What is Big Data?

Before Big Data - SQL
Hadoop, Hive, Pig, etc.

Big Data refers to data sets that are too large or complex to be dealt with by traditional data processing application software

5 Vs of Big Data

1. Volume: The quantity of generated and stored data
2. Variety: The type and nature of data (Structured, Semi Structured, Unstructured)
3. Velocity: The speed at which the data is generated and processed

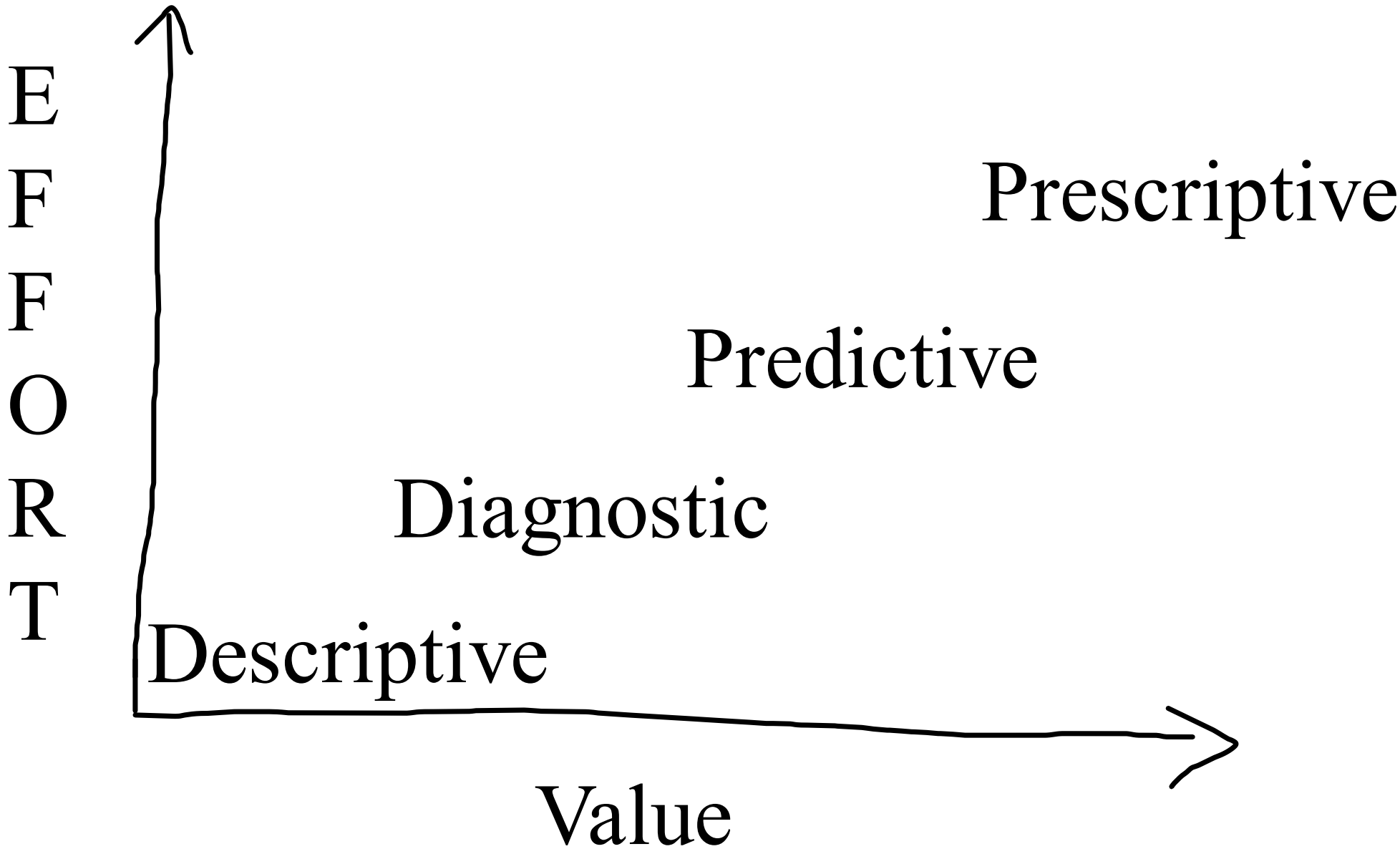
4. Veracity: The truthfulness or reliability of data
5. Value: The worth in information

