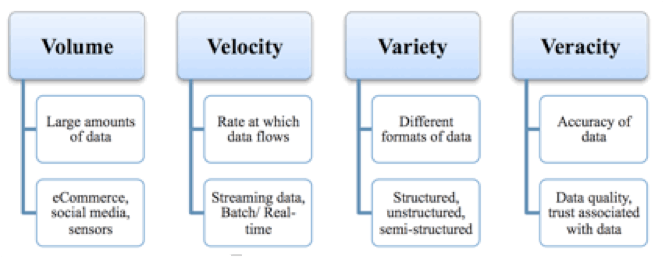
Hadoop Interview Questions & Answers

**Q) MapReduce Vs Spark**

|  |  |
| --- | --- |
| **Hadoop MapReduce Vs Spark** | |
| **Hadoop MapReduce** | **Spark** |
| Open-Source framework for application writing | Open-Source big data framework |
| Process Structured & unstructured data | It is for general purpose data processing |
| It provides batch processing only | Fits for both batch & stream processing |
| It fails in real-time data processing | It can process the real-time data |
| It is difficult to program & require abstraction | It comes with user-friendly APIs of scala. |
| It reads data from disk & after sends results to HDFS | It comes with a graph comutational lib called GraphX |
| It receives fault tolerance via replication | It uses RDD & diff data storage models for fault tolerance |
| It has better security features | Its security is currently infancy. |

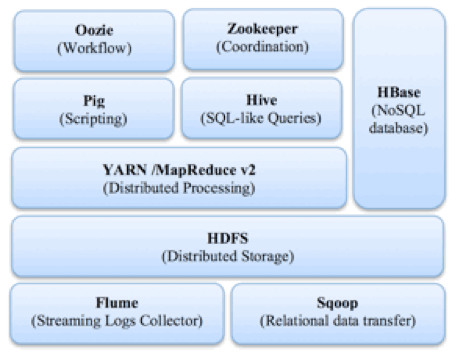
**Q1) What is big data?**

**Big Data** is the really large amount of data that exceeds the processing capacity of conventional database systems and requires special parallel processing mechanism. The data is too big and grows rapidly. This data can be either structural or unstructured data. To retrieve meaningful information from this data, we must choose an alternative way to process it.  
***Characteristics of Big Data:***  
Data that has very large volume, comes from variety of sources and formats and flows into an organization with a great velocity is normally referred to as Big Data.



**Q2) What is Hadoop?**

Hadoop is a framework that allows distributed processing of large data sets across clusters of computers using simple and fault tolerant programming model. It is designed to scale up from a very few to thousands of machines, each machine provides local computation and storage. The Hadoop software library itself is designed to detect and handle failures at the application layer.  
**Hadoop** is written in java by Apache Software Foundation.  It process data very reliably and fault-tolerant manner.  
**Core components of Hadoop:**  
HDFS (Storage) + MapReduce/YARN (Processing)



**Q3) What are the sources generating big data?**

Employers,Users and Machines

* **Employees:** Historically, employees of organizations generated data.
* **Users:**Then a shift occurred where users started generating data. For example, email, social media, photos, videos, audio and e-Commerce.
* **Machines:**Smart phones, intelligent kitchen appliances, CCTV cameras, smart meters, global satellites, and traffic flow sensors.

**Q4) Why do we need a new framework for handling big data?**

Most of the traditional data was organized neatly in relational databases. Data sets now are so large and complex that they are beyond the capabilities of traditional storage and processing systems.  
The following challenges demand cost-effective and innovative forms of handling big data at scale:

**Lots of data**  
Organizations are increasingly required to store more and more data to survive in today’s highly competitive environment. The sheer volume of the data demands lower storage costs as compared to the expensive commercial relational database options.

**Complex nature of data**  
Relational data model has great properties for structured data but many modern systems don’t fit well in row-column format. Data is now generated by diverse sources in various formats like multimedia, images, text, real-time feeds, and sensor streams. Usually for storage, the data is transformed, aggregated to fit into the structured format resulting in the loss of the original raw data.

**New analysis techniques**  
Previously simple analysis (like average, sum) would prove to be sufficient to predict customer behavior. But now complex analysis needs to be performed to gain insightful understanding of data collected. For example, prediction models for effective micro-segmentation needs to analyse the customer’s purchase history, browsing behavior, likes and reviews on social media website to perform micro-segmentation. These advanced analytic techniques need the framework to run on.

**Hadoop to rescue:** Framework that provides low-cost storage and complex analytic processing capabilities

**Q5) Why do we need Hadoop framework, shouldn’t DFS be able to handle large volumes of data already?**

Yes, it is true that when the datasets cannot fit in a single physical machine, then Distributed File System (DFS) partitions the data, store and manages the data across different machines. But, DFS lacks the following features for which we need Hadoop framework:

**Fault tolerant:**  
When a lot of machines are involved chances of data loss increases. So, automatic fault tolerance and failure recovery become a prime concern.

**Move data to computation:**  
If huge amounts of data are moved from storage to the computation machines then the speed depends on network bandwidth.

**Q6) What is the difference between traditional RDBMS and Hadoop?**

|  |  |
| --- | --- |
| **RDBMS** | **Hadoop** |
| Schema on write | Schema on read |
| Scale up approach | Scale out approach |
| Relational tables | Key-value format |
| Structured queries | Function programming |
| Online Transactions | Batch processing |

**Q7) What is HDFS?**

Hadoop Distributed File Systems (HDFS) is one of the core components of Hadoop framework. It is a distributed file system for Hadoop. It runs on top of existing file system (ext2, ext3, etc.)

**Goals:** Automatic recovery from failures, Move Computation than data.

**HDFS features:**

1. Supports storage of very large datasets
2. Write once read many access model
3. Streaming data access
4. Replication using commodity hardware

**Q8) What is difference between regular file system and HDFS?**

|  |  |
| --- | --- |
| **Regular File Systems** | **HDFS** |
| Small block size of data (like 512 bytes) | Large block size (orders of 64mb) |
| Multiple disk seeks for large files | Reads data sequentially after single seek |

**Q9) What HDFS is not meant for?**

HDFS is not good at:

1. Applications that requires low latency access to data (in terms of milliseconds)
2. Lot of small files
3. Multiple writers and file modifications

**Q10) What is HDFS block size and what did you chose in your project?**

By default, the HDFS block size is 64MB. It can be set to higher values as 128MB or 256MB. 128MB is acceptable industry standard.

**Q11) What is the default replication factor?**

Default replication factor is 3

**Q12) What are different hdfs dfs shell commands to perform copy operation?**

$ hadoop fs -copyToLocal  
$ hadoop fs -copyFromLocal  
$ hadoop fs -put

**Q13) What are the problems with Hadoop 1.0?**

1. NameNode: No Horizontal Scalability and No High Availability
2. Job Tracker: Overburdened.
3. MRv1: It can only understand Map and Reduce tasks

**Q14) What comes in Hadoop 2.0 and MapReduce V2 (YARN)?**

NameNode: HA and Federation  
JobTracker: Cluster and application resource

**Q15) What different type of schedulers and type of scheduler did you use?**

**Capacity Scheduler**  
It is designed to run Hadoop applications as a shared, multi-tenant cluster while maximizing the throughput and the utilization of the cluster.

**Fair Scheduler**  
Fair scheduling is a method of assigning resources to applications such that all apps get, on average, an equal share of resources over time.

**Q16) Steps involved in decommissioning (removing) the nodes in the Hadoop cluster?**

1. Update the network addresses in the dfs.exclude and mapred.exclude
2. $ hadoop dfsadmin -refreshNodes  and hadoop mradmin -refreshNodes
3. Check Web UI it will show “Decommissioning in Progress”
4. Remove the Nodes from include file and then run again the step 2 refreshNodes.
5. Remove the Nodes from slave file.

**Q17) Steps involved in commissioning (adding) the nodes in the Hadoop cluster?**

1. Update the network addresses in the dfs.include and mapred.include
2. $ hadoop dfsadmin -refreshNodes  and hadoop mradmin -refreshNodes
3. Update the slave file.
4. Start the DataNode and NodeManager on the added Node.

**Q18) How to keep HDFS cluster balanced?**

Balancer is a tool that tries to provide a balance to a certain threshold among data nodes by copying block data distribution across the cluster.

**Q19) What is distcp?**

1. istcp is the program comes with Hadoop for copying large amount of data to and from Hadoop file systems in parallel.
2. It is implemented as MapReduce job where copying is done through maps that run in parallel across the cluster.
3. There are no reducers.

