Introduction to Big Data with Apache Spark

Spark

databricks™

amplab
UC BERKELEY

Berkeley
UNIVERSITY OF CALIFORNIA

BerkeleyX

pandas: Python Data Analysis Library

- Open source data analysis and modeling library
  - » An alternative to using R

- pandas DataFrame: a table with named columns
  - » The most commonly used pandas object
  - » Represented as a Python Dict (column_name ➔ Series)
  - » Each pandas Series object represents a column
    - 1-D labeled array capable of holding any data type
  - » R has a similar data frame type

Semi-Structured Data in pySpark

- DataFrames introduced in Spark 1.3 as extension to RDDs
- *Distributed* collection of data organized into named columns
  - » Equivalent to Pandas and R DataFrame, but distributed
- Types of columns inferred from values

pySpark and pandas DataFrames

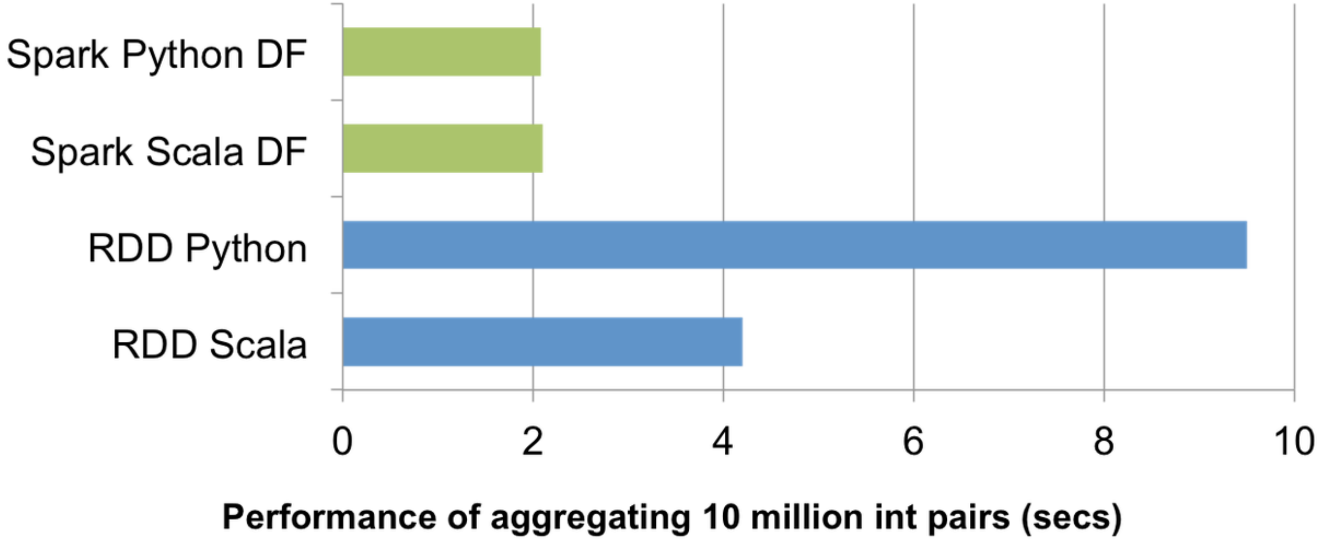
- Easy to convert between Pandas and pySpark
  - » *Note: pandas DataFrame must fit in driver*

```
# Convert Spark DataFrame to Pandas
pandas_df = spark_df.toPandas()

# Create a Spark DataFrame from Pandas
spark_df = context.createDataFrame(pandas_df)
```

pySpark DataFrame Performance

- Almost 5x pySpark performance on a single machine



https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html