

Spark Dataframe

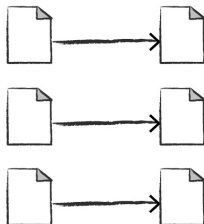
Marco Milanesio

MS Data Science 2020-2021

- `myRange = spark.range(1000).toDF("number")`
- `divisBy2 = myRange.where("number % 2 = 0")`
- Two types of transformations, specifying:
 - **Narrow** dependencies
 - **Wide** dependencies

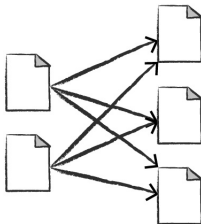
Narrow transformations 1:1

- each input partition will contribute to only one output partition.
- **pipelining**
- `.filter()`, `.where()`, `.map()`, ...



Wide transformations 1:n

- input partitions contributing to many output partitions
- **shuffle**
- `.groupByKey()`, `.reduceByKey()`, `.sort()`, ...



- End2End example
- Summarize most of the steps

Intro & basic imports

```
[1] !apt-get install openjdk-8-jdk-headless -qq > /dev/null
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
!update-alternatives --set java /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java
!java -version
!pip install pyspark
```

```
[3] from google.colab import files
```

```
[4] files.upload()
```

Build session and load the data

```
[5] from pyspark.sql import SparkSession

[7] datafile = '2015-summary.csv'

[8] spark = (SparkSession.builder.master('local').appName('flightsApp').getOrCreate())

[9] sc = spark.sparkContext

[12] flightData2015 = (spark
    .read
    .option('inferSchema', 'true')
    .option('header', 'true')
    .csv(datafile)
)
```

Let's go

Overview of last command in the coding session

