

# INT312 BIG DATA

## MOOC Course Solutions

### Why Big Data and Where Did it Come From?

?

1. Which of the following is an example of big data utilized in action today?

- ☐ Wi-Fi Networks
- ☐ Individual, Unconnected Hospital Databases
- ☐ The Internet
- ☒ Social Media

✓ **Correct**

See [this video](#) for examples of this concept.

2. What reasoning was given for the following: why is the "data storage to price ratio" relevant to big data?

- ☐ Companies can't afford to own, maintain, and spend the energy to support large data storage unless the cost is sufficiently low.
- ☐ Larger storage means easier accessibility to big data for every user because it allows users to download in bulk.
- ☐ It isn't, it was just an arbitrary example of big data usage.
- ☒ Lower prices mean larger storage becomes easier to access for everyone, creating bigger amounts of data for client-facing services to work with.

✓ **Correct**

See [this video](#) to review.

3. What is the best description of personalized marketing enabled by big data?

- ☒ Being able to use personalized data from every single customer for personalized marketing needs.
- ☐ Marketing to each customer on an individual level and suiting to their needs.
- ☐ Being able to obtain and use customer information for groups of consumers and utilize them for marketing needs.

✓ **Correct**

See [this video](#) for examples of this concept.

4. Of the following, which are some examples of personalized marketing related to big data?

- ☐ A survey that asks your age and markets to you a specific brand.
- ☐ News outlets gathering information from the internet in order to report them to the public.
- ☒ Facebook revealing posts that cater towards similar interests.

✓ **Correct**

See [this video](#) [↗](#) for examples of this concept.

5. What is the workflow for working with big data?

- ☒ Big Data -> Better Models -> Higher Precision
- ☐ Extrapolation -> Understanding -> Reproducing
- ☐ Theory -> Models -> Precise Advice

✓ **Correct**

See [this video](#) [↗](#) to review.

6. Which is the most compelling reason why mobile advertising is related to big data?

- ☒ Mobile advertising benefits from data integration with location which requires big data.
- ☐ Mobile advertising allows massive cellular/mobile texting to a wide audience, thus providing large amounts of data.
- ☐ Mobile advertising in and of itself is always associated with big data.
- ☐ Since almost everyone owns a cell/mobile phone, the mobile advertising market is large and thus requires big data to contain all the information.

✓ **Correct**

See [this video](#) [↗](#) for examples of this concept.

7. What are the three types of diverse data sources?

1 / 1 point

- ☐ Information Networks, Map Data, and People
- ☐ Machine Data, Map Data, and Social Media
- ☒ Machine Data, Organizational Data, and People
- ☐ Sensor Data, Organizational Data, and Social Media



Correct

See [this video](#) to review.

8. What is an example of machine data?

1 / 1 point

- ☐ Sorted data from Amazon regarding customer info.
- ☒ Weather station sensor output.
- ☐ Social Media



Correct

See [this video](#) to review.

9. What is an example of organizational data?

1 / 1 point

- ☐ Satellite Data
- ☒ Disease data from Center for Disease Control.
- ☐ Social Media



Correct

See [this video](#) for examples of this concept.

10. Of the three data sources, which is the hardest to implement and streamline into a model?

1 / 1 point

- ☐ Machine Data
- ☐ Organizational Data
- ☒ People

✓ **Correct**  
See [this video](#) to review.

11. Which of the following summarizes the process of using data streams?

1 / 1 point

- ☒ Integration -> Personalization -> Precision
- ☐ Big Data -> Better Models -> Higher Precision
- ☐ Theory -> Models -> Precise Advice
- ☐ Extrapolation -> Understanding -> Reproducing

✓ **Correct**  
See [this video](#) to review.

12. Where does the real value of big data often come from?

1 / 1 point

- ☐ Having data-enabled decisions and actions from the insights of new data.
- ☐ Size of the data.
- ☒ Combining streams of data and analyzing them for new insights.
- ☐ Using the three major data sources: Machines, People, and Organizations.

✓ **Correct**  
See [this video](#) to review.

13. What does it mean for a device to be "smart"?

1 / 1 point

- ☒ Collect data and services autonomously.
- ☐ Having a specific processing speed in order to keep up with the demands of data processing.
- ☐ Must have a way to interact with the user.

✓ **Correct**  
See [this video](#) to review.

14. What does the term "in situ" mean in the context of big data?

1 / 1 point

- ☐ The sensors used in airplanes to measure altitude.
- ☒ Bringing the computation to the location of the data.
- ☐ Accelerometers.
- ☐ In the situation

✓ **Correct**  
See [this video](#) to review.

