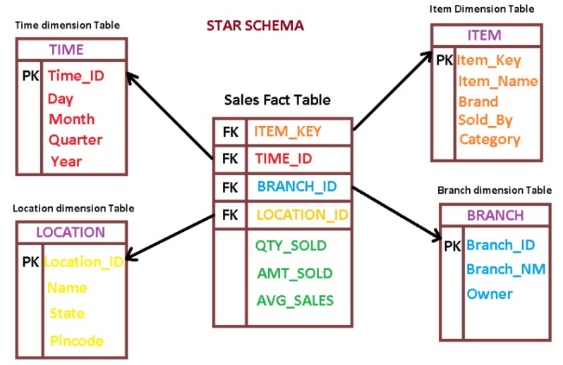
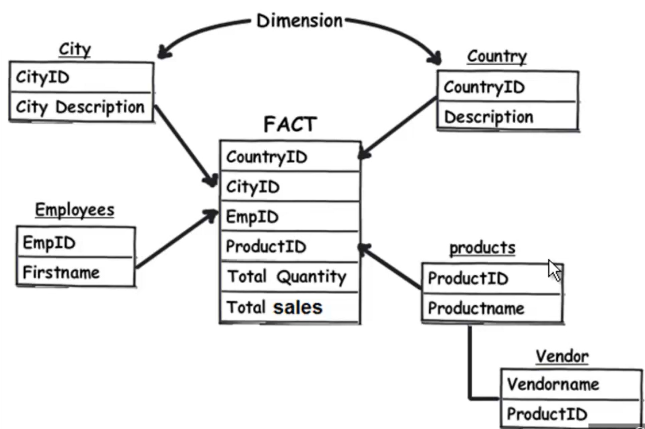
[03] List various types of design schemas in Data Modelling

Remember:

- The most important goal of OLAP (Online Analytical Processing - BI) applications is analysis of historical data.

|  |  |  |
| --- | --- | --- |
| Table | Contains | Description |
| Fact | Measurements/facts  Foreign key to dimension table | Contains all the primary keys of the dimension and associated facts or measures(is a property on which calculations can be made) like quantity sold, amount sold and average sales. |
| Dimension | Dimensions of a fact joined to fact table via a foreign key. | Describe what measures mean. Dimension tables provides descriptive information for all the measurements recorded in fact table. Dimensions are relatively very small as comparison of fact table. Commonly used dimensions are people, products, place and time |

|  |  |  |
| --- | --- | --- |
|  | Snowflake | Star Schema |
| Description | Center fact table, many dim’s and sub dim’s. Data warehouses | Center fact table, many dim’s. Datamarts with simple relationships |
| Normalization | Can have normalized dims’ | Pure denormalized dim’s |
| Redundancy | Chances of data redundancy are low. | Chances of data redundancy are high |
| Maintenance | Less redundancy, less maintenance | More redundancy (denormalized) so more maintenance. |
| Query | Complex queries | Simple queries (denormalized) |
| Joins | More joins (normalization). Provide a faster way for cube processing | Less joins  Cube processing is slow due to the complex join |
| Usage guidelines | Data integrity and less duplication | Speed & performance over data integrity |



[04] Distinguish between structured and unstructured data

|  |  |  |
| --- | --- | --- |
| Parameter | Structured Data | Unstructured Data |
| Storage | DBMS | Unmanaged file structures |
| Standard | ADO.net, ODBC and SQL | STMP. XML. CSV and SMS |
| Integration tool | ELT | Manual data entry or batch processing |
| Scaling | Schema scaling is difficult | Easy |

Structured data

- Structure is defined at design time.

- Data structure is designed in the form of tables.

Semi-structured data.

- Non-relational or NoSQL data (won’t fit neatly into tables, rows, and columns) – Uses tags or keys that organize and provide a hierarchy for the data

Nonstructured data

- Examples of nonstructured data include binary, audio, and image files

- The data structure is defined only when the data is read.

- Nonrelational systems can also support semistructured data such as JSON file formats.

- The schema of unstructured data is typically defined at query time. This means that data can be loaded onto a data platform in its native format.

Implement Relational Data Stores

[05] Explain all components of a Hadoop application

Apache Hadoop framework:

**Hadoop Common**

**Hadoop Distributed File System (HDFS)**

**Hadoop YARN**

**Hadoop MapReduce**

Hadoop Common: It is a common set of utilities and libraries that are utilized by Hadoop.

HDFS: This Hadoop application relates to the file system in which the Hadoop data is stored. It is a distributed file system having high bandwidth.

