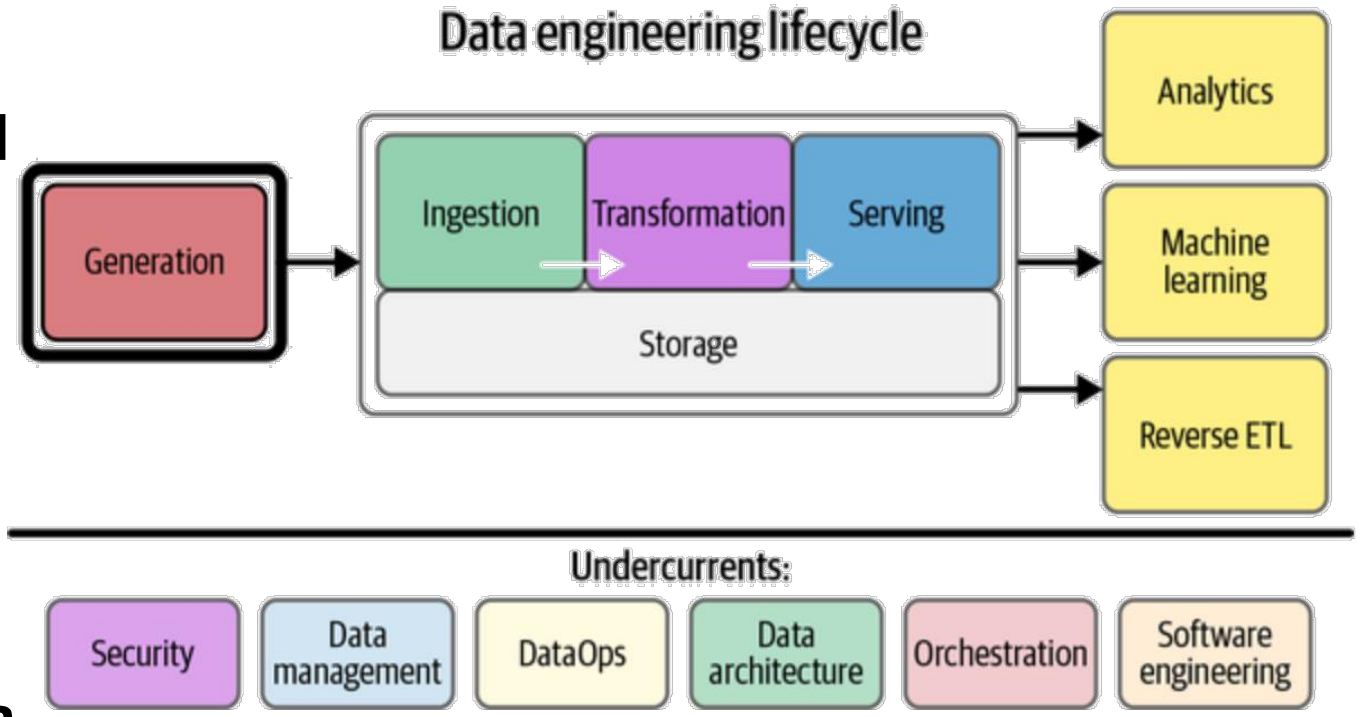
Quiz 2 = tomorrow 21 may

Fundamentals of Data Engineering

Trainer: Pradnyaa S Dindorkar

Data engineering

- Data engineering is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning.
- Data engineer manages data engineering lifecycle, beginning with getting data from source systems & ending with serving data for use cases, such as analysis or machine learning.



https://youtu.be/hZu_87l62J4

Traditional ETL vs Hadoop ELT

- ETL stands for Extract, Transform and Load.
- The ETL process typically extracts data from the source/transactional systems, transforms it to fit the model of data-warehouse and finally loads it to the data warehouse.
- The transformation process involves cleansing, enriching and applying transformations to create desired output.
- Data is usually dumped to a staging area after extraction.

- ELT stands for Extract, Load and Transform.
- As opposed to loading just the transformed data in the target systems, the ELT process loads the entire data into the data lake. This results in faster load times.
- The load process can also perform some basic validations and data cleansing rules.
- The data is then transformed for analytical reporting as per demand.

