IoT Big Data Processing

Apache Spark Session Lab

Albert Bifet and Jacob Montiel

October 4, 2017

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

- Login in the Community Edition of Databricks:
  - `http://community.cloud.databricks.com/`
- Follow and read these two notebooks:
  - Introduction to Apache Spark on Databricks
  - Quick Start DataFrames

- Start creating a new notebook
- Create a cluster
- Attach cluster to notebook

Apache Spark Session Lab

- Download this dataset of tweets:

```
import sys.process._
"wget -P /tmp https://www.datacrucis.com/media/datasets/stratahadoop-BCN-2014.json" !!
```

- Read the file in Spark, and get a DataFrame

类似于数据表的多维表

```
val localpath="file:/tmp/stratahadoop-BCN-2014.json"
dbutils.fs.mkdirs("dbfs:/datasets/")
dbutils.fs.cp(localpath, "dbfs:/datasets/")
display(dbutils.fs.ls("dbfs:/datasets/stratahadoop-BCN-2014.json"))

val df = sqlContext.read.json("dbfs:/datasets/stratahadoop-BCN-2014.json")
```

- 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 notebook on the following tasks, writing the code in Scala:

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