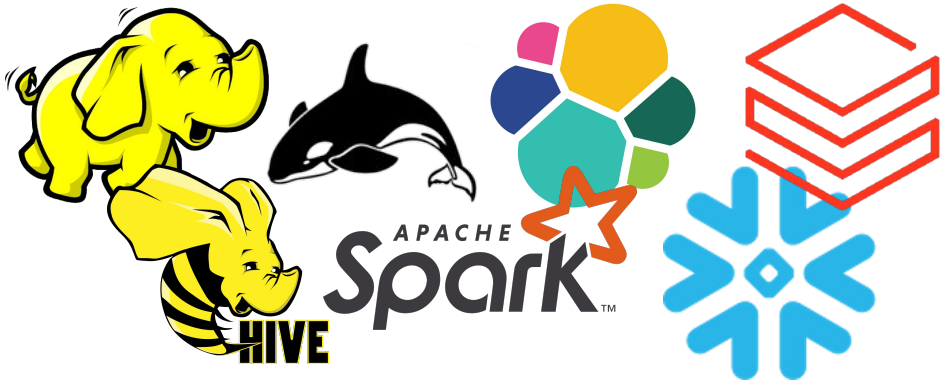
ADALTAS

# Big Data Ecosystem




# 3. The MapReduce Framework

# Reminder: HDFS + YARN architecture

ADALTAS

# MapReduce: a Framework

ADALTAS

MapReduce was made to help people write applications that:

- Process vast amounts of data (TB scale)
- On large clusters (thousands of nodes)
- In a reliable, fault-tolerant manner

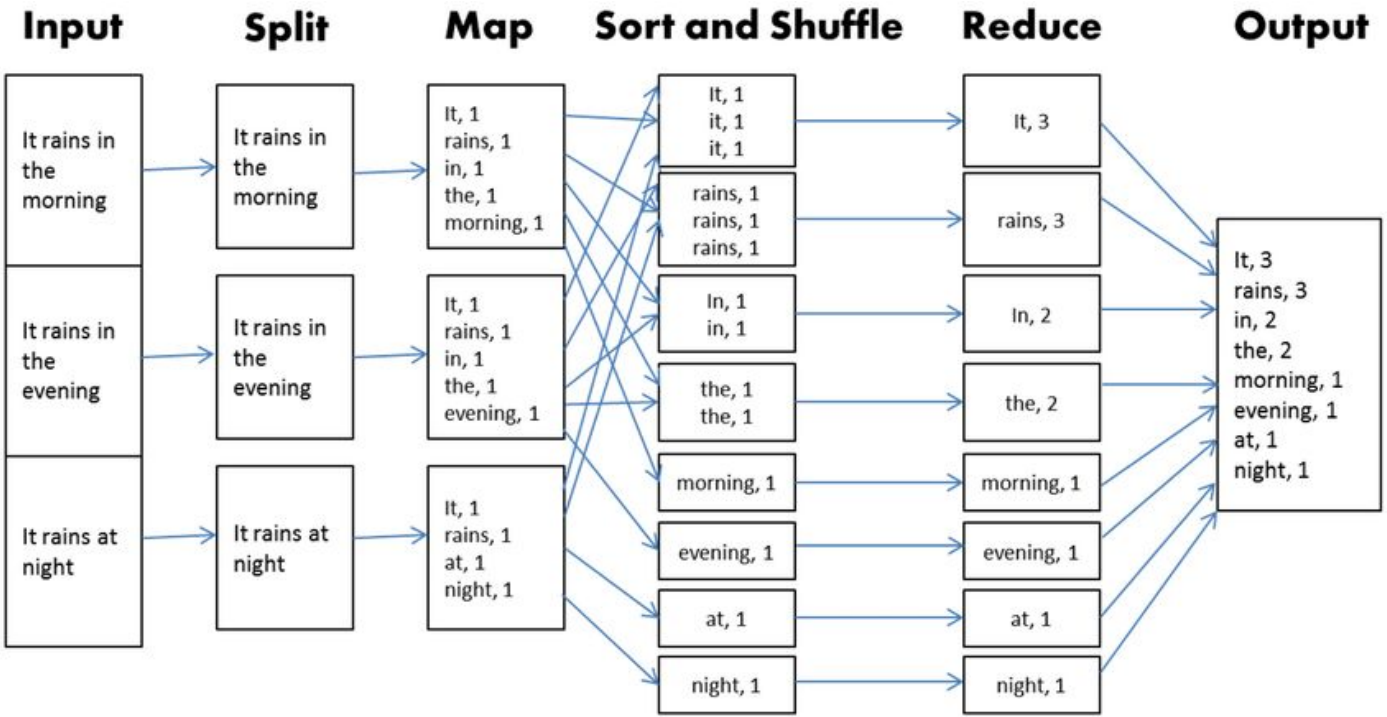
# MapReduce: Application steps

ADALTAS

Input = key/value pairs

1. Map: input k/v pairs to intermediate k/v pairs
2. Shuffle & sort: dispatch the k/v pairs with the same key to the same reduce
3. Reduce: set of values which share a key to smaller set of values

# MapReduce: Word count example



# MapReduce: Distribution on a cluster

ADALTAS

# MapReduce: Important properties

ADALTAS

- Outputs are **written to disk** between each step
- The number of mappers:
  - Depends on the number of blocks
  - Approx. 100 maps / mapper
- The number of reducers depends on the number of workers

# MapReduce vs other frameworks

- MapReduce was the ancestor of YARN
- Today, other frameworks that perform better are used:
  - Apache Tez
  - Apache Spark