Q1: What is Hadoop Big Data Testing?

Big Data means a vast collection of structured and unstructured data, which is very expansive & is complicated to process by conventional database and software techniques. In many organizations, the volume of data is enormous, and it moves too fast in modern days and exceeds current processing capacity. Compilation of databases that are not being processed by conventional computing techniques, efficiently. Testing involves specialized tools, frameworks, and methods to handle these massive amounts of datasets. Examination of Big data is meant to the creation of data and its storage, retrieving of data and analysis them which is significant regarding its volume and variety of speed.

Q2: What do we test in Hadoop Big Data?

In the case of processing of the significant amount of data, performance, and functional testing is the primary key to performance. Testing is a validation of the data processing capability of the project and not the examination of the typical software features.

Q3: How do we validate Big Data?

In Hadoop, engineers authenticate the processing of quantum of data used by Hadoop cluster with supportive elements. Testing of Big data needs asks for extremely skilled professionals, as the handling is swift. Processing is three types namely Batch, Real Time, & Interactive.

Q4: How is data quality being tested?

Along with processing capability, quality of data is an essential factor while testing big data. Before testing, it is obligatory to ensure the data quality, which will be the part of the examination of the database. It involves the inspection of various properties like conformity, perfection, repetition, reliability, validity, completeness of data, etc.

Q5: What do you understand by Data Staging?

The initial step in the validation, which engages in process verification. Data from a different source like social media, RDBMS, etc. are validated, so that accurate uploaded data to the system. We should then compare the data source with the uploaded data into HDFS to ensure that both of them match. Lastly, we should validate that the correct data has been pulled, and uploaded into specific HDFS. There are many tools available, e.g., Talend, Datameer, are mostly used for validation of data staging.

Q6: What is "MapReduce" Validation?

MapReduce is the second phase of the validation process of Big Data testing. This stage involves the developer to verify the validation of the logic of business on every single systemic node and validating the data after executing on all the nodes, determining that:

1. Proper Functioning, of Map-Reduce.
2. Rules for Data segregation are being implemented.
3. Pairing & Creation of Key-value.
4. Correct Verification of data following the completion of Map Reduce.

Q7: What is Output Validation?

Third and the last phase in the testing of bog data is the validation of output. Output files of the output are created & ready for being uploaded on EDW (warehouse at an enterprise level), or additional arrangements based on need. The third stage consists of the following activities.

1. Assessing the rules for transformation whether they are applied correctly
2. Assessing the integration of data and successful loading of the data into the specific HDFS.
3. Assessing that the data is not corrupt by analyzing the downloaded data from HDFS & the source data uploaded.

Q8: What is Architecture Testing?

This pattern of testing is to process a vast amount of data extremely resources intensive. That is why testing of the architectural is vital for the success of any Project on Big Data. A faulty planned system will lead to degradation of the performance, and the whole system might not meet the desired expectations of the organization. At least, failover and performance test services need proper performance in any Hadoop environment.

Q9: What is Performance Testing?

Performance testing consists of testing of the duration to complete the job, utilization of memory, the throughput of data, and parallel system metrics. Any failover test services aim to confirm that data is processed seamlessly in any case of data node failure. Performance Testing of Big Data primarily consists of two functions. First, is Data ingestion whereas the second is Data Processing

Q10: What is Data ingestion?

The developer validates how fast the system is consuming the data from different sources. Testing involves the identification process of multiple messages that are being processed by a queue within a specific frame of time. It also consists of how fast the data gets into a particular data store, e.g., the rate of insertion into the Cassandra & Mongo database.

Q11: What is Data Processing in Hadoop Big data testing?
It involves validating the rate with which map-reduce tasks are performed. It also consists of data testing, which can be processed in separation when the primary store is full of data sets. E.g., Map-Reduce tasks running on a specific HDFS.

Q12: What do you mean by Performance of the Sub - Components?
Systems designed with multiple elements for processing of a large amount of data needs to be tested with every single of these elements in isolation. E.g., how quickly the message is being consumed & indexed, MapReduce jobs, search, query performances, etc.

Q13: What are the general approaches in Performance Testing?
Method of testing the performance of the application constitutes of the validation of large amount of unstructured and structured data, which needs specific approaches in testing to validate such data.

1. Setting up of the Application
2. Designing & identifying the task.
3. Organizing the Individual Clients
4. Execution and Analysis of the workload
5. Optimizing the Installation setup
6. Tuning of Components and Deployment of the system

Q14: What are the Test Parameters for the Performance?
Different parameters need to be confirmed while performance testing which is as follows:

1. Data Storage which validates the data is being stored on various systemic nodes
2. Logs which confirm the production of commit logs.
3. Concurrency establishing the number of threads being performed for reading and write operation
4. Caching which confirms the fine-tuning of "key cache” & "row cache" in settings of the

cache.
5. Timeouts are establishing the magnitude of query timeout.
6. Parameters of JVM are confirming algorithms of GC collection, heap size, and much more.
7. Map-reduce which suggests merging, and much more.
8. Message queue, which confirms the size, message rate, etc

Q15: What are Needs of Test Environment?

