Chapter 3
Key-Value Database

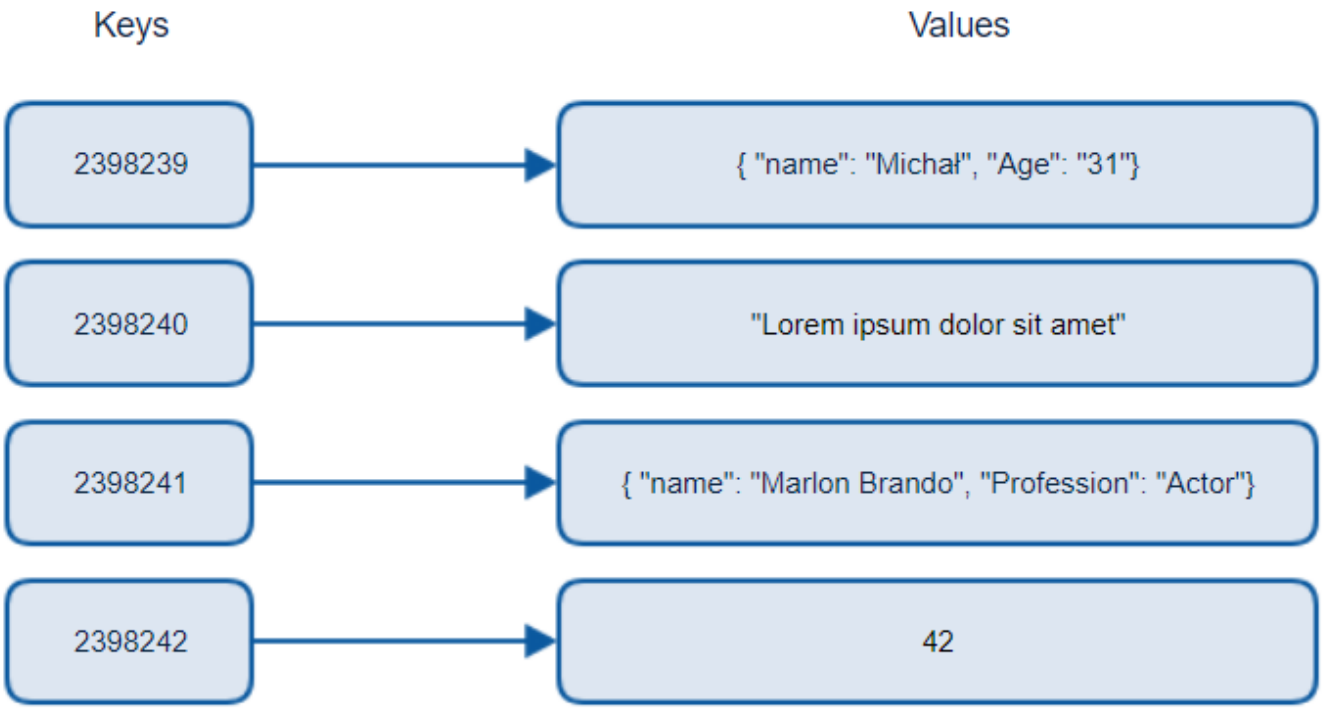
Key-Value Database

- A **Key-Value Database** is a type of **NoSQL database** that stores data as a collection of **key-value pairs**.
- It is designed for :
  - **high-speed lookups,**
  - **scalability,**
  - **and flexibility**,

Key-Value Database

- Each **key** is a unique identifier,
- and its **value** can be anything from a simple string to a complex object (JSON, XML, binary data, etc.).

Key-Value Database



Key-Value Database

- The **key** is used to retrieve the corresponding **value** instantly.
- Data retrieval is done using **hash tables, B-Trees, or in-memory caching**.
- Some key-value databases support **persistence**, replication, and clustering for scalability.

Key-Value Database

- **Advantages of Key-Value Databases**

1. **Easy to Use:** Simple operations like GET, SET, DELETE.
2. **High Performance:** Fast lookups using hash-based access.
3. **Scalability:** Can handle massive amounts of data across distributed nodes.
4. **Flexibility:** Schema-less structure allows storing various data types.
5. **Efficient Caching:** Used for in-memory caching (e.g., Redis, Memcached).

Key-Value Database

- **Use Cases of Key-Value Databases**
- **Session Management:** Storing user session data (e.g., Redis for login sessions).
- **Real-time Analytics:** Handling fast-moving data in stock trading, IoT.
- **Gaming Leaderboards:** Fast retrieval of high-score rankings.
- **Configuration Storage:** Managing settings and configurations in microservices.

Key-Value Database

Example : Calculate total sales per store for the current year?

