Big Data:
Hadoop

Extension to Databases

Professor: Pablo Ramos

pablo.ramos@u-tad.com

INTRODUCTION

- Big data: is a term that describes any type of structured, semi-structured or unstructured data set of large volume and complexity.
  - – Data processing techniques cannot be applied traditional.
  - – It is necessary to use various machines to process the information (cluster)

INTRODUCTION: The 3Vs

- Big data challenges: The 3Vs
  - Volume: Amount of information.
  - Variety: Sources of information.
  - Speed: Speed at which information is created.

INTRODUCTION: The 3Vs

- Volume:
  - Given the nature of the information (e.g. images, videos, etc.) that is stored, the amount of information that must be stored in storage systems is very high.
  - All available information is stored.
  - It is necessary to store information in various locations since it is not possible to do it in just one.
  - Given the volume of information stored, it is possible to apply more than one analysis technique that returns different results.

INTRODUCTION: The 3Vs

- Variety:
  - Originally, information was stored in SQL databases / local file systems or, in the case of needing to store large volumes of information, in expensive databases of large companies. It was necessary for all this information to be properly structured.
  - The current reality is totally different, information comes from various sources in various formats and the objective is to store said information in the original format so as not to lose information.

INTRODUCTION: The 3Vs

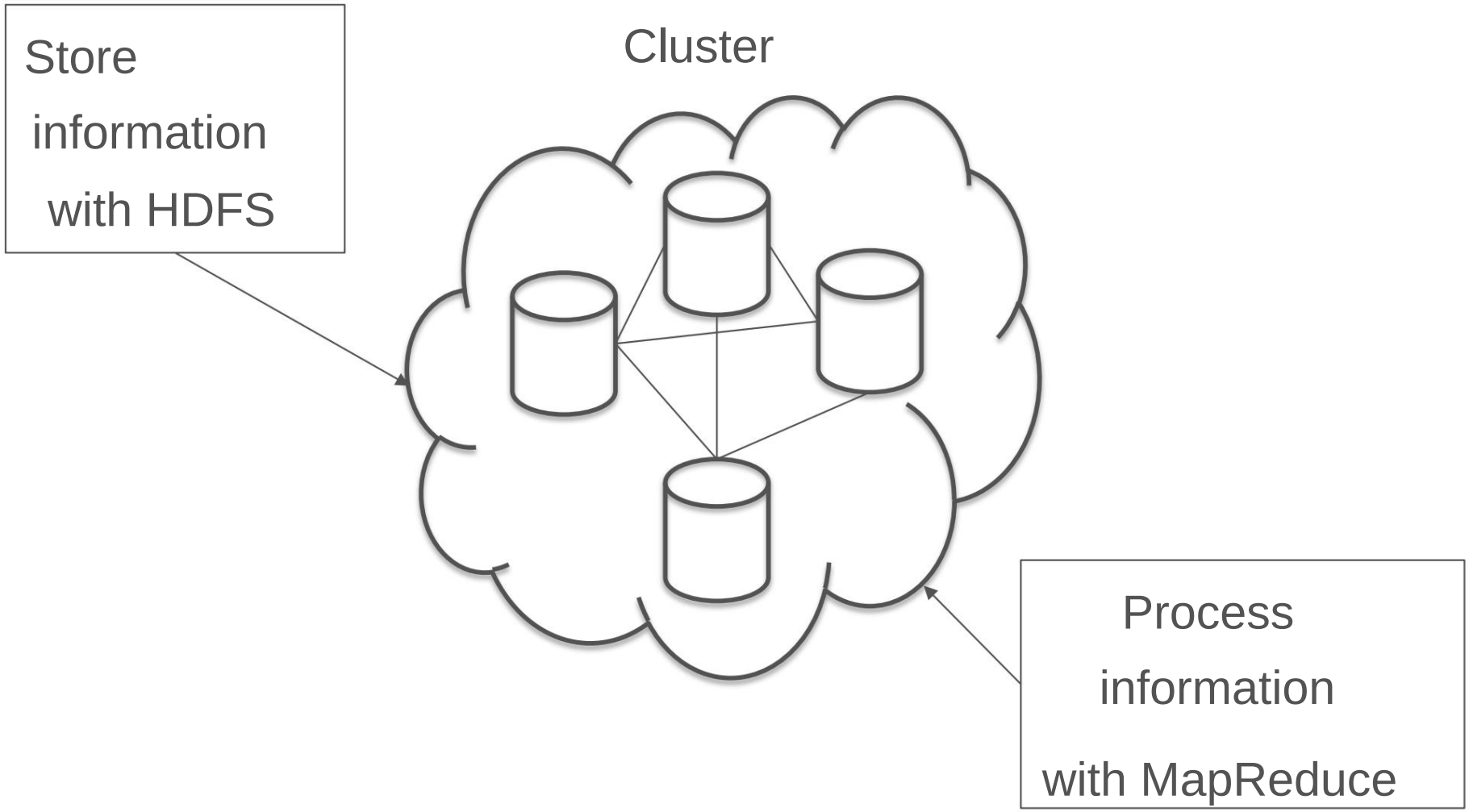
- Speed:
  - Given the nature of the origin of the information that is stores (eg Social Networks), the amount of information that reaches the storage systems per unit of time is very high (real time).
  - The speed at which information is stored (without preprocessing) is critical
  - The speed at which information is processed is also critical. It is not possible to use traditional analysis techniques where processing is centralized.

