CET4001B Big Data Technologies

School of Computer Engineering and Technology

Map-Reduce using Java on Hadoop

LABORATORY ASSIGNMENT NO: 05

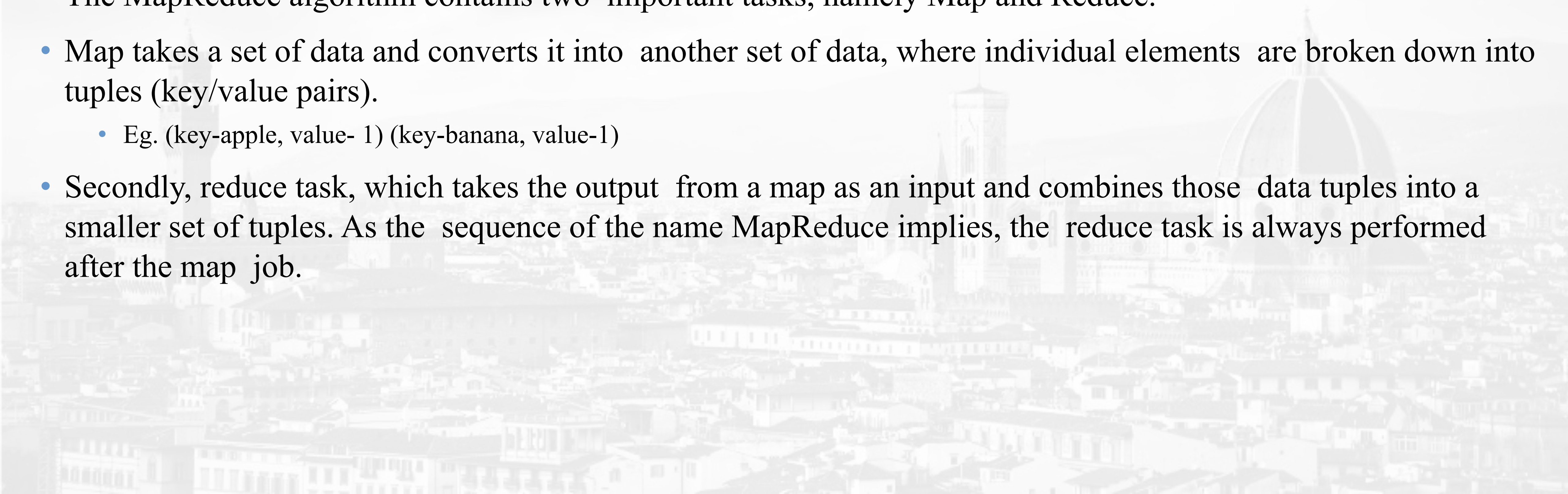
Installation Details

- Installation Details :
  - Hadoop Version : 3.2.4
  - OS: Windows 10 (64 bit)/Linux
  - Mode : Pseudo-distributed(Single node cluster)
- Prerequisites:
  - Java Version 8 SDK (Compatible with Hadoop)
  - Eclipse IDE : for Map-reduce program after installing Hadoop
  - WinZIP/WinRAR or Online tool

Example: git bash for Windows: Only if WinZIP/WinRAR is unable to extract for Extracting Hadoop tar.gz file

Map-Reduce

- MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.
- MapReduce is a processing technique and a program model for distributed computing based on java.
- The MapReduce algorithm contains two important tasks, namely Map and Reduce.
- Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
    - Eg. (key-apple, value- 1) (key-banana, value-1)
- Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.



MapReduce Algorithm

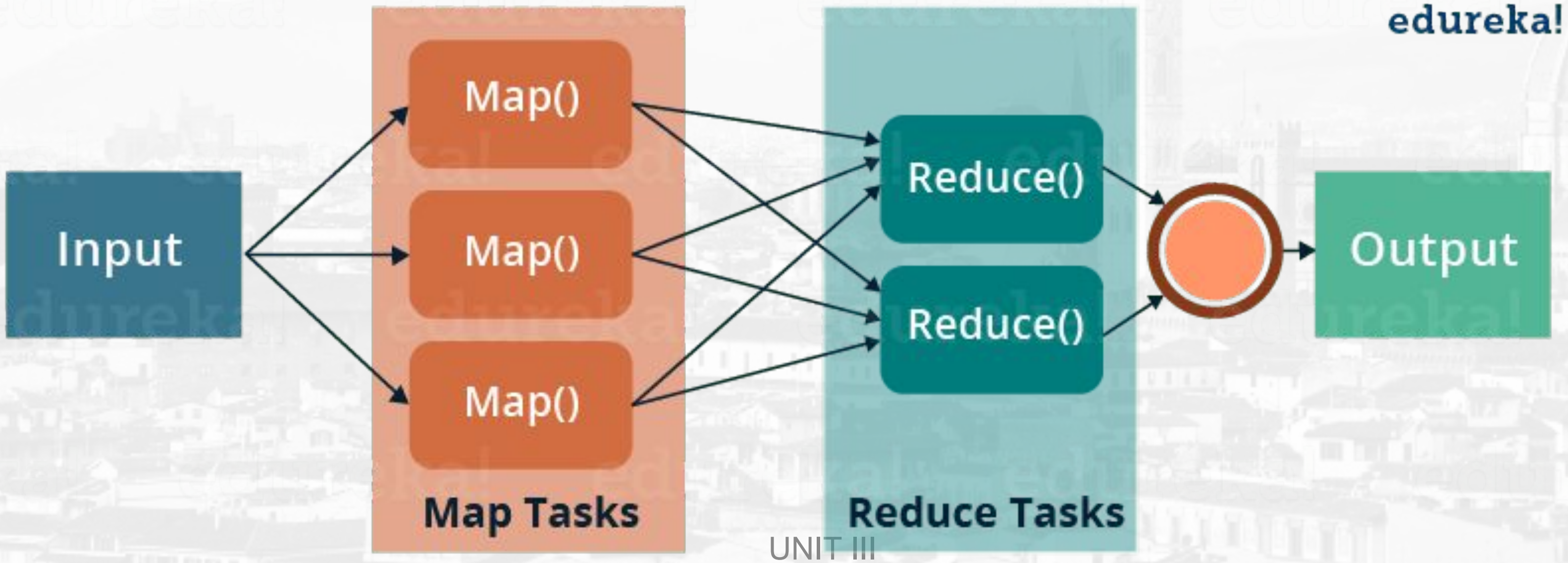
- MapReduce program executes in four stages, namely map stage, reduce stage , shuffle and sort stage

- **Map stage:**
  - The map or mapper’s job is to process the input data.
  - Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS).
  - The input file is passed to the mapper function line by line.
  - The mapper processes the data and creates several small chunks of data.

- **Reduce stage:**
  - This stage is the combination of the Shuffle stage and the Reduce stage.
  - The Reducer’s job is to process the data that comes from the mapper.
  - After processing, it produces a new set of output, which will be stored in the HDFS

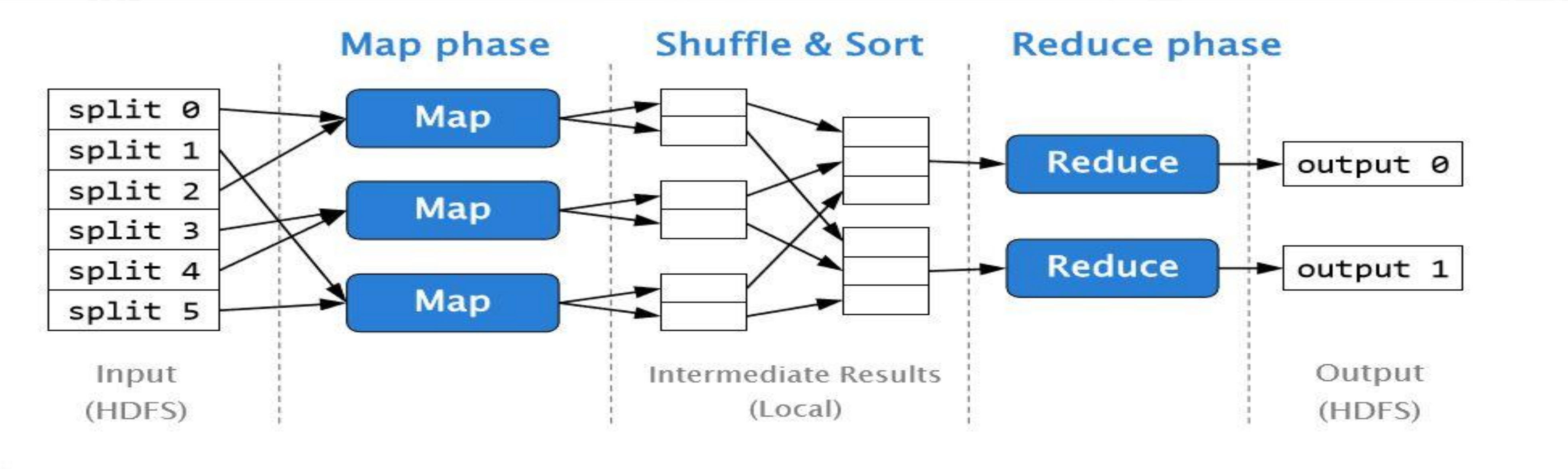
Map-Reduce Workflow

- **Map Phase**
  - Raw data read and converted to key/value pairs
  - Map() function applied to any pair
- **Shuffle and Sort Phase**
  - All key/value pairs are sorted and grouped by their keys
- **Reduce Phase**
  - All values with a the same key are processed by within the same reduce() function



Map-Reduce Programming Model

- Every MapReduce program must specify a **Mapper** and typically a **Reducer**
- The Mapper has a **map()** function that transforms input **(key, value)** pairs into any number of intermediate **(out_key, intermediate_value**) pairs
- The Reducer has a **reduce()** function that transforms intermediate **(out_key, list(intermediate_value))** aggregates into any number of output **(value')** pairs



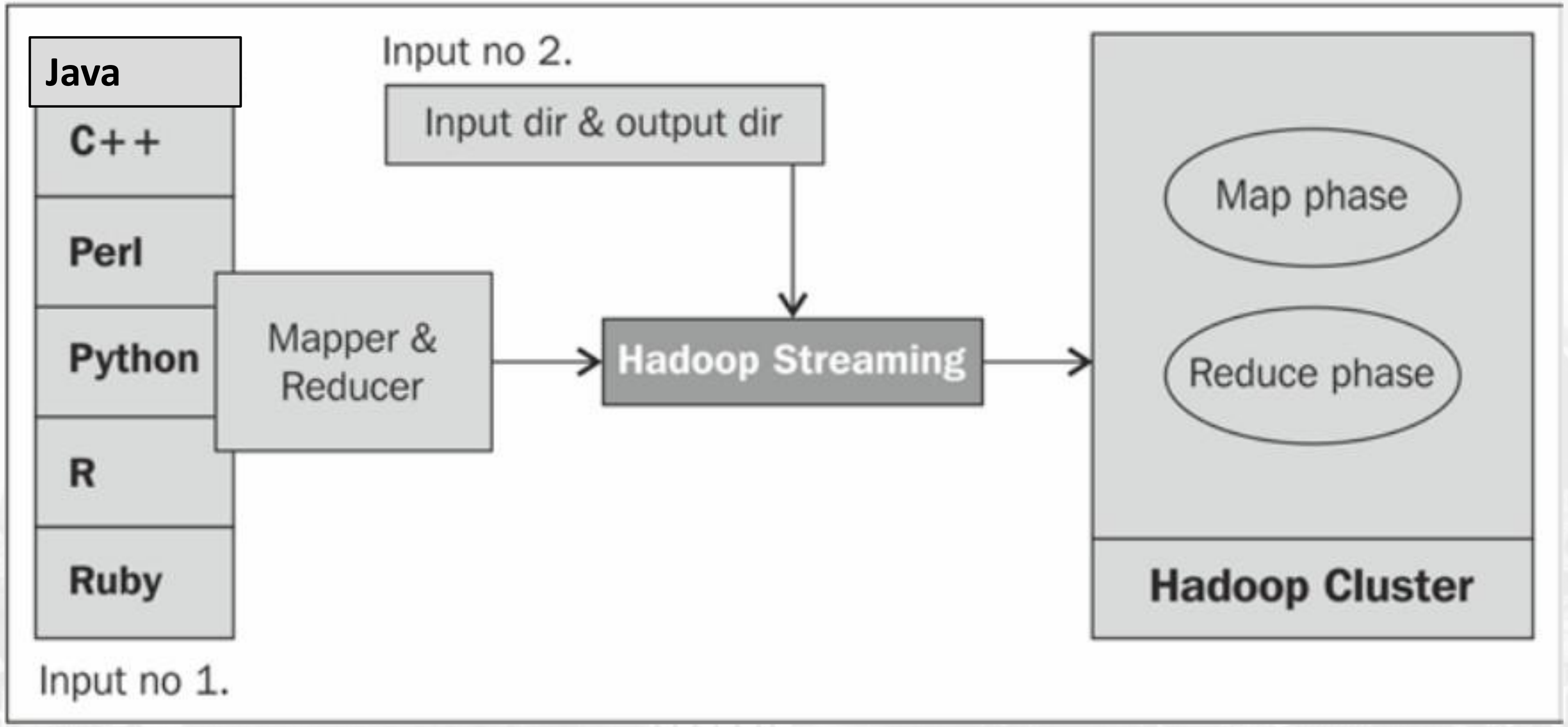
Hadoop Streaming

- Hadoop streaming is a Hadoop utility for running the Hadoop MapReduce job with executable scripts such as Mapper and Reducer.
- This is similar to the pipe operation in Linux.
- With this, the text input file is printed on stream (stdin), which is provided as an input to Mapper and the output (stdout) of Mapper is provided as an input to Reducer; finally, Reducer writes the output to the HDFS directory.

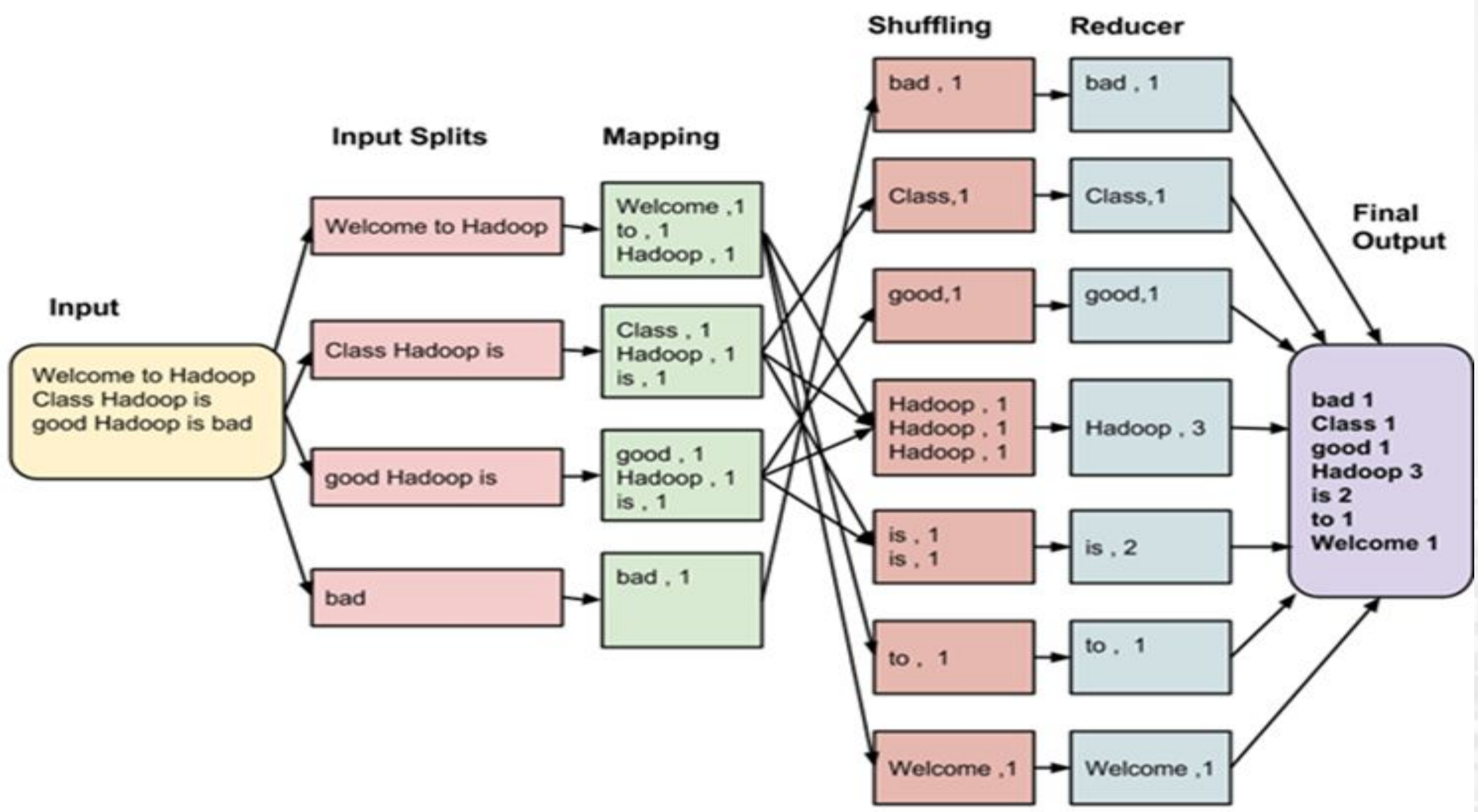
Hadoop Streaming

- The main advantage of the Hadoop streaming utility is that it allows Java as well as non-Java programmed MapReduce jobs to be executed over Hadoop clusters.
- Also, it takes care of the progress of running MapReduce jobs.
- The Hadoop streaming supports the Java, Perl, Python, PHP, R, and C++ programming languages.
- To run an application written in other programming languages, the developer just needs to translate the application logic into the Mapper and Reducer sections with the key and value output elements.