Data storage

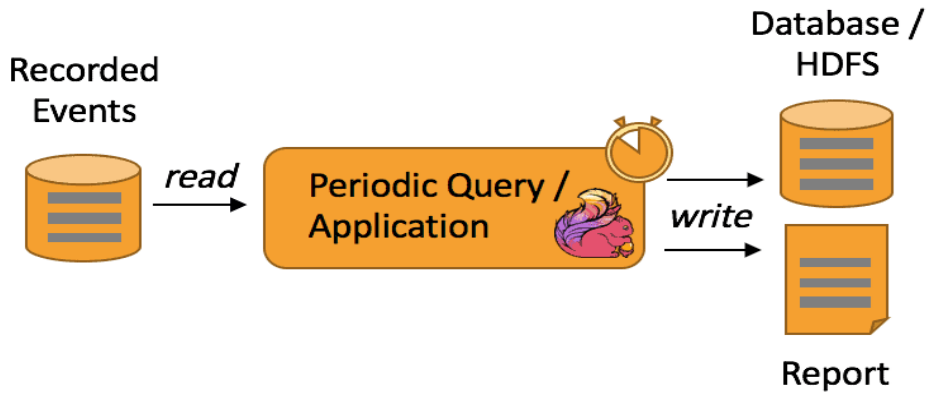
- Data storage is related to multiple stages in data engineering life cycle i.e. ingestion, transformation and serving.
- Storage needs to be selected based on read/write requirement, speed, durability, consistency, availability, scalability, fault tolerance, … factors.
- Storage tradeoffs
  - Local storage vs Distributed storage
  - Strong consistency vs Eventual consistency
- Storage options are: File storage, Local disk storage, Network attached storage (NAS), Cloud file systems (S3/Blob), Block storage, RAID, Storage area network (SAN), Object storage, HDFS, Streaming storage.

Batch processing vs Stream processing

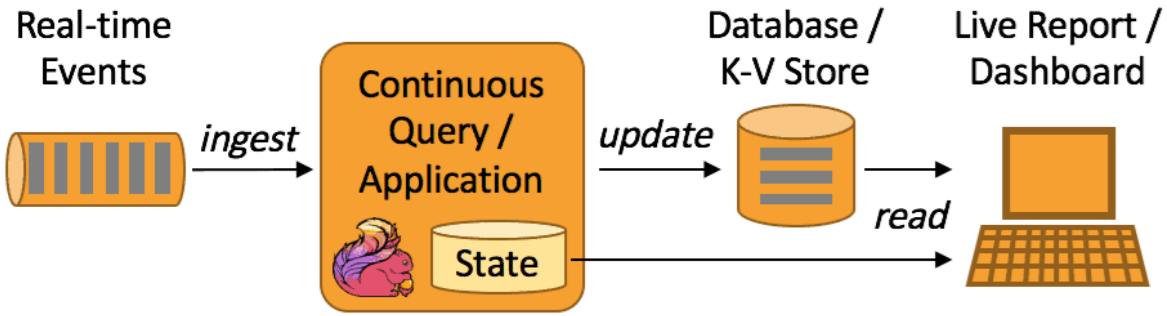
- Processing finite set of data (data at rest).
- Incremental data load is managed by programmer.
- Cluster planned as per data size. High throughput.
- Job run once per batch.

