

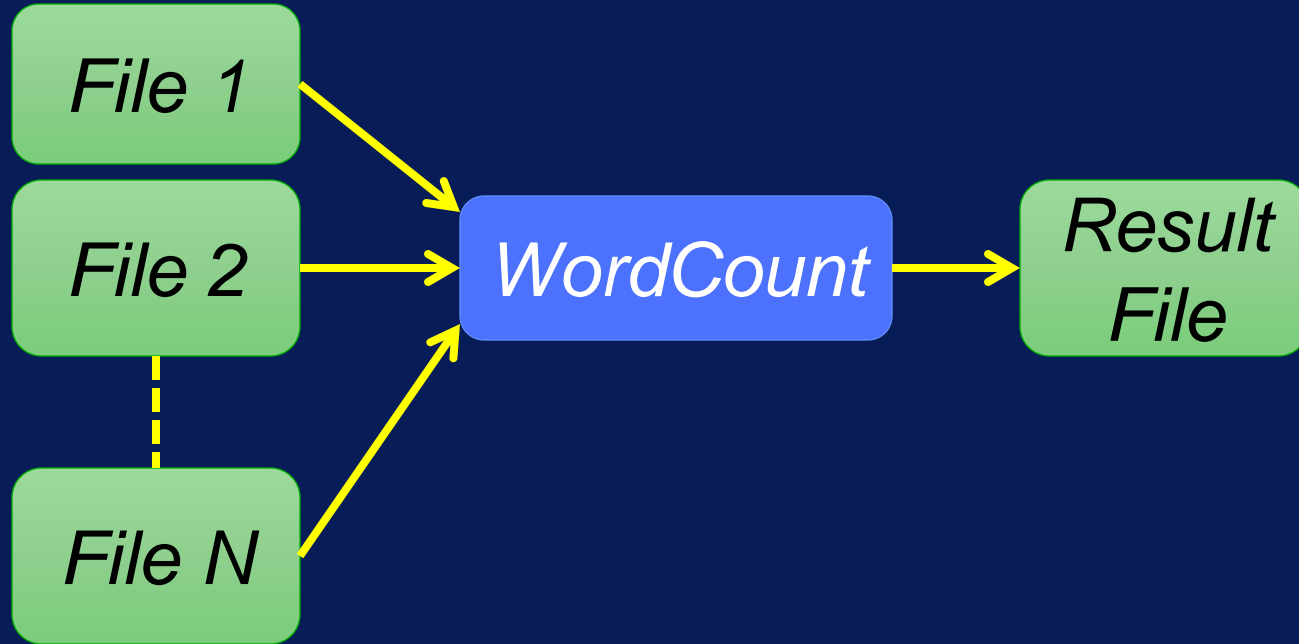
Big Data Processing Pipelines:

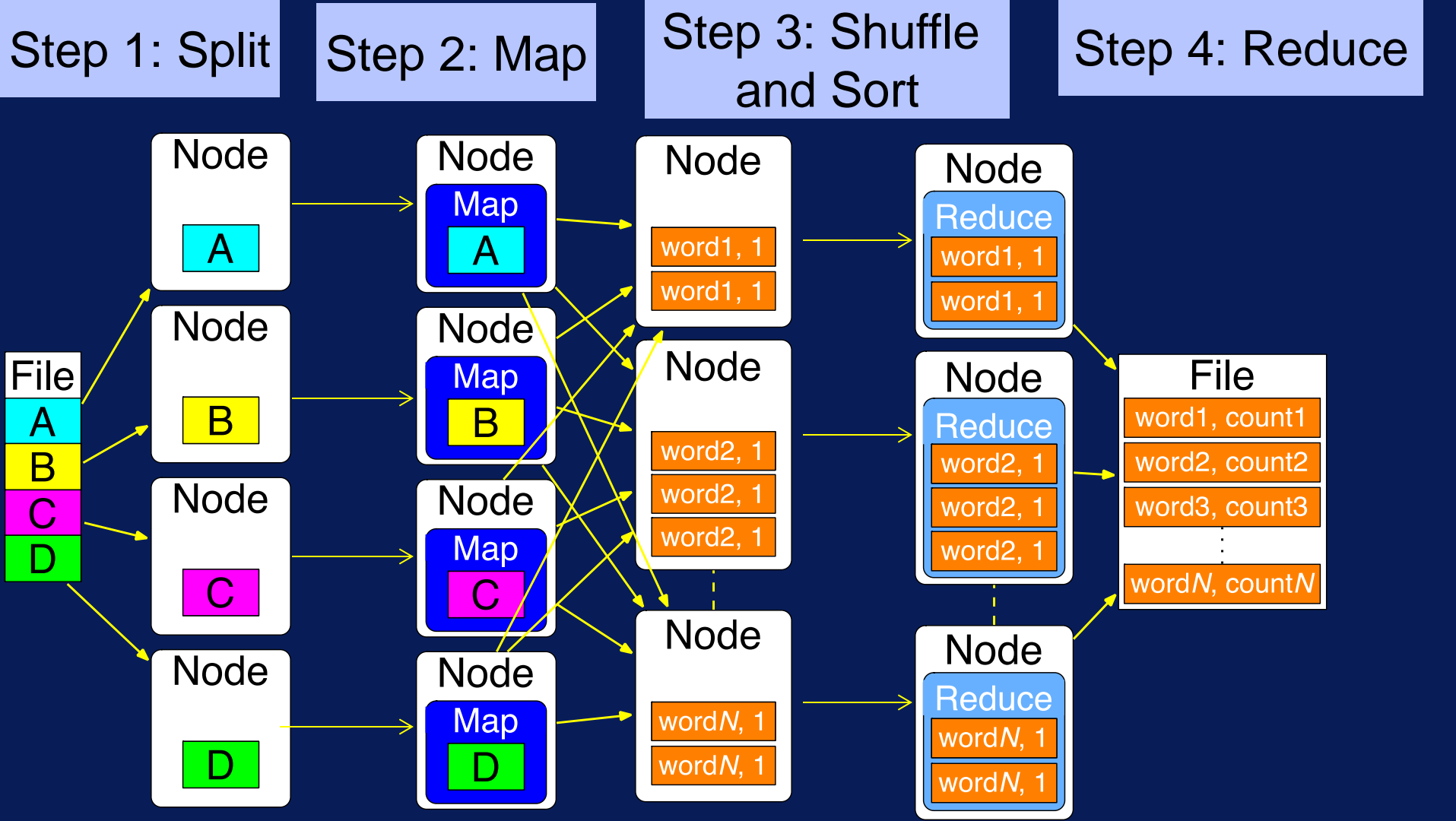
A Dataflow Approach

After this video you will be able to..

- Summarize what dataflow means and its role in data science
- Explain “split->do->merge” big data pipeline with examples
- Define the terms data parallel

Example MapReduce Application: WordCount





Split



Map



Shuffle
and Sort



Reduce






Represents a large
number of applications.

