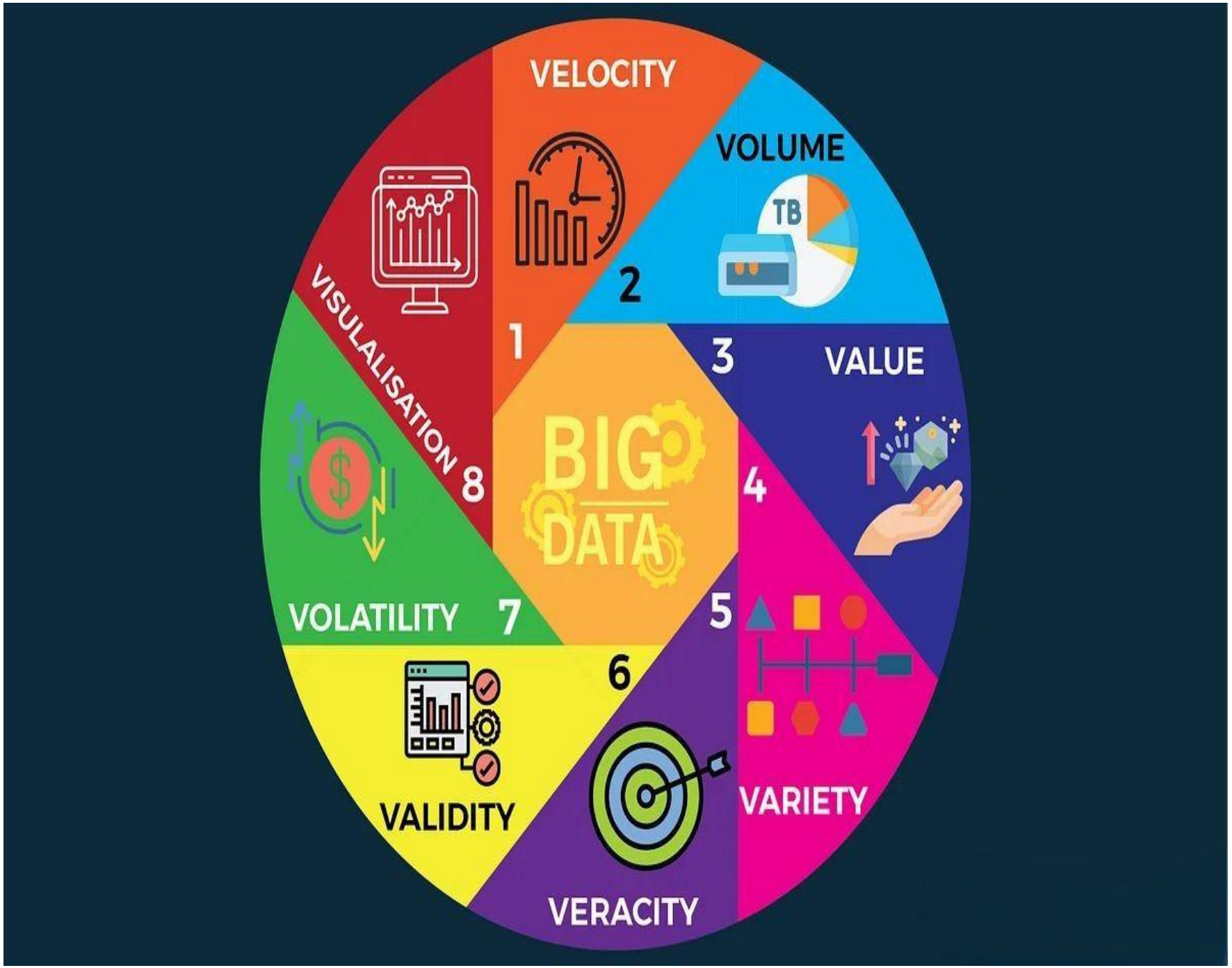
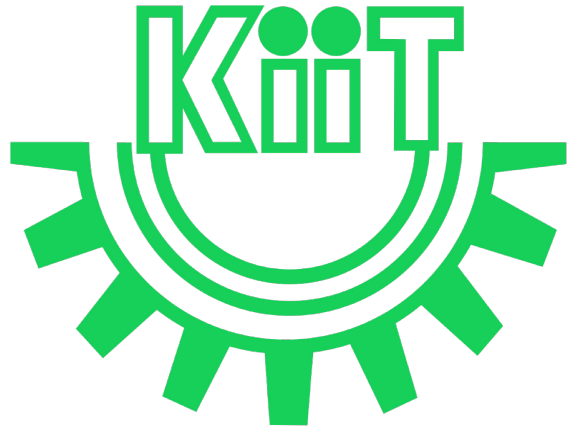


CS 3032: Big Data

Lec-8

In this Discussion . . .

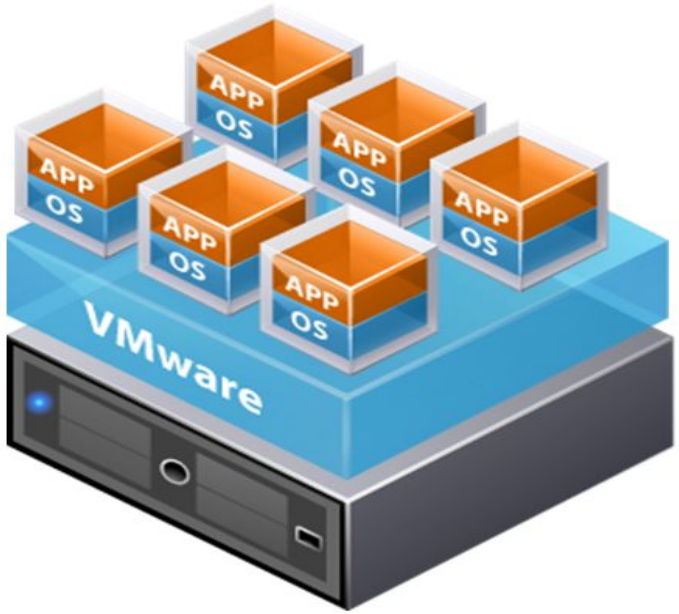
- Virtualization
  - What is Virtualization
  - How Virtualization works
  - Types of Hypervisor
  - Virtualization Types
  - Virtualization Benefits
- Data Streams
  - Common Examples of Data Stream Sources
  - Batch Vs. Real Time Streaming Processing
  - Basic Model of Stream Data



What is Virtualization?

- Virtualization is the process of creating multiple virtual machines/operating system from one physical hardware box
- Each machine works independently from each other and have their own Operating system.

Traditional Architecture

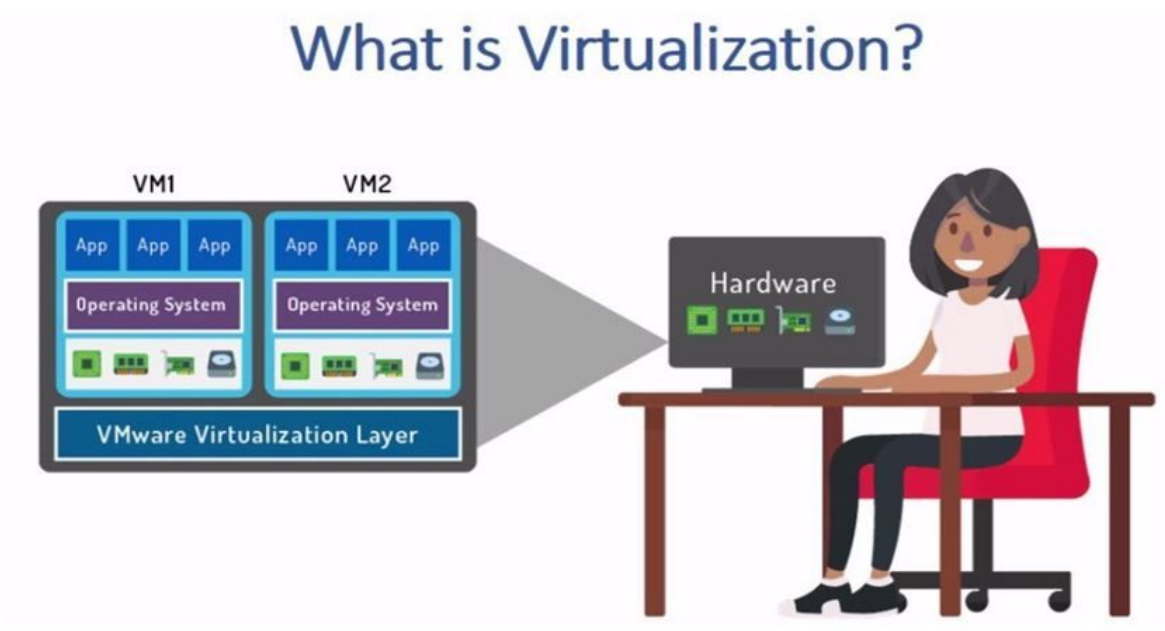


Virtual Architecture

What is Virtualization?

- Virtualization refers to the virtualization of a computer into multiple logical computers through virtualization technology.
- The objects of virtualization can include the virtualization of servers, the Internet, desktops, and archive spaces.
- In simple words- when we run multiple operating systems on single hardware by virtualizing CPU, storage and RAM is called **virtualization**.

Virtualization solutions are being implemented to use less hardware through automation, build new servers in minutes.



What is Virtualization?

- Each Virtual machine has the following Virtual hardware:

Virtual CPU	it is the physical CPU which is allocated to virtual machine
Virtual storage	it is the pooling of multiple physical storage into a single storage source
Virtual network	There are two virtual network: i) One connect the virtual machine inside the hypervisor; ii) protocol base virtual network like VLAN

What is Virtualization ? (Alternate Definition)

- It is a process to run the images of multiple operating systems on a physical computer.
- The images of the operating systems are called virtual machines (VMs).

- A **virtual machine** (VM) is basically a software representation of a physical machine that can execute or perform the same functions as the physical machines.

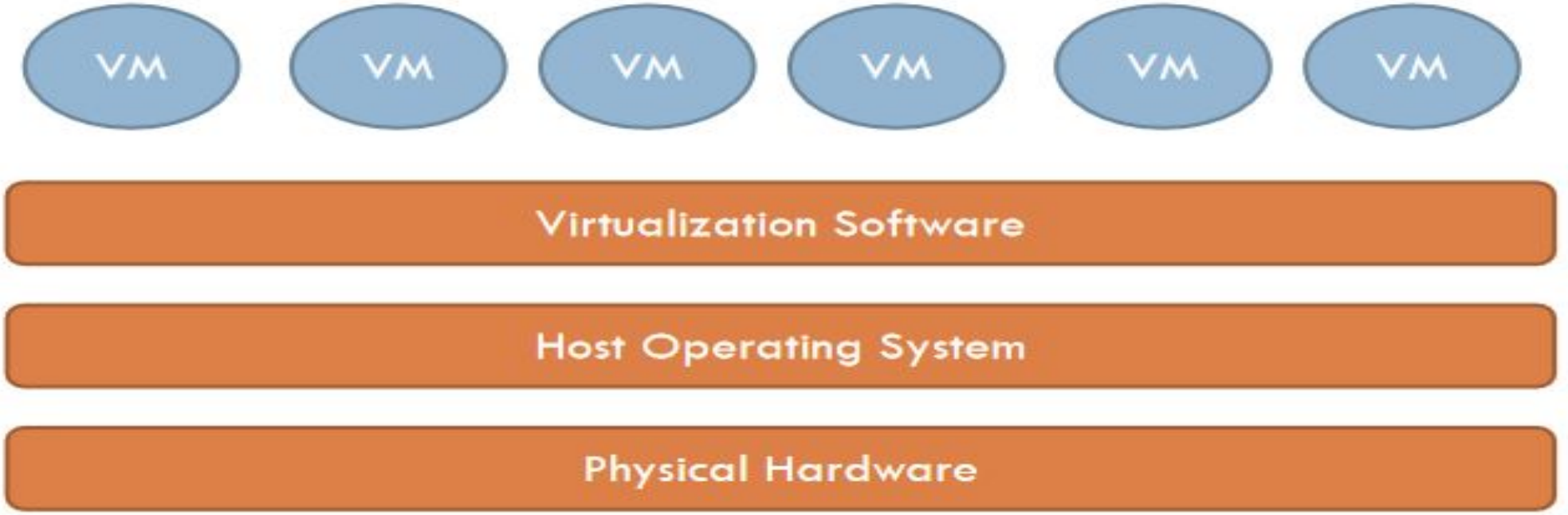
What is Virtualization ?