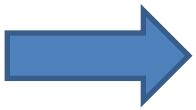
2012-01-01	09:00	San Jose	Men's Clothing	214.05	Amex
2012-01-01	09:00	Fort Worth	Women's Clothing	153.57	Visa
2012-01-01	09:00	San Diego	Music	66.08	Cash
2012-01-01	09:00	Pittsburgh	Pet Supplies	493.51	Discover
2012-01-01	09:00	Omaha	Children's Clothing	235.63	MasterC
2012-01-01	09:00	Stockton	Men's Clothing	247.18	MasterC
2012-01-01	09:00	Austin	Cameras	379.6	Visa
2012-01-01	09:00	New York	Consumer Electronics	296.8	Cash
2012-01-02	15:20	Lincoln	Cameras	242.2	Discover
2012-01-02	15:20	Madison	Baby	254.15	MasterC
2012-01-02	15:20	Wichita	Cameras	446.66	Amex
2012-01-02	15:20	Irvine	Computers	9.23	Discover
2012-01-02	15:20	Anaheim	Cameras	3.64	Visa
2012-01-02	15:21	Birmingham	Music	156.68	Cash

Key-Value Database

- In traditional programming, we will make Hash tables in the form <key-value>

1. For each entry, enter the city and the sale price
2. If we find an entry with a city already entered, we group them by adding the sales

San Jose	214.05
Fort Worth	153.57
San Diego	66.08
New York	55.60
San Jose	100.00



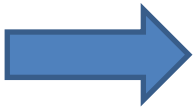
San Jose	31400.05
Fort Worth	15300.57
San Diego	66000.08
New York	55000.60

Key-Value Database

- Millions of rows
- Memory size issues
- Sequential processing ➔ Problem?

- Solution : Map-Reduce of Hadoop

San Jose	214.05
Fort Worth	153.57
San Diego	66.08
New York	55.60
San Jose	100.00



San Jose	31400.05
Fort Worth	15300.57
San Diego	66000.08
New York	55000.60

Key-Value Database
Hadoop and MapReduce

- MapReduce is a **distributed data processing model** and programming framework introduced by Google
- for handling **large-scale datasets** in parallel across multiple nodes in a cluster.
- It is widely used in **Big Data** processing and is the foundation of **Apache Hadoop**.

Key-Value Database
Hadoop and MapReduce

- **MapReduce** follows a **divide-and-conquer** approach by breaking down tasks into two key phases:
- **Map Phase** – Processes and filters data,

  outputting key-value pairs.
- **Reduce Phase** – Aggregates and summarizes results from the Map phase.

Key-Value Database
Hadoop and MapReduce

- **How MapReduce Works?**

1. **Input Data**

- The input dataset is typically stored in **HDFS (Hadoop Distributed File System)** and is split into **chunks** across multiple nodes.

2. **Map Phase**

- The Map function extracts **key-value pairs** from input data.
- The output is **partitioned** and sorted for the next step.

Key-Value Database
Hadoop and MapReduce

- **How MapReduce Works?**

3. Shuffle & Sort (Intermediate Phase)

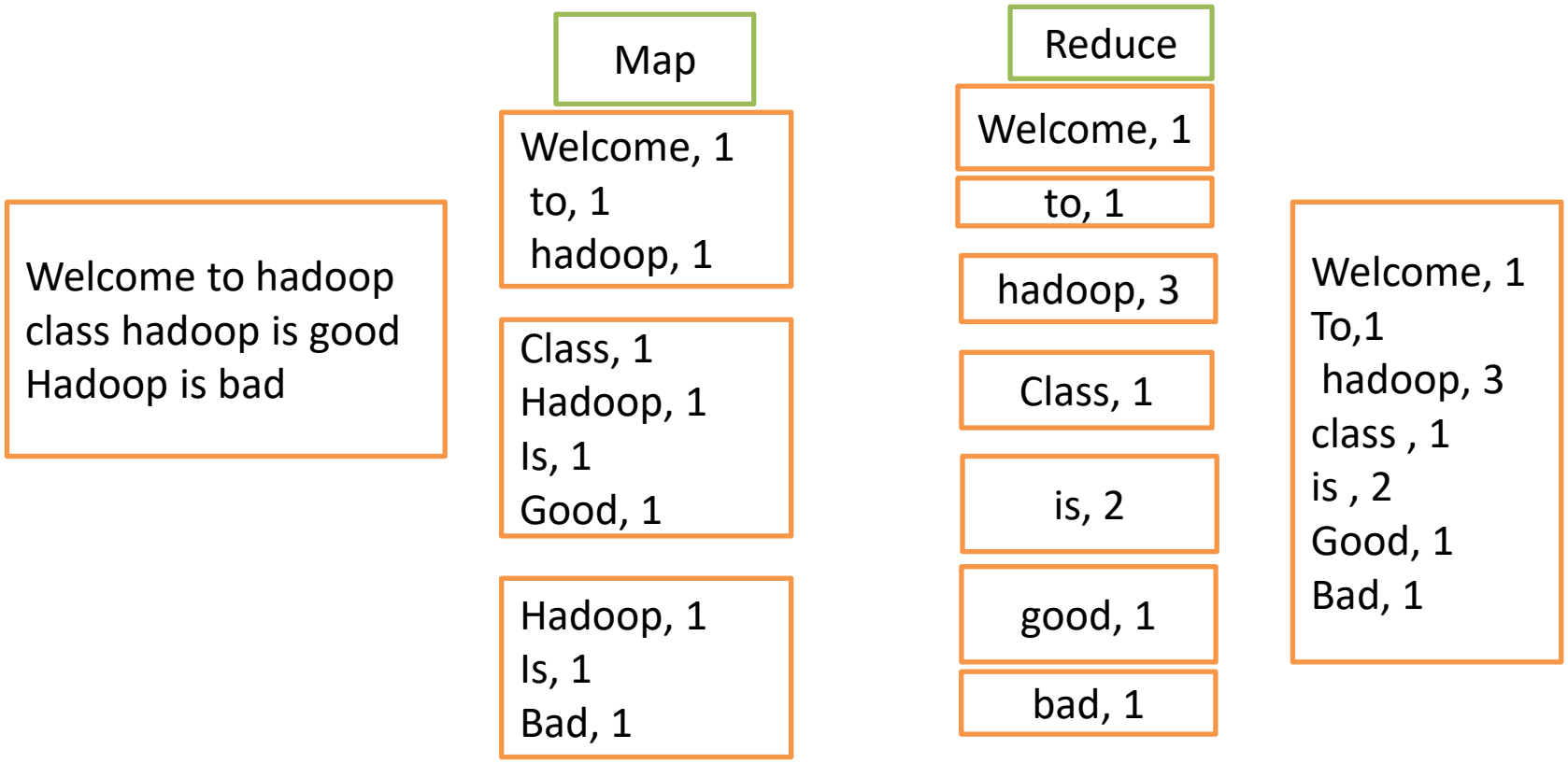
- The framework groups values by key and distributes them to reducers.
- This step ensures that all values for the same key go to the same reducer.

