# 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.
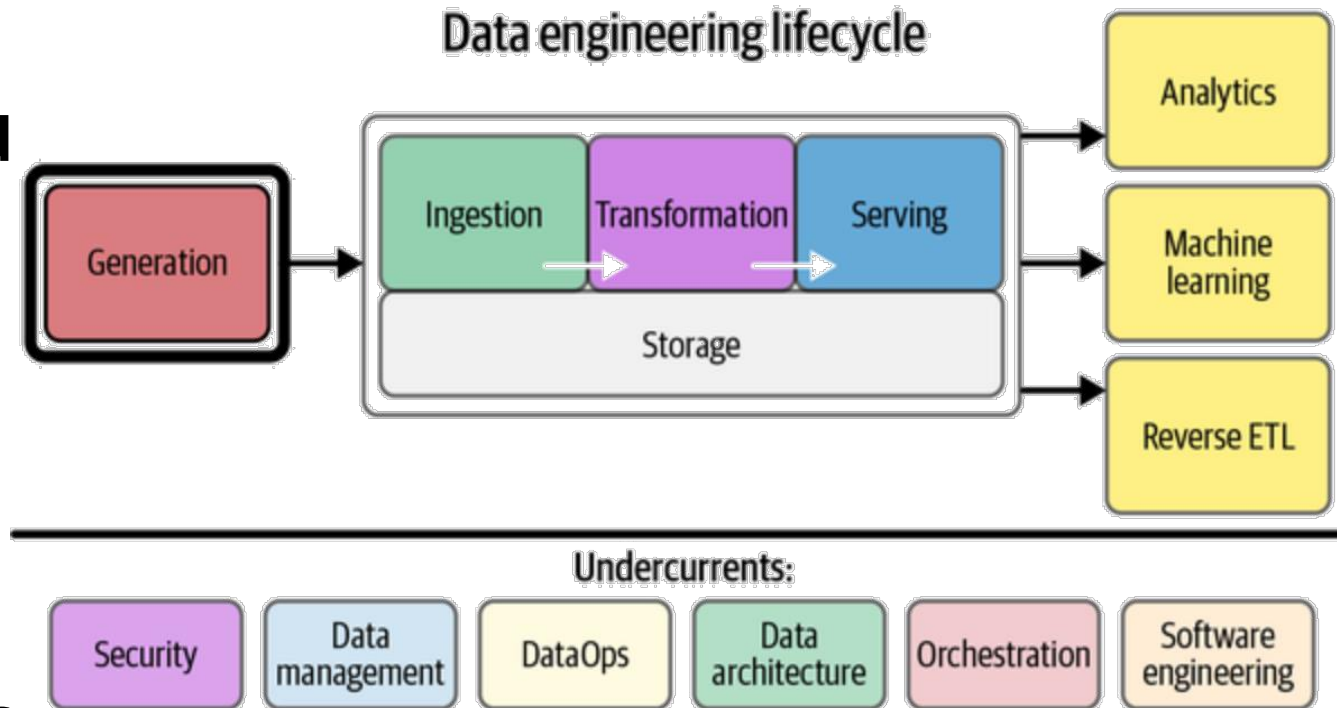
https://youtu.be/hZu_87l62J4

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.**
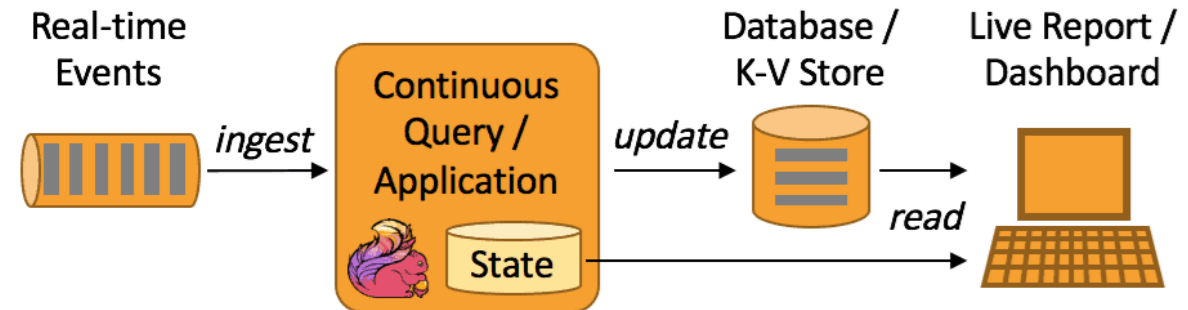
Batch processing vs Stream processing