Test Environment depends on the nature of application being tested. For testing Big data, the environment should cover:
1. Adequate space is available for processing after significant storage amount of test data
2. Data on the scattered Cluster.
3. Minimum memory and CPU utilization for maximizing performance

Q16: What is the difference between the testing of Big data and Traditional database?

>> Developer faces more structured data in case of conventional database testing as compared to testing of Big data which involves both structured and unstructured data.
>> Methods for testing are time-tested and well defined as compared to an examination of big data, which requires R&D Efforts too.
>> Developers can select whether to go for "Sampling" or manual by "Exhaustive Validation" strategy with the help of automation tool.

Q17: What is the difference Big data Testing vs. Traditional database Testing regarding Infrastructure?

A conventional way of a testing database does not need specialized environments due to its limited size whereas in case of big data needs specific testing environment.

Q18: What is the difference Big data Testing vs. Traditional database Testing regarding validating Tools?

1. The validating tool needed in traditional database testing are excel based on macros or automotive tools with User Interface, whereas testing big data is enlarged without having specific and definitive tools.
2. Tools required for conventional testing are very simple and does not require any

specialized skills whereas big data tester need to be specially trained, and updations are needed more often as it is still in its nascent stage.

Q19: What are the tools applied in these scenarios of testing?

Cluster	**Tools**
NoSQL:	Cassandra, CouchDB, DatabasesMongoDB, Redis, HBase. ZooKeeper
MapReduce:	Hadoop, Pig, Hive, Cascading, Kafka, Oozie, S4, Flume, MapR
Storage:	HDFS, S3
Servers:	Elastic, EC2, Heroku
Processing	Mechanical Turk, R, Yahoo! BigSheets.

Big Data Hadoop Testing interview questions for Exprienced

Q20: What are the challenges in Automation of Testing Big data?
Organizational Data, which is growing every data, ask for automation, for which the test of Big Data needs a highly skilled developer. Sadly, there are no tools capable of handling unpredictable issues that occur during the validation process. Lot of Focus on R&D is still going on.

Q30: What are the challenges in Virtualization of Big Data testing?
Virtualization is an essential stage in testing Big Data. The Latency of virtual machine generates issues with timing. Management of images is not hassle-free too.

Q31: What are the challenges in Large Dataset in the testing of Big data?
Challenges in testing are evident due to its scale. In testing of Big Data:

- We need to substantiate more data, which has to be quicker.
- Testing efforts require automation.
- Testing facilities across all platforms require being defined.

Q32: What are other challenges in performance testing?
Big data is a combination of the varied technologies. Each of its sub-elements belongs to a different equipment and needs to be tested in isolation. Following are some of the

different challenges faced while validating Big Data:
>> There are no technologies available, which can help a developer from start-to-finish. Examples are, NoSQL does not validate message queues.
>> Scripting: High level of scripting skills is required to design test cases.
>> Environment: Specialized test environment is needed due to its size of data.
>> Supervising Solution are limited that can scrutinize the entire testing environment
>> The solution needed for diagnosis: Customized way outs are needed to develop and wipe out the bottleneck to enhance the performance.

Q33: What is Query Surge?

Query Surge is one of the solutions for Big Data testing. It ensures the quality of data quality and the shared data testing method that detects bad data while testing and provides an excellent view of the health of data. It makes sure that the data extracted from the sources stay intact on the target by examining and pinpointing the differences in the Big Data wherever necessary.

Q34: What Benefits do Query Surge provides?

1. Query Surge helps us to automate the efforts made by us manually in the testing of Big Data. It offers to test across diverse platforms available like Hadoop, Teradata, MongoDB, Oracle, Microsoft, IBM, Cloudera, Amazon, HortonWorks, MapR, DataStax, and other Hadoop vendors like Excel, flat files, XML, etc.
2. Enhancing Testing speeds by more than thousands times while at the same time offering the coverage of entire data.
3. Delivering Continuously – Query Surge integrates DevOps solution for almost all Build, QA software for management, ETL.
4. It also provides automated reports by email with dashboards stating the health of data.
5. Providing excellent Return on the Investments (ROI), as high as 1,500%

Q35: What is Query Surge's architecture?

Query Surge Architecture consists of the following components:

1. Tomcat - The Query Surge Application Server
2. The Query Surge Database (MySQL)
3. Query Surge Agents – At least one has to be deployed
4. Query Surge Execution API, which is optional.

Q36: What is an Agent?

The Query Surge Agent is the architectural element that executes queries against Source and Target data sources and getting the results to Query Surge.

Q37: How many agents are needed in a Query Surge Trial?

Any Query Surge or a POC, only one agent is sufficient. For production deployment, it is dependent on several factors (Source/data source products / Target database / Hardware Source/ Targets are installed, the style of query scripting), which is best determined as we gain experience with Query Surge within our production environment.

Q39: Do we need to use our database?

Query Surge has its inbuilt database, embedded in it. We need to lever the licensing of a database so that deploying Query Surge does not affect the organization currently has decided to use its services.

Q40: What are the different types of Automated Data Testing available for Testing Big Data?

Following are the various types of tools available for Big Data Testing:

1. Big Data Testing
2. ETL Testing & Data Warehouse
3. Testing of Data Migration
4. Enterprise Application Testing / Data Interface /
5. Database Upgrade Testing