4. Reduce Phase

- The Reduce function processes grouped key-value pairs.
- It applies aggregation, counting, summation, or other computations.
- The final output is written to storage (e.g., HDFS, database).

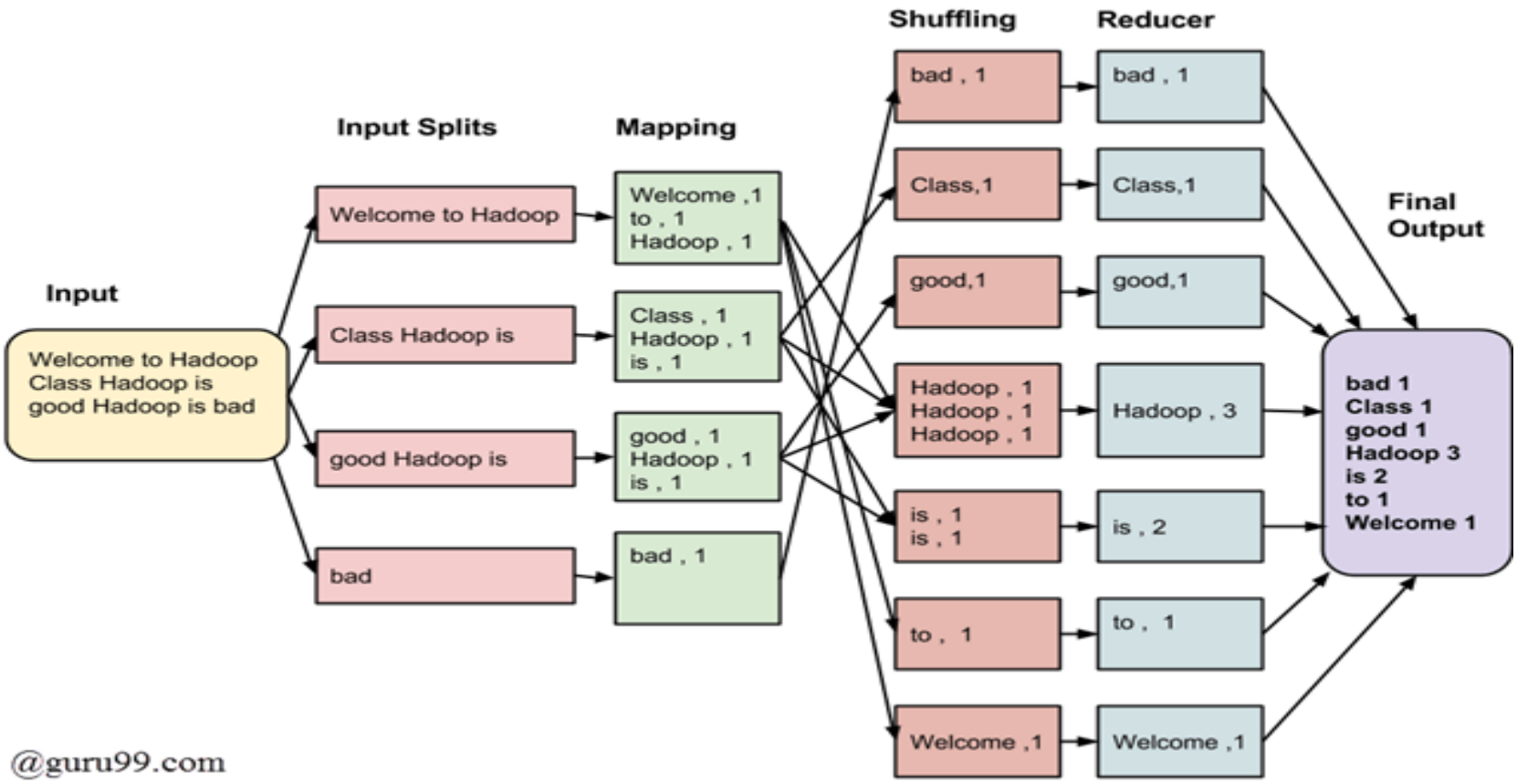
Key-Value Database
Hadoop and MapReduce

MapReduce – (ex: WORD COUNT)



