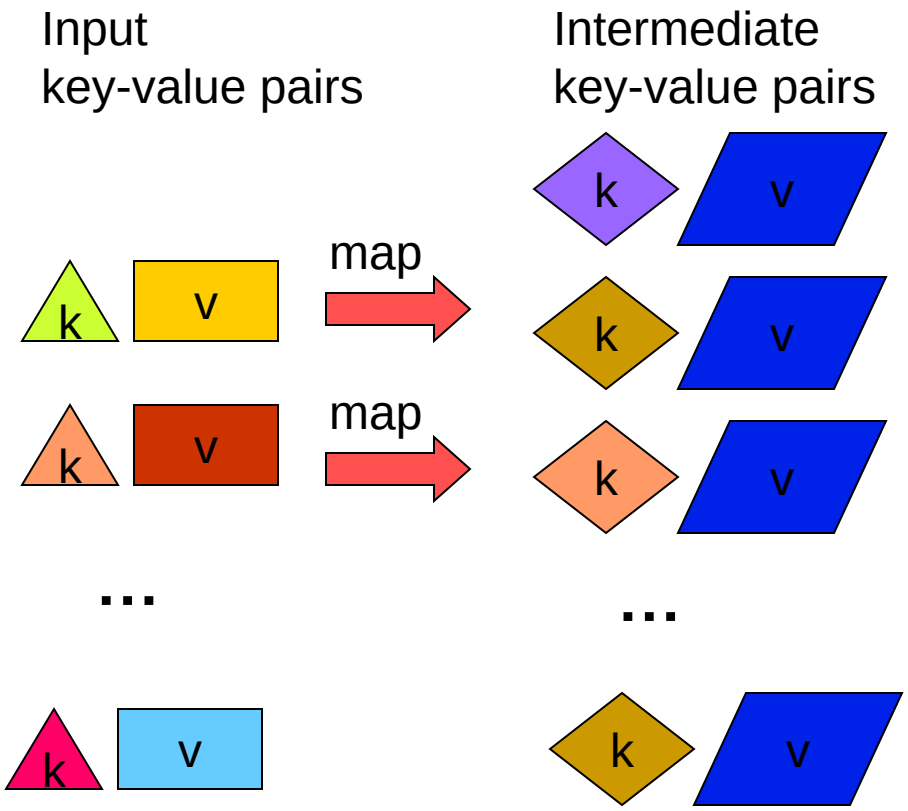
Logistics

- Lab 4 keys and grades will be release today/tomorrow
- Lab 5 material is out, due Thursday 11:59pm
  - JAVA + AWS/EMR
- Lecture 7 coming Monday
- Lab 5 in class coming Wed.
  - Extra time: AWS, Hadoop/EMR setup
- NIST DSE Introduction + QA next Friday
- No office hour next Wed.

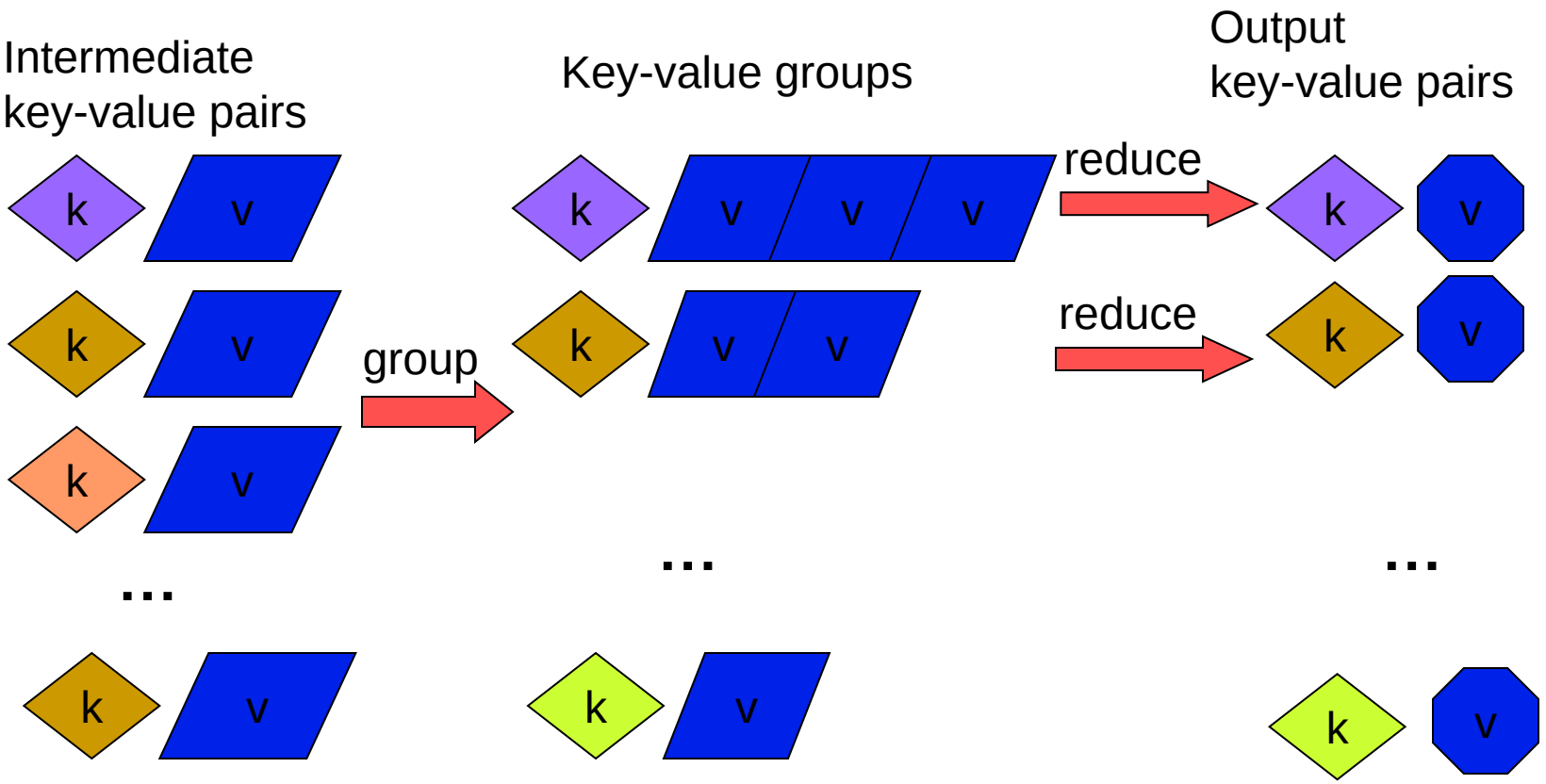
Review

- Large-scale data storage and processing
- Distributed File System
- Map-Reduce programming model for parallel/distributed computing

MapReduce: The Map Step



MapReduce: The Group and Reduce Step

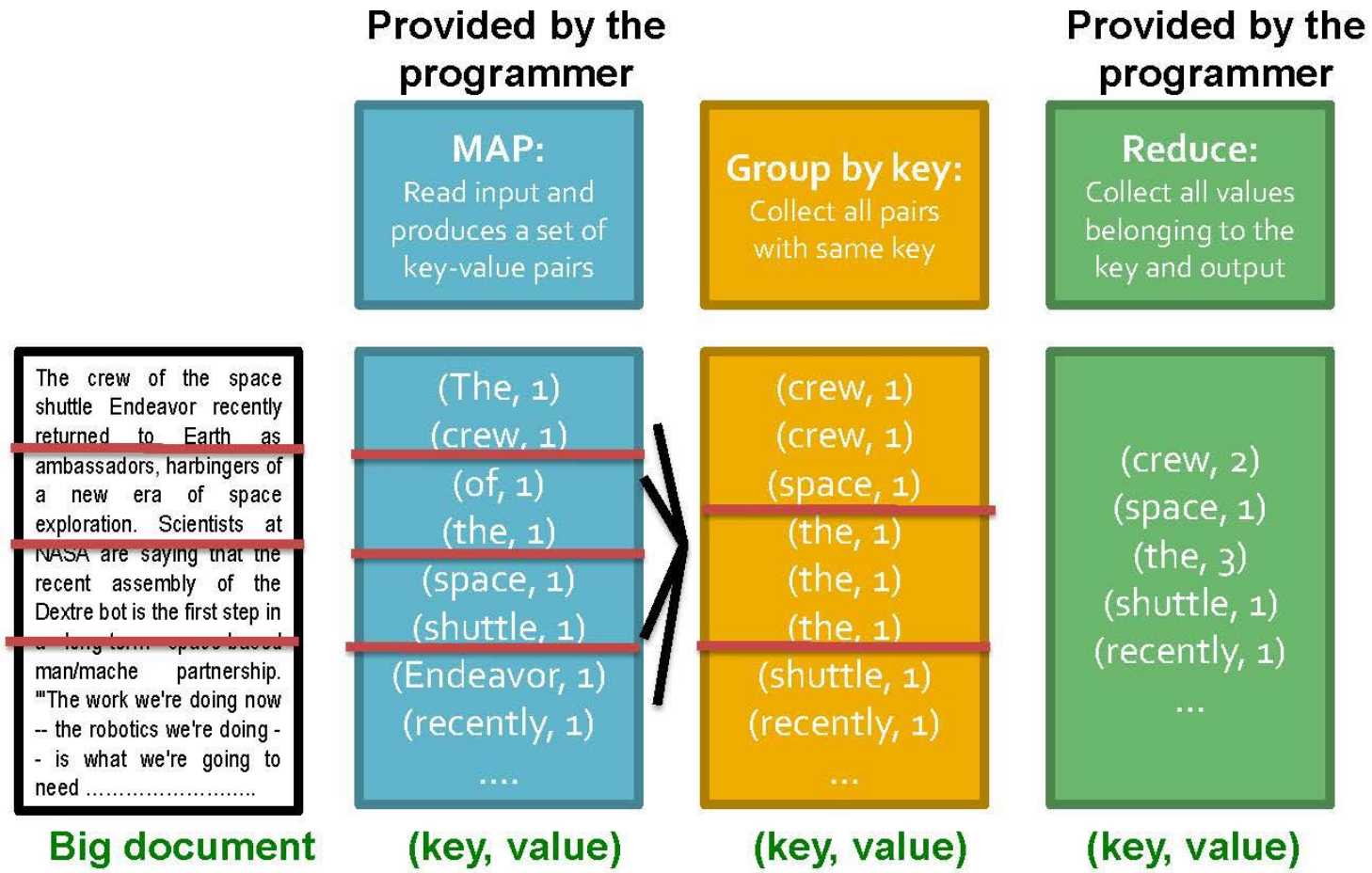


Word Count using MapReduce

```
map(key, value):
// key: document name; value: text of document
    for each word w in value:
        emit(w, 1)


reduce(key, values):
// key: a word; value: an iterator over counts
        result = 0
        for each count v in values:
            result += v
        emit(result)
```

MapReduce: Word Count

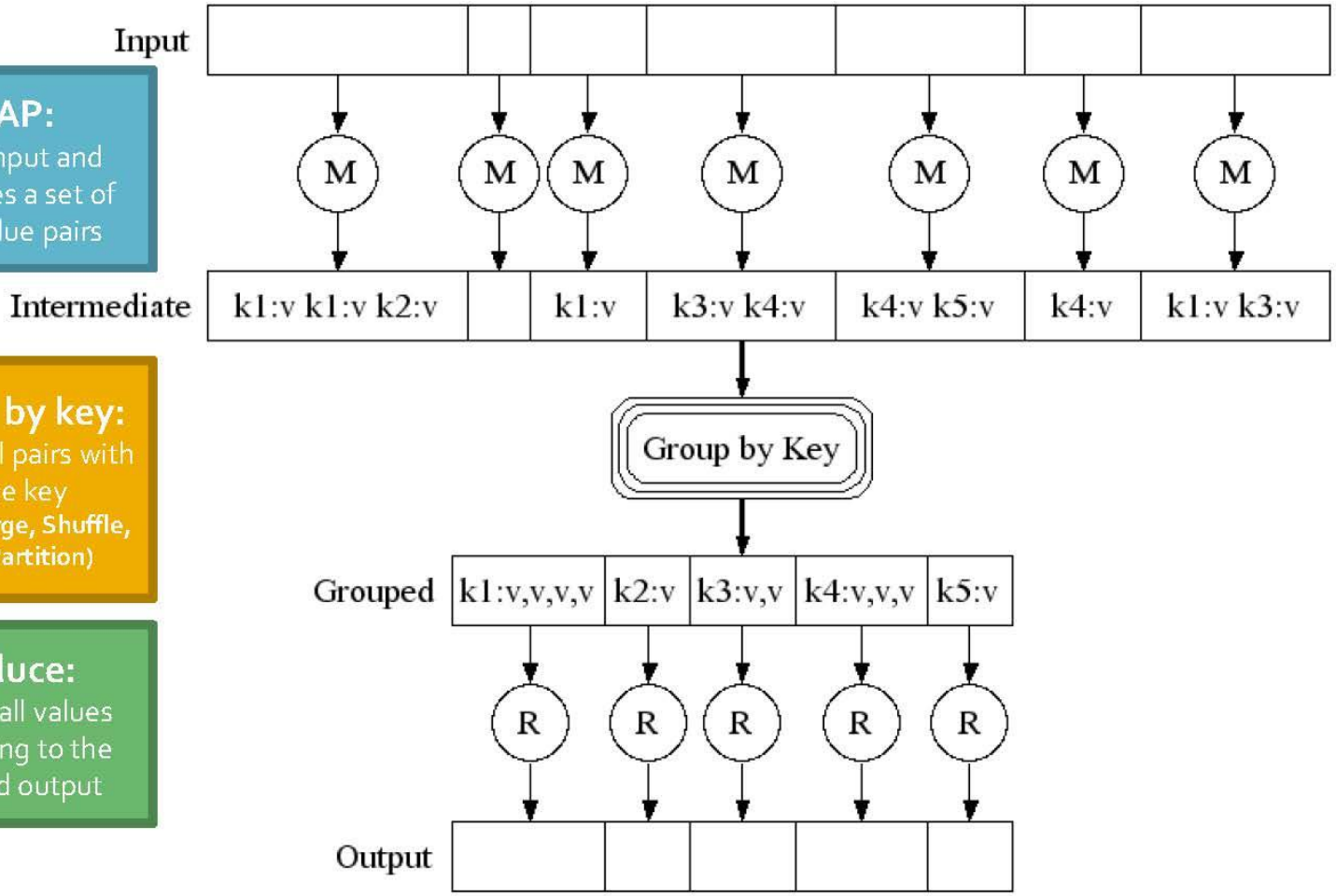


MapReduce Execution

MAP:
Read input and produces a set of key-value pairs

Group by key:
Collect all pairs with same key
(Hash merge, Shuffle, Sort, Partition)

Reduce:
Collect all values belonging to the key and output



Keys to MapReduce Programming

1. Define Input K-V pairs
2. Define Result K-V pairs
3. Specify logic in Mapper
4. Define Intermediate K-V pairs
5. Specify logic in Reducer

MANY OTHER ANALYTICS AND ALGORITHMS CAN BE PARALLELIZED USING MAPREDUCE

Exercise 1: Host size

- Suppose we have a large web corpus
- Let's look at the metadata file
  - Lines of the form (URL, size, date, …)
- For each host, find the total number of bytes
  - i.e., the sum of the page sizes for all URLs from that host

Example 1: Host size (cont.)

- TextInputFormat: (position, "URL, size, data,…")
- KeyValueTextInputFomrat: (URL, "size, data,…")

- Mapper: (position, "URL, size, data,…") -> (hostname, size)
- Mapper: (URL, "size, data,…") -> (hostname, size)

- Reducer: (hostname, list (size)) -> (hostname, totalsize)

Exercise 2: Distributed Grep

- Find all occurrences of a given pattern in a very large set of webpages

- InputFormat <K,V>
  - webpages → (url+offset, single line)
- Result <K,V>
  - (line, N/A) pairs, where line matching the pattern

Exercise 2: Distributed Grep (cont.)

- map(key=url+offset, val=line):
  - If contents matches regexp
  - emit (line, “1”)
- reduce(key=line, values=uniq_counts):
  - Don’t do anything; just emit line

Exercise 3: Graph reversal

- Given a directed graph as an adjacency list:

src1: dest11, dest12, …

src2: dest21, dest22, …

- Construct the graph in which all the links are reversed

Exercise 3: Graph reversal (cont.)

- KayValueTextInputFormat
- Map
  - For each URL linking to target, …
  - Output <target, source> pairs
- Reduce
  - Concatenate list of all source URLs
  - Outputs: <target, ***list*** (source)> pairs

Implementations

- Google
  - Not available outside Google
- Hadoop
  - An open-source implementation in Java
  - Uses HDFS for stable storage
  - Download: http://lucene.apache.org/hadoop/

Summary

- Parallel & Distributed Computing Systems
- Distributed File System
- MapReduce: Parallel programming model over Distributed Systems
- Parallel Analytics using M/R
  - Data mining
  - Index building
  - Aggregation
  - Log Analysis

Midterm

- Lecture 1-7 material
- Assigned reading are supplementary for understanding content in Lecture 1-7
- Main Topics
  - Data types, models and manipulation
    - Unstructured, semi-structured, structured
  - Hypothesis testing
  - Classification and Regression
  - Map-Reduce Programming Model

Midterm Review Example Questions

[Hypothesis Testing] May 2010 Gallup poll of 1029 US adults. When asked if they view divorce as "morally acceptable", 71% of the men and 67% of the women in the sample responded yes. Please describe the steps to conduct a hypothesis testing analysis to indicate if there is a significant difference between men and women in how they view divorce.

Hint: you need to describe H_0 H_A Test statistic, sampling distribution, p-value, significance level, Type I/II error and final conclusion of this hypothesis test

Midterm Review Example Questions I

[Hypothesis Testing] May 2010 Gallup poll of 1029 US adults. When asked if they view divorce as "morally acceptable", 71% of the men and 67% of the women in the sample responded yes. Please describe the steps to conduct a hypothesis testing analysis to indicate if there is a significant difference between men and women in how they view divorce.

Hint: you need to describe H_0 H_A Test statistic, sampling distribution, p-value, significance level, Type I/II error and final conclusion of this hypothesis test

Midterm Review Example Questions II

[Map-Reduce] Given a directed graph as an adjacency list of outlinks – a src to all dest:

src1: dest11, dest12, …

src2: dest21, dest22, …

…

Please use Map-Reduce programming model to construct a same graph as an adjacency list of inlinks from a dest node to all src nodes

Midterm Review Example Questions III

[Classification and Regression] A set of reasonably clean sample records was extracted by Barry Becker from the 1994 Census database. Please describe the steps and techniques to train and test a prediction model as whether a person makes over 50K a year.

Hint: steps include feature extraction/selection, model selection, training, testing, evaluation, analysis/iterate…

The list of attributes are:

Class: >50K, <=50K.
age: continuous.
workclass: Private, Self-emp-not-inc, …

Midterm Review Example Questions III (cont.)

education: Bachelors, Some-college, …

education-num: continuous.
marital-status: Married-civ-spouse, Divorced, …
occupation: Tech-support, Craft-repair, …
relationship: Wife, Husband, Not-in-family, …
race: White, Asian-Pac-Islander, …
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, …

Midterm Review Example Questions IV

[Data types, models and manipulation]

- Please given an example data model and data instance of semi-structured data that can be represented as tree/hierarchically.
- What are the main components of a distributed file system?
- Please give an example of a pivot operation over a simple data cube.
- We have discussed in class different abstractions that is provided by different programming models. Give two examples and describe the utility of the abstraction.