PLTW COMPUTER SCIENCE

Activity 3.1.3

Data Innovations and Parallel Algorithms

Introduction

In the last activity, you considered some of the social issues with collecting and analyzing the explosive amounts of data now being produced. But the era of Big Data also has technical challenges.

The rate at which data have been produced in the last ten years is often described as “explosive,” like this 1991 volcanic explosion of Mt. Pinatubo.

To understand some of these technical issues, think about your school's library. How does the library organize the information stored in books? How do you find information in this library? How does the library prevent books from being lost or stolen? Would these methods work if the library had a million times as many books? Would these methods work if that many new books arrived every day? What problems would arise if the library tried to scale up a millionfold?

These problems describe digital data, too. In the current decade, each year brings an enormous increase in the digital data being collected. Methods that were able to send, store, retrieve, and analyze data ten years ago no longer work. What are the new methods?

Materials

- Writing utensil
- One deck of cards per group of five students (optional)

Procedure

Part I: Big Data

1. So much data is being collected that new technical issues arise. **Big Data** refers to many situations in which data is difficult to manage.

> Refer to your downloadable resources for this material. Interactive content may not be available in the PDF edition of this course.

The definition of Big Data is poorly agreed upon and is also changing as computers get more powerful. Four reasons describe why data might be considered Big Data.

- Volume: Big Data has too many bytes to be stored by one computer.
- Velocity: Big Data is produced faster than one computer can store it.
- Variety: Big Data combines several different and conflicting sources.
- Volcanism: Big Data requires explosive amounts of human or computer processing to be useful.

Consider the following two data sets. For each one, decide whether the data fits the definition of Big Data and explain your reasoning.

- The data from the United States Census in year 2010
- The data relevant to creating family trees for the residents of the United States in year 2010

2. Big Data is measured using metric prefixes that may be unfamiliar. Use decimal or scientific notation to write each quantity in units of bytes.

   A meter long shelf of books: 100 MB = _____ bytes

   Tweets in 2010: 200 GB = _____ bytes

   Weather measurements in 2001: 400 TB = _____ bytes

   Total of all printed books: 2 PB = _____ bytes

   Email sent in 2008: 11 PB = _____ bytes

   All sentences spoken ever: 2 EB = _____ bytes

   New data produced in 2010: 14 EB = _____ bytes

Part II: Parallel Processing in Computers

3. Big Data analysis often depends on innovations in **parallel processing**, in which several central processing units (CPUs) work on a task at once. Google published an important approach in 2004 to programming for parallel processors called **MapReduce**. MapReduce allows many computers to work in parallel and then combine their results.

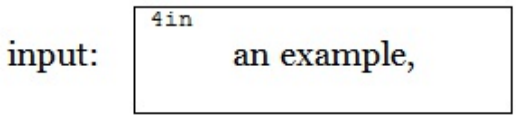
   MapReduce is part of the **Hadoop®** framework. Hadoop is used by the majority of Big Data users. The Hadoop framework uses MapReduce to provide parallel processing. Another important consideration with data is how to make it reliably available and **fault-tolerant**. Fault tolerance means that the system is designed to have reliable storage of data with reliable availability. The Hadoop framework uses the Hadoop Distributed File System to provide **redundant**, distributed data storage. Redundant storage means that multiple copies are maintained; distributed storage means that multiple locations are involved, tolerant against fault in one region.

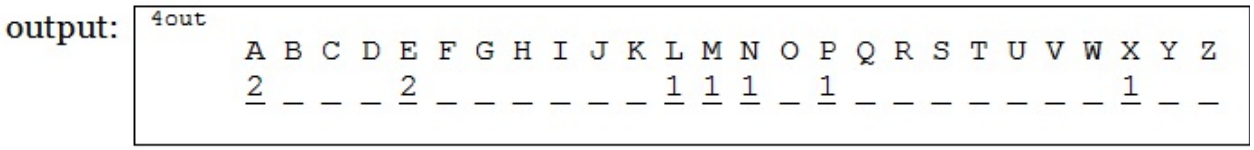
   To program in MapReduce, a software developer writes two functions:

- The `map` function returns a result based on a subset of data.
- The `reduce` function returns a result that combines the output from two previous results of the `map` or `reduce` functions.