**Q20) What are the daemons of HDFS?**

1. NameNode
2. DataNode
3. Secondary NameNode.

**Q21) Command to format the NameNode?**

$ hdfs namenode -format

**Q22) What are the functions of NameNode?**

The NameNode is mainly responsible for:

**Namespace**  
Maintain metadata about the data

**Block Management**  
Processes block reports and maintain location of blocks.  
Supports block related operations  
Manages replica placement

**Q23) What is HDFS Federation?**

* HDFS federation allows scaling the name service horizontally; it uses multiple independent NameNodes for different namespaces.
* All the NameNodes use the DataNodes as common storage for blocks.
* Each DataNode registers with all the NameNodes in the cluster.
* DataNodes send periodic heartbeats and block reports and handles commands from the NameNodes

**Q24) What is HDFS High Availability?**

1. In HDFS High Availability (HA) cluster; two separate machines are configured as NameNodes.
2. But one of the NameNodes is in an **Active** state; other is in a **Standby** state.
3. The Active NameNode is responsible for all client operations in the cluster, while the Standby is simply acting as a slave, maintaining enough state to provide a fast failover if necessary
4. They shared the same storage and all DataNodes connects to both the NameNodes.

**Q25) How client application interacts with the NameNode?**

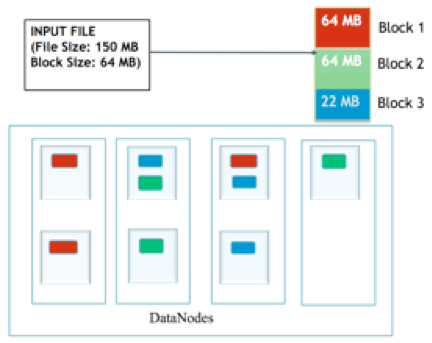
1. Client applications interact using Hadoop HDFS API with the NameNode when it has to locate/add/copy/move/delete a file.
2. The NameNode responds the successful requests by returning a list of relevant DataNode servers where the data is residing.
3. Client can talk directly to a DataNode after the NameNode has given the location of the data

**Q26) What is a DataNode?**

1. A DataNode stores data in the Hadoop File System HDFS is a slave node.
2. On startup, a DataNode connects to the NameNode.
3. DataNode instances can talk to each other mostly during replication.

**Q27) What is rack-aware replica placement policy?**

1. Rack-awareness is used to take a node’s physical location into account while scheduling tasks and allocating storage.
2. Default replication factor is 3 for a data blocks on HDFS.
3. The first two copies are stored on DataNodes located on the same rack while the third copy is stored on a different rack.



**Q28) What is the main purpose of HDFS fsck command?**

fsck a utility to check health of the file system, to find missing files, over-replicated, under-replicated and corrupted blocks.

Command for finding the blocks for a file:

**$ hadoop fsck -files -blocks –racks**

**Q29) What is the purpose of DataNode block scanner?**

1. Block scanner runs on every DataNode, which periodically verifies all the blocks stored on the DataNode.
2. If bad blocks are detected it will be fixed before any client reads.

**Q30) What is the purpose of dfsadmin tool?**

1. It is used to find information about the state of HDFS
2. It performs administrative tasks on HDFS
3. Invoked by hadoop dfsadmin command as superuser

**Q31) What is the command for printing the topology?**

It displays a tree of racks and DataNodes attached to the tracks as viewed by the .hdfs dfsadmin -printTopology

**Q32) What is RAID?**

RAID is a way of combining multiple disk drives into a single entity to improve performance and/or reliability. There are a variety of different levels in RAID  
For example, In RAID level 1 copy of the same data on two disks increases the read performance by reading alternately from each disk in the mirror.

**Q33) Does Hadoop requires RAID?**

1. In DataNodes storage is not using RAID as redundancy can be achieved by replication between the Nodes.
2. In NameNode’s disk RAID is recommended.

**Q34) What are the site-specific configuration files in Hadoop?**

1. conf/core-site.xml
2. conf/hdfs-site.xml
3. conf/yarn-site.xml
4. conf/mapred-site.xml.
5. conf/hadoop-env.sh
6. conf/yarn-env.sh

**Q35) What is MapReduce?**

MapReduce is a programming model for processing on the distributed datasets on the clusters of a computer.

**MapReduce Features:**

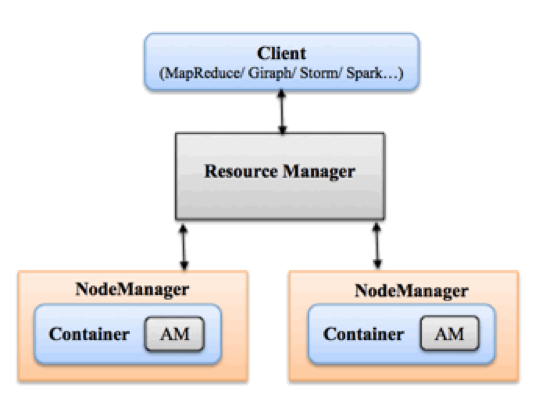
1. Distributed programming complexity is hidden
2. Built in fault-tolerance
3. Programming model is language independent
4. Parallelization and distribution are automatic
5. Enable data local processing

**Q36) What is the fundamental idea behind YARN?**

In YARN (Yet Another Resource Allocator), JobTracker responsibility is split into:

1. Resource management
2. Job scheduling/monitoring having separate daemons.

Yarn supports additional processing models and implements a more flexible execution engine.



**Q37) What MapReduce framework consists of?**

**ResourceManager (RM)**

1. Global resource scheduler
2. One master RM

**NodeManager (NM)**

1. One slave NM per cluster-node.

**Container**

1. RM creates Containers upon request by AM
2. Application runs in one or more containers

**ApplicationMaster (AM)**

1. One AM per application
2. Runs in Container

**Q38) What are different daemons in YARN?**

1. ResourceManager: Global resource manager.
2. NodeManager: One per data node, It manages and monitors usage of the container (resources in terms of Memory, CPU).
3. ApplicationMaster: One per application, Tasks are started by NodeManager

**Q39) What are the two main components of ResourceManager?**  
**Scheduler**

It allocates the resources (containers) to various running applications: Container elements such as memory, CPU, disk etc.

**ApplicationManager**  
It accepts job-submissions, negotiating for container for executing the application specific ApplicationMaster and provides the service for restarting the ApplicationMaster container on failure.

**Q40) What is the function of NodeManager?**

The NodeManager is the resource manager for the node (Per machine) and is responsible for containers, monitoring their resource usage (cpu, memory, disk, network) and reporting the same to the ResourceManager

**Q41) What is the function of ApplicationMaster?**

ApplicationMaster is per application and it has the responsibility of negotiating appropriate resource containers from the Scheduler, tracking their status and monitoring for progress.

**Q42) What are the minimum configuration requirements for a MapReduce application?**

The job configuration requires the

1. input location
2. output location
3. map() function
4. reduce() functions and
5. job parameters.

**Q43) What are the steps to submit a Hadoop job?**

Steps involved in Hadoop job submission:

1. Hadoop job client submits the job jar/executable and configuration to the ResourceManager.
2. ResourceManager then distributes the software/configuration to the slaves.
3. ResourceManager then scheduling tasks and monitoring them.
4. Finally, job status and diagnostic information is provided to the client.

**Q44) How does MapReduce framework view its input internally?**

It views the input as a set of pairs and produces a set of pairs as the output of the job.

**Q45) Assuming default configurations, how is a file of the size 1 GB (uncompressed) stored in HDFS?**

Default block size is 64MB. So, file of 1GB will be stored as 16 blocks. MapReduce job will create 16 input splits; each will be processed with separate map task i.e. 16 mappers.

**Q46) What are Hadoop Writables?**

Hadoop Writables allows Hadoop to read and write the data in a serialized form for transmission as compact binary files. This helps in straightforward random access and higher performance. Hadoop provides in built classes, which implement Writable: Text, IntWritable, LongWritable, FloatWritable, and BooleanWritable.

**Q47) Why comparison of types is important for MapReduce?**

A comparison is important as in the sorting phase the keys are compared with one another. For comparison, the WritableComparable interface is implemented.

**Q48) What is the purpose of RawComparator interface?**

RawComparator allows the implementors to compare records read from a stream without deserialization them into objects, so it will be optimized, as there is not overhead of object creation.

