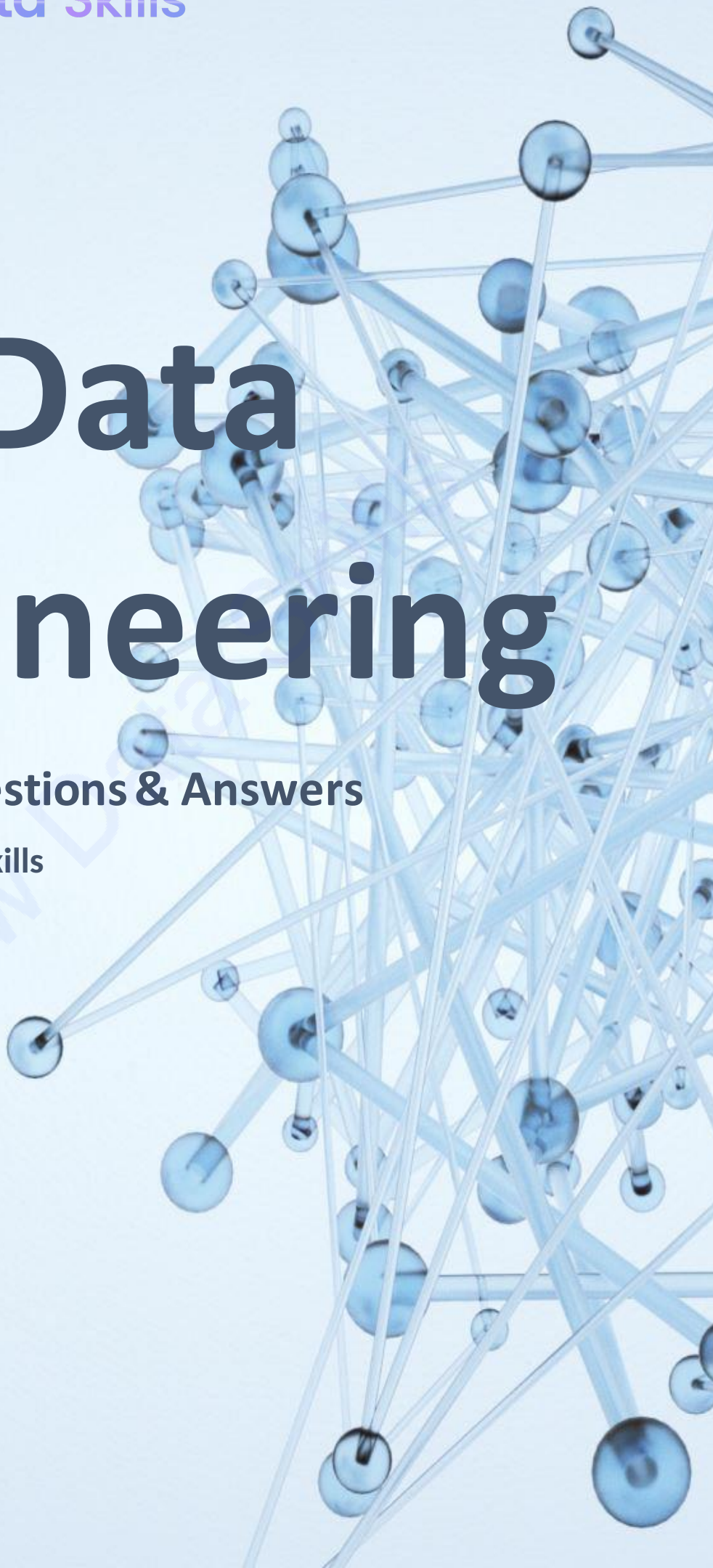
Grow Data Skills



Big Data Engineering

Interview Questions & Answers

Source: GrowDataSkills

TABLES OF CONTENTS

S.No	DESCRIPTION	PAGE No.
1.	SQL INTERVIEW QUESTION	01
2.	SCALA PROGRAMMING	30
3.	SQOOP INTERVIEW QUESTION	54
4.	HIVE INTERVIEW QUESTION	69
5.	SPARK INTERVIEW QUESTION	97
6.	TOP 250+ INTEVIEW QUESTION	130
7.	MISC. INTERVIEW QUESTION	167

SQL Interview Question with Answers

1. What is the SQL server query execution sequence?

○ FROM -> goes to Secondary files via primary file

○ WHERE -> applies filter condition (non-aggregate column) ○ SELECT -> dumps data in tempDB system database ○ GROUP BY -> groups data according to grouping predicate ○ HAVING -> applies filter condition (aggregate function) ○ ORDER BY -> sorts data ascending/descending

2. What is Normalization?

Step by step process to reduce the degree of data redundancy.

Breaking down one big flat table into multiple table based on normalization rules. Optimizing the memory but not in term of performance.

Normalization will get rid of insert, update and delete anomalies.

Normalization will improve the performance of the delta operation (aka. DML operation); UPDATE, INSERT, DELETE

Normalization will reduce the performance of the read operation; SELECT

3. What are the three degrees of normalization and how is normalization done in each degree?

1NF:

A table is in 1NF when: All the attributes are single-valued.

With no repeating columns (in other words, there cannot be two different columns with the same information).

With no repeating rows (in other words, the table must have a primary key).

All the composite attributes are broken down into its minimal component.

There should be SOME (full, partial, or transitive) kind of functional dependencies between non-key and key attributes.

99% of times, it's usually 1NF.

2NF:

A table is in 2NF when: ● It is in 1NF.

● There should not be any partial dependencies so they must be removed if they exist.

3NF:

A table is in 3NF when: ● It is in 2NF.

● There should not be any transitive dependencies so they must be removed if they exist.

BCNF:

■ A stronger form of 3NF so it is also known as 3.5NF

■ We do not need to know much about it. Just know that here you compare between a prime attribute and a prime attribute and a non-key attribute and a non-key attribute.

4. What are the different database objects ?

There are total seven database objects (6 permanent database object + 1 temporary database object)

Permanent DB objects

- Table
- Views
- Stored procedures
- User-defined Functions
- Triggers
- Indexes

Temporary DB object

- Cursors

5. What is collation?

Bigdata Hadoop: SQL Interview Question with Answers

Collation is defined as set of rules that determine how character data can be sorted and compared. This can be used to compare A and, other language characters and also depends on the width of the characters.

ASCII value can be used to compare these character data.

6. What is a constraint and what are the seven constraints?

Constraint: something that limits the flow in a database.

○ 1. Primary key

○ 2. Foreign key

○ 3. Check

■ Ex: check if the salary of employees is over 40,000

○ 4. Default

■ Ex: If the salary of an employee is missing, place it with the default value.

○ 5. Nullability

■ NULL or NOT NULL

○ 6. Unique Key

○ 7. Surrogate Key

■ mainly used in data warehouse

7. What is a Surrogate Key ?

'Surrogate' means 'Substitute'.

Surrogate key is always implemented with a help of an identity column.

Identity column is a column in which the value are automatically generated by a SQL Server based on the seed value and incremental value.

Identity columns are ALWAYS INT, which means surrogate keys must be INT. Identity columns cannot have any NULL and cannot have repeated values. Surrogate key is a logical key.

8. What is a derived column , hows does it work , how it affects the performance of a database and how can it be improved?

The Derived Column a new column that is generated on the fly by applying expressions to transformation input columns.

Ex: FirstName + ' ' + LastName AS 'Full name'

Derived column affect the performances of the data base due to the creation of a temporary new column.

Execution plan can save the new column to have better performance next time.

9. What is a Transaction?

○ It is a set of TSQL statement that must be executed together as a single logical unit. ○ Has ACID properties:

Atomicity: Transactions on the DB should be all or nothing. So transactions make sure that any operations in the transaction happen or none of them do.

Consistency: Values inside the DB should be consistent with the constraints and integrity of the DB before and after a transaction has completed or failed.

Isolation: Ensures that each transaction is separated from any other transaction occurring on the system.

Durability: After successfully being committed to the RDMBS system the transaction will not be lost in the event of a system failure or error.

○ Actions performed on explicit transaction:

BEGIN TRANSACTION: marks the starting point of an explicit transaction for a connection.

COMMIT TRANSACTION (transaction ends): used to end an transaction successfully if no errors were encountered. All DML changes made in the transaction become permanent.

ROLLBACK TRANSACTION (transaction ends): used to erase a transaction which errors are encountered. All DML changes made in the transaction are undone.

SAVE TRANSACTION (transaction is still active): sets a savepoint in a transaction. If we roll back, we can only rollback to the most recent savepoint. Only one save point is possible per transaction. However, if you nest Transactions within a Master Trans, you may put Save points in each nested Tran. That is how you create more than one Save point in a Master Transaction.

10. What are the differences between OLTP and OLAP?

OLTP stands for Online Transactional Processing

OLAP stands for Online Analytical Processing

OLTP:

Normalization Level: highly normalized

Data Usage : Current Data (Database)

Processing : fast for delta operations (DML)

Operation : Delta operation (update, insert, delete) aka DML Terms Used : table, columns and relationships

OLAP:

Normalization Level: highly denormalized

Data Usage : historical Data (Data warehouse)

Processing : fast for read operations

Operation : read operation (select)

Terms Used : dimension table, fact table

11. How do you copy just the structure of a table?

```
SELECT * INTO NewDB.TBL_Structure
FROM OldDB.TBL_Structure
WHERE 1=0 -- Put any condition that does not make any sense.
```

12.What are the different types of Joins?

○ INNER JOIN: Gets all the matching records from both the left and right tables based on joining columns.

○ LEFT OUTER JOIN: Gets all non-matching records from left table & AND one copy of matching records from both the tables based on the joining columns.

○ RIGHT OUTER JOIN: Gets all non-matching records from right table & AND one copy of matching records from both the tables based on the joining columns.

○ FULL OUTER JOIN: Gets all non-matching records from left table & all non-matching records from right table & one copy of matching records from both the tables.

○ CROSS JOIN: returns the Cartesian product.

13. What are the different types of Restricted Joins?

○ SELF JOIN: joining a table to itself

○ RESTRICTED LEFT OUTER JOIN: gets all non-matching records from

left side

○ RESTRICTED RIGHT OUTER JOIN - gets all non-matching records from

right side

○ RESTRICTED FULL OUTER JOIN - gets all non-matching records from left table & gets all non-matching records from right table.

14. What is a sub-query?

○ It is a query within a query

○ Syntax:

SELECT <column_name> FROM <table_name>

WHERE <column_name> IN/NOT IN

(

<another SELECT statement>

)

○ Everything that we can do using sub queries can be done using Joins, but anything that we can do using Joins may/may not be done using Subquery.

○ Sub-Query consists of an inner query and outer query. Inner query is a SELECT statement the result of which is passed to the outer query. The outer query can be SELECT, UPDATE, DELETE. The result of the inner query is generally used to filter what we select from the outer query.

○ We can also have a subquery inside of another subquery and so on. This is called a nested Subquery. Maximum one can have is 32 levels of nested Sub-Queries.

15. What are the SET Operators?

○ SQL set operators allows you to combine results from two or more SELECT statements.

○ Syntax:

SELECT Col1, Col2, Col3 FROM T1 <SET OPERATOR>

SELECT Col1, Col2, Col3 FROM T2

○ Rule 1: The number of columns in first SELECT statement must be same as the number of columns in the second SELECT statement.

○ Rule 2: The metadata of all the columns in first SELECT statement MUST be exactly same as the metadata of all the columns in second SELECT statement accordingly.

○ Rule 3: ORDER BY clause do not work with first SELECT statement. ○ UNION, UNION ALL, INTERSECT, EXCEPT

16. What is a derived table?

○ SELECT statement that is given an alias name and can now be treated as a virtual table and operations like joins, aggregations, etc. can be performed on it like on an actual table.

○ Scope is query bound, that is a derived table exists only in the query in which it was defined.
SELECT temp1.SalesOrderID, temp1.TotalDue FROM

(SELECT TOP 3 SalesOrderID, TotalDue FROM Sales.SalesOrderHeader ORDER BY TotalDue DESC) AS temp1 LEFT OUTER JOIN

(SELECT TOP 2 SalesOrderID, TotalDue FROM Sales.SalesOrderHeader ORDER BY TotalDue DESC) AS temp2 ON temp1.SalesOrderID = temp2.SalesOrderID WHERE temp2.SalesOrderID IS NULL

17. What is a View?

○ Views are database objects which are virtual tables whose structure is defined by underlying SELECT statement and is mainly used to implement security at rows and columns levels on the base table.

○ One can create a view on top of other views.

○ View just needs a result set (SELECT statement).

○ We use views just like regular tables when it comes to query writing. (joins, subqueries, grouping)

○ We can perform DML operations (INSERT, DELETE, UPDATE) on a view. It actually affects the underlying tables only those columns can be affected which are visible in the view.

18. What are the types of views?

1. Regular View:

It is a type of view in which you are free to make any DDL changes on the underlying table.

-- create a regular view

CREATE VIEW v_regular AS SELECT * FROM T1

2. Schemabinding View:

It is a type of view in which the schema of the view (column) are physically bound to the schema of the underlying table. We are not allowed to perform any DDL changes

to the underlying table for the columns that are referred by the schemabinding view structure.

■ All objects in the SELECT query of the view must be specified in two part naming conventions (schema_name.tablename).

■ You cannot use * operator in the SELECT query inside the view (individually name the columns)

■ All rules that apply for regular view.

CREATE VIEW v_schemabound WITH SCHEMABINDING AS SELECT ID, Name

FROM dbo.T2 -- remember to use two part naming convention

3. Indexed View:

19. What is an Indexed View?

○ It is technically one of the types of View, not Index.

○ Using Indexed Views, you can have more than one clustered index on the same table if needed.

○ All the indexes created on a View and underlying table are shared by Query Optimizer to select the best way to execute the query.

○ Both the Indexed View and Base Table are always in sync at any given point.

○ Indexed Views cannot have NCI-H, always NCI-CI, therefore a duplicate set of the data will be created.

20. What does WITH CHECK do?

○ WITH CHECK is used with a VIEW.

○ It is used to restrict DML operations on the view according to search predicate (WHERE clause) specified creating a view.

○ Users cannot perform any DML operations that do not satisfy the conditions in WHERE clause while creating a

view.

○ WITH CHECK OPTION has to have a WHERE clause.

21. What is a RANKING function and what are the four RANKING functions?

Ranking functions are used to give some ranking numbers to each row in a dataset based on some ranking functionality.

Every ranking function creates a derived column which has integer value.

Different types of RANKING function:

ROW_NUMBER(): assigns an unique number based on the ordering starting with 1. Ties will be given different ranking positions.

RANK(): assigns an unique rank based on value. When the set of ties ends, the next ranking position will consider how many tied values exist and then assign the next value a new ranking with consideration the number of those previous ties. This will make the ranking position skip placement. position numbers based on how many of the same values occurred (ranking not sequential).

DENSE_RANK(): same as rank, however it will maintain its consecutive order nature regardless of ties in values; meaning if five records have a tie in the values, the next ranking will begin with the next

ranking position.

Syntax:

<Ranking Function>() OVER(condition for ordering) -- always have to have an OVER clause

Ex:

```
SELECT SalesOrderID, SalesPersonID,
TotalDue,
ROW_NUMBER() OVER(ORDER BY TotalDue), RANK() OVER(ORDER BY TotalDue),
DENSE_RANK() OVER(ORDER BY TotalDue) FROM Sales.SalesOrderHeader
```

■ NTILE(n): Distributes the rows in an ordered partition into a specified number of groups.

22. What is PARTITION BY?

○ Creates partitions within the same result set and each partition gets its own ranking. That is, the rank starts from 1 for each partition.

○ Ex:

SELECT *, DENSE_RANK() OVER(PARTITION BY Country ORDER BY Sales DESC) AS DenseRank FROM SalesInfo

23. What is Temporary Table and what are the two types of it? ○ They are tables just like regular tables but the main difference is its scope.

○ The scope of temp tables is temporary whereas regular tables permanently reside. ○ Temporary table are stored in tempDB.

○ We can do all kinds of SQL operations with temporary tables just like regular tables like JOINs, GROUPING, ADDING CONSTRAINTS, etc.

○ Two types of Temporary Table

■ Local

#LocalTempTableName -- single pound sign

Only visible in the session in which they are created. It is session-bound.

■ Global

##GlobalTempTableName -- double pound sign

Global temporary tables are visible to all sessions after they are created, and are deleted when the session in which they were created in is disconnected.

It is last logged-on user bound. In other words, a global temporary table will disappear when the last user on the session logs off.

24. Explain Variables ?

○ Variable is a memory space (place holder) that contains a scalar value EXCEPT table variables, which is 2D

data.

○ Variable in SQL Server are created using DECLARE Statement. ○ Variables are BATCH-BOUND.

○ Variables that start with @ are user-defined variables.

25. Explain Dynamic SQL (DSQL). ?

Dynamic SQL refers to code/script which can be used to operate on different data-sets based on some dynamic values supplied by front-end applications. It can be used to run a template SQL query

against different tables/columns/conditions.

Declare variables: which makes SQL code dynamic.

Main disadvantage of D-SQL is that we are opening SQL Tool for SQL Injection attacks. You should build the SQL script by concatenating strings and variable.

26. What is SQL Injection Attack?

○ Moderator's definition: when someone is able to write a code at the front end using DSQL, he/she could use malicious code to drop, delete, or manipulate the database. There is no perfect protection from it but we can check if there is certain commands such as 'DROP' or 'DELETE' are included in the command line.

○ SQL Injection is a technique used to attack websites by inserting SQL code in web entry fields.

27. What is SELF JOIN?

○ JOINing a table to itself

○ When it comes to SELF JOIN, the foreign key of a table points to its primary key. ○ Ex: Employee(Eid, Name, Title, Mid)

○ Know how to implement it!!!

28. What is Correlated Subquery?

○ It is a type of subquery in which the inner query depends on the outer query. This means that that the subquery is executed repeatedly, once for each row of the outer query.

○ In a regular subquery, inner query generates a result set that is independent of the outer query.

○ Ex:

```
SELECT *
FROM HumanResources.Employee E
WHERE 5000 IN (SELECT S.Bonus
FROM Sales.SalesPerson S
WHERE S.SalesPersonID = E.EmployeeID)
```

○ The performance of Correlated Subquery is very slow because its inner query depends on the outer query. So the inner subquery goes through every single row of the result of the outer subquery.

29. What is the difference between Regular Subquery and Correlated Subquery?

○ Based on the above explanation, an inner subquery is independent from its outer subquery in Regular Subquery. On the other hand, an inner subquery depends on its outer subquery in Correlated Subquery.

30. What are the differences between DELETE and TRUNCATE .?

Delete:

DML statement that deletes rows from a table and can also specify rows using a WHERE clause. Logs every row deleted in the log file.

Slower since DELETE records every row that is deleted.

DELETE continues using the earlier max value of the identity column. Can have triggers on DELETE.

Truncate:

DDL statement that wipes out the entire table and you cannot delete specific rows.

Does minimal logging, minimal as not logging everything. TRUNCATE will remove the pointers that point to their pages, which are deallocated.

Faster since TRUNCATE does not record into the log file. TRUNCATE resets the identity column.

Cannot have triggers on TRUNCATE.

31. What are the three different types of Control Flow statements?

1. WHILE

2. IF-ELSE

3. CASE

32. What is Table Variable? Explain its advantages and disadvantages.?

○ If we want to store tabular data in the form of rows and columns into a variable then we use a table variable. ○ It is able to store and display 2D data (rows and columns).

○ We cannot perform DDL (CREATE, ALTER, DROP).

Advantages:

■ Table variables can be faster than permanent tables.

■ Table variables need less locking and logging resources.

Disadvantages:

■ Scope of Table variables is batch bound.

■ Table variables cannot have constraints.

■ Table variables cannot have indexes.

■ Table variables do not generate statistics.

■ Cannot ALTER once declared (Again, no DDL statements).

33. What are the differences between Temporary Table and Table Variable?

Temporary Table:

It can perform both DML and DDL Statement. Session bound Scope

Syntax CREATE TABLE #temp

Have indexes

Table Variable:

Can perform only DML, but not DDL Batch bound scope

DECLARE @var TABLE(...)

Cannot have indexes

34. What is Stored Procedure (SP)?

It is one of the permanent DB objects that is precompiled set of TSQL statements that can accept and return multiple variables.

It is used to implement the complex business process/logic. In other words, it encapsulates your entire business process.

Compiler breaks query into Tokens. And passed on to query optimizer. Where execution plan is generated the very 1st time when we execute a stored procedure after creating/altering it and same execution plan is utilized for subsequent executions.

Database engine runs the machine language query and execute the code in 0's and 1's.

When a SP is created all Tsql statements that are the part of SP are pre-compiled and execution plan is stored in DB which is referred for following executions.

Explicit DDL requires recompilation of SP's.

35. What are the four types of SP?

System Stored Procedures (SP_****): built-in stored procedures that were created by Microsoft.

User Defined Stored Procedures: stored procedures that are created by users. Common naming convention (usp_****)

CLR (Common Language Runtime): stored procedures that are implemented as public static methods on a class in a Microsoft .NET Framework assembly.

Extended Stored Procedures (XP_****): stored procedures that can be used in other platforms such as Java or C++.

36. Explain the Types of SP..? ○ SP with no parameters:

○ SP with a single input parameter:

○ SP with multiple parameters:

○ SP with output parameters:

Extracting data from a stored procedure based on an input parameter and outputting them using output variables.

○ SP with RETURN statement (the return value is always single and integer value)

37. What are the characteristics of SP?

○ SP can have any kind of DML and DDL statements. ○ SP can have error handling (TRY … CATCH).

○ SP can use all types of table.

○ SP can output multiple integer values using OUT parameters, but can return only one scalar INT value. ○ SP can take any input except a table variable.

○ SP can set default inputs.

○ SP can use DSQL.

○ SP can have nested SPs.

○ SP cannot output 2D data (cannot return and output table variables).

○ SP cannot be called from a SELECT statement. It can be executed using only a EXEC/EXECUTE statement.

38. What are the advantages of SP?

○ Precompiled code hence faster.

○ They allow modular programming, which means it allows you to break down a big chunk of code into smaller pieces of codes. This way the code will be more readable and more easier to manage.

○ Reusability.

○ Can enhance security of your application. Users can be granted permission to execute SP without having to have direct permissions on the objects referenced in the procedure.

○ Can reduce network traffic. An operation of hundreds of lines of code can be performed through single statement that executes the code in procedure rather than by sending hundreds of lines of code over the network.

○ SPs are pre-compiled, which means it has to have an Execution Plan so every time it gets executed after creating a new Execution Plan, it will save up to 70% of execution time. Without it, the SPs are just like any regular TSQL statements.

39. What is User Defined Functions (UDF)?

○ UDFs are a database object and a precompiled set of TSQL statements that can accept parameters, perform complex business calculation, and return of the action as a value.

○ The return value can either be single scalar value or result set-2D data. ○ UDFs are also pre-compiled and their execution plan is saved.

○ PASSING INPUT PARAMETER(S) IS/ARE OPTIONAL, BUT MUST HAVE A RETURN STATEMENT.

40. What is the difference between Stored Procedure and UDF?

Stored Procedure:

may or may not return any value. When it does, it must be scalar INT. Can create temporary tables.

Can have robust error handling in SP (TRY/CATCH, transactions). Can include any DDL and DML statements.

UDF:

must return something, which can be either scalar/table valued. Cannot access to temporary tables.

No robust error handling available in UDF like TRY/ CATCH and transactions. Cannot have any DDL and can do DML only with table variables.

41. What are the types of UDF?

1. Scalar

Deterministic UDF: UDF in which particular input results in particular output. In other words, the output depends on the input.

Non-deterministic UDF: UDF in which the output does not directly depend on the input.

2. In-line UDF:

UDFs that do not have any function body(BEGIN...END) and has only a RETURN statement. In-line UDF must return 2D data.

3. Multi-line or Table Valued Functions:

It is an UDF that has its own function body (BEGIN … END) and can have multiple SQL

statements that return a single output. Also must return 2D data in the form of table variable.

42. What is the difference between a nested UDF and recursive UDF?

○ Nested UDF: calling an UDF within an UDF

○ Recursive UDF: calling an UDF within itself

43. What is a Trigger?

○ It is a precompiled set of TSQL statements that are automatically executed on a particular DDL, DML or log-on

event.

○ Triggers do not have any parameters or return statement.

○ Triggers are the only way to access to the INSERTED and DELETED tables (aka. Magic Tables). ○ You can DISABLE/ENABLE Triggers instead of DROPPING them:

DISABLE TRIGGER <name> ON <table/view name>/DATABASE/ALL SERVER

ENABLE TRIGGER <name> ON <table/view name>/DATABASE/ALL SERVER

44.What are the types of Triggers?

1. DML Trigger

DML Triggers are invoked when a DML statement such as INSERT, UPDATE, or DELETE occur which modify data in a specified TABLE or VIEW.

A DML trigger can query other tables and can include complex TSQL statements. They can cascade changes through related tables in the database.

They provide security against malicious or incorrect DML operations and enforce restrictions that are more complex than those defined with constraints.

2. DDL Trigger

Pretty much the same as DML Triggers but DDL Triggers are for DDL operations. DDL Triggers are at the database or server level (or scope).

DDL Trigger only has AFTER. It does not have INSTEAD OF.

3. Logon Trigger

Logon triggers fire in response to a logon event.

This event is raised when a user session is established with an instance of SQL server. Logon TRIGGER has server scope.

45. What are 'inserted' and 'deleted' tables (aka. magic tables)?

○ They are tables that you can communicate with between the external code and trigger body.

○ The structure of inserted and deleted magic tables depends upon the structure of the table in a DML statement. ○ UPDATE is a combination of INSERT and DELETE, so its old record will be in the deleted table and its new record will be stored in the inserted table.

46. What are some String functions to remember? LEN(string): returns the length of string.

UPPER(string) & LOWER(string): returns its upper/lower string

LTRIM(string) & RTRIM(string): remove empty string on either ends of the string LEFT(string): extracts a certain number of characters from left side of the string RIGHT(string): extracts a certain number of characters from right side of the string SUBSTRING(string, starting_position, length): returns the sub string of the string REVERSE(string): returns the reverse string of the string

Concatenation: Just use + sign for it

REPLACE(string, string_replaced, string_replace_with)

47. What are the three different types of Error Handling?

1. TRY CATCH

The first error encountered in a TRY block will direct you to its CATCH block ignoring the rest of the code in the TRY block will generate an error or not.

2. @@error

stores the error code for the last executed SQL statement. If there is no error, then it is equal to 0.

If there is an error, then it has another number (error code).

3. RAISERROR() function

A system defined function that is used to return messages back to applications using the same format which SQL uses for errors or warning message.

48. Explain about Cursors ..?

○ Cursors are a temporary database object which are used to loop through a table on row-by-row basis. There are five types of cursors:

■ 1. Static: shows a static view of the data with only the changes done by session which opened the cursor.

■ 2. Dynamic: shows data in its current state as the cursor moves from record-to-record.

■ 3. Forward Only: move only record-by-record

■ 4. Scrolling: moves anywhere.

■ 5. Read Only: prevents data manipulation to cursor data set.

49. What is the difference between Table scan and seek ?

○ Scan: going through from the first page to the last page of an offset by offset or row by row. ○ Seek: going to the specific node and fetching the information needed.

○ 'Seek' is the fastest way to find and fetch the data. So if you see your Execution Plan and if all of them is a seek, that means it's optimized.

50. Why are the DML operations are slower on Indexes?

○ It is because the sorting of indexes and the order of sorting has to be always maintained.

○ When inserting or deleting a value that is in the middle of the range of the index, everything has to be rearranged again. It cannot just insert a new value at the end of the index.

51. What is a heap (table on a heap)?

○ When there is a table that does not have a clustered index, that means the table is on a heap. ○ Ex: Following table 'Emp' is a table on a heap.

SELECT * FROM Emp WHERE ID BETWEEN 2 AND 4 -- This will do scanning.

52. What is the architecture in terms of a hard disk, extents and pages?

○ A hard disk is divided into Extents.

○ Every extent has eight pages.

○ Every page is 8KBs (8060 bytes).

53. What are the nine different types of Indexes?

○ 1. Clustered

○ 2. Non-clustered

○ 3. Covering

○ 4. Full Text Index

○ 5. Spatial

○ 6. Unique

○ 7. Filtered

○ 8. XML

○ 9. Index View

54. What is a Clustering Key?

○ It is a column on which I create any type of index is called a Clustering Key for that particular index.

55. Explain about a Clustered Index.?

○ Unique Clustered Indexes are automatically created when a PK is created on a table.

○ But that does not mean that a column is a PK only because it has a Clustered Index.

○ Clustered Indexes store data in a contiguous manner. In other words, they cluster the data into a certain spot on a hard disk continuously.

○ The clustered data is ordered physically.

○ You can only have one CI on a table.

56. What happens when Clustered Index is created?

○ First, a B-Tree of a CI will be created in the background.

○ Then it will physically pull the data from the heap memory and physically sort the data based on the clustering

key.

○ Then it will store the data in the leaf nodes.

○ Now the data is stored in your hard disk in a continuous manner.

57. What are the four different types of searching information in a table?

○ 1. Table Scan -> the worst way

○ 2. Table Seek -> only theoretical, not possible ○ 3. Index Scan -> scanning leaf nodes

○ 4. Index Seek -> getting to the node needed, the best way

58. What is Fragmentation .?

○ Fragmentation is a phenomenon in which storage space is used inefficiently.

○ In SQL Server, Fragmentation occurs in case of DML statements on a table that has an index.

○ When any record is deleted from the table which has any index, it creates a memory bubble which causes fragmentation.

○ Fragmentation can also be caused due to page split, which is the way of building B-Tree dynamically according to the new records coming into the table.

○ Taking care of fragmentation levels and maintaining them is the major problem for Indexes.

○ Since Indexes slow down DML operations, we do not have a lot of indexes on OLTP, but it is recommended to have many different indexes in OLAP.

59. What are the two types of fragmentation?

1. Internal Fragmentation

It is the fragmentation in which leaf nodes of a B-Tree is not filled to its fullest capacity and contains memory

bubbles.

2. External Fragmentation

It is fragmentation in which the logical ordering of the pages does not match the physical ordering of the pages on the secondary storage device.

60. What are Statistics?

○ Statistics allow the Query Optimizer to choose the optimal path in getting the data from the underlying table. ○ Statistics are histograms of max 200 sampled values from columns separated by intervals.

○ Every statistic holds the following info:

■ 1. The number of rows and pages occupied by a table's data

■ 2. The time that statistics was last updated

■ 3. The average length of keys in a column

■ 4. Histogram showing the distribution of data in column

61. What are some optimization techniques in SQL?

1. Build indexes. Using indexes on a table, It will dramatically increase the performance of your read operation because it will allow you to perform index scan or index seek depending on your search predicates and select predicates instead of table scan.

Building non-clustered indexes, you could also increase the performance further.

2. You could also use an appropriate filtered index for your non clustered index because it could avoid performing

a key lookup.

3. You could also use a filtered index for your non-clustered index since it allows you to create an index on a particular part of a table that is accessed more frequently than other parts.

4. You could also use an indexed view, which is a way to create one or more clustered indexes on the same table.

In that way, the query optimizer will consider even the clustered keys on the indexed views so there might be a possible faster option to execute your query.

5. Do table partitioning. When a particular table as a billion of records, it would be practical to partition a table so that it can increase the read operation performance. Every partitioned

table will be considered as physical smaller tables internally.

6. Update statistics for TSQL so that the query optimizer will choose the most optimal path in getting the data

from the underlying table. Statistics are histograms of maximum 200 sample values from columns separated by

intervals.

7. Use stored procedures because when you first execute a stored procedure, its execution plan is stored and the

same execution plan will be used for the subsequent executions rather than generating an execution plan every

time.

8. Use the 3 or 4 naming conventions. If you use the 2 naming convention, table name and column name, the SQL engine will take some time to find its schema. By specifying the schema name or

even server name, you will be able to save some time for the SQL server.

9. Avoid using SELECT *. Because you are selecting everything, it will decrease the performance. Try to select columns you need.

10. Avoid using CURSOR because it is an object that goes over a table on a row-by-row basis, which is similar to the table scan. It is not really an effective way.

11. Avoid using unnecessary TRIGGER. If you have unnecessary triggers, they will be triggered needlessly. Not only slowing the performance down, it might mess up your whole program as well.

12. Manage Indexes using RECOMPILE or REBUILD.

The internal fragmentation happens when there are a lot of data bubbles on the leaf nodes of the b-tree and the leaf nodes are not used to its fullest capacity. By recompiling, you can push the actual data on the b-tree to the left side of the leaf level and push the memory bubble to the right side. But it is still a temporary solution because the memory bubbles will still exist and won't be still accessed much.

The external fragmentation occurs when the logical ordering of the b-tree pages does not match the physical ordering on the hard disk. By rebuilding, you can cluster them all together, which will solve not only the internal but also the external fragmentation issues. You can check the status of the fragmentation by using Data Management Function, sys.dm_db_index_physical_stats(db_id, table_id, index_id, partition_num, flag), and looking at the columns, avg_page_space_used_in_percent for the internal fragmentation and avg_fragmentation_in_percent for the external fragmentation.

13. Try to use JOIN instead of SET operators or SUB-QUERIES because set operators and sub-queries are slower than joins and you can implement the features of sets and sub-queries using joins.

14. Avoid using LIKE operators, which is a string matching operator but it is mighty slow.

15. Avoid using blocking operations such as order by or derived columns.

16. For the last resort, use the SQL Server Profiler. It generates a trace file, which is a really detailed version of execution plan. Then DTA (Database Engine Tuning Advisor) will take a trace file as its input and analyzes it and gives you the recommendation on how to improve your query further.

