Big Data Platforms

Final Project

Stragglers and serverless computing

Submission due to 09.02.2022

Background: We learned that a single straggler task may badly affect an entire performance of the MapReduce job even job contains thousands of tasks. As example, a MapReduce job with 1000 tasks, while 990 tasks already completed, the completion of the job will be dominated by the remaining 10 tasks which may be slower than other tasks and even some might get stuck for much longer time. There are various root causes for straggler tasks to appear. For example, a worker node with CPU that has a faulty cache, may cause tasks to execute slowly comparing to task executing on other worker nodes. Another example, when tasks interacts with some external service (like database, URL location, etc.) then temporary network latency may cause certain tasks to behave slowly comparing to others. There are various approaches to deal with straggler tasks like backup tasks in Hadoop MapReduce and speculation mode in the Apache Spark.

Problem statement: As serverless computing evolve for variety of new use cases, so the problem of straggler serverless tasks became more critical to address. As example, when Lithops submit a job that contains 1000 invocations and a single execution got stuck, then entire job will stuck as well, until the remaining task is either finished or Lithops reached timeout on waiting. In both cases a straggler task has very negative effect both on the costs of a single job and performance.

Project definition: In this project students will explore effect of straggler tasks in the serverless computing (Function as a service). Students will find relevant background, explain the problem while emphasizing the difference how straggler tasks affects serverless vs cluster based Big Data engines, like Apache Spark. Students will explore how Lithops submit tasks and try to propose various solutions to deal with straggler tasks. A pseudo code should be provided to demonstrate your proposed solutions. Make sure to describe pros and cons of solution you provide. At least 6 different bibliography references should be provided (try to perform various searches like “straggler serverless pdf”, also in https://arxiv.org/corr)

Project scope and template for submission

Project will be submitted as a pdf paper accompanied with a prototype code
The following template should be used for the final report.

Template for the final report

Big Data Platforms

Stragglers and serverless computing

(Names)

(Submission date)

Abstract

Short description of the problem you are solving

Motivation and background

Describe the relevant background, putting emphasize what are struggler tasks, the root cause their appear, provide examples and explain how different Big Data engines deals straggler tasks.

Stragglers in serverless computing

Explain what is serverless computing and the problem of straggler tasks in serverless paradigm. Be specific and address various points on difference and similarity of the straggler tasks in serverless computing vs other cluster based Big Data engines.

Stragglers in jobs submitted by Lithops

Explore and explain how Lithops submit it's job and knows which tasks are finished and which tasks yet running. Explain how Lithops know when entire job is finished and how it waits for the running tasks to complete. Provide two different approaches to extend Lithops with mechanism to prevent jobs to be affected by straggler tasks. Address different aspects of complexity of implementation, costs, performs. Explain positive and negative effects of the approaches you propose.

Prototype

Provide pseudo-code for the approaches you propose. Explain how it works.

Next steps

Suggest next steps to the solutions you proposed

Conclusion

Short conclusion of the work you did

Bibliography

List of all sources you were using in the project. At least 6 different sources
Possible 2 out of 6 sources

1. https://www.cs.fsu.edu/~yuw/pubs/2015-NAS-Yu.pdf
2. https://mjeer.journals.ekb.eg/article_62728_c818f3f951476c6005647f9ba7364efd.pdf

Points on grading

1. Extra points will be given to students who present deep solutions that address broad scope
2. The proposed solutions should address various aspects we learned during the course, like consistency, availability, fault tolerance, etc.
3. All the text students write should be their original and avoid as possible to copy text from other papers. In cases, when text need to be copied from another papers, make sure to reference what text is copied from other source and what is the source. Grade will be affected, If text is copied without proper reference.
4. Make sure you describe strong and weak points of your proposed solutions. Grade will be affected if there are additional weak or strong points than you managed to describe.