OUT OF SYLLABUS

Production-Grade Edition

A NOTE TO JUNIORS

You Don't have to be an 'Expert' in one thing.

Expand your knowledge in as many areas as possible and understand how things work. Adaptability is the key.

Don't by-heart commands / programs.

You are an Engineer, you just have to know how to put things together to make them work for a bigger purpose.

A BASIC DIFFRENCE BETWEEN COLLEGE LIFE AND CORPORATE LIFE

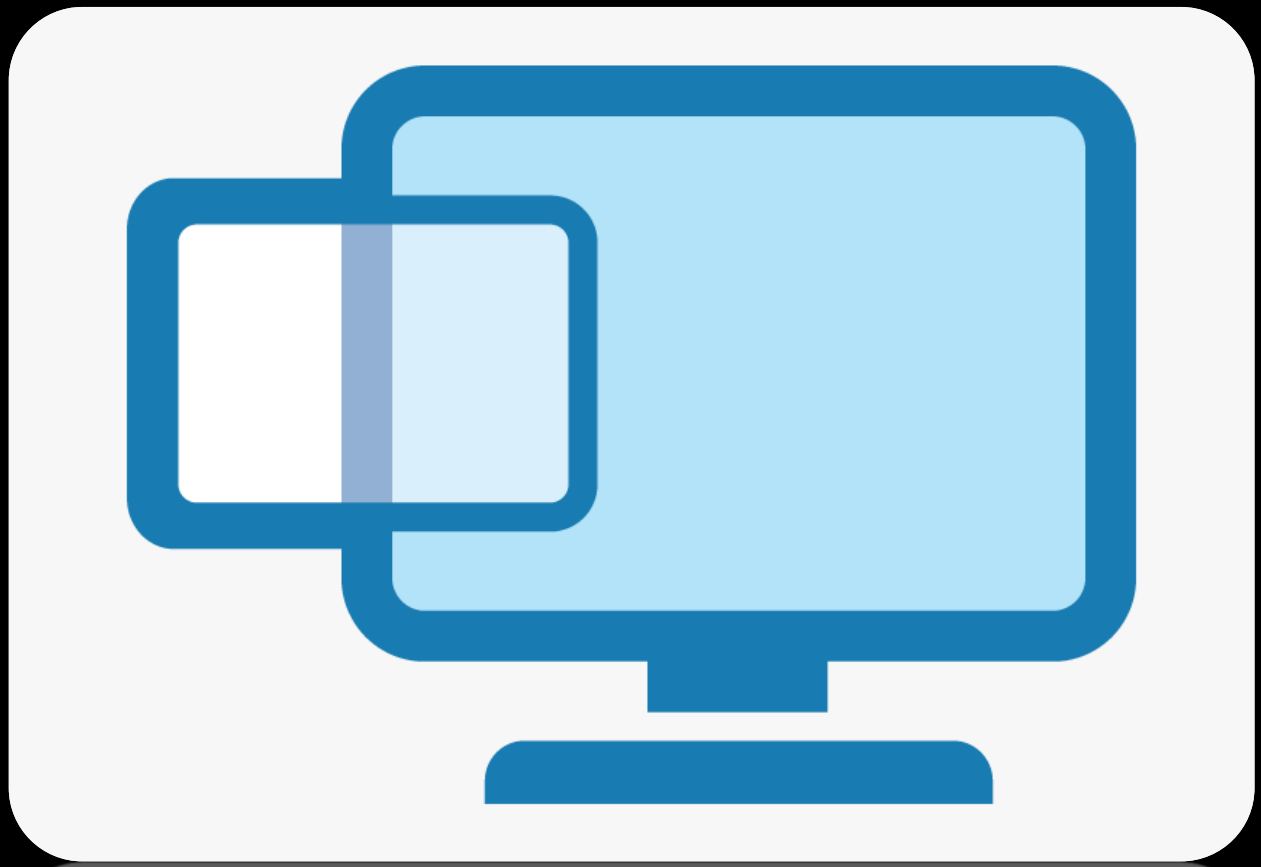
- In College, you think that your job is to know the output of a code snippet.
- In company, you will realize that it is the job of the compiler.