62. How do you present the following tree in a form of a table?

A

/ \

B C

/ \ / \

D E F G

CREATE TABLE tree (node CHAR(1), parent Node CHAR(1), [level] INT) INSERT INTO tree VALUES ('A', null, 1),

('B', 'A', 2),

('C', 'A', 2),

('D', 'B', 3),

('E', 'B', 3),

('F', 'C', 3),

('G', 'C', 3)

SELECT * FROM tree

Result:

A NULL 1

B A 2

C A 2

D B 3

E B 3

F C 3

G C 3

63. How do you reverse a string without using REVERSE ('string') ?

CREATE PROC rev (@string VARCHAR(50)) AS

BEGIN

DECLARE @new_string VARCHAR(50) = ''

DECLARE @len INT = LEN(@string)

WHILE (@len <> 0)

BEGIN

DECLARE @char CHAR(1) = SUBSTRING(@string, @len, 1) SET @new_string = @new_string + @char

SET @len = @len - 1

END

```
PRINT @new_string

END

EXEC rev 'dinesh'
```

64. What is Deadlock?

○ Deadlock is a situation where, say there are two transactions, the two transactions are waiting for each other to release their locks.

○ The SQL automatically picks which transaction should be killed, which becomes a deadlock victim, and roll back the change for it and throws an error message for it.

65. What is a Fact Table?

The primary table in a dimensional model where the numerical performance measurements (or facts) of the

business are stored so they can be summarized to provide information about the history of the operation of an

organization.

We use the term fact to represent a business measure. The level of granularity defines the grain of the fact table.

66. What is a Dimension Table?

Dimension tables are highly denormalized tables that contain the textual descriptions of the business and facts in their fact table.

Since it is not uncommon for a dimension table to have 50 to 100 attributes and dimension tables tend to be relatively shallow in terms of the number of rows, they are also called a wide table.

A dimension table has to have a surrogate key as its primary key and has to have a business/alternate key to link between the OLTP and OLAP.

67. What are the types of Measures?

○ Additive: measures that can be added across all dimensions (cost, sales).

○ Semi-Additive: measures that can be added across few dimensions and not with others.

○ Non-Additive: measures that cannot be added across all dimensions (stock rates).

68. What is a Star Schema?

○ It is a data warehouse design where all the dimensions tables in the warehouse are directly connected to the

fact table.

○ The number of foreign keys in the fact table is equal to the number of dimensions. ○ It is a simple design and hence faster query.

69. What is a Snowflake Schema?

○ It is a data warehouse design where at least one or more multiple dimensions are further normalized. ○ Number of dimensions > number of fact table foreign keys

○ Normalization reduces redundancy so storage wise it is better but querying can be affected due to the excessive joins that need to be performed.

70. What is granularity?

○ The lowest level of information that is stored in the fact table. ○ Usually determined by the time dimension table.

○ The best granularity level would be per transaction but it would require a lot of memory.

71. What is a Surrogate Key?

○ It is a system generated key that is an identity column with the initial value and incremental value and ensures the uniqueness of the data in the dimension table.

○ Every dimension table must have a surrogate key to identify each record!!!

72. What are some advantages of using the Surrogate Key in a Data Warehouse?

○ 1. Using a SK, you can separate the Data Warehouse and the OLTP: to integrate data coming from heterogeneous sources, we need to differentiate between similar business keys from the OLTP. The

keys in OLTP are the alternate key (business key).

○ 2. Performance: The fact table will have a composite key. If surrogate keys are used, then in the fact table, we will have integers for its foreign keys.

■ This requires less storage than VARCHAR.

■ The queries will run faster when you join on integers rather than VARCHAR.

■ The partitioning done on SK will be faster as these are in sequence.

○ 3. Historical Preservation: A data warehouse acts as a repository of historical data so there will be various versions of the same record and in order to differentiate between them, we need a SK then we can keep the history of data.

○ 4. Special Situations (Late Arriving Dimension): Fact table has a record that doesn't have a match yet in the dimension table. Surrogate key usage enables the use of such a 'not found' record as a SK is not dependent on the

ETL process.

73. What is the datatype difference between a fact and dimension tables?

○ 1. Fact Tables

They hold numeric data.

They contain measures.

They are deep.

○ 2. Dimensional Tables

They hold textual data.

They contain attributes of their fact tables.

They are wide.

74. What are the types of dimension tables?

○ 1. Conformed Dimensions

■ when a particular dimension is connected to one or more fact tables. ex) time dimension ○ 2. Parent-child Dimensions

■ A parent-child dimension is distinguished by the fact that it contains a hierarchy based on a recursive

relationship.

■ when a particular dimension points to its own surrogate key to show an unary relationship. ○ 3. Role Playing Dimensions

■ when a particular dimension plays different roles in the same fact table. ex) dim_time and orderDateKey, shippedDateKey…usually a time dimension table.

■ Role-playing dimensions conserve storage space, save processing time, and improve database manageability .

○ 4. Slowly Changing Dimensions: A dimension table that have data that changes slowly that occur by inserting and updating of records.

■ 1. Type 0: columns where changes are not allowed - no change ex) DOB, SSNm

■ 2. Type 1: columns where its values can be replaced without adding its new row - replacement

■ 3. Type 2: for any change for the value in a column, a new record it will be added - historical data. Previous

values are saved in records marked as outdated. For even a single type 2 column, startDate, EndDate, and status are needed.

■ 4. Type 3: advanced version of type 2 where you can set up the upper limit of history which drops the oldest record when the limit has been reached with the help of outside SQL implementation.

■ Type 0 ~ 2 are implemented on the column level.

○ 5. Degenerated Dimensions: a particular dimension that has an one-to-one relationship between itself and the

fact table.

■ When a particular Dimension table grows at the same rate as a fact table, the actual dimension can be removed and the dimensions from the dimension table can be inserted into the actual fact table.

■ You can see this mostly when the granularity level of the the facts are per transaction.

■ E.g. The dimension salesorderdate (or other dimensions in DimSalesOrder would grow everytime a sale is made therefore the dimension (attributes) would be moved into the fact table.

○ 6. Junk Dimensions: holds all miscellaneous attributes that may or may not necessarily belong to any other dimensions. It could be yes/no, flags, or long open-ended text data.

75. What is your strategy for the incremental load?

The combination of different techniques for the incremental load in my previous projects; time stamps, CDC (Change Data Capture), MERGE statement and CHECKSUM() in TSQL, LEFT OUTER JOIN, TRIGGER, the Lookup Transformation in SSIS.

76. What is CDC?

CDC (Change Data Capture) is a method to capture data changes, such as INSERT, UPDATE and DELETE,

happening in a source table by reading transaction log files. Using CDC in the process of an incremental load, you

are going to be able to store the changes in a SQL table, enabling us to apply the changes to a target table incrementally.

In data warehousing, CDC is used for propagating changes in the source system into your data warehouse,

updating dimensions in a data mart, propagating standing data changes into your data warehouse and such.

The advantages of CDC are:

- It is almost real time ETL.

- It can handle small volume of data.

- It can be more efficient than replication.

- It can be auditable.

- It can be used to configurable clean up.

Disadvantages of CDC are:

- Lots of change tables and functions

- Bad for big changes e.g. truncate & reload Optimization of CDC:

- Stop the capture job during load

- When applying changes to target, it is ideal to use merge.

77. What is the difference between a connection and session ?

○ Connection: It is the number of instance connected to the database. An instance is modelized soon as the application is

open again.

○ Session: A session run queries.In one connection, it allowed multiple sessions for one connection.

78. What are all different types of collation sensitivity?

Following are different types of collation sensitivity -

Case Sensitivity - A and a and B and b.

Accent Sensitivity.

Kana Sensitivity - Japanese Kana characters.

Width Sensitivity - Single byte character and double byte character.

79. What is CLAUSE?

SQL clause is defined to limit the result set by providing condition to the query. This usually filters some rows from the whole set of records.

Example - Query that has WHERE condition Query that has HAVING condition.

80. What is Union, minus and Interact commands?

UNION operator is used to combine the results of two tables, and it eliminates duplicate rows from the tables.

MINUS operator is used to return rows from the first query but not from the second query. Matching records of first and second query and other rows from the first query will be displayed as a result set.

INTERSECT operator is used to return rows returned by both the queries.

81.How to fetch common records from two tables?

Common records result set can be achieved by -.

Select studentID from student. INTERSECT Select StudentID from Exam

82.How to fetch alternate records from a table?

Records can be fetched for both Odd and Even row numbers -.

To display even numbers-.

Select studentId from (Select rowno, studentId from student) where mod(rowno,2)=0 To display odd numbers-.

Select studentId from (Select rowno, studentId from student) where mod(rowno,2)=1 from (Select rowno, studentId from student) where mod(rowno,2)=1.[/sql]

83. How to select unique records from a table?

Select unique records from a table by using DISTINCT keyword.

Select DISTINCT StudentID, StudentName from Student.

84.How to remove duplicate rows from table?

Step 1: Selecting Duplicate rows from table

Select rollno FROM Student WHERE ROWID <>

(Select max (rowid) from Student b where rollno=b.rollno);

Step 2: Delete duplicate rows

Delete FROM Student WHERE ROWID <>

(Select max (rowid) from Student b where rollno=b.rollno);

85.What is ROWID and ROWNUM in SQL?

RowID

1.ROWID is nothing but Physical memory allocation

2.ROWID is permanant to that row which identifies the address of that row.

3.ROWID is 16 digit Hexadecimal number which is uniquely identifies the rows.

4.ROWID returns PHYSICAL ADDRESS of that row.

5. ROWID is automatically generated unique id of a row and it is generated at the time of insertion of row.

6. ROWID is the fastest means of accessing data.

ROWNUM:

1. ROWNUM is nothing but the sequence which is allocated to that data retreival bunch.

2. ROWNUM is tempararily allocated sequence to the rows.

3.ROWNUM is numeric sequence number allocated to that row temporarily.

4.ROWNUM returns the sequence number to that row.

5. ROWNUM is an dynamic value automatically retrieved along with select statement output.

6.ROWNUM is not related to access of data.

86. How to find count of duplicate rows?

Select rollno, count (rollno) from Student

Group by rollno Having count (rollno)>1 Order by count (rollno) desc;

87.How to find Third highest salary in Employee table using self-join?

Select * from Employee a Where 3 = (Select Count (distinct Salary) from Employee where a.salary<=b.salary;

88. How to display following using query?

*

**

***

We cannot use dual table to display output given above. To display output use any table. I am using Student

table.

SELECT lpad (‘*’, ROWNUM,’*’) FROM Student WHERE ROWNUM <4;

89. How to display Date in DD-MON-YYYY table?

Select to_date (Hire_date,’DD-MON-YYYY’) Date_Format from Employee;

90. If marks column contain the comma separated values from Student table. How to calculate the count of that comma separated values?

Student Name Marks

Dinesh 30,130,20,4

Kumar 100,20,30

Sonali 140,10

Select Student_name, regexp_count (marks,’,’) + As “Marks Count” from Student;

91. What is query to fetch last day of previous month in oracle?

Select LAST_DAY (ADD_MONTHS (SYSDATE,-1)) from dual;

92. How to display the String vertically in Oracle?

SELECT SUBSTR (‘AMIET’, LEVEL, 1) FROM dual Connect by level <= length (‘AMIET’);

93. How to display departmentwise and monthwise maximum salary?

Select Department_no, TO_CHAR (Hire_date,’Mon’) as Month from Employee group by Department_no, TO_CHAR (Hire_date,’mon’);

94. How to calculate number of rows in table without using count function?

Select table_name, num_rows from user_tables where table_name=’Employee’;

Tip: User needs to use the system tables for the same. So using user_tables user will get the number of rows in the table

95. How to fetch common records from two different tables which has not any joining condition ?

Select * from Table1

Intersect

Select * from Table2;

96. Explain Execution Plan.?

Query optimizer is a part of SQL server that models the way in which the relational DB engine works and comes up with the most optimal way to execute a query. Query Optimizer takes into account amount of resources used, I/O and CPU processing time etc. to generate a plan that will allow query to execute in most efficient and faster manner. This is known as EXECUTION PLAN.

Optimizer evaluates a number of plans available before choosing the best and faster on available. Every query has an execution plan.

Definition by the mod: Execution Plan is a plan to execute a query with the most optimal way which is generated by Query Optimizer. Query Optimizer analyzes statistics, resources used, I/O and CPU

processing time and etc. and comes up with a number of plans. Then it evaluates those plans and the most optimized plan out of the plans is Execution Plan. It is shown to users as a graphical flow chart that should be read from right to left and top to bottom.

Part -2 Scala Programming

1) What is Scala?

Scala is a general-purpose programming language providing support for both functional and Object-Oriented programming.

2. What is tail-recursion in Scala?

There are several situations where programmers have to write functions that are recursive in nature. The main problem with recursive functions is that, it may eat up all the allocated stack space. To overcome this situation, Scala compiler provides a mechanism “tail recursion” to optimize these

recursive functions so that it does not create new stack space, instead uses the current function stack space. To qualify for this, annotation "@annotation.tailrec" has to be used before defining the function and recursive call has to be the last statement, then only the function will compile otherwise, it will give an error.

3. What are 'traits' in Scala?

'Traits' are used to define object types specified by the signature of the supported methods. Scala allows to be partially implemented but traits may not have constructor parameters. A trait consists of method and field definition, by mixing them into classes it can be reused.

4. Who is the father of Scala programming language?

Martin Oderskey, a German computer scientist, is the father of Scala programming language.

5. What are case classes in Scala?

Case classes are standard classes declared with a special modifier case. Case classes export their constructor parameters and provide a recursive decomposition mechanism through pattern matching. The constructor parameters of case classes are treated as public values and can be accessed directly. For a case class, companion objects and its associated method also

get generated automatically. All the methods in the class, as well, methods in the companion objects are generated based on the parameter list. The only advantage of Case class is that it automatically generates the methods from the parameter list.

6. What is the super class of all classes in Scala?

In Java, the super class of all classes (Java API Classes or User Defined Classes) is java.lang.Object. In the same way in Scala, the super class of all classes or traits is "Any" class.

Any class is defined in scala package like "scala.Any".

7. What is a 'Scala Set'? What are methods through which operation sets are expressed?

Scala set is a collection of pairwise elements of the same type. Scala set does not contain any duplicate elements. There are two kinds of sets, mutable and immutable.

8. What is a Scala Map?

Scala Map is a collection of key value pairs wherein the value in a map can be retrieved using the key. Values in a Scala Map are not unique, but the keys are unique. Scala supports two kinds of maps- mutable and immutable. By default, Scala supports immutable map and to make use of the mutable map, programmers must import the scala.collection.mutable.Map class explicitly. When programmers want to use mutable and immutable map together in the same program then the mutable map can be accessed as mutable.map and the immutable map can just be accessed with the name of the map.

9. Name two significant differences between a trait and an abstract class.

Abstract classes have constructors with zero or more parameters while traits do not; a class can extend any number of traits but only one abstract class.

10. What is the use of tuples in Scala?

Scala tuples combine a fixed number of items together so that they can be passed around as whole. A tuple is immutable and can hold objects with different types, unlike an array or list.

11. What do you understand by a closure in Scala?

A closure is also known as an anonymous function whose return value depends upon the value of the variables declared outside the function.

12. What do you understand by Implicit Parameter?

Wherever, we require that function could be invoked without passing all the parameters, we use implicit parameter. We provide the default values for all the parameters or parameters which we want to be used as implicit. When the function is invoked without passing the implicit parameters, local value of that parameter is used. We need to use implicit keyword to make a value, function parameter or variable as implicit.

13. What is the companion object in Scala?

A companion object is an object with the same name as a class or trait and is defined in the same source file as the associated file or trait. A companion object differs from other objects as it has access rights to the class/trait that other objects do not. In particular it can access methods and fields that are private in the class/trait.

14. What are the advantages of Scala Language?

Advantages of Scala Language:-

- Simple and Concise Code
- Very Expressive Code
- More Readable Code
- 100% Type-Safe Language
- Immutability and No Side-Effects
- More Reusable Code
- More Modularity
- Do More with Less Code
- Supports all OOP Features
- Supports all FP Features. Highly Functional.
- Less Error Prone Code
- Better Parallel and Concurrency Programming
- Highly Scalable and Maintainable code
- Highly Productivity
- Distributed Applications
- Full Java Interoperability
- Powerful Scala DSLs available

15. What are the major drawbacks of Scala Language?

Drawbacks of Scala Language:-

- Less Readable Code
- Bit tough to Understand the Code for beginners
- Complex Syntax to learn
- Less Backward Compatibility

16. What is Akka, Play, and Sleek in Scala?

Akka is a concurrency framework in Scala which uses Actor based model for building highly concurrent, distributed, and resilient message-driven applications on the JVM. It uses high-level abstractions like Actor, Future, and Stream to simplify coding for concurrent applications. It also provides load balancing, routing, partitioning, and adaptive cluster management. If you are interested in learning Akka,

17. What is 'Unit' and '()' in Scala?

The 'Unit' is a type like void in Java. You can say it is a Scala equivalent of the void in Java, while still providing the language with an abstraction over the Java platform. The empty tuple '()' is a term representing a Unit value in Scala.

18. What is the difference between a normal class and a case class in Scala?

Following are some key differences between a case class and a normal class in Scala:

- case class allows pattern matching on it.
- you can create instances of case class without using the new keyword
- equals(), hashcode() and toString() method are automatically generated for case classes in Scala
- Scala automatically generate accessor methods for all constructor argument

19. What are High Order Functions in Scala?

High order functions are functions that can receive or return other functions. Common examples in Scala are the filter, map, and flatMap functions, which receive other functions as arguments.

20. Which Scala library is used for functional programming?

Scalaz library has purely functional data structures that complement the standard Scala library. It has pre-defined set of foundational type classes like Monad, Functor, etc.

21. What is the best scala style checker tool available for play and scala based applications?

Scalastyle is best Scala style checker tool available for Play and Scala based applications. Scalastyle observes the Scala source code and indicates potential problems with it. It has three separate plug-ins to supports the following build tools:

SBT

Maven

Gradle

22. What is the difference between concurrency and parallelism?

When several computations execute sequentially during overlapping time periods it is referred to as concurrency whereas when processes are executed simultaneously it is known as parallelism. Parallel collection, Futures and Async library are examples of achieving parallelism in Scala.

23. What is the difference between a Java method and a Scala function?

Scala function can be treated as a value. It can be assigned to a val or var, or even returned from another function, which is not possible in Java. Though Java 8 brings lambda expression which also makes function as a first-class object, which means you can pass a function to a method just like you pass an object as an argument. See here to learn more about the difference between Scala and Java.

24. What is the difference between Function and Method in Scala?

Scala supports both functions and methods. We use same syntax to define functions and methods, there is no syntax difference.

However, they have one minor difference:

We can define a method in a Scala class or trait. Method is associated with an object (An instance of a Class). We can call a method by using an instance of a Class. We cannot use a Scala Method directly without using object.

Function is not associated with a class or trait. It is defined in a Scala Package. We can access functions without using objects, like Java's Static Methods.

25. What is Extractor in Scala?

In Scala, Extractor is used to decompose or disassemble an object into its parameters (or components).

26. Is Scala a Pure OOP Language?

Yes, Scala is a Pure Object-Oriented Programming Language because in Scala, everything is an Object, and everything is a value. Functions are values and values are Objects.

Scala does not have primitive data types and does not have static members.

27. Is Java a pure OOP Language?

Java is not a Pure Object-Oriented Programming (OOP) Language because it supports the following two Non-OOP concepts: Java supports primitive data types. They are not objects.Java supports Static members. They are not related to objects.

28. Does Scala support Operator Overloading? Scala supports Operator Overloading.

Scala has given this flexibility to Developer to decide which methods/functions name should use. When we call 4 + 5 that means '+' is not an operator, it is a method available in Int class (or it's implicit type). Internally, this call is converted into "4.+(5)".

29. Does Java support Operator Overloading?

Java does not support Operator Overloading.

30. What are the default imports in Scala Language?

We know, java.lang is the default package imported into all Java Programs by JVM automatically. We don't need to import this package explicitly.

In the same way, the following are the default imports available in all Scala Programs:

java.lang package

Scala package

scala.PreDef

31. What is an Expression?

Expression is a value that means it will evaluate to a Value. As an Expression returns a value, we can assign it to a variable.

Example:- Scala's If condition, Java's Ternary operator.

32.What is a Statement? Difference between Expression and Statement?

Statement defines one or more actions or operations. That means Statement

performs actions. As it does not return a value, we cannot assign it to a

Variable.

Example:- Java's If condition.

33.What is the difference between Java's "If...Else" and Scala's "If..Else"?

Java's "If..Else":

In Java, "If..Else" is a statement, not an expression. It does not return a value and cannot assign it to a variable.

Example:-

```
int year;

if( count == 0)

year = 2018;
```

else

year = 2017;

Scala's "If..Else":

In Scala, "If..Else" is an expression. It evaluates a value i.e. returns a value. We can assign it to a variable.

val year = if(count == 0) 2018 else 2017

NOTE:-Scala's "If..Else" works like Java's Ternary Operator. We can use Scala's "If..Else" like Java's "If..Else" statement as shown below:

val year = 0

if(count == 0)

year = 2018

else

year = 2017

34. How to compile and run a Scala program?

You can use Scala compiler scalac to compile Scala program (like javac) and scala command to run them (like scala)

35.How to tell Scala to look into a class file for some Java class?

We can use -classpath argument to include a JAR in Scala's classpath, as shown below

$ scala -classpath jar

Alternatively, you can also use CLASSPATH environment variable.

36. What is the difference between a call-by-value and call-by-name parameter?

The main difference between a call-by-value and a call-by-name parameter is that the former is computed before calling the function, and the latter is evaluated when accessed.

37. What exactly is wrong with a recursive function that is not tail-recursive?

Answer: You run the risk of running out of stack space and thus throwing an exception.

38. What is the difference between var and value?

In scala, you can define a variable using either a, val or var keywords. The difference between val and var is, var is much like java declaration, but valis little different. We cannot change the reference to point to another reference, once the variable is declared using val. The variable defined using var keywords are mutable and can be changed any number of times.

39. What is scala anonymous function?

In a source code, anonymous functions are called 'function literals' and at run time, function literals are instantiated into objects called function values. Scala provides a relatively easy syntax for defining anonymous

functions.

40. What is function currying in scala?

Currying is the technique of transforming a function that takes multiple arguments into a function that takes a single argument Many of the same techniques as language like Haskell and LISP are supported by Scala. Function currying is one of the least used and misunderstood one.

41. What do you understand by "Unit" and "()" in Scala?

Unit is a subtype of scala.anyval and is nothing but Scala equivalent of Java void that provides the Scala with an abstraction of the java platform. Empty tuple i.e. () in Scala is a term that represents unit value.

42. What's the difference 'Nil', 'Null', 'None' and 'Nothing' in Scala?

Null - It's a sub-type of AnyRef type in Scala Types hierarchy. As Scala runs on JVM, it uses NULL to provide the compatibility with Java null keyword, or in Scala terms, to provide type for null keyword, Null type exists. It represents the absence of type information for complex types that are inherited from AnyRef.

Nothing - It's a sub-type of all the types exists in Scala Types hierarchy. It helps in providing the return type for the operations that can affect a normal program's flow. It can only be used as a type, as

instantiation of nothing cannot be done. It incorporates all types under AnyRef and AnyVal. Nothing is usually used as a return type for methods that have abnormal termination and result in an exception.

Nil - It's a handy way of initializing an empty list since, Nil, is an object, which extends List [Nothing].

None - In programming, there are many circumstances, where we

unexpectedly received null for the methods we call. In java these are

handled using try/catch or left unattended causing errors in the program.

Scala provides a very graceful way of handling those situations. In cases,

where you don't know, if you would be able to return a value as

expected, we can use Option [T]. It is an abstract class, with just two

sub-classes, Some [T] and none. With this, we can tell users that, the method might return a T of type Some [T] or it might return none.

43. What is Lazy Evaluation?

Lazy Evaluation means evaluating program at run-time on-demand that means when clients access the program then only its evaluated.

The difference between "val" and "lazy val" is that "val" is used to define variables which are evaluated eagerly and "lazy val" is also used to define variables but they are evaluated lazily.

44. What is call-by-name?

Call-by-name means evaluates method/function parameters only when we need them, or we access them. If we don't use them, then it does not evaluate them.

45. Does Scala and Java support call-by-name?

Scala supports both call-by-value and call-by-name function parameters. However, Java supports only call-by-value, but not call-by-name.

46. What is the difference between call-by-value and call-by-name function parameters?

Difference between call-by-value and call-by-name:

The major difference between these two are described below:

In Call-by-name, the function parameters are evaluated only whenever they are needed but not when the function is called.

In Call-by-value, the function parameters are evaluated when the function is called. In Call-by-value, the parameters are evaluated before executing function and they are evaluated only once irrespective of how many times we used them in that function.

In Call-by-name, the parameters are evaluated whenever we access them, and they are evaluated each time we use them in that function.

47. What do you understand by apply and unapply methods in Scala?

Apply and unapply methods in Scala are used for mapping and unmapping data between form and model data.

Apply method - Used to assemble an object from its components. For example, if we want to create an Employee object then use the two components firstName and lastName and compose the Employee object using the apply method.

Unapply method - Used to decompose an object from its components. It follows the reverse process of apply method. So, if you have an employee object, it can be decomposed into two components- firstName and lastName.

48. What is an anonymous function in Scala?

Anonymous Function is also a Function, but it does not have any function name. It is also known as a Function Literal.

49.What are the advantages of Anonymous Function/Function Literal in Scala?

The advantages of Anonymous Function/Function Literal in Scala:

We can assign a Function Literal to variable

We can pass a Function Literal to another function/method

We can return a Function Literal as another function/method result/return

value.

50. What is the difference between unapply and apply, when would you use them?

Unapply is a method that needs to be implemented by an object in order for it to be an extractor. Extractors are used in pattern matching to access an object constructor parameter. It's the opposite of a constructor.

The apply method is a special method that allows you to write someObject(params) instead of someObject.apply(params). This usage iscommon in case classes, which contain a companion object with the apply method that allows the nice syntax to instantiate a new object without the new keyword.

51. What is the difference between a trait and an abstract class in Scala?

Here are some key differences between a trait and an abstract class in Scala:

A class can inherit from multiple traits but only one abstract class.

Abstract classes can have constructor parameters as well as type parameters. Traits can have only type parameters. For example, you can't say trait t(i: Int) {}; the iparameter is illegal.

Abstract classes are fully interoperable with Java. You can call them from Java code without any wrappers. On the other hand, Traits are fully interoperable only if they do not contain any implementation code. See here to learn more about Abstract class in Java and OOP.

52. Can a companion object in Scala access the private members of its companion class in Scala?

According to the private access specifier, private members can be accessed only within that class, but Scala's companion object and class provide special access to private members. A companion object can access all the private members of a companion class. Similarly, a companion class can access all the private members of companion objects.

53. What are scala variables?

Values and variables are two shapes that come in Scala. A value variable is constant and cannot be changed once assigned. It is immutable, while a regular variable, on the other hand, is mutable, and you can change the value.

The two types of variables are var myVar : Int=0;

```
val   myVal: Int=1;
```

54. Mention the difference between an object and a class ?

A class is a definition for a description. It defines a type in terms of methods and composition of other types. A class is a blueprint of theobject. While, an object is a singleton, an instance of a class which is unique. An anonymous class is created for every object in the code, it inherits from whatever classes you declared object to implement.

55. What is the difference between val and var in Scala?

The val keyword stands for value and var stands for variable. You can use keyword val to store values, these are immutable, and cannot change once assigned. On the other hand, keyword var is used to create variables, which are values that can change after being set. If you try to modify a val, the compiler will throw an error. It is like the final variable in Java or const in

C++.

56. What is the difference between Array and List in Scala?

Arrays are always Mutable whereas List is always Immutable.Once created, we can change Array values where as we cannot change List Object.Arrays are fixed-size data structures whereas List is variable-sized data structures. List's size is automatically increased or decreased based on its operations we perform on it.Arrays are Invariants whereas Lists are Covariant.

57. What is "Type Inference" in Scala?

Types can be inferred by the Scala Compiler at compile-time. It is known as "Type Inference". Types means Data type or Result type. We use Types at many places in Scala programs like Variable types, Object types, Method/Function Parameter types, Method/Function return types etc.

In simple words, determining the type of a variable or expression or object etc. at compile-time by compiler is known as "Type Inference".

58. What is Eager Evaluation?

Eager Evaluation means evaluating program at compile-time or program deployment-time irrespective of clients are using that program or not.

59. What is guard in Scala's 'for-Comprehension' construct?

In Scala, for-comprehension construct has an if clause which is used to write a condition to filter some elements and generate new collection. Thisif clause is also known as "Guard".If that guard is true, then add that element to new collection. Otherwise, it does not add that element to original collection 60. Why scala prefers immutability?Scala prefers immutability in design and in many cases uses it as default. Immutability can help when dealing with equality issues or concurrent programs.

61. What are the considerations you need to have when using Scala streams?

Streams in Scala are a type of lazy collection, which are created using starting element and then recursively generated using those elements. Streams are like a List, except that, elements are added only when they are accessed, hence "lazy". Since streams are lazy in terms of adding elements, they can be unbounded also, and once the elements are added, they are cached. Since Streams can be unbounded, and all the values are computed at the time of access, programmers need to be careful on using methods which are not transformers, as it may result in

java.lang.OutOfMemoryErrors. stream.max

stream.size stream.sum

62. Differentiate between Array and List in Scala.

List is an immutable recursive data structure whilst array is a sequential mutable data structure.

Lists are covariant whilst array are invariants.The size of a list automatically increases or decreases based on the operations that are performed on it i.e. a list in Scala is a variable-sized data structure whilst an array is fixed size data structure.

63. Which keyword is used to define a function in Scala?

A function is defined in Scala using the def keyword. This may sound

familiar to Python developers as Python also uses def to define a function.

64. What is Monad in Scala?

A monad is an object that wraps another object in Scala. It helps to perform the data manipulation of the underlying object, instead of manipulating the object directly.

65. Is Scala statically-typed language?

Yes, Scala is a statically-typed language.

66. What is Statically-Typed Language and What is Dynamically-Typed Language?

Statically-Typed Language means that Type checking is done at compile-

time by compiler, not at run-time. Dynamically-Typed Language means that Type checking is done at run-time, not at compile-time by compiler.

67. What is the difference between unapply and apply, when would you use them?

unapply is a method that needs to be implemented by an object in order for it to be an extractor. Extractors are used in pattern matching to access an object constructor parameter. It's the opposite of a constructor.

The apply method is a special method that allows you to write someObject(params) instead of someObject.apply(params). This usage is common in case classes, which contain a companion object with the apply method that allows the nice syntax to instantiate a new object without the new keyword.

68. What is Unit in Scala?

In Scala, Unit is used to represent "No value" or "No Useful value". Unit is a final class defined in "scala" package that is "scala.Unit".

69. What is the difference between Java's void and Scala's Unit? Unit is something like Java's void. But they have few differences. Java's void does not any value. It is nothing.

Scala's Unit has one value ()

() is the one and only value of type Unit in Scala. However, there are no values of type void in Java.

Java's void is a keyword. Scala's Unit is a final class.Both are used to represent a method or function is not returning anything.

70. What is "App" in Scala?

In Scala, App is a trait defined in scala package like "scala.App". It defines main method. If an Object or a Class extends this trait, then they will become as Scala Executable programs automatically because they will inherit main method from Application.

71. What is the use of Scala's App?

The main advantage of using App is that we don't need to write main method. The main drawback of using App is that we should use same name "args" to refer command line argument because scala.App's main() method uses this name.

71. What are option, some and none in scala?

'Option' is a Scala generic type that can either be 'some' generic value or none. 'Queue' often uses it to represent primitives that may be null.

73. What is Scala Future?

Scala Future is a monadic collection, which starts a background task. It is an object which holds the potential value or future value, which would be available after the task is completed. It also provides various operations to further chain the operations or to extract the value. Future also provide various call-back functions like onComplete, OnFailure, onSuccess to name a few, which makes Future a complete concurrent task class.

74. How it differs from java's Future class?

The main and foremost difference between Scala's Future and Java's Future class is that the later does not provide promises/callbacks operations. The only way to retrieve the result is Future.get () in Java.

75. What do you understand by diamond problem and how does Scala resolve this?

Multiple inheritance problem is referred to as the Deadly diamond problem or diamond problem. The inability to decide on which implementation of the method to choose is referred to as the Diamond Problem in Scala. Suppose say classes B and C both inherit from class A, while class D

inherits from both class B and C. Now while implementing multiple inheritance if B and C override some method from class A, there is a confusion and dilemma always on which implementation D should inherit. This is what is referred to as diamond problem. Scala resolves diamond problem through the concept of Traits and class linearization rules.

76. What is the difference between == in Java and Scala?

Scala has more intuitive notion of equality. The == operator will automatically run the instance's equals method, rather than doing Java style comparison to check that two objects are the same reference. By the way, you can still check for referential equality by using eq method. In short, Java == operator compare references while Scala calls the equals() method. You can also read the difference between == and equals() in Java to learn more about how they behave in Java.

77. What is REPL in Scala? What is the use of Scala's REPL?

REPL stands for Read-Evaluate-Print Loop. We can pronounce it as 'ripple'. In Scala, REPL is acts as an Interpreter to execute Scala code from command prompt. That's why REPL is also known as Scala CLI(Command Line Interface) or Scala command-line shell.