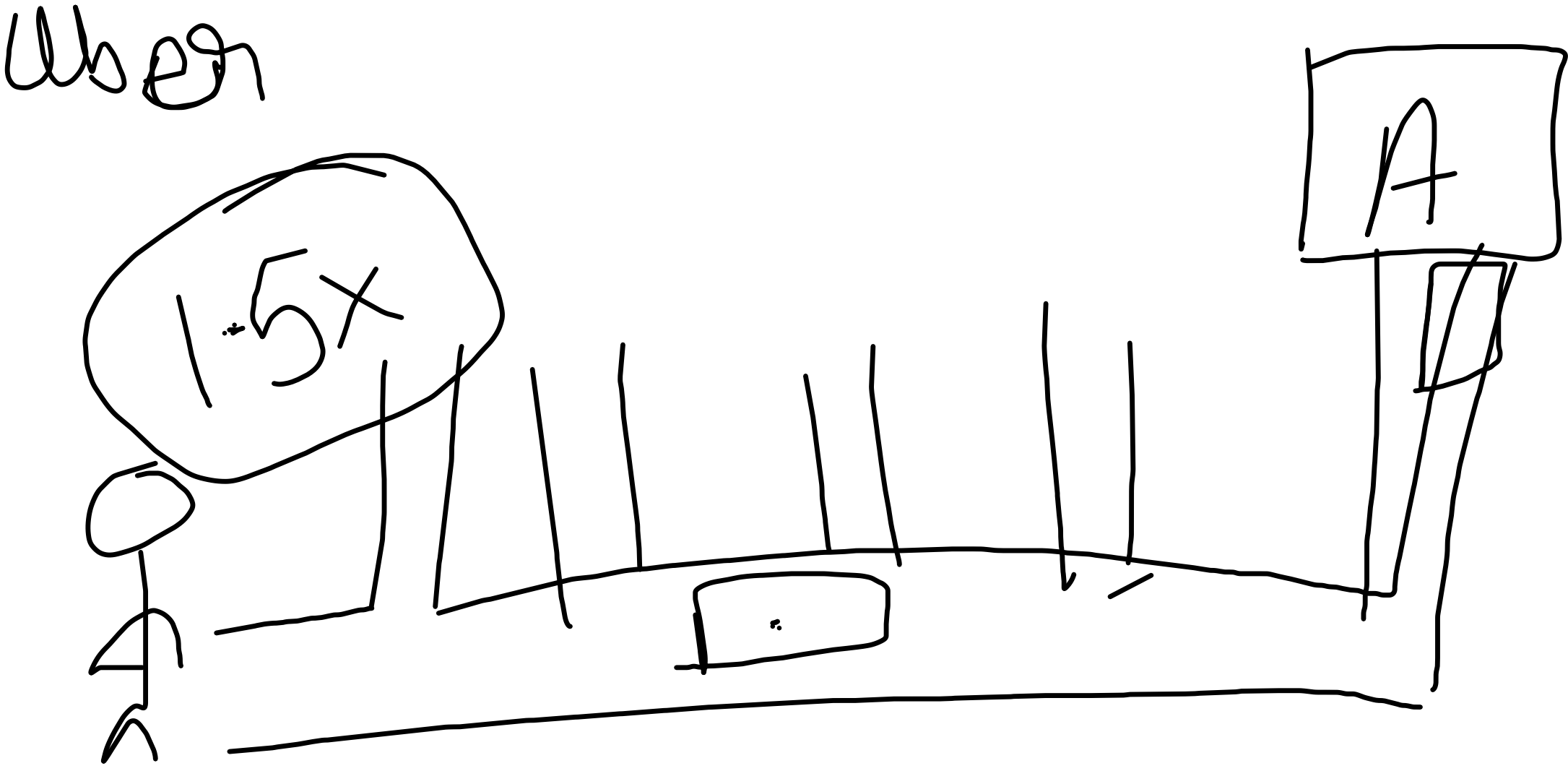
Stages of Big Data Analytics

Business Problem Definition
Data Definition
Data Acquisition and Filtering
Data Extraction
Data Munging
Data Aggregation and Representation
Data Analytics

