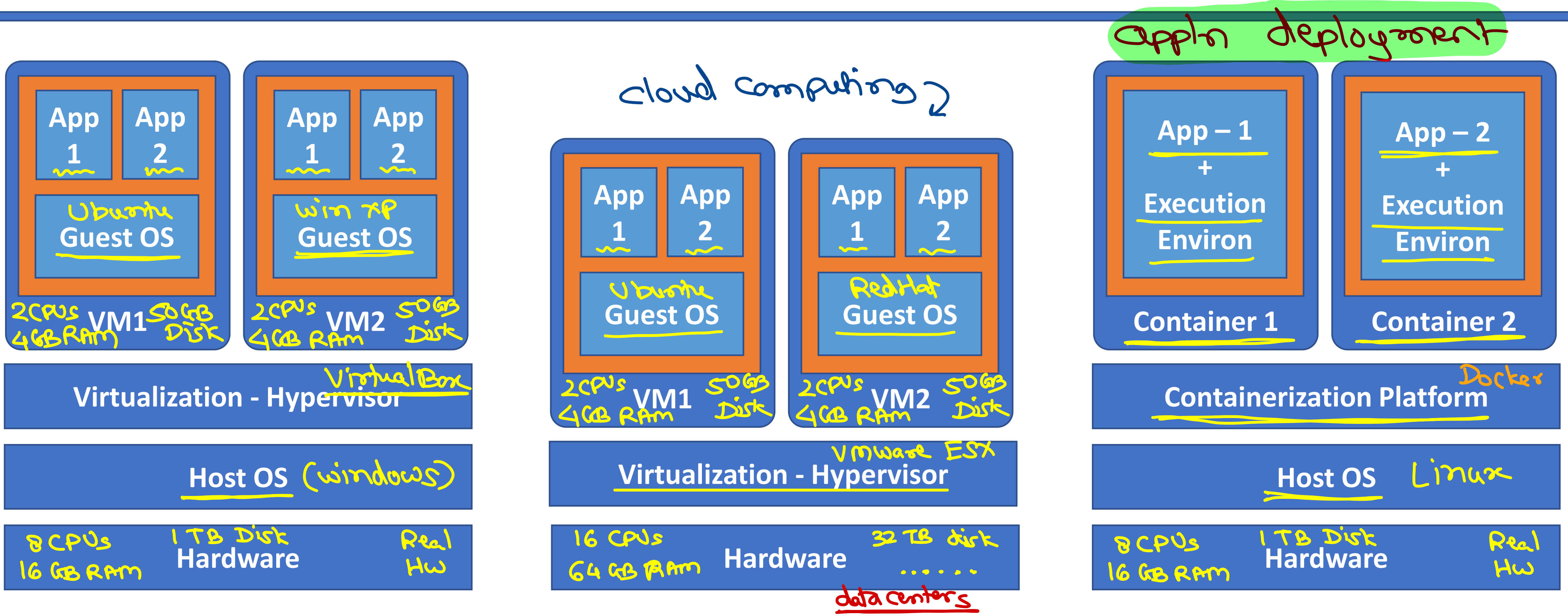
Fundamentals of Data Engineering

Trainer: Nilesh Ghule

Virtualization vs Containerization



appln deployment

cloud computing ?

App 1
App 2
Ubuntu
Guest OS
2CPUs VM1 50GB
4GB RAM Disk

App 1
App 2
win XP
Guest OS
2CPUs VM2 50GB
4GB RAM Disk

Virtualization - Hypervisor VirtualBox

Host OS (windows)