INTRODUCTION: The 3Vs

- Big data challenges: The 3Vs
  - Volume: Amount of information.
  - Variety: Sources of information.
  - Speed: Speed at which information is created.

- Other big data challenges:
  - Variability: Inconsistency of information regarding the time.
  - Veracity: Quality of the stored information.
  - Complexity: Information must be stored in a way that maintains the inherent correlation between data and the full set of information can be leveraged.

HADOOP: Introduction



HADOOP: Introduction

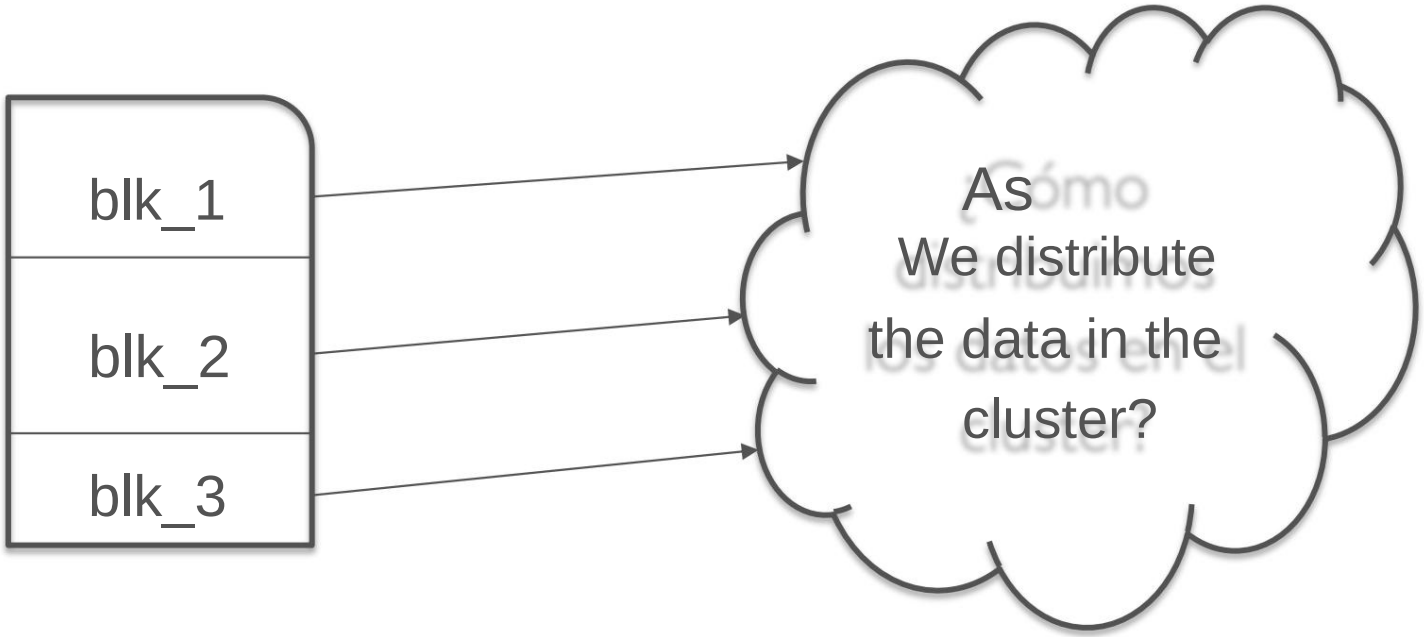
- Hadoop is an open source platform for the
  distributed storage and processing of large volumes of information in clusters of hundreds of nodes.
  - HDFS (Hadoop Distributed File System) for information storage. Based on Google File System.
  - MapReduce for processing. Based on MapReduce for Google
- In addition to these two tools, it implements other utilities that:
  - They work on MapReduce
  - They work on HDFS