Data Visualization
Utilization of the Analysis Result

Types of Big Data Analytics




1.5x
A

Application of Big Data Analytics

Healthcare
Finance and Banking
Space Research
AI
Research and Development
Smart Traffic Systems
Secure Air Traffic Systems

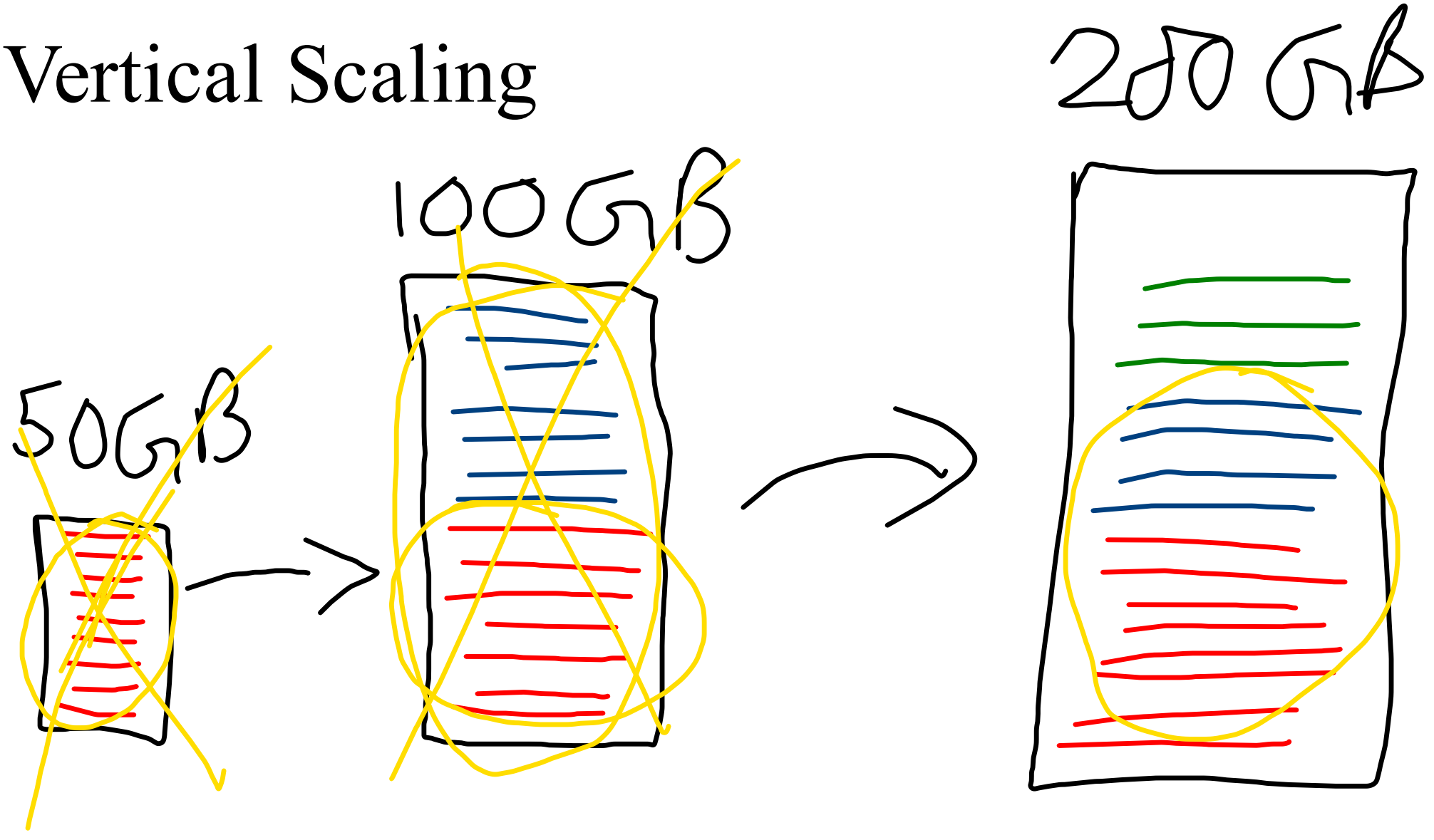
Self Driving Car
Education
Smart Electric Meters
Environment
Virtual Personal Assisstants
IoT
Recommendation

