DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

# Spark SQL – cvičení

Jan Hučín, Adéla Dragounová, Dominik Matula

28. listopadu 2018

# Opakování

- Základní datová struktura Sparku = RDD
- Rozšíření pro práci se sloupci = DataFrame
- DataFrame získáme:
  konverzí RDD, přímým načtením souboru, z Hive

PROHIN

# Přímé načtení CSV → DataFrame

```
DF = sqlContext.read \
     .format("com.databricks.spark.csv") \
     .option("header", "true") \
     .option("delimiter", ",") \
buď  .schema(nazev_schematu) \
nebo .option("inferSchema", "true") \
     .load(cesta)
```

PROHIN

# Jak pracovat s DataFrame?

1. registrovat jako dočasnou tabulku + dotazování SQL
   - `DF.registerTempTable("tabulka")`
   - `sqlContext.sql("sql_dotaz")`
2. pseudo-SQL operace
   - `DF.operace`, např. select, filter, join, groupBy, sort...
3. operace RDD – výsledek může být jen obyčejné RDD
   - např. map, flatMap...
   - řádek v DataFrame je typu **Row** – práce jako s typem **list**

PROFIN

# Pseudo-SQL a další operace

- **select** (omezení na uvedené sloupce)
- **filter** (omezení řádků podle podmínky)
- **join** (připojení jiného DataFrame)
- **groupBy** (seskupení)
- **agg, avg, count** (agregační funkce)
- **toDF** (přejmenování sloupců)
- **withColumn** (transformace sloupců)
- **show** (hezčí výpis obsahu DataFrame)

PROFIN

# Pseudo-SQL

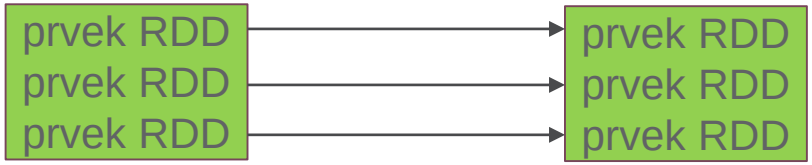
```
DF = DF.filter((DF.mesic>5) & (DF.mesic<9)) \
    .select('stat','tepl').na.drop()
DF = DF.withColumn('tepl', (DF.tepl/10.0-32) * 5/9)
DF = DF.groupBy('stat').avg() \
    .toDF('stat','prum')
DF.sort(DF.prum.desc()).show(1)
```

PROHIN

# Transformace v RDD a v DataFrame

# Transformace v RDD a v DataFrame

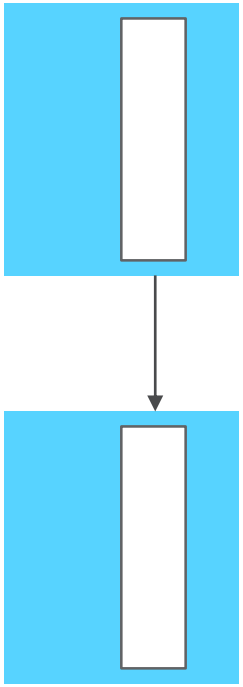
RDD map




transformační funkce:

Python
obvyklé knihovny a objekty
(string.lower, re.search atd.)

DataFrame withColumn



transformační funkce:

Pyspark SQL functions

manipulace se sloupci

nutno importovat, např.
*from pyspark.sql import functions as F*

F.lower, F.regexp_replace atd.

PROFIN

# Pyspark SQL functions – příklady

- **split** (rozdělení řetězce – výsledek je array)
- **size** (počet prvků array – odpovídá funkci len)
- **lower** (na malá písmena)
- **regexp_replace** (náhrada podle regulárního výrazu)
- **udf** (uživatelská funkce – pokud nelze použít funkci Spark SQL)
- **when... otherwise** (ifelse)
- **lit** (konstanta)

**Příklad:**

```
RDD2 = RDD1.map(lambda s: s[2].lower)


from pyspark.sql import functions as F
DF2 = DF1.withColumn('tepl_mala', F.lower(DF1.tepl))
```

PROHIN

# Díky za pozornost

PROFIN

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