SPARK SQL
Data Frames & Data Sets

Genoveva Vargas-Solar

Senior Scientist, French Council of Scientific Research

Javier A. Espinosa Oviedo

Postdoctoral fellow, Barcelona Super Computing Centre

Verano Académico, Télécom SudParis, 26th June 2017
http://vargas-solar.com/datacentric-sciences/

WORKING WITH STRUCTURED DATA

Extends RDD to a "DataFrame" object

- Contains Row objects
- Can run SQL queries
- Has a schema (leading to more efficient storage)
- Read and write to JSON, Hive, parquet
- Communicates with JDBC/ODBC, Tableau

DATA FRAME

- DataSet of Row objects: `DataSet[Row]`
  - - DataSets can explicitly wrap a given struct or type:
    - `DataSet[Person], DataSet[(String, Double)]`
  - A DataSet knows what its columns are from the get-go
- A DataFrame schema is inferred at runtime
- A DataSet can be inferred at compile time
  - Faster detection of errors, and better optimization
  - - RDD's can be converted to DataSets with `.toDS()`

DATASETS ARE THE NEW HOTNESS

The trend in Spark is to use DataSets

DataSets are more efficient

- They can be serialized very efficiently
- Optimal execution plans can be determined at compile time

DataSets allow for better interoperability

- MLLib and Spark Streaming are moving toward using DataSets instead of RDD's for their primary API

DataSets simplify development

- You can perform most SQL operations on a dataset with one line

USING SPARK SQL IN SCALA

In Spark 2.0.0

- Create a SparkSession object instead of a SparkContext when using Spark SQL / DataSets
  - You can get a SparkContext from this session, and use it to issue SQL queries on your DataSets!
  - Stop the session when you're done.

```
myResultDataFrame.show()

myResultDataFrame.select("someFieldName")

myResultDataFrame.filter(myResultDataFrame("someFieldName“) > 200)

myResultDataFrame.groupBy(myResultDataFrame("someFieldName")).mean()

myResultDataFrame.rdd().map(mapperFunction)
```

SHELL ACCESS

Spark SQL exposes a JDBC/ODBC server (if you built Spark with Hive support)

Start it with `sbin/start-thriftserver.sh`

Listens on port 10000 by default

Connect using `bin/beeline -u jdbc:hive2://localhost:10000`

Viola, you have a SQL shell to Spark SQL

You can create new tables, or query existing ones that were cached using `hiveCtx.cacheTable("tableName")`

LET'S PLAY WITH SPARK SQL AND DATAFRAMES

Use the fake social network data provided for the exercise

Query it with SQL, and then use DataSets without SQL

Finally we'll re-do our popular movies example with DataSets, and see how much simpler it is.

Genoveva Vargas-Solar

CR1, CNRS, LIG-LAFMIA

Genoveva.Vargas@imag.fr

http://vargas-solar.com/datascience