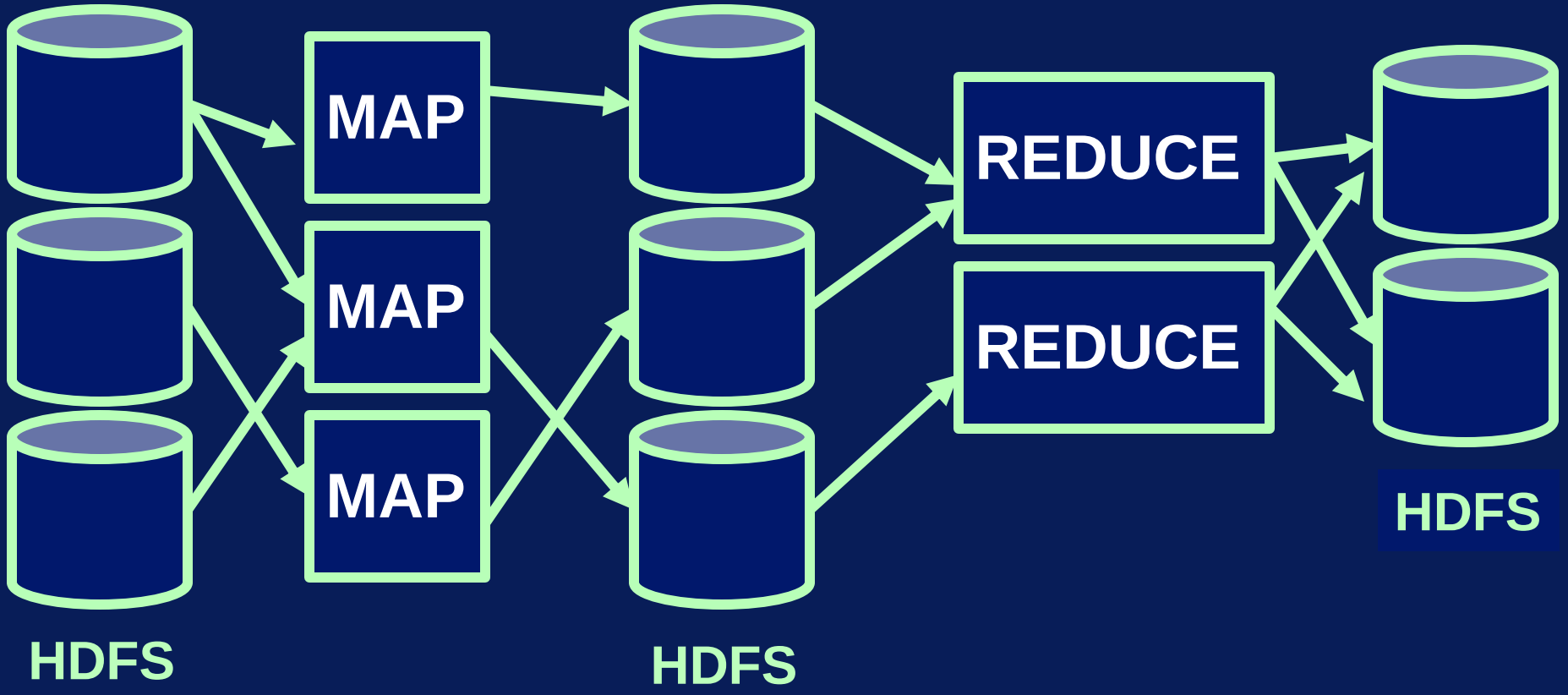
Getting Started with Spark: The Architecture and Basic Concepts

After this video you will be able to..

