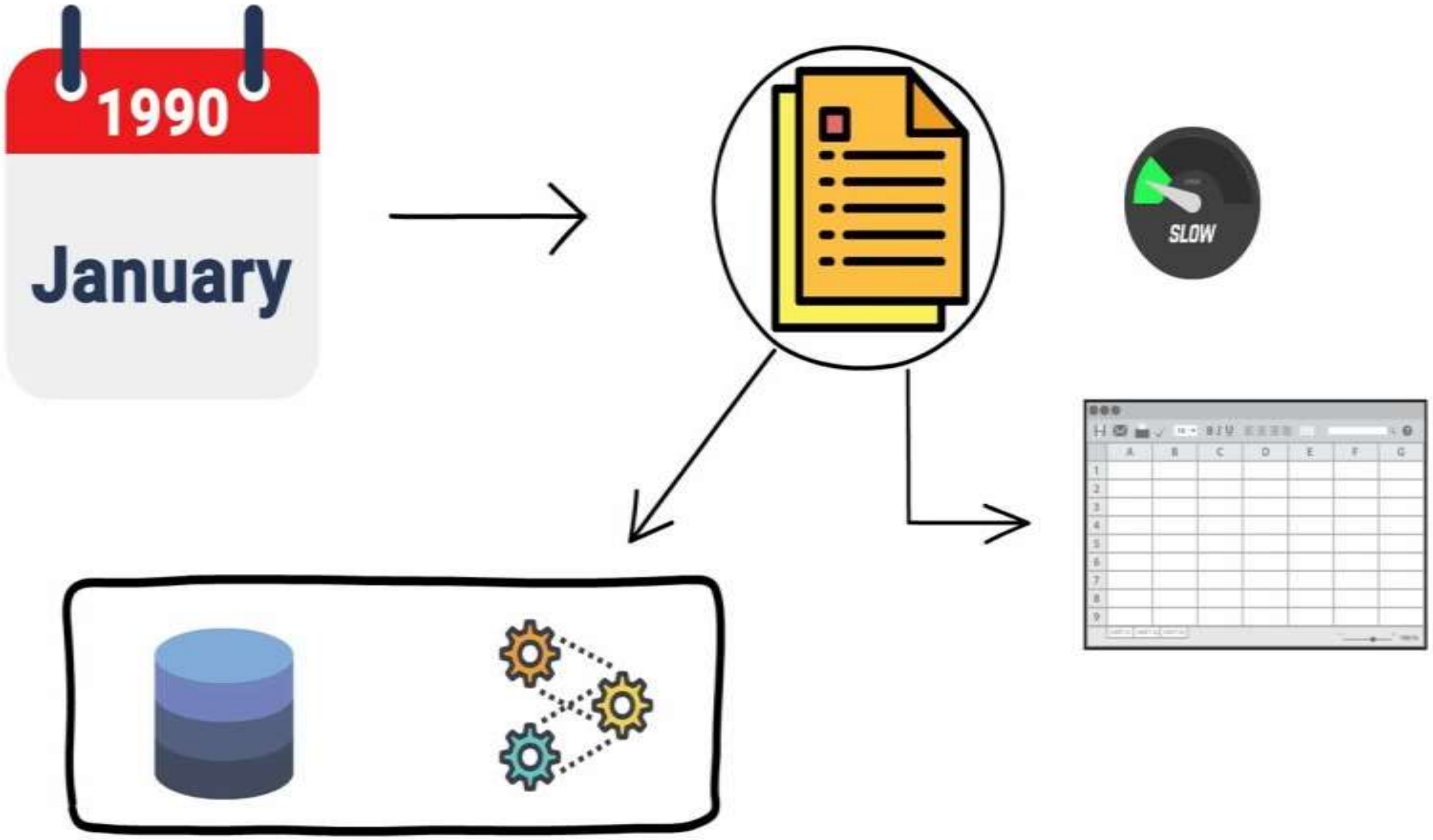
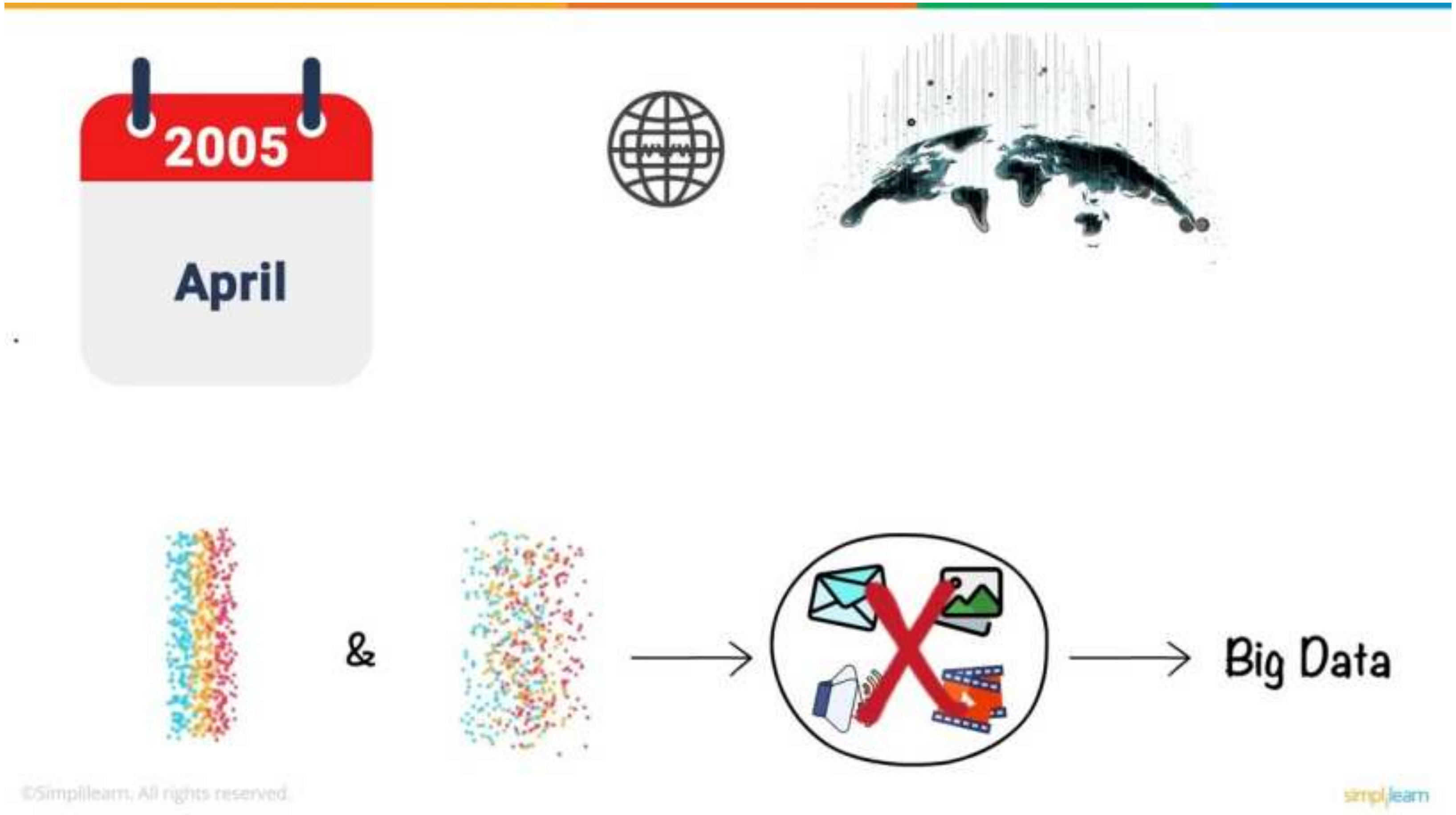
Hadoop

