Spark: Inferring Schema Using Case Classes

To make this recipe one should know about its main ingredient and that is case classes. These are special classes in Scala and the main spice of this ingredient is that all the grunt work which is needed in Java can be done in case classes in one code line. Spark uses reflection on case classes to infer schema.

Recipe for this is given below

1. Start the spark shell and give it some additional memory:

```
$ spark-shell --driver-memory 1G
```

2. Import for the implicit conversations:

```
scala> import sqlContext. implicits._
```

3. Create a person case class:

```
scala> case class Person (first_name:String,last_name:
String,age:Int)
```

4. In another shell, create some sample data to be put in HDFS:

```
$ mkdir person
$ echo "Barack,Obama,53" >>person/person.txt
$ echo "George,Bush,68" >>person/person.txt
$ echo "Bill,Clinton,68" >>person/person.txt
$ hdfs dfs -put person person
```

5. Load the person directly as on RDD:

```
scala> val p = sc.textFile
("hdfs://localhost:9000/user/hduser/person")
```

6. Split each line into an array of string, based on a comma, as a delimiter:

```
val pmap = p.map ( line => line.split (","))
```

7. Convert the RDD of Array[string] into the RDD of person case objects:

```
scala> val personRDD = pmap.map ( p => Person (p(0), p(1),
p(2). toInt))
```

8. Convert the personRDD into the personDF DataFrame:

```
scala> val personDF = personRDD. toDF
```

9. Register the personDF as a table:

```
scala> personDF.registerTempTable ("person")
```

10. Run a SQL query against it:

```
scala> val people = SQL ("select * from person")
```

11. Get the output values from persons:

```
scala> people.collect.foreach (printIn)
```


Recent Posts

Recent Comments

Archives

March 2017

November 2016

September 2016

June 2016

March 2016

February 2016

January 2016

August 2015

July 2015

June 2015

April 2015

March 2015

February 2015

December 2014

November 2014

Categories

Blog

Courses

Job Openings in San Jose, Silicon Valley

Spark Cookbook

Meta

Log in

Entries RSS

Comments RSS

WordPress.org

Infoobjects is a consulting company that helps enterprises transform how and where they run infrastructure and applications. From strategy, to implementation, to ongoing managed services, Infoobjects creates tailored cloud solutions for enterprises at all stages of the cloud journey.

- Our Clients
- Privacy
- Media
- Employment
- Spark on Tap

Recent Posts

- HBase Client in a Kerberos Enabled Cluster
- AWS: CIDR blocks explained
- Amazon DynamoDB Sizing Demystified