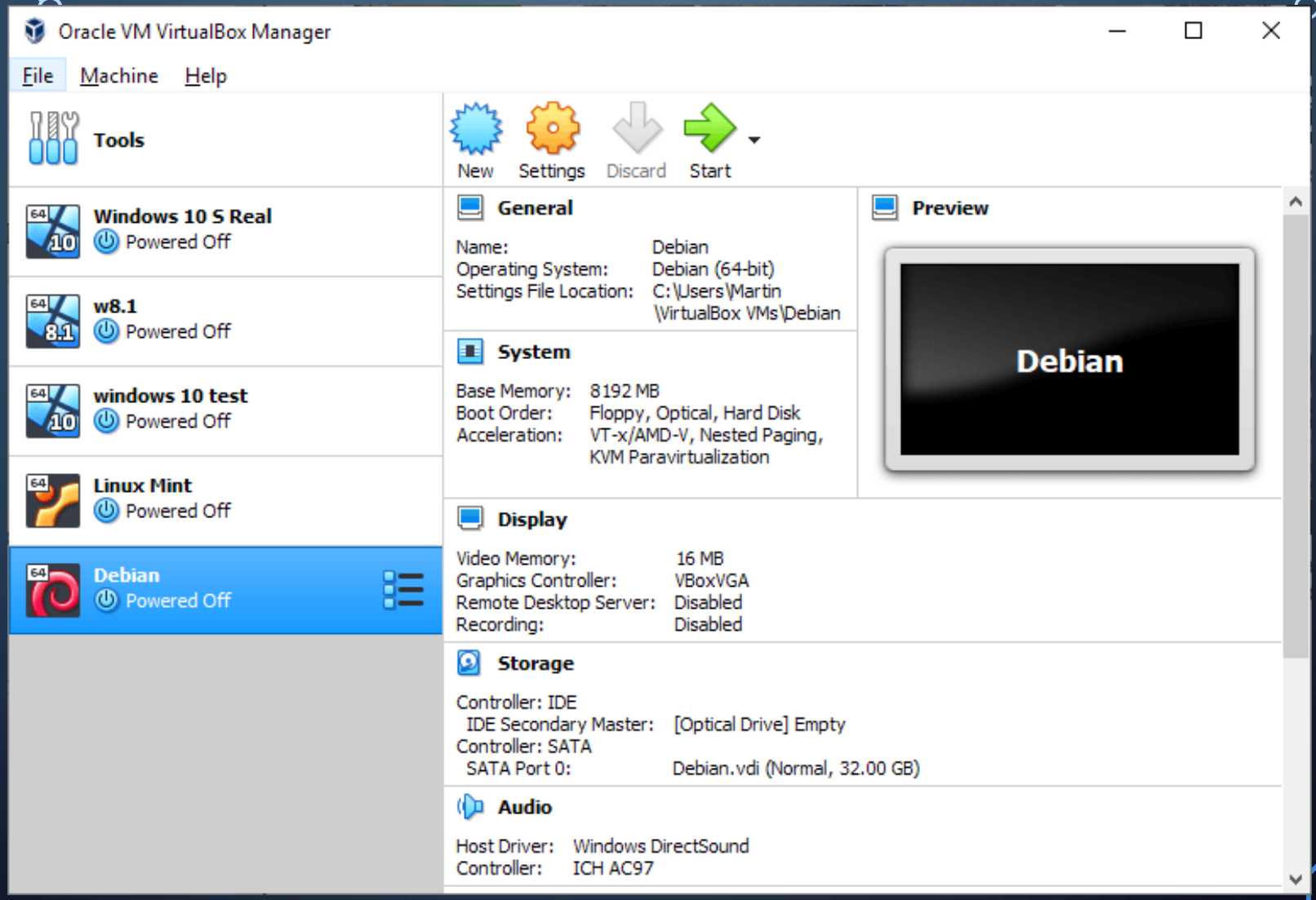
DEMYSTIFYING : CONTAINERIZATION

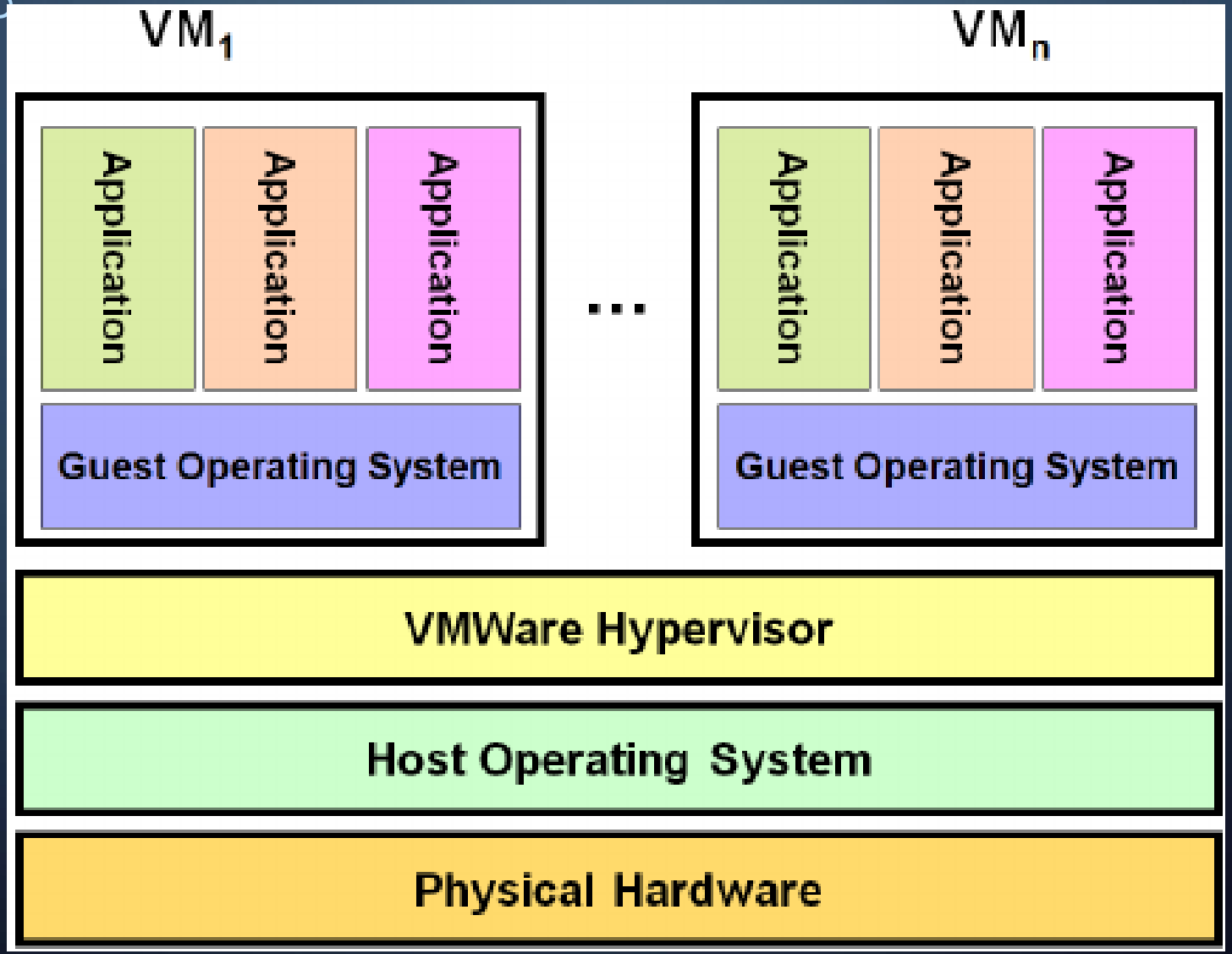
By Chandan Shastri

LET'S START FROM YOUR MINI PROJECTS

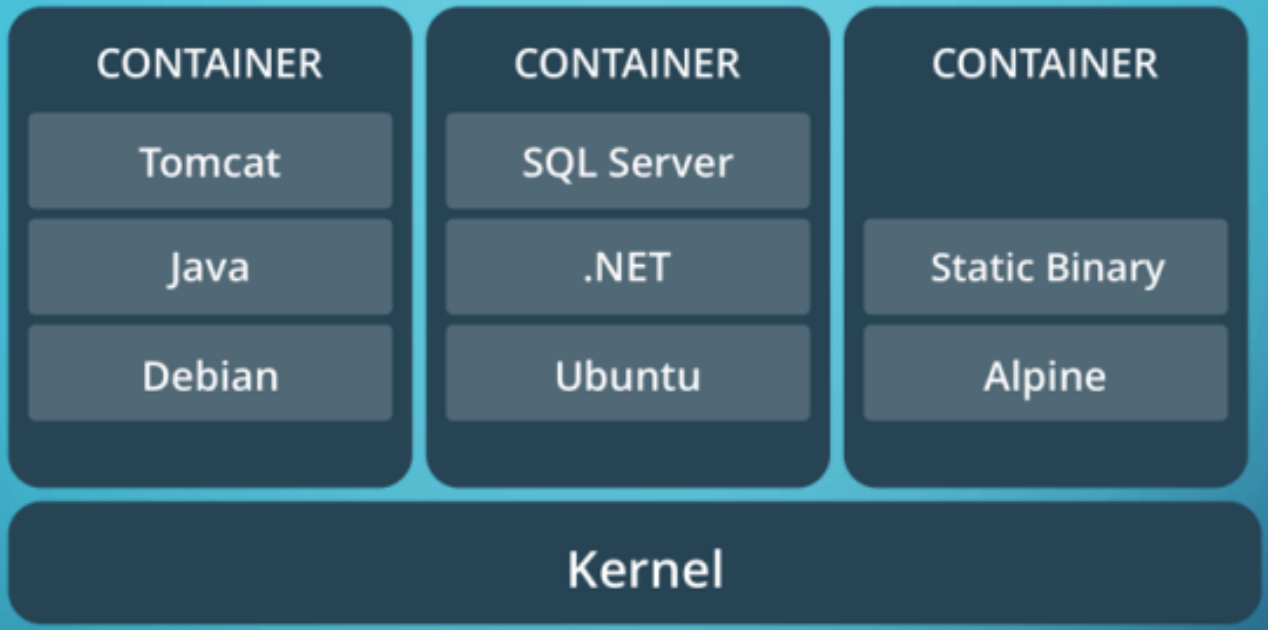
VIRTUAL MACHINES

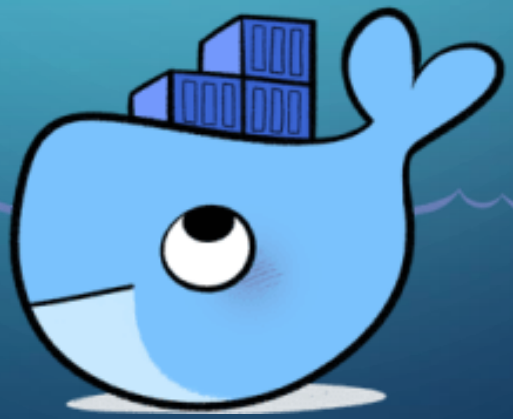



Oracle VM VirtualBox Manager
File
Machine
Help
Tools
Windows 10 S Real
Powered Off
w8.1
Powered Off
windows 10 test
Powered Off
Linux Mint
Powered Off
Debian
Powered Off
New
Settings
Discard
Start
General
Name: Debian
Operating System: Debian (64-bit)
Settings File Location: C:\Users\Martin\VirtualBox VMs\Debian
System
Base Memory: 8192 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization
Preview
Debian
Display
Video Memory: 16 MB
Graphics Controller: VBoxVGA
Remote Desktop Server: Disabled
Recording: Disabled
Storage
Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: Debian.vdi (Normal, 32.00 GB)
Audio
Host Driver: Windows DirectSound
Controller: ICH AC97

