BigData Analytics

Unit 2:
Big Data Framwork and File System

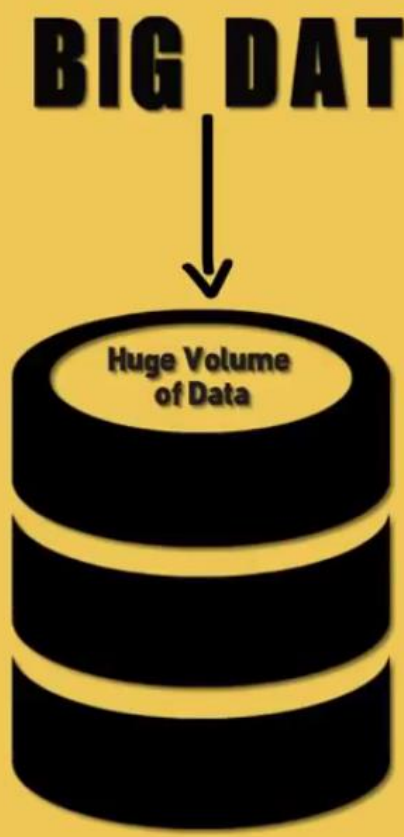
CE:
Hadoop

What is Big Data?

BIG DATA

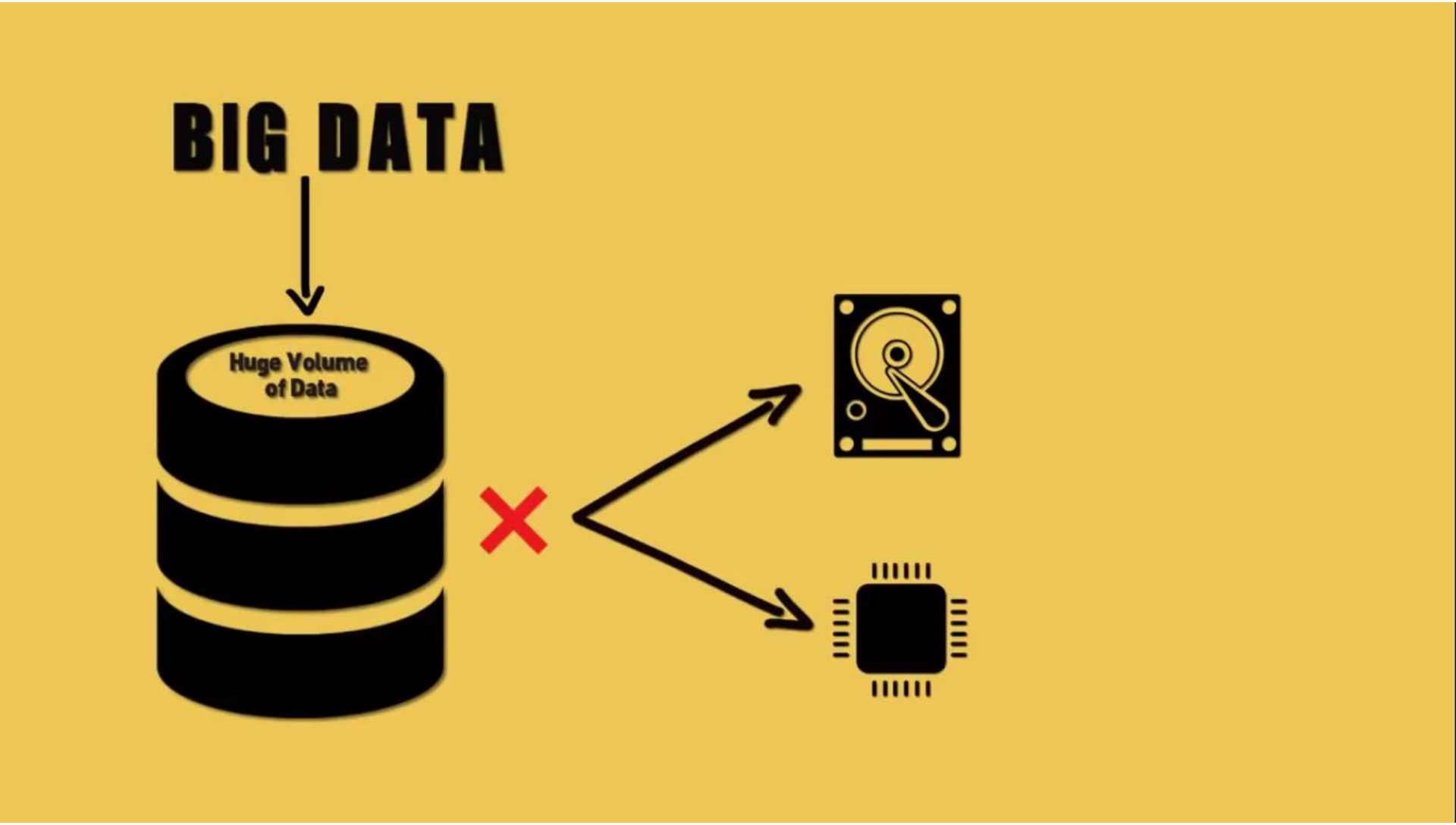
BIG DATA

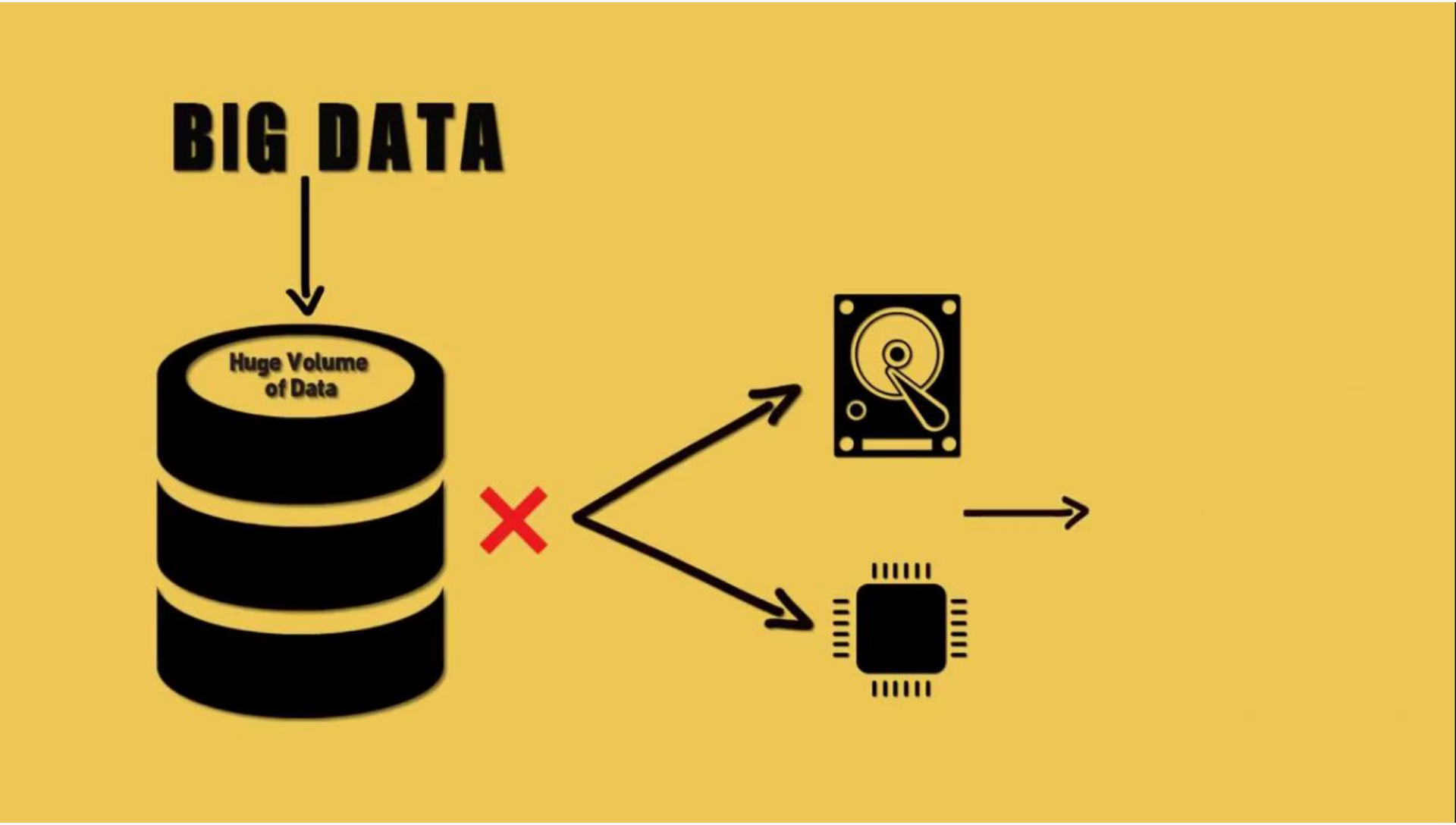
Huge Volume
of Data

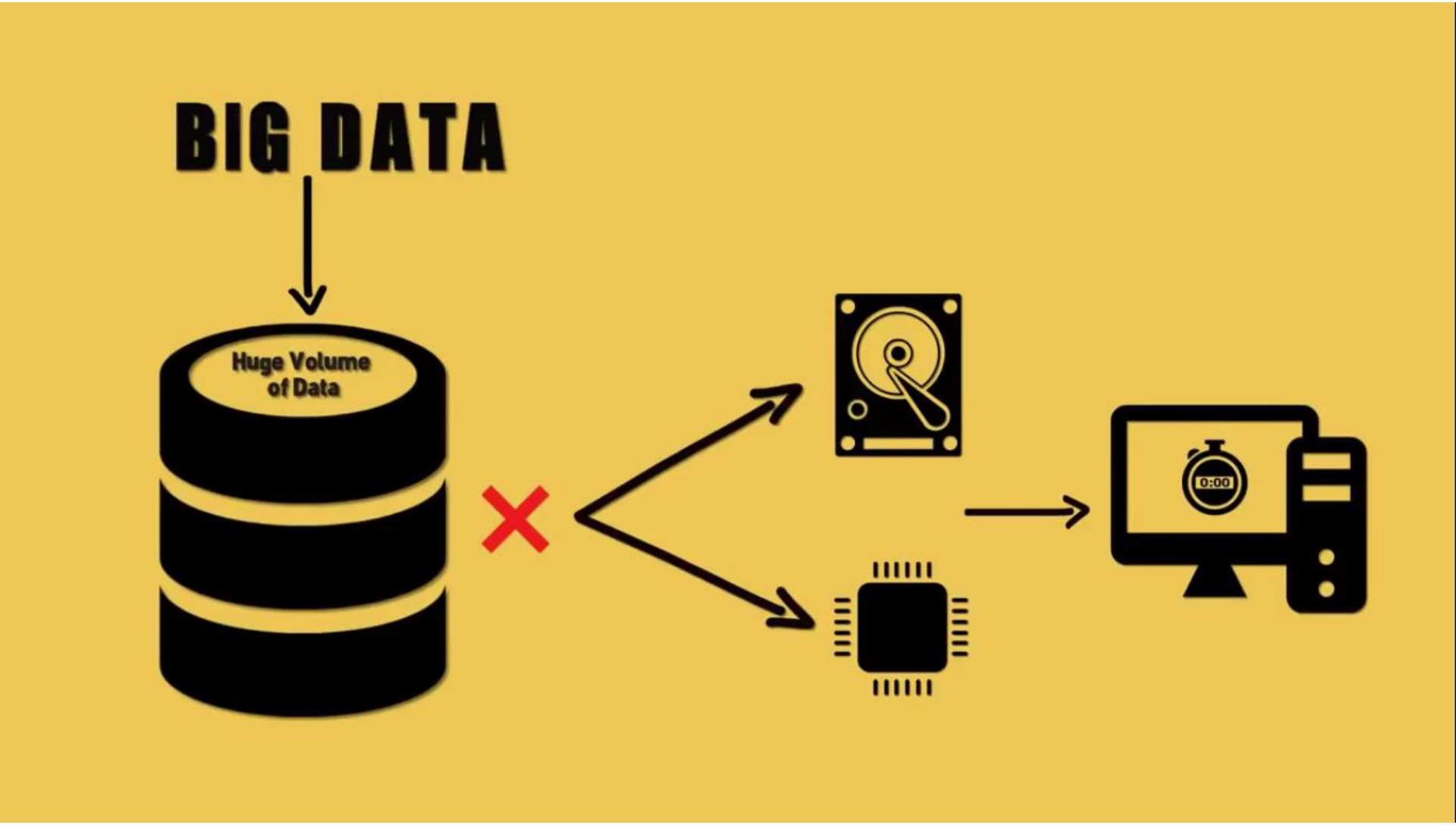

BIG DATA
Huge Volume
of Data


BIG DATA
Huge Volume of Data


BIG DATA
Huge Volume of Data


BIG DATA
Huge Volume of Data


BIG DATA
Huge Volume of Data



hadoop

How huge this Data need to be?


