hadoop Distributed File System (HDFS) [71].

Hadoop soon became the subject of mass-adoption in both industry and academia, being deployed at a fast rate. Development of the Hadoop framework also grew at a fast rate, with new applications related to HDFS and MapReduce being built on top of Hadoop, greatly benefiting the ecosystem as a whole. Some of these applications grew into widely adopted systems in their own right. For example, Hadoop applications such as Apache Pig [39] and Hive [84] allow for easy querying and manipulation of data stored on HDFS, both coming with the addition of their own query languages [66].

As further non-MapReduce model applications became of interest to the Hadoop community, Hadoop soon developed a further abstraction on top of the underlying resources (in most cases, HDFS). The goal of this was to facilitate the development and deployment of many different applications, varying in use-case, which could be run on the Hadoop ecosystem, without forcing developers to fit their application into the MapReduce model. This development was known as Apache Hadoop YARN: Yet Another Resource Negotiator, which can be thought of as an operating system-like abstraction sitting atop of the available Hadoop resources [92]. The abstraction provided by YARN facilitated the development of much more advanced, and non-MapReduce technologies which have since become widely used parts of the Hadoop ecosystem [47].

1.1.3 Realtime data processing

One of the major limitations of Hadoop, and the MapReduce model in general, soon became obvious: MapReduce was designed with the goal of being able to process batches of data, hence, given Hadoop's dominance, batched data processing was the focal point of the entire distributed data processing domain [54]. Essentially, batched data processing is where data gets collected first into large enough batches before being processed all-at-once. The point of processing in such a way is so there would be less overheads than attempting to process each individual datum as it arrives. For a lot of use-cases this was, and still is, fine as there were no other drawbacks apart from a high level of latency between the stages of when the data arrives and when it gets processed. However, for

¹<https://hadoop.apache.org>

other applications, such as stock trading, sensor monitoring, and web traffic processing, a more low-latency, realtime solution was needed [54].

Soon, many solutions, with different use-cases and design goals, were developed in the area of distributed stream processing systems (DSPS). Given the Hadoop ecosystem that was already widely adopted, most of these DSPSs were built upon the still new YARN layer, ensuring overall compatibility with the Hadoop ecosystem, and the underlying HDFS. Some examples of such projects include the beforementioned Apache Storm, currently being used at Twitter, Inc. [88], among many other companies. Also up-and-coming projects, such as Apache Samza which is a recently open-sourced project, currently being used in production at LinkedIn Corporation [11].

1.1.4 Monash IRT Railway Project

The railway project that has been developed at Monash University’s Institute of Railway Technology ² uses numerous sensor technologies on certain train cars, such as the Track Geometry Recording Car (TGRC) and the Instrumented Ore Car (IOC), to monitor railway track conditions and detect track abnormalities [32, 33]. These train cars operate in the Pilbara region of Western Australia, continuously performing round trips from a port to a given loading point, where they are loaded with recently mined minerals and ores.

As is currently the case, data is received and processed using batch data processing technologies. Due to limited coverage of cellular networks in the Pilbara region of Western Australia, in which the trains currently operate, sensor data from a given trip is automatically transmitted in large batches to be received by remote servers once a train has concluded a round trip and arrives back in port from a loading point [83]. Given the current cellular network infrastructure in the region, this is the only feasible option for transmission of data, however Monash IRT have indicated that given future improvements in cellular network infrastructure, streaming the data back to remote servers in realtime is a likely possibility. This leads to the potential possibility of this research project’s outcomes outlined in this thesis.

The form of batch handling and processing performed on the railway data currently leads to a number of limitations and problems. The data is currently stored in a relational database management system (RDBMS), and with the current implementation of the sensors, the data received is not consistently structured. In fact, the only sensor data guaranteed to be received in each batch is geographic location and time. Due to the highly structured nature of RDBMS technology, certain work-arounds need to be performed on the data so that it is compliant to RDBMS schemas, such as the insertions of default values in the case of missing attributes. Furthermore, low query performance has been noted as a problem plaguing the IRT team working with the railway data. They wish to resolve these problems by looking into non-relational models for their data storage and processing systems.

1.2 Research Aim

The main aim of this research project is to develop a fully automated, non-relational big data pipeline to manage the data received from railway car sensors as a part of Monash University’s Institute of Railway Technology project. The intention is to replace their current relational solution, offering at least the same capabilities at a larger scale, and

²<https://platforms.monash.edu/irt/>

with higher performance. The scope of this project is relatively large, and the project, or subproject, outlined in this thesis only covers a portion of the greater aforementioned project.

The main aim of this subproject is to investigate possible solutions of dealing with the railway sensor data being streamed and processed in realtime. This will also involve studying non-relational big data systems, however the details of such systems in the scope of the greater project will be left up to other individuals working on those subprojects.

To allow the possibility of realtime data streaming from the railway sensors, appropriate data stream processing system (DSPS) technology needs to be looked at, tested, and evaluated. This makes up the core part of this subproject's work. These DSPS technologies need to appropriately take, or accept, the data from some specified source, *e.g.* the sensors, apply any realtime processing logic that is required, *e.g.* pre-processing, then forward the data on for handling at another source, *e.g.* long term storage, such as HDFS.

This DSPS system will act as a filtering "pipeline" through which the sensor data flows. By the end of this subproject, we aim to have proof-of-concept implementations of the pipeline working on the National eResearch Collaboration Tools and Resources (NeCTAR) cloud services [8], making use of certain DSPS technologies.

1.3 Research Questions

The following research questions are the main focus points of this subproject:

1. Which existing DSPS technologies can be used to build a data filtering pipeline for the Monash University IRT project?
2. What criteria-based qualitative and performance-based quantitative tests can be performed to recommend a particular DSPS technology for building the pipeline?
3. How can the data filtering pipeline be designed to be extensible, allowing for the addition of future realtime processing requirements?

Additionally, after answering each of these preliminary research questions, we will want to properly implement the theoretical discoveries from each stage. We do this with the goal of achieving a deployable pipeline that can be then be used in the testing and overall evaluation stages.

1.4 Significance and Contribution

1.5 Thesis Structure

The proposed structure of the remainder of the thesis is as follows. §2 looks at prior research and work into the area of big data stream processing, performed both in industry and academia. A clear overview of the DSPS technologies chosen for this project will be given in §3, along with detailing the experiments that will be performed and evaluated. §4 will detail how the experimental systems built for this project were implemented using the DSPS technologies highlighted in §3, with an evaluation of the experiments being detailed in §5. Finally, §6 will conclude the thesis, along with highlighting any further research gaps that have been made apparent as the result of this project.

1.6 Summary

This chapter has introduced the overall project, and in-turn the subproject upon which this thesis will be based. It has described the current state of the Monash University Institute of Railway Technology's project and the current problems that they are faced with, such as the need to deal with non-consistently structured data and low query performance. These problems are exacerbated given the non-realtime nature of the current data, where issues with sensors are only discovered much later after-the-fact. To overcome these issues, and look forward into the hypothetical, but likely, possibility of having the technological infrastructure to support realtime data processing, this research aims to develop a realtime processing pipeline, taking the data straight from the railway sensors and perform any needed realtime processing.

This chapter also outlined a brief context of the research project and big data as a whole, the scope of this subproject, this subproject's research questions. Finally it was concluded with an overview of this thesis' entire structure. The following chapter will look into previous research that has been done on realtime big data processing and handling, along with going into more detail of the railway project's needs.

2 Literature Review

To lay the ground work and context of this research project, this chapter goes into detail on existing research and work done in the areas of big data processing. Both more traditional batch processing of big data will be looked at, along with newer advances in the area of realtime stream processing. Different technologies enabling these processes will be discussed in detail, with the main focus being placed on the widely used open-source projects, such as those encompassed within the Hadoop ecosystem.

The realtime processing of big data is now of more great importance to both academia and industry than it has ever been before. Advancements and progress in modern society can be directly attributed back to data and how we deal with it. The value of data has become increasingly more apparent, and data has become a sort of currency for the information economy [75]. Hence, those in society who realised the value of data early hold immense power over the entire economy, and in turn society, overall [58]. From seemingly inconsequential gains at the macro level, such as the ability to more accurately predict the rise and fall of airline tickets [34], to those of utmost importance for society as a whole, such as predicting and tracking the spread of the Swine Flu Pandemic in 2009 more accurately than the United States Centers for Disease Control and Prevention could using big data processing [69, 61]. It is applications of big data processing such as these that have been recognised by academics and organisations in industry alike, with the last decade seeing a major shift in research and development into new methods for the handling and processing of big data.

This review will be structured in two main sections. In §2.2, an overview and history for the existing major open-source batch processing systems for big data will be given. This will mostly surround the projects which fall under the umbrella of the Hadoop ecosystem. In §2.3, the current state of realtime data processing systems for big data will be discussed, including examples of the most widely used open-source data stream processing systems (DSPS). Comparisons will be drawn between the DSPS technologies and outlined in how they differ. Further discussion and analysis will be given in §2.4, including identifying the gaps in existing literature to which this research project aims to fill, before concluding in §6.

2.1 Big data processing background

Much work has been done in the area of big data processing in the past decade, given the rise of large scale applications and services, a lot of the time involving large numbers of users and their given data. Data processing in general has been of large importance for many years, as no matter the business model, some sort of data has had to be dealt with. Traditionally, this has often been handled through the use of relational models, as introduced by Codd [30], through use of relational database management systems (RDBMS), providing methods for the long term storage of data and functions for performing queries and manipulation on that data [19], such as through use of an implementation of structured query language (SQL) [27].

Given larger scales of data in newer projects, along with the needs to scale horizontally in terms of parallel and distributed processing, the more traditional relational forms of data management have become less attractive due to their complexity in terms of horizontal scaling [17]. Hence, companies dealing with these large scale use cases have started to look outwards to new directions of data management. This unattractiveness of existing

technologies, in terms of scaling and parallelism, arguably led to the need to develop something new with these objectives in mind.

Note that in this chapter, we will refer to the processing of data in realtime as simply “realtime data processing”. This term should be assumed to encompass the meanings that are also often represented through use of terms such as “data stream processing”, “realtime stream processing”, and “stream processing”.

2.2 Batch data processing

Over the last decade, the main “go-to” solution for any sort of processing needed on datasets falling under the umbrella of big data has been the MapReduce programming model on top of some sort of scalable distributed storage system [21]. From a overly simplified functional standpoint, the MapReduce programming model essentially combines the common **Map** and **Reduce** functions (among others), often found in the standard libraries of many functional programming languages, such as Haskell [57], Clojure [48], or even Java 8 [80], to apply a specified type of processing in a highly parallelised and distributed fashion [98].