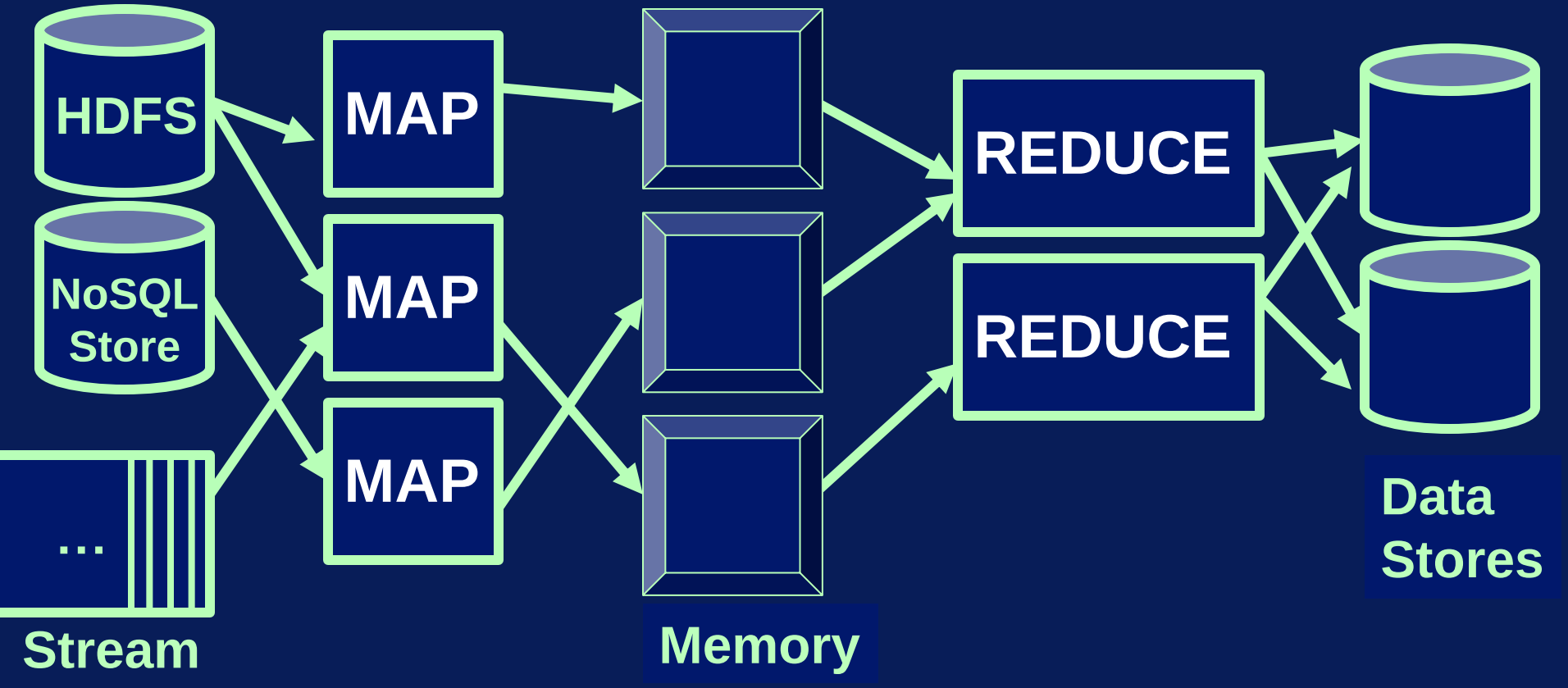
- Describe how Spark does in-memory processing using the RDD abstraction
- Explain the inner workings of the Spark architecture
- Summarize how Spark manages and executes code on Clusters