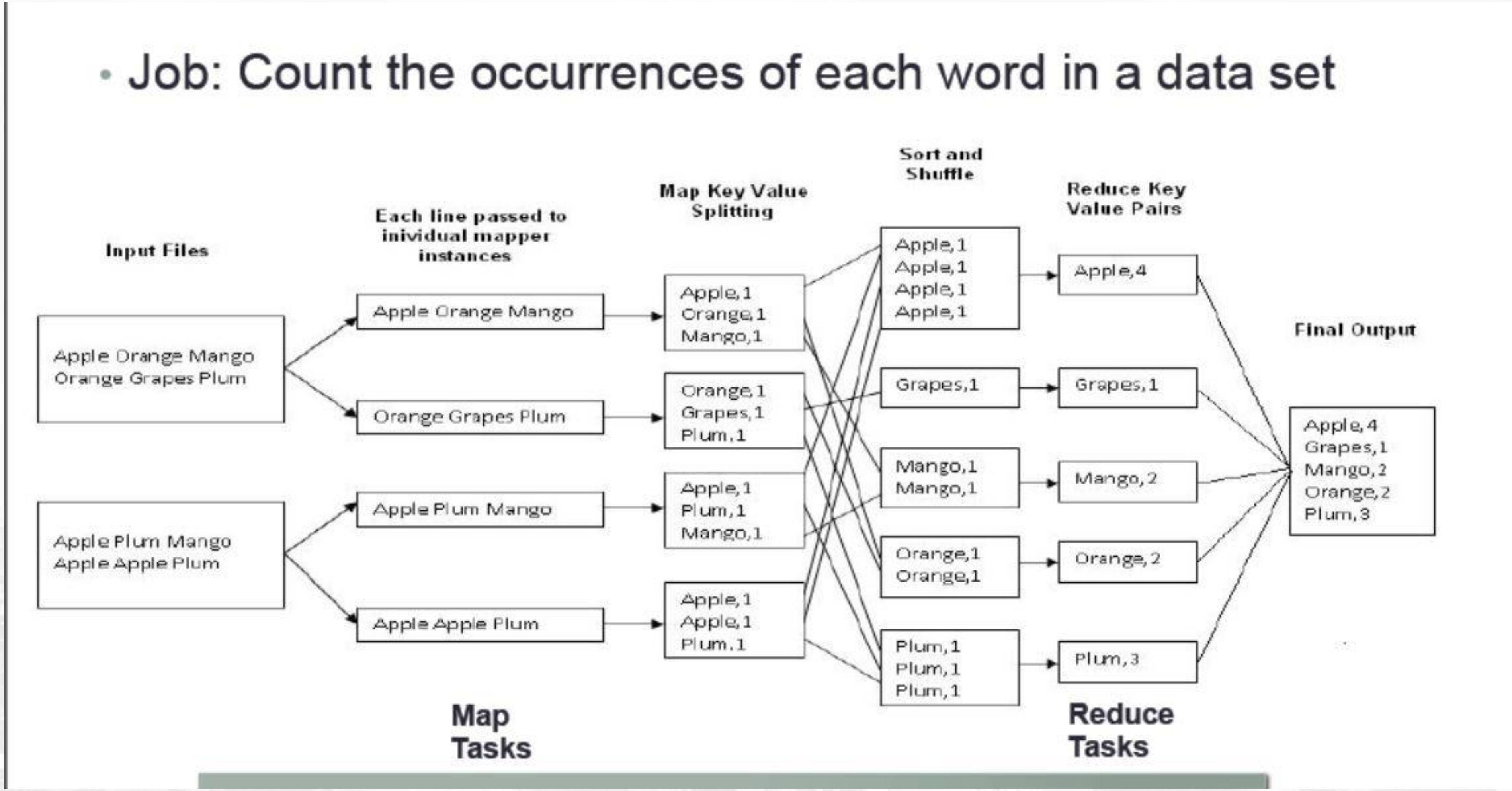
Hadoop Streaming



Example 1: Map-Reduce Programming Model



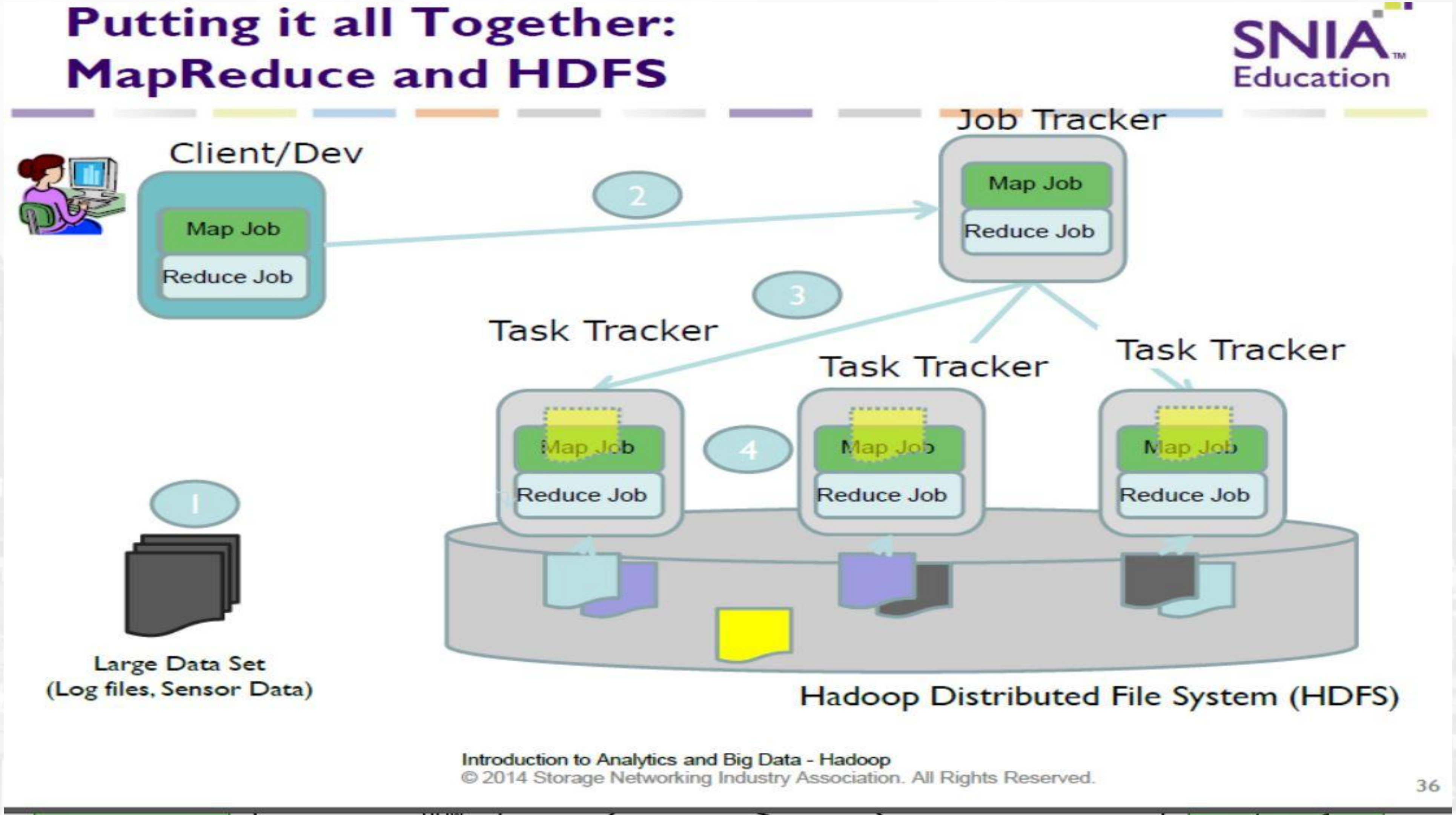
Example 2: Map-Reduce Execution Details



- Job: Count the occurrences of each word in a data set

Word Count Problem uisng MapReduce

Example 2: Map-Reduce Execution Details



Putting it all Together: MapReduce and HDFS

SNIA Education

Introduction to Analytics and Big Data - Hadoop

© 2014 Storage Networking Industry Association. All Rights Reserved.