The MapReduce data processing model specialises in batch mode processing on data which is already available. Batch data processing can be thought of where data needed to be processed is first queued up in “batches” before processing then begins on those batches. Once ready, those batches get fed into the processing system and handled accordingly [31].

2.2.1 MapReduce and GFS

Dean and Ghemawat, in [35], originally presented MapReduce as a technology that had been developed internally at Google, Inc. to be an abstraction to simplify the various computations that engineers were trying to perform on their large datasets. The implementations of these computations, while not complicated functions themselves, were obfuscated by the fact of having to manually parallelise the computations, distribute the data, and handle faults all in an effective manner that met their performance requirements. The MapReduce model that was developed, enabled these computations to be approached from a more simple, high-level manner without the programmer needing to worry about optimising for available resources. In this way, it was more “declarative” while still allowing it to be written in a procedural language, such as Java. Furthermore, the MapReduce abstraction provided high horizontal scalability to differently sized clusters running on modest hardware.

While MapReduce conceptually offers many features that may be found in parallel database technologies, Dean and Ghemawat, in [36] emphasise the differences, and specifically flawed assumptions made about MapReduce in light of parallel databases. Issues in parallel databases such as fault tolerance when it comes to job failure, difficulties with input data from heterogeneous sources, and performing more complicated functions than those supported in SQL, are all fully supported and taken into consideration with MapReduce. Notably, while the programming model allows for further complicated functions, it also remains to be simple with only two major functions that need to be implemented: Map and Reduce [68].

As previously stated, the MapReduce programming model is generally used on top of some sort of distributed storage system. In the previous case at Google, Inc., in the original MapReduce implementation, it was implemented on top of their own proprietary

distributed file system, known as Google File System (GFS). Ghemawat et al., in [41], define GFS to be a “scalable distributed file system for large distributed data-intensive applications”, noting that can be run on “inexpensive commodity hardware”. Note that GFS was designed and in-use at Google, Inc. years before they managed to develop their MapReduce abstraction, and the original paper on MapReduce from Dean and Ghemawat state that GFS was used to manage data and store data from MapReduce [35]. Furthermore, McKusick and Quinlan, in [62], state that, as of 2009, the majority of Google’s data relating to their many web-oriented applications rely on GFS.

2.2.2 Hadoop MapReduce and HDFS

While MapReduce paired with GFS proved to be very successful solution for big data processing at Google, Inc., and there was a considerable amount of notable research published on the technology, it was proprietary in-house software unique to Google, and availability elsewhere was often not an option [44]. Hence, the open-source software community responded in turn with their own implementation of Google’s MapReduce and a distributed file system on which it runs, analogous to the role GFS played. The file system is known as the Hadoop Distributed File System (HDFS). Both of these projects, along with many others to date, make up the Apache Hadoop big data ecosystem³. The Apache Hadoop ecosystem, being a top level Apache Software Foundation open-source project, has been developed by a number of joint contributors from various organisations and institutions such as Yahoo!, Inc., Intel, IBM, UC Berkeley, among others [46].

While Hadoop’s MapReduce implementation very much was designed to be a complete functional replacement for Google’s MapReduce, based upon published works by Google, HDFS is an entirely separate project in its own right. In the original paper from Yahoo! [72], Inc., Shvachko et al. present HDFS as “the file system component of Hadoop” with the intention of being similar to in external appearance to the UNIX file system, however they also state that “faithfulness to standards was sacrificed in favour of improved performance”.

While HDFS was designed with replicating GFS’ functionality in mind, several low-level architectural and design decisions were made that substantially differ to those documented in GFS. For example, in [24], Borthakur documents the method HDFS uses when it comes to file deletion. Borthakur talks about how when a file is deleted in HDFS, it essentially gets moved to a `/trash` directory, much like what happens in a lot of modern operating systems. This `/trash` directory is then purged after a configurable amount of time, the default of which being six hours. To contrast with this, GFS is documented to have a more primitive way of managing deleted files. Ghemawat, et al., in [41], document GFS’ garbage collection implementation. Instead of having a centralised `/trash` storage, deleted files get renamed to a hidden name. The GFS master then, during a regularly scheduled scan, will delete any of these hidden files that have remained deleted for a configurable amount of time, the default being three days. This is by far not the only difference between the two file systems, this is simply an example of a less low-level technical difference.

HDFS, being designed as a standalone technology, is now used as a data storage component for many different big data technologies, not limited to only those within the Hadoop ecosystem [81] but also projects outside [95, 99]. HDFS was designed with general usage in mind and adhered for horizontal scalability being a key asset, hence runs on inexpensive clusters using modest hardware, while still delivering all its potential advantages [25].

³<https://hadoop.apache.org>

2.2.3 Pig and Hive

Given the popularity of Hadoop, there were several early attempts at building further abstractions on top of the MapReduce model, which were met with a high level of success. As highlighted earlier, MapReduce was originally designed to be a nice abstraction on top of the underlying hardware, however according to Thusoo et al., in [85], MapReduce was still too low level resulting in programmers writing programs that are considered to be “hard to maintain and reuse”. Thus, Thusoo et al. at Facebook Inc. built the Hive abstraction on top of MapReduce. Hive allows programmers to write queries in a similarly declarative language to SQL — known affectionately as *HiveQL* — which then get compiled down into MapReduce jobs to run on Hadoop [86].

Hive, along with HiveQL, provide a platform to run ad-hoc queries on top of HDFS stored data, which are run as batch MapReduce jobs. For example, all generated reports on Facebook to their paid advertisers are generated using batch jobs constructed using Hive, along with several other analytics use cases at Facebook [87]. An added benefit shown at Facebook from the use of Hive, over traditional MapReduce, was the users familiarity with SQL-like declarativeness in regards to querying and manipulating data [26].

Another common abstraction that was developed prior to Hive was what is known simply as Pig. Like Hive, Pig attempts to be a further higher level abstraction on top of the MapReduce model, which ultimately compiles down into MapReduce jobs. what differentiates it more from Hive is that instead of offering a solely declarative SQL-like language for data manipulation and querying, it offers a language that takes more of a procedural and functional pipeline paradigm influenced approach [65]. This still allows for relational algebraic constraints to be specified on the data set in relation to defining the result [67]. Olston et al. describe Pig’s language — known as *Pig Latin* — to be what they define as a “dataflow language”, rather than a strictly procedural or declarative language.

Pig works by having the program first parsed into a directed acyclic graph (DAG), which then can be optimised using DAG-related graph theory, before having that DAG being compiled down into native MapReduce batch jobs. The MapReduce stage is also optimised through use of function ordering, before being submitted to run on a Hadoop cluster [76].

Hive, also being a high level abstraction on top of MapReduce, also enable many of their own optimisations to be applied to the underlying MapReduce jobs during the compilation stage [40, 86] as well as having the benefit of being susceptible to manual query optimisations, familiar to programmers familiar with query optimisations from SQL [45].

2.3 Realtime data processing

With HDFS being an open source project with a large range of users [94] and code contributors [46], it has grown as a project in the last few years for uses beyond what it was originally intended for: a backend storage system for Hadoop MapReduce. HDFS is now not only used with Hadoop’s MapReduce but also with a variety of other technologies, a lot of which run as a part of the Hadoop ecosystem, but also those which do not. Big data processing has moved on from the more “traditional” method of processing, involving MapReduce jobs, which were most suitable for the batch processing of data, to those methods which specialise in the realtime processing of streamed data. The main difference of which is that rather than waiting for all the data before processing can be started, in realtime data processing, the data can be streamed into the processing system and processed at any time with whatever data is made available.

The use cases for such realtime data processing differ much from batched processing in ways a lot of the time relating to the availability of data. With batch jobs, generally processing is done after-the-fact, or after data has been collected from various sources and stored in a system, similar to HDFS. With realtime data processing, data can be processed as soon as it is received from sources, without using a storage system, such as HDFS, to first hold it. For example, an array of sensors can stream their data directly to the data stream processing system (DSPS), without any other intermediate step. From there, the DSPS will process the data according to how it is programmed, and either send it on to further systems, or store it for later batch use.

2.3.1 Hadoop YARN

As previously looked at, the focus of the MapReduce model was performing distributed and highly parallelised computations on distributed batches of data. This suited a lot of the big data audience, and hence Hadoop became the dominant method of big data processing for many years [59]. However for some more specialised applications, such as the realtime monitoring of sensors, as touched on earlier, stock trading, and realtime web traffic analytics, the high latency between the data arriving and actual results being generated from the computations was not satisfactory [54]. Furthermore, such use-cases were simply not suited for batch processing.

A recent (2013) industry survey on European company use of big data technology by Bange, Grosser, and Janoschek, noted in [20], shows that over 70% of responders show a need for realtime processing. In that time, there has certainly been a response from the open-source software community, responding with extensions to more traditional batch systems, such as Hadoop MapReduce, along with complete standalone DSPS solutions.

On the Hadoop front, the limitations of the MapReduce model were recognised, and a large effort was made in developing the “next generation” of Hadoop so that it could be properly extensible and used with other programming models not locked into the rigidity of the MapReduce model. This change in Hadoop became known officially known as YARN (Yet Another Resource Negotiator). According to the original developers of YARN, Vavilapalli et al. state that YARN enables Hadoop to become more modular, decoupling the resource management functionality of Hadoop from the programming model (traditionally, MapReduce) [91]. For Hadoop end-users, this decoupling essentially allowed for non-MapReduce technologies to be built on top of the Hadoop ecosystem, which was traditionally a part of MapReduce, allowing for much more flexible applications of big data processing on top of the existing robust framework Hadoop provides.

Examples of such systems now built, or in some cases ported, to run on top of Hadoop, providing alternative processing applications and use cases include:

- Dryad, a general-purpose distributed execution system from Microsoft Research [51]. Dryad is aimed at being high level enough to make it “easy” for developers to write highly distributed and parallel applications.
- Spark, a data processing system, from researchers at UC Berkeley, that focuses on computations that reuse the same working data set in-memory over multiple parallel operations [101]. Spark, and in particular Spark Streaming, will be looked at further in §2.3.3.
- Storm, a realtime stream processing system [63, p. 244]. Storm allows specified processing on an incoming stream of data indefinitely, until stopped. Storm will be looked at further in §2.3.2.

- Tez, an extensible framework which allows for the building of batch and interactive Hadoop applications [82].
- REEF, a YARN-based runtime environment framework [29]. REEF is essentially a further abstraction on top of YARN, with the intention of making a unified big data application server.
- Samza, a relatively new realtime data processing framework from LinkedIn. Discussed further in §2.3.4.

These are just some of the more popular examples of applications built to interact with the Hadoop ecosystem via YARN.

