Module II:
Toolset Introduction

On ApacheSpark, NoSQL, ObjectStorage and the rest…

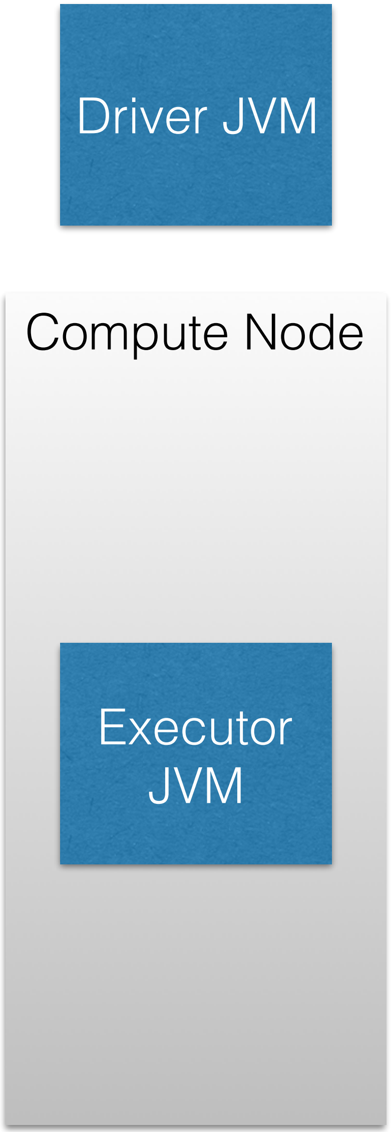
In this Video you will learn…

Introduction to ApacheSpark

JVM
Process

Driver JVM

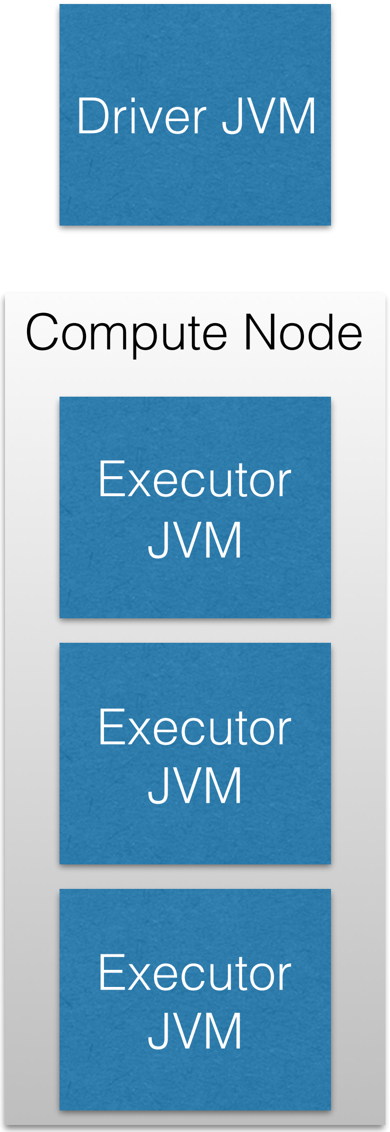
Compute Node

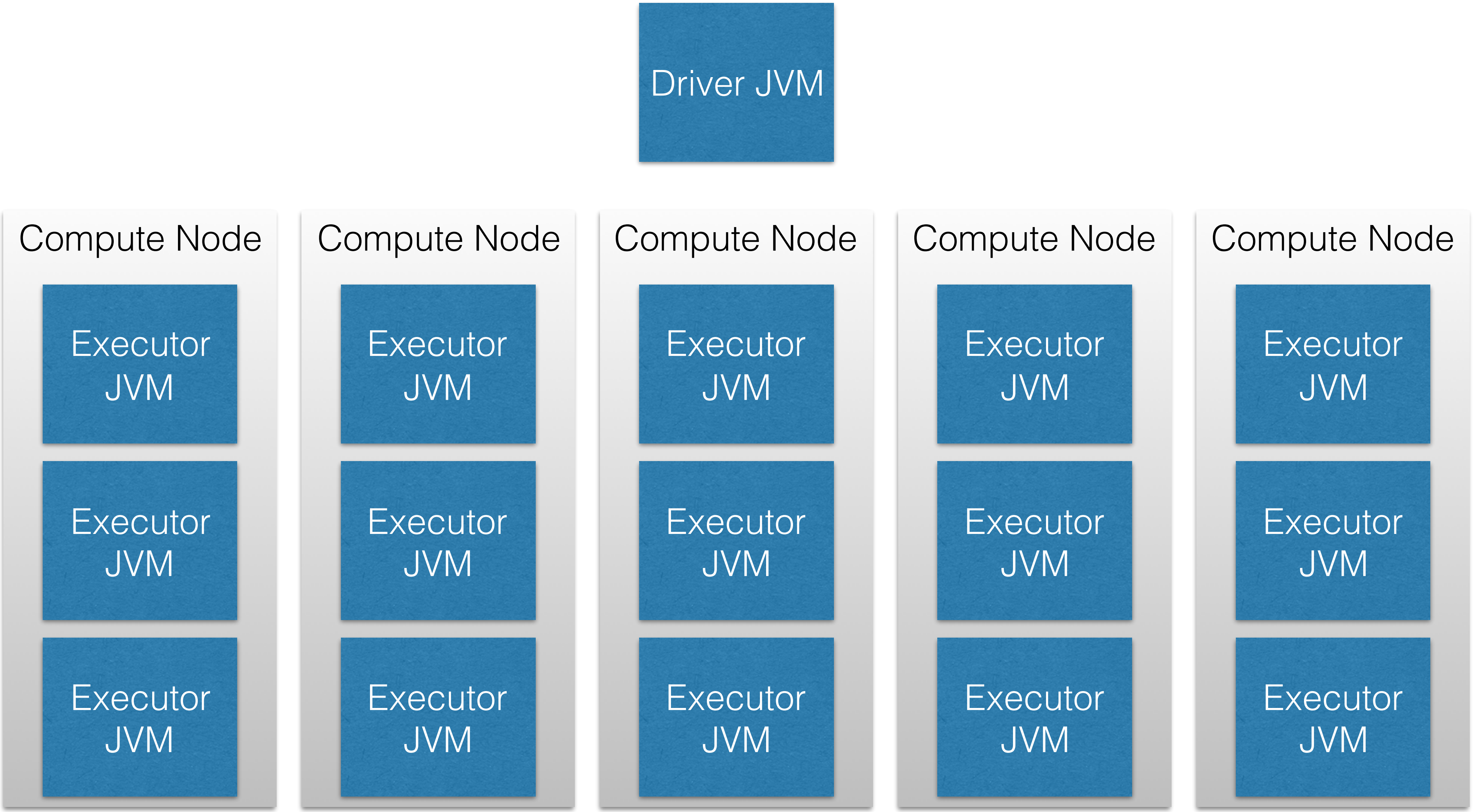


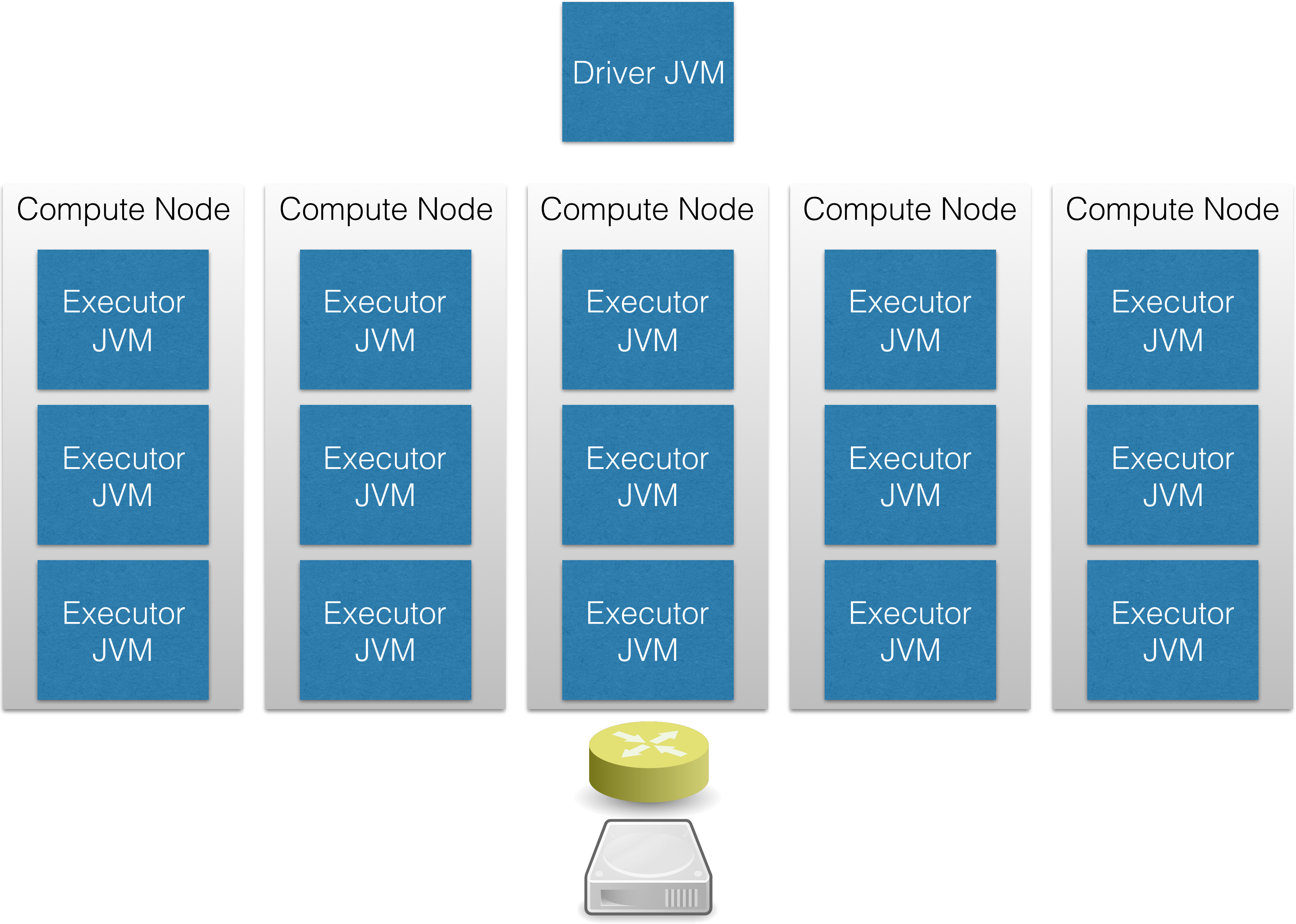
Driver JVM

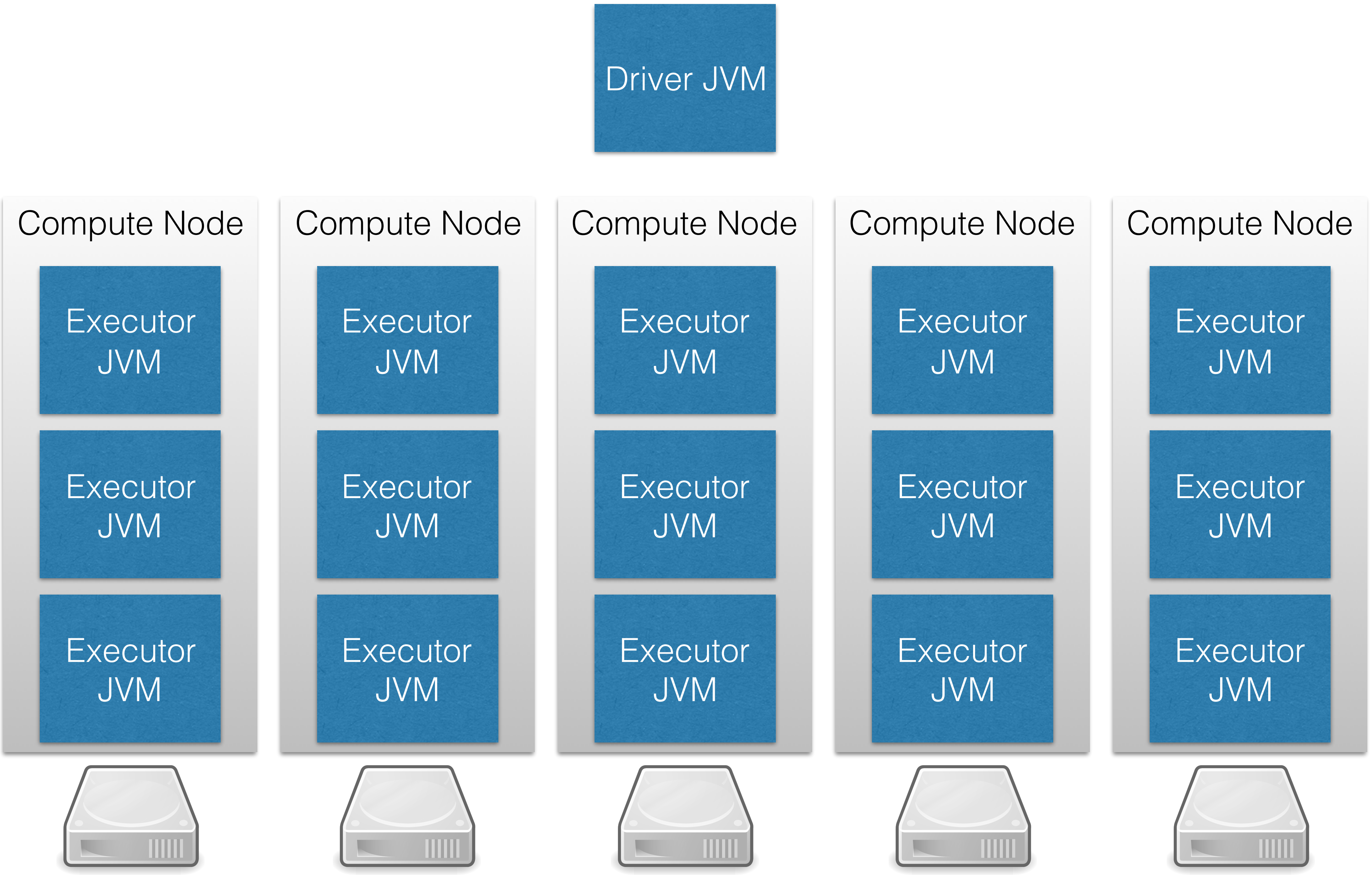
Compute Node