Key-Value Database
Hadoop and MapReduce

MapReduce :**WORD COUNT**



Key-Value Database
Hadoop and MapReduce

- **MapReduce :WORD COUNT**
- **Program : MAP**

public class WordCountMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {

private final IntWritable one = new IntWritable(1); private Text word = new Text();

public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {

String line = value.toString();

StringTokenizer itr = new StringTokenizer(line.toLowerCase());

while(itr.hasMoreTokens())

{ word.set(itr.nextToken()); output.collect(word, **one**); } } }

Key-Value Database
Hadoop and MapReduce

REDUCER Code

```
public class WordCountReducer extends MapReduceBase implements Reducer<Text,
IntWritable, Text, IntWritable> {


   public void reduce(Text key, Iterator<IntWritable> values,  OutputCollector
<Text,  IntWritable> output, Reporter reporter) throws IOException {


 int sum = 0;
      while (values.hasNext()) {
         // replace ValueType with the real type of your value
         IntWritable value = (IntWritable) values.next();
         sum += value.get(); // process value
      }
      output.collect(key, new IntWritable(sum));
   }}
```

Key-Value Database
Hadoop and MapReduce

Calculate total sales per store for the current year?

2012-01-01	09:00	San Jose	Men's Clothing	214.05	Amex
2012-01-01	09:00	Fort Worth	Women's Clothing	153.57	Visa
2012-01-01	09:00	San Diego	Music	66.08	Cash
2012-01-01	09:00	Pittsburgh	Pet Supplies	493.51	Discover
2012-01-01	09:00	Omaha	Children's Clothing	235.63	MasterCard
2012-01-01	09:00	Stockton	Men's Clothing	247.18	MasterCard
2012-01-01	09:00	Austin	Cameras	379.6	Visa
2012-01-01	09:00	New York	Electronics	296.8	Cash
2012-01-02	15:20	Lincoln	Cameras	242.2	Discover
2012-01-02	15:20	Madison	Baby	254.15	MasterCard
2012-01-02	15:20	Wichita	Cameras	446.66	Amex
2012-01-02	15:20	Irvine	Computers	9.23	Discover
2012-01-02	15:20	Anaheim	Cameras	3.64	Visa
2012-01-02	15:21	Birmingham	Music	156.68	Cash

Key-Value Database
Hadoop and MapReduce

Advantages of MapReduce :

1. **Scalability** – Handles petabytes of data across distributed clusters.
2. **Fault Tolerance** – Automatically recovers failed tasks.
3. **Parallel Processing** – Processes data in parallel for speedup.
4. **Flexibility** – Can be implemented in various languages (Java, Python, etc.).

Key-Value Database
Hadoop and MapReduce

Disadvantages of MapReduce :

- **High Latency** – Not suitable for real-time processing.
- **Complex Programming Model** – Requires writing separate Map and Reduce functions.
- **I/O Intensive** – Reads and writes data multiple times, slowing performance.

Key-Value Database
Hadoop and MapReduce

Alternatives to MapReduce :

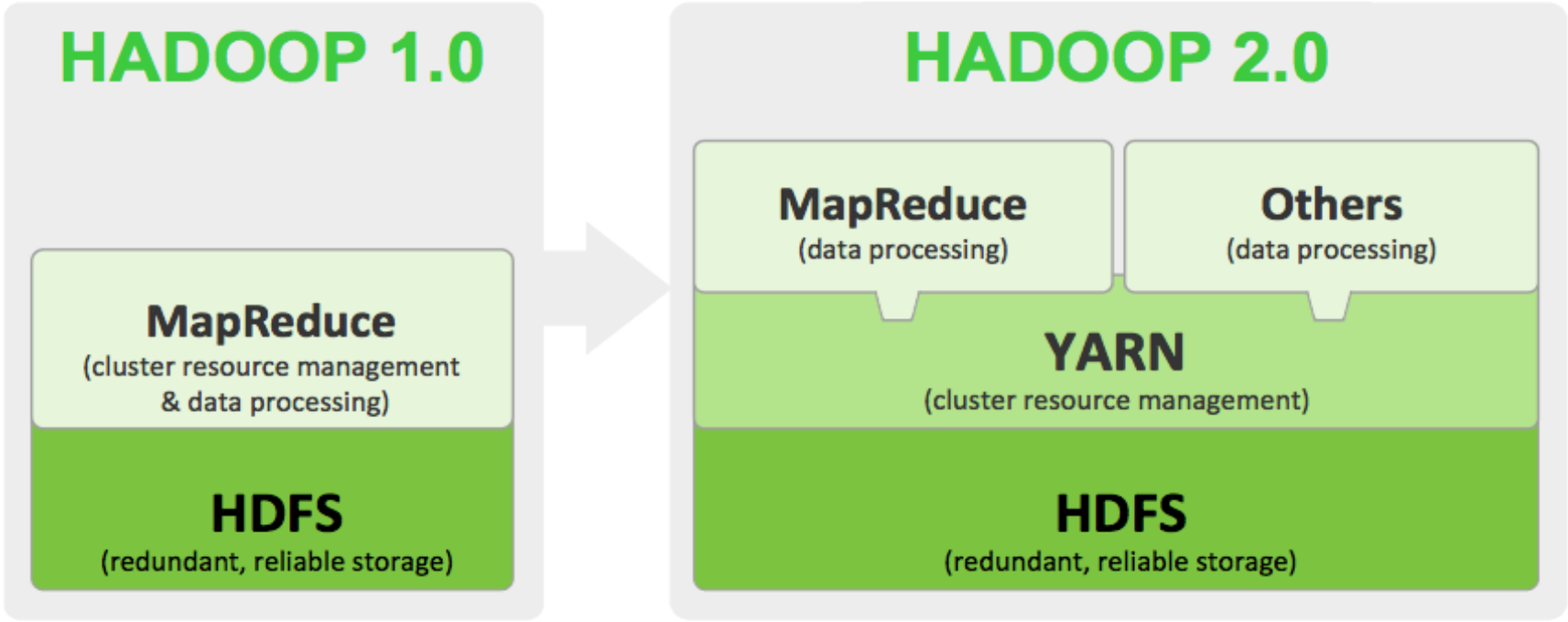
1. **Apache Spark** – Faster, in-memory processing.
2. **Apache Flink** – Real-time stream processing.

