**Hadoop Questions and Answers – Introduction to HDFS**

**This set of Multiple Choice Questions & Answers (MCQs) focuses on “Hadoop Filesystem – HDFS”.**

1. A \_\_\_\_\_\_\_\_ serves as the master and there is only one NameNode per cluster.  
a) Data Node  
**b) NameNode**  
c) Data block  
d) Replication  
**View Answer**

**Answer: b  
Explanation: All the metadata related to HDFS including the information about data nodes, files stored on HDFS, and Replication, etc. are stored and maintained on the NameNode.**

2. Point out the correct statement :  
**a) DataNode is the slave/worker node and holds the user data in the form of Data Blocks**  
b) Each incoming file is broken into 32 MB by default  
c) Data blocks are replicated across different nodes in the cluster to ensure a low degree of fault tolerance  
d) None of the mentioned  
**View Answer**

**Answer: a  
Explanation: There can be any number of DataNodes in a Hadoop Cluster.**

3. HDFS works in a \_\_\_\_\_\_\_\_\_\_ fashion.  
**a) master-worker**  
b) master-slave  
c) worker/slave  
d) all of the mentioned  
View Answer

Answer: a  
Explanation: NameNode servers as the master and each DataNode servers as a worker/slave

4. \_\_\_\_\_\_\_\_ NameNode is used when the Primary NameNode goes down.  
a) Rack  
b) Data  
**c) Secondary**  
d) None of the mentioned  
View Answer

Answer: c  
Explanation: Secondary namenode is used for all time availability and reliability.

5. Point out the wrong statement :  
a) Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level  
b) Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode  
c) User data is stored on the local file system of DataNodes  
**d) DataNode is aware of the files to which the blocks stored on it belong to**View Answer

Answer: d  
Explanation: NameNode is aware of the files to which the blocks stored on it belong to.

6. Which of the following scenario may not be a good fit for HDFS ?  
**a) HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file**b) HDFS is suitable for storing data related to applications requiring low latency data access  
c) HDFS is suitable for storing data related to applications requiring low latency data access  
d) None of the mentioned  
**View Answer**

**Answer: a  
Explanation: HDFS can be used for storing archive data since it is cheaper as HDFS allows storing the data on low cost commodity hardware while ensuring a high degree of fault-tolerance.**

7. The need for data replication can arise in various scenarios like :  
a) Replication Factor is changed  
b) DataNode goes down  
c) Data Blocks get corrupted  
**d) All of the mentioned**View Answer

Answer: d  
Explanation: Data is replicated across different DataNodes to ensure a high degree of fault-tolerance.

8. \_\_\_\_\_\_\_\_ is the slave/worker node and holds the user data in the form of Data Blocks.  
**a) DataNode**  
b) NameNode  
c) Data block  
d) Replication  
View Answer

Answer: a  
Explanation: A DataNode stores data in the [HadoopFileSystem]. A functional filesystem has more than one DataNode, with data replicated across them.

9. HDFS provides a command line interface called \_\_\_\_\_\_\_\_\_\_ used to interact with HDFS.  
a) “HDFS Shell”  
**b) “FS Shell”**  
c) “DFS Shell”  
d) None of the mentioned  
View Answer

Answer: b  
Explanation: The File System (FS) shell includes various shell-like commands that directly interact with the Hadoop Distributed File System (HDFS).

10. HDFS is implemented in \_\_\_\_\_\_\_\_\_\_\_\_\_ programming language.  
a) C++  
**b) Java**c) Scala  
d) None of the mentioned  
View Answer

**Answer: b  
Explanation: HDFS is implemented in Java and any computer which can run Java can host a NameNode/DataNode on it.**

**Hadoop Questions and Answers – Big Data**

**This set of Multiple Choice Questions & Answers (MCQs) focuses on “Big-Data”.**

1. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including:  
a) Improved data storage and information retrieval  
b) Improved extract, transform and load features for data integration  
c) Improved data warehousing functionality  
**d) Improved security, workload management and SQL support**  
**View Answer**

**Answer: d  
Explanation: Adding security to Hadoop is challenging because all the interactions do not follow the classic client- server pattern.**

2. Point out the correct statement :  
a) Hadoop do need specialized hardware to process the data  
**b) Hadoop 2.0 allows live stream processing of real time data**  
c) In Hadoop programming framework output files are divided in to lines or records  
d) None of the mentioned  
View Answer

Answer: b  
Explanation: Hadoop batch processes data distributed over a number of computers ranging in 100s and 1000s.

3. 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  
View Answer

Answer: a  
Explanation: Data warehousing integrated with Hadoop would give better understanding of data.

4. Hadoop is a framework that works with a variety of related tools. Common cohorts include:  
**a) MapReduce, Hive and HBase**  
b) MapReduce, MySQL and Google Apps  
c) MapReduce, Hummer and Iguana  
d) MapReduce, Heron and Trumpet  
View Answer

Answer: a  
Explanation: To use Hive with HBase you’ll typically want to launch two clusters, one to run HBase and the other to run Hive.

5. Point out the wrong statement :  
a) Hardtop’s processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data  
b) Hadoop uses a programming model called “MapReduce”, all the programs should confirms to this model in order to work on Hadoop platform  
**c) The programming model, MapReduce, used by Hadoop is difficult to write and test**  
d) All of the mentioned  
View Answer

Answer: c  
Explanation: The programming model, MapReduce, used by Hadoop is simple to write and test.