- Although virtualization is not a requirement for Big Data analysis, but works efficiently in a virtualized environment.
- **Server virtualization** is the process in which multiple operating systems (OS) and applications run on the same server at the same time, as opposed to one server running one operating system.

What is Server Virtualization ?

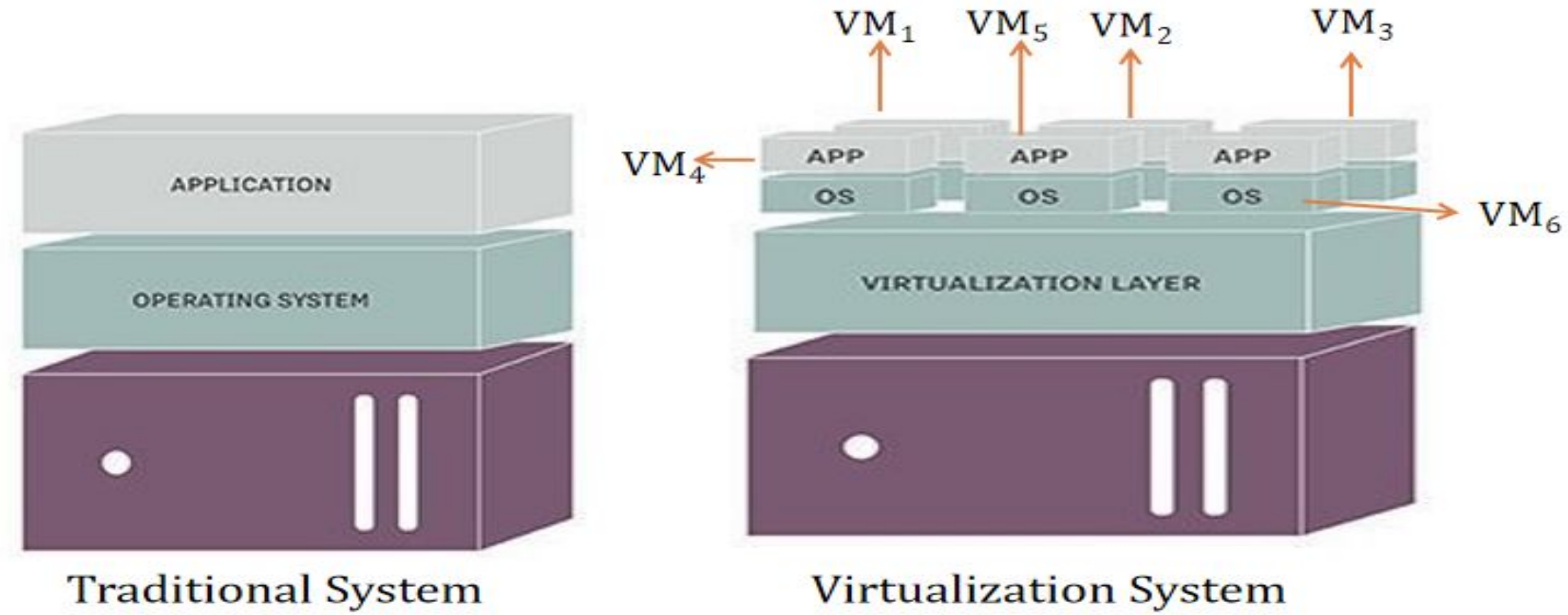
- If that still seems confusing, think of it as one large server being cut into pieces.
- The server then imitates or pretends to be multiple servers on the network when in reality it’s only one.
- This offers companies the capability to utilize their resources efficiently and lowers the overall costs that come with maintaining servers.

Virtualization Environment



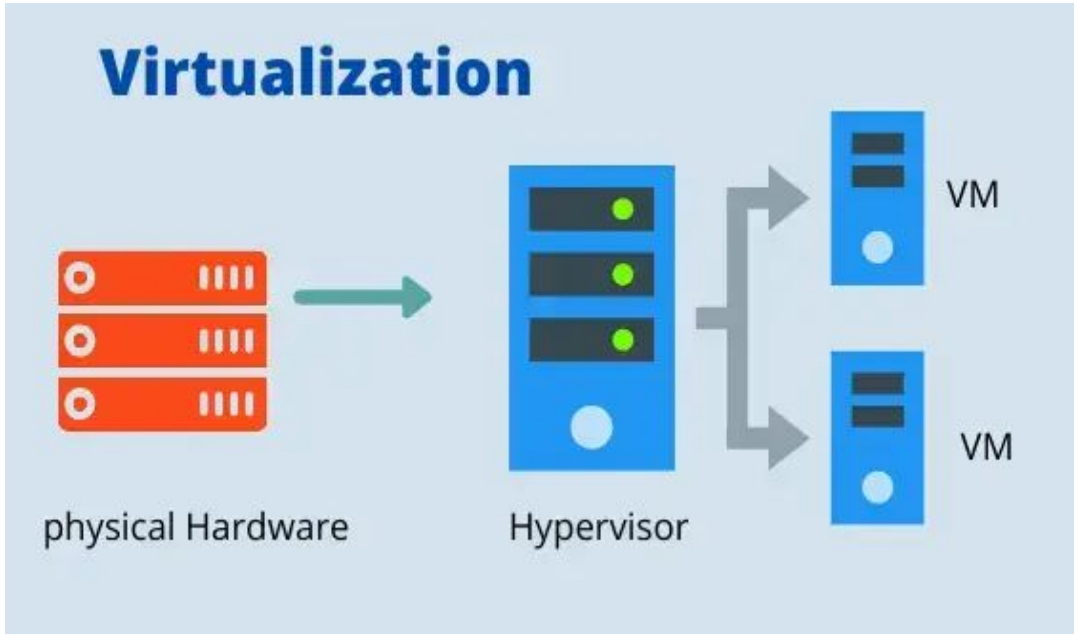
The operating system that runs as a virtual machine is known as the **guest**, while the operating system that runs the virtual machine is known as the **host**. **A guest operating system runs on a hardware virtualization layer, which is at the top of the hardware of a physical machine**

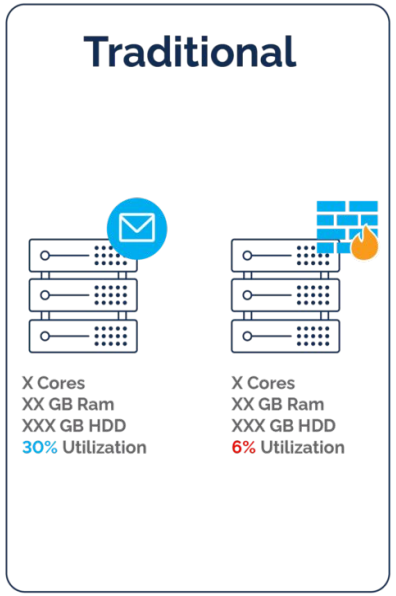
Traditional Vs. Virtualization System

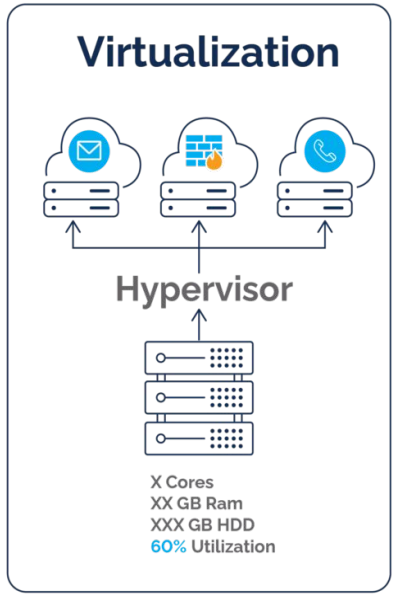


How Virtualization Works

- A software component **hypervisor** is installed on the physical hardware
- Now the hypervisor creates virtual platforms on the physical hardware on top of which multiple Operating system are installed and monitored
- These virtual platforms are called virtual machine (VMs) or virtual hardware







How Virtualization Works

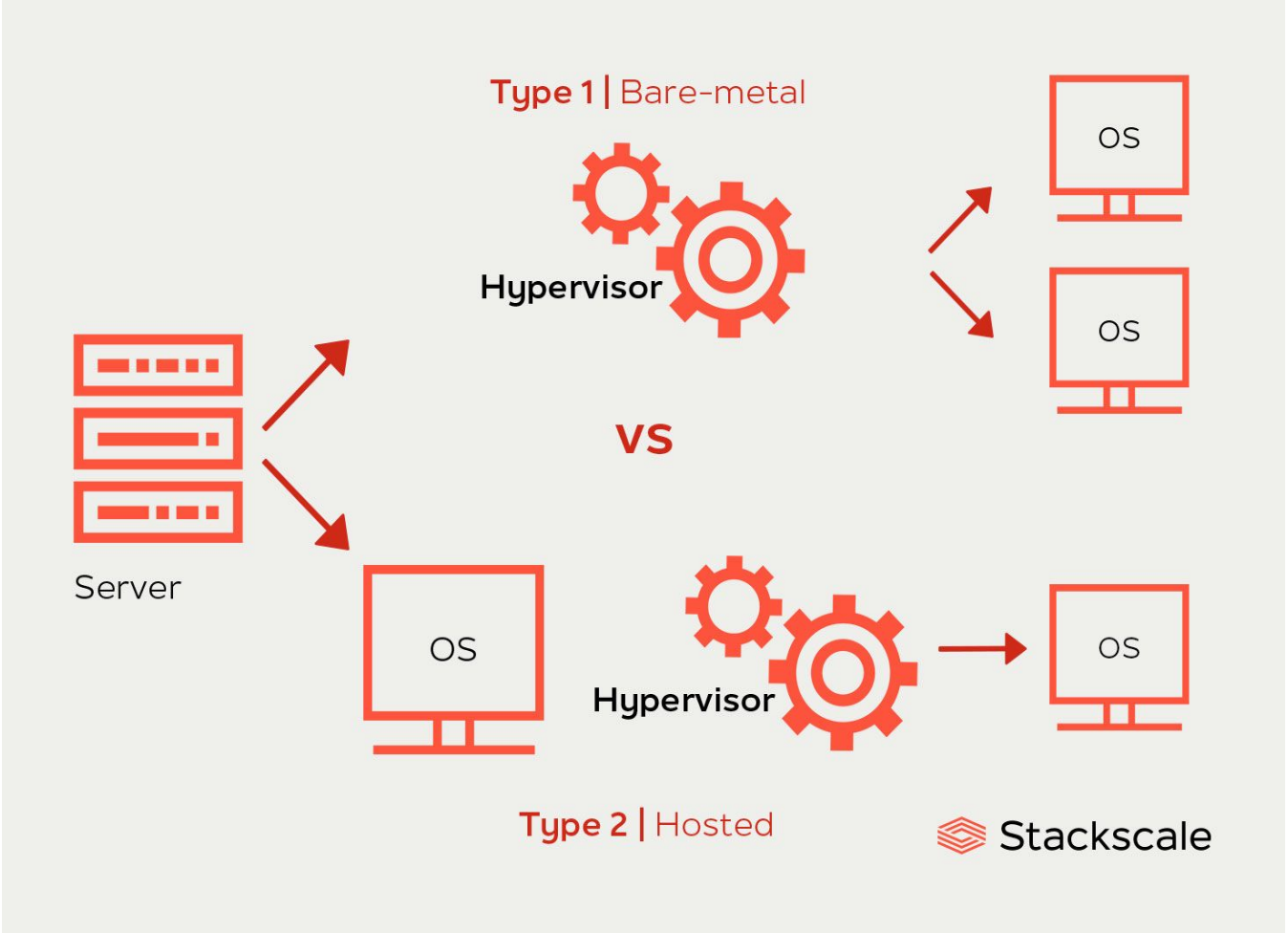
- Hypervisor is also called Virtual machine monitor.
- Guest operating system are then installed on this virtual hardware and the guest OS sees the virtual hardware as if they are its native, actual hardware components
- Next our apps run on these Guest operating system.