- **Processing finite set of data (data at rest).**

- **Incremental data load is managed by programmer.**

- **Cluster planned as per data size. High throughput.**
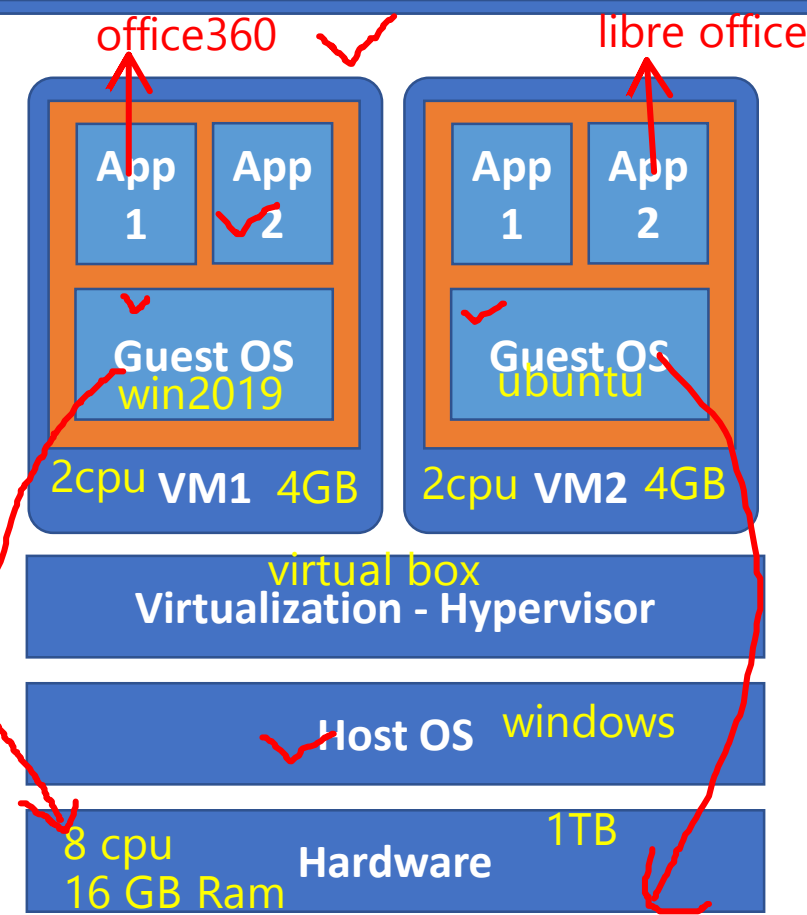
- **Job run once per batch.**
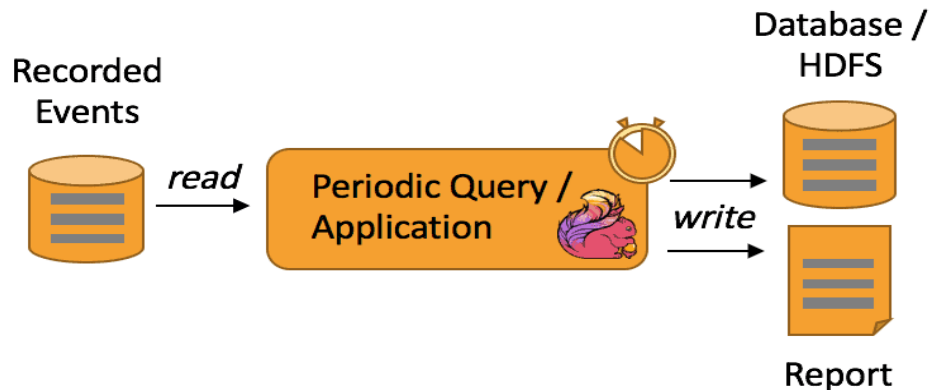
- **Processing live stream of data (data in motion).**

- **Data processing is managed by the framework.**

- **Less throughput.**

- **Job is running forever.**

Batch Processing
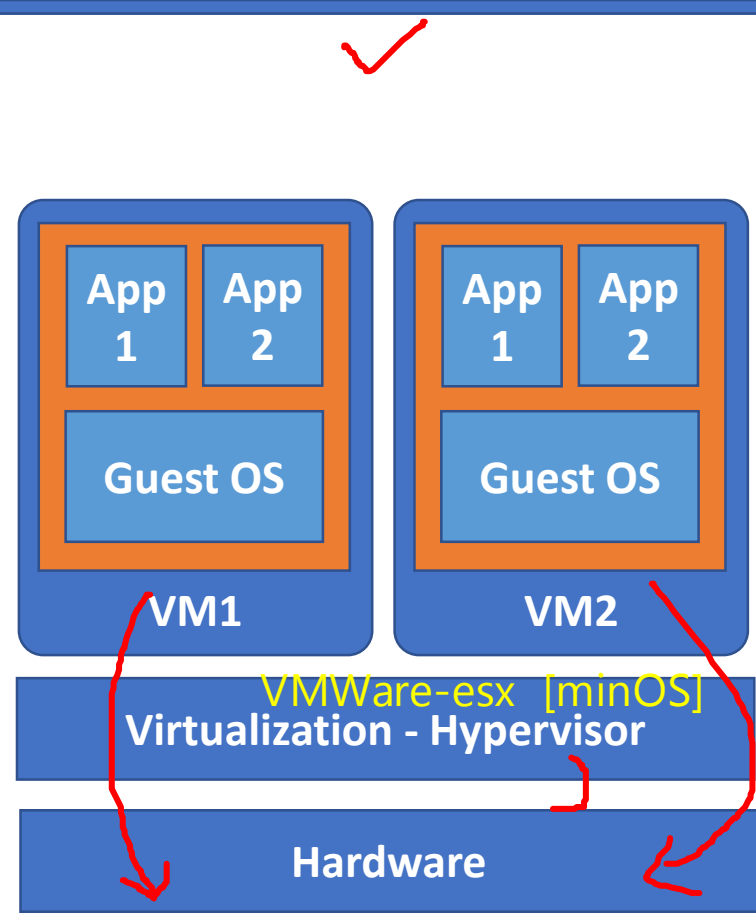
Recorded Events → read → Periodic Query / Application → write → Database / HDFS