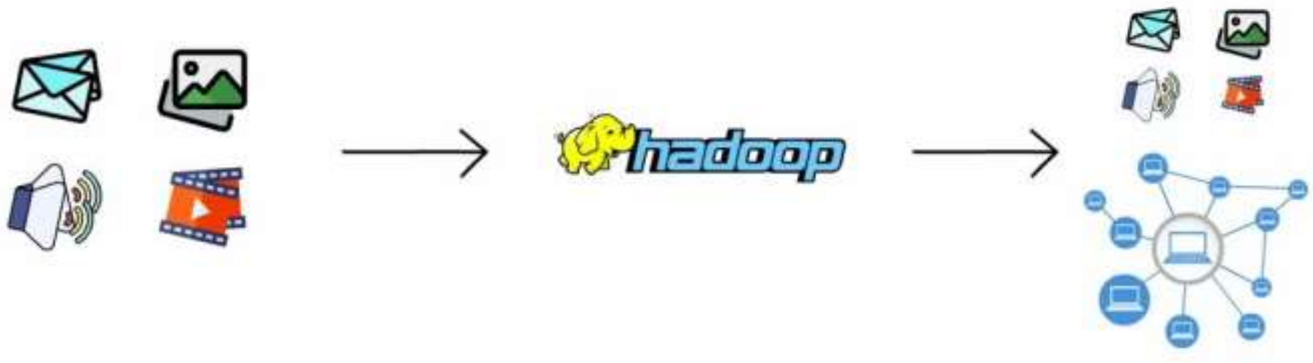
The rapid evolution of the internet has led to an explosion of diverse data types being generated at unprecedented speeds, surpassing the capabilities of traditional storage and processing methods.