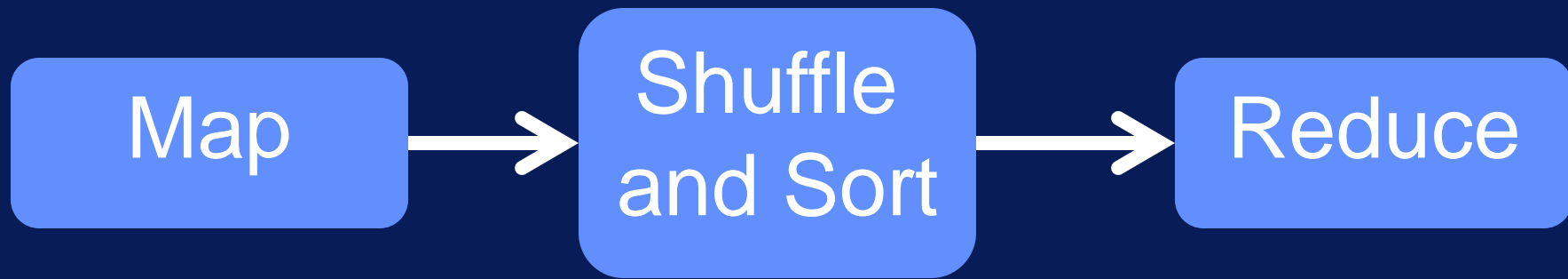


Big Data Pipelines



cat  sort

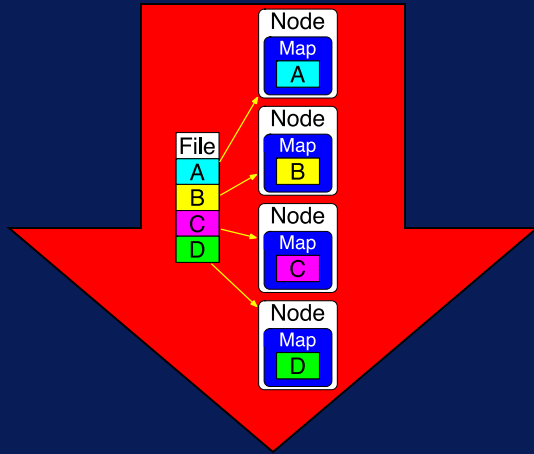
A UNIX pipe provides one-way communication
between two processes on the same computer