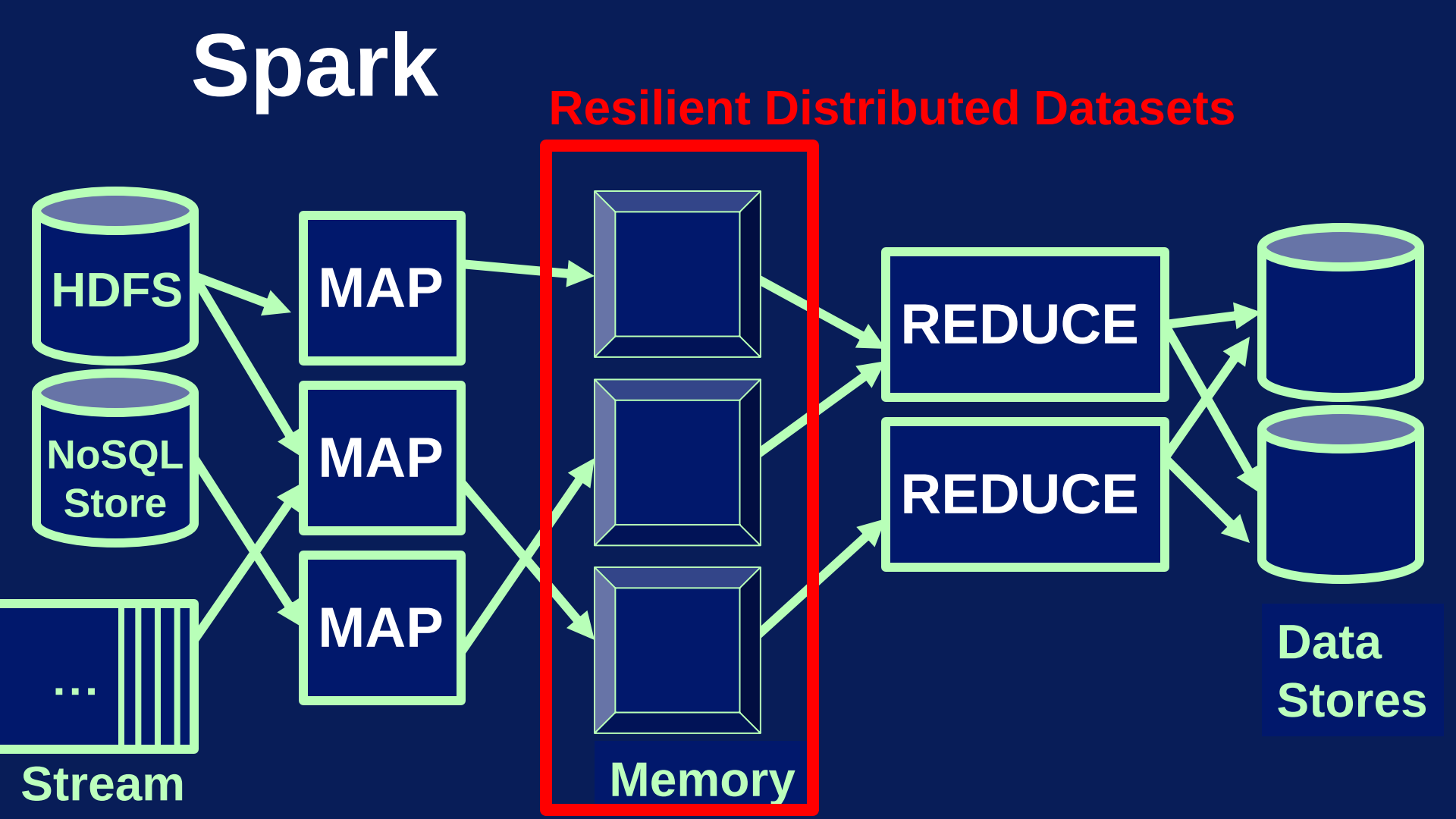
What does in memory processing mean?

MapReduce



Spark


HDFS
NoSQL Store
...
Stream
MAP
MAP
MAP
Memory
REDUCE
REDUCE
Data Stores


Spark
Resilient Distributed Datasets
HDFS
NoSQL Store
...
Stream
MAP
MAP
MAP
Memory
REDUCE
REDUCE
Data Stores

Resilient Distributed Datasets

Dataset

Data storage created from: HDFS, S3, HBase, JSON, text, Local hierarchy of folders

