Hadoop MapReduce "from scratch"

2019 Nov.

Presentation of the Project

Yixiao FEI

2019 Nov.

目录 Contents

目录 Contents

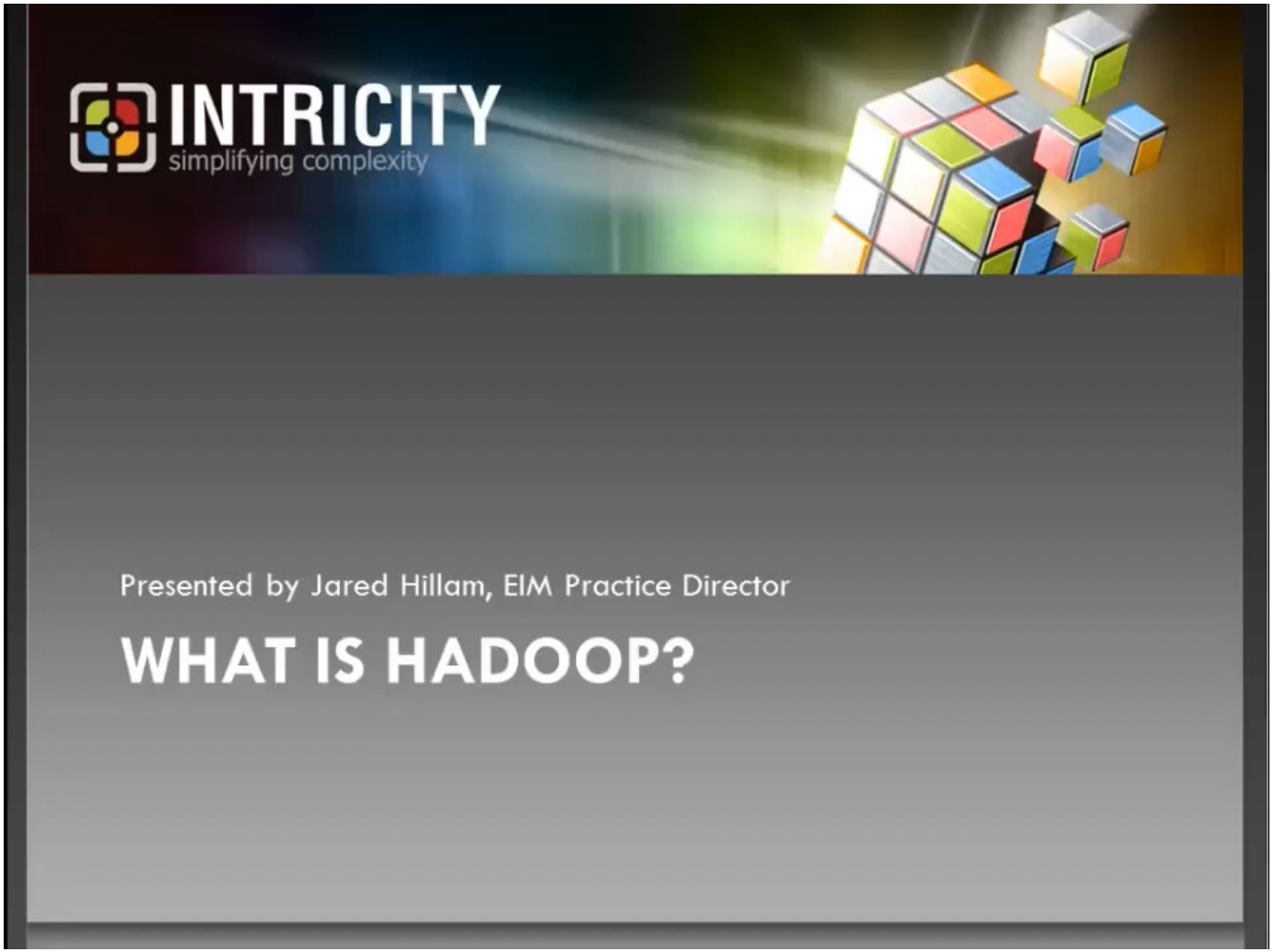
Big Data


BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
MAY
EXAMPLES
SHARED
DEFINITION
CURRENT
MOVING
EVERY
RECONSIDER
WITHIN
ZETTABYTES
PRACTITIONERS
CAPTURE
THOUGHT
INTERNET
BUSINESS
DISTRIBUTED
SETS
GENOMICS
DESCRIBING
RADIO-FREQUENCY
COMPLEXITY
MANAGEMENT
TERABYTES
CASE
ABILITY
PETABYTES
INCLUDE
ELAPSED
SYSTEMS
TOLERABLE
STORAGE
SIZE
MPP
PARALLEL
SAN