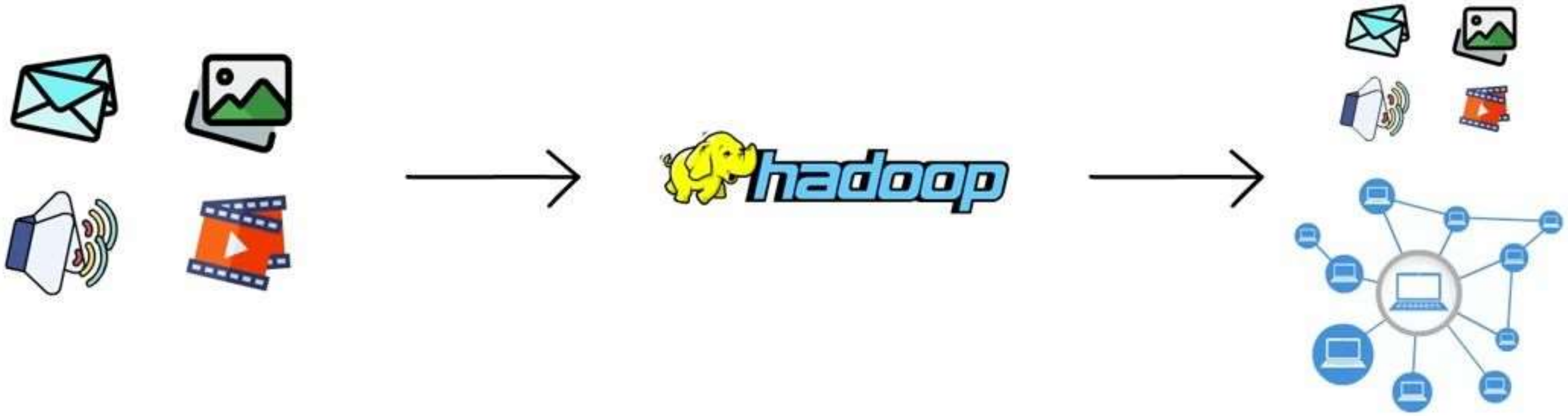


The emergence of semi-structured and unstructured data, such as emails, images, audio, and video, necessitated the use of multiple storage units and processors to effectively manage big data.



