(http://csmatters.org) 5 - 1a

0b101 - 0b1

Manipulating Large Data Sets 2020-2021

Unit 5. Data Manipulation

Revision Date: Feb 06, 2020
Duration: 3 50-minute sessions

Lesson Summary

Summary

Big Data has been defined in many different ways. Easy access to large sets of data and the ability to analyze large data sets changes how people make decisions. Students will explore how Big Data can be used to solve real-world problems in their community. After watching a video that explains how Big Data is different from how we have analyzed and used data in the past, students will explore Big Data techniques in online simulations. Students will identify appropriate data source(s) and formulate solvable questions.

Students are introduced to parallel and distributed programming and to their use with big data.

Learning Objectives

CSP Objectives

- *EU CRD-2 - Developers create and innovate using an iterative design process that is user-focused, that incorporates implementation/feedback cycles, and that leaves ample room for experimentation and risk-taking.*
  - LO CRD-2.J - Identify inputs and corresponding expected outputs or behaviors that can be used to check the correctness of an algorithm or program.
- *EU DAT-2 - Programs can be used to process data, which allows users to discover information and create new knowledge.*
  - LO DAT-2.A - Describe what information can be extracted from data.
  - LO DAT-2.C: - Identify the challenges associated with processing data.
  - LO DAT-2.D - Extract information from data using a program.
  - LO DAT-2.E - Explain how programs can be used to gain insight and knowledge from data.
- *EU CSN-2 - Parallel and distributed computing leverage multiple computers to more quickly solve complex problems or process large data sets.*
  - LO CSN-2.A - For sequential, parallel, and distributed computing: a. Compare problem solutions. b. Determine the efficiency of solutions.
  - LO CSN-2.B - Describe benefits and challenges of parallel and distributed computing.

Key Concepts

Outcomes

- Students will explain how analyzing Big Data is different from the way ordinary data is analyzed.
- Students will provide examples of how the scalability of systems is an important consideration when working with data sets, as the computational capacity of a system affects how data sets can be processed and stored.
- Students will describe how computers can make predictions and answer questions through the use of Big Data, storage of data, and processing data.
- Students will synthesize the relationship(s) between causation and correlation

- Students will describe the benefits and challenges of parallel and distributed computing.

Teacher Resources

Lesson Plan

Session 1 - What is Big Data?

Getting Started (10 min) - Journal

Journal: How can a computer gather data from people? (Think-Pair-Share)

Remind students of the mind guessing game: http://en.akinator.com/ (http://en.akinator.com/) or 20 questions http://www.20q.net/ (http://www.20q.net/)

Discuss: How can the computer learn from people when playing one of these games? How many different answers do you think it could possibly know?

Teacher note: students are not expected to actually play this game during class.

- Students should document in their journal the answer to this question: How does this game store all of the possible answers?
- Ask 3 students to share their answers. (Possible strategies to select a random student: random.com, or pick a random student name stick from a cup.)
- This activity should lead to today's lesson on how large amounts of data are stored and then accessed as needed in a system.

Guided Activities (30 min) - Reading & Video

Read The Rise of Big Data in chunks: An Introduction to "Big Data" (20 mins) Reading can be found at: http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data (http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data)

1. Break students into groups or pairs and jigsaw the seven units of the reading. Each group is to summarize their section in a tweet-sized comment (not more than 140 characters).
2. Share tweets with the class.
3. Explain to students that big data is impacting every area of life. By using more data and processing power we can make better decisions. As an illustration, show a clip from the movie Moneyball: (3 mins) https://www.youtube.com/watch?v=rMObWsKalls
4. After students watch, they create a journal entry explaining at least two ways data was used to better manage the baseball team. Partners discuss journal entries. Share at least one observation with table groups and then share at least one observation from each group with the class.
5. Challenge students to think of evidence that would indicate that the size of a data set affects the amount of information that can be extracted from it. (prompt with counterexamples: if you wanted to predict best movies and only polled senior citizens; if you wanted to offer a new clothing style and only got data from people living in Alaska, etc. why is there more information available the bigger the data set gets?)