With the emergence of YARN, the ability to build DSPS technologies on top of Hadoop led to a lot of interest in realtime processing solutions that was often looked over by previous Hadoop batch users.

2.3.2 Storm

One very notable DSPS technology developed independently of Hadoop, and that is gaining immense popularity and growth in its user base, is the Storm project. Storm was originally developed by a team of engineers lead by Nathan Marz at BackType [78]. BackType has since been acquired by Twitter, Inc. where development has continued. Toshniwal et al. [89] describe Storm, in the context of its use at Twitter, as “a realtime distributed stream data processing engine” that “powers the real-time stream data management tasks that are crucial to provide Twitter services” [89, p. 147]. Since the project’s inception, Storm has seen mass adoption in industry, including among some of the biggest names, such as Twitter, Yahoo!, Alibaba, and Baidu [79].

While Storm does not run on top of YARN, there is currently a large effort from engineers at Yahoo!, Inc. being put into a YARN port for Storm, named “storm-yarn” [42, 56]. This YARN port will allow applications written for Storm to take advantage of the resources managed in a Hadoop cluster by YARN. While still in early stages of development, “storm-yarn” has begun to gain attention in the developer community, through focus from channels such as the Yahoo Developer Network [93] and Hortonworks [37].

The programming model for storm consists of a Storm topology. This topology is made up of various components, which either “transform” the data in some specified way (known as a *bolt*), or input the data from a specified source before outputting it to one of the given bolts (known as a *spout*). The layout of spouts and bolts are then constructed in a separate topology definition, defining the way the data streams (represented as tuples) flow through the topology [52]. An example Storm topology is shown in Figure 1.

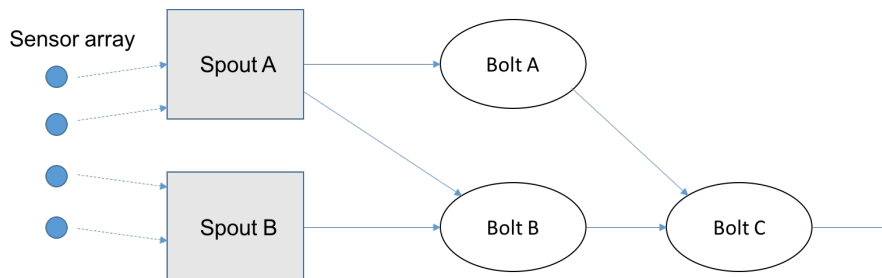


Figure 1: An example Storm topology showing the source of the tuple stream from spouts A and B, gotten from an array of sensors, which then forwards the stream on for processing at bolts A and B. Note that bolts can forward streams onto other bolts.

Storm, while predominantly written in Clojure, running on the JVM, allows topologies to be defined and written in any available programming language through the use of a Thrift definition for topologies [3]. Thrift is a software library, developed at Facebook, Inc., for the purpose of defining datatypes and interfaces for use in multiple programming languages [73]. Storm provide example Storm projects written in Java, Clojure, and Python, however most official Storm documentation, at present, use Java.

2.3.3 Spark Streaming

Spark is another popular big data distributed processing framework, offering of both realtime data processing and more traditional batch mode processing, running on top of YARN [101]. Spark was developed at UC Berkeley, and is notable for its novel approach to in-memory computation, through Spark’s main data abstraction which is termed a *resilient distributed dataset* (RDD). An RDD is a set of data on which computations will be performed, which can be specified to be cached in volatile memory across multiple machines. What this then allows is multiple distributed operations being performed on this same dataset in parallel without such overhead, as secondary storage IO operations, which is noted by Spark as a performance limitation of MapReduce. Spark is designed with highly iterative computations in-mind.

The Spark project has experienced a large growth in terms of project contributions over the past few years, now encompassing a number of subprojects running on top of the Spark engine, all suited for different use cases. These include:

- **Spark SQL:** An SQL-like engine that runs on top of Spark, allowing Spark users to reap the benefits of using a declarative SQL-like language on top of Spark’s functional API [18]. (Previously known as Shark [96]).
- **Spark GraphX:** A graph abstraction on top of the Spark engine, allowing for the efficient expression of distributed graph structures and computations on those structures [95].
- **Spark MLlib:** A distributed scalable machine learning platform on top of Spark, allowing for efficient machine learning capabilities [74].
- **Spark Streaming:** A DSPS platform built on top of the Spark engine, allowing for the processing of data in realtime [102].

Given these different platforms built on top of Spark’s in-memory computation engine, Spark offers solutions for many different big data processing use cases. We are most interested in Spark Streaming, given its realtime DSPS offerings. Spark Streaming uses a different programming model than that which is offered in the base Spark platform, that involves what is labelled as “D-Streams” (discretised streams). D-streams essentially allows for a series of deterministic batch computations be treated as single a realtime data stream [102]. The D-Stream model is specific to the Spark Streaming system — the original batch mode Spark system continues to use the previously mentioned RDD abstraction — and the creators claim performance improvements of being $2 - 5\times$ faster than other DSPS technologies, such as S4 and Storm, given its use of in-memory stream computation [103]. However, this is a topic of debate, with opponents claiming that these Spark performance claims are very much dependent on external factors, such as available resources [43].

Both Spark and Spark Streaming have started to gain notable usage in both industry and research projects in academia in the last few years. Online video distribution company,

Conviva Inc., report to be using Spark for the processing of analytics reports, such as viewer geographical distribution reports [53, 100]. The Mobile Millennium project at UC Berkeley [90], a traffic monitoring system that uses GPS through users' cellular phones for traffic monitoring in the San Francisco Bay Area, has been using Spark for scaling the main algorithm in use for the project: an expectation maximisation (EM) algorithm that has been parallelised by being run on Spark [50].

Spark Streaming, while originally written in Scala, supports official APIs for Java, Scala, and Python. Note that the Python API for Spark has only recently been introduced and, as of yet (version 1.3.1), does not have support for all the data sources which Java and Scala offer [15].

2.3.4 Samza

Samza is a relatively new realtime big data processing framework originally developed in-house at LinkedIn, which has since been open-sourced at the Apache Software Foundation [70]. Samza offers much similar functionality to that of Storm, however instead the running of Samza is highly coupled with the Kafka message broker, which handles the input and output of data streams. Essentially, Kafka is a highly distributed messaging system that focusses on the handling of log data [55], integrating with the Hadoop ecosystem.

While Samza is lacking in maturity and adoption rates, as compared to projects such as Storm, it is built on mature components, such as YARN and Kafka, and thus a lot of crucial features are offloaded onto these platforms. For example, the archiving of data, stream persistence, and imperfection handling is offloaded to Kafka [22]. Likewise, YARN is used for ensuring fault tolerance through the handling of restarting machines that have failed in a cluster, among further resource management [22].

Samza processes streams of tuples through pre-defined jobs which perform any specified operation on those tuples, much like bolts in Storm. The source of streams in Samza comes from Kafka, which can be also used as an output destination for jobs. How the data is received from the actual source (such as a sensor), then sent through Kafka is a matter of further implementation [97]. Note that a layout of jobs in Samza differs considerably from defining a Storm topology layout, in that each individual job defines its input and output sources, rather than jobs being linked together in a separate definition file.

Samza is mainly programmed using Scala, running on the JVM, and allows programming of jobs through a set of Java interfaces [1]. It is officially supported and recommended to program Samza jobs in Java, however it would also be possible to implement the provided interfaces using any JVM language, such as Scala or Clojure.

2.3.5 S4

S4 (Simple Scalable Streaming System) is another realtime big data processing framework that originated at Yahoo!, Inc. that has since been open-sourced [64]. It is a relatively old project compared to the before-mentioned projects, with development becoming less of a priority in the last few years, with the last update being released in 2013. S4 was highly influenced in design by the MapReduce programming model.

Much like what was noted about Samza in §2.3.4, S4 attempts offload several lower level tasks to more mature and established systems specialising in those areas. The logical architecture of S4 lays out its jobs in a network of processing elements (PEs) which are

arranged as a directed acyclic graph. Each of these PEs entail the type of processing to be done on the data at that point in the network. Each of the PEs are assigned to a processing node, a logical host in the cluster. The management and coordination of these processing nodes is offloaded by S4 to ZooKeeper [54]. Much like the before-mentioned Kafka, ZooKeeper in itself is its own complex service used as a part of many different big data infrastructures, including Samza. ZooKeeper specialises in the high-performance coordination of distributed processes inside distributed applications [49].

Note that a number of performance issues have been noted in S4, including non-linear performance scalability, bad resource management, low expectations on fault tolerance, and high reliance on the network [28]. As mentioned earlier, development on S4 has stagnated, adoption of S4 over DSPS platforms, such as Samza, Storm, and Spark Streaming, has not been seen to be major, and literature in big data research have shown a significant lack of focus on S4 when looking at the topics of DSPS technologies.

2.4 Discussion and analysis

From the previously covered literature, it is rather difficult to provide a reasonable comparison for all the different realtime data processing projects. A lot of the claims made in original literature relating to the projects cannot be quantified fairly, as comparisons or tests have not been carried out relating to other projects. Instead, Stonebraker, Çentintemel, and Zdonik proposed what they claim to be the eight requirements for realtime data processing systems [77], which can be used to give an impartial comparison of the previously covered projects. The requirements were defined a number of years prior to the creation of the four main realtime data processing systems that were covered (2005), however are highly cited as being the defining features that the current generation of realtime data processing systems have strived to meet. The requirements, put forward by Stonebraker et al., are summarised as follows:

1: Keep the data moving - This requirement relates to the high mobility of data and importance of low latency in the overall processing. Hence, processing should happen as data moves, rather than storing it first, then processing what is stored.

2: SQL on streams - This requirement states that a high-level SQL-like query language should be available for performing on-the-fly queries on data streams. SQL is given as an example, however it is noted the language's operators should be more oriented to data streams.

3: Handle stream imperfections - Given the high degree of imperfections in data streams, including factors such as missing and out-of-order data, this requirement states that processing systems need to be able to handle these issues. Simply waiting for all data to arrive if some is missing is not acceptable.

4: Generate predictable outcomes - This requirement relates to the determinism associated with the outcomes of specified processes to be applied to data. A realtime processing system should have predictable and repeatable outcomes. Note that this requirement is rather hard to satisfy as in practice data streams are, by character, rather unpredictable. However, the operations performed on given data are required to be predictable.

5: Integrate stored and streamed data - This requirement states that a realtime processing system should provide the capabilities to be able to process both data that is already stored and data that is being delivered in realtime. This should happen seamlessly and provide the same programming interface for either source of data.

6: Guarantee data safety and availability - This requirement states that realtime processing systems should ensure that they have a high level of availability for processing data, and in any cases of failures, the integrity of data should remain consistent.