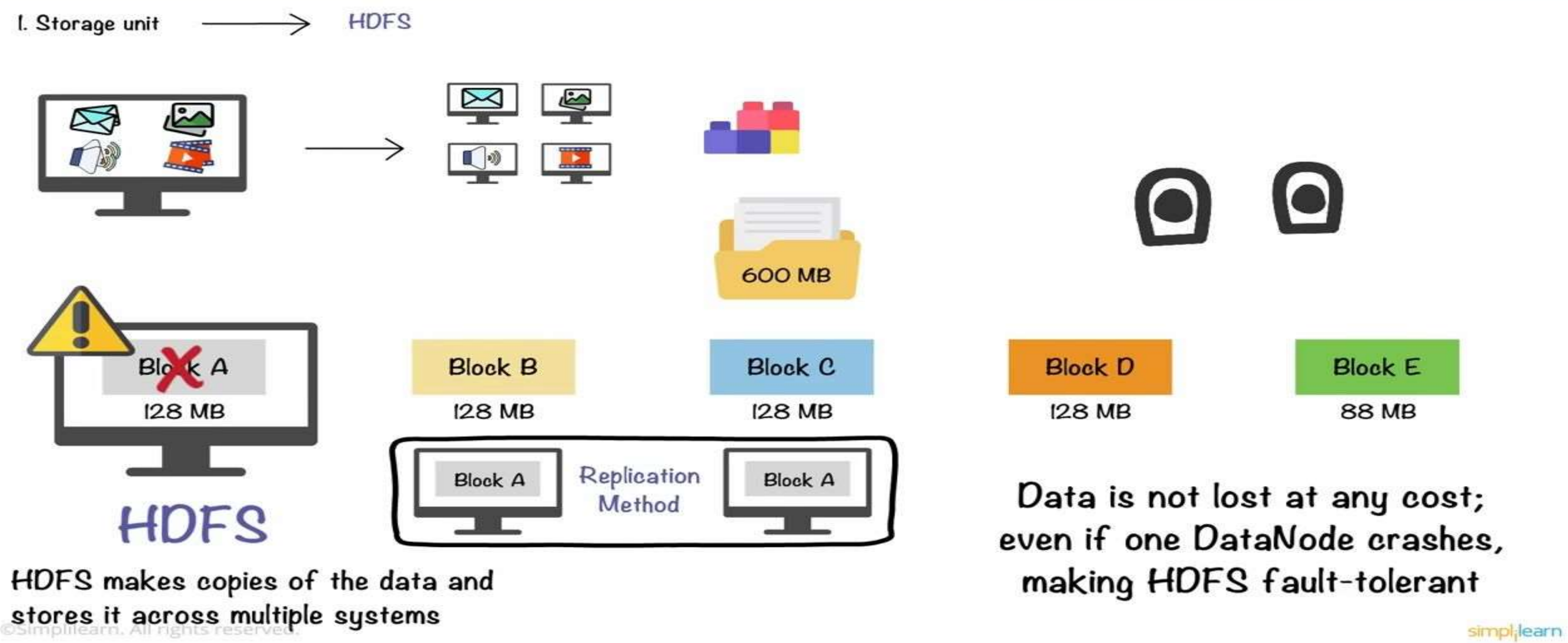
The solution to insufficient storage and processing power was the implementation of multiple storage units and processors, leading to the development of Hadoop, which efficiently stores and processes large amounts of data using a cluster of commodity hardware.





Hadoop consisted of three components that were specifically designed to work on big data