**Q49) What is a NullWritable?**

It is a special type of Writable that has zero-length serialization. In MapReduce, a key or a value can be declared as NullWritable if we don’t need that position, storing constant empty value.

**Q50) What is Avro Serialization System?**

Avro is a language-neutral data serialization system. It has data formats that work with different languages. Avro data is described using a language-independent schema (usually written in JSON). Avro data files support compression and are **splittable**.

Avro provides AvroMapper and AvroReducer to run MapReduce programs.

**Q51) Explain use cases where SequenceFile class can be a good fit?**

When the data is of type binary then SequenceFile will provide a persistent structure for binary key-value pairs. SequenceFiles also work well as containers for smaller files as HDFS and MapReduce are optimized for large files.

**Q52) What is MapFile?**

A MapFile is an indexed SequenceFile and it is used for look-ups by key.

**Q53) What is the core of the job in MapReduce framework?**

The core of a job:  
*Mapper interface*: map method  
*Reducer interface* reduce method

**Q54) What are the steps involved in MapReduce framework?**

1. Firstly, the mapper input key/value pairs maps to a set of intermediate key/value pairs.
2. Maps are the individual tasks that transform input records into intermediate records.
3. The transformed intermediate records do not need to be of the same type as the input records.
4. A given input pair maps to zero or many output pairs.
5. The Hadoop MapReduce framework creates one map task for each InputSplit generated by the InputFormat for the job.
6. It then calls map(WritableComparable, Writable, Context) for each key/value pair in the InputSplit for that task.
7. All intermediate values associated with a given output key are grouped passed to the Reducers.

**Q55) Where is the Mapper Output stored?**

The mapper output is stored on the Local file system of each individual mapper nodes. The intermediate data is cleaned up after the Hadoop Job completes.

**Q56) What is a partitioner and how the user can control which key will go to which reducer?**

Partitioner controls the partitioning of the keys of the intermediate map-outputs by the default. The key to decide the partition uses hash function. Default partitioner is HashPartitioner.  
A custom partitioner is implemented to control, which keys go to which Reducer.

public class SamplePartitioner extends Partitioner {

@Override

public int getPartition(Text key, Text value, int numReduceTasks) {

}

}

**Q57) What are combiners and its purpose?**

1. Combiners are used to increase the efficiency of a MapReduce program. It can be used to aggregate intermediate map output locally on individual mapper outputs.
2. Combiners can help reduce the amount of data that needs to be transferred across to the reducers.
3. Reducer code as a combiner if the operation performed is commutative and associative.
4. Hadoop may or may not execute a combiner.

**Q58) How a number of partitioners and reducers are related?**

The total numbers of partitions are the same as the number of reduce tasks for the job.

**Q59) What is IdentityMapper?**

IdentityMapper implements the mapping inputs directly to output. IdentityMapper.class is used as a default value when JobConf.setMapperClass is not set.

**Q60) What is IdentityReducer?**

In IdentityReducer no reduction is performed, writing all input values directly to the output. IdentityReducer.class is used as a default value when JobConf.setReducerClass is not set

**Q61) What is the reducer and its phases?**

Reducer reduces a set of intermediate values, which has same key to a smaller set of values. The framework then calls reduce().  
**Syntax:**  
reduce(WritableComparable, Iterable, Context) method for each pair in the grouped inputs.  
Reducer has three primary phases:

1. Shuffle
2. Sort
3. Reduce

**Q62) How to set the number of reducers?**

The number of reduces for the user sets the job:

1. Job.setNumReduceTasks(int)
2. -D mapreduce.job.reduces

**Q63) Detail description of the Reducer phases?**

**Shuffle:**  
Sorted output (Mapper) à Input (Reducer). Framework then fetches the relevant partition of the output of all the mappers.

**Sort:**  
The framework groups Reducer inputs by keys. The shuffle and sort phases occur simultaneously; while map-outputs are being fetched they are merged.

**Secondary Sort:**  
Grouping the intermediate keys are required to be different from those for grouping keys before reduction, then Job.setSortComparatorClass(Class).

**Reduce:**  
reduce(WritableComparable, Iterable, Context) method is called for each pair in the grouped inputs.  
The output of the reduce task is typically written using Context.write(WritableComparable, Writable).

**Q64) Can there be no Reducer?**

Yes, the number of reducer can be zero if no reduction of values is required.

**Q65) What can be optimum value for Reducer?**

Value of Reducers can be: 0.95

1. 1.75 multiplied by ( \* < number of maximum containers per node>)

Increasing number of reducers

1. Increases the framework overhead
2. Increases load balancing
3. Lowers the cost of failures

**Q66) What are a Counter and its purpose?**

The counter is a facility for MapReduce applications to report its statistics. They can be used to track job progress in a very easy and flexible manner. It is defined by MapReduce framework or by applications. Each Counter can be of any Enum type. Applications can define counters of type Enum and update them via counters.incrCounter in the map and/or reduce methods.

**Q67) Define different types of Counters?**

**Built in Counters:**

* Map Reduce Task Counters
* Job Counters

**Custom Java Counters:**

* MapReduce allows users to specify their own counters (using Java enums) for performing their own counting operation.

**Q68) Why Counter values are shared by all map and reduce tasks across the MapReduce framework?**

Counters are global so shared across the MapReduce framework and aggregated at the end of the job across all the tasks.

**Q69) Explain speculative execution.**

1. Speculative execution is a way of dealing with individual machine’s performance. As there are lots of machines in the cluster, some machines can have low performance, which affects the performance of the whole job.
2. Speculative execution in Hadoop can run multiple copies of the same map or reduce task on different task tracker nodes and the results from first node to finish are used.

**Q70) What is DistributedCache and its purpose?**

DistributedCache is a facility provided by the MapReduce framework to cache files (text, archives, jars etc.) needed by applications. It distributes application-specific, large, read-only files efficiently. The user needs to use DistributedCache to distribute and symlink the script file.

**Q71) What is the Job interface in MapReduce framework?**

Job is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution. Some basic parameters are configured for example:

1. Job.setNumReduceTasks(int)
2. Configuration.set(JobContext.NUM\_MAPS, int)
3. Mapper
4. Combiner (if any)
5. Partitioner
6. Reducer
7. InputFormat
8. OutputFormat implementations
9. setMapSpeculativeExecution(boolean))/ setReduceSpeculativeExecution(boolean))
10. Maximum number of attempts per task (setMaxMapAttempts(int)/ setMaxReduceAttempts(int)) etc.
11. DistributedCache for large amounts of (read-only) data.

**Q72) What is the default value of map and reduce max attempts?**

The framework will try to execute a map task or reduce task by **default 4 times** before giving up on it.

**Q73) Explain InputFormat?**

InputFormat describes the input-specification for a MapReduce job. The MapReduce framework depends on the InputFormat of the job to:

Checks the input-specification of the job.  
It then splits the input file(s) into logical InputSplit instances, each of which is then assigned to an individual Mapper.

To extract input records from the logical InputSplit for processing by the Mapper it provides the RecordReader implementation.  
Default: **TextInputFormat**

**Q74) What is InputSplit and RecordReader?**

InputSplit specifies the data to be processed by an individual Mapper.  
In general, InputSplit presents a byte-oriented view of the input.

Default: **FileSplit**  
RecordReader reads pairs from an InputSplit, then processes them and presents record-oriented view

**Q75) Explain the Job OutputFormat?**

OutputFormat describes details of the output for a MapReduce job.  
The MapReduce framework depends on the OutputFormat of the job to:  
It checks the job output-specification

To write the output files of the job in the pairs, it provides the RecordWriter implementation.  
Default: **TextOutputFormat**

**Q76) How is the option in Hadoop to skip the bad records?**

Hadoop provides an option where a certain set of bad input records can be skipped when processing map inputs. This feature can be controlled by the **SkipBadRecords** class.

