CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 3: 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:

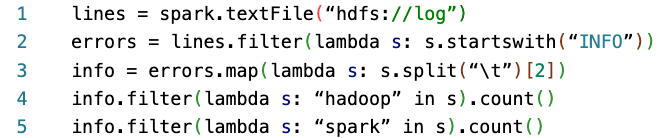
c) In HDFS, each chunk is replicated for three times by default. In contrast, in Spark, RDD uses lineage for reliability. What is a major problem if Spark also uses replications for reliability?

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: