**Hadoop Questions and Answers – Spark with Hadoop-1**

**This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on ” Spark with Hadoop-1″.**

1. Spark was initially started by \_\_\_\_\_\_\_\_\_\_\_\_ at UC Berkeley AMPLab in 2009.  
a) Mahek Zaharia  
**b) Matei Zaharia**  
c) Doug Cutting  
d) Stonebraker  
View Answer

Answer: b  
Explanation: Apache Spark is an open-source cluster computing framework originally developed in the AMPLab at UC Berkeley.

2. Point out the correct statement :  
a) RSS abstraction provides distributed task dispatching, scheduling, and basic I/O functionalities  
**b) For cluster manager, Spark supports standalone Hadoop YARN**  
c) Hive SQL is a component on top of Spark Core  
d) None of the mentioned  
View Answer

Answer: b  
Explanation: Spark requires a cluster manager and a distributed storage system.

3. \_\_\_\_\_\_\_\_\_\_\_\_ is a component on top of Spark Core.  
a) Spark Streaming  
**b) Spark SQL**  
c) RDDs  
d) All of the mentioned  
**View Answer**

**Answer: b  
Explanation: Spark SQL introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.**

4. Spark SQL provides a domain-specific language to manipulate \_\_\_\_\_\_\_\_\_\_\_ in Scala, Java, or Python.  
a) Spark Streaming  
b) Spark SQL  
**c) RDDs**  
d) All of the mentioned  
View Answer

Answer: c  
Explanation: Spark SQL provides SQL language support, with command-line interfaces and ODBC/JDBC server.

5. Point out the wrong statement :  
a) For distributed storage, Spark can interface with a wide variety, including Hadoop Distributed File System (HDFS)  
b) Spark also supports a pseudo-distributed mode, usually used only for development or testing purposes  
c) Spark has over 465 contributors in 2014  
**d) All of the mentioned**  
View Answer

**Answer: d  
Explanation: Spark is the most active project in the Apache Software Foundation and among Big Data open source projects.**

6. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ leverages Spark Core’s fast scheduling capability to perform streaming analytics.  
a) MLlib  
**b) Spark Streaming**  
c) GraphX  
d) RDDs  
View Answer

Answer: b  
Explanation: Spark Streaming ingests data in mini-batches and performs RDD transformations on those mini-batches of data.

7. \_\_\_\_\_\_\_\_\_\_\_\_ is a distributed machine learning framework on top of Spark  
**a) MLlib**  
b) Spark Streaming  
c) GraphX  
d) RDDs  
View Answer

Answer: a  
Explanation: MLlib implements many common machine learning and statistical algorithms to simplify large scale machine learning pipelines.

8. \_\_\_\_\_\_\_\_ is a distributed graph processing framework on top of Spark.  
a) MLlib  
b) Spark Streaming  
**c) GraphX**  
d) All of the mentioned  
View Answer

9. GraphX provides an API for expressing graph computation that can model the \_\_\_\_\_\_\_\_\_\_ abstraction.  
a) GaAdt  
b) Spark Core  
**c) Pregel**  
d) None of the mentioned  
View Answer

Answer: c  
Explanation: GraphX is used for machine learning.

10. Spark architecture is \_\_\_\_\_\_\_\_\_\_\_ times as fast as Hadoop disk-based Apache Mahout and even scales better than Vowpal Wabbit.  
**a) 10**  
b) 20  
c) 50  
d) 100  
View Answer

Answer: a  
Explanation: Spark architecture has proven scalability to over 8000 nodes in production.

**Hadoop Questions and Answers – Spark with Hadoop-2**

**This set of Hadoop Questions for campus interviews focuses on “Spark with Hadoop”.**

1. Users can easily run Spark on top of Amazon’s \_\_\_\_\_\_\_\_\_\_  
a) Infosphere  
**b) EC2**  
c) EMR  
d) None of the mentioned  
View Answer

Answer: b  
Explanation: Users can easily run Spark (and Shark) on top of Amazon’s EC2 either using the scripts that come with Spark.

2. Point out the correct statement :  
**a) Spark enables Apache Hive users to run their unmodified queries much faster**  
b) Spark interoperates only with Hadoop  
c) Spark is a popular data warehouse solution running on top of Hadoop  
d) None of the mentioned  
View Answer

Answer: a  
Explanation: Shark can accelerate Hive queries by as much as 100x when the input data fits into memory, and up 10x when the input data is stored on disk.

3. Spark runs on top of \_\_\_\_\_\_\_\_\_\_\_, a cluster manager system which provides efficient resource isolation across distributed applications  
a) Mesjs  
**b) Mesos**  
c) Mesus  
d) All of the mentioned  
View Answer

**Answer: b  
Explanation: Mesos enables fine grained sharing which allows a Spark job to dynamically take advantage of the idle resources in the cluster during its execution.**

4. Which of the following can be used to launch Spark jobs inside MapReduce ?  
a) SIM  
**b) SIMR(Spark Inside Map-Reduce)**  
c) SIR  
d) RIS  
View Answer

Answer: b  
Explanation: With SIMR, users can start experimenting with Spark and use its shell within a couple of minutes after downloading it.

5. Point out the wrong statement :  
**a) Spark is intended to replace, the Hadoop stack**  
b) Spark was designed to read and write data from and to HDFS, as well as other storage systems  
c) Hadoop users who have already deployed or are planning to deploy Hadoop Yarn can simply run Spark on YARN  
d) None of the mentioned  
**View Answer**

**Answer: a  
Explanation: Spark is intended to enhance, not replace, the Hadoop stack.**

6. Which of the following language is not supported by Spark ?  
a) Java  
**b) Pascal**  
c) Scala  
d) Python  
View Answer

Answer: b  
Explanation: The Spark engine runs in a variety of environments, from cloud services to Hadoop or Mesos clusters.

7. Spark is packaged with higher level libraries, including support for \_\_\_\_\_\_\_\_\_ queries.  
**a) SQL**  
b) C  
c) C++  
d) None of the mentioned  
**View Answer**

**Answer: a  
Explanation: Standard libraries increase developer productivity and can be seamlessly combined to create complex workflows.**

8. Spark includes a collection over \_\_\_\_\_\_\_\_ operators for transforming data and familiar data frame APIs for manipulating semi-structured data.  
a) 50  
b) 60  
c) 70  
**d) 80**  
**View Answer**

**Answer: d  
Explanation: Spark provides easy-to-use APIs for operating on large datasets.**

9. Spark is engineered from the bottom-up for performance, running \_\_\_\_\_\_\_\_\_\_\_ faster than Hadoop by exploiting in memory computing and other optimizations.  
**a) 100x**  
b) 150x  
c) 200x  
d) None of the mentioned  
View Answer

**Answer: a  
Explanation: Spark is fast on disk too; it currently holds the world record in large scale on-disk sorting.**

10. Spark powers a stack of high-level tools including Spark SQL, MLlib for :  
a) regression models  
b) statistics  
**c) machine learning**  
d) reproductive research  
**View Answer**

**Answer: c  
Explanation: Spark is used at a wide range of organizations to process large datasets.**