IT UNIVERSITY OF COPENHAGEN

Clustering Player Behaviours in Data Streams with K-Means in Map-Reduce

by

Sigurdur Karl Magnusson

A thesis submitted in partial fulfillment for the degree of Master of Science in IT

in the

Computer Science

Software Development and Technology

Supervisors

Julian Togelius, Rasmus Pagh

March 2013

"Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius – and a lot of courage – to move in the opposite direction."

Albert Einstein

Abstract

The abstract is a short summary of the thesis. It announces in a brief and concise way the scientific goals, methods, and most important results. The chapter "conclusions" is not equivalent to the abstract! Nevertheless, the abstract may contain concluding remarks. The abstract should not be discursive. Hence, it cannot summarize all aspects of the thesis in very detail. Nothing should appear in an abstract that is not also covered in the body of the thesis itself. Hence, the abstract should be the last part of the thesis to be compiled by the author.

A good abstract has the following properties: *Comprehensive:* All major parts of the main text must also appear in the abstract. *Precise:* Results, interpretations, and opinions must not differ from the ones in the main text. Avoid even subtle shifts in emphasis. *Objective:* It may contain evaluative components, but it must not seem judgemental, even if the thesis topic raises controversial issues. *Concise:* It should only contain the most important results. It should not exceed 300–500 words or about one page. *Intelligible:* It should only contain widely-used terms. It should not contain equations and citations. Try to avoid symbols and acronyms (or at least explain them). *Informative:* The reader should be able to quickly evaluate, whether or not the thesis is relevant for his/her work.

An Example: The objective was to determine whether ... (*question/goal*). For this purpose, ... was ... (*methodology*). It was found that ... (*results*). The results demonstrate that ... (*answer*).

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

Abbreviations

LAH **L**ist **A**bbreviations **H**ere

Chapter 1

Chapter Title Here

1.1 Motivation

We live in a time of data, where data is growing at an amazing rate. Data is everywhere around us and companies are generating more and more data. Storing data is cheaper than ever and many smaller and mid-sized companies are buying storage subscriptions at other bigger companies that offer "Storage as a Service" (SaaS). A SaaS company offers e.g. data reliability and durability by storing critical data in multiple facilities and on multiple devices. Having all this data does not give a meaning by itself, it needs to be analysed and interpret to give a business value.

"Information is the oil of the 21st century, and analytics is the combustion engine."

Peter Sondergaard, senior VP at Gartner

The need to analyse all this data have caused many new companies to rise up that are specialised in analysing large quantities of data, extracting knowledge and converting it to a business value. Many of those companies offer "Analytics as a Service" (AaaS) that offer an analytical software to discover trends and unknown patterns in the data.

1.2 Problem Statement

GameAnalytics.com is a cloud hosted service for collecting, analyzing and reporting game metrics. Working with large quantities of game metric data that needs to be analysed and processed efficiently.

A valuable application for GA is to design and implement a scalable streaming version of a clustering algorithm. That can read a large data in mini-batches into memory and cluster the data incrementally.

The goal would be to incrementally find clusters efficiently in a streaming data, showing predominant characteristics of human behavior, e.g. the hardcore players, casual players, people who didn't understand the game, etc.

1.3 Contributions

- Clustering Player Behaviours using Map-Reduce framework
- Software and knowledge to GameAnalytics for further development

1.4 Project Outline

Chapter 2

Background Theory

2.1 Player Behaviour

2.1.1 Game Metric

2.1.2 Features

2.2 Clustering

2.2.1 K-Means

2.2.2 Streaming and Data Streams

2.3 Map-Reduce and Hadoop

Chapter 3

Related Work

In subsequent sections we will give a overview of related and recent work that we found. We start with looking into Clustering Player Behaviours and then we dive into numerous work related to K-Means when clustering a finite stream of data, endless data streams and finally different K-Means Map-Reduce implementations and results.

3.1 Clustering Player Behaviours

3.2 K-Means Clustering

K-Means is one of the most studied clustering algorithm out there and is still actively researched. It's a simple algorithm that partition the data into k partitions. From it appearance in YEAR CITE, people have been researching and finding different solutions to what K-Means is lacking. The problem with manually set number of partitions, initializing the centers for the k clusters and dealing with outliers in data.

Our work relates to both working with a stream of data of length n that doesn't fit in memory and the incrementally evolving characteristics of clustering endless data

streams. The Map-Reduce framework allows us to implement a scalable clustering algorithm that work on a different segments of the data in parallel.

3.2.1 Streaming

3.2.1.1 Map-Reduce

3.2.2 Data Streams

Chapter 4

Methodology

4.1 Game Events Data

4.1.1 Building Features

4.2 K-Means in Map-Reduce

4.2.1 Euclidean distance

4.2.2 Normalized Histograms

4.3 Experimental Set-Up

Chapter 5

Results and Discussion

5.1 Results

5.2 Discussion

Chapter 6

Conclusions

Appendix A

Appendix Title Here

Write your Appendix content here.