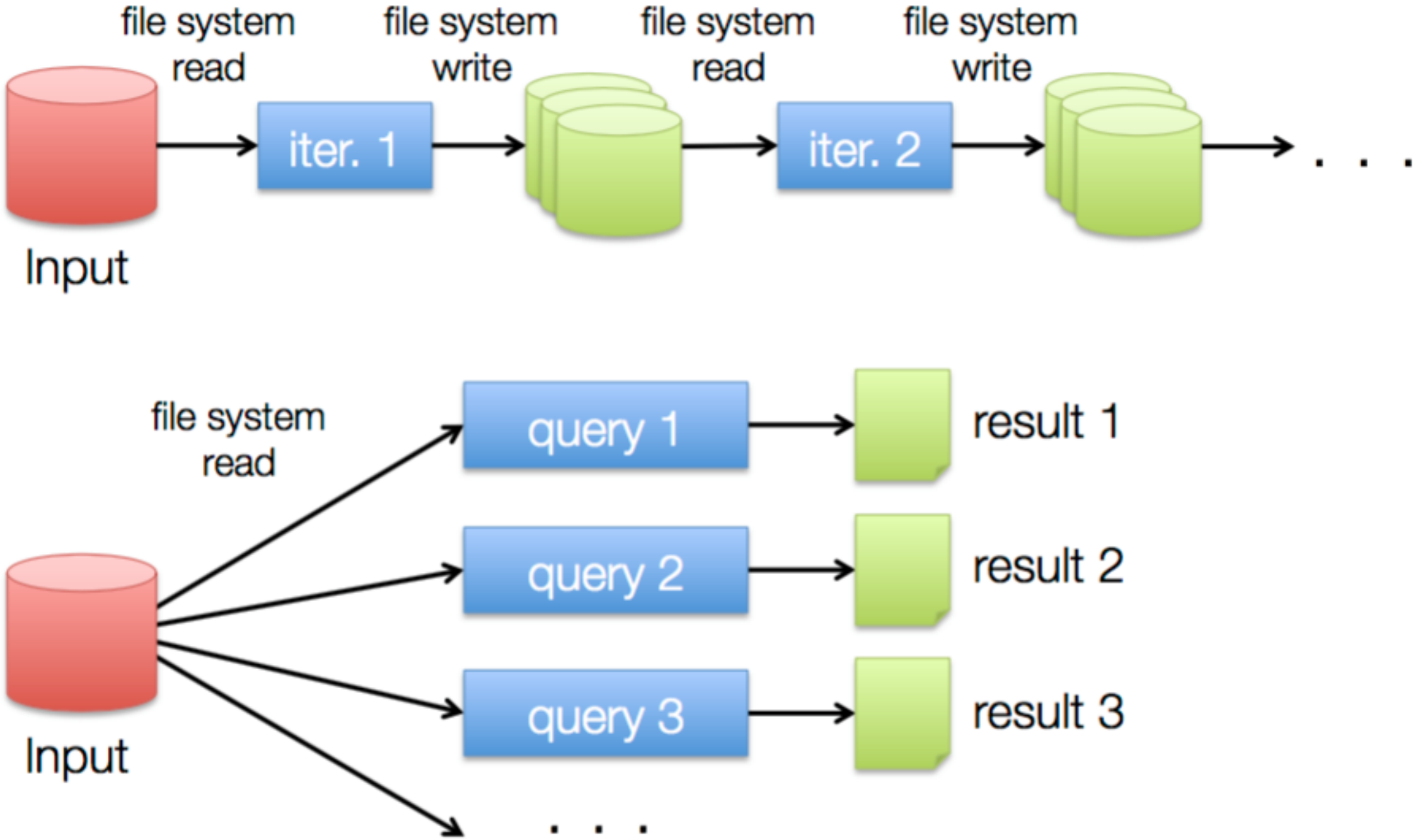
Distributed Systems

Assignment 3

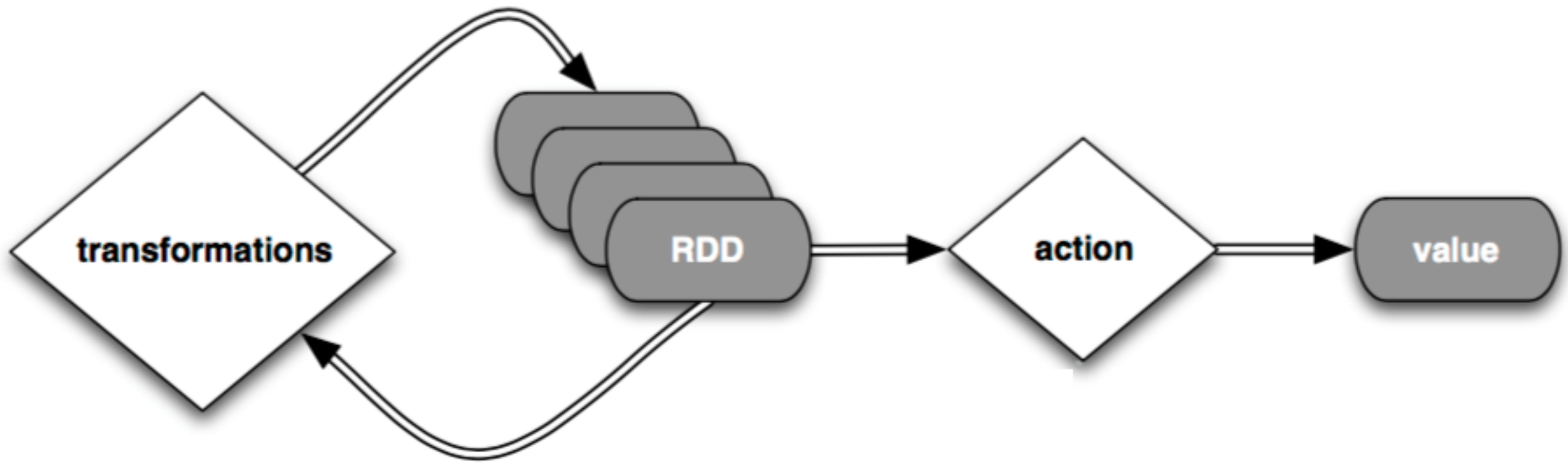
Shen Gao
Bibek Paudel

Apache Spark



Commonly spend 90% of time doing I/O

Picture courtesy of databricks



Picture courtesy of databricks

Assignment 3 tasks

Comparison between Hadoop and Spark.

- Task: Get the Top-10 words (same with task 2.1 of map/reduce programming)
- NO DATA CLEANING.
- 99% done for you in the WordCountExample.java

Log File Analysis

<Events **startTimestamp="1447245354293”** …..>

<Command __id="311" **_type="InsertStringCommand"** repeat="2" timestamp="1186109" timestamp2="1186542">

</Command>

<DocumentChange __id=“308” …..> is NOT a Command!

- We only consider COMMAND as text marked by <Command> <Command/></Command> tags; we do not consider DocumentChange.
- The first line gives the beginning time of the log
- The timestamp field gives the time relative to the beginning timestamp.
- The “_type” field gives the type of a command

Log File Analysis

- Calculating the average time between two commands.
- Tips: 1. Count the total number of commands. 2. Get the beginning and ending time.

Log File Analysis

- Counting the number of commands in every 15 minutes interval.
- Tips: reduce by key, and use the timestamp as the key.
- Output:

0 45

1 32

2 95

Log File Analysis

- Counting and ordering the frequency of all distinct commands in descending order.
- If a command is an eclipse command, we need to count the frequency for each distinct type of eclipse commands.

PasteCommand 98

CopyCommand 87

EclipseCommand:eventLogger.styledTextCommand.DELETE_PREVIOUS 76

FileOpenCommand 65