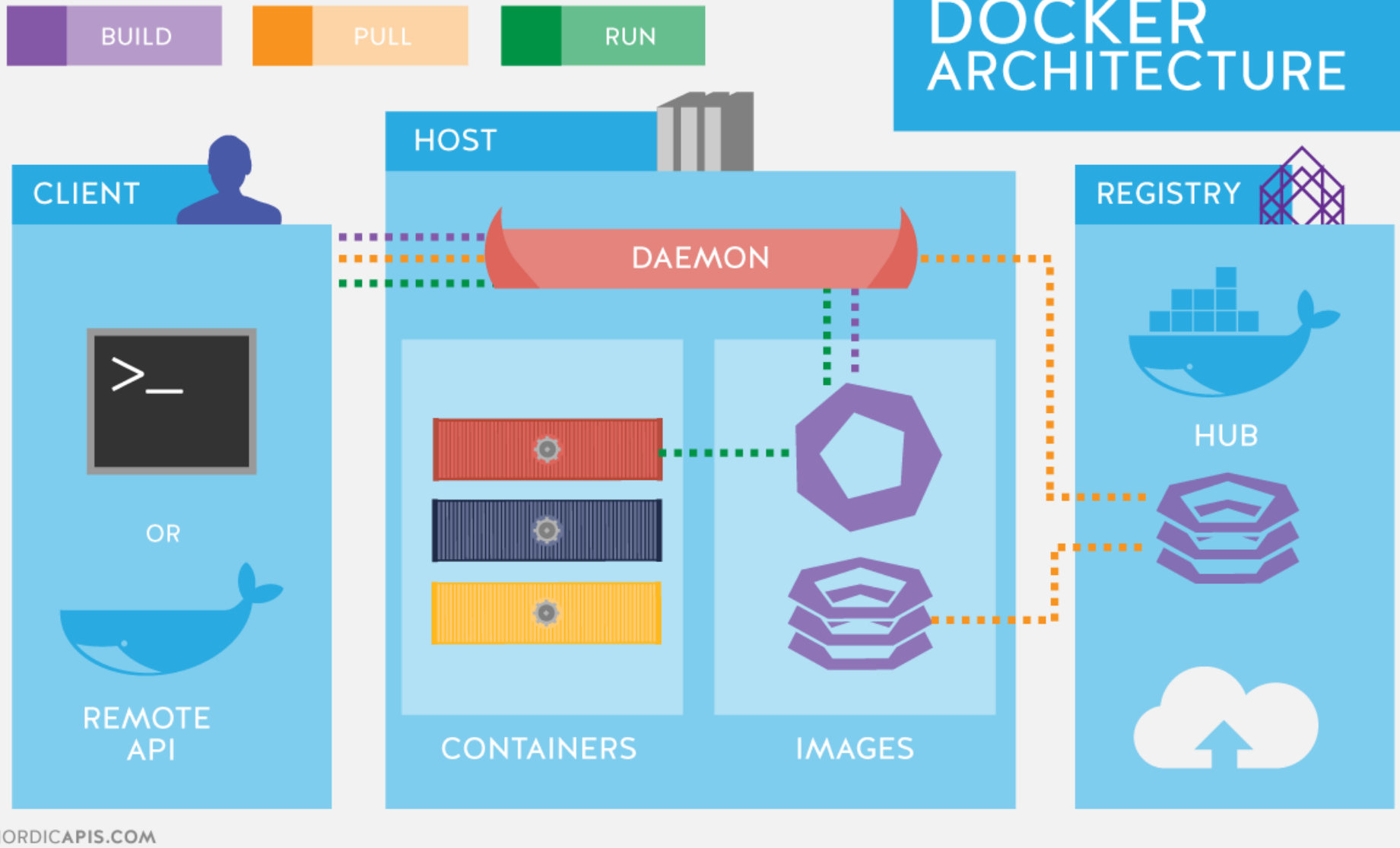

VM1
VMn
Application
Application
Application
Guest Operating System
...
Application
Application
Application
Guest Operating System
VMWare Hypervisor
Host Operating System
Physical Hardware


docker

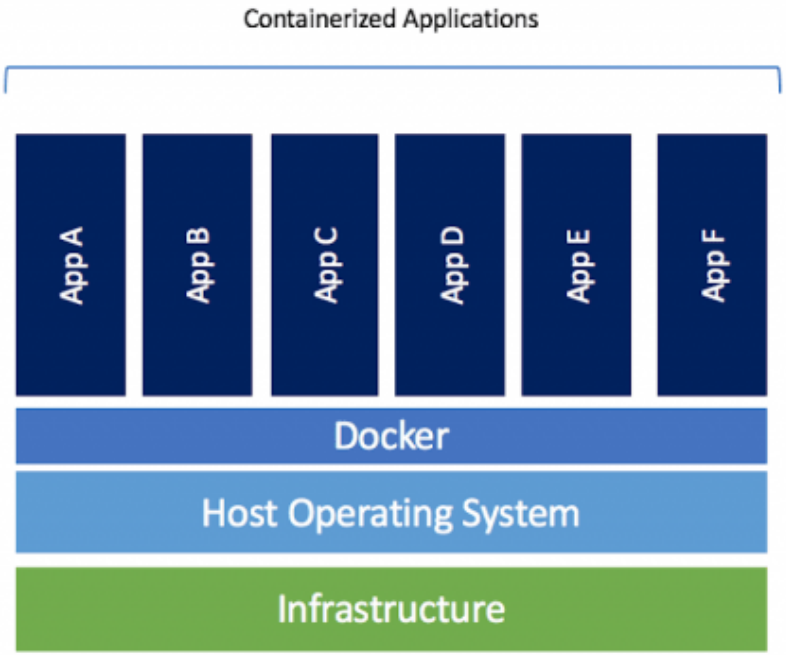

CONTAINER
Tomcat
Java
Debian
CONTAINER
SQL Server
.NET
Ubuntu
CONTAINER
Static Binary
Alpine
Kernel

