Final Report

_____________________

Compressive Sensing Based Network Monitoring For Large-scale Data Centre

Mingjie Shao
C00188468

_____________________

A document submitted in part fulfilment of the degree of

BSc. (Hons.) in Software Development

Supervisor: Dr. Lei Shi

Institute of Technology Carlow

April 18, 2018
Abstract

Most of technique company are using web services nowadays. However, a high availability and high reliability infrastructure is need to provide a consistent support. Therefore, cloud monitoring is presenting an important role in most industries who provide web services. This research project is going to find the possibility of implementing compressive sensing algorithm in cloud monitoring which is common used in image processing and signal processing academic sector.
# Contents

1. Introduction ......................................................... 3
2. Project Description .................................................. 4
3. Conformance to Specification and Design .......................... 6
4. Learning Outcomes ................................................... 7
   4.1 Technique Learnings .............................................. 7
   4.2 Mathematical Learnings ......................................... 9
   4.3 Personal Learnings .............................................. 9
5. Project Review ........................................................ 10
   5.1 What went right? .................................................. 10
   5.2 What went wrong? ............................................... 10
   5.3 Outstanding ..................................................... 10
6. Future Work .......................................................... 11
   6.1 Development Path ............................................... 11
   6.2 Research Path .................................................... 11
7. Acknowledgements .................................................... 12
Chapter 1: Introduction

In this final report, we will summary what we did on this research project during the past seven months.

Firstly, we will talk about the project background, what's the problem that we are going to solve and what we achieved finally. Secondly, we will compare specification and our design and look at conformance, because some of the research targets or focusing areas are changed during the research period. Also, we will illustrate the reason of changing our research target or focusing area. Then, we are going to use a separate chapter talk about what we learned in technique, mathematical and personal perspective. Also, we will review the entire research project in Chapter 5. We will talk about what went right and what went wrong.

Finally, we will discuss the future work in both development and research perspective, express my heartfelt gratitude to everyone who helped me.
Chapter 2: Project Description

The project is a metric backup mechanism for cloud services which allow us compress the unused metric backups under 10% sampling rate. In most of cloud industries, we have to log metric data points every minute, some companies may have one second resolution. So the storage of metrics will be 60 squared depends on the resolution.

However, the metric data points which have been generated a few year ago might be useless, but companies don’t want to delete it. The research project is basically looking for a mechanism of compress the old metric data points, store the sampled data points. Once we need it, we could rebuild the original metric by the sampled data points. The best data points we have achieved is 10%.

Although, the rebuild progress takes a while, but compare with the data storage and the good rebuild quality, we can see the scenarios of implementing this approach in metric backup.

Figure 2.1: Original Metric
During this project, we have test it locally based on a CPU Utilization dataset[1]. After getting good reconstruction result, we have tested it based on 1000 nodes and 8 nodes Kubernetes managed cloud environment. The final achievement is that we can run a CLI program to making backups, we can rebuild the original metric once it needed.
First of all, they are different, I have changed the idea. The specification or the initial idea was designing a straggler detector by using compressive sensing. As the resource of compressive sensing in cloud monitoring is pool, I have to write my own code from scratch. After the second iteration, I have done the algorithm implementation in Python. However, I realized that a real-time streaming straggler detection might be impossible, because the metrics rebuilding took a while, longer than what I expected. Also, the compressed middle data is matrix instead of metric data points, they don't have linear relationship. I have to give up with this idea and think about something else.

One day after the second presentation, I was browsing products on Amazon Web Services. And I found a service called Amazon Glacier which is a secure, durable, and extremely low-cost cloud storage service for data archiving and long-term backup. The thing I achieved after the second iteration is that I can compress the metric under 10% sampling rate and rebuild the original metrics with good quality. So the second idea is compressing old metrics which industries will not use frequently. Once anyone need it, we can rebuild the metric by our compressed metric. It can save 90% storage for old metrics.

The reason of changing idea I think is that the resource of this area is pool. Before making the working algorithm programming, I wouldn’t know the difficulty of rebuilding metrics data points. The good things is that after research on the features I have got, I found the scenarios where we can apply this algorithm.
Chapter 4: Learning Outcomes

4.1 Technique Learnings

4.1.1 Python

During the second and third iteration of this research project, I have used Python as my main language. As I have learned Python fundamentals on third year Paul Barry’s class, the knowledge of Python I have learned new is command line based scripts. For example, the argparse built-in module in Python, it’s a parser for command-line options, arguments and sub-commands. So that I could take the start time, end time and the metric name which customer what to get compressed.

4.1.2 Kubernetes

Kubernetes is an open-source container management platform designed by Google. I have gained a little bit experience last summer during my internship at Amazon. So in this research project, the cluster initialization part is pretty straight forward as I knew the basic concepts of Node, Pod, Deployments in Kubernetes. The knowledge I have learned new in this project would be cAdvisor, Heapster and Grafana, because we are mainly focusing on monitoring topic in this project.

