01_Basic_Part1

March 28, 2020

0.1 Load and show data frames

[1]:
```
from pyspark.sql import SparkSession
```

[2]:
```
spark = SparkSession.builder.appName("Basics01").getOrCreate()
```

Note that after typing the dot, you can press tab to get a list of all available methods. Shift tab for documentation of the function/method

[3]:
```
path = "Python-and-Spark-for-Big-Data-master/Spark_DataFrames/people.json"
```

[4]:
```
df = spark.read.json(path)
```

[5]:
```
df.show()
```

```
+----+-------+
| age|   name|
+----+-------+
|null|Michael|
|  30|   Andy|
|  19| Justin|
+----+-------+
```

[6]:
```
df.printSchema()
```

```
root
 |-- age: long (nullable = true)
 |-- name: string (nullable = true)
```

[7]:
```
df.columns
```

[7]:
```
['age', 'name']
```

[8]:
```
df.describe().show()
```

```
+-------+------------------+-------+
|summary|               age|   name|
+-------+------------------+-------+
```

```
|  count|                 2|      3|
|   mean|              24.5|   null|
| stddev|7.7781745930520225|   null|
|    min|                19|   Andy|
|    max|                30|Michael|
+-------+------------------+-------+
```

[9]:
```
df.describe()
```

[9]:
```
DataFrame[summary: string, age: string, name: string]
```

0.2 Manually define the schema

[10]:
```
from pyspark.sql.types import (StructField, StringType,
                               IntegerType, StructType)
```

[11]:
```
data_schema = [StructField("age", IntegerType(), True),
               StructField("name", StringType(), True)]
```

[12]:
```
final_struc = StructType(fields = data_schema)
```

[13]:
```
df = spark.read.json(path, schema = final_struc)
```

[14]:
```
df.printSchema()
```

```
root
 |-- age: integer (nullable = true)
 |-- name: string (nullable = true)
```

[ ]:
```

```