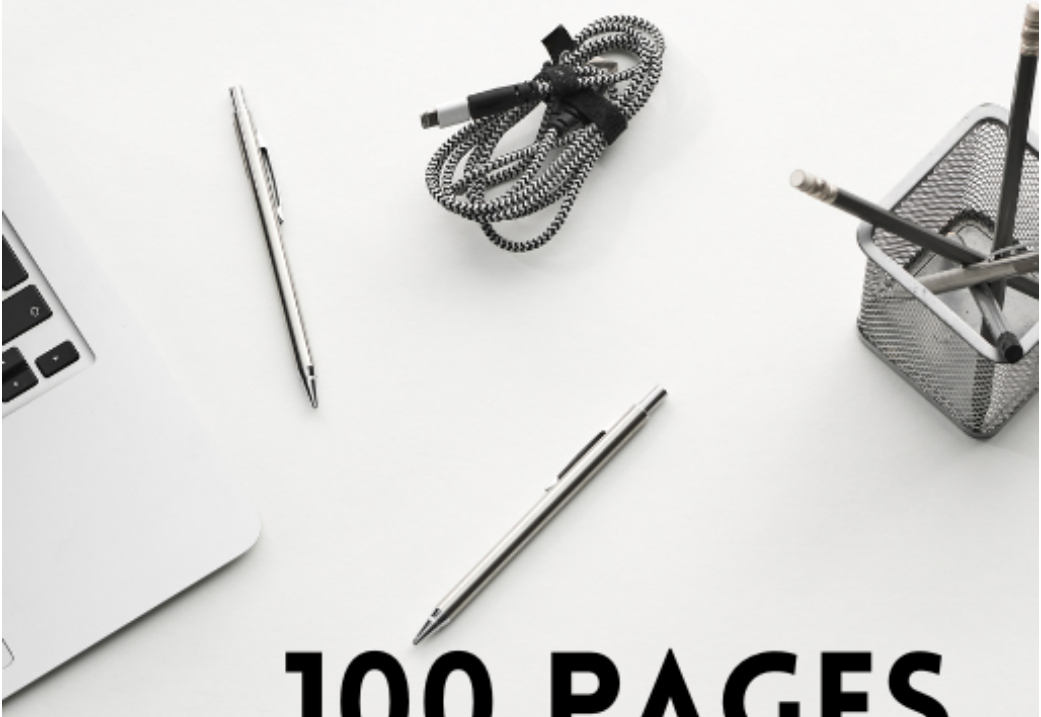

100 PAGES OF DATA ENGINEERING Q&A

All about Data Engineering

Interview Questions

1. What is Hadoop MapReduce?

A.) For processing large datasets in parallel across hadoop cluster, hadoop mapReduce framework is used.

2. What are the difference between relational database and HDFS?

A.) There are 6 major categories we can define RDMBS and HDFS. They are

a. Data Types

b. processing

c. Schema on read Vs Write

d. Read/write speed

e. cost

Best fit use case

RDBMS	HDFS
1. In RDBMS it relies on structured data and schema is always known.	any kind of data can be stored into Hadoop. i.e structured, unstructured, semi-structured.
2. Rdbms provides limited or no processing capabilites.	hadoop allows us to process the data which is distributed across the cluster in a parallel fashion.
3. Rdbms is based schema on write.	Hadoop follows schema on read policy.
4. in rdbms reads are fast because the schema is already known.	writes are fast in hadoop because no schema validation happens during hdfs write.
5. Licensed software, therefore, need to pay for software.	Hadoop is open source framework, hence no need to pay for software.
6. Rdbms is used for OLTP(online transactional processing) system.	Hadoop is used for data discovery, data analytics or OLAP System.

3. Explain Bigdata and explain 5v's of bigdata?

A.) Bigdata is a term for collection of large and complex datasets, that makes it to difficult processing using relational database management tools or traditional data processing applications. It is difficult capture, visualize, curate, store, search, share, transfer and analyze bigdata.

IBM has defined bigdata with 5v's they are.

a. Volume

b. velocity

c. variety

d. veracity

e. Value: It is good to have access to bigdata but unless we turn into value it is useless. which means using bigdata adding benifits to the organizations, and are they seen ROI using Bigdata.

4. What is Hadoop and its components?

A.) When bigdata emerged as a problem, Hadoop evolved as Solution to bigdata. Apache hadoop is a framework which provides us various services or tools, to store and process the bigdata.

It helps in analyzing bigdata and making bussiness decisions out of it, which cannot be done using traditional systems.

Main components in Hadoop are

a. Storage(Namenode, DataNode)

b. Processing framework yarn (resource manager, Node Manager)

5. What are HDFS and Yarn?

A.) HDFS (Hadoop Distributed File System) is the storage unit of Hadoop. It is responsible for storing different kinds of data as blocks in a distributed environment.

It follows master and slave topology.

NameNode: NameNode is the master node in the distributed environment and it maintains the metadata information for the blocks of data stored in HDFS like block location, replication factors etc.

DataNode: DataNodes are the slave nodes, which are responsible for storing data in the HDFS. NameNode manages all the DataNodes.

YARN (Yet Another Resource Negotiator) is the processing framework in Hadoop, which manages resources and provides an execution environment to the processes.

ResourceManager: It receives the processing requests, and then passes the parts of requests to corresponding NodeManagers accordingly, where the actual processing takes place. It allocates resources to applications based on the needs.

NodeManager: NodeManager is installed on every DataNode and it is responsible for the execution of the task on every single DataNode.

6. Tell me about various Hadoop Daemons and their roles in hadoop cluster?

A.) Generally approach this question by first explaining the HDFS daemons i.e. NameNode, DataNode and Secondary NameNode, and then moving on to the YARN daemons i.e. ResorceManager and NodeManager, and lastly explaining the JobHistoryServer.

JobHistoryServer: It maintains information about MapReduce jobs after the Application Master terminates.

7. Compare HDFS with Network attached servive(NAS)?

A.) Network-attached storage (NAS) is a file-level computer data storage server connected to a computer network providing data access to a heterogeneous group of clients. NAS can either be a hardware or software which provides services for storing and accessing files.

Whereas Hadoop Distributed File System (HDFS) is a distributed filesystem to store data using commodity hardware.

In HDFS Data Blocks are distributed across all the machines in a cluster.

Whereas in NAS data is stored on a dedicated hardware.

8. List the difference between Hadoop 1.0 vs Hadoop 2.0?

A.) To answer this we need to highlight 2 important features that are a. PassiveNode b. Processing.

In Hadoop 1.x, “NameNode” is the single point of failure. In Hadoop 2.x, we have Active and Passive “NameNodes”. If the active “NameNode” fails, the passive “NameNode” takes charge. Because of this, high availability can be achieved in Hadoop 2.x.

Also, in Hadoop 2.x, YARN provides a central resource manager. With YARN, you can now run multiple applications in Hadoop, all sharing a common resource. MRV2 is a particular type of distributed application that runs the MapReduce framework on top of YARN. Other tools can also perform data processing via YARN, which was a problem in Hadoop 1.x.

9. What are active and Passive Namenodes?

A.) In a High availability architecture, there are 2 namenodes. i.e.

a. Active “NameNode” is the “NameNode” which works and runs in the cluster.

b. Passive “NameNode” is a standby “NameNode”, which has similar data as active “NameNode”.

When the active “NameNode” fails, the passive “NameNode” replaces the active “NameNode” in the cluster.

10. Why does one remove or add datanodes freaquently?

A.) most attractive features of the Hadoop framework is its utilization of commodity hardware. However, this leads to frequent “DataNode” crashes in a Hadoop cluster.

Another striking feature of Hadoop Framework is the ease of scale in accordance with the rapid growth in data volume. this is why hadoop admins work is to commission or decommission nodes in hadoop cluster.

11.) what happens when two clients tries to access same file in Hdfs?

A.) When first client request for file or data hdfs provides access to write, but when second client request it rejects by saying already another client accessing it.

12. How does nameNOde tackles data node failures?

A.) NameNode periodically receives a Heartbeat (signal) from each of the DataNode in the cluster, which implies DataNode is functioning properly.

13. What will you do when NameNode is down?

A.) Use the file system metadata replica (FsImage) to start a new NameNode.

Then, configure the DataNodes and clients so that they can acknowledge this new NameNode, that is started.

14. What is checkpoint?

A.) Checkpointing” is a process that takes an FsImage, edit log and compacts them into a new FsImage.

Thus, instead of replaying an edit log, the NameNode can load the final in-memory state directly from the FsImage.

15. what is Hdfs fault tolerant?

A.) When data stored in over hdfs, namenode replicates the data in server datanode, which has default value of 3.

if any DataNode fails Namenode automatically copies data to another datanode to makesure data is fault tolerant.

16. can NameNode and dataNode are commodity hardware?

A.) No.. beacuse Namenode built in with high memory space, and with higher quality software.

but datanode built in with cheaper hardware.

17. Why do we use Hdfs for files with large data sets but not when there are lot of small files?

A.) NameNode stores the metadata information regarding the file system in the RAM.

Therefore, the amount of memory produces a limit to the number of files in my HDFS file system. In other words, too many files will lead to the generation of too much metadata.

And, storing these metadata in the RAM will become a challenge. hence hdfs is only works with large datsets instead large no.of small.

18. How do you define block, and what is the default block size?

A.) Block are nothing but smallest continuous locations in harddrive where data is stored.

default block size of hadoop 1 is 64 mb

and hadoop 2 is 128 mb.

19. How do you define Rack awareness in hadoop?

A.) Rack Awareness is the algorithm in which the “NameNode” decides how blocks and their replicas are placed, based on rack definitions to minimize network traffic between “DataNodes” within the same rack.

20. What is the difference between hdfs block, and input split?

A.) The “HDFS Block” is the physical division of the data while “Input Split” is the logical division of the data.

Hdfs block divides data into blocks to store the blocks together processing, where Input split Divides the data into the input split and assign it to the mapper function for processing.

21. Name of three modes which hadoop can run?

A.) Stanalone mode

pseudo- distribution mode

fully distributed mode

29. What do you know about SequenceFileFormat?

A.) “SequenceFileInputFormat” is an input format for reading within sequence files.

It is a specific compressed binary file format which is optimized for passing the data between the outputs of one “MapReduce” job to the input of some other “MapReduce” job.

30. What is Hive?

A.) Apache Hive is a data warehouse system built on top of Hadoop and is used for analyzing structured and semi-structured data developed by Facebook.

Hive abstracts the complexity of Hadoop MapReduce.

31. what is Serde in Hive?

A.) The “SerDe” interface allows you to instruct “Hive” about how a record should be processed.

A “SerDe” is a combination of a “Serializer” and a “Deserializer”.

“Hive” uses “SerDe” (and “FileFormat”) to read and write the table’s row.

31. can the default hive metastore used by multiple users at the same time?

A.) “Derby database” is the default “Hive Metastore”.

Multiple users (processes) cannot access it at the same time.

It is mainly used to perform unit tests.

32. what is the default location for hive to store in table data?

A.) The default location where Hive stores table data is inside HDFS in /user/hive/warehouse.

33. What is Apache Hbase?

A.) HBase is an open source, multidimensional, distributed, scalable and a NoSQL database written in Java.

HBase runs on top of HDFS (Hadoop Distributed File System) and provides BigTable (Google) like capabilities to Hadoop.

It is designed to provide a fault-tolerant way of storing the large collection of sparse data sets.

HBase achieves high throughput and low latency by providing faster Read/Write Access on huge datasets.

34. What are the components of apache Hbase?

A.) HBase has three major components, i.e. HMaster Server, HBase RegionServer and Zookeeper.

Region Server: A table can be divided into several regions. A group of regions is served to the clients by a Region Server.

HMaster: It coordinates and manages the Region Server (similar as NameNode manages DataNode in HDFS).

ZooKeeper: Zookeeper acts like as a coordinator inside HBase distributed environment.

It helps in maintaining server state inside the cluster by communicating through sessions.

35. what are the components of Region server?

A.) WAL: Write Ahead Log (WAL) is a file attached to every Region Server inside the distributed environment. The WAL stores the new data that hasn't been persisted or committed to the permanent storage.

Block Cache: Block Cache resides in the top of Region Server. It stores the frequently read data in the memory.

MemStore: It is the write cache. It stores all the incoming data before committing it to the disk or permanent memory. There is one MemStore for each column family in a region.

HFile: HFile is stored in HDFS. It stores the actual cells on the disk.

36. What is the difference between Hbase and Relation database?

A.) HBase is an open source, multidimensional, distributed, scalable and a NoSQL database written in Java.

Hbase	Relational Database
1. It is schema-less	It is schema-based database.
2. It is column-oriented data store	It is row-oriented data store.
3. It is used to store de-normalized data	It is used to store normalized data.
4. Automated partitioning is done is HBase support for partitioning.	There is no such provision or built-in

35. What is Apache Spark?

A.) Apache Spark is a framework for real-time data analytics in a distributed computing environment.

It executes in-memory computations to increase the speed of data processing.

It is 100x faster than MapReduce for large-scale data processing by exploiting in-memory computations and other optimizations.

36. can you build Spark with any particular Hadoop version?

A.) yes. spark can be built with any version of Hadoop.

37. What is RDD?

A.) RDD is the acronym for Resilient Distribution Datasets – a fault-tolerant collection of operational elements that run parallel.

The partitioned data in RDD are immutable and distributed, which is a key component of Apache Spark.

38. Are Hadoop and Bigdata are co related?

A.) Big Data is an asset, while Hadoop is an open-source software program, which accomplishes a set of goals and objectives to deal with that asset.

Hadoop is used to process, store, and analyze complex unstructured data sets through specific proprietary algorithms and methods to derive actionable insights.

So yes, they are related but are not alike.

39. why is Hadoop used in bigdata analytics?

A.) Hadoop allows running many exploratory data analysis tasks on full datasets, without sampling. Features that make Hadoop an essential requirement for Big Data are –

Data collection

Storage

Processing

Runs independently.

40. Name of some of the important tools used for data analytics?

A.) The important Big Data analytics tools are –

NodeXL

KNIME

Tableau

Solver

OpenRefine

Rattle GUI

Qlikview.

41. what is FSCK?

A.) FSCK or File System Check is a command used by HDFS.