8 CPUs 16 GB RAM
1 TB Disk
Hardware
Real HW

App 1
App 2
Ubuntu
Guest OS
2CPUs VM1 50GB
4GB RAM Disk

App 1
App 2
RedHat
Guest OS
2CPUs VM2 50GB
4GB RAM Disk

VMware ESX
Virtualization - Hypervisor

16 CPUs 64 GB RAM
Hardware
32 TB disk
......
data centers

App – 1
+
Execution
Environ
Container 1

App – 2
+
Execution
Environ
Container 2

Containerization Platform Docker

Host OS Linux

8 CPUs 16 GB RAM
1 TB Disk
Hardware
Real HW

Type-II Virtualization

VMWare, VirtualBox, KVM, …

Type-I Virtualization

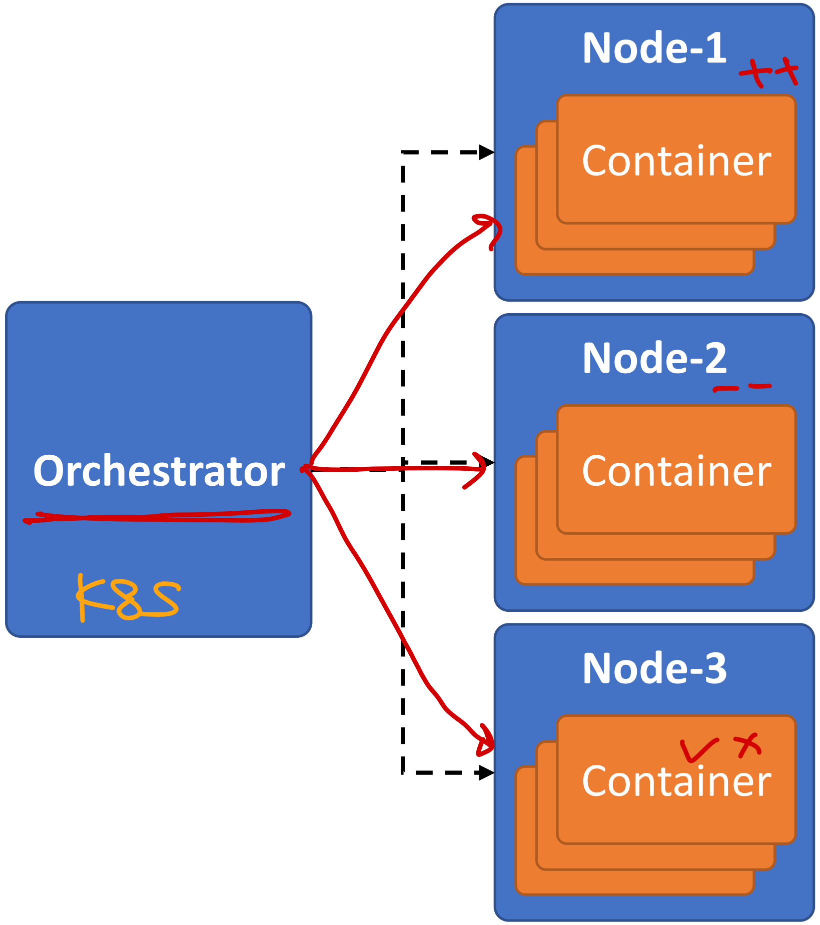
VMWare ESX, XEN, Hyper-V, …

Containerization

Docker, Podman, rkt, …

Orchestration

- Container Orchestration auto increase or decrease containers to handle change in workloads/demands. It also handles container failure (re-start).
- Ex: Docker swarm, Kubernetes, …



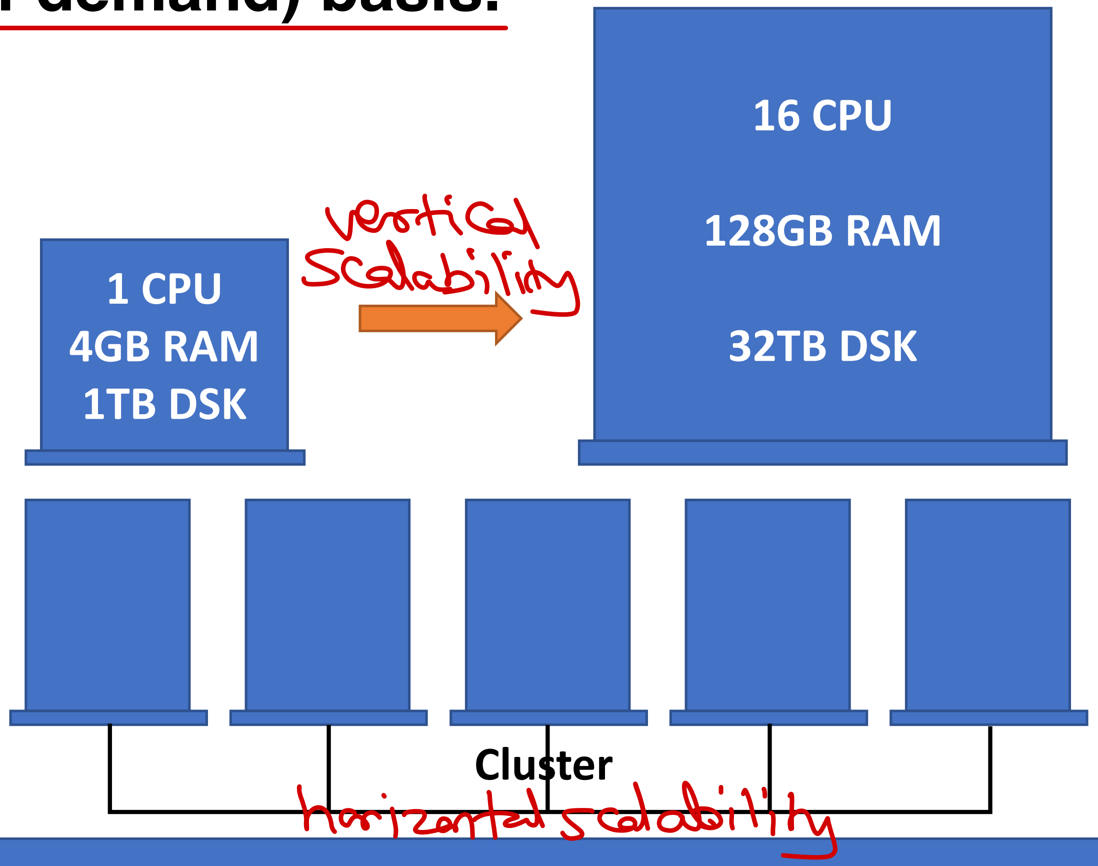
DevOps = Docker + K8S + GIT + Jenkins + …

