CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 6: NoSQL and Spark

1. Short Q&A on NoSQL.

a) MapReduce and the Google File System (GFS) were designed to work well together. What important optimization in MapReduce is enabled by having GFS expose block replica locations via an API?

b) List two features that are originally designed for relational databases and are now integrated into the MapReduce/Hadoop software stack.

c) What are the advantages of adding schema to MapReduce/Hadoop?

d) Why we need YARN since we already have reasonably efficient cluster management systems in MapReduce/Hadoop?

2. Short Q&A on Spark

a) List three of the many development features in Spark that show the trend of convergence between relational databases and Spark.

b) In HDFS, each chunk is replicated for three times by default. In contrast, in Spark, RDD uses lineage for reliability. What are the major problems if Spark also uses replications for reliability?

c) Is it true that in the Spark runtime, RDD cannot reside in the hard disk?

3. NoSQL databases have been a hot research topic. Reading the below paper will help you to have an overview on the NoSQL databases.

Rick Cattell. 2011. Scalable SQL and NoSQL data stores. SIGMOD Rec. 39, 4 (May 2011), 12-27. http://www.cattell.net/datastores/Datastores.pdf

After reading the paper, answer the following questions:

a) Compare ACID and BASE. Why do NoSQL systems choose BASE?

b) If we plan to relational database techniques for supporting the following system, what are the problems? Or, why do we need specialized engines for each workload?

1) key-value stores,

2) document stores.

3) Case studies. Would you please give some examples/case studies in practice on using the NoSQL databases mentioned in the paper, rather than relational databases?