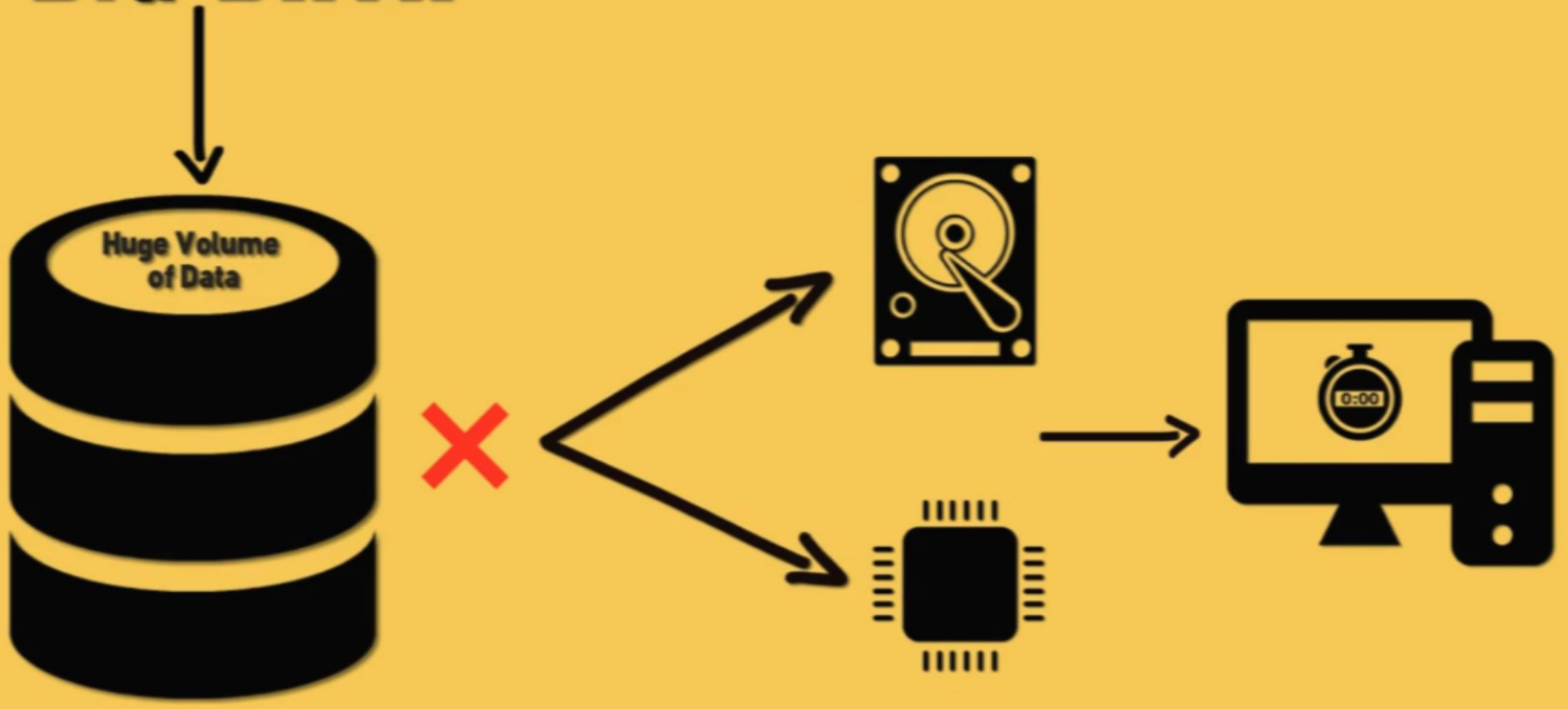
WHAT IS BIGDATA?

KARTHICK SELVAM

WHAT IS BIGDATA?

- Bigdata can be referred to as huge volume of data that cannot be stored and processed in the traditional approach with in the given time frame.

BIG DATA



HOW HUGE THIS DATA NEEDS TO BE?

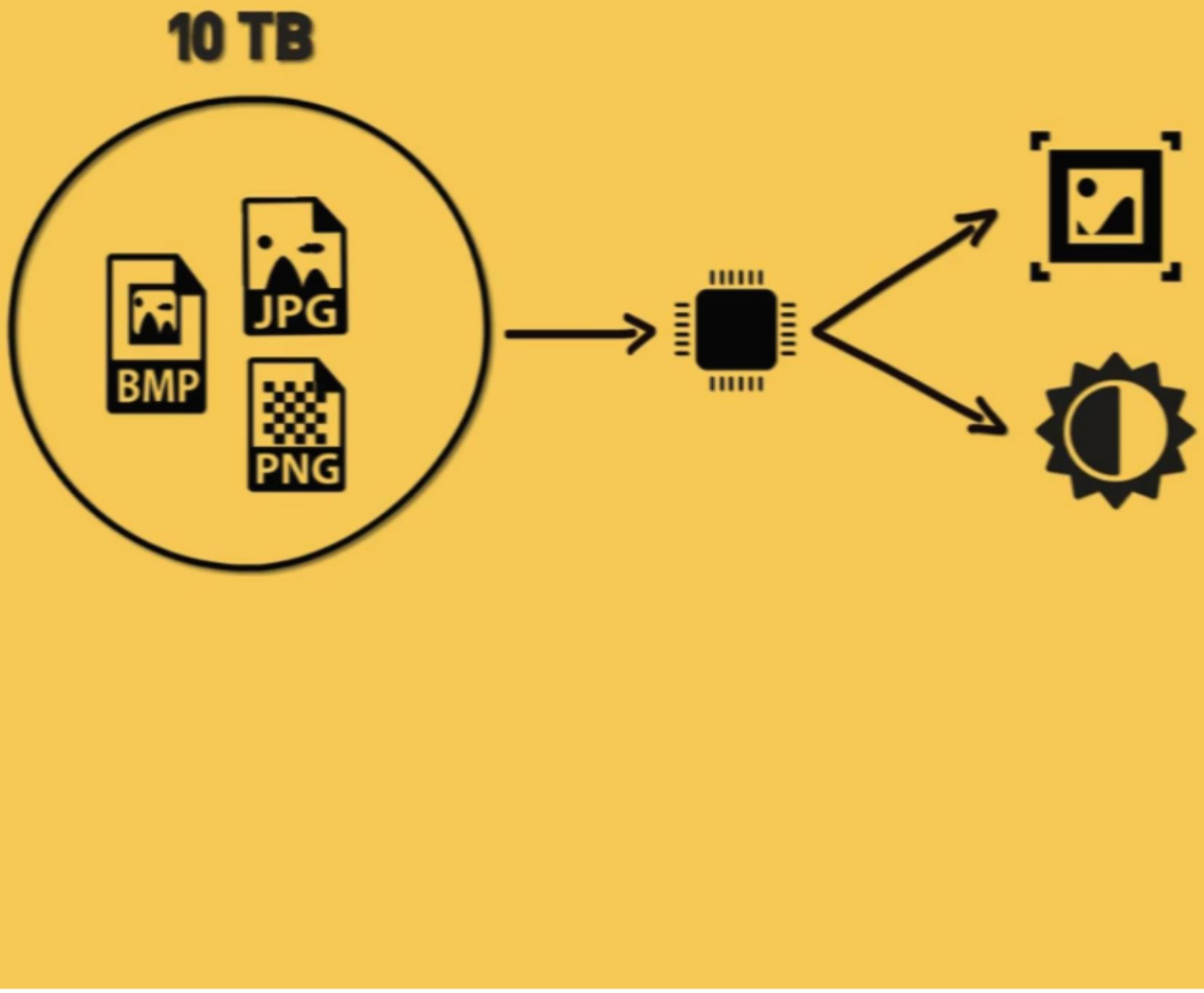
- In ordered to be classified as big data
- There is lot of misconception while referring the term bigdata, we usually used to the term bigdata to refer to the data that is either in GB/TB/PB/EB > in SIZE.
- This is not the term that defines the bigdata completely
- Even a small amount of data is referred to as big data depending on the context of its been used.
- Example
  - For instance we r trying to attach a document that is about 100 mb in size to email, You would not be able to do so, as the email system should not support an attachment of that size , therefor the 100MB of that attachment with respect to email can be referred to as BIGDATA.

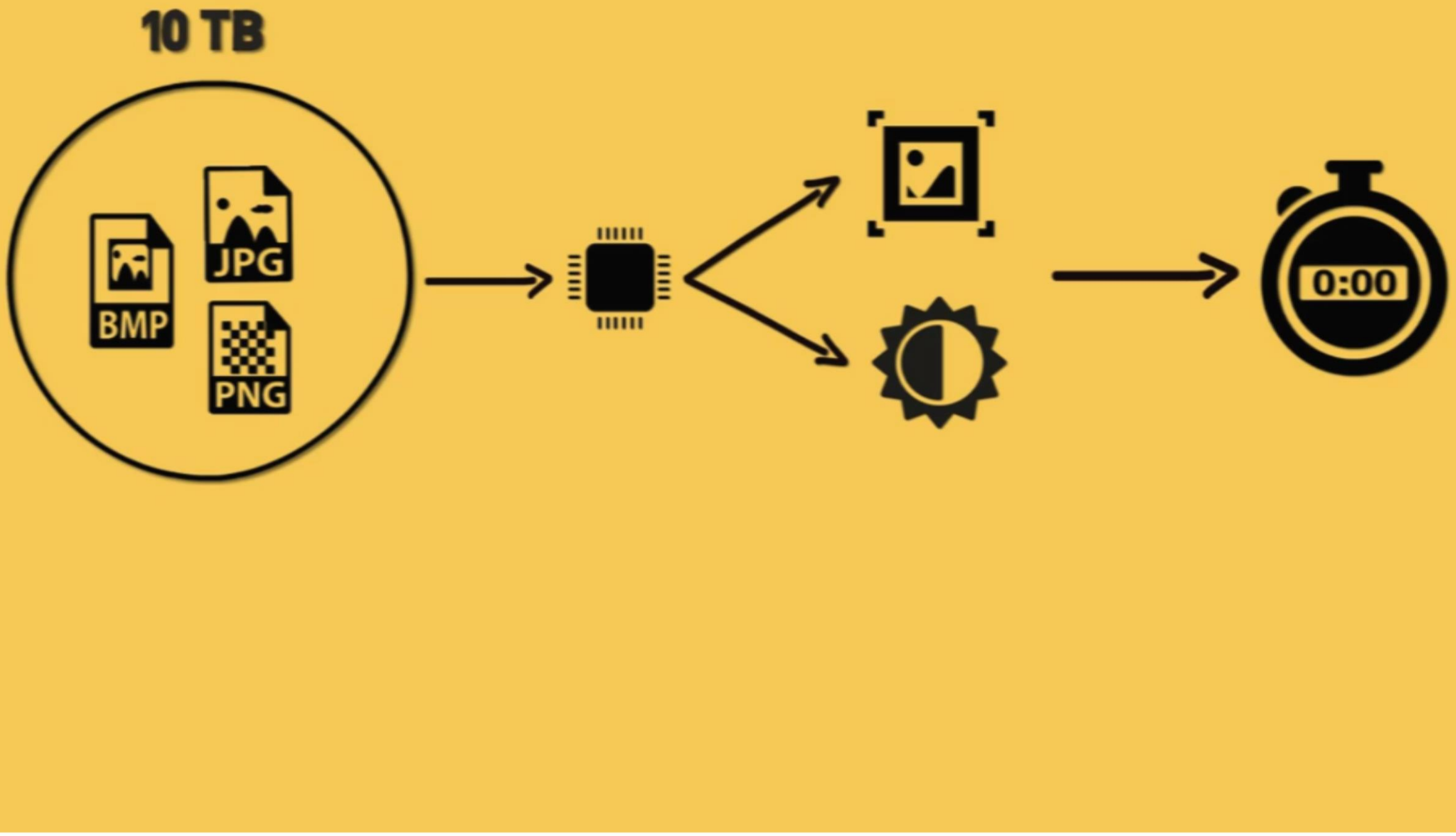

100 MB
BIG DATA

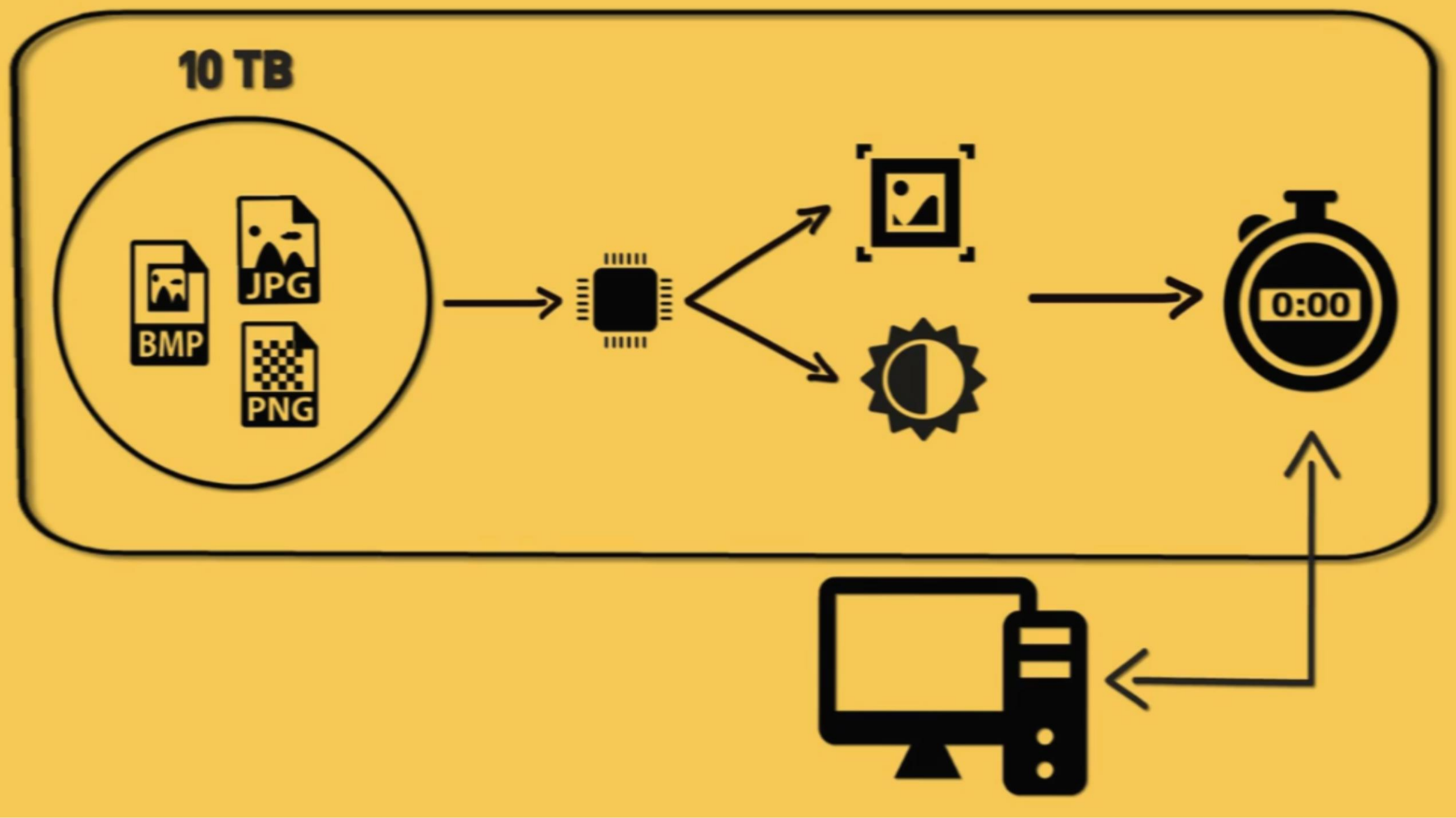
EXAMPLE 1

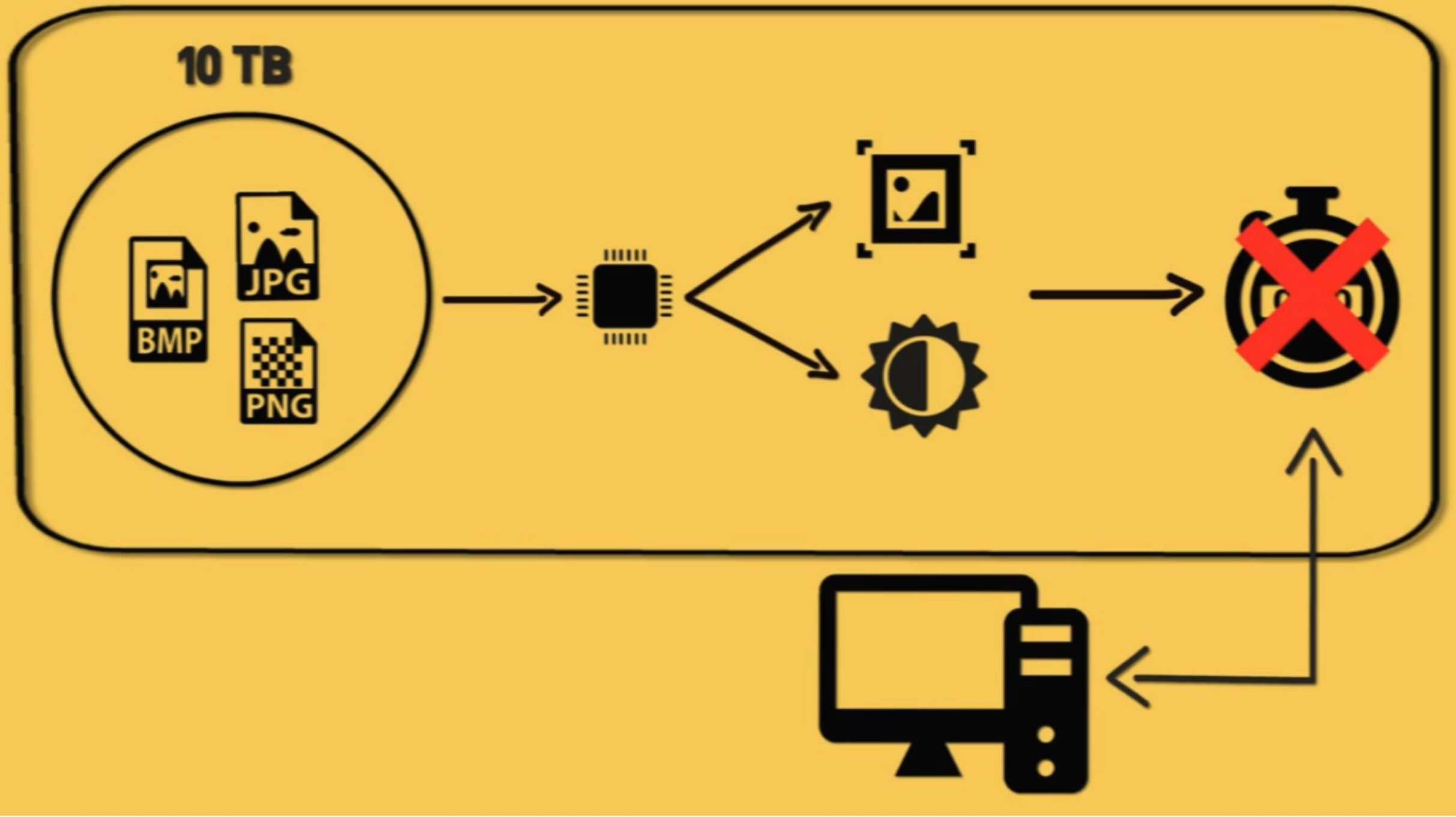
KARTHICK SELVAM

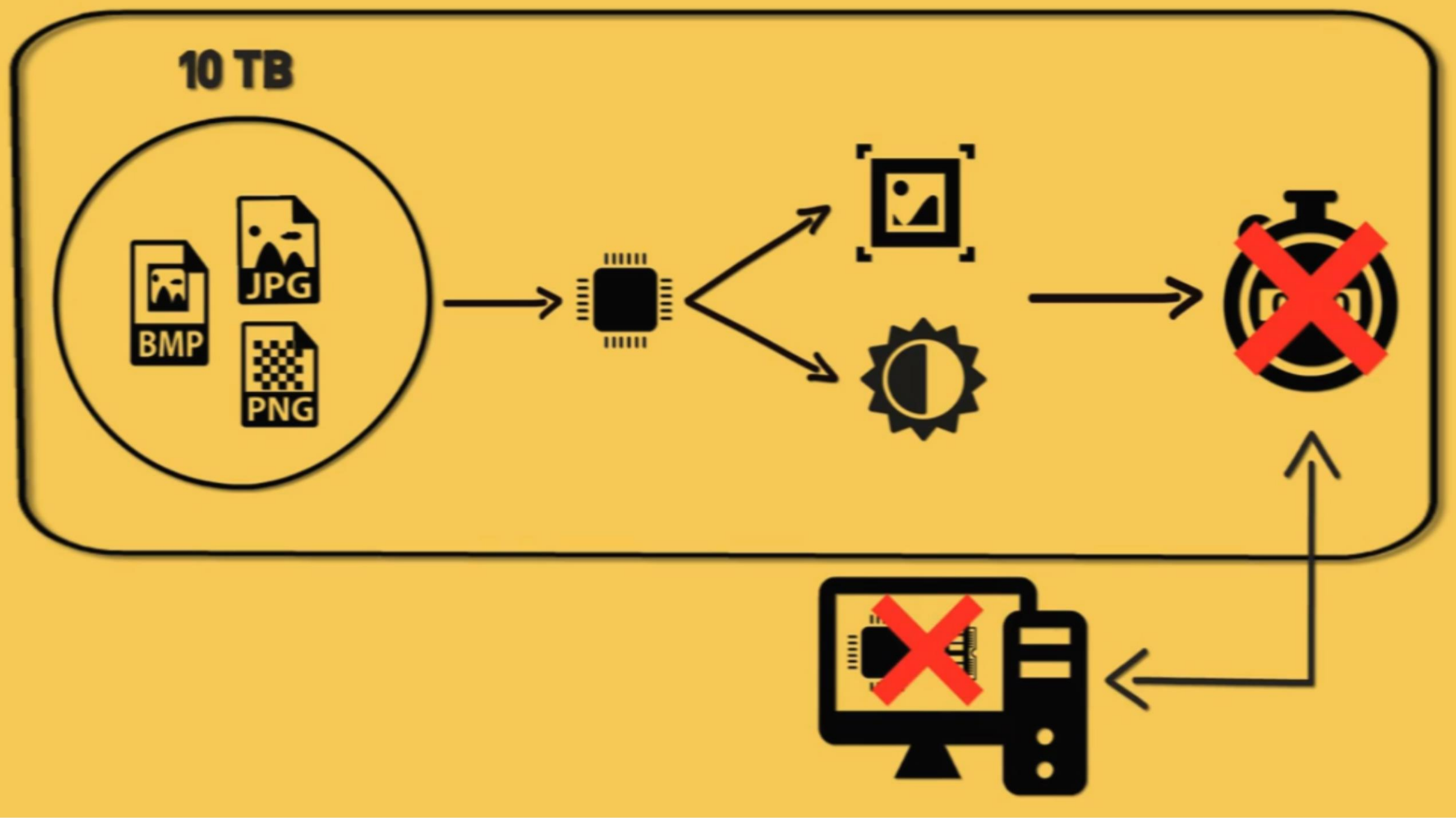
10 TB

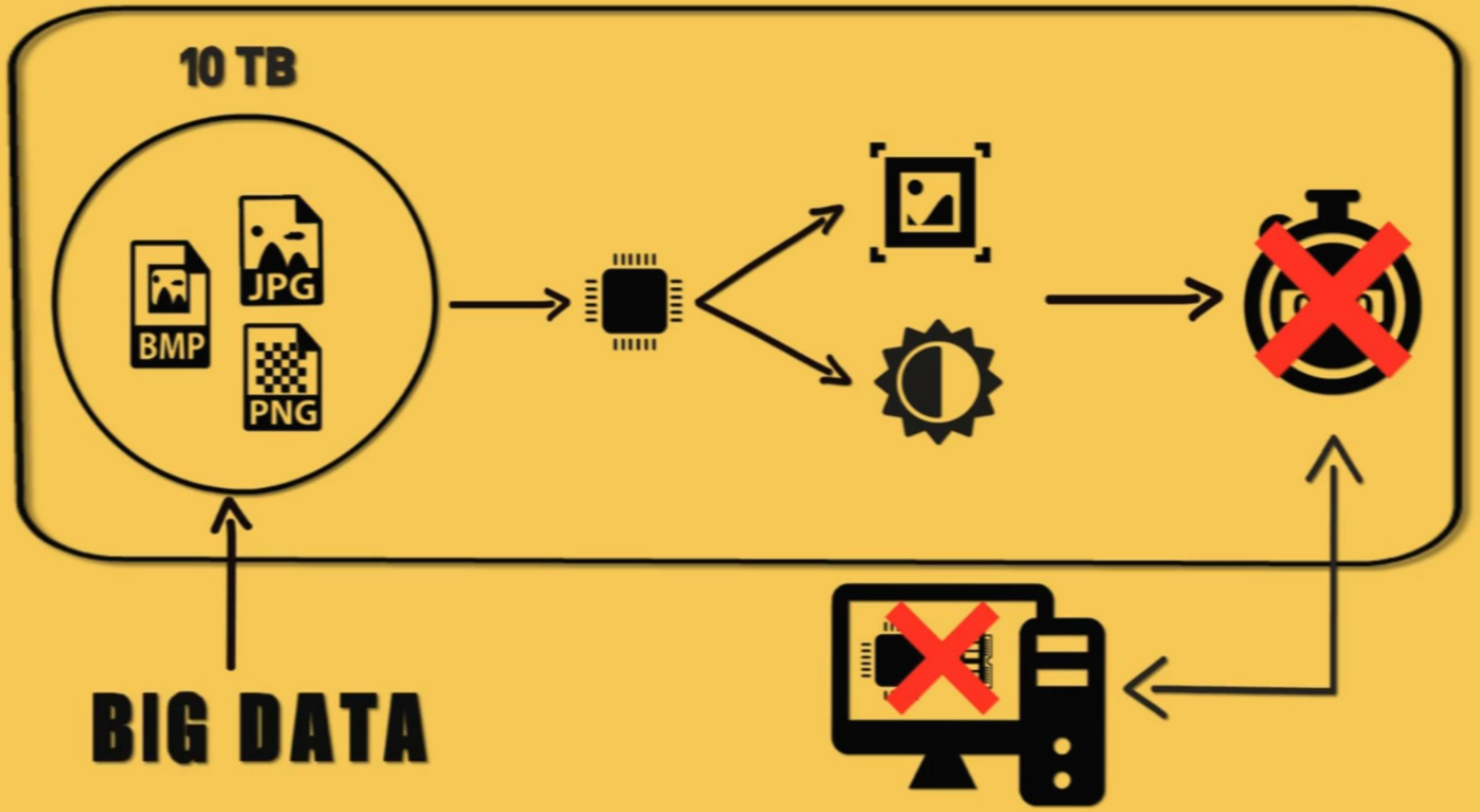

10 TB
BMP
JPG
PNG


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

EXAMPLE 2

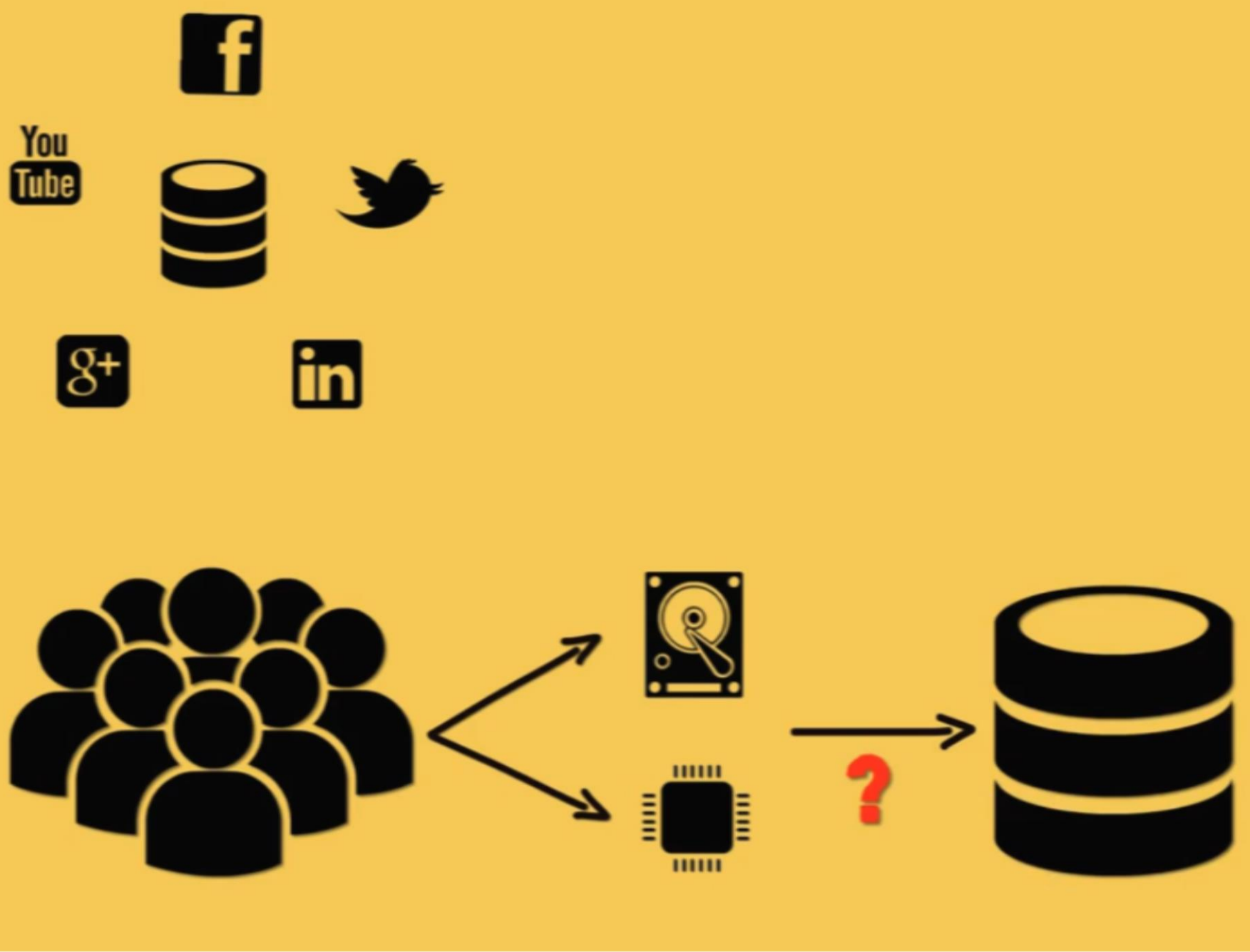
KARTHICK SELVAM

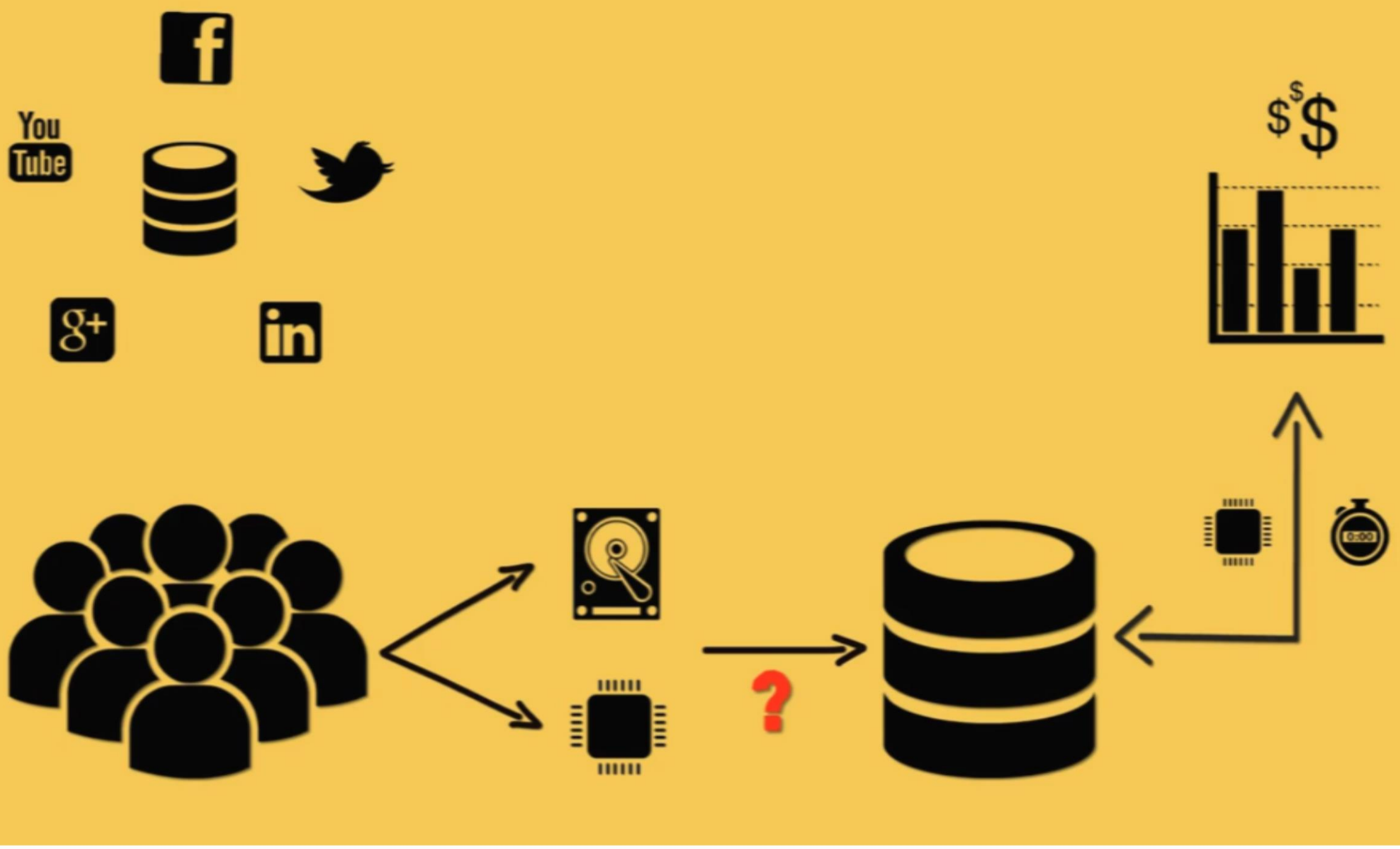

You
Tube

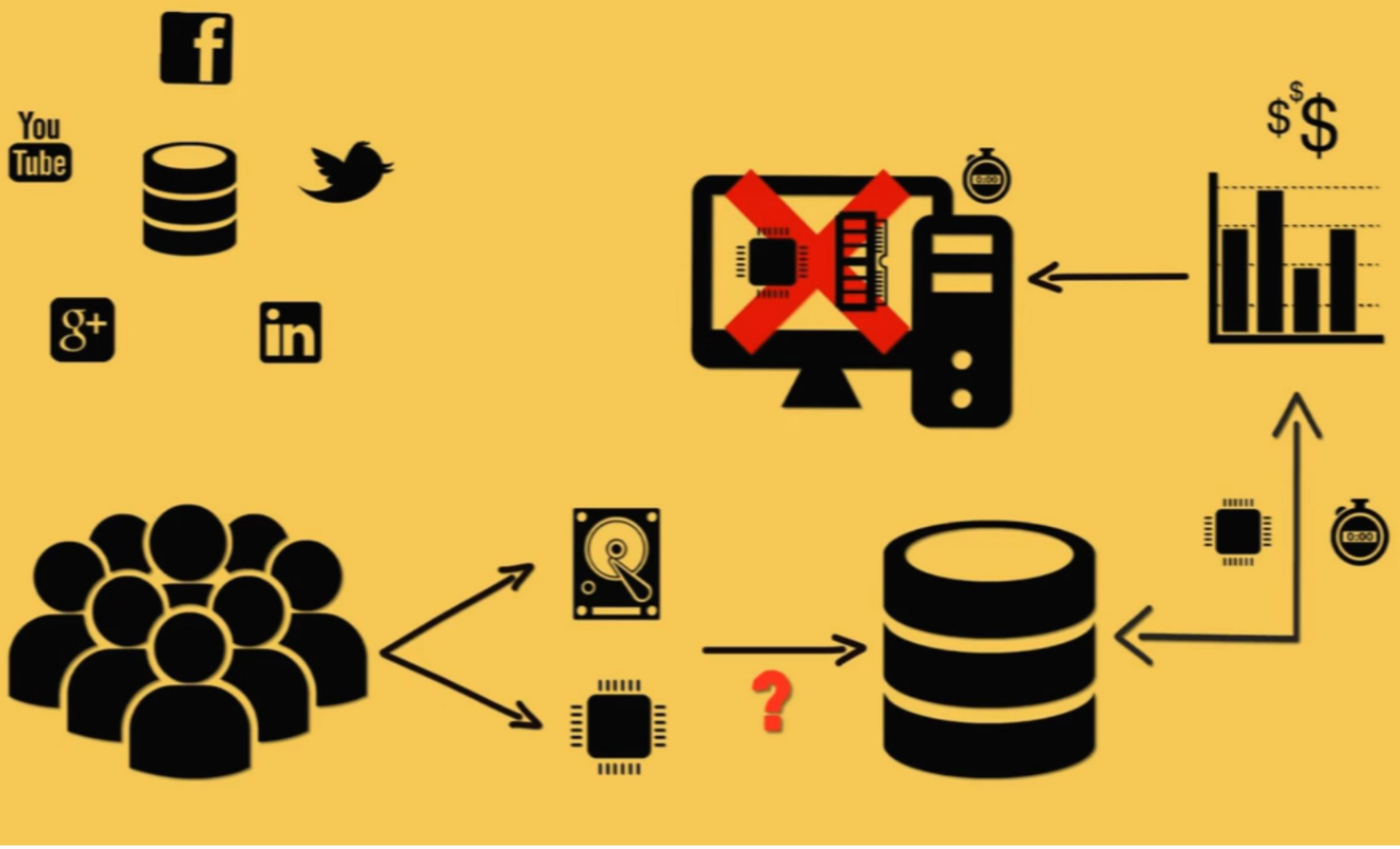
DATA ANALYSIS ON BIGDATA

- 701,389 logins on Facebook
- 150 million emails sent
- 527,760 photos shared on Snapchat
- 347,222 tweets on Twitter
- 28,194 new posts on Instagram
- 1.04 million vine loops
- 2.4 million search queries on Google
- 2.78 million video views on YouTube
- 20.8 million messages on WhatsApp


You
Tube


You
Tube
?


You
Tube
$
$
$
?


You
Tube
$
$
$
?

EXAMPLE 3

KARTHICK SELVAM


BIG DATA
?

CLASSIFICATION OF BIGDATA

KARTHICK SELVAM

STRUCTURED DATA

SEMI-STRUCTURED DATA

UN-STRUCTURED DATA

CLASSIFICATION OF BIGDATA

- Structured data
  - Structured data is a data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concern all data which can be stored in database SQL in table with rows and columns. They have relational key and can easily mapped into pre-designed fields. Today, those data are most processed in development and simplest way to manage information. Example: Relational data.
- semi-structured data
  - Semi-structured data is information that does not reside in a rational database but that have some organizational properties that make it easier to analyze. With some process, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. Example: XML data.
- Unstructured data
  - Unstructured data is a data that is which is not organized in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs

CHARACTERISTICS OF BIGDATA

KARTHICK SELVAM

CHARACTERISTICS OF BIGDATA

- Volume referred to as amount of data that is getting generated.
- Velocity referred to as speed at which that data is getting generated.
- Variety referred to as different type of data that is getting generated.
- Veracity refers to an uncertainty of data available, which makes it harder for the companies to react quickly and make appropriate solutions.