GB / TB / PB / EX
> in Size


GB / TB / PB / EX
> in Size


GB / TB / PB / EX
> in Size

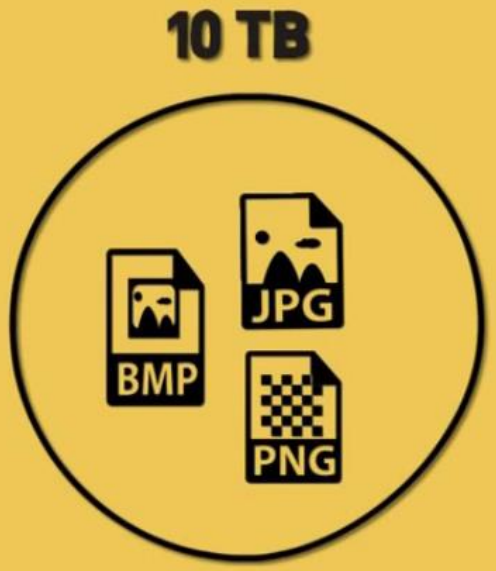
BIG DATA

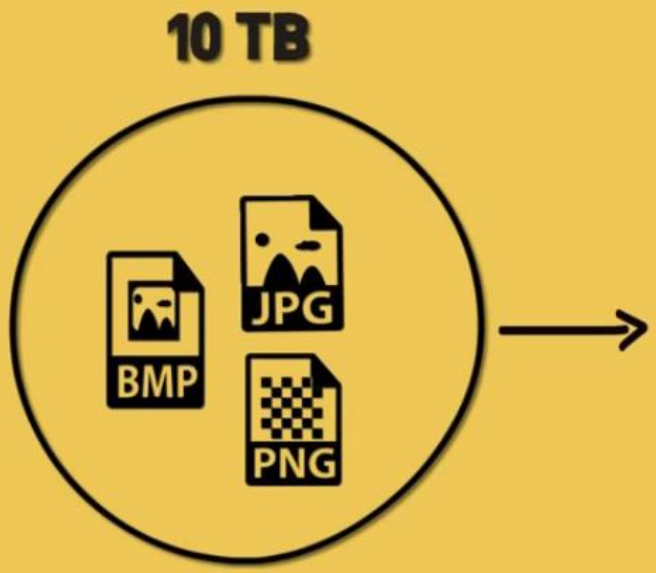

100 MB

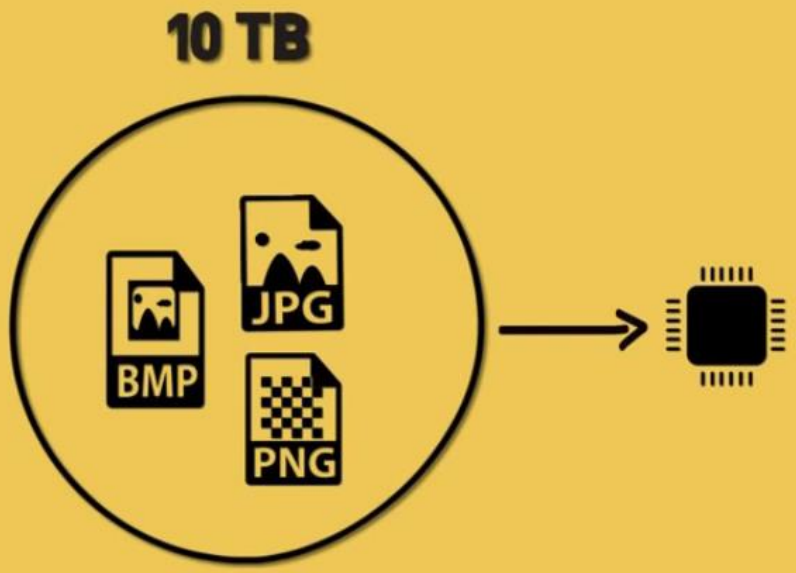

100 MB

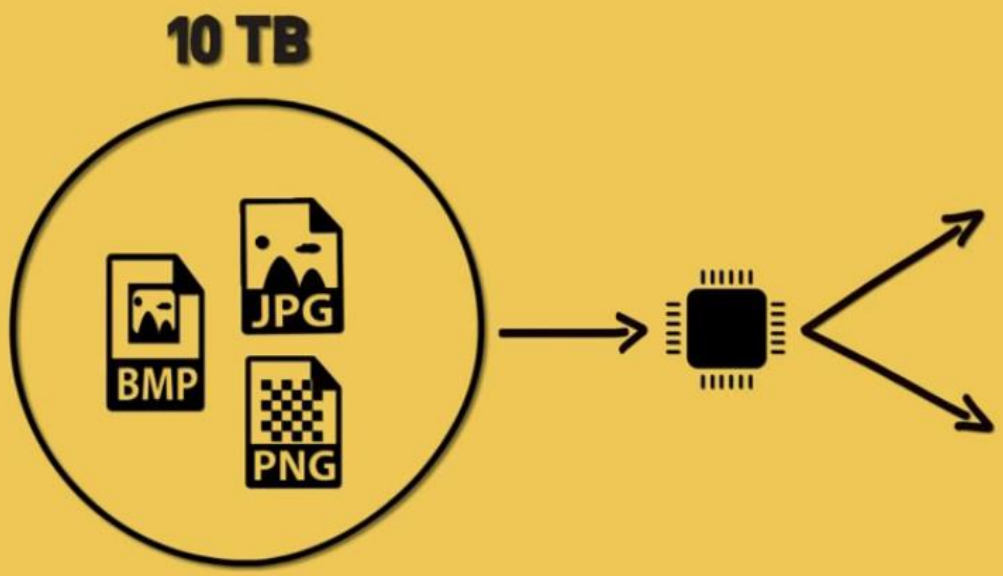

100 MB

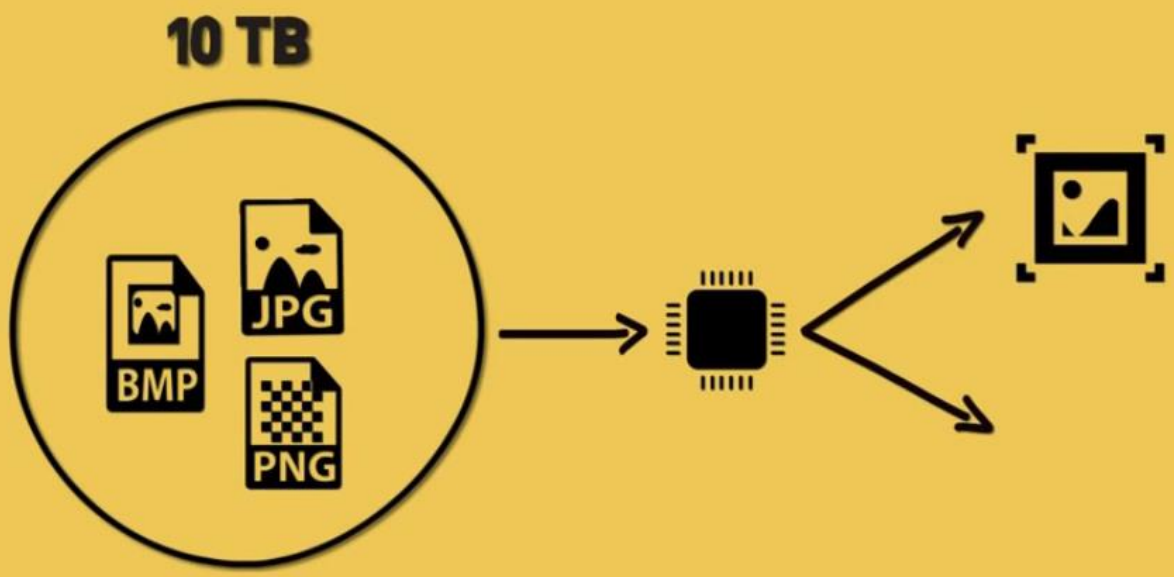

100 MB
BIG DATA

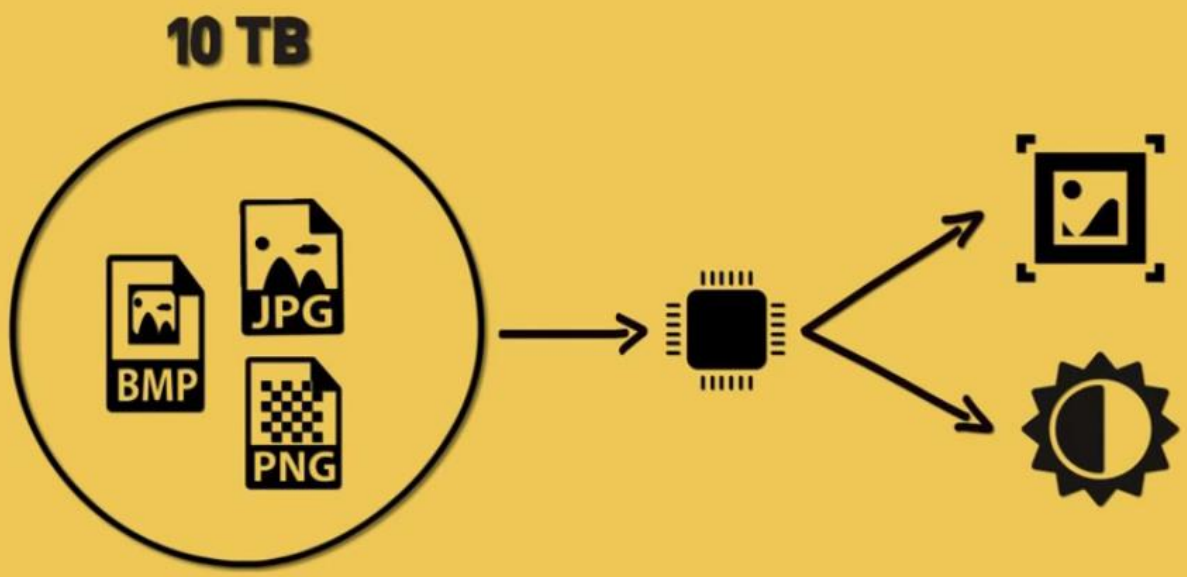

10 TB
BMP
JPG
PNG

