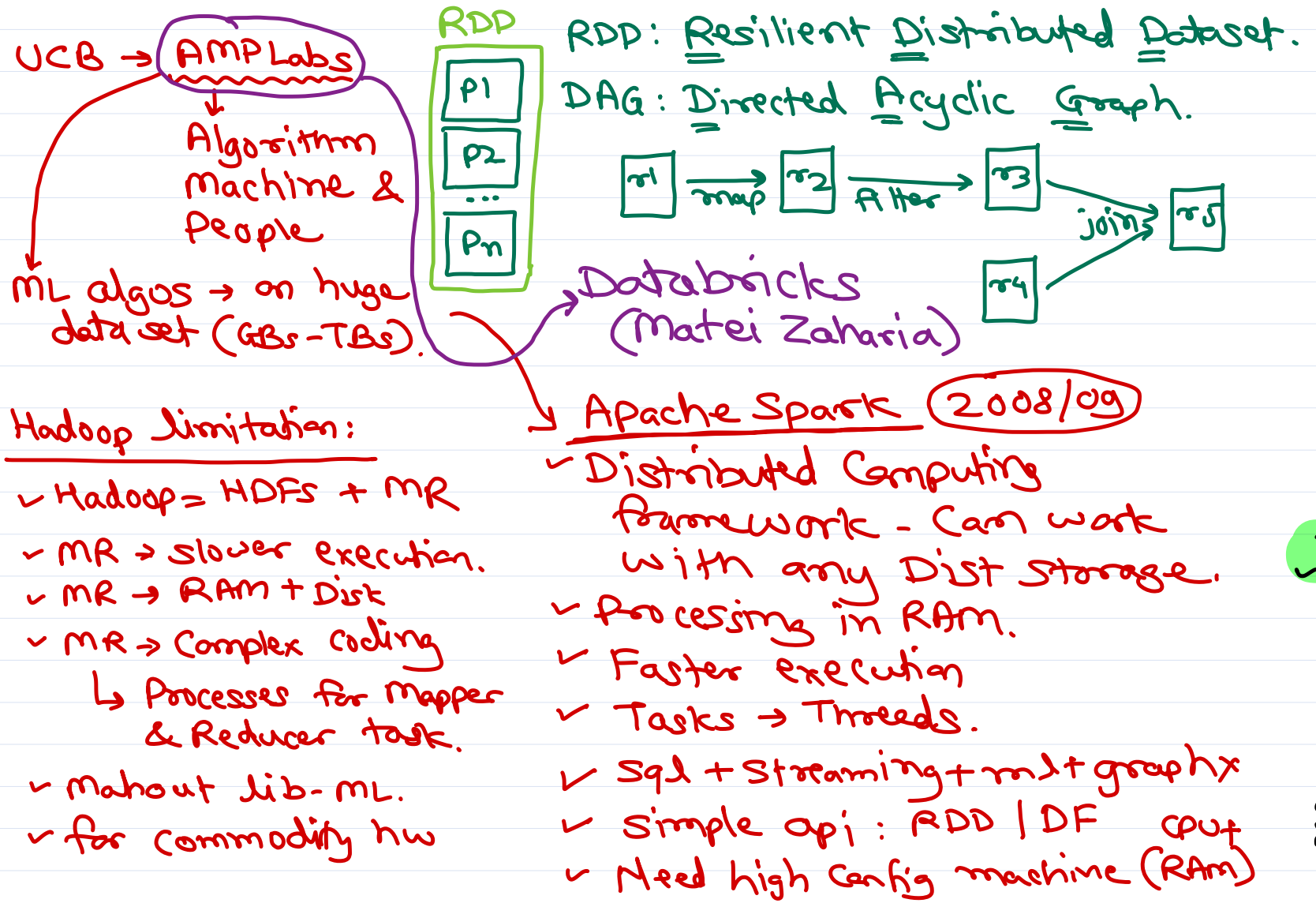
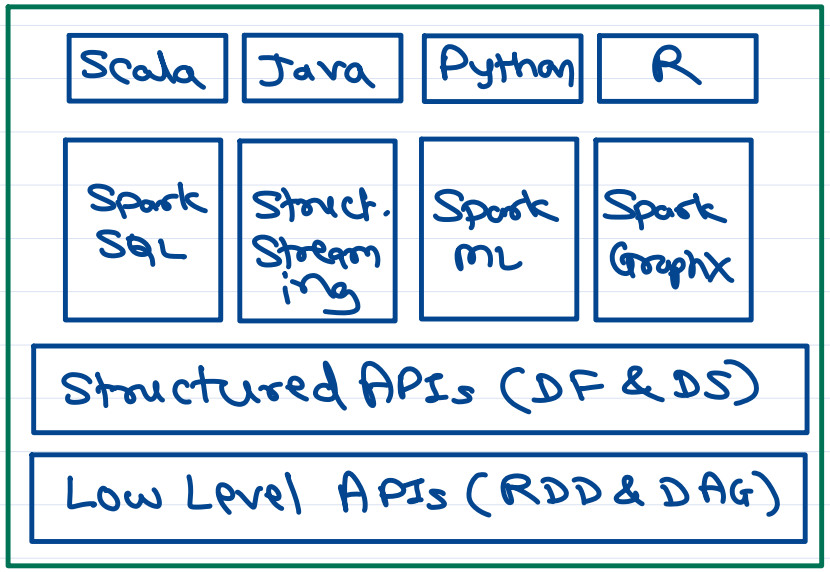
Big Data Technologies