7: Partition and scale applications automatically - This requirement states that the partitioning of and processing of data should be performed transparently and automatically over the hardware on which it is running on. It should also scale to more different levels of hardware without user intervention.

8: Process and respond instantaneously - This requirement relates to delivering highly responsive feedback to end-users, even for high-volume applications.

The previously covered literature has been used to determine whether or not the before-mentioned realtime data processing systems adhere to these requirements. The outcome of this is shown in Table 1.

Note that in Table 1, those cells with “N/A” as the value simply mean that the literature is inconclusive on whether or not they adhere to the particular requirement.

Table 1: **Realtime data processing systems compared**

	Storm	Spark Streaming	S4	Samza
1: Data mobility	Fetch	Micro-batch	Push	Fetch (Kafka)
2: SQL on streams	Extension (Trident)	No	No	No
3: Handling stream imperfections	User responsibility	N/A	No	Yes (Kafka)
4: Deterministic outcomes	N/A	N/A	Depends on operation	N/A
5: Stored and streaming data	Yes (Lambda architecture)	Yes (Spark)	N/A	No
6: High availability	Yes (rollback recovery)	Yes (checkpoint recovery)	Yes (checkpoint recovery)	Yes (rollback recovery)
7: Partition and scaling	User responsibility	N/A	Yes (KVP)	Yes (Kafka topics)
8: Instant response	N/A	N/A	N/A	N/A

2.5 Conclusion

The choosing of appropriate realtime data processing platforms for the processing of a given application and dataset is an important, and often confusing, problem. Most platforms compare themselves in terms of performance with other frameworks, often which are disputed by members in other platform “camps” [43, 38]. Hence, providing an informed recommendation based on the features each platform offers, and an impartial performance evaluation is a much needed and important contribution to address this gap in qualitative and quantitative evaluations of the available platforms.

This research will further the field of realtime big data processing in addressing these shown gaps, and making the decision process far more streamlined for researchers and developers of DSPS technologies alike.

3 Data Stream Processing System Technology Overview

This chapter describes the choices for data stream processing systems (DSPS) in which the pipeline for the Monash University Institute of Railway Technology will be built, along with the methods which will be used for evaluating the different technologies used. As identified in the the previous chapter, a number of relatively new DSPS technologies exist and have been in use in different projects at different companies. Here we will look at more of an overview of the possibilities they offer in terms of programming interfaces. Furthermore, we will outline the testing parameters that will be used in a later chapter to evaluate the usage of each technology, highlighting both quantitative performance-based metrics as well as qualitative metrics looking at the ease of programmability from the point-of-view of the DSPS programmer.

This chapter will first start with a look at the choices in DSPS technology which will be used further in the project, in §3.1, before going into the evaluation methods and approach in §3.4. Finally, we will look at an overview of the testing data in §3.5, before concluding in §3.6.

3.1 Data Stream Processing System Technology Choices

The DSPS technologies that have been chosen to be focused on in this sub-project include the following:

- Samza
- Storm
- Spark Streaming

Literature concerning these DSPS technologies have been covered in §2.3, however we will look into more depth into the systems regarding their usage from the programmer’s point-of-view. Note that in the previous chapter, a further DSPS technology, S4, was covered, however due to its notes decline in usage and development in the chapter, it has been decided to omit the use of the technology from this project.

3.1.1 Samza Programming

Note that most of the content in this section is sourced from the official Samza 0.9 documentation. [14]

Samza defines itself as a realtime data processing system which allows the processing of streams made up of immutable messages of similar type or category. Streams in Samza exist independently of other concepts, allowing themselves to be *consumed* from or *produced* to by various supported systems. The concepts of consuming and producing refer to the actions of getting data from a stream and placing data onto a stream, respectively. Samza defines the systems that can be used for consuming and producing to streams to be any piece of software that implements the stream abstraction. For example, Kafka, a popular distributed messaging queueing system, is often used to consume and produce to Samza streams. Systems such as these can be “plugged” into a Samza project to work with the same streams which native Samza applications are written to work with.

For a Samza programmer, the most important concepts of the overall Samza project to understand are the concepts of the **system** and **task**. In a Samza project, a system exists

to produce a particular stream which then can be processed by a task which consumes it. A task will always then produce output to another stream. Systems and tasks can then be arranged in a pipeline-like structure allowing a specified type of processing on the messages which travel on those streams. Hence, it is easy to think of a system as the source of data for an overall Samza project. A very simple example of such a Samza project is shown in Figure 2.

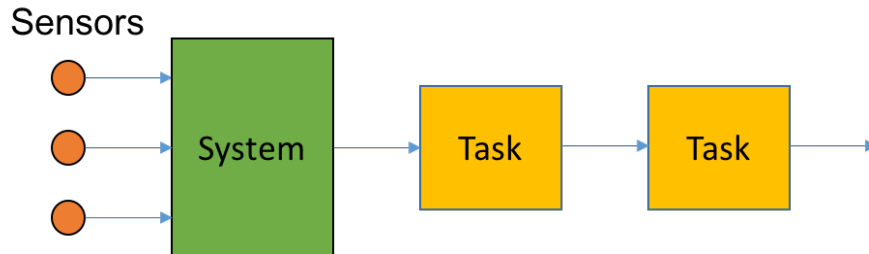


Figure 2: A simple Samza pipeline showing the roles of the system, task, and stream concepts. Streams are represented as arrows.

When programming a Samza project, interfaces exist for both the system and task concepts through the `org.apache.samza.system.SystemConsumer` and `org.apache.samza.task.StreamTask` Java interfaces. Existing systems exist in the Samza project, and may be used to produce streams which can then be processed by an implemented `StreamTask`.

The `StreamTask` interface is arguably the main interface in which the most important processing will be done in a Samza project. It requires implementing a single method, `process`, which passes in an immutable message from the stream on which the task is configured to consume. These messages are then processed as specified by the programmer before the result of the processing is wrapped in a message envelope and produced to a stream specified by the programmer.

The streams on which `StreamTask` implementations consume are specified in a task specific configuration file, independent to that of the implementation source code. These configuration files offer a range of parameters for the execution of tasks, each of which are documented⁴, however can result in quite verbose and hard to understand configurations files. For each task configuration, a system needs to be defined that produces to the stream that the task consumes from. These can either be a self-implemented system, or an existing pluggable system, such as Kafka.

For self-implementation of systems, then `SystemConsumer` interface requires much more understanding and effort to implement than that of the `StreamTask` interface. Methods, such as `start`, `stop`, `register`, and `poll` need to be implemented, allowing for the producing of messages onto a particular stream. The `start` and `stop` methods simply connect and disconnect the system to the underlying system, while the `poll` method takes care of producing any messages from the system. The `register` method is much more complicated in the way that it requires registering the implemented system with lower-level components of Samza to integrate the system into Samza.

Note that, due to Samza’s relative infancy to other DSPS technologies, many of these interfaces either completely lack or offer insufficient documentation. Furthermore, for the same reason, example Samza projects are hard to find. This is a key factor that will be touched upon in later chapters on evaluating the DSPS technologies.

Samza projects are generally compiled and distributed as JAR files, which may or

⁴<https://samza.apache.org/learn/documentation/0.9/jobs/configuration-table.html>

may not contain project dependencies, which then can be run directly on installed Samza distributions. Assembly of these JAR files are conventionally performed through use of Java build automation tools, such as Maven or Gradle.

3.2 Storm Programming

Note that most of the content in this section is sourced from the official Storm documentation. [13]

Storm defines itself as a distributed realtime computation system that provides primitives for performing realtime computations. Storm compares itself to existing big data systems, such as Hadoop MapReduce, which similarly provides primitives for performing batch data computations.

Like Samza, Storm also is built around the key concept of a stream, however Storm's definition of a stream slightly differs to that of Samza's. In Storm, a stream is defined as an unbounded sequence of immutable tuples, rather than a sequence of immutable messages as they are defined in Samza. A tuple is defined as a dynamically typed list of values, in which the values can be of any type. Rather than consuming and producing to streams, Storm approaches computations on streams through *stream transformations* using the previously mentioned realtime computation primitives.

The realtime computation primitives that Storm offers are the components of Storm that are of most interest to the programmer. These primitives are known as the **spout** and **bolt**, previously looked at in §2.3.2. To compare these components to those present in Samza, the spout acts as a source of streams, much like Samza's systems, while the bolts acts as consumers of streams, which process and emit new streams, much like Samza's tasks. Storm offers interfaces for implementation of custom spouts and bolts through the `backtype.storm.spout.ISpout` and `backtype.storm.task.IBolt` interfaces, respectively.

As explained in §2.3.2, spouts and bolts are arranged in a particular way to make up an overall Storm topology, which is used to transform streams in some specified manner. A simple example of such a topology is shown in the previous chapter, in Figure 1.

Basic spouts require implementation of multiple methods, `open`, `nextTuple`, and `declareOutputFields`, which act as a spout initialiser, handler for each tuple to be emitted to a stream, and a tuple type declaration, respectively. Implementation of a simple spout is a much easier task in Storm than a system in Samza, due to different provided interfaces depending on the type of configuration needed for a particular spout. For example, a very basic spout can be made, using a lot of default configuration options, by implementing the `backtype.storm.topology.base.BaseRichSpout` interface.

Bolts also require implementation of three main methods, `prepare`, `execute`, and `declareOutputFields`, allowing bolts to be initialised, process received tuples in a specified manner, and declare tuple value types, respectively.

One major point of different with this components, to their analogous components in Samza, is that the spout and bolt implements are completely independent to the overall Storm topology layout. In Samza, systems and tasks require the output streams to be hardcoded within their respective implementations, however Storm has a further separate source file for defining the topology layout. This topology layout is generally defined in a separate file to all spout and bolt implementations using the provided

`backtype.storm.topology.TopologyBuilder` class. This class exposes methods for defining the topology layout, in terms of spout and bolt placement, along with inputs and outputs for each spout and bolt implementation, as well as overall topology configuration. An obvious downside to this over Samza configuration files is that it requires a recompilation in the case that configuration options need to be tweaked, while Samza configuration files are not part of the compiled source code.

Storm topologies are generally compiled and distributed in the same way as Samza projects; as JAR files containing compiled bytecode, assembled using Java build automation tools.

3.3 Spark Streaming Programming

Note that most of the content in this section is sourced from the official Spark Streaming 1.3.1 documentation. [16]

While many analogies can be made between the key components of Storm and Samza, due to their similarities in design and usage, Spark Streaming presents itself differently. Spark Streaming defines itself as an extension of its parent project, Spark, a big data batch processing system, which affords scalable, high-throughput, fault-tolerant stream processing of live data streams. Data can be streamed into Spark Streaming using a variety of supported sources, from complicated data systems, such as Kafka, to low-level sources, such as TCP sockets.