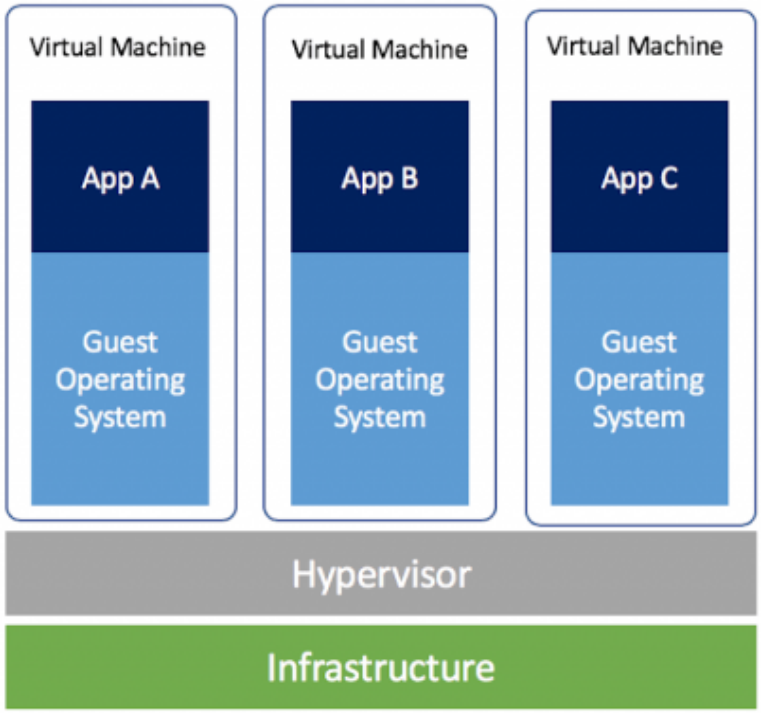


DOCKER ARCHITECTURE



NORDICAPIS.COM

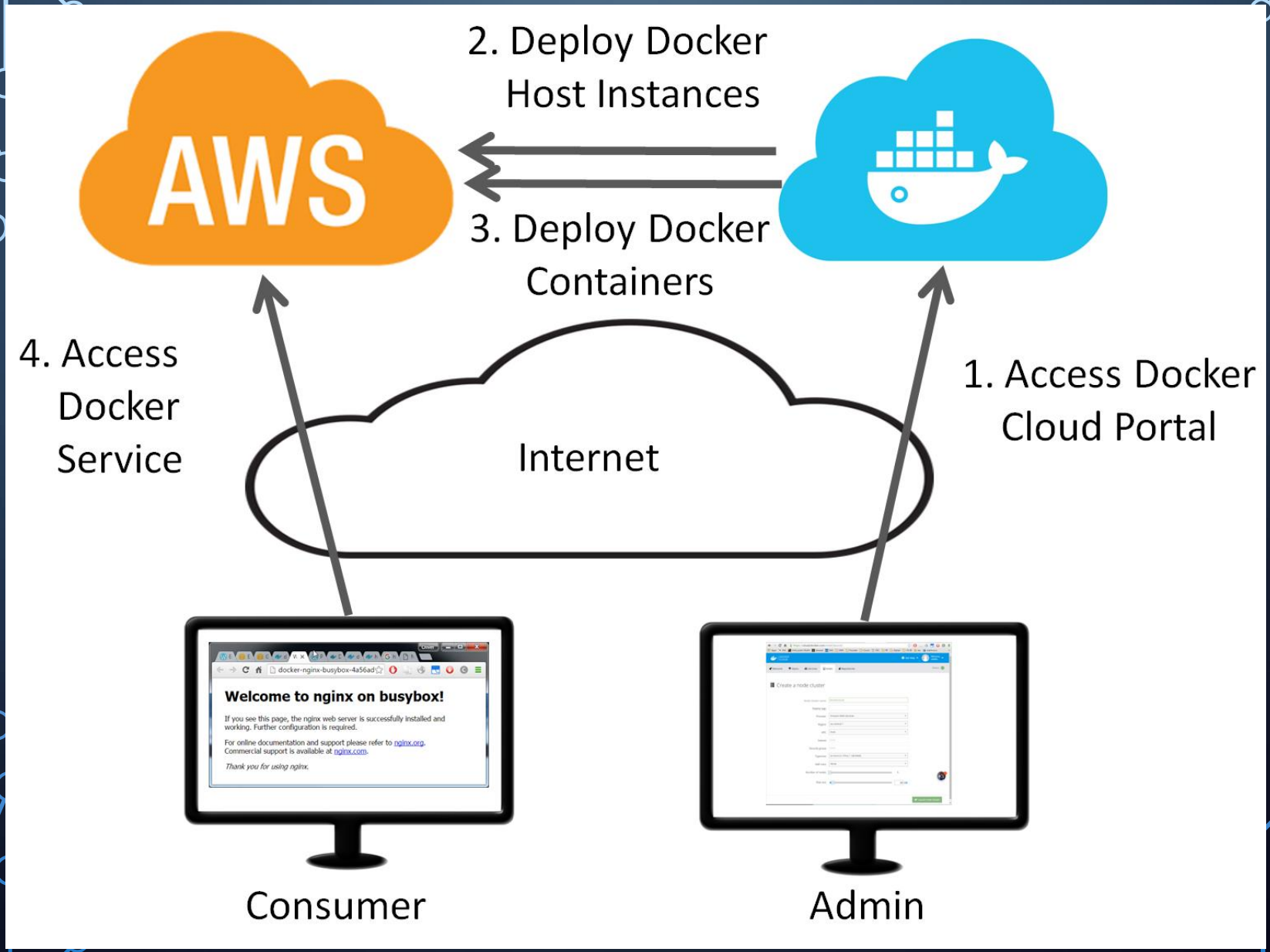

Containerized Applications
App A
App B
App C
App D
App E
App F
Docker
Host Operating System
Infrastructure


Virtual Machine
App A
Guest Operating System
Virtual Machine
App B
Guest Operating System
Virtual Machine
App C
Guest Operating System
Hypervisor
Infrastructure

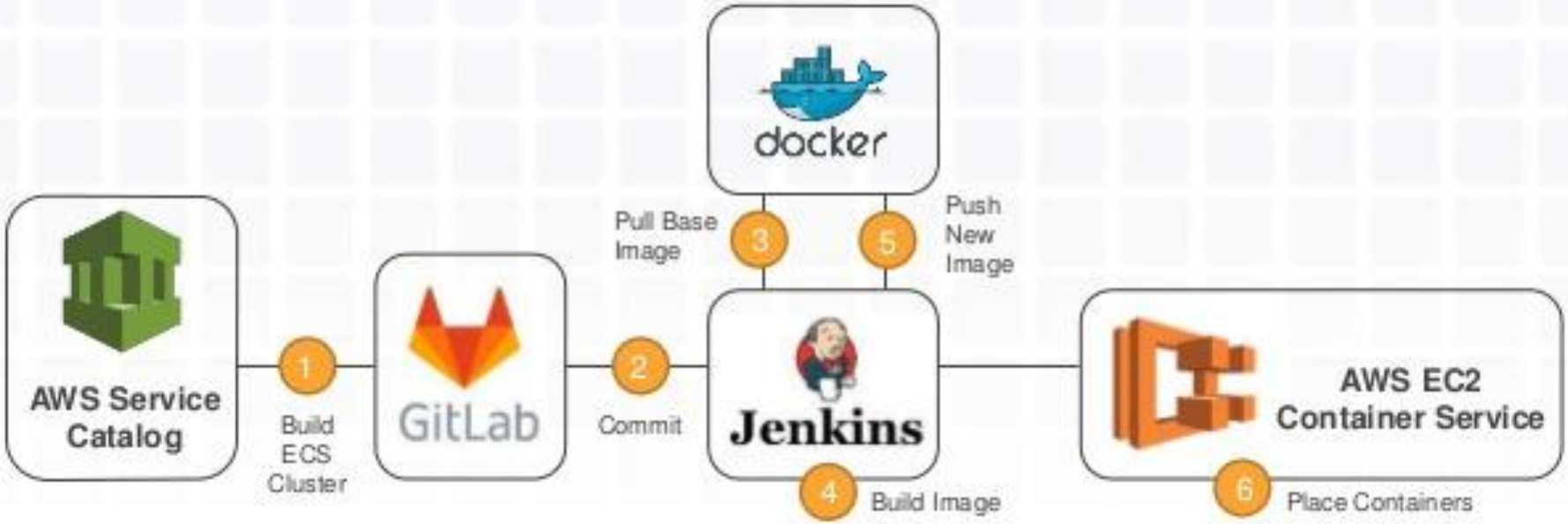
DOCKER PRACTICLES

- Docker Hub
- Docker Pull / Push / Run / Start / Stop
- Exposing Ports
- VS Code Integration

WHY DOCKER ?


