An Improvement of a Checkpoint-based Distributed Testing Technique on a Big Data Environment

Bhuridech Sudsee, Chanwit Kaewkasi Kaewkasi

Suranaree University of Technology, Nakhon Ratchasrima, Thailand, 30000

m5741861@g.sut.ac.th, chanwit@sut.ac.th

***Abstract*— The advancement of storage technologies and the fast-growing number of generated data have made the world moved into the Big Data era. In this past, we had many data mining tools but they are inadequate to process Data-Intensive Scalable Computing workloads. The Apache Spark framework is a popular tool designed for Big Data processing. It leverages in-memory processing techniques that make Spark up to 100 times faster than Hadoop. Testing this kind of Big Data program is time consuming. Unfortunately, developers lack a proper testing framework, which cloud help assure quality of their data-intensive processing programs while saving development time and storage usages.**

We propose Distributed Test Checkpointing (DTC) for Apache Spark. DTC applies unit testing to the Big Data software development life cycle and reduce time spent for each testing loop with checkpoint. By using checkpoint technique, DTC keeps quality of Big Data processing software while keeps an inexpensive testing cost by overriding original Spark mechanism so that developers no pain to learn how to use DTC. Moreover, DTC has no addition abstraction layers. Developers can upgrade to a new version of Spark seamlessly. From the experimental results, we found that in the subsequence rounds of unit testing, DTC dramatically speed the testing time up to 450-500% faster. In case of storage, DTC can cut unnecessary data off and make the storage 19.7 times saver than the original checkpoint of Spark. DTC can be used either in case of JVM termination or testing with random values.

***Keyword*—Distributed Checkpointing; Apache Spark; Big Data Testing; Software Testing;**

Bhuridech Sudsee received B.Eng. in Computer Engineering from Suranaree University of Technology and B.Sc. in Information Technology from Sukhothai Thammathirat Open University, both in Thailand. Currently, he is studying a Master degree in Computer Engineering. His fields of research interests are high-performance computing, distributed computing, data storage, Big Data processing and MapReduce frameworks.

Chanwit Kaewkasi received his PhD in Computer Science from the University of Manchester, United Kingdom in 2010. He is currently an Assistant Professor at School of Computer Engineering, Suranaree University of Technology, Thailand. Dr. Kaewkasi is actively researching in the areas of Low-Power Clusters, Cloud Computing, Big Data and Software Container Technologies.