Internally, Spark Streaming processes data quite differently to Samza and Storm, due to being an extension to the base Spark batch processing system. Data streams are fed into Spark Streaming, which automatically partitions the streams into small batches which are then sent to the underlying Spark batch processing engine. The Spark batch processing engine then outputs a stream of results made up of multiple batches of data. A diagram illustrating this process is shown in Figure ??.

This way of processing streams as a sequence of batches is all enabled through Spark Streaming's stream abstraction, discretised streams, or DStreams. It differs to how streams are conceptualised in Samza and Storm in the way that while streams are still, in essence, an unbounded sequence of data structures, being tuples in Storm, and message envelopes in Samza, these data structures are the same data structure which is used to represent data batches in Spark. These are, of course, the same resilient distributed dataset (RDD) data structure as explained in the previous chapter, in §2.3.3. Hence, Spark Streaming DStreams can be decomposed into a sequence of RDDs, making them fully compatible with batch-mode Spark.

For programming a Spark Streaming job, there are not a set of provided interfaces expected to be implemented, like in Samza or Storm. Instead, all of the computations to be performed on streams are performed on an instantiation of the `org.apache.spark.streaming.dstream.DStream` class returned from a DStream producing function from the `org.apache.spark.streaming.StreamingContext` class. Processing performed on DStreams are generally programmed in a dataflow fashion, with each function performed on the DStream returning a new DStream.

This way of operating on DStreams in a Spark Streaming program leads to all processing being contained in a single file, whereas in Storm or Samza processing would be split between tasks or bolts written in different files. This leads to more straightforward control over DStream transformations when programming, however also leads to much more tightly coupled logic. This will be a topic of interest later when evaluating the different DSPS technologies.

Spark Streaming programs are compiled and distributed in the same fashion in which Samza and Storm projects are. However, for Spark Streaming programs written in Python using their relatively new PySpark API, compilation is unnecessary and an entire Spark Streaming Python project can be distributed and submitted and run on different installations of Spark Streaming straight from the Python source file. This greatly eases the distribution and deployment aspects of Spark Streaming, when using PySpark, in relation to Storm topologies and Samza projects.

3.4 Evaluation Method & Approach

Evaluation of each of the chosen DSPS technologies will be performed using both qualitative and quantitative testing methods of evaluation. The quantitative evaluation methods used will focus on the benchmarking of various features that are common to each of the technologies, and the comparison of the features that each technology supports. The qualitative evaluation methods will focus on looking at the differences in ease-of-use, support for different programming languages and features, and complexity of code written to implement the pipeline on each of the different DSPS technologies.

With looking at the decision of giving a clear recommendation for a particular technology out of the choices stated, we think it is import to look at both qualitative and quantitative aspects for comparison. These systems are significantly non-trivial and vastly different in design and usage. However, as they still afford the same possible functionality, it is very possible to give a properly constructed evaluation of them. Not only is a factor such as performance a critical factor in forming a recommendation, but also are the possibilities each technology affords, in terms of programming and extensibility, when recommending a technology for use in such a large, and long production-life project.

3.4.1 Quantitative methods

The quantitative methods to be used to evaluate the candidate DSPS technologies include mostly performance-based metrics. Each of the metrics that will be used is as follows:

- DSPS start-up time
- Response time to single data value
- Throughput time for large amounts of streamed data
- System resource usage
- Available DSPS features

DSPS start-up time refers to the amount of time between submitting a ready job to a DSPS technology to the time when the system is at a ready state to begin processing incoming data streams. We consider this to be a less important performance metric to assess, however it is still of interest to see the differences present in each candidate technology. It is less important as, for the most part, it will not affect overall performance of each processing system as they will generally be running at all times, waiting for data to be streamed in. This metric will be consistently measured using the internal logs for each DSPS technology.

Response time to a single data value refers to the amount of time taken for each given DSPS technology to completely process a given value sent onto the processing stream through the entire implemented pipeline. The same data value of “42” will be sent onto

the stream, and we will be expecting a result of `(42, true)`. This result will be further explained in the next chapter as we look at the design and implementation of the processing pipeline and simple task it will need to perform. Response time to a single data value is quite an important performance metric to consider as in the real use of the pipeline, as, most of the time, single data values, representing sensor readings, will be placed onto the respective DSPS technology’s processing pipeline. The speed at which each technology can process these data values is important to the overall performance of the system as it will enable more data to be processed without forcing large queues of pending data.

Throughput time for large amounts of streamed data refers to the amount of time taken to process large batches of data delivered at once. While this is less of a realistic use-case for testing in relation to the Monash IRT project, as data is not expected to be delivered in such large amounts, it is aimed at looking to stress-test each of the candidate DSPS technologies to see how they perform and scale to deal with different high-stress data loads. For this stress-test, we will look at five different ranks of tests, using data sets containing 100,000, 500,000, 1,000,000, 5,000,000, and 10,000,000 data values. These data sets will be generated using sample data, received from the Monash IRT team, of prior sensor readings in the IRT project. Throughput time will be measured from the time processing begins on the incoming data, to the time in which all data values in the datasets have finished being processed.

System resource usage refers to the peak amount of resident system memory that is being used by each of the candidate DSPS technologies during processing. We are to measure the peak resident memory usage as this gives an accurate overall indication of how stressful each DSPS technology is to host the operating system. Other system resource metrics, such as CPU usage, have been omitted from tests as each system uses all available CPU processing power to process its data. Resident memory usage will be measured using the built-in UNIX `ps` system utility⁵.

Available DSPS features refers to the features each candidate DSPS technology supports as they relate to Stonebraker, Çentintemel, and Zdonik’s eight requirements for realtime data processing systems, as looked at in the previous chapter in §2.4. The supported features for each technology will be compared to show the processing possibilities available in each technology.

3.4.2 Qualitative methods

The qualitative methods used to evaluate the candidate DSPS systems include mostly programmer-centric metrics, showing the programming possibilities for each technology. Each of the metrics that will be compared is as follows:

- Ease of programmability
- Ease of deployment
- Ease of pipeline extensibility

Ease of programmability refers to a number of different features including official programming language support, complexity of required interfaces to be implemented, and complexity of required system configuration. To summarise, we will look at the difficulty and learning-curve that may need to be overcome to program each candidate DSPS technology. Measuring this presents a more difficult, and somewhat subjective, task compared

⁵<http://unixhelp.ed.ac.uk/CGI/man-cgi?ps>

to the metrics used in the quantitative metrics, however these will be measured using impartial comparisons of the methods which each technology affords to be programmed.

Ease of deployment refers to the methods each candidate technology affords its projects to be deployed and distributed. Similar to with ease of programmability, this metric will be measured by an impartial comparison of each possible method of project deployment and distribution.

Ease of pipeline extensibility refers to the methods used to extend an existing pipeline, or project, existing in each of the candidate technologies. For example, if an existing project is already in production, we will be measuring the ease of which each technology allows that project to be extended in the ways of adding further processing tasks to the existing system, or restructuring an existing project processing layout (for example, a Storm topology layout). Parameters we will be looking at include whether or not an entire project needs to be recompiled and deployed to be extended, and the ease of modifying overall system configurations.

3.5 Overview of the data

As this sub-project focuses on the realtime processing of streaming data, data streams will have to be simulated from datasets acquired from the Monash IRT team. An initial dataset has been given that consists of sensor data in the categories as shown in Table 2.

Table 2: The different types of sensor data received from the Monash University Institute of Railway Technology team.

Name	Data type	Description
Time	Floating point value	Time in seconds since the start of the test.
Speed	Floating point value	Speed in kilometres per hour.
LoadEmpty	-1, 0, 1	1 if load is present, 0 if empty, or -1 if unknown/in-port.
km	Floating point value	Elapsed kilometres travelled since the start of the test.
Lat	Floating point value from -90 to +90	Geographic latitude coordinate at given reading.
Lon	Floating point value from -180 to +180	Geographic longitude coordinate at given reading.
Track	Integer value from 1 to 213841	Given track in which the train is located.
SND1	Floating point value	Spring nest deflection count in millimetres at car corner.
SND2	Floating point value	Spring nest deflection count in millimetres at car corner.
SND3	Floating point value	Spring nest deflection count in millimetres at car corner.
SND4	Floating point value	Spring nest deflection count in millimetres at car corner.
CouplerForce	Floating point value	The amount of force couple detected in the train car.
LateralAccel	Floating point value	The amount of lateral acceleration detected in the train car.
AccLeft	Floating point value	The amount of acceleration detected by the left sensor.
AccRight	Floating point value	The amount of acceleration detected by the right sensor.
BounceFront	Floating point value	The amount of bounce detected at the front of the car.
BounceRear	Floating point value	The amount of bounce detected at the rear of the car.
RockFront	Floating point value	The amount of rock detected at the front of the car.
RockRear	Floating point value	The amount of rock detected at the rear of the car.

The sample data acquired from the IRT team includes 99 999 rows of data recorded from each of the shown sensors, organised in a CSV file with headers. This is a general example of how data is currently received and handled by the IRT team, however in the case of realtime data processing, data would be received in quite a different manner. Rather than being received in large batches of readings, such as the sample dataset acquired, streams of data may be created from each particular sensor, delivering data values to the DSPS systems along input streams a single value at a time. Data is streamed in an asynchronous fashion, and can be simply thought of as a given DSPS system listens on an incoming stream, and acts upon any data value in a specified fashion whenever they may be received.

Hence, to simulate streaming data from the data samples we have acquired, the DSPS

pipeline is constructed to listen on a particular TCP socket connection for any incoming data. Once a data value is encountered on the connection, it is fed into the pipeline and processed accordingly while the pipeline continues to listen on the socket for further incoming data. As the pipeline is constructed in such a way, we can simulate the data being sent out from sensors using a program such as `netcat`⁶, where we can pipe data from a file to a particular socket over a connection. We can also have time intervals between each data value, simulating time between sensor readings.

For the test data for the quantitative tests, as each sensor's data reading values are of the same format (a floating point number), the specific sensor data to be used in the tests is not of major concern. Instead, we have chosen to focus on generating data based of the values from the speed sensor. As specified in Table 2, each value represents a speed reading formatted in kilometres per hour as a floating point number. To generate the different datasets for the data batches for the throughput time quantitative test, we have extracted the speed readings from the provided sample dataset and replicated them a number of times to generate the large batches. Finally, this leaves us with five datasets containing speed readings of sizes 100,000, 500,000, 1,000,000, 5,000,000, and 10,000,000 data values, respectively.