AWS
2. Deploy Docker Host Instances
3. Deploy Docker Containers
4. Access Docker Service
1. Access Docker Cloud Portal
Internet
Welcome to nginx on busybox!
If you see this page, the nginx web server is successfully installed and working. Further configuration is required.
For online documentation and support please refer to nginx.org.
Commercial support is available at nginx.com.
Thank you for using nginx.
Consumer
Admin

Our Deployment Pipeline



DEMYSTIFYING : DATA ENGINEERING

By Chandan Shastri

ROLES IN BIG DATA PROJECTS

Data Analyst

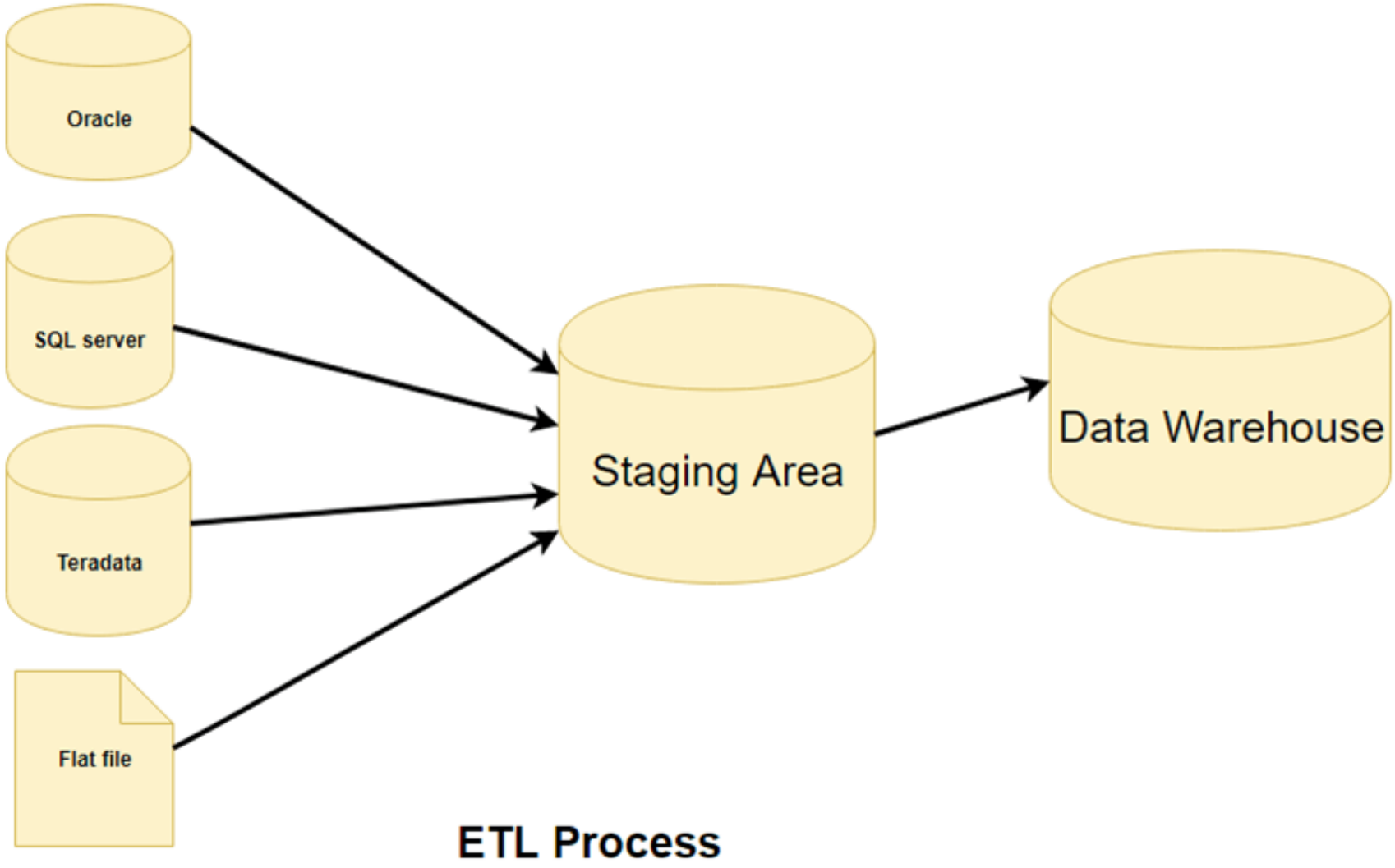
Data Engineer

Data Scientist
(Extremely
Advanced Tools)

BIG DATA IS OF COURSE BIG.

- But how it gets so Big ?