CI/CD pipeline

Scalability and Elasticity

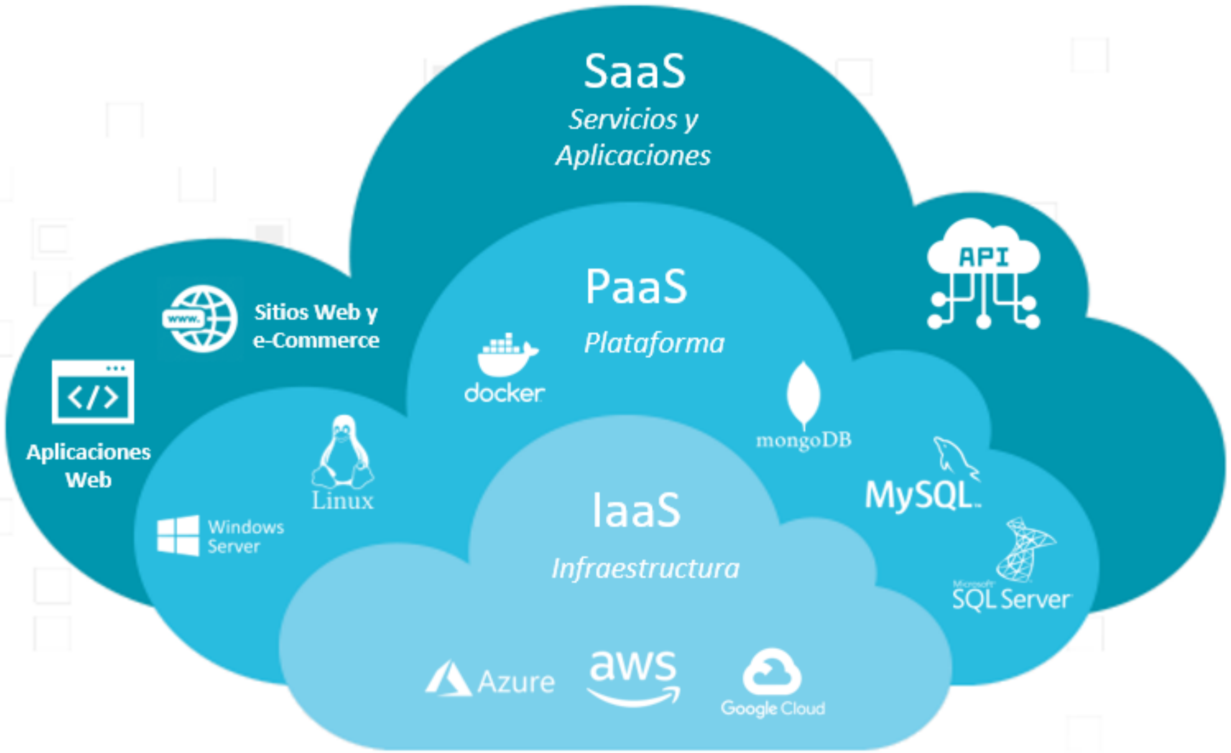
- Scalability is "ability of system / application to perform well under an increased or expanding workload".
- The resource usage is increased or decreased as per workload.
- Vertical scalability / Up scaling:
  - Increasing single system (hardware) resources in order to handle higher loads.
  - Need to handle SPOF (single point of failure) by adding backup system.
- Horizontal scalability / Out scaling:
  - Adding new systems/notes into the cluster in order to handle higher loads.
  - More economical solution with higher complexity.