Problems of Big Data Analytics

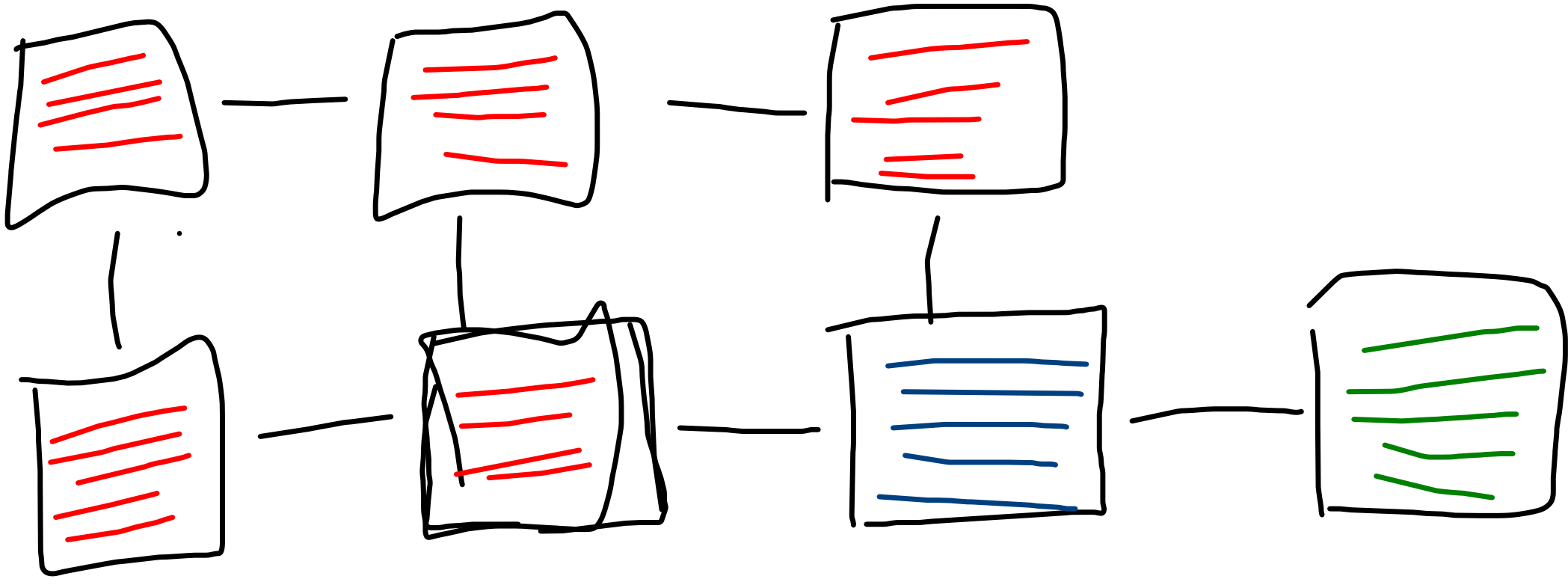
Unstructured Data

Hadoop does not enforce a schema on the data.

