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:

BASE = Basically Available, Soft state, Eventually consistent

ACID = Atomicity, Consistency, Isolation, and Durability

The idea is that by giving up ACID constraints, one can improve performance.

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

Answer: Less expensive, using commodity servers.

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:

Key-value store:

* Improves scalability and efficiency – writing or reading user pages is faster.
* No need for complex queries or based on the content of user pages – just reads and writes.
* May be acceptable for user pages to be slightly stale – thus eventual consistency is acceptable

Document store:

* Flexible schema may be beneficial (e.g. special types of vehicles may require different sets of fields)
* Unlike key-value stores, document stores allow for queries based on fields of a document

**2. Spark**

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

Answer: Spark stores most of its intermediate results in memory, making it much faster, especially for iterative processing.

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:

* Line 1: narrow transformation
* Line 2: narrow transformation
* Line 3: wide transformation
* Line 4: action

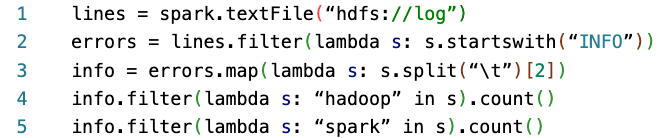
c) In HDFS, each chunk is replicated for three times by default. In contrast, in Spark, RDD uses lineage for reliability. What is a major problem if Spark also uses replications for reliability?

Answer: Consumes a lot of memory; memory is much more scarce than disk space

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

Answer: False. RDD can also be in the disk if out of memory.

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



Answer: First, let’s understand what the program does:

Line 1: Reads a file from HDFS

Line 2: Filter to extract lines start with INFO

Line 3: Split the string by tab and extract the 2nd component.

Line 4: Count the number of lines with “hadoop”

Line 5: Count the number of lines with “spark”

How to speed-up: we should add info.cache() (or info.persist()) before line 4, to cache the RDD in memory (or hard disk) so it doesn’t have to be re-computed in line 5.