```
'General Questions'
-----------------------
1. Tell About Yourself
2. Rate Yourself in
    Spark
    Scala
    Hive
    Sql
    Unix
3. Tell About your current Project
4. How many years of experience in Spark and Big data Ecosystem
5. What are roles and Responsibility of you in your team
6. Explain your Dev and Production cluster
7. What version that your using for
    Spark
    Hive
    Scala
    Hadoop


'Spark '
--------

1. What is RDDs and why they are immutable
2. What is Data Frame
3. What is Data Set
4. Difference between RDDs and Data Frame
5. Difference between Spark 1.0 and Spark 2.0
6. Difference Between Repartitions and coalsec
7. Different kinds of Transformation and Different types of Transformation
8. Different Actions
9. Features of RDD
10. Performance Tuning in Spark
11. Difference between Persist vs cache
12. What is Spark SQL
13. How Fault tolerant achieved in Spark
14. What version you are using in Spark
15. Code Sample 1.x and 2.x
16. What is Lineage Graph in Spark and how does it helps in fault tolerant
17. Why Data Set are faster than Data Frame
18. Role of Encoder and working or Encoder
19. How Spark is Better than Hadoop
20. Explain Spark Architecture and Spark Ecosystem
21. What is Main Abstraction of Spark
22. How to Integrate Hive and Spark ? And What are its advantages
23. Pair RDD and Differenet Transformation
24. lazy Evaluation in Spark and its benefits
25. Json in Hive and Spark
26. Join Example using Spark Core and Spark SQL
27. What is Project Tungsten in Spark
28. Why we wont use collect() in production code
29. Does Spark Requires Hadoop or not ? Explain
30. What is Broadcast Variable and Accumulators and What are its usage
31. Where you used Apache Spark in your Project
32. Explain Catalyst Framework
33. What are advantages of Parquet File format
34. Why kairo Serialization is better the Default Java Serialization
35. Checkpointing in Spark
36. MLib in your Project ?
37. Fold Operation in Spark
38. How Spark Can you be used for Data Extraction from RDBMS,
    How it is better than Sqoop
39. Roles and Responsibility of
    1. Driver
    2. Executor
    3. Worker Node
```

```
40. Spark Submit Job Command
41. Explain Apache Streaming and How it is Achieved
42. Explain D-Stream
43. What is Speculative Execution in Spark
44. What are the Machine Learning algorithm is possible in Spark
45. Difference between Spark Session and Spark Context
46. How do you do logging in Spark Job and how to retrieve
47. Difference Betwen
    a. SoryByKey vs distributeByKey
    b. Map vs Map Partition
    c. Map Partition vs Map Partition with Index
    d. Repartitions vs coalsec
48. How to Identify shuffling in spark
49. Common Mistake developers make when it comparately
50. Difference between Spark SQL and Hive
51. Explain sliding window operations
52. Why there are no indexes in spark Sql
53. How Memory Handled in Data Sets
54. What is Data Piping
55. How Data Security Achieved in Spark
56. Explain Kerberos Security
57. How Execution Starts and Ends of Spark
58. MEMORY_ONLY_2 (2 MEANS WHAT )
59. Dependencies in RDD
60. What is DAGSchedular
61. What is task with respect to Spark Job Execution
62. Explain Data Locality with respect to Spark



'Scala'
--------

1. Features of Scala
2. What is closure
3. What is currying
4. Method Overiding and Method overloading
5. Difference between val and var
6. How Exception can be handled in Scala
7. What are different transformation in scala
8. What is Higher Order Functions
9. What do you mean by First class Functions
10. How to process XMLs in Scala
11. Advantages of Scala over other Languages
12. What is differenec between concurrency and parallilism
13. What is Difference between Nil,Null,None,Nothing
14. Explain Data types in Scala
15. Explain
    a. Singleton Object
    b. class
    c. traits
16. Recursion problem in scala
17. What do you understand by case class in scala
18. Advantages of Having immutability in scala
19. Why Scala preferred than python
20. Explain scala collection
21. Explain Object Main Extends App means
22. what is unit in Java
23. Program to Explain
    a. If Else
    b. For Loop
    c. case statement
24. How does yield work
25. Explain fold left and fold right
26. How do you handle regular expression in scala
27. What Testing framework that you use in scala
```

```
28. Explain Scala Collections
    a. Sets
    b. Map
29. Main Advantage of Scala
30. Explain Annotations
31. Explain Singleton and Companion objects
32. Explain String Interpolation
33. Explain Exception Handling in Scala
34. Write a Producer and Combiner code in scala



'Hive'
--------

1. What is Difference between partition and bucketing
2. what is different join operations avaialble in Hive
3. What is static and Dynamic partition
4. What is Different Join
    a. Map Side join
    b. Bucket Map Join
    c. ??