HADOOP: Introduction

- In addition to these two tools, it implements other utilities:
  - They work on MapReduce
    - Hive: Allows you to create MapReduce instances from SQL queries.
    - Pig: It is a high-level programming language that allows you to easily create MapReduce instances.
  - They work on HDFS
    - Impala: It is a SQL query engine that allows parallel information processing.
    - Hbase: Column-based NoSQL database designed for real-time data processing.

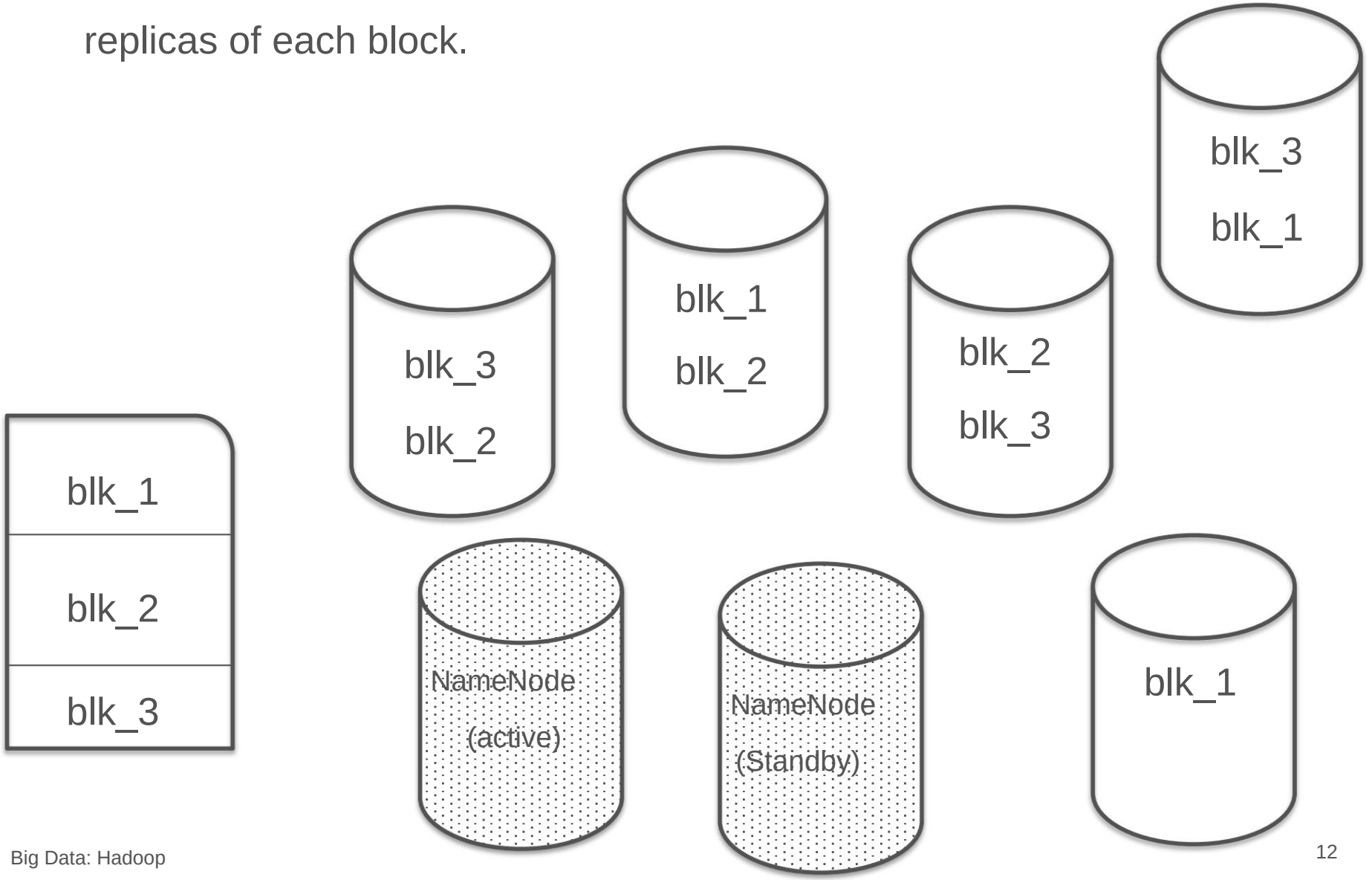
HADOOP: HDFS

- HDFS: Hadoop Distributed File System
  - Suppose we want to store a file in our distributed file system.
    - For better data handling and to improve data integrity, we split the file into sub-blocks.



HADOOP: HDFS

- HDFS: Hadoop Distributed File System – 3 replicas of each block.



HADOOP: HDFS

- HDFS has a master/slave architecture.
  - An HDFS cluster is composed of:
    - A NameNode that manages access to the file system. Typically there is one active NameNode and one in standby mode in case the first one fails.
    - A set of DataNodes (as many as the cluster has nodes) that manage data storage.
  - HDFS file system allows you to store files of all types. HDFS divides files into sub-blocks of a predefined size (64MB) and distributes them across DataNodes according to the replication factor.

HADOOP: HDFS

- HDFS has a master/slave architecture.
  - NameNode:
    - HDFS allows files to be structured hierarchically in a manner very similar to traditional file systems. All information regarding the structure and namespace is stored and managed by the NameNode.
    - Information regarding the replication factor and the Block placement is also managed by the NameNode.
      - Typically distributes one replica per rack.
      - When a file is consulted, the following are returned:

        blocks closest to the client.