6. What was Hadoop named after?  
a) Creator Doug Cutting’s favorite circus act  
b) Cutting’s high school rock band  
**c) The toy elephant of Cutting’s son**  
d) A sound Cutting’s laptop made during Hadoop’s development  
View Answer

**Answer: c  
Explanation: Doug Cutting, Hadoop’s creator, named the framework after his child’s stuffed toy elephant.**

7. All of the following accurately describe Hadoop, EXCEPT:  
a) Open source  
**b) Real-time**  
c) Java-based  
d) Distributed computing approach  
View Answer

Answer: b  
Explanation: Apache Hadoop is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware.

8. \_\_\_\_\_\_\_\_\_\_ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.  
**a) MapReduce**  
b) Mahout  
c) Oozie  
d) All of the mentioned  
**View Answer**

**Answer: a  
Explanation: MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm.**

9. \_\_\_\_\_\_\_\_\_\_ has the world’s largest Hadoop cluster.  
a) Apple  
b) Datamatics  
**c) Facebook**  
d) None of the mentioned  
**View Answer**

**Answer: c  
Explanation: Facebook has many Hadoop clusters, the largest among them is the one that is used for Data warehousing.**

10. Facebook Tackles Big Data With \_\_\_\_\_\_\_ based on Hadoop.  
**a) ‘Project Prism’**  
b) ‘Prism’  
c) ‘Project Big’  
d) ‘Project Data’  
View Answer

Answer: a  
Explanation: Prism automatically replicates and moves data wherever it’s needed across a vast network of computing facilities.

**Hadoop Questions and Answers – Hadoop Ecosystem**

**This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Hadoop Ecosystem”.**

1. \_\_\_\_\_\_\_\_ is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.  
a) Pig Latin  
b) Oozie  
**c) Pig**  
d) Hive  
View Answer

Answer: c  
Explanation: Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs.

2. Point out the correct statement :  
**a) Hive is not a relational database, but a query engine that supports the parts of SQL specific to querying data**  
b) Hive is a relational database with SQL support  
c) Pig is a relational database with SQL support  
d) All of the mentioned  
View Answer

Answer: a  
Explanation: Hive is a SQL-based data warehouse system for Hadoop that facilitates data summarization, ad hoc queries, and the analysis of large datasets stored in Hadoop-compatible file systems.

3. \_\_\_\_\_\_\_\_\_ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.  
a) Scalding  
b) HCatalog  
**c) Cascalog**  
d) All of the mentioned  
**View Answer**

**Answer: c  
Explanation: Cascalog also adds Logic Programming concepts inspired by Datalog. Hence the name “Cascalog” is a contraction of Cascading and Datalog.**

4. Hive also support custom extensions written in :  
a) C#  
**b) Java**  
c) C  
d) C++  
**View Answer**

**Answer: b  
Explanation: Hive also support custom extensions written in Java, including user-defined functions (UDFs) and serializer-deserializers for reading and optionally writing custom formats.**

5. Point out the wrong statement :  
**a) Elastic MapReduce (EMR) is Facebook’s packaged Hadoop offering**  
b) Amazon Web Service Elastic MapReduce (EMR) is Amazon’s packaged Hadoop offering  
c) Scalding is a Scala API on top of Cascading that removes most Java boilerplate  
d) All of the mentioned  
**View Answer**

**Answer: a  
Explanation: Rather than building Hadoop deployments manually on EC2 (Elastic Compute Cloud) clusters, users can spin up fully configured Hadoop installations using simple invocation commands, either through the AWS Web Console or through command-line tools.**

6. \_\_\_\_\_\_\_\_ is the most popular high-level Java API in Hadoop Ecosystem  
a) Scalding  
b) HCatalog  
c) Cascalog  
**d) Cascading**  
View Answer

Answer: d  
Explanation: Cascading hides many of the complexities of MapReduce programming behind more intuitive pipes and data flow abstractions.

7. \_\_\_\_\_\_\_\_\_\_\_ is general-purpose computing model and runtime system for distributed data analytics.  
**a) Mapreduce**  
b) Drill  
c) Oozie  
d) None of the mentioned  
View Answer

**Answer: a  
Explanation: Mapreduce provides a flexible and scalable foundation for analytics, from traditional reporting to leading-edge machine learning algorithms.**

8. The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to :  
**a) SQL**  
b) JSON  
c) XML  
d) All of the mentioned  
View Answer

Answer: a  
Explanation: Pig Latin, in essence, is designed to fill the gap between the declarative style of SQL and the low-level procedural style of MapReduce.

9. \_\_\_\_\_\_\_ jobs are optimized for scalability but not latency.  
a) Mapreduce  
b) Drill  
c) Oozie  
**d) Hive**  
View Answer

Answer: d  
Explanation: Hive Queries are translated to MapReduce jobs to exploit the scalability of MapReduce.

10. \_\_\_\_\_\_ is a framework for performing remote procedure calls and data serialization.  
a) Drill  
b) BigTop  
**c) Avro**  
d) Chukwa  
View Answer

**Answer: c  
Explanation: In the context of Hadoop, Avro can be used to pass data from one program or language to another.**