**According to analysts, for what can traditional IT systems provide a foundation when they’re integrated with big data technologies like Hadoop?**

**(A)** Big data management and data mining

**(B)** Data warehousing and business intelligence

**(C)** Management of Hadoop clusters

**(D)** Collecting and storing unstructured data

Answer

**A**

MCQ No - 2

**What are the main components of Big Data?**

**(A)** MapReduce

**(B)** HDFS

**(C)** YARN

**(D)** All of these

Answer

**D**

MCQ No - 3

**What are the different features of Big Data Analytics?**

**(A)** Open-Source

**(B)** Scalability

**(C)** Data Recovery

**(D)** All the above

Answer

**D**

MCQ No - 4

**According to analysts, for what can traditional IT systems provide a foundation when they’re integrated with big data technologies like Hadoop?**

**(A)** Big data management and data mining

**(B)** Data warehousing and business intelligence

**(C)** Management of Hadoop clusters

**(D)** Collecting and storing unstructured data

Answer

**A**

MCQ No - 5

**What are the four V’s of Big Data?**

**(A)** Volume

**(B)** Velocity

**(C)** Variety

**(D)** All the above

Answer

**D**

**[Big Data Analytics(2180710)](http://gtu-mcq.com/BE/Computer-Engineering/Semester-8/2180710/Units" \o "Click Here to Get Units ofBig Data Analytics(2180710))**

* [1. INTRODUCTION TO BIG DATA](http://gtu-mcq.com/BE/Computer-Engineering/Semester-8/2180710/_5TbpQVPx_qYPVF_FlWmxA/MCQs" \o "MCQs of the unit INTRODUCTION TO BIG DATA)
* [2. INTRODUCTION TO HADOOP AND HADOOP ARCHITECTURE](http://gtu-mcq.com/BE/Computer-Engineering/Semester-8/2180710/0oVObO6nFaD0nKDOd35Ujg/MCQs" \o "MCQs of the unit INTRODUCTION TO HADOOP AND HADOOP ARCHITECTURE)
* [3. HDFS, HIVE AND HIVEQL, HBASE](http://gtu-mcq.com/BE/Computer-Engineering/Semester-8/2180710/4Jj28xrRw3q9ida-tI3RvQ/MCQs" \o "MCQs of the unit HDFS, HIVE AND HIVEQL, HBASE)
* [4. SPARK](http://gtu-mcq.com/BE/Computer-Engineering/Semester-8/2180710/sK38IwVWkrKxMmuZg82BRQ/MCQs" \o "MCQs of the unit SPARK)
* [5. NoSQL](http://gtu-mcq.com/BE/Computer-Engineering/Semester-8/2180710/XQpwjq2bsy0ovxbcl5IZ6Q/MCQs" \o "MCQs of the unit NoSQL)
* [6. Data Base for the Modern Web](http://gtu-mcq.com/BE/Computer-Engineering/Semester-8/2180710/zh571mghkBh4GKk7rTAeqA/MCQs" \o "MCQs of the unit Data Base for the Modern Web)

#### ****All of the following accurately describe Hadoop, EXCEPT:****

**(A)** Open-source

**(B)** Real-time

**(C)** Java-based

**(D)** Distributed computing approach

Answer

**B**

MCQ No - 7

#### ****\_\_\_\_\_\_\_\_\_\_\_ is general-purpose computing model and runtime system for distributed data analytics.****

**(A)** Mapreduce

**(B)** Drill

**(C)** Oozie

**(D)** None of the above

Answer

**A**

MCQ No - 8

#### ****The examination of large amounts of data to see what patterns or other useful information can be found is known as****