Types of Hypervisors

- There are two ways in hypervisor can be installed:

Bare metal or Native or type-1 Hypervisor:

- ❖ In this case, the hypervisor runs directly on the physical machine, which creates and monitors the guest operating system.
- ❖ Guest operating runs on the separate layer above the hypervisor.
- ❖ Ex- VMware ESX and ESXi, Oracle VM

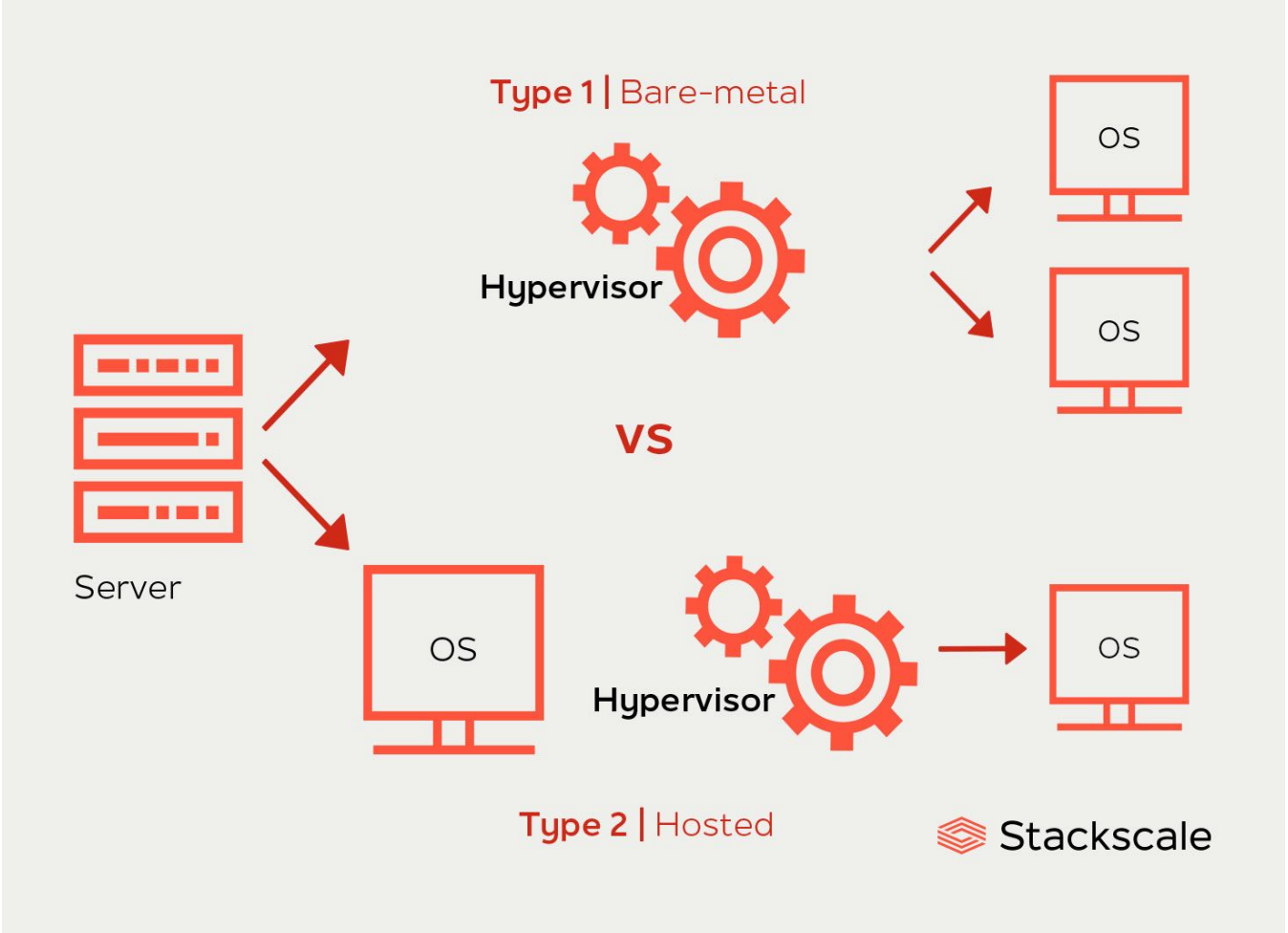


Types of Hypervisors

- There are two ways in hypervisor can be installed:

Hosted hypervisor or type-2 hypervisor:

- ❖ In this case, the hypervisor is installed on the host operating system.
- ❖ Example- VMware Workstation or Fusion or Player, Oracle VM VirtualBox



Virtualization Features

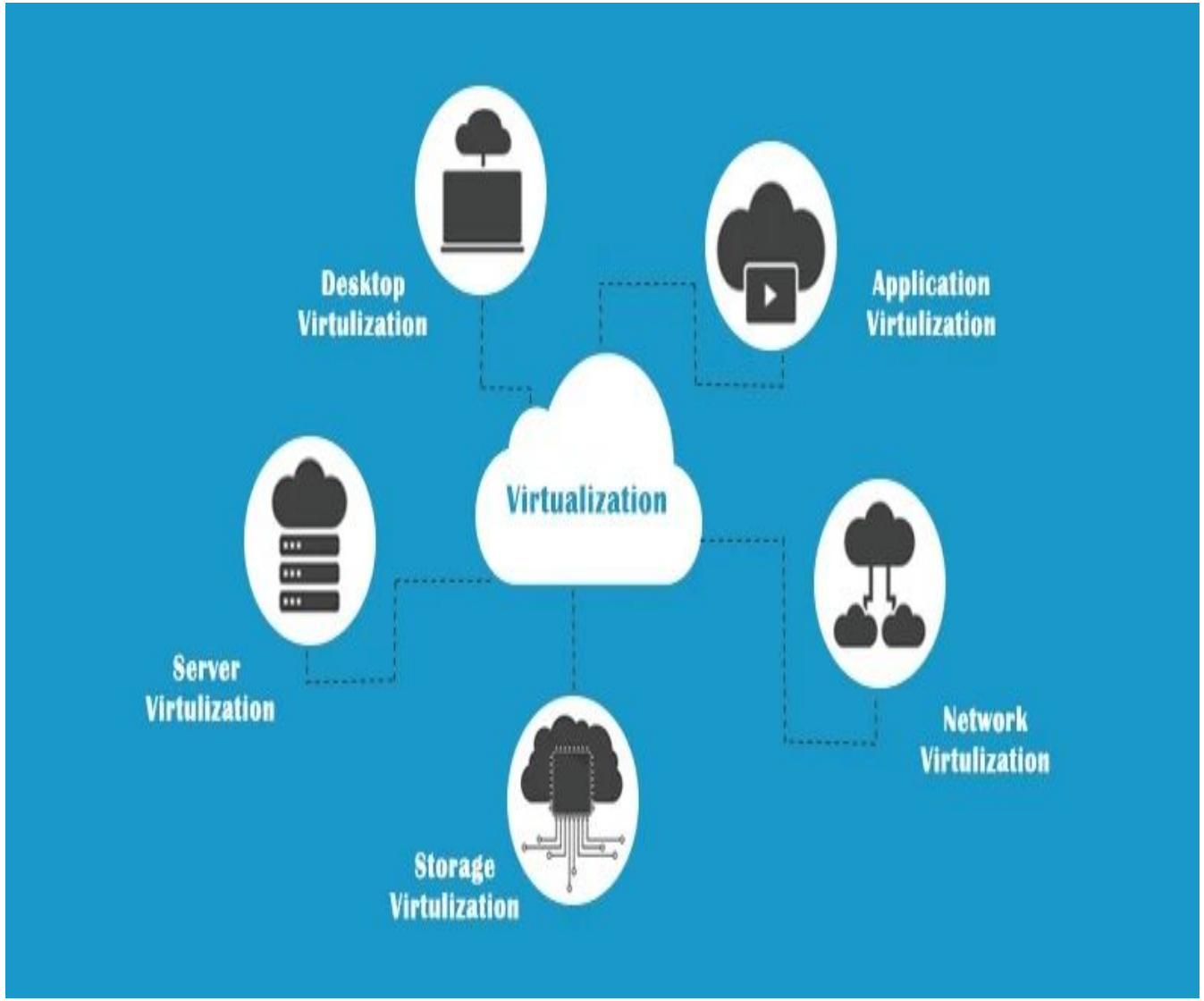
- **Partitioning:** Multiple applications and operating systems are supported by a single physical system by partitioning the available resources.
- **Isolation:** Each VM runs in an isolated manner from its host physical system and other VMs. If one VM crashes, the other VMs and the host system are not affected.

Virtualization Features

- **Encapsulation:** Each VM encapsulates its state as a file system i.e. it can be copied or moved like a simple file.
- **Interposition:** Generally in a VM, all the new guest actions are performed through the monitor (Hypervisor). A monitor can inspect, modify or deny operations such as compression, encryption, profiling, and translation. Such types of actions are done without the knowledge of the operating system.

