1. Spark is initially developed by which university
   Ans) Berkley
2. What are the characteristics of Big Data?
   Ans) Volume, Velocity and Variety
3. The main focus of Hadoop ecosystem is on
   Ans) Batch Processing
4. Streaming data tools available in Hadoop ecosystem are?
   Ans) Apache Spark and Storm
5. Spark has API's in? How many languages it supports
   Ans) Java, Scala, R and Python
6. Which kind of data can be processed by spark?
   Ans) Stored Data and Streaming Data
7. Spark can store its data in?
   Ans) HDFS, MongoDB and Cassandra

8. How spark engine runs?

Ans) Integrating with Hadoop and Standalone

9. In spark data is represented as?

Ans) RDDs

10. Which kind of data can be handled by Spark ?

Ans) Structured, Unstructured and Semi-Structured

11. Which among the following are the challenges in Map reduce?

Ans) Every Problem has to be broken into Map and Reduce phase

Collection of Key / Value pairs

High Throughput

12. Apache spark is a framework with?

Ans) Scheduling, Monitoring and Distributing Applications

13. Which of the features of Apache spark

Ans) DAG, RDDs and In- Memory processing

14) How much faster is the processing in spark when compared to Hadoop?

ANS) 10-100X

15) In spark data is represented as?

Ans) RDDs

16) List of Transformations

Ans) map(), filter(), flatmap(), groupBy(), groupByKey(), sample(), union(), join(), distinct(), keyBy(), partitionBy and zip().

17) List of Actions

Ans) getNumPartitions(), collect(), reduce(), aggregate(), max(), sum(), mean(), stdev(), countByKey().

18. Spark is developed in

Ans) Scala

19. Which type of processing Apache Spark can handle

Ans) Batch Processing, Interactive Processing, Stream Processing and Graph Processing