The main purpose of REPL is that to develop and test small snippets of Scala code for practice purpose. It is very useful for Scala Beginners to practice basic programs.

78. What are the similarities between Scala's Int and Java's java.lang.Integer?

Similarities between Scala's Int and Java's java.lang.Integer are Both are classes.Both are used to represent integer numbers. Both are 32-bit signed integers.

79.What are the differences between Scala's Int and Java's java.lang.Integer?

Differences between Scala's Int and Java's java.lang.Integer are Scala's Int class does not implement Comparable interface. Java's java.lang.Integer class implements Comparable interface.

80. What is the relationship between Int and RichInt in Scala?

Java's Integer is something like Scala's Int and RichInt. RichInt is a final class defined in scala.runtime package like "scala.runtime.RichInt".

In Scala, the Relationship between Int and RichInt is that when we use Int in a Scala program, it will automatically convert into RichInt to utilize all methods available in that Class. We can say that RichInt is an Implicit class of Int.

81. What is the best framework to generate rest api documentation for scala-based applications?

Swagger is the best tool for this purpose. It is very simple and open-source tool for generating REST APIs documentation with JSON for Scala-based applications.

If you use Play with Scala to develop your REST API, then use play-

swagger module for REST API documentation.

If you use Spray with Scala to develop your REST API, then use spray-

swagger module for REST API documentation.

82. What is the use of Auxiliary Constructors in Scala?

Auxiliary Constructor is the secondary constructor in Scala declared using the keywords "this" and "def". The main purpose of using auxiliary constructors is to overload constructors. Just like in Java, we can provide implementation for different kinds of constructors so that the right one is invoked based on the requirements. Every auxiliary constructor in Scala should differ in the number of parameters or in data types.

83. How does yield work in Scala?

The yield keyword if specified before the expression, the value returned from every expression, will be returned as the collection. The yield keyword is very useful, when there is a need, you want to use the return value of expression. The collection returned can be used the normal collection and iterate over in another loop.

84. What are the different types of Scala identifiers? There four types of Scala identifiers

Alpha numeric identifiers

Operator identifiers

Mixed identifiers

Literal identifiers

85. What are the different types of Scala literals?

The different types of literals in scala are

Integer literals

Floating point literals Boolean literals

Symbol literals

Character literals

String literals

Multi-Line strings

86. What is SBT? What is the best build tool to develop play and scala applications?

SBT stands for Scala Build Tool. Its a Simple Build Tool to develop Scala-

based applications.

Most of the people uses SBT Build tool for Play and Scala Applications. For example, IntelliJ IDEA Scala Plugin by default uses SBT as Build tool for this purpose.

87. What is the difference between :: and ::: in Scala?

:: and ::: are methods available in List class.

:: method is used to append an element to the beginning of the list.

And ::: method is used to concatenate the elements of a given list in front of this list.

:: method works as a cons operator for List class. Here 'cons' stands for construct.

::: method works as a concatenation operator for List class.

88. What is the difference between #:: and #::: in Scala?

#:: and #::: are methods available in Stream class

#:: method words as a cons operator for Stream class. Here 'cons' stands for construct.

#:: method is used to append a given element at beginning of the stream.

#::: method is used to concatenate a given stream at beginning of the

stream.

89. What is the use of '???' in Scala-based Applications?

This '???' three question marks is not an operator, a method in Scala. It is used to mark a method which is 'In Progress' that means Developer should provide implementation for that one.

90. What is the best Scala style checker tool available for Play and Scala based applications?

Scalastyle is best Scala style checker tool available for Play and Scala based applications. Scalastyle observes our Scala source code and indicates potential problems with it. It has three separate plug-ins to supports the following build tools:

SBT

Maven

Gradle

It has two separate plug-ins to supports the following two IDEs:

IntelliJ IDEA

Eclipse IDE

91. How Scala supports both Highly Scalable and Highly Performance applications?

As Scala supports Multi-Paradigm Programming(Both OOP and FP) and uses Actor Concurrency Model, we can develop very highly Scalable and high-performance applications very easily.

92. What are the available Build Tools to develop Play and Scala based Applications?

The following three are most popular available Build Tools to develop Play and Scala Applications:

SBT

Maven

Gradle

93. What is Either in Scala?

In Scala, either is an abstract class. It is used to represent one value of two possible types. It takes two type parameters: Either[A,B].

94. What are Left and Right in Scala? Explain Either/Left/Right Design Pattern in Scala?

It exactly has two subtypes: Left and Right. If Either[A,B] represents an instance A that means it is Left. If it represents an instance B that means it is Right.

This is known as Either/Left/Right Design Pattern in Scala.

95. How many public class files are possible to define in Scala source file?

In Java, we can define at-most one public class/interface in a Source file.

Unlike Java, Scala supports multiple public classes in the same source file.

We can define any number of public classes/interfaces/traits in a Scala Source file.

96. What is Nothing in Scala?

In Scala, nothing is a Type (final class). It is defined at the bottom of the Scala Type System that means it is a subtype of anything in Scala. There are no instances of Nothing.

97.What's the difference between the following terms and types in Scala: 'Nil', 'Null', 'None', and 'Nothing' in Scala?

Even though they look similar, there are some subtle differences between them, let's see them one by one:

Nil represents the end of a List.

Null denotes the absence of value but in Scala, more precisely, Null is a type that represents the absence of type information for complex types that are inherited from AnyRef. It is different than null in Java. None is the value of an Option if it has no value in it.

Nothing is the bottom type of the entire Scala type system, incorporating all types under AnyVal and AnyRef. Nothing is commonly used as a return type from a method that does not terminate normally and throws an exception.

98. How to you create Singleton classes in Scala?

Scala introduces a new object keyword, which is used to represent Singleton classes. These are the class with just one instance and their method can be thought of as like Java's static methods. Here is a Singleton class in Scala:

```
package test

object Singleton{

def sum(l: List[Int]): Int = l.sum

}
```

This sum method is available globally, and can be referred to, or imported, as the test. Singleton.sum. A singleton object in Scala can also extend classes and traits.

99. What is 'Option' and how is it used in Scala?

The 'Option' in Scala is like Optional of Java 8. It is a wrapper type that avoids the occurrence of a NullPointerException in your code by giving you default value in case object is null. When you call get() from Option it can return a default value if the value is null. More importantly, Option provides the ability to differentiate within the type system those values that can be nulled and those that cannot be nulled.

100. What is the difference between a call-by-value and call-by-name parameter?

The main difference between a call-by-value and a call-by-name parameter is that the former is computed before calling the function, and the later is evaluated when accessed.

101. What is default access modifier in Scala? Does Scala have "public" keyword?

In Scala, if we don't mention any access modifier to a method, function, trait, object or class, the default access modifier is "public". Even for Fields also, "public" is the default access modifier.

Because of this default feature, Scala does not have "public" keyword.

102. Is Scala an Expression-Based Language or Statement-Based Language?

In Scala, everything is a value. All Expressions or Statements evaluates to a Value. We can assign Expression, Function, Closure, Object etc. to a Variable. So, Scala is an Expression-Oriented Language.

103. Is Java an Expression-Based Language or Statement-Based Language?

In Java, Statements are not Expressions or Values. We cannot assign them to a Variable. So, Java is not an Expression-Oriented Language. It is a Statement-Based Language.

104. Mention Some keywords which are used by Java and not required in Scala?

Java uses the following keywords extensively:

'public' keyword - to define classes, interfaces, variables etc. 'static' keyword - to define static members.

105. Why Scala does not require them?

Scala does not require these two keywords. Scala does not have 'public' and 'static' keywords.

In Scala, default access modifier is 'public' for classes, traits, methods/functions, fields etc. That's why, 'public' keyword is not required.

To support OOP principles, Scala team has avoided 'static' keyword. That's why Scala is a Pure-OOP Language. It is very tough to deal static members in Concurrency applications.

Part -3 SQOOP Interview Questions with answer

1) What is SQOOP..?

This is the short meaning of (SQl+HadOOP =SQOOP)

It is a tool designed to transfer data between Hadoop and relational databases or mainframes. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle or a mainframe into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.

Sqoop automates most of this process, relying on the database to describe the schema for the data to be imported. Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance.

The Sqoop main intended for:

System and application programmers

System administrators Database administrators Data analysts

Data engineers

2) Why is the default maximum mappers are 4 in Sqoop?

As of my knowledge, the default number of mapper 4 is followed by minimum concurrent task for one machine. We will lead to set a higher number of concurrent tasks, which can result in faster job completion.

3) is it possible set speculative execution in Sqoop ..?

In sqoop by default speculative execution is off, because if Multiple mappers run for single task, we get duplicates of data in HDFS. Hence to avoid this decrepency it is off. Also number of reducers for sqoop job is 0, since it is merely a job running a MAP only job that dumps data into HDFS. We are not aggregating anything.

4) What causes of hadoop throw ClassNotFoundException while sqoop integration ..?

The most causes of that the supporting library (like connectors) was not updated in sqoop's library path, so we need to update it on that specific path.

5) How to view all the databases and tables in RDBMS from SQOOP..?

Using below commands we can,

sqoop-list-databases

sqoop-list-tables

6) How to view table columns details in RDBMS from SQOOP..?

Unfortunately we don't have any commands like sqoop-list-columns, But we can achieve via free form query to check the information schema for the particular RDBMS tables.here is an example:

$ sqoop eval --connect 'jdbc:mysql://nameofmyserver;' database=nameofmydatabase; username=dineshkumar;

password=dineshkumar --query "SELECT column_name, DATA_TYPE FROM

INFORMATION_SCHEMA.Columns WHERE table_name='mytableofinterest'

7) I am getting FileAlreadyExists exception error in Sqoop while importing data from RDBMS to a hive table.? So How do we resolve it.?

you can specify the --hive-overwrite option to indicate that existing table in hive must be replaced. After your data is imported into HDFS or this step is omitted

8) What is the default file format to import data using Apache Sqoop?

Sqoop allows data to be imported using two file formats

i) Delimited Text File Format

This is the default file format to import data using Sqoop. This file format can be explicitly specified using the -

as-textfile argument to the import command in Sqoop. Passing this as an argument to the command will produce

the string based representation of all the records to the output files with the delimited characters between rows

and columns.

ii) Sequence File Format

It is a binary file format where records are stored in custom record-specific data types which are shown as Java classes. Sqoop automatically creates these data types and manifests them as java classes.

9) How do I resolve a Communications Link Failure when connecting to MySQL?

Verify that you can connect to the database from the node where you are running Sqoop: $ mysql --host=IP Address --database=test --user=username --password=password

Add the network port for the server to your my.cnf file

Set up a user account to connect via Sqoop. Grant permissions to the user to access the database over the

network:

Log into MySQL as root mysql -u root -p ThisIsMyPassword

Issue the following command: mysql> grant all privileges on test.* to 'testuser'@'%' identified by 'testpassword'

10) How do I resolve an IllegalArgumentException when connecting to Oracle?

This could be caused a non-owner trying to connect to the table so prefix the table name with the schema, for example SchemaName.OracleTableName.

11) What's causing this Exception in thread main java.lang.IncompatibleClassChangeError when running non-CDH Hadoop with Sqoop?

Try building Sqoop 1.4.1-incubating with the command line property -Dhadoopversion=20.

12) I have around 300 tables in a database. I want to import all the tables from the database except the tables named Table298, Table 123, and Table299. How can I do this without having to import the tables one by one?

This can be accomplished using the import-all-tables import command in Sqoop and by specifying the exclude-tables option with it as follows-

sqoop import-all-tables--connect -username -password --exclude-tables Table298, Table 123, Table 299

13) Does Apache Sqoop have a default database?

Yes, MySQL is the default database.

bigdatascholars.blogspot.com/2018/08/sqoop-interview-question-and-answers.html

14) How can I import large objects (BLOB and CLOB objects) in Apache Sqoop?

Apache Sqoop import command does not support direct import of BLOB and CLOB large objects. To import large objects, I Sqoop, JDBC based imports have to be used without the direct argument to the import utility.

15) How can you execute a free form SQL query in Sqoop to import the rows in a sequential manner?

This can be accomplished using the -m 1 option in the Sqoop import command. It will create only one MapReduce task which will then import rows serially.

16) What is the difference between Sqoop and DistCP command in Hadoop?

Both distCP (Distributed Copy in Hadoop) and Sqoop transfer data in parallel but the only difference is that distCP command can transfer any kind of data from one Hadoop cluster to another whereas Sqoop transfers data between RDBMS and other components in the Hadoop ecosystem like HBase, Hive, HDFS, etc.

17) What is Sqoop metastore?

Sqoop metastore is a shared metadata repository for remote users to define and execute saved jobs created using sqoop job defined in the metastore. The sqoop -site.xml should be configured to connect to the metastore.

18) What is the significance of using -split-by clause for running parallel import tasks in Apache Sqoop?

--Split-by clause is used to specify the columns of the table that are used to generate splits for data imports. This clause specifies the columns that will be used for splitting when importing the data into the Hadoop cluster. — split-by clause helps achieve improved performance through greater parallelism. Apache Sqoop will create splits based on the values present in the columns specified in the -split-by clause of the import command.

If the -split-by clause is not specified, then the primary key of the table is used to create the splits while data import. At times the primary key of the table might not have evenly distributed values between the minimum and maximum range. Under such circumstances -split-by clause can be used to

specify some other column that has even distribution of data to create splits so that data import is efficient.

19) You use -split-by clause but it still does not give optimal performance then how will you improve the performance further.

Using the -boundary-query clause. Generally, sqoop uses the SQL query select min (), max () from to find out the boundary values for creating splits. However, if this query is not optimal then using the -boundary-query argument any random query can be written to generate two numeric columns.

20) During sqoop import, you use the clause -m or -numb-mappers to specify the number of mappers as 8 so that it can run eight parallel MapReduce tasks, however, sqoop runs only four parallel MapReduce tasks. Why?

Hadoop MapReduce cluster is configured to run a maximum of 4 parallel MapReduce tasks and the sqoop import can be configured with number of parallel tasks less than or equal to 4 but not more than4.

21) You successfully imported a table using Apache Sqoop to HBase but when you query the table it is found that the number of rows is less than expected. What could be the likely reason?

If the imported records have rows that contain null values for all the columns, then probably those records might have been dropped off during import because HBase does not allow null values in all the columns of a record.

22) The incoming value from HDFS for a particular column is NULL. How will you load that row into RDBMS in which the columns are defined as NOT NULL?

Using the -input-null-string parameter, a default value can be specified so that the row gets inserted with the default value for the column that it has a NULL value in HDFS.

23) How will you synchronize the data in HDFS that is imported by Sqoop?

Data can be synchronised using incremental parameter with data import ---Incremental parameter can be used with one of the two options-

i) append-If the table is getting updated continuously with new rows and increasing row id values then incremental import with append option should be used where values of some of the columns are checked (columns to be checked are specified using -check-column) and if it discovers any modified value for those columns then only a new row will be inserted.

ii) lastmodified - In this kind of incremental import, the source has a date column which is checked for. Any records that have been updated after the last import based on the lastmodifed column in the source, the values would be updated.

24) What are the relational databases supported in Sqoop?

Below are the list of RDBMSs that are supported by Sqoop Currently.

MySQL

PostGreSQL

Oracle

Microsoft SQL

IBM's Netezza

Teradata

25) What are the destination types allowed in Sqoop Import command?

Currently Sqoop Supports data imported into below services.

HDFS

Hive

HBase

HCatalog

Accumulo

26) Is Sqoop similar to distcp in hadoop?

Partially yes, hadoop's distcp command is similar to Sqoop Import command. Both submits parallel map-only jobs.

But distcp is used to copy any type of files from Local FS/HDFS to HDFS and Sqoop is for transferring the data records only between RDMBS and Hadoop eco system services, HDFS, Hive and HBase.

27)What are the majorly used commands in Sqoop?

In Sqoop Majorly Import and export commands are used. But below commands are also useful some times.

codegen

eval

import-all-tables

job

list-databases

list-tables

merge

metastore

28) While loading tables from MySQL into HDFS, if we need to copy tables with maximum possible speed, what can you do ?

We need to use -direct argument in import command to use direct import fast path and this -direct can be used only with MySQL and PostGreSQL as of now.

29) While connecting to MySQL through Sqoop, I am getting Connection Failure exception what might be the root cause and fix for this error scenario?

This might be due to insufficient permissions to access your MySQL database over the network. To confirm this we can try the below command to connect to MySQL database from Sqoop's client machine.

$ mysql --host=MySql node > --database=test --user= --password=

If this is the case then we need grant permissions user @ sqoop client machine as per the answer to Question 6 in this post.

30) What is the importance of eval tool?

It allow users to run sample SQL queries against Database and preview the result on the console.

31) What is the process to perform an incremental data load in Sqoop?

The process to perform incremental data load in Sqoop is to synchronize the modified or updated data (often referred as delta data) from RDBMS to Hadoop. The delta data can be facilitated through the incremental load

command in Sqoop.

Incremental load can be performed by using Sqoop import command or by loading the data into hive without overwriting it. The different attributes that need to be specified during incremental load in Sqoop are-

1)Mode (incremental) -The mode defines how Sqoop will determine what the new rows are. The mode can have value as Append or Last Modified.

2)Col (Check-column) -This attribute specifies the column that should be examined to find out the rows to be

imported.

3)Value (last-value) -This denotes the maximum value of the check column from the previous import operation.

32) What is the significance of using -compress-codec parameter?

To get the out file of a sqoop import in formats other than .gz like .bz2 compressions when we use the -compress

-code parameter.

33) Can free form SQL queries be used with Sqoop import command? If yes, then how can they be used?

Sqoop allows us to use free form SQL queries with the import command. The import command should be used with the -e and - query options to execute free form SQL queries. When using the -e and -query options with the import command the -target dir value must be specified.

34) What is the purpose of sqoop-merge?

The merge tool combines two datasets where entries in one dataset should overwrite entries of an older dataset preserving only the newest version of the records between both the data sets.

35) How do you clear the data in a staging table before loading it by Sqoop?

By specifying the -clear-staging-table option we can clear the staging table before it is loaded. This can be done again and again till we get proper data in staging.

36) How will you update the rows that are already exported?

The parameter -update-key can be used to update existing rows. In a comma-separated list of columns is used which uniquely identifies a row. All of these columns is used in the WHERE clause of the generated UPDATE query. All other table columns will be used in the SET part of the query.

37) What is the role of JDBC driver in a Sqoop set up?

To connect to different relational databases sqoop needs a connector. Almost every DB vendor makes this connecter available as a JDBC driver which is specific to that DB. So Sqoop needs the JDBC driver of each of the database it needs to interact with.

38) When to use --target-dir and --warehouse-dir while importing data?

To specify a particular directory in HDFS use --target-dir but to specify the parent directory of all the sqoop jobs use --warehouse-dir. In this case under the parent directory sqoop will create a directory with the same name as the table.

39) When the source data keeps getting updated frequently, what is the approach to keep it in sync with the data in HDFS imported by sqoop?

sqoop can have 2 approaches.

To use the --incremental parameter with append option where value of some columns are checked and only in case of modified values the row is imported as a new row.

To use the --incremental parameter with lastmodified option where a date column in the source is checked for records which have been updated after the last import.

40) Is it possible to add a parameter while running a saved job?

Yes, we can add an argument to a saved job at runtime by using the --exec option sqoop job --exec jobname -- -- newparameter

41) sqoop takes a long time to retrieve the minimum and maximum values of columns mentioned in -split-by parameter. How can we make it efficient?

We can use the --boundary -query parameter in which we specify the min and max value for the column based on which the split can happen into multiple mapreduce tasks. This makes it faster as the query inside the -

boundary-query parameter is executed first and the job is ready with the information on how many mapreduce tasks to create before executing the main query.

42) How will you implement all-or-nothing load using sqoop ?

Using the staging-table option we first load the data into a staging table and then load it to the final target table only if the staging load is successful.

43) How will you update the rows that are already exported ?

The parameter --update-key can be used to update existing rows. In it a comma-separated list of columns is used which uniquely identifies a row. All of these columns is used in the WHERE clause of the generated UPDATE query. All other table columns will be used in the SET part of the query.

44) How can you sync a exported table with HDFS data in which some rows are deleted.?

Truncate the target table and load it again.

45) How can we load to a column in a relational table which is not null but the incoming value from HDFS has a null value.?

By using the -input-null-string parameter we can specify a default value and that will allow the row to be inserted into the target table.

46) How can you schedule a sqoop job using Oozie?

Oozie has in-built sqoop actions inside which we can mention the sqoop commands to be executed.

47) Sqoop imported a table successfully to HBase but it is found that the number of rows is fewer than expected. What can be the cause?

Some of the imported records might have null values in all the columns. As Hbase does not allow all null values in a row, those rows get dropped.

48) How can you force sqoop to execute a free form Sql query only once and import the rows serially. ?

By using the -m 1 clause in the import command, sqoop creates only one mapreduce task which will import the rows sequentially.

49) In a sqoop import command you have mentioned to run 8 parallel Mapreduce task but sqoop runs only 4. What can be the reason?

The Mapreduce cluster is configured to run 4 parallel tasks. So the sqoop command must have number of parallel tasks less or equal to that of the MapReduce cluster.

50) What happens when a table is imported into a HDFS directory which already exists using the -append parameter?

Using the --append argument, Sqoop will import data to a temporary directory and then rename the files into the normal target directory in a manner that does not conflict with existing filenames in that directory.

51) How to import only the updated rows form a table into HDFS using sqoop assuming the source has last update timestamp details for each row?

By using the lastmodified mode. Rows where the check column holds a timestamp more recent than the timestamp specified with --last-value are imported.

52) Give a Sqoop command to import all the records from employee table divided into groups of records by the values in the column department_id.

$ sqoop import --connect jdbc:mysql://DineshDB --table EMPLOYEES --split-by dept_id -m2

53) What does the following query do?

$ sqoop import --connect jdbc:mysql://DineshDB --table sometable --where "id > 1000" --target-dir "/home/dinesh/sqoopincremental" --append

It performs an incremental import of new data, after having already imported the first 1000 rows of a table

54) What is the importance of $CONDITIONS in Sqoop..?

Sqoop performs highly efficient data transfers by inheriting Hadoop's parallelism.

To help Sqoop split your query into multiple chunks that can be transferred in parallel, you need to include the $CONDITIONS placeholder in the where clause of your query.

Sqoop will automatically substitute this placeholder with the generated conditions specifying which slice of data should be transferred by each individual task.

While you could skip $CONDITIONS by forcing Sqoop to run only one job using the --num-mappers 1 param- eter, such a limitation would have a severe performance impact.

For example:-

If you run a parallel import, the map tasks will execute your query with different values substituted in for $CONDITIONS. one mapper may execute "select * from TblDinesh WHERE (salary>=0 AND salary < 10000)", and the next mapper may execute "select * from TblDinesh WHERE (salary >= 10000 AND salary < 20000)" and so on.

55) can sqoop run without a hadoop cluster.?

To run Sqoop commands, Hadoop is a mandatory prerequisite. You cannot run sqoop commands without the Hadoop libraries.

56) Is it possible to import a file in fixed column length from the database using sqoop import?

Importing column of a fixed length from any database you can use free form query like below

sqoop import --connect jdbc:oracle:* --username Dinesh --password pwd

-e "select substr(COL1,1,4000),substr(COL2,1,4000) from table where \$CONDITIONS"

--target-dir /user/dineshkumar/table_name --as-textfile -m 1

57) How to use Sqoop validation?

You can use this parameter (--validate) to validate the counts between what's imported/exported between RDBMS and HDFS.

58) How to pass Sqoop command as file arguments in Sqoop.?

specify an options file, simply create an options file in a convenient location and pass it to the command line via -

-options-file argument.

eg: sqoop --options-file /users/homer/work/import.txt --table TEST

59) is it possible to import data apart from HDFS and Hive.?

Sqoop supports additional import targets beyond HDFS and Hive. Sqoop can also import records into a table in HBase and Accumulo.

60) is it possible to use sqoop --direct command in Hbase .?

This function is incompatible with direct import. But Sqoop can do bulk loading as opposed to direct writes. To use bulk loading, enable it using --hbase-bulkload.

61) Can I configure two sqoop command so that they are dependent on each other? Like if the first sqoop job is successful, second gets triggered. If first fails, second should not run..?

No, using sqoop commands it is not possible, but You can use oozie for this. Create an oozie workflow. Execute the second action only if the first action succeeds.

62) What is UBER mode and where is the settings to enable in Hadoop .?

Normally mappers and reducers will run by ResourceManager (RM), RM will create separate container for mapper and reducer. Uber configuration, will allow to run mapper and reducers in the same process as the ApplicationMaster (AM).

Part -4 Hive Interview Questions with answer

1) What is Hive ?

Apache Hive is an open source for data warehouse system. Its similar like SQL Queries. We can use Hive for analyzing and querying in large data sets on top of Hadoop.

2) Why do we need Hive?

Hive is a tool in Hadoop ecosystem which provides an interface to organize and query data in a databse like fashion and write SQL like queries. It is suitable for accessing and analyzing data in Hadoop using SQL syntax.

3) What is a metastore in Hive?

It is a relational database storing the metadata of hive tables, partitions, Hive databases etc...

When you create a table, this metastore gets updated with the information related to the new table which gets queried when you issue queries on that table.

4) Is Hive suitable to be used for OLTP systems? Why?

No, Hive does not provide insert and update at row level. So it is not suitable for OLTP system.

5) Can you explain about ACID transactions in Hive?

Hive supports ACID transactions: The full form of ACID is Atomicity, Consistency, Isolation and Durability. ACID transactions are provided at the row levels, there are Insert, Delete, and Update options so that Hive supports ACID transaction. Insert, Delete and Update.

6) What are the types of tables in Hive?

There are two types of tables in Hive : Internal Table(aka Managed Table) and External table.

7) What kind of data warehouse application is suitable for Hive?

Hive is not considered as a full database. The design rules and regulations of Hadoop and HDFS put restrictions on what Hive can do.Hive is most suitable for data warehouse applications.

Where Analyzing the relatively static data, Less Responsive time and No rapid changes in data.

Hive does not provide fundamental features required for OLTP (Online Transaction Processing). Hive is suitable for data warehouse applications in large data sets.

8) Explain what is a Hive variable. What do we use it for?

Hive variable is basically created in the Hive environment that is referenced by Hive scripting languages. It provides to pass some values to the hive queries when the query starts executing. It uses the source command.

9) How to change the warehouse.dir location for older tables?

To change the base location of the Hive tables, edit the hive.metastore.warehouse.dir param. This will not affect the older tables. Metadata needs to be changed in the database (MySQL or Derby). The location of Hive tables is in table SDS and column LOCATION.

10) What are the types of metastore available in Hive ?

There are three types of meta stores available in Hive.

Embedded Metastore (Derby)

Local Metastore

Remote Metastore.

11) Is it possible to use same metastore by multiple users, in case of embedded hive?

No, it is not possible to use metastores in sharing mode. It is recommended to use standalone real database like MySQL or PostGresSQL.

12) If you run hive server, what are the available mechanism for connecting it from application?

There are following ways by which you can connect with the Hive Server

1. Thrift Client: Using thrift you can call hive commands from a various programming languages e.g. C++, Java, PHP, Python and Ruby.

2. JDBC Driver : It supports for the Java protocal.

3. ODBC Driver: It supports ODBC protocol.

13) What is SerDe in Apache Hive ?

A SerDe is a short name for a Serializer Deserializer.

Hive uses SerDe as FileFormat to read and write data from tables. An important concept behind Hive is that it DOES NOT own the Hadoop File System format that data is stored in. Users are able to write files to HDFS with whatever tools or mechanism takes their fancy (CREATE EXTERNAL TABLE or LOAD DATA INPATH) and use Hive to correctly parse that file format in a way that can be used by Hive. A SerDe is a powerful and customizable mechanism that Hive uses to parse data stored in HDFS to be used by Hive.

14)Which classes are used by the Hive to Read and Write HDFS Files ?

Following classes are used by Hive to read and write HDFS files

TextInputFormat or HiveIgnoreKeyTextOutputFormat: These 2 classes read/write data in plain text file format.

SequenceFileInputFormat or SequenceFileOutputFormat: These 2 classes read/write data in hadoop SequenceFile format.

15) Give examples of the SerDe classes which hive uses to Serialize and Deserialize data ?

Hive currently use these SerDe classes to serialize and Deserialize data:

MetadataTypedColumnsetSerDe: This SerDe is used to read/write delimited records like CSV, tab-separated control-A separated records (quote is not supported yet.)

ThriftSerDe: This SerDe is used to read or write thrift serialized objects. The class file for the Thrift object must be loaded first.

DynamicSerDe: This SerDe also read or write thrift serialized objects, but it understands thrift DDL so the schema of the object can be provided at runtime. Also it supports a lot of different protocols, including TBinaryProtocol, TJSONProtocol, TCTLSeparatedProtocol(which writes data in delimited records).

16) How do you write your own custom SerDe and what is the need for that?

In most cases, users want to write a Deserializer instead of a SerDe, because users just want to read their own data format instead of writing to it.

For example, the RegexDeserializer will deserialize the data using the configuration parameter regex, and possibly a list of column names.

If your SerDe supports DDL (basically, SerDe with parameterized columns and column types), you probably want to implement a Protocol based on DynamicSerDe, instead of writing a SerDe from scratch. The reason is that the framework passes DDL to SerDe through thrift DDL format, and its non-trivial to write a thrift DDL parser.

Depending on the nature of data the user has, the inbuilt SerDe may not satisfy the format of the data. So users need to write their own java code to satisfy their data format requirements.

17) What is ObjectInspector functionality?

Hive uses ObjectInspector to analyze the internal structure of the row object and also the structure of the individual columns.

ObjectInspector provides a uniform way to access complex objects that can be stored in multiple formats in the memory, including:

Instance of a Java class (Thrift or native Java)

A standard Java object (we use java.util.List to represent Struct and Array, and use java.util.Map to represent Map)

A lazily-initialized object (For example, a Struct of string fields stored in a single Java string object with starting offset for each field)

A complex object can be represented by a pair of ObjectInspector and Java Object. The ObjectInspector not only tells us the structure of the Object, but also gives us ways to access the internal fields inside the Object.

In simple terms, ObjectInspector functionality in Hive is used to analyze the internal structure of the columns, rows, and complex objects. It allows to access the internal fields inside the objects.

18) What is the functionality of Query Processor in Apache Hive?

This component implements the processing framework for converting SQL to a graph of map or reduce jobs and the execution time framework to run those jobs in the order of dependencies and the help of metastore details.

19) What is the limitation of Derby database for Hive metastore?

With derby database, you cannot have multiple connections or multiple sessions instantiated at the same time.

Derby database runs in the local mode and it creates a log file so that multiple users cannot access Hive simultaneously.

20) What are managed and external tables?

We have got two things, one of which is data present in the HDFS and the other is the metadata, present in some database. There are two categories of Hive tables that is Managed and External Tables.

In the Managed tables, both the data and the metadata are managed by Hive and if you drop the managed table, both data and metadata are deleted.There are some situations where your data will be controlled by some other application and you want to read that data but you must allow Hive to delete that data. In such case, you can create an external table in Hive.

In the external table, metadata is controlled by Hive but the actual data will be controlled by some other application. So, when you delete a table accidentally, only the metadata will be lost and the actual data will reside wherever it is.

21) What are the complex data types in Hive?

MAP: The Map contains a key-value pair where you can search for a value using the key.

STRUCT: A Struct is a collection of elements of different data types. For example, if you take the address, it can have different data types. For example, pin code will be in Integer format.

ARRAY: An Array will have a collection of homogeneous elements. For example, if you take your skillset, you can have N number of skills

UNIONTYPE: It represents a column which can have a value that can belong to any of the data types of your choice.

22) How does partitioning help in the faster execution of queries?

With the help of partitioning, a sub directory will be created with the name of the partitioned column and when you perform a query using the WHERE clause, only the particular sub-directory will be scanned instead of scanning the whole table. This gives you faster execution of queries.

23) How to enable dynamic partitioning in Hive?

Related to partitioning there are two types of partitioning Static and Dynamic. In the static partitioning, you will specify the partition column while loading the data.

Whereas in dynamic partitioning, you push the data into Hive and then Hive decides which value should go into which partition. To enable dynamic partitioning, you have set the below property

set hive.exec.dynamic.parition.mode = nonstrict;

Example: insert overwrite table emp_details_partitioned partition(location) select * from emp_details;

24) What is bucketing ?

The values in a column are hashed into a number of buckets which is defined by user. It is a way to avoid too many partitions or nested partitions while ensuring optimizes query output.

25) How does bucketing help in the faster execution of queries?

If you have to join two large tables, you can go for reduce side join. But if both the tables have the same number of buckets or same multiples of buckets and also sorted on the same column there is a possibility of SMBMJ in which all the joins take place in the map phase itself by matching the corresponding buckets. Buckets are basically files that are created inside the HDFS directory.

There are different properties which you need to set for bucket map joins and they are as follows:

set hive.enforce.sortmergebucketmapjoin = false;

set hive.auto.convert.sortmerge.join = false;

set hive.optimize.bucketmapjoin = ture;

set hive.optimize.bucketmapjoin.sortedmerge = true;

26) How to enable bucketing in Hive?

By default bucketing is disabled in Hive, you can enforce to enable it by setting the below property

set hive.enforce.bucketing = true;

27) What are the different file formats in Hive?

Every file format has its own characteristics and Hive allows you to choose easily the file format which you wanted to use.