15. Which of the following are reasons mentioned for why data generated by people are hard to process? Choose all that apply.

1 / 1 point

- ☒ The velocity of the data is very high.

✓ **Correct**  
See [this video](#) to review.

- ☒ Skilled people to analyze the data are hard to come by.

✓ **Correct**  
See [this video](#) to review.

- ☐ They cannot be modeled and stored.

☒ Very unstructured data.

✓ **Correct**

See [this video](#) to review.

16. What is the purpose of retrieval and storage; pre-processing; and analysis in order to convert multiple data sources into valuable data?

1 / 1 point

- ☐ Since the multi-layered process is built into the Neo4j database connection.
- ☐ To enable ETL methods.
- ☒ To allow scalable analytical solutions to big data.
- ☐ Designed to work like the ETL process.

✓ **Correct**

See [this video](#) to review.

17. Which of the following are benefits of organization-generated data? Choose all that apply.

1 / 1 point

☒ Better Profit Margins

✓ **Correct**

See [this video](#) to review.

☒ Improved Safety

✓ **Correct**

See [this video](#) to review.

☒ Customer Satisfaction

✓ **Correct**

See [this video](#) to review.

☒ Higher Sales

18. What are data silos and why are they bad?

1 / 1 point

- ☐ Highly unstructured data. Bad because it does not provide meaningful results for organizations.
- ☐ A giant centralized database to house all the data produces within an organization. Bad because it is hard to maintain as highly structured data.
- ☐ A giant centralized database to house all the data production within an organization. Bad because it hinders opportunity for data generation.
- ☒ Data produced from an organization that is spread out. Bad because it creates unsynchronized and invisible data.



Correct

See [this video](#) to review.

19. Which of the following are benefits of data integration? Choose all that apply.

1 / 1 point

- ☒ Increase data availability.



Correct

See [this video](#) to review.

- ☒ Reduce data complexity.



Correct

See [this video](#) to review.

- ☐ Monitoring of data.

- ☒ Unify your data system.



Correct

See [this video](#) to review.

- ☒ Adds value to big data.



Correct

See [this video](#) to review.

- ☒ Increase data collaboration.


## V for the V's of Big Data

1.

1 / 1 point

Amazon has been collecting review data for a particular product. They have realized that almost 90% of the reviews were mostly a 5/5 rating. However, of the 90%, they realized that 50% of them were customers who did not have proof of purchase or customers who did not post serious reviews about the product. Of the following, which is true about the review data collected in this situation?

- ☐ Low Valence
- ☒ Low Veracity
- ☐ High Valence
- ☐ High Veracity
- ☐ Low Volume
- ☐ High Volume

 **Correct**See [this video](#)  for examples of this concept.

2. As mentioned in the slides, what are the challenges to data with a high valence?

1 / 1 point

- ☒ Complex Data Exploration Algorithms
- ☐ Difficult to Integrate
- ☐ Reliability of Data

 **Correct**See [this video](#)  to review.



3. Which of the following are the 6 V's in big data?

1 / 1 point

☒ Veracity

✓ **Correct**

See [this video](#)  to review.

☐ Vision

☒ Volume

✓ **Correct**

See [this video](#)  to review.

☒ Velocity

✓ **Correct**

See [this video](#)  to review.

☒ Valence

✓ **Correct**

See [this video](#)  to review.

☒ Variety

✓ **Correct**

See [this video](#)  to review.

☒ Value

✓ **Correct**

See [this video](#)  to review.

4. What is the veracity of big data?

1 / 1 point

- ☐ The size of the data.
- ☐ The speed at which data is produced.
- ☐ The connectedness of data.
- ☒ The abnormality or uncertainties of data.

✓ Correct

See [this video](#) to review.

5. What are the challenges of data with high variety?

1 / 1 point

- ☒ Hard to integrate.
- ☐ Hard in utilizing group event detection.
- ☐ The quality of data is low.
- ☐ Hard to perform emergent behavior analysis.

✓ Correct

See [this video](#) to review.

6. Which of the following is the best way to describe why it is crucial to process data in real-time?

1 / 1 point

- ☒ Prevents missed opportunities.
- ☐ More expensive to batch process.
- ☐ Batch processing is an older method that is not as accurate as real-time processing.
- ☐ More accurate.

✓ Correct

See [this video](#) to review.

7. What are the challenges with big data that has high volume?

1 / 1 point

- ☐ Effectiveness and Cost
- ☐ Storage and Accessibility
- ☐ Speed Increase in Processing
- ☒ Cost, Scalability, and Performance

✓ Correct

See [this video](#) to review.

