Procedia Computer Science

Volume 80, 2016, Pages 439–449

ICCS 2016. The International Conference on Computational Science

An Evaluation of Data Stream Processing Systems for Data Driven Applications

Jonathan Samosir, Maria Indrawan-Santiago and Pari Delir Haghighi
Faculty of IT, Monash University, Australia.

Abstract
Real-time data stream processing technologies play an important role in enabling time-critical decision making in many applications. This paper aims at evaluating the performance of platforms that are capable of processing streaming data. Candidate technologies include Storm, Samza, and Spark Streaming. To form the recommendation, a prototype pipeline is designed and implemented in each of the platforms using data collected from sensors used in monitoring heavy-haul railway systems. Through the testing and evaluation of each candidate platform, using both quantitative and qualitative metrics, the paper describes the findings, where Storm is found to be the most appropriate candidate.

Keywords: Big data, real-time data stream processing, Storm, Spark, Samza, Hadoop ecosystems

1 Introduction

Data intensive science, as the fourth paradigm of science [1], involves the process of capturing, curating, analysing and communicating data for scientific discovery. With the proliferation of sensor technologies as a tool to capture physical measurements, there is a need to support real-time decisions based on intermittent data availability, in addition to the ability to process large volumes of data. Current distributed technologies, such as Hadoop, have been used to address these use-cases, however they are often not appropriate for use-cases involving non-deterministically available data, or dynamic real-time data. There are currently numerous real-time data processing platforms that are in development and in production use, both in industry and academia. In this paper, we present our experience in evaluating and ultimately selecting a data stream processing platform for processing sensor data collected from the monitoring of heavy-haul railway systems.

To evaluate the candidate real-time data streaming platforms, and forming a recommendation for the most suitable solution, appropriate data stream processing system (DSPS) technologies need to be looked at, tested, and evaluated. We selected three data stream platforms for evaluation and implemented each of the platforms to test their suitability for our project's data processing needs. The platform was designed as a data processing "pipeline" that allows data to be streamed from the railway and processed, in real-time, according to the given processing requirements.

From describing our experience, we provide an insight on the capabilities of the chosen platforms. Each of the platforms is evaluated based on a set of quantitative and qualitative criteria. Both the

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.322

quantitative and qualitative performance report can be used to guide others when selecting DSPS technologies for scientific projects that match the use-case.

2 Real-time Data Processing of Big Data

"Traditional" methods of processing big data, involving MapReduce jobs to batch process data, are not completely suitable for real-time use-cases involving processing data with non-deterministic availability. Real-time data processing, enabled by DSPS technologies, allows data to be processed as soon as it is made available, without the need of a storage system, such as HDFS. A 2013 industry survey on European company use of big data technology shows that over 70% of responders show a need for real-time processing [2]. Since that time, there has certainly been a response from the open-source software community, through the rapid development of modern DSPS technologies.

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 [3]. Storm was originally developed by a team of engineers lead by Nathan Marz at BackType. Toshniwal et al. [4] describe Storm, in the context of its use at Twitter, as "a real-time distributed stream data processing engine" that "powers the real-time stream data management tasks that are crucial to provide Twitter services" [3, pg. 147]. Since the project's inception, Storm has seen mass adoption in industry, including amongst some of the biggest names, such as Twitter, Yahoo!, Alibaba, and Baidu [5].

Spark [6] is another popular big data distributed processing framework, offering of both real-time data processing and more traditional batch mode processing, running on top of Hadoop YARN [6]. Spark is notable for its novel approach to in-memory computation, through Spark's main data abstraction, the resilient distributed dataset (RDD). Spark Streaming [8], built on the Spark engine, affords the processing of streamed data.

Samza is a relatively new real-time big data processing framework originally developed in-house at LinkedIn, which has since been open-sourced at the Apache Software Foundation [9]. Samza offers much similar functionality to that of Storm. 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, which many core features are offloaded to. Samza processes streams of data through pre-defined jobs, which perform any specified operation on the data streams. The source of streams in Samza comes from Kafka, which can be used as an output destination for jobs for further pipeline-style processing.

3 Data Filtering Pipeline Design

The proposed data filtering pipeline is 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. Figure 1 shows an overview of the pipeline.

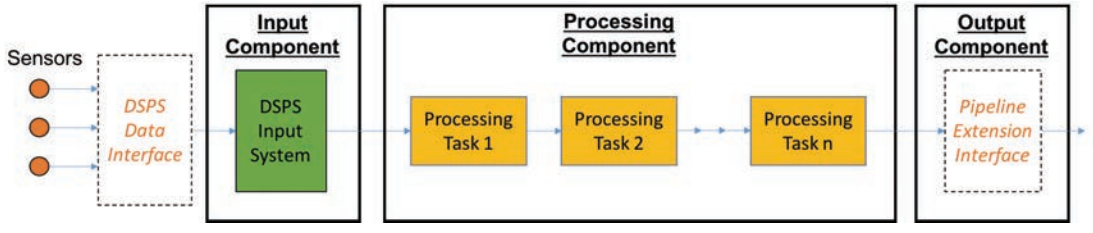

Figure 1: An overview of the data filtering pipeline components