There are different file formats supported by Hive

1. Text File format
2. Sequence File format
3. Parquet
4. Avro
5. RC file format
6. ORC

28) How is SerDe different from File format in Hive?

SerDe stands for Serializer and Deserializer. It determines how to encode and decode the field values or the column values from a record that is how you serialize and deserialize the values of a column. But file format determines how records are stored in key value format or how do you retrieve the records from the table.

29) What is RegexSerDe?

Regex stands for a regular expression. Whenever you want to have a kind of pattern matching, based on the pattern matching, you have to store the fields.

RegexSerDe is present in org.apache.hadoop.hive.contrib.serde2.RegexSerDe.

In the SerDeproperties, you have to define your input pattern and output fields. For example, you have to get the column values from line xyz/pq@def if you want to take xyz, pq and def separately.

To extract the pattern, you can use:

input.regex = (.*)/(.*)@(.*)

To specify how to store them, you can use

output.format.string = %1$s%2$s%3$s;

30) How is ORC file format optimised for data storage and analysis?

ORC stores collections of rows in one file and within the collection the row data will be stored in a columnar format. With columnar format, it is very easy to compress, thus reducing a lot of storage cost.

While querying also, it queries the particular column instead of querying the whole row as the records are stored in columnar format.

ORC has got indexing on every block based on the statistics min, max, sum, count on columns so when you query, it will skip the blocks based on the indexing.

31) How to access HBase tables from Hive?

Using Hive-HBase storage handler, you can access the HBase tables from Hive and once you are connected, you can query HBase using the SQL queries from Hive. You can also join multiple tables in HBase from Hive and retrieve the result.

32) When running a JOIN query, I see out-of-memory errors.?

This is usually caused by the order of JOIN tables. Instead of [FROM tableA a JOIN tableB b ON], try [FROM tableB b JOIN tableA a ON] NOTE that if you are using LEFT OUTER JOIN, you might want to change to RIGHT OUTER JOIN. This trick usually solve the problem the rule of thumb is, always put the table with a lot of rows having the same value in the join key on the rightmost side of the JOIN.

33) Did you used Mysql as Metatstore and faced errors like com.mysql.jdbc.exceptions.jdbc4. CommunicationsException: Communications link failure ?

This is usually caused by MySQL servers closing connections after the connection is idling for some time. Run the following command on the MySQL server will solve the problem [set global wait_status=120]

When using MySQL as a metastore I see the error [com.mysql.jdbc.exceptions.MySQLSyntaxErrorException: Specified key was too long; max key length is 767 bytes].

This is a known limitation of MySQL 5.0 and UTF8 databases. One option is to use another character set, such as latin1, which is known to work.

34) Does Hive support Unicode?

You can use Unicode string on data or comments, but cannot use for database or table or column name.

You can use UTF-8 encoding for Hive data. However, other encodings are not supported (HIVE 7142 introduce encoding for LazySimpleSerDe, however, the implementation is not complete and not address all cases).

35) Are Hive SQL identifiers (e.g. table names, columns, etc) case sensitive?

No, Hive is case insensitive.

36) What is the best way to load xml data into hive ?

The easiest way is to use the Hive XML SerDe (com.ibm.spss.hive.serde2.xml.XmlSerDe), which will allow you to directly import and work with XML data.

37) When Hive is not suitable?

It does not provide OLTP transactions support only OLAP transactions.If application required OLAP, switch to NoSQL database.HQL queries have higher latency, due to the mapreduce.

38) Mention what are the different modes of Hive?

Depending on the size of data nodes in Hadoop, Hive can operate in two modes. These modes are, Local mode and Map reduce mode

39) Mention what is (HS2) HiveServer2?

It is a server interface that performs following functions.

It allows remote clients to execute queries against Hive Retrieve the results of mentioned queries

Some advanced features Based on Thrift RPC in its latest version include Multi-client concurrency

Authentication.

40) Mention what Hive query processor does?

Hive query processor convert graph of MapReduce jobs with the execution time framework. So that the jobs can be executed in the order of dependencies.

41) Mention what are the steps of Hive in query processor?

The components of a Hive query processor include,

1. Logical Plan Generation
2. Physical Plan Generation Execution Engine
3. Operators
4. UDFs and UDAFs
5. Optimizer
6. Parser
7. Semantic Analyzer
8. Type Checking

42) Explain how can you change a column data type in Hive?

You can change a column data type in Hive by using command,

ALTER TABLE table_name CHANGE column_name column_name new_datatype;

43) Mention what is the difference between order by and sort by in Hive?

SORT BY will sort the data within each reducer. You can use any number of reducers for SORT BY operation.

ORDER BY will sort all of the data together, which has to pass through one reducer. Thus, ORDER BY in hive

uses a single.

44) Explain when to use explode in Hive?

Hadoop developers sometimes take an array as input and convert into a separate table row. To convert complex data types into desired table formats, then we can use explode function.

45) Mention how can you stop a partition form being queried?

You can stop a partition form being queried by using the ENABLE OFFLINE clause with ALTER TABLE statement.

46) Can we rename a Hive table ?

yes, using below command Alter Table table_name RENAME TO new_name

47) What is the default location where hive stores table data?

hdfs://namenode_server/user/hive/warehouse

48) Is there a date data type in Hive?

Yes. The TIMESTAMP data types stores date in java.sql.timestamp format

49) Can we run unix shell commands from hive? Give example.

Yes, using the ! mark just before the command.

For example !pwd at hive prompt will list the current directory.

50) Can hive queries be executed from script files? How?

Using the source command.

Example –

Hive> source /path/to/file/file_with_query.hql

51) What is the importance of .hiverc file?

It is a file containing list of commands needs to run when the hive CLI starts. For example setting the strict mode to be true etc.

52) What are the default record and field delimiter used for hive text files?

The default record delimiter is − \n

And the filed delimiters are − \001,\002,\003

53) What do you mean by schema on read?

The schema is validated with the data when reading the data and not enforced when writing data.

54) How do you list all databases whose name starts with p?

SHOW DATABASES LIKE 'p.*'

55) What does the "USE" command in hive do?

With the use command you fix the database on which all the subsequent hive queries will run.

56) How can you delete the DBPROPERTY in Hive?

There is no way you can delete the DBPROPERTY.

57) What is the significance of the line?

set hive.mapred.mode = strict;

It sets the mapreduce jobs to strict mode.By which the queries on partitioned tables can not run without a WHERE clause. This prevents very large job running for long time.

58) How do you check if a particular partition exists?

This can be done with following query

SHOW PARTITIONS table_name PARTITION(partitioned_column='partition_value')

59) Which java class handles the Input and Output records encoding into files in Hive tables ?

For Input: org.apache.hadoop.mapred.TextInputFormat

For Output: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

60) What is the significance of 'IF EXISTS" clause while dropping a table?

When we issue the command DROP TABLE IF EXISTS table_name

Hive throws an error if the table being dropped does not exist in the first place.

61) When you point a partition of a hive table to a new directory, what happens to the data?

The data stays in the old location. It has to be moved manually.

Write a query to insert a new column(new_col INT) into a hive table (htab) at a position before an existing column (x_col)

ALTER TABLE table_name

CHANGE COLUMN new_col INT BEFORE x_col

62) Does the archiving of Hive tables, it saves any spaces in HDFS?

No. It only reduces the number of files which becomes easier for namenode to manage.

63) While loading data into a hive table using the LOAD DATA clause, how do you specify it is a hdfs file and not a local file ?

By Omitting the LOCAL CLAUSE in the LOAD DATA statement.

64) If you omit the OVERWRITE clause while creating a hive table,what happens to file which are new and files which already exist?

The new incoming files are just added to the target directory and the existing files are simply overwritten. Other files whose name does not match any of the incoming files will continue to exist. If

you add the OVERWRITE clause then all the existing data in the directory will be deleted before new data is written.

65) What does the following query do?

INSERT OVERWRITE TABLE employees PARTITION (country, state)

SELECT ..., se.cnty, se.st

FROM staged_employees se;

It creates partition on table employees with partition values coming from the columns in the select clause. It is called Dynamic partition insert.

66) What is a Table generating Function on hive?

A table generating function is a function which takes a single column as argument and expands it to multiple column or rows. Example exploe().

67) How can Hive avoid MapReduce?

If we set the property hive.exec.mode.local.auto to true then hive will avoid mapreduce to fetch query results.

68) What is the difference between LIKE and RLIKE operators in Hive?

The LIKE operator behaves the same way as the regular SQL operators used in select queries. Example – street_name like '%Chi'

But the RLIKE operator uses more advance regular expressions which are available in java

Example – street_name RLIKE '.*(Chi|Oho).*' which will select any word which has either chi or oho in it.

69) Is it possible to create Cartesian join between 2 tables, using Hive?

No. As this kind of Join can not be implemented in map reduce

70) What should be the order of table size in a join query?

In a join query the smallest table to be taken in the first position and largest table should be taken in the last position.

71) What is the usefulness of the DISTRIBUTED BY clause in Hive?

It controls how the map output is reduced among the reducers. It is useful in case of streaming data

72) How will you convert the string '51.2' to a float value in the price column?

Select cast(price as FLOAT)

73) What will be the result when you do cast('abc' as INT)?

Hive will return NULL

74) Can we LOAD data into a view?

No. A view can not be the target of a INSERT or LOAD statement.

75) What types of costs are associated in creating index on hive tables?

Indexes occupies space and there is a processing cost in arranging the values of the column on which index is cerated.

Give the command to see the indexes on a table. SHOW INDEX ON table_name

This will list all the indexes created on any of the columns in the table table_name.

76) What does /*streamtable(table_name)*/ do?

It is query hint to stream a table into memory before running the query. It is a query optimization Technique.

77) Can a partition be archived? What are the advantages and Disadvantages?

Yes. A partition can be archived. Advantage is it decreases the number of files stored in namenode and the archived file can be queried using hive. The disadvantage is it will cause less efficient query and does not offer any space savings.

78) What is a generic UDF in hive?

It is a UDF which is created using a java program to server some specific need not covered under the existing functions in Hive. It can detect the type of input argument programmatically and provide appropriate response.

79) The following statement failed to execute. What can be the cause?

LOAD DATA LOCAL INPATH '${env:HOME}/country/state/'

OVERWRITE INTO TABLE address;

The local inpath should contain a file and not a directory. The $env:HOME is a valid variable available in the hive

environment.

80) How do you specify the table creator name when creating a table in Hive?

The TBLPROPERTIES clause is used to add the creator name while creating a table. The TBLPROPERTIES is added like −

TBLPROPERTIES('creator'= 'Joan')

81) Which method has to be overridden when we use custom UDF in Hive?

Whenever you write a custom UDF in Hive, you have to extend the UDF class and you have to override the evaluate() function.

82) Suppose I have installed Apache Hive on top of my Hadoop cluster using default metastore configuration. Then, what will happen if we have multiple clients trying to access Hive at the same time?

The default metastore configuration allows only one Hive session to be opened at a time for accessing the metastore. Therefore, if multiple clients try to access the metastore at the same time, they will get an error. One has to use a standalone metastore, i.e. Local or remote metastore configuration in Apache Hive for allowing access to multiple clients concurrently.

Following are the steps to configure MySQL database as the local metastore in Apache Hive: One should make the following changes in hive-site.xml:

1. javax.jdo.option.ConnectionURL property should be set to jdbc:mysql://host/dbname?createDataba seIfNotExist=true.
2. javax.jdo.option.ConnectionDriverName property should be set to com.mysql.jdbc.Driver. One should also set the username and password as:
3. javax.jdo.option.ConnectionUserName is set to desired username.
4. javax.jdo.option.ConnectionPassword is set to the desired password.

The JDBC driver JAR file for MySQL must be on the Hive classpath, i.e. The jar file should be copied into the Hive lib directory.

Now, after restarting the Hive shell, it will automatically connect to the MySQL database which is running as a standalone metastore.

83) Is it possible to change the default location of a managed table?

Yes, it is possible to change the default location of a managed table. It can be achieved by using the clause LOCATION [hdfs_path].

84) When should we use SORT BY instead of ORDER BY?

We should use SORT BY instead of ORDER BY when we have to sort huge datasets because SORT BY clause sorts the data using multiple reducers whereas ORDER BY sorts all of the data together using a single reducer. Therefore, using ORDER BY against a large number of inputs will take a lot of time to execute.

85) What is dynamic partitioning and when is it used?

In dynamic partitioning values for partition columns are known in the runtime, i.e. It is known during loading of the data into a Hive table.

One may use dynamic partition in following two cases:

1. Loading data from an existing non-partitioned table to improve the sampling and therefore, decrease the

query latency.

2. When one does not know all the values of the partitions before hand and therefore, finding these partition values manually from a huge data sets is a tedious task.

86) Suppose, I create a table that contains details of all the transactions done by the customers of year 2016: CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ,; Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?

We can solve this problem of query latency by partitioning the table according to each month. So, for each month we will be scanning only the partitioned data instead of whole data sets.

As we know, we can not partition an existing non-partitioned table directly. So, we will be taking following steps to solve the very problem:

- 1.Create a partitioned table, say partitioned_transaction:
- CREATE TABLE partitioned_transaction (cust_id INT, amount FLOAT, country STRING) PARTITIONED BY (month STRING) ROW FORMAT DELIMITED FIELDS
- TERMINATED BY (,) ;
- 2. Enable dynamic partitioning in Hive:
- SET hive.exec.dynamic.partition = true;
- SET hive.exec.dynamic.partition.mode = nonstrict;
- 3. Transfer the data from the non - partitioned table into the newly created partitioned table:

INSERT OVERWRITE TABLE partitioned_transaction PARTITION (month) SELECT cust_id, amount, country, month FROM transaction_details;

Now, we can perform the query using each partition and therefore, decrease the query time.

87) How can you add a new partition for the month December in the above partitioned table?

For adding a new partition in the above table partitioned_transaction, we will issue the command give below:

ALTER TABLE partitioned_transaction ADD PARTITION (month=Dec) LOCATION /partitioned_transaction;

88) What is the default maximum dynamic partition that can be created by a mapper/reducer? How can you change it?

By default the number of maximum partition that can be created by a mapper or reducer is set to 100. One can change it by issuing the following command:

SET hive.exec.max.dynamic.partitions.pernode = value

89) I am inserting data into a table based on partitions dynamically. But, I received an error FAILED ERROR IN SEMANTIC ANALYSIS: Dynamic partition strict mode requires at least one static partition column. How will you remove this error?

To remove this error one has to execute following commands:

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

90) Suppose, I have a CSV file sample.csv present in temp directory with the following entries:

id first_name last_name email gender ip_address

1 Hugh Jackman hughjackman@cam.ac.uk Male 136.90.241.52

2 David Lawrence dlawrence1@gmail.com Male 101.177.15.130

3 Andy Hall andyhall2@yahoo.com Female 114.123.153.64

4 Samuel Jackson samjackson231@sun.com Male 89.60.227.31

5 Emily Rose rose.emily4@surveymonkey.com Female 119.92.21.19

How will you consume this CSV file into the Hive warehouse using built SerDe?

SerDe stands for serializer or deserializer. A SerDe allows us to convert the unstructured bytes into a record that we can process using Hive. SerDes are implemented using Java. Hive comes with several built-in SerDes and many other third-party SerDes are also available.

Hive provides a specific SerDe for working with CSV files. We can use this SerDe for the sample.csv by issuing following commands:

CREATE EXTERNAL TABLE sample (id int, first_name string,last_name string, email string,

gender string, ip_address string)

ROW FORMAT SERDE org.apache.hadoop.hive.serde2.OpenCSVSerde STORED AS TEXTFILE LOCATION temp;

Now, we can perform any query on the table sample:

SELECT first_name FROM sample WHERE gender = male;

91) Suppose, I have a lot of small CSV files present in input directory in HDFS and I want to create a single Hive table corresponding to these files. The data in these files are in the format: {id, name, e-mail, country}. Now, as we know, Hadoop performance degrades when we use lots of small files.

So, how will you solve this problem where we want to create a single Hive table for lots of small files without degrading the performance of the system?

One can use the SequenceFile format which will group these small files together to form a single sequence file. The steps that will be followed in doing so are as follows:

1.Create a temporary table:

CREATE TABLE temp_table (id INT, name STRING, e-mail STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY , STORED AS TEXTFILE;

Load the data into temp_table:

LOAD DATA INPATH input INTO TABLE temp_table;

2.Create a table that will store data in SequenceFile format:

CREATE TABLE sample_seqfile (id INT, name STRING, e-mail STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY , STORED AS SEQUENCEFILE; Transfer the data from the temporary table into the sample_seqfile table:

INSERT OVERWRITE TABLE sample SELECT * FROM temp_table;

Hence, a single SequenceFile is generated which contains the data present in all of the input files and therefore, the problem of having lots of small files is finally eliminated.

92) Can We Change settings within Hive Session? If Yes, How?

Yes, we can change the settings within Hive session, using the SET command. It helps to change Hive job settings for an exact query.

Example: The following commands shows buckets are occupied according to the table definition.
hive> SET hive.enforce.bucketing=true;

We can see the current value of any property by using SET with the property name. SET will list all the properties with their values set by Hive.

hive> SET hive.enforce.bucketing;

hive.enforce.bucketing=true

And this list will not include defaults of Hadoop. So we should use the below like

SET -v

It will list all the properties including the Hadoop defaults in the system.

93) Is it possible to add 100 nodes when we have 100 nodes already in Hive? How?

Yes, we can add the nodes by following the below steps.

Take a new system create a new username and password.

Install the SSH and with master node setup ssh connections. Add ssh public_rsa id key to the authorized keys file.

Add the new data node host name, IP address and other details in /etc/hosts slaves file 192.168.1.102 slave3.in slave3.

Start the Data Node on New Node.

Login to the new node like suhadoop or ssh -X hadoop@192.168.1.103.

Start HDFS of a newly added slave node by using the following command ./bin/hadoop-daemon.sh start data node.

Check the output of jps command on a new node

94) Explain the concatenation function in Hive with an example?

Concatenate function will join the input strings.We can specify the N number of strings separated by a comma.

Example:

CONCAT (It,-,is,-,a,-,eLearning,-,provider);

Output:

It-is-a-eLearning-provider

So, every time we set the limits of the strings by -. If it is common for every strings, then Hive provides another

command

CONCAT_WS. In this case,we have to specify the set limits of operator first. CONCAT_WS (-,It,is,a,eLearning,provider);

Output: It-is-a-eLearning-provider.

95) Explain Trim and Reverse function in Hive with examples?

Trim function will delete the spaces associated with a string.

Example:

TRIM(BHAVESH);

Output:

BHAVESH

To remove the Leading space LTRIM(BHAVESH);

To remove the trailing space

RTRIM(BHAVESH);

In Reverse function, characters are reversed in the string.

Example:

REVERSE(BHAVESH);

Output:

HSEVAHB

96) Explain process to access sub directories recursively in Hive queries?

By using below commands we can access sub directories recursively in Hive hive> Set mapred.input.dir.recursive=true;

hive> Set hive.mapred.supports.subdirectories=true;

Hive tables can be pointed to the higher level directory and this is suitable for the directory structure which is like /data/country/state/city/

97) How to skip header rows from a table in Hive?

Header records in log files

1. System=
2. Version=
3. Sub-version=

In the above three lines of headers that we do not want to include in our Hive query. To skip header lines from our tables in the Hive,set a table property that will allow us to skip the header lines.

CREATE EXTERNAL TABLE employee (name STRING,job STRING, dob STRING, id INT,

salary INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY STORED AS TEXTFILE LOCATION /user/data

TBLPROPERTIES(skip.header.line.count=2);

98) What is the maximum size of string data type supported by hive? Mention the Hive support binary formats ?

The maximum size of string data type supported by hive is 2 GB.

Hive supports the text file format by default and it supports the binary format Sequence files, ORC files, Avro Data files, Parquet files.

Sequence files: Splittable, compressible and row oriented are the general binary format.

ORC files: Full form of ORC is optimized row columnar format files. It is a Record columnar file and column oriented storage file. It divides the table in row split. In each split stores that value of the first row in the first column and followed sub subsequently.

AVRO datafiles: It is same as a sequence file splittable, compressible and row oriented, but except the support of schema evolution and multilingual binding support.

99) What is the precedence order of HIVE configuration?

We are using a precedence hierarchy for setting the properties SET Command in HIVE

The command line -hiveconf option Hive-site.XML

1. Hive-default.xml
2. Hadoop-site.xml
3. Hadoop-default.xml

100) If you run a select * query in Hive, Why does it not run MapReduce?

The hive.fetch.task.conversion property of Hive lowers the latency of mapreduce overhead and in effect when executing queries like SELECT, LIMIT, etc., it skips mapreduce function

101) How Hive can improve performance with ORC format tables?

We can store the hive data in highly efficient manner in the Optimized Row Columnar file format. It can simplify many Hive file format limitations. We can improve the performance by using ORC files while reading, writing and processing the data.

Set hive.compute.query.using.stats-true; Set hive.stats.dbclass-fs;

CREATE TABLE orc_table (idint,name string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY \: LINES TERMINATED BY \n STORES AS ORC;

102) Explain about the different types of join in Hive?

HiveQL has 4 different types of joins -

JOIN- Similar to Outer Join in SQL

FULL OUTER JOIN - Combines the records of both the left and right outer tables that fulfil the join condition.

LEFT OUTER JOIN- All the rows from the left table are returned even if there are no matches in the right table.

RIGHT OUTER JOIN-All the rows from the right table are returned even if there are no matches in the left table.

103) How can you configure remote metastore mode in Hive?

To configure metastore in Hive, hive-site.xml file has to be configured with the below property -

hive.metastore.uris

thrift: //node1 (or IP Address):9083

IP address and port of the metastore host

104) What happens on executing the below query? After executing the below query, if you modify the column how will the changes be tracked?

Hive> CREATE INDEX index_bonuspay ON TABLE employee (bonus)

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

The query creates an index named index_bonuspay which points to the bonus column in the employee table. Whenever the value of bonus is modified it will be stored using an index value.

105) How to load Data from a .txt file to Table Stored as ORC in Hive?

LOAD DATA just copies the files to hive datafiles. Hive does not do any transformation while loading data into tables.

So, in this case the input file /home/user/test_details.txt needs to be in ORC format if you are loading it into an ORC table.

A possible workaround is to create a temporary table with STORED AS TEXT, then LOAD DATA into it, and then copy data from this table to the ORC table.

Here is an example:

CREATE TABLE test_details_txt(visit_id INT, store_id SMALLINT) STORED AS TEXTFILE;

CREATE TABLE test_details_orc(visit_id INT, store_id SMALLINT) STORED AS ORC;

Load into Text table

LOAD DATA LOCAL INPATH /home/user/test_details.txt INTO TABLE test_details_txt; Copy to ORC table

INSERT INTO TABLE test_details_orc SELECT * FROM test_details_txt;

106) How to create HIVE Table with multi character delimiter

FILELDS TERMINATED BY does not support multi-character delimiters. The easiest way to do this is to use RegexSerDe:

CREATE EXTERNAL TABlE tableex(id INT, name STRING)

ROW FORMAT org.apache.hadoop.hive.contrib.serde2.RegexSerDe WITH SERDEPROPERTIES (input.regex = ^(\\d+)~*(.*)$)

STORED AS TEXTFILE LOCATION /user/myusername;

107) Is there any way to get the column name along with the output while execute any query in Hive?

If we want to see the columns names of the table in HiveQl, the following hive conf property should be set to true. hive> set hive.cli.print.header=true;

If you prefer to see the column names always then update the $HOME/.hiverc file with the above setting in the first line.

Hive automatically looks for a file named .hiverc in your HOME directory and runs the commands it contains, if any.

108) How to Improve Hive Query Performance With Hadoop?

1.Use Tez Engine

Apache Tez Engine is an extensible framework for building high-performance batch processing and interactive data processing. It is coordinated by YARN in Hadoop. Tez improved the MapReduce paradigm by increasing the processing speed and maintaining the MapReduce ability to scale to petabytes of data.

Tez engine can be enabled in your environment by setting hive.execution.engine to tez:

set hive.execution.engine=tez;

2.Use Vectorization

Vectorization improves the performance by fetching 1,024 rows in a single operation instead of fetching single row each time. It improves the performance for operations like filter, join, aggregation, etc.

Vectorization can be enabled in the environment by executing below commands.

- set hive.vectorized.execution.enabled=true;
- set hive.vectorized.execution.reduce.enabled=true;

3.Use ORCFile

Optimized Row Columnar format provides highly efficient ways of storing the hive data by reducing the data storage format by 75% of the original. The ORCFile format is better than the Hive files

format when it comes to reading, writing, and processing the data. It uses techniques like predicate push-down, compression, and more to improve the performance of the query.

4.Use Partitioning

With partitioning, data is stored in separate individual folders on HDFS. Instead of querying the whole dataset, it will query partitioned dataset.

1)Create Temporary Table and Load Data Into Temporary Table

2)Create Partitioned Table

3)Enable Dynamic Hive Partition

4)Import Data From Temporary Table To Partitioned Table

5.Use Bucketing

The Hive table is divided into a number of partitions and is called Hive Partition. Hive Partition is further subdivided into clusters or buckets and is called bucketing or clustering.

Cost-Based Query Optimization

Hive optimizes each querys logical and physical execution plan before submitting for final execution. However, this is not based on the cost of the query during the initial version of Hive.

During later versions of Hive, query has been optimized according to the cost of the query (like which types of join to be performed, how to order joins, the degree of parallelism, etc.).

109) How do I query from a horizontal output to vertical output?

There should be easier way to achieve that using explode function and selecting separately data for prev and next columns.

110) Is there a simple way to replace non numeric characters hive excluding - to allow only -ve and +ve numbers.?

we can try to use regexp_extract instead:

regexp_extract('abcd-9090','.*(-[0-9]+)',1)

111) What is Hive Tablename maximum character limit.?

Metastore service uses a RDBMS as back-end, and you can see that the standard MySQL schema defines

CREATE TABLE IF NOT EXISTS TBLS (...

TBL_NAME varchar(128) CHARACTER SET latin1 COLLATE latin1_bin DEFAULT NULL,

112) How can I convert a date in string format ("April 25, 2018") to timestamp in hive?

use from_unixtime in conjunction with unix_timestamp.

select from_unixtime(unix_timestamp(`date`,'MMM dd, yyyy'),'yyyy-MM-dd')

113) How to drop the hive database whether it contains some tables.?

Use cascade command while drop the database.

Example:

hive> drop database sampleDB cascade;

114) I Dropped and recreated hive external table, but no data shown, So what should I do?

This is because the table you created is a partitioned table. The insert you ran would have created a partition for partition_column='abcxyz'. When you drop and re-create the table, Hive looses the information about the partition, it only knows about the table.

Run the command below to get hive to re-create the partitions based on the data.

MSCK REPAIR TABLE user_dinesh.sampletable;

Part -5 Spark Interview Questions with Answers

1.Spark Architecture

Apache Spark follows a master/slave architecture with two main daemons and a cluster manager -

i. Master Daemon - (Master/Driver Process) ii. Worker Daemon -(Slave Process)

A spark cluster has a single Master and any number of Slaves/Workers. The driver and the executors run their

individual Java processes and users can run them on the same horizontal spark cluster or on separate machines i.e. in a vertical spark cluster or in mixed machine configuration.

2.Explain about Spark submission

The spark-submit script in Spark's bin directory is used to launch applications on a cluster. It can use all of Spark's supported cluster managers through a uniform interface so you don't have to configure your application especially for each one.

youtube link: https://youtu.be/t84cxWxiiDg

Example code:

```
./bin/spark-submit \
--class org.apache.spark.examples.SparkPi \
--master spark://207.184.161.138:7077 \
--deploy-mode cluster \
--supervise \
--executor-memory 20G \
--total-executor-cores 100 \ /path/to/examples.jar
arguments1
```

3. Difference Between RDD, Dataframe, Dataset?

Resilient Distributed Dataset (RDD)

RDD was the primary user-facing API in Spark since its inception. At the core, an RDD is an immutable distributed collection of elements of your data, partitioned across nodes in your cluster that can be operated in parallel with a low-level API that offers transformations and actions.

DataFrames (DF)

Like an RDD, a DataFrame is an immutable distributed collection of data. Unlike an RDD, data is organized into named columns, like a table in a relational database. Designed to make large data sets processing even easier, DataFrame allows developers to impose a structure onto a distributed collection of data, allowing higher-level abstraction; it provides a domain specific language API to manipulate your distributed data.

Datasets (DS)

Starting in Spark 2.0, Dataset takes on two distinct APIs characteristics: a strongly-typed API and an untyped API, as shown in the table below. Conceptually, consider DataFrame as an alias for a collection of generic objects Dataset[Row], where a Row is a generic untyped JVM object.

4. When to use RDDs?

Consider these scenarios or common use cases for using RDDs when:

1. you want low-level transformation and actions and control on your dataset; your data is unstructured, such as media streams or streams of text;
2. you want to manipulate your data with functional programming constructs than domain specific expressions;
3. you don't care about imposing a schema, such as columnar format, while processing or accessing data attributes by name or column; and
4. you can forgo some optimization and performance benefits available with DataFrames and Datasets for structured and semi-structured data.

5. What are the various modes in which Spark runs on YARN? (Client vs Cluster Mode)

YARN client mode: The driver runs on the machine from which client is connected

YARN Cluster Mode: The driver runs inside cluster

6. What is DAG - Directed Acyclic Graph?

Directed Acyclic Graph - DAG is a graph data structure which has edge which are directional and does not have any loops or cycles. It is a way of representing dependencies between objects. It is widely used in computing.

7. What is a RDD and How it works internally?

RDD (Resilient Distributed Dataset) is a representation of data located on a network

which is Immutable - You can operate on the rdd to produce another rdd but you can't alter it.

Partitioned / Parallel - The data located on RDD is operated in parallel. Any operation on RDD is done using multiple nodes.

Resilience - If one of the node hosting the partition fails, other nodes takes its data.

You can always think of RDD as a big array which is under the hood spread over many computers which is completely abstracted. So, RDD is made up many partitions each partition on different computers.

8. What do we mean by Partitions or slices?

Partitions also known as 'Slice' in HDFS, is a logical chunk of data set which may be in the range of Petabyte, Terabytes and distributed across the cluster.

By Default, Spark creates one Partition for each block of the file (For HDFS)

Default block size for HDFS block is 64 MB (Hadoop Version 1) / 128 MB (Hadoop Version 2) so as

the split size.

However, one can explicitly specify the number of partitions to be created. Partitions are basically used to speed up the data processing.

If you are loading data from an existing memory using sc.parallelize(), you can enforce your number of partitions by passing second argument.

You can change the number of partitions later using repartition().

If you want certain operations to consume the whole partitions at a time, you can use: mappartition().

9.What is the difference between map and flatMap?

Map and flatmap both function are applied on each element of RDD. The only difference is that the function that is applied as part of map must return only one value while flatmap can return a list of values.

So, flatmap can convert one element into multiple elements of RDD while map can only result in equal number of elements.

So, if we are loading rdd from a text file, each element is a sentence. To convert this RDD into an RDD of words, we will have to apply using flatmap a function that would split a string into an array of words. If we have just to cleanup each sentence or change case of each sentence, we would be using map instead of flatmap.

10.How can you minimize data transfers when working with Spark?

The various ways in which data transfers can be minimized when working with Apache Spark are:

1.Broadcast Variable- Broadcast variable enhances the efficiency of joins between small and large RDDs.

2.Accumulators - Accumulators help update the values of variables in parallel while executing.

3.The most common way is to avoid operations ByKey, repartition or any other operations which trigger shuffles.

11.Why is there a need for broadcast variables when working with Apache Spark?

These are read only variables, present in-memory cache on every machine. When working with Spark, usage of broadcast variables eliminates the necessity to ship copies of a variable for every task, so data can be processed faster. Broadcast variables help in storing a lookup table inside the memory which enhances the retrieval efficiency when compared to an RDD lookup ().

12.How can you trigger automatic clean-ups in Spark to handle accumulated metadata?

You can trigger the clean-ups by setting the parameter "spark.cleaner.ttl" or by dividing the long running jobs into different batches and writing the intermediary results to the disk.

13. Why is BlinkDB used?

BlinkDB is a query engine for executing interactive SQL queries on huge volumes of data and renders query results marked with meaningful error bars. BlinkDB helps users balance 'query accuracy' with response time.

14. What is Sliding Window operation?

Sliding Window controls transmission of data packets between various computer networks. Spark Streaming library provides windowed computations where the transformations on RDDs are applied over a sliding window of data. Whenever the window slides, the RDDs that fall within the particular window are combined and operated upon to produce new RDDs of the windowed DStream.

15. What is Catalyst Optimiser?

Catalyst Optimizer is a new optimization framework present in Spark SQL. It allows Spark to automatically transform SQL queries by adding new optimizations to build a faster processing system.

16. What do you understand by Pair RDD?

Paired RDD is a distributed collection of data with the key-value pair. It is a subset of Resilient Distributed Dataset. So it has all the feature of RDD and some new feature for the key-value pair. There are many transformation operations available for Paired RDD. These operations on Paired RDD are very useful to solve many use cases that require sorting, grouping, reducing some value/function.

Commonly used operations on paired RDD are: groupByKey() reduceByKey() countByKey() join()

etc

17. What is the difference between persist() and cache()?

persist () allows the user to specify the storage level whereas cache () uses the default storage level(MEMORY_ONLY).

18. What are the various levels of persistence in Apache Spark?

Apache Spark automatically persists the intermediary data from various shuffle operations, however it is often suggested that users call persist () method on the RDD in case they plan to reuse it. Spark has various persistence levels to store the RDDs on disk or in memory or as a combination of both with different replication levels.