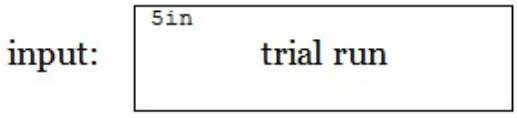
As a class, students will simulate the hardware and software in a Hadoop cluster, making data available, assigning processing tasks, and executing processor instructions on data. The result will be the letter frequencies in the text of this step. Recall from the activity on encryption that the frequency of a letter is the number of times it is used. What is the frequency of the letter “p” in this paragraph?

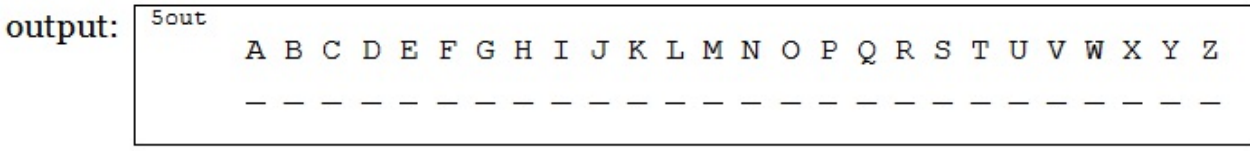
- We will be executing a single job: finding the letter frequencies in the text of the previous step. The `map` function for this job outputs the frequencies of letters in the input. An example is shown here, representing the input and output files as boxes. The input file `4in` contains the text “an example,” and the output file `4out` contains the frequencies of letters that were contained in the input.





Perform the `map` function on input `5in` by hand. Reproduce the output file `5out` by writing down the letters and frequencies.





4. The `reduce` function for this job outputs the sum of the frequencies from two previous outputs. An example is shown here for input files `4out` and `7out`, creating the output file `37out`.

Explain why the values for M, N, and O in the output are all 2.

5. You will be assigned one of the following roles as directed by your teacher. We are using these names for the roles because these are the names of the object-oriented classes in the Hadoop framework.

Name of Role	What They Do
Job Tracker *(1 person)*	Assign Tasks
Data Node + Task Tracker *(half the class)*	Store Data Execute Tasks
Task Tracker *(rest of class)*	Execute Tasks

   - The student with the Job Tracker role should have the slips of paper labeled `out1, out2, ..., out49,` cut into individual strips with one alphabet each. The Job Tracker will hand out individual slips as a way to assign `map` tasks. The Job Tracker should also have the remaining slips of paper `out50, out51, ..., out100` as a way to assign `reduce` tasks.
   - Students with the Data Node + Task Tracker role should have the Activity 3.1.1 Supplement: Data sheet on paper or on a monitor.
   - Students with the Task Tracker role will need to access data by asking a Data Node.
6. Read all the directions in this step and then race as a class to complete the entire job.
   - The Job Tracker will assign one `map` task to each Task Tracker, including Task Trackers that are also Data Nodes.
     - To assign a map task, the Job Tracker hands an output file to the Task Tracker and issues a command. For example, to tell a Task Tracker to record frequencies of input `4`, the Job Tracker says `map(4)` and hands the Task Tracker a slip of paper for the output. It might help if the output file has the number corresponding to the input file, but it is not necessary.
     - As soon as a Task Tracker finishes their `map` assignment, they should hold up the output file to be collected.

- The Job Tracker will collect output files. When the Job Tracker has two completed output files, the Job Tracker should assign a `reduce` task to a Task Tracker.
    - To assign a reduce task, the Job Tracker gives a Task Tracker two completed output files (which become the input for the `reduce` procedure) and a blank output file.
    - As soon as a Task Tracker finishes their `reduce` assignment, they should rip the two input files in half and hold up the output file to be collected.
    - The Job Tracker should continue assigning `reduce` tasks until only one output exists. The job is then complete.

7. (Optional) In groups of five, design an algorithm that will allow all five people working together to sort a standard shuffled deck of cards into any final sorted sequence agreed upon beforehand. After all groups of five have decided on a strategy, compete to determine which group wins with their algorithm.
8. Until recently, most computers had only one processor. Programmers sometimes fail to think about parallel computing until too late in their thinking process. To take advantage of modern computing, especially to process Big Data, the programmer must think about their algorithm using separate, independent tasks. These independent parts of a job are called **threads**.

