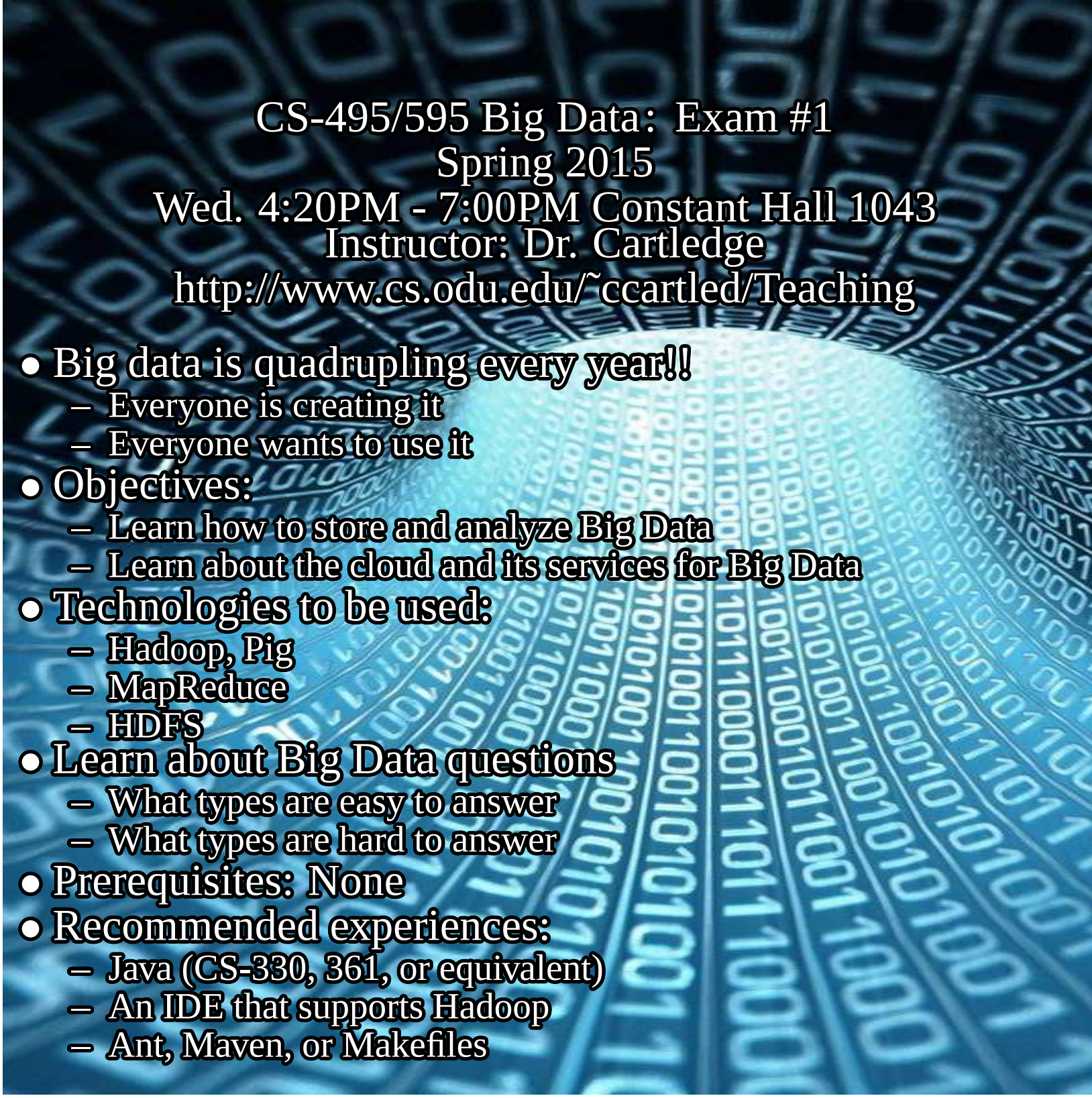


Other notes:

- This is not a programming course. You will learn to use an existing framework, not a new language.
- There will be approx. three programming assignments. Graduate students will expand on the basic assignment with analysis and presentations.
- Text:Hadoop: The Definitive Guide, ISBN-13: 9781449311520

Contents

1 Directions

1. Sign and date the Honor Code acknowledgement.
2. Undergraduate students – Answer the undergraduate part of each question. In addition, you **MAY** answer the graduate part. Incorretly answering the graduate part will not affect your grade. Answering the graduate part correctly will be counted as extra credit.
3. Graduate students – Answer both undergraduate and graduate parts of 10 questions. You may answer more questions (both undergrade and graduate parts), be sure to mark the extra questions as extra credit so the grader will know how to treat them.