Problem Statement

Perform Map-Reduce processing for Applications like Weather Monitoring/Finance/E-Commerce/Agriculture/Healthcare

BATCH-1 EXERCISE

Problem Statement :

Write a map-reduce program to display the product wise total sales.

Input File : 1

SalesJan2009.csv file which contains following data :

Transaction_date	Product	**Price**	**Payment_Type**	Name	City	State	Country	Account_Created	Last_Login	Latitude	Longitude
1/2/2009 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	1/2/2009 6:00	1/2/2009 6:08	51.5	-1.11667
1/2/2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United States	1/2/2009 4:42	1/2/2009 7:49	39.195	-94.6819
1/2/2009 13:08	Product1	1200	Mastercard	Federica e Andrea	Astoria	OR	United States	1/1/2009 16:21	1/3/2009 12:32	46.18806	-123.83
1/3/2009 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/2005 21:13	1/3/2009 14:22	-36.1333	144.75

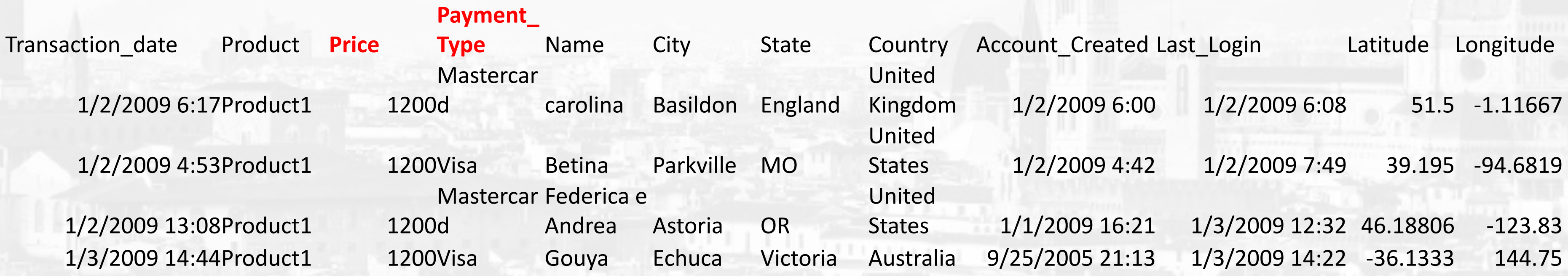
BATCH-2 EXERCISE

Problem Statement :