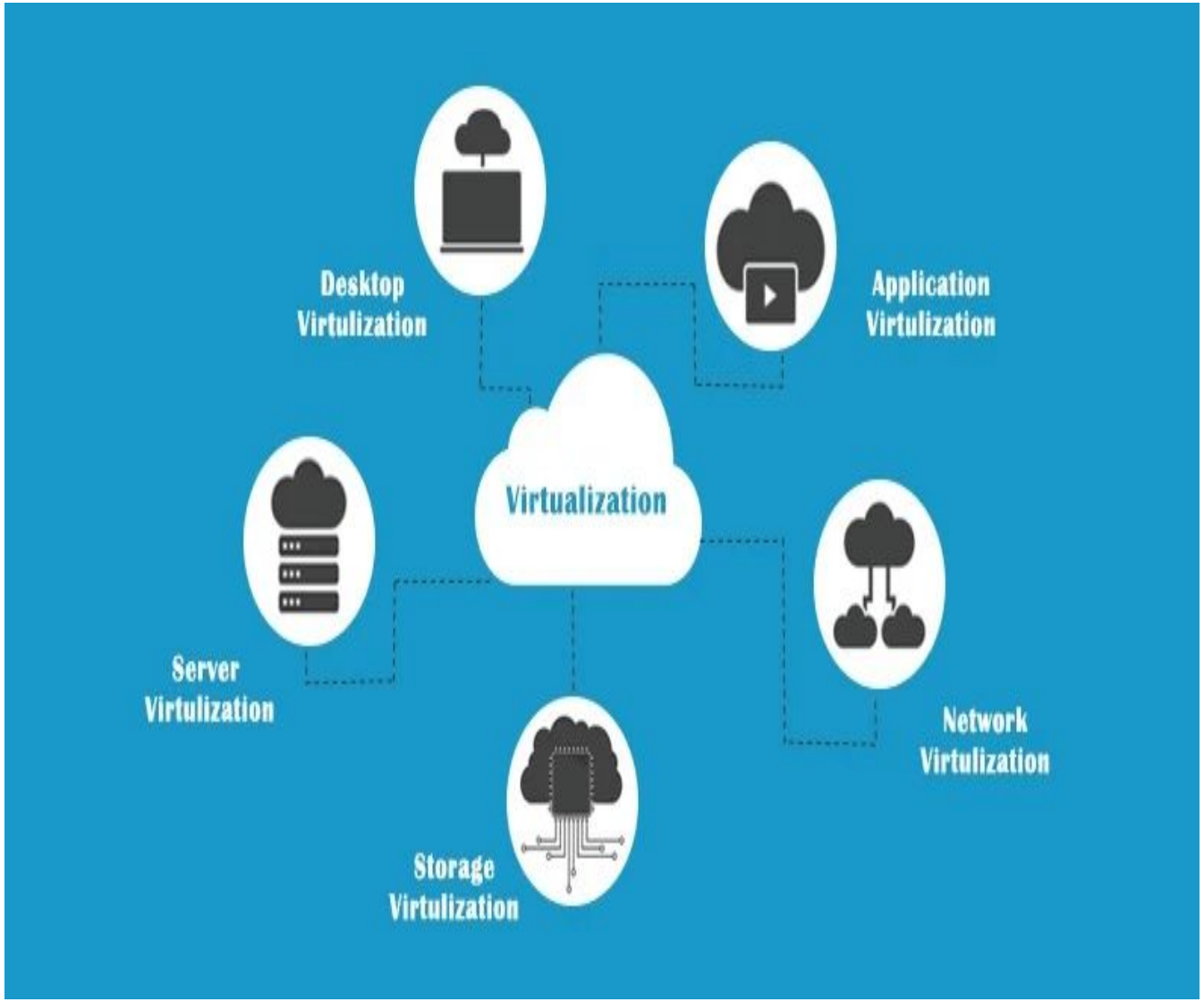
Virtualization Types

I. **Server virtualization:**
   A. Here a single physical server is partitioned into multiple virtual server.
   B. Each virtual server has its own hardware and related resources, such as RAM, CPU, hard drive and network controllers.
   C. A thin layer of software is also inserted with the hardware which consists of a VM monitor, also called hypervisor to manage the traffic between VMs and the physical machine.



Virtualization Types

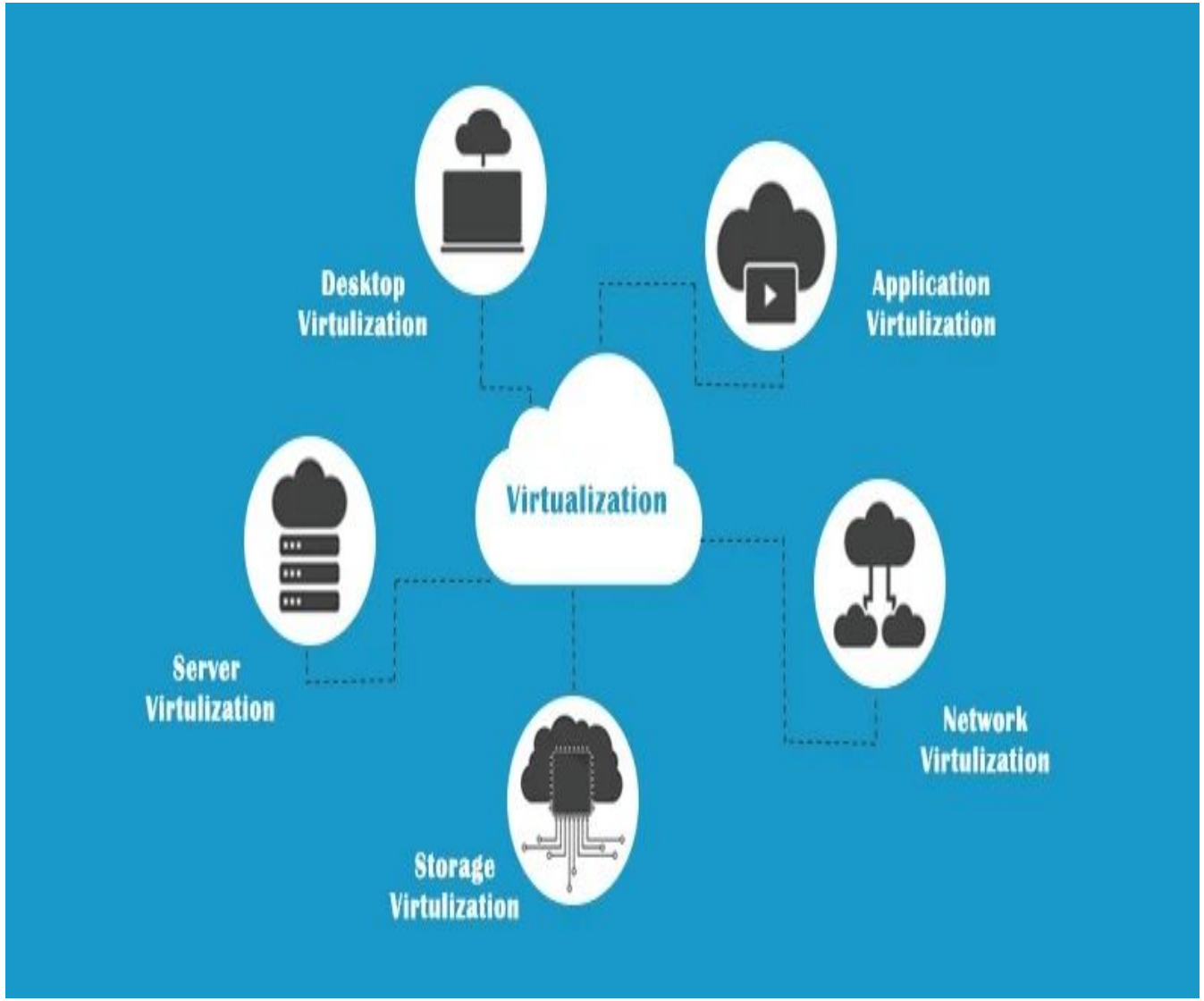
II. **Application virtualization:**
   A. Allows the user to access the application, not from their workstation, but from a remotely located server.
   B. The server stores all personal information and other characteristics of the application, but can still run on a local workstation.
   C. Technically, the application is not installed, but acts like it is.



Virtualization Types

III. Data and Storage virtualization:

A. Is the process of grouping the physical storage from multiple network storage devices so that it acts as if it's on one storage device.