3.6 Conclusion

In conclusion, we have examined the three chosen candidate DSPS technologies in detail in terms of the possibilities they afford for the programmer, outlined the testing parameters which will be used in the overall evaluation of candidate technologies, and the data in which will be used in the highlighted tests. These components lay down a framework for implementation and testing which will be looked at in the following two chapters, respectively.

⁶<http://nc110.sourceforge.net/>

4 Pipeline Implementation

This chapter describes the design and implementation of the prototype data filtering pipeline to be built. This pipeline prototype will be used for the purpose of testing and evaluating the different DSPS technologies, allowing a recommendation to be made for the Monash University Institute of Railway Technology's (IRT) project. Each of the candidate DSPS technologies have been detailed in-depth in previous chapters, and all exhibit the capabilities needed to implement the intended prototype pipeline.

This chapter will first go into the design of the data filtering pipeline, in §4.1, before detailing how each data stream processing system was set up, in §4.2.2. In §4.3, the implementation of the prototype pipeline will be detailed in each of the DSPS systems. Finally, the chapter will be concluded in §??.

4.1 Data Filtering Pipeline Design

The overall design of the data filtering pipeline will be designed in such a way that readings from sensors can be fed into the pipeline, processed in some specified manner, then output for further use, including storage for batch processing, or discarded in the case of noisy data. Thinking of the pipeline in such a way, we can decompose the design into three main components:

1. Sensor data input component
2. Sensor data processing component
3. Results output component

Each component would also need to connect somehow to form the entire pipeline design. The components are listed in order of their precedence of their role in the pipeline. Hence, the first component's output would be the input to the second component, of which the output would be the input to the third component. This creates the dataflow behaviour exhibited in pipelines. This basic design that we have so far is illustrated in Figure 3.

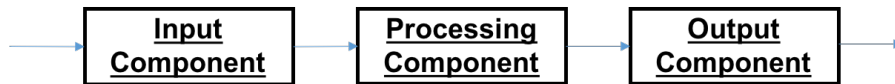


Figure 3: A simplistic overview of our pipeline design. Note the dataflow behaviour exhibited with each component output being an input of the following component.

We now have a simple design of the pipeline, however we are left with the questions of the output of the **Output Component**, the input of the **Input Component**, and the contents of each component.

The output of the **Output Component** will simply be whatever extensions are needed to be made to the pipeline. Hence, rather than acting a sink, this component will be an abstract interface which any extension needed will be able to adhere to to extend the pipeline. Designing the pipeline in this manner directly addresses our third research question touched upon in Chapter 1, in §1.3. This allows the output data to be dealt with in any manner required by the IRT team, whether it be stored for later batch processing, or forwarded on for further realtime processing.

The input of the **Input Component** will be the data read from the sensors to which the pipeline is connected to. The sensors generally would not be directly connected to

the pipeline, so we could have another abstract interface here that any sensor connects to, then the interface sends the data to the actual component. This allows data to be fed into the pipeline from any type of sensor that is able to adhere to the interface.

The contents of each component is a larger task. For the **Processing Component**, it will want to be consisting of a number of tasks, each of which are tasked to process the data in some manner before forwarding it onto whatever the next specified task is. This, in itself, acts as a mini-pipeline where the core of processing will be done. Here we will say that the **Processing Component** consists of n sequential processing tasks. For the **Input Component**, the contents will consist of a task-like component tasked with passing data onto the processing tasks which make up the **Processing Component**. The contents of the **Output Component** will be as described previously: an abstract interface allowing for the extension of the pipeline.

This leaves us with a final pipeline design as illustrated in Figure 4.

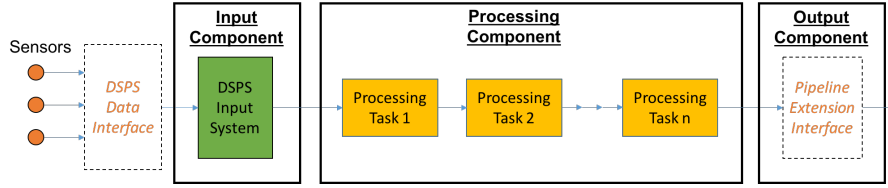


Figure 4: A complete overview of our pipeline design.

4.2 Testing Environment Details

4.2.1 Details on Host System

The testing environment on which the prototype pipelines were implemented using each of the candidate DSPS technologies is the same National eResearch Collaboration Tools and Resources (NeCTAR) cloud computing environment touched upon briefly in Chapter 1. This cloud service platform is provided as an initiative by the Australian Government for researchers that require the infrastructure⁷.

Our environment consists of a single NeCTAR cloud instance of which the details are specified in Table 4.

Table 3: Control system used for testing pipelines.

Distribution	Ubuntu GNU/Linux 14.04.2 LTS
Kernel	Linux 3.13.0-36-generic
Architecture	x86_64
Virtual CPUs	16
Available RAM	64 GB
Available Disk Space	490 GB

The NeCTAR cloud instance utilises a single node upon which all processing is performed. In a real deployment, it is common to use a full cluster upon which the pipeline would be deployed, where processing is shared between nodes. Due to limited testing resources, all testing has been performed on a single node cluster, however this remains a constant feature between all implementation and tests performed in the project. Furthermore, the pipelines built upon each of the candidate DSPS technologies are freely

⁷<https://www.nectar.org.au/about-nectar>

configurable to be deployed to run on arbitrarily sized clusters without requiring changes to the underlying source code.

4.2.2 Details on Data Stream Processing System Technologies

While most details of each of the candidate DSPS technologies has previously been given in the last chapter, it is important to outline the specific versions of each technology, as well as the programming languages used for implementation.

Version 0.9.0 of Samza was used for implementation of the prototype pipeline in Samza. At the time of implementation, this was the major stable version of Samza, accepted as a top-level Apache project. Java, the only official programming language supported by the Samza project, was used, using JDK 1.8.0 as the target implementation.

Version 0.9.4 of Storm was used for implementation of the prototype pipeline in Storm. As with Samza, this was also the major stable version of Storm available. Java 8 was used as the implementation language, also compiling for the JDK 1.8.0.

Version 1.3.0 of Spark and the Spark Streaming extension were used for the Spark Streaming implementations. Prototype pipelines were implemented in both Scala 2.9.2, targeting the JDK 1.8.0, and Python 2.7, which is made possible through the relatively new (as of Spark 1.2.0) PySpark subsystem. Implementations for the Spark Streaming pipeline were built targeting both Spark JVM and PySpark due to numerous undocumented claims that performance differs on both Spark implementations. These claims are addressed further in the following chapter.

The specific JDK implementation for all tests was the Intel 64-bit OpenJDK 1.8.0 release 45 built and packaged for the GNU/Linux distribution of Ubuntu 14.04.2 LTS.

4.3 Implementation of Pipelines in Data Stream Processing System Technologies

4.3.1 Samza

4.3.2 Storm

4.3.3 Spark Streaming (JVM Spark)

4.3.4 Spark Streaming (PySpark)

4.4 Conclusion

In conclusion this chapter has touched upon all details of implementation for each of the prototype pipelines. We have detailed how each prototype data filtering pipeline has been built for each candidate system, ready for subsequent tests and evaluation, of which will be discussed in the following chapter. Furthermore, the overall design of the pipeline that has been implemented has been detailed along with the details of the overall testing environment upon which the pipelines are to be implemented.

5 Discussion and Evaluation

In order to evaluate the three different data stream processing system technologies used to build the proposed pipeline, we compare them in terms of both quantitative and qualitative measures. The quantitative measures used to contrast each technology include aspects such as start-up time, data response time, data throughput allowance, and system resource usage. The qualitative measures used the contrast each technology will highlight different aspects such as ease-of-programmability, available features, ease-of-deployment, and ease-of-extensibility.

The control for each of the evaluation measures will be the system on which the different technologies are deployed. Each identically-functioning pipeline built using different DSPS technologies will be deployed in the same NeCTAR cloud instance, as highlighted in the previous chapter in §4.2. The control NeCTAR cloud instance system is touched upon further in §5.1.

This chapter begins with an overview of the NeCTAR testing environment, in §5.1, upon which all tests are performed. Quantitative evaluations and results are then described in §5.2, while qualitative evaluations and results are described in §5.3. Discussion of the evaluations and subsequent DSPS technology recommendations will be given in §5.5, before concluding in §3.6.

5.1 Testing environment

The testing environment upon which all needed tests will be performed also acts as the major control constant between each of the different DSPS systems that have been built. The environment is based on a single NeCTAR cloud instance, as detailed in. The system details of the given instance are detailed in Table 4.

Table 4: Control system used for testing pipelines.

Distribution	Ubuntu GNU/Linux 14.04.2 LTS
Kernel	Linux 3.13.0-36-generic
Architecture	x86_64
Virtual CPUs	16
Available RAM	64 GB
Available Disk Space	490 GB

The NeCTAR cloud instance utilises a single node upon which all processing is performed. In a real deployment, it is common to use a full cluster upon which the pipeline would be deployed, where processing is shared between nodes. Due to limited testing resources, all testing has been performed on a single node cluster, however this is still a constant feature between all tests, and the pipelines built upon each of the existing technologies are freely configurable to be deployed to run on arbitrarily sized clusters without needing changes to the underlying source code.

5.2 Quantitative Evaluations

5.2.1 Testing data

The testing data for the quantitative stage of evaluation includes data sourced from actual readings from the speed sensors attached to train cars in the Monash University Institute

of Railway Technology data project. We use a total of 99 999 discrete speed readings, some of which are altered to give bad values. The base function of the built pipelines is to detect bad, or noisy, values and flag them before forwarding all values on for further processing in the pipeline, as explained in §4. This is an example of actual realtime data processing that the Monash IRT team have indicated they would be interested in performing.

Hence, the test data is all known. The majority of the test data readings are valid speed readings, however the pipelines built will detect all those which are not valid and flag them.

5.2.2 Tests to be performed

The quantitative tests to be performed and contrasted will include the following:

- System start-up time
- Response time to streaming data
- Data throughput allowance
- System resource usage

Each of these can be quantified and compared easily via tracking various aspects of the system.

Start-up time will be measured from the time each DSPS technology is started via the system shell until the time the DSPS technology is ready to accept incoming data streams. This can be measured using the UNIX `time` command, and allowing the DSPS technology terminate upon start-up, as well as the timestamps recorded in the output logs for each of the DSPS technologies.

Response time will be measured by the amount of time a ready DSPS technology takes from the time a data value is pushed onto the input stream until the time that the result of that given data value is pushed onto the output stream. This can be measured using the output logs for each of the DSPS technologies.

Data throughput allowance will be measured by comparing how much time it takes for the DSPS technology to process the entirety of our test data. This can be measured using the output logs for each of the DSPS technologies.