Or created transforming another RDD

Resilient Distributed Datasets

Distributed

Distributed across the cluster of machines

Divided in partitions, atomic chunks of data

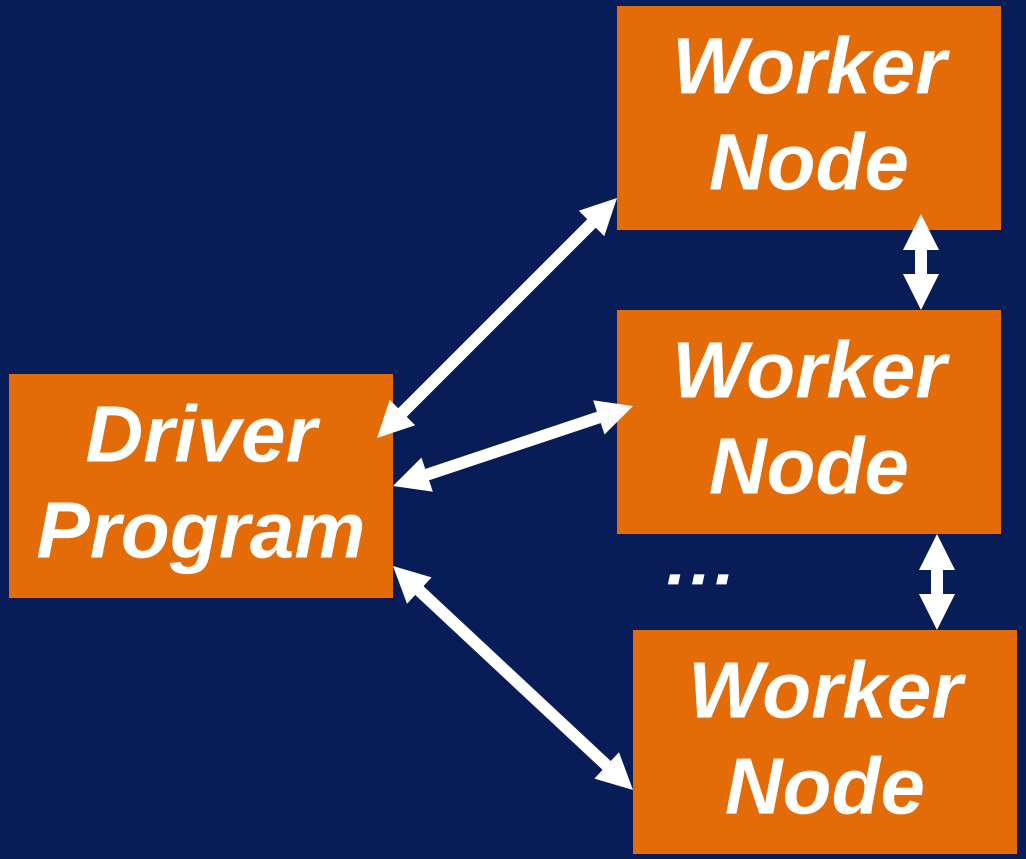