At the end of the exam, you will have at least 20 answers.

2 Academic Integrity / Honor Code

By attending Old Dominion University you have accepted the responsibility to abide by the honor code. If you are uncertain about how the honor code applies to any course activity, you should request clarification from the instructor. The honor pledge is as follows:

> "*I pledge to support the honor system of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagiarism. I am aware that as a member of the academic community, it is my responsibility to turn in all suspected violators of the honor system. I will report to Honor Council hearings if I am summoned.*"

_____ _____

Signature Printed name

_____ _____

UIN Date

I am taking this course as (check one):

Undergraduate ☐
Graduate ☐
Neither ☐

3 Exam

1. When dealing with LARGE data sets, you can expect to find inconsistencies, values that are missing, values that are out of scale, values that are illogical.

   (a) Undergraduate question: What are two terms for making the data ready for analysis?
   The basic answer is:

      i. Data munging
      ii. Data wrangling

   (b) Graduate question: Assignments 1 and 2 had to be “readied” for processing. Name three techniques or filtering that was applied to the data.
   The basic answer is:

      i. Case folding
      ii. Stemming
      iii. Stop words
      iv. Punctuation removal

2. In class we talked about four different ways to "transmit" code to a reader in a soft document.

   (a) Undergraduate question: What are the four ways?
   The basic answer is:

      i. *Copy and paste*
      ii. *Download from somewhere*
      iii. *Encode the software in the document*
      iv. *Attach the software to the PDF*

   (b) Graduate question: Explain the limitations of each way.
   The basic answer is:

      i. *Copy and paste* — the reader highlights the code and then uses whatever copy command or keystrokes the operating system provides to copy and paste the code into the editor. This works well if:
         A. The listing is small and doesn't cross page boundaries, or
         B. The listing has not been reformatted to "look good" on the page (pretty printing and so on), or
         C. The listing is a stand alone entity and doesn't depend on other pieces of code.
      ii. *Download from somewhere* — the reader has access to the site to download a complete soft copy of the listing. This works well if:
         A. The software remains at particular location on the site, and
         B. The site is still alive, and
         C. The user has access to the site. Downloading may be problematic if some sort of authentication is required.
      iii. *Encode the software in the document* — the reader can copy and paste an encoded version of the listing from the document, "decode" it, and process it an editor. This works well if:
         A. Detailed decoding instructions are available, and
         B. The reader has access to the decoding tools, and
         C. The "copy and paste" procedure is accurate and complete. This is not necessarily true.
      iv. *Attach the software to the PDF* — a PDF file can have attachments and the reader can save those attachments to their local file system. This works well if:
         A. Detailed instructions are available to the reader about how to access attachments, and
         B. The PDF reader permits the attachments to be saved to the local file system. By default, certain file extensions can not be saved locally.

3. As part of the homework submissions, six questions were to be answered.

   (a) Undergraduate question: What are the six questions?
   The basic answer is:

      i. What is the question?
      ii. Why is it important?
      iii. What have others done to try and solve the question?
      iv. What will I do to solve the problem?
      v. What will I do to prove that I have solved the problem?
      vi. What is the conclusion?

   (b) Graduate question: How much should be written to answer each question?
   The basic answer is:

      i. What is the question? — This should be exactly one sentence in the form of a question.
      ii. Why is it important? — This should be a short paragraph.
      iii. What have others done to try and solve the question? — A short couple of paragraphs.
      iv. What will I do to solve the problem? — As much as necessary, but not any more than necessary.
      v. What will I do to prove that I have solved the problem? — Justify your results.
      vi. What is the conclusion? — A paragraph to wrap things up.