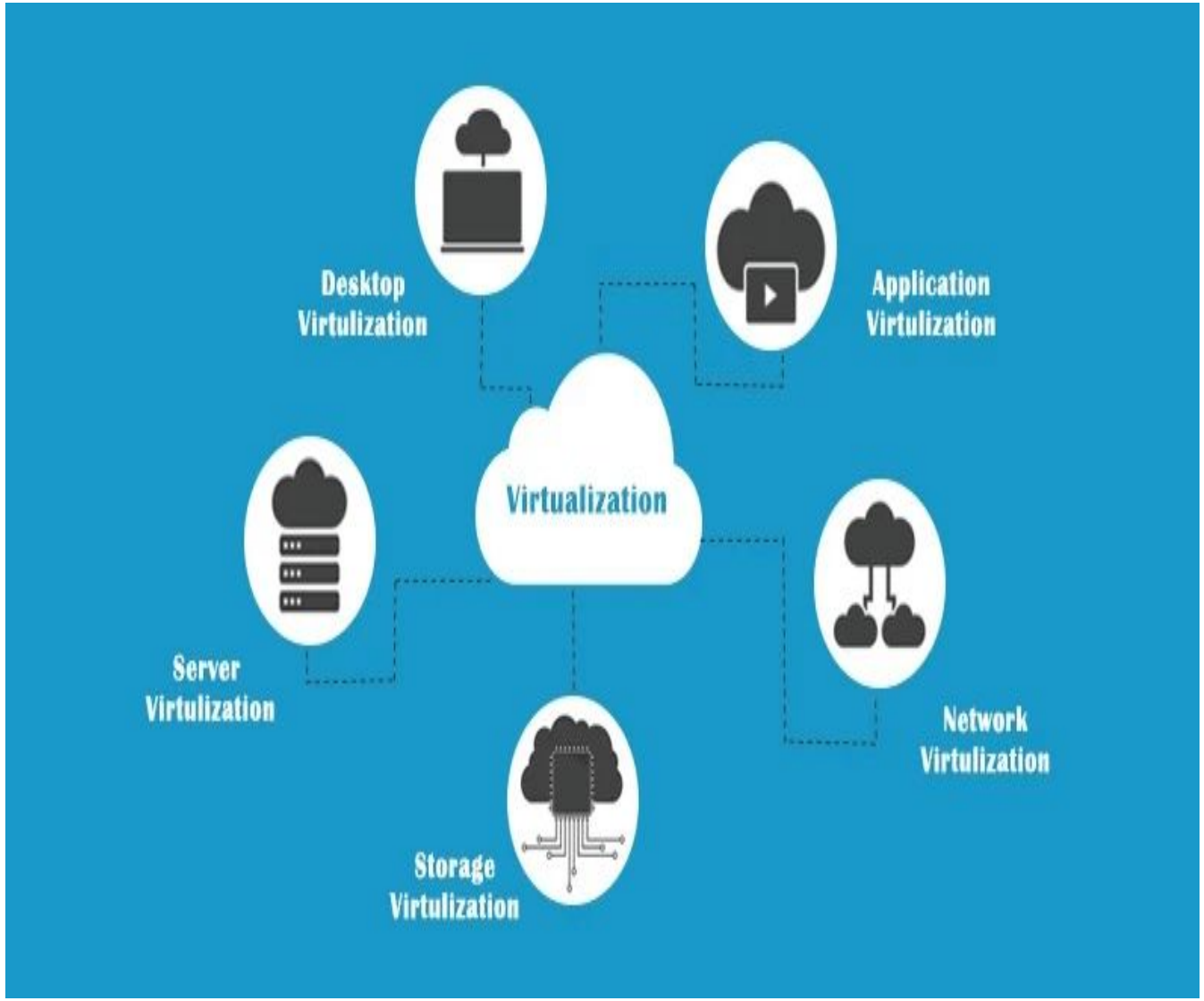
Virtualization Types

IV. Network virtualization:

A. It means using virtual networking as a pool of connection resources.

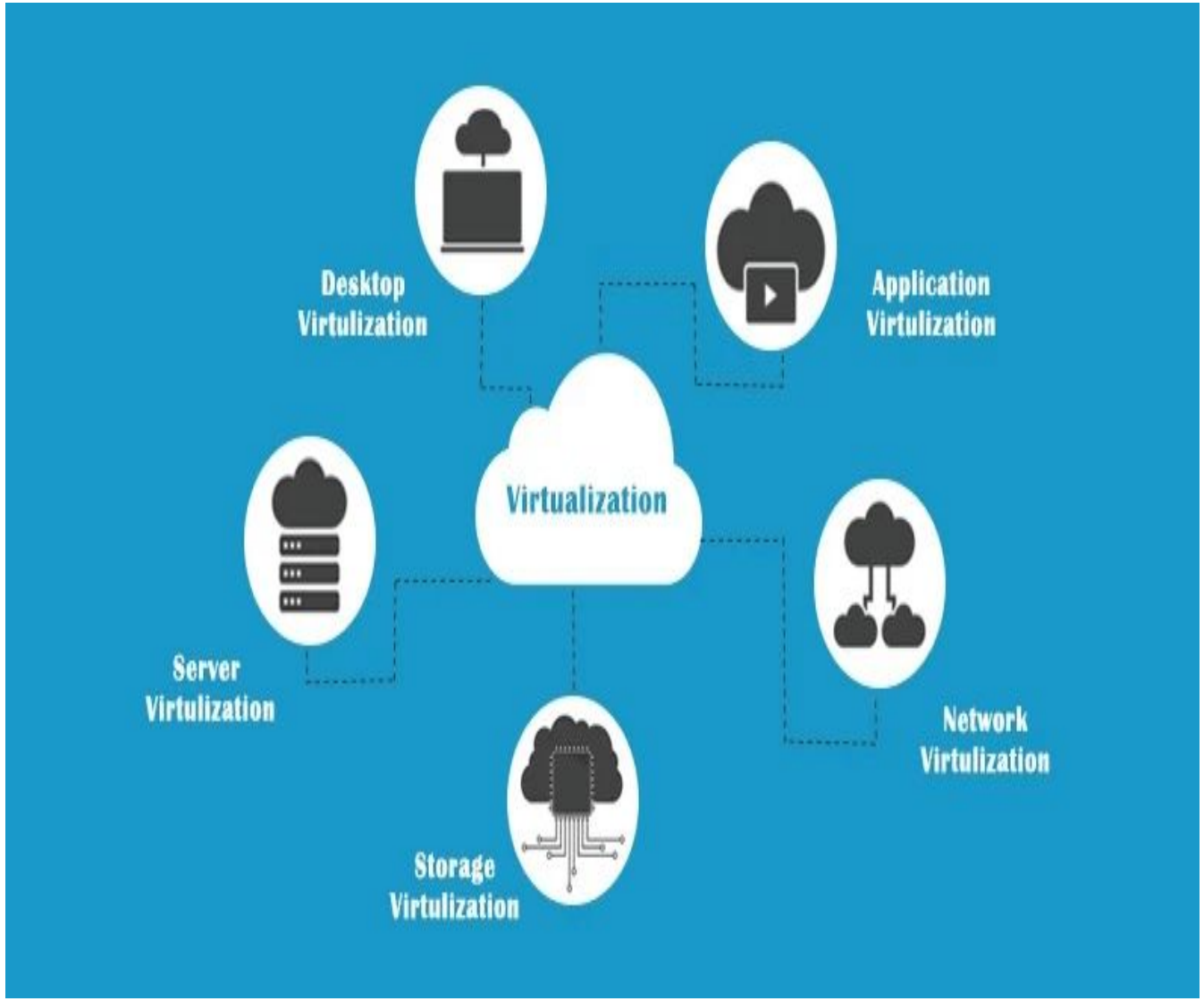
B. During implementing such virtualization, physical network is to be relied for managing traffic between connections.

C. Several virtual networks can be created from a single physical implementation.



Virtualization Types

V. **Desktop virtualization:**
   A. Technology that lets users simulate a workstation load to access a desktop from a connected device.
   B. It separates the desktop environment and its applications from the physical client device used to access it.



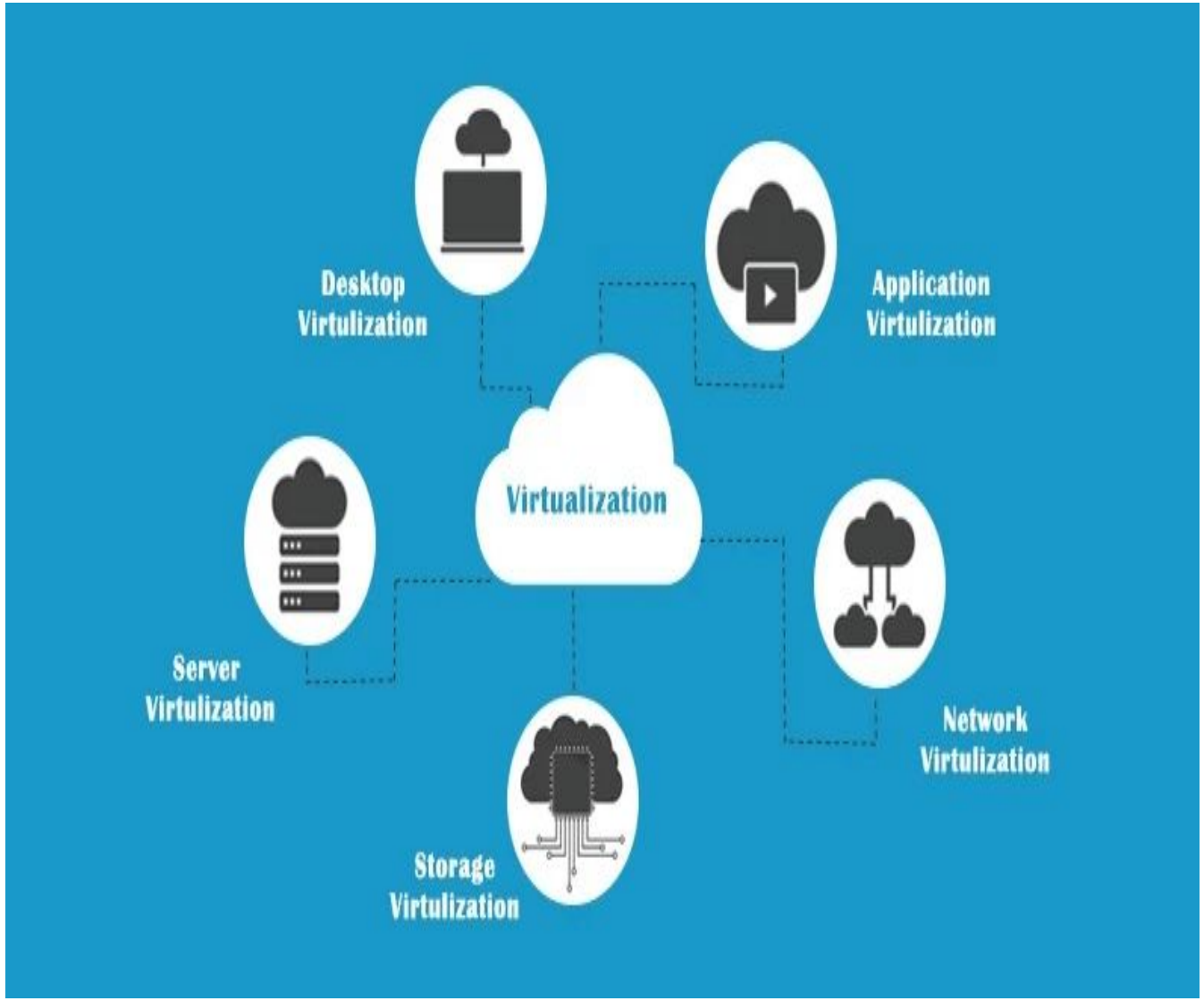
Virtualization Types

VI. Processor and Memory virtualization:

A. It decouples memory from the server and optimize the power of the processor and maximizes its performance.

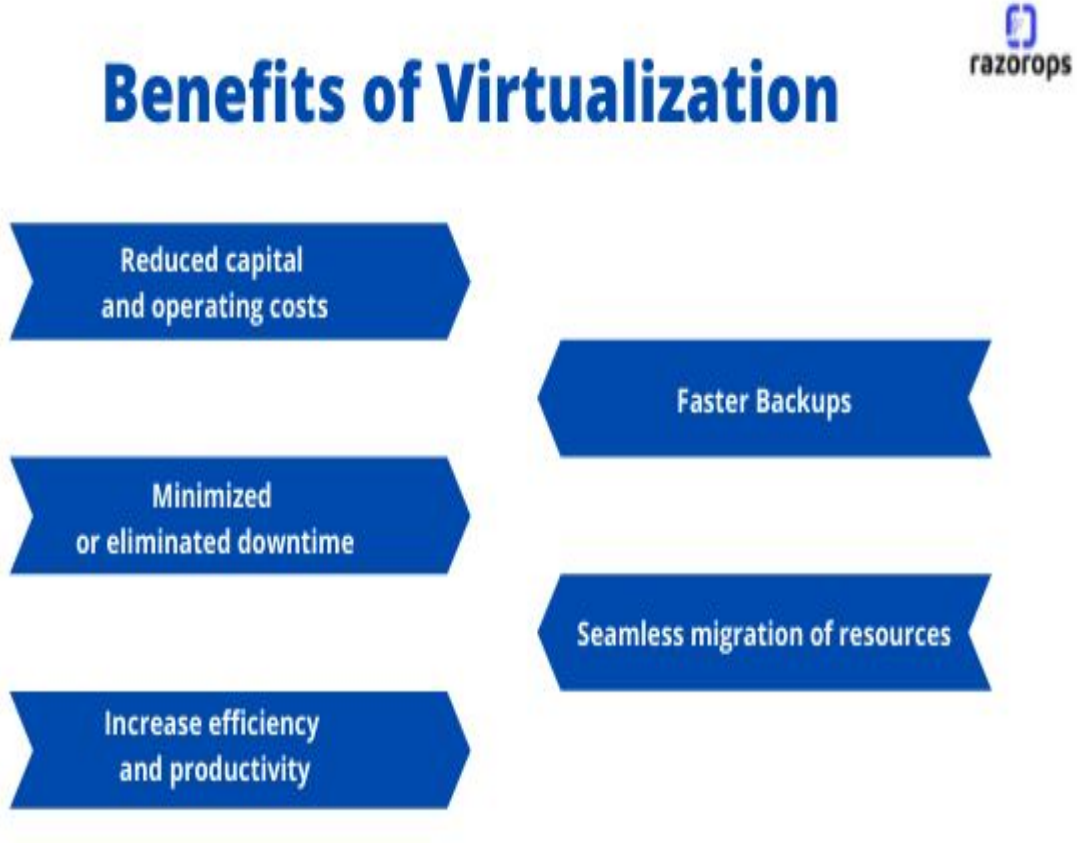
B. Big data analysis needs systems to have high processing power (CPU) and memory (RAM) for performing complex computations.