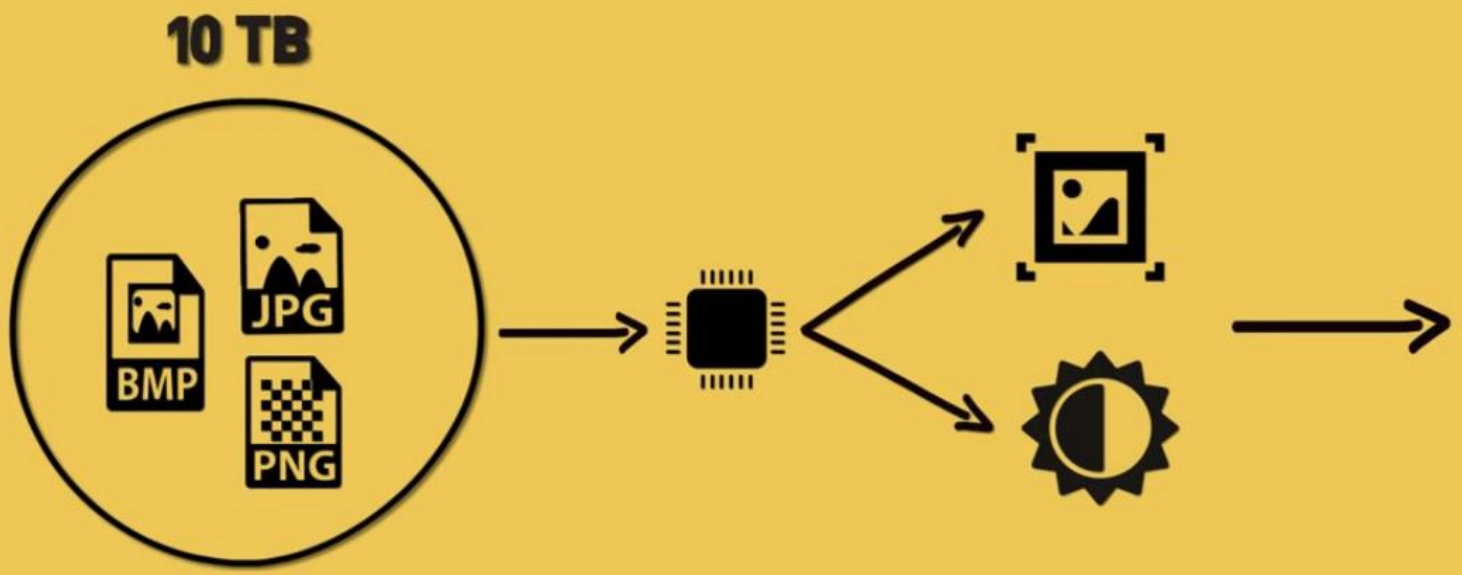

10 TB
BMP
JPG
PNG

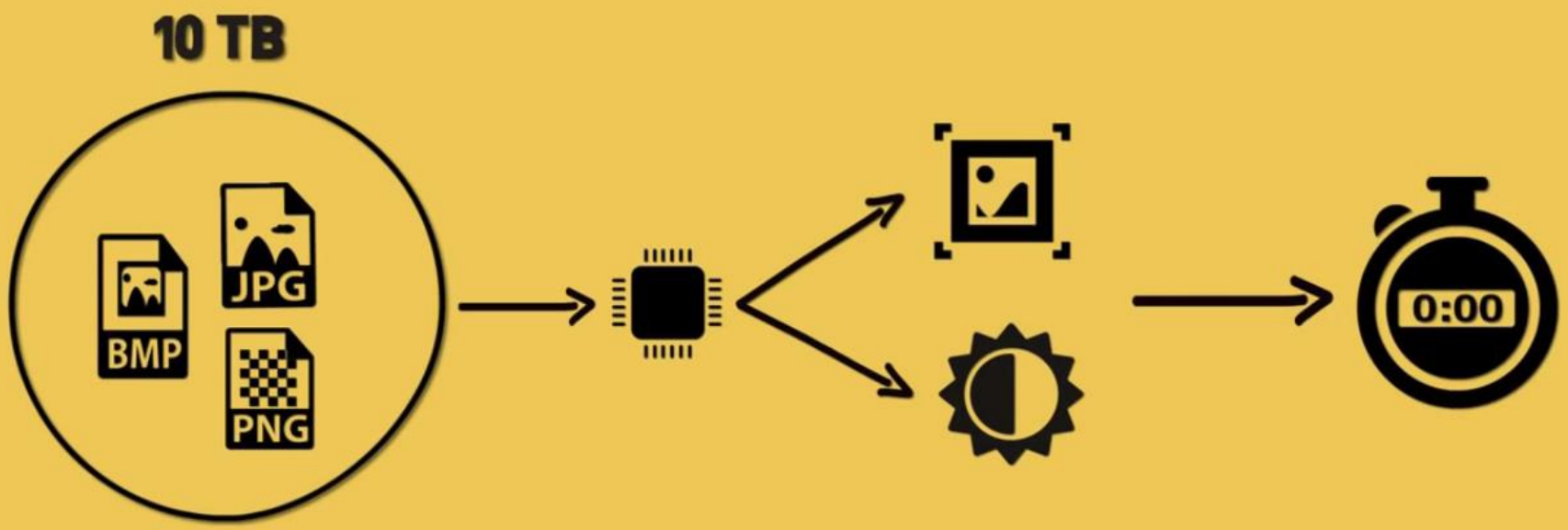

10 TB
BMP
JPG
PNG

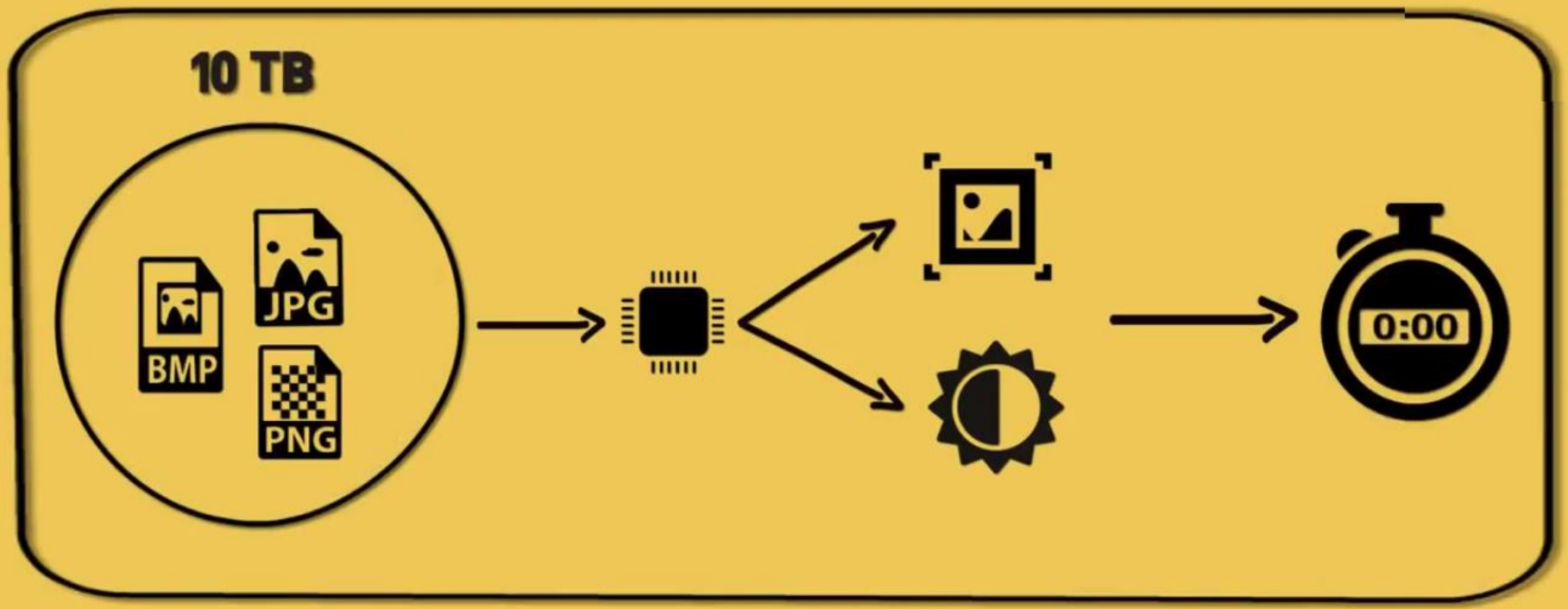

10 TB
BMP
JPG
PNG

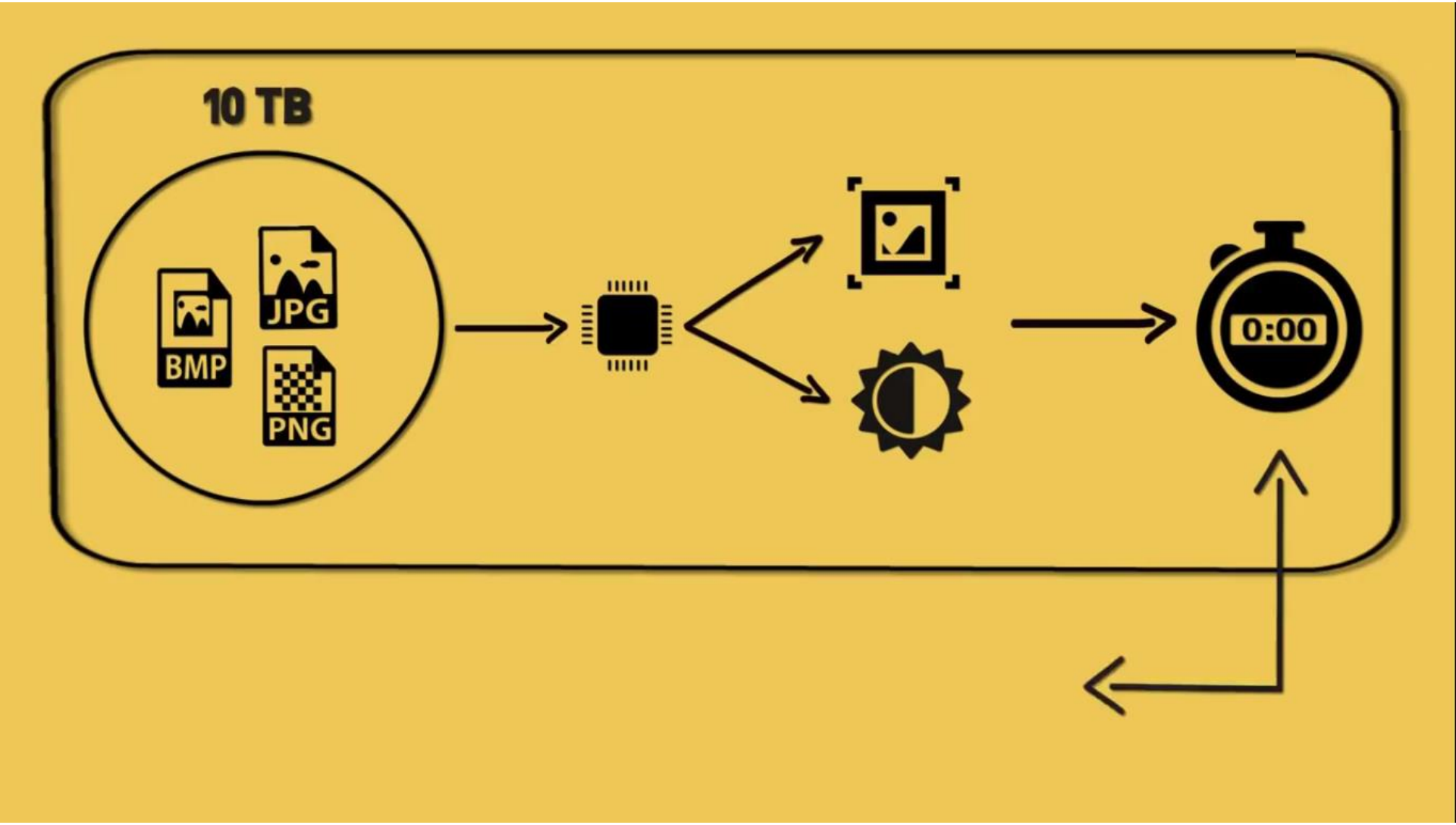

10 TB
BMP
JPG
PNG

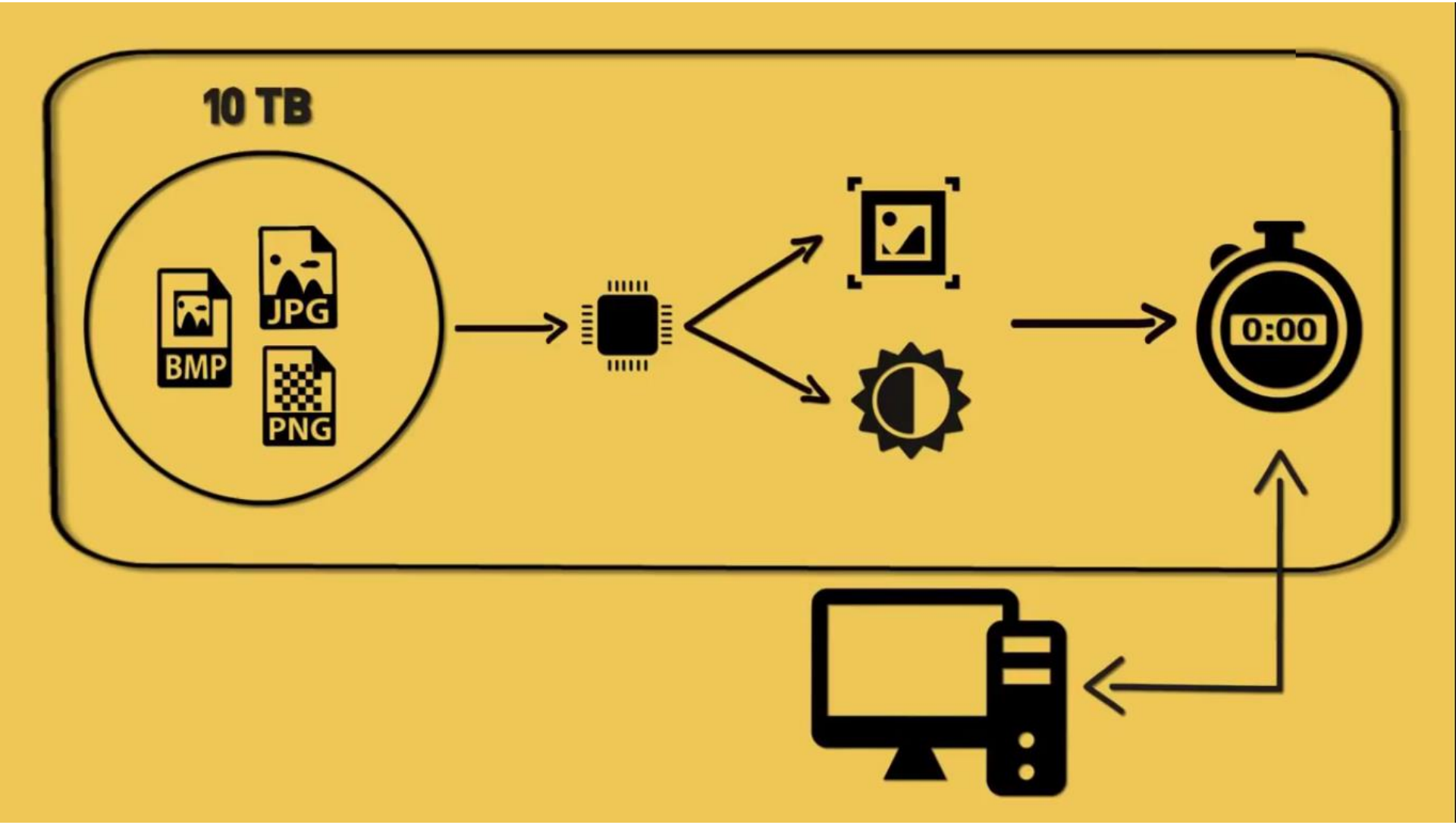

10 TB
BMP
JPG
PNG

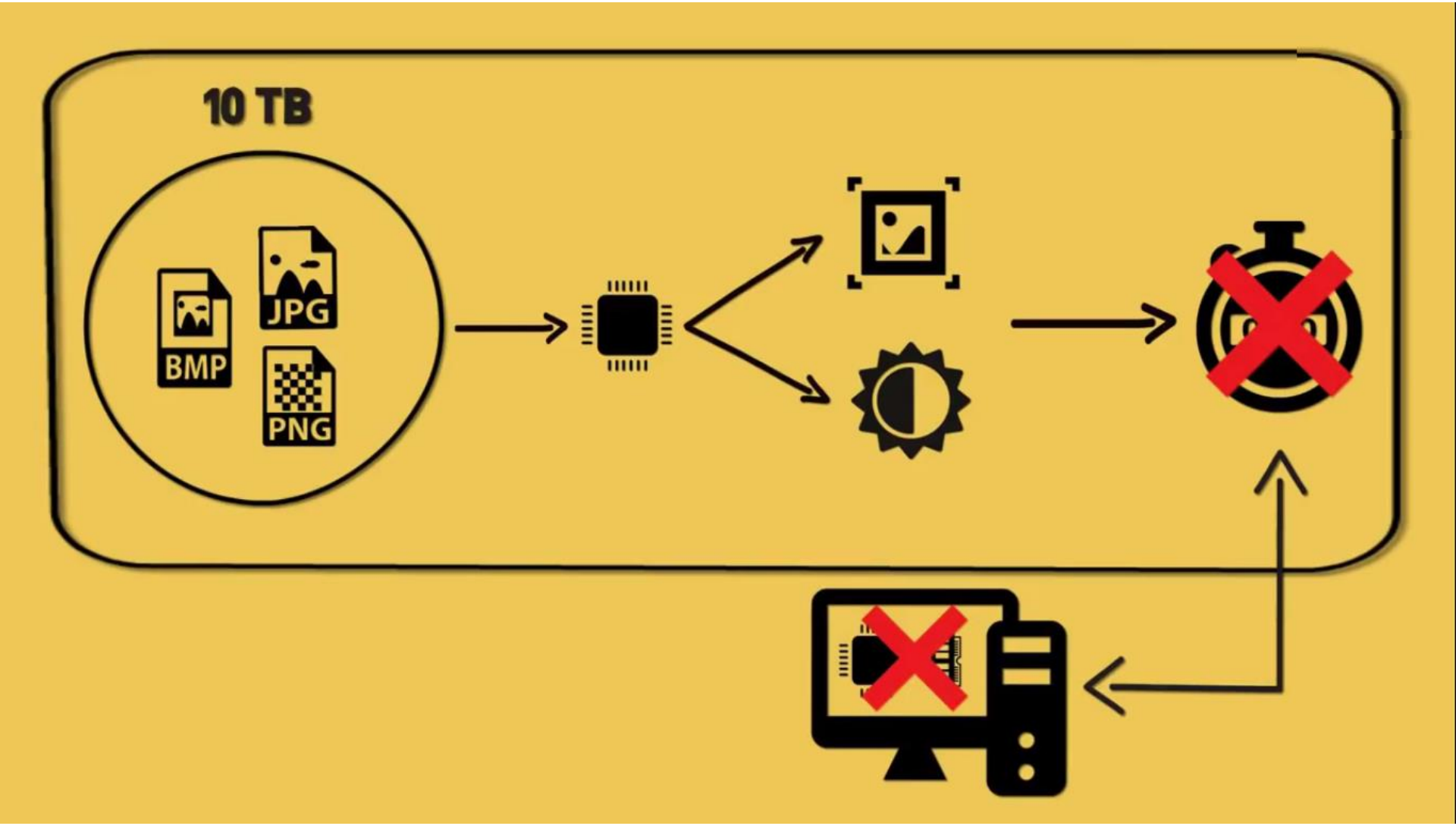

10 TB
BMP
JPG
PNG

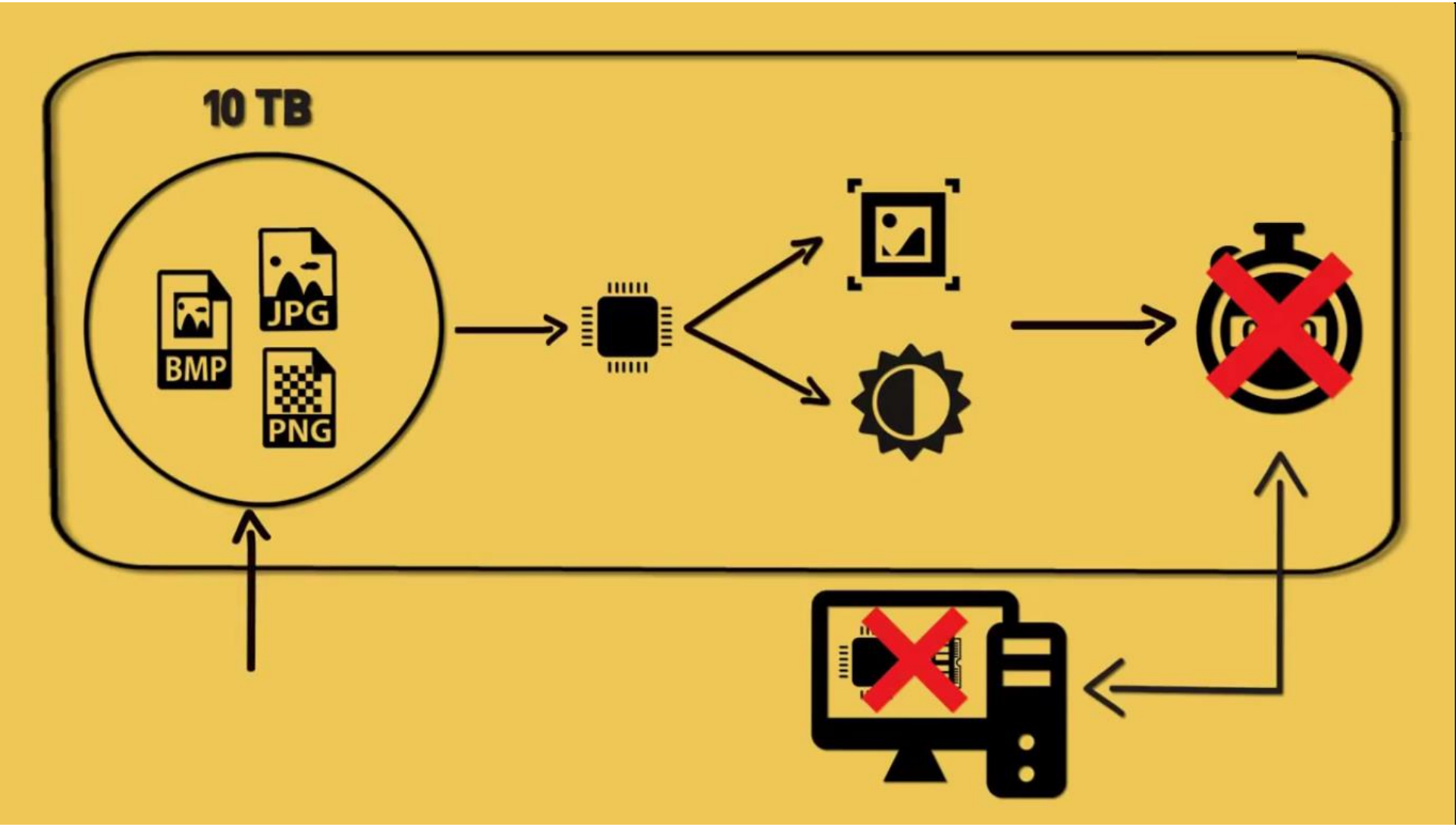

10 TB