   If the threads perform the same instructions on different data, the approach is called **data parallel**. If the threads have different instructions, the approach is called **task parallel**. MapReduce is both. In what ways is MapReduce data parallel, and in what ways is it task parallel?

9. In the last few years, most new computers have been manufactured with dual or quad processors, and they mostly use a task parallel approach by working on different threads. The graphics card in a computer uses hundreds of parallel processors collectively called the **graphics processing unit (GPU)**. The GPU processors share an instruction bus but operate on different streams of data. The Cuda platform allows you to write programs for the Central Processing Unit (CPU) that will assign data-parallel processing tasks to the GPU. Read **https://developer.nvidia.com/about-cuda**. That web site's list of application domains of parallel processing is repeated here. Pick one and skim a Wikipedia page about the application. With your partner, summarize what is being accomplished with parallel processing in that application. Record your summary as directed by your teacher.
    - Bioinformatics
    - Computational Structural Mechanics
    - Electronic Design Automation
    - Medical Imaging
    - Computational Finance
    - Defense
    - Numerical Analytics
    - Computational Fluid Dynamics
    - Data Science
    - Computational Chemistry
    - Electronic Design Automation
    - Weather and Climate Simulation

Part III: Parallel Processing in Brains

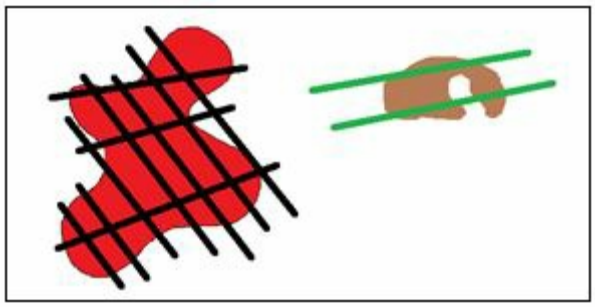
11. Refer to your downloadable resources for this material. Interactive content may not be available in the PDF edition of this course.

The retina is light-sensitive brain tissue at the back of each eyeball. Light is directed by the lens to create an image on the retina. The neurons in the retina include light sensitive rods and red-, green-, or blue-sensitive cones**.** These neurons turn the image into a set of electrical signals. The signals are processed by another layer of neurons and are then sent to the back of the brain. The signal is sent through the optic nerve to the occipital lobe, the back of the brain. Before reaching the occipital lobe, the image is processed to identify dots, lines, and faces in all areas of the field of vision.

- The parallel processing from retina to occipital lobe identifies dots, lines, and faces in all areas of the field of vision. In what way is that description describing a task parallel approach? In what way is the approach data parallel?
- The retina, optic nerve, and occipital lobe are made of neurons, the type of cell that sends electrical output to muscles or other neurons based on sensory input or electrical input from other neurons. Neurons transmit about 10 bits per second.

  Each human retina has about 126 million light sensitive neurons. The signal from the retina is processed and carried by only 1 million neurons in the optic nerve. Estimate the image compression ratio as the percentage decrease in the size of the image in bytes.

12. Data visualization combines the most powerful aspects of artificial computers with human intelligence. Processing an image or discovering knowledge from a data visualization is very difficult for artificial intelligence. To demonstrate the power of the human brain to use information presented through the retina, consider the image below. The image contains two shapes, both obscured by lines.



- Estimate the ratio of the big shape's area to the small shape's area. How many of the small shape, chopped up as needed, could you fit in the large shape?
- Describe the algorithm you would use to do the same calculation for this image using a computer.
- Contrast how well you think a human and a computer could calculate area ratios in images like this. Is the human or computer faster? Is the human or computer more accurate? Is the human or computer more flexible when the image is changed slightly by using different colors or shapes, or by adding new distracting elements like the lines?
- Why do we need computers to create data visualizations? Why do we need humans to

interpret data visualizations?