VOLUME

VELOCITY

VARIETY

CHALLENGES ASSOCIATED WITH BIGDATA

KARTHICK SELVAM

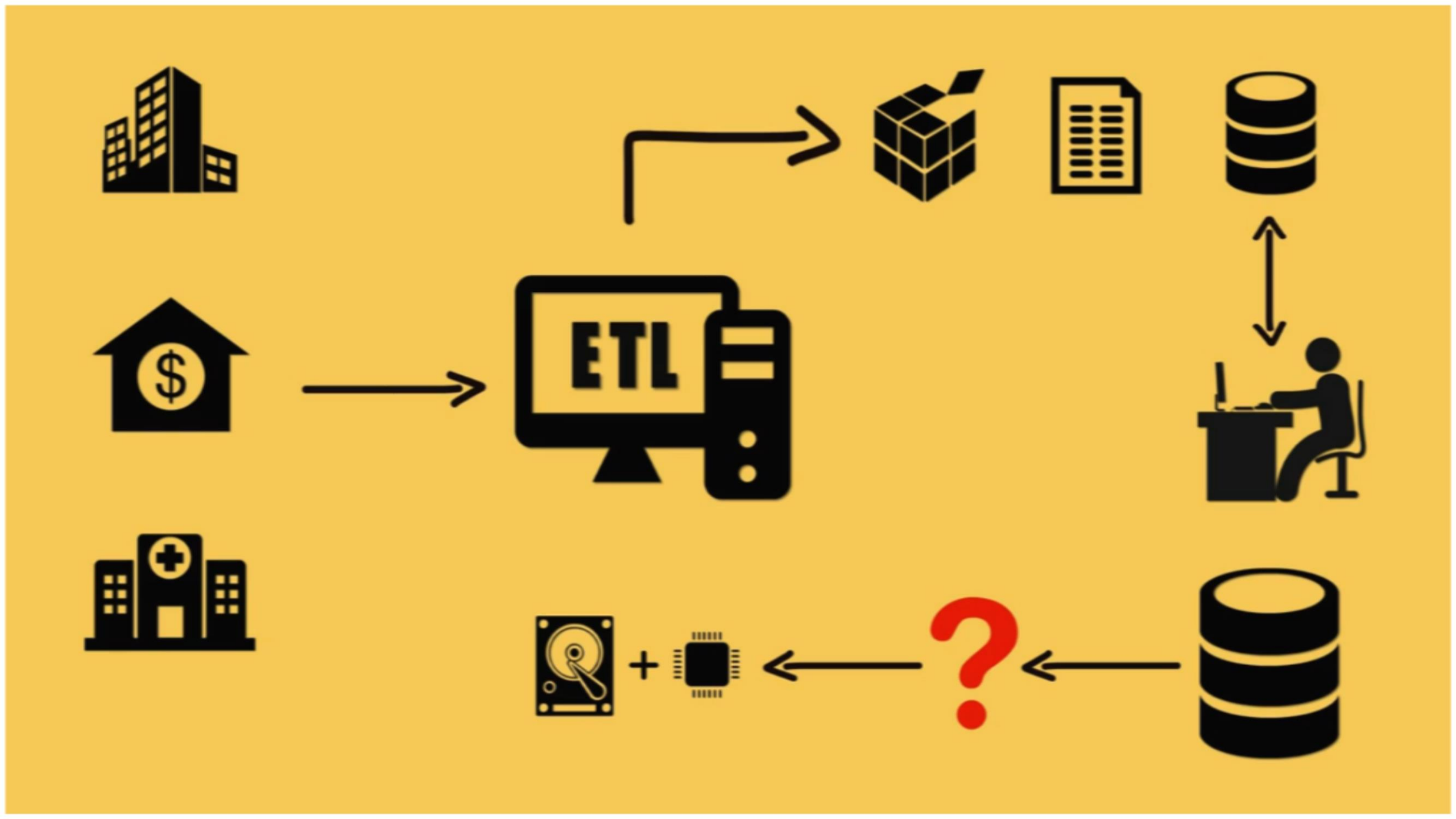
CHALLENGES ASSOCIATED WITH BIGDATA

- 2 main challenges associated with bigdata
  - How do we store and manage such a huge volume of data efficiently.
  - How do we process and extract valuable information's from such a huge volume of data with in a given time frame.