BMP
JPG
PNG
0:00

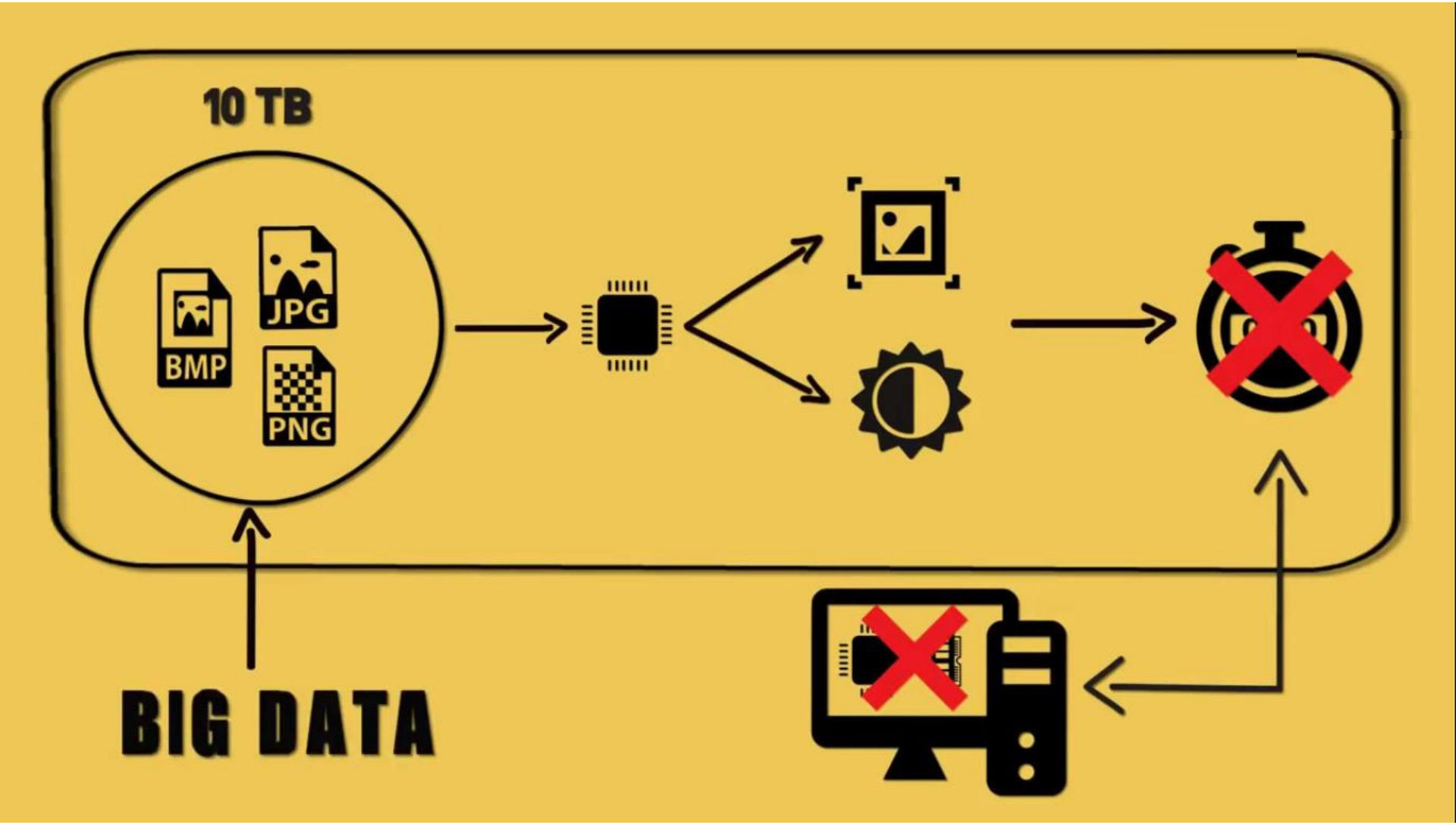

10 TB
BMP
JPG
PNG
0:00


10 TB
BMP
JPG
PNG
0:00


10 TB
BMP
JPG
PNG
0:00


10 TB
BMP
JPG
PNG


10 TB
BMP
JPG
PNG


10 TB
BMP
JPG
PNG
BIG DATA


You
Tube


You
Tube

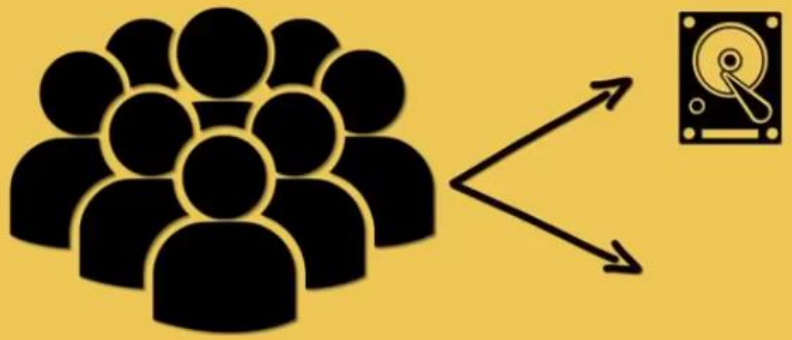

hadoop


You
Tube

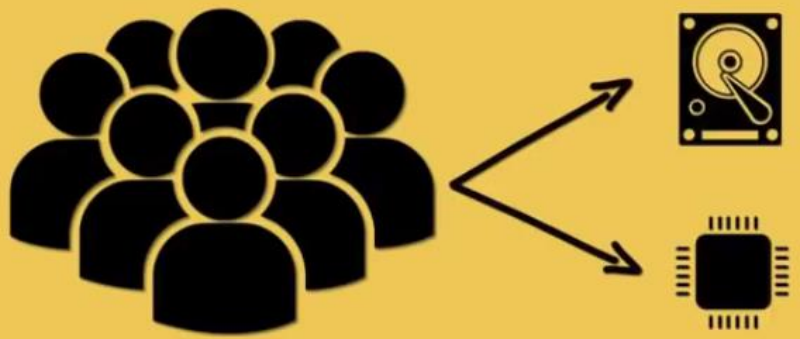

You
Tube

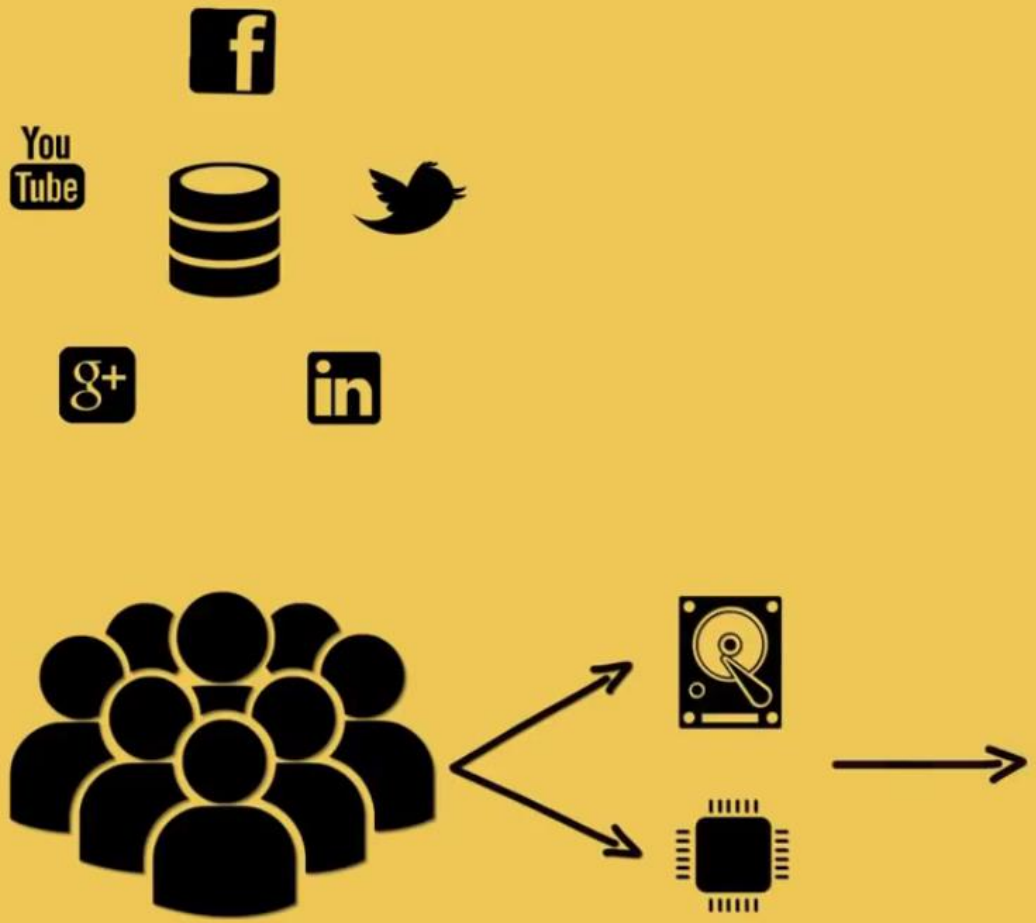

You
Tube

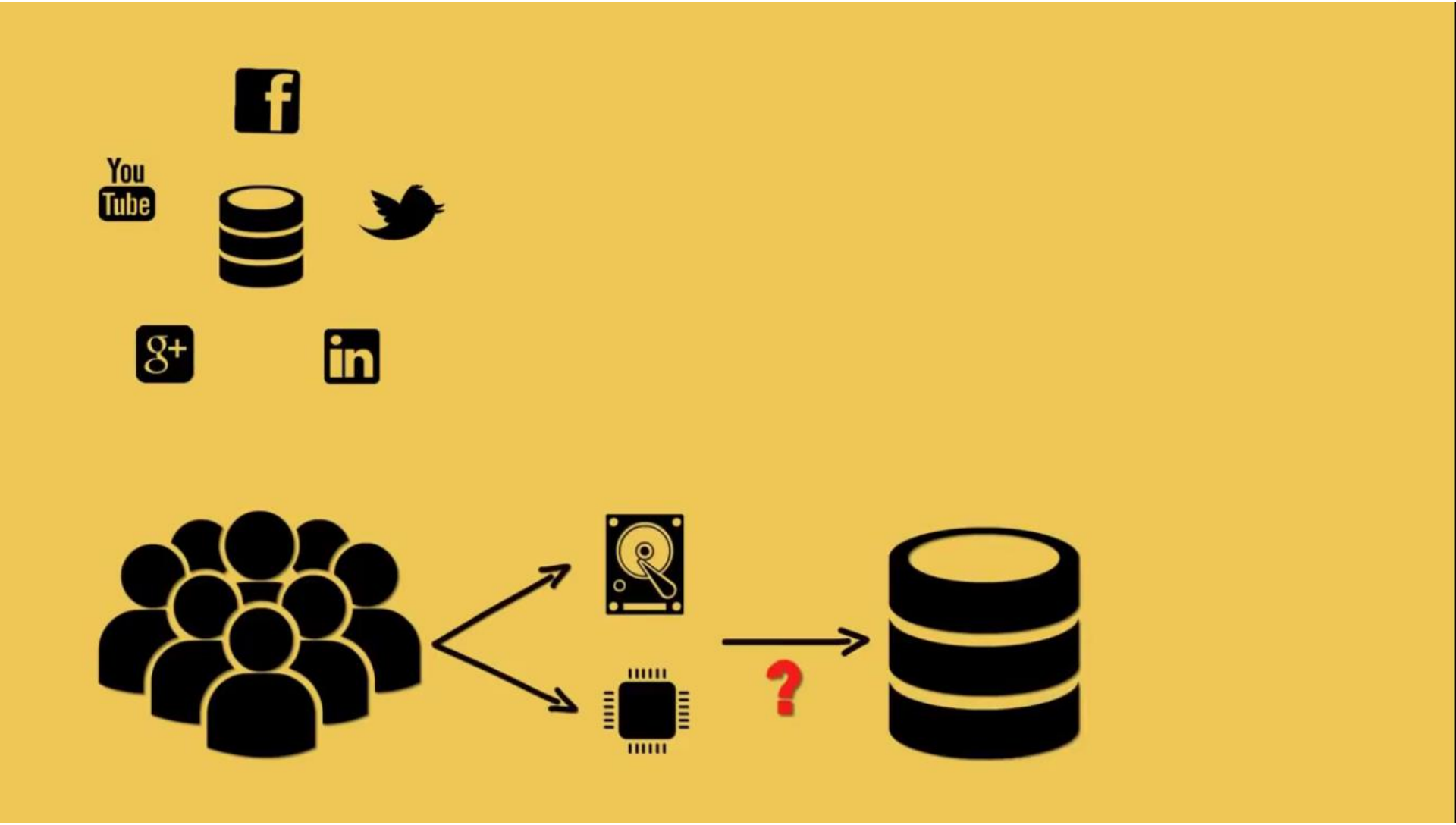

