1. You have a company and you have an information system to deploy. When do you choose to move to the cloud?

- Cost Reduction: You want to reduce upfront costs by avoiding buying and maintaining hardware.
- Improved Accessibility: You want better accessibility so employees can access the system from anywhere.
- Managed Maintenance: You prefer the cloud provider to handle maintenance, updates, and security.

2. Assume now that you have a server and two virtual machines are created on this server using VMWare.

a. Why do we call it a virtual machine?

It's called a virtual machine because it's not a real physical computer. It's a software-based copy created by a program like VMWare. It runs inside a real computer and uses part of its resources (like CPU, RAM, storage) to act like a separate computer.

b. In order for the server to be virtualized and accessed by several VM, what is the main condition for the server to satisfy?

The **main condition** is that the **server hardware supports virtualization** so the hypervisor can create multiple VMs.

c. Is the above virtualization called OS or Hardware virtualization? Explain.

Hardware virtualization. Because VMWare creates multiple virtual machines on one physical server using a hypervisor that manages hardware resources.

d. Given that the virtualization is hosted - full virtualization. How does an instruction access the CPU hardware?

In full virtualization, non-critical instructions run directly on the CPU hardware without interruption. Note : if the next part do not ask about critical instructions , talk about critical instructions here.

e. In case the instruction to be run is critical, how does the call access the hardware, explain the whole mechanism. (5 pts)

In full virtualization, critical instructions are trapped by the hypervisor (VMM), which then emulates or handles it before allowing access to the CPU.

f. What is the part of the server responsible of managing the memory access of a virtual machine? (5 pts)

Virtual Machine Monitor (VMM), also called the Hypervisor.

3. Give an example of one PaaS, one IaaS, one SaaS. When will you need to use each one of them? For each one of them, who is the Cloud provider and who is the Customer/user.

SaaS (Software as a Service) :

- Example: Microsoft Teams
- Cloud Provider: Microsoft
- Users: Students, employees
- When to use: When we want to use ready-made software without installing or managing it.

Paas (Platform as a Service) :

- Example: Google App Engine
- Cloud Provider: Google
- Users: Developers
- When to use: When we want to build and deploy applications without managing the underlying servers

IaaS (Infrastructure as a Service) :

- Example: Amazon EC2 (Elastic Compute Cloud)
- Cloud Provider: Amazon (AWS)
- Users: IT admins, developers
- When to use: When we want to control and manage servers, storage, and networking without buying physical hardware

4. Cloud Storage:

a. List three kinds of risks that exist in the architecture of cloud storage models. How you can protect your cloud against them?

Risks:

- Unauthorized Access to Data
- Data Loss
- Downtime or Unavailability

Protections:

- Use data encryption
- Set up automated backups and data replication
- Add failover systems to switch to backup servers automatically.

b. What are the two types of storage that exist in in the cloud? Explain them briefly.

- Object Storage: which is a file repository and is used to store individual files.
- Volume Storage: which is a virtual hard drive and where the volumes attach to virtual machines for use just like a physical hard drive or array.

5. What is the difference between a “Public Cloud” and a “Private Cloud”?

- Public Cloud: Cloud services are offered over the internet, accessible by anyone, and shared among multiple users. It is cheap, scalable, but offers less control over security and customization.
- Private Cloud: Cloud infrastructure is dedicated to a single organization, providing more control, privacy, and security, but it is more expensive.

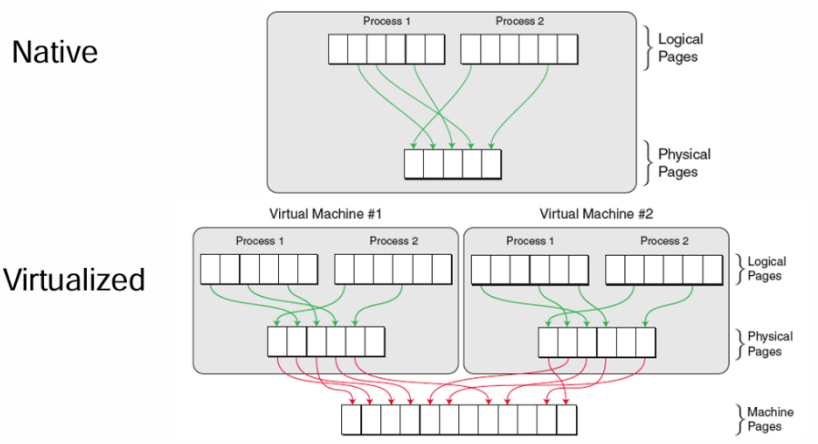
6. Assuming a traditional computer is composed of a hardware, host OS, and application. After virtualization, cite the components that will be added or modified and the role of these components.

