16.2.1

MapReduce Process

Now that you and Jennifer have an understanding of big data, the next step is figuring out how to process it. Jennifer is starting to get excited about big data, so you decide that you will handle half of the dataset while Jennifer will handle the other half. MapReduce is a common tool for splitting up large datasets, so you sit down to find out if it will work for you.

MapReduce is used as a means for distributing and processing data on your cluster. MapReduce is built on the process of **mapping**—the process of assigning the same job to each of the computers—and **reducing,** which is when you come back together to combine the results. **Mapping** - processing input and producing small chunks of data across each computer- and **reducing**, which is when you come back together to combine the results.

In this case, you and Jennifer have decided to count all the reviews of video games in different categories across a dataset. To save time, you decide to take the first half, and Jennifer takes the second half. There are 2,800 rows of data to comb through, and we want a total for reviews of sports, fantasy, and role-playing games.

By splitting up the data, the time to analyze it has been reduced by half. In the same way, MapReduce works on divided datasets in smaller batches so that we can work faster.

Once you are done, you will come together and combine the tallies for both.

During the mapping stage the map function, which is the function applied to each computer, takes a small piece of the input and then converts the data into **key-value pairs**, with key identifiers and associated values. For this project, you'll count all of the reviews from the first set of data and store it, as shown below:

Key	**Value**
sports	500
fantasy	300
role_playing	600

Jennifer does the same for her set and stores it as follows:

Key	**Value**
sports	1000
fantasy	100
role_playing	300

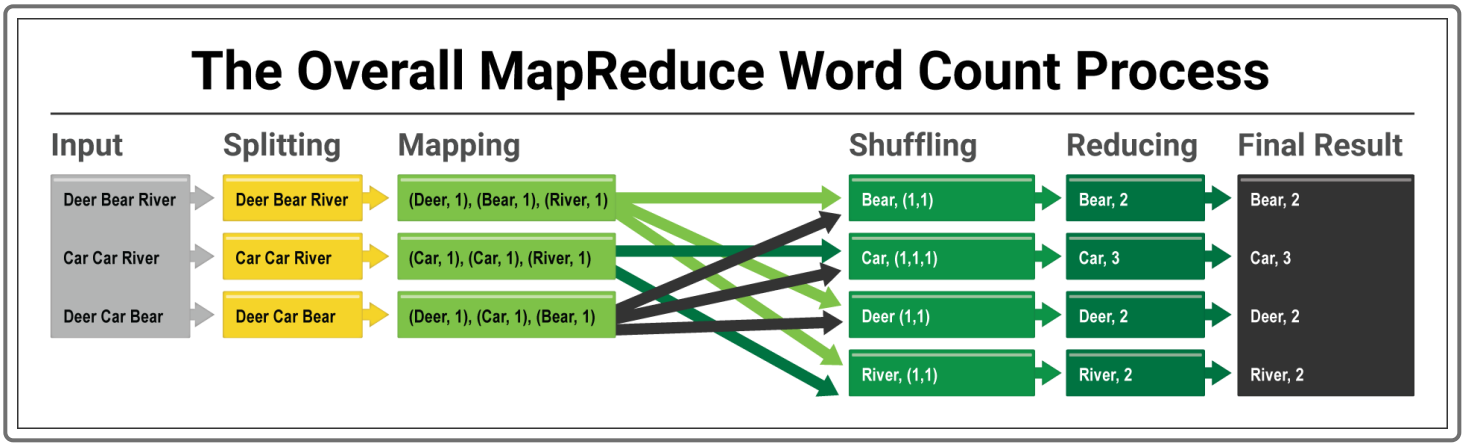
You have now mapped out the review datasets.

Reducing is when you aggregate the results, in this case, by adding up your figures:

Key	**Value**
sports	1500
fantasy	400
role_playing	900

You have just reduced the two results into one.

Another example would be running a word count on a document. The following image shows the MapReduce process for running word count on an input:



Let's break down each part of the word count process:

- **Input:** The entire file is fed to the word counter.
- **Splitting:** Each line of text is separated.
- **Mapping:** Each word in every line is assigned a value of 1.
- **Shuffling:** The words are combined and organized alphabetically, creating a list of the words' values.
- **Reducing:** The list of values are summed for each word.
- **Final Result:** The complete list of words and value (counts) are displayed.

MapReduce is an important tool for handling big data. Next, we'll look at a Python library that will allow us to practice this process at a smaller scale.