It checks if any file is corrupt,or if there are some missing blocks for a file. FSCK generates a summary report, which lists the overall health of the file system.

42. what are the different core methods of Reducer?

A.) There are three core methods of a reducer-

setup() – It helps to configure parameters like heap size, distributed cache, and input data size.

reduce() – Also known as once per key with the concerned reduce task. It is the heart of the reducer.

cleanup() – It is a process to clean up all the temporary files at the end of a reducer task.

43. what are the most common Input fileformats in Hadoop?

A.) The most common input formats in Hadoop are –

Key-value input format

Sequence file input format

Text input format.

44. what are the different fileformats that can be used in Hadoop?

A.) File formats used with Hadoop, include –

CSV

JSON

Columnar

Sequence files

AVRO

Parquet file.

45. what is commodity hardware?

A.) Commodity hardware is the basic hardware resource required to run the Apache Hadoop framework.

It is a common term used for affordable devices, usually compatible with other such devices.

46. what do you mean by logistic regression?

A.) Also known as the logit model, logistic regression is a technique to predict the binary result from a linear amalgamation of predictor variables.

47. Name the port number for namenode, task tracker, job tracker?

A.) NameNode – Port 50070

Task Tracker – Port 50060

Job Tracker – Port 50030.

48. Name the most popular data management tools that used with edge nodes in hadoop?

A.) The most commonly used data management tools that work with Edge Nodes in Hadoop are –

Oozie

Ambari

Pig

Flume.

49. what is block in Hadoop distributed file system?

A.) When the file is stored in HDFS, all file system breaks down into a set of blocks.

50. what is the functionality of jps command?

A.) The ‘jps’ command enables us to check if the Hadoop daemons like namenode, datanode, resourcemanager, nodemanager, etc. are running on the machine.

51. what types of biases can happen through sampling?

A.) Three types of biases can happen through sampling, which are –

Survivorship bias

Selection bias

Under coverage bias.

52. what is the difference between Sqoop and distcp?

A.) DistCP is used for transferring data between clusters, while Sqoop is used for transferring data between Hadoop and RDBMS, only.

53.How much data is enough to get a valid outcome?

A.) The amount of data required depends on the methods you use to have an excellent chance of obtaining vital results.

54. Is Hadoop is different from other parallel computing systems? How?

A.) Yes, it is. Hadoop is a distributed file system. It allows us to store and manage large amounts of data in a cloud of machines, managing data redundancy.

The main benefit of this is that since the data is stored in multiple nodes, it is better to process it in a distributed way. Each node is able to process the data stored on it instead of wasting time moving the data across the network.

In contrast, in a relational database computing system, we can query data in real-time, but it is not efficient to store data in tables, records, and columns when the data is huge.

Hadoop also provides a schema for building a column database with Hadoop HBase for run-time queries on rows.

55. What is BackUp Node?

A.) Backup Node is an extended checkpoint node for performing checkpointing and supporting the online streaming of file system edits.

Its functionality is similar to Checkpoint, and it forces synchronization with NameNode.

56. what are the common data challenges?

A.) The most common data challenges are –

Ensuring data integrity

Achieving a 360-degree view

Safeguarding user privacy

Taking the right business action with real-time resonance.

57. How do you overcome above mentioned data challenges?

A.) Data challenges can be overcome by –

Adopting data management tools that provide a clear view of data assessment

Using tools to remove any low-quality data

Auditing data from time to time to ensure user privacy is safeguarded

Using AI-powered tools, or software as a service (SaaS) products to combine datasets and make them usable.

58. What is the hierarachical Clustering algorithm?

A.) The hierarchical grouping algorithm is the one that combines and divides the groups that already exist.

58. what is K- Mean clustering?

A.) K mean clustering is a method of vector quantization.

59. can you mention the crieteria for good data model?

A.) A good data model –

It should be easily consumed

Large data changes should be scalable

Should offer predictable performances

Should adapt to changes in requirements.

60. Name the different commands for starting up and shutting down the hadoop daemons?

A.) To start all the daemons:

./sbin/start-all.sh

To shut down all the daemons:

./sbin/stop-all.sh

61. Talk about the different tombstone markers used for deletion purpose in Hbase?

A.) There are three main tombstone markers used for deletion in HBase. They are-

Family Delete Marker – For marking all the columns of a column family.

Version Delete Marker – For marking a single version of a single column.

Column Delete Marker – For marking all the versions of a single column.

62. How can bigdata add value to bussinesses?

A.) ig Data Analytics helps businesses to transform raw data into meaningful and actionable insights that can shape their business strategies.

The most important contribution of Big Data to business is data-driven business decisions.

63. How do you deploy bigdata solution?

A.) we can deploy bigdata in 3 stages. they are

Data Ingestion: begin by collecting data from multiple sources, be it social media platforms, log files, business documents, anything relevant to your business.

Data can either be extracted through real-time streaming or in batch jobs.

Data storage: Once the data extracted, it can be stored in Hbase, or Hdfs.

While HDFS storage is perfect for sequential access, HBase is ideal for random read/write access.

Data processing: Usually, data processing is done via frameworks like Hadoop, Spark, MapReduce, Flink, and Pig, to name a few.

64. List the different file permissions in hdfs files or directory levels?

A.) There are three user levels in HDFS – Owner, Group, and Others. For each of the user levels, there are three available permissions:

read (r)

write (w)

execute(x).

65. Elaborate on the process that overwrite the replication factor in Hdfs?

A.) In HDFS, there are two ways to overwrite the replication factors – on file basis and on directory basis.

66. Explain overFitting?

A.) Overfitting refers to a modeling error that occurs when a function is tightly fit (influenced) by a limited set of data points.

67. what is feature selection?

A.) Feature selection refers to the process of extracting only the required features from a specific dataset.

When data is extracted from disparate sources.

Feature selection can be done via three techniques.

a. filters method

b. wrappers method

c. Embedded method.

68. Define OUtliers?

A.) outliers are the values that are far removed from the group, they do not belong to any specific cluster or group in the dataset.

The presence of outliers usually affects the behavior of the model.

Here are six outlier detection methods:

1. Extreme value analysis

2. probabilistic analysis

3. linear models

4. information-theoretic models

5. High-dimensional outlier detection.

70. How can you handle missing values in Hadoop?

A.) there are different ways to estimate the missing values.

These include regression, multiple data imputation, listwise/pairwise deletion, maximum likelihood estimation, and approximate Bayesian bootstrap.

MapReduce Interview Questions:

1. Compare MapReduce and SPark?

A.) there are 4 crieteria to be followed to compare MR with spark. they are.

1. processing speeds

2. standalone mode

3. Ease of use

4. versatility

MapReduce	Spark
1. processing speed is good	It is execeptional
2. standalone mode needs hadoop	it can work independently
3. it needs extensive java program	API for python & scala& java
4. It is optimized real time machine-learning applications.	not optimized real time & mL applications.

2. what is MapReduce?

A.) It is a framework/a programming model that is used for processing large data sets over a cluster of computers using parallel programming.

3. State the reason why we can't perform aggregation in mapper? why do we need reducer for this?

A.) We cannot perform “aggregation” (addition) in mapper because sorting does not occur in the “mapper” function, as sorting occurs only on reducer.

During “aggregation”, we need the output of all the mapper functions which may not be possible to collect in the map phase as mappers may be running on the different machine.

4. What is the recordReader in Hadoop?

A.) The “RecordReader” class loads the data from its source and converts it into (key, value) pairs suitable for reading by the “Mapper” task.

5. Explain Distributed cache in MapReduce Framework?

A.) Distributed Cache is a dedicated service of the Hadoop MapReduce framework, which is used to cache the files whenever required by the applications.

This can cache read-only text files, archives, jar files, among others, which can be accessed and read later on each data node where map/reduce tasks are running.

6. How do reducers communicate with each other?

A.) The “MapReduce” programming model does not allow “reducers” to communicate with each other.

6. What does mapReduce partitioner do?

A.) A “MapReduce Partitioner” makes sure that all the values of a single key go to the same “reducer”, thus allowing even distribution of the map output over the “reducers”.

It redirects the “mapper” output to the “reducer” by determining which “reducer” is responsible for the particular key.

7. How will you write custom partitioner?

A.) custom partitioner can be written in following ways.

Create a new class that extends Partitioner Class

Override method – getPartition, in the wrapper that runs in the MapReduce.

Add the custom partitioner to the job by using method set Partitioner.

8. What is combiner?

A.) A “Combiner” is a mini “reducer” that performs the local “reduce” task.

It receives the input from the “mapper” on a particular “node” and sends the output to the “reducer”.

9. what are main components of MapReduce?

A.) Main Driver Class: providing job configuration parameters

Mapper Class: must extend org.apache.hadoop.mapreduce.Mapper class and performs execution of map() method

Reducer Class: must extend org.apache.hadoop.mapreduce.Reducer class.

10. What is Shuffling and Sorting in MapReduce?

A.) Shuffling and Sorting are two major processes operating simultaneously during the working of mapper and reducer.

The process of transferring data from Mapper to reducer is Shuffling.

In MapReduce, the output key-value pairs between the map and reduce phases (after the mapper) are automatically sorted before moving to the Reducer.

11. What is identity mapper and Chain mapper?

A.) Identity Mapper is the default Mapper class provided by Hadoop.

It only writes the input data into output and do not perform and computations and calculations on the input data.

Chain mapper: Chain Mapper is the implementation of simple Mapper class through chain operations across a set of Mapper classes, within a single map task.

In this, the output from the first mapper becomes the input for second mapper

12. What main configuration parameters are specified in Mapreduce?

A.) following configuration parameters to perform the map and reduce jobs:

The input location of the job in HDFs.

The output location of the job in HDFS.

The input’s and output’s format.

The classes containing map and reduce functions, respectively.

The .jar file for mapper, reducer and driver classes.

13. Name Job control options specified by mapreduce?

A.) Since this framework supports chained operations wherein an input of one map job serves as the output for other.

The various job control options are:

Job.submit() : to submit the job to the cluster and immediately return

Job.waitforCompletion(boolean) : to submit the job to the cluster and wait for its completion.

14. What is inputFormat in hadoop?

A.) inputformat defines the input specifications for a job.it peforms following instructions.

1. validates input-specifications of job.

2. Split the input file(s) into logical instances called InputSplit.

3. Provides implementation of RecordReader to extract input records from the above instances for further Mapper processing.

15. What is the difference between Hdfs block and inputsplit?

A.) An HDFS block splits data into physical divisions while InputSplit in MapReduce splits input files logically.

16. what is the text inputformat?

A.) TextInputFormat, files are broken into lines, wherein key is position in the file and value refers to the line of text.

Programmers can write their own InputFormat.

17. what is role of job Tracker?

A.) The primary function of the JobTracker is resource management, which essentially means managing the TaskTrackers.

Apart from this, JobTracker also tracks resource availability and handles task life cycle management.

18. Explian jobconf in mapreduce?

A.) It is a primary interface to define a map-reduce job in the Hadoop for job execution.

JobConf specifies mapper, Combiner, partitioner, Reducer,InputFormat , OutputFormat implementations

19. what is output committer?

A.) OutPutCommitter describes the commit of MapReduce task.

FileOutputCommitter is the default available class available for OutputCommitter in MapReduce.

20. what is map in Hadoop?

A.) In Hadoop, a map is a phase in HDFS query solving.

A map reads data from an input location, and outputs a key value pair according to the input type.

21. what is reducer in hadoop?

A.) In Hadoop, a reducer collects the output generated by the mapper, processes it, and creates a final output of its own.

22. what are the parameters of mappers and reducers?

A.) The four parameters for mappers are:

LongWritable (input)

text (input)

text (intermediate output)

IntWritable (intermediate output)

The four parameters for reducers are:

Text (intermediate output)

IntWritable (intermediate output)

Text (final output)

IntWritable (final output)

23. What is partitioning?

A.) Partitioning is a process to identify the reducer instance which would be used to supply the mappers output.

Before mapper emits the data (Key Value) pair to reducer, mapper identify the reducer as an recipient of mapper output.

24. what is mapreduce used for-by company?

A.) construction of index for google search: The process of constructing a positional or nonpositional index is called index construction or indexing.

article clustering for google news: For article clustering, the pages are first classified according to whether they are needed for clustering.

statistical machine transalation.

25 what are the mapreduce design goals?

A.) scalability to large data volumes

cost-efficiency.

26. what are the challenges of Mapreduce?

A.) cheap node fails, specially if having many.

a commodity network is equao to implies.

programming distributed systems are hard.

27. what is the mapreduce programming model?

A.) MapReduce programming model is based on a concept called key-value records.

It also provides paradigms for parallel data processing.

28. what are the mapreduce execution details?

A.) In the case of MapReduce execution, a single master controls job execution on multiple slaves.

29. Mention benifits of Mapreduce?

A,) Highly scalable

cost-effective

secure.

30. is the renaming the output file possible?

A.) yes, the implementation of multiple format output class makes it possible to rename the output file.

Apache Sqoop Interview Questions:

1. Mention the best features of Apache Sqoop?

A.) Apache sqoop is a tool in hadoop ecosystem have several advantages. i.e.

Parallel import/export

Connectors for all major RDBMS Databases

Import results of SQL query

Incremental Load

Full Load

Kerberos Security Integration

Load data directly into Hive / HBase

Compression

Support for Accumulo

2. How can you import large objects like BLOB and CLOB in sqoop?

A.) The direct import function is not supported by Sqoop in case of CLOB and BLOB objects. Hence, if you have to import large purposes, you can use JDBC based imports.

This can be done without introducing the direct argument of the import utility.

3. What is default database of Apache sqoop?

A.) The default database of apache sqoop is MySQL.

4. Describe the process of executing free-form SQL query to import rows?

A.) To achieve a free-form SQL query, you have to use the –m1 option. This would create only one Mapreduce task.

This would then import the rows directly.

5. Describe the importance of using compress-codec parameter?

A.) The –compress-codec parameter can be used to get the export file of the Sqoop import in the required formats.

6. What is the significance of Sqoop eval tool?

A.) Sqoop eval can be against the database as it can preview the results that are displayed on the console. Interestingly, with the help of the Eval tool, you would be well aware of the fact that the desired data can be imported correctly or not.

7. What is the meaning of free form import in sqoop?

A.) With the use of Sqoop, one can import the relational database query. This can be done using column and table name parameters.