- Hypervisor (VMM): Added between hardware and OS; manages virtual machines and hardware sharing.
- Guest Operating Systems: Multiple OSes running inside virtual machines instead of just one host OS.
- Virtual Machines (VMs): Isolated environments that emulate complete computers on the same physical hardware.

7. In Docker, What is an image? What is the difference between an image and a container? Explain.

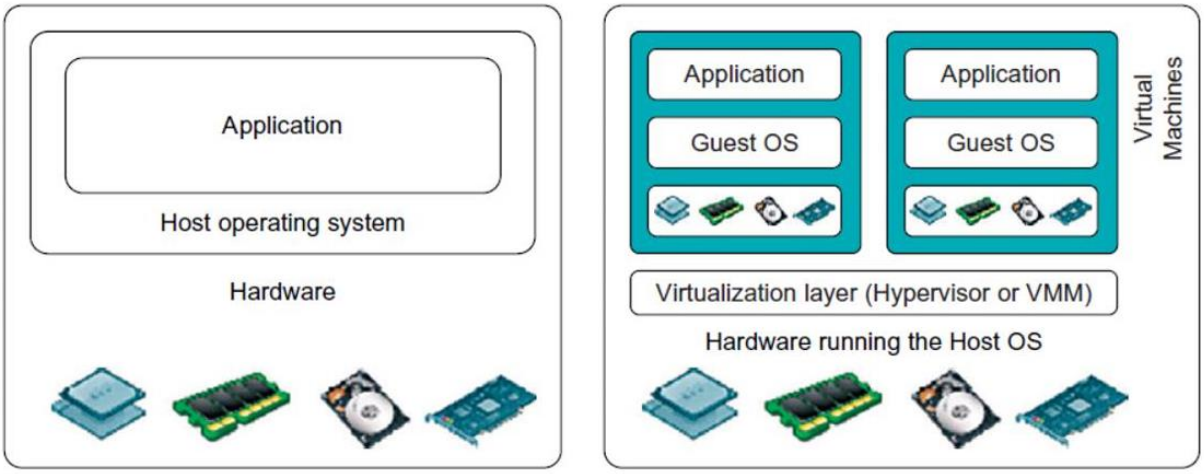
An image is a blueprint used to create containers. A container is a running instance of an image , it's the actual live environment where your app runs.

8. How can a memory be shared between two virtual machines? Explain with a figure.



9. What is the difference between a virtual machine and a real machine? Enhance your explanation with a simple figure.

- A real machine is a physical computer with actual hardware components like CPU, memory, and storage.
- A **Virtual Machine (VM)** is a **software-based emulation** of a physical computer. It is created and managed by a **hypervisor**, and runs its **own operating system and applications**, just like a real machine.



10. What is the role of the each of the following:

a. Cloud provider : They own and manage the cloud infrastructure and services, ensuring availability, security, and maintenance of resources that customers use.

b. Customer / user : They use the cloud services to run applications, store data, and access computing resources without managing the underlying infrastructure.

Part 2 : Big Data

1. Explain the key components of the Hadoop ecosystem and their roles in the data processing pipeline.

- HDFS: Stores big data across many computers.
- MapReduce: Processes the big data by splitting work into small parts.
- Hive: Lets you use SQL to ask questions about the data.
- Pig: Helps write easy scripts to handle data instead of coding.
- HBase: A fast database to read and write data quickly on Hadoop.

2. Discuss the MapReduce programming model and its advantages in processing large-scale data.

- Map: Reads and splits the data into key-value pairs.
- Reduce: Aggregates or processes all values that have the same key to produce the final output.

Advantages :

- Can handle very big data easily by splitting work across many machines.
- Works in parallel, so it's fast and efficient.

3. Illustrate the concept of data shuffling and sorting in MapReduce. Explain why it is a crucial step in the MapReduce process.

- After the Map step, data is shuffled so that all values with the same key go to the same place.
- The data is also sorted by key during this move.
- This helps the Reduce step to easily combine all values with the same key.
  Why important: Because it groups the data so the reducer can work on it correctly.