Map

Shuffle
and Sort

Reduce

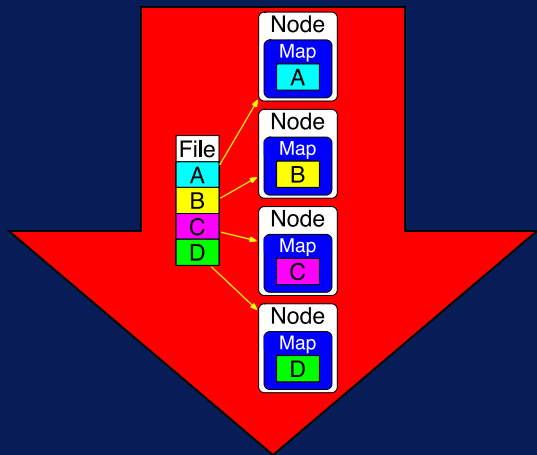


Parallelization
over the input

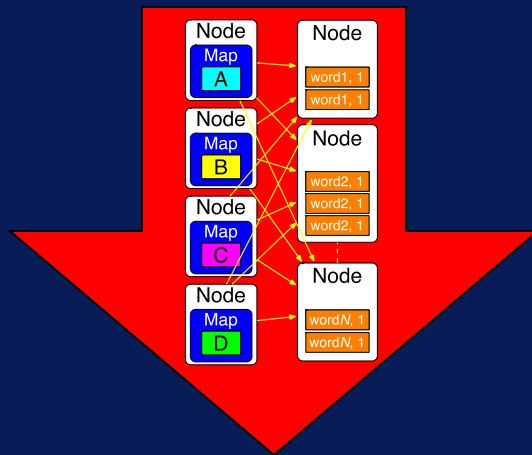
Map

Shuffle
and Sort

Reduce



Parallelization
over the input

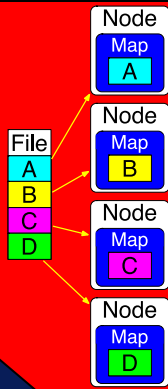


Parallelization
data sorting

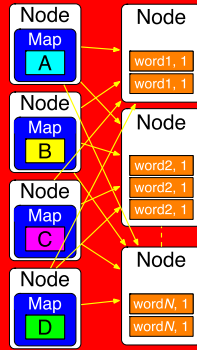
Map

Shuffle
and Sort

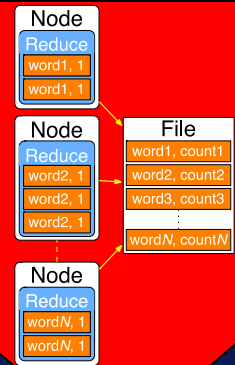
Reduce



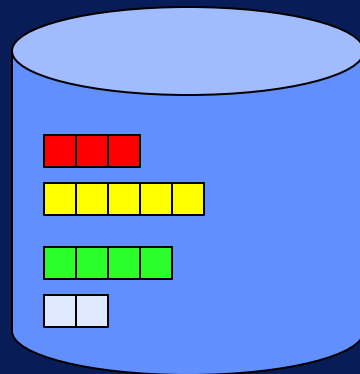
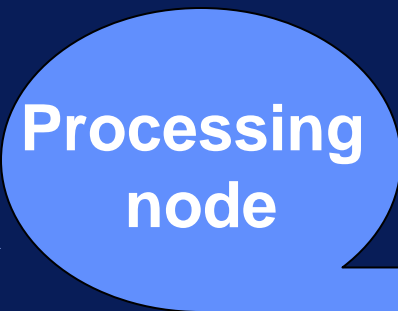
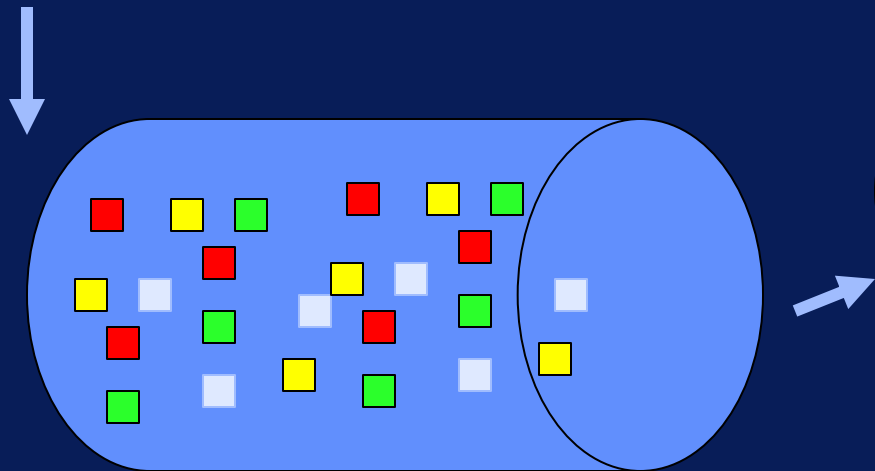
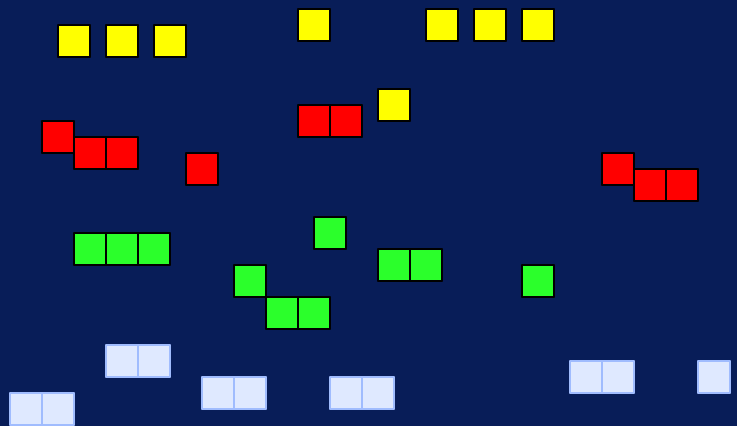
Parallelization
over the input

