

How MapReduce Works

School of Information Studies
Syracuse University

MapReduce

A programming model for large-scale distributed data processing.

Foundations in functional programming, LISP.

- Map \rightarrow Apply a transformation to a dataset.
- Shuffle → Transfer output from mapper to reducer nodes.
- Reduce \rightarrow Aggregate items into a single result.
- Combine → Output of reducer nodes into single output.

Hadoop 1.0 had only MapReduce.

Hadoop 2.0, MapReduce programs use YARN.



WordCount in MapReduce



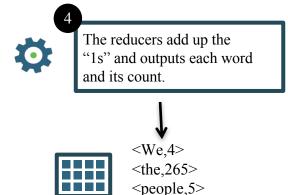


HDFS

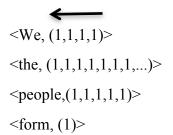
HDFS

The mappers read the file's blocks from HDFS line by line.

We the people, in order to form a...



<form, 1>



The lines of text are split into words and output to the reducers.



The shuffle/sort phase combines pairs with the same key.



MapReduce: Total Orders by State