# Data Science 101

1. Which of the following are parts of the 5 P's of data science and what is the additional P introduced in the slides?

☒ Process

✓ **Correct**

See [this video](#)  to review.

☐ Perception

☒ Purpose

✓ **Correct**

See [this video](#)  to review.

☒ People

✓ **Correct**

See [this video](#)  to review.

☒ Product

✓ **Correct**

See [this video](#)  to review.

☒ Programmability

✓ **Correct**

See [this video](#)  to review.

☒ Platforms

✓ **Correct**

See [this video](#)  to review.

2. Which of the following are part of the four main categories to acquire, access, and retrieve data?

☒ Remote Data

✓ **Correct**

See [this video](#) [↗](#) to review.

☐ Web Services

☒ Text Files

✓ **Correct**

See [this video](#) [↗](#) to review.

☒ Traditional Databases

✓ **Correct**

See [this video](#) [↗](#) to review.

☒ NoSQL Storage

✓ **Correct**

See [this video](#) [↗](#) to review.

3. What are the steps required for data analysis?

☐ Investigate, Build Model, Evaluate

☐ Regression, Evaluate, Classification

☒ Select Technique, Build Model, Evaluate

☐ Classification, Regression, Analysis

✓ **Correct**

See [this video](#) [↗](#) to review.

4. Of the following, which is a technique mentioned in the videos for building a model?

- ☒ Analysis
- ☐ Investigation
- ☐ Validation
- ☐ Evaluation

✓ **Correct**

See [this video](#)  to review.

5. What is the first step in finding a right problem to tackle in data science?

- ☐ Ask the Right Questions
- ☐ Define Goals
- ☒ Define the Problem
- ☐ Assess the Situation

✓ **Correct**

See [this video](#)  to review.

6. What is the first step in determining a big data strategy?

- ☐ Organizational Buy-In
- ☒ Business Objectives
- ☐ Build In-House Expertise
- ☐ Collect Data

✓ **Correct**

See [this video](#)  to review.

7. According to Ilkay, why is exploring data crucial to better modeling?

Data exploration... <complete the sentence>

- ☒ leads to data understanding which allows an informed analysis of the data.
- ☐ enables understanding of general trends, correlations, and outliers.
- ☐ enables histograms and others graphs as data visualization.
- ☐ enables a description of data which allows visualization.

✓ **Correct**

See [this video](#)  to review.

8. Why is data science mainly about teamwork?

- ☐ Engineering solutions are preferred.
- ☐ Exhibition of curiosity is required.
- ☐ Analytic solutions are required.
- ☒ Data science requires a variety of expertise in different fields.

✓ **Correct**

See [this video](#)  to review.

9. What are the ways to address data quality issues?

- ☒ Remove outliers.

✓ **Correct**

See [this video](#)  to review.

- ☒ Generate best estimates for invalid values.

✓ **Correct**

See [this video](#)  to review.

---

9. What are the ways to address data quality issues?

☒ Remove outliers.

✓ **Correct**

See [this video](#)  to review.

☒ Generate best estimates for invalid values.

✓ **Correct**

See [this video](#)  to review.

☐ Data Wrangling

☒ Merge duplicate records.

✓ **Correct**

See [this video](#)  to review.

☒ Remove data with missing values.

✓ **Correct**

See [this video](#)  to review.

10. What is done to the data in the preparation stage?

☐ Select Analytical Techniques

☐ Identify Data Sets and Query Data

☐ Build Models

☐ Retrieve Data

☒ Cleaning, Integrating, and Packaging

✓ **Correct**

See [this video](#)  to review.

# Foundations for Big Data

1. Which of the following is the best description of why it is important to learn about the foundations for big data?

1 / 1 point

- ☐ Foundations help you revisit calculus concepts required in the understanding of big data.
- ☐ Foundations is all that is required to show a mastery of big data concepts.
- ☐ Foundations stand the test of time.
- ☒ Foundations allow for the understanding of practical concepts in Hadoop.

✓ Correct

See [this video](#) to review.

2. What is the benefit of a commodity cluster?

1 / 1 point

- ☐ Much faster than a traditional super computer
- ☐ Prevents individual component failures
- ☐ Prevents network connection failure
- ☒ Enables fault tolerance

✓ Correct

See [this video](#) to review.

3. What is a way to enable fault tolerance?

1 / 1 point

- ☒ Data-Parallel Job Restart
- ☐ Distributed Computing
- ☐ Better LAN Connection
- ☐ System Wide Restart

✓ Correct

See [this video](#) to review.



---

4. What are the specific benefit(s) to a distributed file system?

1 / 1 point

☒ High Fault Tolerance



Correct

See [this video](#) to review.