8. Describe the advantage of utilizing --password-file rather than -p option?

A.) The –password-file option is usually used inside the Sqoop script file. On the other hand, the –P option is able to read the standard input along with the column name parameters.

9. Is the JDBC driver fully capable to connect sqoop on the databases?

A.) The JDBC driver is not capable to connect Sqoop on the databases. This is the reason that Sqoop requires both the connector and JDBC driver.

10. what is the meaning of input split in Hadoop?

A.) Input Split is that kind of a function which is associated with splitting the input files into various chunks. These chunks can also assign each split to a mapper in the ongoing process of data correction.

11. Illustrate the utility of --help command in sqoop?

A.) The command in the sqoop can be utilized to list the various available commands.

12. what is Codegen commnad in sqoop?

A.) The Codegen command is associated with the generation of code so that it can appropriately interact with the database records.

13. Describe the procedure involved in executing an incremental data load in sqoop?

A.) the process of performing additional data load is to update the uploaded data. This data is often referred to as delta data. In Sqoop, this delta data can be altered with the use of incremental load command.

14. What is the default file format in order to import data with the utilization of apache sqoop?

A.) Delimiting the text File Format.

15. List all basic sqoop commands along with their properties?

A.) The basic controls in Apache Sqoop along with their uses are:

1. Export: This function helps to export the HDFS directory into a database table

2. List Tables: This function would help the user to list all tables in a particular database.

3. Codegen: This function would help you to generate code so that you can interact with varied types of database records.

4. Create: This function allows a user to import the table definition within the hive of databases.

5. Eval: This function would always help you to assess the SQL statement and display the results.

6. Version: This function would help you to depict the information related to the text of the database.

7. Import all tables: This function would help a user to import all the tables from a database to HDFS.

8. List all the databases: This function would assist a user to create a list of the available databases on a particular server.

16. what are the limitations of importing the RDBMS tables into Hcatlog directly?

A.) In order to import the tables into the Hcatalog in a direct manner, you have to make sure that you are using the –Hcatalog database option. However, in this process, you would face a limitation of importing the tables.

It is in the form of the fact that this option do not supports a plethora of arguments like –direct, –as-Avro file and -export-dir.

17. what is the procedure of updating the rows that have been directly uploaded?

A.) In order to update the existing rows that have been exported, we have to use parameter is in the form of update key

18. What is the significance of sqoop import Mainframe tool?

A.) The Sqoop Import Mainframe tool can also be used to import all the important datasets which lies in a partitioned dataset.

This tool would always help you to make sure that you are importing the right types of data tools and that too in a proper manner.

19. Define Metastore?

A.) It is also known as a shared metadata repository with the help of which the local users can execute and define various types of list tables.

In order to connect to the metastore, you have to make changes to the Sqoop –site.xml.

20. Does sqoop uses MapReduce Function?

A.) Apache Sqoop also uses the Map-Reduce function of Hadoop to obtain data from the relational databases.

During the process of importing data, Sqoop controls the mappers and their numbers.

21. Compare Sqoop and Flume?

A.) criteria	Sqoop	Flume
Application	Importing data from RDBMS	Moving bulk streaming data into HDFS
Architecture	Connector-connecting to respective data	Agent – fetching of the right data
Loading of data	Event driven	Not event driven

22. How can we import data from particular row or column?

A.) Sqoop allows to Export and Import the data from the data table based on the where clause.

23. Role of JDBC driver in sqoop setup?

A.) Sqoop needs a connector to connect the different relational databases. Almost all Database vendors make a JDBC connector available specific to that Database, Sqoop needs a JDBC driver of the database for interaction.

No, Sqoop needs JDBC and a connector to connect a database.

24. Using Sqoop command how can we control the number of Mappers?

A.) We can control the number of mappers by executing the parameter –num-mapers in sqoop command.

25. What is the purpose of Sqoop-merge?

A.) This tool combines 2 datasets where entries in one dataset overwrite entries of an older dataset preserving only the new version of the records between both the data sets.

26. Explain the Saved Job process in Sqoop?

A.) Sqoop allows us to define saved jobs which make this process simple. A saved job records the configuration information required to execute a Sqoop command at a later time.

sqoop-job tool describes how to create and work with saved jobs.

27. Sqoop is Which type of tool and main use of Sqoop?

A.) And Sqoop is a data transfer tool.

The main use of Sqoop is to import and export the large amount of data from RDBMS to HDFS and vice versa.

28. I am getting connection failure exception during connecting to Mysql through Sqoop, what is the root cause and fix for this error scenario?

A.) This will happen when there is lack of permissions to access our Mysql database over the network.

We can try the below command to confirm the connect to Mysql database from aSqoop client machine.

$ mysql –host=MySqlnode> –database=test –user= –password=

We can grant the permissions with below commands.

mysql> GRANT ALL PRIVILEGES ON *.* TO ‘%’@’localhost’;

mysql> GRANT ALL PRIVILEGES ON *.* TO ‘ ’@’localhost’;

29. I am getting java.lang.IllegalArgumentException: during importing tables from oracle database.what might be the root cause and fix for this error scenario?

A.) Sqoop commands are case- sensitive of table names and user names.

By specifying the above two values in UPPER case, it will resolve the issue.

30. Is Sqoop same as to distcp in Hadoop?

A.) No. Because the only the distcp import command is same as Sqoop import command and both the commands submit parallel map-only jobs but both command functions are different.

Distcp is used to copy any type of files from Local filesystem to HDFS and Sqoop is used for transferring the data records between RDBMS and Hadoop eco- system service.

31. I am having around 500 tables in a database. I want to import all the tables from the database except the tables named Table 498, Table 323, and Table 199. How can we do this without having to import the tables one by one?

A.) This can be proficient 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 Table498, Table 323, Table 199

32. Explian the significance of using -split-by clause in Sqoop?

A.) split-by is a clause, it is used to specify the columns of the table which are helping to generate splits for data imports during importing the data into the Hadoop cluster.

This clause specifies the columns and helps to improve the performance via greater parallelism

33. If the source data gets updated every now and then, how will you synchronize the data in HDFS that is imported by Sqoop?

A.) By using incremental parameter we can syncronize the data.

and also we can use append, and lastmodified modes to update existing data.

34. When to use target-dir and when to use warehouse-dir in sqoop?

A.) we use –target-dir to specify a particular directory in HDFS.

Whereas we use –warehouse-dir to specify the parent directory of all the sqoop jobs.

35. what is the purpose of validation in sqoop?

A.) In Sqoop, validating the data copied is Validation’s main purpose.

Basically, either Sqoop import or Export by comparing the row counts from the source as well as the target post copy.

36. what is accumulo in sqoop?

A.) The Accumulo in sqoop is a sorted, distributed key and value store. It provides robust, extensible data storage and retrieves data.

37. Explain relaxed isolation in sqoop?

A.) This is used to import the data which is read uncommitted for mappers.

The sqoop transfer committed data relational database to the Hadoop file system but with this argument, we can transfer uncommitted data in the isolation level.

38. What are reducers in Sqoop?

A.) The reducer is used for accumulation or aggregation.

In the sqoop there is no reducer because import and export work parallel in sqoop.

39. What is boundary query in sqoop?

A.) The boundary query is used for splitting the value according to id_no of the database table.

To boundary query, we can take a minimum value and maximum value to split the value.

To make split using boundary queries, we need to know all the values in the table.

40. What is sqoop?

A.) The sqoop is an acronym of SQL-TO-HADOOP.

Sqoop is a tool used to transfer to data between Hadoop and relational databases or vice versa.

Interview Questions

1. what is spark?

A.) Spark is General purpose, in memory compute engine.

General purpose: it can support any storage, any compute engine

in memory: Spark save storage in memory rather disc in MapReduce.

compute engine: It is plug and play compute engine.

2. Difference between spark & MR?

A.) Performance: Spark was designed to be faster than MapReduce, and by all accounts, it is; in some cases, Spark can be up to 100 times faster than MapReduce.

Operability: Spark is easier to program than MapReduce.

Data Processing: MapReduce and Spark are both great at different types of data processing tasks.

Failure Recovery

Security.

3. Explain the architecture of spark?

A.) Spark Architecture. The Spark follows the master-slave architecture. Its cluster consists of a single master and multiple slaves. The Spark architecture depends upon two abstractions:

Resilient Distributed Dataset (RDD) Directed Acyclic Graph (DAG) Resilient Distributed Datasets (RDD)

In spark there are 2 kinds of operations.

1. Transformations.

2. Actions.

transformations are lazy which means when we execute the below lines, no actual computation has happened but a diagram will be created.

but actions are not.

A DAG is generated when we compute spark statements.

Execution happens when action is encountered before that only entries are made into DAG.

4. What is RDD?

A.) Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects.

Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

Resilient: ability to quickly recover from failures.

5. How spark achieves fault tolerance?

A.) Rdd Provides Fault tolerance through lineage graph.

A lineage graph keeps a track of transformations to be executed after action has been called.

6. What is Transformations & action in spark?

A.) In spark there are 2 kinds of operations.

1. Transformations.

2. Actions.

transformations are lazy which means when we execute the below lines, no actual computation has happend but a diagram will be created.

but actions are not.

if transformations are not lazy, we just need query simple function to large file.

7. Difference between Narrow & wide transformations?

A.) A narrow transformation is one in which a single input partition maps to a single output partition.

example filter,map, flatmap

A wide transformation is a much more expensive operation and is sometimes referred to as a shuffle in Spark.

Ex: ReduceByKey, Groupby

8. what is difference between DAG & Lineage?

A.) DAG:A DAG is generated when we compute spark statements.

Execution happens when action is encountered before that only entries are made into DAG.

Lineage: Rdd Provides Fault tolerance through lineage graph.

A lineage graph keeps a track of transformations to be executed after action has been called.

9. What is partition and how spark Partitions the data?

A.) A Partition in simple terms is a split in the input data, so partitions in spark are basically smaller logical chunks or divisions of the input data.

Spark distributes this partitioned data among the different nodes to perform distributed processing on the data.

10. what is spark core?

A.) spark provides distributed task scheduling, and basic I/O functionalities.

Spark uses a specialized fundamental data structure known as RDD (Resilient Distributed Datasets) that is a logical collection of data partitioned across machines.

11. what is spark driver or driver program?

A.) A Spark driver is the process that creates and owns an instance of SparkContext.

It is the cockpit of jobs and tasks execution (using DAGScheduler and Task Scheduler).

12. What is spark executors?

A.) Executors are worker nodes' that running individual tasks in a given Spark job.

They are launched at the beginning of a Spark application and typically run for the entire lifetime of an application.

13. what is worker node?

A.) Worker Node is the Slave Node. Master node assign work and worker node actually perform the assigned tasks.

Worker node processes the data stored on the node.

14. what is lazy evaluation in spark?

A.) As the name itself indicates its definition, lazy evaluation in Spark means that the execution will not start until an action is triggered.

In Spark, the picture of lazy evaluation comes when Spark transformations occur.

15. what is pair RDD in spark?

A.) Paired RDD in Spark is an RDD with the distributed collection of objects containing key-value pairs.

Paired RDDs is a very powerful data structure because it supports to act on each key operation in parallel or re-group data across the network.

16. Difference between persist() and cache() in spark?

A.) Both persist () and cache () are the Spark optimization technique, used to store the data, but only difference is cache () method by default stores the data in-memory (MEMORY_ONLY) whereas in persist () method developer can define the storage level to in-memory or in-disk.

17. what is serialization and deserialization?

A.) serialization can be defined as which is converting from object to bytes which can be sent over network.

deserialization can be defined as converting the data to a form that can be stored sent over the network into a form which can be read.

18. Avoid returning null, in scala code,

using None which in turn can very well handled by getOrElse

19. Diamond problem in scala occurs when child class/object tries to refer?

A.) multiple parent classes having same method name.

20. Singleton, Lazy initialization design patterns to for respectively if we have

memory constraints, defer expensive computation.

21. For the following code in scala: lazy val output = {println("Hello"); 1} println("Learning Scala") println(output). What can be the result, in proper order?

A.) Learning scala, Hello,1

22. Suppose we have a series of 9 Mapreduce Jobs, then how many Disk I/Os are needed in total?

A.) 18

23. Which operations is not lazy?

A.) collect, take

24. Suppose while running spark on hadoop2, input file is of size 800 MB. How many RDD partitions will be created in all ?

A.) 7 partitions

25. which will help Rdds to achieve resiliency?

A.) RDDs maintain a Lineage graph

RDD contents cannot be changed

26. RDDs says materialized in which condition?

A.) when action is called to execute file with collect.

27. flatmap does not provide always multiple inputs to get multiple outputs

reduceByKey is not an action

reduceByKey cannot take two or more parameters.

we cannot create spark-context in spark-shell

28. Actions are functions applied on RDD, resulting into another RDD.

A.) true

29. Spark transformations & actions are evaluated lazily?

A.) False

30. what is higher order functions?

A.) map(), reduce(), foreach()

31. Which of the below gives one to one mapping between input & output. *?

A.) map

32. by default spark UI is available on which port?

A.) port 4040

33. what is broadcast variable?

A.) Broadcast variables in Apache Spark is a mechanism for sharing variables across executors that are meant to be read-only. Without broadcast variables these variables would be shipped to each executor for every transformation and action, and this can cause network overhead.

34. what is accumulator?

A.) it is a shared variable a single file kept in driver program and remaining executor update it.

None of the executors can read the value of accumulator, but it can only update it.

these are similar to counters in mapRduce.

19. Difference between map() and flatmap()?

A.) Map () operation applies to each element of RDD and it returns the result as new RDD. In the Map, operation developer can define his own custom business logic. While FlatMap () is similar to Map, but FlatMap allows returning 0, 1 or more elements from map function.

20. what are the various level of persistance in spark?

A.) Spark has various persistence levels to store the RDDs on disk or in memory or as a combination of both with different replication levels namely: MEMORY_ONLY; MEMORY_ONLY_SER; MEMORY_AND_DISK; MEMORY_AND_DISK_SER, DISK_ONLY; OFF_HEAP

21. What is accumulator in spark?

A.) sparkContext.accumulator () is used to define accumulator variables.

value property on the accumulator variable is used to retrieve the value from the accumulator.

Accumulators are variables that are used for aggregating information across the executors.