Vertical Scaling

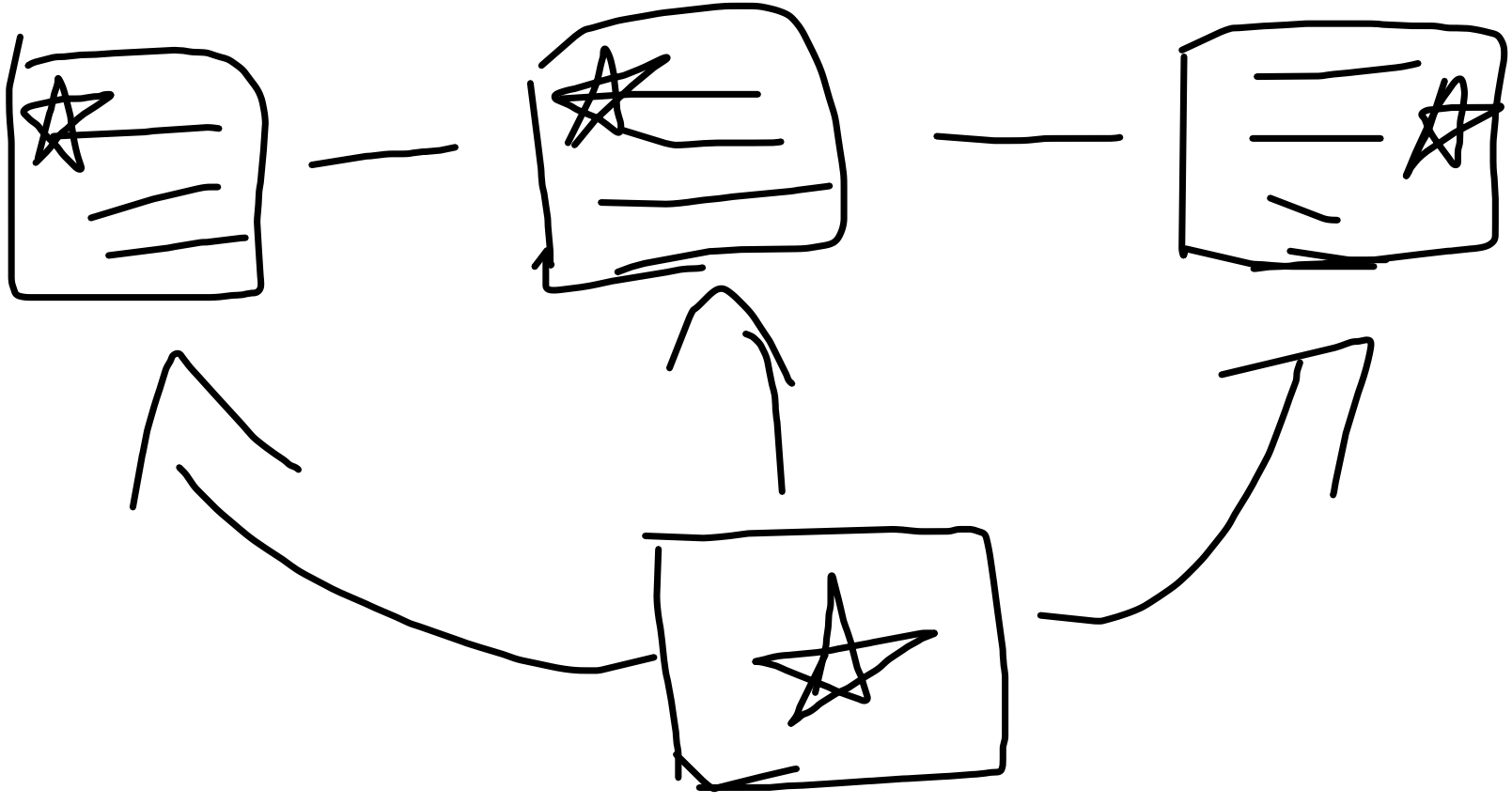


Horizontal Scaling



Processing

Parallel Processing / Distributed Processing



What is Hadoop?

Apache Hadoop is a collection of open source software utilities that facilitates using a network of computers to solve problems involving massive amounts of data and computation.

Hadoop Architecture

Hadoop Commons: contains libraries and utilities needed by other hadoop modules.