20. List two statements of Spark

Ans) Spark can run on the top of Hadoop

Spark can process data stored in HDFS

Spark can use Yarn as resource management layer

21. Spark's core is a batch engine? True OR False

Ans) True

22) Spark is 100x faster than MapReduce due to

Ans) In-Memory Computing

23) MapReduce program can be developed in spark? T / F

Ans) True

24. Programming paradigm used in Spark

Ans) Generalized

25. Spark Core Abstraction

Ans) RDD

26. Choose correct statement about RDD

Ans) RDD is a distributed data structure

27. RDD is

Ans) Immutable, Recomputable and Fault-tolerant

28. RDD operations

Ans) Transformation, Action and Caching

29. We can edit the data of RDD like conversion to uppercase? T/F

Ans) False

30. Identify correct transformation

Ans) Map, Filter and Join

31. Identify Correct Action

Ans) Reduce

32. Choose correct statement

Ans) Execution starts with the call of Action

33. Choose correct statement about Spark Context

Ans) Interact with cluster manager and Specify spark how to access cluster

34. Spark cache the data automatically in the memory as and when needed? T/F

Ans) False

35. For resource management spark can use

Ans) Yarn, Mesos and Standalone cluster manager

36. RDD can not be created from data stored on

Ans) Oracle

37. RDD can be created from data stored on

Ans) Local FS, S3 and HDFS

38. Who is father of Big data Analytics

- ✓ **Doug Cutting**

39. What are major Characteristics of Big Data

- ✓ **Volume, Velocity and Variety(3 V's)**

40. What is Apache Hadoop

- ✓ **Open-source Software Framework**

41. Who developed Hadoop

- ✓ **Doug Cutting**

42. Hadoop supports which programming framework

- ✓ **Java**

43. What is the heart of Hadoop

- ✓ **MapReduce**

44. What is MapReduce

- ✓ **Programming Model for Processing Large Data Sets.**

45. What are the Big Data Dimensions

- ✓ **4 V's**

46. What is the caption of Volume

- ✓ **Data at Scale**

47. What is the caption of Velocity

- ✓ **Data in Motion**

48. What is the caption of Variety

- ✓ **Data in many forms**

49. What is the caption of Veracity

- ✓ **Data Uncertainty**

50. What is the biggest Data source for Big Data

- ✓ **Transactions**

51. What is the biggest Analytic capability for Big Data

- ✓ **Query and Reporting**

52. What is the biggest Infrastructure for Big Data

- ✓ **Information integration**

53. What are the Big Data Adoption Stages

- ✓ **Educate, Explore, Engage and Execute**

54. What is Mahout

- ✓ **Algorithm library for scalable machine learning on Hadoop**

55. What is Pig

- ✓ **Creating MapReduce programs used with Hadoop.**

56. What is HBase

- ✓ **Non-Relational Database**

57. What is the biggest Research Challenge for Big Data

- ✓ **Heterogeneity , Incompleteness and Security**

58. What is Sqoop

- ✓ **Transferring bulk data between Hadoop to Structured data.**

59. What is Oozie

- ✓ **Workflow scheduler system to manage Hadoop jobs.**

60. What is Hue

- ✓ **Web interface that supports Apache Hadoop and its ecosystem**

61. What is Avro

- ✓ **Avro is a data serialization system.**

62. What is Giraph

- ✓ **Iterative graph processing system built for high scalability.**

63. What is Cassandra

- ✓ **Cassandra does not support joins or sub queries, except for batch analysis via Hadoop**

64. What is Chukwa

- ✓ **Chukwa is an open source data collection system for monitoring large distributed systems**

65. What is Hive

- ✓ **Hive is a data warehouse on Hadoop**

66. What is Apache drill

- ✓ **Apache Drill is a distributed system for interactive analysis of large-scale datasets.**

67. What is HDFS

- ✓ **Hadoop Distributed File System (HDFS)**

68. Facebook generates how much data per day

- ✓ **25TB**

69. What is BIG DATA?

- ✓ **Big Data is nothing but an assortment of such a huge and complex data that it becomes very tedious to capture, store, process, retrieve and analyze it with the help of on-hand database management tools or traditional data processing techniques.**

70. What is HUE expansion

- ✓ **Hadoop User Interface**

71.Can you give some examples of Big Data?

- ✓ There are many real life examples of Big Data! Facebook is generating 500+ terabytes of data per day, NYSE (New York Stock Exchange) generates about 1 terabyte of new trade data per day, a jet airline collects 10 terabytes of censor data for every 30 minutes of flying time.

72. Can you give a detailed overview about the Big Data being generated by Facebook?

- ✓ As of December 31, 2012, there are 1.06 billion monthly active users on Facebook and 680 million mobile users. On an average, 3.2 billion likes and comments are posted every day on Facebook. 72% of web audience is on Facebook. And why not! There are so many activities going on Facebook from wall posts, sharing images, videos, writing comments and liking posts, etc.

73. What are the three characteristics of Big Data?

- ✓ The three characteristics of Big Data are: **Volume:** Facebook generating 500+ terabytes of data per day**. Velocity:** Analyzing 2 million records each day to identify the reason for losses. **Variety:** images, audio, video, sensor data, log files, etc.

74. How Big is 'Big Data'?

- ✓ With time, data volume is growing exponentially. Earlier we used to talk about **Megabytes** or **Gigabytes.** But time has arrived when we talk about data volume in terms of **terabytes, petabytes** and also **zettabytes**! Global data volume was around 1.8ZB in 2011 and is expected to be 7.9ZB in 2015.

75. How analysis of Big Data is useful for organizations?

- ✓ Effective analysis of Big Data provides a lot of business advantage as organizations will learn which areas to focus on and which areas are less important.

76. Who are 'Data Scientists'?

- ✓ Data scientists are experts who find solutions to analyze data. Just as web analysis, we have data scientists who have good business insight as to how to handle a business challenge.

