# School of Computing

**UNIVERSITY OF LEEDS**

## COMP5200M Project Specification

***NOTE to student***: *ensure you have discussed the content with the supervisor. Submit only an **electronic version** of this form in pdf via the COMP5200M module in the VLE; with filename of the format <surname><year>-Spec ( e.g. SMITH17-Spec.pdf).*

| |
|---|
| **Student Name: Xuyang Cao** |
| **Programme of Study: Advanced Computer Science (Data Analytics) MSc** |
| **Supervisor Name: Jie Xu** |
| **Name of External Company** (if any)**: Edgetic Ltd** |
| **Type of Project: Exploratory software** |
| **Provisional Title of Project:**<br><br>Big Data Analytics for Cloud Computing Datacenters |
| **Aim of Project:**<br><br>**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.** |

**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.

**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.