Laboratorio di Reti e Sistemi Distribuiti

19: Map Reduce

Roberto Marino, PhD[1]
roberto.marino@unime.it

[1]Dipartimento di Matematica, Informatica, Fisica e Scienze della Terra
Future Computing Research Laboratory
Università di Messina

Last Update: 27th May 2025

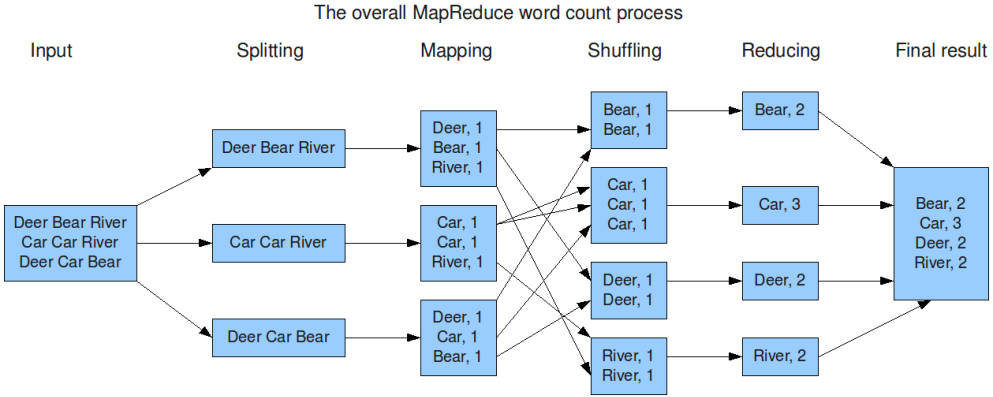
Cos'è MapReduce?

- Paradigma di programmazione distribuita per l'elaborazione di grandi volumi di dati.
- Introdotto da Google nel 2004.
- Composto da due funzioni principali:
  - **Map**: elabora dati in parallelo.
  - **Reduce**: aggrega i risultati.
- Implementato in Apache Hadoop, Spark, ecc.

Architettura di MapReduce

- **Input**: file diviso in blocchi.
- **Mapper**: applica la funzione Map su ogni blocco.
- **Shuffle & Sort**: raggruppa dati per chiave.
- **Reducer**: aggrega valori per chiave.
- **Output**: risultato finale scritto su disco.

Architettura di MapReduce



La funzione Map

- Riceve una coppia chiave-valore in input.
- Emette una lista di nuove coppie chiave-valore.

Esempio:

```
("ciao mondo") diventa [("ciao", 1), ("mondo", 1)]

```

Shuffle & Sort

- Raggruppa le coppie emesse dai Mapper.
- Tutti i valori associati alla stessa chiave vengono raccolti.
- Esempio:

```
[("ciao", 1), ("mondo", 1), ("ciao", 1)]
diventa ("ciao", [1, 1]), ("mondo", [1])

```

La funzione Reduce

- Riceve una chiave e una lista di valori.
- Esegue un'aggregazione (somma, media, ecc.).

Esempio:

```
("ciao", [1, 1]) diventa ("ciao", 2)
("mondo", [1]) diventa ("mondo", 1)

```

Vantaggi di MapReduce

- Alta scalabilità su cluster distribuiti.
- Tolleranza ai guasti.
- Modello semplice per il programmatore.
- Adatto a elaborazione batch di grandi dataset.