Spark SQL

Ladle Patel

DataFrames

- Spark SQL is a Spark module for structured data processing.
- A DataFrame is a Dataset organized into named columns. It is conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer optimizations under the hood.
- DataFrames can be constructed from a wide array of sources such as: structured data files, tables in Hive, external databases, or existing RDDs.
- The DataFrame API is available in Scala, Java, Python, and R.

DataSets

- A Dataset is a distributed collection of data.
- Dataset is a new interface added in Spark 1.6 that provides the benefits of RDDs (strong typing, ability to use powerful lambda functions) with the benefits of Spark SQL's optimized execution engine.
- The Dataset API is available in Scala and Java.
- Python does not have the support for the Dataset API. But due to Python's dynamic nature.

SparkSession

- The entry point to programming Spark with the Dataset and DataFrame API.

- To create a basic SparkSession, just use SparkSession.builder()

- import org.apache.spark.sql.SparkSession

val spark = SparkSession.builder().appName("Spark SQL basic example").config("spark.some.config.option", "some-value").getOrCreate()

- // For implicit conversions like converting RDDs to DataFrames

- import spark.implicits._

Operations

```
val df = spark.read.json("examples/src/main/resources/people.json")

df.show()

df.printSchema()

df.select("name").show()

df.select("name", "age" + 1).show()

df.filter("age" > 21).show()

df.groupBy("age").count().show()
```

Operations Cont..

```
df.filter("age" > 21).show()

df.groupBy("age").count().show()

df.createOrReplaceTempView("people")

val sqlDF = spark.sql("SELECT * FROM people")

sqlDF.show()
```

Thank You.