77. What is Hadoop?

- ✓ Hadoop is a framework that allows for distributed processing of large data sets across clusters of commodity computers using a simple programming model.

78. Why the name 'Hadoop'?

- ✓ Hadoop doesn't have any expanding version like 'OOPS'. The charming yellow elephant you see is basically named after Doug's son's toy elephant!

79. Why do we need Hadoop?

- ✓ Everyday a large amount of unstructured data is getting dumped into our machines.

80. What are some of the characteristics of Hadoop framework?

- ✓ Hadoop framework is written in Java. It is designed to solve problems that involve analyzing large data (e.g. petabytes). The programming model is based on Google's MapReduce. The infrastructure is based on Google's Big Data and Distributed File System.

81. Give a brief overview of Hadoop history.

- ✓ In 2002, Doug Cutting created an open source, web crawler project. In 2004, Google published MapReduce, GFS papers.
- ✓ In 2006, Doug Cutting developed the open source, MapReduce and HDFS project. In 2008, Yahoo ran 4,000 node Hadoop cluster and Hadoop won terabyte sort benchmark.
- ✓ In 2009, Facebook launched SQL support for Hadoop.

82. Give examples of some companies that are using Hadoop structure?

- ✓ A lot of companies are using the Hadoop structure such as Cloudera, EMC, MapR, Horton works, Amazon, Facebook, eBay, Twitter, Google and so on.

83. What is the basic difference between traditional RDBMS and Hadoop?

- ✓ RDBMS is used for transactional systems to report and archive the data.
- ✓ Hadoop is an approach to store huge amount of data in the distributed file system and process it.
- ✓ RDBMS will be useful when you want to seek one record from Big data, whereas.
- ✓ Hadoop will be useful when you want Big data in one shot and perform analysis on that later.

84. What is structured and unstructured data?

- ✓ Structured data is the data that is easily identifiable as it is organized in a structure. The most common form of structured data is a database where specific information is stored in tables, that is, rows and columns.
- ✓ Unstructured data refers to any data that cannot be identified easily. It could be in the form of images, videos, documents, email, logs and random text.

85. What are the core components of Hadoop?

- ✓ Core components of Hadoop are HDFS and MapReduce. HDFS is basically used to store large data sets and MapReduce is used to process such large data sets.

86. What is HDFS?

- ✓HDFS is a file system designed for storing very large files with streaming data access patterns, running clusters on commodity hardware.

87. What are the key features of HDFS?

- ✓HDFS is highly fault-tolerant, with high throughput, suitable for applications with large data sets, streaming access to file system data and can be built out of commodity hardware.

88. What is Fault Tolerance?

- ✓ Suppose you have a file stored in a system, and due to some technical problem that file gets destroyed. Then there is no chance of getting the data back present in that file.

89. Replication causes data redundancy then why is pursued in HDFS?

- ✓ HDFS works with commodity hardware (systems with average configurations) that has high chances of getting crashed any time. Thus, to make the entire system highly fault-tolerant, HDFS replicates and stores data in different places.

90. Since the data is replicated thrice in HDFS, does it mean that any calculation done on one node will also be replicated on the other two?

✓ Since there are 3 nodes, when we send the MapReduce programs, calculations will be done only on the original data. The master node will know which node exactly has that particular data.

91. What is throughput? How does HDFS get a good throughput?

✓ Throughput is the amount of work done in a unit time. It describes how fast the data is getting accessed from the system and it is usually used to measure performance of the system.

92. What is streaming access?

- ✓ As HDFS works on the principle of 'Write Once, Read Many', the feature of streaming access is extremely important in HDFS. HDFS focuses not so much on storing the data but how to retrieve it at the fastest possible speed, especially while analyzing logs.

93. What is a commodity hardware? Does commodity hardware include RAM?

- ✓ Commodity hardware is a non-expensive system which is not of high quality or high-availability. Hadoop can be installed in any average commodity hardware.

