School of Computing

UNIVERSITY OF LEEDS

COMP5200M Scoping and Planning Document

Student Name: **Xuyang Cao**
Programme of Study: **Advanced Computer Science (Data Analytics) MSc**
Provisional Title of Project: **Big Data Analytics for Cloud Computing Datacenters**
Name of External Company (if any)**:** Edgetic Ltd
Supervisor Name: Jie Xu
Type of Project: Exploratory software
NOTE to student*: ensure you have discussed the content with the supervisor well in advance of the deadline for submission. Submit an* ***electronic version*** *of this report in pdf via the appropriate link in Minerva; with filename of the format <surname><year>-SP (e.g. SMITH15-SP.pdf).*

Signature of Student: **Xuyang Cao**	**Date:** 10 May 2019

Assessor *(leave blank)***:** ***NOTE to assessor:*** *feedback form is available to download from Minerva in Resources → 'Quick links'. On completion, please upload the feedback form to Minerva.*

Contents

1. Background Research for the project

1.1 Context

Emerging technologies like cloud computing allow data to be captured from interactive devices in large various formats. The volume of data is growing at outstanding speed. As a matter of fact, the size of data plays a very crucial role. The executive chairman of Google Eric Schmidt says, “From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days ... and the pace is accelerating.” Facebook which still have 2 billion active users, nowadays is the largest social media platform. There are some more intriguing Facebook statistics: 1.5 billion people are active on Facebook daily, Europe has more than 307 million people on Facebook, every minute there are 510,000 comments posted and 293,000 statuses updated, etc (Marr, 2019).

Big technology companies like Google, Yahoo and Facebook run a large scale of computing applications. YouTube receives over 1 billion unique users each month and 100 hours of video are uploaded to YouTube every minute (YouTube.com, 2019). With the explosive amount of data, storage capacity and data processing have become issues. As one of Big Data characteristics, Velocity refers to the speed of incoming data and time it takes to process. Like mentioned above, massive data is produced by large amount of applications, the data need to be processed in different ways to meet the different demands, for instance, batch processing and real-time processing etc. Facebook need to process approximately 510,000 comments posted and 293,000 statuses updated (Marr, 2019).

1.2 Problem statement

Recent years, the volumes of data have been growing up. Many challenges are emerging.

- Handle increasing volumes of data

  In terms of data storage, traditional data systems, such as relational databases and data warehouses have been used for years to store and analyse data. Due to the high speed of data generated and the variety of data, such as structured, semi-structured and unstructured data, relational databases which is designed to work with structured data is hard to deal with unstructured data. Besides, relational databases which has schema-on-write characteristics that requires data be validated against a

schema before it can be written to disk (Trujillo et al., 2019) cannot handle the high speed of incoming data.

- Scalability

  Limits on scalability is also an issue for processing and storing data. Software all has limited scalability and resource usage limitations. Single server cannot handle large amount of data and streaming data in a short time. Scalable applications could meet the high demands of storage of rapid growth in data volume, meanwhile could also decrease resources during off-peak time. On-demand applications is important for processing unpredictable growth of data and reducing usage of resources to save much energy.

- High availability

  High availability is also an important principle for designing big data systems. If the server is running in standalone mode, the server goes down and the service will be unable to handle incoming data which will fail to guarantee the integrity of data. Meanwhile, hard disks are not always trustable. Because if the hard disks are broken or not available, it will lead to failed storage. If the data is not backed up, it will be lost permanently. It is not acceptable especially for bank systems.

1.3 Possible solution

Data centres are the infrastructure which could be an effective solution to the issues above. Data centres which are complex systems-of-systems are built to underpin modern distributed service-oriented systems. IoT devices, social media, and streaming video services, these applications are typically hosted in data centres which could offer high available, scalable and secure services (Townend et al., 2019).

1.4 Issues About the Possible Solution

Although data centre is an effective solution to deal with challenges of the increasing data volumes, there are some issues about it. Data centres are digital factories that process electronical power into digital services and generate large quantity of waste heat which needs additional power (Townend et al., 2019). Therefore, how to make use of these resources efficiently is a key factor.

1.5 How to demonstrate the quality of the solution

This paper basically focuses on software-based solutions for improving efficiency in data centres and evaluate a scheduling system taking into account software models. The criteria to evaluate the quality of the solution is judge the results of the evaluation and apply the results to the real scheduling systems to improve the efficiency.

2. Scope for this project

2.1 Aim

The aim of the project is the overall top-level goal. It might be helpful to consider this in conjunction with the project title.