4. We talked about some types of questions that are paralizable.

   (a) Undergraduate question: Which of these computations can be paralized?
      1. $a[i] = b[i] + c[i]$
      2. $a[i] = f(b)$
      3. $a[i] = a[i-1] + b[i-1]$
      4. $a = b + c$
      5. Compute the value of π

      The basic answer is:
      1. $a[i] = b[i] + c[i]$ — Yes
      2. $a[i] = f(b)$ — Yes
      3. $a[i] = a[i-1] + b[i-1]$ — No
      4. $a = b + c$ — depends on what the factors are
      5. Compute the value of π — partially

   (b) Graduate question: Explain your answers.
      The basic answer is:
      1. $a[i] = b[i] + c[i]$ — all factors are independent
      2. $a[i] = f(b)$ — all factors are independent
      3. $a[i] = a[i-1] + b[i-1]$ — factors are dependent
      4. $a = b + c$ — depends on what the factors are
      5. Compute the value of π — generating test data is, the division is not

5. Amdahl's Law sets the upper limit the speed improvement that can be expected by doing parts of an algorithm in parallel.

   (a) Undergraduate question: What is Amdahl's equation?
   The basic answer is:
   $T(n) = T(1) * (B + \frac{1}{n}(1 - B))$

   (b) Graduate question: Explain each term in the equation.
   The basic answer is:

   - Time for serial execution $\stackrel{\text{def.}}{=} T(1)$
   - Portion that can NOT be paralyzed $\stackrel{\text{def.}}{=} B \in [0, 1]$
   - Number of parallel resources $\stackrel{\text{def.}}{=} n$

6. In assignment #1, our English department professor was interested in studying the vocabularies of various plays. She wanted the vocabulary treated in special ways.

   (a) Undergraduate question: What are the four different ways she wanted the vocabulary manipulated before analysis?
   The basic answer is:
   She is not interested in these types of words:
      i. Articles (sometimes called “stopwords”)
      ii. Plurals of the same word (stem and stems would be treated as the same)
      iii. Names of the actors
      iv. Different cases of the same word

   (b) Graduate question: Is there a universal collection of“high frequency” words, and why do they deserve special attention?
   The basic answer is:
   They are called “stop words” and they are domain specific. Stop words skew the vocabulary distributions and tend to hide/drown out contributions from other words.

7. The definition of “Big Data” varies from context to context. In one sense, Big Data is any collection of data that can not be handled by a single machine, at a single time.

   (a) Undergraduate question: What are the classic three Vs that describe Big Data and what do they mean?
   The basic answer is:

   - *Volume*: lots of data
   - *Velocity*: data is created fast
   - *Variety*: data has different origins

   (b) Graduate question: Sometimes three other Vs are used to describe Big Data. What are they and what do they mean?
   The basic answer is:

   - *Veracity*: is the data trustworthy?
   - *Value*: how “good” is the data?
   - *Variability*: is the data consistent?

8. Big Data is everywhere. We are “swimming” in it, as well as contributing our data to others.

   (a) Undergraduate question: There are three Big Data players/roles. What are they?
   The basic answer is:

   - Brokers
   - Scientists
   - Visionaries

   (b) Graduate question: What are their functions?
   The basic answer is:

   - Brokers — collectors of data and make data available
   - Scientists — figure out what is in the data and how to get information out
   - Visionaries — have a vision of what is in the data, but do not have the skills to extract the information

9. In the realm of statistics, there are a number of different ways to talk about *n*. Some of these ways are directly applicable to Big Data, data processing, and inferences that can be made.

   (a) Undergraduate question: What are two different views on *n* relative to Big Data?
   The basic answer is:
   n=1,000 vs. *n=all*, or
   - Classical (frequentism)
   - Bayesian

   (b) Graduate question: Big data focuses on one type of statistical relationship. What is it and why?
   The basic answer is:
   Correlation. Items/events/things that occur together, but do not indicate that one causes another. Causation is not indicated or implied.

10. Luis von Ahn came up with the idea for *captcha* processing. Later he came up with *recaptcha*.

    (a) Undergraduate question: There are two services that are provided by a *recaptcha*. What are they?
    The basic answer is:

        i. Keep robots and automated tools from getting past a page.
        ii. Have humans solve an OCR problem that software can not.

    (b) Graduate question: How does Amdahl's Law apply to *recaptcha*?
    The basic answer is:
    Harnesses minimal human processing that when aggregated is a substantial effort. Humans can easily solve some simple problems that are very difficult for computers and software.