- Elastic: Cloud systems are designed to increase/decrease ~~load~~ resources as per workload.
- Cloud payments are usually pay-per-use (on-demand) basis.



Cloud Service Models

AWS (Amazon), GCP (Google), Azure (Microsoft), Smart Cloud (IBM), ...



- IaaS: Infrastructure as-a Service
  - AWS EC2, S3, VPC
    vm storage network
- PaaS: Platform as-a Service
  - Beanstalk, SageMaker,
- SaaS: Software as-a Service
  - Gmail, Drive, Facebook, LinkedIn, Netflix
- DaaS: Database as-a Service
  - RDS, Aurora, Atlas, DynamoDb
- FaaS: Function as-a Service
  - Lambda, Google functions

Data centers → Huge infrastructure with lots of real computers and network, ...

→ Computer/Machine
→ Network
→ Storage

Big Data & Analytics Spectrum

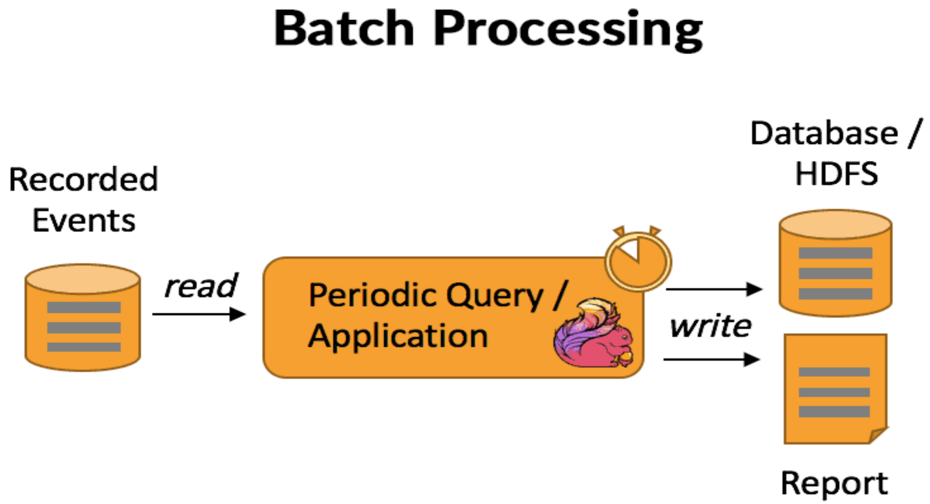
PG-DBDA course

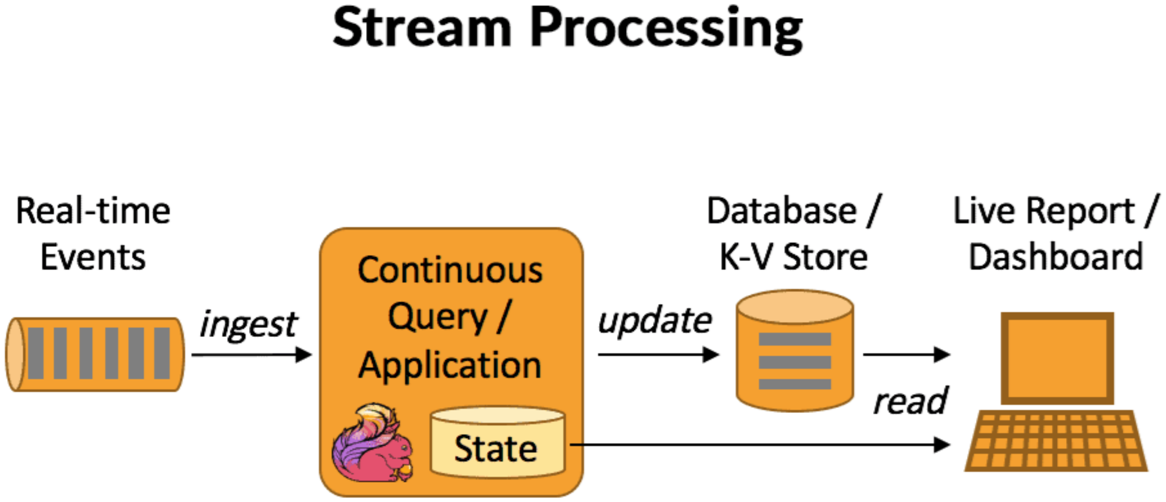
- **Data storage**
  - RDBMS & NoSQL databases ✓
  - Data warehouse ✓
  - S3, DFS, … ✓
- **Data Analysis & visualizations**
  - Data Visualizations ✓
  - Business reports ✓
- **Artificial Intelligence, Data Science & Data mining**
  - Mathematics, Statistics ✓ & Computer algorithms
  - Machine learning & Deep learning ✓
  - R Programming, Python ✓
- **(Big) Data Engineering**
  - Hadoop ✓, Hive ✓, Spark ✓, Kafka ✓, BigTable ✓, …
  - Java ✓, Scala ✓, Python ✓, SQL
- **Infrastructure**
  - Linux ✓, Cloud Computing ✓

Batch processing vs Stream processing

- Processing finite set of data (data at rest).
- Incremental data load is managed by programmer.
- Cluster planned as per data size. High throughput.
- Job run once per batch.