3. **Google BigQuery** – Serverless Big Data analytics.

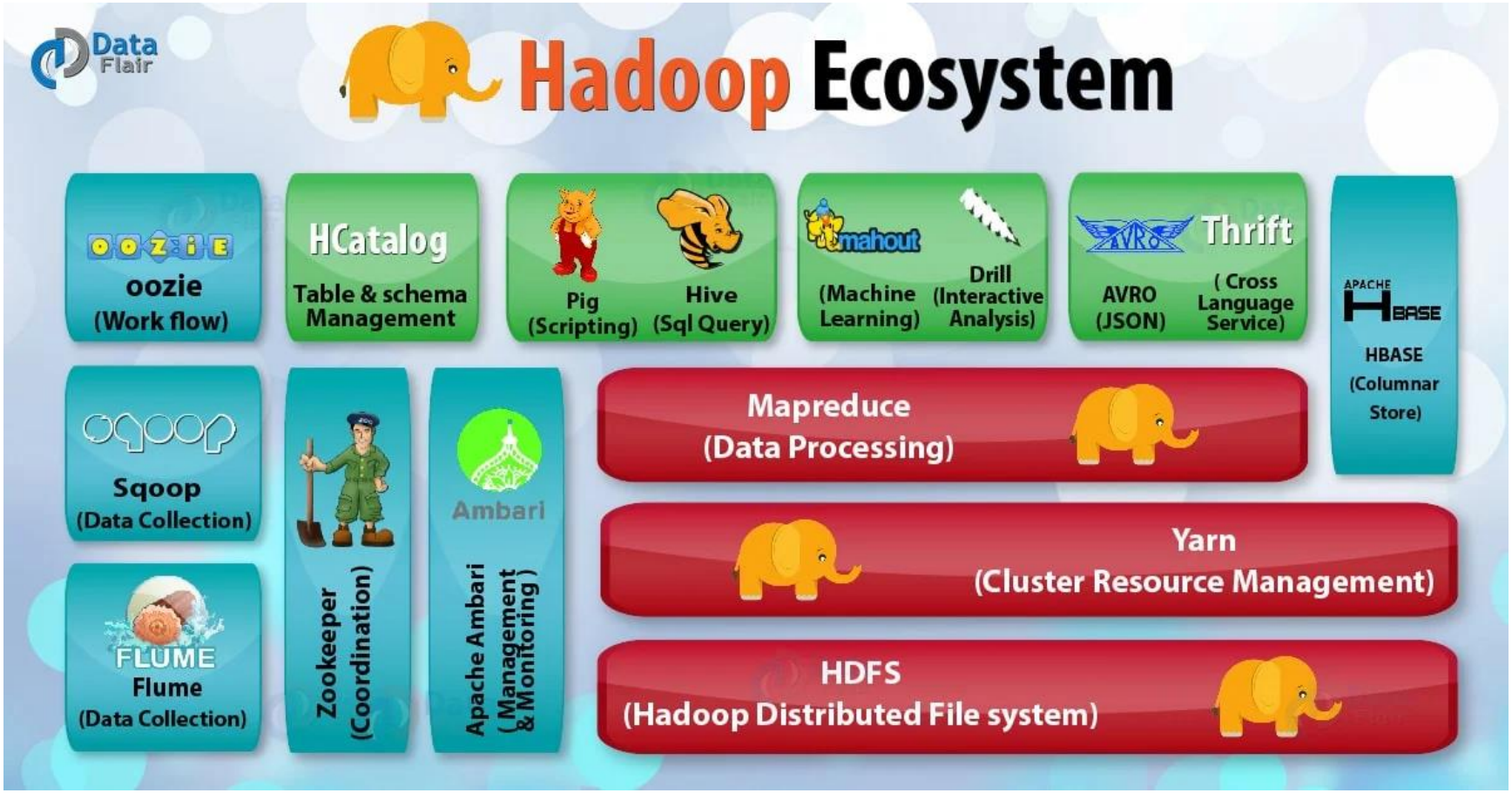
Key-Value Database
Hadoop and MapReduce

MapReduce is a programming model available in Hadoop environments

Used to access big data stored in the Hadoop File System (HDFS)