Show the first 3-5 minutes of this clip. (It becomes a bit dry, so just show the amount that is appropriate for your students to get the idea): https://www.youtube.com/watch?v=7D1CQ_LOizA (https://www.youtube.com/watch?v=7D1CQ_LOizA)

- **What are the 3 V's? List some details about each V.**
- (at 4:40) What is Hadoop and how is it used?
- Identify appropriate data source and form questions
- Extract data source into a format supported by underlying tools
- Normalize data (remove redundancies, irrelevant details)

- Make data uniform (change abbreviations, typos, capitalization inconsistencies, etc.)
- Import data into a tool
- Perform analysis
- Visualize results

Some examples of how big data is used appropriately:

- Netflix and Amazon use it to improve user recommendations
- Dominos used it to determine that more people order pizza when it is raining so they now base some of their ad campaigns around weather patterns
- Police use it to predict when and where crimes will appear

Some examples of how big data was inappropriately used:

- In 2012 Target store's “outing” a teenager’s pregnancy
- In 2012 Google spent 22.5 million on a settlement over allegations that they secretly tracked user’s web surfing
- In 2012 Facebook spent 20 million to settle a lawsuit that alleged that they used user pictures without the user’s knowledge to endorse products that they “liked”
- In 2013 the revelation of the NSA using big data for national security concerns

Wrap Up (10 min) – Group Review

- Place students in groups of 4 where the first student is A, the next is B, etc. Each group creates a single sheet of paper with the letters across the bottom and the numbers 1 - 5 to demonstrate the level of understanding of each concept. Students should plot their understanding of each respective concept (see list below) on the graph. (The file "BigDataSampleDotGraph" (https://drive.google.com/open?id=1mVPZpte0c9cbwMVODlPd5r3dMu7aLhwdfutKgmxvT30) in the lesson resources folder shows an example.)
- Collect this graph as a level of student’s understanding of the concepts in the video.
- Review each concept using the notes below.
    - Big data is kind of like drinking water from a fire hose. It's too much to process for a small pipeline…
    - A. The three “Vs”: Volume, Variety, and Velocity
    - B. Big Data processing steps:
    - C. Tools for processing big data:
        - Microsoft Excel (or some type of spreadsheet tool - i.e. Calc is another one)
        - Hadoop - a well known big data tool, provided by Apache, requires extensive programming knowledge to set up and use
        - SAS - provides a more intuitive interface and better graphical representations of data
        - Google Prediction API - takes advantage of machine learning to extract meaning from data
        - BitDeli - lightweight, easier to use version of Hadoop
    - D. Very few restrictions on the use of big data
        - Companies collect *large* amounts of data on their customers
        - Can be sold to other companies
        - Can be sold to the government
        - Can be used to “de-anonymize” someone

Homework

Students are to pick three topics they want to research that use big data. It is preferred that these topics relate to something learned this year in the course (e.g., the need for IPv6). Tomorrow, as the students enter class, they will sign up on a list with their chosen topic. Since the students will have three options, it is likely they will get one of their selected topics to research.

Session 2 - Where can Big Data be used?

Getting Started (5 min) - Journal

Journal: Think about your daily and weekly activities. What types of data are being stored about you?

Remind students to think about what they do online, in stores, while in a car, etc.

Guided Activities (10 min) – Processing Big Data

Review the steps to processing Big Data:

1. Identify appropriate data source and form questions

2. Extract data source into a format supported by underlying tools
3. Normalize data (remove redundancies, irrelevant details, etc. as needed)
4. Make data uniform (user-entered data may include abbreviations, spelling errors, or inconsistent capitalization without changing the meaning of the data)
5. Import data into a tool
6. Perform analysis
7. Visualize results

As a class, walk through these steps using the two files in the lesson resources folder (https://drive.google.com/open?id=0By2KZS8SzSUcMkJxbTIzMm9RRm8)(FailedBanklist.csv (https://drive.google.com/open?id=0B2umpE0UajfYaFB6bVlGTmhXYzQ) & Consumer_Complaints.csv (https://drive.google.com/open?id=0B2umpE0UajfYQ05rcEg5MlJhbUE))

Step 1.

Demonstrate how files such as these can be obtained at http://catalog.data.gov/dataset (http://catalog.data.gov/dataset)

Formulate questions such as:

1. How does the size of a data set affect the amount of information that can be extracted from it?
2. Are there any banks that are on both the complaint list and the failed bank list?
3. Can we make some deductions about banks that may be on both lists? If so, what deductions can we make?

Step 2.

Extract data source into a format supported by underlying tools

Open one of these files in Notepad (or some simple editing program such as Notepad++) and demonstrate how the actual data itself is separated by commas, thus the file name "csv" for comma-separated value.

Open both files in Microsoft Excel. Complete a find for the bank name "Banco Popular de Puerto Rico" on both lists. You may want to first sort the data by bank name to find this bank or you can use CTRL + F to find the bank name (see screenshots below).