Resilient Distributed Datasets

Resilient

Recover from errors, e.g. node failure, slow processes

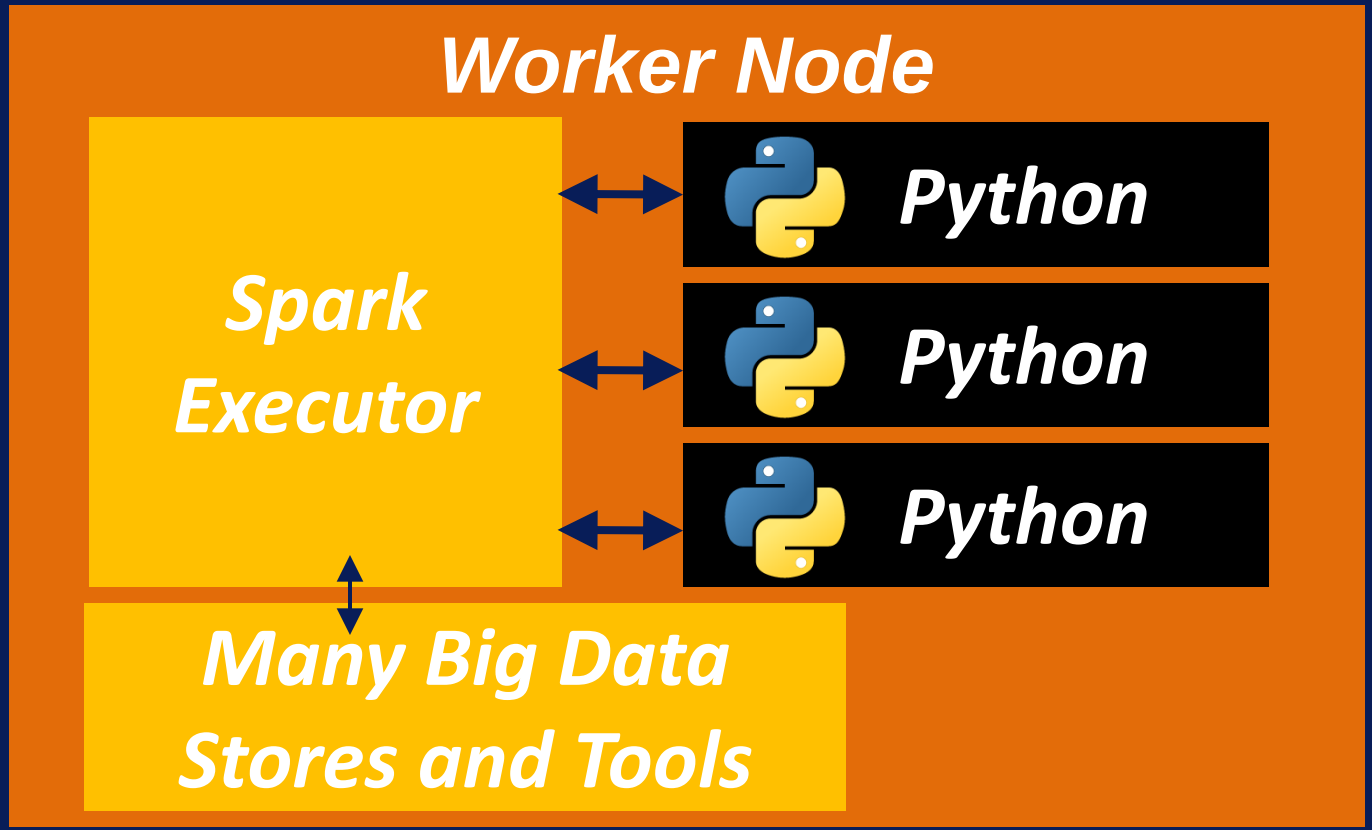
Track history of each partition, re-run

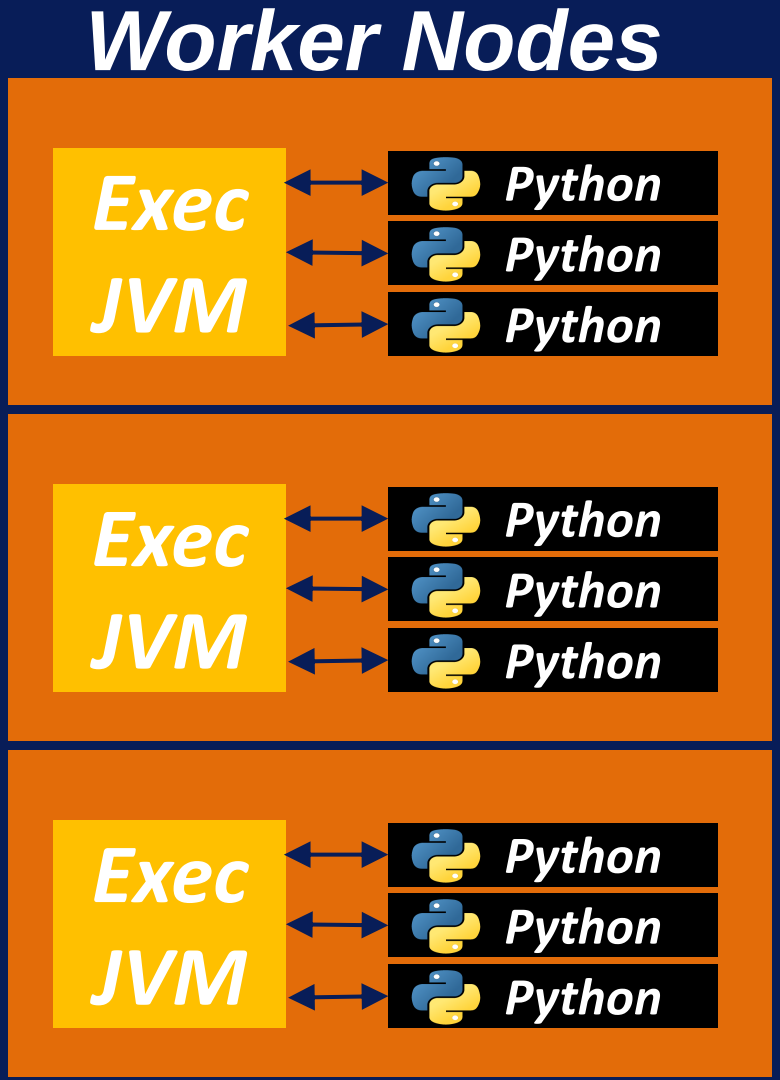