94. What is a Name node?

- ✓ Name node is the master node on which job tracker runs and consists of the metadata. It maintains and manages the blocks which are present on the data nodes.

95. Is Name node also a commodity?

- ✓ No. Name node can never be a commodity hardware because the entire HDFS rely on it. It is the single point of failure in HDFS. Name node has to be a high-availability machine.

96. What is a metadata?

- ✓ Metadata is the information about the data stored in data nodes such as location of the file, size of the file and so on.

97. What is a Data node?

- ✓ Data nodes are the slaves which are deployed on each machine and provide the actual storage. These are responsible for serving read and write requests for the clients.

98. Why do we use HDFS for applications having large data sets and not when there are lot of small files?

- ✓ HDFS is more suitable for large amount of data sets in a single file as compared to small amount of data spread across multiple files.

99. What is a daemon?

- ✓ Daemon is a process or service that runs in background. In general, we use this word in UNIX environment.

100. What is a job tracker?

- ✓ Job tracker is a daemon that runs on a name node for submitting and tracking MapReduce jobs in Hadoop. It assigns the tasks to the different task tracker.

101. What is a task tracker?

- ✓ Task tracker is also a daemon that runs on data nodes. Task Trackers manage the execution of individual tasks on slave node.

102. Is Name node machine same as data node machine as in terms of hardware?

- ✓ It depends upon the cluster you are trying to create. The Hadoop VM can be there on the same machine or on another machine.

103. What is a heartbeat in HDFS?

- ✓ A heartbeat is a signal indicating that it is alive. A data node sends heartbeat to Name node and task tracker will send its heart beat to job tracker.

104. Are Name node and job tracker on the same host?

- ✓ No, in practical environment, Name node is on a separate host and job tracker is on a separate host.

105. What is a ‘block’ in HDFS?

- ✓ A ‘block’ is the minimum amount of data that can be read or written. In HDFS, the default block size is 64 MB as contrast to the block size of 8192 bytes in Unix/Linux.

106. What are the benefits of block transfer?

- ✓ A file can be larger than any single disk in the network. Blocks provide fault tolerance and availability.

107. If we want to copy 10 blocks from one machine to another, but another machine can copy only 8.5 blocks, can the blocks be broken at the time of replication?

- ✓ In HDFS, blocks cannot be broken down. Before copying the blocks from one machine to another, the Master node will figure out what is the actual amount of space required, how many block are being used, how much space is available, and it will allocate the blocks accordingly.

108. How indexing is done in HDFS?

- ✓ Hadoop has its own way of indexing. Depending upon the block size, once the data is stored, HDFS will keep on storing the last part of the data which will say where the next part of the data will be.

109. If a data Node is full how it's identified?

- ✓ When data is stored in data node, then the metadata of that data will be stored in the Name node. So Name node will identify if the data node is full.

110.If data nodes increase, then do we need to upgrade Name node?

- ✓ While installing the Hadoop system, Name node is determined based on the size of the clusters.

111. Are job tracker and task trackers present in separate machines?

✓ Yes, job tracker and task tracker are present in different machines. The reason is job tracker is a single point of failure for the Hadoop MapReduce service.

112. When we send a data to a node, do we allow settling in time, before sending another data to that node?

✓ Yes, we do.

113. Does Hadoop always require digital data to process?

✓ Yes. Hadoop always require digital data to be processed.

114. On what basis Name node will decide which data node to write on?

✓ As the Name node has the metadata (information) related to all the data nodes, it knows which data node is free.

115. Doesn't Google have its very own version of DFS?

✓ Yes, Google owns a DFS known as "Google File System (GFS)" developed by Google Inc. for its own use.

116. Who is a 'user' in HDFS?

✓ A user is like you or me, who has some query or who needs some kind of data.

117. Is client the end user in HDFS?

✓ No, Client is an application which runs on your machine, which is used to interact with the Name node (job tracker) or data node (task tracker).