Report

Stream Processing

Real-time Events → ingest → Continuous Query / Application (State) → update → Database / K-V Store → read → Live Report / Dashboard

Virtualization vs Containerization

Type-II Virtualization: office360, libre office, App 1, App 2, Guest OS win2019, Guest OS ubuntu, 2cpu VM1 4GB, 2cpu VM2 4GB, virtual box, Virtualization - Hypervisor, Host OS windows, 8 cpu 16 GB Ram, Hardware, 1TB

Type-I Virtualization: App 1, App 2, Guest OS, Guest OS, VM1, VM2, VMWare-esx [minOS], Virtualization - Hypervisor, Hardware

Containerization: App – 1 + Execution Environ, App – 2 + Execution Environ, Container 1, Container 2, docker, Containerization Platform, Host OS linux, cpu RAM Hardware Disk
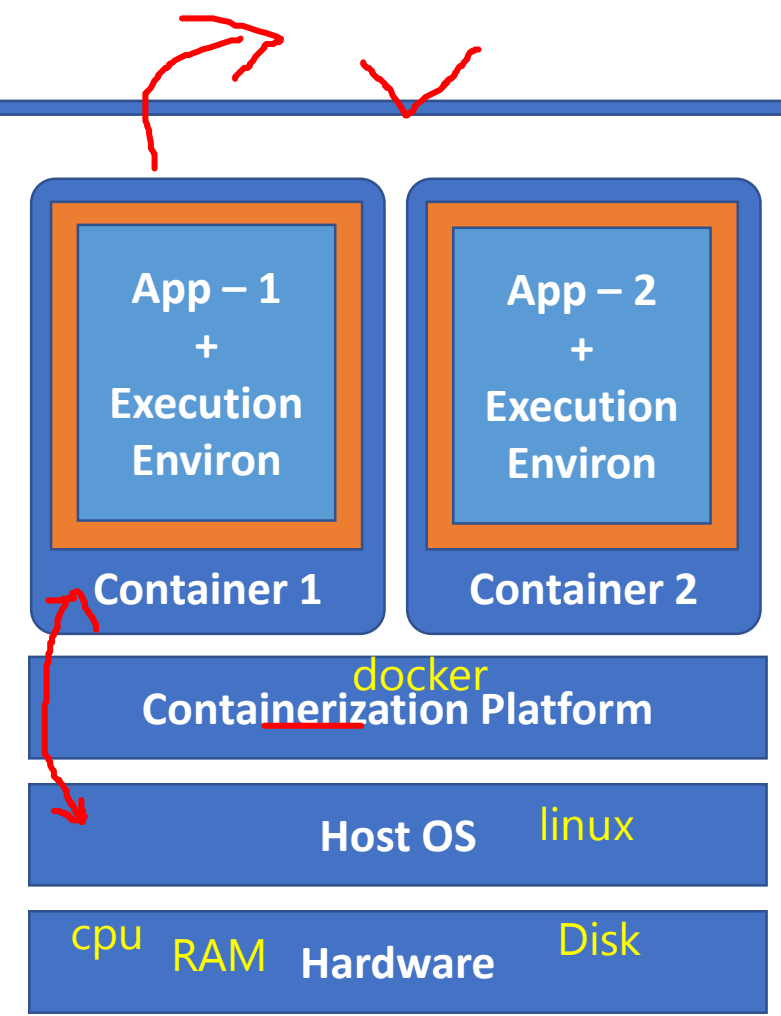
Type-II Virtualization