- Processing live stream of data (data in motion).
- Data processing is managed by the framework.
- Less throughput.
- Job is running forever.

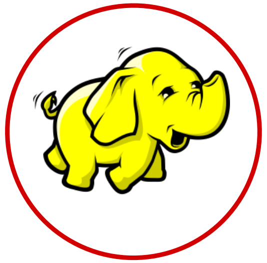


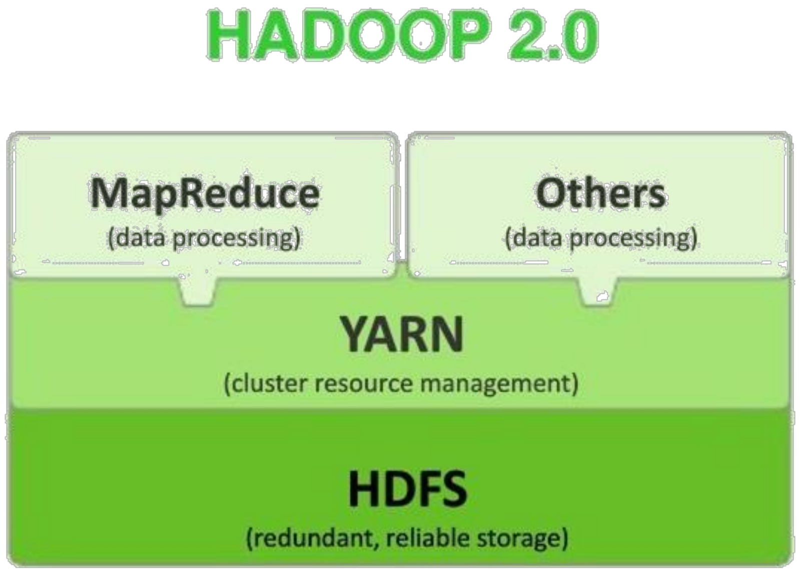


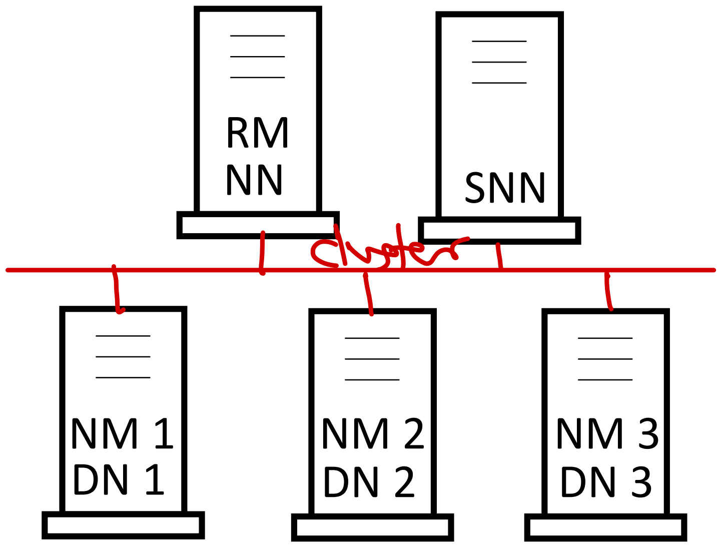
Apache Hadoop

Google → Distributed Storage = Google File System - 2003
→ Distributed Computing = Map Reduce - 2004

- Hadoop is developed by Doug cutting.
  - Web crawler – Nutch
  - Distributed computing and storage needed to process huge data produced by the crawler.
  - Joined Yahoo. Developed and open sourced under Apache license. → 2006
- Hadoop
  - Distributed storage: HDFS → Hadoop Distributed File System (like GFS)
  - Distributed computing: Map-reduce
  - Cluster manager: YARN - Yet Another Resource Negotiator
- Hadoop is like a Kernel/Platform on which many different applications are built (eco-systems).