Hadoop Distributed File System (HDFS): a distributed file system that stores data on commodity machines

Mapreduce: Programming framework that enables parallel processing

Map Phase: Create key value pairs
Shuffle Phase: Data is sorted by keys
Reduce Phase: Aggregate the values based on keys

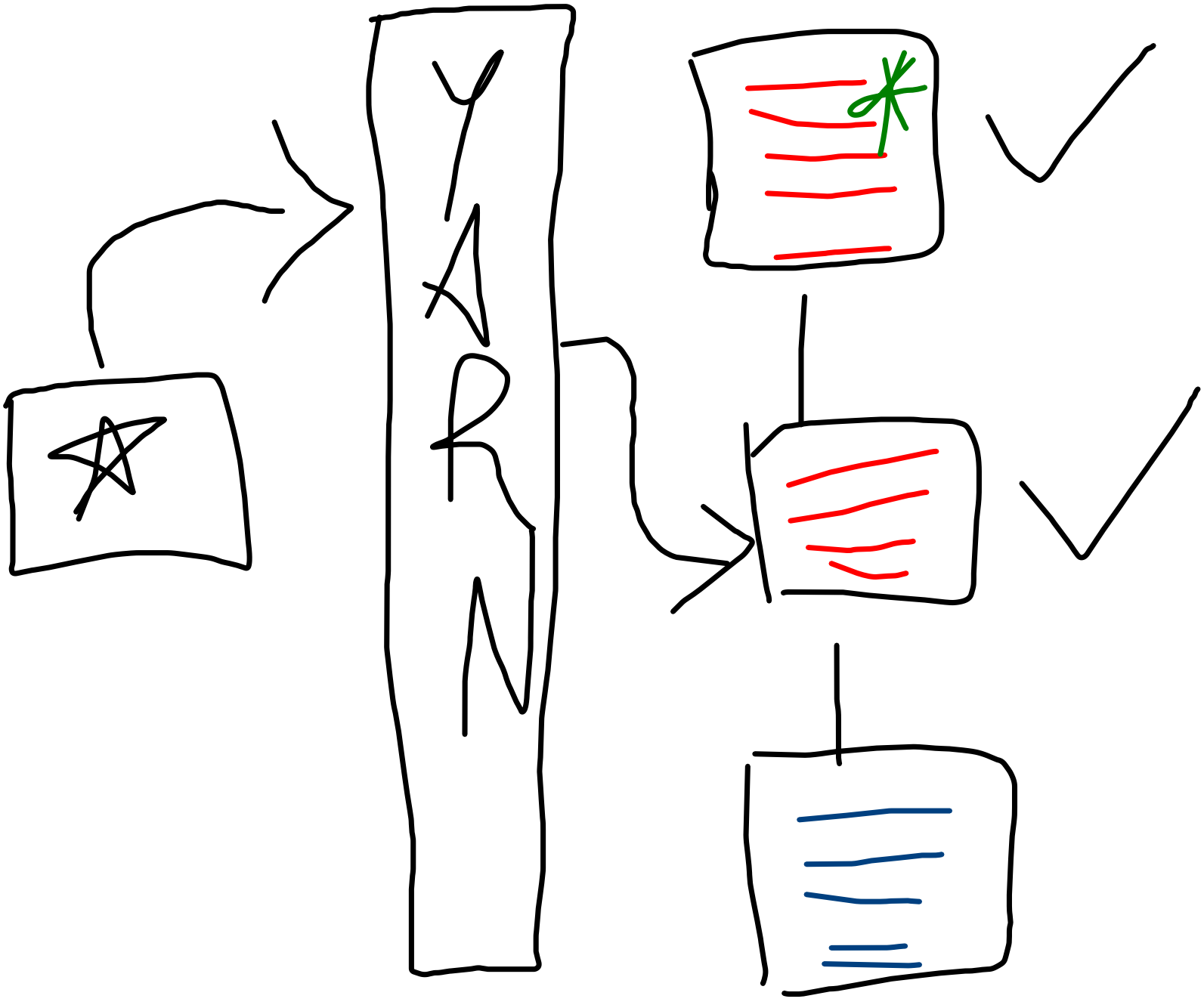
Wordcount

hello world hello how are you world

Map Phase: (hello, 1), (world, 1), (hello, 1), (how, 1), (are, 1), (you, 1), (world, 1)