VMWare, VirtualBox, KVM, …

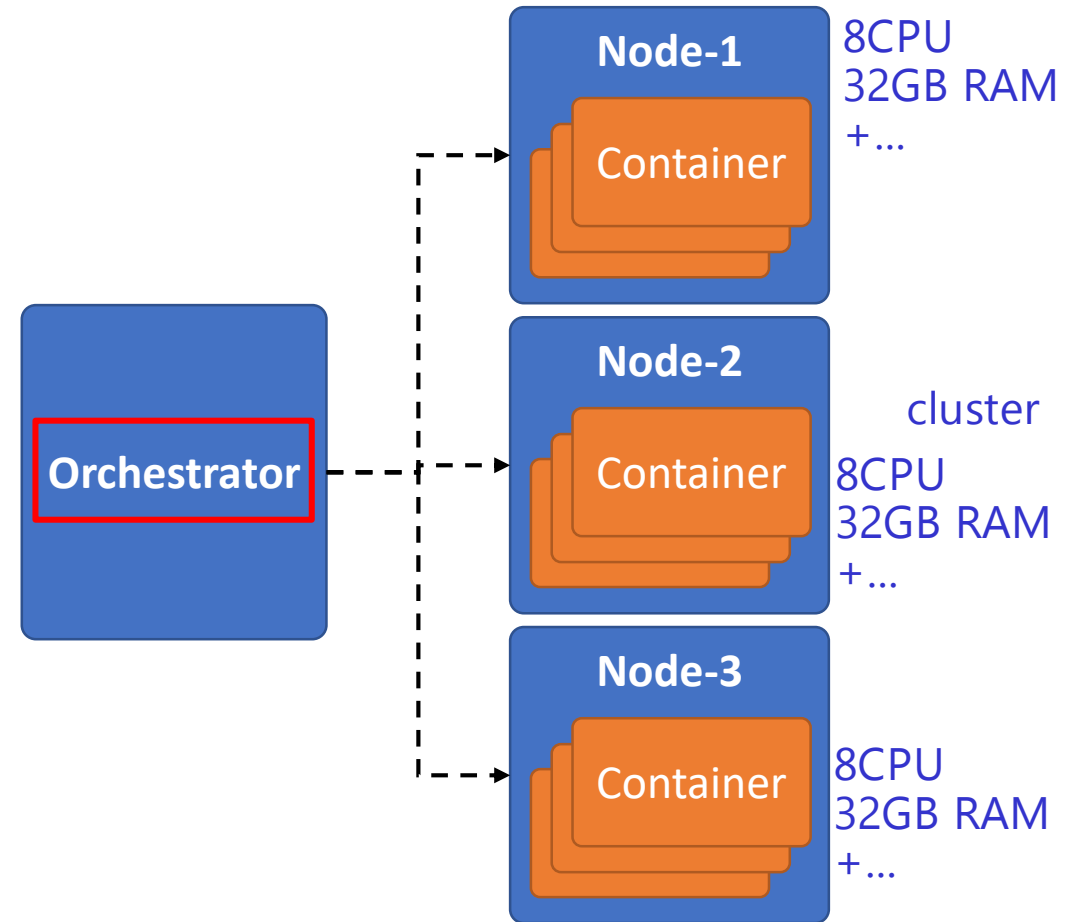**Type-I Virtualization**

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, …**

Node-1 8CPU 32GB RAM +...

Container

Orchestrator

cluster

Node-2

Container

8CPU 32GB RAM +...

Node-3

Container

8CPU 32GB RAM +...
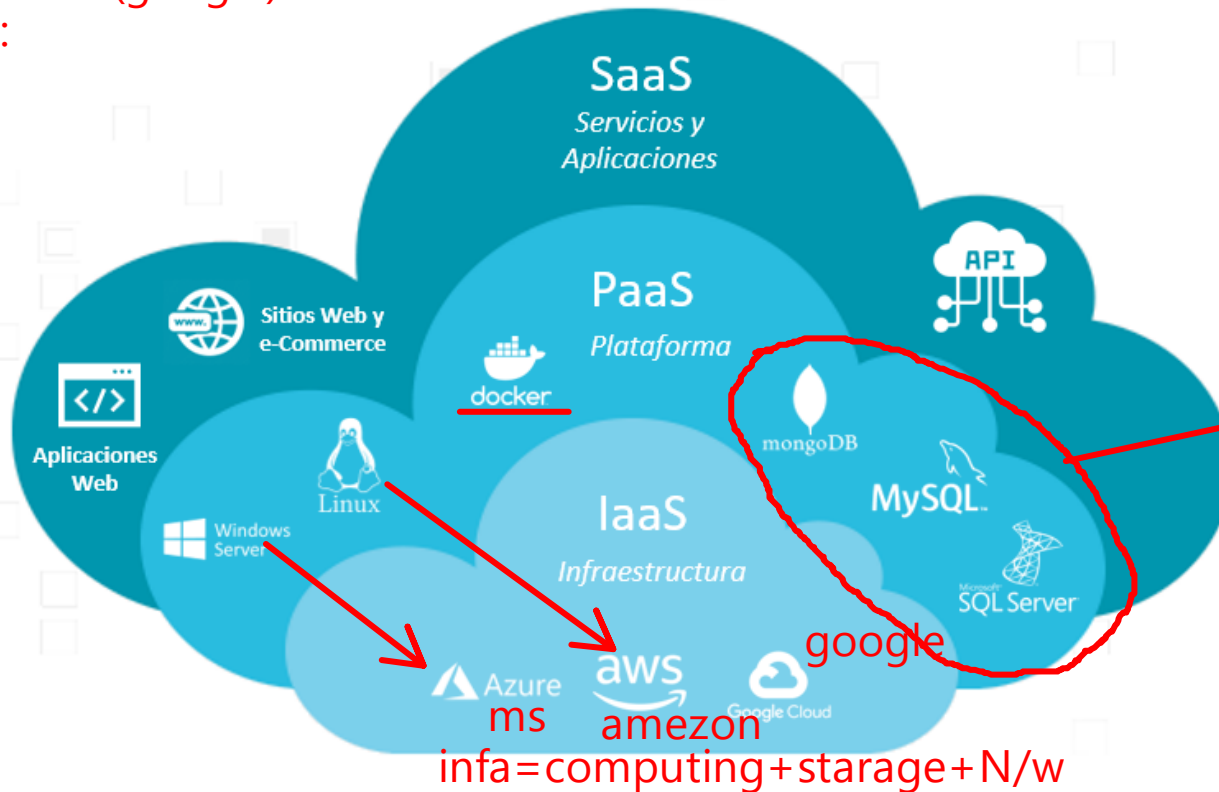
Scalability and Elasticity