Executor JVM


Driver JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM

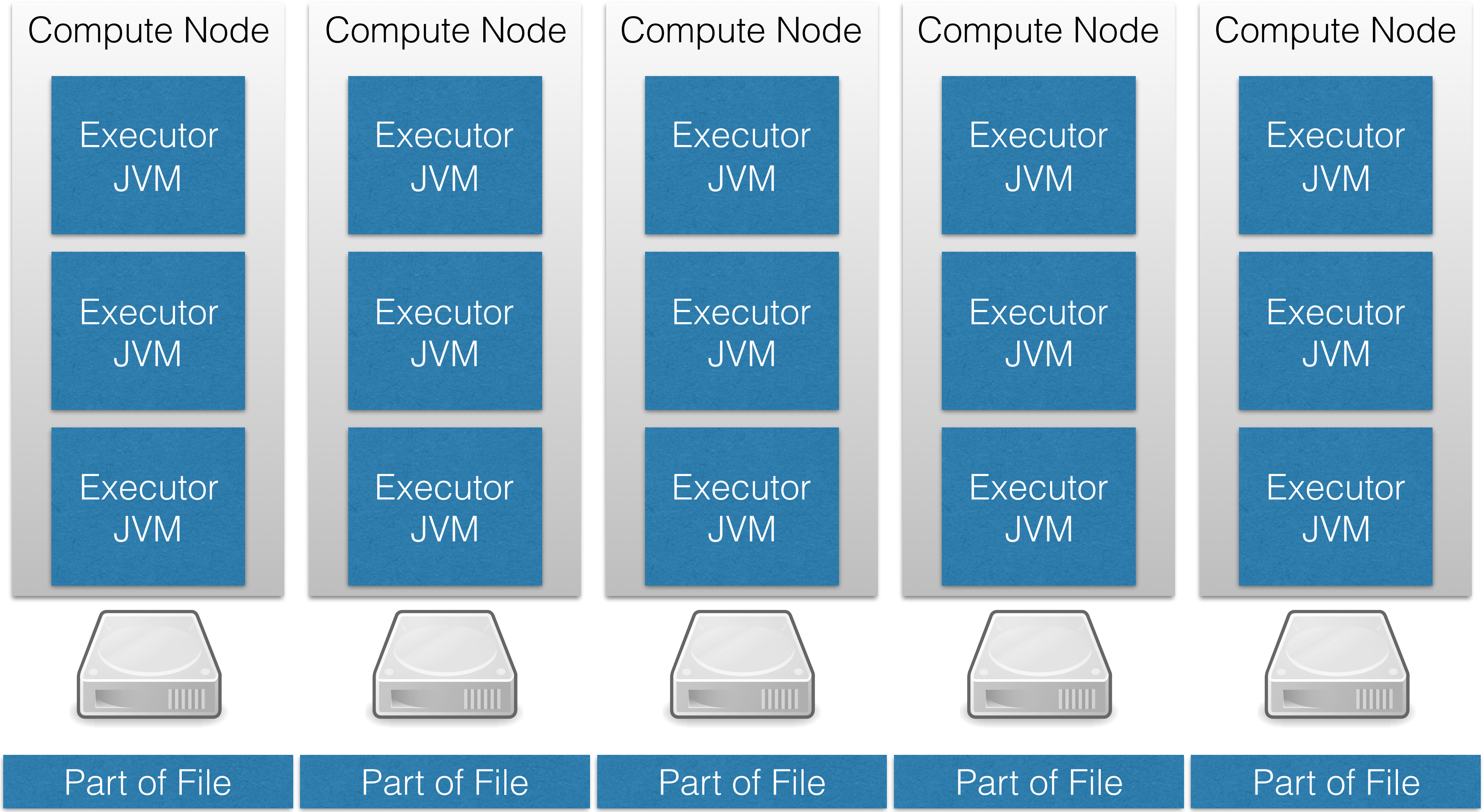

Driver JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM


Driver JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM



Driver JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM

HDFS



Compute Node

Executor JVM

Executor JVM

Executor JVM

Part of File

Compute Node

Executor JVM