The various storage/persistence levels in Spark are -

- MEMORY_ONLY
- MEMORY_ONLY_SER
- MEMORY_AND_DISK

• MEMORY_AND_DISK_SER, DISK_ONLY

• OFF_HEAP

19.Does Apache Spark provide check pointing?

Lineage graphs are always useful to recover RDDs from a failure but this is generally time consuming if the RDDs have long lineage chains. Spark has an API for check pointing i.e. a REPLICATE flag to persist. However, the decision on which data to checkpoint - is decided by the user. Checkpoints are useful when the lineage graphs are long and have wide dependencies.

20.What do you understand by Lazy Evaluation?

Spark is intellectual in the manner in which it operates on data. When you tell Spark to operate on a given dataset, it heeds the instructions and makes a note of it, so that it does not forget - but it does nothing, unless asked for the final result. When a transformation like map () is called on a RDD-the operation is not performed immediately. Transformations in Spark are not evaluated till you perform an action. This helps optimize the overall data processing workflow.

21.What do you understand by SchemaRDD?

An RDD that consists of row objects (wrappers around basic string or integer arrays) with schema information about the type of data in each column. Dataframe is an example of SchemaRDD.

22.What are the disadvantages of using Apache Spark over Hadoop MapReduce?

Apache spark does not scale well for compute intensive jobs and consumes large number of system resources. Apache Spark's in-memory capability at times comes a major roadblock for cost efficient processing of big data. Also, Spark does have its own file management system and hence needs to be integrated with other cloud based data platforms or apache hadoop.

23.What is "Lineage Graph" in Spark?

Whenever a series of transformations are performed on an RDD, they are not evaluated immediately, but lazily(Lazy Evaluation). When a new RDD has been created from an existing RDD, that new RDD contains a pointer to the parent RDD. Similarly, all the dependencies between the RDDs will be logged in a graph, rather than the actual data. This graph is called the lineage graph.

Spark does not support data replication in the memory. In the event of any data loss, it is rebuilt using the "RDD Lineage". It is a process that reconstructs lost data partitions.

24.What do you understand by Executor Memory in a Spark application?

Every spark application has same fixed heap size and fixed number of cores for a spark executor. The heap size is what referred to as the Spark executor memory which is controlled with the spark.

executor.memory property of the -executor-memory flag. Every spark application will have one executor on each worker node. The executor memory is basically a measure on how much memory of the worker node will the application utilize.

25. What is an "Accumulator"?

"Accumulators" are Spark's offline debuggers. Similar to "Hadoop Counters", "Accumulators" provide the number of "events" in a program.

Accumulators are the variables that can be added through associative operations. Spark natively supports accumulators of numeric value types and standard mutable collections. "AggregrateByKey()" and "combineByKey()" uses accumulators.

26.What is SparkContext .?

sparkContext was used as a channel to access all spark functionality.

The spark driver program uses spark context to connect to the cluster through a resource manager (YARN or Mesos.). sparkConf is required to create the spark context object, which stores configuration parameter like appName (to identify your spark driver), application, number of core and memory size of executor running on worker node.

In order to use APIs of SQL, HIVE, and Streaming, separate contexts need to be created.

Example:

creating sparkConf :

val conf = new SparkConf().setAppName("Project").setMaster("spark://master:7077") creation of sparkContext: val sc = new SparkContext(conf)

27.What is SparkSession ?

SPARK 2.0.0 onwards, SparkSession provides a single point of entry to interact with underlying Spark functionality and it allows Spark programming with DataFrame and Dataset APIs. All the functionality available with sparkContext are also available in sparkSession.

In order to use APIs of SQL, HIVE, and Streaming, no need to create separate contexts as sparkSession includes all the APIs.

Once the SparkSession is instantiated, we can configure Spark's run-time config properties.

Example:Creating Spark session:

val spark = SparkSession.builder.appName("WorldBankIndex").getOrCreate() Configuring properties:

spark.conf.set("spark.sql.shuffle.partitions", 6)

```
spark.conf.set("spark.executor.memory", "2g")
```

28. Why RDD is an immutable ?

Following are the reasons:

Immutable data is always safe to share across multiple processes as well as multiple threads.

Since RDD is immutable we can recreate the RDD any time. (From lineage graph). If the computation is time-consuming, in that we can cache the RDD which result in performance improvement.

29. What is Partitioner?

A partitioner is an object that defines how the elements in a key-value pair RDD are partitioned by key, maps each key to a partition ID from 0 to numPartitions - 1. It captures the data distribution at the output. With the help of partitioner, the scheduler can optimize the future operations. The contract of partitioner ensures that records for a given key have to reside on a single partition.

We should choose a partitioner to use for a cogroup-like operations. If any of the RDDs already has a partitioner, we should choose that one. Otherwise, we use a default HashPartitioner.

There are three types of partitioners in Spark :

a) Hash Partitioner :- Hash- partitioning attempts to spread the data evenly across various partitions based on the key.

b) Range Partitioner :- In Range- Partitioning method , tuples having keys with same range will appear on the same machine.

c) Custom Partitioner

RDDs can be created with specific partitioning in two ways :

i) Providing explicit partitioner by calling partitionBy method on an RDD

ii) Applying transformations that return RDDs with specific partitioners .

30. What are the benefits of DataFrames ?

1.DataFrame is distributed collection of data. In DataFrames, data is organized in named column.

2. They are conceptually similar to a table in a relational database. Also, have richer optimizations.

3. Data Frames empower SQL queries and the DataFrame API.

4. we can process both structured and unstructured data formats through it. Such as: Avro, CSV, elastic search, and Cassandra. Also, it deals with storage systems HDFS, HIVE tables, MySQL, etc.

5. In Data Frames, Catalyst supports optimization(catalyst Optimizer). There are general libraries available to represent trees. In four phases, DataFrame uses Catalyst tree transformation:

- Analyze logical plan to solve references

- Logical plan optimization

- Physical planning

- Code generation to compile part of a query to Java bytecode.

6. The Data Frame API's are available in various programming languages. For example Java, Scala, Python, and R.

7. It provides Hive compatibility. We can run unmodified Hive queries on existing Hive warehouse.

8. It can scale from kilobytes of data on the single laptop to petabytes of data on a large cluster.

9. DataFrame provides easy integration with Big data tools and framework via Spark core.

31. What is Dataset ?

A Dataset is an immutable collection of objects, those are mapped to a relational schema. They are strongly-typed in nature.

There is an encoder, at the core of the Dataset API. That Encoder is responsible for converting between JVM objects and tabular representation. By using Spark's internal binary format, the tabular representation is stored that allows to carry out operations on serialized data and improves memory utilization. It also supports automatically generating encoders for a wide variety of types, including primitive types (e.g. String, Integer, Long) and Scala case classes. It offers many functional transformations (e.g. map, flatMap, filter).

32. What are the benefits of Datasets?

1. 1)Static typing- With Static typing feature of Dataset, a developer can catch errors at compile time (which saves time and costs).
2. 2)Run-time Safety:- Dataset APIs are all expressed as lambda functions and JVM typed objects, any mismatch of typed-parameters will be detected at compile time. Also, analysis error can be detected at compile time too,when using Datasets, hence saving developer-time and costs.
3. 3)Performance and Optimization- Dataset APIs are built on top of the Spark SQL engine, it uses Catalyst to generate an optimized logical and physical query plan providing the space and speed efficiency.

4. 4) For processing demands like high-level expressions, filters, maps, aggregation, averages, sum,SQL queries, columnar access and also for use of lambda functions on semi-structured data, DataSets are best.
5. 5) Datasets provides rich semantics, high-level abstractions, and domain-specific APIs

33. What is Shared variable in Apache Spark?

Shared variables are nothing but the variables that can be used in parallel operations.

Spark supports two types of shared variables: broadcast variables, which can be used to cache a value in memory on all nodes, and accumulators, which are variables that are only "added" to, such as counters and sums.

34.How to accumulated Metadata in Apache Spark?

Metadata accumulates on the driver as consequence of shuffle operations. It becomes particularly tedious during long-running jobs.

To deal with the issue of accumulating metadata, there are two options:

First, set the spark.cleaner.ttl parameter to trigger automatic cleanups. However, this will vanish any persisted RDDs.

The other solution is to simply split long-running jobs into batches and write intermediate results to

disk. This facilitates a fresh environment for every batch and don't have to worry about metadata build-up.

35. What is the Difference between DSM and RDD?

On the basis of several features, the difference between RDD and DSM is:

i. Read

RDD - The read operation in RDD is either coarse-grained or fine-grained. Coarse-grained meaning we can transform the whole dataset but not an individual element on the dataset. While fine-grained means we can transform individual element on the dataset.

DSM - The read operation in Distributed shared memory is fine-grained.

ii. Write

RDD - The write operation in RDD is coarse-grained.

DSM - The Write operation is fine grained in distributed shared system.

iii. Consistency

RDD - The consistency of RDD is trivial meaning it is immutable in nature. We can not realtor the content of RDD i.e. any changes on RDD is permanent. Hence, The level of consistency is very high.

DSM - The system guarantees that if the programmer follows the rules, the memory will be consistent. Also, the results of memory operations will be predictable.

iv. Fault-Recovery Mechanism

RDD - By using lineage graph at any moment, the lost data can be easily recovered in Spark RDD. Therefore, for each transformation, new RDD is formed. As RDDs are immutable in nature, hence, it is easy to recover.

DSM - Fault tolerance is achieved by a checkpointing technique which allows applications to roll back to a recent checkpoint rather than restarting.

v. Straggler Mitigation

Stragglers, in general, are those that take more time to complete than their peers. This could happen due to many reasons such as load imbalance, I/O blocks, garbage collections, etc.

An issue with the stragglers is that when the parallel computation is followed by synchronizations such as reductions that causes all the parallel tasks to wait for others.

RDD - It is possible to mitigate stragglers by using backup task, in RDDs. DSM - To achieve straggler mitigation, is quite difficult.

vi. Behavior if not enough RAM

RDD - As there is not enough space to store RDD in RAM, therefore, the RDDs are shifted to disk. DSM - If the RAM runs out of storage, the performance decreases, in this type of systems.

36. What is Speculative Execution in Spark and how to enable it?

One more point is, Speculative execution will not stop the slow running task but it launch the new task in parallel.

Tabular Form :

Spark Property >> Default Value >> Description

spark.speculation >> false >> enables (true) or disables (false) speculative execution of tasks.

spark.speculation.interval >> 100ms >> The time interval to use before checking for speculative tasks.

spark.speculation.multiplier >> 1.5 >> How many times slower a task is than the median to be for speculation.

spark.speculation.quantile >> 0.75 >> The percentage of tasks that has not finished yet at which to start speculation.

37. How is fault tolerance achieved in Apache Spark?

The basic semantics of fault tolerance in Apache Spark is, all the Spark RDDs are immutable. It remembers the dependencies between every RDD involved in the operations, through the lineage graph created in the DAG, and in the event of any failure, Spark refers to the lineage graph to apply the same operations to perform the tasks.

There are two types of failures - Worker or driver failure. In case if the worker fails, the executors in that worker node will be killed, along with the data in their memory. Using the lineage graph, those tasks will be accomplished in any other worker nodes. The data is also replicated to other worker nodes to achieve fault tolerance. There are two cases:

- 1.Data received and replicated - Data is received from the source, and replicated across worker nodes. In the case of any failure, the data replication will help achieve fault tolerance.
- 2.Data received but not yet replicated - Data is received from the source but buffered for replication. In the case of any failure, the data needs to be retrieved from the source.

For stream inputs based on receivers, the fault tolerance is based on the type of receiver:

1.Reliable receiver - Once the data is received and replicated, an acknowledgment is sent to the source. In case if the receiver fails, the source will not receive acknowledgment for the received data. When the receiver is restarted, the source will resend the data to achieve fault tolerance.

2.Unreliable receiver - The received data will not be acknowledged to the source. In this case of any failure, the source will not know if the data has been received or not, and it will nor resend the data, so there is data loss.

To overcome this data loss scenario, Write Ahead Logging (WAL) has been introduced in Apache Spark 1.2. With WAL enabled, the intention of the operation is first noted down in a log file, such that if the driver fails and is restarted, the noted operations in that log file can be applied to the data. For sources that read streaming data, like Kafka or Flume, receivers will be receiving the data, and those will be stored in the executor's memory. With WAL enabled, these received data will also be stored in the log files.

WAL can be enabled by performing the below:

Setting the checkpoint directory, by using streamingContext.checkpoint(path)

Enabling the WAL logging, by setting spark.stream.receiver.WriteAheadLog.enable to True.

38.Explain the difference between reduceByKey, groupByKey, aggregateByKey and combineByKey?

1.groupByKey:

groupByKey can cause out of disk problems as data is sent over the network and collected on the reduce workers.

Example:-

```
sc.textFile("hdfs://").flatMap(line => line.split(" ") ).map(word => (word,1))
.groupByKey().map((x,y) => (x,sum(y)) )
```

2.reduceByKey:

Data is combined at each partition , only one output for one key at each partition to send over network. reduceByKey required combining all your values into another value with the exact same type.

Example:-

sc.textFile("hdfs://").flatMap(line => line.split(" ")).map(word => (word,1)) .reduceByKey((x,y)=> (x+y))

3.aggregateByKey:

same as reduceByKey, which takes an initial value.

3 parameters as input 1). initial value 2). Combiner logic function 3).merge Function

Example:-

val inp =Seq("dinesh=70","kumar=60","raja=40","ram=60","dinesh=50","dinesh=80","kumar=40"

,"raja=40")

val rdd=sc.parallelize(inp,3)

val pairRdd=rdd.map(_.split("=")).map(x=>(x(0),x(1)))

val initial_val=0

val addOp=(intVal:Int,StrVal: String)=> intVal+StrVal.toInt val mergeOp=(p1:Int,p2:Int)=>p1+p2

val out=pairRdd.aggregateByKey(initial_val)(addOp,mergeOp) out.collect.foreach(println)

4.combineByKey:

combineByKey values are merged into one value at each partition then each partition value is merged into a single value. It's worth noting that the type of the combined value does not have to match the type of the original value and often times it won't be.

3 parameters as input

1. create combiner

2. mergeValue

3. mergeCombiners

Example:

val inp = Array(("Dinesh", 98.0), ("Kumar", 86.0), ("Kumar", 81.0), ("Dinesh", 92.0), ("Dinesh", 83.0),

("Kumar", 88.0))

val rdd = sc.parallelize(inp,2)

//Create the combiner

```
val combiner = (inp:Double) => (1,inp)


//Function to merge the values within a partition.Add 1 to the # of entries and inp to the existing inp
val mergeValue = (PartVal:(Int,Double),inp:Double) =>{

(PartVal._1 + 1, PartVal._2 + inp)

}


//Function to merge across the partitions

val mergeCombiners = (PartOutput1:(Int, Double) , PartOutput2:(Int, Double))=>{

(PartOutput1._1+PartOutput2._1 , PartOutput1._2+PartOutput2._2)

}

//Function to calculate the average.Personinps is a custom type val CalculateAvg = (personinp:(String, (Int, Double)))=>{

val (name,(numofinps,inp)) = personinp

(name,inp/numofinps)

}


val rdd1=rdd.combineByKey(combiner, mergeValue, mergeCombiners)
rdd1.collect().foreach(println)

val rdd2=rdd.combineByKey(combiner, mergeValue, mergeCombiners).map( CalculateAvg)
rdd2.collect().foreach(println)
```

39. Explain the mapPartitions() and mapPartitionsWithIndex() ?

mapPartitions() and mapPartitionsWithIndex() are both transformation.

mapPartitions() :

It runs one at a time on each partition or block of the Rdd, so function must be of type iterator<T>. It improves performance by reducing creation of object in map function.

mapPartitions() can be used as an alternative to map() and foreach() .

mapPartitions() can be called for each partitions while map() and foreach() is called for each elements

in an RDD

Hence one can do the initialization on per-partition basis rather than each element basis

MappartionwithIndex():

It is similar to MapPartition but with one difference that it takes two parameters, the first parameter is the index and second is an iterator through all items within this partition (Int, Iterator<T>).

mapPartitionsWithIndex is similar to mapPartitions() but it provides second parameter index which keeps the track of partition.

40.Explain fold() operation in Spark?

fold() is an action. It is wide operation (i.e. shuffle data across multiple partitions and output a single

value)It takes function as an input which has two parameters of the same type and outputs a single value of the input type.

It is similar to reduce but has one more argument 'ZERO VALUE' (say initial value) which will be used in the initial call on each partition.

def fold(zeroValue: T)(op: (T, T) ⇒ T): T

Aggregate the elements of each partition, and then the results for all the partitions, using a given associative function and a neutral "zero value". The function op(t1, t2) is allowed to modify t1 and return it as its result value to avoid object allocation; however, it should not modify t2.

This behaves somewhat differently from fold operations implemented for non-distributed collections in functional languages like Scala. This fold operation may be applied to partitions individually, and then fold those results into the final result, rather than apply the fold to each element sequentially in some defined ordering. For

functions that are not commutative, the result may differ from that of a fold applied to a non-distributed

collection.

zeroValue: The initial value for the accumulated result of each partition for the op operator, and also the initial value for the combine results from different partitions for the op operator - this will typically be the neutral element (e.g. Nil for list concatenation or 0 for summation)

Op: an operator used to both accumulate results within a partition and combine results from different partitions

Example :

val rdd1 = sc.parallelize(List(1,2,3,4,5),3) rdd1.fold(5)(_+_)

Output : Int = 35

val rdd1 = sc.parallelize(List(1,2,3,4,5)) rdd1.fold(5)(_+_)

Output : Int = 25

val rdd1 = sc.parallelize(List(1,2,3,4,5),3) rdd1.fold(3)(_+_)

Output : Int = 27

41.Difference between textFile Vs wholeTextFile ?

Both are the method of org.apache.spark.SparkContext.

textFile() :

def textFile(path: String, minPartitions: Int = defaultMinPartitions): RDD[String]

Read a text file from HDFS, a local file system (available on all nodes), or any Hadoop-supported file system URI, and return it as an RDD of Strings

For example sc.textFile("/home/hdadmin/wc-data.txt") so it will create RDD in which each individual line an element.

Everyone knows the use of textFile.

wholeTextFiles() :

def wholeTextFiles(path: String, minPartitions: Int = defaultMinPartitions): RDD[(String, String)]

Read a directory of text files from HDFS, a local file system (available on all nodes), or any Hadoop-supported file system URI.Rather than create basic RDD, the wholeTextFile() returns pairRDD.

For example, you have few files in a directory so by using wholeTextFile() method,

it creates pair RDD with filename with path as key,and value being the whole file as string.

Example:-

val myfilerdd = sc.wholeTextFiles("/home/hdadmin/MyFiles") val keyrdd = myfilerdd.keys

keyrdd.collect

val filerdd = myfilerdd.values

filerdd.collect

42.What is cogroup() operation.?

> It's a transformation.

> It's in package org.apache.spark.rdd.PairRDDFunctions

def cogroup[W1, W2, W3](other1: RDD[(K, W1)], other2: RDD[(K, W2)], other3: RDD[(K, W3)]): RDD[(K, (Iterable[V], Iterable[W1], Iterable[W2], Iterable[W3]))]

For each key k in this or other1 or other2 or other3, return a resulting RDD that contains a tuple with the list of values for that key in this, other1, other2 and other3.

Example:

val myrdd1 = sc.parallelize(List((1,"spark"),(2,"HDFS"),(3,"Hive"),(4,"Flink"),(6,"HBase")))

val myrdd2= sc.parallelize(List((4,"RealTime"),(5,"Kafka"),(6,"NOSQL"),(1,"stream"),(1,"MLlib")))
val result = myrdd1.cogroup(myrdd2)

result.collect

Output :

Array[(Int, (Iterable[String], Iterable[String]))] =

Array((4,(CompactBuffer(Flink),CompactBuffer(RealTime))),

(1,(CompactBuffer(spark),CompactBuffer(stream, MLlib))),

(6,(CompactBuffer(HBase),CompactBuffer(NOSQL))),

(3,(CompactBuffer(Hive),CompactBuffer())),

(5,(CompactBuffer(),CompactBuffer(Kafka))),

(2,(CompactBuffer(HDFS),CompactBuffer())))

43.Explain pipe() operation ?

Return an RDD created by piping elements to a forked external process.

def pipe(command: String): RDD[String]

In general, Spark is using Scala, Java, and Python to write the program. However, if that is not enough, and one want to pipe (inject) the data which written in other languages like 'R', Spark provides general mechanism in the form of pipe() method

Spark provides the pipe() method on RDDs.

With Spark's pipe() method, one can write a transformation of an RDD that can read each element in the RDD from standard input as String.

It can write the results as String to the standard output.

Example:

test.py

#!/usr/bin/python

import sys

for line in sys.stdin:

print "hello " + line

spark-shell Scala:

val data = List("john","paul","george","ringo") val dataRDD = sc.makeRDD(data)

val scriptPath = "./test.py"

val pipeRDD = dataRDD.pipe(scriptPath) pipeRDD.foreach(println)

44.Explain coalesce() operation.?

It's in a package org.apache.spark.rdd.ShuffledRDD

def coalesce(numPartitions: Int, shuffle: Boolean = false, partitionCoalescer: Option[PartitionCoalescer] = Option.empty)(implicit ord: Ordering[(K, C)] = null): RDD[(K, C)]

Return a new RDD that is reduced into numPartitions partitions.

Example:

val myrdd1 = sc.parallelize(1 to 1000, 15) myrdd1.partitions.length

val myrdd2 = myrdd1.coalesce(5,false) myrdd2.partitions.length

Int = 5

45. Explain the repartition() operation?

> repartition() is a transformation.

> This function changes the number of partitions mentioned in parameter numPartitions(numPartitions : Int)

> It's in package org.apache.spark.rdd.ShuffledRDD

def repartition(numPartitions: Int)(implicit ord: Ordering[(K, C)] = null): RDD[(K, C)]

Return a new RDD that has exactly numPartitions partitions.

Can increase or decrease the level of parallelism in this RDD. Internally, this uses a shuffle to redistribute data.

If you are decreasing the number of partitions in this RDD, consider using coalesce, which can avoid performing a shuffle.

Example :

val rdd1 = sc.parallelize(1 to 100, 3) rdd1.getNumPartitions

val rdd2 = rdd1.repartition(6)

rdd2.getNumPartitions

46. Explain the top() and takeOrdered() operation.?

Both top() and takeOrdered() are actions.

Both returns then elements of RDD based on default ordering or based on custom ordering provided by user.

def top(num: Int)(implicit ord: Ordering[T]): Array[T]

Returns the top k (largest) elements from this RDD as defined by the specified implicit Ordering[T] and maintains the ordering. This does the opposite of takeOrdered.

def takeOrdered(num: Int)(implicit ord: Ordering[T]): Array[T]

Returns the first k (smallest) elements from this RDD as defined by the specified implicit Ordering[T] and maintains the ordering. This does the opposite of top.

Example :

val myrdd1 = sc.parallelize(List(5,7,9,13,51,89))

myrdd1.top(3) //Array[Int] = Array(89, 51, 13)

myrdd1.takeOrdered(3)//Array[Int] = Array(5, 7, 9)

myrdd1.top(3) //Array[Int] = Array(89, 51, 13)

47.Explain the lookup() operation.?

> It is an action

> It returns the list of values in the RDD for key 'key'

Example:

val rdd1 = sc.parallelize(Seq(("myspark",78),("Hive",95),("spark",15),("HBase",25),("spark",39), ("BigData",78),("spark",49)))

rdd1.lookup("spark")

rdd1.lookup("Hive")

rdd1.lookup("BigData")

Output:

Seq[Int] = WrappedArray(15, 39, 49)

Seq[Int] = WrappedArray(95)

Seq[Int] = WrappedArray(78)

48. How to Kill Spark Running Application.?

Get the application Id from the spark scheduler, for instance application_743159779306972_1234 and then, run the command in terminal like below yarn application -kill

application_743159779306972_1234

49.How to stop INFO messages displaying on spark console?

Edit spark conf/log4j.properties file and change the following line:

log4j.rootCategory=INFO, console

to

log4j.rootCategory=ERROR, console

for Spark 2.X please import below commands,

import org.apache.log4j.{Logger,Level}

Logger.getLogger("org").setLevel(Level.ERROR)

for Python: spark.sparkContext.setLogLevel("ERROR")

50.Where the logs are available in Spark on YARN?

we can access logs through the command

Syntax:-

yarn logs -applicationId <application ID>

51.How to find out the different values in between two spark dataframes.?

Simply we can achieve by except operation

Example:-

```
scala> customerDF.show
+---+--------+---+------+
|cId|    name|age|gender|
+---+--------+---+------+
|  1|   James| 21|     M|
|  2|     Liz| 25|     F|
|  3|    John| 31|     M|
|  4|Jennifer| 45|     F|
|  5|  Robert| 41|     M|
```

```
|          6|  Sandra| 45|     F|
+---+--------+---+------+

scala> custDF.show
+---+-----+---+------+
|cId| name|age|gender|
+---+-----+---+------+
|          1|James| 21|     M|
|          2|  Liz| 25|     F|
+---+-----+---+------+

scala> customerDF.except(custDF).show +---+--------+---+------+
|cId|    name|age|gender|
+---+--------+---+------+
|          5|  Robert| 41|     M|
|          3|    John| 31|     M|
|          4|Jennifer| 45|     F|
|          6|  Sandra| 45|     F|
+---+--------+---+------+
```

52.What are security options in Apache Spark?

Spark currently supports authentication via a shared secret. Authentication can be configured to be on via the spark.authenticate configuration parameter. This parameter controls whether the Spark communication protocols do authentication using the shared secret. This authentication is a basic handshake to make sure both sides have the same shared secret and are allowed to communicate. If the shared secret is not identical they will not be allowed to communicate. The shared secret is created as follows:

For Spark on YARN deployments, configuring spark.authenticate to true will automatically handle generating and distributing the shared secret. Each application will use a unique shared secret.

For other types of Spark deployments, the Spark parameter spark.authenticate.secret should be configured on each of the nodes. This secret will be used by all the Master/Workers and applications.

53. What is Scala?

Scala is a modern multi-paradigm programming language designed to express common programming patterns in a concise, elegant, and type-safe way. It smoothly integrates features of object-oriented and functional languages

54. What are the Types of Variable Expressions available in Scala?

val (aka Values):

You can name results of expressions with the val keyword.Once refer a value, it does not re-compute it.

Example:

val x = 1 + 1

x = 3 // This does not compile.

var (aka Variables):

Variables are like values, except you can re-assign them. You can define a variable with the var keyword.

Example:

var x = 1 + 1

x = 3 // This can compile.

55. What is the difference between method and functions in Scala..?

Methods:-

Methods look and behave very similar to functions, but there are a few key differences between them.

Methods are defined with the def keyword. def is followed by a name, parameter lists, a return type, and a body.

Example:

def add(x: Int, y: Int): Int = x + y println(add(1, 2)) // 3

Functions:-

Functions are expressions that take parameters.

Bigdata Hadoop: Spark Interview Questions with Answers

You can define an anonymous function (i.e. no name) that returns a given integer plus one: (x: Int) => x + 1

You can also name functions. like

val addOne = (x: Int) => x + 1

println(addOne(1)) // 2

56. What is case classes in Scala?

Scala has a special type of class called a “case” class. By default, case classes are immutable and compared by value. You can define case classes with the case class keywords.

Example:

case class Point(x: Int, y: Int)

val point = Point(1, 2)

val anotherPoint = Point(1, 2)

val yetAnotherPoint = Point(2, 2)

57. What is Traits in Scala?

Traits are used to share interfaces and fields between classes. They are similar to Java 8’s interfaces. Classes and objects can extend traits but traits cannot be instantiated and therefore have no parameters.

Traits are types containing certain fields and methods. Multiple traits can be combined.

A minimal trait is simply the keyword trait and an identifier:

Example:

trait Greeter { def greet(name: String): Unit = println("Hello, " + name + "!") }

58. What is singleton object in scala?

An object is a class that has exactly one instance is called singleton object. Here’s an example of a singleton object with a method:

object Logger {

def info(message: String): Unit = println("Hi i am Dineshkumar")

}

59. What is Companion objects in scala?

An object with the same name as a class is called a companion object. Conversely, the class is the object’s companion class. A companion class or object can access the private members of its companion. Use a companion object for methods and values which are not specific to instances of the companion class.

Example:

import scala.math._

case class Circle(radius: Double) { import Circle._

```
def area: Double = calculateArea(radius)
}

object Circle {
private def calculateArea(radius: Double): Double = Pi * pow(radius, 2.0)
}

val circle1 = new Circle(5.0)
circle1.area
```

60. What are the special datatype available in Scala?

Any:

Any is the supertype of all types, also called the top type. It defines certain universal methods such as equals, hashCode, and toString. Any has two direct subclasses: AnyVal and AnyRef.

Sample Example:

```
val list: List[Any] = List(
"a string",
732,  // an integer
'c',        // a character
true, // a boolean value
() => "an anonymous function returning a string"
)
list.foreach(element => println(element))
```

AnyVal:

AnyVal represents value types. There are nine predefined value types and they are non-nullable: Double, Float, Long, Int, Short, Byte, Char, Unit, and Boolean. Unit is a value type which carries no meaningful information. There is exactly one instance of Unit which can be declared literally like so: (). All functions must return something so sometimes Unit is a useful return type.

AnyRef:

AnyRef represents reference types. All non-value types are defined as reference types. Every user-defined type in Scala is a subtype of AnyRef. If Scala is used in the context of a Java runtime environment, AnyRef corresponds to java.lang.Object.

Nothing:

Nothing is a subtype of all types, also called the bottom type. There is no value that has type Nothing. A common use is to signal non-termination such as a thrown exception, program exit, or an infinite loop (i.e., it is the type of an expression which does not evaluate to a value, or a method that does not return normally).

Null:

Null is a subtype of all reference types (i.e. any subtype of AnyRef). It has a single value identified by the keyword literal null. Null is provided mostly for interoperability with other JVM languages and should almost never be used in Scala code. We'll cover alternatives to null later in the tour.

61. What is Higher order functions in Scala?

Higher order functions take other functions as parameters or return a function as a result. This is possible

because functions are first-class values in Scala. The terminology can get a bit confusing at this point, and we use

the phrase "higher order function" for both methods and functions that take functions as parameters or that

return a function.

In Higher Order function will make the features are, a Functions that accept another functions & A Functions that return to a functions to reduce redundant code.

One of the most common examples is the higher-order function map which is available for collections in Scala. val salaries = Seq(20000, 70000, 40000)

val doubleSalary = (x: Int) => x * 2

val newSalaries = salaries.map(doubleSalary) // List(40000, 140000, 80000)

62. What is Currying function or multiple parameter lists in Scala?

Methods may define multiple parameter lists. When a method is called with a fewer number of parameter lists,

then this will yield a function taking the missing parameter lists as its arguments. This is formally known as

currying.

Example:

def foldLeft[B](z: B)(op: (B, A) => B): B

val numbers = List(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

val res = numbers.foldLeft(0)((m, n) => m + n) print(res) // 55

63. What is Pattern matching in Scala?

Pattern matching is a mechanism for checking a value against a pattern. A successful match can also deconstruct a value into its constituent parts. It is a more powerful version of the switch statement in Java and it can likewise be used in place of a series of if/else statements.

Syntax:

import scala.util.Random

val x: Int = Random.nextInt(10)

x match {

case 0 => "zero"

case 1 => "one"

case 2 => "two"

case _ => "many"

}

def matchTest(x: Int): String = x match {

case 1 => "one"

case 2 => "two"

case _ => "many"

}

matchTest(3) // many

matchTest(1) // one

64. What are the basic properties avail in Spark?

It may be useful to provide some simple definitions for the Spark nomenclature:

Worker Node: A server that is part of the cluster and are available to run Spark jobs Master Node: The server that coordinates the Worker nodes.

Executor: A sort of virtual machine inside a node. One Node can have multiple Executors.

Driver Node: The Node that initiates the Spark session. Typically, this will be the server where context is located.

Driver (Executor): The Driver Node will also show up in the Executor list.

65. What are the Configuration properties in Spark?

spark.executor.memory:- The maximum possible is managed by the YARN cluster whichcannot exceed the actual RAM available.

spark.executor.cores:- Number of cores assigned per Executor which cannot be higher than the cores available in each worker.

spark.executor.instances:- Number of executors to start. This property is acknowledged by the cluster if spark.dynamicAllocation.enabled is set to “false”.

spark.memory.fraction:- The default is set to 60% of the requested memory per executor.

spark.dynamicAllocation.enabled:- Overrides the mechanism that Spark provides to dynamically adjust resources. Disabling it provides more control over the number of the Executors that can be started, which in turn impact the amount of storage available for the session. For more information, please see the Dynamic Resource Allocation page in the official Spark website.

66. What is Sealed classes?

Traits and classes can be marked sealed which means all subtypes must be declared in the same file.

This is useful for pattern matching because we don’t need a “catch all” case. This assures that all subtypes are known.

Example:

sealed abstract class Furniture

case class Couch() extends Furniture

case class Chair() extends Furniture

def findPlaceToSit(piece: Furniture): String = piece match { case a: Couch => "Lie on the couch"

case b: Chair => "Sit on the chair"

}

67. What is Type Inference?

The Scala compiler can often infer the type of an expression so you don’t have to declare it explicitly.

Example:

val Name = "Dineshkumar S" // it consider as String val id = 1234 // considered as int

68. When not to rely on default type inference?

The type inferred for obj was Null. Since the only value of that type is null, So it is impossible to assign a different value by default.