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

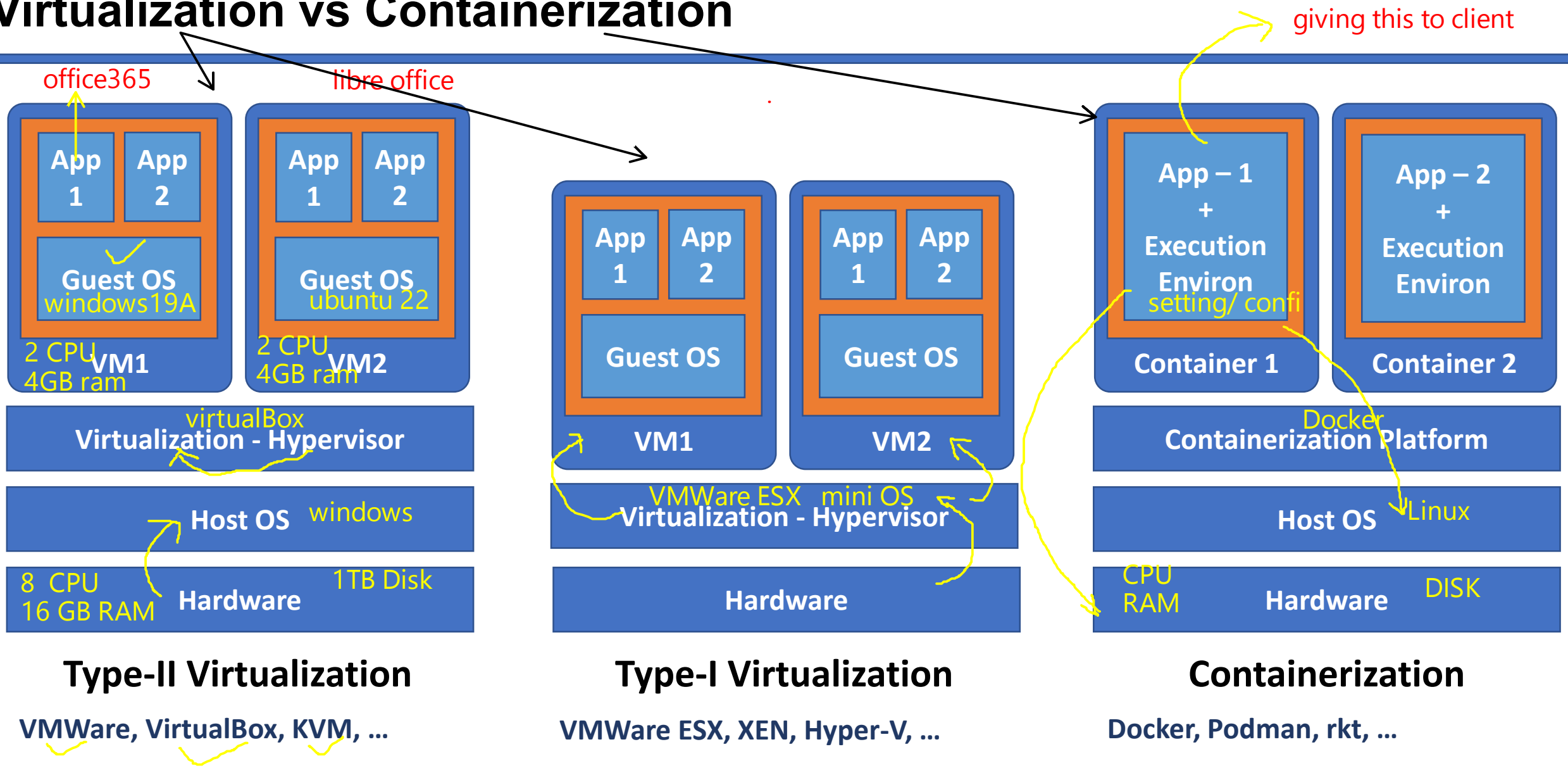
Batch Processing



Stream Processing



Virtualization vs Containerization



Type-II Virtualization

VMWare, VirtualBox, KVM, ...

Type-I Virtualization

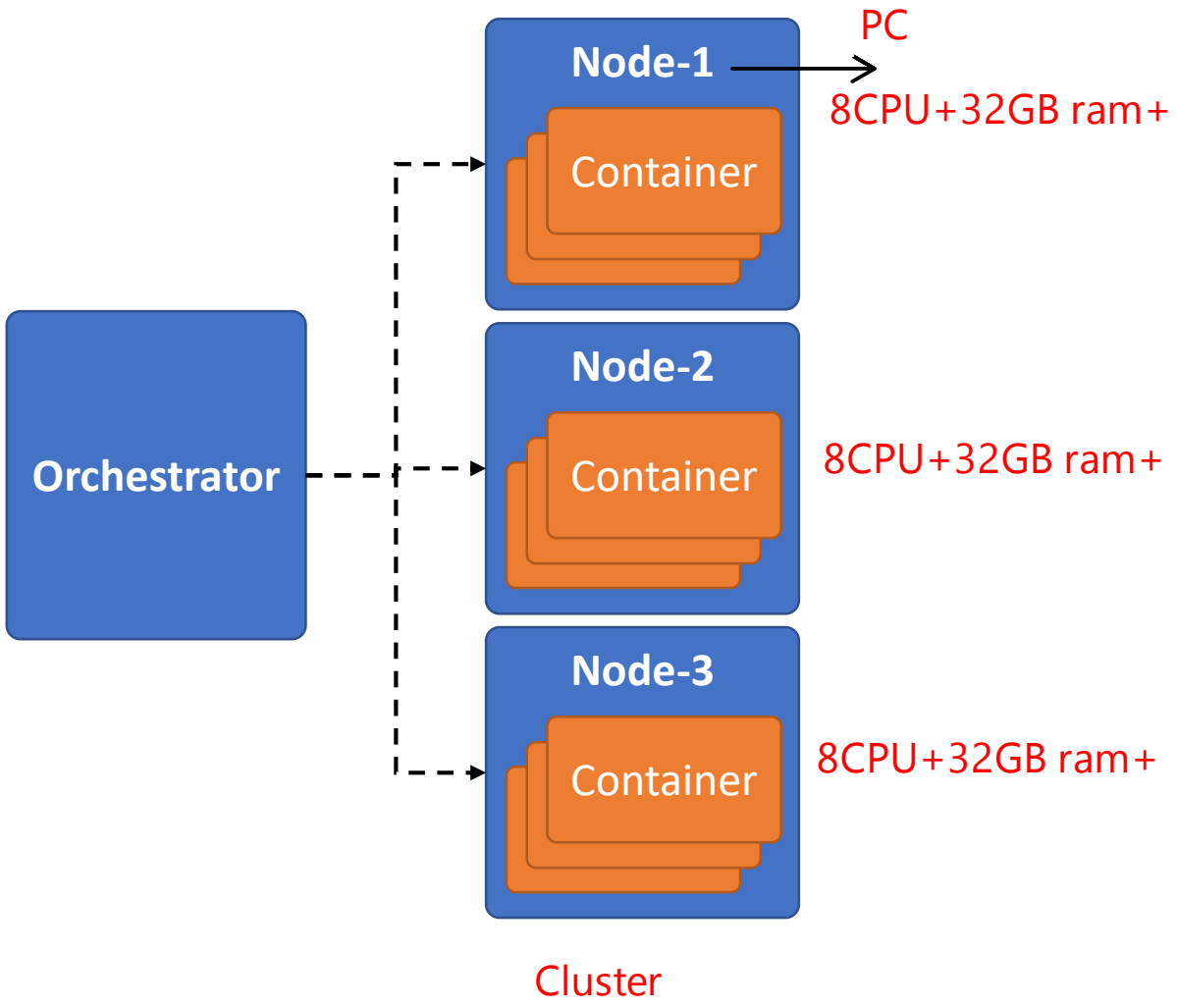
VMWare ESX, XEN, Hyper-V, ...