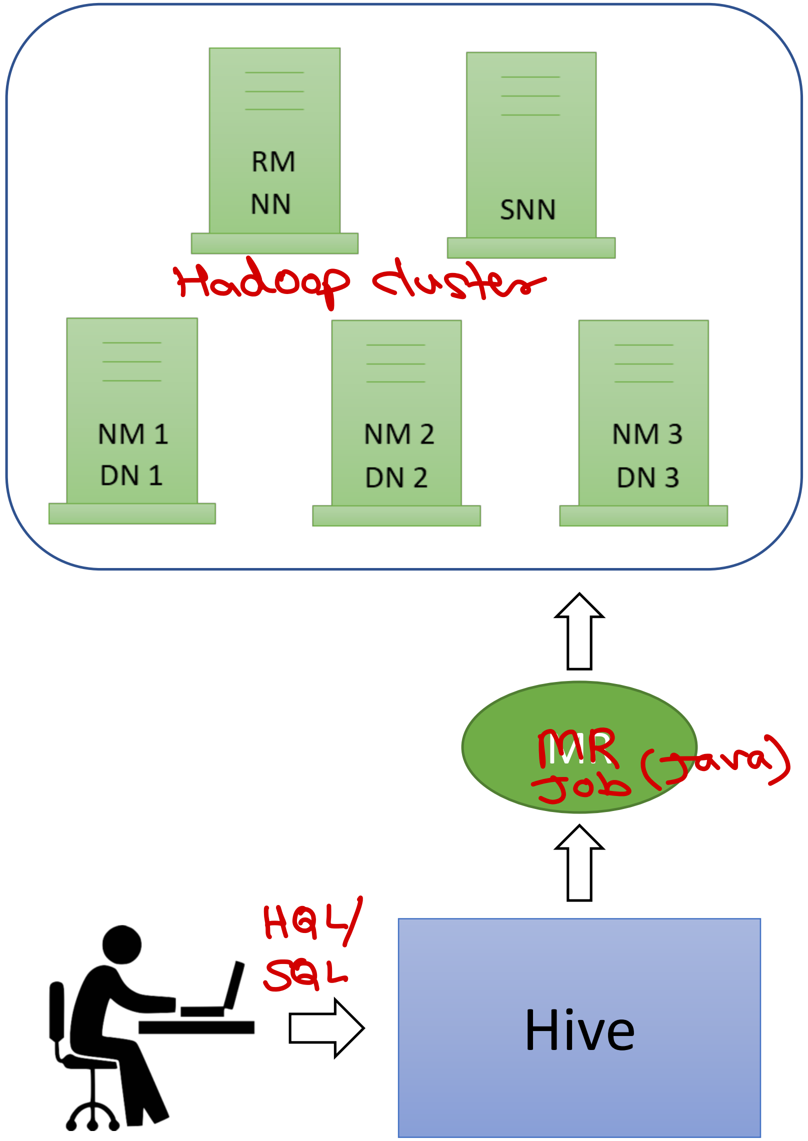


Apache Hive

- Developed by Facebook (2007)
- Client software that convert Hive QL queries to MapReduce.
- Hive QL is similar to SQL with many extended features.
- Hive manages structured data.
- Hive is data warehouse (OLAP) built for Hadoop.
  - Data storage = HDFS
  - Metadata = RDBMS
  - Data processing = Map-reduce or Spark or Tez.

$\overbrace{\text{Spark or Tez}}^{\text{High speed execution}}$

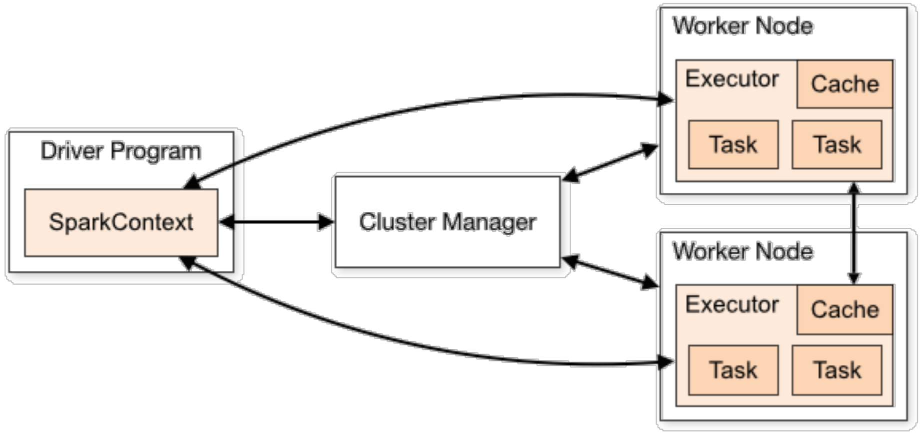
✓ ✓ ✓



Apache Spark

can work with any storage: HDFS, S3, Local, ...