Hadoop MapReduce: It is based according to the algorithm for the provision of large-scale data processing.

Hadoop YARN: It is used for resource management within the Hadoop cluster. It can also be used for task scheduling for users.

[06] What is namenode?

Master/ name node

-- slave/ data node

-- slave/ data node

-- slave/ data node

It is the centerpiece of HDFS. It stores data of HDFS and tracks various files across the clusters. Here, the actual data is not stored. The data is stored in DataNodes.

**NameNode**

An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. It manages the metadata i.e. the data of the files is not stored on the NameNode but rather it has the directory tree of all the files present in the HDFS file system on a hadoop cluster

**DataNode**

The DataNodes are responsible for serving read and write requests from the file system's clients.

**DataNode**

The DataNodes perform block creation, deletion, and replication upon instruction from the NameNode.

Note: HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates ac

[07] Define Hadoop streaming

It is a utility which allows for the creation of the map and Reduces jobs and submits them to a specific cluster.

[09] Define Block and Block Scanner in HDFS

Blocks are the smallest unit of a data file. Hadoop automatically splits huge files into small pieces.

Block - The minimum amount of data that can be read or written is generally referred to as a “block” in HDFS. The default size of a block in HDFS is 64MB.

Block Scanner verifies the list of blocks that are presented on a DataNode.

Block Scanner - Block Scanner tracks the list of blocks present on a DataNode and verifies them to find any kind of checksum errors. Block Scanners use a throttling mechanism to reserve disk bandwidth on the datanode

[10] What are the steps that occur when Block Scanner detects a corrupted data block?

1) First of all, when Block Scanner find a corrupted data block, DataNode report to NameNode

2) NameNode start the process of creating a new replica using a replica of the corrupted block.

3) Replication count of the correct replicas tries to match with the replication factor. If the match found corrupted data block will not be deleted.

[11] Name two messages that NameNode gets from DataNode?

1) Block report

2) Heartbeat.

[12] List out various XML configuration files in Hadoop?

There are five XML configuration files in Hadoop:

Mapred-site

Core-site

HDFS-site

Yarn-site

[14] Explain the features of Hadoop

* It is an open-source framework that is available freeware.
* Hadoop is compatible with the many types of hardware and easy to access new hardware within a specific node.
* Hadoop supports faster-distributed processing of data.
* It stores the data in the cluster, which is independent of the rest of the operations.
* Hadoop allows creating 3 replicas for each block with different nodes.

[15]Explain the main methods of Reducer

**setup ()**: It is used for configuring parameters like the size of input data and distributed cache.

**cleanup()**: This method is used to clean temporary files.

**reduce()**: It is a heart of the reducer which is called once per key with the associated reduced task

[17]

[18] How to deploy a big data solution?

1) Integrate data using data sources like RDBMS, SAP, MySQL, Salesforce

2) Store data extracted data in either NoSQL database or HDFS.

3) Deploy big data solution using processing frameworks like Pig, Spark, and MapReduce.

[19] Explain FSCK

File System Check or FSCK is command used by HDFS. FSCK command is used to check inconsistencies and problem in file.

[20]

[21]

[22] Explain Hadoop distributed file system

Hadoop works with scalable distributed file systems like S3, HFTP FS, FS, and HDFS. Hadoop Distributed File System is made on the Google File System. This file system is designed in a way that it can easily run on a large cluster of the computer system.

[25] List various modes in Hadoop

1) Standalone mode

2) Pseudo distributed mode

3) Fully distributed mode.

[26] How to achieve security in Hadoop?

1) The first step is to secure the authentication channel of the client to the server. Provide time-stamped to the client.

2) In the second step, the client uses the received time-stamped to request TGS for a service ticket.

3) In the last step, the client use service ticket for self-authentication to a specific server.

[27] What is Heartbeat in Hadoop?

In Hadoop, NameNode and DataNode communicate with each other. Heartbeat is the signal sent by DataNode to NameNode on a regular basis to show its presence.

[28] Distinguish between NAS and DAS in Hadoop

|  |  |
| --- | --- |
| NAS | DAS |
| Storage capacity is to bytes | Storage capacity is bytes |
| Management cost per GB is moderate | Management cost per GB is high. |
| Transmit data using Ethernet or TCP/IP. | Transmit data using IDE/ SCSI |

[31] What is FIFO scheduling?

It is a Hadoop Job scheduling algorithm. In this FIFO scheduling, a reporter selects jobs from a work queue, the oldest job first

[32] Mention default port numbers on which task tracker, NameNode, and job tracker run in Hadoop

Task tracker runs on 50060 port

NameNode runs on 50070 port

Job Tracker runs on 50030 port

[33] How to disable Block Scanner on HDFS Data Node

In order to disable Block Scanner on HDFS Data Node, set dfs.datanode.scan.period.hours to 0.

