DATA 100: Vitamin 12 Solutions

May 7, 2019

1 Data Warehouse

A data warehouse ____ and organizes historical data from ____ source(s).

☑ collects, multiple

□ distributes, a single

□ gathers, large

□ collects, water

2 ETL

The acronym ETL stands for:

□ extra, transform, load

☑ extract, transform, load

□ extract, transition, load

□ extra, transform, look

3 Data Lake

Which of the following issues are associated with data lakes?

□ Standardized schemas

□ Careful governance and planning

☑ Large amounts of uncleaned data

☑ Limited compatibility with existing software tools

4 Distributed File Systems

Distributed file systems need to be fault tolerant because:

□ massive files are spread across multiple machines.

□ files are processed in parallel.

☑ they are often built using cheap commodity hardware.

5 Parallelism

Which of the following programming paradigms enable programs to be executed in parallel across large datasets using data processing frameworks like Spark?

□ object oriented programming

☑ functional programming

☑ declarative programming