- 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.

16 CPU

128GB RAM

32TB DSK

1 CPU
4GB RAM
1TB DSK

Cluster
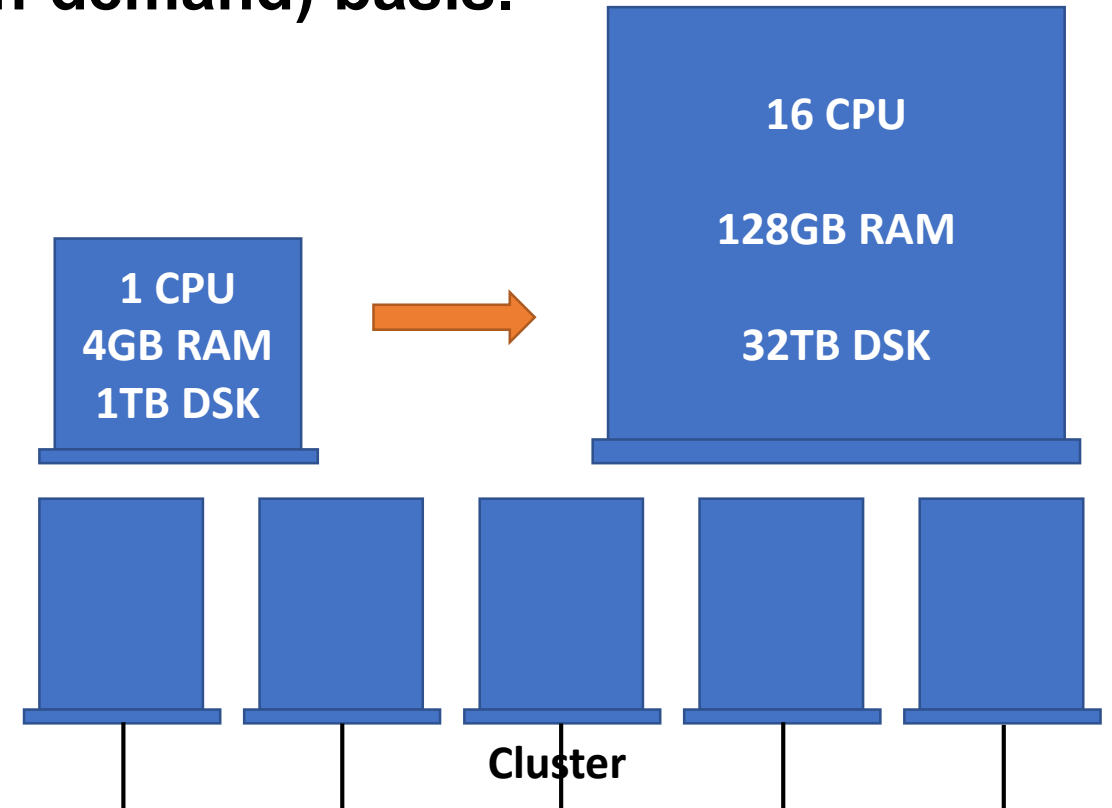
Cloud Service Models

cloud provider
1: AWS (Amezon)
2: Azure (MC)
3: GCP(google)
4:

ms   amezon   google
infa=computing+starage+N/w

VM
- **IaaS: Infrastructure as-a Service**
  - **AWS EC2, S3, VPC**

- **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**

RDBMS   mySQL   AWS

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

mongoDB
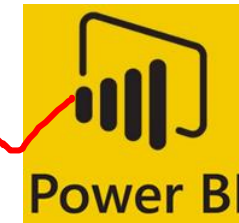
Big Data & Analytics Spectrum

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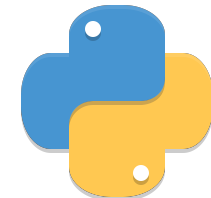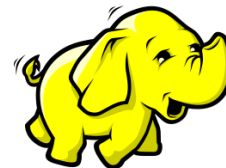 - Business reports
- **Artificial Intelligence, Data Science & Data mining**
  - Mathematics, Statistics & Computer algorithms
  - Machine learning & Deep learning
  - R Programming, Python
- **(Big) Data Engineering**
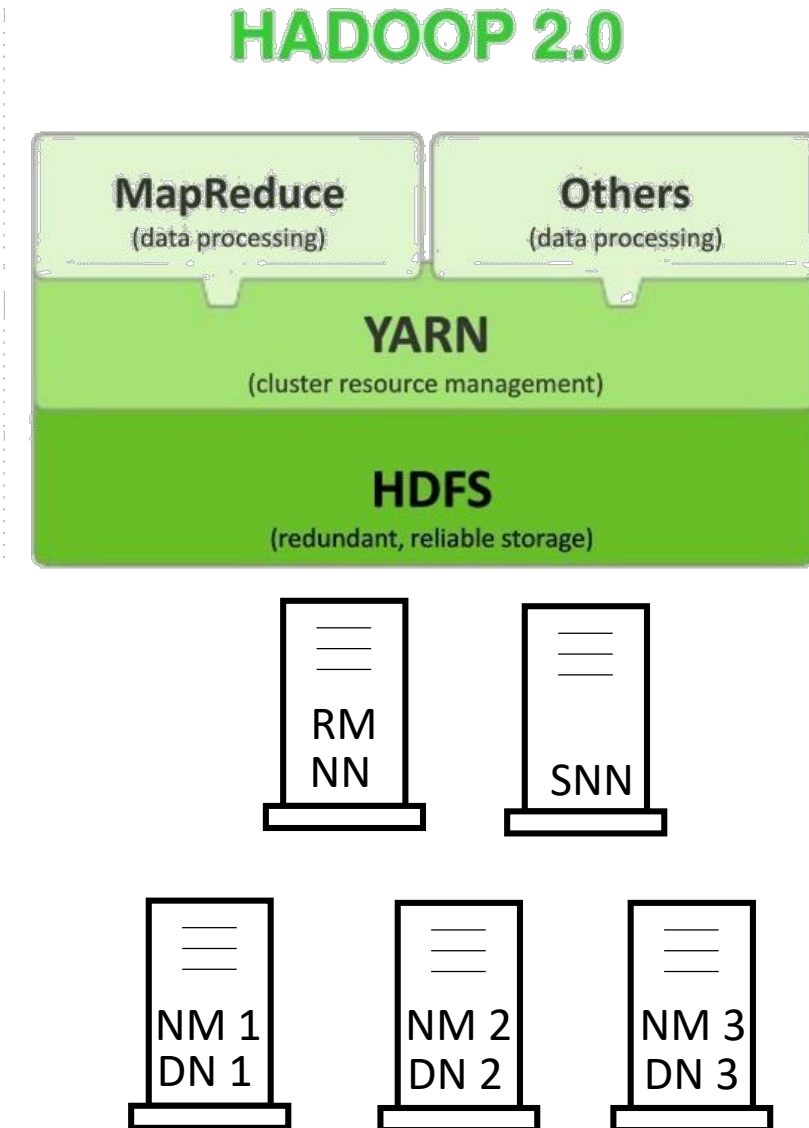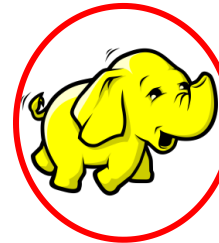 - Hadoop, Hive, Spark, Kafka, BigTable, …
  - Java, Scala, Python.
- **Infrastructure**
  - Linux, Cloud Computing
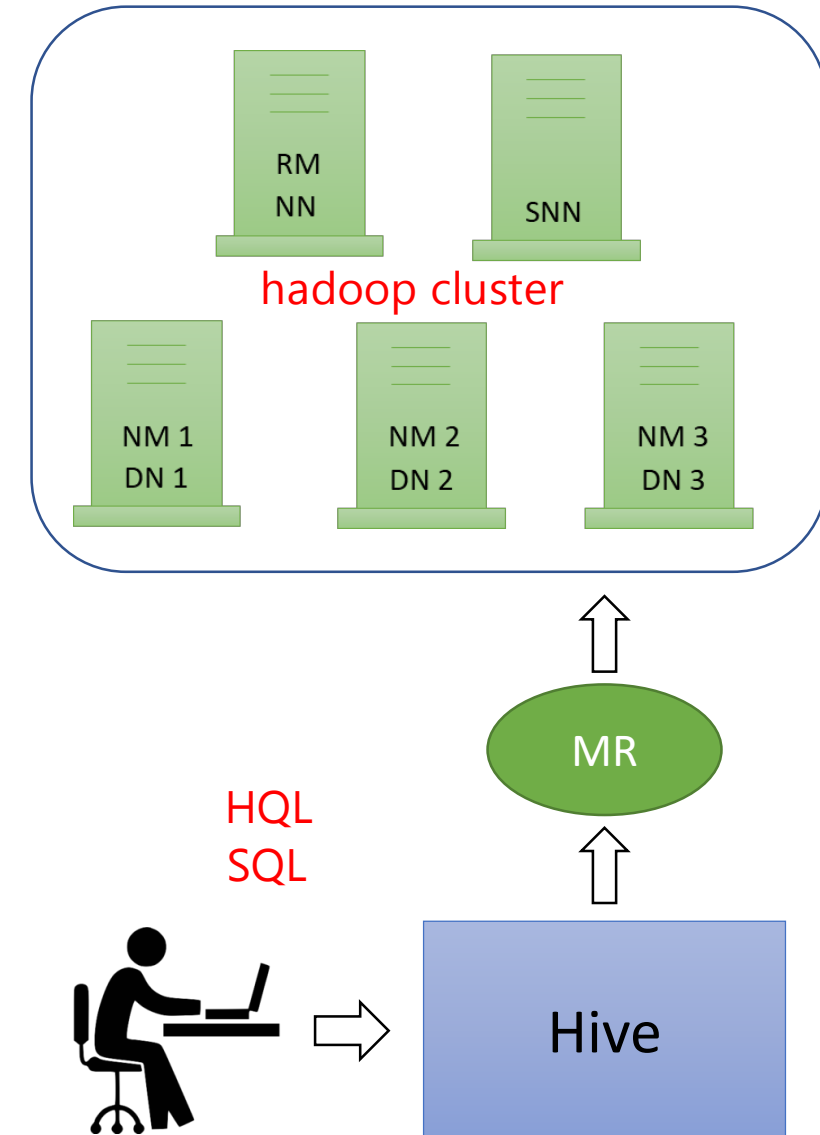
Apache Hadoop