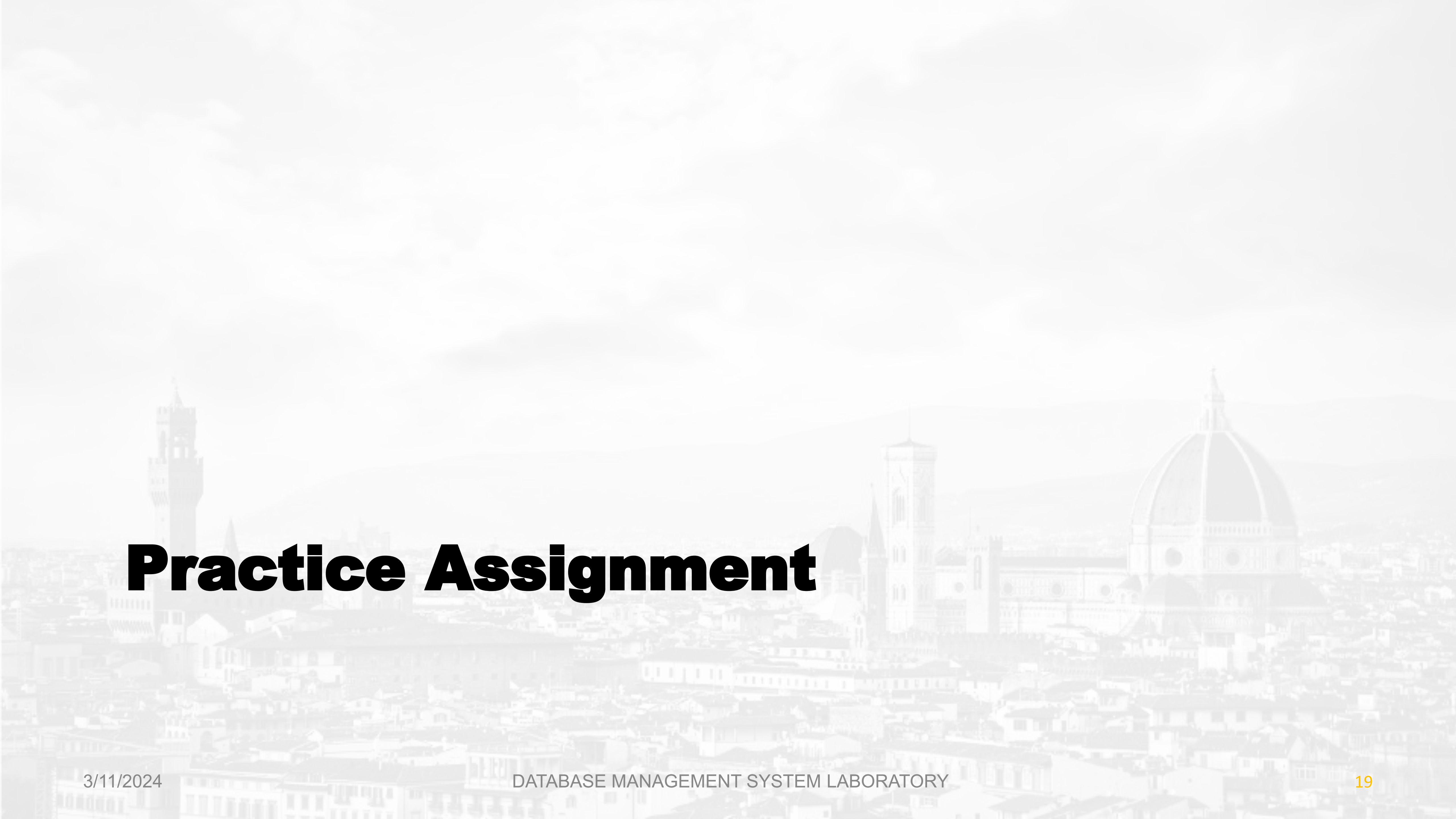
Write a map-reduce program to display the state wise total sales.

Input File : 1

SalesJan2009.csv file which contains following data :



Transaction_date	Product	**Price**	**Payment_Type**	Name	City	State	Country	Account_Created	Last_Login	Latitude	Longitude
1/2/2009 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	1/2/2009 6:00	1/2/2009 6:08	51.5	-1.11667
1/2/2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United States	1/2/2009 4:42	1/2/2009 7:49	39.195	-94.6819
1/2/2009 13:08	Product1	1200	Mastercard	Federica e Andrea	Astoria	OR	United States	1/1/2009 16:21	1/3/2009 12:32	46.18806	-123.83
1/3/2009 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/2005 21:13	1/3/2009 14:22	-36.1333	144.75



Practice Assignment

Problem Statement :

Write a map-reduce program to display the payment type wise total sales.

Input File : 1

SalesJan2009.csv file which contains following data :

Transaction_date	Product	**Price**	**Payment_Type**	Name	City	State	Country	Account_Created	Last_Login	Latitude	Longitude
1/2/2009 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	1/2/2009 6:00	1/2/2009 6:08	51.5	-1.11667
1/2/2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United States	1/2/2009 4:42	1/2/2009 7:49	39.195	-94.6819
1/2/2009 13:08	Product1	1200	Mastercard	Federica e Andrea	Astoria	OR	United States	1/1/2009 16:21	1/3/2009 12:32	46.18806	-123.83
1/3/2009 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/2005 21:13	1/3/2009 14:22	-36.1333	144.75

Practice Assignment

Below given example shows the document structure of a library having book documents queries.

```
book1 = {name : "Understanding JAVA", pages : 100,author:  ,publisher:o’Relly}
book2 = {name : "Understanding JSON", pages : 200 ,author:  ,publisher:Mc’Graw
Hill}
```

Crate CSV file for it and Write MapReduce program to find the number of books having pages less 250 pages and greater than that.