Containerization

Docker, Podman, rkt, ...

Orchestration

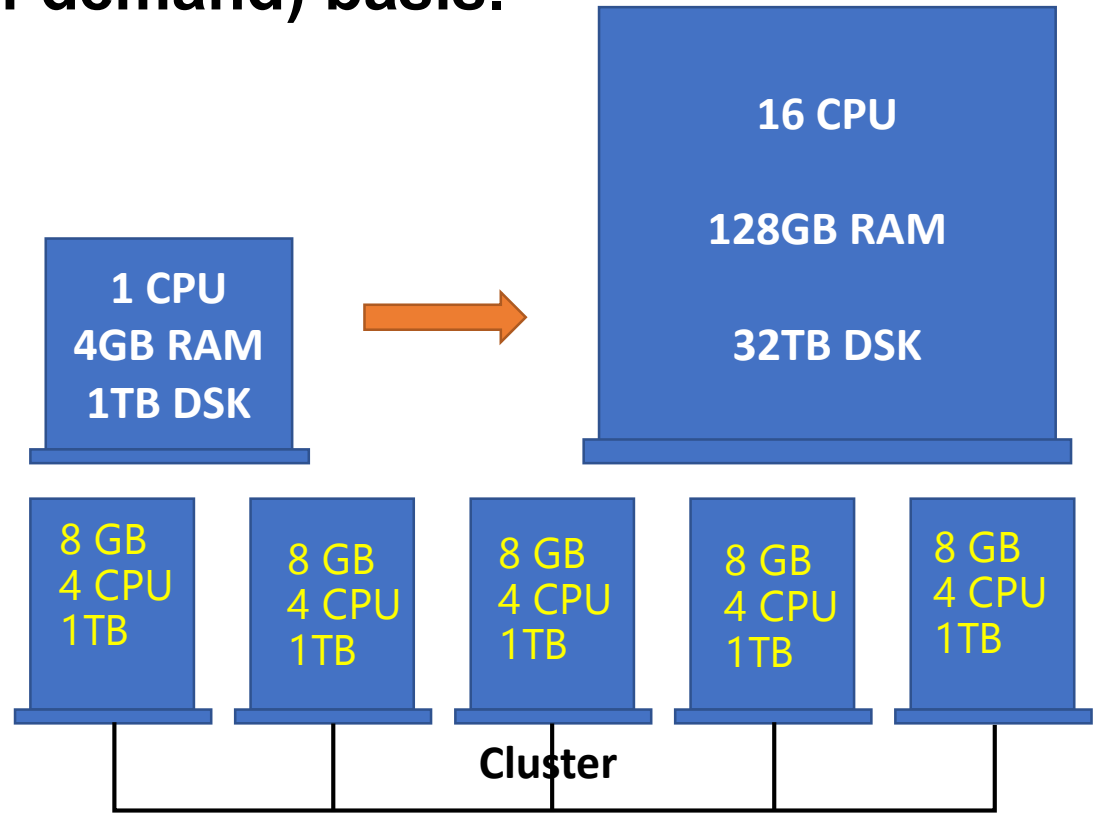
- **Container Orchestration auto increase or decrease containers to handle change in workloads/demands. It also handles container failure (re-start).**
- **Ex: Docker swarm, Kubernetes, …**