Spark Architecture

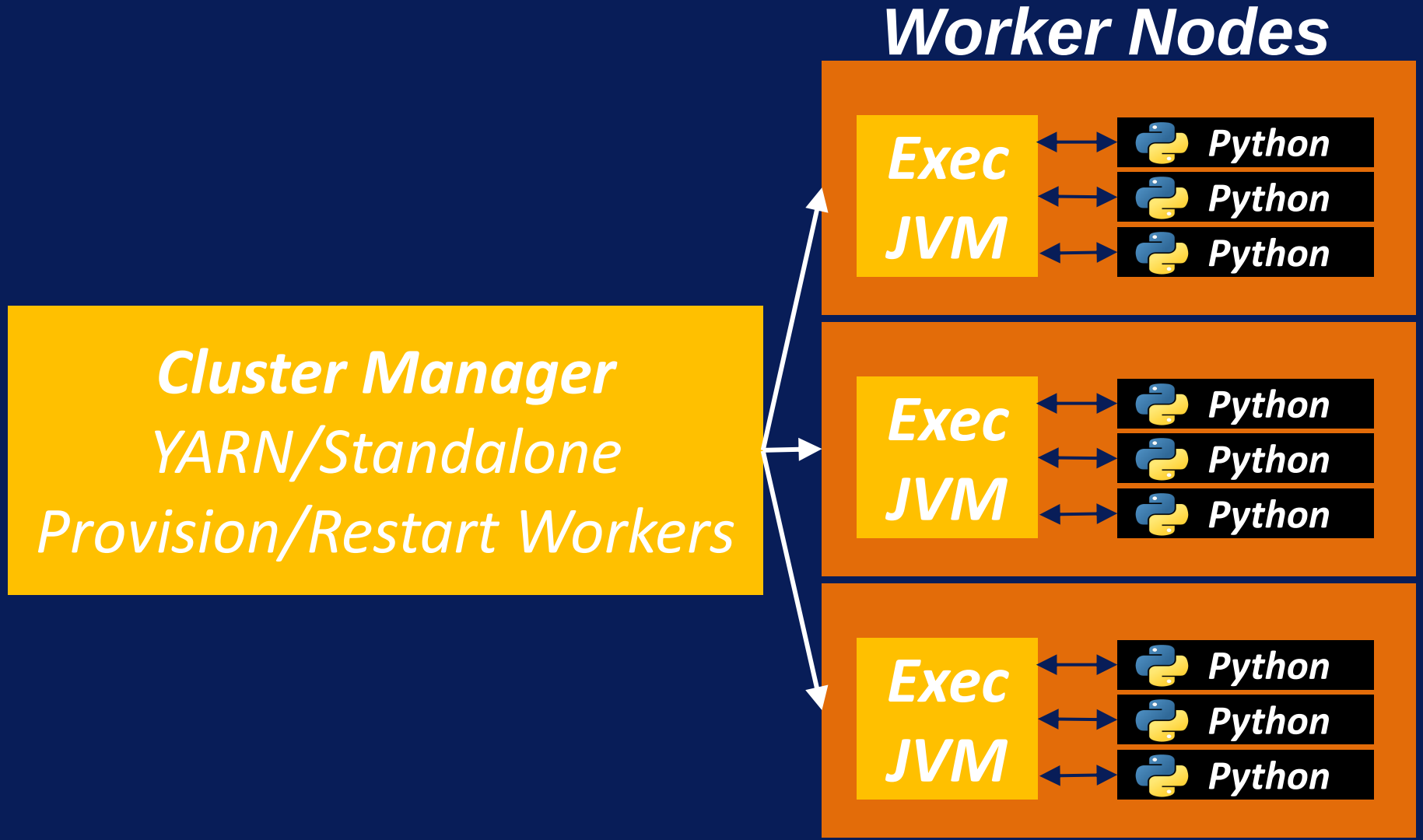

Worker Node
Worker Node
...
Worker Node
Driver Program

Driver Program

```
In [1]: lines = sc.textFile("hdfs:/user/cloudera/words.txt")
```