11. “This, in a nutshell, is what Hadoop provides: a reliable shared storage and analysis system. The storage is provided by HDFS and analysis by MapReduce. There are other parts to Hadoop, but these capabilities are its kernel.” is a quote from the textbook.

    (a) Undergraduate question: What are the four main things that Hadoop “brings to the table” for Big Data processing?
    The basic answer is:

    - Paralizable
    - Data locality
    - Coordination
    - Output

    (b) Graduate question: How are each of these things important to Big Data processing?
    The basic answer is:

    - Paralizable – number of mappers determined by the size of the input (growth in linear)
    - Data locality – HDFS distributes input data to processing nodes
    - Coordination – starts, monitors, stops, restarts parallel tasks
    - Output – intermediate output is local, global is in HDFS

12. Hadoop, a Java application claims to be almost language agnostic.

    (a) Undergraduate question: Name three languages that Hadoop supports.
    The basic answer is:
    C, C++, Ruby, Python, . . .

    (b) Graduate question: What are the minimum requirements for any language that Hadoop runs?
    The basic answer is:
    Supports STDIN and STDOUT.

13. The Hadoop Distributed File System (HDFS) is an integral part of the Hadoop ecosystem. It is robust, and reliable.

    (a) Undergraduate question: What is HDFS good at?
    The basic answer is:

    - LARGE files (> terabyte)
    - Streaming access (WORM)
    - Commodity hardware (failures are common)

    (b) Graduate question: What is HDFS **not** good at, and why?
    The basic answer is:

    - Low-latency data access (HDFS is based on an RPC model)
    - Lots of small files (overhead per file is constant)
    - Multiple writers, writes only at the end of a file

14. The HDFS has two basic types of nodes that do the “heavy” lifting to ensure a robust and reliable environment.

    (a) Undergraduate question: What are these nodes called, and what do they contain/maintain?
    The basic answer is:

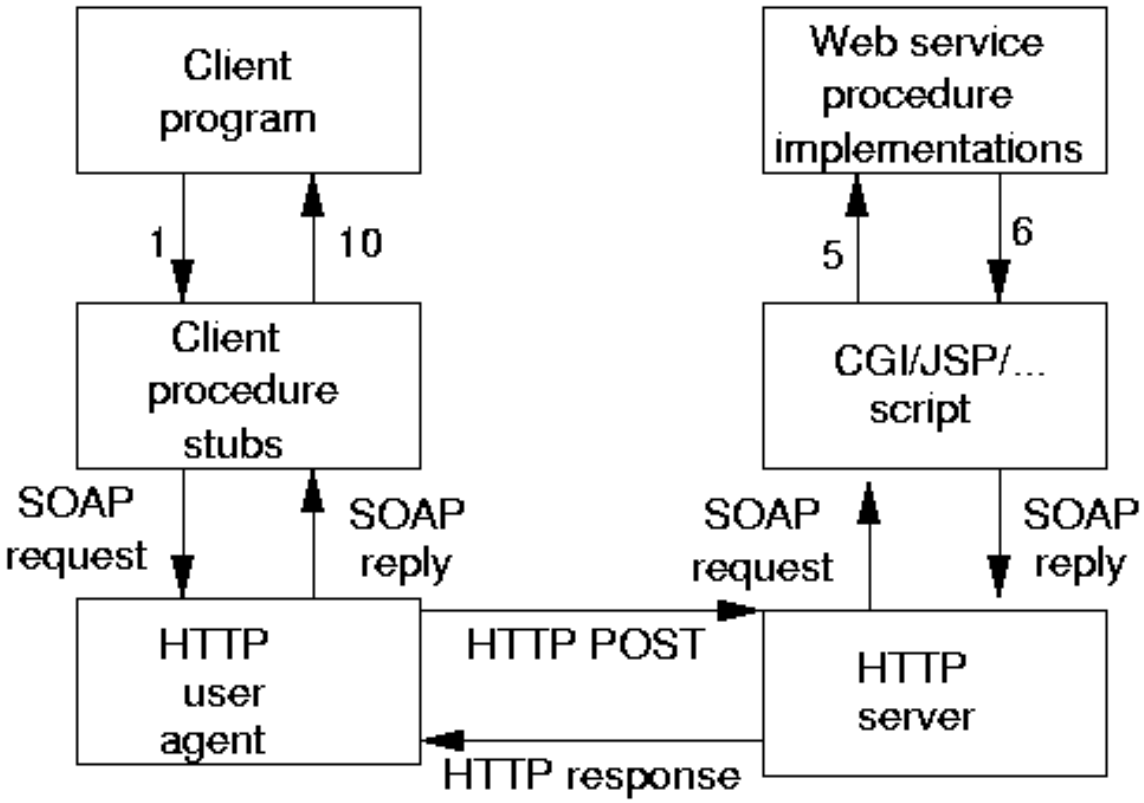
    - Namenode contains meta-data about files
    - Datanodes contain blocks