You
Tube

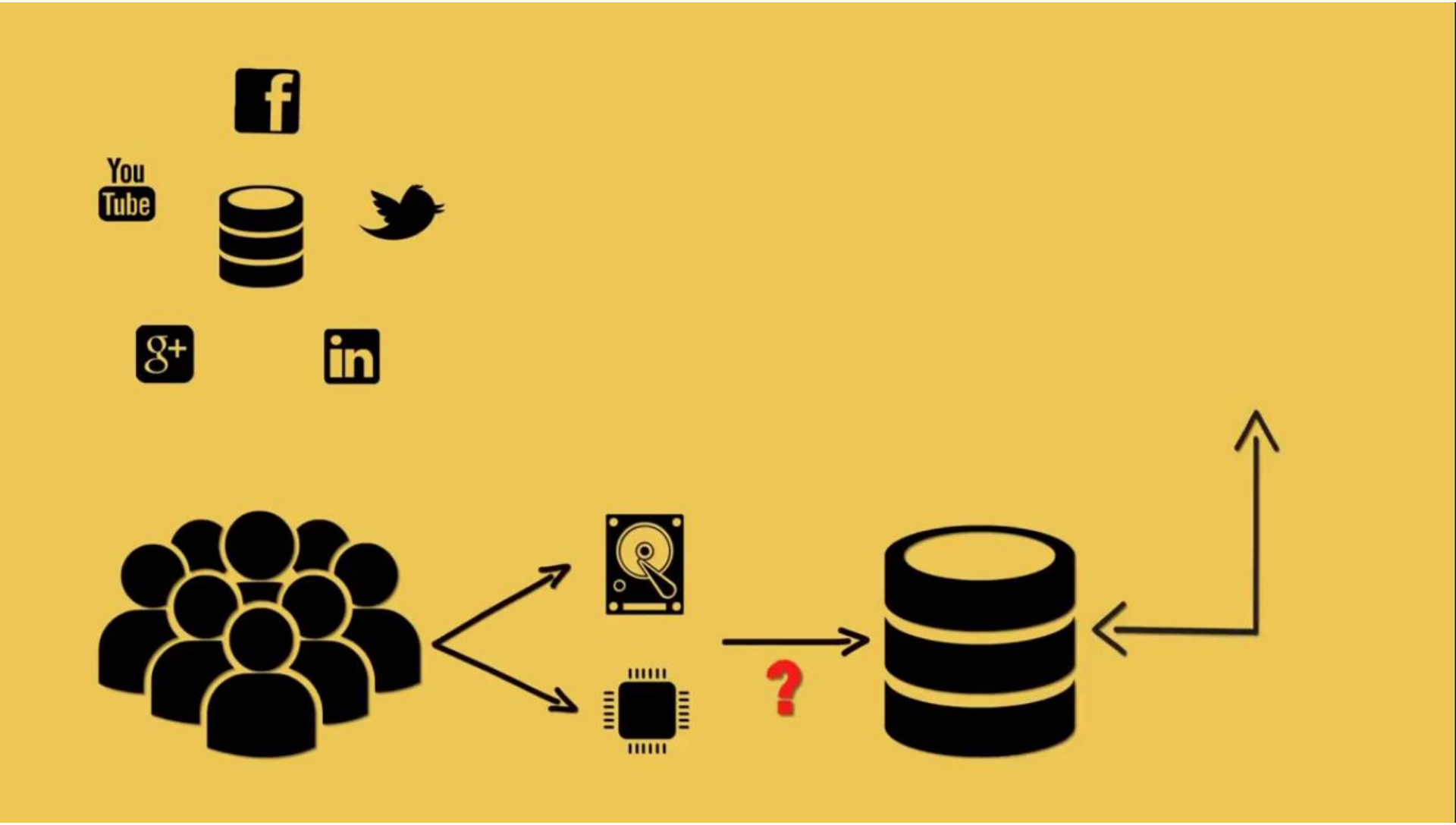


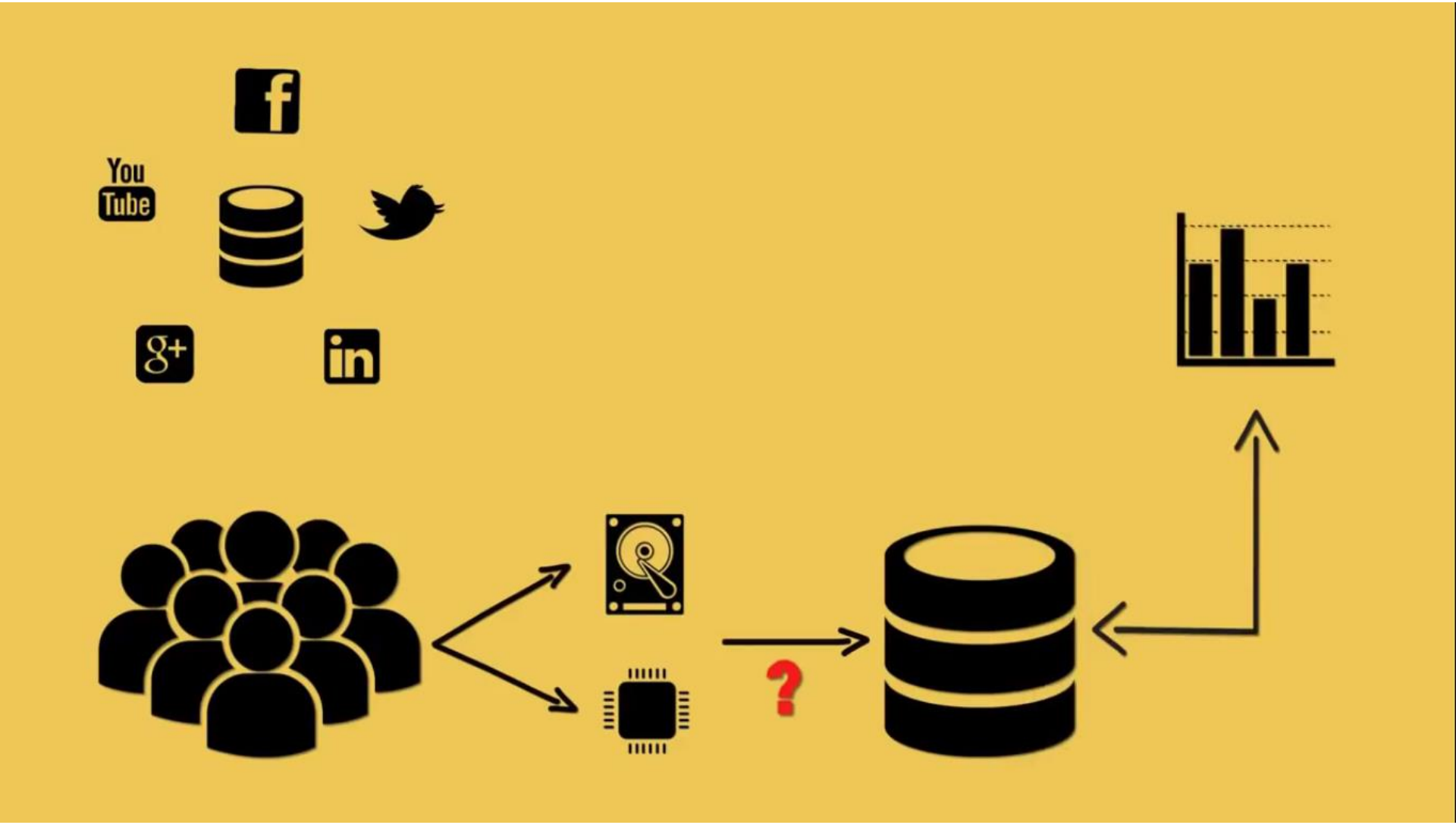

You
Tube

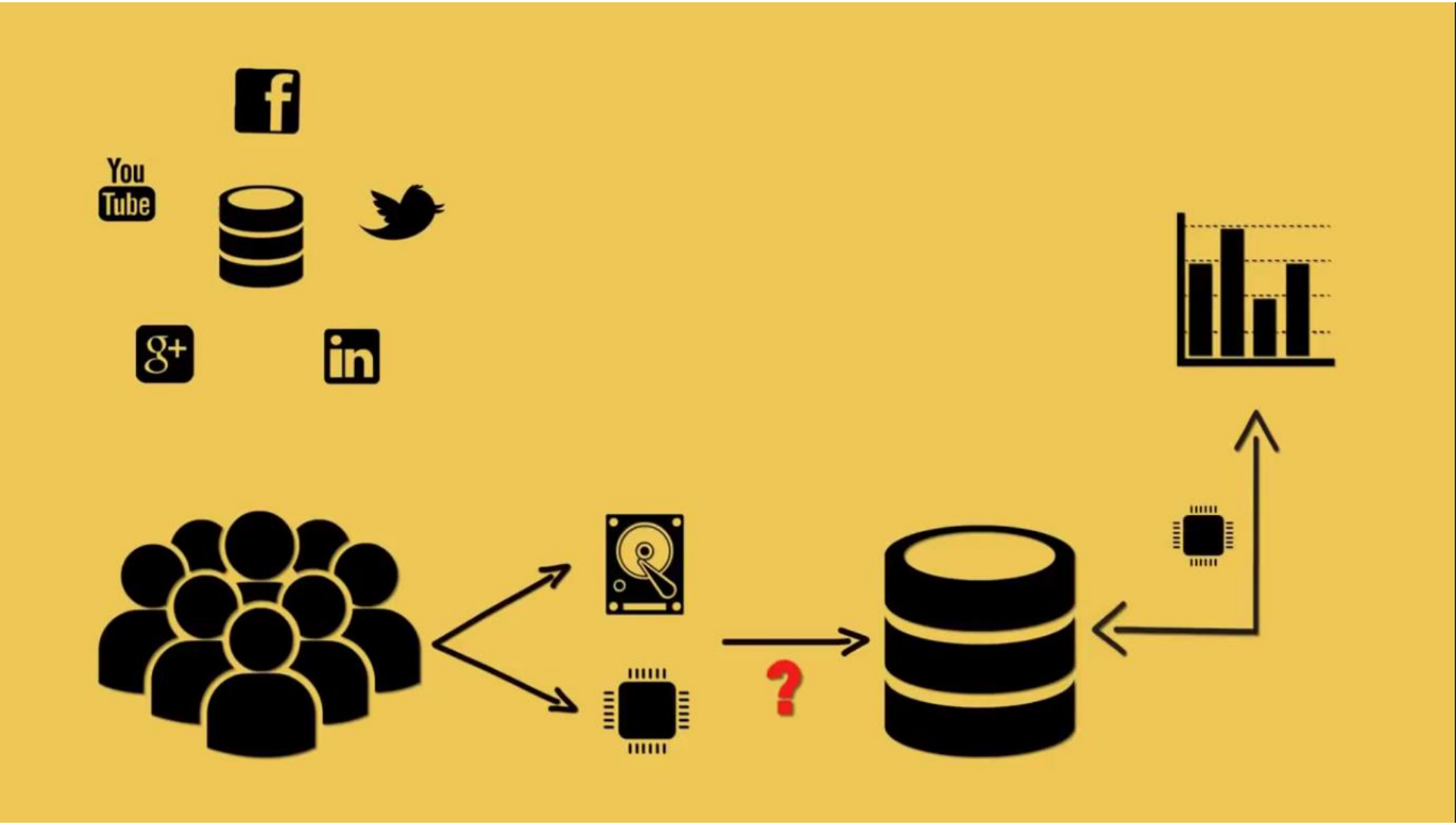


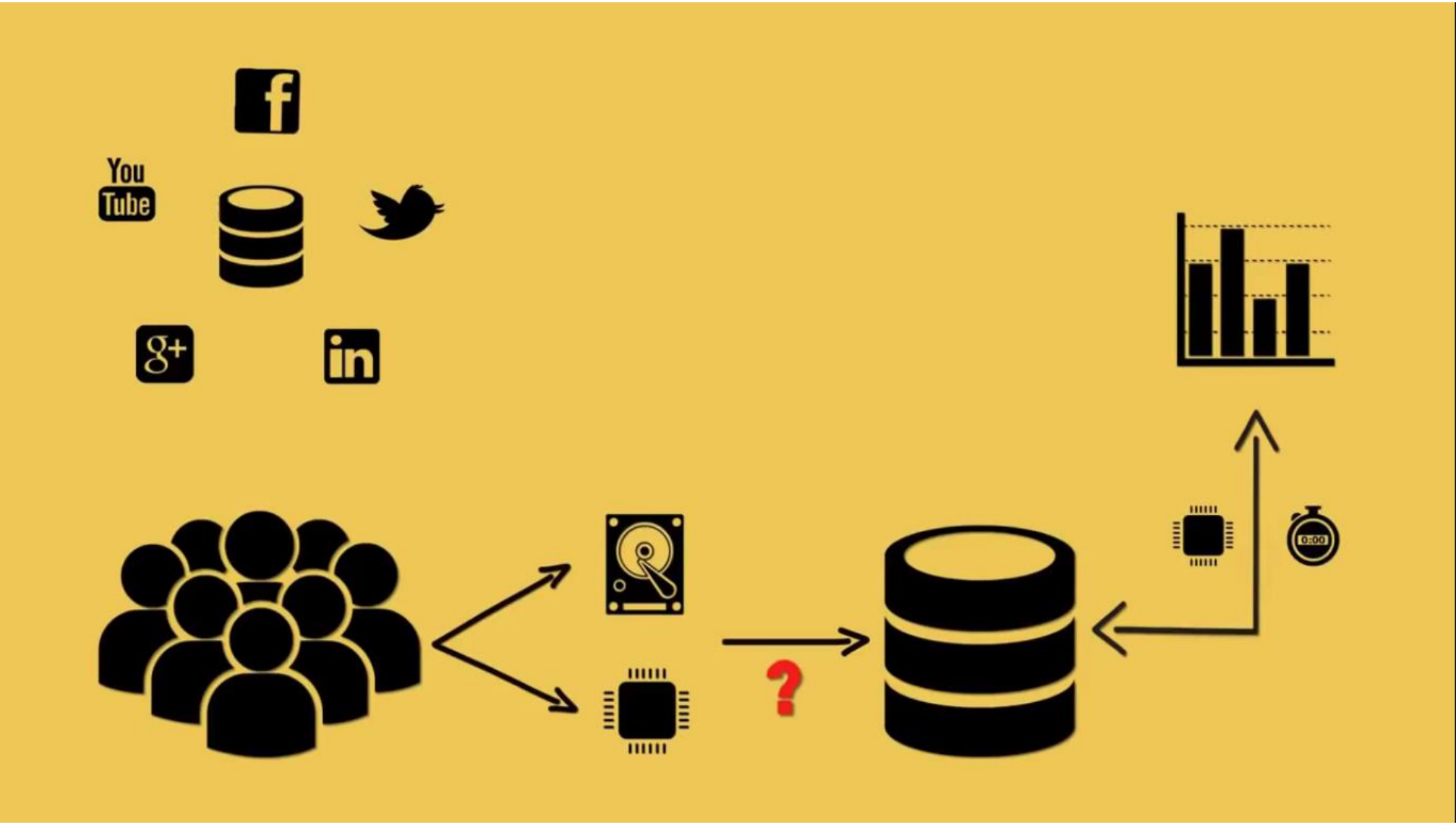

You
Tube

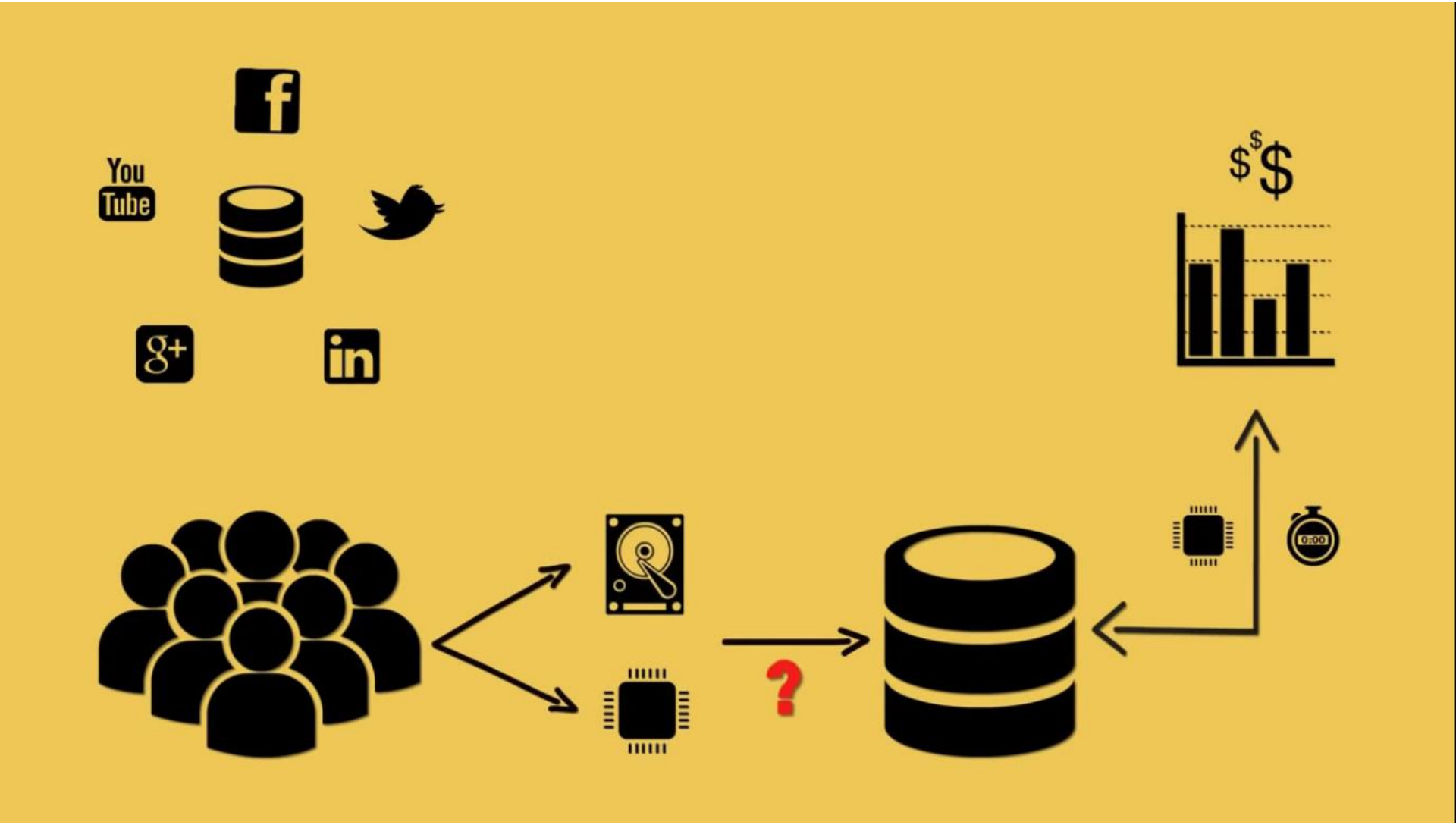

You
Tube
?


You
Tube
?

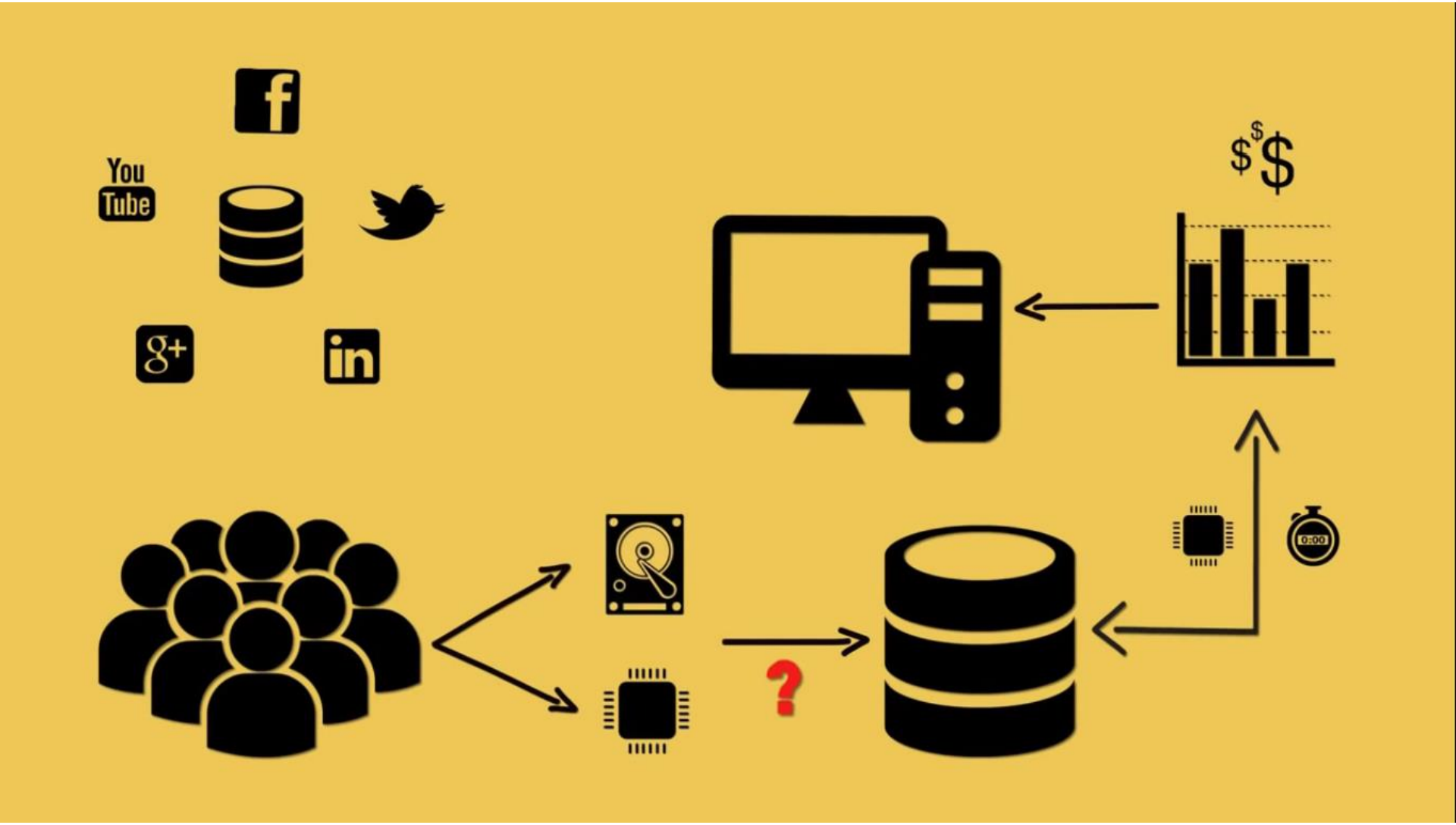

You
Tube


You
Tube

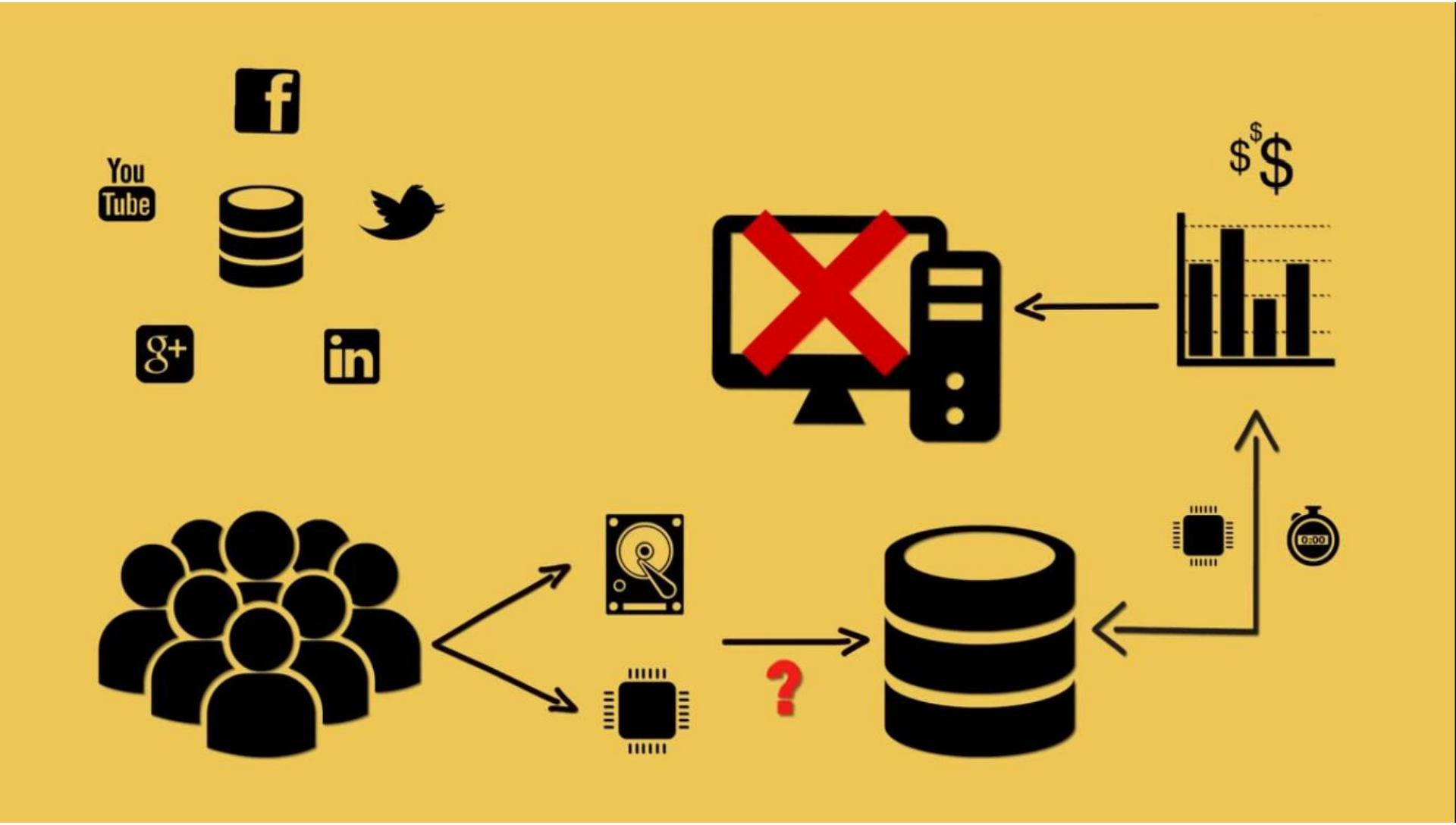

You
Tube
g+
in
0:00
?