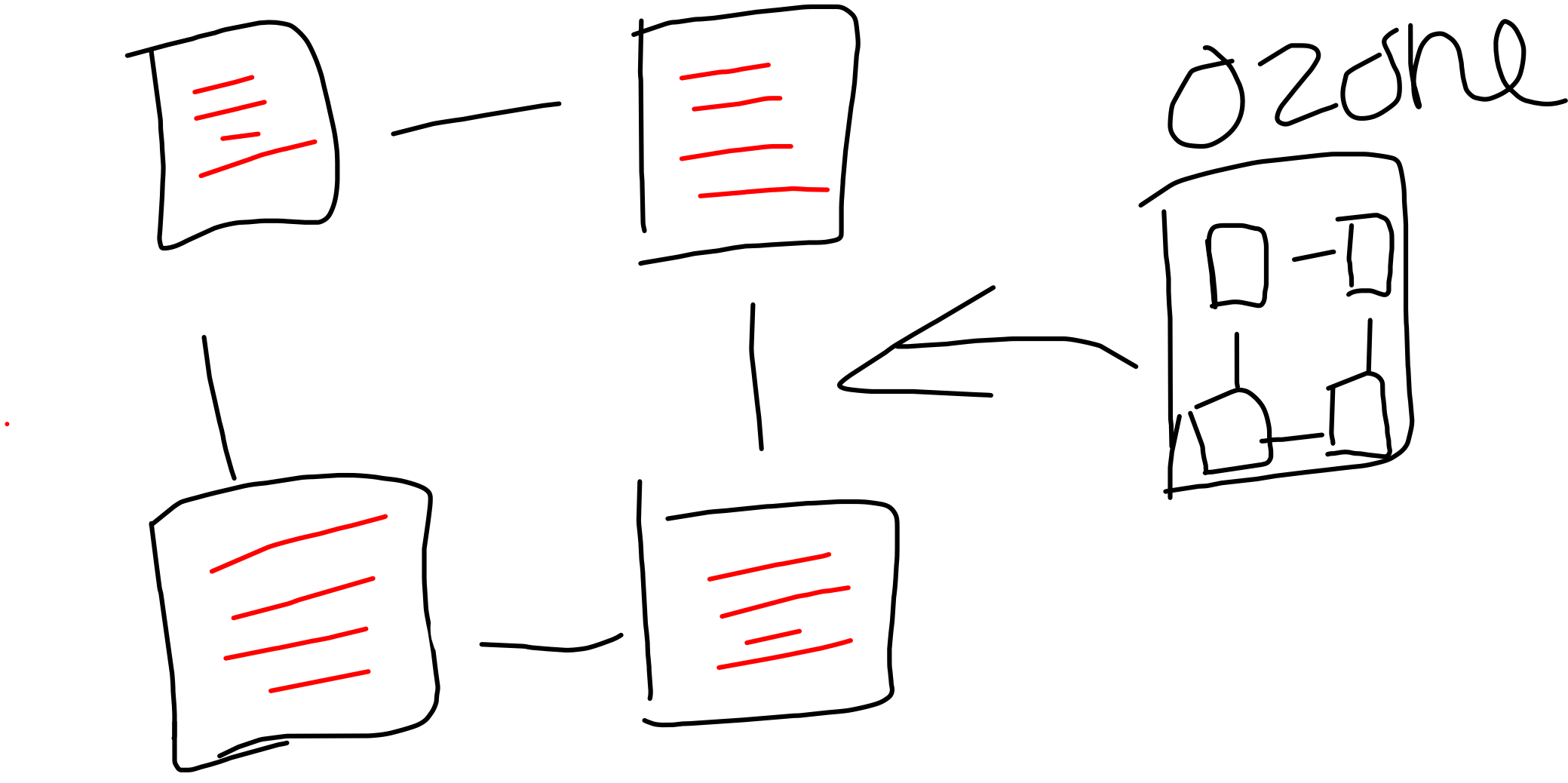
Shuffle Phase: (are, 1), (hello, 1), (hello, 1), (how, 1), (world, 1), (world, 1), (you, 1)