5. Difference between order by , sort by , distribute by , cluster by
6. How do we intergrate Hive with Spark
7. Difference between Managed Tables and External Tables
8. Different indexes in Hive
9. How to create a Schema for the Data in Hive
10. What are different Data types in Hive
11. How to Select Complex Data Types in Hive
12. How to create Partition Table for Date column
13. Why Hive is not suitable for OLTP Applications
14. What is Metastore in Hive & What is the Metastore in that you used. And How do you configure
15. When you should use Sort by instead of Order by
16. What is Partitioning and when do you perform Partitioning
17. What is bucketing and when do you use bucketing
18. Explain Hive Indexing
19. Explain Different types of Joins in Hive
20. Explain
    a. Bucket Map Join
    b. Skew Join
    c. Sort Merge Bucket Join
21. Explain SORT BY, ORDER BY, DISTRIBUTE BY and CLUSTER BY with Example
22. How do process query for
    a. XML
    b. Json
    c. CSV
23. What are complex data types and how do you query Hive Collections
24. Explain What are the Optimization Technique Avaialble in Hive
25. Explain Views in Hive
26. Did you used UDFs in Hive
27. What is Beelime
28. What version of Hive you used in your organization
29. What is Impala
30. Explain Different SET Operations in Hive
31. Why do you drop a External Table
32. Explain Serde in Hive
33. What are File Formats supported by Hive
34. Explain variables in Hive
35. Explain How do you insert Date in Hive Table
36. Explain Analytical functions in Hive
37. How do you delete Duplicates in Hive
38. Explain Architecture of Hive
39. What is Apache HCatalog
40. What is Hive Current Version and What is Hive stable Version
41. Difference between SQL and HQL
42. How do you pull the Oracle data into Hive
```

```
43. How to integrate Hive with Spark



'Sqoop'
--------
1. How to Import Query data into HDFS
2. How to Import Data from Oracle to Hive Table or Hive Partitions
3. How to do incremental import using sqoop
4. How to craeate job or store the last value and retrieve in sqoop
5. How to set the boundry in sqoop
6. How to import data into HBase
7. Boundary Query
8. $CONDITIONS
9. --where
10. Append and overwrite Directo
ry (overwrite doesnot exist, we need to handle separatlely in shell)
11. How to do Incremental load or delta load
12. Insert/update in Sqoop Incremental
    Why update not work in sqoop
13. Integeration of Hive with Sqoop
14. How you query using sqoop
15. How to pull all the tables using sqoop
16. What are file formats supported by sqoop
17. Does Sqoop supports CLOB Columns
18. Different Options avaialbe in sqoop
19. What is better sqoop or Spark pull
20. How you do incremental pull using sqoop job
21. How to Handle Null in sqoop import
22. Explain --append option in sqoop
23. Explain free form query in sqoop
24. Difference between --target-dir --warehouse-dir
25. How to store and use last value in sqoop job
26. How to used password file
27. where you should copy the jars
28. How to exclude table in import all
29. How to increase number of mappers
30. how to do compression
31. Is it possible to update record using sqoop
32. Export and Import Data from and to Oracle
33. Export and Import Data from and to Hive
34. Export and Import Data from and to Hbase
35. Export and Import Data from and to Hive
36. The nine functions of Sqoop?
    A.  Full Load
    B.  Incremental Load
    C.  Parallel import/export
    D.  Import results of SQL query
    E.  Compression
    F.  Connectors for all major RDBMS Databases
    G.  Kerberos Security Integration
    H.  Load data directly into Hive/Hbase
    I.  Support for Accumulo
37. Default number of parallel jobs
38. Explain

    --append
    --as-avrodatafile
    --as-sequencefile
    --as-textfile
    --boundary-query
    --columns
    --direct
    --direct-split-size
    --inline-lob-limit
    ---m
    --e,--query
    --split-by
    --table
    --target-dir
    --warehouse-dir
    --where
    --compress
    --compression-codec
    --null-string
    --null-non-string

'HDFS'