118. What is the communication channel between client and name node/ data node?

✓ The mode of communication is SSH.

119. What is a rack?

✓ Rack is a storage area with all the data nodes put together. Rack is a physical collection of data nodes which are stored at a single location.

120. On what basis data will be stored on a rack?

✓ When the client is ready to load a file into the cluster, the content of the file will be divided into blocks.

121. Do we need to place 2nd and 3rd data in rack 2 only?

✓ Yes, this is to avoid data node failure.

122. What if rack 2 and datanode fails?

- ✓ If both rack2 and datanode present in rack 1 fails then there is no chance of getting data from it.

123. What is a Secondary Namenode? Is it a substitute to the Namenode?

- ✓ The secondary Namenode constantly reads the data from the RAM of the Namenode and writes it into the hard disk or the file system.

124. What is the difference between Gen1 and Gen2 Hadoop with regards to the Namenode?

- ✓ In Gen 1 Hadoop, Namenode is the single point of failure. In Gen 2 Hadoop, we have what is known as Active and Passive Namenodes kind of a structure.

125. What is MapReduce?

- ✓ Map Reduce is the ‘heart‘ of Hadoop that consists of two parts – ‘map’ and ‘reduce’. Maps and reduces are programs for processing data.

126. Can you explain how do ‘map’ and ‘reduce’ work?

- ✓ Name node takes the input and divide it into parts and assign them to data nodes.

127. What is ‘Key value pair’ in HDFS?

- ✓ Key value pair is the intermediate data generated by maps and sent to reduces for generating the final output.

128. What is the difference between MapReduce engine and HDFS cluster?

✓ HDFS cluster is the name given to the whole configuration of master and slaves where data is stored. Map Reduce Engine is the programming module which is used to retrieve and analyze data.

129. Do we require two servers for the Name node and the data nodes?

✓ Yes, we need two different servers for the Name node and the data nodes.

130. Why are the number of splits equal to the number of maps?

✓ The number of maps is equal to the number of input splits because we want the key and value pairs of all the input splits.

131. Which are the two types of ‘writes’ in HDFS?

✓ There are two types of writes in HDFS: posted and non-posted write. Posted Write is when we write it and forget about it, without worrying about the acknowledgement.

132. Why ‘Reading‘ is done in parallel and ‘Writing‘ is not in HDFS?

✓ Reading is done in parallel because by doing so we can access the data fast. But we do not perform the write operation in parallel. The reason is that if we perform the write operation in parallel, then it might result in data inconsistency.

133. Is a job split into maps?

- ✓ No, a job is not split into maps. Spilt is created for the file. The file is placed on data nodes in blocks. For each split, a map is needed.

134.Can Hadoop be compared to NOSQL database like Cassandra?

- ✓ Though NOSQL is the closet technology that can be compared to Hadoop, it has its own pros and cons. There is no DFS in NOSQL. Hadoop is not a database.

135. How can I install Cloudera VM in my system?

- ✓ When you enroll for the Hadoop course at Edureka, you can download the Hadoop Installation steps.pdf file from our dropbox.

Relational DB's vs. Big Data (Spark)

1. It deals with Giga Bytes to Terabytes
2. It is centralized
3. It deals with structured data
4. It is having stable Data Model
5. It deals with known complex inter relationships
6. Tools are Relational DB's: SQL,MYSQL,DB2.
7. Access is Interactive and batch.
8. Updates are Read and write many times.
9. Integrity is high
10. Scaling is Nonlinear

1. It deals with Petabytes to Zettabytes
2. It is distributed
3. It deals with semi-structured and unstructured
4. It is having unstable Data Model
5. It deals with flat schemas and few Interrelationships
6. Tools are Hadoop,R,Mahout
7. Access is Batch
8. Updates are Write once, read many times.
9. Integrity is low
10. Scaling is Linear