HADOOP: MapReduce

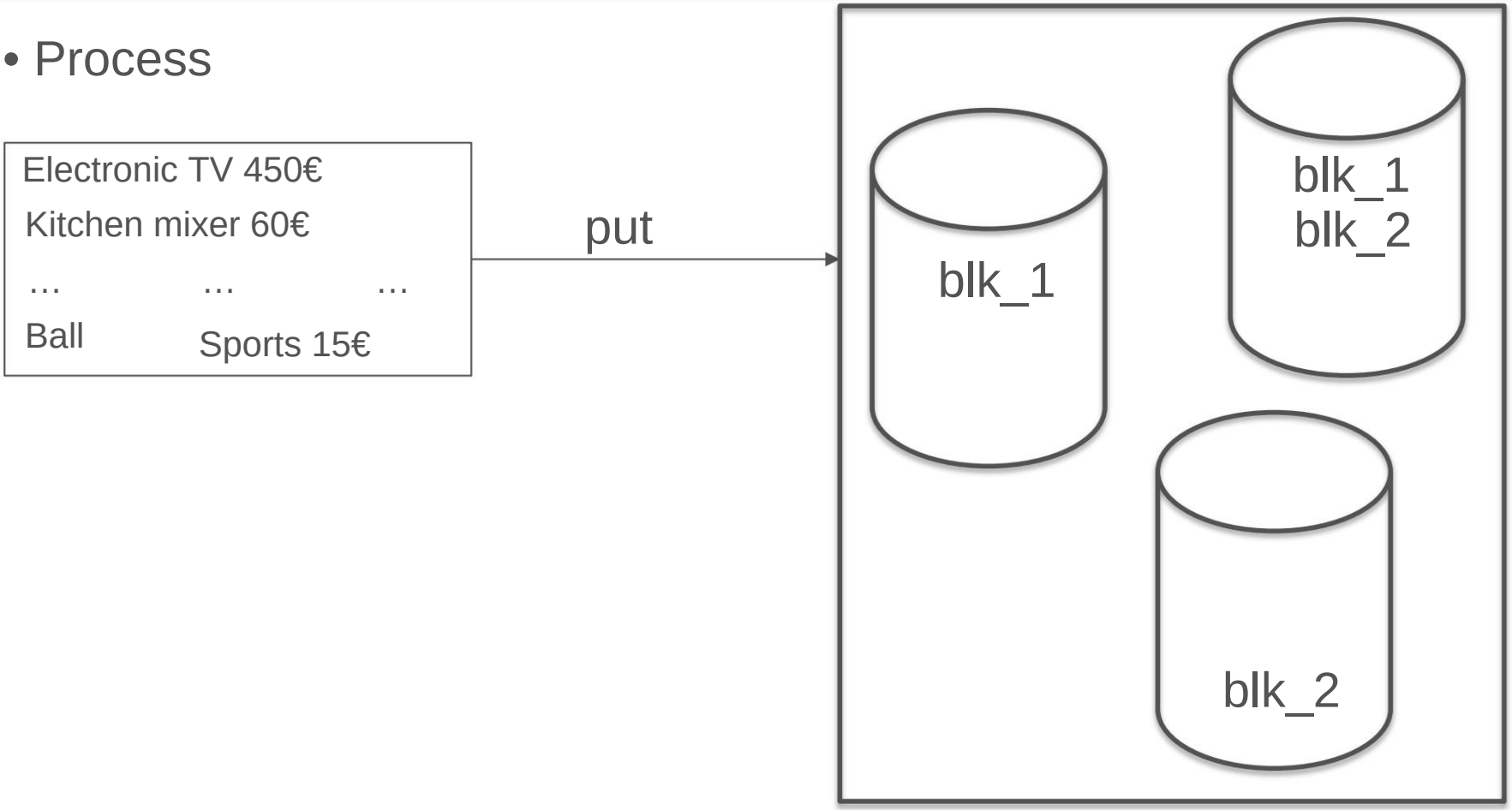
- MapReduce is a technique for parallel processing of large amounts of information in clusters.
  - It is based on the functional programming functions map and reduces.
  - However, the real potential of this technique comes from the parallel use of these functions. Execution of this technique in a single thread would not offer any substantial advantage.
  - This technique is only efficient for processing large amounts of information. For small data sets, sequential execution is more efficient.

HADOOP: MapReduce

- MapReduce consists of three steps:
  - Map: Performs data filtering and sorting tasks.
    - Each node stores part of the information process, runs a map task on your data and saves the result temporarily.
  - Shuffle & Sort: Rearranges the results.
    - The results are reorganized so that the results belonging to the same key are physically located in the same node.
  - Reduce: Performs data aggregation tasks.
    - Grouped results for the same key are processed by a reduce task that obtains the final result for that key.