22. what is broadcast variable in spark?

A.) Broadcast variables in Apache Spark is a mechanism for sharing variables across executors that are meant to be read-only. Without broadcast variables these variables would be shipped to each executor for every transformation and action, and this can cause network overhead.

23. what is checkpointing in spark?

A.) Checkpointing stores the rdd physically to hdfs and destroys the lineage that created it.

The checkpoint file won't be deleted even after the Spark application terminated.

Checkpoint files can be used in subsequent job run or driver program

24. what is spark context?

A.) SparkContext is an entry point to Spark and defined in org.apache.spark package and used to programmatically create Spark RDD, accumulators, and broadcast variables on the cluster. Its object sc is default variable available in spark-shell and it can be programmatically created using SparkContext class.

25. what is Executor memory in spark?

A.) 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.

26. Explain spark stages?

A.) Spark stages are the physical unit of execution for the computation of multiple tasks. The Spark stages are controlled by the Directed Acyclic Graph (DAG) for any data processing and transformations on the resilient distributed datasets (RDD).

Basically, there are two types of stages in spark- ShuffleMapstage and ResultStage.

27. what is spark SQL?

A.) Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and data.

28. Difference between RDD vs Dataframe & Dataset in spark?

A.) The schema gives an expressive way to navigate inside the data. RDD is a low level API whereas DataFrame/Dataset are high level APIs. With RDD, you have more control on what you do. A DataFrame/Dataset tends to be more efficient than an RDD. What happens inside Spark core is that a DataFrame/Dataset is converted into an optimized RDD.

29. How spark SQL is different from HQL & SQL?

A.) Spark-sql: SparkSQL is a special component on the sparkCore engine that support SQL and HiveQueryLanguage without changing any syntax

SQL - SQL is a traditional query language that directly interacts with RDBMs

HQL - HQL is a JAVA-based OOP language that uses the Hibernate interface to convert the OOP code into query statements and then interacts with databases.

30. What is catalyst Optimizer?

A.) Catalyst optimizer makes use of some advanced programming language features to build optimized queries. Catalyst optimizer was developed using programming construct Scala. Catalyst Optimizer allows both Rule Based optimization and Cost Based optimization.

31. what is spark streaming?

A.) Spark Streaming is an extension of the core Spark API that allows data engineers and data scientists to process real-time data from various sources including (but not limited to) Kafka, Flume, and Amazon Kinesis. This processed data can be pushed out to file systems, databases.

32. What is DStream?

A.) DStream is a continuous stream of data. It receives input from various sources like Kafka, Flume, Kinesis, or TCP sockets. It can also be a data stream generated by transforming the input stream. At its core, DStream is a continuous stream of RDD (Spark abstraction).

33. How to create Micro batch and its benifit?

A.) It allows developers to write code, and influence the architecture.

Microservices are small applications that your development teams create independently.

34. what is windowing in spark streaming?

A.) Window The simplest windowing function is a window, which lets you create a new DStream, computed by applying the windowing parameters to the old DStream.

35. what is Scala programming Languages & its advantages?

A.) Easy to Pick Up

Pretty Good IDE support

IntelliJ IDEA. Most developers consider this to be the best IDE for Scala. It has great UI, and the editor is pretty...

Scalability

36. What is the difference between Statically typed & Dynamically typed language?

A.) In statically typed languages type checking happens at compile time. Whereas, in dynamically typed languages, type checking happens at run-time.

37. what is the difference var and val in scala?

A.) The keywords var and val both are used to assign memory to variables.

var keyword initializes variables that are mutable, and the val keyword initializes variables that are immutable.

38. what is the difference between == in java and scala?

A.) In Java, C++, and C# the == operator tests for reference, not value equality.

in Scala, == is testing for value equality.

39. What is Typesafe in scala?

A.) Type-safe means that the set of values that may be assigned to a program variable must fit well-defined and testable criteria.

40. what is type inference in scala?

A.) With Scala type inference, Scala automatically detects the type of the function without explicitly specified by the user.

41. what is Unit in scala? what is difference between java void's and scala unit?

A.) Unit is a final class defined in “scala” package that is “scala.Unit”. Unit is something similar to 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.

42. what is scala singleton object?

A.) Scala Singleton Object is an object which is declared by using object keyword instead by class. No object is required to call methods declared inside singleton object. In scala, there is no static concept.

43. what is companion object in scala?

A.) A companion object in Scala is an object that’s declared in the same file as a class, and has the same name as the class. For instance, when the following code is saved in a file named Pizza.scala, the Pizza object is considered to be a companion object to the Pizza class:

44. Difference between and singleton object and class in scala?

A.) An object is a singleton -- an instance of a class which is guaranteed to be unique. For every object in the code, an anonymous class is created, which inherits from whatever classes you declared object to implement.

Classes have fields and methods

45. what is scala Map?

A.) Scala map is a collection of key/value pairs. Any value can be retrieved based on its key. Keys are unique in the Map, but values need not be unique.

46. what is scala set?

A.) Set is a collection that contains no duplicate elements. There are two kinds of Sets, the immutable and the mutable.

47. what is the use of Tuples in scala?

A.) A tuple is a data structure which can store elements of the different data type. It is also used for storing and retrieving of data. In scala, tuples are immutable in nature and store heterogeneous types of data.

48. what is Scala case class?

A.) Scala Case Class is like a regular class, except it is good for modeling immutable data. It also serves useful in pattern matching, such a class has a default apply () method which handles object construction.

49. what is scala option?

A.) Scala Option [T] is a container for zero or one element of a given type. An Option [T] can be either Some [T] or None object, which represents a missing value.

50. what is use case of App class in scala?

A.) Scala provides a helper class, called App, that provides the main method. Instead of writing your own main method, classes can extend the App class to produce concise and executable applications in Scala.

51. Difference between terms & types in scala? Nill, NUll, None, Nothing?

A.) Null– Its a Trait. null– Its an instance of Null- Similar to Java null.

Nil– Represents an empty List of anything of zero length.

Nothing is a Trait. Its a subtype of everything. But not superclass of anything. There are no instances of Nothing.

None– Used to represent a sensible return value. Just to avoid null pointer.

52. what is traits in scala?

A.) In scala, trait is a collection of abstract and non-abstract methods. You can create trait that can have all abstract methods or some abstract and some non-abstract methods. A variable that is declared either by using val or var keyword in a trait get internally implemented in the class that implements the trait.

53. Difference between Traits and abstract class in scala?

A.) Traits	Abstract Class
Allow multiple inheritances.	Do not Allow multiple inheritances.
Constructor parameters are not allowed in Trait.	Constructor parameter are allowed in Abstract Class.
The Code of traits is interoperable until it is implemented.	The code of abstract class is fully interoperable.
Traits can be added to an object instance in Scala.	Abstract classes cannot be added to object instance in Scala.

54. Difference between Call-by-value and call-by-name parameter?

A.) call-by-value is The same value will be used all throughout the function. Whereas in a Call by Name, the expression itself is passed as a parameter to the function and it is only computed inside the function, whenever that particular parameter is called.

55. what are Higher order functions in scala?

A.) Scala Higher Order Functions is a function that either takes a function as argument or returns a function. In other words we can say a function which works with function is called higher order function. Higher order function allows you to create function composition, lambda function or anonymous function etc.

56. What is Pure function in scala?

A.) A function is called pure function if it always returns the same result for same argument values and it has no side effects like modifying an argument (or global variable) or outputting something.

57. Explain scala anonymous function in scala?

A.) In Scala, An anonymous function is also known as a function literal. A function which does not contain a name is known as an anonymous function. An anonymous function provides a lightweight function definition.

58. what is closure in scala?

A.) A closure is a function, whose return value depends on the value of one or more variables declared outside this function.

59. what is currying in scala?

A.) Currying is the process of converting a function with multiple arguments into a sequence of functions that take one argument.

60. what is option in scala? why do we use it?

A.) Scala Option[T] is a container for zero or one element of a given type. An Option[T] can be either Some[T] or None object, which represents a missing value.

Option type is used frequently in Scala programs and you can compare this with the null value

61. what is tail recursion in scala?

A.) Recursion is a method which breaks the problem into smaller subproblems and calls itself for each of the problems.

62. What is yield in scala?

A.) For each iteration of your for loop, yield generates a value which is remembered by the for loop (behind the scenes)

63. can we able to do datasets in python?

A.) A simple way to get sample datasets in Python is to use the pandas 'read_csv' method to load them directly from the internet.

64. How to join two tables using dataframes?

A.) empDF. join (deptDF, empDF ("emp_dept_id") === deptDF ("dept_id"),"inner").show(false)

65. How to remove duplicates records in dataframe?

A.) Use distinct () – Remove Duplicate Rows on DataFrame.

Use dropDuplicate () – Remove Duplicate Rows on DataFrame.

66. How to add columns in Dataframe?

A.) Using withColumn () to Add a New Column. Here, we have added a new column.

67. SQL basics concepts such as Rank, Dense Rank, Row Number?

A.) RANK() Returns the rank of each row in the result set of partitioned column. select Name,Subject,Marks, RANK()

DENSE_RANK() This is same as RANK() function. Only difference is returns rank without gaps.

ROW_NUMBER will always generate unique values without any gaps, even if there are ties.

68. Query to find 2nd largest number in the table?

A.) SELECT MAX(sal) as Second_Largest FROM emp_test WHERE sal < (SELECT MAX(sal) FROM emp_test)

69. To find duplicate record in table?

A.) select a.* from Employee a where rowid != (select max(rowid) from Employee b where a.Employee_num =b.Employee_num;

70. Difference between list and Tuple?

A.) LIST TUPLE

Lists are mutable Tuples are immutable

Implication of iterations is Time-consuming. The implication of iterations is comparatively Faster

The list is better for performing operations, such as insertion and deletion. Tuple data type is appropriate for accessing the elements

Lists consume more memory Tuple consume less memory as compared to the list

Lists have several built-in methods Tuple does not have many built-in methods.

The unexpected changes and errors are more likely to occur In tuple, it is hard to take place.

71. Difference between def and Lambda?

A.) lambda is a keyword that returns a function object and does not create a 'name'. Whereas def creates name in the local namespace

lambda functions are good for situations where you want to minimize lines of code as you can create function in one line of python code. ...

lambda functions are somewhat less readable for most Python users.

72. Why bigdata on cloud preferred these days?

A.) Big Data Cloud brings the best of open source software to an easy-to-use and secure environment that is seamlessly integrated and serverless.

73. What is aws EMR?

A.) Amazon Elastic MapReduce (EMR) is an Amazon Web Services (AWS) tool for big data processing and analysis. Amazon EMR offers the expandable low-configuration service as an easier alternative to running in-house cluster computing.

74. How to write a UDF in hive?

A.) By writing UDF (User Defined function) hive makes it easy to plug in your own processing code and invoke it from a Hive query. UDF's have to be writhen in Java, the Language that Hive itself is written in.

Create a Java class for the User Defined Function which extends ora.apache.hadoop.hive.sq.exec.UDF and implements more than one evaluate () methods. Put in your desired logic and you are almost there.

Package your Java class into a JAR file (I am using Maven)

Go to Hive CLI, add your JAR, and verify your JARs is in the Hive CLI classpath

CREATE TEMPORARY FUNCTION in Hive which points to your Java class

Use it in Hive SQL

75. file formats row based vs column based?

A.) In a row storage format, each record in the dataset has to be loaded, parsed into fields and then the data for Name is extracted. With the column-oriented format, it can directly go to the Name column as all the values for that column are stored together. It doesn't need to go through the whole record.

76. What is RDD?

A.) Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

77. how to join two larger tables in spark?

A.) Spark uses SortMerge joins to join large table. It consists of hashing each row on both table and shuffle the rows with the same hash into the same partition. There the keys are sorted on both side and the sortMerge algorithm is applied.

78. what is Bucketed tables?

A.) In the table directory, the Bucket numbering is 1-based and every bucket is a file. Bucketing is a standalone function. This means you can perform bucketing without performing partitioning on a table. A bucketed table creates nearly equally distributed data file sections.

79. How to read the parquet file format in spark?

A.) Similar to write, DataFrameReader provides parquet () function (spark.read.parquet) to read the parquet files and creates a Spark DataFrame.

80. How to tune spark executor, cores and executor memory?

A.) Number of cores = Concurrent tasks as executor can run

total available cores in one Node(CPU) - we come to

3 executors per node.

The property spark.executor.memory specifies the amount of memory to allot to each executor.

81. Default partition size in spark?

A.) default number of partitions is based on the following. On the HDFS cluster, by default, Spark creates one Partition for each block of the file. In Version 1 Hadoop the HDFS block size is 64 MB and in Version 2 Hadoop the HDFS block size is 128 MB

82. is there any use of running spark program on single machine?

A.) Spark also provides a simple standalone deploy mode. You can launch a standalone cluster either manually, by starting a master and workers by hand, or use our provided launch scripts. It is also possible to run these daemons on a single machine for testing.

83. how to find how many resources are available in YARN?

A.) yarn.resource-types.memory-mb.increment-allocation: The fairscheduler grants memory in increments of this value. If you submit a task with resource request that is not a multiple of memory-mb.increment-allocation, the request will be rounded up to the nearest increment. Defaults to 1024 MB. yarn.resource-types.vcores.increment-allocation

84. Differences between cluster and client Mode?

A.) In cluster mode, the driver will get started within the cluster in any of the worker machines. So, the client can fire the job and forget it. In client mode, the driver will get started within the client. So, the client has to be online and in touch with the cluster.

85. Explain about the dynamic allocation in spark?

A.) Spark dynamic allocation is a feature allowing your Spark application to automatically scale up and down the number of executors. And only the number of executors not the memory size and not the number of cores of each executor that must still be set specifically in your application or when executing spark-submit command.

86. Difference between partition by and cluster by in hive?

A.) In Hive partitioning, the table is divided into the number of partitions, and these partitions can be further subdivided into more manageable parts known as Buckets/Clusters. Records with the same bucketed column will be stored in the same bucket. “clustered by” clause is used to divide the table into buckets.

Cluster By is a short-cut for both Distribute By and Sort By. Hive uses the columns in Distribute By to distribute the rows among reducers. All rows with the same Distribute By columns will go to the same reducer. However, Distribute By does not guarantee clustering or sorting properties on the distributed keys.

