Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Feb 12th. 2019

Please note that assignments in this class may be submitted to

www.turnitin.com, a web-based anti-plagiarism system, for an evaluation of their originality.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Reading review**

**Apache Spark: A Unified Engine for Big Data Processing**

The growth in the amount of data in industry and research has brought enormous opportunities and enormous computing challenges.

The universality of Spark has 3 important benefits:

1. Applications are easier to develop because they use a unified API
2. Combined processing tasks are more effective
3. Support for new applications that were not possible with previous systems.

As parallel data processing becomes commonplace, the combinable lines of processing functions will be one of the most important concerns for usability and performance.

This article introduces Spark’s key model: RDD. And introduces its development and advantages. As the key programming model in Spark, RDD is a cluster of fault-tolerant sets of partitioned objects that can be operated in parallel. Data sharing is the main difference between Spark and previous computing models. Data sharing provides significant acceleration for interactive queries and iterative algorithms. In addition, RDD has the advantage of being able to automatically recover from a fault.

Much like Google’s MapReduce, Spark is designed to work with multiple external systems for persistent storage. With RDD, we can build a variety of higher-level libraries on Spark for the use of many specialized computing engines.

Apache Spark is used in a variety of programs and involves many areas. The main applications are:

1. Batch processing
2. Interactive query
3. Stream processing
4. Scientific application

Finally, the article compares the Spark and MapReduce models. It also depicts the 4 main tasks that the fast-growing Apache Spark is doing:

1. Data Frames and more declarative APIs
2. Performance optimization
3. R language support
4. Research library

In this article, the author describes the importance of scalable data processing for next-generation computer applications and the benefits of Spark, as well as the hope for the future of Spark.