[34] How to define the distance between two nodes in Hadoop?

The distance is equal to the sum of the distance to the closest nodes. The method getDistance() is used to calculate the distance between two nodes.

[36] Define replication factor in HDFS

Replication factor is a total number of replicas of a file in the system.

[37] What data is stored in NameNode?

Namenode stores the metadata for the HDFS like block information, and namespace information.

**NameNode**

An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients.

**DataNode**

The DataNodes are responsible for serving read and write requests from the file system's clients.

**DataNode**

The DataNodes perform block creation, deletion, and replication upon instruction from the NameNode.

[38] What do you mean by Rack Awareness?

In Haddop cluster, Namenode uses the Datanode to improve the network traffic while reading or writing any file that is closer to the nearby rack to Read or Write request. Namenode maintains the rack id of each DataNode to achieve rack information. This concept is called as Rack Awareness in Hadoop.

[39] What are the functions of Secondary NameNode?

FsImage which stores a copy of EditLog and FsImage file.

NameNode crash: If the NameNode crashes, then Secondary NameNode's FsImage can be used to recreate the NameNode.

Checkpoint: It is used by Secondary NameNode to confirm that data is not corrupted in HDFS.

Update: It automatically updates the EditLog and FsImage file. It helps to keep FsImage file on Secondary NameNode updated.

[40] What happens when NameNode is down, and the user submits a new job?

NameNode is the single point of failure in Hadoop so the user can not submit a new job cannot execute. If the NameNode is down, then the job may fail, due to this user needs to wait for NameNode to restart before running any job.

[41] What are the basic phases of reducer in Hadoop?

1. Shuffle: Here, Reducer copies the output from Mapper.

2. Sort: In sort, Hadoop sorts the input to Reducer using the same key.

3. Reduce: In this phase, output values associated with a key are reduced to consolidate the data into the final output.

[42] Why Hadoop uses Context object?

Context Object is used to help the mapper interact with other Hadoop systems. Context Object can be used for updating counters, to report the progress and to provide any application level status updates. ContextObject has the configuration details for the job and also interfaces, that helps it to generating the output.

Hadoop framework uses Context object with the Mapper class in order to interact with the remaining system. Context object gets the system configuration details and job in its constructor.

We use Context object in order to pass the information in setup(), cleanup() and map() methods. This object makes vital information available during the map operations.

main methods of Reducer

setup (): It is used for configuring parameters like input data size, distributed cache, heap size, etc.

cleanup(): This method is used to clean temporary files.

reduce(): It is a heart of the reducer which is called once per key with the associated reduced task

[43] Define Combiner in Hadoop

It is an optional step between Map and Reduce. Combiner takes the output from Map function, creates key value pairs, and submit to Hadoop Reducer. Combiner's task is to summarize the final result from Map into summary records with an identical key.

[44] What is the default replication factor available in HDFS What it indicates?

Default replication factor in available in HDFS is three. Default replication factor indicates that there will be three replicas of each data

[45] What do you mean Data Locality in Hadoop?

In a Big Data system, the size of data is huge, and that is why it does not make sense to move data across the network. Now, Hadoop tries to move computation closer to data. This way, the data remains local to the stored location.

[46] Define Balancer in HDFS

In HDFS, the balancer is an administrative used by admin staff to rebalance data across DataNodes and moves blocks from overutilized to underutilized nodes.

[47] Explain Safe mode in HDFS

It is a read-only mode of NameNode in a cluster. Initially, NameNode is in Safemode. It prevents writing to file-system in Safemode. At this time, it collects data and statistics from all the DataNodes.

[48] What is the importance of Distributed Cache in Apache Hadoop?

Hadoop has a useful utility feature so-called Distributed Cache which improves the performance of jobs by caching the files utilized by applications. An application can specify a file for the cache using JobConf configuration.

Hadoop framework makes replica of these files to the nodes one which a task has to be executed. This is done before the execution of task starts. Distributed Cache supports the distribution of read only files as well as zips, and jars files.

[49] What is Metastore in Hive?

It stores schema as well as the Hive table location.

Hive table defines, mappings, and metadata that are stored in Metastore. This can be stored in RDBMS supported by JPOX.

[50] What do mean by SerDe in Hive?

SerDe is a short name for Serializer or Deserializer. In Hive, SerDe allows to read data from table to and write to a specific field in any format you want.