69. How can we debug spark application locally?

Actually we can doing that in local debugging, setting break points, inspecting variables, etc. set spark submission in deploy mode like below -

spark-submit --name CodeTestDinesh --class DineshMainClass --master local[2] DineshApps.jar

then spark driver to pause and wait for a connection from a debugger when it starts up, by adding an option like -

the following:

--conf spark.driver.extraJavaOptions=-

agentlib:jdwp=transport=dt_socket,server=y,suspend=y,address=5005

where agentlib:jdwp is the Java Debug Wire Protocol option, followed by a comma-separated list of sub-

options:

1. transport: defines the connection protocol used between debugger and debuggee -- either socket or "shared
2. memory" -- you almost always want socket (dt_socket) except I believe in some cases on Microsoft Windows
3. server: whether this process should be the server when talking to the debugger (or conversely, the client) -- you always need one server and one client. In this case, we're going to be the server and wait for a connection from the debugger
4. suspend: whether to pause execution until a debugger has successfully connected. We turn this on so the driver won't start until the debugger connects
5. address: here, this is the port to listen on (for incoming debugger connection requests). You can set it to any available port (you just have to make sure the debugger is configured to connect to this same port)

70. Map collection has Key and value then Key should be mutable or immutable?

Behavior of a Map is not specified if value of an object is changed in a manner that affects equals comparison while object with the key. So Key should be an immutable.

71. What is OFF_HEAP persistence in spark?

One of the most important capabilities in Spark is persisting (or caching) datasets in memory across operations. Each persisted RDD can be stored using a different storage level. One of the possibilities is to store RDDs in serialized format off-heap. Compared to storing data in the Spark JVM, off-heap storage reduces garbage collection overhead and allows executors to be smaller and to share a pool of memory. This makes it attractive in environments with large heaps or multiple concurrent applications.

72. What is the difference between Apache Spark and Apache Flink?

1.Stream Processing:

While Spark is a batch oriented system that operates on chunks of data, called RDDs, Apache Flink is a stream processing system able to process row after row in real time.

2.Iterations:

By exploiting its streaming architecture, Flink allows you to natively iterate over data, something Spark also supports only as batches

3.Memory Management:

Spark jobs have to be optimized and adapted to specific datasets because you need to manually control partitioning and caching if you want to get it right

4.Maturity:

Flink is still in its infancy and has but a few production deployments

5.Data Flow:

In contrast to the procedural programming paradigm Flink follows a distributed data flow approach. For data set operations where intermediate results are required in addition to the regular input of an operation, broadcast variables are used to distribute the pre calculated results to all worker nodes.

73. How do we Measuring the impact of Garbage Collection?

GC has happened due to use too much of memory on a driver or some executors or it might be where garbage collection becomes extremely costly and slow as large numbers of objects are created in the JVM. You can do by this validation '-verbose:gc -XX:+PrintGCDetails-XX:+PrintGCTimeStamps' to Spark's JVM options using the `spark.executor.extraJavaOptions` configuration parameter.

74. Apache Spark vs. Apache Storm?

Apache Spark is an in-memory distributed data analysis platform-- primarily targeted at speeding up batch analysis jobs, iterative machine learning jobs, interactive query and graph processing.

One of Spark's primary distinctions is its use of RDDs or Resilient Distributed Datasets. RDDs are great for pipelining parallel operators for computation and are, by definition, immutable, which allows Spark a unique form of fault tolerance based on lineage information. If you are interested in, for example, executing a Hadoop

MapReduce job much faster, Spark is a great option (although memory requirements must be considered).

Apache Storm is focused on stream processing or what some call complex event processing. Storm implements a fault tolerant method for performing a computation or pipelining multiple computations on an event as it flows into a system. One might use Storm to transform unstructured data as it flows into a system into a desired format.

Storm and Spark are focused on fairly different use cases. The more "apples-to-apples" comparison would be between Storm Trident and Spark Streaming. Since Spark's RDDs are inherently immutable, Spark Streaming implements a method for "batching" incoming updates in user-defined time intervals that get transformed into their own RDDs. Spark's parallel operators can then perform computations on these RDDs. This is different from Storm which deals with each event individually.

One key difference between these two technologies is that Spark performs Data-Parallel computations while Storm performs Task-Parallel computations. Either design makes trade offs that are worth knowing.

75. How to overwrite the output directory in spark?

refer below command using Dataframes,

df.write.mode(SaveMode.Overwrite).parquet(path)

76. How to read multiple text files into a single RDD?

You can specify whole directories, use wildcards and even CSV of directories and wildcards like below.

Eg.:

val rdd = sc.textFile("file:///D:/Dinesh.txt, file:///D:/Dineshnew.txt")

77. Can we run SPARK without base of HDFS?

Apache Spark is a fast and general-purpose cluster computing system. It is not a data storage system. It uses external storage system for storing and reading data. So we can run Spark without HDFS in distributed mode using any HDFS compatible file systems like S3, GPFS, GlusterFs, Cassandra, and etc.

There is another file system called Tachyon. It is a in memory file system to run spark in distributed mode.

78. Define about generic classes in scala?

Generic classes are classes which take a type as a parameter. They are particularly useful for collection classes.

Generic classes take a type as a parameter within square brackets []. One convention is to use the letter A as type parameter identifier, though any parameter name may be used.

Example: The instance stack can only take Int values. val stack = new Stack[Int]

stack.push(1)

stack.push(2)

println(stack.pop) // prints 2

println(stack.pop) // prints 1

79. How to enable tungsten sort shuffle in Spark 2.x?

SortShuffleManager is the one and only ShuffleManager in Spark with the short name sort or tungsten-sort.

In other words, there's no way you could use any other ShuffleManager but SortShuffleManager (unless you enabled one using spark.shuffle.manager property).

80. How to prevent Spark Executors from getting Lost when using YARN client mode?

The solution if you're using yarn was to set

--conf spark.yarn.executor.memoryOverhead=600,

alternatively if your cluster uses mesos you can try

--conf spark.mesos.executor.memoryOverhead=600 instead.

81. What is the relationship between the YARN Containers and the Spark Executors.?

First important thing is this fact that the number of containers will always be the same as the executors created by a Spark application e.g. via --num-executors parameter in spark-submit.

Set by the yarn.scheduler.minimum-allocation-mb every container always allocates at least this amount of memory. This means if parameter --executor-memory is set to e.g. only 1g but yarn.scheduler.minimum-allocation-mb is e.g. 6g, the container is much bigger than needed by the Spark application.

The other way round, if the parameter --executor-memory is set to somthing higher than the yarn.scheduler.minimum-allocation-mb value, e.g. 12g, the Container will allocate more memory dynamically, but only if the requested amount of memory is smaller or equal to yarn.scheduler.maximum-allocation-mb value.

The value of yarn.nodemanager.resource.memory-mb determines, how much memory can be allocated in sum by all containers of one host!

So setting yarn.scheduler.minimum-allocation-mb allows you to run smaller containers e.g. for smaller executors (else it would be waste of memory).

Setting yarn.scheduler.maximum-allocation-mb to the maximum value (e.g. equal to yarn.nodemanager.resource.memory-mb) allows you to define bigger executors (more memory is

allocated if needed, e.g. by --executor-memory parameter).

82. How to allocate the memory sizes for the spark jobs in cluster?

Before we are answering this question we have to concentrate the 3 main features.

which are –

1. NoOfExecutors,

2. Executor-memory and

3. Number of executor-cores

Lets go with example now, let's imagine we have a cluster with six nodes running NodeManagers, each with 16 cores and 64GB RAM. The NodeManager sizes, yarn.nodemanager.resource.memory-mb and yarn.nodemanager.resource.cpu-vcores, should be set to 63 * 1024 = 64512 (megabytes) and 15 respectively. We never provide 100% allocation of each resources to YARN containers because the node needs some resources to run the OS processes and Hadoop. In this case, we leave a gigabyte and a core for these system processes.

Cloudera Manager helps by accounting for these and configuring these YARN properties automatically. So the allocation likely matched as --num-executors 6 --executor-cores 15 --executor-memory 63G.

However, this is the wrong approach because: 63GB more on the executor memory overhead won't fit within the 63GB RAM of the NodeManagers. The application master will cover up a core on one of the nodes, meaning that there won't be room for a 15-core executor on that node. 15 cores per executor can lead to bad HDFS I/O throughput. So the best option would be to use --num-executors 17 --executor-cores 5 --executor-memory 19G.

This configuration results in three executors on all nodes except for the one with the Application Master, which

will have two executors. --executor-memory was derived as (63/3 executors per node) = 21. 21 * 0.07 = 1.47. 21 -

1.47 ~ 19.

83. How autocompletion tab can enable in pyspark.?

Please import the below libraries in pyspark shell

import rlcompleter, readline

readline.parse_and_bind("tab: complete")

84. can we execute two transformations on the same RDD in parallel in Apache Spark?

All standard RDD methods are blocking (with exception to AsyncRDDActions) so actions will be evaluated sequentially.

It is possible to execute multiple actions concurrently using non-blocking submission (threads, Futures) with correct configuration of in-application scheduler or explicitly limited resources for each action.

Example:

val df = spark.range(100000)

val df1= df.filter('id < 1000)

val df2= df.filter('id >= 1000)

print(df1.count() + df2.count()) //100000

Regarding cache it is impossible to answer without knowing the context. Depending on the cluster configuration, storage, and data locality it might be cheaper to load data from disk again, especially when resources are limited, and subsequent actions might trigger cache cleaner.

85. Which cluster type should I choose for Spark?

- Standalone - meaning Spark will manage its own cluster
- YARN - using Hadoop's YARN resource manager
- Mesos - Apache's dedicated resource manager project

Start with a standalone cluster if this is a new deployment. Standalone mode is the easiest to set up and will provide almost all the same features as the other cluster managers if you are only running Spark.

If you would like to run Spark alongside other applications, or to use richer resource scheduling capabilities (e.g. queues), both YARN and Mesos provide these features. Of these, YARN will likely be preinstalled in many

Hadoop distributions.

One advantage of Mesos over both YARN and standalone mode is its fine-grained sharing option, which lets interactive applications such as the Spark shell scale down their CPU allocation between commands. This makes it attractive in environments where multiple users are running interactive shells.

In all cases, it is best to run Spark on the same nodes as HDFS for fast access to storage. You can install Mesos or

the standalone cluster manager on the same nodes manually, or most Hadoop distributions already install YARN

and HDFS together.

86. What is DStreams in Spark Streaming?

Spark streaming uses a micro batch architecture where the incoming data is grouped into micro batches called Discretized Streams (DStreams) which also serves as the basic programming abstraction.

The DStreams internally have Resilient Distributed Datasets (RDD) and as a result of this standard RDD transformations and actions can be done.

87. What is Stateless Transformation .?

The processing of each batch has no dependency on the data of previous batches called Stateless

Transformation. Stateless transformations are simple RDD transformations. It applies on every batch meaning every RDD in a DStream. It includes common RDD transformations like map(), filter(), reduceByKey() etc.

88. What is Stateful Transformation .?

The uses data or intermediate results from previous batches and computes the result of the current

batch called Stateful Transformation. Stateful transformations are operations on DStreams that track data across time. Thus it makes use of some data from previous batches to generate the results for a new batch.

In streaming if we have a use case to track data across batches then we need state-ful DStreams.

For example we may track a user's interaction in a website during the user session or we may track a particular twitter hashtag across time and see which users across the globe is talking about it.

Types of state-ful transformation.

1.State-ful DStreams are of two types - window based tracking and full session tracking.

For stateful tracking all incoming data should be transformed to key-value pairs such that the key states can be tracked across batches. This is a precondition.

2.Window based tracking

In window based tracking the incoming batches are grouped in time intervals, i.e. group batches every 'x' seconds. Further computations on these batches are done using slide intervals.

Part – 6 TOP 250+ Interviews Questions on AWS

Q1) What is AWS?

Answer:AWS stands for Amazon Web Services. AWS is a platform that provides on-demand resources for hosting web services, storage, networking, databases and other resources over the internet with a pay-as-you-go pricing.

Q2) What are the components of AWS?

Answer:EC2 – Elastic Compute Cloud, S3 – Simple Storage Service, Route53, EBS – Elastic Block Store, Cloudwatch, Key-Paris are few of the components of AWS.

Q3) What are key-pairs?

Answer:Key-pairs are secure login information for your instances/virtual machines. To connect to the instances we use key-pairs that contain a public-key and private-key.

Q4) What is S3?

Answer:S3 stands for Simple Storage Service. It is a storage service that provides an interface that you can use to store any amount of data, at any time, from anywhere in the world. With S3 you pay only for what you use and the payment model is pay-as-you-go.

Q5) What are the pricing models for EC2instances?

Answer:The different pricing model for EC2 instances are as below,

- On-demand
- Reserved
- Spot
- Scheduled
- Dedicated

Q6) What are the types of volumes for EC2 instances?

Answer:

- There are two types of volumes,
- Instance store volumes
- EBS – Elastic Block Stores

Q7) What are EBS volumes?

Answer:EBS stands for Elastic Block Stores. They are persistent volumes that you can attach to the instances. With EBS volumes, your data will be preserved even when you stop your instances, unlike your instance store volumes where the data is deleted when you stop the instances.

Q8) What are the types of volumes in EBS?

Answer:Following are the types of volumes in EBS,

- General purpose
- Provisioned IOPS
- Magnetic

- Cold HDD
- Throughput optimized

Q9) What are the different types of instances?

Answer: Following are the types of instances,

- General purpose
- Computer Optimized
- Storage Optimized
- Memory Optimized
- Accelerated Computing

Q10) What is an auto-scaling and what are the components?

Answer: Auto scaling allows you to automatically scale-up and scale-down the number of instances depending on the CPU utilization or memory utilization. There are 2 components in Auto scaling, they are Auto-scaling groups and Launch Configuration.

Q11) What are reserved instances?

Answer: Reserved instances are the instance that you can reserve a fixed capacity of EC2 instances. In reserved instances you will have to get into a contract of 1 year or 3 years.

Q12)What is an AMI?

Answer: AMI stands for Amazon Machine Image. AMI is a template that contains the software configurations, launch permission and a block device mapping that specifies the volume to attach to the instance when it is launched.

Q13) What is an EIP?

Answer: EIP stands for Elastic IP address. It is designed for dynamic cloud computing. When you want to have a static IP address for your instances when you stop and restart your instances, you will be using EIP address.

Q14) What is Cloudwatch?

Answer: Cloudwatch is a monitoring tool that you can use to monitor your various AWS resources. Like health check, network, Application, etc.

Q15) What are the types in cloudwatch?

Answer: There are 2 types in cloudwatch. Basic monitoring and detailed monitoring. Basic monitoring is free and detailed monitoring is chargeable.

Q16) What are the cloudwatch metrics that are available for EC2 instances?

Answer: Diskreads, Diskwrites, CPU utilization, networkpacketsIn, networkpacketsOut, networkIn, networkOut, CPUCreditUsage, CPUCreditBalance.

Q17) What is the minimum and maximum size of individual objects that you can store in S3

Answer: The minimum size of individual objects that you can store in S3 is 0 bytes and the maximum bytes that you can store for individual objects is 5TB.

Q18) What are the different storage classes in S3?

Answer: Following are the types of storage classes in S3,

- Standard frequently accessed
- Standard infrequently accessed
- One-zone infrequently accessed.
- Glacier
- RRS – reduced redundancy storage

Q19) What is the default storage class in S3?

Answer: The default storage class in S3 in Standard frequently accessed.

Q20) What is glacier?

Answer: Glacier is the back up or archival tool that you use to back up your data in S3.

Q21) How can you secure the access to your S3 bucket?

Answer: There are two ways that you can control the access to your S3 buckets,

- ACL – Access Control List
- Bucket polices

Q22) How can you encrypt data in S3?

Answer: You can encrypt the data by using the below methods,

- Server Side Encryption – S3 (AES 256 encryption)
- Server Side Encryption – KMS (Key management Service)
- Server Side Encryption – C (Client Side)

Q23) What are the parameters for S3 pricing?

Answer: The pricing model for S3 is as below,

- Storage used
- Number of requests you make
- Storage management
- Data transfer
- Transfer acceleration

Q24) What is the pre-requisite to work with Cross region replication in S3?

Answer: You need to enable versioning on both source bucket and destination to work with cross region replication. Also both the source and destination bucket should be in different region.

Q25) What are roles?

Answer: Roles are used to provide permissions to entities that you trust within your AWS account. Roles are users in another account. Roles are similar to users but with roles you do not need to create any username and password to work with the resources.

Q26) What are policies and what are the types of policies?

Answer: Policies are permissions that you can attach to the users that you create. These policies will contain that access that you have provided to the users that you have created. There are 2 types of policies.

- Managed policies
- Inline policies

Q27) What is cloudfront?

Answer: Cloudfront is an AWS web service that provided businesses and application developers an easy and efficient way to distribute their content with low latency and high data transfer speeds. Cloudfront is content delivery network of AWS.

Q28) What are edge locations?

Answer: Edge location is the place where the contents will be cached. When a user tries to access some content, the content will be searched in the edge location. If it is not available then the content will be made available from the origin location and a copy will be stored in the edge location.

Q29) What is the maximum individual archive that you can store in glacier?

Answer: You can store a maximum individual archive of upto 40 TB.

Q30) What is VPC?

Answer: VPC stands for Virtual Private Cloud. VPC allows you to easily customize your networking configuration. VPC is a network that is logically isolated from other network in the cloud. It allows you to have your own IP address range, subnets, internet gateways, NAT gateways and security groups.

Q31) What is VPC peering connection?

Answer: VPC peering connection allows you to connect 1 VPC with another VPC. Instances in these VPC behave as if they are in the same network.

Q32) What are NAT gateways?

Answer: NAT stands for Network Address Translation. NAT gateways enables instances in a private subnet to connect to the internet but prevent the internet from initiating a connection with those instances.

Q33) How can you control the security to your VPC?

Answer: You can use security groups and NACL (Network Access Control List) to control the security to your

VPC.

Q34) What are the different types of storage gateway?

Answer: Following are the types of storage gateway.

- File gateway
- Volume gateway
- Tape gateway

Q35) What is a snowball?

Answer: Snowball is a data transport solution that used source appliances to transfer large amounts of data into and out of AWS. Using snowball, you can move huge amount of data from one place to another which reduces your network costs, long transfer times and also provides better security.

Q36) What are the database types in RDS?

Answer: Following are the types of databases in RDS,

- Aurora
- Oracle
- MYSQL server
- Postgresql
- MariaDB
- SQL server

Q37) What is a redshift?

Answer: Amazon redshift is a data warehouse product. It is a fast and powerful, fully managed, petabyte scale data warehouse service in the cloud.

Q38) What is SNS?

Answer: SNS stands for Simple Notification Service. SNS is a web service that makes it easy to notifications from the cloud. You can set up SNS to receive email notification or message notification.

Q39) What are the types of routing polices in route53?

Answer: Following are the types of routing policies in route53,

- Simple routing
- Latency routing
- Failover routing
- Geolocation routing
- Weighted routing
- Multivalue answer

Q40) What is the maximum size of messages in SQS?

Answer: The maximum size of messages in SQS is 256 KB.

Q41) What are the types of queues in SQS?

Answer: There are 2 types of queues in SQS.

- Standard queue
- FIFO (First In First Out)

Q42) What is multi-AZ RDS?

Answer: Multi-AZ (Availability Zone) RDS allows you to have a replica of your production database in another availability zone. Multi-AZ (Availability Zone) database is used for disaster recovery. You will have an exact copy of your database. So when your primary database goes down, your application will automatically failover to the standby database.

Q43) What are the types of backups in RDS database?

Answer: There are 2 types of backups in RDS database.

- Automated backups
- Manual backups which are known as snapshots.

Q44) What is the difference between security groups and network access control list?

Answer:

Security Groups	Network access control list
Can control the access at the instance level	Can control access at the subnet level
Can add rules for “allow” only	Can add rules for both “allow” and “deny”
Evaluates all rules before allowing the traffic	Rules are processed in order number when allowing traffic.
Can assign unlimited number of security groups	Can assign upto 5 security groups.
Statefull filtering	Stateless filtering

Q45) What are the types of load balancers in EC2?

Answer: There are 3 types of load balancers,

- Application load balancer
- Network load balancer
- Classic load balancer

Q46) What is and ELB?

Answer: ELB stands for Elastic Load balancing. ELB automatically distributes the incoming application traffic or network traffic across multiple targets like EC2, containers, IP addresses.

Q47) What are the two types of access that you can provide when you are creating users?

Answer: Following are the two types of access that you can create.

- Programmatic access
- Console access

Q48) What are the benefits of auto scaling?

Answer: Following are the benefits of auto scaling

- Better fault tolerance
- Better availability
- Better cost management

Q49) What are security groups?

Answer: Security groups acts as a firewall that contains the traffic for one or more instances. You can associate one or more security groups to your instances when you launch then. You can add rules to each security group that allow traffic to and from its associated instances. You can modify the rules of a security group at any time, the new rules are automatically and immediately applied to all the instances that are associated with the security group

Q50) What are shared AMI's?

Answer: Shared AMI's are the AMI that are created by other developed and made available for other developed to use.

Q51)What is the difference between the classic load balancer and application load balancer?

Answer: Dynamic port mapping, multiple port multiple listeners is used in Application Load Balancer, One port one listener is achieved via Classic Load Balancer

Q52) By default how many Ip address does aws reserve in a subnet?

Answer: 5

Q53) What is meant by subnet?

Answer: A large section of IP Address divided in to chunks are known as subnets

Q54) How can you convert a public subnet to private subnet?

Answer: Remove IGW & add NAT Gateway, Associate subnet in Private route table

Q55) Is it possible to reduce a ebs volume?

Answer: no it's not possible, we can increase it but not reduce them

Q56) What is the use of elastic ip are they charged by AWS?

Answer: These are ipv4 address which are used to connect the instance from internet, they are charged if the instances are not attached to it

Q57) One of my s3 is bucket is deleted but i need to restore is there any possible way?

Answer: If versioning is enabled we can easily restore them

Q58) When I try to launch an ec2 instance i am getting Service limit exceed, how to fix the issue?

Answer: By default AWS offer service limit of 20 running instances per region, to fix the issue we need to contact AWS support to increase the limit based on the requirement

Q59) I need to modify the ebs volumes in Linux and windows is it possible

Answer: yes its possible from console use modify volumes in section give the size u need then for windows go to disk management for Linux mount it to achieve the modification

Q60) Is it possible to stop a RDS instance, how can I do that?

Answer: Yes it's possible to stop rds. Instance which are non-production and non multi AZ's

Q61) What is meant by parameter groups in rds. And what is the use of it?

Answer: Since RDS is a managed service AWS offers a wide set of parameter in RDS as parameter group which is modified as per requirement

Q62) What is the use of tags and how they are useful?

Answer: Tags are used for identification and grouping AWS Resources

Q63) I am viewing an AWS Console but unable to launch the instance, I receive an IAM Error how can I rectify it?

Answer: As AWS user I don't have access to use it, I need to have permissions to use it further

Q64) I don't want my AWS Account id to be exposed to users how can I avoid it?

Answer: In IAM console there is option as sign in url where I can rename my own account name with AWS account

Q65) By default how many Elastic Ip address does AWS Offer?

Answer: 5 elastic ip per region

Q66) You are enabled sticky session with ELB. What does it do with your instance?

Answer: Binds the user session with a specific instance

Q67) Which type of load balancer makes routing decisions at either the transport layer or the Application layer and supports either EC2 or VPC.

Answer: Classic Load Balancer

Q68) Which is virtual network interface that you can attach to an instance in a VPC?

Answer: Elastic Network Interface

Q69) You have launched a Linux instance in AWS EC2. While configuring security group, you Have selected SSH, HTTP, HTTPS protocol. Why do we need to select SSH?

Answer: To verify that there is a rule that allows traffic from EC2 Instance to your computer

Q70) You have chosen a windows instance with Classic and you want to make some change to the

Security group. How will these changes be effective?

Answer: Changes are automatically applied to windows instances

Q71) Load Balancer and DNS service comes under which type of cloud service?

Answer: IAAS-Storage

Q72) You have an EC2 instance that has an unencrypted volume. You want to create another

Encrypted volume from this unencrypted volume. Which of the following steps can achieve this?

Answer: Create a snapshot of the unencrypted volume (applying encryption parameters), copy the. Snapshot and create a volume from the copied snapshot

Q73) Where does the user specify the maximum number of instances with the auto scaling Commands?

Answer: Auto scaling Launch Config

Q74) Which are the types of AMI provided by AWS?

Answer: Instance Store backed, EBS Backed

Q75) After configuring ELB, you need to ensure that the user requests are always attached to a Single instance. What setting can you use?

Answer: Sticky session

Q76) When do I prefer to Provisioned IOPS over the Standard RDS storage?

Answer:If you have do batch-oriented is workloads.

Q77) If I am running on my DB Instance a Multi-AZ deployments, can I use to the stand by the DB Instance for read or write a operation along with to primary DB instance?

Answer: Primary db instance does not working.

Q78) Which the AWS services will you use to the collect and the process e-commerce data for the near by real-time analysis?

Answer: Good of Amazon DynamoDB.

Q79) A company is deploying the new two-tier an web application in AWS. The company has to limited on staff and the requires high availability, and the application requires to complex

queries and table joins. Which configuration provides to the solution for company's requirements?

Answer: An web application provide on Amazon DynamoDB solution.

Q80) Which the statement use to cases are suitable for Amazon DynamoDB?

Answer:The storing metadata for the Amazon S3 objects& The Running of relational joins and complex an updates.

Q81) Your application has to the retrieve on data from your user's mobile take every 5 minutes and then data is stored in the DynamoDB, later every day at the particular time the data is an extracted into S3 on a per user basis and then your application is later on used to visualize the data to user. You are the asked to the optimize the architecture of the backend system can to lower cost, what would you recommend do?

Answer: Introduce Amazon Elasticache to the cache reads from the Amazon DynamoDB table and to reduce the provisioned read throughput.

Q82) You are running to website on EC2 instances can deployed across multiple Availability Zones with an Multi-AZ RDS MySQL Extra Large DB Instance etc. Then site performs a high number of the small reads and the write per second and the relies on the eventual consistency model. After the comprehensive tests you discover to that there is read contention on RDS MySQL. Which is the best approaches to the meet these requirements?

Answer:The Deploy Elasti Cache in-memory cache is running in each availability zone and Then Increase the RDS MySQL Instance size and the Implement provisioned IOPS.

Q83) An startup is running to a pilot deployment of around 100 sensors to the measure street noise and The air quality is urban areas for the 3 months. It was noted that every month to around the 4GB of sensor data are generated. The company uses to a load balanced take auto scaled layer of the EC2 instances and a RDS database with a 500 GB standard storage. The pilot was success and now they want to the deploy take atleast 100K sensors.let which to need the supported by backend. You need to the stored data for at least 2 years to an analyze it. Which setup of following would you be prefer?

Answer: The Replace the RDS instance with an 6 node Redshift cluster with take 96TB of storage.

Q84) Let to Suppose you have an application where do you have to render images and also do some of general computing. which service will be best fit your need?

Answer:Used on Application Load Balancer.

Q85) How will change the instance give type for the instances, which are the running in your applications tier and Then using Auto Scaling. Where will you change it from areas?

Answer: Changed to Auto Scaling launch configuration areas.

Q86) You have an content management system running on the Amazon EC2 instance that is the approaching 100% CPU of utilization. Which option will be reduce load on the Amazon EC2 instance?

Answer: Let Create a load balancer, and Give register the Amazon EC2 instance with it.

Q87) What does the Connection of draining do?

Answer: The re-routes traffic from the instances which are to be updated (or) failed an health to check.

Q88) When the instance is an unhealthy, it is do terminated and replaced with an new ones, which of the services does that?

Answer: The survice make a fault tolerance.

Q89) What are the life cycle to hooks used for the AutoScaling?

Answer: They are used to the put an additional taken wait time to the scale in or scale out events.

Q90) An user has to setup an Auto Scaling group. Due to some issue the group has to failed for launch a single instance for the more than 24 hours. What will be happen to the Auto Scaling in the condition?

Answer: The auto Scaling will be suspend to the scaling process.

Q91) You have an the EC2 Security Group with a several running to EC2 instances. You changed to the Security of Group rules to allow the inbound traffic on a new port and protocol, and then the launched a several new instances in the same of Security Group.Such the new rules apply?

Answer:The Immediately to all the instances in security groups.

Q92) To create an mirror make a image of your environment in another region for the disaster recoverys, which of the following AWS is resources do not need to be recreated in second region?

Answer: May be the selected on Route 53 Record Sets.

Q93) An customers wants to the captures all client connections to get information from his load balancers at an interval of 5 minutes only, which cal select option should he choose for his application?

Answer: The condition should be Enable to AWS CloudTrail for the loadbalancers.

Q94) Which of the services to you would not use to deploy an app?

Answer: Lambda app not used on deploy.

Q95) How do the Elastic Beanstalk can apply to updates?

Answer: By a duplicate ready with a updates prepare before swapping.

Q96) An created a key in the oregon region to encrypt of my data in North Virginia region for security purposes. I added to two users to the key and the external AWS accounts. I wanted to encrypt an the object in S3, so when I was tried, then key that I just created is not listed.What could be reason&solution?

Answer:The Key should be working in the same region.

Q97) As a company needs to monitor a read and write IOPS for the AWS MySQL RDS instances and then send real-time alerts to the operations of team. Which AWS services to can accomplish this?

Answer:The monitoring on Amazon CloudWatch

Q98) The organization that is currently using the consolidated billing has to recently acquired to another company that already has a number of the AWS accounts. How could an Administrator to ensure that all the AWS accounts, from the both existing company and then acquired company, is billed to the single account?

Answer: All Invites take acquired the company's AWS account to join existing the company's of organization by using AWS Organizations.

Q99) The user has created an the applications, which will be hosted on the EC2. The application makes calls to the Dynamo DB to fetch on certain data. The application using the DynamoDB SDK to connect with the EC2 instance. Which of respect to best practice for the security in this scenario?

Answer: The user should be attach an IAM roles with the DynamoDB access to EC2 instance.

Q100) You have an application are running on EC2 Instance, which will allow users to download the files from a private S3 bucket using the pre-assigned URL. Before generating to URL the Q101) application should be verify the existence of file in S3. How do the application use the AWS credentials to access S3 bucket securely?

Answer:An Create an IAM role for the EC2 that allows list access to objects in S3 buckets. Launch to instance with this role, and retrieve an role's credentials from EC2 Instance make metadata.

Q101) You use the Amazon CloudWatch as your primary monitoring system for web application. After a recent to software deployment, your users are to getting Intermittent the 500 Internal Server to the Errors, when you using web application. You want to create the CloudWatch alarm, and notify the on-call engineer let when these occur. How can you accomplish the using the AWS services?

Answer: An Create a CloudWatch get Logs to group and A define metric filters that assure capture 500 Internal Servers should be Errors. Set a CloudWatch alarm on the metric and By Use of Amazon Simple to create a Notification Service to notify an the on-call engineers when prepare CloudWatch alarm is triggered.

Q102) You are designing a multi-platform of web application for the AWS. The application will run on the EC2 instances and Till will be accessed from PCs, tablets and smart phones.Then Supported accessing a platforms are Windows, MACOS, IOS and Android. They Separate sticky sessions and SSL certificate took setups are required for the different platform types. Which do describes the most cost effective and Like performance efficient the architecture setup?

Answer:Assign to multiple ELBs an EC2 instance or group of EC2 take instances running to common component of the web application, one ELB change for each platform type.Take Session will be stickiness and SSL termination are done for the ELBs.

Q103) You are migrating to legacy client-server application for AWS. The application responds to a specific DNS visible domain (e.g. www.example.com) and server 2-tier architecture, with

multiple application for the servers and the database server. Remote clients use to TCP to connect to the application of servers. The application servers need to know the IP address of clients in order to the function of properly and are currently taking of that information from TCP socket. A Multi-AZ RDS MySQL instance to will be used for database. During the migration you change the application code but you have file a change request. How do would you implement the architecture on the AWS in order to maximize scalability and high availability?

Answer: File a change request to get implement of Proxy Protocol support in the application. Use of ELB with TCP Listener and A Proxy Protocol enabled to distribute the load on two application servers in the different AZs.

Q104) Your application currently is leverages AWS Auto Scaling to the grow and shrink as a load Increases/decreases and has been performing as well. Your marketing a team expects and steady ramp up in traffic to follow an upcoming campaign that will result in 20x growth in the traffic over 4 weeks. Your forecast for approximate number of the Amazon EC2 instances necessary to meet peak demand is 175. What should be you do avoid potential service disruptions during the ramp up traffic?

Answer: Check the service limits in the Trusted Advisors and adjust as necessary, so that forecasted count remains within the limits.

Q105) You have a web application running on the six Amazon EC2 instances, consuming about 45% of resources on the each instance. You are using the auto-scaling to make sure that a six instances are running at all times. The number of requests this application processes to consistent and does not experience to spikes. Then application are critical to your business and you want to high availability for at all times. You want to the load be distributed evenly has between all instances. You also want to between use same Amazon Machine Image (AMI) for all instances. Which are architectural choices should you make?

Answer: Deploy to 3 EC2 instances in one of availability zone and 3 in another availability of zones and to use of Amazon Elastic is Load Balancer.

Q106) You are the designing an application that a contains protected health information. Security and Then compliance requirements for your application mandate that all protected to health information in application use to encryption at rest and in the transit module. The application to uses an three-tier architecture. where should data flows through the load balancers and is stored on the Amazon EBS volumes for the processing, and the results are stored in the Amazon S3 using a AWS SDK. Which of the options satisfy the security requirements?

