Analyse d'une trace de données de Google

Big Data Frameworks - TP noté

ING3-HPDA

Année 2024–2025

1 Introduction

La trace d'utilisation d'un cluster de Google [1] est un ensemble de données collecté par Google sur l'un de leurs clusters pendant le mois de mai de 2011. Contenant une quantité considérable d'information (41 GBytes), elle a été mise à disposition de la communauté scientifique afin d'encourager la recherche sur l'analyse des données des grands centres de calcul.

2 Ressources

La trace est disponible sur le *github* de Google [2]. La documentation du format et des spécifications de la trace sont aussi en ligne [3]. De plus, il existe un forum en Internet [4] où vous pouvez trouver des réponses à plusieurs questions. Finalement, il existe une considérable quantité d'articles scientifiques sur l'analyse de la trace [5].

À cause de sa taille, la trace ne peut être analysée confortablement qu'en utilisant des infrastructures adaptées pour le Big Data.

3 Modalités

Ce projet est à effectuer par groupes de 4 ou 5, la composition des groupes est libre. Les choix algorithmiques et d'implémentation sont libres. Vous devez utiliser les outils d'analyse vu en cours (Pandas, Spark, etc.) mais, en général, vous avez la liberté d'utiliser en plus ce que vous considérez nécessaire.

4 Objectifs

Le but de ce projet est d'analyser la trace de Google et caractériser certains aspects de l'utilisation du cluster. Au moins, les questions suivantes doivent être traitées :

1. Identification des travaux et tâches (*jobs* et *tasks* dans la nomenclature de Google) les plus prenants en CPU, en valeurs maximales et en moyenne pendant la durée de la trace. Ces travaux seront appelés désormais *jobs dominants*.
2. Identification des jobs et tasks les plus prenants en mémoire vive, en valeurs maximales et en moyenne pendant la durée de la trace.

En plus, d'autres actions pourront être envisagées :

1. Classification des jobs dominants par classe de priorité.
2. Pour les jobs dominants, étude de la corrélation entre la consommation de CPU et de la mémoire vive.

5 Déroulement des séances

Vous devez vous assurer d'avoir acquis une bonne compréhension du problème et de la stratégie que vous allez adopter. Votre rapport devra contenir, dans une section de *Méthodologie*, les informations suivantes :

— Votre compréhension sur votre renseignement initial du format de la trace.

— L'élection du langage et des outils d'analyse.

— Le nettoyage initial des données que vous envisagez nécessaire.

— Le traitement que vous avez fait des données.

Si vous avez des questions, vous devez faire le point avec votre encadrant pour discuter et réorienter (s'il y en a besoin) votre stratégie.

6 Rendu du projet et notation

Un rapport en PDF ainsi qu'un ou plusieurs liens vers vos codes devront être fournis. Les codes pourront être fournis en format notebook, comme un développement sur la Google Cloud Platform, etc. Seront notés la cohérence de la réalisation, les techniques et méthodes d'implémentation et, bien entendu, la réponse apportée à l'objectif attendu.

Le rendu final en PDF et le code seront envoyés à votre encadrant au plus tard à 23h59 le dimanche **02/02/2025**. Il devra contenir un lien vers le GitLab ou Github où vous avez déposé votre code source et/ou notebook.

Références

[1] J. Wilkes, *More Google cluster data, 2011, (Google research blog).*, 2011. Posted at http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html.

[2] J. Wilkes and C. Reiss, *ClusterData2011_2 traces*, 2011. https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md.

[3] J. H. C. Reiss, J. Wilkes, *Google cluster-usage traces: format + schema, Technical Report, Google Inc., Mountain View, CA, USA, 2011 . Revised 2012.03.20.* Posted at https://drive.google.com/open?id=0B5g07T_gRDg9Z0lsSTEtTWtpOW8&authuser=0.

[4] *Google cluster data - discussions.* https://groups.google.com/forum/#!forum/googleclusterdata-discuss.

[5] *Google cluster traces bibliography.* https://github.com/google/cluster-data/blob/master/bibliography.bib.