1. What is Data Locality
2. Difference between 1.0 vs 2.0
3. Explain the Architecture of 2.0
4. Explain the role of YARN
5. What is the Issue with Hadoop 1.0.
6. How Name node single point of failure is rectified in Hadoop 2.0
7. Why block size is 128 KB in Hadoop
8. Exaplain
    a. Edit logs
    b. FSImage
9. Explain how fault tolerant is achieved in Hadoop
10. Why Hadoop
11. Explain Heartbeat in Hadoop
12. Explain the replication factor in Hadoop
13. Explain Safe mode in Hadoop
14. Explain Small file problem in Hadoop
15. Why Hadoop is less costly
16. Explain Rack Awareness in Hadoop
17. Explain the Daemons of Hadoop
18. What are 4 configuration files in Hadoop
19. Commands
    a. copyFromLocal
    b. moveFromLocal
    c. put
    d. get
    e. copyToLocal
    f. moveToLocal
    g. get
    h. put
    i. mkdir
    j. ls
    h. append
    i. setrep
    j. mv
    k. put
    l. rm
    m. fsck
20. What do you know abou Speculative Execution


'MR'
1. In Map Reduce ideally how many mappers should be configured on a slave
2. How to set no of Mappers in Map Reduce
3. Where is output of Mappers Stored
4. What is Partitioner and Combiner
5. Explain shuffling and sorting
6. Explain input split
7. Explain Record Reader
8. Explain Reducer
9. Is map only job possible
10. Explain Distrubuted Cache
11. Write a word count problem in Map reduce
```

```
'KAFKA' https://mindmajix.com/apache-kafka-interview-questions
-------
https://data-flair.training/blogs/kafka-interview-questions/



1. Explain Different components of KAFKA
2. Explain role of offsetin Kafka
3. Explain consumer group
4. Explain role of zookeeper
5. Explain the term of leader and follower in Kafka Environment
6. Why Replications are important in Kafka
7. Explain Kafka Architecture
8. Explain Partitioning Key
9. Advantages of Kafka
10. Explain
    a. Producer
    b. Consumer
    c. Broker
    d. topic
    e. partition
11. Main components where the data is processed seamlessly in kakka
12. Difference between Kafka and flume
13. Why Kafka is better than flume
14. ISR in Kafka
15. Key advantages of Kafka
16. How to create a topic in kafka
17. how to start zookeeper
18. What is default retension period of Kafka Broker
19. How do intergrate Spark Streaming with Kafka
20. How to make RDBMS or Producer
    and RDBMS as consumer



'PIG'
-----

1. Difference between PIG and Hive
2. Explain ( ILLUSTRATE,DESCRIBE,EXPLAIN,Define)
3. What are the Data types avaialble in PIG
4. Explain What are the transformation avaialble in PIG
    a. Distinct
    b. filter
    c. for each
    d. order by
    e. group
    f. cogroup
    g. Join
        join
        left outer Join
        Right outer Join
        Full outer join
        cross
    h. limit
    i. Union
    j. split
5. Explain Data types avaialble in PIG
6. Explain Flatten in PIG
7. How do you process below formats using PIG
    a. JSON
    b. CSV
    c. XML
8. Scenerios that we can you PIG
9. Explain Tuple ,Bag and Map
```

```
10. Is PIG case sensitive
11. Explain Architecture of PIG
12. Use Cases of PIG
13. How fileds are referenced in PIG when schema is not avaialble
14. What are Different in-built functions avaialble in PIG
15. Difference between group and cogroup
16. How to get the metadata
17. UDFx in Pig
18. How do you create pig script and run
19. How to read and store the data
20. How do you store processed data in Hive


'SQL Questions '

1. What is Different types of SQL Statement
2. What are the different Database objects you know
3. What is View ? Types ? and how it is different from Table
4. What is Materialized view and What are the types of refreshed method
5. Difference between view and MV
6. What is Partition and what are different types of partion can be added to table
7. Explain advantage of Using Partitioning in Oracle
8. Exaplain use of Indexes and Different types of Indexes
9. Difference between B-tree and Bitmap Index
10. What do you mean by local and global index
11. What is Synonym and what are the types of synonoyms
12. What you mean by DB-link
13. What are the Data Dictionary tables avaialble in Oracle
14. What are the Different constraints available in Oracle
15. What is different between Table level and column level constraint
16. Use of Sequences
17. What is the Oracle version that you are currently using
18. Explain
    a. DDL
    b. DML
    c. DRL
    d. DCL
19. What are the pre-defined data types avaialble in oracle
    a. Character
    b. Numberic
    c. Date
    d. What are aggregate function
20. Explain working of
    a.  Co-related sub queries
    b.  group by query
21. Explain Different types of Joins available in Oracle
22. How do you delete duplicates from the table
23. Explain Locking mechanism in oracle
24. Explain Use of Global Temporary table (GTT)
25. Difference between  Rank() and Dense_Rank()
26. Explain Use of RowNumber() and Rowid
27. Practice Hierarchiel queries
28. Use of LISTAGG() Queries -- Practice 3 Queries
29. Difference between RowNumber() and rownum
30. Explain the working for B-tree
31. Difference between Delete, Truncate and Drop
32. Explain ACID properties
33. Explain use of Decode() and case
34. Difference between SGA and PGA
35. Explain Complete flow of
     select * from emp ;
36. Explain complete working of
     update emp set ename='VISHAL' where empno=7900;