Scalability and Elasticity

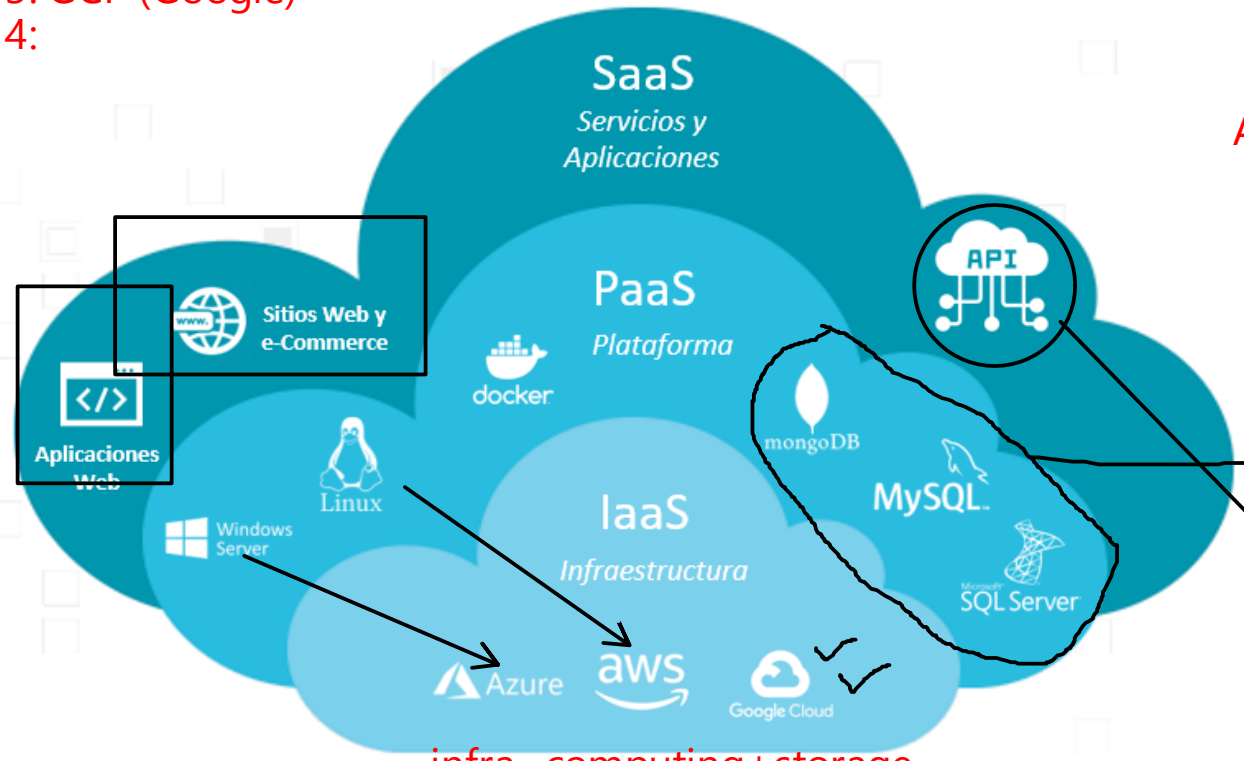
- Scalability is “ability of system / application to perform well under an increased or expanding workload”.
- The resource usage is increased or decreased as per workload.
- Vertical scalability / Up scaling:
  - Increasing single system (hardware) resources in order to handle higher loads.
  - Need 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 as per workload.
- Cloud payments are usually pay-per-use (on-demand) basis.



Cloud Service Models

Cloud providers
1: AWS (Amezon)
2: Azure (microsoft)
3: GCP (Google)
4:



infra=computing+storage
+networking

VM

- **IaaS: Infrastructure as-a Service**
  - **AWS EC2, S3, VPC**
- **PaaS: Platform as-a Service**
  - **Beanstalk, SageMaker,** => ML deploy

AWS app deploy

- **SaaS: Software as-a Service**
  - **Gmail, Drive, Facebook, LinkedIn, Netflix**
- **DaaS: Database as-a Service**
  - **RDS, Aurora, Atlas, DynamoDb** kv nosql

RDBms MySQl

- **FaaS: Function as-a Service**
  - **Lambda, Google functions**

AWS

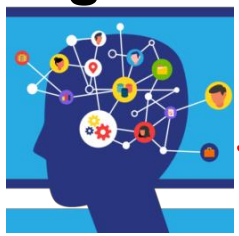
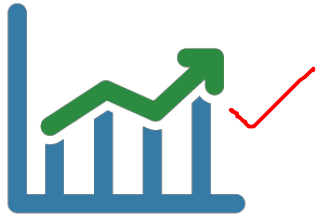
mongoDB

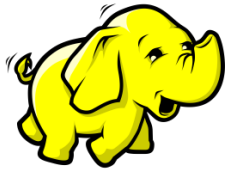
Big Data & Analytics Spectrum

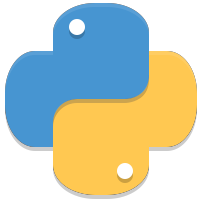
- **Data storage**
  - RDBMS & NoSQL databases
  - Data warehouse
  - S3, DFS, …
- **Data Analysis & visualizations**
  - Data Visualizations
  - Business reports
- **Artificial Intelligence, Data Science & Data mining**
  - Mathematics, Statistics & Computer algorithms
  - Machine learning & Deep learning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