Key-Value Database
Hadoop and MapReduce



Key-Value Database
Hadoop and MapReduce

- A Hadoop cluster of 20,000 servers (standard and inexpensive servers) with 256 MB data blocks can process about 5 TB of data.
- With MapReduce, you can therefore reduce the processing time compared to sequential processing of such a large dataset.
- Google's cluster contains 10,000,000 servers

Key-Value Database
Hadoop and MapReduce

- With Hadoop and MapReduce, rather than sending the data to where the application or algorithms are located,
- The algorithms are executed on the server where the data already resides, which has the effect of speeding up processing.

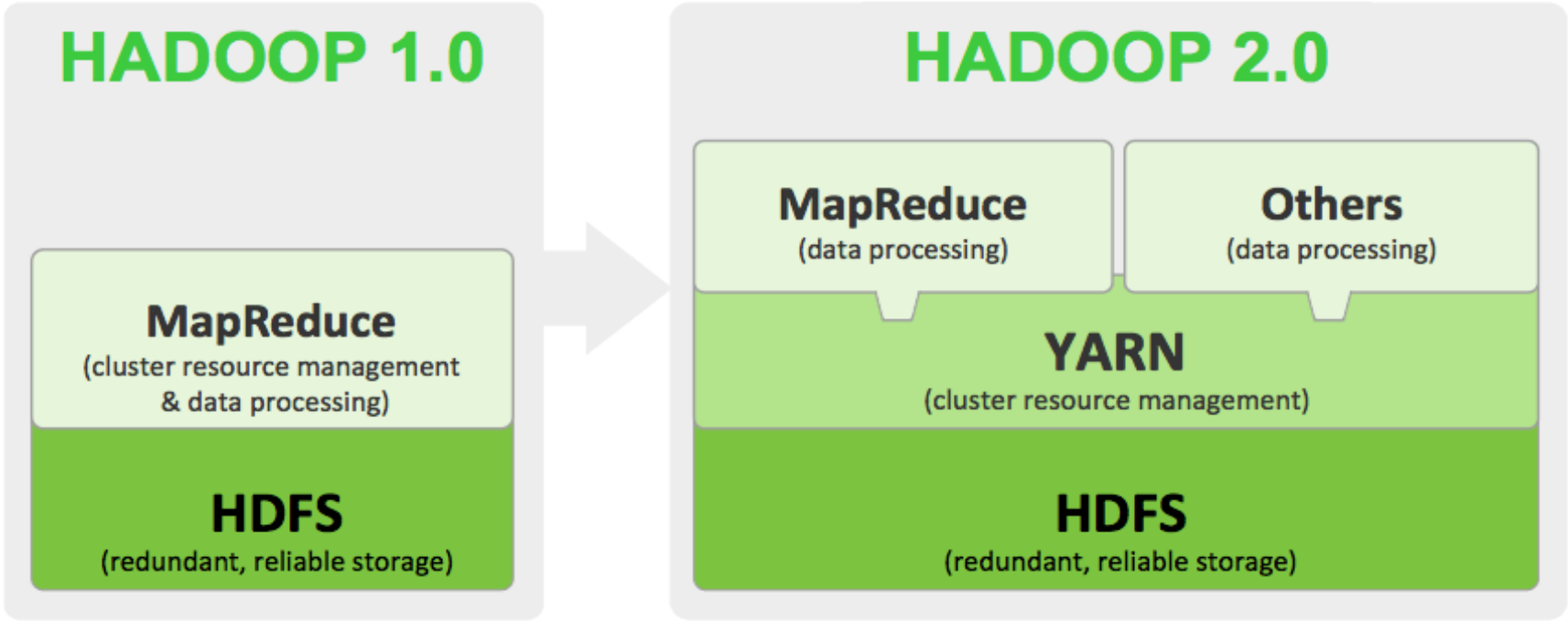
Key-Value Database
Hadoop and MapReduce

Distributed storage:

- Hadoop Distributed File System (HDFS),
- Amazon S3,
- Google Cloud Storage

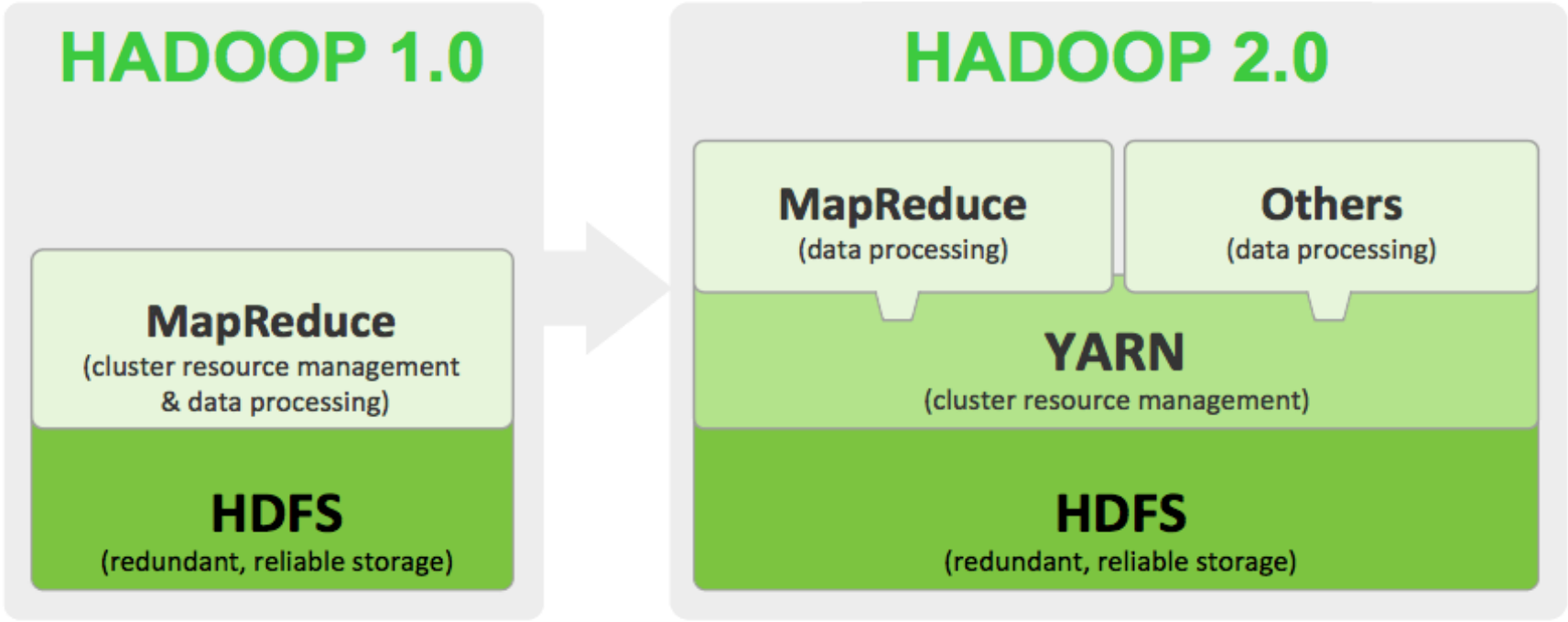
Key-Value Database
Hadoop Distributed File System (HDFS)

- Is a distributed file system that manages large data sets



Key-Value Database
Hadoop Distributed File System (HDFS)

- HDFS is one of the core components of Apache Hadoop, (along with MapReduce and YARN).
- Running on commodity hardware



Key-Value Database
Hadoop and MapReduce

- HDFS objectives :

1. Fast recovery from hardware failures :
   - HDFS can include thousands of servers failure of at least one server is inevitable
   - HDFS was designed to detect faults and automatically recover quickly

Key-Value Database
Hadoop and MapReduce

- HDFS objectives :

