Architectures for massive data management

Apache Spark Session Lab

Albert Bifet
albert.bifet@telecom-paristech.fr

October 20, 2015

Twitter

- Tweets are public
- Tweets are a data stream that can be read using a Twitter API
- The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet.
- A trend on Twitter refers to a hashtag-driven topic that is immediately popular at a particular time.

Apache Spark Session Lab

- Download Apache Spark
  - `http://spark.apache.org/downloads.html`
- Follow the Apache Spark Quick Start Tutorial
  - `http://spark.apache.org/docs/latest/quick-start.html`
- Read the Apache Spark RDD Guide
  - `http://spark.apache.org/docs/latest/programming-guide.html#resilient-distributed-datasets-rdds`
- Read the Apache Spark DataFrame Guide
  - `http://spark.apache.org/docs/latest/sql-programming-guide.html#dataframes`

Apache Spark Session Lab

- Download this dataset of tweets:
  - `http://www.datacrucis.com/datasets/stratahadoop-barcelona-2014-tweets.html`
- Start the Spark Shell
- Read the file in Spark, and get a DataFrame
  - `val df = sqlContext.read.json("filename")`
- Get an RDD with the text of the tweets
  - `val rdd = df.select("text").rdd.map(row => row.getString(0))`
- Count words
  - `val wordCounts = rdd.flatMap(_.split(" ")).map(word => (word,1)).reduceByKey((a,b) => a+b)`
- Show 10 word counts
  - `wordCounts.take(10).foreach(println)`

Apache Spark Session Lab Assignment

Write a report on the following tasks, writing the code in Scala and using Apache Spark:

1. Find hashtags on tweets
2. Count hashtags on tweets
3. Select the 10 most frequent hashtags
4. Select the 10 users with more tweets
5. Detect trending topics