1.what shell scripting

A shell script is program composed of a series of operating system commands that are executed by the command-line interpreter or the shell

2. hive table

3.sqoop increment loading

4.self join

5.if table with n rows and other table with m rows then table rows will be m\*n rows

m x n rows table will come

6 how can you fetch 20 th row from the file

sed -n 20p student

20 th line will from file name student

7, what happen i delete file in hdfs

file will be deleted ,it will not be in trash because trash we disable bydefault

In [**Hadoop**](http://data-flair.training/blogs/hadoop-introduction-tutorial-quick-guide/) to enable the trash feature and set the time delay for the trash removal, we can set the **fs.trash.interval** property in **core-site.xml** to the delay (in minutes).  
For example, if you want users to have 24 hours (1,440 minutes) to restore a deleted file, you should specify following configuration parameter in the core-site.xml

<property>

<name>fs.trash.interval</name>

<value>1440</value>

</property>

8, what happen if i delete table in hive

temporary will be deleted automatically once we close the terminal

managed table data and metastore will delete permanently

external table meta store will be deleted but data will be in hdfs along with path at the time of creation

9.how rum queris automatically

crontab and oozie

crontab -e open crontab

min hour data month weekday <Command to be execete> we have to mention

OOzie

We have to write workflow.xml file

then mention in OOzie job.properties with time and date mentioning

10.what happen when we submit hive job end to end

* when User Interface (UI) calls the execute interface to the Driver.
* The driver creates a session handle for the query. Then it sends the query to the compiler to generate an execution plan.
* The compiler needs the metadata. So it sends a request for *getMetaData*. Thus receives the *sendMetaData* request from Metastore.
* Now compiler uses this metadata to type check the expressions in the query. The compiler generates the plan which is **DAG** of stages with each stage being either a [**map/reduce job**](http://data-flair.training/blogs/hadoop-mapreduce-job-execution-flow/), a metadata operation or an operation on [**HDFS**](http://data-flair.training/blogs/apache-hadoop-hdfs-introduction-tutorial/). The plan contains map operator trees and a reduce operator tree for map/reduce stages.
* Now execution engine submits these stages to appropriate components. After in each task the deserializer associated with the table or intermediate outputs is used to read the rows from HDFS files. Then pass them through the associated operator tree. Once it generates the output, write it to a temporary HDFS file through the serializer. Now temporary file provides the subsequent map/reduce stages of the plan. Then move the final temporary file to the table’s location for DML operations.
* Now for queries, execution engine directly read the contents of the temporary file from HDFS as part of the fetch call from the Driver.

11, hive optimisation

1. map-side join

2.proper using of partitions and bucketization

3.enabling vetorization

4.storing orc format

5.parallel execution

 set hive.exec.parallel=true;

 set hive.exec.parallel.thread.number=8;

6.using hive indexing

CREATE INDEX olympic\_index

 ON TABLE olympic (age)

 AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'

 WITH DEFERRED REBUILD;

7.use tez engine for preocessing

set hive.execution.engine=tez;

8.increasing split size

9. spark support run hive on spark and using set hive.execution.engine =spark

12,what schedule used in your job

capacity ,

13 when i delete file from hdfs then how to restore it , if yes then where it store

it delete permanently but if trash enable then file be in trash for particular time interval

if we restore then it restore in same place

14.what happened in spark when we submit job

1.spark-submit then it lunches drive program and main method invoke by the user

2. driver program ask for cluster manager for launch executors

3.excutores launched y cluster manager behalf of driver program

4.driver program process the application like transformation and action

5.driver sent work to executors in form of task

6.executor execution and store in the hard disk( output path)

7 driver exit from the main method

15 how many language support spark

4 languages scal java python R

4.A **self join** is a **join** in which a table is joined with itself (which is also called Unary relationships), especially when the table has a FOREIGN KEY which references its own PRIMARY KEY. To **join** a table itself means that each row of the table is combined with itself and with every other row of the table.

1. scripts are considered to be [scripting languages](https://en.wikipedia.org/wiki/Scripting_language). Typical operations performed by shell scripts include file manipulation, program execution, and printing text. A script which sets up the environment, runs the program,

2.hive has 3 types of tables temporary ,managed, external tables

:temporary tables is session based table, tables will be deleted once we close the terminal

:managed tables is default table in hive ,its accessed by only one user who created that table

if we delete the managed table then metadata and table will be deleted permanently

:external table is the accessed by all user in cluster, we have to mention external key word while creating table ,

if we delete the external table then meta data will be deleted and data will be there in hdfd with path

3: we use the below statements

incremental append

check-column <id> incremental column

last-value 0

we create the sqoop job with sqoop job -- create

then we exec the sqoop job with sqoop job --exec

that key word we mention in crontab or oozie for run regularly