Trainer: Mr. Nilesh Ghule.

Spark





Spark 2.x+ toolkit

pyspark

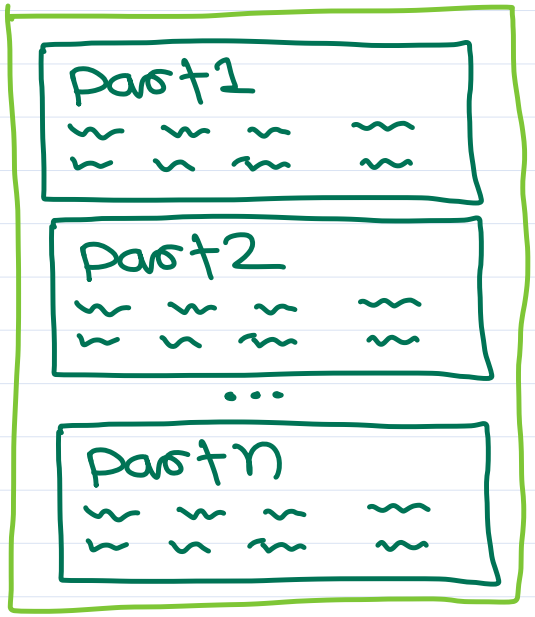
- ✓ Spark programming in Python.
- ✓ pyspark - pip pkg for Spark dev.

Spark-R

- ✓ Spark prog in R.

Spark

DataFrame → abstraction over RDD.

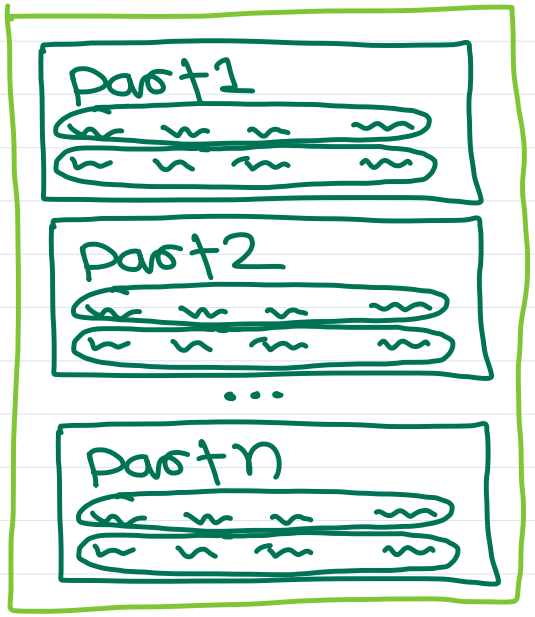


RDD → Functional programming.

DF → SQL style programming (Rows & cols)

Ds → collection of immutable objects. Available only in Scala (not in Python).

DataSet



RDD

Spark Vendors

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

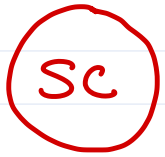
Spark Philosophy

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

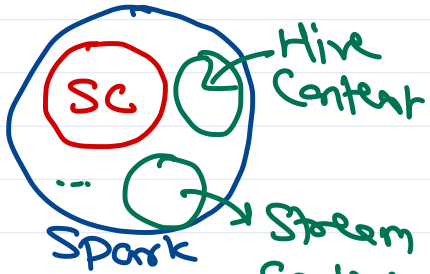
Spark

SparkContext: Main obj in Spark apphn.
- Responsible for RDD creation & Job execution (DAG)
- Web UI (port 4040).

Low level API.



SessionSession: Main obj in Spark apphn.
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 apphn.



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

For standalone apphns, they must be created at the start of apphn.

DataFrameReader

```
df = spark.read
    .option("header", True)
    .option("inferSchema", True)
    .csv("/path/of/csv")
```

Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>