[51] List components available in Hive data model

Tables

Partitions

Buckets

[52] 52) Explain the use of Hive in Hadoop eco-system.

Hive provides an interface to manage data stored in Hadoop eco-system. Hive is used for mapping and working with HBase tables. Hive queries are converted into MapReduce jobs in order to hide the complexity associated with creating and running MapReduce jobs.

[53] List various complex data types/collection are supported by Hive

Hive supports the following complex data types:

Map

Struct

Array

Union

[54] Explain how .hiverc file in Hive is used?

In Hive, .hiverc is the initialization file. This file is initially loaded when we start Command Line Interface (CLI) for Hive. We can set the initial values of parameters in .hiverc file.

[55] Is it possible to create more than one table in Hive for a single data file?

Yes, we can create more than one table schemas for a data file. Hive saves schema in Hive Metastore. Based on this schema, we can retrieve dissimilar results from same Data.

[56] Explain different SerDe implementations available in Hive

There are many SerDe implementations available in Hive. You can also write your own custom SerDe implementation. Following are some famous SerDe implementations:

OpenCSVSerde

RegexSerDe

DelimitedJSONSerDe

ByteStreamTypedSerDe

[57] List table generating functions available in Hive

Following is a list of table generating functions:

Explode(array)

JSON\_tuple()

Stack()

Explode(map)

[58] What is a Skewed table in Hive?

A Skewed table is a table that contains column values more often. In Hive, when we specify a table as SKEWED during creation, skewed values are written into separate files, and remaining values go to another file.

[59] List out objects created by create statement in MySQL.

Objects created by create statement in MySQL are as follows:

Database

Index

Table

User

Procedure

Trigger

Event

View

Function

[60] How to see the database structure in MySQL?

In order to see database structure in MySQL, you can use

DESCRIBE command. Syntax of this command is DESCRIBE Table name;

[61] How to search for a specific String in MySQL table column?

Use regex operator to search for a String in MySQL column. Here, we can also define various types of regular expression and search for using regex.

[LI] When should you use HBase and what are the key components of HBase?

HBase should be used when the big data application has –

1)A variable schema

2)When data is stored in the form of collections

3)If the application demands key based access to data while retrieving.

Key components of HBase are –

Region- This component contains memory data store and Hfile.

Region Server-This monitors the Region.

HBase Master-It is responsible for monitoring the region server.

Zookeeper- It takes care of the coordination between the HBase Master component and the client.

Catalog Tables-The two important catalog tables are ROOT and META.ROOT table tracks where the META table is and META table stores all the regions in the system.

[LI] Explain about some important Sqoop commands other than import and export.

**Create Job (--create)**

Here we are creating a job with the name my job, which can import the table data from RDBMS table to HDFS. The following command is used to create a job that is importing data from the employee table in the db database to the HDFS file.

$ Sqoop job --create myjob \

--import \

--connect jdbc:mysql://localhost/db \

--username root \

--table employee --m 1

**Verify Job (--list)**

‘--list’ argument is used to verify the saved jobs. The following command is used to verify the list of saved Sqoop jobs.

$ Sqoop job --list

**Inspect Job (--show)**

‘--show’ argument is used to inspect or verify particular jobs and their details. The following command and sample output is used to verify a job called myjob.

$ Sqoop job --show myjob

**Execute Job (--exec)**

‘--exec’ option is used to execute a saved job. The following command is used to execute a saved job called myjob.

$ Sqoop job --exec myjob

[LI] Explain about the core components of Flume.

Event- The single log entry or unit of data that is transported.

Source- This is the component through which data enters Flume workflows.

Sink-It is responsible for transporting data to the desired destination.

Channel- it is the duct between the Sink and Source.

Agent- Any JVM that runs Flume.

Client- The component that transmits event to the source that operates with the agent.

[LI] Can Apache Kafka be used without Zookeeper?

It is not possible to use Apache Kafka without Zookeeper because if the Zookeeper is down Kafka cannot serve client request.

[LI] What do you mean by a bag in Pig?

Collection of tuples is referred as a bag in Apache Pig

[LI] What is a Hive Metastore?

Hive Metastore is a central repository that stores metadata in external database.

LinkedIn Assessment

[01] Partitioner controls the partitioning of what data?

final keys

final values

**intermediate keys**

intermediate values

[02] SQL Windowing functions are implemented in Hive using which keywords?

UNION DISTINCT, RANK

**OVER, RANK**

OVER, EXCEPT

UNION DISTINCT, RANK

[03] Rather than adding a Secondary Sort to a slow Reduce job, it is Hadoop best practice to perform which optimization?