C. Processor and Memory virtualization can increase the speed of processing and the analysis results sooner.



Benefits of Virtualization

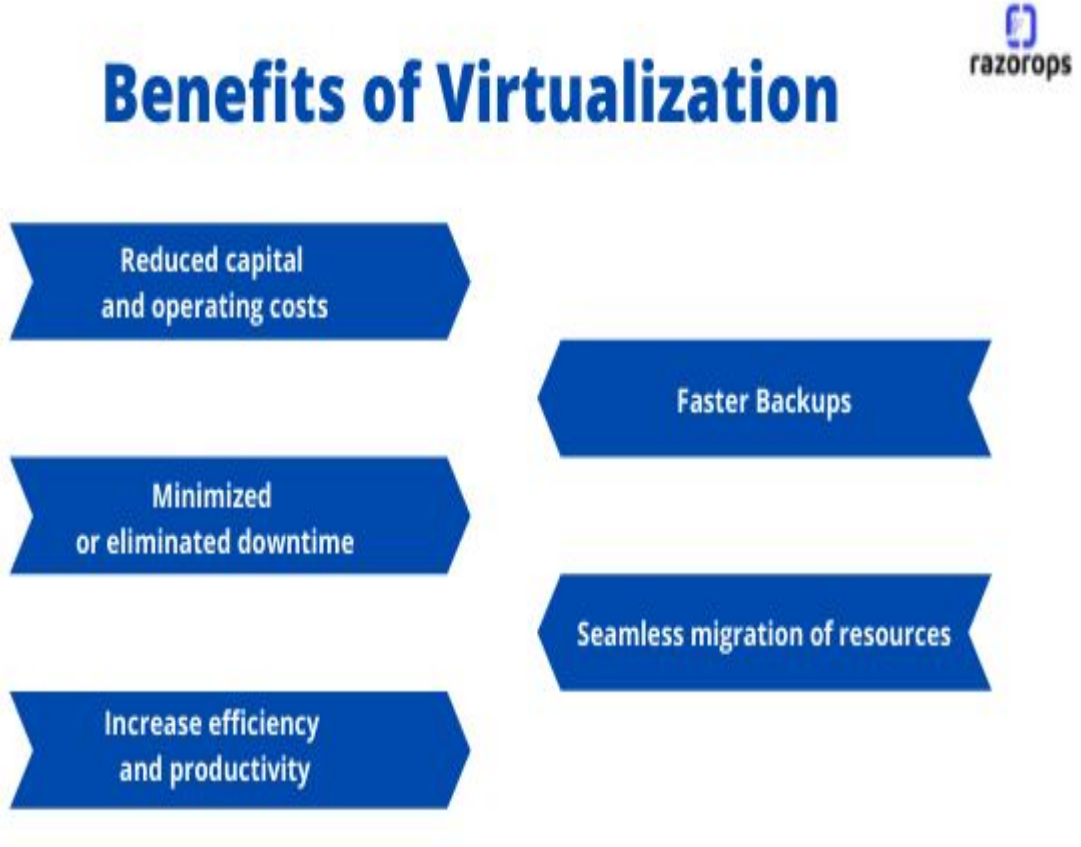
Resource Optimization: We can create multiple virtual machines on the unused and utilized hardware without need to buy new hardware
Consolidation: If an organization has multiple application running on multiple hardware, we can consolidate into one single machine and creating multiple virtual machines to host that application. This results in less physical space, A/C resources, and other data center resources.



Benefits of Virtualization

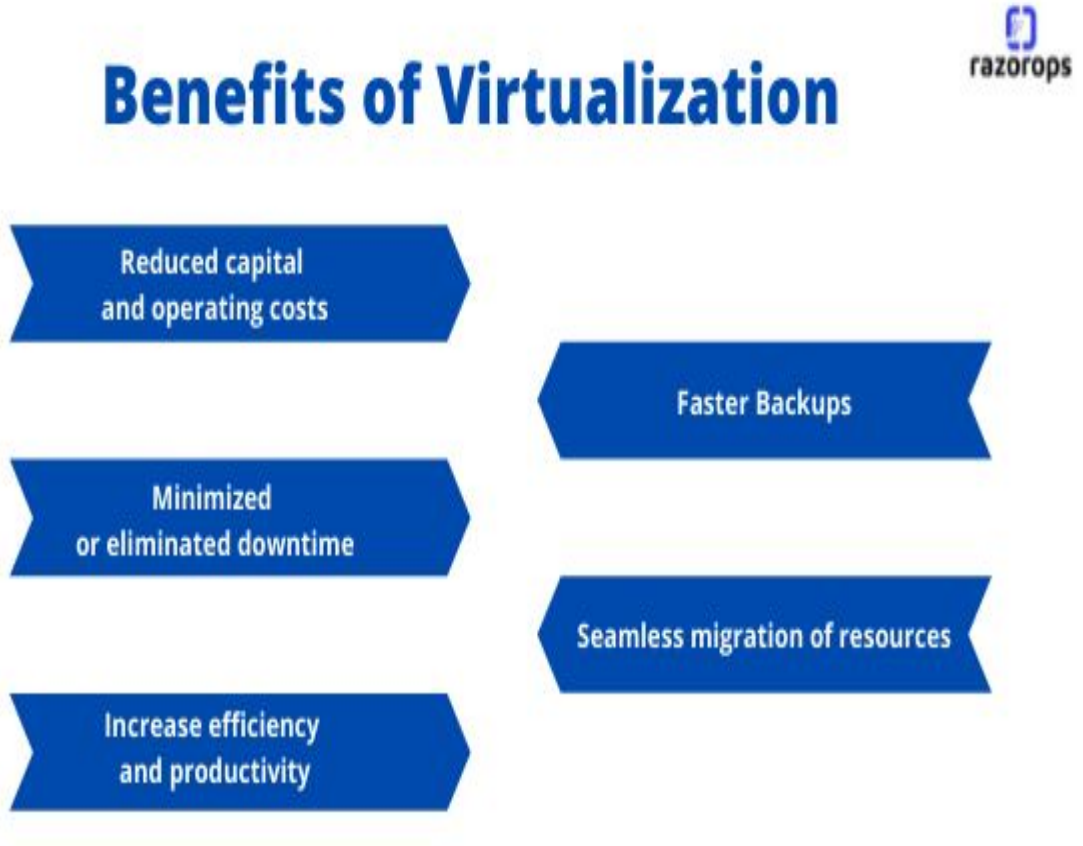
Maximize Uptime: With virtualization, we

- can spin off virtual machine easily.
- Reconfiguration of computer resources without impacting users.
- Guaranteed uptime of server and applications
- Elasticity: we can bring up resource as and when required.
- Speedy disaster recovery



Benefits of Virtualization

Easily Migrate workload: We can move virtual machine from one physical box to another box regardless of the their of difference in hardware configuration. So it increase reliability and availability

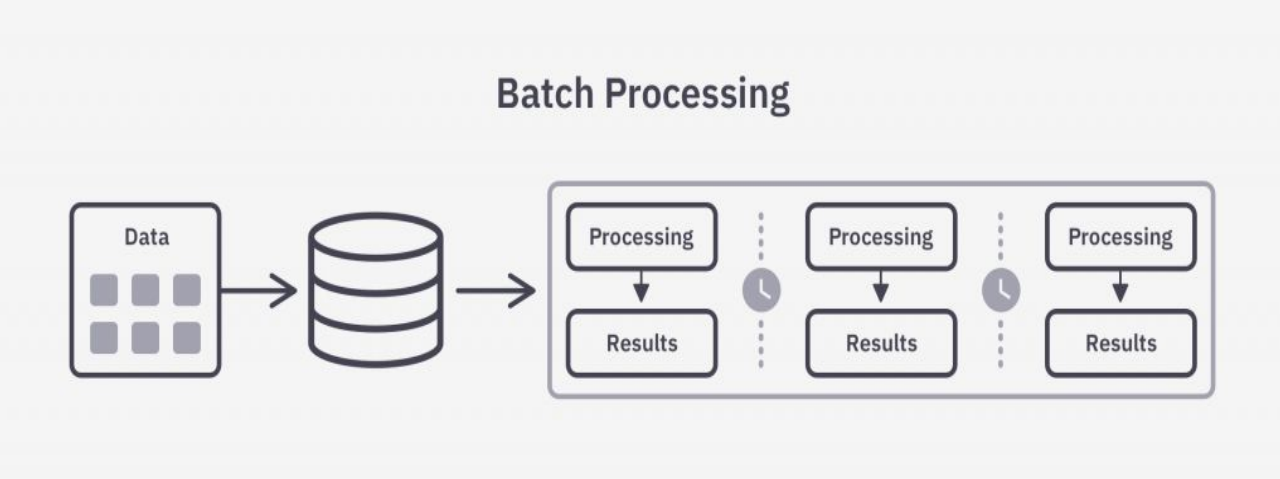


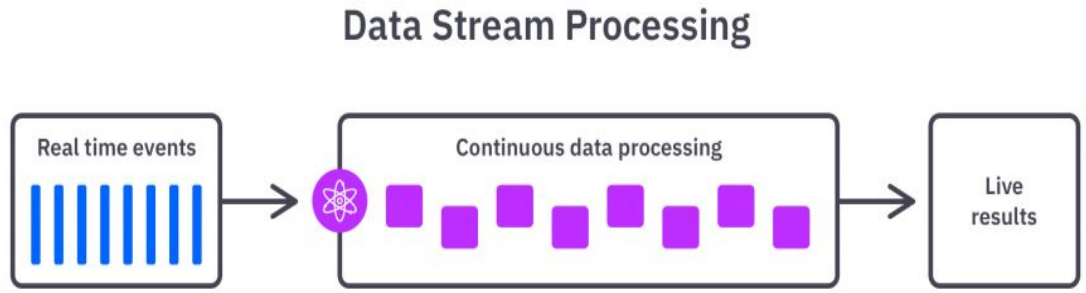
Data Stream

- In 2020, the total amount of data generated by every person around the world was 1.7 megabytes per second, totaling 44 zettabytes.
- By 2025, the amount of stream data generated globally is estimated to reach an outstanding 463 zettabytes.
- This tremendous amount of data being generated has prompted many organizations to ditch batch processing and adopt real-time data streams in an effort to stay up-t-date with the ever-changing business needs.

Batch Vs. Real Time Streaming Processing

- In batch data processing, data is downloaded in batches before being processed, stored, and analyzed.
- On the other hand, stream data ingest data continuously, allowing it to be processed simultaneously and in real-time.





Data Stream (Contd.)