- Spark is Distributed computing framework, that can process huge amount of data.
- Spark can be used as eco-system of Hadoop or can be used as independent distributed computing framework.
- Developed by UCB AMPlabs division.
  - Algorithms,
  - Machines,
  - People
- Further developed/maintained by DataBricks.
- Popular Spark vendors
  - DataBricks, AWS EMR, Cloudera, MapR
- Spark Toolkit



② Scala * | ③ Python * | ④ R | ⑤ Java

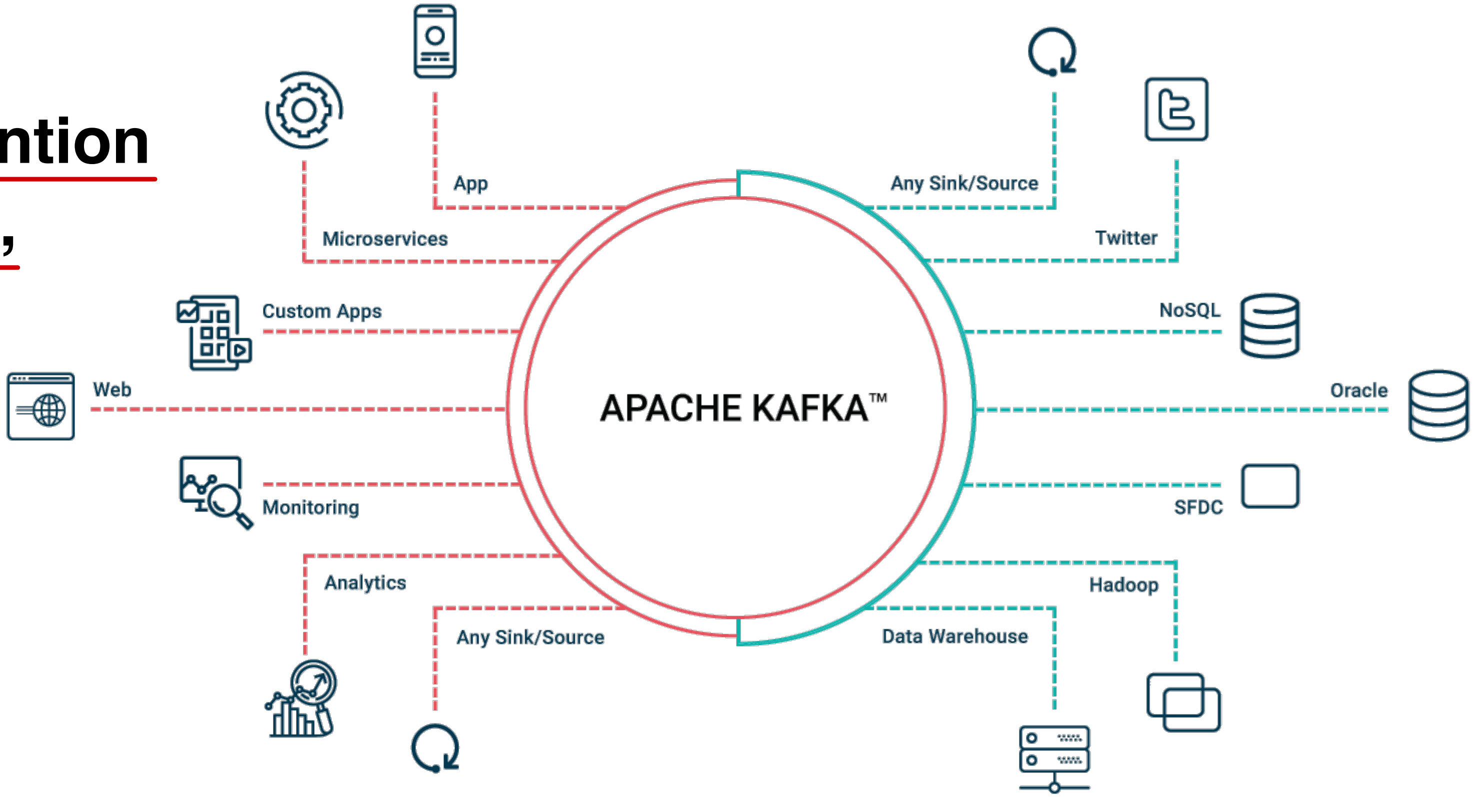
① Spark SQL | Spark Streaming | Spark ML | Spark GraphX

wrapper on RDD — Spark High Level API (Dataframes)