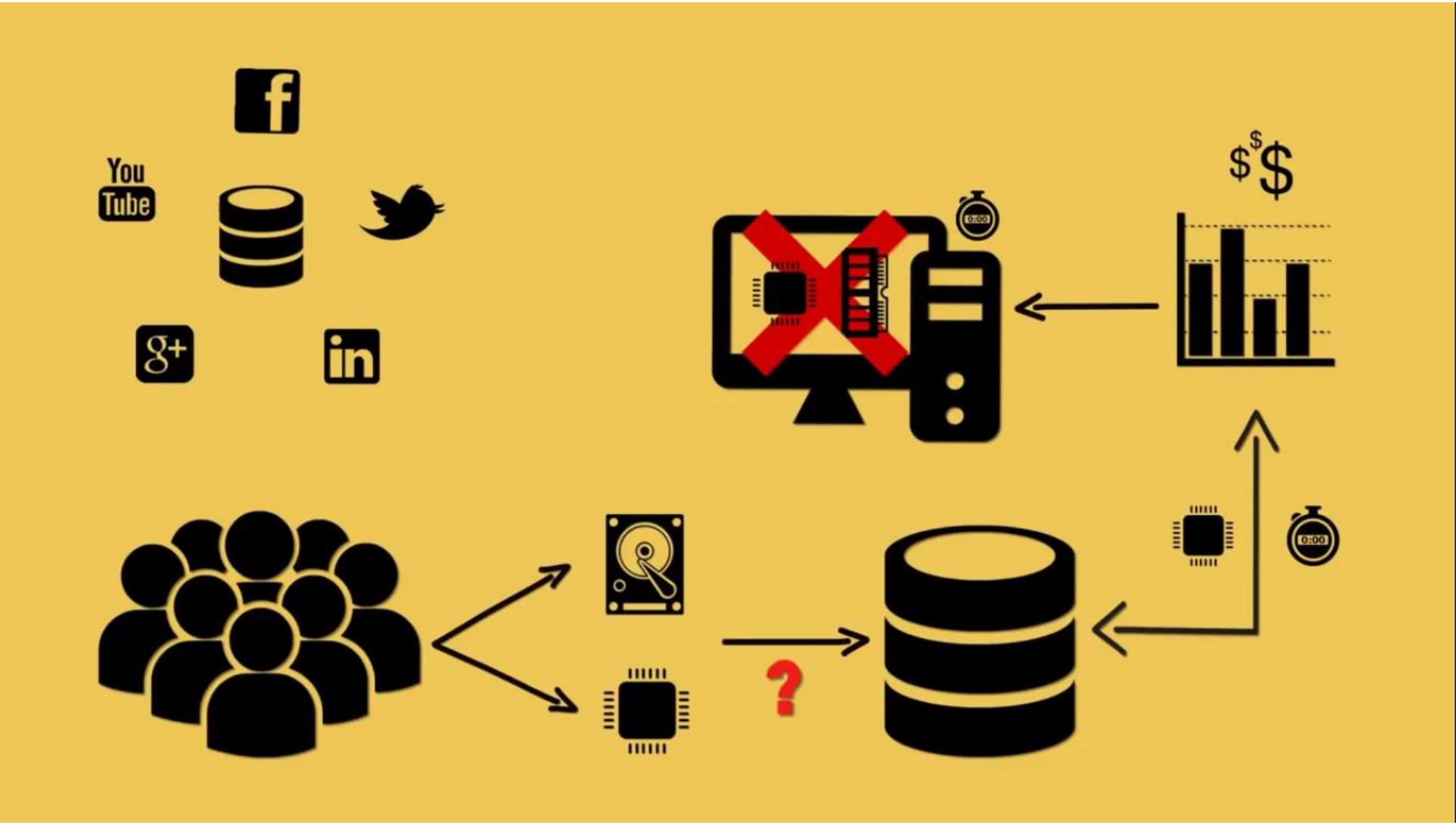


hadoop



hadoop


You
Tube
$
$
$
0:00
?



hadoop


BIG DATA
You Tube
$ $ $
?

CE: What is Big Data?

Rewind…

- Big Data is a collection of large datasets that cannot be processed using traditional computing techniques.
- It is not a single technique or a tool, rather it involves many areas of business and technology.

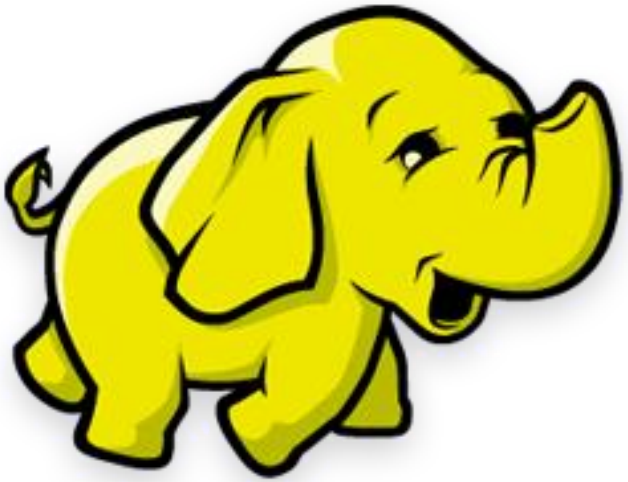
CE: Big Problems with Big Data

Rewind…

- Big Data is …
  - ✓ Unstructured
  - ✓ Unprocessed
  - ✓ Un-aggregated
  - ✓ Un-filtered
  - ✓ Repetitive
  - ✓ Low quality
  - ✓ And generally messy.

Oh, and there is a lot of it.

~: Solution :~



hadoop

CE: 2.4.0. Hadoop

What is Hadoop?

- ***Hadoop** is a **Big Data Solution**.*

✓ An enterprise will have a computer to store and process big data.

✓ For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc.

✓ This approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data.

✓ But when it comes to dealing with huge amounts of *scalable data*, it is a wild task to process such data through a single database bottleneck (block/jam).

CE: 2.4.0. Hadoop (Conti...)

- For this **Google** is the solution.
  - ✓ Google solved this problem using an algorithm called **MapReduce**.
  - ✓ This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.

✓ Using the solution provided by **Google**, Doug Cutting and his team developed an **Open Source Project** called **HADOOP**.

✓ Hadoop runs applications using ***the MapReduce algorithm***, where the data is processed in parallel with others.

- In short, ***Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.***

CE: 2.4.0. Hadoop (Conti...)

- ***Hadoop is an Apache open source framework written in java*** that allows distributed processing of large datasets across clusters of computers using simple programming models.

- Hadoop is most industry accepted standard / tool.

- It is currently used by…
  - ✓ Google,
  - ✓ Facebook,
  - ✓ LinkedIn,
  - ✓ Yahoo,
  - ✓ Twitter, etc.

CE: 2.4.0. Hadoop (Conti...)

- In simple words,

  Hadoop is a "software library", that allows its users to process large datasets across distributed clusters of computers, which enables to gather, store and analyze huge sets of data.

or

- In another simple words…

  "*Hadoop* is a framework that allows us to store and process large data sets in parallel and distributed fashion."

- ***Hadoop provides*** *various* ***tools and technologies***, collectively ***termed as the Hadoop ecosystem***, to enable development and deployment of Big Data solutions.

CE: 2.4.0. Hadoop (Conti...)

Why is Hadoop?

- Flexible
- Scalable
- Building more efficient data economy
- Robust Ecosystem
- Hadoop is getting more “Real-Time”!
- Cost Effective
- Upcoming Technologies using Hadoop
- Hadoop is getting cloudy!

Big Data Framework and File System

2.1. Distributed and Parallel Computing for Big Data

2.2. Cloud Computing versus Big Data

2.3. Cloud Providers in Big Data Market

2.4. Introduction to Distributed File System Framework

2.5. Ecosystem Components

2.6. Distributed File System Architecture

2.7. Processing Data with Framework

2.8. Managing Resources and Applications with YARN

CE: 2.4. Introduction to Distributed File System Framework

CE: 2.4. Introduction to Distributed File System Framework

- A ***Distributed File System (DFS)*** as the name suggests, is a file system that is ***distributed on multiple file servers*** or ***multiple locations***.
- It allows programs to *access* or *store* isolated files as they do with the local ones, *allowing programmers to access files from any network* or *computer*.
- The main purpose of DFS is to allows users of physically distributed systems *to share their data and resources* by using a *Common File System*.
- A ***collection of workstations*** and ***mainframes connected by a Local Area Network (LAN)*** is a configuration on DFS.
- A DFS is executed as a part of the operating system.
- In DFS, a ***namespace*** is created and it process is transparent for the clients.

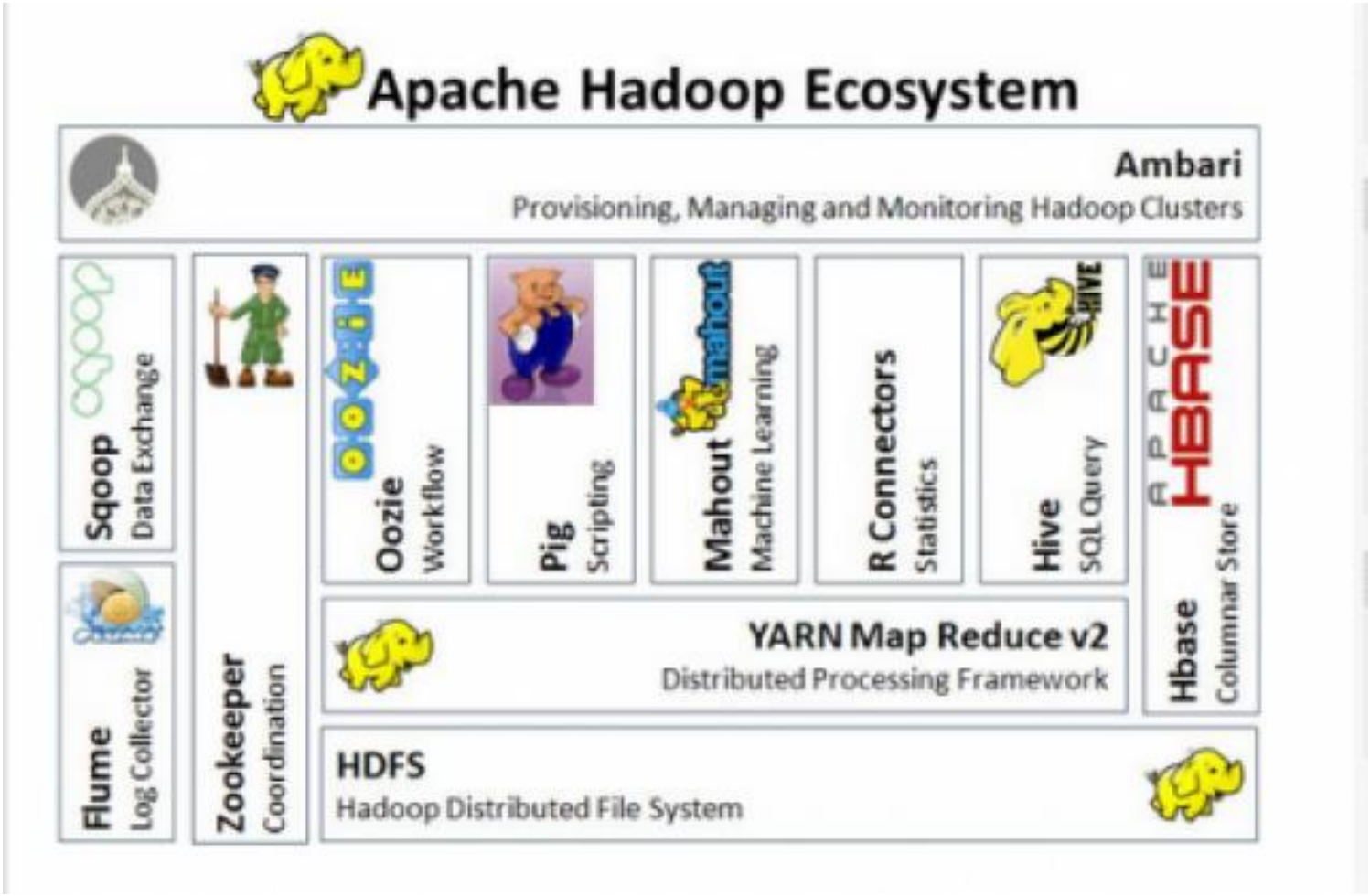
CE: 2.4. Introduction to Distributed File System Framework (Conti…)

The challenges associated with them compared to traditional file systems are as follows −

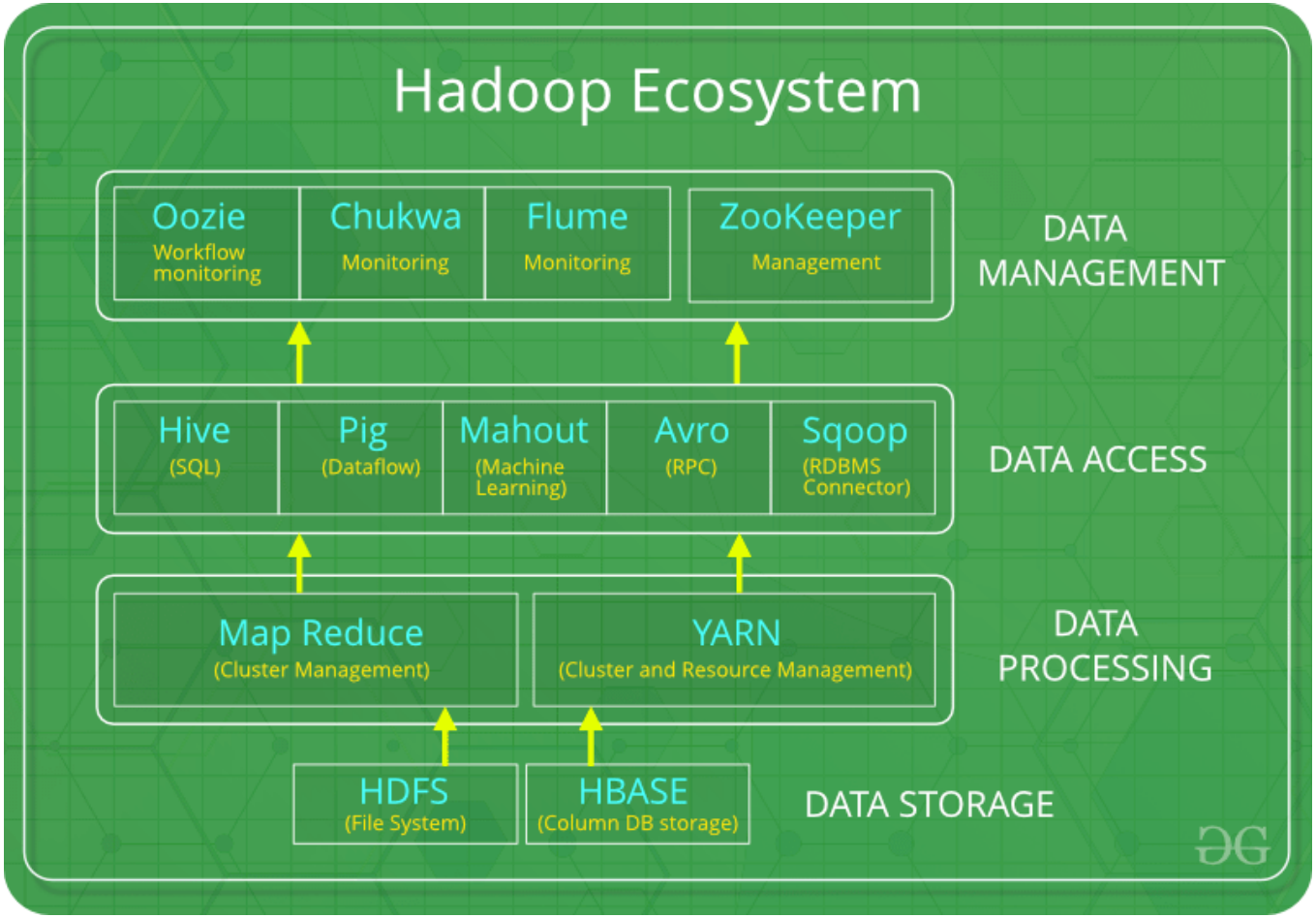
- Data redundancy and inconsistency.
- Difficulty in accessing data.
- Data isolation
- Integrity problems
- Unauthorized access is not restricted.
- It coordinates only physical access.