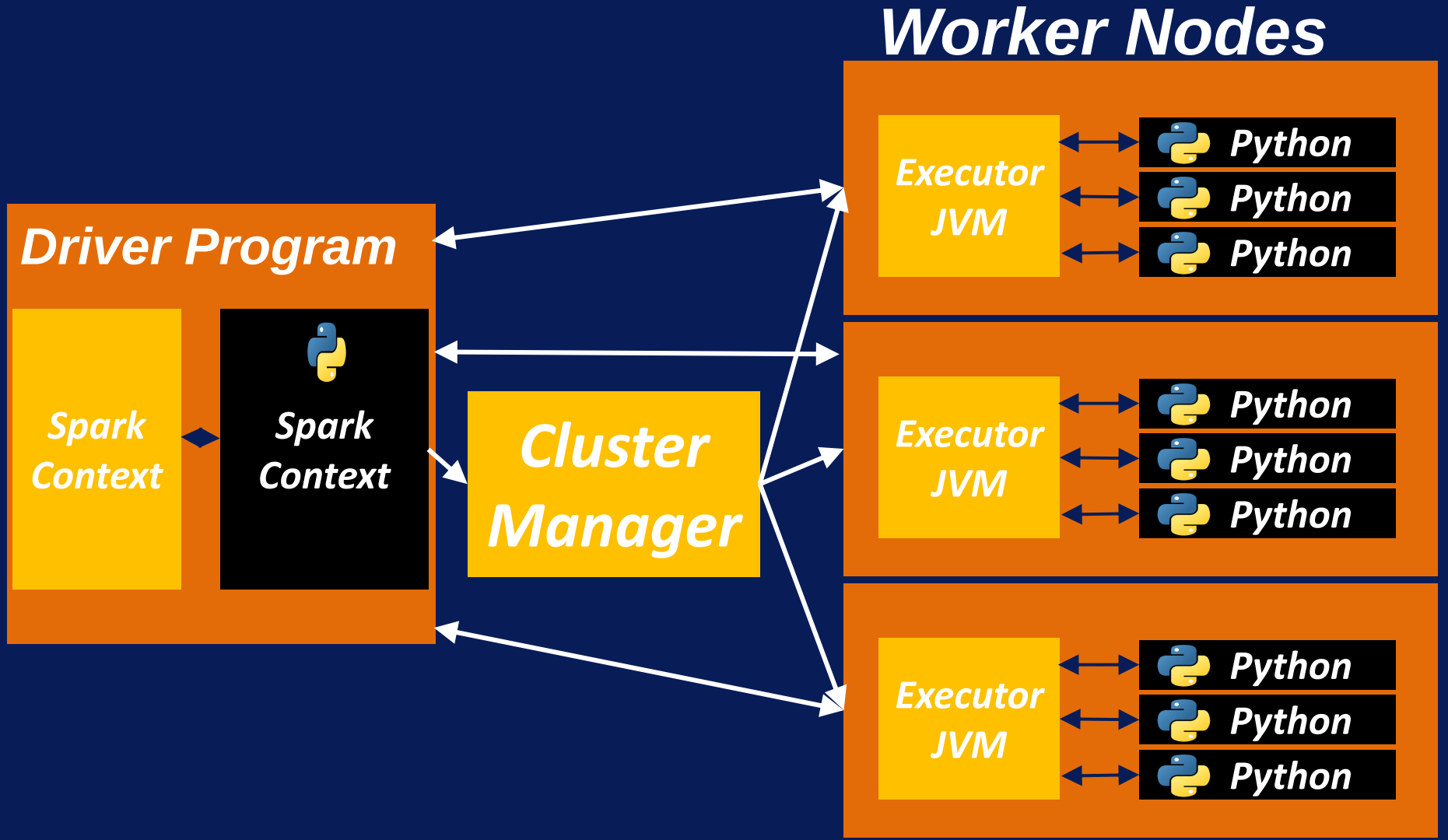

Worker Node
Spark Executor
Python
Python
Python
Many Big Data Stores and Tools


Worker Nodes
Exec
JVM
Python
Python
Python
Exec
JVM
Python
Python
Python
Exec
JVM
Python
Python
Python

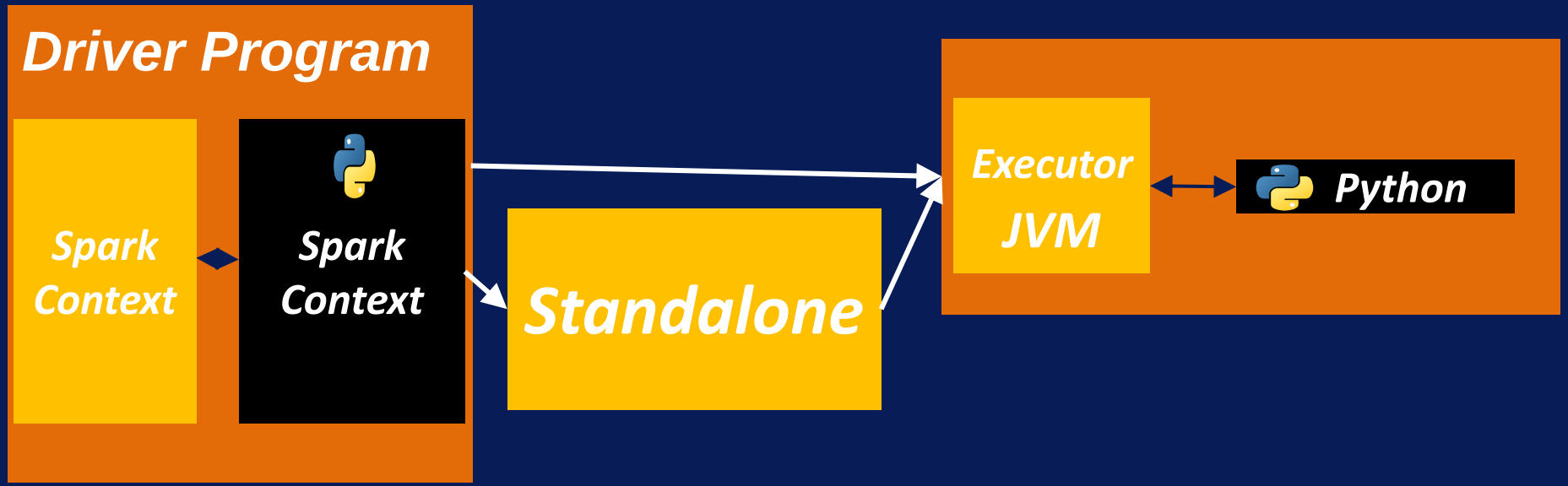

Worker Nodes
Cluster Manager
YARN/Standalone
Provision/Restart Workers
Exec JVM
Python
Python
Python
Exec JVM
Python
Python
Python
Exec JVM
Python
Python
Python

Which cluster manager?

http://www.agildata.com/apache-spark-cluster-managers-yarn-mesos-or-standalone/


Worker Nodes
Driver Program
Spark Context
Spark Context
Cluster Manager
Executor JVM
Python
Python
Python
Executor JVM
Python
Python
Python
Executor JVM
Python
Python
Python

Cloudera VM


Driver Program
Spark Context
Spark Context
Standalone
Executor JVM
Python