System resource usage will be measured by comparing the amount of system memory *reserved* for use for each of the running DSPS technologies. These are expected to vary during runtime, so samples will be taken at the peak system resource usage of each DSPS technology during the processing of our test data. Measurements will be taken using the UNIX tool `ps` to collect readings for the amount of CPU usage and virtual memory reserved for use by the DSPS technology. Note that the amount of memory reported by `ps` is simply the amount of virtual memory reserved for use by the process, as opposed to the actual amount of memory being used.

5.2.3 Outcomes of Tests Performed

5.3 Qualitative Evaluations

5.4 Tests to be performed

The qualitative tests to be performed and contrasted will include the following:

- Ease of programmability
- Available DSPS features
- Ease of deployment
- Ease of extensibility

As each of the underlying systems are considerably different in

5.5 DSPS Technology Recommendations

6 Conclusions

6.1 Research Contribution

6.2 Future Work

References

- [1] Samza - api overview. <https://samza.apache.org/learn/documentation/0.9/api/overview.html>. (Visited on 2015-05-12).
- [2] Storm, distributed and fault-tolerant realtime computation. <http://storm-project.net>, November 2013.
- [3] About storm. <https://storm.apache.org/about/multi-language.html>, 2014. (Visited on 2015-05-12).
- [4] Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more. <http://www.amazon.com>, August 2014.
- [5] Amplab - UC Berkeley — Algorithms, Machines and People Lab. <https://amplab.cs.berkeley.edu>, 2014.
- [6] Apache SparkTM— Lightning-Fast Cluster Computing. <https://spark.apache.org>, 2014.
- [7] Australian Synchrotron. <https://www.synchrotron.org.au>, August 2014.
- [8] home — NeCTAR. <http://www.nectar.org.au>, August 2014.
- [9] Institute of railway technology. <https://platforms.monash.edu/irt/>, March 2014. (Visited on 2015-05-09).
- [10] The Large Hadron Collider — CERN. <http://home.web.cern.ch/topics/large-hadron-collider>, August 2014.
- [11] Samza. <http://samza.incubator.apache.org>, 2014.
- [12] Spark Streaming — Apache Spark. <https://spark.apache.org/streaming/>, 2014.
- [13] Storm Documentation. <https://storm.apache.org/documentation/Home.html>, 2014. (Visited on 06/01/2015).
- [14] Samza - Documentation. <https://samza.apache.org/learn/documentation/0.9/>, 2015. (Visited on 06/01/2015).
- [15] Spark streaming - spark 1.3.1 documentation. <https://spark.apache.org/docs/1.3.1/streaming-programming-guide.html>, 2015. (Visited on 2015-05-12).
- [16] Spark Streaming - Spark 1.3.1 Documentation. <https://spark.apache.org/docs/latest/streaming-programming-guide.html>, 2015. (Visited on 06/01/2015).
- [17] AGRAWAL, D., DAS, S., AND EL ABBADI, A. Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology* (2011), ACM, pp. 530–533.
- [18] ARMBRUST, M., XIN, R. S., LIAN, C., HUAI, Y., LIU, D., BRADLEY, J. K., MENG, X., KAFTAN, T., FRANKLIN, M. J., GHODSI, A., ET AL. Spark sql: Relational data processing in spark.
- [19] ASTRAHAN, M. M., BLASGEN, M. W., CHAMBERLIN, D. D., ESWARAN, K. P., GRAY, J., GRIFFITHS, P. P., KING, W. F., LORIE, R. A., MCJONES, P. R., MEHL, J. W., ET AL. System r: relational approach to database management. *ACM Transactions on Database Systems (TODS)* 1, 2 (1976), 97–137.

- [20] BANGE, C., GROSSER, T., AND JANOSCHEK, N. Big data survey eu-
rope - usage, technology and budgets in european best-practice compa-
nies. [http://www.pmone.com/fileadmin/user_upload/doc/study/BARC_BIG-
DATA_SURVEY_EN_final.pdf](http://www.pmone.com/fileadmin/user_upload/doc/study/BARC_BIG_DATA_SURVEY_EN_final.pdf), 2013.
- [21] BIFET, A. Mining big data in real time. *Informatica* 37, 1 (Mar. 2013), 15+.
- [22] BOCKERMANN, C. A survey of the stream processing landscape.
- [23] 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.
- [24] BORTHAKUR, D. The hadoop distributed file system: Architecture and design.
Hadoop Project Website 11 (2007), 21.
- [25] BORTHAKUR, D. Hdfs architecture guide. *Hadoop Apache Project* (2008), 53.
- [26] BORTHAKUR, D., GRAY, J., SARMA, J. S., MUTHUKKARUPPAN, K., SPIEGEL-
BERG, N., KUANG, H., RANGANATHAN, K., MOLKOV, D., MENON, A., RASH,
S., ET AL. Apache hadoop goes realtime at facebook. In *Proceedings of the 2011
ACM SIGMOD International Conference on Management of data* (2011), ACM,
pp. 1071–1080.
- [27] CHAMBERLIN, D. D., AND BOYCE, R. F. Sequel: A structured english query
language. In *Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop
on Data description, access and control* (1974), ACM, pp. 249–264.
- [28] CHAUHAN, J., CHOWDHURY, S. A., AND MAKAROFF, D. Performance evaluation
of yahoo! s4: A first look. In *P2P, Parallel, Grid, Cloud and Internet Computing
(3PGCIC), 2012 Seventh International Conference on* (2012), IEEE, pp. 58–65.
- [29] CHUN, B.-G., CONDIE, T., CURINO, C., DOUGLAS, C., MATUSEVYCH, S., MY-
ERS, B., NARAYANAMURTHY, S., RAMAKRISHNAN, R., RAO, S., ROSEN, J.,
ET AL. Reef: Retainable evaluator execution framework. *Proceedings of the VLDB
Endowment* 6, 12 (2013), 1370–1373.
- [30] CODD, E. F. A relational model of data for large shared data banks. *Communica-
tions of the ACM* 13, 6 (1970), 377–387.
- [31] CONDIE, T., CONWAY, N., ALVARO, P., HELLERSTEIN, J. M., GERTH, J., TAL-
BOT, J., ELMELEEGY, K., AND SEARS, R. Online aggregation and continuous
query support in mapreduce. In *Proceedings of the 2010 ACM SIGMOD Interna-
tional Conference on Management of data* (2010), ACM, pp. 1115–1118.
- [32] DARBY, M., ALVAREZ, E., MCLEOD, J., TEW, G., AND CREW, G. The devel-
opment of an instrumented wagon for continuously monitoring track condition. In
AusRAIL PLUS 2003, 17-19 November 2003, Sydney, NSW, Australia (2003).
- [33] DARBY, M., ALVAREZ, E., MCLEOD, J., TEW, G., CREW, G., ET AL. Track
condition monitoring: the next generation. In *Proceedings of 9th International Heavy
Haul Association Conference* (2005), vol. 1, pp. 1–1.
- [34] DARLIN, D. Airfares made easy (or easier). *The New York Times* 1 (2006), B1.
- [35] DEAN, J., AND GHEMAWAT, S. MapReduce: Simplified data processing on large
clusters. *Commun. ACM* 51, 1 (Jan. 2008), 107–113.

- [36] DEAN, J., AND GHEMAWAT, S. Mapreduce: a flexible data processing tool. *Communications of the ACM* 53, 1 (2010), 72–77.
- [37] EVANS, B., AND FENG, A. Storm-YARN Released as Open Source. <https://developer.yahoo.com/blogs/ydn/storm-yarn-released-open-source-143745133.html>, 2013.
- [38] EVANS, B., AND GRAVES, T. Yahoo compares Storm and Spark. <http://www.slideshare.net/ChicagoHUG/yahoo-compares-storm-and-spark>, 2014.
- [39] GATES, A. F., NATKOVICH, O., CHOPRA, S., KAMATH, P., NARAYANAMURTHY, S. M., OLSTON, C., REED, B., SRINIVASAN, S., AND SRIVASTAVA, U. Building a high-level dataflow system on top of map-reduce: The pig experience. *Proc. VLDB Endow.* 2, 2 (Aug. 2009), 1414–1425.
- [40] GATES, A. F., NATKOVICH, O., CHOPRA, S., KAMATH, P., NARAYANAMURTHY, S. M., OLSTON, C., REED, B., SRINIVASAN, S., AND SRIVASTAVA, U. Building a high-level dataflow system on top of Map-Reduce: the Pig experience. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1414–1425.
- [41] GHEMAWAT, S., GOBIOFF, H., AND LEUNG, S.-T. The google file system. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles* (New York, NY, USA, 2003), SOSP ’03, ACM, pp. 29–43.
- [42] GITHUB. yahoo/storm-yarn. <https://github.com/yahoo/storm-yarn>, 2014.
- [43] GOETZ, P. T. Apache Storm Vs. Spark Streaming. <http://www.slideshare.net/ptgoetz/apache-storm-vs-spark-streaming>, 2014.
- [44] GROSSMAN, R. L., AND GU, Y. On the varieties of clouds for data intensive computing. *IEEE Data Eng. Bull.* 32, 1 (2009), 44–50.
- [45] GRUENHEID, A., OMIECINSKI, E., AND MARK, L. Query optimization using column statistics in hive. In *Proceedings of the 15th Symposium on International Database Engineering & Applications* (2011), ACM, pp. 97–105.
- [46] HADOOP.APACHE.ORG. Who we are. <https://hadoop.apache.org/who.html#Hadoop+Committers>, 2014.
- [47] HARRISON, G. Hadoop’s next-generation YARN. *Database Trends and Applications* 26, 4 (Dec. 2012), 39.
- [48] HICKEY, R. The clojure programming language. In *Proceedings of the 2008 symposium on Dynamic languages* (2008), ACM, p. 1.
- [49] HUNT, P., KONAR, M., JUNQUEIRA, F. P., AND REED, B. Zookeeper: Wait-free coordination for internet-scale systems. In *USENIX Annual Technical Conference* (2010), vol. 8, p. 9.
- [50] HUNTER, T., MOLDOVAN, T., ZAHARIA, M., MERZGUI, S., MA, J., FRANKLIN, M. J., ABBEEL, P., AND BAYEN, A. M. Scaling the mobile millennium system in the cloud. In *Proceedings of the 2nd ACM Symposium on Cloud Computing* (2011), ACM, p. 28.
- [51] ISARD, M., BUDIU, M., YU, Y., BIRRELL, A., AND FETTERLY, D. Dryad: distributed data-parallel programs from sequential building blocks. In *ACM SIGOPS Operating Systems Review* (2007), vol. 41, ACM, pp. 59–72.