CE: 2.5.
Hadoop Ecosystem

CE: 2.5. Hadoop Ecosystem



CE: 2.5. Hadoop Ecosystem



Hadoop Ecosystem Elements at various Stages of Data Processing

CE: 2.5. Hadoop Ecosystem

- HDFS -> Hadoop Distributed File System
- HBase -> NoSQL Database
- MapReduce -> Data processing using programming
- YARN -> Yet Another Resource Negotiator
- HIVE, PIG -> Data Processing Services using Query (SQL-like)
- Mahout -> Machine Learning
- Spark -> In-memory Data Processing
- Sqoop -> To Transfer Bulk Data between Hadoop & RDBMS.
- Oozie -> Job Scheduling
- Chukwa -> Managing large distributed systems.
- Flume -> for efficiently collecting, aggregating, and moving large amounts of log data.
- Zookeeper -> Managing Cluster
- Ambari -> Provision, Monitor and Maintain cluster

- Apache Drill -> SQL on Hadoop

CE: 2.5. Hadoop Ecosystem

- HDFS -> Hadoop Distributed File System
- HBase -> NoSQL Database
- MapReduce -> Data processing using pro
- YARN -> Yet Another Resource Nego
- HIVE, PIG -> Data Processing Ser
- Mahout -> Machine Learning
- Spark -> In-memory Data Processing
- Sqoop -> To Transfer Bulk Data between Ha
- Oozie -> Job Scheduling
- Chukwa -> Managing large distributed syste
- Flume -> for efficiently collecting, aggregati of log data.
- Zookeeper -> Managing Cluster
- Ambari -> Provision, Monitor and Maintain cluster

- Apache Drill -> SQL on Hadoop

All these elements enable uses to process large datasets in real time and provide tools to support various types of Hadoop projects, schedule jobs, and manage cluster resources.

CE: 2.5. Hadoop Ecosystem

But….
What is Hadoop Ecosystem?

- **Apache Hadoop ecosystem** refers to the various components of the Apache Hadoop software library.
- It includes open source projects as well as a complete range of complementary tools.
- Some of the most well-known tools of the *Hadoop ecosystem* include… HDFS, Hive, Pig, YARN, MapReduce, Spark, HBase, Oozie, Sqoop, Zookeeper, etc.

CE: 2.5. Hadoop Ecosystem (Conti…)

- **MapReduce** and **H**adoop **D**istributed **F**ile **S**ystem (**HDFS**) … are two core components of the Hadoop ecosystem, that provide a great starting point to manage Big Data.

- Both, MapReduce and HDFS provide the necessary services and basic structure to deal with the core requirements of Big Data solutions.

- Other services and tools of the ecosystem provide the environment and components required to build and manage purpose-driven Big Data applications.

CE: 2.6. Hadoop Distributed File System

- It is the most important component of Hadoop Ecosystem.
- HDFS, is one of the largest Apache project and primary storage system of Hadoop.
- It is a java based file system that provides scalable, fault tolerance, reliable and cost efficient data storage for Big data.
- It is a distributed file system that runs on commodity (product/service) hardware.
- HDFS is already configured with default configuration for many installations. Most of the time for large clusters configuration is needed.
- Hadoop interact directly with HDFS by shell-like commands.

CE: 2.6. Hadoop Distributed File System (Conti…)

- HDFS has a master-slave architecture.
- It consists of a single **NameNode** and a number of **DataNodes**.

CE: 2.6. Hadoop Distributed File System (Conti…)

1. NameNode:

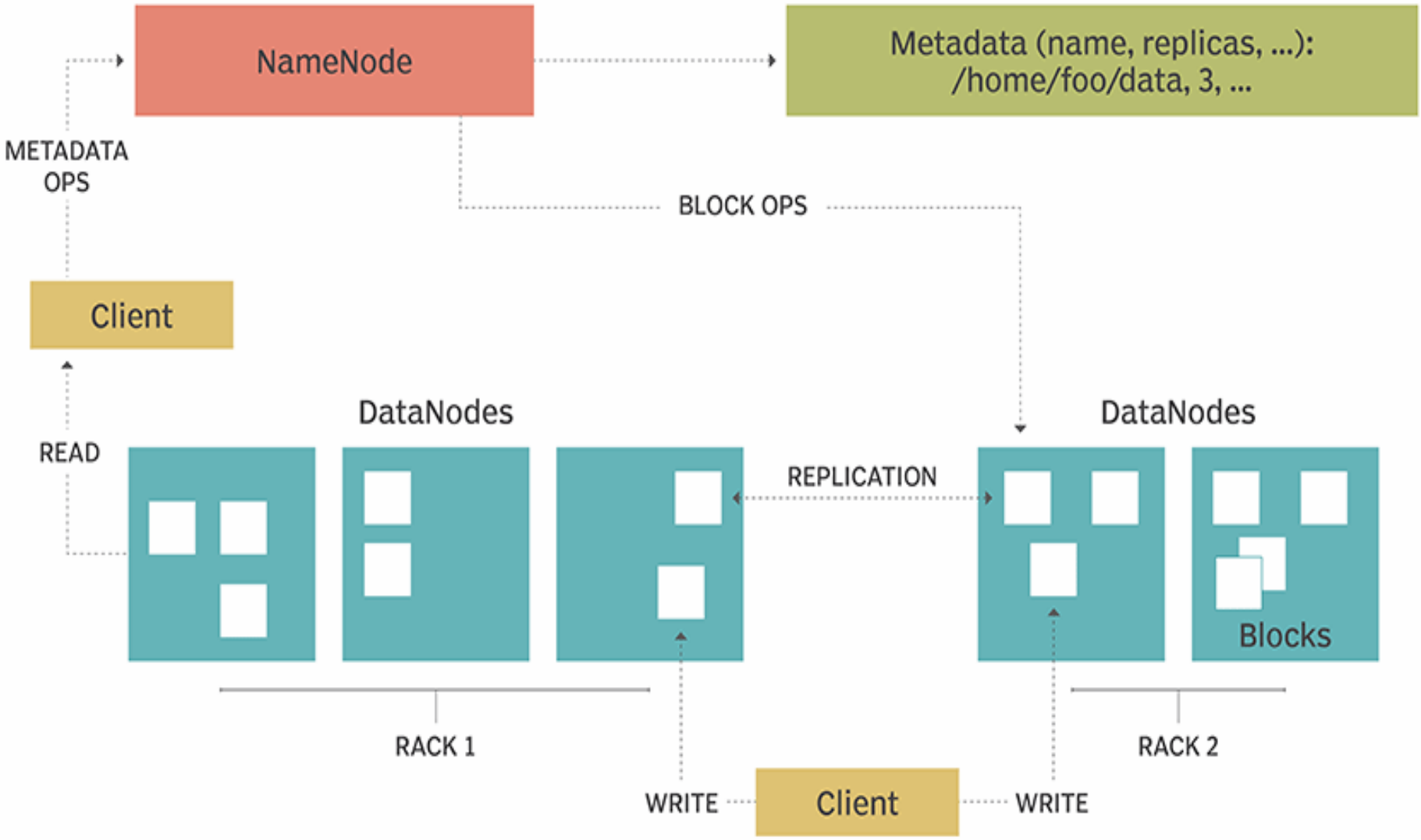
- It is also known as ***Master node***.
- NameNode does not store actual data or dataset.
- NameNode stores Metadata i.e. number of blocks, their location, on which ***Rack,*** Datanode data is stored and other details.
- It consists of files and directories.
- Tasks of HDFS NameNode:
  - ✓ Manage file system namespace.
  - ✓ Regulates client's access to files.
  - ✓ Executes file system execution such as naming, closing, opening files and directories.

CE: 2.6. Hadoop Distributed File System (Conti…)

2. DataNode:

- It is also known as ***Slave node***.
- It is responsible for storing actual data in HDFS.
- Data node performs read and write operation as per the request of the client0s.
- Replica block of Datanode consists of 2 files on the file system. The 1st file is for data and 2nd file is for recording the block's metadata.
- HDFS Metadata includes checksums for data.
- At startup, each Datanode connects to its corresponding Namenode and does handshaking. Verification of namespace ID and software version of DataNode take place by handshaking. At the time of mismatch found, DataNode goes down automatically.

HDFS architecture



CE: 2.6. Hadoop Distributed File System (Conti…)

- Internally, a file gets **split** into a number of ***data blocks*** and stored on a group of ***slave machines***.
- *Namenode manages* modifications to file system namespace. These are actions like the opening, closing and renaming files or directories.
- *NameNode also keeps track of mapping of blocks to DataNodes*. This DataNodes serves read/write request from the file system's client.
- *DataNode* also creates, deletes and replicates blocks on demand from NameNode.

CE: 2.6. Hadoop Distributed File System (Conti…)

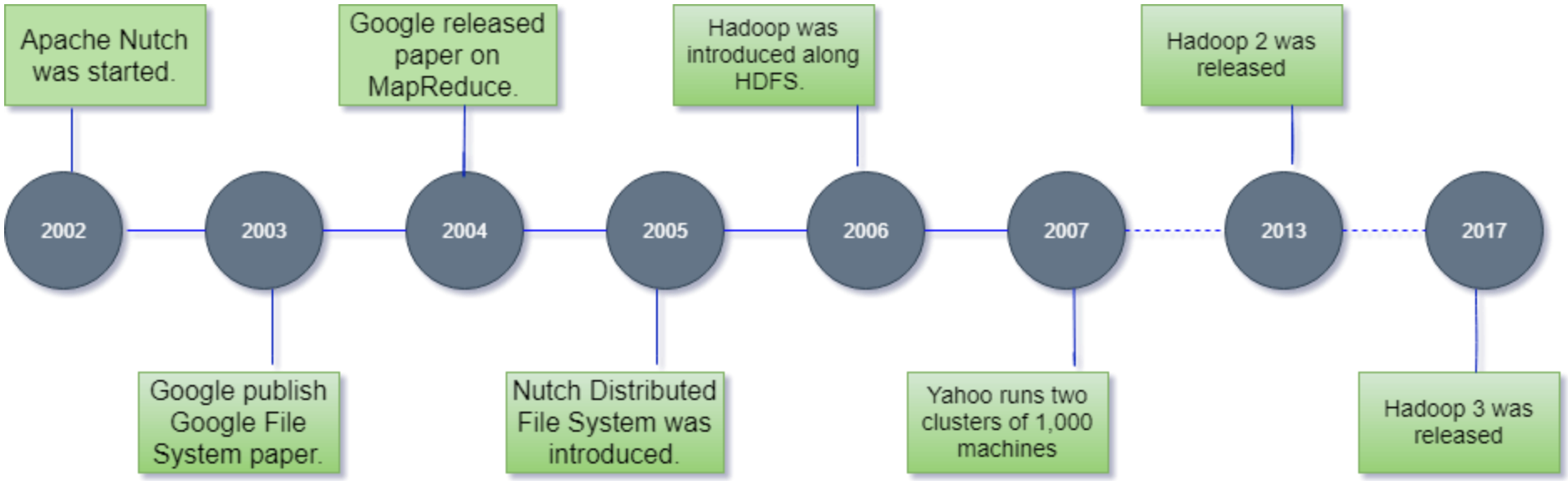
- Tasks of HDFS DataNode:
  - ✓ DataNode performs operations like block replica creation, deletion, and replication according to the instruction of NameNode.
  - ✓ DataNode manages data storage of the system.

CE: 2.6. Hadoop Distributed File System (Conti…)

What is Block in HDFS?

- Block is nothing but the smallest unit of storage on a computer system.
- It is the smallest contiguous storage allocated to a file.
- In Hadoop, we have a default block size of **128MB**.

CE: Hadoop History



CE: 2.7. Processing Data with Framework - MapReduce

CE: 2.7. MapReduce

- ***Hadoop MapReduce*** is the *core Hadoop ecosystem component* which provides data processing.

- MapReduce is a ***software framework*** and ***programming model*** used for processing huge amount of *structured* and *unstructured* data stored in the Hadoop Distributed File system.

- As the processing component, ***MapReduce is the heart of Apache Hadoop***.

- The term "**MapReduce**" refers to two separate phases:
  1. Map - deal with **splitting** and **mapping** of data.
  2. Reduce - deal with **shuffling** and **reducing** the data.

CE: 2.7. MapReduce (Conti…)

- As the sequence of the name, MapReduce implies…..
  the reduce job is always performed after the map job.

- ***Hadoop is capable of running MapReduce programs*** written in various languages: Java, Ruby, Python, and C++.

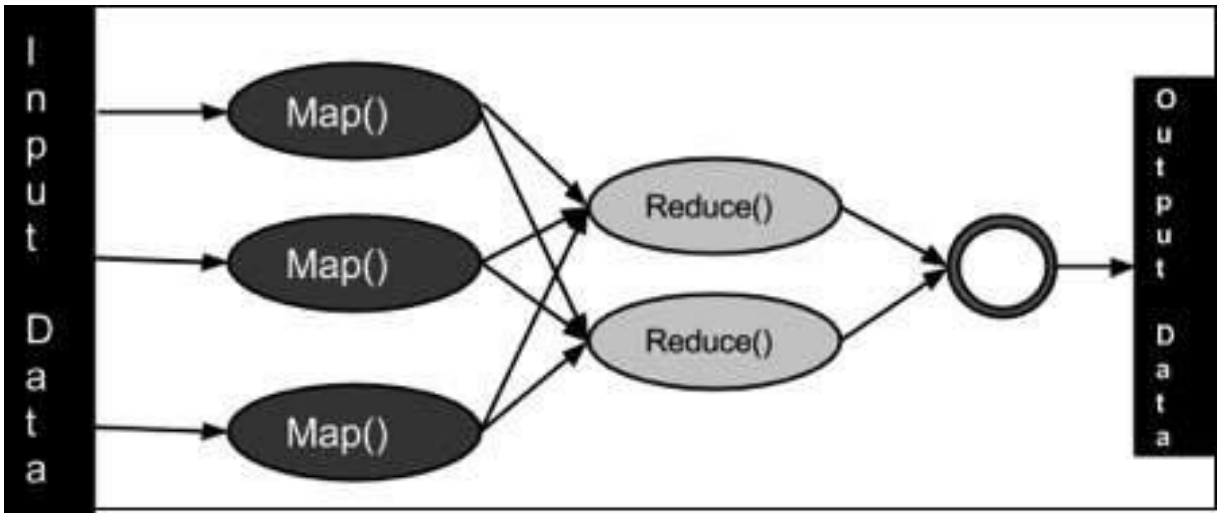
- The programs of *Map Reduce in cloud computing are parallel in nature*, thus are very useful for performing large-scale data analysis using multiple machines in the cluster.

- The input to each phase is ***key-value*** pairs.

- In addition, every programmer needs to specify two functions:
  1. Map()
  2. reduce ()

CE: 2.7. MapReduce (Conti…)

- A simplified flow diagram for the MapReduce program is….

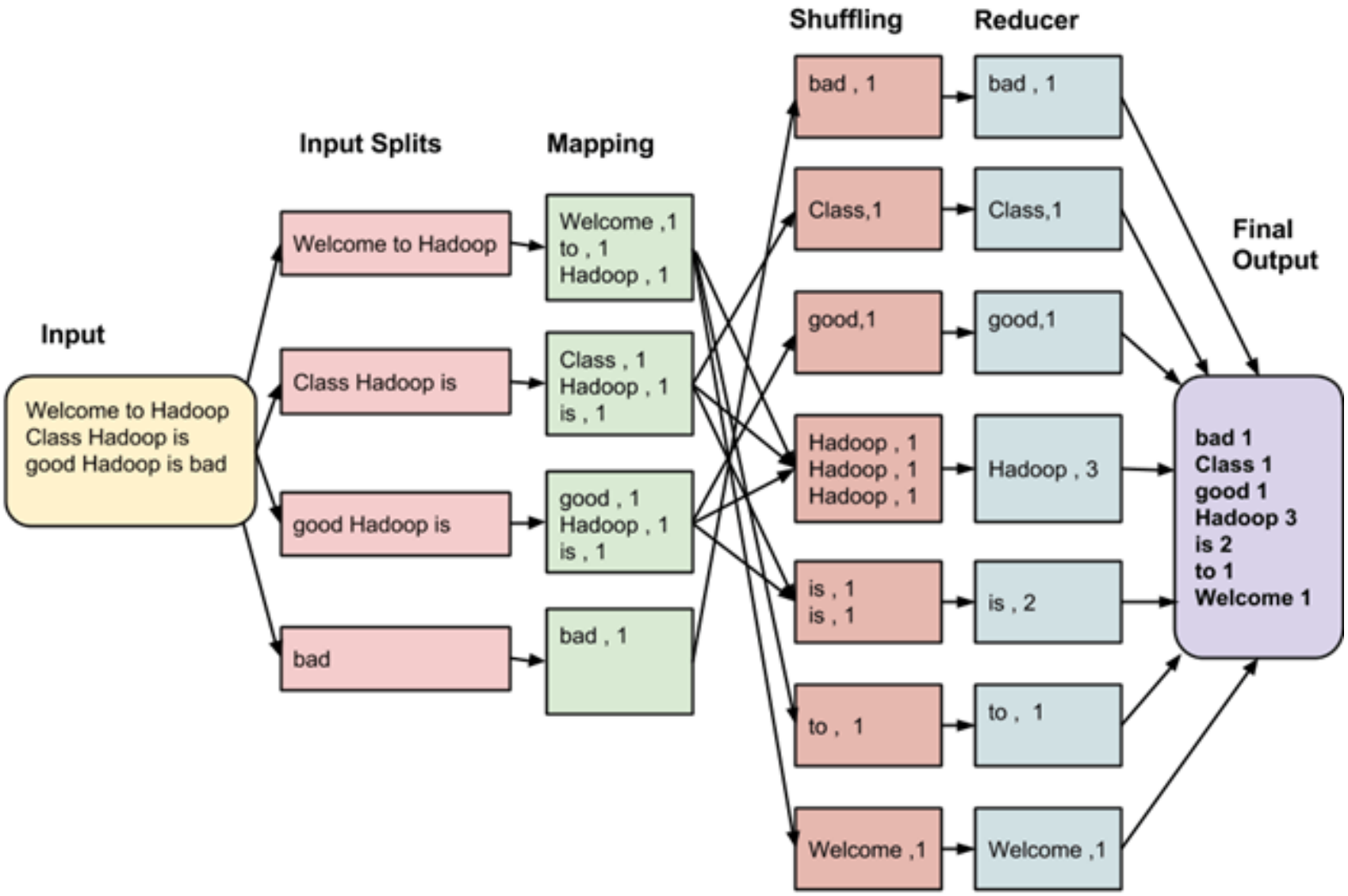


CE: 2.7. MapReduce (Conti…)

- The whole process goes through four phases of execution… i.e.
  - ✓ **splitting,**
  - ✓ **mapping,**
  - ✓ **shuffling, and**
  - ✓ **reducing**.