- # Hadoop is developed by <u>Doug cutting.</u>
  - ## 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 dist. file system
  - ## Distributed storage: HDFS
  java
  - ## 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).
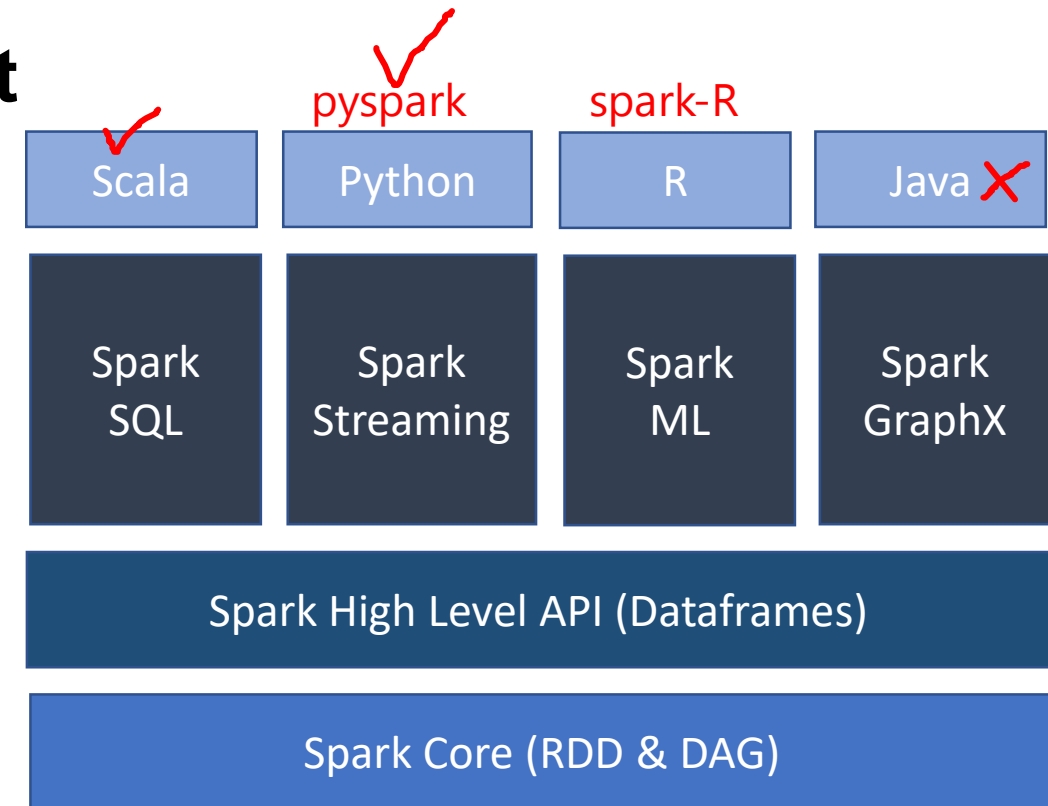
HADOOP 2.0

| MapReduce (data processing) | Others (data processing) |

YARN (cluster resource management)