WORMSWORLD.US
#SUPERLIKE


Oracle
SQL server
Teradata
Flat file
Staging Area
Data Warehouse

ETL Process

The Big Data technology stack is evolving rapidly

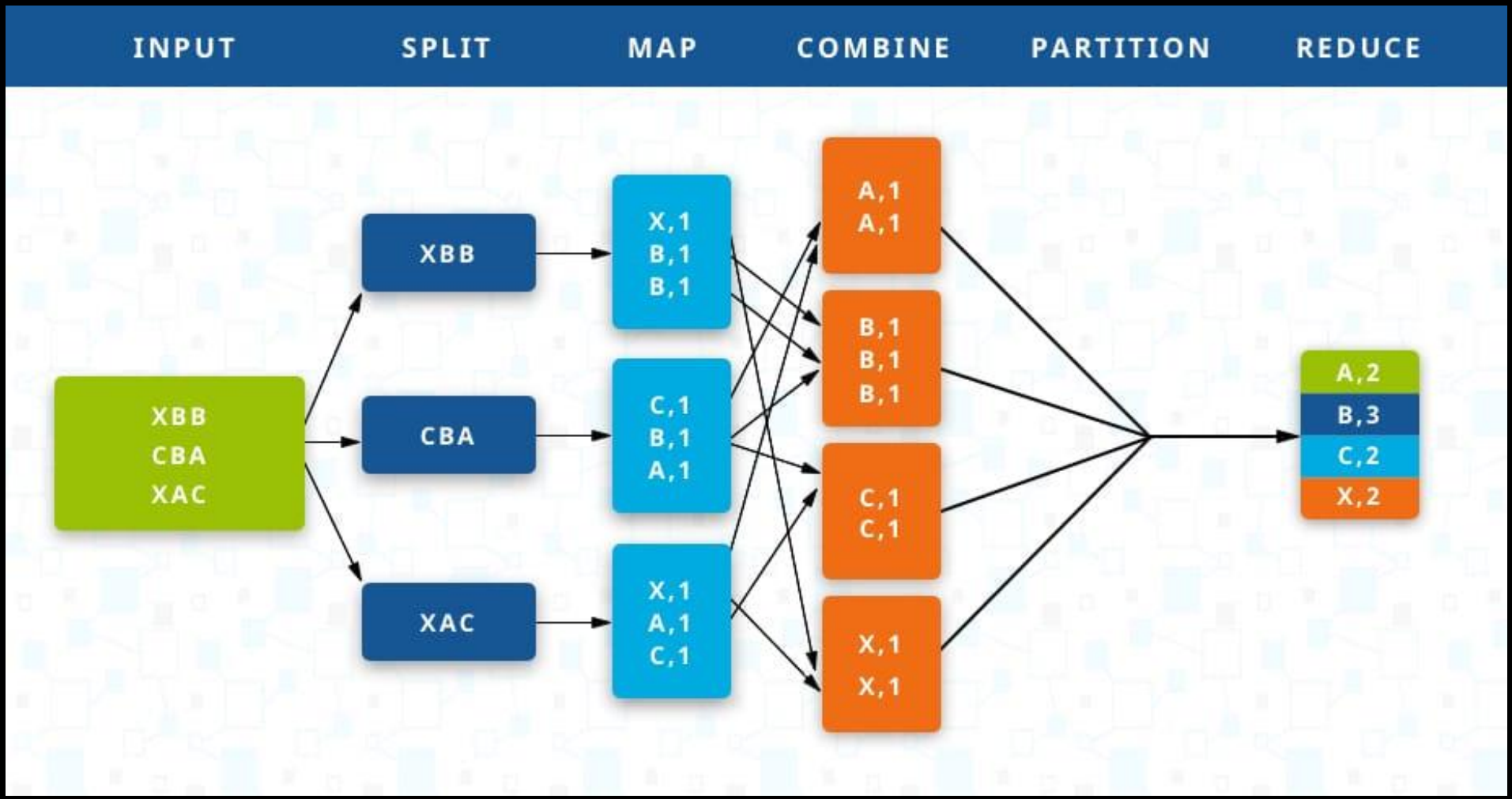
LET'S SEE ONLY THREE OF THEM FOR NOW

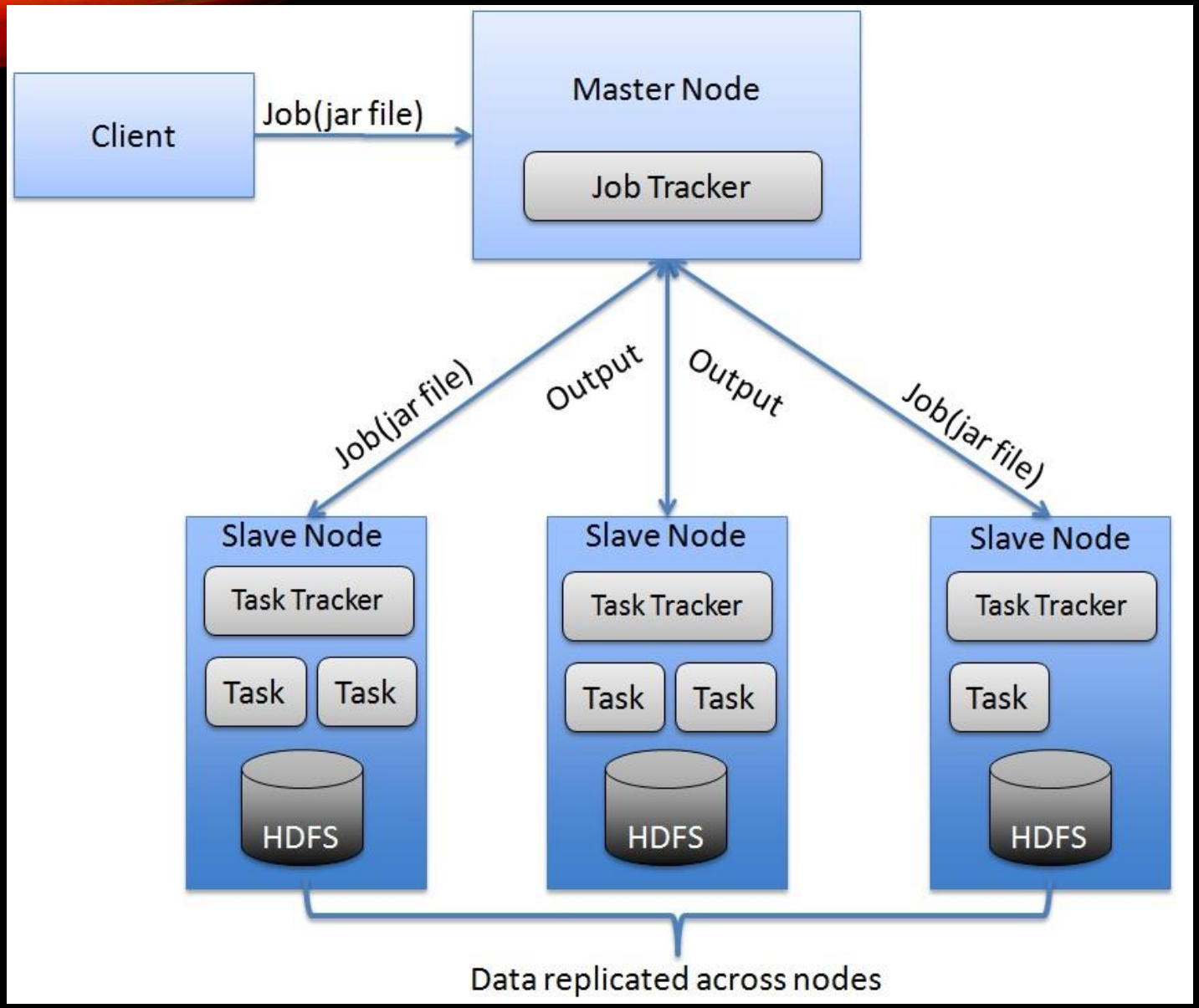
THE TRENDING STACK

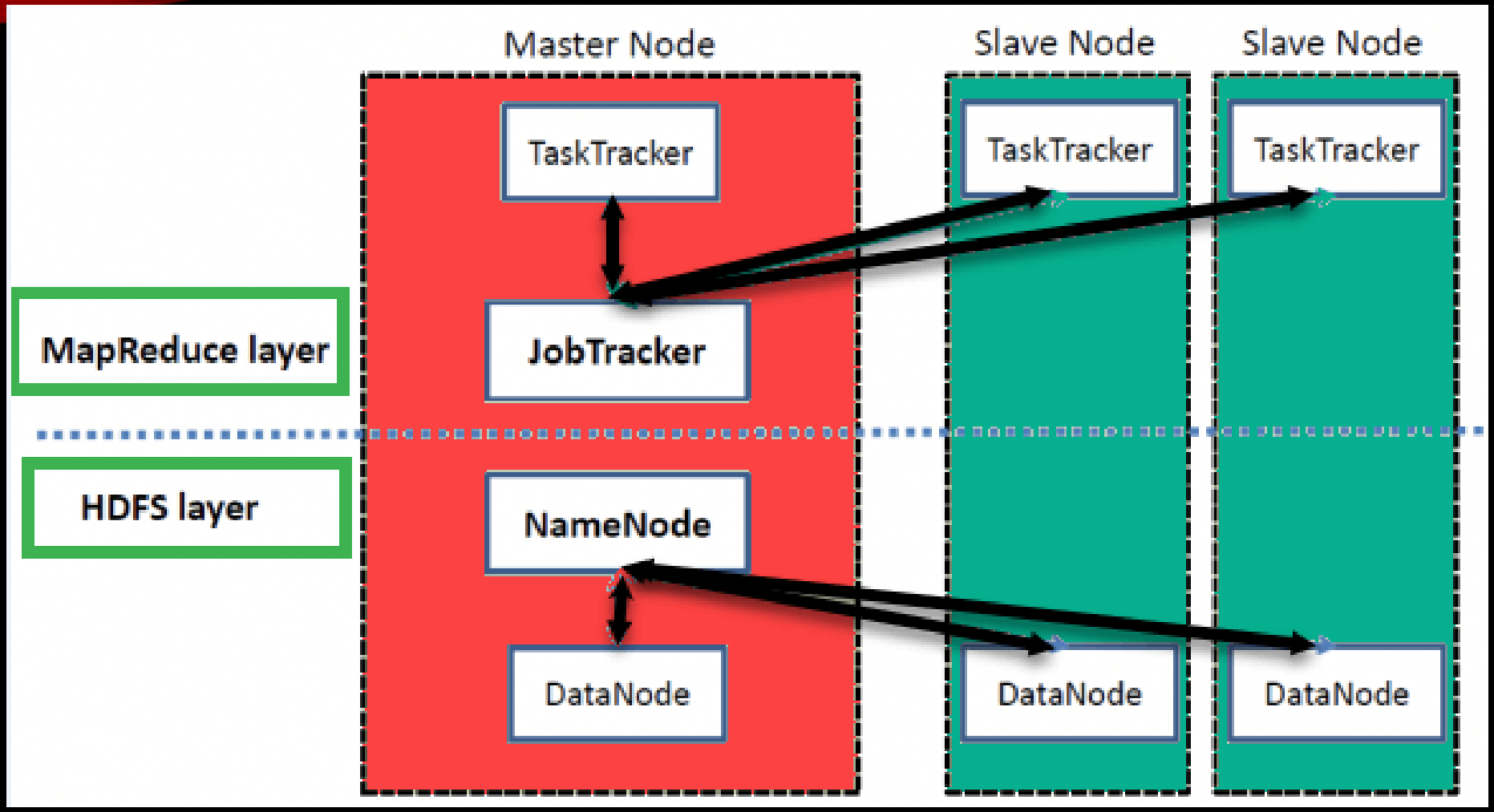

hadoop
HDFS

DISTRIBUTED COMPUTING

MAP - REDUCE

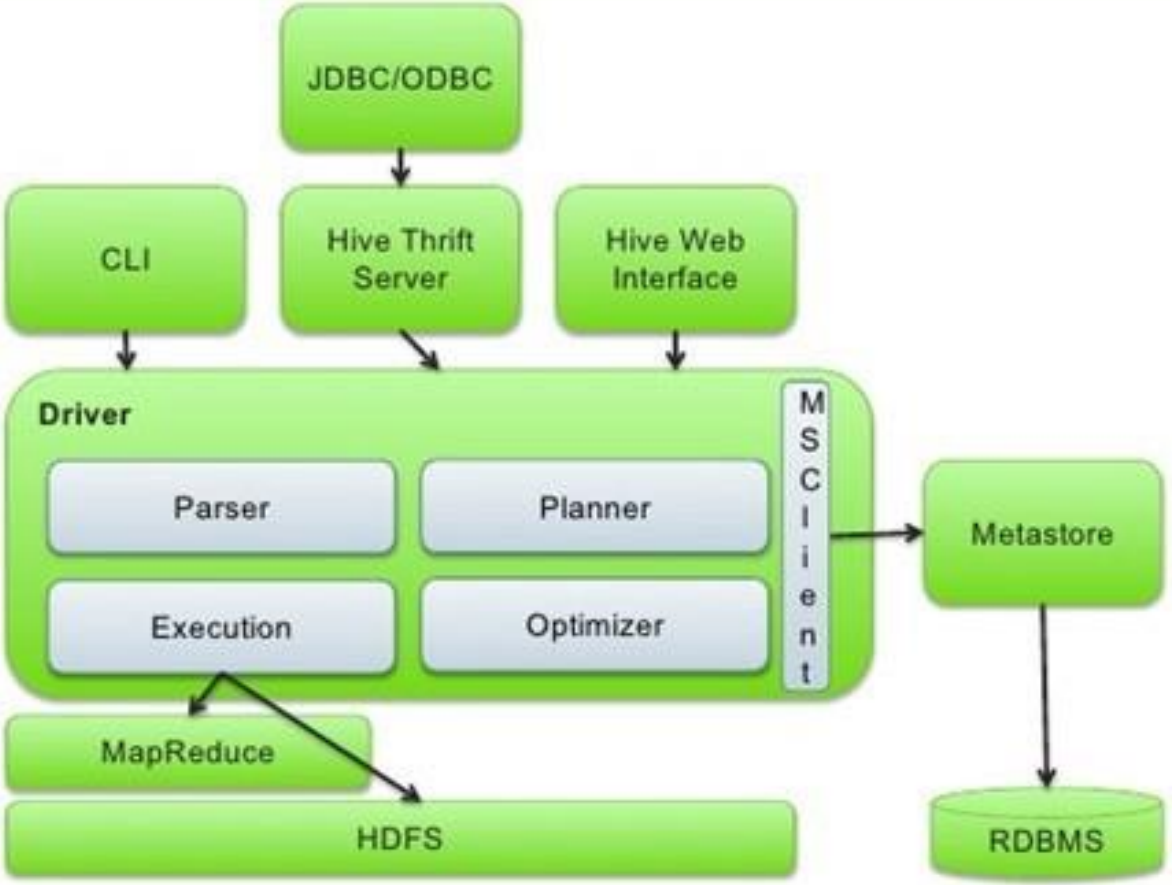

INPUT
SPLIT
MAP
COMBINE
PARTITION
REDUCE
XBB
CBA
XAC
XBB
CBA
XAC
X,1
B,1
B,1
C,1
B,1
A,1
X,1
A,1
C,1
A,1
A,1
B,1
B,1
B,1
C,1
C,1
X,1
X,1
A,2
B,3
C,2
X,2


Client
Job(jar file)
Master Node
Job Tracker
Job(jar file)
Output
Output
Job(jar file)
Slave Node
Task Tracker
Task
Task
HDFS
Slave Node
Task Tracker
Task
Task
HDFS
Slave Node
Task Tracker
Task
HDFS
Data replicated across nodes


Master Node
Slave Node
Slave Node
TaskTracker
TaskTracker
TaskTracker
JobTracker
MapReduce layer
HDFS layer
NameNode
DataNode
DataNode
DataNode