    (b) Graduate question: What is the effect if one of these nodes is lost?
    The basic answer is:

    - Namenode contains meta-data about files
    - Datanodes contain blocks
    - Loss of a datanode can be detected and a replicate created
    - Loss of a namenode is catastrophic, the entire HFDS is lost.

15. The HFDS uses remote procedure call (RPC) technology to perform its functions. There are numerous steps involved in using RPC architecture.

(a) Undergraduate question: Draw a diagram of an RPC invocation.
The basic answer is:



(b) Graduate question: During an RPC invocation, data will be serialized and deserialized. What does that mean?
The basic answer is:
During serialization, complex memory structures are converted to a byte stream and then sent to another program. The receiving program converts the byte stream into its own memory structure for processing.

16. The HDFS has two primary concerns.

    (a) Undergraduate question: What are they?
    The basic answer is:

    - Data integrity – ensuring that data is complete and intact
    - Data compression

    (b) Graduate question: Give two ways that the HDFS meets each of its primary concerns.
    The basic answer is:

    - Data integrity – ensuring that data is complete and intact
        i. Checks CRCs
        ii. Bit rot
        iii. Creates new replications when and where necessary
    - Data compression
        i. Minimizing data size
        ii. Network activity
        iii. Adds processing time

17. The HDFS will split large files across the network to meet its requirements for robustness and resilience. A large file (1 GByte) will be split into many pieces. One mapper will read and process from each split.

    (a) Undergraduate question: What size are the splits?
    The basic answer is:
    64 MBytes

    (b) Graduate question: How are compressed files split?
    The basic answer is:
    Depends on the type of compression. Not all compressed files are splitable.

Table 4-1. A summary of compression formats

Compression format	Tool	Algorithm	Filename extension	Splittable
DEFLATE[a]	N/A	DEFLATE	*.deflate*	No
gzip	*gzip*	DEFLATE	*.gz*	No
bzip2	*bzip2*	bzip2	*.bz2*	Yes
LZO	*lzop*	LZO	*.lzo*	No[b]
Snappy	N/A	Snappy	*.snappy*	No

18. All Hadoop jobs are executed based on a job scheduler. There are three different types of schedulers available that can be used. Each has different performance characteristics based on the type of job that is submitted.

    (a) Undergraduate question: What are the three types?
    The basic answer is:

    - FIFO
    - Fair
    - Capacity

    (b) Graduate question: How are they different?
    The basic answer is:

    - FIFO – default in Hadoop ver. 1 – first come first served
    - Fair – also available – jobs placed in user pool, one job per pool is scheduled
    - Capacity – default in Hadoop ver. 2 – similar to Fair, but adds priorities and relationships between pools

19. A notional view of a computer has a few functional components. Each component is separated by the next component by a layer. These layers allow components to be "swapped" in and out without affecting other components. Understanding the layers is the key to "virtualizing" a computer or a file system.

    (a) Undergraduate question: What are the four functional components that we talked about in class?
    The basic answer is:

    - User
    - Application
    - Hardware
    - Operating system

    (b) Graduate question: What do these components do?
    The basic answer is:

    - User — the person (or thing) that want's something done
    - Application — the program that does the work
    - Hardware — the silicone, copper, other tangibles that generate heat
    - Operating system — schedules access to the hardware and CPU functions.

20. Computer hardware virtualization means (among other things), that a data center can support the same user base with fewer physical computers.

    (a) Undergraduate question: What are three benefits from virualization?
    The basic answer is:

        i. Lower power
        ii. Lower cooling
        iii. Cheaper upgrade path

    (b) Graduate question: What is the ultimate limit on the number of virtual machines a physical machine can support?
    The basic answer is:
    The hardware (CPU cores, RAM, disk space, etc.).