FACULTY OF DATA
AND DECISION SCIENCES
MAKING SENSE OF IT

Tutorial 2

MapReduce

Anis Barhoum

Recap of the last tutorial


BIG DATA . BIG. HUGE!

- 5 V's of Big Data:
  - Volume – amount of data
  - Velocity – the speed at which we receive and process the data
  - Value – comes from the hidden insights from the data
  - Variety – diversity of the data
  - Veracity – the accuracy or truth from the data

What now?

Big Data introduced significant challenges in **storing**, **processing**, and **managing** huge datasets efficiently.

Traditional systems struggled to scale, leading to the need for distributed solutions like Hadoop.

What is Hadoop?

Apache Hadoop Software is an open-source framework that:

- Allows for the distributed storage and processing of large datasets.
- Scales up from a single computer to thousands of clustered computers.
- Provides fault-tolerant distributed storage (through HDFS – Hadoop Distributed File System).

A short video: https://www.youtube.com/watch?v=aReuLtY0YMI (not affiliated with this course or any of the staff members)

What is MapReduce?

- A distributed data processing **programming paradigm**.
- Processes Large-Scale Datasets across multiple machines.
- Has two main steps: **Mapping** and **Reducing,** which we will define.

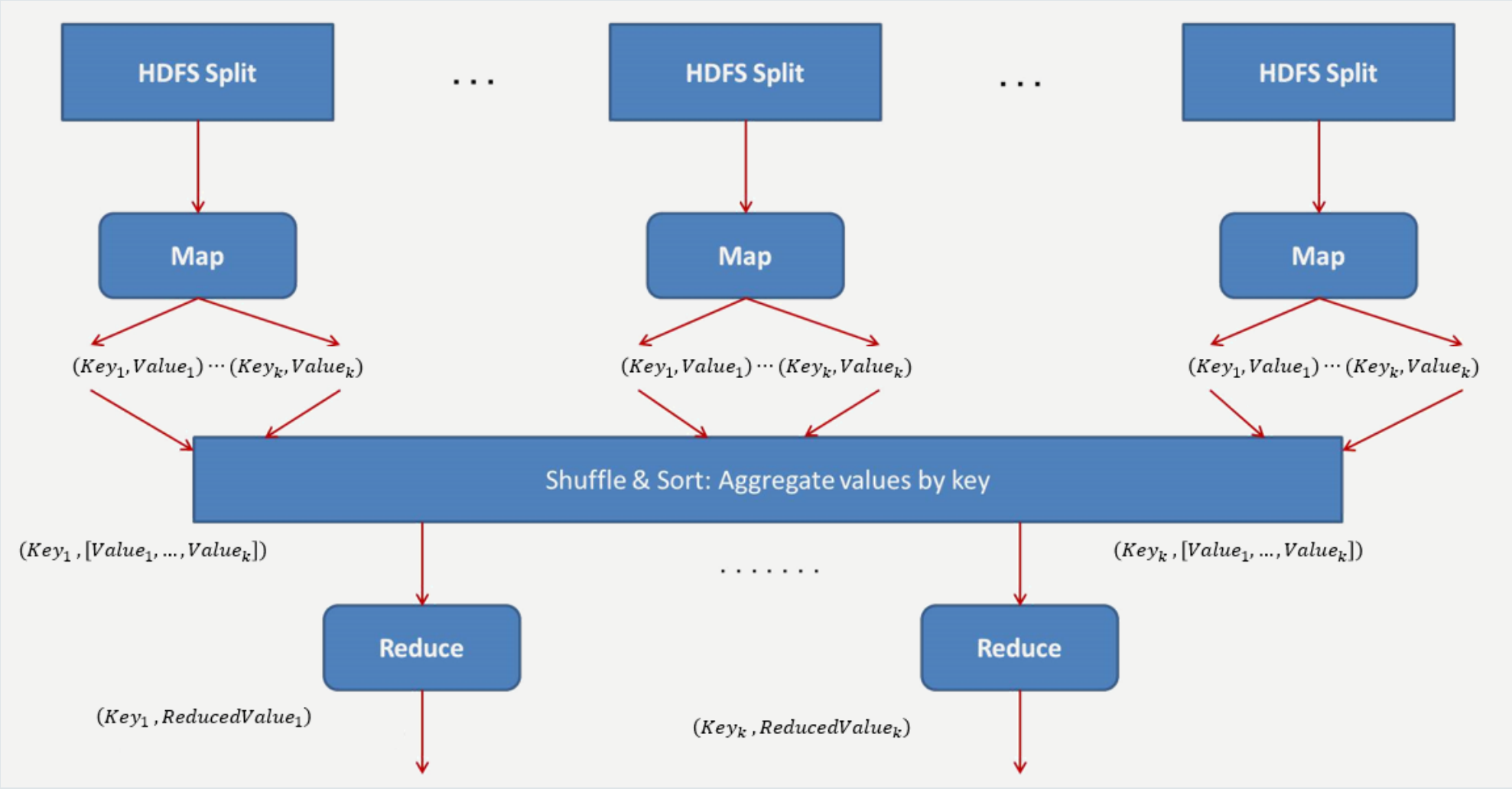
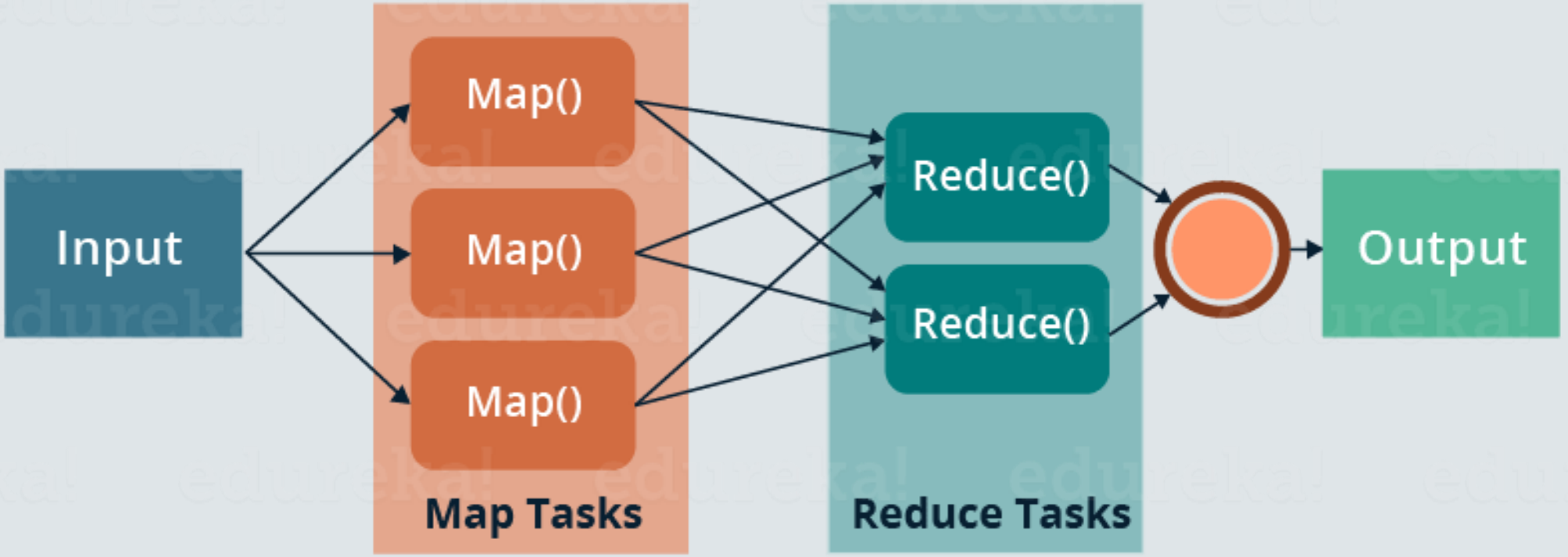