HIVE

Apache Hive Architecture



Hive Vs SQL

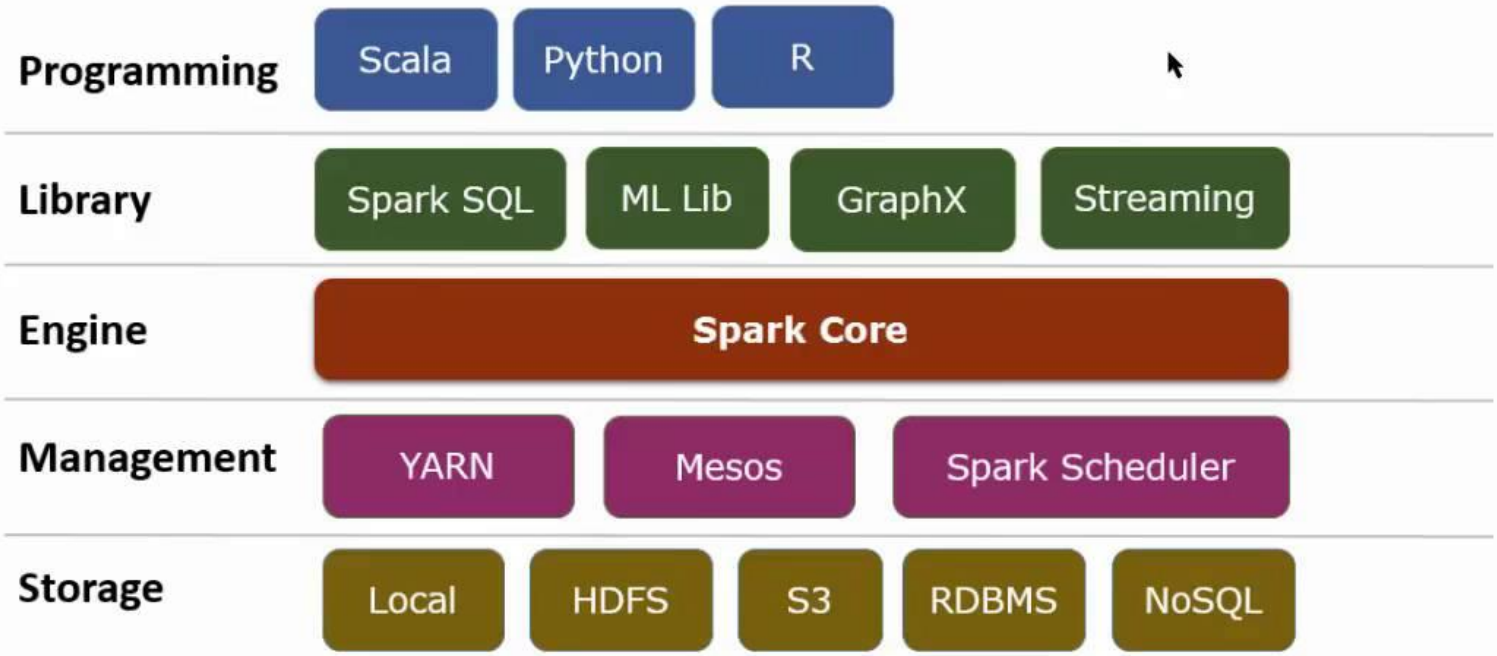
	RDBMS	HIVE
Language	SQL-92 standard (maybe)	Subset of SQL-92 plus Hive-specific extension
Update Capabilities	INSERT, UPDATE and DELETE	INSERT but not UPDATE or DELETE
Transactions	Yes	No
Latency	Sub-Second	Minutes or more
Indexes	Any number of indexes, very important for performance	No indexes, data is always scanned (in parallel)
Data size	TBs	PBs
Data per query	GBs	PBs

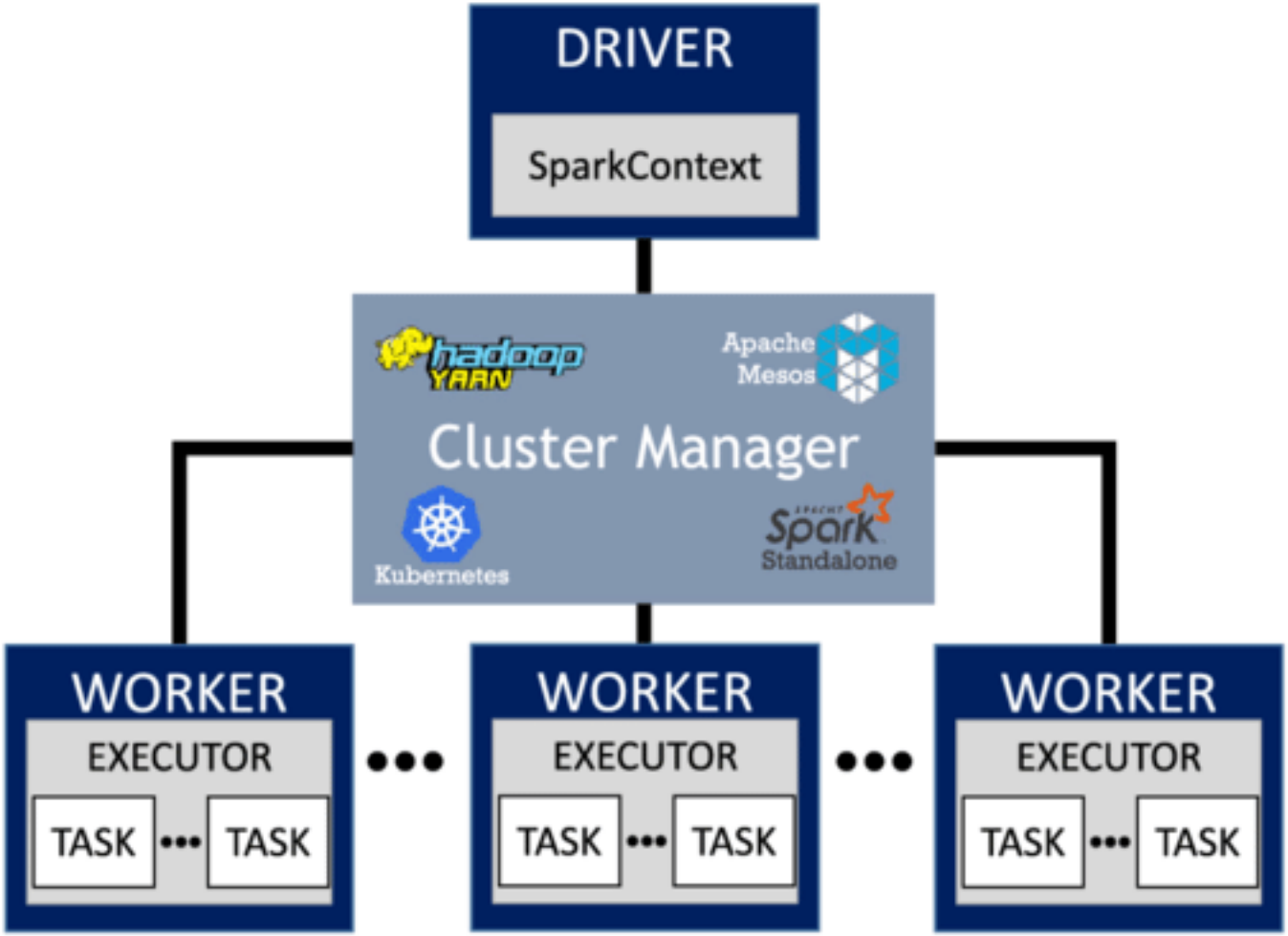
Augment MySQL Deployments, Sarah [illegible], Cloudera, 2010


APACHE
Spark
TM

V2 Maestros
The Data Science Experts

Spark Framework




DRIVER
SparkContext
hadoop YARN
Apache Mesos
Cluster Manager
Kubernetes
Spark Standalone
WORKER
EXECUTOR
TASK
TASK
WORKER
EXECUTOR
TASK
TASK
WORKER
EXECUTOR
TASK
TASK