Add a partitioned shuffle to the Map job.

**Add a partitioned shuffle to the Reduce job.**

Break the Reduce job into multiple, chained Reduce jobs.

Break the Reduce job into multiple, chained Map jobs.

[04] Hadoop Auth enforces authentication on protected resources. Once authentication has been established, it sets what type of authenticating cookie?

encrypted HTTP

unsigned HTTP

compressed HTTP

**signed HTTP**

[05] MapReduce jobs can be written in which language?

**Java or Python**

SQL only

SQL or Java

Python or SQ

[06] To perform local aggregation of the intermediate outputs, MapReduce users can optionally specify which object?

Reducer

**Combiner**

Mapper

Counter

[07] To verify job status, look for the value \_\_\_ in the \_\_\_.

SUCCEEDED; syslog

**SUCCEEDED; stdout**

DONE; syslog

DONE; stdout

[08] Which line of code implements a Reducer method in MapReduce 2.0?

**public void reduce(Text key, Iterator values, Context context){…}**

public static void reduce(Text key, IntWritable[] values, Context context){…}

public static void reduce(Text key, Iterator values, Context context){…}

public void reduce(Text key, IntWritable[] values, Context context){…}

[09] To get the total number of mapped input records in a map job task, you should review the value of which counter?

FileInputFormatCounter

FileSystemCounter

JobCounter

**TaskCounter (NOT SURE)**

[10] Hadoop Core supports which CAP capabilities?

**A, P**

C, A

C, P

C, A, P

[11] What are the primary phases of a Reducer?

combine, map, and reduce

**shuffle, sort, and reduce**

reduce, sort, and combine

map, sort, and combine

[12] To set up Hadoop workflow with synchronization of data between jobs that process tasks both on disk and in memory, use the \_\_\_ service, which is \_\_\_.

Oozie; open source

Oozie; commercial software

Zookeeper; commercial software

**Zookeeper; open source**

[13] For high availability, use multiple nodes of which type?

data

**name**

memory

worker

[14] DataNode supports which type of drives?

**hot swappable**

cold swappable

warm swappable

non-swappable

[15] Which method is used to implement Spark jobs?

on disk of all workers

on disk of the master node

in memory of the master node

**in memory of all workers**

[16] In a MapReduce job, where does the map() function run?

on the reducer nodes of the cluster

**on the data nodes of the cluster (NOT SURE)**

on the master node of the cluster

on every node of the cluster

[17] To reference a master file for lookups during Mapping, what type of cache should be used?

**distributed cache**

local cache

partitioned cache

cluster cache

[18] Skip bad records provides an option where a certain set of bad input records can be skipped when processing what type of data?

cache inputs

reducer inputs

intermediate values

**map inputs**

[19] Which command imports data to Hadoop from a MySQL database?

spark import --connect jdbc:mysql://mysql.example.com/spark --username spark --warehouse-dir user/hue/oozie/deployments/spark

sqoop import --connect jdbc:mysql://mysql.example.com/sqoop --username sqoop --warehouse-dir user/hue/oozie/deployments/sqoop

**sqoop import --connect jdbc:mysql://mysql.example.com/sqoop --username sqoop --password sqoop --warehouse-dir user/hue/oozie/deployments/sqoop**

spark import --connect jdbc:mysql://mysql.example.com/spark --username spark --password spark --warehouse-dir user/hue/oozie/deployments/spark

[20] In what form is Reducer output presented?

**compressed (NOT SURE)**

sorted

not sorted

encrypted

[21] Which library should be used to unit test MapReduce code?

JUnit

XUnit

**MRUnit**

HadoopUnit

[22] If you started the NameNode, then which kind of user must you be?

hadoop-user

**super-user**

node-user

admin-user

[23] State \_\_ between the JVMs in a MapReduce job

can be configured to be shared

is partially shared

is shared

**is not shared (**[**https://www.lynda.com/Hadoop-tutorials/Understanding-Java-virtual-machines-JVMs/191942/369545-4.html**](https://www.lynda.com/Hadoop-tutorials/Understanding-Java-virtual-machines-JVMs/191942/369545-4.html)**)**

[24] To create a MapReduce job, what should be coded first?

a static job() method

**a Job class and instance (NOT SURE)**

a job() method

a static Job class

[25] To connect Hadoop to AWS S3, which client should you use?

**S3A**

S3N

S3

the EMR S3

[26] HBase works with which type of schema enforcement?

schema on write

no schema

external schema

schema on read

[27] HDFS file are of what type?

read-write

read-only

write-only

append-only

[28] A distributed cache file path can originate from what location?

hdfs or top

http

hdfs or http

hdfs

[29] Which library should you use to perform ETL-type MapReduce jobs?

