05: Spark on Zeppelin – semi-structured log file

Posted on July 28, 2018

This tutorial extends the series: Spark on Apache Zeppelin Tutorials.

Step 1: Pull apache/zeppelin image from the docker hub, and build the image with the following command.

```
$ docker pull apache/zeppelin:0.7.3
```

"docker images" will show the image that was created.

300+ Java Interview FAQs

300+ Java FAQs 🔥

16+ Java Key Areas Q&As

150+ Java Architect FAQs

80+ Java Code Quality Q&As

150+ Java Coding Q&As

300+ Big Data Interview FAQs

300+ Big Data FAQs 🔥

Tutorials - Big Data

TUT - ⏯ Starting Big Data

TUT - Starting Spark & Scala

Step 2: Run the above image to create a container with the following command.

```
$ docker run -it -p 8080:8080 apache/zeppelin:0.7.
```

Step 3: Go to a browser and type: “http://locahost:8080” to verify if zeppelin note book is accessible.

Step 4: Create a ssh session into the docker container from a terminal:

Firstly, get the container id with:

```
$ docker ps
```

Use the container id E.g. 89bce1cdc799 to ssh.

```
$ docker exec -it 89bce1cdc799 /bin/bash
```

Step 5: Once inside the docker container running on ubuntu, install “vim” editor.

```
$ apt-get update
$ apt-get install vim
```

Step 6: Create a temp.log file with vim.

```
$ vim temp.log
```

TUT - Starting with Python
TUT - Kafka
TUT - Pig
TUT - Apache Storm
TUT - Spark Scala on Zeppelin
TUT - Cloudera
TUT - Cloudera on Docker
TUT - File Formats
TUT - Spark on Docker
TUT - Flume
TUT - Hadoop (HDFS)
TUT - HBase (NoSQL)
TUT - Hive (SQL)
TUT - Hadoop & Spark
TUT - MapReduce
TUT - Spark and Scala
TUT - Spark & Java
TUT - PySpark on Databricks
TUT - Zookeeper

Enter the following text by pressing “i” to insert.

```
ERROR 2009-09-19 09:56:01 main RDFDefaultErrorHan
INFO 2009-09-13 09:56:01 login Some.java:44 Some
INFO 2009-09-11 09:56:01 login Some.java:44 Some
WARNING 2009-09-09 09:56:01 main MarketData.java:
INFO 2009-09-06 09:56:01 login Some.java:44 Some
INFO 2009-09-05 09:56:01 login Some.java:44 Some
INFO 2009-09-02 09:56:01 login Some.java:44 Some
INFO 2009-09-01 09:56:01 login Some.java:44 Some
ERROR 2009-08-25 09:56:01 main RDFDefaultErrorHan
```

exit vim by pressing “wq!”.

Step 7: On zeppelin note book on the browser http:.localhost:8080, create a note book and use “%sh” to list the files.

```
%sh

ls -ltr
pwd
```

Step 8: Create a spark RDD by reading this temp.log file. The path to temp.log file is “file:///zeppelin/temp.log”

```
%spark

val logsRdd = sc.textFile("file:///zeppelin/temp.l
logsRdd.take(10).mkString("n")
```

Step 9: Create a DataFrame as shown below:

```
```

```
%spark

import java.text.SimpleDateFormat
import java.sql.Date

val DATE_FORMAT = new SimpleDateFormat("yyyy-mm-d

case class Log(level: String, date: java.sql.Date

val logsDf = logsRdd.map { line =>

    val s =  line.split(" ")
    val logLevel = s(0)
    val dateTime = DATE_FORMAT.parse(s(1) + " " +
    val fileName = s(3).split(":")(0)
    Log(logLevel,new Date(dateTime.getTime()), fi

}.toDF()

logsDf.show()

```

Output:

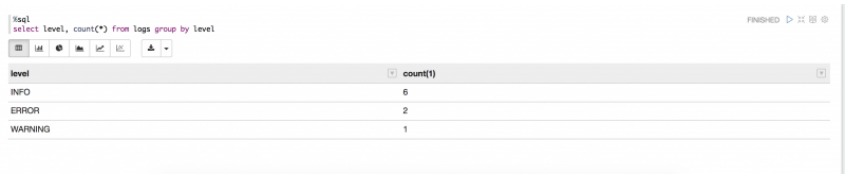
```

+-------+----------+--------+
|  level|      date|fileName|
+-------+----------+--------+
|  ERROR|2009-01-19|    main|
|   INFO|2009-01-13|   login|
|   INFO|2009-01-11|   login|
|WARNING|2009-01-09|    main|
|   INFO|2009-01-06|   login|
|   INFO|2009-01-05|   login|
|   INFO|2009-01-02|   login|
|   INFO|2009-01-01|   login|
|  ERROR|2009-01-25|    main|
+-------+----------+--------+

```

Step 10: Register as temp table so that you can run SQL queries to analyze the data.

```

%spark

logsDf.registerTempTable("logs")

```

Step 11: Analyse with SQL – tabular and graph.

```
%sql

select level, count(*) from logs group by level

```



zeppelin log file analysis

‹ 04: Spark on Zeppelin – DataFrame joins in Scala

40+ Apache Spark best practices & optimisation interview FAQs – Part-1 ›

Disclaimer