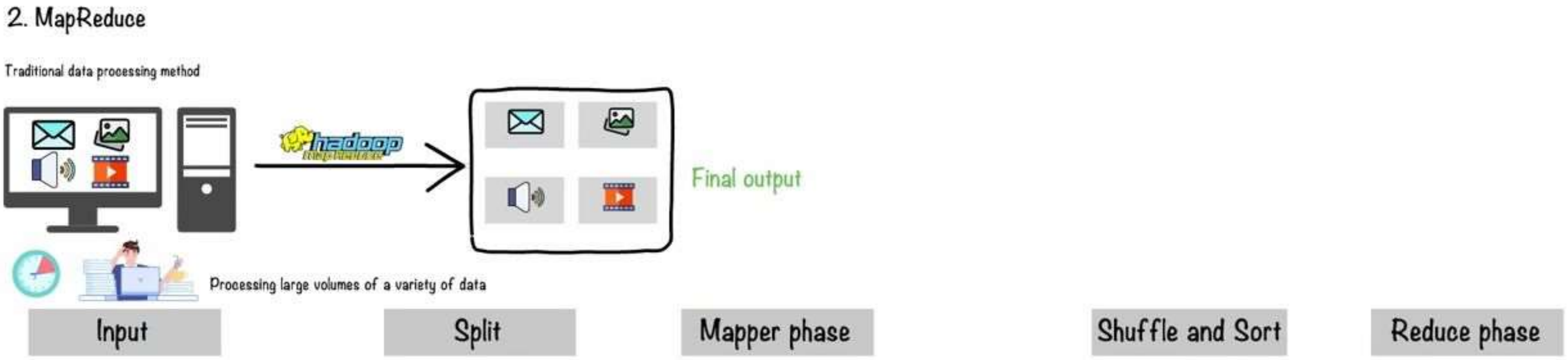
HDFS: Hadoop Distributed File System



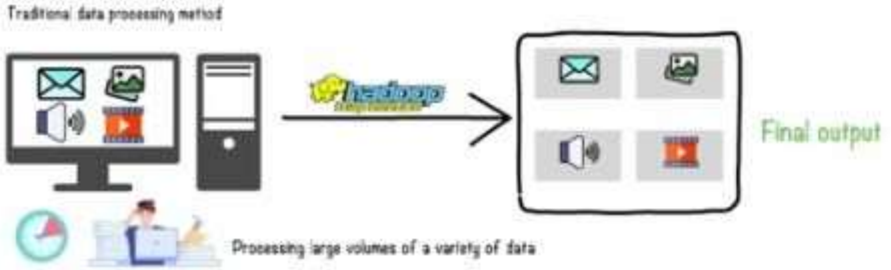
HDFS makes copies of the data and stores it across multiple systems