87. How to choose partitioning column in hive? and which column shouldn't use partition and why?

A.) When the column with a high search query has low cardinality. For example, if you create a partition by the country name then a maximum of 195 partitions will be made and these number of directories are manageable by the hive. On the other hand, do not create partitions on the columns with very high cardinality.

88. how to transfer data from unix system to HDFS?

A.) hdfs dfs -put test /hadoop ubuntu@ubuntu-VirtualBox

89. can we extract only different data from two different tables?

A.) using Join column we can extract data.

SELECT tablenmae1.colunmname, tablename2.columnnmae

FROM tablenmae1

JOIN tablename2

ON tablenmae1.colunmnam = tablename2.columnnmae

ORDER BY columnname;

90. What is the difference between SQL vs NoSQL?

A.) SQL databases are vertically scalable while NoSQL databases are horizontally scalable. SQL databases have a predefined schema whereas NoSQL databases use dynamic schema for

unstructured data. SQL requires specialized DB hardware for better performance while NoSQL uses commodity hardware.

91. how to find particular text name in HDFS?

A.) You can use cat command on HDFS to read regular text files. hdfs dfs -cat /path/to/file.csv

92. Explain about sqoop ingestion process?

A.) Apache Sqoop is a data ingestion tool designed for efficiently transferring bulk data between Apache Hadoop and structured data-stores such as relational databases, and vice-versa.

93. Explain about sort Merge Bucket Join?

A.) Sort Merge Bucket (SMB) join in hive is mainly used as there is no limit on file or partition or table join. SMB join can best be used when the tables are large. In SMB join the columns are bucketed and sorted using the join columns. All tables should have the same number of buckets in SMB join.

94. Explain about tungsten?

A.) Tungsten is a Spark SQL component that provides increased performance by rewriting Spark operations in bytecode, at runtime.

95. How can we join two bigger tables in spark?

A.) either using Sort Merge Joins if we are joining two big tables, or Broadcast Joins if at least one of the datasets involved is small enough to be stored in the memory of the single all executors.

A ShuffleHashJoin is the most basic way to join tables in Spark

96. Explain about left outer join?

A.) The left outer join returns a resultset table with the matched data from the two tables and then the remaining rows of the left table and null from the right table's columns.

97. How to count the lines in a file by using linux command?

A. using -wc

98. How to achieve map side joins in hive?

A.) only possible since the right table that is to the right side of the join conditions, is lesser than 25 MB in size. Also, we can convert a right-outer join to a map-side join in the Hive.

99. when we use select command does it goes to reducer in Hive?

A.) We can use reducer if and there is no aggregation of data in mapside only then it uses reducer.

100. How to validate the data once the ingestion is done?

A.) data validation is used as a part of processes such as ETL (Extract, Transform, and Load) where you move data from a source database to a target data warehouse so that you can join it with other data for analysis. Data validation helps ensure that when you perform analysis, your results are accurate.

Steps to data validation:

a. Determine data sample: validate a sample of your data rather than the entire set.

b. Validate the database: Before you move your data, you need to ensure that all the required data is present in your existing database

c. Validate the data format: Determine the overall health of the data and the changes that will be required of the source data to match the schema in the target.

Methods for data validation:

a. Scripting: Data validation is commonly performed using a scripting language

b. Enterprise tools: Enterprise tools are available to perform data validation.

c. open source tools: Open source options are cost-effective, and if they are cloud-based, can also save you money on infrastructure costs.

101. what is the use of split by command in sqoop?

A.) split-by in sqoop is used to create input splits for the mapper. It is very useful for parallelism factor as splitting imposes the job to run faster. Hadoop MAP Reduce is all about divide and conquer. When using the split-by option, you should choose a column which contains values that are uniformly distributed.

102. Difference between dataframe vs datasets?

A.) DataFrames gives a schema view of data basically, it is an abstraction. In dataframes, view of data is organized as columns with column name and types info. In addition, we can say data in dataframe is as same as the table in relational database.

As similar as RDD, execution in dataframe too is lazy triggered.

In Spark, datasets are an extension of dataframes. Basically, it earns two different APIs characteristics, such as strongly typed and untyped. Datasets are by default a collection of strongly typed JVM objects, unlike dataframes. Moreover, it uses Spark's Catalyst optimizer.

103. Difference between schema on read vs schema on write?

A.) Schema on read differs from schema on write because you create the schema only when reading the data. Structured is applied to the data only when it's read, this allows unstructured data to be stored in the database.

The main advantages of schema on write are precision and query speed.

104. Different types of partition in hive?

A.) Types of Hive Partitioning

Static Partitioning

Dynamic Partitioning

105. How to find counts based on age group?

A.) SELECT Col1, COUNT(*) FROM Table GROUP BY Col1.

106. How to find a word in a log file by using pyspark?

A.) input_file = sc.textFile("/path/to/text/file")

map = input_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1))

counts = map.reduceByKey(lambda a, b: a + b)

counts.saveAsTextFile("/path/to/output/")

107. Explain about Executor node in spark?

A.) Executors are worker nodes' processes in charge of running individual tasks in a given Spark job. They are launched at the beginning of a Spark application and typically run for the entire lifetime of an application. Once they have run the task they send the results to the driver.

108. Difference between Hadoop & spark?

A.) biggest difference is that it works in memory. Whereas Hadoop reads and writes files to HDFS, Spark processes data in RAM using a concept known as an RDD, Resilient Distributed Dataset.

Hadoop has its own storage system HDFS while Spark requires a storage system like HDFS which can be easily grown by adding more nodes. They both are highly scalable as HDFS storage can go more than hundreds of thousands of nodes. Spark can also integrate with other storage systems like S3 bucket.

109. Query to find duplicate value in SQL?

A.) Using the GROUP BY clause to find the duplicate values

use the GROUP BY clause to group all rows by the target column, which is the column that you want to check duplicate. Then, use the COUNT ()

110. Difference between Row number and Dense Rank in SQL?

A.) Rank () SQL function generates rank of the data within ordered set of values but next rank after previous rank is row_number of that particular row. On the other hand, Dense_Rank () SQL function generates next number instead of generating row_number. Below is the SQL example which will clarify the concept

111. What are the various hive optimization techniques?

A.) Tez-Execution Engine in Hive:Tez Execution Engine – Hive Optimization Techniques, to increase the Hive performance.

Usage of Suitable File Format in Hive

Hive Partitioning.

Bucketing in Hive.

112. How Mapreduce will work? Explain?

A.) MapReduce can perform distributed and parallel computations using large datasets across a large number of nodes. A MapReduce job usually splits the input datasets and then process each of them independently by the Map tasks in a completely parallel manner. The output is then sorted and input to reduce tasks.

it uses key value pair.

113. How many types of tables have in Hive?

A.) Hive knows two different types of tables: Internal table and the External table. The Internal table is also known as the managed table.

114. How to drop table in HBase?

A.) Dropping a Table using HBase Shell Using the drop command, you can delete a table. Before dropping a table, you have to disable it. hbase (main):018:0> disable 'emp' 0 row (s) in 1.4580 seconds hbase (main):019:0> drop 'emp' 0 row (s) in 0.3060 seconds

115. What is the difference between Batch and real time processing?

A.) In real time processing processor needs to very responsive and active all the time.

In batch processing processor only needs to busy when work is assigned to it.

Real-time processing needs high computer architecture and high hardware specification.

Normal computer specification can also work with batch processing.

Time to complete the task is very critical in real-time.

Real-time processing is expensive.

Batch processing is cost effective

116. How can Apache spark be used alongside Hadoop?

A.) Running Alongside Hadoop You can run Spark alongside your existing Hadoop cluster by just launching it as a separate service on the same machines. To access Hadoop data from Spark, just use a hdfs:// URL (typically hdfs://<namenode>:9000/path, but you can find the right URL on your Hadoop Namenode's web UI).

117. What is Data explode and lateral view in Hive?

A.) In Hive, lateral view explode the array data into multiple rows. In other word, lateral view expands the array into rows.

Hive has way to parse array data type using LATERAL VIEW. Use LATERAL VIEW with UDTF to generate zero or more output rows for each input row. Explode is one type of User Defined Table Building Function.

118. What is combiner, shuffling, sorting in Mapreduce?

A.) Shuffling is the process by which it transfers mappers intermediate output to the reducer. Reducer gets 1 or more keys and associated values on the basis of reducers.

In Sort phase merging and sorting of map output takes place.

The combiner should combine key/value pairs with the same key. Each combiner may run zero, once, or multiple times.

119. what is the difference between reduceByKey and GroupByKey?

A.) The groupByKey method operates on an RDD of key-value pairs, so key a key generator function is not required as input. What is reduceByKey? The reduceByKey is a higher-order method that takes associative binary operator as input and reduces values with the same key. This function merges the values of each key using the reduceByKey method in Spark.

120. what is static and dynamic partition in Hive?

A.) Usually dynamic partition load the data from non partitioned table. Dynamic Partition takes more time in loading data compared to static partition. When you have large data stored in a table then Dynamic partition is suitable.

Static partitions are preferred when loading big data in Hive table and it saves your time in loading data compared to dynamic partition. In static partitioning, we need to specify the partition column value in each and every LOAD statement.

121. Udf example in Hive?

A.) A UDF processes one or several columns of one row and outputs one value. For example : SELECT lower(str) from table For each row in "table," the "lower" UDF takes one argument, the value of "str", and outputs one value, the lowercase representation of "str".

122. what is Serde in Hive?

A.) In SerDe interface handles both serialization and deserialization and also interpreting the results of serialization as individual fields for processing. It allows Hive to read data from a table, and write it back to HDFS in any format, user can write data formats.

123. How to check the file size in Hadoop?

A.) You can use the hadoop fs -ls -h command to check the size. The size will be displayed in bytes.

124. How to submit the spark Job?

A.) Using --deploy-mode, you specify where to run the Spark application driver program.

cluster Mode: In cluster mode, the driver runs on one of the worker nodes, and this node shows as a driver on the Spark Web UI of your application. cluster mode is used to run production jobs.

Client mode: In client mode, the driver runs locally where you are submitting your application from. client mode is majorly used for interactive and debugging purposes.

Using --master option, you specify what cluster manager to use to run your application. Spark currently supports Yarn, Mesos, Kubernetes, Stand-alone, and local.

125. what is vectorization and why it used?

A.) Vectorization is the process of converting an algorithm from operating on a single value at a time to operating on a set of values at one time

126. what are Complex data types in Hive?

A.) ARRAY

Struct

Map

127. What is sampling in Hive?

A.) Prepare the dataset.

Create a Hive Table and Load the Data

Sampling using Random function.

Create a Bucketed table

Load data into Bucketed table

Sampling using bucketing.

Block sampling in hive.

128. what are different type of xml files in Hadoop?

A.) resource-types.xml

node-resources.xml

yarn-site.xml

129. what is case class?

A.) A Case Class is just like a regular class, which has a feature for modeling unchangeable data. It is also constructive in pattern matching.

It has been defined with a modifier case , due to this case keyword, we can get some benefits to stop oneself from doing a sections of codes that have to be included in many places with little or no alteration.

130. Difference between broadcast and accumulators?

A.) Broadcast variables − used to efficiently, distribute large values. Accumulators − used to aggregate the information of particular collection.

131. what is the difference between spark context and spark session?

A.) SparkContext has been available since Spark 1.x versions and it’s an entry point to Spark when you wanted to program and use Spark RDD. Most of the operations/methods or functions we use in Spark are comes from SparkContext for example accumulators, broadcast variables, parallelize and more.

SparkSession is essentially combination of SQLContext, HiveContext and future StreamingContext.

132. what are the operation of dataframe?

A.) A spark data frame can be said to be a distributed data collection that is organized into named columns and is also used to provide the operations such as filtering, computation of aggregations, grouping and also can be used with Spark SQL.

133. Explain why spark preferred over mapreduce?

A.) the benefits of Apache Spark over Hadoop MapReduce are given below: Processing at high speeds: The process of Spark execution can be up to 100 times faster due to its inherent ability to exploit the memory rather than using the disk storage.

134. what is the difference between partitioning and Bucketing?

A.) Bucketing decomposes data into more manageable or equal parts. With partitioning, there is a possibility that you can create multiple small partitions based on column values. If you go for bucketing, you are restricting number of buckets to store the data.

135. How to handle incremental data in bigdata?

A.) Move existing HDFS data to temporary folder

Run last modified mode fresh import

Merge with this fresh import with old data which saved in temporary folder.

136. Explain the Yarn Architecture?

A.) Apache YARN framework contains a Resource Manager (master daemon), Node Manager (slave daemon), and an Application Master.

YARN is the main component of Hadoop v2.0. YARN helps to open up Hadoop by allowing to process and run data for batch processing, stream processing, interactive processing and graph processing which are stored in HDFS. In this way, It helps to run different types of distributed applications other than MapReduce.

137. what is incremental sqoop?

A.) Incremental imports mode can be used to retrieve only rows newer than some previously-imported set of rows. Why Append mode ?? works for numerical data that is incrementing over time, such as auto-increment keys,

138. Why we use Hbase and how it store data?

A.) HBase provides a flexible data model and low latency access to small amounts of data stored in large data sets HBase on top of Hadoop will increase the throughput and performance of distributed cluster set up. In turn, it provides faster random reads and writes operations

139. What are various optimization technique in hive?

A.) Apache Hive Optimization Techniques — 1. Partitioning. Bucketing.

140. Sqoop command to exclude tables while retrieval?

A.) sqoop import --connect jdbc:mysql://localhost/sqoop --username root --password hadoop --table <tablename> --target-dir '/Sqoop21/AllTables' --exclude-tables <table1>,<tables2>.

141. how to create sqoop password alias?

A.) sqoop import --connect jdbc:mysql://database.example.com/employees \ --username dbuser --password-alias mydb.password.alias. Similarly, if the command line option is not preferred, the alias can be saved in the file provided with --password-file option.

142. what is sqoop job optimization?

A.) To optimize performance, set the number of map tasks to a value lower than the maximum number of connections that the database supports.

143. what is sqoop boundary quieries and split by usage?

A.) The boundary query is used for splitting the value according to id_no of the database table. To boundary query, we can take a minimum value and maximum value to split the value.

split-by in sqoop is used to create input splits for the mapper. It is very useful for parallelism factor as splitting imposes the job to run faster.