- For example:
  - ✓ Consider that the following input data for MapReduce in Big data Program:
    - Welcome to Hadoop Class
    - Hadoop is good
    - Hadoop is bad

CE: 2.7. MapReduce (Conti…)

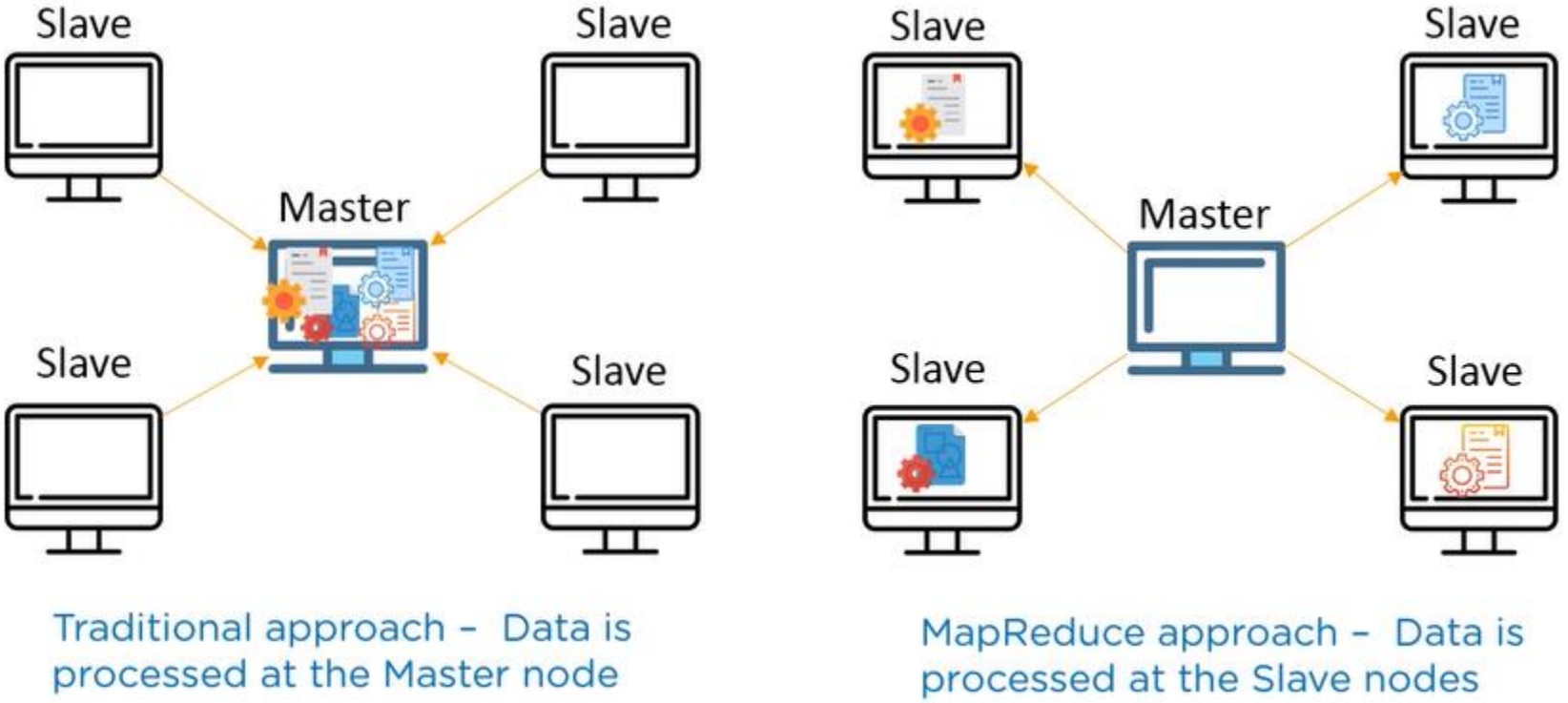


CE: 2.7. MapReduce (Conti…)

- MapReduce is used for parallel processing, of the Big Data, which is stored in HDFS.

CE: 2.7. MapReduce (Conti…)

- In MapReduce approach, ***processing is done at the slave nodes*** and the ***final result is sent to the master node***.



CE: Applications of MapReduce

❖ E-commerce:

- E-commerce companies such as ***Walmart, E-Bay, and Amazon*** use MapReduce to analyze buying behavior.
- MapReduce provides meaningful information that is used as the basis for developing product recommendations.

❖ Social networks:

- The MapReduce programming tool can evaluate certain information on social media platforms such as ***Facebook, Twitter, and LinkedIn***.
- It can evaluate important information such as who liked your status and who viewed your profile.

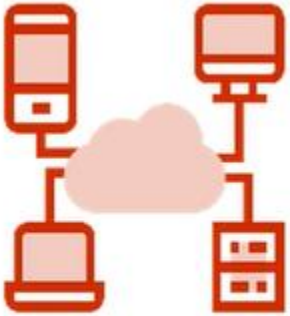
❖ Entertainment:

- ***Netflix*** uses MapReduce to analyze the clicks and logs of online customers.
- This information helps the company suggest movies based on customers' interests and behavior.

CE: 2.8
Managing Resources and Applications with YARN

CE: 2.8.1. Managing Resources with YARN

- YARN stands for "**Y**et **A**nother **R**esource **N**egotiator".
- It is the resource management layer.
- YARN is responsible for *resource allocation* and *job scheduling*.



CE: 2.8.1. Managing Resources with YARN (Conti…)

- Before 2012, users could write MapReduce programs using *scripting languages* such as *Java*, *Python*, and *Ruby*. They could also use *Pig*, a language used to transform data. No matter what language was used, its implementation depended on the MapReduce processing model.

- In May 2012, during the release of Hadoop version 2.0, YARN was introduced. It was no longer limited to working with the MapReduce framework anymore as *YARN supports multiple processing models in addition to MapReduce*, such as ***Spark***.

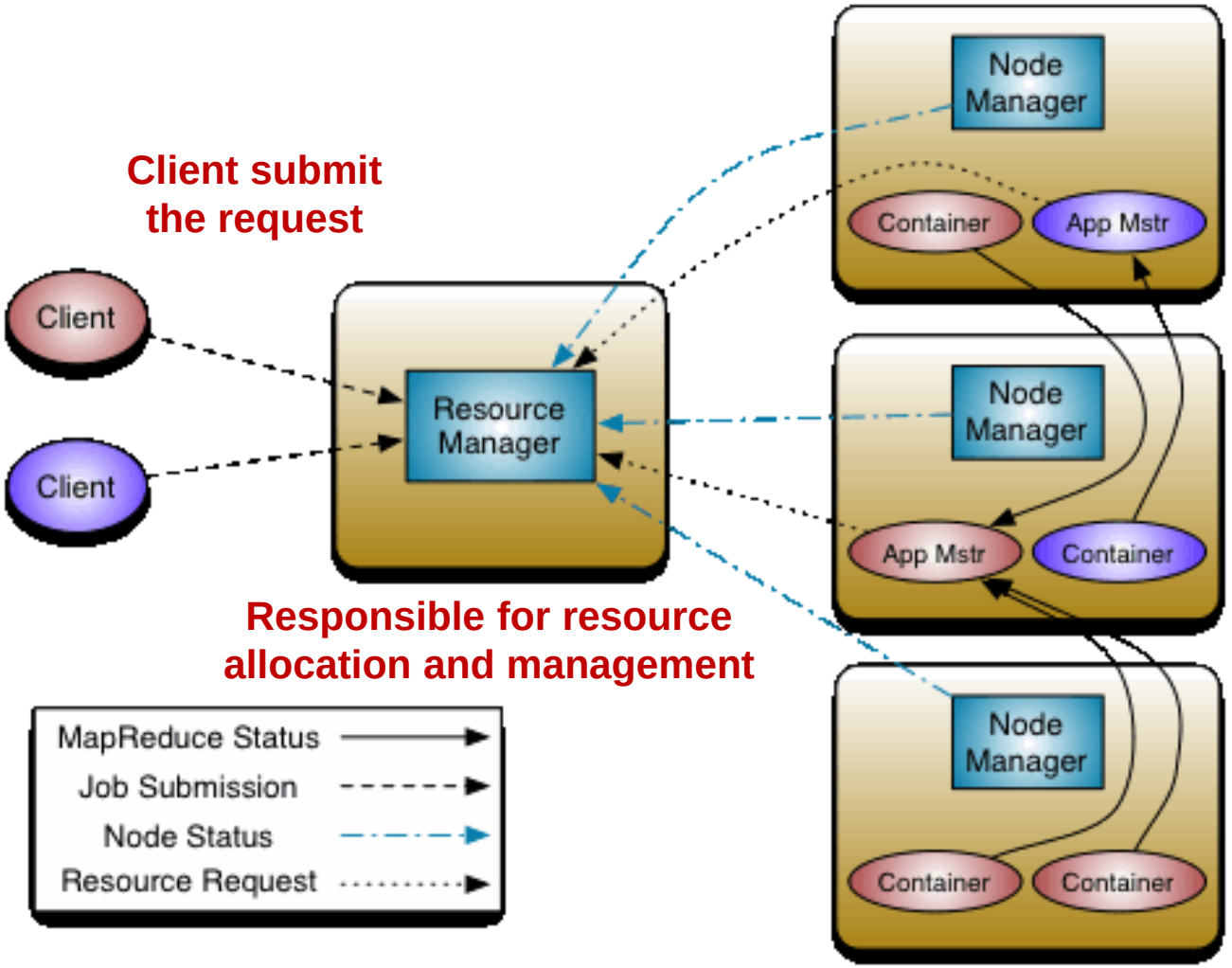
- Other features of YARN include significant performance improvement and a flexible execution engine.

CE: 2.8.1. Managing Resources with YARN (Conti…)

- YARN also allows different data processing engines like…
  - graph processing,
  - interactive processing,
  - stream processing
  - as well as batch processing
- to run and process data stored in HDFS (Hadoop Distributed File System), thus making the system much more efficient.

- Through its various components, it can dynamically allocate various resources and schedule the application processing.

CE: 2.8. Hadoop YARN Architecture

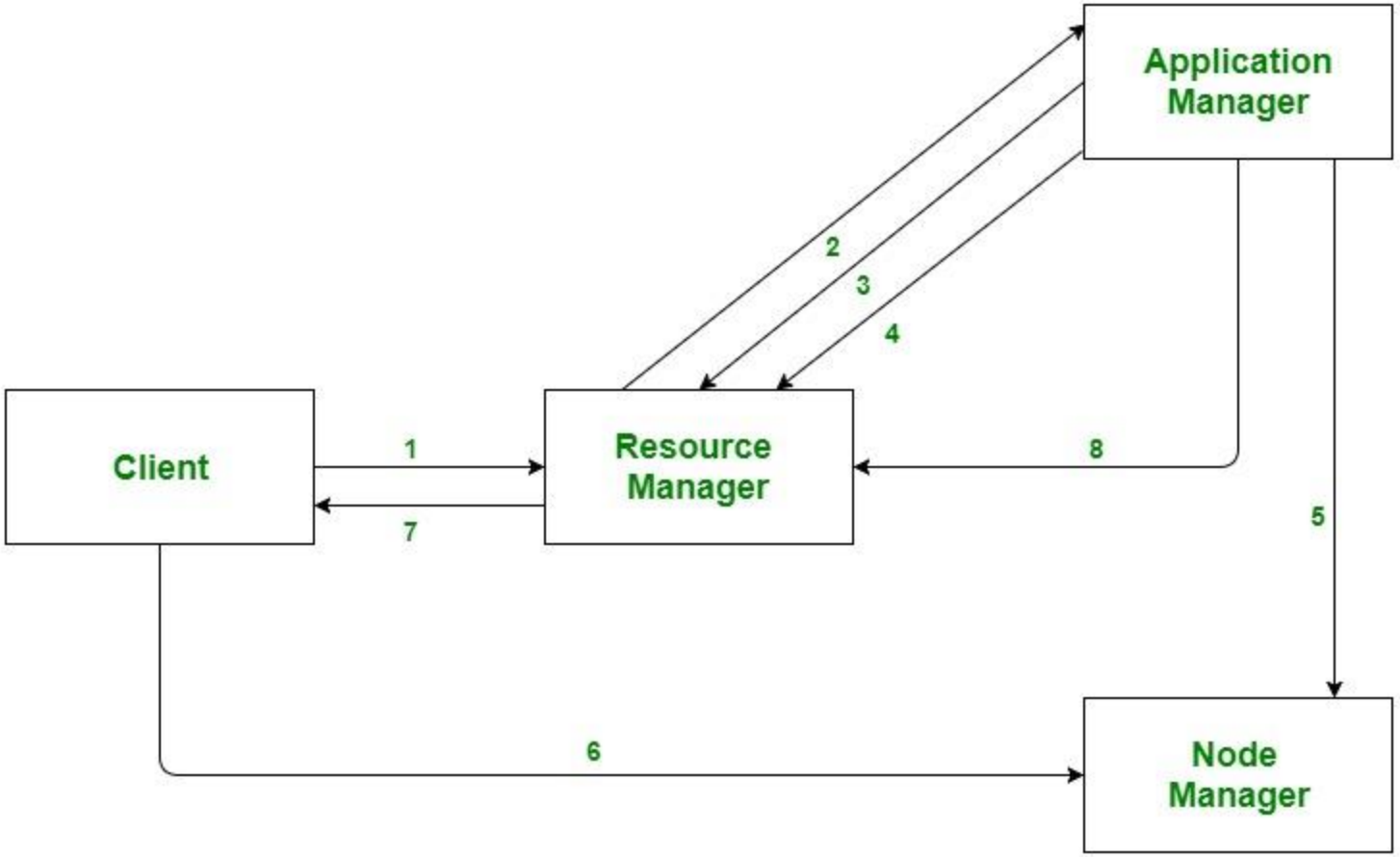


Node Manager
Manages the nodes and monitor resource usage

Container
is a collection of physical resources such as RAM, CPU

App Master
requests container from the NodeManager

CE: 2.8. Hadoop YARN Architecture (Conti…)



Application workflow in Hadoop YARN

CE: 2.8. Hadoop YARN Architecture (Conti…)

- Application workflow in Hadoop YARN:

1. Client submits an application
2. The Resource Manager allocates a container to start the Application Manager
3. The Application Manager registers itself with the Resource Manager
4. The Application Manager negotiates containers from the Resource Manager
5. The Application Manager notifies the Node Manager to launch containers
6. Application code is executed in the container
7. Client will get response from Resource Manager/Application Manager to monitor application's status
8. Once the processing is complete, the Application Manager un-registers with the Resource Manager

CE: 2.8.2. Hadoop YARN Features

❖ Scalability:

- The scheduler in Resource manager of YARN architecture allows Hadoop to extend and manage thousands of nodes and clusters.

❖ Compatibility:

- YARN supports the existing map-reduce applications without disruptions thus making it compatible with Hadoop 1.0 as well.

❖ Cluster Utilization:

- Since YARN supports Dynamic utilization of cluster in Hadoop, which enables optimized Cluster Utilization.

❖ Multi-tenancy:

- It allows multiple engine access thus giving organizations a benefit of multi-tenancy.

YARN vs. MapReduce

How the two technologies differ on cluster resource management

YARN	MapReduce
Supports a variety of processing engines and applications	Supported its own batch processing applications only
Separates its duties across multiple components	Consolidated most of its work in a single component
Can dynamically allocate pools of resources to applications	Provided static allocations of resources for designated tasks


Thank
You


hadoop