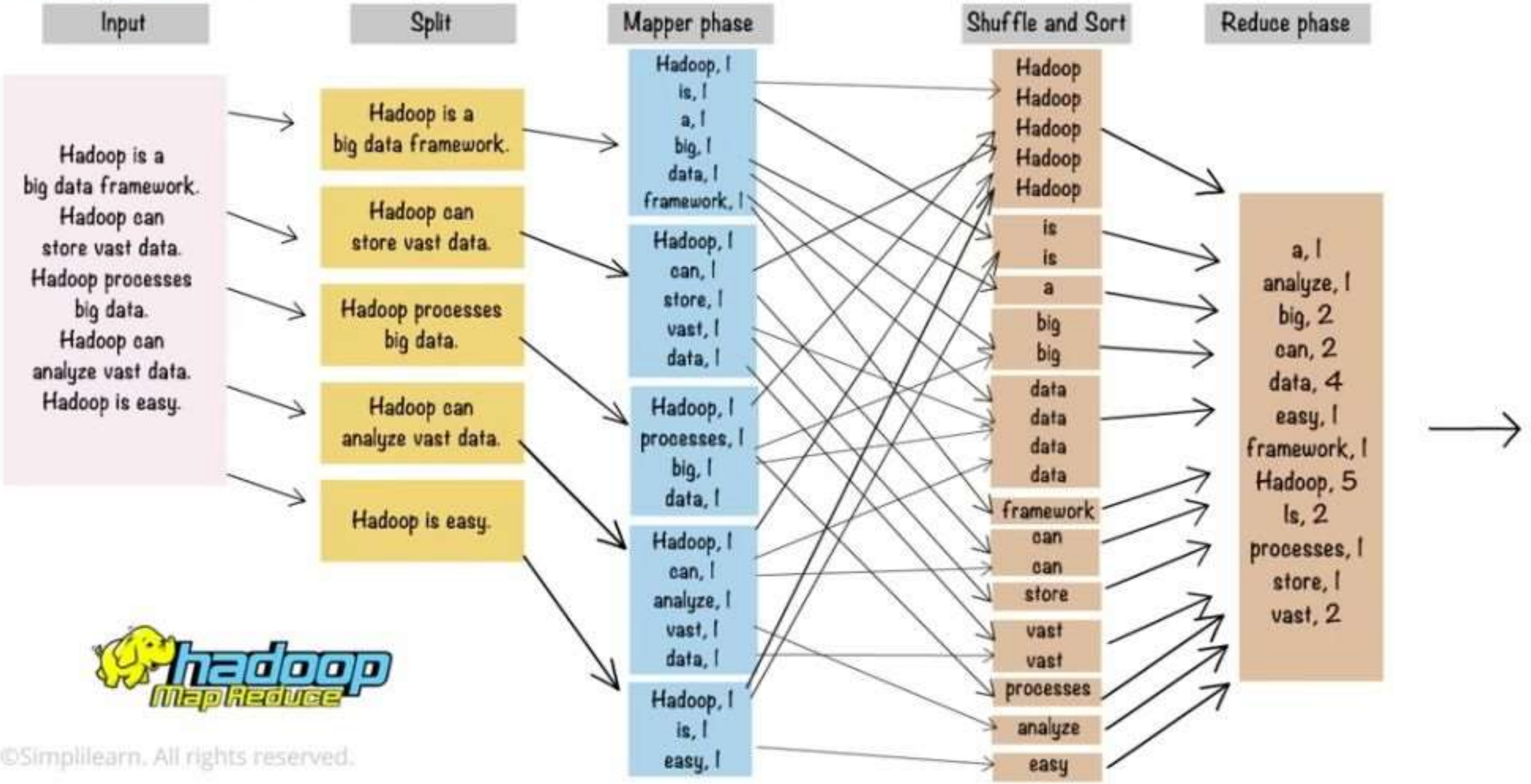
Data is not lost at any cost; even if one DataNode crashes, making HDFS fault-tolerant

MapReduce improves data processing efficiency by splitting large datasets into parts for parallel processing on multiple nodes, followed by aggregation of the results.



2. MapReduce


Traditional data processing method
Final output
Processing large volumes of a variety of data


Input

Hadoop is a
big data framework.
Hadoop can
store vast data.
Hadoop processes
big data.
Hadoop can
analyze vast data.
Hadoop is easy.

Split

Hadoop is a
big data framework.

Hadoop can
store vast data.

Hadoop processes
big data.

Hadoop can
analyze vast data.

Hadoop is easy.