[Check Out Hadoop Tutorials](https://mindmajix.com/hadoop)

**Q77) Different ways of debugging a job in MapReduce?**

1. Add debug statement to log to standard error along with the message to update the task’s status message. Web UI makes it easier to view.
2. Create a custom counter, it gives valuable information to deal with the problem dataset
3. Task page and task detailed page
4. Hadoop Logs
5. MRUnit testing

**PROGRAM 1: Counting the number of words in an input file**

**Introduction**  
This section describes how to get the word count of a sample input file.

**Software Versions**  
The software versions used are:  
VirtualBox: 4.3.20  
CDH 5.3: Default MapReduce Version  
hadoop-core-2.5.0  
hadoop-yarn-common-2.5.0

**Steps**  
1. **Create the input file**  
Create the input.txt file with sample text.  
$ vi input.txt  
Thanks Lord Krishna for helping us write this book  
Hare Krishna Hare Krishna Krishna Krishna Hare Hare  
Hare Rama Hare Rama Rama Rama Hare Hare

2. **Move the input file into HDFS**  
Use the –put or –copyFromLocal command to move the file into HDFS  
$ hadoop fs -put input.txt

3. **Code for the MapReduce program**  
Java files:  
WordCountProgram.java  // Driver Program  
WordMapper.java         // Mapper Program  
WordReducer.java        // Reducer Program  
————————————————–  
WordCountProgram.java File: Driver Program  
————————————————–  
import org.apache.hadoop.conf.Configuration;  
import org.apache.hadoop.conf.Configured;  
import org.apache.hadoop.fs.Path;  
import org.apache.hadoop.io.IntWritable;  
import org.apache.hadoop.io.Text;  
import org.apache.hadoop.mapreduce.Job;  
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;  
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;  
import org.apache.hadoop.util.Tool;  
import org.apache.hadoop.util.ToolRunner;  
public class WordCountProgram extends Configured implements Tool{  
@Override  
public int run(String[] args) throws Exception {  
Configuration conf = new Configuration();  
Job job = new Job(conf, “wordcountprogram”);  
job.setJarByClass(getClass());  
// Configure output and input source  
TextInputFormat.addInputPath(job, new Path(args[0]));  
job.setInputFormatClass(TextInputFormat.class);  
job.setMapperClass(WordMapper.class);  
job.setReducerClass(WordReducer.class);  
// Configure output  
TextOutputFormat.setOutputPath(job, new Path(args[1]));  
job.setOutputFormatClass(TextOutputFormat.class);  
job.setOutputKeyClass(Text.class);  
job.setOutputValueClass(IntWritable.class);  
return job.waitForCompletion(true) ? 0 : 1;  
}  
public static void main(String[] args) throws Exception {  
int exitCode = ToolRunner.run(new WordCountProgram(), args);  
System.exit(exitCode);  
}  
}  
————————————————–  
WordMapper.java File: Mapper Program  
————————————————–  
import java.io.IOException;  
import java.util.StringTokenizer;  
import org.apache.hadoop.io.IntWritable;  
import org.apache.hadoop.io.LongWritable;  
import org.apache.hadoop.io.Text;  
import org.apache.hadoop.mapreduce.Mapper;  
public class WordMapper extends Mapper {  
private final static IntWritable count = new IntWritable(1);  
private final Text nameText = new Text();  
@Override  
protected void map(LongWritable key, Text value, Context context) throws IOException,  
InterruptedException {  
StringTokenizer tokenizer = new StringTokenizer(value.toString(),” “);  
while (tokenizer.hasMoreTokens()) {  
nameText.set(tokenizer.nextToken());  
context.write(nameText, count);  
}  
}  
}  
———————————————–  
WordReducer.java file: Reducer Program  
————————————————–  
import java.io.IOException;  
import org.apache.hadoop.io.IntWritable;  
import org.apache.hadoop.io.Text;  
import org.apache.hadoop.mapreduce.Reducer;  
public class WordReducer extends Reducer {  
@Override  
protected void reduce(Text t, Iterable counts, Context context)  
throws IOException, InterruptedException {  
int sum = 0;  
for (IntWritable count : counts) {  
sum += count.get();  
}  
context.write(t, new IntWritable(sum));  
}  
}  
4. **Run the MapReduce program**  
Create the jar of the Code in Step 3 and use the following command to run the MapReduce program  
$ hadoop jar WordCount.jar WordCountProgram input.txt output1  
Here,  
WordCount.jar: Name of jar exported having the all the methods.  
WordCountProgram: Driver Program having the entire configuration  
input.txt: Input file  
output1: Output folder where the output file will be stored  
5. **View the Output**  
View the output in the output1 folder  
$ hadoop fs -cat /user/cloudera/output1/part-r-00000  
Hare  8  
Krishna     5  
Lord  1  
Rama  4  
Thanks      1  
book  1  
for   1  
helping     1  
this  1  
us    1  
write 1

**Q78) What problem does Apache Flume solve?**

**Scenario:**

1. There are several services producing a huge number of logs that run in different servers. These logs need to be accumulated, stored and analyzed together.
2. Hadoop has emerged as a cost effective and scalable framework for storage and analysis for big data.

**Problem:**

1. How can these logs be collected, aggregated and stored to a place where Hadoop can process them?
2. Now there is a requirement for a reliable, scalable, extensible and manageable solution.

**Q79) What is Apache Flume?**

Apache Flume is a distributed data collection service that gets flows of data (like logs) from the systems that generate them and aggregates them to a centralized data store where they can be processed together.

**Goals:** reliability, recoverability, and scalability

**Flume features:**

1. Ensures guaranteed data delivery
2. Gather high volume data streams in real time
3. Streaming data is coming from multiple sources into Hadoop for analysis
4. Scales horizontally

**Q80) How is Flume-NG different from Flume 0.9?**

**Flume 0.9:**

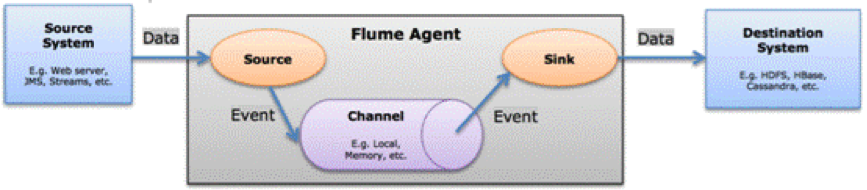
Centralized configuration of the agents handled by Zookeeper.  
Input data and writing data are handled by same thread.  
Flume 1.X (Flume-NG):  
No centralized configuration. Instead a simple on-disk configuration file is used.  
Different threads called **runners** handle input data and writing data.

**Q90) What is the problem with HDFS and streaming data (like logs)?**

1. In a regular filesystem when you open a file and write data, it exists on disk even before it is closed.
2. Whereas in HDFS, the file exists only as a directory entry of zero length till it is closed. This implies that if data is written to a file for an extended period without closing it, you may be left with an empty file if there is a network disconnect with the client.
3. It is not a good approach to close the files frequently and create smaller files as this leads to poor efficiency in HDFS.

**Q91) What are core components of Flume?**

**Flume architecture:**



**Flume Agent:**

1. An agent is a daemon (physical Java virtual machine) running Flume.
2. It receives and stores the data until it is written to a next destination.
3. Flume source, channel and sink run in an agent.

**Source:**

1. A source receives data from some application that is producing data.
2. A source writes events to one or more channels.
3. Sources either poll for data or wait for data to be delivered to them.
4. **For Example:** log4j, Avro, syslog, etc.

**Sink:**

1. A sink removes the events from the agent and delivering it to the destination.
2. The destination could be different agent or HDFS, HBase, Solr etc.
3. **For Example:**Console, HDFS, HBase, etc.

**Channel:**

1. A channel holds events passing from a source to a sink.
2. A source ingests events into the channel while sink removes them.
3. A sink gets events from one channel only.
4. **For Example:** Memory, File, JDBC etc.

**Q92) Explain a common use case for Flume?**

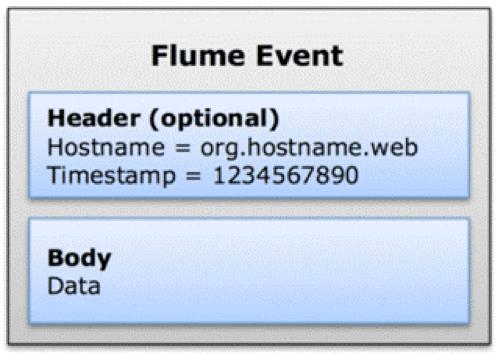
***Common Use case:***Receiving web logs from several sources into HDFS.  
Web server logs → Apache Flume → HDFS (Storage) → Pig/Hive (ETL) → HBase (Database) → Reporting (BI Tools)

1. Logs are generated by several log servers and saved in local hard disks, which need to be pushed into HDFS using Flume framework.
2. Flume agents, which are running on, log servers collect the logs, which are pushed into HDFS.
3. Data analytics tools like Pig or Hive then process this data.
4. The analysed data is stored in structured format in HBase or other database.
5. Business intelligence tools will then generate reports on this data.

