DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

Spark SQL

Jan Hučín

December 4th, 2019

Agenda

1. Why Spark SQL?
2. RDD and DataFrame: what‘s the same and what‘s different
3. RDD to DataFrame conversion
4. DataFrame transformations and actions

PROFINIT

Why Spark SQL?

Spark SQL and DataFrames (DataSets)

- Enhances the classical RDD approach
- Data structure **DataFrame** = „RDD with columns“
  - similar to database relation table
  - with metadata (field names, types)
  - works with columns
  - using SQL-like syntax or directly SQL

1;Andrea;35;64.3;Praha
2;Martin;43;87.1;Ostrava
3;Simona;18;57.8;Brno

id	name	age	weight	city
1	Andrea	35	64.3	Praha
2	Martin	42	87.1	Ostrava
3	Simona	18	57.8	Brno

PROFINIT

Spark SQL – pros and requirements

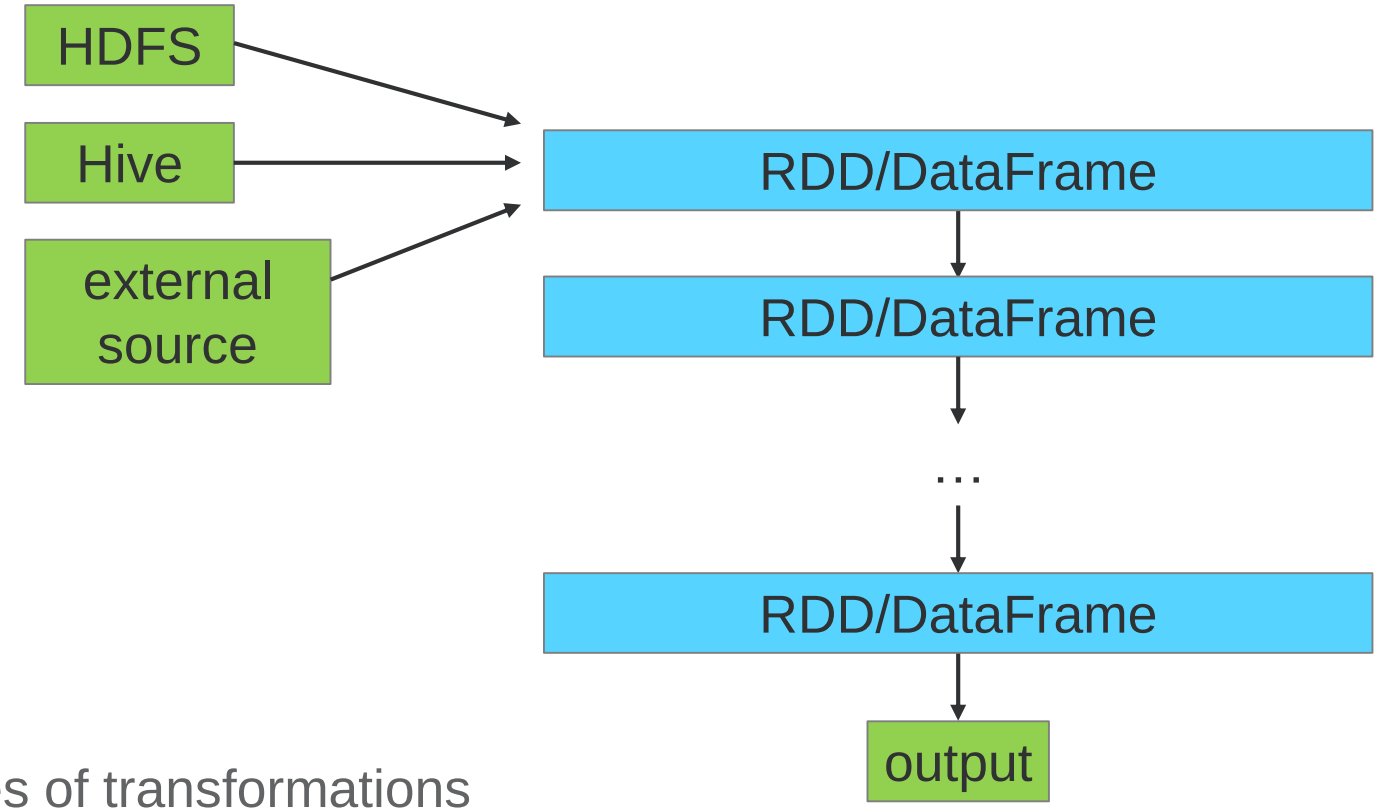
- Advantage over Spark RDD:
  - shorter and easier code
  - Hive usage
  - easier and better optimized
  - ⇨ runs faster

