Pyspark

Monitoria 2 de
Python contra o tempo

HDI SEGUROS

O que é PySpark?

O Apache Spark é escrito na linguagem de programação Scala. PySpark é uma API em Python para executar o Spark e foi lançado para oferecer suporte à colaboração entre Apache Spark e Python.

O PySpark também oferece suporte à interface do Apache Spark com conjuntos de dados distribuídos resilientes (RDDs) na linguagem de programação Python. Isso é obtido aproveitando a biblioteca Py4J.

Py4J

“A Bridge between Python and Java”

Py4J é uma biblioteca popular incorporada ao PySpark que permite a interface dinâmica com objetos na JVM do Python.

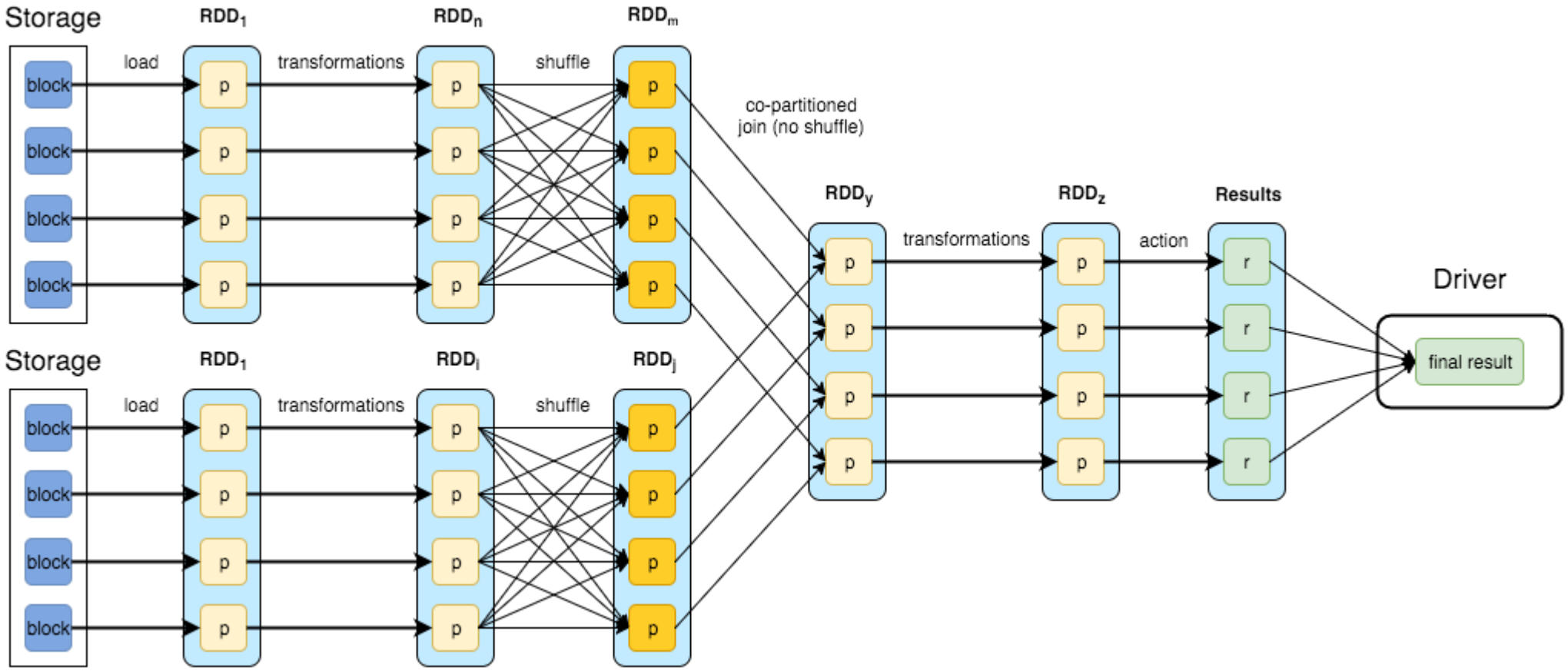
O PySpark possui muitas implementações de bibliotecas para programação eficiente e também possui bibliotecas externas compatíveis.