Hive

Pig

Impala

Mahout

[30] What is the output of the Reducer?

a relational table

an update to the input file

a single, combined list

**a set of <key, value> pairs**

map function processes a certain key-value pair and emits a certain number of key-value pairs and the Reduce function processes values grouped by the same key and emits another set of key-value pairs as output.

[31] To optimize a Mapper, what should you perform first?

Override the default Partitioner.

Skip bad records.

Break up Mappers that do more than one task into multiple Mappers.

Combine Mappers that do one task into large Mappers.

[32] When implemented on a public cloud, with what does Hadoop processing interact?

**files in object storage**

graph data in graph databases

relational data in managed RDBMS systems

JSON data in NoSQL databases

[33] In the Hadoop system, what administrative mode is used for maintenance?

data mode

**safe mode**

single-user mode

pseudo-distributed mode

Tutorialspoint mock tests

<https://www.tutorialspoint.com/hadoop/pdf/hadoop_mock_test_i.pdf>

<https://www.tutorialspoint.com/hadoop/pdf/hadoop_mock_test_ii.pdf>

<https://www.tutorialspoint.com/hadoop/pdf/hadoop_mock_test_iii.pdf>

<https://www.tutorialspoint.com/hadoop/pdf/hadoop_mock_test_iv.pdf>

[01] Which of the following is not a scheduling option available in YARN

**A - Balanced scheduler**

B - Fair scheduler

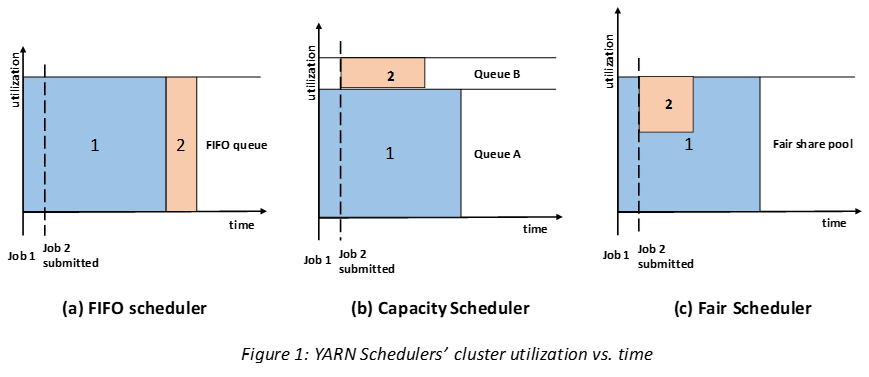
C - Capacity scheduler

D - FiFO schesduler.

FIFO, a small job blocks until the large job complete.

Capacity scheduler maintains a separate queue for small jobs in order to start them as soon a request initiates.

Fair scheduler does not have any requirement to reserve capacity. It dynamically balances the resources into all accepted jobs. When a job starts — if it is the only job running — it gets all the resources of the cluster



[02] Which of the following technologies is a document store database?

A - HBase

B - Hive

C - Cassandra

**D – CouchDB**

Apache Cassandra: A NoSQL database distributed database from Apache that is highly scalable and designed to manage very large amounts of structured data

Apache Hive: SQL-like interface called HiveQL. Facilitates reading, writing, and managing large datasets residing in distributed storage and queried using SQL syntax. Built on top of Hadoop.

Apache HBase: A NoSQL database built on Hadoop that provides random access and strong consistency for large amounts of unstructured and semi-structured data–potentially billions of rows times millions of columns

It's commonly used for search engines, and has automatic failover

Apache CouchDB: A clustered database that allows you to run a single logical database server on any number of servers or VMs.

[03] When archiving Hadoop files, which of the following statements are true? (Choose two answers)

Archived files will display with the extension .arc.

**Many small files will become fewer large files.**

**MapReduce processes the original files names even after files are archived.**

Archived files must be UN archived for HDFS and MapReduce to access the original, small files.

Archive is intended for files that need to be saved but no longer accessed by HDFS.

Hadoop Archives or HAR files, are a file archiving facility that packs files into HDFS blocks more efficiently, there by reducing Name Node memory usage while still allowing transparent access to files,

• Hadoop Archives can be used as input to map reduce.

[04] Which scenario demands highest bandwidth for data transfer between nodes in Hadoop?

A - Different nodes on the same rack

B - Nodes on different racks in the same data center.

**C - Nodes in different data centers**

D - Data on the same node.

