LOG8415
Concepts avancés en infonuagique

Foutse Khomh
S. Amirhossein Abtahizadeh
Département Génie Informatique et Génie Logiciel
École Polytechnique de Montréal, Québec, Canada
foutse.khomh[at]polymtl.ca
a.abtahizadeh[at]polymtl.ca

Octobre 8, 2021

1 Identification

Student's name: Yanis Toubal

Date of the reading note: 2015

Author(s): Andrey Goder, Alexey Spiridonov, Yin Wang

Title of the article: Bistro: Scheduling Data-Parallel Jobs Against Live Production Systems

Publication: Goder, A. & Spiridonov, A. & Wang, Y.. (2015). Bistro: Scheduling data-parallel jobs against live production systems. 2015 USENIX Annual Technical Conference (USENIX ATC 15). 459-471.

2 Article

Keywords: Big Data, Batch Job, Online Workloads, Bistro, Batch Scheduler, Facebook

Concepts and definitions:

- Big Data: data from various sources that increase in volume at an ever-increasing velocity that can be analyzed to extract relevant informations. The data is very voluminous, is complex and can be found in various formats.
- Batch Job: Scheduled job in the form of a program that can be run without user interactions. They are frequently used for automating tasks.
- Batch Scheduler: System that control batch jobs and enables the automatic execution of batch jobs at a scheduled time.

- ad hoc: Generally speaking, it means conceiving an improvised solution to solve a particular problem. In programming particularly, it means writing quick code as a temporary solution to solve a problem or a set of problems.
- Live Workload: A set of tasks that must be completed before a specific deadline. Time is an important constraint with this kind of workload.
- Offline Workload: A set of tasks that uses a part of the data already present in the system. Usually used for complex analysis.

Summary: One important issue when dealing with Big Data is the fact that most large-scale data processing systems use copy of offline data as an input which usually causes a lot of issues since copying a large amount of data can be tedious. The solution of running the online data as batch jobs isn't suited with existing batch schedulers systems that are designed for offline data. Facebook which originally used Hadoop, faced many of these issues.

To overcome these problems, Facebook created many ad hoc (in-house) schedulers to tackle specific types of job which temporarily solved the issues but were hard to maintain and scale. Using the experience from these ad hoc applications, they acquired a deep understanding of the requirements of such a system and started to think about creating a new system to solve these problems more efficiently. The most crucial and surprising discovery they made was that batch jobs where mostly parallel(map-only).

From this newly acquired knowledge, Bistro came to light in Facebook. Bistro is a scheduler for batch jobs which as the main purpose of replacing Hadoop and ad hoc scheduler by letting batch jobs (offline data) share ressources with online workloads in an efficient manner. This system uses a tree-based scheduling algorithm which is more efficient then the old brute-force scheduling algorithm which is used as the baseline. The different modules of Bistro work in an asynchronous manner and communique through snapshots and queues.

So far, Bistro has been introduced to many production systems as the general purpose scheduler and its efficienty has been shown when compared to the old system used.

Research contributions: They described in great details Bistro that efficiently run batch jobs in a live workload system. They explained the reasons that led to Bistro being created namely the limitations of Hadoop, the temporary ad hoc schedulers and the need to manage efficiently live data. They also described the architecture in depth using a high level figure. They showed how they tested Bistro and provided interesting data they collected by comparing it's performance with the old system. They also showed many use cases of Bistro in the Facebook infrastructure.

3 Analysis

Quality:

General organization:	Language and style:	Technique:	Bibliography:
☐ Very good;	☐ Very good;	☐ Very good;	☐ Very good;
☒ Good;	☐ Good;	☐ Good;	☒ Good;
☐ Medium;	☐ Medium;	☒ Medium;	☐ Medium;
☐ Bad;	☒ Bad;	☐ Bad;	☐ Bad;
☐ Very bad.	☐ Very bad.	☐ Very bad;	☐ Very bad;
		☐ N/A.	

Forces of the message:

- The authors had in depth knowledge of the subject they were describing especially the Bistro system where they provided many detailed explanations about how it works and the role of each of his component.
- A lot of examples field examples from Facebook were used when talking about the theoretical aspects. It made it easier to relate the practical and the theoretical aspects because of that which increased the overrall comprehension of the text.
- The tables (Section 4) used were very relevant and presented the key informations in a concise and precise way. It also made good comparaisons between the systems.

Weaknesses of the message:

- The figure used in the text were hard to understand even when explained in the text. Most of them contained a lot of unnecessary informations and weren't well presented.
- The overrall language of the text made it very hard to digest especially for someone who doesn't have a background in big data processing systems. Most technical terms weren't defined properly and some phrases were hard to understand even after reading them multiple times.
- The position of the figures in the text didn't make much sense. They seemed out of place and most of them weren't referenced in the same page. They were also not referenced a lot in the text and they presentation in the text was very minimal.
- The authors didn't go too much details when talking about the related work of this system and the future outlooks in the section 5. They mostly repeated what they had previously explained and only talked about a few other technologies which were undirectly linked to Bistro.

Future directions: This article provides an in depth description of the Bistro system which is now widely used in Facebook to handle the offline aswell as the online data.

This way of handling the necessary data could pave the way to new big data technologies that are more performant and efficient. Also, since Bistro is open source, it can be used by anybody. Also, anybody can contribute to the project so future improvements on the efficienty of the system and more features are to be expected.

The system presented seems very promising, but since this article was written in 2015, it would be a good idea to write a follow up article and see what have changed and if some massive improvements happened since then. Also, it would be great to see if other systems were inspired from it and if so compare them with Bistro.

Other important articles:

- Job scheduling under the portable batch system.
  In Job scheduling strategies for parallel processing,
  pp. 279-294, 1995

- Mesos: A platform for fine-grained resource sharing in the data center
  In NSDI
  2011

- Theory and practice in parallel job scheduling.
  In Job scheduling strategies for parallel processing
  pp. 1-34, 1997

- Mapreduce: Simplified data processing on large clusters
  In OSDI
  2004