QUALITIES
SINCE
MASSIVELY
GROW
SEARCH
DIFFICULTY
GARTNER
WORKING
OPPORTUNITIES
AMOUNT
CURRENTLY
SOLID
TYPES
RECORDS
NETWORKS
DATABASES
HUNDREDS
DESKTOP
CAPACITY
BURIED
WORLD'S
NOW
BIOLOGICAL
PROCESSING
UBIQUITOUS
ANALYTICS
TENS
COMPLEX
CITATION
INDEXING
LARGER
CONTINUES
SOCIAL
USE
SET
ORGANIZATIONS
RELATIONAL
CONNECTOMICS

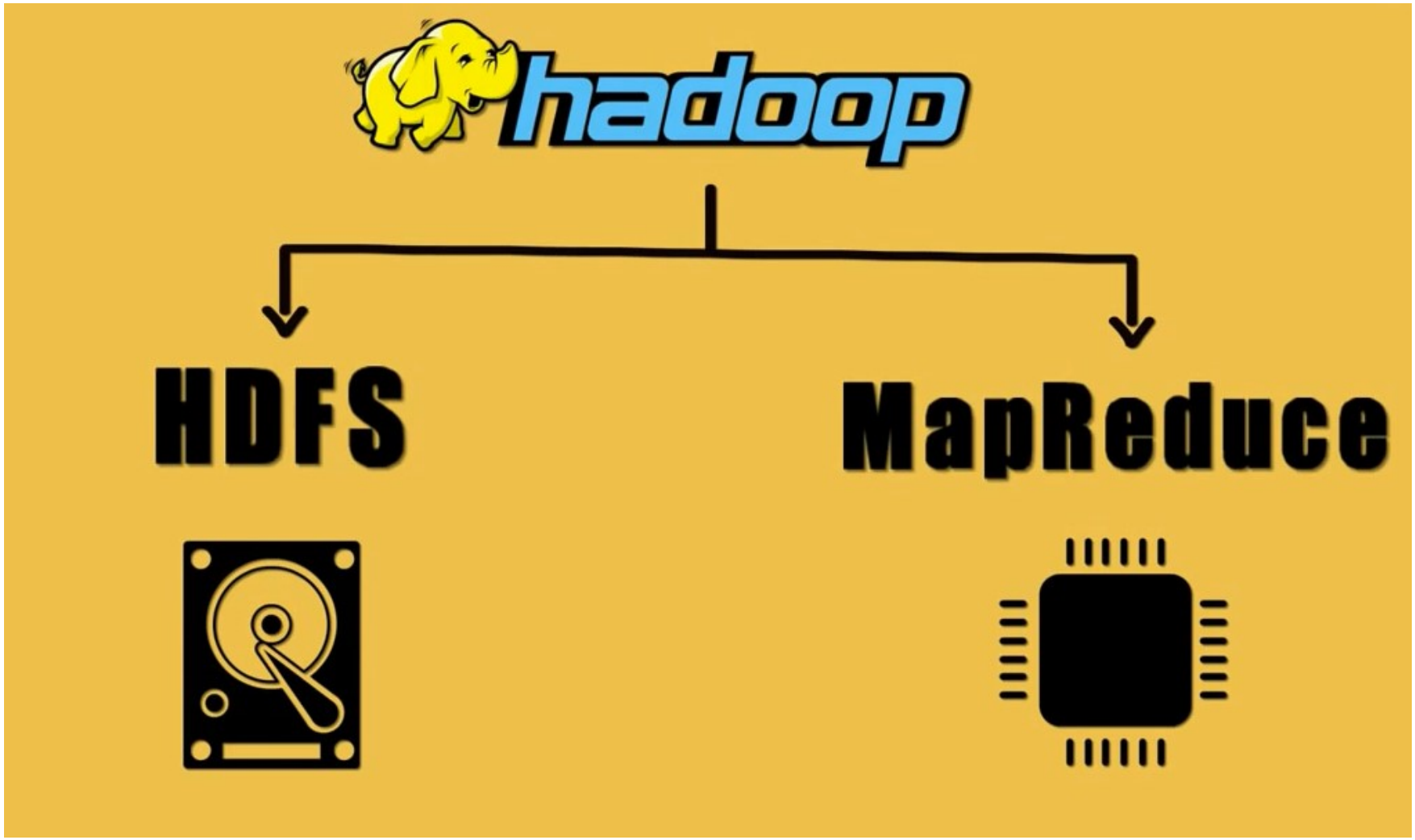
Hadoop

- Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

Hadoop


INTRICITY
simplifying complexity
Presented by Jared Hillam, EIM Practice Director
WHAT IS HADOOP?

