CITS3402 High Performance Computing Assignment 2

An essay on MapReduce,Hadoop and Spark

The total marks for this assignment is 15, the assignment can be done in groups of two, or individually.

The materials: The primary materials for this assignment are two classic papers:

1. J. Dean and S. Ghemawat, MapReduce: Simplified data processing on large clusters
2. M. Zaharia, M. Chowdhury, M.J.Franklin, S. Shenker and I. Stoica, Spark: Cluster computing with working sets

You can also use any additional materials through your own research. Here are some suggestions:

- The lecture slides posted in the lecture page of this unit.
- A set of slides on MapReduce and Hadoop by M. Zaharia.
- A set of slides on Spark from databricks

Tasks: The tasks are to summarize and explain in your own words certain parts of these two papers as explained below:

- You have to summarize sections 1-3 of the paper by Dean and Ghemawat. The emphasis would be to explain the MapReduce paradigm clearly; explain the examples given in section 2.3 clearly with further clarifications; explain the implementation of the MapReduce paradigm.
- The important parts of the paper by Zaharia et al. are: sections 1-2, section 3.1, Section 4 (the last part on "Interpreter Integration" is excluded), and section 6. The emphasis would be to explain the limitations of hadoop and how Spark overcomes those limitations; the Spark programming model; resilient distributed dataset; the example in section 3.1; implementation of Spark; and the comparison of Spark with distributed shared memory models and Hadoop.

Marking guideline: The marks for the essay will be based on coverage of the various topics (8 marks), quality of technical presentation (4 marks), and correctness and appropriate usage of English (3 marks). As such there is no length limit, so long as the presentation is clear and of good quality.

Deadline: The submission deadline is 11:59 pm on October 10, through cssubmit. This project carries 15% of the total marks for this unit.

Amitava Datta
September 2018

School of Computer Science & Software Engineering
The University of Western Australia
Crawley, Western Australia, 6009.
Phone: +61 8 9380 2716 - Fax: +61 8 9380 1089.
CRICOS provider code 00126G