Executor JVM

Executor JVM

Part of File

Compute Node

Executor JVM

Executor JVM

Executor JVM

Part of File

Compute Node

Executor JVM

Executor JVM

Executor JVM

Part of File

Compute Node

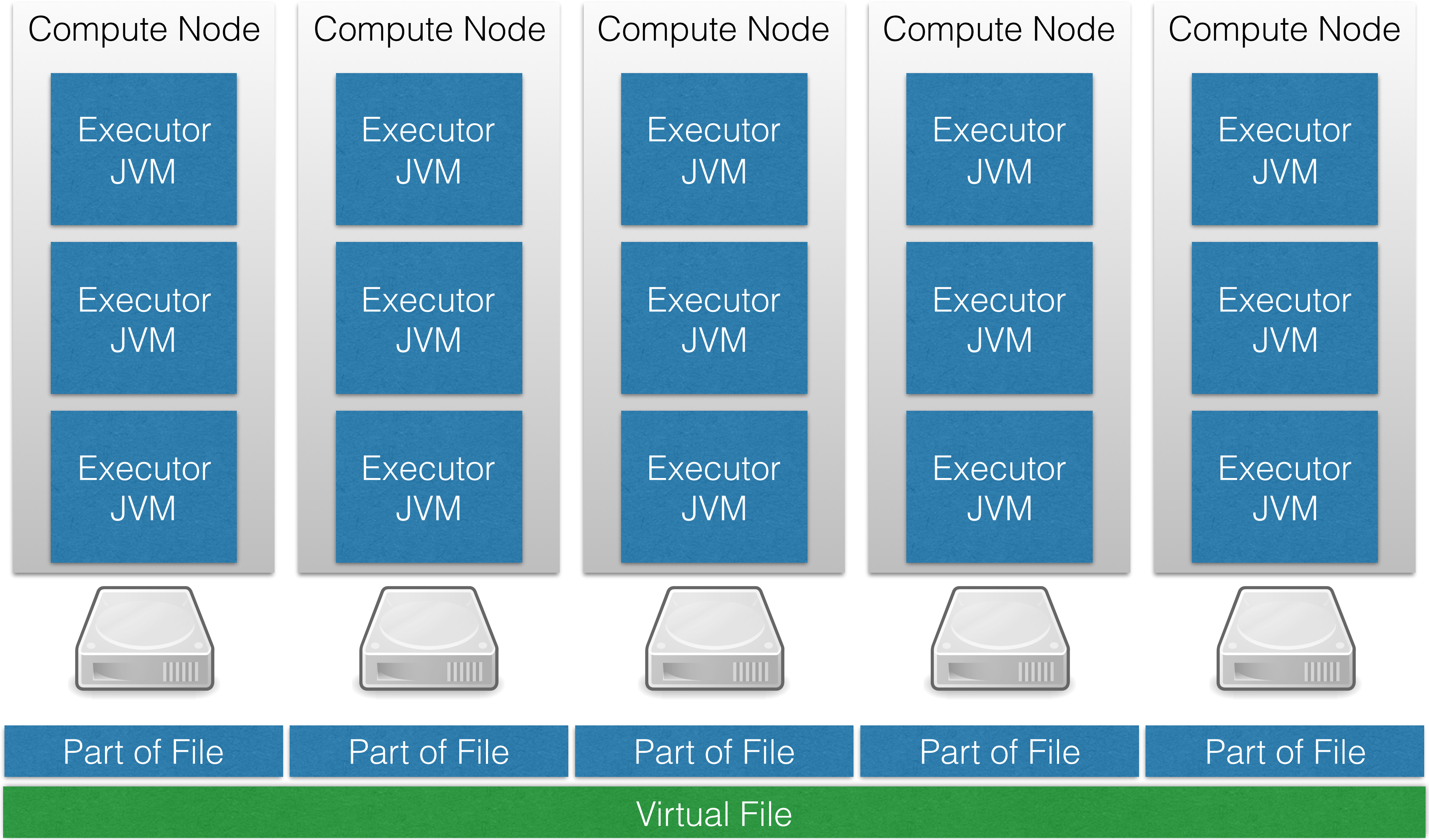
Executor JVM

Executor JVM

Executor JVM

Part of File

HDFS



Compute Node
Executor JVM
Executor JVM
Executor JVM

Compute Node
Executor JVM
Executor JVM
Executor JVM

Compute Node
Executor JVM
Executor JVM
Executor JVM

Compute Node
Executor JVM
Executor JVM
Executor JVM

Compute Node
Executor JVM
Executor JVM
Executor JVM

Part of File
Part of File
Part of File
Part of File
Part of File

Virtual File

RDD

- “Resilient Distributed Dataset“

- Distributed, immutable collection of data

- Created from HDFS, ObjectStore, Cloudant NoSQL, dashDB SQL, simple files, …

- In-memory, but spillable to disk

- lazy

Summary

- ApacheSpark programs are implicitly parallel
- Same code can process 1 KB or 1 PB
- RDD central API
- Data and task distribution transparent

Quiz

- What is the main advantage in implementing data analysis workflows using the RDD API?

  - RDD has functions for data analysis no other framework provides
    False: Similar functions can be found in the Pandas Data Frame for example

  - If you use the RDD API your code implicitly is getting executed in parallel on the ApacheSpark cluster
    Correct

  - RDD has been invented because the expressiveness of SQL was not sufficient
    False: Although nearly everything can be expressed using the RDD API the expressiveness of SQL is also very high and usually sufficient for data analysis tasks

The next Video covers…

Programming Language Options