Lab10

CS562 - Advanced Database Applications

- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

- It is made by apache software foundation in 2011.

- Written in JAVA.

Why Hadoop?

- Processing big data is very difficult in relational database.
- It would take too much time to process data and cost.

Hadoop solves this problem with distributed computing!

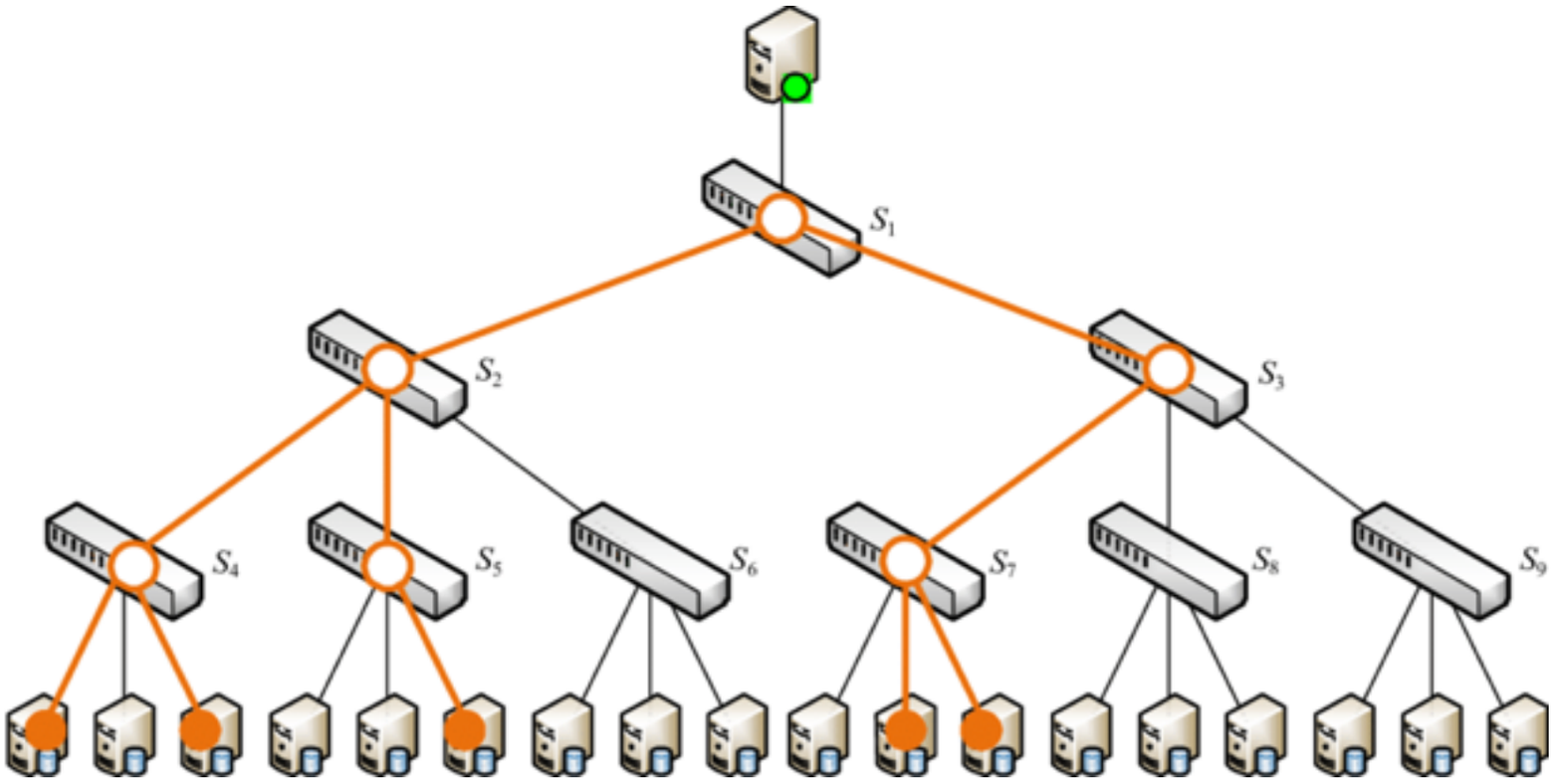
Pros of Hadoop

1. Computing power
2. Flexibility
3. Fault Tolerance
4. Low Cost
5. Scalability

Hadoop components

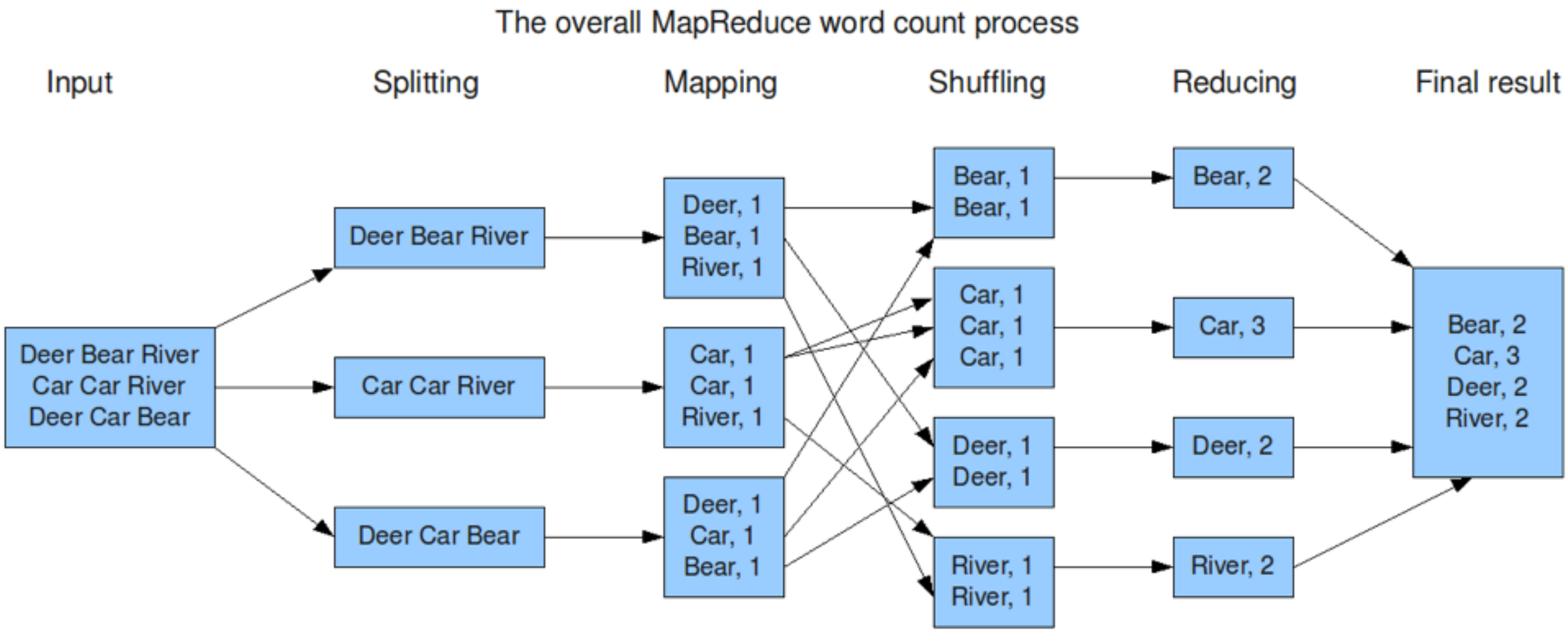
- Hadoop Distributed File System (HDFS)
- Data Processing Framework & MapReduce

Hadoop File Systems (HDFS)



The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. It employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.

Map - Reduce


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Hadoop with PIG **vs** Hadoop with Java

Basis for comparison	PIG	MapReduce
Operations	• Dataflow language. • High-Level Language. • Performing join operations in a pig are simple	• Data processing language. • Low-level language • Quite difficult to perform the join operations.
Lines of code and verbosity	Multi-query approach, thereby reducing the length of the codes.	require almost 10 times more the number of lines to perform the same task.
Compilation	No need for compilation. On execution, every Apache Pig operator is converted internally into a MapReduce job.	MapReduce jobs have a long compilation process.
Code portability	Works with any of the versions in Hadoop	No guarantee that supports with every version in Hadoop

Demo