



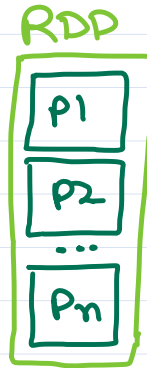
# Big Data Technologies

Trainer: Mr. Nilesh Ghule.



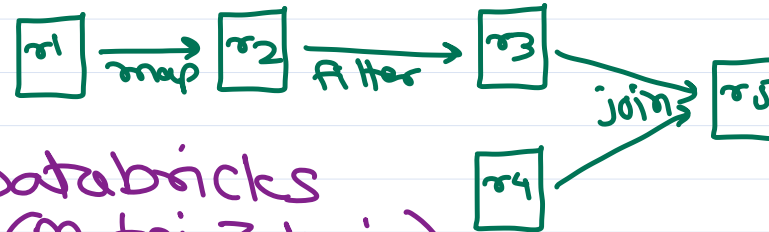
# Spark

UCB → AMPLabs  
↓  
Algorithm machine & People  
ML algos → on huge data set (GBs-TBs).



RDD: Resilient Distributed Dataset.

DAG: Directed Acyclic Graph.



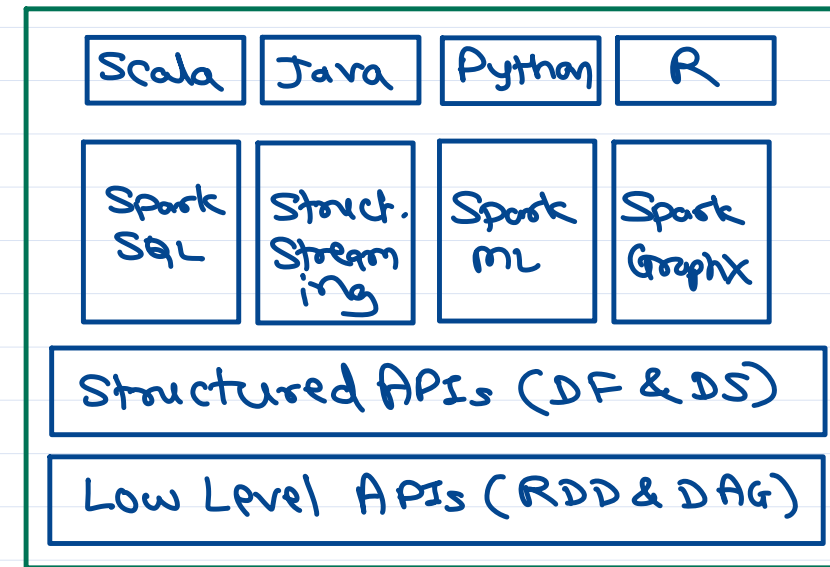
Databricks  
(Matei Zaharia)

Hadoop Limitation:

- ✓ Hadoop = HDFS + MR
- ✓ MR → slower execution.
- ✓ MR → RAM + Disk
- ✓ MR → Complex coding
  - ↳ Processes for Mapper & Reducer task.
- ✓ Mahout lib-ML.
- ✓ for commodity hw

→ Apache Spark (2008/09)

- ✓ Distributed Computing Framework - Can work with any Dist Storage.
- ✓ Processing in RAM.
- ✓ Faster execution
- ✓ Tasks → Threads.
- ✓ SQL + Streaming + ML + GraphX
- ✓ Simple api: RDD | DF
- ✓ Need high config machine (RAM)



Spark 2.x+ toolkit

pyspark

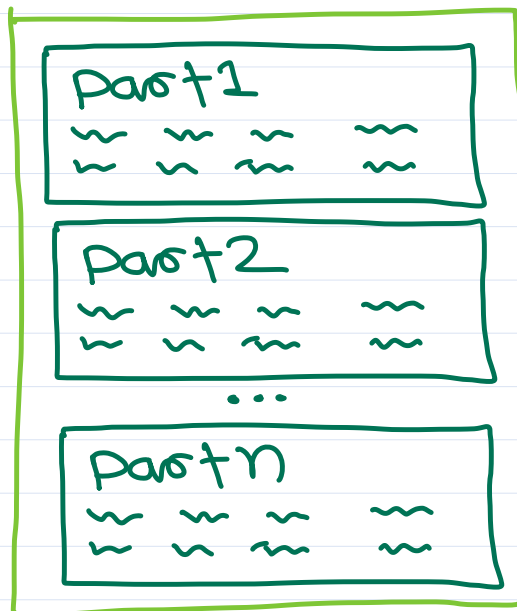
- ✓ Spark programming in Python.
- ✓ pyspark - pip pkg for Spark dev.