- **Data Stream is a continuous, fast-changing, and ordered chain of data transmitted at a very high speed. It is an ordered sequence of information for a specific interval.**
- The sender's data is transferred from the sender's side and immediately shows in data streaming at the receiver's side.
- **Streaming does not mean downloading the data or storing the information on storage devices.**
- **A stream data source is characterized by continuous time-stamped logs that document events in real time.**

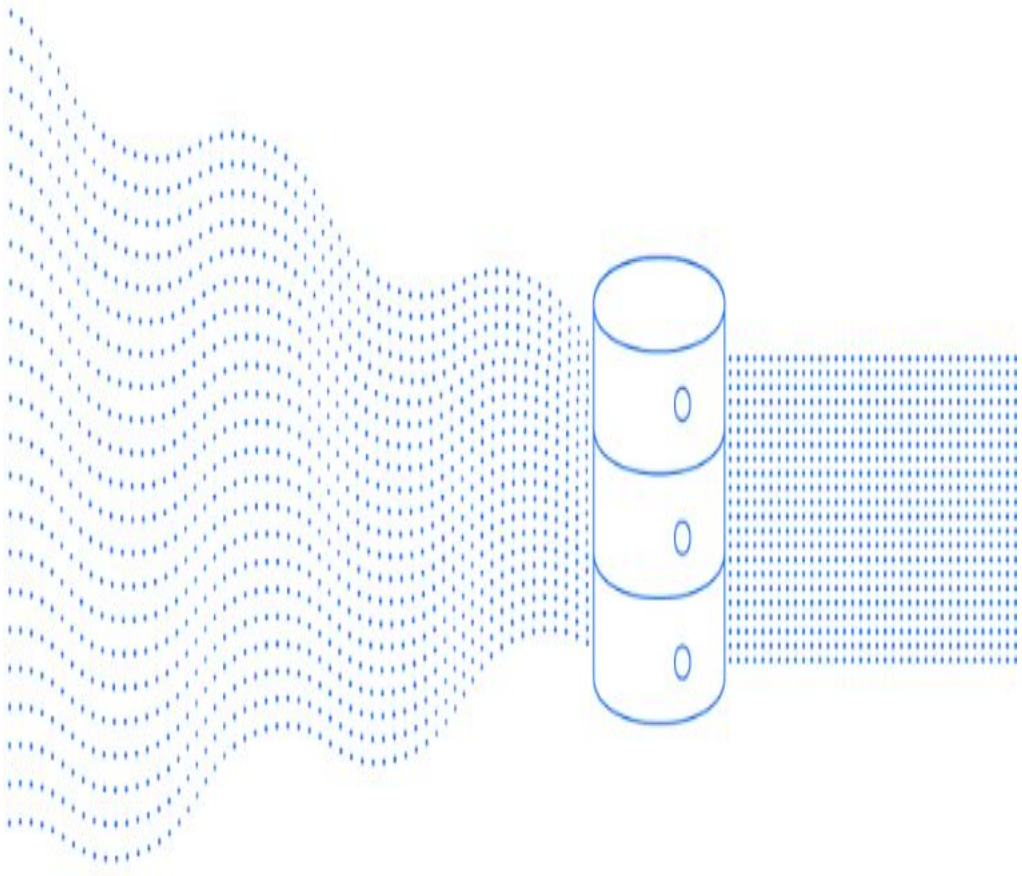
Common Examples of Data Stream Sources

Data streaming is used for data that is generated in small batches and continuously transmitted – such as from IoT sensors, server and security logs, real-time advertising platforms, click-stream data from apps & websites, Real-time ATM transaction Live event data, Call records, Satellite data, Audio listening, Real-time surveillance systems, Online transactions etc.

Data Stream (Alternate Definitions)

Data Stream: Usually signifies large data volume, likely unstructured and structured arriving at a very high rate, which requires real time/near real time analysis for effective decision making.

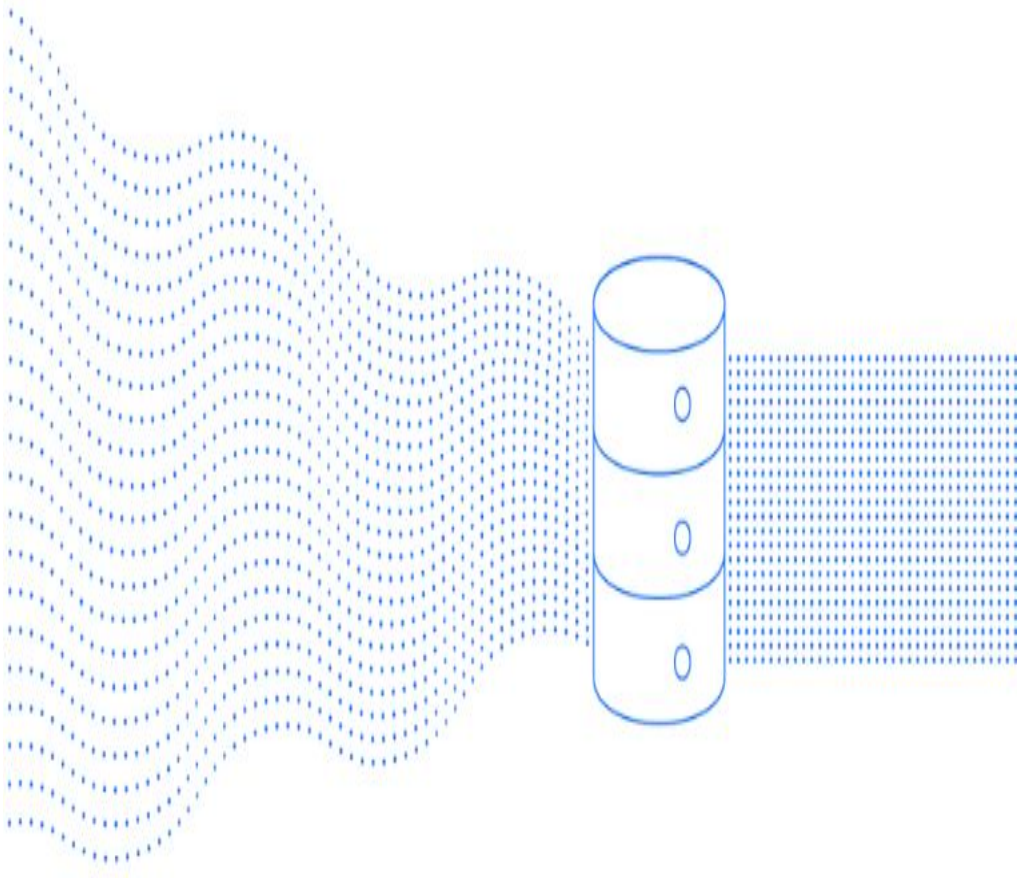
- It is basically continuously generated data and arrives in a stream (sequence of data elements made available over time).
- It is generally time-stamped and geo-tagged (in the form of latitude and longitude)



Data Stream

Data Stream: Usually signifies large data volume, likely unstructured and structured arriving at a very high rate, which requires real time/near real time analysis for effective decision making.

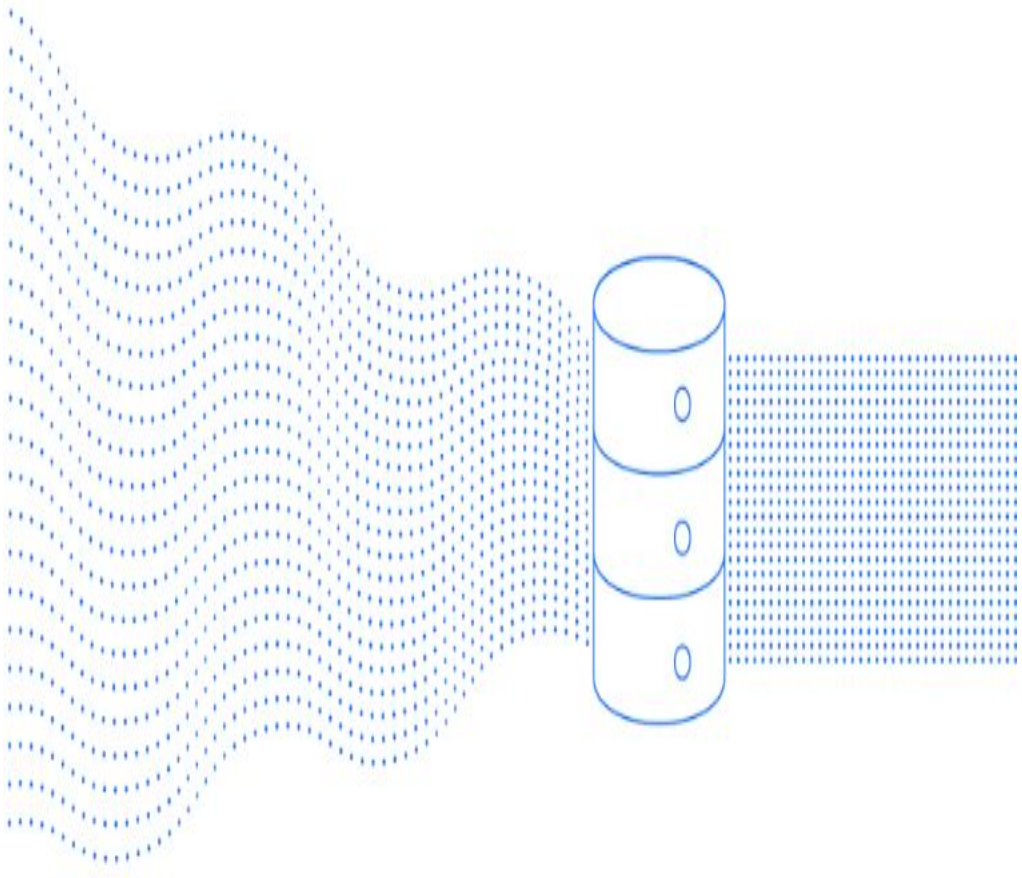
- Stream is composed of synchronized sequence of elements or events.
- If it is not processed immediately, then it is lost forever (No Real Time Significance).



Data Stream

Data Stream: Usually signifies large data volume, likely unstructured and structured arriving at a very high rate, which requires real time/near real time analysis for effective decision making.

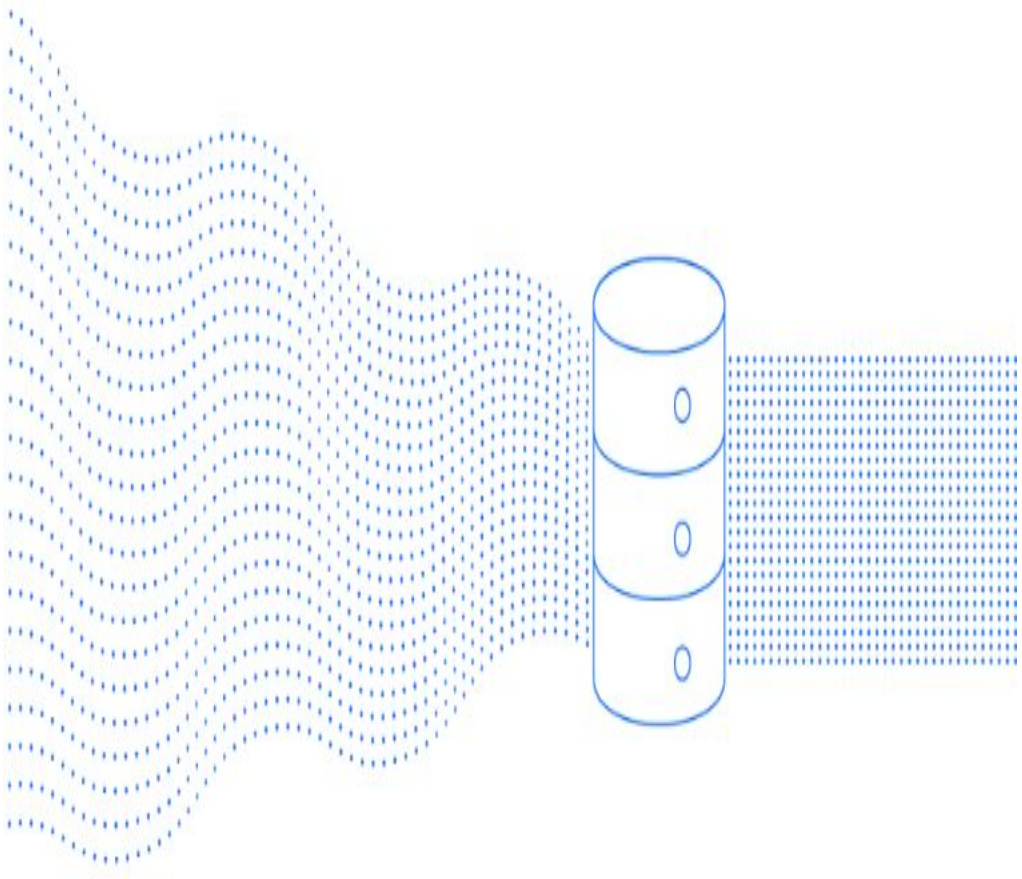
- In general, such data is generated as part of application logs, events, or collected from a large pool of devices continuously generating events such as ATM or PoS