Steps 3 & 4.

Normalize data (remove redundancies, irrelevant details)

> In this step, there is technically no need to remove redundancies or irrelevant details but you can show the students how you could remove data or limit the data to a particular data set. For example, if were to want to look at only the banks from Maryland, you can use the filter tool to only view those banks from MD.

Make data uniform (user-entered data may include abbreviations, spelling errors, or inconsistent capitalization without changing the meaning of the data)

> Cleaning data: Depending on how the data was collected, it may not be uniform. Therefore the data may need to be cleaned before it can be processed. Cleaning the data is the process that makes the data uniform without changing its meaning. For example, replacing all equivalent abbreviations, spellings, and capitalizations of the same word.

Step 5.

Import data into the tool

Right now the file type is as a csv file. By resaving the file as a .xlsx file it becomes a true spreadsheet file.

Step 6.

Perform analysis

We have determined that the bank "Banco Popular de Puerto Rico" is on both lists. Now ask the students "Why is this bank on both lists?" Note: On the Failed Bank list the Banco Popular de Puerto Rico is actually an acquiring institution. By looking more closely at the dates of the acquisition of the failed bank "Westernbank Puerto Rico" one can formulate some possible deductions that maybe the reason "Banco Popular de Puerto Rico" is on the complaint list is because they had recently taken over a failed bank. It could be possible that some of these complaints were related to this recent acquisition.

Step 7.

Visualize Results

Explain to students that they will learn more about visualizing their results in Unit 6. They can complete graph visualization in excel. Show them the website: http://www.gapminder.org/ (http://www.gapminder.org/). Explain that even though a visualization in excel is not interactive like http://www.gapminder.org/ (http://www.gapminder.org/), they can complete some form of visualizing their data by using a spreadsheet. Note: http://www.gapminder.org/ (http://www.gapminder.org/) is VERY attention-grabbing. Only briefly show the students what they can do with it (see how data changes over time, look at many different data sets and download data in different forms - including csv and xlsx formats).

Independent Activity (30 min) – Online Research

Students should research their selected topics from homework. Some possible websites for finding data are listed above under “Possible good resource(s) for data collection.”

Students are to get your approval for a topic and then use the Big Data Sets Worksheet (https://drive.google.com/open?id=1q83owjEkA_RTcmqp5DwaR0zzUEHwrilxiNvzgTbjBAI) in the Lesson Resource Folder (https://drive.google.com/open?id=0By2KZS8SzSUcMkJxbTIzMm9RRm8) to find big data sets that are related to the approved topic.

Wrap Up (5 min) – Exit Slip

Students are to review using http://www.gapminder.org/ (http://www.gapminder.org/) looking specifically at life expectancy. Students will write *one* question after “playing” the timeline of life expectancy using gapminder on an exit slip before leaving class. For example, one may write “Why is the life expectancy of countries such as Denmark, Sweden, & Norway typically higher than other countries throughout most of the timeline?”

Session 3 How can big data be processed?

Getting Started

Challenge students to work in teams of four to find:

- All of the US States Capital locations (latitude and longitude of each)
- All of the Consumer Products Currently made by Apple
- The Top Five Social media Websites in terms of daily users
- The 10 most common baby names in each of the last five years

After 2 min stop students and ask them to share how they approached the problem.

Say: You split the task into pieces, and each person worked at the same time to get the job done more quickly than would be possible by yourself. This is parallelism. In computing, parallelism can be defined as the use of multiple processing units working together to complete some task. There are many different kinds of hardware that can serve as a “processing unit”, but the principle is the same: a task is broken into pieces in some way, and the processing units cooperate on those pieces to get the task done. Basic of Processes with Python (http://selkie.macalester.edu/csinparallel/modules/ParallelProcessesInPython/build/html/ProcessesBasics.html)

Students launch a Python 2 IDE such repl.it or JDoodle (https://www.jdoodle.com/python-programming-online) and paste the following code adapted from the Basics of Processes with Python web page.

```
from multiprocessing import *
def sayHi():
print "Hi from process", current_process().pid
def procEx():
print "Hi from process", current_process().pid, "(parent process)"
otherProcess1 = Process(target=sayHi, args=())
otherProcess1.start()

otherProcess2 = Process(target=sayHi, args=())
otherProcess2.start()

otherProcess3 = Process(target=sayHi, args=())
otherProcess3.start()
### execute
procEx()
```

Students execute the code - and debug if needed.