**(A)** Data examination

**(B)** Information analysis

**(C)** Big data analytics

**(D)** Data analysis

Answer

**C**

MCQ No - 9

#### ****Big data analysis does the following except****

**(A)** Collects data

**(B)** Spreads data

**(C)** Organizes data

**(D)** Analyzes data

Answer

**B**

MCQ No - 10

#### ****What makes Big Data analysis difficult to optimize?****

**(A)** Big Data is not difficult to optimize

**(B)** Both data and cost effective ways to mine data to make business sense out of it

**(C)** The technology to mine data

**(D)** All of the above

Answer

**B**

#### ****The new source of big data that will trigger a Big Data revolution in the years to come is****

**(A)** Business transactions

**(B)** Social media

**(C)** Transactional data and sensor data

**(D)** RDBMS

Answer

**C**

MCQ No - 12

#### ****The unit of data that flows through a Flume agent is****

**(A)** Log

**(B)** Row

**(C)** Event

**(D)** Record

Answer

**C**

MCQ No - 13

#### ****Listed below are the three steps that are followed to deploy a Big Data Solution except****

**(A)** Data Ingestion

**(B)** Data Processing

**(C)** Data dissemination

**(D)** Data Storage

Answer

**C**

MCQ No - 14

#### ****Check below the best answer to "which industries employ the use of so-called "Big Data" in their day to day operations?****

**(A)** Weather forecasting

**(B)** Marketing

**(C)** Healthcare

**(D)** All of the above

Answer

**D**

MCQ No - 15

#### ****There are almost as many bits of information in the digital universe as there are stars in the actual universe?****

**(A)** True

**(B)** False

Answer

**A**

MCQ No - 16

#### ****The word 'Big data' was coined by****

**(A)** Roger Mougalas

**(B)** John Philips

**(C)** Simon Woods

**(D)** Martin Green

Answer

**A**

MCQ No - 17

#### ****The word 'Big Data' was coined in the year****

**(A)** 2000

**(B)** 1970

**(C)** 1998

**(D)** 2005

Answer

**C**

MCQ No - 18

#### ****Concerning the Forms of Big Data, which one of these is odd?****

**(A)** Structured

**(B)** Unstructured

**(C)** Processed

**(D)** Semi-Structured

Answer

**C**

MCQ No - 19

#### ****Big Data applications benefit the media and entertainment industry by****

**(A)** Predicting what the audience wants

**(B)** Ad targeting

**(C)** Scheduling optimization

**(D)** All of the above

Answer

**D**

MCQ No - 20

#### ****The feature of big data that refers to the quality of the stored data is \_\_\_\_\_\_****

**(A)** Variety

**(B)** Volume

**(C)** Variability

**(D)** Veracity

Answer

**D**

### Question 1

What is the difference between interval/ratio and ordinal variables?

**a)**The distance between categories is equal across the range of interval/ratio data.

**Feedback:**

The data that we gather varies from person to person. People are of different ages, have different income levels and prefer to do some things more than other people. We call these things variables just because their values vary from person to person. Analysis of quantitative data starts by trying to understand what kinds of variables we are dealing with. A person's age is an example of an interval/ratio variable, because ages are measured in years. We can do a lot of statistical analysis on this kind of variable because the interval (one year) is the same for everybody in our data-set. Some variables are called 'dichotomous', meaning all possible answers are of one of two types (male/female, for example). We call those variables 'nominal', which we can, literally, only "name", like many types of job occupation, for example. Finally, we refer to some variables as 'ordinal', which means we can only place the values in an order of first, second, third and so on, without considering the gap between the first and second, or whether it was the same as between second and third. Apart from dichotomous variables, all others can be rank-ordered.

### Question 2

What is the difference between a bar chart and a histogram?

**c)**There are no gaps between the bars on a histogram.

**Feedback:**

Histograms are used to display interval/ratio variables, which involve a continuous range of values, and so there are no gaps between the bars that represent each category. Bar charts, on the other hand, display nominal or ordinal data, which fall into discrete categories.

### Question 3

What does the term 'outlier' mean?