Data Stream

Data Stream:
Example:
Data Center:

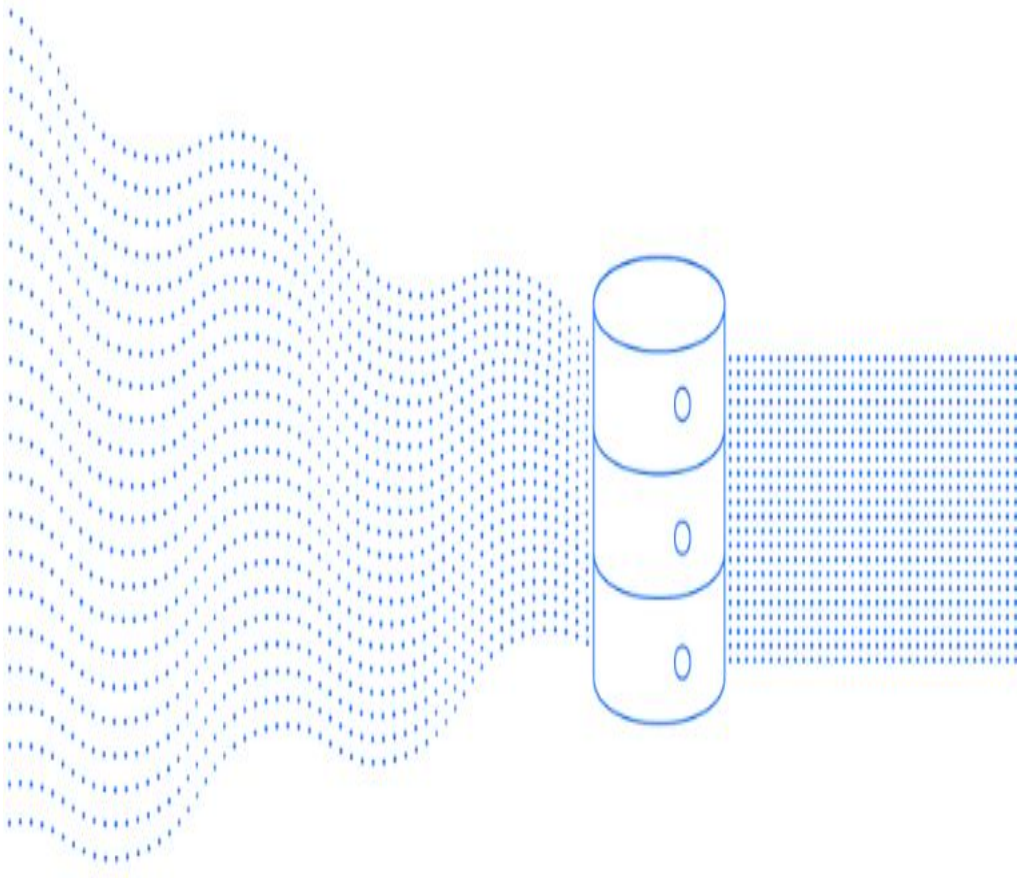
- Large network deployment of a data center with hundreds of servers, switches, routers and other devices in the network.
- The event logs from all these devices at real time create a stream of data.
- This data can be used to prevent failures in the data center and automate triggers so that the complete data center is fault tolerant



Data Stream

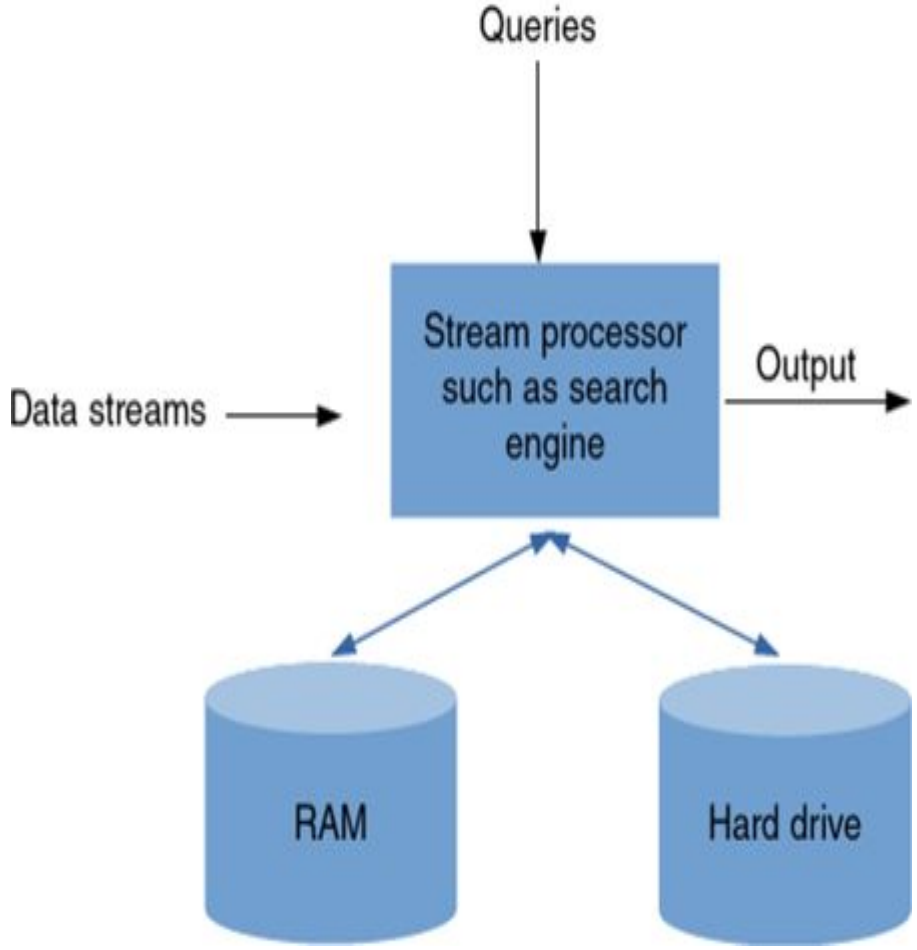
Data Stream:
Example:
Stock Market:

- The data generated here is a stream of data where a lot of events are happening in real-time.
- The price of stock are continuously varying.
- These are large continuous data streams which needs analysis in real-time for better decisions on trading

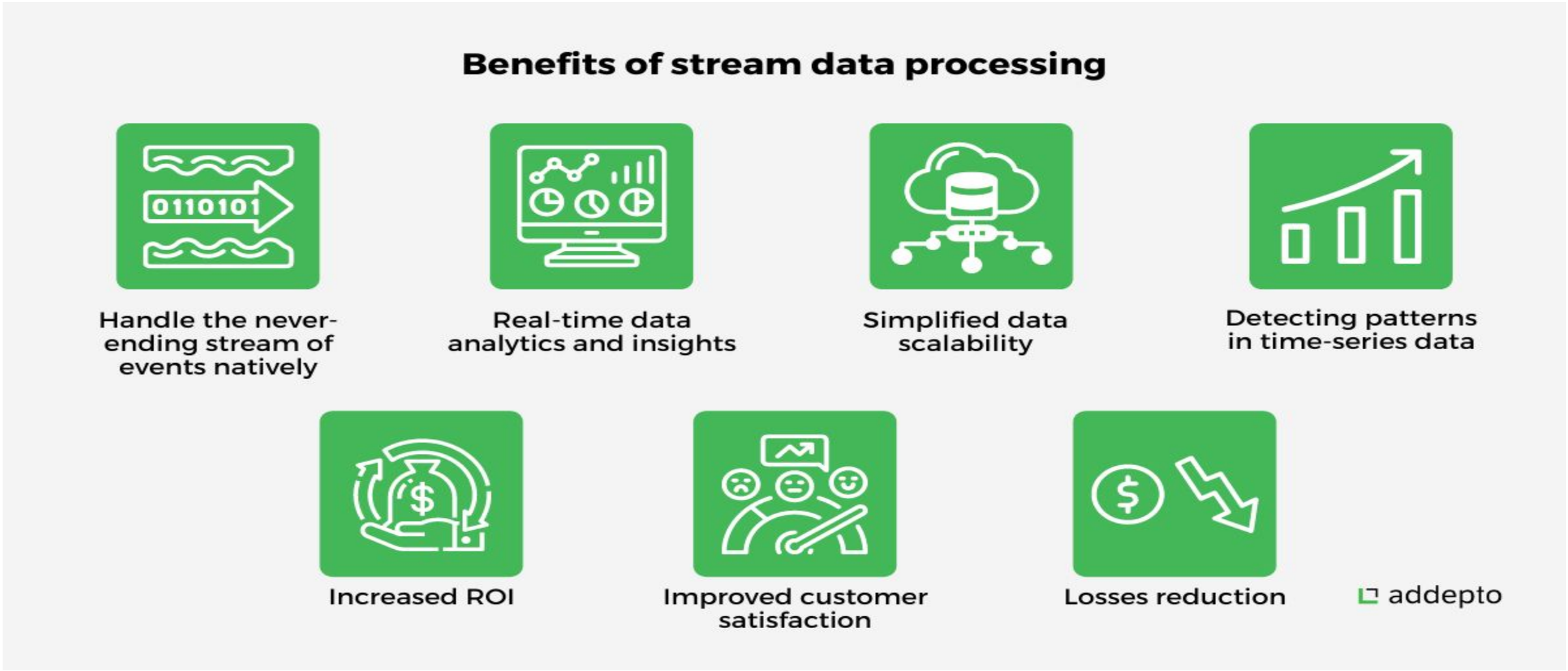


Basic Model of Stream Data

- Input data rapidly and streams needn't have the same data rates or data types
- The system cannot store the data entirely
- Queries tends to ask information about recent data
- The scan never turn back



Stream Data Processing Advantages



Big Data Streaming

References

1. https://www.engadget.com/2016-09-30-how-to-get-started-with-virtualization.html
2. https://linux.how2shout.com/what-is-virtualization-technology-and-its-advantages/
3. https://www.parkplacetechnologies.com/blog/what-is-hypervisor-types-benefits/
4. https://www.stackscale.com/blog/hypervisors/
5. https://www.dnsstuff.com/what-is-vm-virtual-machine
6. https://www.redswitches.com/blog/virtualization-types-cloud-computing/
7. https://hevodata.com/learn/data-streams-in-data-mining/
8. https://www.citrix.com/solutions/vdi-and-daas/what-is-desktop-virtualization.html#:~:text=Desktop%20virtualization%20is%20technology%20that,device%20used%20to%20access%20it.
9. Ibm.com. Digitization: A Climate Sinner or Savior? URL: https://www.ibm.com/blogs/nordic-msp/digitization-and-the-climate/. Accessed Aug 18, 2023
10. Visualcapitalist.com. How Much Data is Generated Each Day? URL: https://www.visualcapitalist.com/how-much-data-is-generated-each-day/. Accessed Aug 18, 2023

References

11. https://dsstream.com/streaming-data-architecture/
12. https://addepto.com/blog/stream-data-model-and-architecture/
13.