**Q93) What are Flume events?**

**Flume events:**

1. Basic **payload of data** transported by Flume (typically a single log entry)
2. It has zero or more headers and a body



Event Headers are key-value pairs that are used to make routing decisions or carry other structured information like:

1. Timestamp of the event
2. Hostname of the server where event has originated

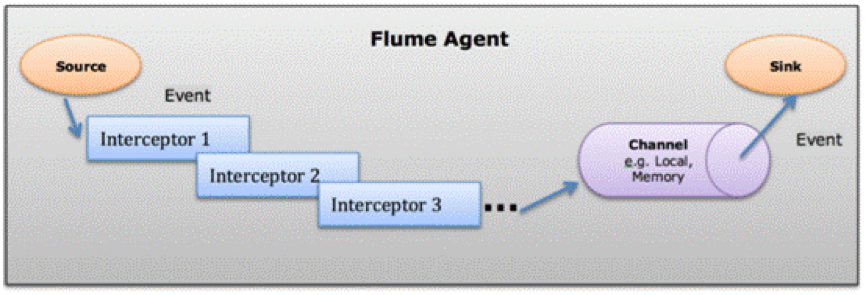
**Event Body**

Event Body is an array of bytes that contains the actual payload.

**Q94) Can we change the body of the flume event?**

Yes, editing Flume Event using interceptors can change its body.

**Q95) What are interceptors?**



**Interceptor**

An interceptor is a point in your data flow where you can inspect and alter flume events. After the source creates an event, there can be zero or more interceptors tied together before it is delivered to sink.

**Q96) What are channel selectors?**

**Channel selectors:**

Channel selectors are responsible for how an event moves from a source to one or more channels.  
Types of channel selectors are:

1. **Replicating Channel Selector:** This is the default channel selector that puts a copy of event into each channel
2. **Multiplexing Channel Selector:** Routes data into different channel depending on header information and/or interceptor logic

**Q97) What are sink processors?**

**Sink processor:**

Sink processor is a mechanism for failover and load balancing events across multiple sinks from a channel

**Q98) How to Configure an Agent?**

1. An agent is configured using a simple Java property file of key/value pairs
2. This configuration file is passed as an argument to the agent upon startup.
3. You can configure multiple agents in a single configuration file. It is required to pass an agent identifier (called a name).
4. Each agent is configured starting with:
   * agent.sources=
   * agent.channels=
   * agent.sinks=
5. Each source, channel and sink also has a distinct name within the context of that agent.

**Q99) Explain the “Hello world” example in flume.**

In the following example, the source listens on a socket for network clients to connect and sends event data. Those events were written to an in-memory channel and then fed to a log4j sink to become output.  
Configuration file for one agent (called a1) that has a source named s1, a channel named c1 and a sink named k1.  
# netcatAgent.conf: Logs the netcat events to console  
# Name of the components on this agent  
a1.sources=s1  
a1.channels=c1  
a1.sinks=k1  
# Configure the source  
a1.sources.s1.type=Netcat  
   
**Q100) What is Hive?**

Hive is a Hadoop based system for querying and analyzing large volumes of Structured data which is stored on HDFS or in other words Hive is an query engine built to work on top of Hadoop that can compile queries into Map Reduce jobs and run them on the cluster.

**Q101) In which scenario Hive is good fit?**

1. Data warehousing applications where more static data is analyzed.
2. Fast response time is not the criteria.
3. Data is not changing rapidly.
4. An abstract to underlying MapReduce programs
5. Like SQL

**Q102) What are the limitations of Hive?**

Hive does not provide:

1. Record-level operations like INSERT, DELETE or UPDATE.
2. Cannot be used for low latency jobs.
3. Transaction.

**Q103) What are the differences between Hive and RDBMS?**

HIVE:

* Schema on Read
* Batch processing jobs
* Data stored on HDFS
* Processed using MapReduce

RDBMS:

* Schema on write
* Real time jobs
* Data stored on internal structure
* Processed using database

**Q104) What are the components of Hive architecture?**

* Hive Driver
* Metastore
* Hive CLI/HUE/HWI

**Q105) What is the purpose of Hive Driver?**

Hive Driver is responsible for compiling, optimizing and then executing the HiveQL.

**Q106) What is a Metastore and what it stores?**

1. It is a database by default Derby SQL server
2. Holds metadata about table definition, column, and data types partitioning information,
3. It can be stored in MySQL, derby, oracle etc.

**Q107) What is the purpose of storing the metadata?**

People want to read the dataset with a particular schema in mind.  
For e.g.: BA and CFO of a company look at the data with a particular schema.  
BA may be interested in say IP addresses and timings of the clicks in a weblog while the CFO may be interested in say the clicks that were direct clicks on the website or from paid Google adds.  
Underneath it’s the same dataset that is accessed. This schema is used again and again. So it makes sense to store this schema in a RDBMS.

**Q108) List the various options available with the Hive command.**

**Syntax:**

$ ./hive –service serviceName  
where  
serviceName options are:  
cli  
help  
hiveserver  
hwi  
jar  
lineage  
metastore  
rcfile

**Q109) Explain the different services that can be invoked using the Hive command.**

cli

* default service
* used to define tables, run queries, etc.

hiveserver

* aemon that listens for Thrift connections from other processes

hwi

* Simple web interface for running queries

jar

* Extension of the hadoop jar command

metastore

* External Hive metastore service to support multiple clients

rcfile

* Tool for printing the contents of an RFFile

**Q110) Can you execute Hadoop dfs Commands from Hive CLI? How?**

Hadoop dfs commands can be run from within the hive CLI by dropping the hadoop work from the command and adding a semicolon in the end.

For Example:

Hadoop dfs command:  
hadoop dfs -ls /  
From within hive  
hive > dfs -ls / ;

**Q111) How to give multiline comments in Hive Scripts?**

Hive does not support multiline comments. All lines of comments should start with the string —  
For e.g.  
— This is first line of comment  
— This is second line of comment !!

**Q112) What is the reason for creating a new metastore\_db whenever Hive query is run from a different directory?**

**Embedded mode:**  
Whenever Hive runs in embedded mode, it checks whether the metastore exists. If the metastore does not exist then it creates the local metastore.

**Property:** Default value

**javax.jdo.option.ConnectionURL** = “jdbc:derby:;databaseName=metastore\_db;create=true”

**Q113) When Hive is run in embedded mode, how to share the metastore within multiple users?**

No.  
For sharing use the standalone database (like MySQL, PostGresQL) for metastore

**Q114) How can an application connect to Hive run as a server?**

Thrift Client: Hive commands can be called hive command from programming languages like Java, PHP, Python, Ruby, C++  
JDBC Driver: Type 4 (pure Java) JDBC Driver  
ODBC driver:  ODBC protocol

**Q115) List the Primitive Data Types?**

**DataTypes:**  
TINYINT

**Q116) What problem does Apache Pig solve?**

**Scenario**  
1. MapReduce paradigm presented by Hadoop is low level and rigid so developing can be challenging.  
2. Jobs are (mainly) in Java where developer needs to think in terms of map and reduce

**Problem**  
1. Many common operations like filters, projections, joins requires a custom code  
2. Not everyone is a Java expert!!!  
3. MapReduce has a long development cycle

**Q117) What is Apache Pig?**

Apache Pig is a platform for analyzing large data sets that consists high-level language for expressing data analysis programs, with infrastructure for evaluating these programs.  
**Goals:** Ease of programming, Improved Code readability, Flexible, Extensible  
**Pig Features:**  
Ease of programming:

* Generates MapReduce programs automatically
* Fewer lines of code

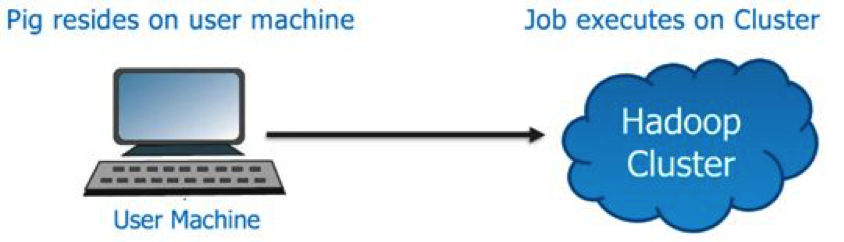
Flexible:

* Metadata is optional

Extensible:

* Easy extensible by UDFs

Resides on the client machine



**Q118) In which scenario MapReduce is a better fit than Pig?**

Some problems are harder to express in Pig. For example:

1. Complex grouping or joins
2. Combining lot of datasets
3. Replicated join
4. Complex cross products

In such cases, Pig’s MAPREDUCE relational operator can be used which allows plugging in Java MapReduce job.

