databricks (/)

Find posts, topics, and users...

Home (/) /

Query a MongoDB collection using Pyspark

(/users/2726/rporwal.html) pyspark (/topics/pyspark.html) apache spark (/topics/apache+spark.html) mongodb (/topics/mongodb.html)

1

Question by rporwal (/users/2726/rporwal.html) · Jul 27, 2016 at 01:12 PM ·

I am building an application utilizing PHP with **MongoDB** as a database.One of collection across DB has massive volume of data i.e **8GB** data. I perform **aggregate** operation on data stored in MongoDB collection and accordingly generate statistics . But processing huge volume of data takes a long duration.Hence I opted for Apache spark (http://spark.apache.org/) to process data stored in **MongDB collection** I have configured MongoDB spark connector (https://docs.mongodb.com/spark-connector/) and executed a demo script in **python** to fetch data from mongo collection through spark.

Following is python code snippet

```
<code>from pyspark importSparkContext,SparkConf
from pyspark.sql importSQLContext
conf=SparkConf()
conf.set('spark.mongodb.input.uri','mongodb://[host]/db.collection')
conf.set('spark.mongodb.output.uri','mongodb://[host]/db.collection')
sc =SparkContext(conf=conf)
sqlContext =SQLContext(sc)
df = sqlContext.read.format("com.mongodb.spark.sql.DefaultSource").load()
df.printSchema()
df.registerTempTable("mycollection")
result_data=sqlContext.sql("SELECT * from mycollection limit 10")
result_data.show()
```

In above code snippet I have utilized pyspark.sql module (https://spark.apache.org/docs/1.6.1/api/python/pyspark.sql.html#module-pyspark.sql) to generate **RDD**. But generation of **RDD** incurs reading of all data from collection which takes a long duration to read massive volume of data as opposed to principle on which **Apache Spark** works. Hence suggest me an appropriate solution to **read data from Mongo collection using pyspark with optimal performance** and also if any **alternate package in Apache spark** exists to **communicate with MongoDB**.

Add comment

1 Answer

Sort

Answer by girivaratharajan (/users/1077/girivaratharajan.html) · Jul 28, 2016 at 10:59 PM

(/users/1077/girivaratharajan.html)

@rporwal (/users/2726/rporwal.html) -

1

You can try with “com.stratio.datasource” % “spark-mongodb_2.10” % “0.11.1” package as well.

```
import com.stratio.datasource.mongodb.
import com.stratio.datasource.mongodb.config.
import com.stratio.datasource.mongodb.config.MongodbConfig._
```

```
val builder = MongodbConfigBuilder(Map(Host -> List(“hostname:27017”), Database -> “test”, Collection ->”testcollection”, SamplingRatio -> 1.0, WriteConcern -> “normal”))
val readConfig = builder.build()
val mongoRDD = sqlContext.fromMongoDB(readConfig)
mongoRDD.registerTempTable(“testtable”)
val df = sqlContext.sql(“SELECT * FROM testtable”)
//df.printSchema()
df.take(30).foreach(println)
```

You can also convert your collection to a BSON Document and read through pymongo spark package available in. This method is pretty faster when compared to other approaches.

https://github.com/mongodb/mongo-hadoop/blob/master/spark/src/main/python/README.rst (https://github.com/mongodb/mongo-hadoop/blob/master/spark/src/main/python/README.rst)

```
import pymongo_spark
pymongo_spark.activate()
bsonFileRdd = sc.BSONFileRDD(“/home/vgiri/nettuts.bson”)
bsonFileRdd.take(5)
```

More important is mongo-hadoop-spark.jar and setup.py (http://setup.py) file should be added in the class path while executing this.

Add comment · Share

Your answer

Fill in the details...

Hint: You can notify a user about this post by typing @username

□ I'm not a robot

reCAPTCHA
Privacy - Terms

Post Answer

Follow this Question

9 People are following this question.

(/us [illegible] krishna-
(/users/ [illegible] rajan.html)

Related Questions

How to execute python script using spark submit in php? (/questions/9467/how-to-execute-python-script-through-spark-submit.html) 1 Answer

reading and writing using Spark (R & python) from Hdfs (/questions/10561/reading-and-writing-using-spark-r-python-from-hdfs.html) 2 Answers

How to configure an external package into Apache Spark? (/questions/9531/how-to-configure-an-external-package-into-apache-s.html) 0 Answers

How to move files to other storage from s3 ? (/questions/10076/how-to-move-files-to-other-storage-from-s3.html) 2 Answers

Creating a subclass of RDD and DStream in PySpark? (/questions/10179/creating-a-subclass-of-rdd-and-dstream-in-pyspark.html) 0 Answers

Databricks Inc.
160 Spear Street, 13th Floor
San Francisco, CA 94105

info@databricks.com (mailto:info@databricks.com)
1-866-330-0121