CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 2: NoSQL and Spark

**1. NoSQL**

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?

Answer:

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

Answer:

c) 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.

Answer:

**2. Spark**

a) Why Spark is more suitable for iterative processing compared to Hadoop?

Answer:

b) In the below spark code block, please indicate which lines are transformation and which lines are action. For transformation, please also indicate whether it is a narrow transformation or wide transformation.

Text, letter

Description automatically generated

Answer:

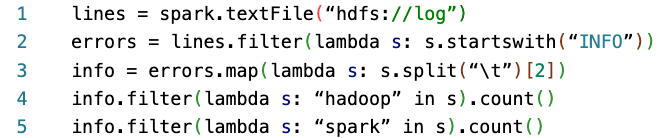
c) In HDFS, each chunk is replicated for three times by default. In contrast, in Spark, RDD uses lineage for reliability. What is a major problem if Spark also uses replications for reliability?

Answer:

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

Answer:

e) Explain how the following program can be sped up.



Answer: