CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 2: NoSQL and Spark

**1. MapReduce/Hadoop**

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.

**2. Spark**

a) What are the advantages of using schema in Spark?

b) List three of the many common development features or considerations between relational databases and Spark.

c) In HDFS, each chunk is replicated for three times by default. In contrast, in Spark, RDD uses lineage for reliability. What is a major problem if Spark also uses replications for reliability?

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

**3. NoSQL**

NoSQL databases have been a hot research topic.

The following questions relate to the trade-offs between relational and NoSQL systems. A more detailed discussion can be found in this paper (not required reading for the class, but still a useful summary if you are interested):

*Rick Cattell. 2011. Scalable SQL and NoSQL data stores. SIGMOD Rec. 39, 4 (May 2011), 12-27.*

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

b) Why do we need specialized engines (e.g. document stores) in NoSQL systems, as compared to relational databases?

c) What is a practical reason to prefer horizonal scalability over vertical scalability?

d) In the paper, they have shared suitable applications for key-value stores and document stores:

Application of key-value store: Application of document store:

Text

Description automatically generatedText

Description automatically generated

Discuss some factors that make these applications suitable for key-value stores and document stores respectively.