**Q119) In which scenario Pig is better fit than MapReduce?**

Pig provides common data operations (joins, filters, group by, order by, union) and nested data types (tuple, bag and maps), which are missing from MapReduce.

**Q120) Where not to use Pig?**

1. Completely unstructured data. For example: images, audio, video
2. When more power to optimize the code is required
3. Retrieving a single record in a very large dataset

**Q121) What can be feed to Pig?**

We can input structured, semi-structured or unstructured data to Pig.  
*For example*, CSV’s, TSV’s, Delimited Data, Logs

**Q122) What are the components of Apache Pig platform?**

**Pig Engine**  
Parser, Optimizer and produces sequences of MapReduce programs

**Grunt**  
Pig’s interactive shell  
It allows users to enter Pig Latin interactively and interact with HDFS

**Pig Latin**  
High level and easy to understand dataflow language  
Provides ease of programming, extensibility and optimization.

**Q123) What are the execution modes in Pig?**

Pig has **two** execution modes:

**Local mode**  
No Hadoop / HDFS installation is required  
All processing takes place in only one local JVM  
Used only for quick prototyping and debugging Pig Latin script  
pig -x local

**MapReduce mode (Default)**  
Parses, checks and optimizes locally

1. Plans execution as one MapReduce job
2. Submits job to Hadoop
3. Monitors job progress

pig or pig -x mapreduce

**Q124) Different running modes for running Pig?**

Pig has **two** running modes:

**Interactive mode**  
Pig commands runs one at a time in the grunt shell

**Batch mode**  
Commands are in pig script file.

**Q125) What are the different ways to develop PigLatin scripts?**

Plugins are available which features such as syntax/error highlighting, auto completion etc.  
Eclipse plugins

1. PigEditor
2. PigPen
3. Pig-Eclipse

Vim, Emacs, TextMate plugins also available

**Q126) What are the Data types in Pig?**

**Scalar Types**  
Int, long, float, double, chararray, bytearray, boolean (since Release 0.10.0)

**Complex Types**  
Map, Tuple, Bag

**Q127) Which type in Pig is not required to fit in Memory?**

1. Bag is the type not required to fit in memory, as it can be quite large.
2. It can store bags to disk when necessary.

**Q128) What is a Map in Pig?**

Map is a chararray to data element mapping, where data element be of any Pig data type.  
It can also be called as a set of key-value pairs where  
     Keys → chararray and Values → any pig data type  
For example [‘student’#’Mahi’, ’Rank’#1]

**Q129) What is a Tuple in Pig? (~ RDBMS row in a table)**

A tuple is an ordered set of fields; fields can be of any data type.  
It can also be called as a sequence of fields of any type.

Explore Hadoop Sample Resumes! Download & Edit, Get Noticed by Top Employers![**Download Now!**](https://mindmajix.com/hadoop-sample-resumes)

**Q130) HDFS and Mapreduce Features in Hadoop Versions:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **1.x** | **0.22** | **2.x** |
| Secure Authentication | Yes | No | Yes |
| Old Configuration Names | Yes | Deprecate | Deprecate |
| New Configuration Names | No | Yes | Yes |
| Old Mapreduce API | Yes | Yes | Yes |
| New Mapreduce API | Yes | Yes | Yes |
| Mapreduce 1 Runtime | No | No | Yes |
| Mapreduce 2 Runtime | No | No | Yes |
| HDFS Federation | No | No | Yes |
| HDFS High Availability | No | No | Yes |

**Q131) What is Cluster Rebalancing?**

* The architecture of HDFS is in flow with data rebalancing schemes.
* A scheme automatically move data from one data node into another data node.

If in case, there is a sudden rise in particular file, a scheme dynamically creates additional replies and rebalance the data within cluster. This type of data rebalancing schemes are yet to be positioned.

**Q132) What is Data Integrity?**

1. Data Integrity is a method where a block of data is fetched from a datanode, but it comes corrupted. This corruption is due to faults in storage device, network and buggy software.
2. HDFS software employs checksum checking on the HDFS file contents.
3. When a client handles file contents, it verifies that data retrieved matches the checksum of the relevant checksum file.
4. If not, then the client may opt to retrieve the block from variant datanode that replicates the other block.

**Q133) What is Hadoop File System?**

1. Hadoop File System indicating the compiler to interact with Linux local environment to HDFS environment.
2. Hadoop File System is not support the -vi command. Because HDFS is write once.
3. Hadoop File System is support for -touchz command.
4. Hadoop File System looks the only HDFS directory but not local directory.
5. We can not create a file on top of HDFS.
6. We can not create the file on local.
7. We can not update a file on top of HDFS. We can updations in local, after that file is put into the HDFS.
8. Hadoop file system does not support hard links (or) soft links.
9. Hadoop File System does not implement user quotas.
10. Error implementation is sent to stderr & output is sent to stdout.

Display detailed help for a command:  
Hadoop fs - help

**Q134) User Command Archive?**

* Hadoop stores the small files inefficiently such as each file get stored in a  block & namenode has to keep the metadata information in memory so with this reason most of the namenode memory will get eat up this small files only which results in a wastage of memory.
* To avoid the same problem we use hadoop archives (or) har files (.har a the extension for all the archive files).
* When creating archive directory the input is converted to mapreduce jobs, so we can call hadoop archives as a input for our mapreduce programming.

1. Hadoop archives are special format archives.  
2. Hadoop archive maps to a file system directory.  
3. Hadoop archive always has a .har extension  
4. Hadoop archive directory contains metadata.

**Q135) What is Serialization?**

1. Serialization transforms objects into bytes
2. Hadoop utilizes PR6 for transmitting across the network.
3. Hadoop employs a very own serialization format which is writable
4. Comparison types are crucial
5. Hadoop enables a Raw comparator, that abolishes deserialization
6. External frameworks are enables via : **enter Avro**

**Q136) Datanode block scanner?**

All the datanodes runs the block scanner, which periodically verifies all the blocks stored on the datanode. This allows bad blocks to be detected and fixed before they are read by clients.  
It maintains

1. A list of blocks to verify
2. It scans them one by one for checksum errors.
3. Block scanner report can be verify by visiting http://datanode:50075/blockScannerReport.

**Q137) What is HBASE Data Storage?**

HBASE is column oriented data storage

**Column-Oriented:**

1. The reason to store values on a per column basis instead is based on the assumption
2. That for specific queries, not all of the values are needed
3. Reduced I/O
4. The data of column-oriented databases is saved in the way grouped by columns and the following column values are stored on the contiguous disk locations. This is quite different from the conventional approach followed by the traditional databases which stores all the rows contiguously.
   1. **Storefile**: Store File for each state for each region for the table.
   2. **Block**: Blocks within a store file within a store for each region for the table
5. Hlog used for recovering
   1. Send heartbeat(loadinfo) to master
   2. Write requests handle
   3. Read request handle
   4. Flush
   5. Compaction
   6. Region Splits(Manage)

**Q138) What is Hadoop Streaming?**

A utility to enable Mapreduce code in any language: C, Perl, Python, C++, Bash etc. The examples include a python mapper and an AWK reducer.

**Q139) What is the difference between Base & NOSQL?**

Favours consistencies are availability (but availability is good in practice)  
Great hadoop integration (very efficient bulk loads, Mapreduce Analysis)  
Ordered range partitions(not hash)  
Automatically shards/scales (just run on more servers)  
Sparse column stronge(not key value)

**Q140) What is HBASE Client?**

The HBase client finds the HRegion servers that serve the specific row range of interest. The HBase client, on instantiation, exchanges information with the HBase Master to locate the ROOT region. The client communicates with the region server of interest once the ROOT region is located and scans it to locate the META region that contains the user region’s location which consists of the desired row range.

**Q141) Why HBASE?**

We can infrastructure, no usage limits  
Data Model  
Semistructured data in Hbase  
Time series ordered  
Scaling is built-in (Just add more servers)  
But extra indexing is DIY  
Very active developer community  
Established, mature project (in relative terms)  
Matches our own toolset (Java/Linux based)

**Q142) What is ZOOKEEPER?**

Master election and server availability  
cluster management: Assignment transaction state management  
Client contacts zookeeper to bootstrap connection to the HBase cluster.  
Region key ranges, region server address  
Guarantees consistency of data across clients.