Figure 4.1: Kubernetes on AWS
4.1.3 Amazon Web Services

Amazon Web Services is the biggest cloud service provider, it has bigger market share compared with Google Cloud and Microsoft Azure. Amazon Web Services offers a broad set of global cloud-based products including compute, storage, databases, analytics, networking, mobile, developer tools, management tools, IoT, security and enterprise applications. These services help organizations move faster, lower IT costs, and scale.

The services involved in this research project include: Amazon EC2, Amazon CloudWatch, Amazon Simple Storage Service(S3), Amazon EC2 Auto Scaling, Amazon VPC. I have used Amazon EC2 before, but they were on-demand instances and reserved instances. During this research project, in order to reduce the cost of EC2 instances, I have used spot instances which can give us 70% off based on the lowest price setting. In addition, I have used Amazon CloudWatch Python SDK in my python script by importing AWS Boto3 module. I have heavily used Get-Metric-Statistics method during the third iteration.

![Kubernetes Cluster](image)

Figure 4.2: Kubernetes Cluster consists of 8 EC2 t2.micro instances and 1 EC2 m4.xlarge instance

4.1.4 Kubectl & Kops

Kubectl and Kops are also my best friend during operating with Kubernetes. Kubectl is Kubernetes built-in command line interface for running commands against Kubernetes clusters. However, the functionality of this CLI is not enough especially when I was trying to create a cluster. Kops is also a command line interface for Kubernetes. It provides an easier way to get a production grade Kubernetes cluster up and running. It supports Amazon Web Services, so I just need to specify a S3 bucket location for it to store all configuration files and status.
4.2 Mathematical Learnings

4.2.1 Compressive Sensing

At the beginning of this research project, I have no idea what compressive sensing algorithm is. By search on Google, I could briefly understand what it is, and it’s commonly used in image compression and signal sampling. Also, by searching compressive sensing on Google scholar, I realized that it became a very popular algorithm that lots of people are researching on this topic. This algorithm has been implemented into lots of academic sector and researchers are still looking for the possibility of implementing this algorithm in more areas e.g cloud computing, wireless signals, 5th generation mobile networks and etc. However, the papers in cloud monitoring is pool. It brings difficulties to me during the research. What I did was basically read books: A Mathematical Introduction to Compressive Sensing[2] and start from scratch.

4.2.2 Gaussian Distribution

Random matrix design is very important in compressive sensing, a good random matrix can give us good metric reconstruction. During this research project, I have gained some statistics knowledge and linear algebra knowledge.

4.3 Personal Learnings

The reason of choosing this research project is I love programming, I have confidence of making softwares. However, Do I like research? I have no idea at the beginning. The main reason of choosing research is that I want to have research experience for making decision of studying taught master or research master. One of my friend has finished his first year of a PhD degree and he gave up one year later. The reason is he doesn’t like research, he prefer programming. Personally, I want to give it a try and gain the research experience to see do I like it.

The outcome is good when I look backward. I have learned how to read a paper and how to start researching on a topic: get to know the background of your research area, understand the problem you gonna solve, read papers to know what other researchers did on this academic sector and what you are going to do to make contributions in this academic sector.

As I have changed my initial target after the second iteration, I learned how to support yourself under the heavy pressure. There were two months left only and I have to give up on my initial idea. Fortunately, by research on the result I have got on the second iteration, I found we can used it for compress metric backups.
5.1 What went right?

Firstly, the compressive sensing algorithm does have the ability to get a sample of metrics and rebuild the original metrics by using the sample. The best sampling rate I have achieved is 10%, which means a 1TB metric can be store as 100MB backup.

The second right thing I did should be moving my target from building a streaming fast straggler detection to compressed metric backup mechanism.

5.2 What went wrong?

At the beginning, I was thinking to build a straggler detection mechanism for cloud monitoring in Kubernetes by using compressive sensing algorithm. However, I found the metrics reconstruction took a pretty long time. So it might be impossible to achieve streaming straggler detection.

5.3 Outstanding

The metric backup is not fully automotive at the moment. What I want to achieve is that having a Kubernetes add-on, so that every operations engineer can easily edit the configuration file of the add-on and deploy it on Kubernetes. All the metric backups compression jobs will be executed automatically depends on their configurations.
Chapter 6: Future Work

There are two paths that we would expect to carry on. One path is development which means more programming. The other path is research, as the random measurement matrix can be self designed which probably give use a better matrix reconstruction correction rate.

6.1 Development Path

At the end of this research project, we can ssh to the master node of Kubernetes, get metrics, sampling and put it back to the storage as a compressed backup. However, a fully automated compressed backup mechanism would be a big advantage for Kubernetes. We will expect developing a Kubernetes add-on which can achieve automated backup compression by using compressive sensing algorithm.

6.2 Research Path

For the research perspective, we have used Gaussian random matrix for the metrics sampling. However, the random matrix design is very important, because it will affect the accuracy of metrics reconstruction. So a self designed based on cloud monitoring metrics behavior will give a better accuracy.
Chapter 7: Acknowledgements

I would like to say thank you to my supervisor Dr. Lei Shi who have instructed and favored me in the course of writing this paper. Also, I would like to express my heartfelt gratitude to my housemates who give me support, encourage me and give me a quite environment during writing this paper.
Bibliography