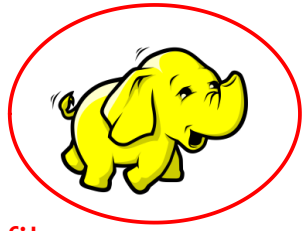


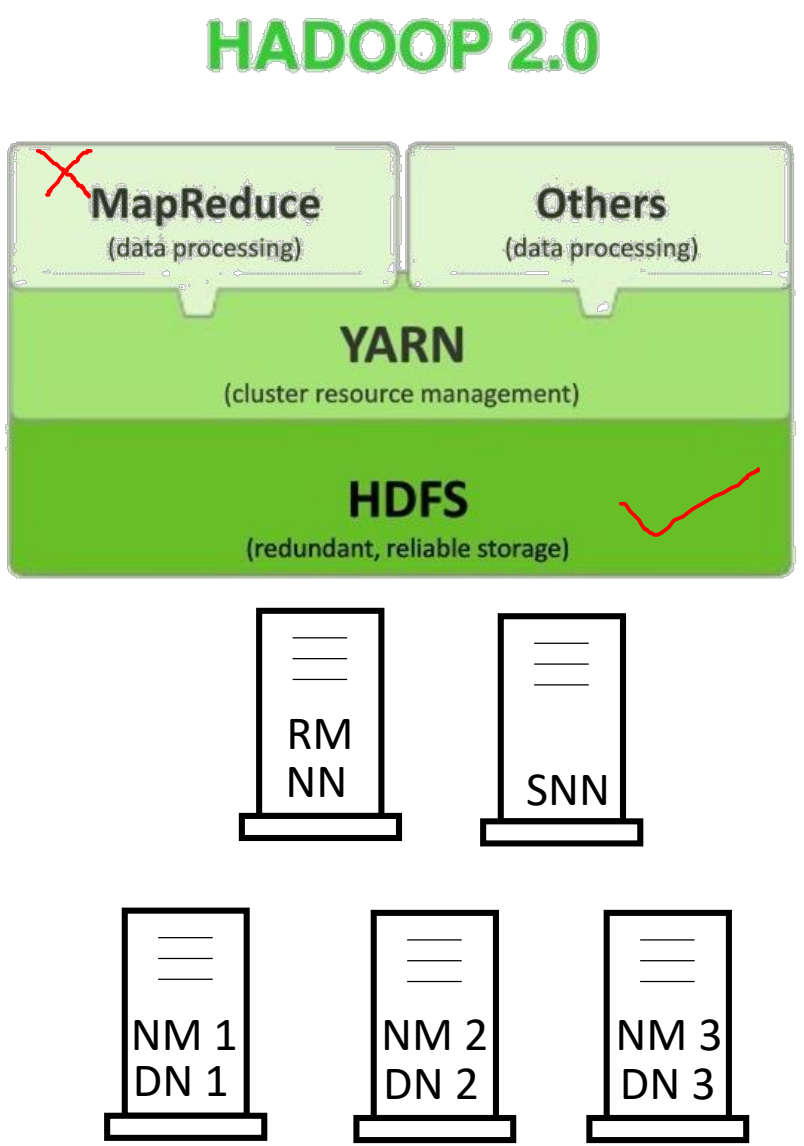
Apache Hadoop

@2003 - GFS -> google file system -> 1st distributed file system
@2004 - MapReduce -> distributed processing

3.3

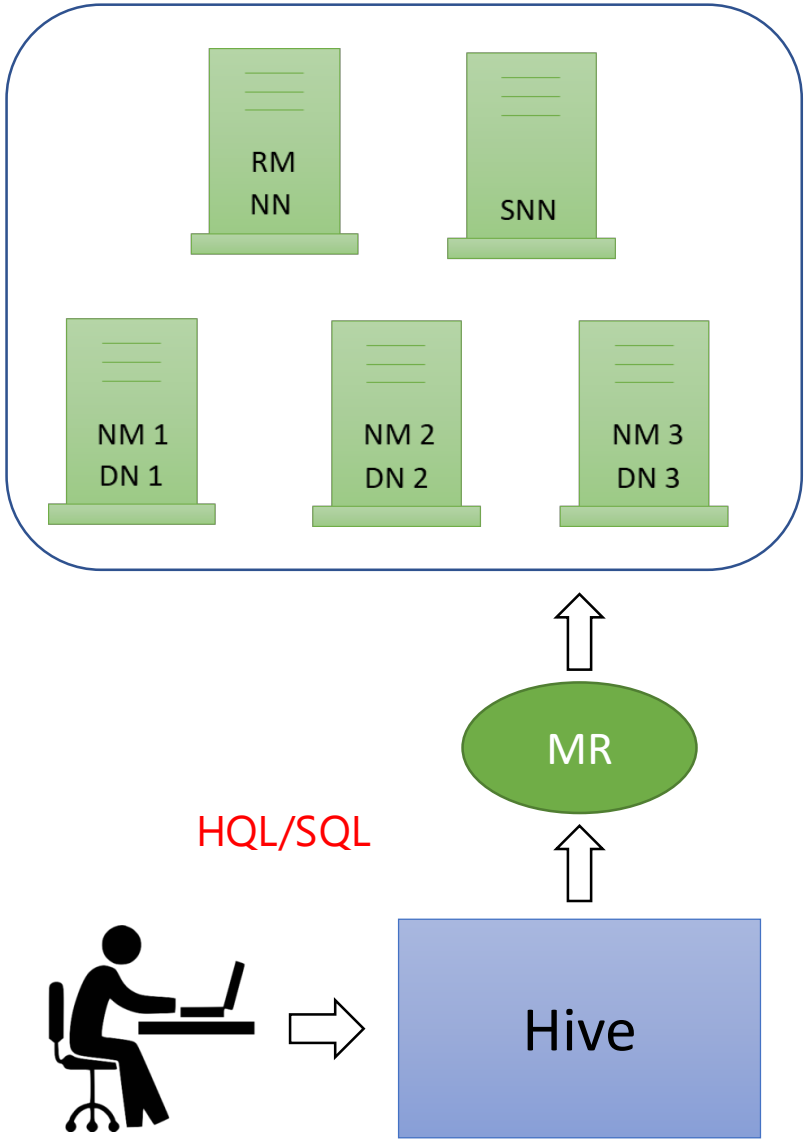
- **Hadoop is developed by Doug cutting.**
  - **Web crawler – Nutch**
  - **Distributed computing and storage needed to process huge data produced by the crawler.**
  - **Joined Yahoo. Developed and open sourced under Apache license.** @2006
- **Hadoop** Hadoop distributed file system
  - **Distributed storage: HDFS**
  - **Distributed computing Map-reduce** java 70-> 5files
  - **Cluster manager: YARN** Yet another resource negotiator
- **Hadoop is like a Kernel/Platform on which many different applications are built (eco-systems).**