**List of Other Big Data Courses:**

|  |  |
| --- | --- |
| [Hadoop Adminstartion](https://mindmajix.com/hadoop-administration-training) | [MapReduce](https://mindmajix.com/mapreduce-training) |
| [Big Data On AWS](https://mindmajix.com/big-data-on-aws-training) | [Informatica Big Data Integration](https://mindmajix.com/informatica-big-data-integration-training) |
| [Bigdata Greenplum DBA](https://mindmajix.com/bigdata-greenplum-dba-training) | [Informatica Big Data Edition](https://mindmajix.com/informatica-big-data-edition-training) |
| [Hadoop Hive](https://mindmajix.com/hadoop-hive-training) | [Impala](https://mindmajix.com/impala-training) |
| [Hadoop Testing](https://mindmajix.com/hadoop-testing-training) | [Apache Mahout](https://mindmajix.com/apache-mahout-training) |

############################################################################################dyzer#####

**What kind of issues you’re facing while using cluster?**

1. Lack of configuration management.  
2. Poor allocation of resources.  
3. Lack of a dedicated network.  
4. Lack of monitoring and metrics.  
5. Ignorance of what log files contain what information.  
6. Drastic measures to address simple problems.  
7. Inadvertent introduction of single points of failure.  
8. Over reliance on defaults  
Cluster issues are somehow related to Admin team. Other task that need to be manage daily are  
1. Managing space between application users.  
2. Distcp – Data back ups and migration.  
3. Managing Services and adding nodes using Ambari .  
4. Changing cluster capacity .  
5. user/group permission management.  
6. Alerts and Notifications.  
7. Script configuration

**Mention recommend hard-disk and ram size?**

**What kind of jobs have you used can you explain?**  
Mostly we use it to schedule job at cluster node instead of running manual script each time.  
1. Alert mails are triggered when threshold value is reached.

**What trouble shooting issues you faced?**  
1. Issues can be related to cluster or logs like  
2. IO exception error  
3. Cluster in safe mode  
4. Host unreachable,  
5. Change in host identification

**Cluster maintenance & backup?**  
1. FileSystem Checksrecursively Health check up  
2. sudo -u hdfs hadoop fsck /  
3. HDFS Balancer utility  
4. sudo -u hdfs hdfs balancer -threshold <threshold-value>  
5. Adding or Decommissioning nodes to the cluster  
6. Node Failures  
7. Database and Metadata Backups for individual database dumps.  
8. Purging older log files  
9. Plan unplaned downtime.  
10. Network issue (host unreachable)

**Have you used any monitoring tools?**  
Ganglia is a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids. We haven’t used  
Ganglia is more concerned with gathering metrics and tracking them over time while Nagios has focused on being an alerting mechanism.

#########################################################################################

# Real Time Hadoop Interview Questions From Different Readers

*This entry was posted in*[*Hadoop Interview Questions for experienced and freshers*](http://hadooptutorial.info/category/interview-questions/hadoop-interview-questions-for-experienced-and-freshers/)[*Java Interview Questions*](http://hadooptutorial.info/category/interview-questions/java-interview-questions/)[*MapReduce Interview Questions*](http://hadooptutorial.info/category/interview-questions/mapreduce-interview-questions/)[*Pig Interview Questions for experienced and freshers*](http://hadooptutorial.info/category/interview-questions/pig-interview-questions-for-experienced-and-freshers/)*on*[*July 17, 2015*](http://hadooptutorial.info/2015/07/17/)*by*[*Siva*](http://hadooptutorial.info/author/siv535gmail-com/)

**Real Time Hadoop Interview Questions from Various interviews**

1. Hive – Where do you use Internal or Managed table? What scenarios?
2. In your resume, what do you mean by, “monitoring & managing MapReduce jobs"? Explain?
3. Interviewer’s Project: How to modify the RDBMs’ Nested SQL queries into Hadoop framework using Pig.
4. Sqoop: Need to know very well. Some of the current projects are importing data from other RDBMs sources into HDFS.
5. Can you join or transform tables/columns when importing using Sqoop?
6. Can you do the above with different RDBMs (not clear)?
7. How do you transfer flat files from Unix systems?
8. What is your Pig/Hive programming level (1- 10)? (Almost all interviewers asked this.)
9. Learn Scala! – Interviewer repeatedly told me.

**Other Interview Questions:**

1. Hive – Interval vs External How do you save your files in Hive
2. Sqoop – Incremental vs hast modified relate to your project
3. Sqoop – How to check if RDBMS Table Columns added/removed and how to incorporate these changes into the import job.
4. What are the challenges you’ve faced in your project? Give 2 examples.
5. How do you check Data Integrity (log files)
6. How to improve performance in your script (PIG)?
7. Tell me about your project? work.
8. How do you use Partitioning/Bucketing in your project? (Examples from your project)
9. Where do you look for answers? (user groups, Apache Web, stack overflow)
10. NOSQL- HBase – Unstructured data storage?
11. How to debug Production issue? Give example. (logs, script counters, JVM)
12. Data Ingestion
13. What is the file size you’ve used?

Dev. Environmet

Production Environmet

1. Does Hive support indexing? (How does this relate to Partition and Bucketing)
2. Pig support Conditional Loop?
3. Hive – What type of data stored?
4. Recruiter: In your experience, what is the jump from DB developer to Hadoop without Java experience?

**More Technical type Interview Questions:**

1. What functions did you use in PIG?
2. Filter – What did you filter out?
3. Join – What did you join?

**Sponsored Content**

**[Dandruff can be the major reason for your hair loss. Try this and kill the Dandruff in one shot!](https://www.1shotdandruffkiller.com/product/?utm_source=Outbrain&utm_medium=Campaign2Creative1&utm_campaign=ProductLanding&obOrigUrl=true" \t "_blank)**[1 Shot Dandruff Killer](https://www.1shotdandruffkiller.com/product/?utm_source=Outbrain&utm_medium=Campaign2Creative1&utm_campaign=ProductLanding&obOrigUrl=true" \t "_blank)

**[Easy and delicious method to lose weight without diets](http://plaingage-relance.com/1928609e-41da-49fb-a974-6b06de3d6c88?ad_id=00e479f82aff30a44d5f6878377ee28042&ad_title=Why+Every+Judge+On+Dragons%27+Lair+Backed+This+%C2%A33+Product&doc_title=Easy+and+delicious+method+to+lose+weight+without+diets&origsrcid=$origsrcid$&origsrcname=$origsrcname$&obOrigUrl=true" \t "_blank)**[ablog.website](http://plaingage-relance.com/1928609e-41da-49fb-a974-6b06de3d6c88?ad_id=00e479f82aff30a44d5f6878377ee28042&ad_title=Why+Every+Judge+On+Dragons%27+Lair+Backed+This+%C2%A33+Product&doc_title=Easy+and+delicious+method+to+lose+weight+without+diets&origsrcid=$origsrcid$&origsrcname=$origsrcname$&obOrigUrl=true" \t "_blank)

**[Watch Diana discovering her Full Powers in Wonder Woman. Start Your 30 Day Free Trial.](https://www.primevideo.com/detail/0G0X3Y41F8HX0BHVRJWXWNKZ3Q/ref=dvm_web_3pd_in_pl_s_600005102_WW_D_T1?obOrigUrl=true" \t "_blank)**[Amazon Prime Video](https://www.primevideo.com/detail/0G0X3Y41F8HX0BHVRJWXWNKZ3Q/ref=dvm_web_3pd_in_pl_s_600005102_WW_D_T1?obOrigUrl=true" \t "_blank)

**[An Easy Method to Regrow Your Lost Hair Naturally Sitting at Home](http://www.custom-trackx.com/465f4caf-325e-47d7-a0b6-fdc70c3542f2?Adid=00c95ba639be17250a962ab888406232fa&Orgsrc=$origsrcid$&section=$section_name$&obOrigUrl=true" \t "_blank)**[fitandhealth.life](http://www.custom-trackx.com/465f4caf-325e-47d7-a0b6-fdc70c3542f2?Adid=00c95ba639be17250a962ab888406232fa&Orgsrc=$origsrcid$&section=$section_name$&obOrigUrl=true" \t "_blank)

