Python Biella Group

PySpark on Databricks

Data engineering with Python on large scale

a cura di: **Marco Santoni**

flowe

ABOUT ME

Work @ Flowe Data Platform

intervista
pythonista

Co-host of Intervista Pythonista

HOW DID WE GET HERE

flowe

3V (2001)

SPARK (2009)

DWH (late 90s)

HADOOP (2005)

CLOUD

(BIG QUERY 2011, SNOWFLAKE 2012, DATABRICKS 2013)

SPARK

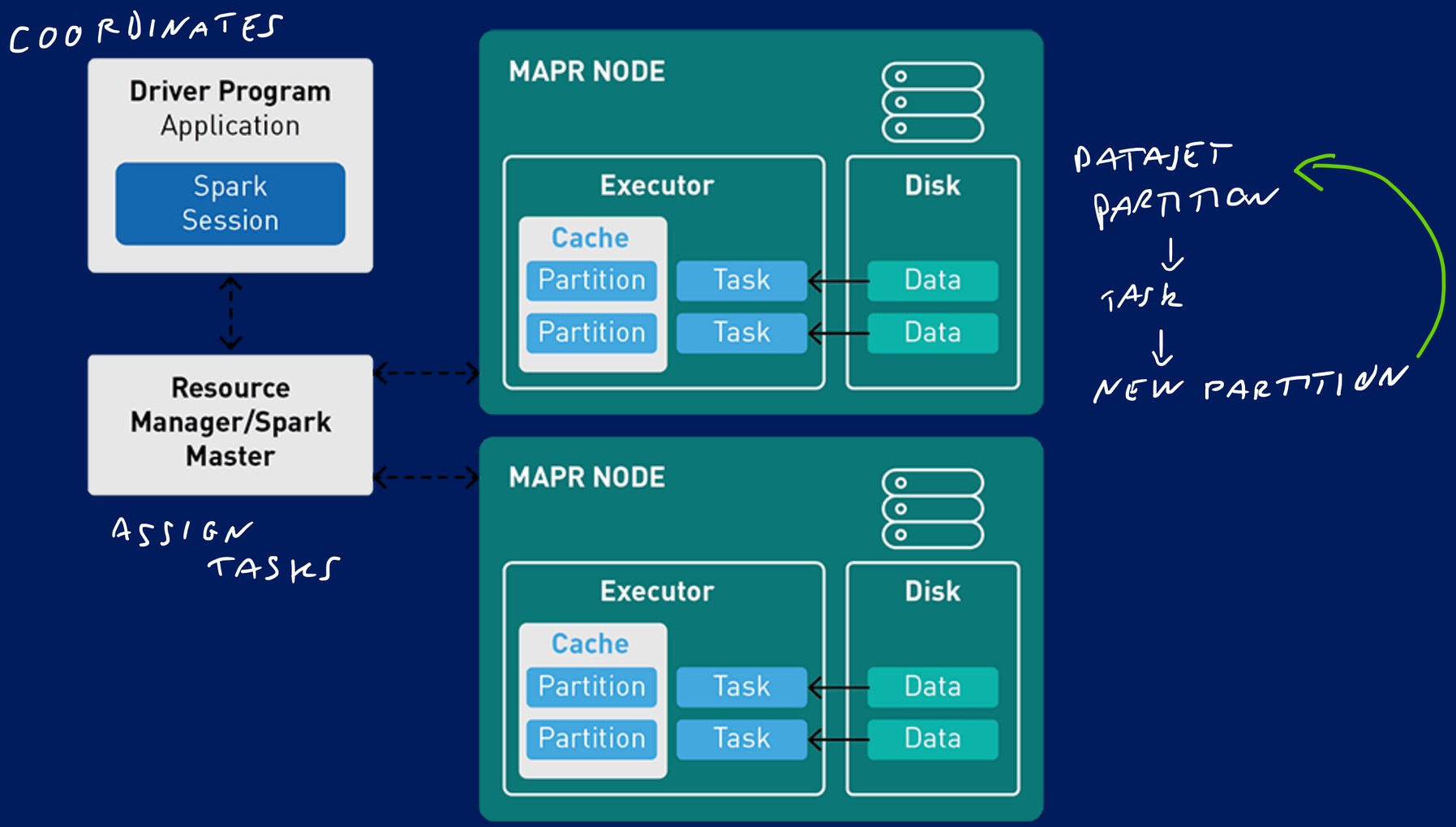

Image from Spark 101

WHY SPARK

flowe

Simplicity

Speed

Support

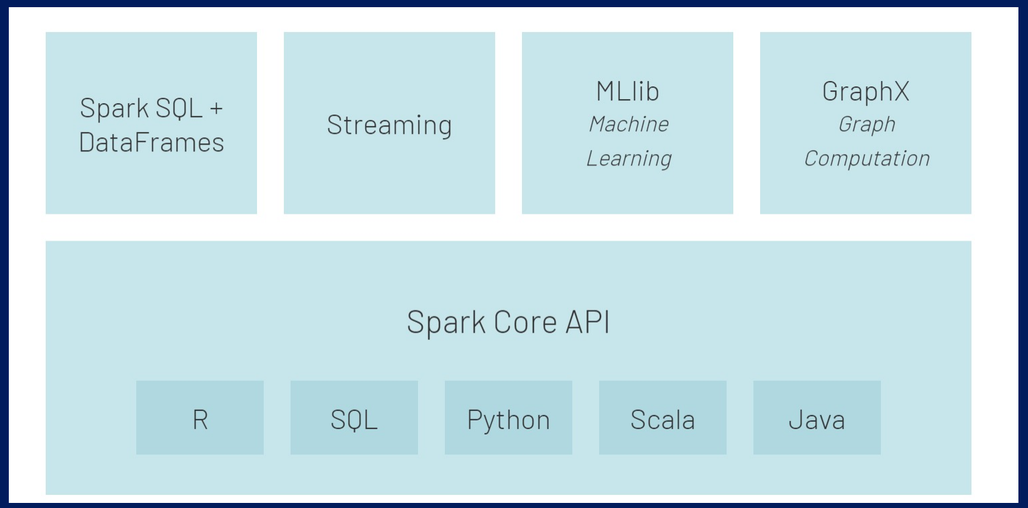

Image from Gettting Started with Spark

WHY DATABRICKS

flowe

By creators of Apache Spark™

Easy to setup clusters

One platform

Image from imgflip

DEMO

flowe

Databricks workspace

Reading, transforming and writing

Querying and analyzing

Scheduling jobs

REFERENCES

flowe

Data lakes history

Spark 101

Getting starte with Apacke Spark

Demo notebook