Part IV: Big Data and Societal Impact

13. Data scientists use **exploratory data analysis**, in which the strengths of human and computer intelligence are used together. By using computational power, you can create a data visualization that displays the data in a graphic which the human brain can use to identify patterns. These patterns suggest hypotheses that can be tested with **inferential statistics** in which data are used to draw conclusions that will probably apply to additional data that has not yet been collected.
    - In the rest of this lesson, you will learn about different kinds of data csvisualizations. What kind of visualization would you use to see patterns in the data you produced about letter frequencies?
    - In the last activity, your class brainstormed a list of questions about medicine, education, or retail consumption that might be answered by Big Data. With your partner, produce an artifact in which you:
        - State the question to be answered by data.
        - Sketch a data visualization that would support one answer to the research question.

14. For the AP CS Principles *Explore* Performance Task you must find three recent, credible sources of information about a computing innovation that:
    - Has or could benefit and harm society, economy, or culture
    - Consumes, produces, or transforms data
    - Raises a storage, privacy, or security concern regarding data.

    In each unit of this course, you will investigate particular impacts of computing innovations on society. In this activity, find one or more articles referenced in the ACM TechNews archive **http://technews.acm.org/archives.cfm** or another credible source about the impact of a computing innovation related to one these topics:

    - data collection
    - data processing

    Technical news is often reported in secondary sources such as newspapers and magazines. You can often get more detailed and accurate information by following references. For summaries from the ACM TechNews , be certain to follow the references to the original (but still usually secondary) articles being summarized by the ACM TechNews.

    Other topics may be explored at the discretion of your teacher. Find relevant summaries of news article from the ACM TechNews and read the original articles being summarized. Complete some portions of the *Explore* task described below as directed by your teacher.

    Task part 1. Create an audio, video, or visual artifact that illustrates, represents, or explains the computing innovation's purpose, function, or effect. (3 page/1 minute/30MB max)

    Task part 2. Essays

    - Name the innovation and its purpose and function. Describe how your artifact illustrates, represents, or explains the computing innovation's purpose, function, or effect.

(Approximately 100 words.)
- Describe the tools, technique, and process you used to produce the artifact. (Approximately 100 words.)
- Explain the beneficial AND harmful effect(s) the innovation has or could have on society, economy, or culture. (Approximately 250 words.)
- Describe the data; the consumption, production, or transformation of data; and the storage, privacy, or security concern(s) directly related to the innovation. (Approximately 250 words.)
- Use APA-style citations to correctly reference the article(s).

Note: This step is adapted from the official College Board Explore Performance Task but it does not duplicate the content of College Board Task or Rubric. The task provided here contains elements that are different than the College Board Performance Task and Rubric. Please reference official College Board materials.

Conclusion

1. Estimate how many bytes of data are collected about you and your immediate surroundings per day. Consider cameras, microphones, keyboard and mouse input, car sensors, etc.
2. The introduction to the first activity in this lesson described the computing revolution as occurring in four categories:
    - Connectivity through the Internet (Unit 2)
    - Large-scale data collection and analysis (Unit 3)
    - Robotics and physical automation (Unit 4)
    - Simulation and modeling (Unit 4)

    These four categories are expected to have their greatest impact on our world roughly in the order they are listed, although all four categories of impact are changing our world at once.

    The following events were some milestones in the first category, Connectivity through the Internet.

    - 1969: The first packet transmitted from one computer to another
    - 1982: The TCP/IP protocol agreed upon, with 235 computers connected to Internet
    - 1989: First Internet Service Provider offers Internet to home users
    - 1991: First web page
    - 1995: First commercial use of Internet

    The Internet was the primary way in which computing revolutionized the world in the 1990s and 2000s.

    - There are four criteria for calling data “Big Data” given in Step 8. By this definition, connectivity through the Internet had to occur first, before Big Data. Why? Explain the relationship between Big Data and connectivity through the Internet.
    - Use your imagination and make some predictions. What will be the milestones on the timeline of Big Data, and when will they occur?
3. The entire human brain, and not just the eye, is both task parallel and data parallel. Consider sensation from the other senses or output to muscles. Describe ways in which the brain is task parallel and ways in which it is data parallel for senses other than vision or for motor output to

muscles.