Lecture 15.1
Hadoop!

EN 600.320/420

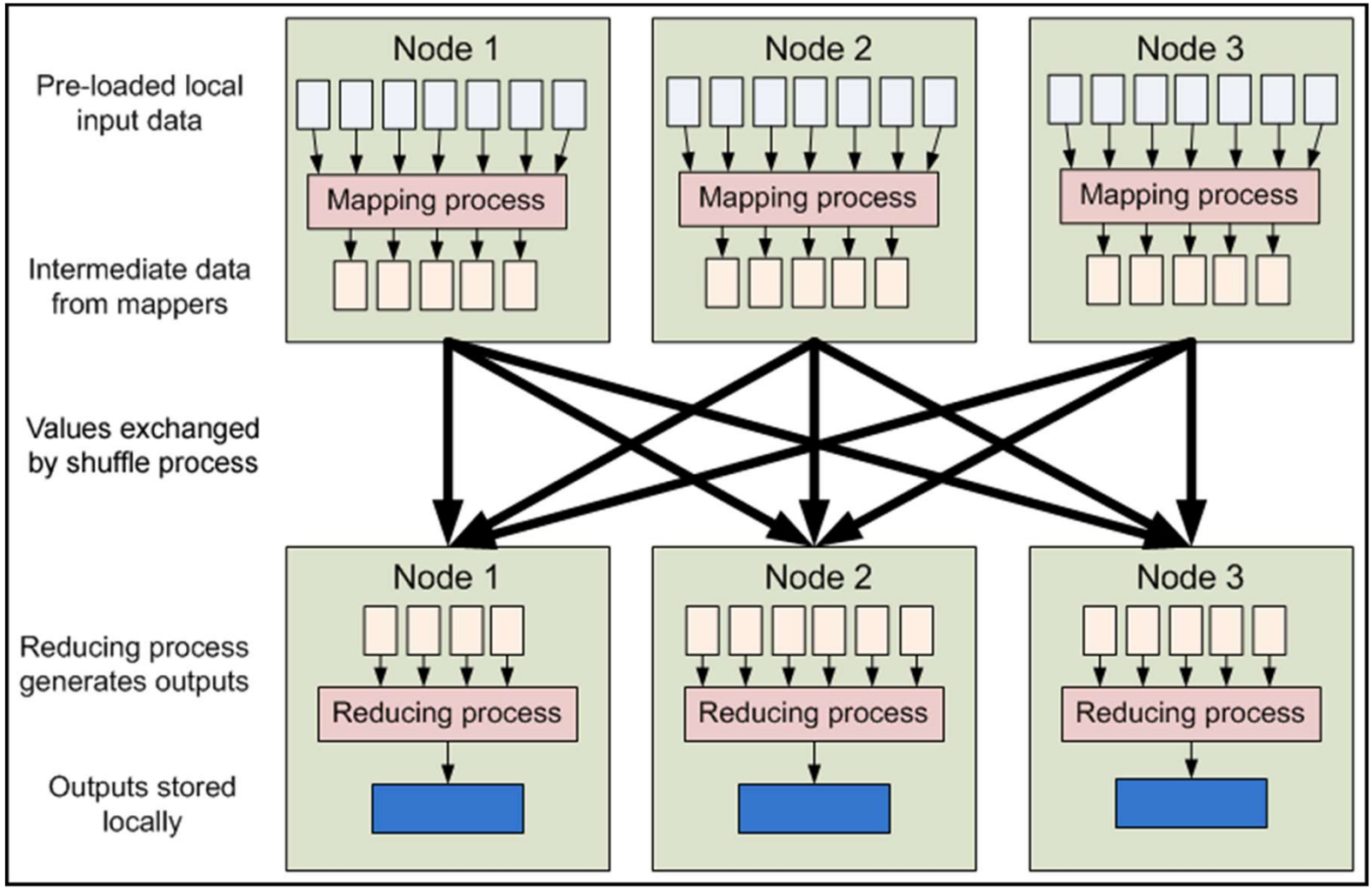
Instructor: Randal Burns

4 April 2018

Department of Computer Science, *Johns Hopkins University*

Hadoop!

- Open source reimplementation of Map/Reduce



Tutorial

- Official
  - http://hadoop.apache.org/docs/r2.7.3/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
- If you like Eclipse/VM/Windows etc.
  - This is older, I'm not sure if it's still relevant
  - http://developer.yahoo.com/hadoop/tutorial/module3.html
- Bad news
  - Long
- You all should do the tutorial

Define map() and reduce()

- Java package org.apache.hadoop.io.mapreduce
  - Mapper (interface for mapper function)
  - Reducer (interface for reducer function)

- Paradigm
  - Implement the interfaces
  - Called by the Hadoop! runtime

Mapper Code

- See example changes every year

Observations

- Types:
  - Hadoop! wraps all types that are to be input/output
  - Must use IntWritable() for output, not int
  - Text is a Hadoop! class for strings

- Collector paradigm for I/O:
  - Output of the map, input to shuffle
  - Output of the reducer, into partitions

Reducer Code

- See example changes every year

Hadoop! has Schemas

- For type checking
- Mapper – specifies
  - Input key and value type
  - Output key and value type
- Reducer – specifies same
  - DANGER: reduces is not a transformation, so you cannot change the key type
  - Doing so will break the system (silently?? Used to be)
  - Seems like a poor design

Configure and Launch (Driver)

- Configure a job: a class with "public static void main(..)" entry point to be run by Hadoop!
- Assign, output types (seems redundant)
- Assign input and output directories
- Configure
  - Mapper, reducer, **combiner**
- Create a client to manipulate the running job
- LAUNCH!

Driver Code

- See example (changes every year)