**d)**An extreme value at either end of a distribution

**Feedback:**

When we calculate a simple average, the 'arithmetic mean', we have to remember that a wide range of values can give the same average as a narrow range and that extreme values could make a simple average fairly meaningless. These values are called 'outliers', extremely high or low values in a distribution that threaten to skew the results. The 'median' is useful, in this regard, because it simply identifies the mid-point in a whole array of values, giving us a measure of the significance of the arithmetic mean.

### Question 4

What is the function of a contingency table, in the context of bivariate analysis?

**Correct answer:**

**b)**It summarizes the frequencies of two variables so that they can be compared.

**Feedback:**

'Bivariate' analysis means that we are analysing two variables together, usually to see if any co-relation exists between them. There are various techniques available for this, one of which is a contingency table. This technique is principally used to compare nominal variables with another type, where the frequencies (in numbers or percentages) of the two different variables are simultaneously analysed to identify patterns of association between them.

### Question 5

If there were a perfect positive correlation between two interval/ratio variables, the Pearson's *r* test would give a correlation coefficient of:

**Correct answer:**

**b)**+1

**Feedback:**

A coefficient is a measure of the degree to which two sets of numbers co-relate. If the variables always move in 'lock-step' with each other, we call that a 'perfect' correlation. Sometimes the variables move in the same direction as each other, a 'positive' correlation and sometimes in the opposite direction, a 'negative' correlation. Pearson's *r* test gives an answer of +1 when there is a perfect positive correlation between interval/ratio variables and -1 when there is a perfect negative correlation between them.

### Question 6

What is the name of the test that is used to assess the relationship between two ordinal variables?

**Correct answer:**

**a)**Spearman's *rho*

**Feedback:**

Pearson's *r* test is extremely valuable but limited to assessing correlations between interval/ratio variables. Spearman's *rho* test is a very similar technique which can be used on pairs of variables when either both are ordinal or one is ordinal and the other is interval/ratio. The result will lie between -1 and +1, indicating the range of possible correlation, from perfectly negative to perfectly positive. The phi coefficient is used for dichotomous variables and Cramer's *V* is a test of the strength of the relationship between nominal variables. Chi square, in brief, tests for the likelihood of relationships existing through mere chance, so is usually used in conjunction with the tests discussed in this question.

### Question 7

When might it be appropriate to conduct a multivariate analysis test?

**Correct answer:**

**d)**All of the above.

**Feedback:**

Multivariate analysis involves the analysis of three or more variables, and tends to be used when we have reason to suspect the nature of the relationship between two variables. Two variables can, indeed, be related to each other but perhaps in a more complex way than appears at first sight. Perhaps when a number of factors co-exist the relationship between any two of them is strong. Multivariate analysis enables us to test for many types of cross-relationships between a number of variables, at once.

### Question 8

What is meant by a "spurious" relationship between two variables?

**Correct answer:**

**c)**A relationship that appears to be true because each variable is related to a third one.

**Feedback:**

One of the conditions under which it is appropriate to use multivariate analysis is when the relationship between two variables might be spurious: this means that the relationship, which seemed to exist, doesn't exist in reality. A third variable turns out, perhaps, to be responsible for the variation in both sets of values, and so they are not really related to each other, so their relationship was "spurious".

### Question 9

A test of statistical significance indicates how confident the researcher is about:

**Correct answer:**

**d)**generalising their findings from the sample to the population.

**Feedback:**

Tests of statistical significance allow the researcher to estimate how confident they can be that there is a real relationship between the variables they are studying and thus that their results can be generalised from the sample to the target population.

### Question 10

Setting the *p* level at 0.01 increases the chances of making a:

**Correct answer:**

**b)**Type II error

**Feedback:**

The *p* value represents the level of probability that an apparently significant relationship between variables was really just due to chance. If *p* is set at 0.01, this means that we would expect such a result in only 1 in 100 cases. This is a very stringent level, and while it means that the researcher can be more confident about a significant result if they find one, it also increases the chance of making a Type II error: confirming the null hypothesis when it should be rejected.