Spark-R

- ✓ Spark prog in R.

# Spark

DataFrame → abstraction over RDD.

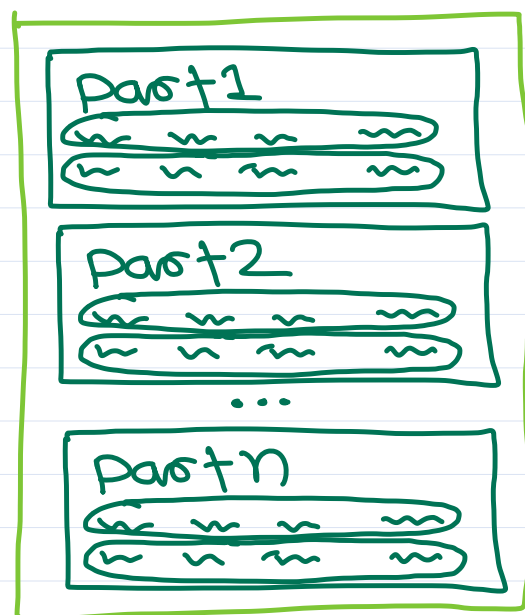


RDD → Functional programming.

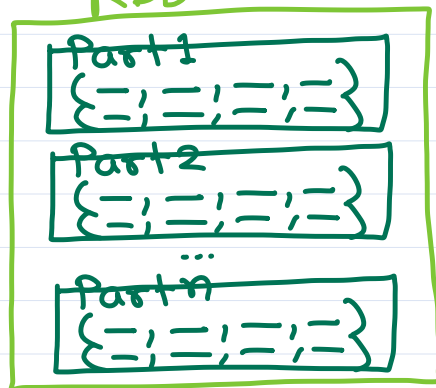
DF → SQL style programming  
(Rows & cols)

DS → collection of immutable objects. Available only in

DataSet Scala (not in Python).



RDD



Spark Vendors

- ① Databricks
- ② AWS EMR
- ③ Glue
- ④ GCP
- ⑤ Azure
- ⑥ MapR

Spark Philosophy

① Unified

↳ Same APIs in all langs.  
↳ Similar perform.

② Compute Engine

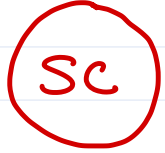
↳ Dist Computing with any Storage

③ Libraries

↳ SQL, ML, Stream, Graph.  
↳ spark-packages.org

# Spark

SparkContext: Main obj in Spark appn.



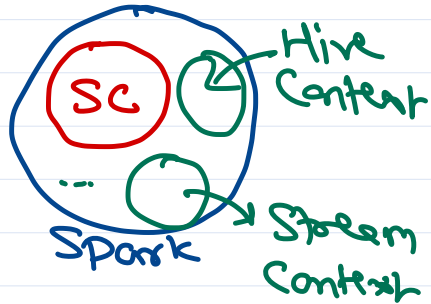
- Responsible for RDD creation & Job execution (DAG)
- Web UI (port 4040).

Low level API.

On spark shells, SparkContext (sc) and SparkSession (spark) is auto created.

For standalone appns, they must be created at the start of appn.

Session Session: Main obj in Spark appn.



High level API

Wrapper on SparkContext.

Responsible for DataFrame creation & Job execution  
Web UI (port 4040).

Encapsulate additional ctxs if needed

Singleton pattern: one for whole appn.

df = spark.read ← DataFrame Reader

- option("header", True)
- option("inferSchema", True)
- csv("/path/of/csv")





Thank you!

*Nilesh Ghule <nilesh@sunbeaminfo.com>*