0:00

TRADITIONAL APPROACH OF STORING AND PROCESSING BIGDATA.

KARTHICK SELVAM

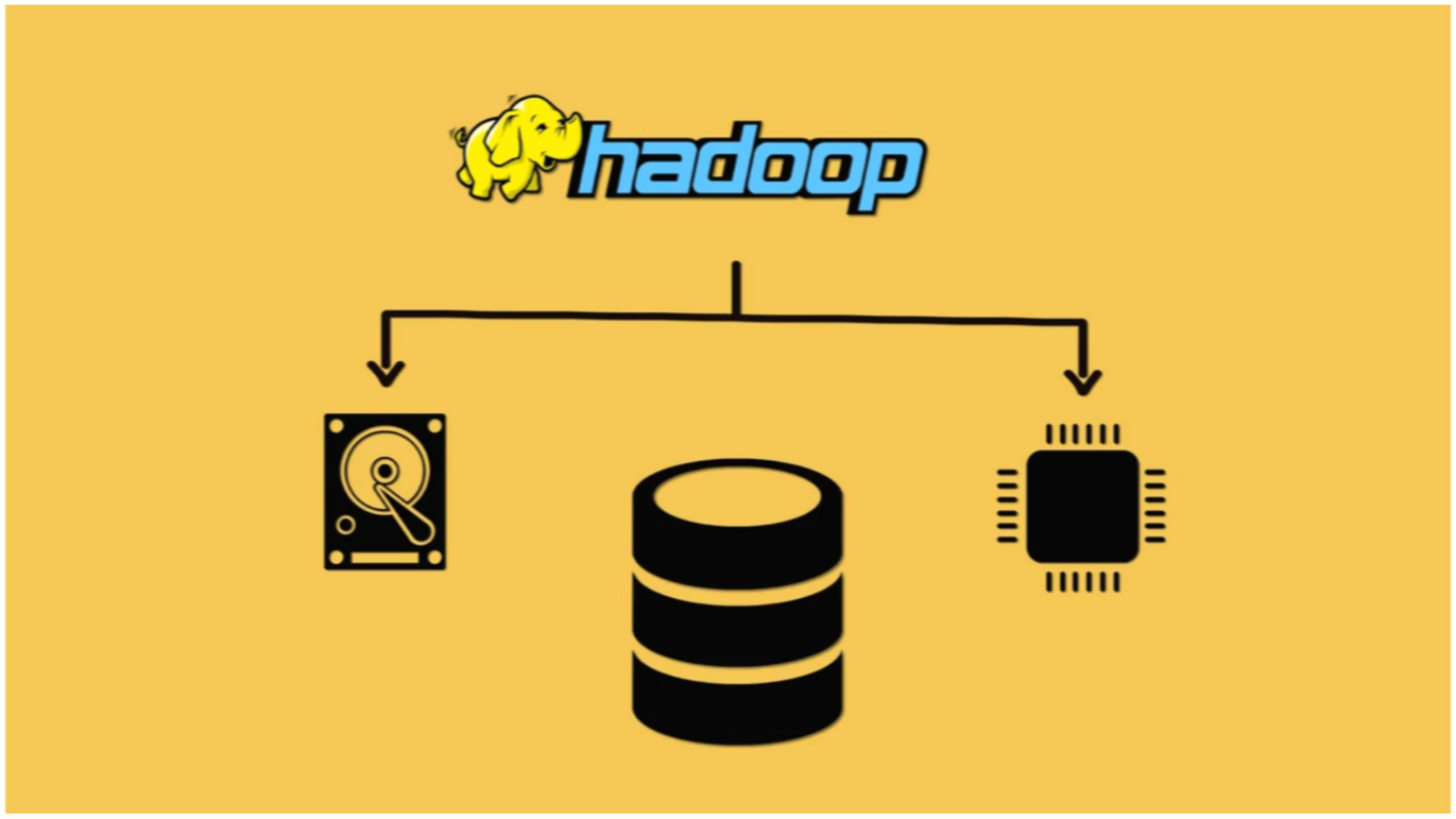

ETL

DRAWBACKS OF USING TRADITIONAL APPROACH


hadoop


TM
open source


The Apache
Software Foundation
http://www.apache.org/


hadoop

HDFS

MapReduce

HADOOP CLUSTER

NameNode

JobTracker

STORAGE NODE

COMPUTE NODE

DataNode

TaskTracker