[05] The default replication factor for HDFS file system in hadoop is

A - 1

B - 2

**C - 3**

D – 4

Replication Factor is nothing but it is a process of making replicate or duplicate’s of data. 3 Replica or copy of each file block is made

[06] The property used to set the default filesystem for Hadoop in core-site.xml is-

A - filesystem.default

**B - fs.default**

C - fs.defaultFS

D - hdfs.default

[07] What mechanisms Hadoop uses to make namenode resilient to failure.

A - Take backup of filesystem metadata to a local disk and a remote NFS mount.

**B - Store the filesystem metadata in cloud.**

C - Use a machine with at least 12 CPUs

D - Using expensive and reliable hardware.

[08] If the IP address or hostname of a datanode changes

A - The namenode updates the mapping between file name and block name

**B - The namenode need not update mapping between file name and block name**

C - The data in that data node is lost forever

D - There namenode has to be restarted

[09] The command used to copy a directory form one node to another in HDFS is

A - rcp

B - dcp

C - drcp

**D – distcp**

[20] The number of tasks a task tracker can accept depends on

A - Maximum memory available in the node

B - Not limited

**C - Number of slots configured in it**

D - As decided by the jobTracker

[20] hadoop fs –expunge

A - Gives the list of datanodes

B - Used to delete a file

C - Used to exchange a file between two datanodes.

**D - Empties the trash.**

[21] The role of a Journal node is to

A - Report the location of the blocks in a data node

**B - Report the edit log information of the blocks in the data node.**

C - Report the Schedules when the jobs are going to run

D - Report the activity of various components handled by resource manager

[22] The heartbeat signal are sent from

A - Jobtracker to Tasktracker

B - Tasktracker to Job tracker

C - Jobtracker to namenode

D - Tasktracker to namenode

[23] In HDFS the files cannot be

A - read

B - deleted

**C - executed**

D – Archived

[24] Which of the below apache system deals with ingesting streaming data to hadoop

A – Ozie workflow scheduler

B – Kafka event streaming

**C – Flume simple and flexible architecture based on streaming data flows.**

D – Hive

[25] Which one of the following is not a main component of HBase?

A - Region Server.

**B - Nagios.**

C - ZooKeeper.

D - Master Server.

[26] The client reading the data from HDFS filesystem in Hadoop

A - gets the data from the namenode

B - gets the block location from the datanode

**C - gets only the block locations from the namenode**

D - gets both the data and block location from the namenode

[27] In a Hadoop cluster, what is true for a HDFS block that is no longer available due to disk corruption or machine failure?

A - It is lost for ever

**B - It can be replicated form its alternative locations to other live machines.**

C - The namenode allows new client request to keep trying to read it.

D - The Mapreduce job process runs ignoring the block and the data stored in it.

[28] In the secondary namenode the amount of memory needed is

A - Similar to that of primary node

B - Should be at least half of the primary node

C - Must be double of that of primary node

D - Depends only on the number of data nodes it is going to handle

[29] You have loads of data that can be processed by you MRjobs. However you need the data to be available to Analysts and Scients in you organisation. What is the best format to represent the input?

A - Sequence File.

**B - Avro.**

C - XML.

D - HTML.

[30] Which of the following is not a goal of HDFS?

A - Fault detection and recovery

B - Handle huge dataset

**C - Prevent deletion of data**

D - Provide high network bandwidth for data movement

[31] HDFS files are designed for

A - Multiple writers and modifications at arbitrary offsets.

B - Only append at the end of file

C - Writing into a file only once.

D - Low latency data access.

[32] The command “hadoop fs -test -z URI “ gives the result 0 if

A - if the path is a directory

B - if the path is a file

C - if the path is not empty

**D - if the file is zero length**

[33] The Zookeeper

**A - Detects the failure of the namenode and elects a new namenode.**

B - Detects the failure of datanodes and elects a new datanode.

C - Prevents the hardware from overheating by shutting them down.

D - Maintains a list of all the components IP address of the Hadoop cluster.

[34] The org.apache.hadoop.io.Writable interface declares which two methods? (Choose 2 answers.)

**public void readFields(DataInput).**

public void read(DataInput).

public void writeFields(DataOutput).

**public void write(DataOutput).**

[35] As compared to RDBMS, Hadoop

A - Has higher data Integrity.

B - Does ACID transactions

C - IS suitable for read and write many times

**D - Works better on unstructured and semi-structured data.**

[36] In Hadoop 2.x release HDFS federation means

A - Allowing namenodes to communicate with each other.

B - Allow a cluster to scale by adding more datanodes under one namenode.