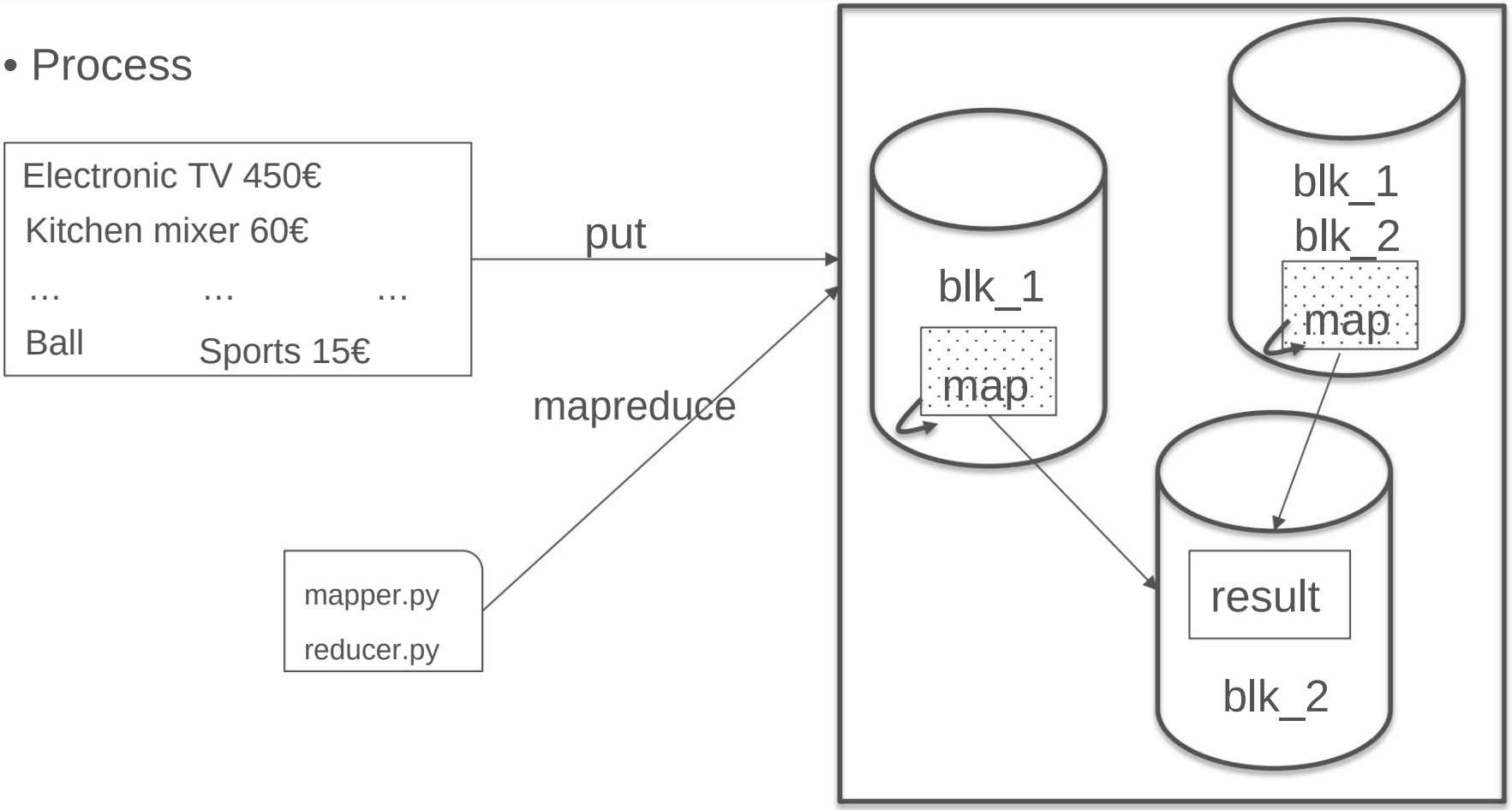
HADOOP: MapReduce

- Process