Aqui estão alguns exemplos:

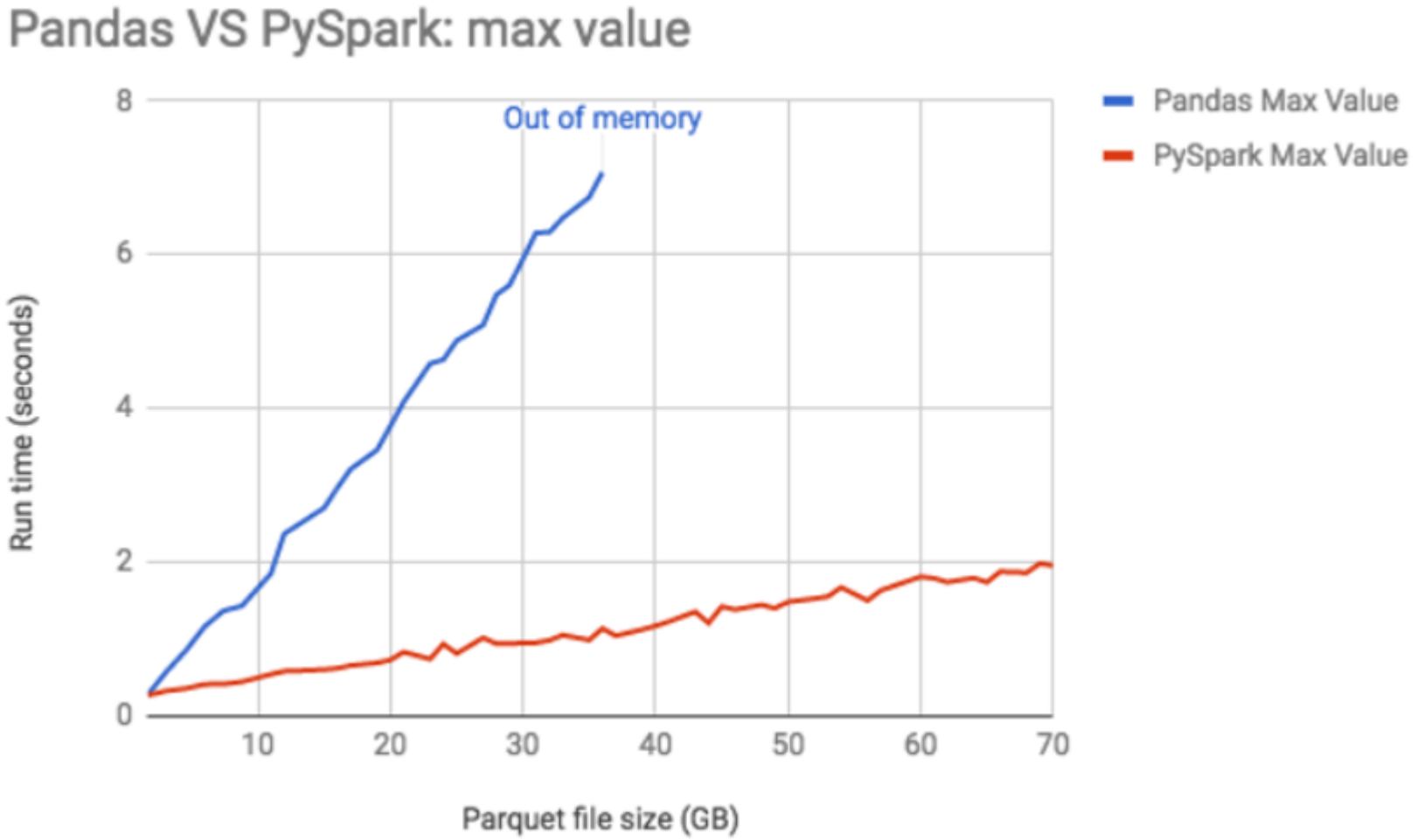
- **PySparkSQL**
- **MLlib**
- **GraphFrames**

PySparkSQL	MLlib	GraphFrames
PySparkSQL é uma biblioteca PySpark para análises semelhantes a SQL em grandes quantidades de dados estruturados e semiestruturados	MLlib é um wrapper para PySpark e a biblioteca de machine learning (ML) do Spark. Esta biblioteca usa a técnica de paralelismo de dados para armazenar e trabalhar com dados	GraphFrames é uma biblioteca de processamento de dados gráficos que usa PySpark Core e PySparkSQL para fornecer um conjunto de APIs para análise de gráficos eficiente. Também é otimizado para computação distribuída rápida. Vantagens de usar o PySpark: Python é extremamente fácil aprender e implementar.

Exemplo funcionamento Pyspark



Pyspark vs Pandas



🚫 Quando não usar

- **Processamento de dados pequenos Processamento de nó único Consultas SQL tradicionais**
- **Pc’s Fracos e com pouca memória**
- **Ambiente de baixa performace**

Para casos assim ainda aprenderemos o Dask, que pode ser um ferramenta mais adequada para determinados casos

Mãos à obra!

Bora para o Jupyter.