144. what is hbase compaction technique and write operation hbase using spark??

A.) HBase Minor Compaction The procedure of combining the configurable number of smaller HFiles into one large HFile is what we call Minor compaction.

hbase-spark connector which provides HBaseContext to interact Spark with HBase. HBaseContext pushes the configuration to the Spark executors and allows it to have an HBase Connection per Spark Executor.

145. what are hive managed Hbase tables and how to create that?

A.) Hive tables Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. External tables are tables where Hive has loose coupling with the data. Replication Manager replicates external tables successfully to a target cluster.

CREATE [EXTERNAL] TABLE foo(…) STORED BY ‘org.apache.hadoop.hive.hbase.HBaseStorageHandler’ TBLPROPERTIES (‘hbase.table.name’ = ‘bar’); [/sql]

146. How Hbase can be a Distributed database?

A.) Hbase is one of NoSql column-oriented distributed database available in apache foundation. HBase gives more performance for retrieving fewer records rather than Hadoop or Hive. It's very easy to search for given any input value because it supports indexing, transactions, and updating.

147. What is hive metastore and how to access that?

A.) Metastore is the central repository of Apache Hive metadata. It stores metadata for Hive tables (like their schema and location) andpartitions in a relational database. It provides client access to this information by using metastore service API. Hive metastore consists of two fundamental units: 1. A service that provides metastore access to other Apache Hive services. 2. Disk storage for the Hive metadata which is separate from HDFSstorage.

probably get most of the information you need through HCatalog, without direct access to the metastore tables.

148. What are the ways to remove duplicates in hive?

A.) Use Insert Overwrite and DISTINCT Keyword

GROUP BY Clause to Remove Duplicate

Use Insert Overwrite with row_number () analytics functions

148. What is Hive Managed and External tables?

A.) Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. External tables are tables where Hive has loose coupling with the data. Replication Manager replicates external tables successfully to a target cluster. The managed tables are converted to external tables after replication.

149. How partition can be restored?

A.) using MSCK REPAIR

150. what is data loading in hive?

A.) Hive provides us the functionality to load pre-created table entities either from our local file system or from HDFS. The LOAD DATA statement is used to load data into the hive table.

151. How to automate Hive jobs?

A.) Like you can also use Hive CLI and its very ease to do such jobs. You can write shell script in Linux or .bat in Windows. In script you can simply go like below entries. $HIVE_HOME/bin/hive -e 'select a.col from tab1 a'; or if you have file : $HIVE_HOME/bin/hive -f /home/my/hive-script.sql Make sure you have set $HIVE_HOME in your env.

152. where do we run job in spark?

A.) The spark-submit script in Spark's bin directory is used to launch applications on a cluster.

153. How to allocate resources in spark?

A.) Resources allocation - Dynamic/Static; Upstream or Downstream application .

154. Difference between Edge node vs Data Node?

A.) The majority of work is assigned to worker nodes. Worker node store most of the data and perform most of the calculations Edge nodes facilitate communications from end users to master and worker nodes.

155. what is Hive context?

A.) HiveContext is a super set of the SQLContext. Additional features include the ability to write queries using the more complete HiveQL parser, access to Hive UDFs, and the ability to read data from Hive tables. And if you want to work with Hive you have to use HiveContext, obviously.

156. How to read file from hdfs or other sources in spark?

A.) Use textFile() and wholeTextFiles() method of the SparkContext to read files from any file system and to read from HDFS, you need to provide the hdfs path as an argument to the function.

157. How to add custom schema to rdd?

A.) In spark, schema is array StructField of type StructType. Each StructType has 4 parameters. Column Name ; Data type of that column; Boolean value indication if values in this column can be null or not; Metadata column - this is optional column which can be used to add additional information about column

158. How to convert dataframe to rdd?

A.) To convert a dataframe back to rdd simply use the .rdd method: rdd = df.rdd But the setback here is that it may not give the regular spark RDD, it may return a Row object. In order to have the regular RDD format run the code below: rdd = df.rdd.map(tuple)

159. Difference between case class and class?

A.) A class can extend another class, whereas a case class can not extend another case class (because it would not be possible to correctly implement their equality).

160. what is optimization technique in spark?

A.) Spark optimization techniques are used to modify the settings and properties of Spark to ensure that the resources are utilized properly and the jobs are executed quickly. All this ultimately helps in processing data efficiently. The most popular Spark optimization techniques are listed below:

popular Spark optimization techniques are listed:

1. Data Serialization: Here, an in-memory object is converted into another format that can be stored in a file or sent over a network.

a. Java serialization: The ObjectOutputStream framework is used for serializing objects.

b. kyro serialization : To improve the performance, the classes have to be registered using the registerKryoClasses method.

2. caching: This is an efficient technique that is used when the data is required more often. Cache() and persist() are the methods used in this technique.

3. Data structure tuning: We can reduce the memory consumption while using Spark, by tweaking certain Java features that might add overhead.

4. Garbage collection optimization: G1 and GC must be used for running Spark applications. The G1 collector manages growing heaps. GC tuning is essential according to the generated logs, to control the unexpected behavior of applications.

161. What is unit data type in scala?

A.) The Unit type in Scala is used as a return statement for a function when no value is to be returned. Unit type can be e compared to void data type of other programming languages like Java.

162. What is boundary query in sqoop?

A.) The boundary query is used for splitting the value according to id_no of the database table. To boundary query, we can take a minimum value and maximum value to split the value. To make split using boundary queries, we need to know all the values in the table.

163. What is the use of sqoop eval command?

A.) It allows users to execute user-defined queries against respective database servers and preview the result in the console. So, the user can expect the resultant table data to import. Using eval, we can evaluate any type of SQL query that can be either DDL or DML statement.

164. How can we decide number of bucketing?

A.) The number of buckets is determined by hashFunction (bucketingColumn) mod numOfBuckets numOfBuckets is chose when you create the table with partitioning. The hash function output depends on the type of the column choosen.

165. Is it possible to bucketing and partitioning on same column?

A.) Yes.

Partitioning is you data is divided into number of directories on HDFS. Each directory is a partition.

166. How to do optimized joins in Hive?

A.) Use Tez to Fasten the execution Apache TEZ is an execution engine used for faster query execution.

Enable compression in Hive Compression techniques reduce the amount of data being transferred

Use ORC file format ORC (optimized record columnar) is great when it comes to hive performance tuning.

167. How to optimize join of 2 big tables?

A.) use the Bucket Map Join. For that the amount of buckets in one table must be a multiple of the amount of buckets in the other table. It can be activated by executing set hive.optimize.bucketmapjoin=true; before the query. If the tables don't meet the conditions, Hive will simply perform the normal Inner Join.

If both tables have the same amount of buckets and the data is sorted by the bucket keys, Hive can perform the faster Sort-Merge Join

168. what are major issues faced in spark development?

A.) Debugging - Spark although can be written in Scala, limits your debugging technique during compile time. You would encounter many run-time exceptions while running the Spark job. This is, many a times, because of the data. Sometimes because of the data type mismatch (there is dynamic data type inference) and sometimes data having null values and all. So there will be lot of iterations of run-time debugging.

Optimization - Optimizing a Spark code is a job to do. You need to optimize from the code side and from the resource allocation side too. A very well written code with good logic often performs badly because of badly done resource allocation.

169. what is dynamic allocation?

A.) Dynamic partitions provide us with flexibility and create partitions automatically depending on the data that we are inserting into the table.

170. What types of transformations do we perform in spark?

A.) Narrow transformation

Wide transformation

171. how to load data in hive table?

A.) Using Insert Command

table to table load

172. Difference between Map Vs Map Partition?

A.) Mappartitions is a transformation that is similar to Map. In Map, a function is applied to each and every element of an RDD and returns each and every other element of the resultant RDD. In the case of mapPartitions, instead of each element, the function is applied to each partition of RDD

mapPartitions exercises the function at the partition level

173. If we have some header information in a file how to read from it, and how to convert it to dataset or dataframe?

A.) we can add the option like header is true in while reading the file.

174. Difference between case class vs Struct type?

A.) Structs are value types and are copied on assignment. Structs are value types while classes are reference types. Structs can be instantiated without using a new operator. A struct cannot inherit from another struct or class, and it cannot be the base of a class.

175. What is sort by vs Order by in hive?

A.) Hive sort by and order by commands are used to fetch data in sorted order. The main differences between sort by and order by commands are given below. Sort by. hive> SELECT E.EMP_ID FROM

Employee E SORT BY E.empid; hive> SELECT E.EMP_ID FROM Employee E SORT BY E.empid; May use multiple reducers for final output.

176. How to increase the performance of Sqoop?

A.) Controlling the amount of parallelism that Sqoop will use to transfer data is the main way to control the load on your database.

Using more mappers will lead to a higher number of concurrent data transfer tasks, which can result in faster job completion.

However, it will also increase the load on the database as Sqoop will execute more concurrent queries.

177. While sqooping some data loss. how to handle that?

A.) Some lost data is recoverable, but this process often requires the assistance of IT professionals and costs time and resources your business could be using elsewhere. In other instances, lost files and information cannot be recovered, making data loss prevention even more essential.

Reformatting can also occur during system updates and result in data loss.

178. How to update record in Hbase table?

A.) Using put command you can insert a record into the HBase table easily. Here is the HBase Create data syntax. We will be using Put command to insert data into HBase table

179. what happens when sqoop fails in between the large data import job?

A.) sqoop import - job failure between data import, due to insert collisions in some cases, or lead to duplicated data in others. Since Sqoop breaks down export process into multiple transactions, it is possible that a failed export job may result in partial data being committed to the database.

180. what are hadoop components and their services?

A.) HDFS: Hadoop Distributed File System is the backbone of Hadoop which runs on java language and stores data in Hadoop applications. They act as a command interface to interact with Hadoop. the two components of HDFS – Data node, Name Node. Name node manages file systems and operates all data nodes and maintains records of metadata updating. In case of deletion of data, they automatically record it in Edit Log. Data Node (Slave Node) requires vast storage space due to reading and writing operations.

Yarn: It’s an important component in the ecosystem and called an operating system in Hadoop which provides resource management and job scheduling task.

Hbase: It is an open-source framework storing all types of data and doesn't support the SQL database. They run on top of HDFS and written in java language.

HBase master, Regional Server. The HBase master is responsible for load balancing in a Hadoop cluster and controls the failover. They are responsible for performing administration role. The regional server's role would be a worker node and responsible for reading, writing data in the cache.

Sqoop: It is a tool that helps in data transfer between HDFS and MySQL and gives hand-on to import and export data

Apache spark: It is an open-source cluster computing framework for data analytics and an essential data processing engine. It is written in Scala and comes with packaged standard libraries.

Apache Flume: It is a distributed service collecting a large amount of data from the source (webserver) and moves back to its origin and transferred to HDFS. The three components are Source, sink, and channel.

MapReduce: It is responsible for data processing and acts as a core component of Hadoop. Map Reduce is a processing engine that does parallel processing in multiple systems of the same cluster.

Apache Pig: Data Manipulation of Hadoop is performed by Apache Pig and uses Pig Latin Language. It helps in the reuse of code and easy to read and write code.

Hive: It is an open-source platform for performing data warehousing concepts; it manages to query large data sets stored in HDFS. It is built on top of the Hadoop Ecosystem. the language used by Hive is Hive Query language.

Apache Drill: Apache Drill is an open-source SQL engine which process non-relational databases and File system. They are designed to support Semi-structured databases found in Cloud storage.

Zookeeper: It is an API that helps in distributed Coordination. Here a node called Znode is created by an application in the Hadoop cluster.

Oozie: Oozie is a java web application that maintains many workflows in a Hadoop cluster. Having Web service APIs controls over a job is done anywhere. It is popular for handling Multiple jobs effectively.

181. What are important configuration files in Hadoop?

A.) HADOOP-ENV.sh ->>It specifies the environment variables that affect the JDK used by Hadoop Daemon (bin/hadoop). We...

CORE-SITE.XML ->>It is one of the important configuration files which is required for runtime environment settings of...

HDFS-SITE.XML ->>It is one of the important configuration files which is required for runtime environment settings of...

MAPRED-SITE.XML ->>It is one of the important configuration files which is required for runtime environment.

182. what is rack awareness?

A.) With the rack awareness policy's we store the data in different Racks so no way to lose our data. Rack awareness helps to maximize the network bandwidth because the data blocks transfer within the Racks. It also improves the cluster performance and provides high data availability.

183. problem with having lots of small files in HDFS? and how to overcome?

A.) Problems with small files and HDFS A small file is one which is significantly smaller than the HDFS block size (default 64MB). If you're storing small files, then you probably have lots of them (otherwise you wouldn't turn to Hadoop), and the problem is that HDFS can't handle lots of files.

Hadoop Archive

Sequence files

184. Main difference between Hadoop 1 and Hadoop 2?

A.) Hadoop 1.x System is a Single Purpose System. We can use it only for MapReduce Based Applications. If we observe the components of Hadoop 1.x and 2.x, Hadoop 2.x Architecture has one extra and new component that is : YARN (Yet Another Resource Negotiator).

185. What is block scanner in hdfs?

A.) Block Scanner is basically used to identify corrupt datanode Block. During a write operation, when a datanode writes in to the HDFS, it verifies a checksum for that data. This checksum helps in verifying the data corruptions during the data transmission.

186. what do you mean by high availability of name node? How is it achieved?

A.) In hadoop version 2.x there are two namenodes one of which is in active state and the other is in passive or standby state at any point of time.

187. Explain counters in MapReduce?

A.) A Counter in MapReduce is a mechanism used for collecting and measuring statistical information about MapReduce jobs and events. Counters keep the track of various job statistics in MapReduce like number of operations occurred and progress of the operation. Counters are used for Problem diagnosis in MapReduce.

188. Why the output of map tasks are spilled to local disk and not in hdfs?

A.) Execution of map tasks results into writing output to a local disk on the respective node and not to HDFS. Reason for choosing local disk over HDFS is, to avoid replication which takes place in case of HDFS store operation. Map output is intermediate output which is processed by reduce tasks to produce the final output.

189. Define Speculative execution?

A.) Speculative execution is an optimization technique where a computer system performs some task that may not be needed. Work is done before it is known whether it is actually needed, so as to prevent a delay that would have to be incurred by doing the work after it is known that it is needed.