Review the definition of scalability. Ask how the scalability of systems is an important consideration when working with data sets, as the computational capacity of a system affects how data sets can be processed and stored.

Ask: What do the numbers mean in each line of output?

Guided Activities

Activity 1 Parallel Programming

Explain:

1. Parallel processes do not always finish in the same order they started.
   - The code above follows a common pattern: a parent process creates one or more child processes to do some task. In this example, suppose we call procEx. The first line in that function prints a simple message about what process is running. This is done by calling the function current_process that is defined in the multiprocessing module. The current_process function returns a Process object representing the currently running process.
2. Every Process object has a public field called "pid", which stands for “process identifier”.
   - Thus current_process().pid returns the pid for the currently running process. This is what gets printed. By calling the Process constructor, we created a Process object, but we have not yet started a new process. That is, the process exists but is not available to be run yet. The process is actually started with the last line of procEx.
3. There are two steps to make a child process do some task:
   - A Process object is created using the constructor, and then the Process object is started using the start method. When more than one process is started the processes can run at the same time - in parallel.
4. All of our programs before today used sequential computing. Sequential computing is a computational model in which operations are performed in order one at a time. Sequential computing solutions take as much time to run as the sum of the time taken by each of their steps. With parallel computing, the total time needed is a combination of a sequential part and just the longest time taken by each of the parallel parts. as a result, solutions that use parallel computing can scale more effectively than solutions that use sequential computing. Show this Udacity video on the **advantages of parallel processing. (https://youtu.be/pa_785ZqrC4?t=18)**

Discuss:

Use the questions below as prompts. Students share until the idea that parallel solutions scale more effectively is discussed.

Say: Computer scientists use the term scaling to mean how a process responds to increases in the size of the project.

- How might a parallel computing solution make a process more scalable?
- Why does a search engine such as Google would use parallel processing to rank the web pages in their database?

Activity 2 Distributed Programming

Say: In the first activity we saw an example of a program where more than one computer processor on a computer could be used at a time. Sharing a task among the processors on many computers is an example of distributed problem-solving. Distributed computing shares components of a software system among multiple computers so a large task can be done in less time using the resources available to each of the computers - including both processing and storage resources. As a result, Distributed computing allows problems to be solved that could not be solved on a single computer because of either the processing time or storage needs involved.

Students read the first 5 paragraphs of the OpenScientist (http://www.openscientist.org/p/distributed-computing-project-open-for.html).

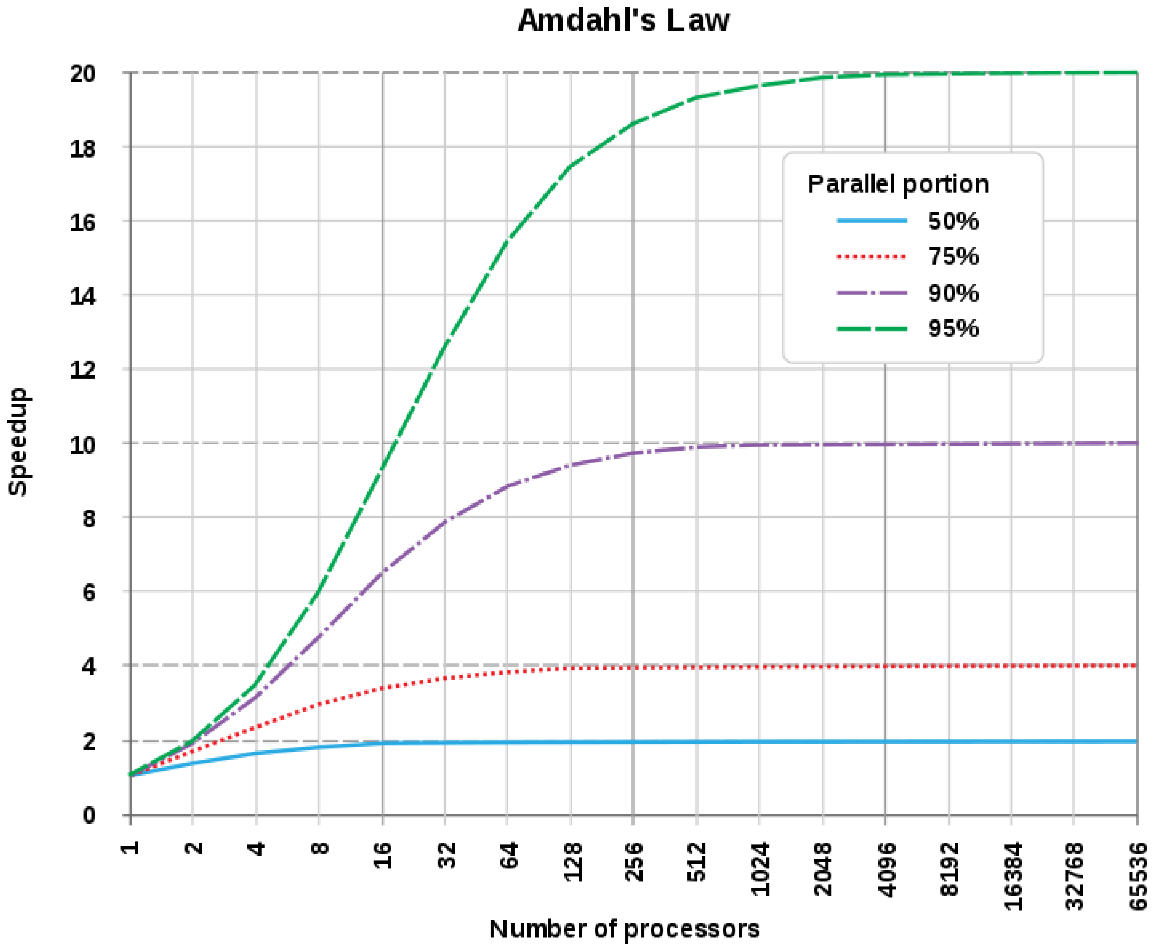
Discuss: With elbow partners, discuss what are the advantages of distributed computing and choose the top three projects that interest you from the list on the OpenScientist (http://www.openscientist.org/p/distributed-computing-project-open-for.html).