**[Shop worry-free with up to Rs. 500\* refunded returns](https://ad.doubleclick.net/ddm/trackclk/N426203.3027775OUTBRAININDIA/B21103698.221454641;dc_trk_aid=419837778;dc_trk_cid=101562614;dc_lat=;dc_rdid=;tag_for_child_directed_treatment=?obOrigUrl=true" \t "_blank)**[PayPal](https://ad.doubleclick.net/ddm/trackclk/N426203.3027775OUTBRAININDIA/B21103698.221454641;dc_trk_aid=419837778;dc_trk_cid=101562614;dc_lat=;dc_rdid=;tag_for_child_directed_treatment=?obOrigUrl=true" \t "_blank)

**[Skip login and entering card details every time you pay.](https://ad.doubleclick.net/ddm/trackclk/N426203.3027775OUTBRAININDIA/B21103698.221620911;dc_trk_aid=419808908;dc_trk_cid=101457715;dc_lat=;dc_rdid=;tag_for_child_directed_treatment=?obOrigUrl=true" \t "_blank)**[PayPal](https://ad.doubleclick.net/ddm/trackclk/N426203.3027775OUTBRAININDIA/B21103698.221620911;dc_trk_aid=419808908;dc_trk_cid=101457715;dc_lat=;dc_rdid=;tag_for_child_directed_treatment=?obOrigUrl=true" \t "_blank)

[Recommended by](http://www.outbrain.com/what-is/default/en)

1. What is your cluster size?
2. What is the file size for production environment?
3. How long does it take to run your script in Production cluster?
4. Are you planning for anything to improve the performance?
5. What size of file do you use for Development?
6. What did you work on HBase?
7. Why Hadoop? compare to RDBMS.
8. Hive – What did you do to increase the performance.
9. PIG – what did you do to increase the performance
10. What Java UDF did you write?
11. What scenario do you think you can use Java for?
12. You can process log files in RDBMS too. Why Hadoop?
13. Hive partitioning – your project example? Why?

1. Hive – What file format do you use in your work? (Avro, Parquet, Sequence file)
2. Hadoop – What is the challenge or difficulty you’ve faced?
3. PIG – What is the challenge or difficulty you’ve faced?
4. Flume – What is the challenge or difficulty you’ve faced?
5. Sqoop – What is the challenge or difficulty you’ve faced? (he didn’t ask this question)
6. How experienced are you in Linux?
7. What shell type do you use?
8. How about your experience in Cloudera Manager?
9. How about your experience in Cloudera Manager?
10. Do you use Impala? (I compared it with Hive and explained in more details)
11. How do you select the Eco system tools for your project?

**InfoSys – Interview Questions:**

As you can see, questions are mostly based on theory.

1. Why Hadoop? (Compare to RDBMS)
2. What would happen if NameNode failed? How do you bring it up?
3. What details are in the “fsimage" file?
4. What is SecondaryNameNode?
5. Explain the MapReduce processing framework? (start to end)
6. What is Combiner? Where does it fit and give an example? Preferably from your project.
7. What is Partitioner? Why do you need it and give an example? Preferably from your project.
8. Oozie – What are the nodes?
9. What are the actions in Action Node?
10. Explain your Pig project?
11. What log file loaders did you use in Pig?
12. Hive Joining? What did you join?
13. Explain Partitioning & Bucketing (based on your project)?
14. Why do we need bucketing?
15. Did you write any Hive UDFs?
16. Filter – What did you filter out?
17. HBase?
18. Flume?
19. Sqoop?
20. Zookeeper?
21. Impala? Explain the use of Impala?
22. Cassandra? What do you know about Cassandra?
23. ClickStream.
24. What is your cluster size?
25. What is the DataNode configurations? (RAM, CPU core, Disk size)
26. What is the NameNode configurations? (RAM, CPU core, Disk size)
27. How many Map slots & reducer slots configured in each DataNode? (he didn’t ask this)
28. How do you copy file from cluster to cluster?
29. What commands do you use to check to check system health, jobs, etc.?
30. Do you use Cloudera Manager to monitor and manage the jobs, cluster, etc.?
31. What is Speculative execution?
32. What do you know about Scala? (interviewer asked about the skills that I listed in my resume)

**Java Interview Questions:**

Had an array of the follwing elements: [29 12 24 18 -11 -5]  
Need an O/P of sorting of arrays ,== [12 18 24 29 -5 -11]  
Need an O/P of even and odd numbers in array ,==[12 18 24] && [29 -5 -11]//Declaring an araylist  
ArrayList<Integer> arraylist = new ArrayList<Integer>();/\* Sorting of arraylist using Collections.sort\*/Collections.sort(arraylist);

for(int counter: arraylist)  
{  
System.out.println(counter);  
}

/\*Sort array in reverse order\*/

Collections.reverseOrder(arraylist);

System.out.println(“\*\*\*\*\*\* Reverse Sorted String Array \*\*\*\*\*\*\*”);  
for (int i : stringArray)  
{  
System.out.println(i);  
}

/\* sort an array to even numbers and odd numbers\*/  
public class SortNumbers  
{  
private static int[] array = {12 18 24 29 -5 -11};  
private static List<Integer> even = new ArrayList<>();  
private static List<Integer> odd = new ArrayList<>();

public static void even(int[] arr, List even , List odd)  
{  
for(int i = 0 ; i < arr.length ; i++)  
{  
if(arr[i] % 2 == 0)  
even.add(arr[i]);  
else  
odd.add(arr[i]);  
}  
}

//To Display the even and Odd numbers  
public static void display(List<Integer> list)  
{  
for(Integer i : list)  
System.out.println(i);  
}

public static void main(String[] args){  
classify(array,even,odd);  
display(even);  
}  
}  
}

2)How to make your class compatible with Java Hash Maps?  
Overriding hashcode() and equals() method.

3)You have two tables Employee and Dept with the below columns.Select Maximum salary by Department.

Employee—–EMPID NAME SAL DEPTID  
DEPT—–dept\_id dept\_name

SELECT d.dept\_name, MAX(e.SAL) FROM Employee e,Dept d where (d.dept\_id = e.dept\_id) group by

**On 07/28/2015**

1. Tell me some List implementations?

ArrayList

Linkedlist

2. In what Purposes you use ArrayList and Linkedlist?

ArrayList for fast searching,

LinkedList,for more insertions/deletes

3. In both Arraylist and Linkedlist, which is faster?

ArrayList is faster as it containis duplicates, no sorting

Linkedlist is slow as it contains adding and removing of elements

4. Tell me some Map implementations?

HashMap (unsorted)

TreeMap (Sorted values)

LinkedHashMap( if you want near-HashMap performance and insertion-order iteration)

5. Which of the Map implementations is faster and why?

Hash map is fast as there is no need of extra burden in sorting values…

6. What Happens in Shuffle Phase in Map Reduce?

All the part files will be exchanged between reduce tasks

part files will be generated by partitioners

map output will be transferred over network…

7. What is the Fundemental Data Structure inside a HashMap?

Integer, For calculating hash value for all keys stored into buckets….Buckets are used as storage

locations…Usually Buckets are array….

8. How do you use Map Reduce methods?

map is method to parse the input records

reduce for aggregating the results reading input from map

9. What are the parameters in Mapper class

map(key, value, context)

10. What is the interface on a Main function on a Mapper?

In Mapper Class you write…..

setup()

map(key, value, context)—-( return type of map method is void…but it writes output to context)

cleanup()

11. Is it possible to get multiple Key,value pairs from the Map phase?

Yes, by concatenating two or more fields into same field.

12. Imagine you have a Server Class Computer, If you have two files of 1 GB each on Hard disk,

These files consists of Integers from smaller to larger, how do you Merge the files into one File

and generate an output of Sorted Order? Tell me the Logic

Read record by record from each file and compare first record from first file with first record in 2nd

file and same way with 2nd record b/w the files….

If first rec in 1st file < 1st record 2nd file then i will emit 1st record in 1st file and i will move cursor

of first file to 2nd record in the first file then check with 1st record in 2nd file and so on…

13. What if there are no records in one of the files in the above Scenario?

I will copy records from the remaining file as it is without comparing

14. What is the execution time of the above Program?

1-2 minutes…in Hadoop

15. If you have two files of 1 TB on two disks, you should Merge the files into one File and generate

an output of Sorted Order? What will you do?

Write all the above logic in map method of map reduce job….or reduce method

16. How the records of the two files are compared in the Map Reducer Phase?

If one of the file is small then i can read that into memory through distributed cache in setup

method of mapper class

17. What Problems you face in the Reducer Phase?

Out of Memory Problem (To overcome this problem increase the heap size

mapreduce.child.java.opts)