Comparative Evaluation of Spark and Flink Stream Processing

February 19, 2017

Contents

Abstract

Recent years have witnessed a big development improvements of Big Data frameworks such as Apache Spark, Apache Flink. For example, both frameworks support real-time streaming processing. While which platform is the best still an open question. In this paper, we provide a performance comparison of streaming data processing in Spark and Flink, by measuring different performance metrics, namely latency and throughput. The experiment is done using multiple streaming workloads over real-world datasets.

1   Introduction

Big Data is now one of the hot topic in Information Technology, it has gained increasing attention of professionals in industry and academics, since the amount of data we generate is increasing. Moreover, recent years showed a rapid development of Big Data frameworks such Hadoop, Spark, Flink.

The Big Data is a method to process not just only just large datasets (volume), but also that arrives in high rate (velocity), and it comes in all kind of formats (variety)[1].

The purpose of this paper is to provide a comparative experimental evaluation of throughput, latency of stream processing in Spark and Flink. We use a dataset of airplanes trajectories provided in the context of Datacron project[1], in order to simulate real-world use cases. Using multiple streaming data workloads we try to cover the main API differences between the both frameworks.

2   Technical Background

This section presents the main characteristics, APIs, and main data abstraction of Spark and Flink.

3   Experiment Setup and Implementation

This section describes the experiment workloads and their implementation in each framework.

4   Performance Results

This section provides and compares the performance results of the different workloads.

---

[1] http://www.datacron\discretionary{-}{}{}project.eu/

5   Conclusion

This section gives the overall conclusion.

References

[1]   Laney, D 2001 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group.