- [52] JONES, M. T. Process real-time big data with twitter storm. *IBM Technical Library* (2013).
- [53] JOSEPH, D. Using Spark and Hive to process BigData at Conviva. <http://www.conviva.com/using-spark-and-hive-to-process-bigdata-at-conviva/>, 2011.
- [54] KAMBURUGAMUVE, S., FOX, G., LEAKE, D., AND QIU, J. Survey of distributed stream processing for large stream sources.
- [55] KREPS, J., NARKHEDE, N., RAO, J., ET AL. Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB* (2011).
- [56] KUMAR, R., GUPTA, N., CHARU, S., AND JANGIR, S. K. Architectural paradigms of big data. In *National Conference on Innovation in Wireless Communication and Networking Technology-2014, Association with THE INSTITUTION OF ENGINEERS (INDIA)*.
- [57] LÄMMEL, R. Google’s mapreduce programming model—revisited. *Science of computer programming* 70, 1 (2008), 1–30.
- [58] LIEVESLEY, D. Increasing the value of data. *BLRD REPORTS* 6122 (1993), 205–205.
- [59] 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.
- [60] MARZ, N. *Big data : principles and best practices of scalable realtime data systems*. O’Reilly Media, [S.l.], 2013.
- [61] MAYER-SCHÖNBERGER, V., AND CUKIER, K. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. An Eamon Dolan book. Houghton Mifflin Harcourt, 2013.
- [62] MCKUSICK, M. K., AND QUINLAN, S. GFS: Evolution on Fast-forward. *ACM Queue* 7, 7 (2009), 10.
- [63] MURTHY, A., VAVILAPALLI, V. K., EADLINE, D., MARKHAM, J., AND NIEMIEC, J. *Apache Hadoop YARN: Moving Beyond MapReduce and Batch Processing with Apache Hadoop 2*. Pearson Education, 2013.
- [64] NEUMEYER, L., ROBBINS, B., NAIR, A., AND KESARI, A. S4: Distributed stream computing platform. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on* (2010), IEEE, pp. 170–177.
- [65] OLSTON, C., CHIOU, G., CHITNIS, L., LIU, F., HAN, Y., LARSSON, M., NEUMANN, A., RAO, V. B., SANKARASUBRAMANIAN, V., SETH, S., ET AL. Nova: continuous pig/hadoop workflows. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (2011), ACM, pp. 1081–1090.
- [66] OLSTON, C., REED, B., SRIVASTAVA, U., KUMAR, R., AND TOMKINS, A. Pig latin: A not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2008), SIGMOD ’08, ACM, pp. 1099–1110.
- [67] OLSTON, C., REED, B., SRIVASTAVA, U., KUMAR, R., AND TOMKINS, A. Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM*

- SIGMOD international conference on Management of data* (2008), ACM, pp. 1099–1110.
- [68] PAVLO, A., PAULSON, E., RASIN, A., ABADI, D. J., DEWITT, D. J., MADDEN, S., AND STONEBRAKER, M. A comparison of approaches to large-scale data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (2009), ACM, pp. 165–178.
 - [69] RITTERMAN, J., OSBORNE, M., AND KLEIN, E. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media* (2009), vol. 9.
 - [70] SAMZA.INCUBATOR.APACHE.ORG. Samza. <https://samza.incubator.apache.org/>, 2014.
 - [71] SHVACHKO, K., KUANG, H., RADIA, S., AND CHANSLER, R. The hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)* (May 2010), pp. 1–10.
 - [72] SHVACHKO, K., KUANG, H., RADIA, S., AND CHANSLER, R. The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on* (2010), IEEE, pp. 1–10.
 - [73] SLEE, M., AGARWAL, A., AND KWIATKOWSKI, M. Thrift: Scalable cross-language services implementation. *Facebook White Paper* 5, 8 (2007).
 - [74] SPARKS, E. R., TALWALKAR, A., SMITH, V., KOTTALAM, J., PAN, X., GONZALEZ, J., FRANKLIN, M. J., JORDAN, M. I., AND KRASKA, T. Mli: An api for distributed machine learning. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on* (2013), IEEE, pp. 1187–1192.
 - [75] ST AMANT, K., AND ULIJN, J. M. Examining the information economy: Exploring the overlap between professional communication activities and information-management practices. *Professional Communication, IEEE Transactions on* 52, 3 (2009), 225–228.
 - [76] STEWART, R. J., TRINDER, P. W., AND LOIDL, H.-W. Comparing high level mapreduce query languages. In *Advanced Parallel Processing Technologies*. Springer, 2011, pp. 58–72.
 - [77] STONEBRAKER, M., ÇETINTEMEL, U., AND ZDONIK, S. The 8 requirements of real-time stream processing. *ACM SIGMOD Record* 34, 4 (2005), 42–47.
 - [78] STORM.APACHE.ORG. Storm, distributed and fault-tolerant realtime computation. <https://storm.apache.org/>, 2014.
 - [79] STORM.APACHE.ORG. Storm documentation. <https://storm.apache.org/documentation/Powered-By.html>, 2014.
 - [80] SU, X., SWART, G., GOETZ, B., OLIVER, B., AND SANDOZ, P. Changing Engines in Midstream: A Java Stream Computational Model for Big Data Processing. *Proceedings of the VLDB Endowment* 7, 13 (2014).
 - [81] TAYLOR, R. C. An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. *BMC bioinformatics* 11, Suppl 12 (2010), S1.
 - [82] TEZ.APACHE.ORG. Apache Tez – Welcome to Apache Tez. <http://tez.apache.org>, 2014.

- [83] THOMAS, S., HARDIE, G., AND THOMPSON, C. Taking the guesswork out of speed restriction. In *CORE 2012: Global Perspectives; Conference on railway engineering, 10-12 September 2012, Brisbane, Australia* (2012), Engineers Australia, p. 707.
- [84] THUSOO, A., SARMA, J., JAIN, N., SHAO, Z., CHAKKA, P., ZHANG, N., ANTONY, S., LIU, H., AND MURTHY, R. Hive - a petabyte scale data warehouse using hadoop. In *2010 IEEE 26th International Conference on Data Engineering (ICDE)* (Mar. 2010), pp. 996–1005.
- [85] THUSOO, A., SARMA, J. S., JAIN, N., SHAO, Z., CHAKKA, P., ANTHONY, S., LIU, H., WYCKOFF, P., AND MURTHY, R. Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1626–1629.
- [86] THUSOO, A., SARMA, J. S., JAIN, N., SHAO, Z., CHAKKA, P., ZHANG, N., ANTONY, S., LIU, H., AND MURTHY, R. Hive-a petabyte scale data warehouse using hadoop. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on* (2010), IEEE, pp. 996–1005.
- [87] THUSOO, A., SHAO, Z., ANTHONY, S., BORTHAKUR, D., JAIN, N., SEN SARMA, J., MURTHY, R., AND LIU, H. Data warehousing and analytics infrastructure at facebook. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (2010), ACM, pp. 1013–1020.
- [88] 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 RYABOV, 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.
- [89] TOSHNIWAL, A., TANEJA, S., SHUKLA, A., RAMASAMY, K., PATEL, J. M., KULKARNI, S., JACKSON, J., GADE, K., FU, M., DONHAM, J., ET AL. Storm@twitter. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (2014), ACM, pp. 147–156.
- [90] TRAFFIC.BERKELEY.EDU. History of the Project — Mobile Millennium. <http://traffic.berkeley.edu/project>, 2014.
- [91] VAVILAPALLI, V. K., MURTHY, A. C., DOUGLAS, C., AGARWAL, S., KONAR, M., EVANS, R., GRAVES, T., LOWE, J., SHAH, H., SETH, S., ET AL. Apache Hadoop YARN: Yet another resource negotiator. In *Proceedings of the 4th annual Symposium on Cloud Computing* (2013), ACM, p. 5.
- [92] VAVILAPALLI, V. K., MURTHY, A. C., DOUGLAS, C., AGARWAL, S., KONAR, M., EVANS, R., GRAVES, T., LOWE, J., SHAH, H., SETH, S., SAHA, B., CURINO, C., O'MALLEY, O., RADIA, S., REED, B., AND BALDESCHWIELER, E. Apache hadoop YARN: Yet another resource negotiator. In *Proceedings of the 4th Annual Symposium on Cloud Computing* (New York, NY, USA, 2013), SOCC '13, ACM, pp. 5:1–5:16.
- [93] WALKER, J. Streaming IN Hadoop: Yahoo! release Storm-YARN - Hortonworks. <http://hortonworks.com/blog/streaming-in-hadoop-yahoo-release-storm-yarn/>, 2013.
- [94] WIKI.APACHE.ORG. PoweredBy - Hadoop Wiki. <https://wiki.apache.org/hadoop/PoweredBy>, 2014.

- [95] XIN, R. S., GONZALEZ, J. E., FRANKLIN, M. J., AND STOICA, I. Graphx: A resilient distributed graph system on spark. In *First International Workshop on Graph Data Management Experiences and Systems* (2013), ACM, p. 2.
- [96] XIN, R. S., ROSEN, J., ZAHARIA, M., FRANKLIN, M. J., SHENKER, S., AND STOICA, I. Shark: Sql and rich analytics at scale. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of data* (2013), ACM, pp. 13–24.
- [97] YANG, F., MERLINO, G., AND LÉAUTÉ, X. The radstack: Open source lambda architecture for interactive analytics.
- [98] YANG, H.-C., DASDAN, A., HSIAO, R.-L., AND PARKER, D. S. Map-reduce-merge: simplified relational data processing on large clusters. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (2007), ACM, pp. 1029–1040.
- [99] YANG, W., LIU, X., ZHANG, L., AND YANG, L. T. Big data real-time processing based on storm. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on* (2013), IEEE, pp. 1784–1787.
- [100] ZAHARIA, M., CHOWDHURY, M., DAS, T., DAVE, A., MA, J., MCCAULEY, M., FRANKLIN, M., SHENKER, S., AND STOICA, I. Fast and interactive analytics over hadoop data with spark. USENIX.
- [101] ZAHARIA, M., CHOWDHURY, M., FRANKLIN, M. J., SHENKER, S., AND STOICA, I. Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing* (2010), pp. 10–10.
- [102] ZAHARIA, M., DAS, T., LI, H., HUNTER, T., SHENKER, S., AND STOICA, I. Discretized streams: A fault-tolerant model for scalable stream processing. Tech. rep., DTIC Document, 2012.
- [103] ZAHARIA, M., DAS, T., LI, H., HUNTER, T., SHENKER, S., AND STOICA, I. Discretized streams: Fault-tolerant streaming computation at scale. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles* (2013), ACM, pp. 423–438.