190. is it legal to set the number of reducer tasks to zero?

A.) Yes, It is legal to set the number of reduce-tasks to zero if there is no need for a reducer. In this case the outputs of the map task is directly stored into the HDFS which is specified in the setOutputPath

191. where does the data of hive table gets stored?

A.) Hive stores data at the HDFS location /user/hive/warehouse folder if not specified a folder using the LOCATION clause while creating a table.

192. Why hdfs is not used by hive metastore for storage?

A.) Because HDFS is slow, and due to it's distributed and dynamic nature, once something is stored in HDFS, it would be really hard to find it without proper metadata… So the metadata is kept in memory in a special (usually dedicated) server called the namenode ready to be queried.

193. when should we use sort by and order by?

A.) When there is a large dataset then one should go for sort by as in sort by , all the set reducers sort the data internally before clubbing together and that enhances the performance. While in Order by, the performance for the larger dataset reduces as all the data is passed through a single reducer which increases the load and hence takes longer time to execute the query.

194. How hive distribute in the rows into buckets?

A.) Distribute BY clause used on tables present in Hive. Hive uses the columns in Distribute by to distribute the rows among reducers. All Distribute BY columns will go to the same reducer.

195. what do you mean by data locality?

A.) In Hadoop, Data locality is the process of moving the computation close to where the actual data resides on the node, instead of moving large data to computation. This minimizes network congestion and increases the overall throughput of the system.

196. what are the installation modes in Hadoop?

A.) Standalone Mode.

Pseudo-Distributed Mode.

Fully Distributed Mode.

197. what is the role of combiner in hadoop?

A.) Combiner that plays a key role in reducing network congestion. The main job of Combiner a "Mini-Reducer is to handle the output data from the Mapper, before passing it to Reducer.

198. what is the role of partitoner in hadoop?

A.) The Partitioner in MapReduce controls the partitioning of the key of the intermediate mapper output. By hash function, key (or a subset of the key) is used to derive the partition. A total number of partitions depends on the number of reduce task.

199. Difference between Rdbms and noSql?

A.) RDBMS is called relational databases while NoSQL is called a distributed database. They do not have any relations between any of the databases. When RDBMS uses structured data to identify the primary key, there is a proper method in NoSQL to use unstructured data. RDBMS is scalable vertically and NoSQL is scalable horizontally.

200. what is column family?

A.) A column family is a database object that contains columns of related data. It is a tuple (pair) that consists of a key-value pair, where the key is mapped to a value that is a set of columns.

201. How do reducers communicate with each other?

A.) Yes, reducers can communicate with each other by dispatching intermediate key value pairs that get shuffled to another reduce C. Yes, reducers running on the same machine can communicate with each other through shared memory, but not reducers on different machines.

202. Name the components of spark Ecosystem?

A.) Spark Core

2.Spark SQL

3.Spark Streaming

4.MLlib

5.GraphX

203. what is block report in spark?

A.) Block or report user Block or report isspark. Block user. Prevent this user from interacting with your repositories and sending you notifications.

204. what is distributed cache?

A.) In computing, a distributed cache is an extension of the traditional concept of cache used in a single locale. A distributed cache may span multiple servers so that it can grow in size and in transactional capacity.

205. Normalization vs Denormalization?

A.) Normalization is the process of dividing larger tables in to smaller ones reducing the redundant data, while denormalization is the process of adding redundant data to optimize performance. – Normalization is carried out to prevent databases anomalies.

206. how can you optimize the mapreduce jobs?

A.) Proper configuration of your cluster.

LZO compression usage.

Proper tuning of the number of MapReduce tasks.

Combiner between Mapper and Reducer.

207. what are the advantages of combiner?

A.) Use of combiner reduces the time taken for data transfer between mapper and reducer.

Combiner improves the overall performance of the reducer.

It decreases the amount of data that reducer has to process.

208. what are different schedulers in yarn?

A.) There are three types of schedulers available in YARN: FIFO, Capacity and Fair. FIFO (first in, first out) is the simplest to understand and does not need any configuration. It runs the applications in submission order by placing them in a queue.

209. Explain Hive metastore and Warehouse?

A.) A Hive metastore warehouse (aka spark-warehouse) is the directory where Spark SQL persists tables whereas a Hive metastore (aka metastore_db) is a relational database to manage the metadata of the persistent relational entities, e.g. databases, tables, columns, partitions.

210. Difference between Hive vs beeline?

A.) The primary difference between the Hive CLI & Beeline involves how the clients connect to ApacheHive. The Hive CLI, which connects directly to HDFS and the Hive Metastore, and can be used only on a host with access to those services. Beeline, which connects to HiveServer2 and requires access to only one .jar file: hive-jdbc-version-standalone.jar

211. what are temporary tables in hive?

A.) temporary table is a convenient way for an application to automatically manage intermediate data generated during a complex query. Rather than manually deleting tables needed only as temporary data in a complex query, Hive automatically deletes all temporary tables at the end of the Hive session in which they are created.

212. what is lateral view?

A.) The LATERAL VIEW statement is used with user-defined table generating functions such as EXPLODE() to flatten the map or array type of a column.The explode function can be used on both ARRAY and MAP with LATERAL VIEW.

213. what is the purpose of view in hive?

A.) Views are similar to tables, which are generated based on the requirements. We can save any result set data as a view in Hive ; Usage is similar to as views used in SQL

214. Handling nulls while importing data?

A.) To force Sqoop to leave NULL value blank during import, put the following options in the Sqoop command line: –null-string The string to be written for a null value for string columns. –null-non-string The string to be written for a null value for non-string columns.

215. how is spark better than Hive?

A.) Hive is the best option for performing data analytics on large volumes of data using SQLs. Spark, on the other hand, is the best option for running big data analytics. It provides a faster, more modern alternative to MapReduce.

216. Processing of big tables in spark?

A.) Spark uses SortMerge joins to join large table. It consists of hashing each row on both table and shuffle the rows with the same hash into the same partition. There the keys are sorted on both side and the sortMerge algorithm is applied.

217. Benifits of window function in spark?

A.) Spark Window functions are used to calculate results such as the rank, row number e.t.c over a range of input rows and these are available to you by importing org.apache.spark.sql.functions._

218. Difference between window functions and group by?

A.) GROUP BY functionality only offers aggregate functions; whereas Window functions offer aggregate, ranking, and value functionalities. SQL Window function is efficient and powerful. It not only offers GROUP BY aggregate functionality but advanced analytics with ranking and value options.

219. How can we add a column to dataframe?

A.) use withColumn () transformation function.

220. Difference between logical and physical plan?

A.) Logical Plan just depicts what I expect as output after applying a series of transformations like join, filter, where, groupBy, etc clause on a particular table. Physical Plan is responsible for deciding the type of join, the sequence of the execution of filter, where, groupBy clause, etc. This is how SPARK SQL works internally!

221. Benifits of scala over python?

A.) Scala is a statically typed language that provides an interface to catch the compile time errors. Thus refactoring code in Scala is much easier and ideal than Python. Being a dynamic programming language, testing process, and its methodologies are much complex in Python

222. How to enforce schema on a data frame?

A.) What Is Schema Enforcement? Schema enforcement, also known as schema validation, is a safeguard in Delta Lake that ensures data quality by rejecting writes to a table that do not match the table's schema.

223. Benifits of enforce schema over default schema?

A.) Because objects are no longer tied to the user creating them, users can now be defined with a default schema. The default schema is the first schema that is searched when resolving unqualified object names.

224. what are the challenges faced in spark?

A.) No space left on device:

This is primarily due to executor memory, try increasing the executor memory. Example --executor-memory 20G

Caused by: org.apache.spark.SparkException: Exception thrown in awaitResult:

The default spark.sql.broadcastTimeout is 300 Timeout in seconds for the broadcast wait time in broadcast joins.

To overcome this problem increase the timeout time as per required example

--conf "spark.sql.broadcastTimeout= 1200"

225.How is scala is different from other languages?

A) Though there are a lot of similarities between the two, there are many more differences between them. Scala, when compared to Java, is relatively a new language. It is a machine-compiled language, whereas Java is object-oriented. Scala has enhanced code readability and conciseness.

226. what is functional programming in scala?

A.) Functional programming is a programming paradigm that uses functions as the central building block of programs.

In functional programming, we strive to use pure functions and immutable values.

227. what is spark config?

A.) SparkConf allows you to configure some of the common properties (e.g. master URL and application name), as well as arbitrary key-value pairs through the set () method.

228. can we configure cpu cores in spark context?

A.) The more cores we have, the more work we can do. In spark, this controls the number of parallel tasks an executor can run. From the driver code, SparkContext connects to cluster manager (standalone/Mesos/YARN).

229. how does partition happen while creating RDD?

A.) In case of compressed file you would get a single partition for a single file (as compressed text files are not splittable). When you call rdd.repartition (x) it would perform a shuffle of the data from N partititons you have in rdd to x partitions you want to have, partitioning would be done on round robin basis.

230. To rename a column in Dataframe to some other name? how to achieve that?

A.) Using Spark withColumn() function we can add , rename , derive, split etc a Dataframe Column.There are many other things which can be achieved using withColumn() which we will check one by one with suitable examples.]

231. Difference between spark 1.6 and 2.x?

A.) Even though Spark is very faster compared to Hadoop, Spark 1.6x has some performance issues which are corrected in Spark 2.x.

they are

sparkSession

Faster analysis

added SQL features

MLib improvements

New streaming module

Unified dataset and data frame API's

232. How do you decide number of executors?

A.) Number of executors is related to the amount of resources, like cores and memory, you have in each worker.

233. How to remove duplicates from an array of elemets?

A.) The ways for removing duplicate elements from the array:

Using extra space

Constant extra space

Using Set

Using Frequency array

Using HashMap

234. what is diamond problem in spark and how to resolve it?

A.) Diamond problem occurs when we use multiple inheritance in programming languages like C++ or Java.

The solution to the diamond problem is default methods and interfaces. We can achieve multiple inheritance by using these two things. The default method is similar to the abstract method. The only difference is that it is defined inside the interfaces with the default implementation.

Pyspark Interview Questions

1. What's the difference between an RDD, a DataFrame, and a DataSet?

RDD-

It is Spark's structural square. RDDs contain all datasets and dataframes.

If a similar arrangement of data needs to be calculated again, RDDs can be efficiently reserved.

It's useful when you need to do low-level transformations, operations, and control on a dataset.

It's more commonly used to alter data with functional programming structures than with domain-specific expressions.

DataFrame-

It allows the structure, i.e., lines and segments, to be seen. You can think of it as a database table.

Optimized Execution Plan- The catalyst analyzer is used to create query plans.

One of the limitations of dataframes is Compile Time Wellbeing, i.e., when the structure of information is unknown, no control of information is possible.

Also, if you're working on Python, start with DataFrames and then switch to RDDs if you need more flexibility.

DataSet (A subset of DataFrames)-

It has the best encoding component and, unlike information edges, it enables time security in an organized manner.

If you want a greater level of type safety at compile-time, or if you want typed JVM objects, Dataset is the way to go.

Also, you can leverage datasets in situations where you are looking for a chance to take advantage of Catalyst optimization or even when you are trying to benefit from Tungsten's fast code generation.

2. How can you create a DataFrame a) using existing RDD, and b) from a CSV file?

Here's how we can create DataFrame using existing RDDs-

The toDF() function of PySpark RDD is used to construct a DataFrame from an existing RDD. The DataFrame is constructed with the default column names "_1" and "_2" to represent the two columns because RDD lacks columns.

```
dfFromRDD1 = rdd.toDF()
```

```
dfFromRDD1.printSchema()
```

Here, the printSchema() method gives you a database schema without column names-

```
root
```

|-- _1: string (nullable = true)

|-- _2: string (nullable = true)

Use the toDF() function with column names as parameters to pass column names to the DataFrame, as shown below.-

```
columns = ["language","users_count"]

dfFromRDD1 = rdd.toDF(columns)

dfFromRDD1.printSchema()
```

The above code snippet gives you the database schema with the column names-

```
root

 |-- language: string (nullable = true)

 |-- users: string (nullable = true)
```

3. Explain the use of StructType and StructField classes in PySpark with examples.

The StructType and StructField classes in PySpark are used to define the schema to the DataFrame and create complex columns such as nested struct, array, and map columns. StructType is a collection of StructField objects that determines column name, column data type, field nullability, and metadata.

PySpark imports the StructType class from pyspark.sql.types to describe the DataFrame's structure. The DataFrame's printSchema() function displays StructType columns as "struct."

To define the columns, PySpark offers the pyspark.sql.types import StructField class, which has the column name (String), column type (DataType), nullable column (Boolean), and metadata (MetaData).

Example showing the use of StructType and StructField classes in PySpark-

```
import pyspark

from pyspark.sql import SparkSession

from pyspark.sql.types import StructType,StructField, StringType, IntegerType

spark = SparkSession.builder.master("local[1]") \

                    .appName('ProjectPro') \

                    .getOrCreate()

data = [("James","","William","36636","M",3000),

   ("Michael","Smith","","40288","M",4000),

   ("Robert","","Dawson","42114","M",4000),

   ("Maria","Jones","39192","F",4000)

 ]

schema = StructType([ \

   StructField("firstname",StringType(),True), \

   StructField("middlename",StringType(),True), \
```

```
    StructField("lastname",StringType(),True), \
    StructField("id", StringType(), True), \
    StructField("gender", StringType(), True), \
    StructField("salary", IntegerType(), True) \
  ])

df = spark.createDataFrame(data=data,schema=schema)

df.printSchema()

df.show(truncate=False)
```

4. What are the different ways to handle row duplication in a PySpark DataFrame?

There are two ways to handle row duplication in PySpark dataframes. The distinct() function in PySpark is used to drop/remove duplicate rows (all columns) from a DataFrame, while dropDuplicates() is used to drop rows based on one or more columns.