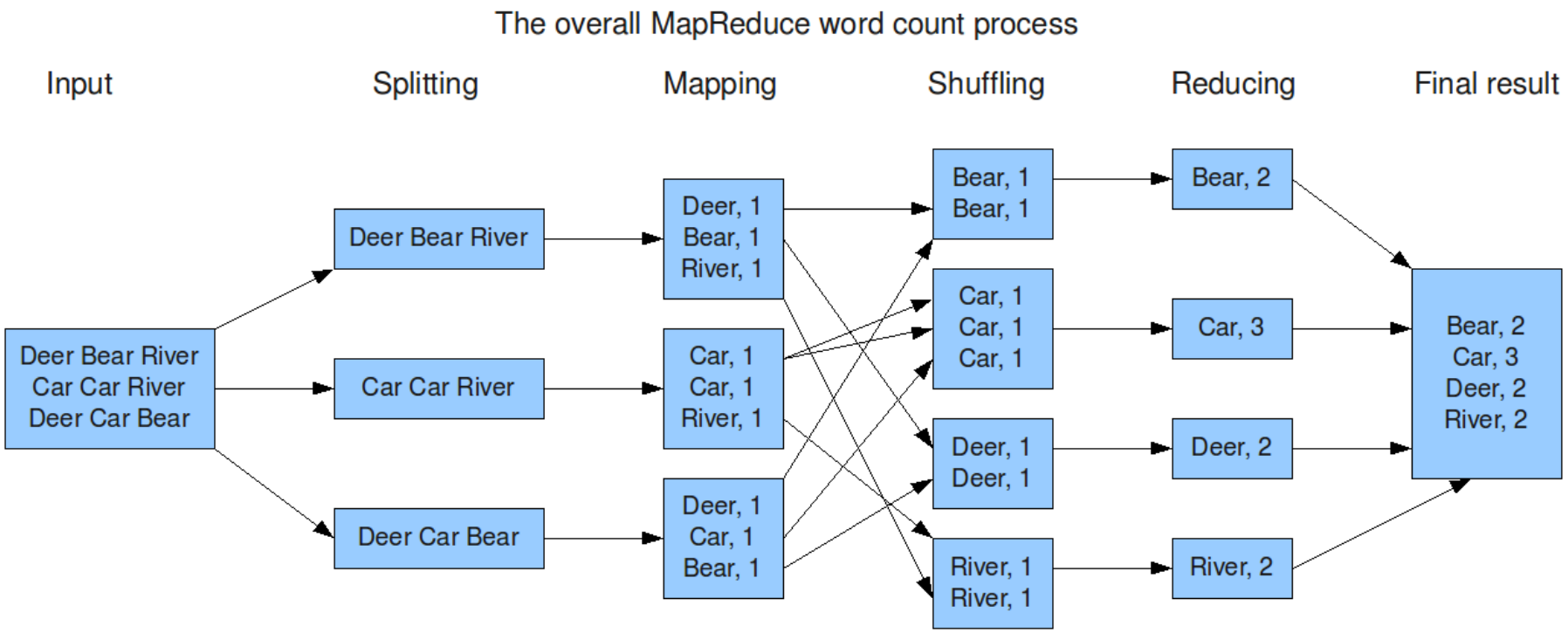
Hadoop



目录 Contents

Hadoop Cluster

MapReduce


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

目录 Contents

Word Count

- Trivial Method
- MapReduce
- Experiments

Important Notes

- Work: 3 people/group
- Deadline: 12:00, 24 Dec. 2019
- Filename: MP-GroupNumber.zip (eg. IM-Group1.zip) including:
    1) source code
    2) compiled runnable Jar and a Readme file
    3) report (your names and IDs)
- TA’s contact : destiny_fyx@outlook.com

Evaluation

- Functionality (50%), Report (20%), Presentation(30%)
- Delay of submission will cost 30% of points per day
- The report includes at least:
  - ✓ System diagram
  - ✓ Function description
  - ✓ How to build/run (some demo)
  - ✓ Experiment (comparison and reflection)
  - ✓ Good written ability

No Tolerance for Plagiarism

Things that might Help

- Eclipse
- Maven
- Docker
- Linux Commands
- Git

Thanks !