HDFS (redundant, reliable storage)

RM NN

SNN

NM 1 DN 1

NM 2 DN 2

NM 3 DN 3

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.**

RM
NN

SNN

hadoop cluster

NM 1
DN 1

NM 2
DN 2

NM 3
DN 3

MR

HQL
SQL

Hive

Apache Spark

- **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.**

- **Further developed/maintained by DataBricks.**

- **Popular Spark vendors**
  - **DataBricks, AWS EMR, Cloudera, MapR**

- **Spark Toolkit**

pyspark   spark-R

Scala	Python	R	Java ✗
Spark SQL	Spark Streaming	Spark ML	Spark GraphX

Spark High Level API (Dataframes)

Spark Core (RDD & DAG)

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.**
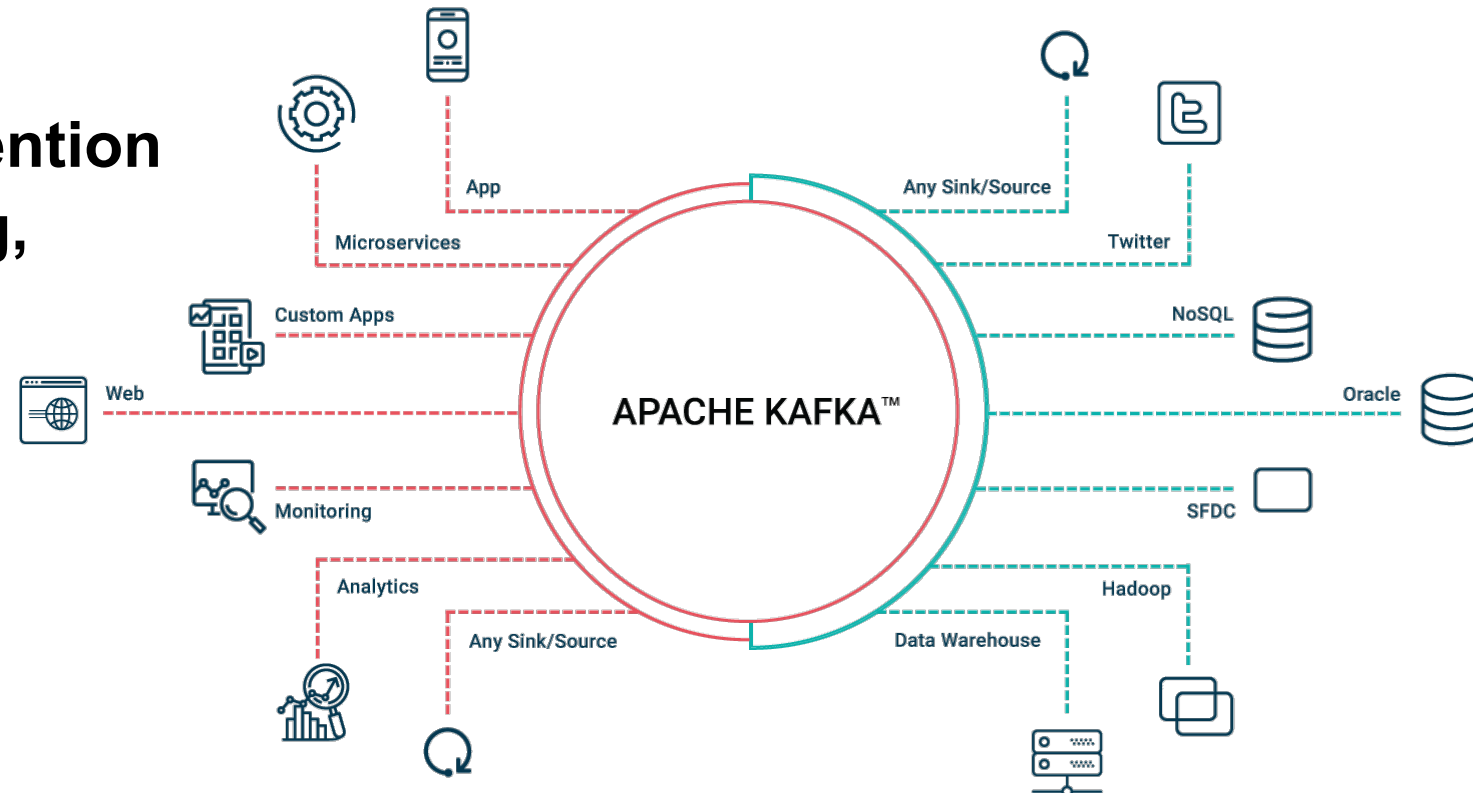
Big Data domains & opportunities

- **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

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

A. Batch

B. Stream

C. Unit

D. Query

MCQ

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

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

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

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

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 : 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. Splunk

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

A. Hadoop

B. Spark

C. Hive

D. Splunk

Thank you!

Pradnyaa S Dindorkar <pradnya@sunbeaminfo.com>