Answer: Use TCP load balancing on load balancer system, SSL termination on Amazon to create EC2 instances, OS-level disk take encryption on Amazon EBS volumes, and The amazon S3 with server-side to encryption and Use the SSL termination on load balancers, an SSL listener on the Amazon to create EC2 instances, Amazon EBS encryption on the EBS volumes containing the PHI, and Amazon S3 with a server-side of encryption.

Q107) An startup deploys its create photo-sharing site in a VPC. An elastic load balancer distributes to web traffic across two the subnets. Then the load balancer session to stickiness is configured to use of AWS-generated session cookie, with a session TTL of the 5 minutes. The web server to change Auto Scaling group is configured as like min-size=4, max-size=4. The startup is the preparing for a public launchs, by running the load-testing software installed on the single Amazon Elastic Compute Cloud (EC2) instance to running in us-west-2a. After 60 minutes of load-testing, the web server logs of show the following:WEBSERVER LOGS | # of

HTTP requests to from load-tester system | # of HTTP requests to from private on beta users || webserver #1 (subnet an us-west-2a): | 19,210 | 434 | webserver #2 (subnet an us-west-2a): | 21,790 | 490 || webserver #3 (subnet an us-west-2b): | 0 | 410 || webserver #4 (subnet an us-west-2b): | 0 | 428 |Which as recommendations can be help of ensure that load-testing HTTP requests are will evenly distributed across to four web servers?

Answer:Result of cloud is re-configure the load-testing software to the re-resolve DNS for each web request.

Q108) To serve the Web traffic for a popular product to your chief financial officer and IT director have purchased 10 m1.large heavy utilization of Reserved Instances (RIs) evenly put spread across two availability zones: Route 53 are used to deliver the traffic to on Elastic Load Balancer (ELB). After the several months, the product grows to even more popular and you need to additional capacity As a result, your company that purchases two c3.2xlarge medium utilization RIs You take register the two c3.2xlarge instances on with your ELB and quickly find that the ml of large instances at 100% of capacity and the c3.2xlarge instances have significant to capacity that's can unused Which option is the most of cost effective and uses EC2 capacity most of effectively?

Answer: To use a separate ELB for the each instance type and the distribute load to ELBs with a Route 53 weighted round of robin.

Q109) An AWS customer are deploying an web application that is the composed of a front-end running on the Amazon EC2 and confidential data that are stored on the Amazon S3. The customer security policy is that all accessing operations to this sensitive data must authenticated and authorized by centralized access to management system that is operated by separate security team. In addition, the web application team that be owns and administers the EC2 web front-end instances are prohibited from having the any ability to access data that circumvents this centralized access to management system. Which are configurations will support these requirements?

Answer:The configure to the web application get authenticate end-users against the centralized access on the management system. Have a web application provision trusted to users STS tokens an entitling the download of the approved data directly from a Amazon S3.

Q110) A Enterprise customer is starting on their migration to the cloud, their main reason for the migrating is agility and they want to the make their internal Microsoft active directory available to the many applications running on AWS, this is so internal users for only have to remember one set of the credentials and as a central point of user take control for the leavers and joiners. How could they make their actions the directory secures and the highly available with minimal on-premises on infrastructure changes in the most cost and the time-efficient way?

Answer: By Using a VPC, they could be create an the extension to their data center and to make use of resilient hardware IPSEC on tunnels, they could then have two domain consider to controller instances that are joined to the existing domain and reside within the different subnets in the different availability zones.

Q111) What is Cloud Computing?

Answer:Cloud computing means it provides services to access programs, application, storage, network, server over the internet through browser or client side application on your PC, Laptop, Mobile by the end user without installing, updating and maintaining them.

Q112) Why we go for Cloud Computing?

Answer:

- Lower computing cost
- Improved Performance
- No IT Maintenance
- Business connectivity
- Easily upgraded
- Device Independent

Q113) What are the deployment models using in Cloud?

Answer:

- Private Cloud
- Public Cloud
- Hybrid cloud
- Community cloud 4

Q114) Explain Cloud Service Models?

Answer: SAAS (Software as a Service): It is software distribution model in which application are hosted by a vendor over the internet for the end user freeing from complex software and hardware management. (Ex: Google drive, drop box)

PAAS (Platform as a Service): It provides platform and environment to allow developers to build applications. It frees developers without going into the complexity of building and maintaining the infrastructure. (Ex: AWS Elastic Beanstalk, Windows Azure)

IAAS (Infrastructure as a Service): It provides virtualized computing resources over the internet like cpu, memory, switches, routers, firewall, Dns, Load balancer (Ex: Azure, AWS)

Q115) What are the advantage of Cloud Computing?

Answer:

- Pay per use
- Scalability
- Elasticity
- High Availability
- Increase speed and Agility
- Go global in Minutes

Q116) What is AWS?

Answer: Amazon web service is a secure cloud services platform offering compute, power, database, storage, content delivery and other functionality to help business scale and grow.

AWS is fully on-demand

AWS is Flexibility, availability and Scalability

AWS is Elasticity: scale up and scale down as needed.

Q117) What is mean by Region, Availability Zone and Edge Location?

Answer: Region: An independent collection of AWS resources in a defined geography. A collection of Data centers (Availability zones). All availability zones in a region connected by high bandwidth.

Availability Zones: An Availability zone is a simply a data center. Designed as independent failure zone. High speed connectivity, Low latency.

Edge Locations: Edge location are the important part of AWS Infrastructure. Edge locations are CDN endpoints for cloud front to deliver content to end user with low latency

Q118) How to access AWS Platform?

Answer:

- AWS Console
- AWS CLI (Command line interface)
- AWS SDK (Software Development Kit)

Q119) What is EC2? What are the benefits in EC2?

Amazon Elastic compute cloud is a web service that provides resizable compute capacity in the cloud.AWS EC2 provides scalable computing capacity in the AWS Cloud. These are the virtual servers also called as an instances. We can use the instances pay per use basis.

Benefits:

- Easier and Faster
- Elastic and Scalable
- High Availability
- Cost-Effective

Q120) What are the pricing models available in AWS EC2?

Answer:

- On-Demand Instances
- Reserved Instances
- Spot Instances
- Dedicated Host

Q121) What are the types using in AWS EC2?

Answer:

- General Purpose
- Compute Optimized
- Memory optimized
- Storage Optimized
- Accelerated Computing (GPU Based)

Q122) What is AMI? What are the types in AMI?

Answer:

Amazon machine image is a special type of virtual appliance that is used to create a virtual machine within the amazon Elastic compute cloud. AMI defines the initial software that will be in an instance when it is launched.

Types of AMI:

- Published by AWS
- AWS Marketplace
- Generated from existing instances
- Uploaded virtual server

Q123) How to Addressing AWS EC2 instances?

Answer:

- Public Domain name system (DNS) name: When you launch an instance AWS creates a DNS name that can be used to access the
- Public IP: A launched instance may also have a public ip address This IP address assigned from the address reserved by AWS and cannot be specified.
- Elastic IP: An Elastic IP Address is an address unique on the internet that you reserve independently and associate with Amazon EC2 instance. This IP Address persists until the customer release it and is not tried to

Q124) What is Security Group?

Answer: AWS allows you to control traffic in and out of your instance through virtual firewall called Security groups. Security groups allow you to control traffic based on port, protocol and source/Destination.

Q125) When your instance show retired state?

Answer:Retired state only available in Reserved instances. Once the reserved instance reserving time (1 yr/3 yr) ends it shows Retired state.

Q126) Scenario: My EC2 instance IP address change automatically while instance stop and start. What is the reason for that and explain solution?

Answer:AWS assigned Public IP automatically but it's change dynamically while stop and start. In that case we need to assign Elastic IP for that instance, once assigned it doesn't change automatically.

Q127) What is Elastic Beanstalk?

Answer:AWS Elastic Beanstalk is the fastest and simplest way to get an application up and running on AWS.Developers can simply upload their code and the service automatically handle all the details such as resource provisioning, load balancing, Auto scaling and Monitoring.

Q128) What is Amazon Lightsail?

Answer:Lightsail designed to be the easiest way to launch and manage a virtual private server with AWS.Lightsail plans include everything you need to jumpstart your project a virtual machine, ssd based storage, data transfer, DNS Management and a static ip.

Q129) What is EBS?

Answer:Amazon EBS Provides persistent block level storage volumes for use with Amazon EC2 instances. Amazon EBS volume is automatically replicated with its availability zone to protect component failure offering high availability and durability. Amazon EBS volumes are available in a variety of types that differ in performance characteristics and Price.

Q130) How to compare EBS Volumes?

Answer: Magnetic Volume: Magnetic volumes have the lowest performance characteristics of all Amazon EBS volume types.

EBS Volume size: 1 GB to 1 TB Average IOPS: 100 IOPS Maximum throughput: 40-90 MB

General-Purpose SSD: General purpose SSD volumes offers cost-effective storage that is ideal for a broad range of workloads. General purpose SSD volumes are billed based on the amount of data space provisioned regardless of how much of data you actually store on the volume.

EBS Volume size: 1 GB to 16 TB Maximum IOPS: upto 10000 IOPS Maximum throughput: 160 MB

Provisioned IOPS SSD: Provisioned IOPS SSD volumes are designed to meet the needs of I/O intensive workloads, particularly database workloads that are sensitive to storage performance and consistency in random access I/O throughput. Provisioned IOPS SSD Volumes provide predictable, High performance.

EBS Volume size: 4 GB to 16 TB Maximum IOPS: upto 20000 IOPS Maximum throughput: 320 MB

Q131) What is cold HDD and Throughput-optimized HDD?

Answer: Cold HDD: Cold HDD volumes are designed for less frequently accessed workloads. These volumes are significantly less expensive than throughput-optimized HDD volumes.

EBS Volume size: 500 GB to 16 TB Maximum IOPS: 200 IOPS Maximum throughput: 250 MB

Throughput-Optimized HDD: Throughput-optimized HDD volumes are low cost HDD volumes designed for frequent access, throughput-intensive workloads such as big data, data warehouse.

EBS Volume size: 500 GB to 16 TB Maximum IOPS: 500 IOPS Maximum throughput: 500 MB

Q132) What is Amazon EBS-Optimized instances?

Answer: Amazon EBS optimized instances to ensure that the Amazon EC2 instance is prepared to take advantage of the I/O of the Amazon EBS Volume. An amazon EBS-optimized instance uses an optimized configuration stack and provide additional dedicated capacity for Amazon EBS I/When you select Amazon EBS-optimized for an instance you pay an additional hourly charge for that instance.

Q133) What is EBS Snapshot?

Answer:

- It can back up the data on the EBS Volume. Snapshots are incremental backups.
- If this is your first snapshot it may take some time to create. Snapshots are point in time copies of volumes.

Q134) How to connect EBS volume to multiple instance?

Answer: We can't able to connect EBS volume to multiple instance, but we can able to connect multiple EBS Volume to single instance.

Q135) What are the virtualization types available in AWS?

Answer: Hardware assisted Virtualization: HVM instances are presented with a fully virtualized set of hardware and they executing boot by executing master boot record of the root block device of the image. It is default Virtualization.

Para virtualization: This AMI boot with a special boot loader called PV-GRUB. The ability of the guest kernel to communicate directly with the hypervisor results in greater performance levels than other virtualization approaches but they cannot take advantage of hardware extensions such as networking, GPU etc. Its customized Virtualization image. Virtualization image can be used only for particular service.

Q136) Differentiate Block storage and File storage?

Answer:

Block Storage: Block storage operates at lower level, raw storage device level and manages data as a set of numbered, fixed size blocks.

File Storage: File storage operates at a higher level, the operating system level and manage data as a named hierarchy of files and folders.

Q137) What are the advantage and disadvantage of EFS? Advantages:

Answer:

- Fully managed service
- File system grows and shrinks automatically to petabytes
- Can support thousands of concurrent connections
- Multi AZ replication
- Throughput scales automatically to ensure consistent low latency Disadvantages:
- Not available in all region
- Cross region capability not available
- More complicated to provision compared to S3 and EBS

Q138) what are the things we need to remember while creating s3 bucket?

Answer:

- Amazon S3 and Bucket names are
- This means bucket names must be unique across all AWS
- Bucket names can contain upto 63 lowercase letters, numbers, hyphens and
- You can create and use multiple buckets
- You can have upto 100 per account by

Q139) What are the storage class available in Amazon s3?

Answer:

- Amazon S3 Standard
- Amazon S3 Standard-Infrequent Access
- Amazon S3 Reduced Redundancy Storage
- Amazon Glacier

Q140) Explain Amazon s3 lifecycle rules?

Answer: Amazon S3 lifecycle configuration rules, you can significantly reduce your storage costs by automatically transitioning data from one storage class to another or even automatically delete data after a period of time.

- Store backup data initially in Amazon S3 Standard
- After 30 days, transition to Amazon Standard IA
- After 90 days, transition to Amazon Glacier
- After 3 years, delete

Q141) What is the relation between Amazon S3 and AWS KMS?

Answer: To encrypt Amazon S3 data at rest, you can use several variations of Server-Side Encryption. Amazon S3 encrypts your data at the object level as it writes it to disks in its data centers and decrypt it for you when you access it'll SSE performed by Amazon S3 and AWS Key Management Service (AWS KMS) uses the 256-bit Advanced Encryption Standard (AES).

Q142) What is the function of cross region replication in Amazon S3?

Answer: Cross region replication is a feature allows you asynchronously replicate all new objects in the source bucket in one AWS region to a target bucket in another region. To enable cross-region replication, versioning must be turned on for both source and destination buckets. Cross region replication is commonly used to reduce the latency required to access objects in Amazon S3

Q143) How to create Encrypted EBS volume?

Answer: You need to select Encrypt this volume option in Volume creation page. While creation a new master key will be created unless you select a master key that you created separately in the service. Amazon uses the AWS key management service (KMS) to handle key management.

Q144) Explain stateful and Stateless firewall.

Answer:

Stateful Firewall: A Security group is a virtual stateful firewall that controls inbound and outbound network traffic to AWS resources and Amazon EC2 instances. Operates at the instance level. It supports allow rules only. Return traffic is automatically allowed, regardless of any rules.

Stateless Firewall: A Network access control List (ACL) is a virtual stateless firewall on a subnet level. Supports allow rules and deny rules. Return traffic must be explicitly allowed by rules.

Q145) What is NAT Instance and NAT Gateway?

Answer:

NAT instance: A network address translation (NAT) instance is an Amazon Linux machine Image (AMI) that is designed to accept traffic from instances within a private subnet, translate the source IP address to the Public IP address of the NAT instance and forward the traffic to IWG.

NAT Gateway: A NAT gateway is an Amazon managed resources that is designed to operate just like a NAT instance but it is simpler to manage and highly available within an availability Zone. To allow instance within a private subnet to access internet resources through the IGW via a NAT gateway.

Q146) What is VPC Peering?

Answer: Amazon VPC peering connection is a networking connection between two amazon vpc's that enables instances in either Amazon VPC to communicate with each other as if they are within the same network. You can create amazon VPC peering connection between your own Amazon VPC's or Amazon VPC in another AWS account within a single region.

Q147) What is MFA in AWS?

Answer: Multi factor Authentication can add an extra layer of security to your infrastructure by adding a second method of authentication beyond just password or access key.

Q148) What are the Authentication in AWS?

Answer:

- User Name/Password
- Access Key
- Access Key/ Session Token

Q149) What is Data warehouse in AWS?

Data ware house is a central repository for data that can come from one or more sources. Organization typically use data warehouse to compile reports and search the database using highly complex queries. Data warehouse also typically updated on a batch schedule multiple times per day or per hour compared to an OLTP (Online Transaction Processing) relational database that can be updated thousands of times per second.

Q150) What is mean by Multi-AZ in RDS?

Answer: Multi AZ allows you to place a secondary copy of your database in another availability zone for disaster recovery purpose. Multi AZ deployments are available for all types of Amazon RDS Database engines. When you create s Multi-AZ DB instance a primary instance is created in one Availability Zone and a secondary instance is created by another Availability zone.

Q151) What is Amazon Dynamo DB?

Answer: Amazon Dynamo DB is fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. Dynamo DB makes it simple and Cost effective to store and retrieve any amount of data.

Q152) What is cloud formation?

Answer: Cloud formation is a service which creates the AWS infrastructure using code. It helps to reduce time to manage resources. We can able to create our resources Quickly and faster.

Q153) How to plan Auto scaling?

Answer:

- Manual Scaling
- Scheduled Scaling
- Dynamic Scaling

Q154) What is Auto Scaling group?

Answer: Auto Scaling group is a collection of Amazon EC2 instances managed by the Auto scaling service. Each auto scaling group contains configuration options that control when auto scaling should launch new instance or terminate existing instance.

Q155) Differentiate Basic and Detailed monitoring in cloud watch?

Answer:

Basic Monitoring: Basic monitoring sends data points to Amazon cloud watch every five minutes for a limited number of preselected metrics at no charge.

Detailed Monitoring: Detailed monitoring sends data points to amazon CloudWatch every minute and allows data aggregation for an additional charge.

Q156) What is the relationship between Route53 and Cloud front?

Answer: In Cloud front we will deliver content to edge location wise so here we can use Route 53 for Content Delivery Network. Additionally, if you are using Amazon CloudFront you can configure Route 53 to route Internet traffic to those resources.

Q157) What are the routing policies available in Amazon Route53?

Answer:

- Simple
- Weighted
- Latency Based
- Failover
- Geolocation

Q158) What is Amazon ElastiCache?

Answer: Amazon ElastiCache is a web services that simplifies the setup and management of distributed in memory caching environment.

- Cost Effective
- High Performance
- Scalable Caching Environment
- Using Memcached or Redis Cache Engine

Q159)What is SES, SQS and SNS?

Answer: SES (Simple Email Service): SES is SMTP server provided by Amazon which is designed to send bulk mails to customers in a quick and cost-effective manner.SES does not allows to configure mail server.

SQS (Simple Queue Service): SQS is a fast, reliable and scalable, fully managed message queuing service. Amazon SQS makes it simple and cost Effective. It's temporary repository for messages to waiting for processing and acts as a buffer between the component producer and the consumer.

SNS (Simple Notification Service): SNS is a web service that coordinates and manages the delivery or sending of messages to recipients.

Q160) How To Use Amazon Sqs? What Is Aws?

Answer:Amazon Web Services is a secure cloud services stage, offering compute power, database storage, content delivery and other functionality to help industries scale and grow.

Q161) What is the importance of buffer in AWS?

Answer:low price – Consume only the amount of calculating, storage and other IT devices needed. No long-term assignation, minimum spend or up-front expenditure is required.

Elastic and Scalable – Quickly Rise and decrease resources to applications to satisfy customer demand and control costs. Avoid provisioning maintenance up-front for plans with variable consumption speeds or low lifetimes.

Q162) What is the way to secure data for resounding in the cloud?

Answer:

- Avoid storage sensitive material in the cloud. …
- Read the user contract to find out how your cloud service storing works. …
- Be serious about passwords. …
- Encrypt. …
- Use an encrypted cloud service.

Q163) Name The Several Layers Of Cloud Computing?

Answer:Cloud computing can be damaged up into three main services: Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS). PaaS in the middle, and IaaS on the lowest

Q164) What Is Lambda edge In Aws?

Answer:Lambda Edge lets you run Lambda functions to modify satisfied that Cloud Front delivers, executing the functions in AWS locations closer to the viewer. The functions run in response to Cloud Front events, without provisioning or managing server.

Q165) Distinguish Between Scalability And Flexibility?

Answer:Cloud computing offers industries flexibility and scalability when it comes to computing needs:

Flexibility. Cloud computing agrees your workers to be more flexible – both in and out of the workplace. Workers can access files using web-enabled devices such as smartphones, laptops and notebooks. In this way, cloud computing empowers the use of mobile technology.

One of the key assistances of using cloud computing is its scalability. Cloud computing allows your business to easily expensive or downscale your IT requests as and when required. For example, most cloud service workers will allow you to increase your existing resources to accommodate increased business needs or changes. This will allow you to support your commercial growth without exclusive changes to your present IT systems.

Q166) What is IaaS?

Answer:IaaS is a cloud service that runs services on "pay-for-what-you-use" basis

IaaS workers include Amazon Web Services, Microsoft Azure and Google Compute Engine

Users: IT Administrators

Q167) What is PaaS?

Answer:PaaS runs cloud platforms and runtime environments to develop, test and manage software

Users: Software Developers

Q168) What is SaaS?

Answer:In SaaS, cloud workers host and manage the software application on a pay-as-you-go pricing model

Users: End Customers

Q169) Which Automation Gears Can Help With Spinup Services?

Answer:The API tools can be used for spin up services and also for the written scripts. Persons scripts could be coded in Perl, bash or other languages of your preference. There is one more option that is flowery management and stipulating tools such as a dummy or improved descendant. A tool called Scalar can also be used and finally we can go with a controlled explanation like a Right scale. Which automation gears can help with pinup service.

Q170) What Is an Ami? How Do I Build One?

Answer:An Amazon Machine Image (AMI) explains the programs and settings that will be applied when you launch an EC2 instance. Once you have finished organizing the data, services, and submissions on your ArcGIS Server instance, you can save your work as a custom AMI stored in Amazon EC2. You can scale out your site by using this institution AMI to launch added instances

Use the following process to create your own AMI using the AWS Administration Console:

*Configure an EC2 example and its attached EBS volumes in the exact way you want them created in the custom AMI.

1. Log out of your instance, but do not stop or terminate it.

2. Log in to the AWS Management Console, display the EC2 page for your region, then click Instances.
3. Choose the instance from which you want to create a custom AMI.
4. Click Actions and click Create Image.
5. Type a name for Image Name that is easily identifiable to you and, optionally, input text for Image Description.
6. Click Create Image.

Read the message box that appears. To view the AMI standing, go to the AMIs page. Here you can see your AMI being created. It can take a though to create the AMI. Plan for at least 20 minutes, or slower if you've connected a lot of additional applications or data.

Q171) What Are The Main Features Of Amazon Cloud Front?

Answer:Amazon Cloud Front is a web service that speeds up delivery of your static and dynamic web content, such as .html, .css, .js, and image files, to your users.CloudFront delivers your content through a universal network of data centers called edge locations

Q172) What Are The Features Of The Amazon Ec2 Service?

Answer:Amazon Elastic Calculate Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale cloud calculating easier for designers. Amazon EC2's simple web serviceinterface allows you to obtain and configure capacity with minimal friction.

Q173) Explain Storage For Amazon Ec2 Instance.?

Answer:An instance store is a provisional storing type located on disks that are physically attached to a host machine. … This article will present you to the AWS instance store storage type, compare it to AWS Elastic Block Storage (AWS EBS), and show you how to backup data stored on instance stores to AWS EBS

Amazon SQS is a message queue service used by scattered requests to exchange messages through a polling model, and can be used to decouple sending and receiving components

Q174) When attached to an Amazon VPC which two components provide connectivity with external networks?

Answer:

- Internet Gateway {IGW)
- Virtual Private Gateway (VGW)

Q175) Which of the following are characteristics of Amazon VPC subnets?

Answer:

- Each subnet maps to a single Availability Zone.
- By defaulting, all subnets can route between each other, whether they are private or public.

Q176) How can you send request to Amazon S3?

Answer:Every communication with Amazon S3 is either genuine or anonymous. Authentication is a process of validating the individuality of the requester trying to access an Amazon Web Services (AWS) product. Genuine requests must include a autograph value that authenticates the request sender. The autograph value is, in part, created from the requester's AWS access keys (access key identification and secret access key).

Q177) What is the best approach to anchor information for conveying in the cloud ?

Answer:Backup Data Locally. A standout amongst the most vital interesting points while overseeing information is to guarantee that you have reinforcements for your information,

- Avoid Storing Sensitive Information. …
- Use Cloud Services that Encrypt Data. …
- Encrypt Your Data. …
- Install Anti-infection Software. …
- Make Passwords Stronger. …
- Test the Security Measures in Place.

Q178) What is AWS Certificate Manager ?

Answer:AWS Certificate Manager is an administration that lets you effortlessly arrangement, oversee, and send open and private Secure Sockets Layer/Transport Layer Security (SSL/TLS) endorsements for use with AWS administrations and your inward associated assets. SSL/TLS declarations are utilized to anchor arrange interchanges and set up the character of sites over the Internet and additionally assets on private systems. AWS Certificate Manager expels the tedious manual procedure of obtaining, transferring, and reestablishing SSL/TLS endorsements.

Q179) What is the AWS Key Management Service

Answer:AWS Key Management Service (AWS KMS) is an overseen benefit that makes it simple for you to make and control the encryption keys used to scramble your information. … AWS KMS is additionally coordinated with AWS CloudTrail to give encryption key use logs to help meet your inspecting, administrative and consistence needs.

Q180) What is Amazon EMR ?

Answer:Amazon Elastic MapReduce (EMR) is one such administration that gives completely oversaw facilitated Hadoop system over Amazon Elastic Compute Cloud (EC2).

Q181) What is Amazon Kinesis Firehose ?

Answer:Amazon Kinesis Data Firehose is the least demanding approach to dependably stack gushing information into information stores and examination devices. … It is a completely overseen benefit that consequently scales to coordinate the throughput of your information and requires no continuous organization

Q182) What Is Amazon CloudSearch and its highlights ?

Answer:Amazon CloudSearch is a versatile cloud-based hunt benefit that frames some portion of Amazon Web Services (AWS). CloudSearch is normally used to incorporate tweaked seek abilities into different applications. As indicated by Amazon, engineers can set a pursuit application up and send it completely in under 60 minutes.

Q183) Is it feasible for an EC2 exemplary occurrence to wind up an individual from a virtual private cloud?

Answer:Amazon Virtual Private Cloud (Amazon VPC) empowers you to characterize a virtual system in your very own consistently disengaged zone inside the AWS cloud, known as a virtual private cloud (VPC). You can dispatch your Amazon EC2 assets, for example, occasions, into the subnets of your VPC. Your VPC nearly looks like a conventional system that you may work in your very own server farm, with the advantages of utilizing adaptable foundation from AWS. You can design your VPC; you can choose its IP address extend, make subnets, and arrange course tables, organize portals, and security settings. You can interface occurrences in your VPC to the web or to your own server farm

Q184) Mention crafted by an Amazon VPC switch.

Answer:VPCs and Subnets. A virtual private cloud (VPC) is a virtual system committed to your AWS account. It is consistently segregated from other virtual systems in the AWS Cloud. You can dispatch your AWS assets, for example, Amazon EC2 cases, into your VPC.

Q185) How would one be able to associate a VPC to corporate server farm?

Answer:AWS Direct Connect empowers you to safely associate your AWS condition to your on-premises server farm or office area over a standard 1 gigabit or 10 gigabit Ethernet fiber-optic association. AWS Direct Connect offers committed fast, low dormancy association, which sidesteps web access suppliers in your system way. An AWS Direct Connect area gives access to Amazon Web Services in the locale it is related with, and also access to different US areas. AWS Direct Connect enables you to consistently parcel the fiber-optic associations into numerous intelligent associations called Virtual Local Area Networks (VLAN). You can exploit these intelligent associations with enhance security, separate traffic, and accomplish consistence necessities.

Q186) Is it conceivable to push off S3 with EC2 examples ?

Answer:Truly, it very well may be pushed off for examples with root approaches upheld by local event stockpiling. By utilizing Amazon S3, engineers approach the comparative to a great degree versatile, reliable, quick, low-valued information stockpiling substructure that Amazon uses to follow its own overall system of sites. So as to perform frameworks in the Amazon EC2 air, engineers utilize the instruments giving to stack their Amazon Machine Images (AMIs) into Amazon S3 and to exchange them between Amazon S3 and Amazon EC2. Extra use case may be for sites facilitated on EC2 to stack their stationary substance from S3.

Q187) What is the distinction between Amazon S3 and EBS ?

Answer:EBS is for mounting straightforwardly onto EC2 server examples. S3 is Object Oriented Storage that isn't continually waiting be gotten to (and is subsequently less expensive). There is then much less expensive AWS Glacier which is for long haul stockpiling where you don't generally hope to need to get to it, however wouldn't have any desire to lose it.

There are then two principle kinds of EBS – HDD (Hard Disk Drives, i.e. attractive turning circles), which are genuinely ease back to access, and SSD, which are strong state drives which are excessively quick to get to, yet increasingly costly.

- Finally, EBS can be purchased with or without Provisioned IOPS.
- Obviously these distinctions accompany related estimating contrasts, so it merits focusing on the distinctions and utilize the least expensive that conveys the execution you require.

Q188) What do you comprehend by AWS?

Answer:This is one of the generally asked AWS engineer inquiries questions. This inquiry checks your essential AWS learning so the appropriate response ought to be clear. Amazon Web Services (AWS) is a cloud benefit stage which offers figuring power, investigation, content conveyance, database stockpiling, sending and some different administrations to help you in your business development. These administrations are profoundly versatile, solid, secure, and cheap distributed computing administrations which are plot to cooperate and, applications in this manner made are further developed and escalade.

Q189) Clarify the principle components of AWS?

Answer:The principle components of AWS are:

Highway 53: Route53 is an exceptionally versatile DNS web benefit.

Basic Storage Service (S3): S3 is most generally utilized AWS stockpiling web benefit.

Straightforward E-mail Service (SES): SES is a facilitated value-based email benefit and enables one to smoothly send deliverable messages utilizing a RESTFUL API call or through an ordinary SMTP.

Personality and Access Management (IAM): IAM gives enhanced character and security the board for AWS account.

Versatile Compute Cloud (EC2): EC2 is an AWS biological community focal piece. It is in charge of giving on-request and adaptable processing assets with a “pay as you go” estimating model.

Flexible Block Store (EBS): EBS offers consistent capacity arrangement that can be found in occurrences as a customary hard drive.

CloudWatch: CloudWatch enables the controller to viewpoint and accumulate key measurements and furthermore set a progression of cautions to be advised if there is any inconvenience.

This is among habitually asked AWS engineer inquiries questions. Simply find the questioner psyche and solution appropriately either with parts name or with the portrayal alongside.

Q190) I’m not catching your meaning by AMI? What does it incorporate?

Answer:You may run over at least one AMI related AWS engineer inquiries amid your AWS designer meet. Along these lines, set yourself up with a decent learning of AMI.

AMI represents the term Amazon Machine Image. It’s an AWS format which gives the data (an application server, and working framework, and applications) required to play out the dispatch of an occasion. This AMI is the duplicate of the AMI that is running in the cloud as a virtual server. You can dispatch occurrences from the same number of various AMIs as you require. AMI comprises of the followings:

A pull volume format for a current example

Launch authorizations to figure out which AWS records will inspire the AMI so as to dispatch the occasions

Mapping for square gadget to compute the aggregate volume that will be appended to the example at the season of dispatch

Q191) Is vertically scale is conceivable on Amazon occurrence?

Answer:Indeed, vertically scale is conceivable on Amazon example.

This is one of the normal AWS engineer inquiries questions. In the event that the questioner is hoping to find a definite solution from you, clarify the system for vertical scaling.

Q192) What is the association among AMI and Instance?

Answer:Various sorts of examples can be propelled from one AMI. The sort of an occasion for the most part manages the equipment segments of the host PC that is utilized for the case. Each kind of occurrence has unmistakable registering and memory adequacy.

When an example is propelled, it gives a role as host and the client cooperation with it is same likewise with some other PC however we have a totally controlled access to our occurrences. AWS engineer inquiries questions may contain at least one AMI based inquiries, so set yourself up for the AMI theme exceptionally well.

Q193) What is the distinction between Amazon S3 and EC2?

Answer:The contrast between Amazon S3 and EC2 is given beneath:

Amazon S3

Amazon EC2

The significance of S3 is Simple Storage Service. The importance of EC2 is Elastic Compute Cloud.

It is only an information stockpiling administration which is utilized to store huge paired files. It is a cloud web benefit which is utilized to have the application made.

It isn't required to run a server. It is sufficient to run a server.

It has a REST interface and utilizations secure HMAC-SHA1 validation keys. It is much the same as a tremendous PC machine which can deal with application like Python, PHP, Apache and some other database.

When you are going for an AWS designer meet, set yourself up with the ideas of Amazon S3 and EC2, and the distinction between them.

Q194) What number of capacity alternatives are there for EC2 Instance?

Answer:There are four stockpiling choices for Amazon EC2 Instance:

- Amazon EBS
- Amazon EC2 Instance Store
- Amazon S3
- Adding Storage

Amazon EC2 is the basic subject you may run over while experiencing AWS engineer inquiries questions. Get a careful learning of the EC2 occurrence and all the capacity alternatives for the EC2 case.

Q195) What are the security best practices for Amazon Ec2 examples?

Answer:There are various accepted procedures for anchoring Amazon EC2 occurrences that are pertinent whether occasions are running on-preface server farms or on virtual machines. How about we view some broad prescribed procedures:

Minimum Access: Make beyond any doubt that your EC2 example has controlled access to the case and in addition to the system. Offer access specialists just to the confided in substances.

Slightest Privilege: Follow the vital guideline of minimum benefit for cases and clients to play out the capacities. Produce jobs with confined access for the occurrences.

Setup Management: Consider each EC2 occasion a design thing and use AWS arrangement the executives administrations to have a pattern for the setup of the occurrences as these administrations incorporate refreshed enemy of infection programming, security highlights and so forth.

Whatever be the activity job, you may go over security based AWS inquiries questions. Along these lines, motivate arranged with this inquiry to break the AWS designer meet.

Q196) Clarify the highlights of Amazon EC2 administrations.