- Make a list of the 10 most interesting projects to the class.
- Ask: Why these projects are better solved using Distributed Computing rather than a single computer?
- Students share until both time and storage constraints of using a single computer are discussed.

Activity 3 Speedup

Say: Parallel and distributed computing are powerful tools but they have their limits in terms of increasing efficiency.

Show this graph of Amdahl's Law by Daniels220 at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6678551 (https://commons.wikimedia.org/w/index.php?curid=6678551).



Ask:

- At what number of processors, does the blue line stop indicating an increase in speed?
- At what number of processors, does the red line stop indicating an increase in speed?
- At what number of processors, does the purple line stop indicating an increase in speed?
- At what number of processors, does the green line stop indicating an increase in speed?

Explain: The "speedup" of a parallel solution is measured in the time it took to complete the task sequentially divided by the time it took to complete the task when done in parallel. The 4 different colored lines represent four different processes with increasing portions that can be done in parallel. At first, they all speed up as more processors are used. After a time, adding more processors no longer speeds up each process. The sequential parts of processing limit the overall time.

Discuss: With elbow partners, students discuss why they think this is the case. Students share until these key points are verbalized:

- a large solution consists of parallelizable parts and non-parallelized (sequential) parts
- at some point adding more parallel parts no longer adds to the efficiency of the solution

Wrap up:

Students return to the OpenScientist (http://www.openscientist.org/p/distributed-computing-project-open-for.html) and choose the project that most interests them.

Say: Parallel computing consists of a parallel portion and a sequential portion.

Students identify a part of their chosen project that is done in serial and a part that is done in parallel.

Evidence of Learning

Formative Assessment

Students are to submit a document stating their topic for research using Big Data. This document should answer the questions:

Topic:

How is Big Data used to solve or remedy the topic?

Link(s) used to find Big Data? (i.e. data.gov, etc)

Summative Assessment

How has the transformation of data storage affected how data itself is used?

Answer: Storage and processing of large digital data enables us to analyze large data sets quickly rather than small sampling sizes as used before.

How can a computer use Big Data to make predictions?

Answer: Computers can use smart algorithms, powerful processors, and clever software to make inferences and predictions for solvable questions.

How can parallel processing help scale a solution to a Big Data problem?

Larger problems often have more parallelizable segments. The more processes that can be done in parallel the faster the overall process will become.

(http://www.umbc.edu/) (http://www.umd.edu/) (http://www.nsf.gov/)

Authored by: CS Matters in Maryland
Website: csmatters.org (http://csmatters.org)
Email: csmattersinmaryland@gmail.com (mailto:csmattersinmaryland@gmail.com)