雲端運算與大數據應用
第九單元
pySpark

國立高雄第一科技大學

資訊管理系

黃文楨

學習目標

- 瞭解Spark的主要特性
- 了解何謂Hadoop
- 了解何謂map/reduce
- 了解Spark和 hadoop的主要差異
- 了解如何實作pySpark的word count範例

主題

- 學習目標
- Spark的主要特性
- Hadoop
- map/reduce
- Spark和 hadoop的主要差異
- 實作pySpark的word count範例
- 相關的學習資源
- 參考資料

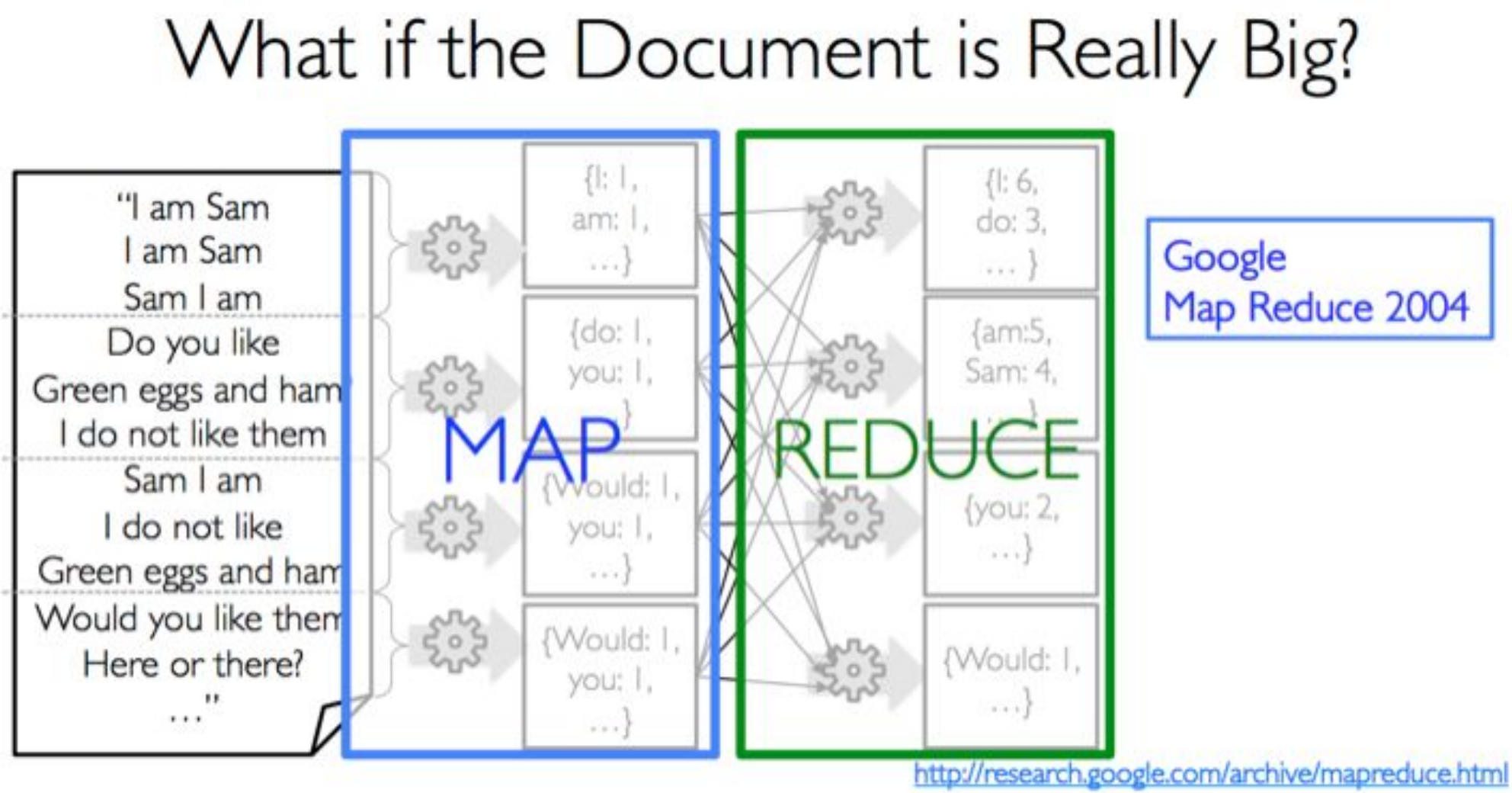
Spark的主要特性

http://spark.apache.org/

Hadoop

http://hadoop.apache.org/

map/reduce


What if the Document is Really Big?
"I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
I do not like them
Sam I am
I do not like
Green eggs and ham
Would you like them
Here or there?
..."
{I: 1, am: 1, ...}
{do: 1, you: 1, ...}
{Would: 1, you: 1, ...}
{Would: 1, you: 1, ...}
MAP
REDUCE
{I: 6, do: 3, ... }
{am:5, Sam: 4, ...}
{you: 2, ...}
{Would: 1, ...}
Google
Map Reduce 2004
http://research.google.com/archive/mapreduce.html

Spark和 hadoop的主要差異

執行結果 in memory vs HD

Hadoop (Java) vs Spark (Scala, Java, Python, R)

實作pySpark的word count範例

- py_word_count.ipynb
- shakespeare.txt
- kitmatic (beta)
- docker for mac (beta)
- jupyter_all_spark_notebook

相關的學習資源

https://www.youtube.com/watch?v=KzFe4T0PwQ8

https://www.youtube.com/watch?v=8VecnBTRxdg

https://www.youtube.com/watch?v=mL5dQ_1gkiA

参考資料

- http://spark.apache.org/
- http://hadoop.apache.org/
- https://en.wikipedia.org/wiki/MapReduce
-