Final Examination

CSC230 Database Technologies for Analytics

22 November 2016

1. Read the LinkedIn pages of the alumni of our college whose names you will find on the Moodle page for our course.

   (a) Where did Cody Dunne earn his Ph.D.?

   (b) Where did Zaw Win earn his MBA?

   (c) At which Silicon Valley firm did Sarah Gilliland work?

   (d) What is the name of the company that Jennifer Chang has created?

   (e) How did David Zabner continue his study of programming following the completion of his studies at Cornell College?

   (f) For which public agency does Myka Forrest work?

   (g) What is Brian McMillin's role at the Meredith Corporation?

   (h) Where did Thao Nguyen find an internship?

   (a) Cody Dunne earned his Ph.D. at the University of Maryland.

   (b) Zaw Win earned his MBA at Bentley University in Waltham, Massachusetts.

   (c) Sarah Gilliland worked at Facebook.

   (d) Jennifer Chang is a co-founder of Complex Computations.

   (e) David Zabner completed the program at Startup Institute Boston and now teaches others how to build Web sites and applications at the Startup Institute.

   (f) Myka Forrest develops software for the Grant Wood Area Agency. She has supported the education of students with special needs through her work.

   (g) Brian McMillin is the Director of Strategy and Development at the Meredith Corporation.

(h) Thao Nguyen completed an internship at Aegon Americas. She used SQL in her analysis of investments.

2. What do each of these Linux commands accomplish?

    (a)   `scp opportunities.html isaacasimov@cs1.cornellcollege.edu:`

    (b)   `ssh isaacasimov@cs1.cornellcollege.edu`

    (c)   `mv opportunities.html /var/www/html/cs230/isaac`

    (d)   `chmod ugo+x np.bash`

(a) The command copies a file (`opportunities.html`) to a server (`cs1.cornellcollege.edu`) in the home folder of the account that belongs to `isaacasimov`.

(b) The command logs a user into the account that is owned `isaacasimov`.

(c) The command moves a file to the folder to which the Apache HTTP server can find it and send it to a client's Web browser.

(d) The command changes the permission on a file. The file's name is `np.bash`. The command makes this file executable (x) by all people. This includes the owner (user:u) of the file, members of the owner's team (group:g), and everyone else (others:o) who use the computer.

3. Here are the contents of a CSV file:

```
Acadia National Park, Bar Harbor, Maine
Biscayne National Park, Homestead, Florida
Voyageurs National Park, International Falls, Minnesota
```

(a) If you edit this file in Emacs, position the cursor at the beginning of the first line, and type the following sequence of commands, what is the result?

```
C-x ( C-s , C-b ' ESC F ESC B ' C-a, C-n C-x )
```

(b) Now what happens if you type `C-x e`?

(a) The macro places a single quotation mark at the end of the name of the national park (Acadia National Park) and another single quotation mark at the start of the name of the city (Bar Harbor).

(b) This command executes the macro again on the next line.

4. How might we justify the effort required to learn the commands of Linux or Emacs?

The creators of this software were computer scientists who wanted to make other developers of software efficient in their work. The commands in this software are short sequences of key strokes. They are difficult to learn, but once learned, easy to type quickly.

5. *The Internet of Things: The Death of a Traditional Database?* is an article about the future of the technology that we have studied this month. What does the author mean by "self-*" characteristics?

The author described software that is self-managing, self-repairing, and self-tuning.

6. The white paper from MongoDB that we read distinguishes between operational and analytical uses of databases. What is that distinction?

Operational uses typically provide services directly to an organization's clients. For example, students enroll in a course, report hours worked, check grades by connecting to a college's database. Many operational uses provide services 24 hours per day and 7 days per week.

Analytical uses allow experts in an organization to measure performance, detect trends, and make predictions. For example, a prospect researcher in the college's fund-raising office examines records of past giving to develop models that will allow the identification of alumni who are most likely to respond to invitations to make gifts to the college.

7. The white paper from MongoDB contains the word "idempotent." What does that word mean?

A command is idempotent if successive executions of the command produce the same result.

8. Michael Stonebraker developed the PostgreSQL database management system.

   (a) What kinds of data did PostgreSQL store that previous database management system could not store?

   (b) For what kinds of applications did this make PostgreSQL suited?

   (a) Previous database management systems could work with text and numbers. PostgreSQL can work with geometric data, including points, line segments, and polygons.

   (b) Geometric data is important in computer-aided design systems and geographic information systems.

9. What is a principal purpose of normalization of databases?

Normalization removes redundancy. This makes it easier to assure consistency in the data. Redundancy (the storage of the same value in multiple locations) creates the possiblity of updating a one copy of the data but not another.

10. It is often helpful to look at a design problem in several different ways. Designers of databases often use a diagramatic notation in their work.

What is the name of that kind of diagram?

Designers of databases draw entity-relation diagrams.

11. JSON is a popular format for representing, storing, and sharing data.

    What is the JSON representation of this table?

peak	elevation
Katahdin	5267
Gannett	13809

    ```
    [
      { "peak" : "Katahdin", "elevation" : 5267 },
      { "peak" : "Gannett", "elevation" : 13809 }
    ]
    ```

12. XML is, like JSON, a popular format for representing, storing, and sharing data. Like JSON, XML data is also self-describing.

    Find a small example of small XML on the Web.

    Use your example to explain what "self-describing" means.

    Here is an example of XML.

    ```
    <?xml version="1.0" encoding="UTF-8"?>
    <note>
      <to>Tove</to>
      <from>Jani</from>
      <heading>Reminder</heading>
      <body>Don't forget me this weekend!</body>
    </note>
    ```

    This XML describes a note. We know this because it starts with a tag that contains the word "note." Like other notes that we have seen, it contains a part ("to") to hold the name of the person to whom the note is addressed, a part ("from") to hold the name of the author of the note, a part ("heading") to hold the subject of the note, and a part ("body") to hold the message.

5

13. To what part (or parts) of the ACID criteria might a NoSQL database management system not fully conform?

Because the data could be distributed over a very large number of nodes, because those nodes could be separated over a very large geographic region, and because the system might have to handle a very large number of transactions at a very rapid rate, the system might not be able to keep the data consistent at all times.

The "C" in ACID stands for "consistent."

Some NoSQL database systems do not assure consistency, but only eventual consistency.

14. How should we expect the form or nature of data in our databases to change in the future?

We are used to working with structured data. We can expect demands to work with semi-structured and unstructured data.

We are used to complete data. We can expect demands to work with incomplete data.

We are used to working with certain values. We can expect demands to work with data whose values are known only with probabilities that are less than certain.

We are used to working with a finite range of possible values and uniform representations of values.

There will be a greater variety of data in the future.

15. How should we expect the sources of data to change in the future?

We are used to data that people enter through forms.

We can expect data from very large numbers of sensors and embedded computers. These devices will provide great volumes of data at great velocities.