4. Map Reduce Code for a given question (Session 2024) :

```python
def mapper():
    for line in sys.stdin:
        data = line.strip().split("\t")

        if len(data) == 5:
            transactionId, customerId, productId, purchaseQuantity, timestamp = data
            print(f"{productId}\t{purchaseQuantity}")
```

```python
def reducer():
    purchaseQuantity = 0
    oldKey = None
    count = 0

    for line in sys.stdin:
        data = line.strip().split("\t")

        if len(data) != 2:
            continue

        thisKey, thisQuantity = data
        thisQuantity = int(thisQuantity)

        if oldKey and oldKey != thisKey:
            average = purchaseQuantity / count
            print(f"{oldKey}\t{average:.2f}")
            purchaseQuantity = 0
            count = 0

        oldKey = thisKey
        purchaseQuantity += thisQuantity
        count += 1

    # Print last key average
    if oldKey is not None:
        average = purchaseQuantity / count
        print(f"{oldKey}\t{average:.2f}")
```

5. Commands used in assignement :

```
1. Command to Create a Directory in HDFS :
- hdfs dfs -mkdir /input

2. Command to Upload a File into HDFS :
(Assume the file is data.txt on your local machine)
- hdfs dfs -put data.txt /input

3. Run the MapReduce Job Using Hadoop Streaming :
(Assuming you have mapper.py and reducer.py in your local directory)
- hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
    -input /input \
    -output /output \
    -mapper mapper.py \
    -reducer reducer.py \
    -file mapper.py \
    -file reducer.py

4. List the Output Files
- hdfs dfs -ls /output

5. View the Output Result
- hdfs dfs -cat /output/part-00000
```

6. Explain the purpose and functionality of Hadoop Distributed File System (HDFS). How does it provide fault tolerance and data reliability?

   HDFS is a distributed file system that stores large files across multiple machines in a Hadoop cluster. It splits files into blocks and stores multiple copies of each block on different DataNodes.

   Fault Tolerance and Data Reliability: HDFS replicates each data block (usually 3 copies) on different nodes so if one node fails, the data can be read from another node .

7. Discuss the role of NameNode and DataNode in HDFS architecture. How do they collaborate to manage the file system?

   NameNode: The master nodes that manages the file system's metadata (file names, directories, and block locations). It keeps track of where all data blocks are stored in the cluster.

   DataNode: The worker nodes that actually store the data blocks. They handle read and write requests from clients.

   Collaboration : The NameNode tells clients where data blocks are, and DataNodes store the actual data and report back to the NameNode.

8. Then list three problems that you can face with them and how we can solve them?

   DataNode Failure:

   - **Problem:** If a DataNode fails, some data blocks are lost.
   - **Solution:** Hadoop keeps **3 copies** of each block. If one is lost, it copies from another.

   NameNode Disk Failure

   - **Problem:** If the NameNode's hard disk fails, metadata is lost.
   - **Solution:** Save metadata on a **Network File System (NFS)** as a backup.

NameNode Crash (Single Point of Failure)

- **Problem:** If the NameNode stops working, the whole system goes down.
- **Solution:** Use **2 NameNodes** (active + standby) for **High Availability**.

9. Describe the Hadoop cluster setup process, including the configuration of master and worker nodes. What factors should be considered when determining the cluster size?

Cluster Setup Process:

- Install Hadoop and JAVA on machine.
- Edit configuration files (core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml).
- Designate a **master node** (NameNode, ResourceManager) and **worker nodes** (DataNodes, NodeManagers).
- Format the HDFS using **hdfs namenode -format** and start Hadoop daemons **using start-dfs.sh** and **start-yarn.sh.**

Factors :

Amount of data to process and store. Speed and number of tasks needed. Available hardware (CPU, memory, storage).

10. Discuss the benefits and challenges of processing unstructured data using Hadoop MapReduce.

Benefits:

- Can handle very large and diverse datasets without needing a fixed format.
- Fault tolerant due to data replication.
- Supports **parallel processing**, allowing tasks to run simultaneously on different nodes, improving performance

Challenges:

- Writing MapReduce code for unstructured data can be complex.
- Processing can be slower because of multiple read/write steps.
- Requires understanding the structure of data for accurate results.

11. When you talk about Big Data, you often hear the term “three Vs”. What is the meaning of each V and to what it refers?

Three Vs of Big Data:

- **Volume:** How much data there is.
- **Velocity:** How fast the data is created and used.
- **Variety:** Different types of data (like text, images, videos).