BigData Analytics Technologies

Coursework

Module Learning Outcomes Assessed:

1. Knowledge of big data technologies and principles
2. Apply knowledge on MapReduce, Spark, and NoSQL data stores to provide scalable solutions for important bigdata problems
3. Knowledge on monitoring and improving performance of bigdata processing applications
4. Build up skills on learning new technologies and trends on bigdata processing

Tasks:

The course work comprises four tasks which are shown below. The course work should be implemented and tested in a Linux Ubuntu environment. Detailed report needs to be prepared with the steps followed and the results obtained for tasks 1 - 3. Task 4 is a self study which requires detailed study and summarization of key findings. Please submit a single report with your answers for the four tasks.

Task 1 : Implement Graph Triangle Counting

Use three open source data processing systems Apache Hadoop, Apache Spark, and JasmineGraph to store graph data, and count the number of triangles available on each of the following data sets.

Name	URL
Epinions social network	https://snap.stanford.edu/data/soc-Epinions1.html
Youtube Social Network	https://snap.stanford.edu/data/com-Youtube.html
Google Web Graph	https://snap.stanford.edu/data/web-Google.html

Run the three triangle counting setups on a single system. Report the number of triangles found from each of the three setups on the three data sets. Report the performance behavior of your setups in terms of elapsed time as well as any other relevant performance metrics such as memory, CPU, disk usage, etc. Compare and contrast the performance results obtained from the three setups. Next, scale your triangle counting setup for a larger graph data set LiveJournal social network. Report the elapsed time, and other performance metrics used with the above three graph data sets.

https://snap.stanford.edu/data/soc-LiveJournal1.html

Task 2 : Log Analysis using ELK Stack

Setup ELK stack locally on your system. Point to a log source such as Apache web server and visualize the logs on a Kibana dashboard. Configure an email alert where log entries exceed a certain threshold.

Task 3 : Incremental Checkpointing of Data Stream Processing Application

Setup Apache Flink stream processor on a single system. Setup NEXMARK stream processing benchmark on this Flink system. Run the NEXMARK benchmark and report the performance metrics observed.
Enable incremental checkpointing of the Apache Flink setup and use RocksDB as the state backend. Configure RocksDB state backend to conduct checkpointing with best performance figures for your system. Explain the rationale behind your choices.

https://github.com/nexmark/nexmark

Task 4 : Self Study

Investigate the latest research and trends on information security aspects of bigdata analytics. In particular explain approaches for privacy preserving bigdata analytics. Prepare a summary report based on the results found refering the latest research papers published in top international conferences.