Masteroppgaven min

Et resultat av to års arbeid

Roger Bløtekjær

Thesis submitted for the degree of
Master of science in Informatikk: språkteknologi
30 credits

Institutt for informatikk
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2018

Masteroppgaven min

Et resultat av to års arbeid

Roger Bløtekjær

Masteroppgaven min

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

0.1 Summary

0.2 Foreword

I got nothing. Put this on a different page though, maybe with a dramatic font or something.

Contents

Chapter 1

Introduction/Background ?

1.1 Target group

This thesis covers the deep technical aspects of big data analysis and genetic algorithms. All techniques used will be explained in detail, but it is advised to have a certain degree of technical insight before reading.

1.2 Area of research

Hadoop and MapReduce are already well established technologies employed in countless applications around the world. I propose a new method of implementing Hadoop clusters, with an out-of-the-box approach, meaning that this thesis purely covers the implementation aspect.

1.3 Personal motivation

The subject for this master thesis is a natural continuation of my previous work *Hadoop MapReduce Scheduling Paradigms*, published in 2017, in the 2nd IEEE International Conference on Cloud Computing and Big Data Analysis (ISSS-BDA 2017). Back then the topic was haphazardly picked from a list of eligible ones, but the more I read into it - the more I understood the incredible use cases for Hadoop within the massive industries that are driving forces for our technological advancements. It felt like an awakening when I realized the potential implications of future IoT, advanced data analysis and business intelligence.

1.4 Research method in brief

Throughout this thesis I will develop an entire suite of tools centered around a genetic algorithm for automatically optimizing a Hadoop configuration, given a representative generated data set. There will be a defined optimal result scoring based on the configuration parameters fed to the Hadoop cluster, that will be compared to my algorithms automatic configuration. Relevant scoring metrics are speed of algorithm compared to manual setup or other frameworks, and naturally the relevancy and completeness of data collected.

1.5 Most relevant previous findings

1.6 Why is this worthwhile

During my previous research the motivation for all the papers surveyed was mostly the same. **First and foremost, it was well acknowledged that Hadoop clusters show very suboptimal performance with out-of-the-box settings. Secondly, it was shown that finding optimal parameters for a Hadoop cluster is very time consuming and hard. As organizations often run with a lack of resources, time and domain-specific engineers, this is a field ripe for improvement.**

1.7 How far will this advance the field?

Hopefully this will provide a fully functional, open source light weight framework allowing companies to easily deploy Hadoop Clusters without worrying about tailoring the solution or suboptimal performance. If the task proves to be too big, this will lay the foundation for further work to make a de facto solution. Or something.

1.8 Structure of the report

Add comments about every chapter here.

Chapter 2

Related literature and theoretical focus

2.1 Gunther

Gunther[1] has done something very very similar. Need to speak to my supervisors to sort this out.

Chapter 3

Presentation of domain where technology is used

BMC Software, Inc states the following about Apache Hadoop:
Financial services companies use analytics to assess risk, build investment models, and create trading algorithms; Hadoop has been used to help build and run those applications. Retailers use it to help analyze structured and unstructured data to better understand and serve their customers. In the asset-intensive energy industry Hadoop-powered analytics are used for predictive maintenance, with input from Internet of Things (IoT) devices feeding data into big data programs. Telecommunications companies can adapt all the aforementioned use cases. For example, they can use Hadoop-powered analytics to execute predictive maintenance on their infrastructure. Big data analytics can also plan efficient network paths and recommend optimal locations for new cell towers or other network expansion. To support customer-facing operations telcos can analyze customer behavior and billing statements to inform new service offerings. Examples of Hadoop There are numerous public sector programs, ranging from anticipating and preventing disease outbreaks to crunching numbers to catch tax cheats. Hadoop is used in these and other big data programs because it is effective, scalable, and is well supported by large vendor and user communities. Hadoop is a de facto standard in big data.[2]

Chapter 4

Method

Chapter 5

Results

Chapter 6

Discussion

Chapter 7

Conclusion

Bibliography

[1] Liao, G., Datta, K., & Willke, T. L. (2013, August). *Gunther: Search-based auto-tuning of mapreduce.* In European Conference on Parallel Processing (pp. 406-419). Springer Berlin Heidelberg.

[2] BMC Software, Inc (2018 January) *http://www.bmc.com/guides/hadoop-examples.html*