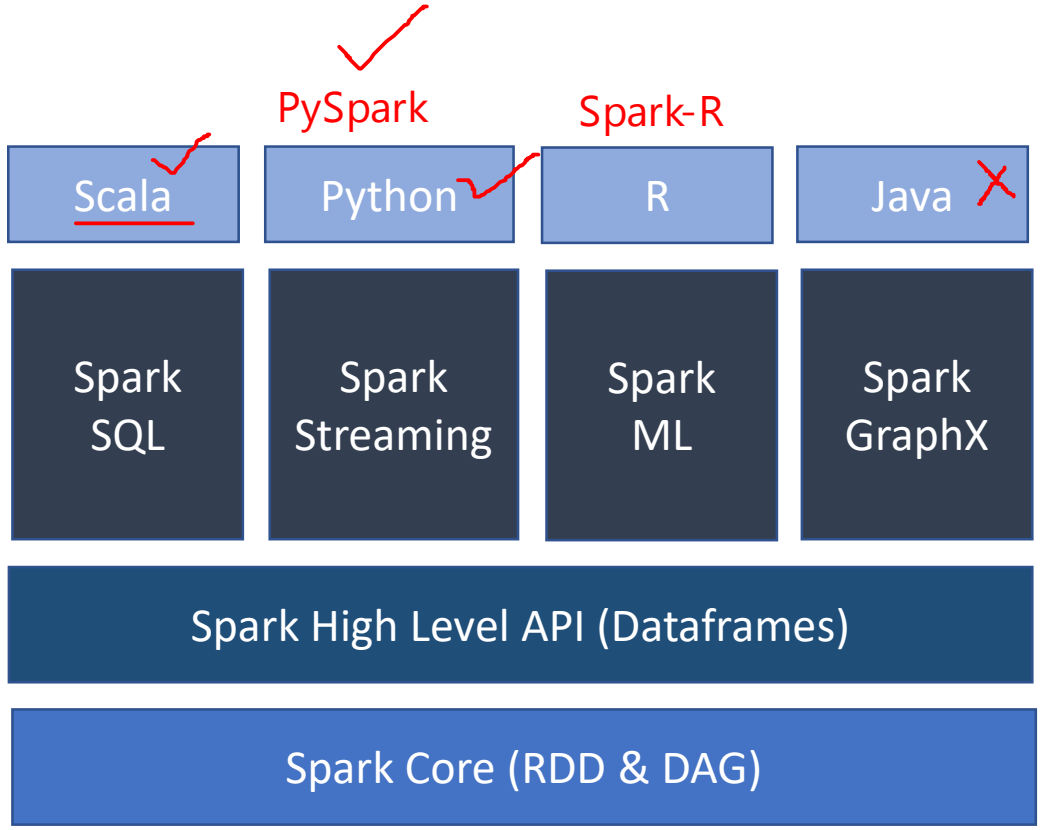
Apache Hive

- 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.