The aim of the project to find the patterns of the resources allocated, the relationships between of the usage of the resources and the time of running jobs, the behaviours of the node in the scheduling system etc. Besides, modelling the relationships between the resources of nodes consuming and the types of tasks submitted, such as batch processing and long-running application etc. Finally, analysed the gap between the real resources consumed and the actual resources requested in different size of data.

2.2 Objectives

- Research solutions to solve problems in data storage and processing.
- Find patterns of data, e.g. resource utilization, behaviours of node-level, time series, based on the timeline during analysis of amounts of logs.
- Locate the problems, e.g. straggler, failure, in a scheduling system.
- Improve efficiency in scheduling virtual resources.

2.3 Deliverables

- A thesis focusing on doing research on how to improve the efficiency of scheduling system in data centres based on software solutions.
- A model of relationships between the resources of nodes consuming and the types of tasks processed.

- The results of analysing the patterns of the allocations and the usage of the resources changing with the time of running jobs in a scheduling system.

3. Project schedule

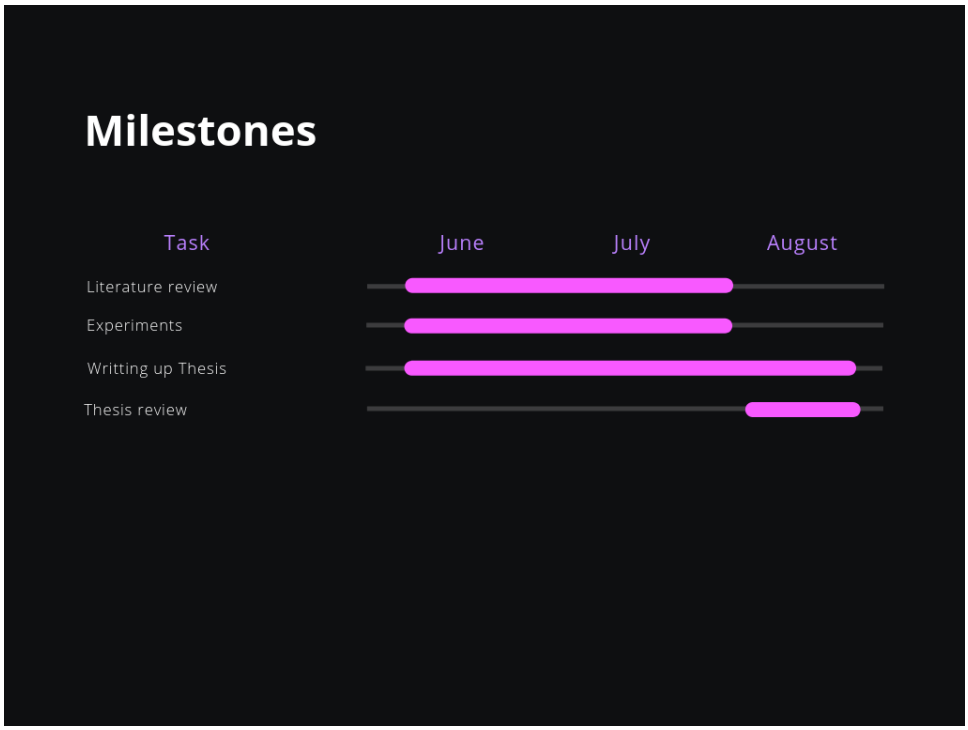
3.1 Methodology

Monday.com is a good tool based on browser for splitting up the tasks, estimating the hours of finishing the tasks and addressing the due date.

XMind is a useful tool for drawing diagrams of the architecture of the methods.

Toggle.com is a useful tool for recording the time consuming in multiple sub tasks.

3.2 Tasks, milestones and timeline



References

Marr. (2019). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. [online] Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#e93f38b60ba9 [Accessed 15 Jun. 2019].

Youtube.com. (2019). Press - YouTube. [online] Available at: https://www.youtube.com/yt/about/press/ [Accessed 15 Jun. 2019].

Trujillo, G., Garcia, R., Jones, S., Kim, C. and Murray, J. (2019). Traditional Data Systems | Understanding the Big Data World | Pearson IT Certification. [online] Pearsonitcertification.com. Available at: http://www.pearsonitcertification.com/articles/article.aspx?p=2427073&seqNum=2 [Accessed 15 Jun. 2019].

Townend P., Clement S., Burdett D., Yang R., Shaw J., Slater B., Xu J. (2019). *Improving Data Center Efficiency Through Holistic Scheduling In Kubernetes.* Leeds: University of Leeds.

Appendix A. How ethical issues are addressed

This paper focuses on improving the efficiency of a scheduling system in a data centre. All data acquired is approved in advance. This does not involve any ethical issues.