CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 3: 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.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

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

图示

描述已自动生成

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

图形用户界面, 文本, 应用程序

描述已自动生成

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

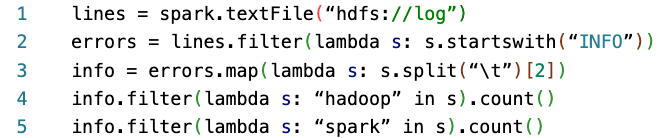
Line 1: Reads a file from HDFS

Line 2: Filter to extract lines starting with “INFO”

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

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

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



Answer:

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.