2. Access to streaming data

3. Hosting of large data sets.

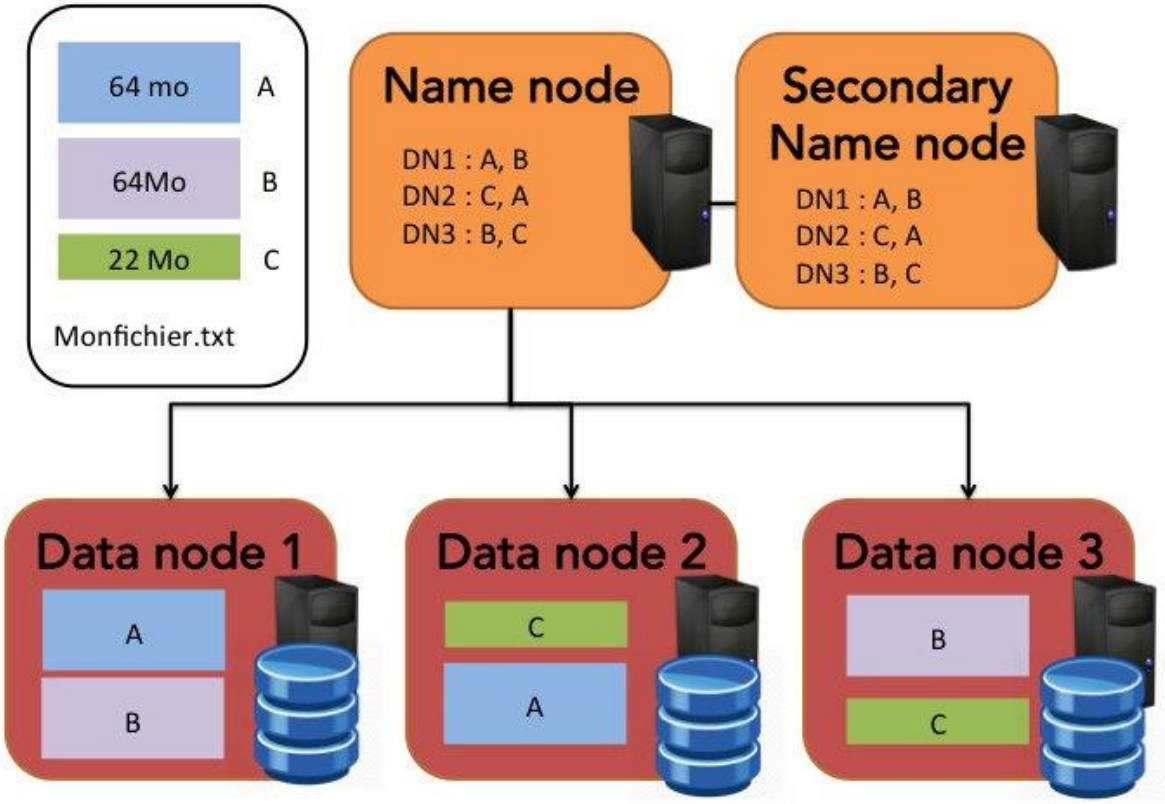
4. Portability

Key-Value Database
Hadoop and MapReduce

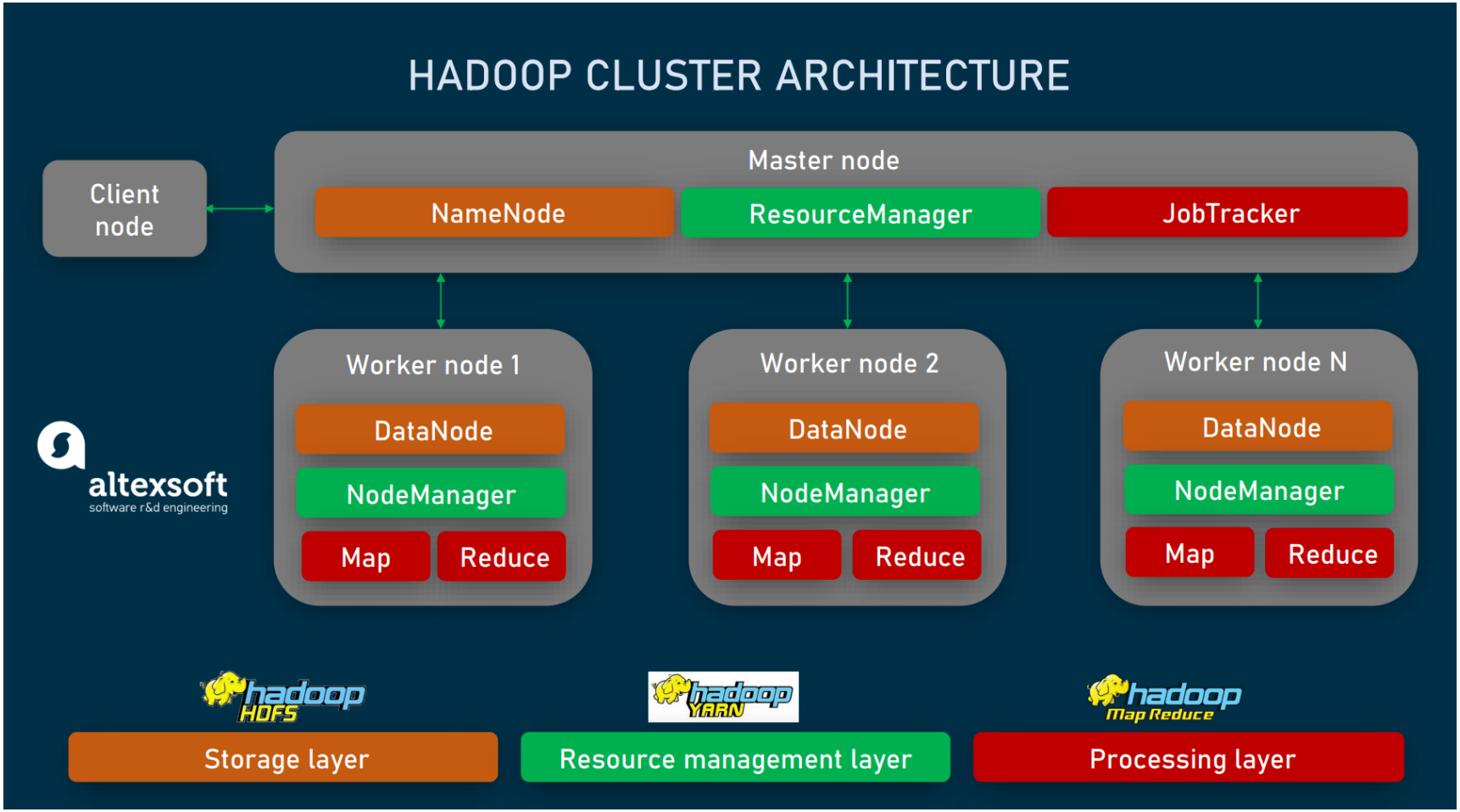
- A cluster of several machines Master/slave
- Principle data is stored on datanodes (slaves)
- Metadata on data blocks is managed by the namenode (master)

Key-Value Database
Hadoop and MapReduce

- Each data file is broken down into blocks.
- Default size 64 MB
- With replication principle



Key-Value Database
Hadoop and MapReduce



Key-Value Database
Hadoop and MapReduce