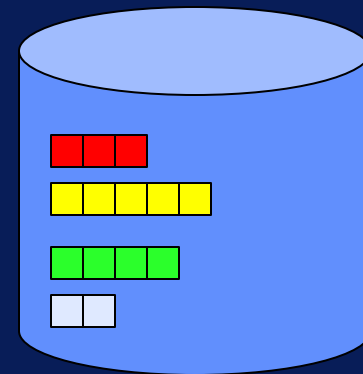
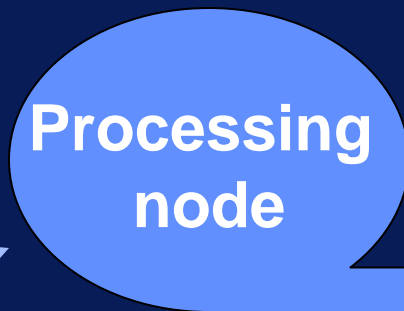
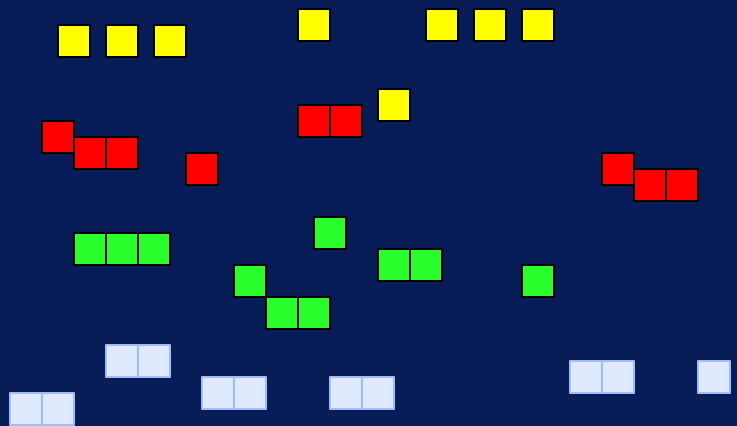


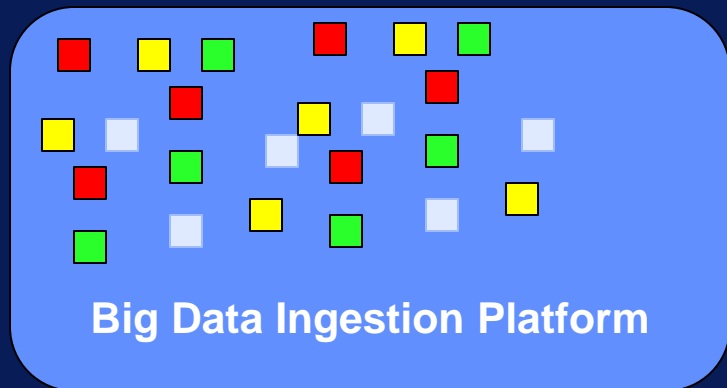
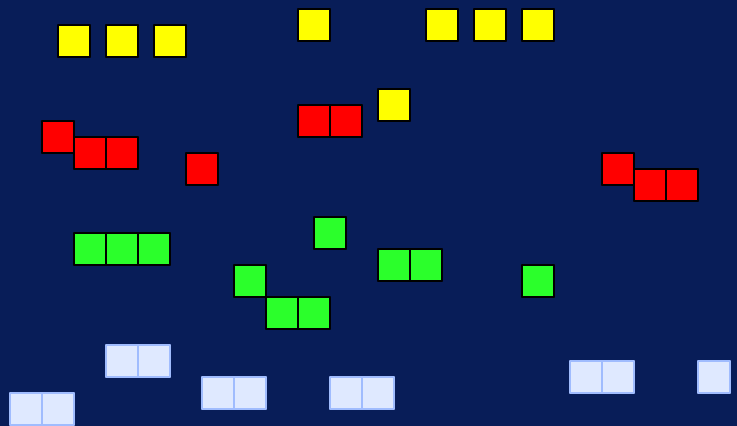
Parallelization over
intermediate data



Parallelization
over data groups







**Streaming Data
Processing Platform**