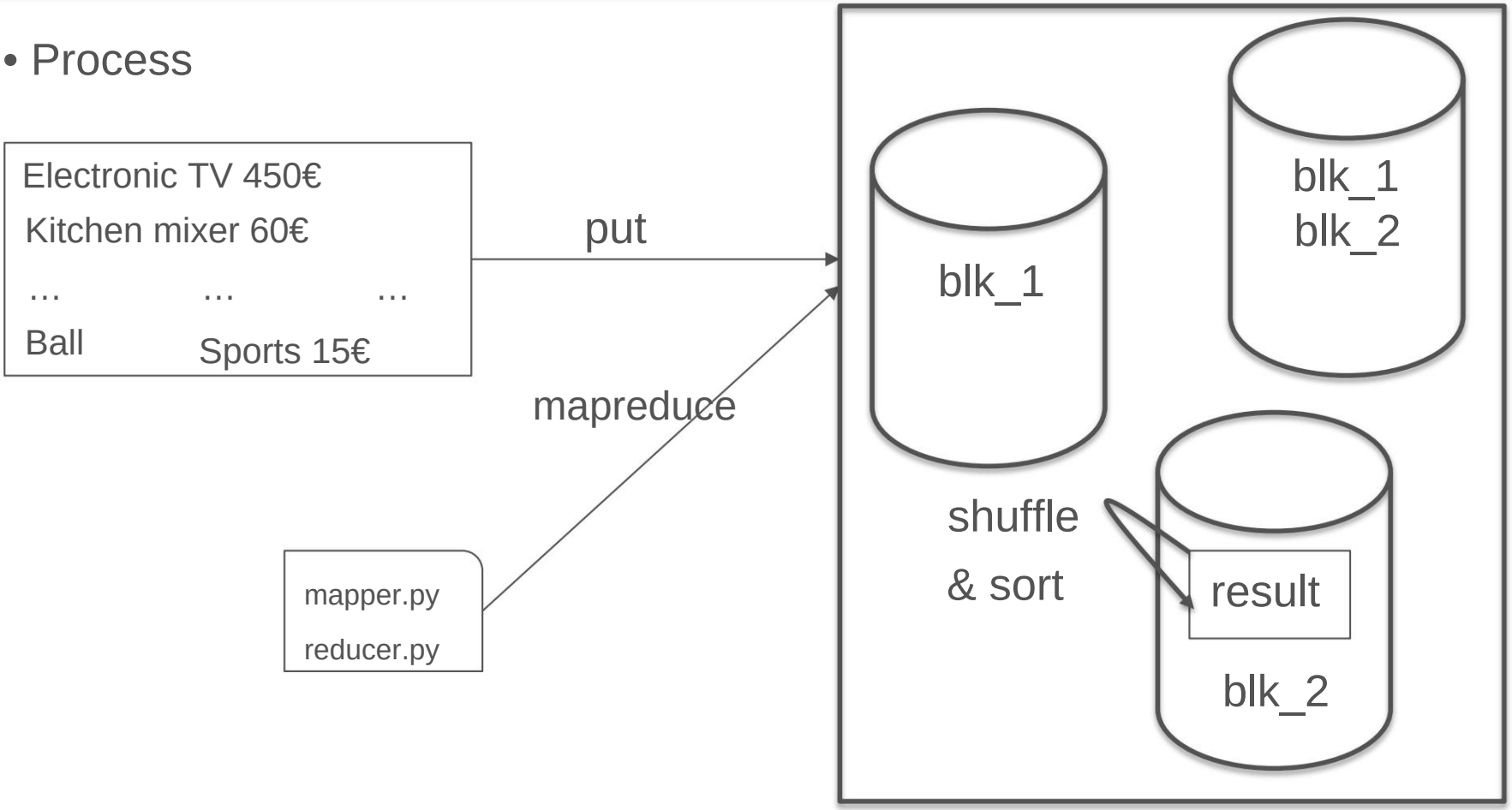
HADOOP: MapReduce

- Process



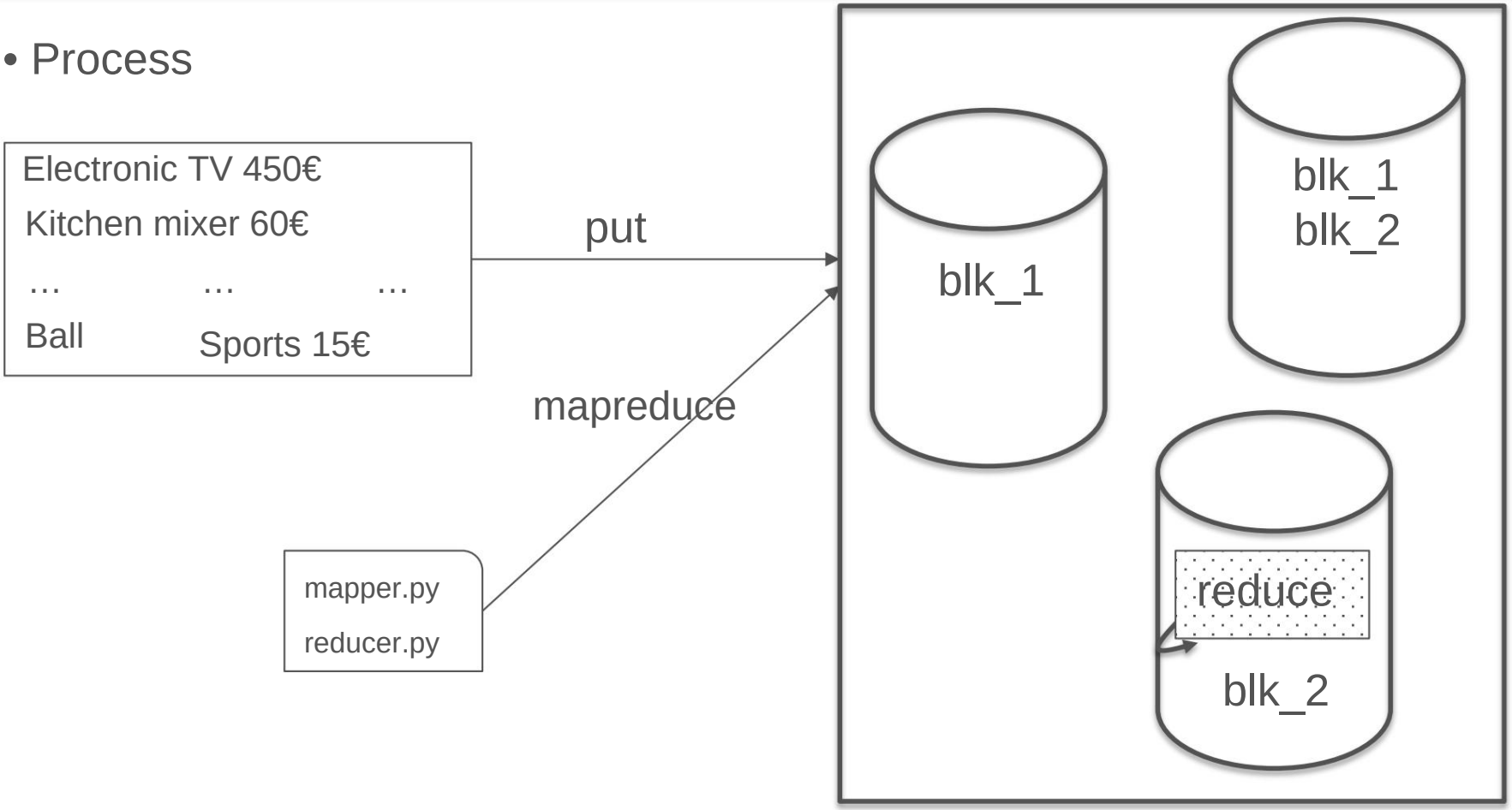
HADOOP: MapReduce

- Process



HADOOP: MapReduce

- Process



HADOOP: MapReduce

- Process