37. Explain Merge Operation in Oracle.
38. Explain Current of Operation in Oracle
39. Explain types of Sub-Query in Oracle
```

```
40. Explain On Delete null and On delete cascade.
41. Difference between varchar vs varchar2 vs Nvarchar2
42. Explain Pseudo Columns in Oracle
43. Explain Sub-partitioning in Oracle.
44. Explain
    a. Hard Parse
    b. soft parse
45. Explain with respect to oracle Architecture
    a. Blocks
    b. segments
    c. Extents
    d. Data Files
    e. Tablespace
46. Various Hints in Oracle
47. Page no 148 to 185
48. How do you create table faster in Oracle
49. Basic checks you do to improve performance of query
50. Normalization and its Types.
51. Nth Highest Paid Employee
52. Employees with Maximum salary in Each Department
53. Explain
    a. Union
    b. Union all
    c. Intersection
    d. Minus
54. Difference betweeen user_*, all_* and dba_* data Dictionary objects
55. Explain Difference Keys in Oracle


PL/SQL
-------

1. What is the Use of PL/SQL ? What are the Advantages
2. Write an annonyms blocks to update an Employee
3. What are
    a. Procedure
    b. Functions
    c. Packages
    and what are scenarios that above are used
4. Difference between Functions and Procedure
5. What is context switching
6. What is Bulk collect and Bulk Exception
    And when it is used and what is its significance
7. What is Trigger and what are the different types of triggers
8. What is mutating table error
9. Can we use commit in trigger ? Justify the Answer
10. What is Cursor and its types
11. Explain Parameterized cursor
12. What is Ref-Cursor
13. What are Excpetion ? List pre-defined Exception
14. Explain Raise vs Raise Application Error
15. Use of SQLCODE , SQLERRM
16. How do you find the line no Error in PL/SQL -->DBMS_SQLBACKTRACE
17. Collections in PL/SQL
18. Explain Pragma Autonomous Transaction
19. Use of Pragma Exception_INT
20. Modes of Paramter
    a. In
    b. In-out
    c. out
21. Types of Notations
22. Explain Overloaing Procedurs
23. Explain Dynmaic SQL in PL/SQL
24. How do you perform DDL in PL/SQL
25. Check SQL%ROW_COUNT Usage in PL/SQL
26. What are PL/SQL Datatypes
```

```
27. Difference between %ROWTYPE AND %TYPE
    AND Explain both
28. Practice Example
    a. Function
    b. Procedure
    c. Package
    d. Bulk Collect
    e. Bulk collect with Exception
    f. Collectiosn
    g. Cursor
    h. Excpetion
    g. Autonomous Transaction
    h. Dynamic SQL
    i. IF , IF-ELSE
    J. for loop
29. Check Error logging mechanism in Exception from Steven Feuerstein.
30. DBMS Scheduler Jobs in Oracle
31. Doing Activities Fast , Read more on it

    a. Create table with parallel 32 and nologging
    b. Insert /*+ Append*/
    c. create index with parallel 32 and nologging
    d. Disable any triggers while loading any data into table
    e. Parallel session using Shell script and primary key columns


'Data Warehouse'
---------------------
1. What is Surrogate Key
2. What is Normalization and its types
3. What is SCD ? Type 1 and Type 2 Dimention
4. Explain Star Schema
5. Explain Snowflake Schema
6. Explain
    a. Junk Dimentions
    b. Confimed Dimensions
    c. Denerated Dimensions
7. What is ETL


'UNIX'
```