- Requirements
  - enhanced API: **`sqlContext`** object

- When not to use
  - tasks unproper for SQL (unstructured data) ⇨ Spark RDD
  - taks with big memory demands ⇨ map-reduce, Hive

PROFINIT

Spark RDD and SQL



- series of transformations finished by action
- RDD can be transformed to DataFrame and back

PROFINIT

Example

- Which USA state has the highest average temperature in summer?
  (solved by Hive)

Data sample:

```
station,month,day,hour,temperature,flag,latit,longit,elevation,state,name
AQW00061705,1,1,1,804,P,-14.3306,-170.7136,3.7,AS,PAGO PAGO WSO AP
AQW00061705,1,2,1,804,P,-14.3306,-170.7136,3.7,AS,PAGO PAGO WSO AP
AQW00061705,1,3,1,803,P,-14.3306,-170.7136,3.7,AS,PAGO PAGO WSO AP
AQW00061705,1,4,1,802,P,-14.3306,-170.7136,3.7,AS,PAGO PAGO WSO AP
AQW00061705,1,5,1,802,P,-14.3306,-170.7136,3.7,AS,PAGO PAGO WSO AP
```

PROFINIT

Solution by RDD only - uncomfortable

```
tp_raw = sc.textFile('/user/pascepet/teplota')

tp_raw = tp_raw.filter(lambda r:
   (r.split(',')[1] in set('678')) & (r.split(',')[4] != ''))

tp = tp_raw.map(adjust_row)

tp_st = tp.reduceByKey(sums) \

    .map(lambda x: (x[0], x[1][0]/x[1][1])) \

    .sortBy(lambda y: y[1], False)

tp_st.take(1)
```

```
adjust_row

AQW00061705,7,30,4,804,P,-14.3306,-170.7136,3.7,AS,PAGO PAGO WSO AP
                                  ↓
                          (AS, (26.89, 1))
```

PROFINIT

How to get a DataFrame?

- transformation from existing RDD
  - if convertable
  - **`sqlContext.createDataFrame(`*`RDD, schema`*`)`**
- direct input of file
  - schema may be defined (Parquet, ORC) or inferred (CSV)
  - **`sqlContext.read.format(`*`format`*`).load(`*`path`*`)`**
- Hive query
  - **`sqlContext.sql(`*`sql_query`*`)`**

PROFINIT

Example: direct input of CSV → DataFrame

```
tpDF2 = sqlContext.read \
        .format("com.databricks.spark.csv") \
        .option("header", "true") \
        .option("delimiter", ",") \
        .option("inferSchema", "true") \
        .load("/user/pascepet/teplota")
```

Example: Hive query → DataFrame

```
tpDF3 = sqlContext.sql('select * from temperature')
```

PROFINIT

How to work with a DataFrame?

1. registration of temporary table + SQL querying
   - `DF.registerTempTable("table")`
   - `sqlContext.sql("select * from table")`
2. SQL-like operations
   - `DF.operations`; select, filter, join, groupBy, sort...
3. RDD operation (map, flatMap, …) the results is RDD, not DataFrame

PROFINIT

SQL-like operations

- **select**
- **filter**
- **join**
- **groupBy**
- **agg, avg, count**
- **toDF** (columns renaming)
- **withColumn** (columns transformation)

The result is DataFrame.

Action example: **show** (nice printing)

PROFINIT

Solutions: temp. table + SQL; SQL-like operations

```
tpDF.registerTempTable("t")

tp_stDF = sqlContext.sql("""select state, avg(temperature) as
temp_avg from t
group by state order by temp_avg desc""")

tp_stDF.show(1)
```

```
tpDF2 = tpDF2.filter((tpDF2.mesic>5) & (tpDF2.mesic<9)) \
    .select('stat','tepl').na.drop()
tpDF2 = tpDF2.groupBy('stat').avg() \
    .toDF('stat',,tepl_prum')
tpDF2.sort(tpDF2.tepl_prum.desc()).limit(1).show()
```

PROFINIT

Thank you for your attention

PROFINIT

Profinit, s.r.o.
Tychonova 2, 160 00 Praha 6

Telefon
+ 420 224 316 016

Web
www.profinit.eu

LinkedIn
linkedin.com/company/profinit

Twitter
twitter.com/Profinit_EU