Answer:Amazon EC2 administrations have following highlights:

- Virtual Computing Environments
- Proffers Persistent capacity volumes
- Firewall approving you to indicate the convention
- Pre-designed layouts
- Static IP address for dynamic Cloud Computing

Q197) What is the system to send a demand to Amazon S3?

Answer: Reply: There are 2 different ways to send a demand to Amazon S3 –

- Using REST API
- Using AWS SDK Wrapper Libraries, these wrapper libraries wrap the REST APIs for Amazon

Q198) What is the default number of basins made in AWS?

Answer**:**This is an extremely straightforward inquiry yet positions high among AWS engineer inquiries questions. Answer this inquiry straightforwardly as the default number of pails made in each AWS account is 100.

Q199) What is the motivation behind T2 examples?

Answer:T2 cases are intended for

Providing moderate gauge execution

Higher execution as required by outstanding task at hand

Q200) What is the utilization of the cradle in AWS?

Answer:This is among habitually asked AWS designer inquiries questions. Give the appropriate response in straightforward terms, the cradle is primarily used to oversee stack with the synchronization of different parts i.e. to make framework blame tolerant. Without support, segments don't utilize any reasonable technique to get and process demands. Be that as it may, the cushion makes segments to work in a decent way and at a similar speed, hence results in quicker administrations.

Q201) What happens when an Amazon EC2 occurrence is halted or ended?

Answer:At the season of ceasing an Amazon EC2 case, a shutdown is performed in a typical way. From that point onward, the changes to the ceased state happen. Amid this, the majority of the Amazon EBS volumes are stayed joined to the case and the case can be begun whenever. The occurrence hours are not included when the occasion is the ceased state.

At the season of ending an Amazon EC2 case, a shutdown is performed in an ordinary way. Amid this, the erasure of the majority of the Amazon EBS volumes is performed. To stay away from this, the estimation of credit deleteOnTermination is set to false. On end, the occurrence additionally experiences cancellation, so the case can't be begun once more.

Q202) What are the mainstream DevOps devices?

Answer:In an AWS DevOps Engineer talk with, this is the most widely recognized AWS inquiries for DevOps. To answer this inquiry, notice the well known DevOps apparatuses with the kind of hardware –

- Jenkins – Continuous Integration Tool
- Git – Version Control System Tool
- Nagios – Continuous Monitoring Tool
- Selenium – Continuous Testing Tool
- Docker – Containerization Tool
- Puppet, Chef, Ansible – Deployment and Configuration Administration Tools.

Q203) What are IAM Roles and Policies, What is the difference between IAM Roles and Policies.

Answer:Roles are for AWS services, Where we can assign permission of some AWS service to other Service.

Example – Giving S3 permission to EC2 to access S3 Bucket Contents.

Policies are for users and groups, Where we can assign permission to user's and groups.

Example – Giving permission to user to access the S3 Buckets.

Q204) What are the Defaults services we get when we create custom AWS VPC?

Answer:

- Route Table

- Network ACL
- Security Group

Q205) What is the Difference Between Public Subnet and Private Subnet ?

Answer:Public Subnet will have Internet Gateway Attached to its associated Route Table and Subnet, Private Subnet will not have the Internet Gateway Attached to its associated Route Table and Subnet

Public Subnet will have internet access and Private subnet will not have the internet access directly.

Q206) How do you access the Ec2 which has private IP which is in private Subnet ?

Answer: We can access using VPN if the VPN is configured into that Particular VPC where Ec2 is assigned to that VPC in the Subnet. We can access using other Ec2 which has the Public access.

Q207) We have a custom VPC Configured and MYSQL Database server which is in Private Subnet and we need to update the MYSQL Database Server, What are the Option to do so.

Answer:By using NAT Gateway in the VPC or Launch a NAT Instance (Ec2) Configure or Attach the NAT Gateway in Public Subnet (Which has Route Table attached to IGW) and attach it to the Route Table which is Already attached to the Private Subnet.

Q208) What are the Difference Between Security Groups and Network ACL

Answer:

Security Groups	**Network ACL**
Attached to Ec2 instance	Attached to a subnet.
Stateful – Changes made in incoming rules is automatically applied to the outgoing rule	Stateless – Changes made in incoming rules is not applied to the outgoing rule
Blocking IP Address can't be done	IP Address can be Blocked
Allow rules only, by default all rules are denied	Allow and Deny can be Used.

Q209) What are the Difference Between Route53 and ELB?

Answer:Amazon Route 53 will handle DNS servers. Route 53 give you web interface through which the DNS can be managed using Route 53, it is possible to direct and failover traffic. This can be achieved by using DNS Routing Policy.

One more routing policy is Failover Routing policy. we set up a health check to monitor your application endpoints. If one of the endpoints is not available, Route 53 will automatically forward the traffic to other endpoint.

Elastic Load Balancing

ELB automatically scales depends on the demand, so sizing of the load balancers to handle more traffic effectively when it is not required.

Q210) What are the DB engines which can be used in AWS RDS?

Answer:

- MariaDB
- MYSQL DB
- MS SQL DB
- Postgre DB
- Oracle DB

Q211) What is Status Checks in AWS Ec2?

Answer: System Status Checks – System Status checks will look into problems with instance which needs AWS help to resolve the issue. When we see system status check failure, you can wait for AWS to resolve the issue, or do it by our self.

- Network connectivity
- System power
- Software issues Data Centre's
- Hardware issues
- Instance Status Checks – Instance Status checks will look into issues which need our involvement to fix the issue. if status check fails, we can reboot that particular instance.
- Failed system status checks
- Memory Full
- Corrupted file system
- Kernel issues

Q212) To establish a peering connections between two VPC's What condition must be met?

Answer:

- CIDR Block should overlap
- CIDR Block should not overlap
- VPC should be in the same region
- VPC must belong to same account.
- CIDR block should not overlap between vpc setting up a peering connection . peering connection is allowed within a region , across region, across different account.

Q213) Troubleshooting with EC2 Instances:
Answer: Instance States

- If the instance state is 0/2- there might be some hardware issue
- If the instance state is ½-there might be issue with OS.
  Workaround-Need to restart the instance, if still that is not working logs will help to fix the issue.

Q214) How EC2instances can be resized.

Answer: EC2 instances can be resizable(scale up or scale down) based on requirement

Q215) EBS: its block-level storage volume which we can use after mounting with EC2 instances.

Answer:For types please refer AWS Solution Architect book.

Q216) Difference between EBS,EFS and S3

Answer:

- We can access EBS only if its mounted with instance, at a time EBS can be mounted only with one instance.
- EFS can be shared at a time with multiple instances
- S3 can be accessed without mounting with instances

Q217) Maximum number of bucket which can be crated in AWS.

Answer:100 buckets can be created by default in AWS account.To get more buckets additionally you have to request Amazon for that.

Q218) Maximum number of EC2 which can be created in VPC.

Answer:Maximum 20 instances can be created in a VPC. we can create 20 reserve instances and request for spot instance as per demand.

Q219) How EBS can be accessed?

Answer:**EBS** provides high performance block-level storage which can be attached with running EC2 instance. Storage can be formatted and mounted with EC2 instance, then it can be accessed.

Q220)Process to mount EBS to EC2 instance

Answer:

- Df –k
- mkfs.ext4 /dev/xvdf
- Fdisk –l
- Mkdir /my5gbdata
- Mount /dev/xvdf /my5gbdata

Q221) How to add volume permanently with instance.

Answer:With each restart volume will get unmounted from instance, to keep this attached need to perform below step

Cd /etc/fstab

/dev/xvdf /data ext4 defaults 0

0 <edit the file system name accordingly>

Q222) What is the Difference between the Service Role and SAML Federated Role.

Answer: Service Role are meant for usage of AWS Services and based upon the policies attached to it,it will have the scope to do its task. Example : In case of automation we can create a service role and attached to it.

Federated Roles are meant for User Access and getting access to AWS as per designed role.
Example : We can have a federated role created for our office employee and corresponding to that a Group will be created in the AD and user will be added to it.

Q223) How many Policies can be attached to a role.

Answer: 10 (Soft limit), We can have till 20.

Q224) What are the different ways to access AWS.

Answer:3 Different ways (CLI, Console, SDK)

Q225) How a Root AWS user is different from in IAM User.

Answer: Root User will have acces to entire AWS environment and it will not have any policy attached to it. While IAM User will be able to do its task on the basis of policies attached to it.

Q226) What do you mean by Principal of least privilege in term of IAM.

Answer: Principal of least privilege means to provide the same or equivalent permission to the user/role.

Q227)What is the meaning of non-explicit deny for an IAM User.

Answer: When an IAM user is created and it is not having any policy attached to it,in that case he will not be able to access any of the AWS Service until a policy has been attached to it.

Q228) What is the precedence level between explicit allow and explicit deny.

Answer: Explicit deny will always override Explicit Allow.

Q229) What is the benefit of creating a group in IAM.

Answer:Creation of Group makes the user management process much simpler and user with the same kind of permission can be added in a group and at last addition of a policy will be much simpler to the group in comparison to doing the same thing manually.

Q230) What is the difference between the Administrative Access and Power User Access in term of pre-build policy.

Answer: Administrative Access will have the Full access to AWS resources. While Power User Access will have the Admin access except the user/group management permission.

Q231) What is the purpose of Identity Provider.

Answer: Identity Provider helps in building the trust between the AWS and the Corporate AD environment while we create the Federated role.

Q232) What are the benefits of STS (Security Token Service).

Answer: It help in securing the AWS environment as we need not to embed or distributed the AWS Security credentials in the application. As the credentials are temporary we need not to rotate them and revoke them.

Q233) What is the benefit of creating the AWS Organization.

Answer: It helps in managing the IAM Policies, creating the AWS Accounts programmatically, helps in managing the payment methods and consolidated billing.

Q234) What is the maximum file length in S3?

Answer: utf-8 1024 bytes

Q235) which activity cannot be done using autoscaling?

Answer:Maintain fixed running of ec2

Q236) How will you secure data at rest in EBS?

Answer: EBS data is always secure

Q237) What is the maximum size of S3 Bucket?

Answer: 5TB

Q238) Can objects in Amazon s3 be delivered through amazon cloud front?

Answer:Yes

Q239) which service is used to distribute content to end user service using global network of edge location?

Answer: Virtual Private Cloud

Q240) What is ephemaral storage?

Answer: Temporary storage

Q241) What are shards in kinesis aws services?

Answer: Shards are used to store data in Kinesis.

Q242) Where can you find the ephemeral storage?

Answer: In Instance store service.

Q243) I have some private servers on my premises also i have distributed some of My workload on the public cloud,what is the architecture called?

Answer:Virtual private cloud

Q244)Route 53 can be used to route users to infrastructure outside of aws.True/false?

Answer: False

Q245) Is simple workflow service one of the valid Simple Notification Service subscribers?

Answer: No

Q246) which cloud model do Developers and organizations all around the world leverage extensively?

Answer: IAAS-Infrastructure as a service.

Q247) Can cloud front serve content from a non AWS origin server?

Answer: No

Q248) Is EFS a centralised storage service in AWS?

Answer: Yes

Q249) Which AWS service will you use to collect and process ecommerce data for near real time analysis?

Answer: Both Dynamo DB & Redshift

Q250)An high demand of IOPS performance is expected around 15000.Which EBS volume type would you recommend?

Answer: Provisioned IOPS.

Part – 7 : Interview Questions

1. How much data you are processing everyday?

2. On what types of data you are working?

3. How many tables you are having in your RDBMS and what is the size of each table on a average?

4. How you use sqoop incremental load and how to stop the sqoop incremental job?

5. How many rows you are getting after doing "select * from table_name"?

6. How much time it is taking to process the data in hive and spark?

7. How much data is getting appended everyday?

8. On what frequency means how many times you are using your sqoop job in a day or in a week?

9. How you are reusing the RDD(RDD transformations) with scenarios.

10. What types of processing you are doing through your hive and Spark?

11. When to use RDD and Dataframe?

https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-datafram

es-and-datasets.html

12. Why RDD is immutable?

https://www.quora.com/Why-is-RDD-immutable-in-Spark

13. What is the difference between spark context and Spark session?

http://data-flair.training/forums/topic/sparksession-vs-sparkcontext-in-apache-spark

14. What is the difference between partitioning,bucketing,repartitioning and coalesce?

https://data-flair.training/blogs/hive-partitioning-vs-bucketing/

https://stackoverflow.com/questions/31610971/spark-repartition-vs-coalesce

15. How to debug your spark code and how to build the jar?

16. How are you scheduling your jobs using oozie?

17. How to select data from a hive external partitioned table?

http://blog.zhengdong.me/2012/02/22/hive-external-table-with-partitions/

18. When to use spark client and cluster mode?

19. What is your cluster configuration and versions of each component?

20. Real time use cases of companion objects and trait ?

21. How to fetch data from a scala list in a parallel manner?

Not exactly the answer, but helpful

https://alvinalexander.com/scala/how-to-use-parallel-collections-in-scala-performance

22. How to increase spark executor memory and hive utilisation memory?

https://stackoverflow.com/questions/26562033/how-to-set-apache-spark-executor-memo ry

23. Different types of Nosql databases.What is the difference between hbase and Cassandra.

https://www.3pillarglobal.com/insights/exploring-the-different-types-of-nosql-databases

https://data-flair.training/blogs/hbase-vs-cassandra/

24. Questions regarding different file formats of Hadoop and when to use?

Useful blog:

https://community.hitachivantara.com/community/products-and-solutions/pentaho/blog/2

017/11/07/hadoop-file-formats-its-not-just-csv-anymore

25. What is the difference between hive map join and hive bucket join?

https://data-flair.training/blogs/map-join-in-hive/

https://data-flair.training/blogs/bucket-map-join/

26. Performance optimization techniques in sqoop,hive and spark.

Hive - https://hortonworks.com/blog/5-ways-make-hive-queries-run-faster/

27. End to end project flow and the usage of all the hadoop ecosystem components.

28. Why does Apache Spark exist and how PySpark fits into the picture?

29. What is the file size you are using in your development and production environment?

30. Use cases of accumulators and broadcast variables?

31. Explain the difference between internal and external tables?

Ans: 1.Dropping table: if you drop internal table it will remove both schema in metastore & data in Hdfs. If you drop external table, only schema will be removed from metastore whereas data still exists in Hdfs

When to use?

1. Use internal table if it's data won't be used by other bigdata Ecosystems.

2. Use external table if it's data would be used by other bigdata ecosystems as it won't have any impact just in case of table drop operation

32. How did you run Hive load scripts in Production?

Ans: All the hive commands were kept in .sql files (for ex - load ordersdata.sql) and these files were invoked in in Unix shell script through command: hive -f ordersdata.sql. These unix scripts had few other HDFS commands as well. For ex - To load data into HDFS, make backup on local file system, send email once load was done etc. etc.). These unix scripts were called through Enterprise scheduler (Control M or Autosys or Zookeeper).

33. Why does hive doesn't store metadata in HDFS?

Ans: 1.Storing metadata in HDFS results in high latency/delay considering the fact of sequential access in HDFS for read/write operations. So it's evident to store metadata in Metastore to achieve low latency because of random access in metastore(MySQL dB)

34. Which file format works best with hive tables ? Why?

Ans: Usually Columnar formats are efficient in terms of both file size and query performance. ORC and parquet both are Columnar formats.

Compared to row oriented data, file sizes are smaller in Columnar formats since values from one column to another are stored next to each other.

Query performance is also improved since a query engine can ignore columns that's not required to answer a query.

Key ability of parquet: it stores the data that has deeply nested structures in a Columnar fashion (columns within columns). In the real world, this might be the scenario.

On the other hand, If we take another comparison parameter - compression: then ORC with zlib leads compared to parquet with Snappy.

In the end, it depends on the structure of your data. If you have nested structured data then parquet should be the option. If not (flat data) then go with ORC as it also offers better compression ratio than parquet.

35. How To append files of various DF's?

Ans:DF.write.option("basePath","/working/DIR").mode("append").parquet("/working/DIR")

--option takes (key, value), mode needs to be "append" Use option stating which would be base directory which takes (k,v) pair Mode as append. Mode is dataframewriter which take string as argument

36. Hive Scenario based interview questions:

1.Suppose, I create a table that contains details of all the transactions done by the customers of year 2016: CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;

Now, after inserting 50,000 tuples in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query.

How will you solve this problem and list the steps that I will be taking in order to do so?

Ans:

We can solve this problem of query latency by partitioning the table according to each month. So, for each month we will be scanning only the partitioned data instead of whole data sets.

As we know, we can't partition an existing non-partitioned table directly. So, we will be taking following steps to solve the very problem:

Create a partitioned table, say partitioned_transaction:

CREATE TABLE partitioned_transaction (cust_id INT, amount FLOAT, country STRING) PARTITIONED BY (month STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;

2. Enable dynamic partitioning in Hive:

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

3. Transfer the data from the non - partitioned table into the newly created partitioned table:

INSERT OVERWRITE TABLE partitioned_transaction PARTITION (month) SELECT cust_id, amount, country, month FROM transaction_details;

Now, we can perform the query using each partition and therefore, decrease the query time.

37. Why mapreduce will not run if you run select * from table in hive?

Whenever you run a normal 'select *', a fetch task is created rather than a mapreduce task which just dumps the data as it is without doing anything on it. This is equivalent to a:

hadoop fs -cat $file_name

In general, any sort of aggregation, such as min/max/count is going to require a MapReduce job

38. How to import first 10 records from a RDBMS table into HDFS using sqoop?How to import all the records except first 20 rows/records and also last 50 records using sqoop import?

39. What is the difference between Kafka and flume?

40. How to change the number of replication factors and how to change the number of mappers and reducers ?

41. How the number of partitions and stages get decided in spark?

42. What is the default number of mappers and reducers in map reduce job?

43. How to change the block size while importing the data into HDFS?

44. What setting need to be done while doing dynamic partition and bucketing.

45. How to run mapreduce and spark job?

46. What is data sets in spark and how to create and use it?

47. What are the difference between hive and hbase,hive and RDBMS,No SQL and RDBMS?

48. What are the difference between Hadoop and RDBMS?

49. What are the difference between Hadoop and spark?

50. What are the difference between Scala and Java.

51. What are the advantages and disadvantages of functional programming?

52. What are the advantages of Hadoop over distributed file systems?

53. Core concept of map reduce internal architecture and job flow.

54. Architecture of Hadoop,Yarn and Spark.

55. What are the advantages of using Yarn as cluster manager than Mesos and Spark standalone CM?

Company Specific Questions

Company: Fedility Date: 07-Aug-2018

1. What security authentication you are using. how you are managing?

2. about Centry, security authentication ?

3. how do you do schedule the jobs in Fair scheduler

4. prioritizing jobs

5. how you are doing Accenterl control for HDFS?

6. Disaster Recovery activities

7. what issues you are faced so far

8. do you know about puppet

9. hadoop development activities

Company: Accenture Dt: 06-July-2018

1) What are your daily activities? And What are your roles and responsibilities in your current project? What are the services that are implemented in your current project?

2) What have you done for performance tunning??

3) What is the block size in your project?

4) Explain your current project process

5) Have you used Storm Kafka or Solr services in your project?

6) Have you used puppet tool

7) Have you used security in your project? Why do you use security in your cluster?

8) Explain how kerberos authentication happens?

9) What is your cluster size and what are the services you are using? 10) Do you have good hands on experience in Linux

11) Have you used Flume or Storm in your Project?

Company: ZNA 04-July-2018

1)Roles and responsibilities in current project

2)What do you monitor in cluster i.e; What do you monitor to ensure that cluster is in healthy

state?

3)Are you involved in planning and implementation of Hadoop cluster. What are the components that need to keep in mind while planning hadoop Cluster.

4)You are given 10 Empty boxes with 256GB RAM and good Hardware conditions, How will you plan your cluster with these 10 boxes when there is 100GB of data to come per day.(Steps right from Begining i.e; chosing OS, chossing Softwares to be installed on empty boxes, Installation steps to install RedHat Linux)

5) Steps to install Cloudera Hadoop.

6) What is JVM?

7) What is Rack awareness??

8) What is Kerberos security and how will you install and enable it using CLI and how to integrate it with Cloudera manager.

9) What is High Availability? How do you implement High availability on a pre existing cluster with single node? What are the requirements to implement HA.

10) What is HIVE? How do you install and configure from CLI. 11) What is Disc Space and Disc Quota

12) How to add data nodes to your cluster without using Cloudera Manager.

13) How to add Disk space to Datanode which is already added to cluster. And how to format the disk before adding it to cluster.

14) How good r u at shell scripting? Have you used shell scripting to automate any of your

activities.

What are the activities that r automated using shell scripting in your current project? 15) What are the benefits of YARN compare to Hadoop-1.

17) Difference between MR1 and MR2?

18) Most challenges that you went through in your project. 19) Activities performed on Cloudera Manager

20) How you will know about the threshold, do you check manually every time. Do you know about puppet etc.,

21) How many clusters and nodes are present in your project.

22) You got a call when u r out of office saying there is no enough space i.e., HDFS threshold has been reached. What is the your approach to resolve this issue.

23) Heat beat messages, Are they sequential processing or parallel processing. 24) What is the volume of data you receive to your cluster every day.

25) What is HDFS?

26) How do you implement SSH, SCP and SFTP in Linux 27) What are the services used for HA.

28) Do you have experience on HBASE.

29) Does HA happen automatically.

Company: Infosys (Secound Round) Dt: 04-April-2018

1. what is distribution you use and how did you upgrade from 5.3 to 5.4
2. are you upgrading in node.. how?
3. How do you copy config files to other nodes
4. what security system you follows, what is diff with out kerberos
5. What is JN, HA
6. what is usage of SNN
7. usage of Automatic failover , how you do ? what all r other methods?
8. How do you load data for teradata to Hadoop
9. Are you using IMpala?
10. what is cluster size
11. How do you install the cloudera manager
12. what is iQuery
13. You already had dev exp, going to ask question n Deve
14. What Unix your using and how to find the OS full details.

Company: Cognizant (CTS) Dt: 04-Nov-2017

1)how you will give access to Hue

2)what is rebalancing

3)what will be needed from user for Karbarose

4)Java heap issue

5)explain about sqoop

6)Expain about oozie

7)where log files wil be stored

tar -cvf

8)what is Master and region server

9)What is Edge node

10)expalin yarn

11)High availability

12)what is the responsability of zookeeper

13)What needs to be done in order to run the standby node 14)Decommission of datanode

15)Cluster details

16)Scalability

17)How you will check the upgradation is successful 18)schedulers

19)what will be the steps you perform when a process got failed 20)recent issues you got faced

21)what are the recent issues you faced 22)Shell scripting

23)what will be the precations you will take in order to avoid single point of failure 24)what is your backup plan

25)how will you upgrade the cloudera manager from 5.3 to 5.4

Company: EMC (Duration was 45 Mins) Dt: 04-Dec-2017

01) Could you explain your big data experience.

02) Could explain about your environment, how many clusters. 03) What is the size of your cluster.

04) How is data loaded into HIVE.

05) What is the configuration of nodes.

06) What do you do for map reduce performance tuning.

07) What are the parameters and values used for tuning.

08) What will happen, when you change those values.

09) What else are used for tuning, other than reducer.

10) which components are there between mapper and reducer. 11) What are the steps to install Kerberos.

12) How do you integrate Kerberos in Tableau.

13) Do you have idea about SQOOP, FLUME.

14) What type of files come into your application.

15) Have you worked on un-structured files.

16) What type of tables you are using in HIVE, internal or external tables. 17) Do you have idea about HUE.

18) Where HUE is installed.

19) How do you give access to HUE and how Kerberos is integrated in HUE. 20) Do you have idea about SPARK, SPLUNK.

21) Could you explain unix scripts you have developed till now.

22) What are the routine unix command you use.

23) How do you check I/O operations in unix.

24) What is the size of the container in your project.

25) What is the architecture of your project, how does data comes. 26) Do you have experience on Teradata.

27) What is the difference between Teradata and Oracle. 28) What are the utilities used in teradata.

Company: Wipro (Duration was15 Mins) Dt: 20-Feb-2015

1) What is your experiance in big data space.

2) What are your day to day activities.

3) Responsibilities you are saying should be automated by now, what is your actual work in it.

4) Have you seen a situation, where mapreduce program is not performing well which used to execute properly before. What is your approch to resolve the issue.

5) Do you came accrosee the issue, where sort and suffle was causing issue in mapreduce

program.

6) Have you worked on Kafka.

7) What are the reporting toole you are using.

8) Any experience on spark.

9) What are the chanllenges you faced.

10) I will inform employer, he will notify next steps

INTERVIEW QUESTIONS

Company: Impetus 21Oct2017

1) What ate your day to day activities.

2) What is the difference between root user and normal user.

3) Is your cluster on cloud. Do you have idea about cloud.

4) Are you racks present in any data center.

5) What Hadoop version you are using.

6) What is the process to add node to cluster. Do you have any standard process. Do you see physical servers.

7) What do you do for Tableau installation and integration.

8) What schedulers you suing in your project.

9) What is your cluster Size.

10) What issue you faced in your project. Do you login frequently.

11) How jobs are handled. Do developers take care of it or you involve. 12) Have you worked on sqoop and Oozie.

13) What are the echo systems you have worked. 14) Do you know about sentry.

15) Looks like, you have worked on Cloudera Manager. What is comfort level on manual and Hortonworks.

16) Have you done any scripting.

Company: Tata Consultancy Services TCS Dt: 18-Oct-2017 (25Mins)

1) Hi, Where are you located. Are you fine to relocate to CA.

2) How much experience you have in big data area.

3) Could you give me your day to day activities?

4) What is the process to upgrade HIVE.

5) What is the way to decommission multiple data nodes.

6) Have you used rsync command.

7) How do you decommission a data node.

8) What is the process to integratemetastore for HIVE. Could you explain the process?

9) Do you have experience on scripting. If yes, is it Unix or python. 10) Have you worked on puppet.

11) Have you worked on other distributions like Horton works. 12) How do you delete files which are older than 7 days.

13) what is the way to delete tmp files from nodes. If there are 100 nodes, do you do it

manually.

14) Have you involved in migration from CDH1 to CDH2.

15) If there is 20TB data in CHD1, What is the way to move it to CDH2.' 16) Have you worked on HBASE.

17) Do you know about Nagios and Ganglia. How graphs are used.

18) In Nagios, what are different options (conditions) to generate alerts. 19) Have you worked on Kerberos.

20) What is command for balancing the datanodes.

Company: DishNET Dt: 15-Oct-2017 (30 Mins)

1) Tell me about yourself.

2) What is meant by High availability.

3) Does HA happen automatically.

4) What are the services used for HA.

5) What are the benefits of YARN compare to Hadoop-1.

6) Have you done integration of map reduce to run HIVE.

7) Do you have experience on HBASE.

8) Could you explain the process of integration on Tableau.

9) What is the process of upgrading data node. 10) what are the schedulers used in hadoop. 11) How do you do load balancing.

12) when you add data node to cluster, how data will be copied to new datanode. 13) How you can remove 5 data nodes from cluster. Can you do it all at same time. 14) How do you give authorization to users.

15) How do you give permissions to a file like write access to one group and read access to other group.

16) How do you authenticate to HIVE tables for users. 17) How do you give LDAP access to HIVE for users. 18) Do you know about Kerberos.

19) Have you done upgrade CDH.

20) Do you need to bring down for CDH upgrade.

21) Have you worked on performance tuning of HIVE queries. 22) What type of performance tunings you have done.

23) Do you have idea about impala.

24) Do you know, how hadoop supports real time activity. 25) How do you allocate resource pool.

26) How do you maintain data in multiple disks on datanode.

27) Will there be any performance issue, if data is in different disks on datanode.

Company: Hexaware Dt: 10-Aug-2018 (41 Mins)

1) Tell me your day to day activities.

2) When adding datanode, do you bring down cluster.

3) What are the echo systems you have on your cluster.

4) Have you involved in cluster planning.

5) Who will take decision to add new data node.

6) Have you involved in planning for adding datanodes.

7) How do you do upgrades, is there any window.

8) When you are adding datanode, what is the impact of new blocks created by running jobs.

9) Do you have any idea about check pointing.

10) For check pointing, do Admin need to any activity or it is automatically taken care by cloudera.

11) Do you know about Ambari. Have you ever worked on Ambari or HortonWorks.

12) Do developers use map reduce programming on the cluster you are working.

13) Do you know, what type of data is coming from different systems to your cluster and what type of analysis is done on the same.

14) Do you have scala and strom in your application. 15) Do you use any oozie scheduler in the project. 16) What type of unix scripting is done.

17) whether your cluster is on any cloud.

18) When you are adding any datanode, do you do anything with configuration files. 19) How much experience you have on linux and scripting. How is your comfort level. 20) Do you have idea about data warehouse.

21) Have you worked on data visualization.

22) Who takes care of copying data from unix to HDFS, whether there is any automation.

23) Looks like, you joined on project which is already configured. Do you have hands-on on configuration cluster from scratch.

24) Have you ever seen hardware of nodes in the cluster. What is the configuration. 25) Have you used, Sqoop to pull data from different databases.

26) What is your cluster size.

Company: Initial Screening by Vendor for VISA Client Date: 5th-Oct-2017

1) What are your day to day activities.

2) How do you add datanode to the cluster.

3) Do you have any idea about dfs.name.dir?

4) What will happend when data node is down.

5) How you will test, whether datanode is working or not.

6) Do you have idea about Zoombie process.

7) How namenode will be knowing datanode is down.

Nagios alert, admin -report (command), cloudera manage

8) Heat beat, whether it is sequential processing or parallel processing.

9) What is the volume of data you receive to the cluster.

40 to 50GB

10) How do you receive data to your cluster. 11) What is your cluster size.

12) What is the port number of namenode.

13) What is the port number of Job tracker.

14) How do you install hive, pig, hbase.

15) What is JVM?

16) How do you do rebalancing.

Company: Verizon 02-Oct-2017

1)How do you dopaswordless SSH in hadoop.

2) Upgrades (Have you done anytime).

3) ClouderaManager port number.

4) what is your cluster size.

5) Versions

6) Map reduce version.

7) Daily activities.

8) What operations, you normally use in cloudera manager.

9) is internet connected to your nodes.

10) Do you have different cloudera managers for dev and production. 11) what are installation steps

Company: HCL 22-Sep-2017

1) Daily activities.

2) versions.

3) What is decommissioning.

4) What is the procedure to decommission datanode.

5) Difference between MR1 and MR2.

6) Difference between Hadoop1 and Hadoop2.

7) Difference between RDBMS and No-SQL.

8) What is the use of Nagios.

Company: Collabera Date: 14-Mar-2018

1) Provide your roles and responsibilities.

2) What do you do for cluster management.

3) At midnight, you got a call saying there is no enough space i.e., HDFS threshold has been reached. What is the your approach to resolve this issue.

4) How many clusters and nodes are present in your project.

5) How you will know about the threshold, do you check manually every time. Do you know about puppet etc.,

6) Code was tested successfully in Dev and Test. When deployed to Productions it is failing. As an admin, how do you track the issue?

7) If namenode is down, whole cluster will be down. What is the approach to bring it back.

8) what is decommissioning?

9) You have decommissioned a node, can you add it back to cluster again. What about the data present in datanode when decommissioned.

10) Node is having different version of software, can we add it to cluster.

More Questions from Collabera Vendor:

1) Activities performed on Cloudera Manager

2) How to start & stop namenode services

3) Most challenges that you went thru in your project

4) How do you install Cloudera and namenode

5) Background of current project

6) If datanode is down, what will be the solution (situation based question)

7) More questions can be expected for Linux &Hadoop administration

SOME BIG DATA REAL TIME PRODUCTION LEVEL QUESTIONS

1)what is the file size you've used?

2)How long does it take to run your script in production cluster?

3)what is the file size for production environment?

4) Are you planning for anything to improve the performance?

5)what size of file do you use for Development?

6)what did you do to increase the performance(Hive,pig)?

7)what is your cluster size?

8)what are the challenges you have faced in your project? Give 2 examples?

9)How to debug production issue?(logs, script counters, JVM)

10)how do you select the eco system tools for your project?

11)How many nodes you are using currently?

12)what is the job scheduler you use in production cluster?

More question

1) What are your day to day activities.

2) How do you add datanode to the cluster.

3) Do you have any idea about dfs.name.dir?

4) What will happend when data node is down.

5) How you will test, whether datanode is working or not.

6) Do you have idea about Zoombie process.

7) How namenode will be knowing datanode is down.

Nagios alert, admin -report (command), cloudera manage

8) Heat beat, whether it is sequential processing or parallel processing.

9) What is the volume of data you receive to the cluster.

40 to 50GB

10) How do you receive data to your cluster. 11) What is your cluster size.

12) What is the port number of namenode.

13) What is the port number of Job tracker.

14) How do you install hive, pig, hbase.

15) What is JVM?

16) How do you do rebalancing.

Company: Verizon 02-Oct-2017

1)How do you dopaswordless SSH in hadoop.

2) Upgrades (Have you done anytime).

3) ClouderaManager port number.

4) what is your cluster size.

5) Versions

6) Map reduce version.

7) Daily activities.

8) What operations, you normally use in cloudera manager.

9) is internet connected to your nodes.

10) Do you have different cloudera managers for dev and production. 11) what are installation steps

Company: HCL 22-Sep-2017

1) Daily activities.

2) versions.

3) What is decommissioning.

4) What is the procedure to decommission datanode.

5) Difference between MR1 and MR2.

6) Difference between Hadoop1 and Hadoop2.

7) Difference between RDBMS and No-SQL.