CS 644: Homework 1

Instructions. Answer the following multiple choice questions by selecting all correct choices. Some of the questions will have more than one correct choice.

Select all correct choices to receive full credit!

1. (6 points) **Big data properties**.

   (a) "Big Data" concerns which of the following types of data?

   ☐ structured ☐ semi-structured ☐ unstructured ☐ all of these

   (b) JSON and XML are examples of which type of data?

   ☐ structured ☐ unstructured ☐ semi-structured ☐ none of these

   (c) Which two of the following statements are true of unstructured data?

   ☐ It is generally easier to analyze than other types of data.
   ☐ It is often referred to as "messy" data.
   ☐ It fits neatly into a schema.
   ☐ It is the most widespread type of data.
   ☐ It is usually found in tables.

2. (6 points) **Hardware and Architecture**.

   (a) What kind of hardware is typically used for big data applications?

   ☐ High-performance PCs
   ☐ Low-cost, commodity hardware
   ☐ Dumb terminal
   ☐ None of the above

   (b) What is "commodity" hardware?

   ☐ High-performance hardware
   ☐ Discarded or second-hand hardware
   ☐ Generic, low-specification, industry-grade hardware
   ☐ Hardware used for trading commodities (e.g., gold, silver, soy-beans)

   (c) Which of the following describes a drawback of traditional relational database management system (or RDBMS) when used for big data applications?

   ☐ RDBMS cannot easily handle the massive volumes of data that have become common in the past two decades.
   ☐ RDBMS for big data requires more processors and memory, which is expensive to scale.
   ☐ Most data found in the wild is semi-structured or unstructured which must be curated and structured before it can be stored in an RDBMS.
   ☐ RDBMS cannot capture the data coming in at high velocity.
   ☐ All of the above.

3. (8 points) **ETL**.

(a) The process that corrects errors and inconsistencies is called *data* _____.

☐ *aggregation* ☐ *cleaning* ☐ *integration* ☐ *transformation* ☐ *reduction*

(b) The process of combining data from different sources into a unified data view is called *data* _____.

☐ *aggregation* ☐ *cleaning* ☐ *integration* ☐ *transformation* ☐ *reduction*

(c) Modifying and converting data into a format acceptable for inserting in a database is called *data* _____.

☐ *aggregation* ☐ *cleaning* ☐ *integration* ☐ *transformation* ☐ *reduction*

(d) The process of collecting the raw data, transmitting the data to a storage platform and preprocessing them is called *data* _____.

☐ *aggregation* ☐ *cleaning* ☐ *integration* ☐ *transformation* ☐ *reduction*

4. (6 points) Miscellany.

(a) What are the "big three" cloud storage service providers? (Select three.)

☐ Amazon **AWS S3**
☐ Facebook **Facespace**
☐ Google **GCP**
☐ Microsoft **Azure**
☐ Twitter **Birdhouse**

(b) Which of the following are programming paradigms? (Select three.)

☐ Declarative
☐ Functional
☐ Hadoop
☐ Imperative (Procedural)
☐ MapReduce
☐ NoSQL

(c) What three concepts characterize a purely functional programming language?

☐ immutability
☐ input/output (I/O)
☐ no side effects
☐ procedural
☐ referential transparency