Exercise set 2

Hou, Jue

014695647

* HDFS: It stands for Hadoop Distributed File System. It shares some similar properties as other distributed file system, but it also has its own features. The system is built on commodity hardware and is assumed to experience hardware failure often. It supports streaming data access instead of interactive access. Also, HDFS is adjusted to support the applications that have an enormous amount of data.
* Spark Tachyon: It is a project implemented inside of the Spark ecosystem. Essentially, Tachyon is a distributed in-memory file system, which is placed between distributed computation system and file system. It split the memory storage function from Spark and make Spark can focus on computation. Tachyon reduces the stress of Spark on memory storage and enhances the speed of massive amount data reading and writing for Spark.
* Remote direct memory access (RDMA): RDMA is a technique that allows data exchange between application memory directly without getting any CPU of computers, cache and OS involved. In this way, it can get data access just like local memory. Since it will get less resource involved, it improves I/O throughput and performance. When requested, the data will be delivered to the network, reducing latency and enabling faster data transfer.
* Spark DataFrames: It is a dataset which is organized into named columns. The dataframe is conceptually equivalent to relational database and is operated in a similar way. But a dataframe has more optimization in its implementation and get full use of RDBM optimization techniques. Dataframe can support better analytics algorithm. In addition, it can be constructed from different sources such as structured data and files.