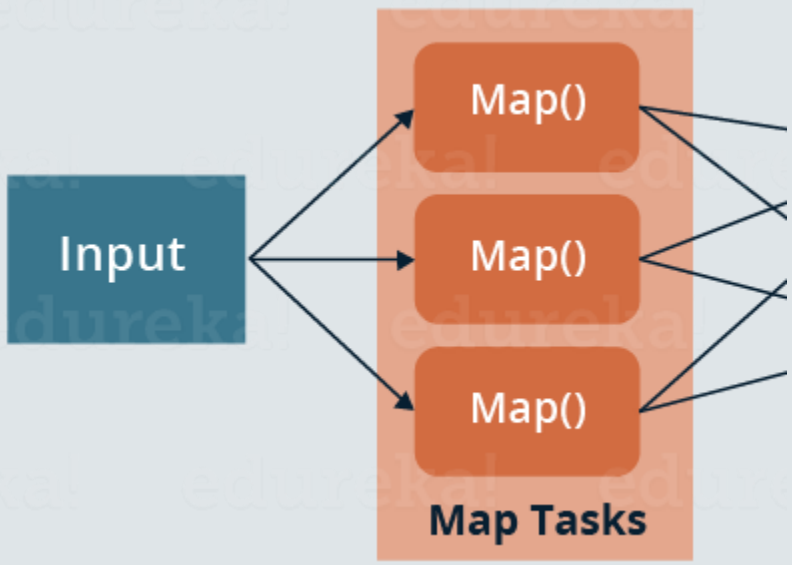
Figure 1: MapReduce Overview

MapReduce Components



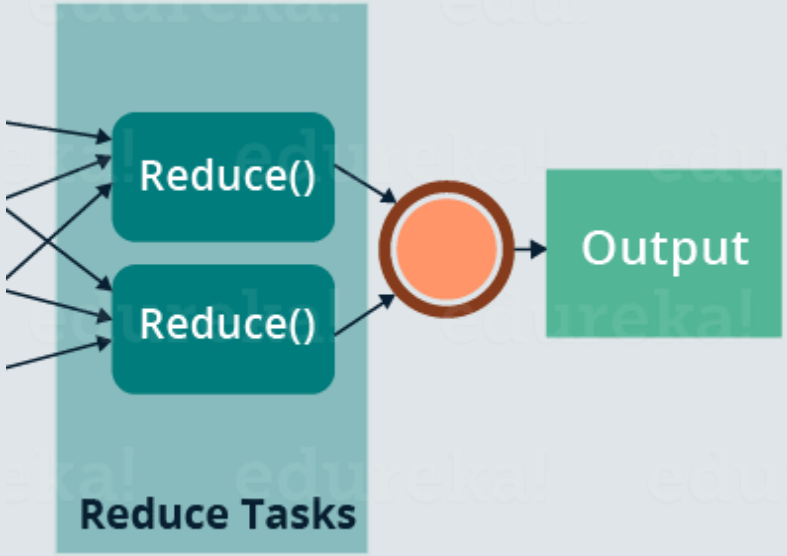
Pic Credit: https://www.edureka.co/blog/mapreduce-tutorial/

MapReduce Components: The Mapper



- **Maps** key-value pair to a set of intermediate key/value pairs.
- A **user defined function** (UDF), that iterates over all data.
- Intermediate pair can be different than input.
- Intermediate results are shuffled and sorted upon saving to disk (partition phase).

MapReduce Components: The Reducer



- **Aggregates** intermediate values with same key to a smaller set of values.
- A **user defined function** (UDF), with user defined logic.
- Output is stored in HDFS.

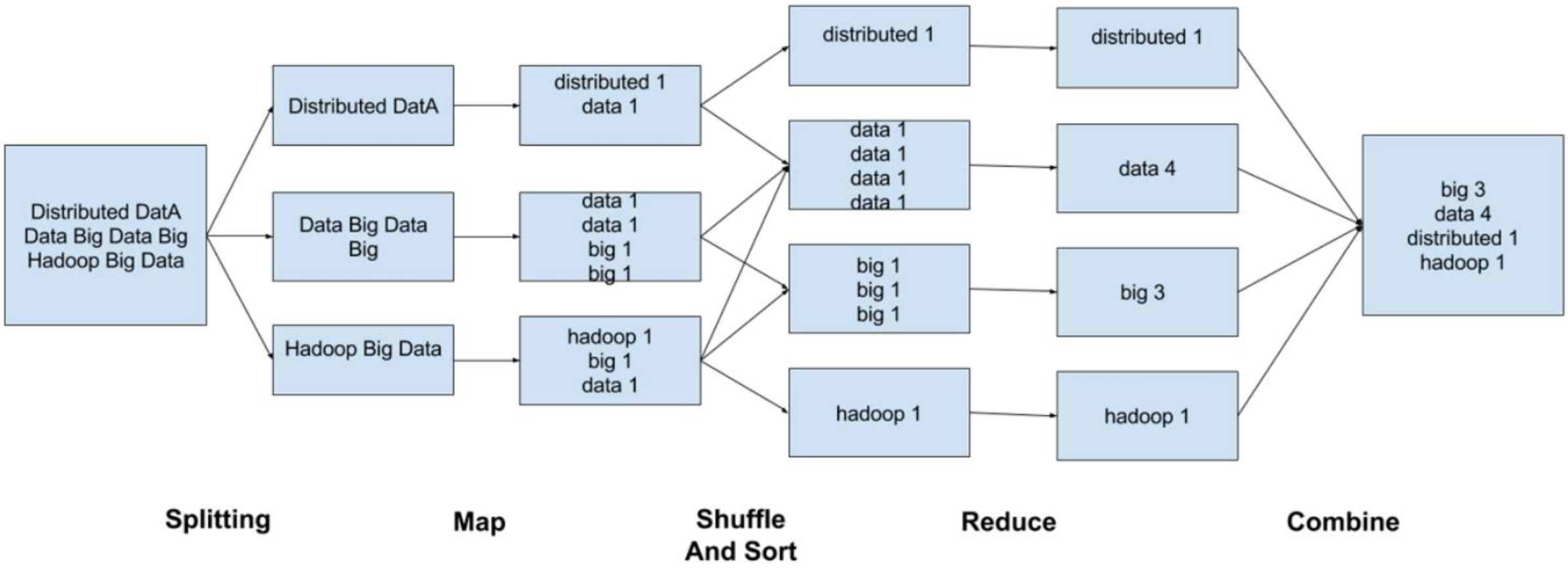


Figure 2: Word Count Example

Advantages

- **Parallel Processing**: Splits large tasks into smaller ones and running them simultaneously. (Divide and Conquer Paradigm).
- **No Bottlenecks**: Tasks are distributed among multiple machines, eliminating single points of failure.
- **Data Locality**: MapReduce moves computation to where the data is stored instead of moving the data.

Where is MapReduce Used in the Real World?

How many times does Ice Spice say “**Grrah**” in her songs?

Let’s answer this question using **MapReduce**