**C - Allow a cluster to scale by adding more namenodes.**

D - Adding more physical memory to both namenode and datanode.

[37] Which of the below property gets configured on hadoop-env.sh?

A - Replication factor

B - Directory names to store hdfs files

C - Host and port where MapReduce task runs

D - Java Environment variables.

[38] What is the main problem faced while reading and writing data in parallel from multiple disks?

A - Processing high volume of data faster.

**B - Combining data from multiple disks.**

C - The software required to do this task is extremely costly.

D - The hardware required to do this task is extremely costly.

[39] When a jobTracker schedules a task is first looks for

A - A node with empty slot in the same rack as datanode

B - Any node on the same rack as the datanode

C - Any node on the rack adjacent to rack of the datanode

D - Just any node in the cluster

[40] When a client contacts the namenode for accessing a file, the namenode responds with

A - Size of the file requested.

B - Block ID of the file requested.

C - Block ID and hostname of any one of the data nodes containing that block.

**D - Block ID and hostname of all the data nodes containing that block.**

[41] As part of the HDFS high availability a pair of primary namenodes are configured. What is true for them?

A - When a client request comes, one of them chosen at random serves the request.

B - One of them is active while the other one remains powered off.

C - Datanodes send block reports to only one of the namenodes.

**D - The standby node takes periodic checkpoints of active namenode’s namespace.**

[42] When the namenode finds that some blocks are over replicated, it

A - Stops the replication job in the entire hdfs file system.

B - It slows down the replication process for those blocks

**C - It deletes the extra blocks.**

D - It leaves the extra blocks as it is.

[43] What is HBASE?

A - Hbase is separate set of the Java API for Hadoop cluster.

**B - Hbase is a part of the Apache Hadoop project that provides interface for scanning large amount of data using Hadoop infrastructure.**

C - Hbase is a "database" like interface to Hadoop cluster data.

D - HBase is a part of the Apache Hadoop project that provides a SQL like interface for data processing.

[44] The decommission feature in hadoop is used for

A - Decommissioning the namenode

B - Decommissioning the data nodes

C - Decommissioning the secondary namenode.

D - Decommissioning the entire Hadoop cluster.

[45] When running on a pseudo distributed mode the replication factor is set to

A - 2

**B - 1**

C - 0

D - 3

[46] When a backup node is used in a cluster there is no need of

**A - Check point node**

B - Secondary name node

C - Secondary data node

D - Rack awareness

[46] The input split used in MapReduce indicates

A - The average size of the data blocks used as input for the program

**B - The location details of where the first whole record in a block begins and the last whole record in the block ends.**

C - Splitting the input data to a MapReduce program into a size already configured in the mapred-site.xml

D - None of these

[46] When

Data Engineering Summary

[01] Explain Data Engineering.

Data engineering is a term used in big data. It focuses on the application of data collection and research. The data generated from various sources are just raw data. Data engineering helps to convert this raw data into useful information.

[02] What is Data Modelling?

Data modelling is the method of documenting complex software design as a diagram so that anyone can easily understand. It is a conceptual representation of data objects that are associated between various data objects and the rules.

[08] What is the full form of HDFS?

HDFS stands for Hadoop Distributed File System.

[13] What are four V's of big data?

**Velocity**

**Variety**

**Volume**

**Veracity**

Value

[16] What is the abbreviation of COSHH?

The abbreviation of COSHH is Classification and Optimization based Schedule for Heterogeneous Hadoop systems.

[23] Explain the main responsibilities of a data engineer

Data engineers have many responsibilities. They manage the source system of data. Data engineers simplify complex data structure and prevent the reduplication of data. Many times they also provide ELT and data transformation.

[24] What is the full form of YARN?

The full form of YARN is Yet Another Resource Negotiator.

[29] List important fields or languages used by data engineer

Probability as well as linear algebra

Machine learning

Trend analysis and regression

Hive QL and SQL databases

[26] What is Big Data?

It is a large amount of structured and unstructured data, that cannot be easily processed by traditional data storage methods. Any data that cannot be stored into traditional RDBMS is termed as Big Data. Data engineers are using Hadoop to manage big data.

[35] Why use commodity hardware in Hadoop?

Commodity hardware is easy to obtain and affordable. It is a system that is compatible with Windows, MS-DOS, or Linux.

[26] List

[26] List

[62] Explain how data analytics and big data can increase company revenue?

Following are the ways how data analytics and big data can increase company revenue:

Use data efficiently to make sure that business growth.

Increase customer value.

Turning analytical to improve staffing levels forecasts.

Cutting down the production cost of the organization