Mapper phase

Hadoop, 1
is, 1
a, 1
big, 1
data, 1
framework, 1

Hadoop, 1
can, 1
store, 1
vast, 1
data, 1

Hadoop, 1
processes, 1
big, 1
data, 1

Hadoop, 1
can, 1
analyze, 1
vast, 1
data, 1

Hadoop, 1
is, 1
easy, 1

Shuffle and Sort

Hadoop
Hadoop
Hadoop
Hadoop
Hadoop

is
is

a

big
big

data
data
data
data

framework

can
can

store

vast
vast

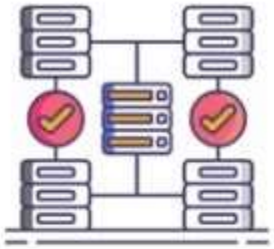
processes

analyze

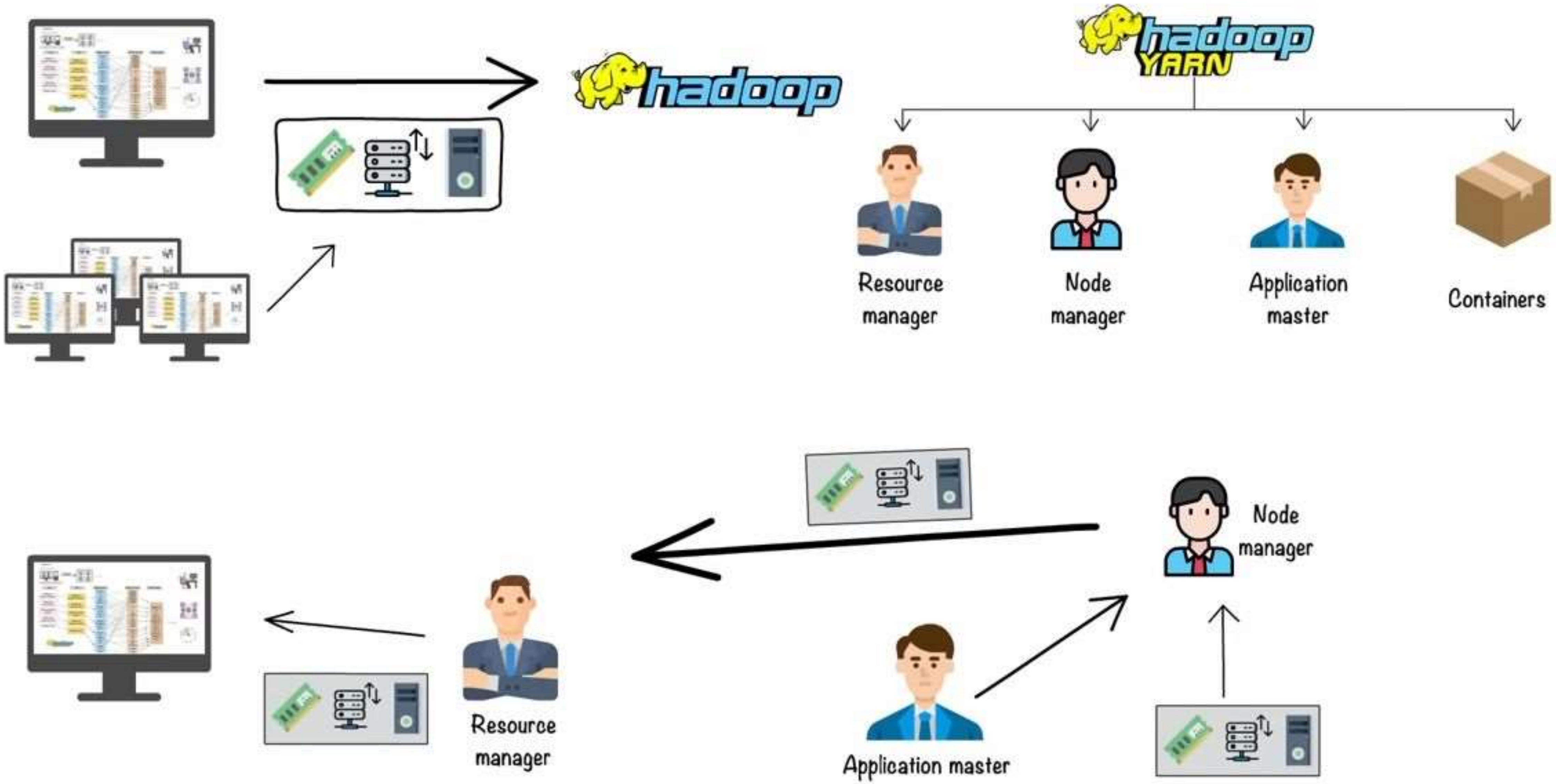
easy

Reduce phase

a, 1
analyze, 1
big, 2
can, 2
data, 4
easy, 1
framework, 1
Hadoop, 5
Is, 2
processes, 1
store, 1
vast, 2



3. YARN



YARN processes job requests and manages cluster resources

What is the advantage of the 3x replication schema in HDFS?

a) Supports parallel processing

b) Faster data analysis

c) Ensures fault tolerance

d) Manages cluster resources