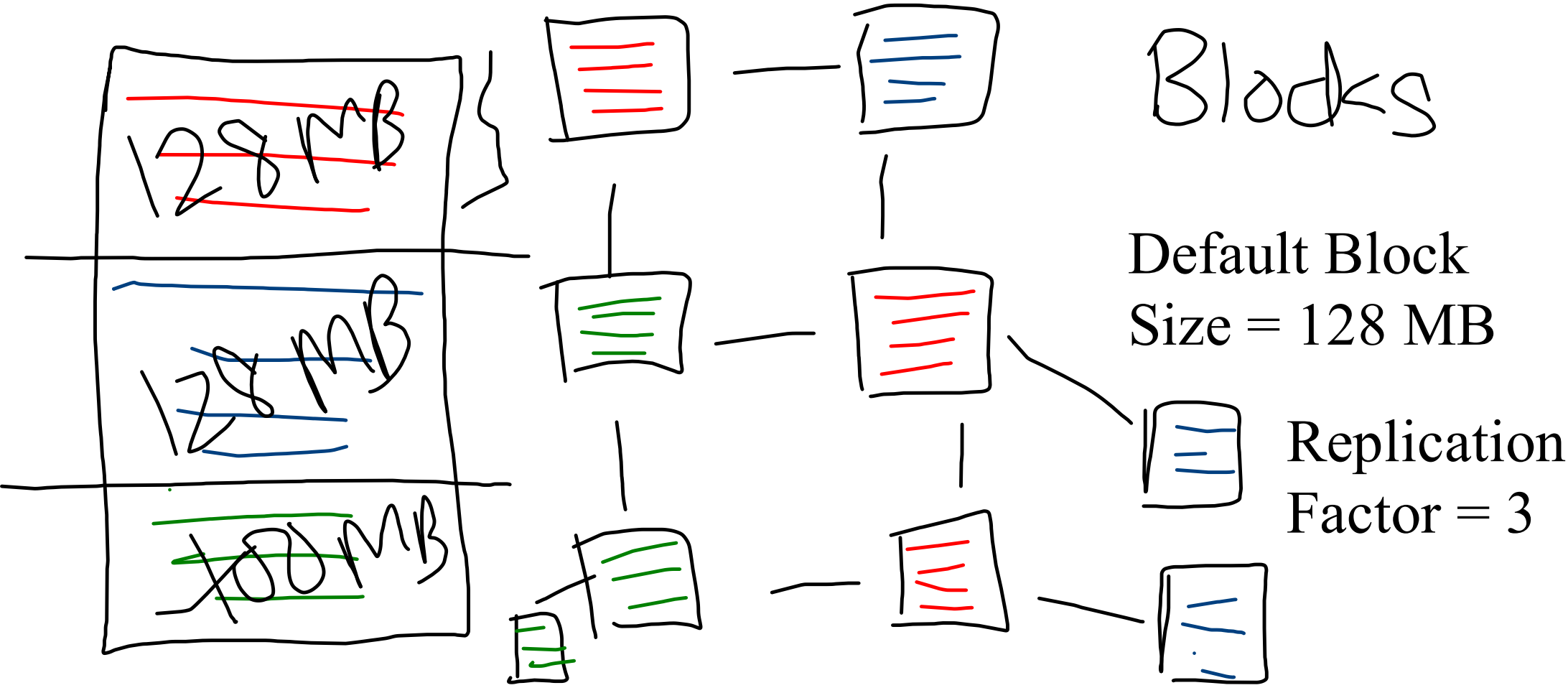
Reduce Phase: (are, 1), (hello, 2), (how, 1), (world, 2), (you, 1)



YARN (Yet Another Resource Negotiator): Introduced in 2012. Responsible for managing computing resources in the cluster in an efficient way

Hadoop Ozone: Introduced in 2020.An object store for Hadoop


ozone


128 MB
128 MB
100 MB
Blocks
Default Block
Size = 128 MB
Replication
Factor = 3