

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

(a) “Big Data” concerns which of the following types of data? (Choose one.)

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

Explanation. Big Data is a blanket term for the data that are too large in size and complex in nature, and which may be structured, unstructured, or semi-structured, and may also be arriving at high velocity.

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

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

Explanation. Semi-structured data are that which have a structure but do not fit into the relational database. Semi-structured data are organized, which makes it easier for analysis when compared to unstructured data. JSON and XML are examples of semi-structured data.

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

- ☐ 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? (Choose one.)

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

Explanation. Big data uses low-cost commodity hardware to make cost-effective solutions.

(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)

Explanation. Commodity hardware is a low-cost, low performance, and low specification functional hardware with no distinctive features.

- (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

Explanation. The data-cleaning process fills in the missing values, corrects the errors and inconsistencies, and removes redundancy in the data to improve the data quality.

- (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

Explanation. Data transformation refers to transforming or consolidating the data into an appropriate format that is acceptable by the big data database and converting them into logical and meaningful information for data management and analysis.

- (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***

FOR PERSONAL USE ONLY – DO NOT DISTRIBUTE