☐ Large Storage

☒ Data Scalability



Correct

See [this video](#) to review.

☒ High Concurrency



Correct

See [this video](#) to review.

5. Which of the following are general requirements for a programming language in order to support big data models?

1 / 1 point

☐ Utilize Map Reduction Methods

☒ Optimization of Specific Data Types



Correct

See [this video](#) to review.

☒ Enable Adding of More Racks



Correct

See [this video](#) to review.

☒ Handle Fault Tolerance



Correct

See [this video](#) to review.

☒ Support Big Data Operations

---

# Intro to Hadoop

---

1. What does IaaS provide?

1 / 1 point

- ☒ Hardware Only
- ☐ Computing Environment
- ☐ Software On-Demand

✓ **Correct**  
See [this video](#) to review.

2. What does PaaS provide?

1 / 1 point

- ☒ Computing Environment
- ☐ Hardware Only
- ☐ Software On-Demand

✓ **Correct**  
See [this video](#) to review.

3. What does SaaS provide?

1 / 1 point

- ☐ Computing Environment
- ☒ Software On-Demand
- ☐ Hardware Only

✓ **Correct**  
See [this video](#) to review.

4. What are the two key components of HDFS and what are they used for?

1 / 1 point

- ☐ NameNode for block storage and Data Node for metadata.
  - ☒ NameNode for metadata and DataNode for block storage.
-

5. What is the job of the NameNode?

1 / 1 point

- ☒ Coordinate operations and assigns tasks to Data Nodes
- ☐ Listens from DataNode for block creation, deletion, and replication.
- ☐ For gene sequencing calculations.

✓ **Correct**  
See [this video](#) to review.

6. What is the order of the three steps to Map Reduce?

1 / 1 point

- ☐ Map -> Reduce -> Shuffle and Sort
- ☐ Shuffle and Sort -> Reduce -> Map
- ☐ Shuffle and Sort -> Map -> Reduce
- ☒ Map -> Shuffle and Sort -> Reduce

✓ **Correct**  
See [this video](#) to review.

7. What is a benefit of using pre-built Hadoop images?

1 / 1 point

- ☐ Less software choices to choose from.
- ☐ Quick prototyping, deploying, and guaranteed bug free.
- ☐ Guaranteed hardware support.
- ☒ Quick prototyping, deploying, and validating of projects.

✓ **Correct**  
See [this video](#) to review.

8. What are some examples of open-source tools built for Hadoop and what does it do?

1 / 1 point

- ☐ Zookeeper, analyze social graphs.
- ☐ Giraph, for SQL-like queries.
- ☒ Zookeeper, management system for animal named related components.
- ☐ Pig, for real-time and in-memory processing of big data.

✓ Correct

See [this video](#) to review.

9. What is the difference between low level interfaces and high level interfaces?

1 / 1 point

- ☐ Low level deals with interactivity while high level deals with storage and scheduling.
- ☒ Low level deals with storage and scheduling while high level deals with interactivity.

✓ Correct

See [this video](#) to review.

10. Which of the following are problems to look out for when integrating your project with Hadoop?

1 / 1 point

✓ Random Data Access

✓ Correct

See [this video](#) to review.

✓ Advanced Algorithms

✓ Correct

See [this video](#) to review.

✓ Infrastructure Replacement

✓ Correct

☒ Task Level Parallelism



Correct

See [this video](#) to review.

☐ Data Level Parallelism

11. As covered in the slides, which of the following are the major goals of Hadoop?

1 / 1 point

☒ Facilitate a Shared Environment



Correct

See [this video](#) to review.

☐ Latency Sensitive Tasks

☒ Provide Value for Data



Correct

See [this video](#) to review.

☒ Optimized for a Variety of Data Types



Correct

See [this video](#) to review.

☒ Enable Scalability



Correct

See [this video](#) to review.

☒ Handle Fault Tolerance



Correct

See [this video](#) to review.

12. What is the purpose of YARN?

1 / 1 point

- ☒ Allows various applications to run on the same Hadoop cluster.
- ☐ Enables large scale data across clusters.
- ☐ Implementation of Map Reduce.

✓ Correct

See [this video](#) to review.

13. What are the two main components for a data computation framework that were described in the slides?

1 / 1 point

- ☐ Resource Manager and Container
- ☐ Node Manager and Container
- ☒ Resource Manager and Node Manager
- ☐ Node Manager and Applications Master
- ☐ Applications Master and Container

✓ Correct

See [this video](#) to review.

## Running Hadoop MapReduce Programs Quiz

1. Ans 6

2. Ans 4