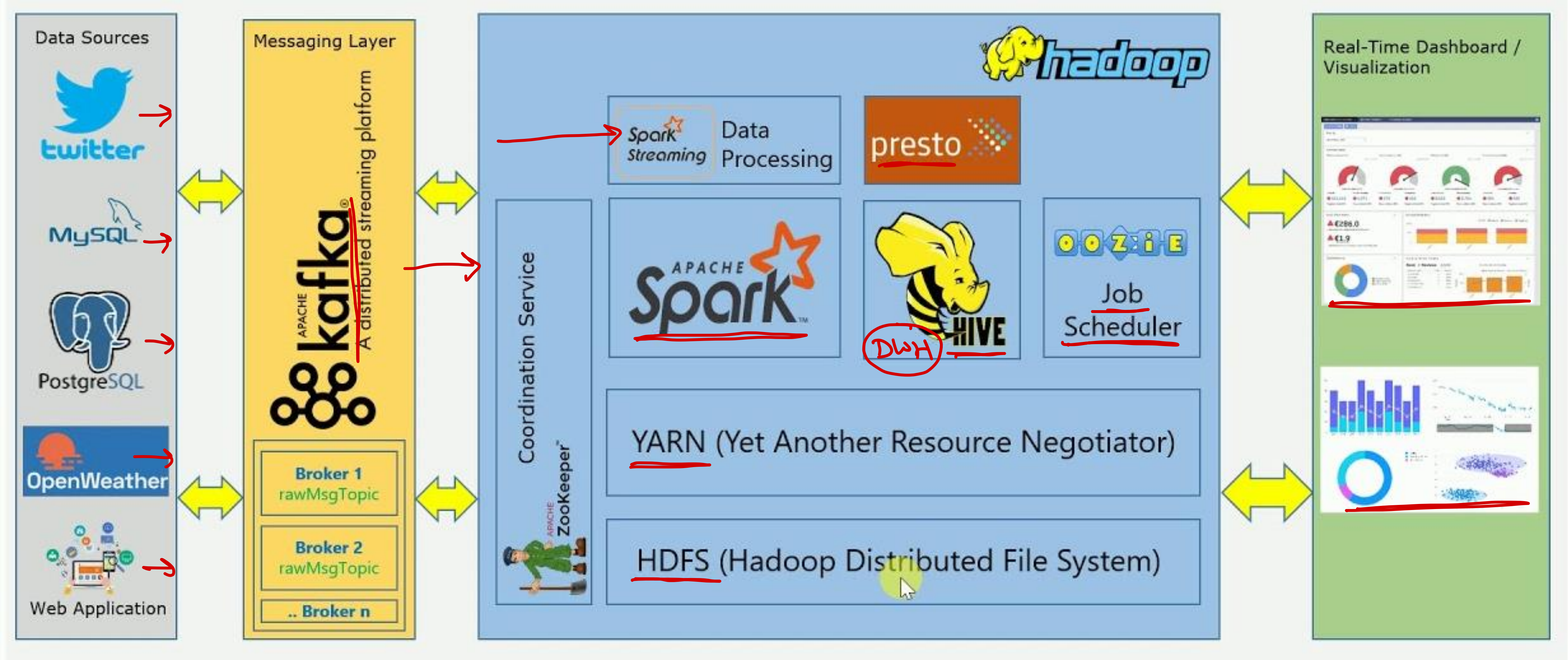
Resilient Distributed Dataset ← Spark Core (RDD & DAG) → Directed Acyclic Graph

Apache Kafka

- Kafka is a distributed messaging system.
- Developed at LinkedIn and open sourced in 2011.
- Used by LinkedIn, Twitter, Uber, airbnb, …
- Advantages
  - Scalable, Durable, Finite retention
  - Low latency, Strong ordering,
  - Exact once delivery
- Applications
  - Stream processing
  - Notifications.



Real time dashboard reference architecture



Big Data domains & opportunities

- Domains: Health-care, Retails, Trading/Share market, Finance, Security, Fraud, Search engines, Log Analysis, Telecom, Traffic Control, Manufacturing and lot more.
- Big Data is all about :- Think, Collect, Manage, Analyze, Summarize, Visualize, Discover Knowledge and Take Decisions.
- Job profiles:
  - Business Analyst/Intelligence
  - Database engineer / DWH
  - Big Data engineer
  - Data operations
  - Big Data Architect

- The sexiest job in the 21st century require a mixture of multidisciplinary abilities and suitable candidates must be prepared to learn & develop constantly.

-Ronald Van Loon

https://www.youtube.com/live/BxwpqnQ6BgQ?si=55cmOUDfiIGDAsLY

SunBeam
Big Data
Webinar

Important.

Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>