NYU | Center for Data Science

Lab 3: Spark and Parquet

Why Spark?

- Speed
- Ease of Use
- Generality
- Scales up (cluster) and down (local)

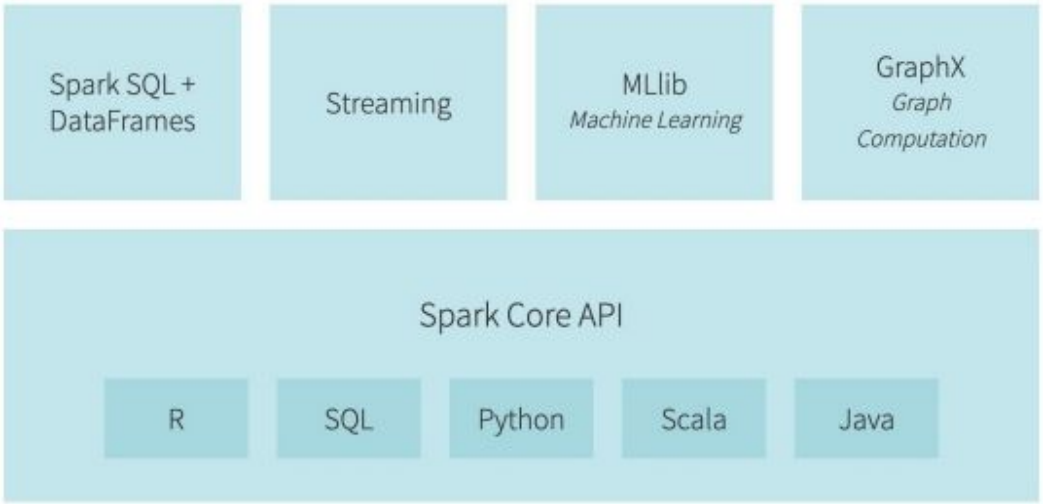
Lab 3 Assignment

- On Brightspaces Assignments — Lab 3
- Instructions for the assignment are provided in the README file of the repository.
- Submission deadline for this assignment is 04/01/2022.

Lab 3 overview

- Part 1:
  - Running spark jobs on peel
  - DataFrame processing with Spark and SparkSQL
- Part 2:
  - Benchmarking CSV vs Parquet
  - Optimizing storage for specific query loads

Spark Ecosystem



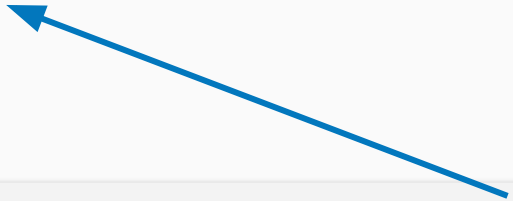
Running Spark on Peel

- **lab_3_starter_code.py** provides the skeleton code to get started
- Data files must be transferred to HDFS before you start.
- To run from peel: **spark-submit lab_3_starter_code.py**
- Spark operations go through the **SparkSession** object:
  - **spark** = SparkSession.builder.appName('part1').getOrCreate()
  - **spark**.read.csv('myfile.csv')

Retrieving output on peel

Script **output log** needs to be retrieved from **yarn** (cluster scheduler)

yarn logs -applicationId **<your_application_id>** -log_files **stdout**



Example:

tracking URL: http://horton.hpc.nyu.edu:8088/proxy/**application_1613664569968_2108**/

Transformations and Actions

Transformation:

Operations which transform one RDD or DataFrame into another. Computation is deferred.

Action:

RDD operations that give non-RDD values.
They trigger deferred computation.

Transformations	Actions
select	show
distinct	count
groupBy	collect
sum	save
orderBy	take
filter	
limit	

CSV vs Parquet

Spark Format Showdown		File Format		
		CSV	JSON	Parquet
Attribute	Columnar	No	No	Yes
	Compressable	Yes	Yes	Yes
	Splittable	Yes*	Yes**	Yes
	Human Readable	Yes	Yes	No
	Nestable	No	Yes	Yes
	Complex Data Structures	No	Yes	Yes
	Default Schema: Named columns	Manual	Automatic (full read)	Automatic (instant)
	Default Schema: Data Types	Manual (full read)	Automatic (full read)	Automatic (instant)

Lab 3 Assignment - Tips

- It is possible to obtain and use Spark/Pyspark locally!
- You can install and use the Pyspark interactive terminal locally to check your answers.
- However you must submit a working lab_3_starter_code.py (can work on Peel) for the first part of the assignment.
- In addition, all the queries in part 2 must also work on Peel.

Questions?