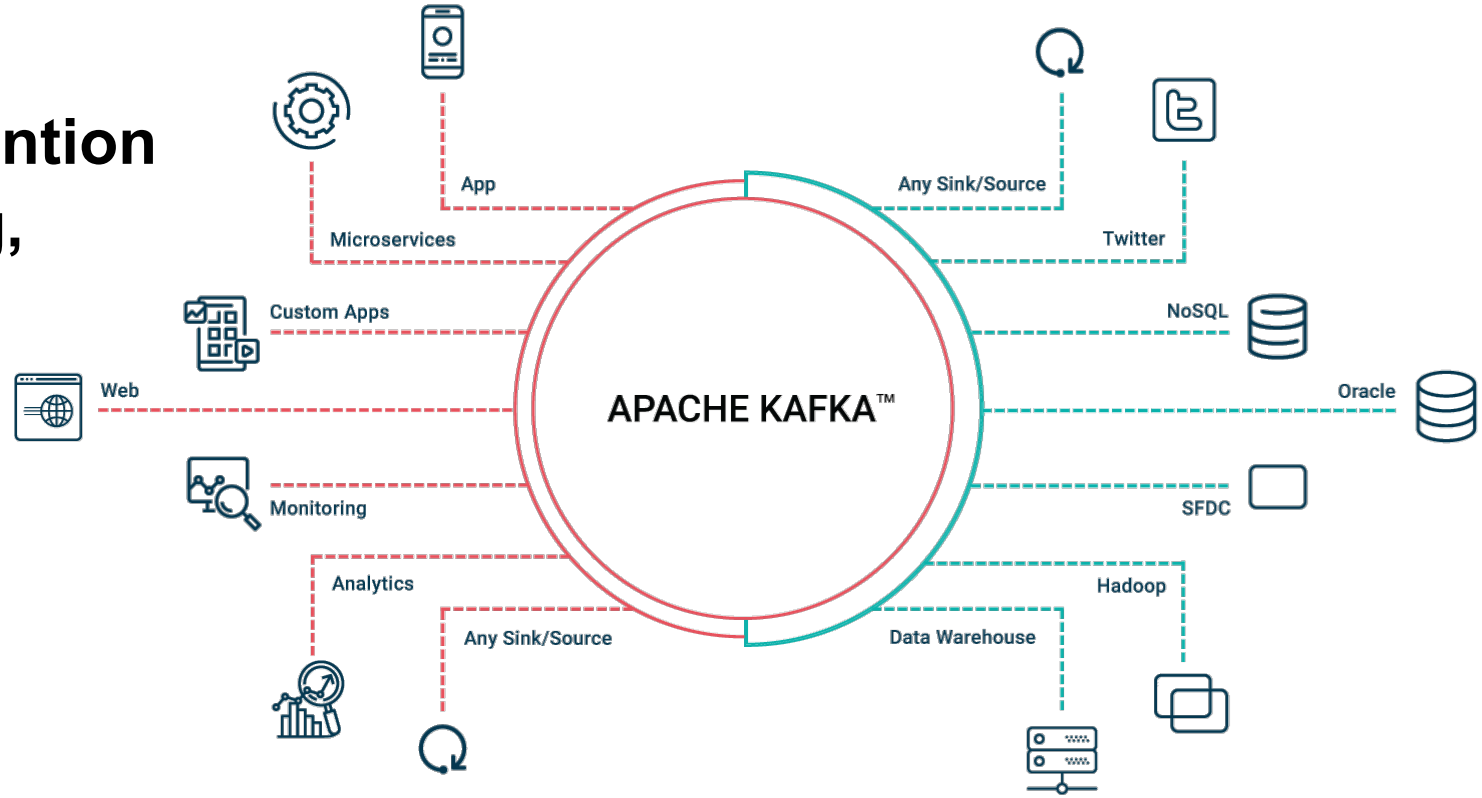
Apache Spark

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



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.



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**






Que : Which of the following is used for live data Processing?

A. Batch

B. Stream

C. Unit

D. Query

MCQ

Que : Which of the following is used for live data Processing?

A. Batch

B. Stream

C. Unit

D. Query

Que : In Cloud computing _____ type of Virtualization is used.

A. Type I
B. Type II
C. Type A
D. Type B

MCQ

Que : In Cloud computing _____ type of Virtualization is used.

A. Type I
B. Type II
C. Type A
D. Type B

MCQ

Que : PaaS stands for _____

A. Python as-a Service

B. Platform as-a Service

C. Program as-a Service

D. Process as-a Service

MCQ

Que : PaaS stands for _____

A. Python as-a Service

B. Platform as-a Service

C. Program as-a Service

D. Process as-a Service

Que : Which of the programming language is not used for big data processing?

A. Python

B. Java

C. Scala

D. mongo

Que : Which of the programming language is not used for big data processing?

A. Python

B. Java

C. Scala

D. mongo

Que : Which of the following are incorrect Big Data Technologies?

A. Hadoop

B. Spark

C. Hive

D. SplunkLine

Que : Which of the following are incorrect Big Data Technologies?

A. Hadoop

B. Spark

C. Hive

D. Splunk Line

Thank you!

Pradnyaa S Dindorkar <pradnya@sunbeaminfo.com>