Here's an example showing how to utilize the distinct() and dropDuplicates() methods-

First, we need to create a sample dataframe.

```
import pyspark

from pyspark.sql import SparkSession

from pyspark.sql.functions import expr

spark = SparkSession.builder.appName('ProjectPro').getOrCreate()
```

```
data = [("James", "Sales", 3000), \
    ("Michael", "Sales", 4600), \
    ("Robert", "Sales", 4100), \
    ("Maria", "Finance", 3000), \
    ("James", "Sales", 3000), \
    ("Scott", "Finance", 3300), \
    ("Jen", "Finance", 3900), \
    ("Jeff", "Marketing", 3000), \
    ("Kumar", "Marketing", 2000), \
    ("Saif", "Sales", 4100) \
  ]

column= ["employee_name", "department", "salary"]

df = spark.createDataFrame(data = data, schema = column)

df.printSchema()

df.show(truncate=False)
```

```
#Distinct

distinctDF = df.distinct()

print("Distinct count: "+str(distinctDF.count()))

distinctDF.show(truncate=False)

#Drop duplicates

df2 = df.dropDuplicates()

print("Distinct count: "+str(df2.count()))

df2.show(truncate=False)

#Drop duplicates on selected columns

dropDisDF = df.dropDuplicates(["department","salary"])

print("Distinct count of department salary : "+str(dropDisDF.count()))

dropDisDF.show(truncate=False)

}
```

5. Explain PySpark UDF with the help of an example.

The most important aspect of Spark SQL & DataFrame is PySpark UDF (i.e., User Defined Function), which is used to expand PySpark's built-in capabilities. UDFs in PySpark work similarly to UDFs in conventional databases. We write a Python function and wrap it in PySpark SQL udf() or register it as udf and use it on DataFrame and SQL, respectively, in the case of PySpark.

Example of how we can create a UDF-

First, we need to create a sample dataframe.

```
spark = SparkSession.builder.appName('ProjectPro').getOrCreate()

column = ["Seqno","Name"]

data = [("1", "john jones"),

    ("2", "tracey smith"),

    ("3", "amy sanders")]

df = spark.createDataFrame(data=data,schema=column)

df.show(truncate=False)
```

Output-

The next step is creating a Python function. The code below generates the convertCase() method, which accepts a string parameter and turns every word's initial letter to a capital letter.

```
def convertCase(str):

    resStr=""

    arr = str.split(" ")

    for x in arr:

      resStr= resStr + x[0:1].upper() + x[1:len(x)] + " "
```

```
    return resStr
```

The final step is converting a Python function to a PySpark UDF.

By passing the function to PySpark SQL udf(), we can convert the convertCase() function to UDF(). The org.apache.spark.sql.functions.udf package contains this function. Before we use this package, we must first import it.

The org.apache.spark.sql.expressions.UserDefinedFunction class object is returned by the PySpark SQL udf() function.

```
""" Converting function to UDF """

convertUDF = udf(lambda z: convertCase(z),StringType())
```

6. Discuss the map() transformation in PySpark DataFrame with the help of an example.

PySpark map or the map() function is an RDD transformation that generates a new RDD by applying 'lambda', which is the transformation function, to each RDD/DataFrame element. RDD map() transformations are used to perform complex operations such as adding a column, changing a column, converting data, and so on. Map transformations always produce the same number of records as the input.

7. What do you mean by 'joins' in PySpark DataFrame? What are the different types of joins?

Joins in PySpark are used to join two DataFrames together, and by linking them together, one may join several DataFrames. INNER Join, LEFT OUTER Join, RIGHT OUTER Join, LEFT ANTI Join, LEFT SEMI Join, CROSS Join, and SELF Join are among the SQL join types it supports.

PySpark Join syntax is-

```
join(self, other, on=None, how=None)
```

The join() procedure accepts the following parameters and returns a DataFrame-

‘other’: The join's right side;

‘on’: the join column's name;

‘how’: default inner (Options are inner, cross, outer, full, full outer, left, left outer, right, right outer, left semi, and left anti.)

8. What is PySpark ArrayType? Explain with an example.

PySpark ArrayType is a collection data type that extends PySpark's DataType class, which is the superclass for all kinds. The types of items in all ArrayType elements should be the same. The ArraType() method may be used to construct an instance of an ArrayType. It accepts two arguments: valueType and one optional argument valueContainsNull, which specifies whether a value can accept null and is set to True by default. valueType should extend the DataType class in PySpark.

```
from pyspark.sql.types import StringType, ArrayType

arrayCol = ArrayType(StringType(),False)
```

9. What do you understand by PySpark Partition?

Using one or more partition keys, PySpark partitions a large dataset into smaller parts. When we build a DataFrame from a file or table, PySpark creates the DataFrame in memory with a specific number of divisions based on specified criteria. Transformations on partitioned data run quicker since each partition's transformations are executed in parallel. Partitioning in memory (DataFrame) and partitioning on disc (File system) are both supported by PySpark.

10. What is meant by PySpark MapType? How can you create a MapType using StructType?

PySpark MapType accepts two mandatory parameters- keyType and valueType, and one optional boolean argument valueContainsNull.

Here’s how to create a MapType with PySpark StructType and StructField. The StructType() accepts a list of StructFields, each of which takes a fieldname and a value type.

```
from pyspark.sql.types import StructField, StructType, StringType, MapType

schema = StructType([
```

```
    StructField('name', StringType(), True),
    StructField('properties', MapType(StringType(),StringType()),True)
])
```

Now, using the preceding StructType structure, let's construct a DataFrame-

```
spark= SparkSession.builder.appName('PySpark StructType StructField').getOrCreate()
dataDictionary = [
        ('James',{'hair':'black','eye':'brown'}),
        ('Michael',{'hair':'brown','eye':None}),
        ('Robert',{'hair':'red','eye':'black'}),
        ('Washington',{'hair':'grey','eye':'grey'}),
        ('Jefferson',{'hair':'brown','eye':''})
        ]
df = spark.createDataFrame(data=dataDictionary, schema = schema)
df.printSchema()
df.show(truncate=False)
```

11. How can PySpark DataFrame be converted to Pandas DataFrame?

First, you need to learn the difference between the PySpark and Pandas. The key difference between Pandas and PySpark is that PySpark's operations are quicker than Pandas' because of its distributed nature and parallel execution over several cores and computers.

In other words, pandas use a single node to do operations, whereas PySpark uses several computers.

You'll need to transfer the data back to Pandas DataFrame after processing it in PySpark so that you can use it in Machine Learning apps or other Python programs.

To convert a PySpark DataFrame to a Python Pandas DataFrame, use the toPandas() function. toPandas() gathers all records in a PySpark DataFrame and delivers them to the driver software; it should only be used on a short percentage of the data. When using a bigger dataset, the application fails due to a memory error.

12. What is the function of PySpark's pivot() method?

The pivot() method in PySpark is used to rotate/transpose data from one column into many Dataframe columns and back using the unpivot() function (). Pivot() is an aggregation in which the values of one of the grouping columns are transposed into separate columns containing different data.

To determine the entire amount of each product's exports to each nation, we'll group by Product, pivot by Country, and sum by Amount.

```
pivotDF = df.groupBy("Product").pivot("Country").sum("Amount")
```

```
pivotDF.printSchema()
```

```
pivotDF.show(truncate=False)
```

This will convert the nations from DataFrame rows to columns, resulting in the output seen below. Wherever data is missing, it is assumed to be null by default.

13. In PySpark, how do you generate broadcast variables? Give an example.

Broadcast variables in PySpark are read-only shared variables that are stored and accessible on all nodes in a cluster so that processes may access or use them. Instead of sending this information with each job, PySpark uses efficient broadcast algorithms to distribute broadcast variables among workers, lowering communication costs.

The broadcast(v) function of the SparkContext class is used to generate a PySpark Broadcast. This method accepts the broadcast parameter v.

Generating broadcast in PySpark Shell:

```
broadcastVariable = sc.broadcast(Array(0, 1, 2, 3))

broadcastVariable.value
```

PySpark RDD Broadcast variable example

```
spark=SparkSession.builder.appName('SparkByExample.com').getOrCreate()

states = {"NY":"New York", "CA":"California", "FL":"Florida"}

broadcastStates = spark.sparkContext.broadcast(states)

data = [("James","Smith","USA","CA"),

   ("Michael","Rose","USA","NY"),

   ("Robert","Williams","USA","CA"),

   ("Maria","Jones","USA","FL")

 ]
```

14. Under what scenarios are Client and Cluster modes used for deployment?

Cluster mode should be utilized for deployment if the client computers are not near the cluster. This is done to prevent the network delay that would occur in Client mode while communicating between executors. In case of Client mode, if the machine goes offline, the entire operation is lost.

Client mode can be utilized for deployment if the client computer is located within the cluster. There will be no network latency concerns because the computer is part of the cluster, and the cluster's maintenance is already taken care of, so there is no need to be concerned in the event of a failure.

15. Write a spark program to check whether a given keyword exists in a huge text file or not?

```
def keywordExists(line):

  if (line.find(“my_keyword”) > -1):

     return 1

  return 0

lines = sparkContext.textFile(“sample_file.txt”);

isExist = lines.map(keywordExists);

sum=isExist.reduce(sum);

print(“Found” if sum>0 else “Not Found”)
```

16. What is meant by Executor Memory in PySpark?

Spark executors have the same fixed core count and heap size as the applications created in Spark. The heap size relates to the memory used by the Spark executor, which is controlled by the -executor-memory flag's property spark.executor.memory. On each worker node where Spark operates, one executor is assigned to it. The executor memory is a measurement of the memory utilized by the application's worker node.

17. List some of the functions of SparkCore.

The core engine for large-scale distributed and parallel data processing is SparkCore. The distributed execution engine in the Spark core provides APIs in Java, Python, and Scala for constructing distributed ETL applications.

Memory management, task monitoring, fault tolerance, storage system interactions, work scheduling, and support for all fundamental I/O activities are all performed by Spark Core. Additional libraries on top of Spark Core enable a variety of SQL, streaming, and machine learning applications.

They are in charge of:

Fault Recovery

Interactions between memory management and storage systems

Monitoring, scheduling, and distributing jobs

Fundamental I/O functions

18. What are some of the drawbacks of incorporating Spark into applications?

Despite the fact that Spark is a strong data processing engine, there are certain drawbacks to utilizing it in applications.

When compared to MapReduce or Hadoop, Spark consumes greater storage space, which may cause memory-related issues.

Spark can be a constraint for cost-effective large data processing since it uses "in-memory" calculations.

When working in cluster mode, files on the path of the local filesystem must be available at the same place on all worker nodes, as the task execution shuffles across different worker nodes based on resource availability. All worker nodes must copy the files, or a separate network-mounted file-sharing system must be installed.

19. How can data transfers be kept to a minimum while using PySpark?

The process of shuffling corresponds to data transfers. Spark applications run quicker and more reliably when these transfers are minimized. There are quite a number of approaches that may be used to reduce them. They are as follows:

Using broadcast variables improves the efficiency of joining big and small RDDs.

Accumulators are used to update variable values in a parallel manner during execution.

Another popular method is to prevent operations that cause these reshuffles.

20. What are Sparse Vectors? What distinguishes them from dense vectors?

Sparse vectors are made up of two parallel arrays, one for indexing and the other for storing values. These vectors are used to save space by storing non-zero values. E.g.- val sparseVec: Vector = Vectors.sparse(5, Array(0, 4), Array(1.0, 2.0))

The vector in the above example is of size 5, but the non-zero values are only found at indices 0 and 4.

When there are just a few non-zero values, sparse vectors come in handy. If there are just a few zero values, dense vectors should be used instead of sparse vectors, as sparse vectors would create indexing overhead, which might affect performance.

The following is an example of a dense vector:

val denseVec = Vectors.dense(4405d,260100d,400d,5.0,4.0,198.0,9070d,1.0,1.0,2.0,0.0)

The usage of sparse or dense vectors has no effect on the outcomes of calculations, but when they are used incorrectly, they have an influence on the amount of memory needed and the calculation time.

21. What role does Caching play in Spark Streaming?

The partition of a data stream's contents into batches of X seconds, known as DStreams, is the basis of Spark Streaming. These DStreams allow developers to cache data in memory, which may be particularly handy if the data from a DStream is utilized several times. The cache() function or the persist() method with proper persistence settings can be used to cache data. For input streams receiving data through networks such as Kafka, Flume, and others, the default persistence level setting is configured to achieve data replication on two nodes to achieve fault tolerance.

Cache method-

val cacheDf = dframe.cache()

Persist method-

val persistDf = dframe.persist(StorageLevel.MEMORY_ONLY)

The following are the key benefits of caching:

Cost-effectiveness: Because Spark calculations are costly, caching aids in data reuse, which leads to reuse computations, lowering the cost of operations.

Time-saving: By reusing computations, we may save a lot of time.

More Jobs Achieved: Worker nodes may perform/execute more jobs by reducing computation execution time.

22. What API does PySpark utilize to implement graphs?

Spark RDD is extended with a robust API called GraphX, which supports graphs and graph-based calculations. The Resilient Distributed Property Graph is an enhanced property of Spark RDD that is a directed multi-graph with many parallel edges. User-defined characteristics are associated with each edge and vertex. Multiple connections between the same set of vertices are shown by the existence of parallel edges. GraphX offers a collection of operators that can allow graph computing, such as subgraph, mapReduceTriplets, joinVertices, and so on. It also offers a wide number of graph builders and algorithms for making graph analytics chores easier.

23. What is meant by Piping in PySpark?

According to the UNIX Standard Streams, Apache Spark supports the pipe() function on RDDs, which allows you to assemble distinct portions of jobs that can use any language. The RDD transformation may be created using the pipe() function, and it can be used to read each element of the RDD as a String. These may be altered as needed, and the results can be presented as Strings.

24. What steps are involved in calculating the executor memory?

Suppose you have the following details regarding the cluster:

No. of nodes = 10

No. of cores in each node = 15 cores

RAM of each node = 61GB

We use the following method to determine the number of cores:

No. of cores = How many concurrent tasks the executor can handle.

As a rule of thumb, 5 is the best value.

Hence, we use the following method to determine the number of executors:

No. of executors = No. of cores/Concurrent Task

= 15/5

= 3

No. of executors = No. of nodes * No. of executors in each node

= 10 * 3

= 30 executors per Spark job

25. Do we have a checkpoint feature in Apache Spark?

Yes, there is an API for checkpoints in Spark. The practice of checkpointing makes streaming apps more immune to errors. We can store the data and metadata in a checkpointing directory. If there's a failure, the spark may retrieve this data and resume where it left off.

In Spark, checkpointing may be used for the following data categories-

Metadata checkpointing: Metadata rmeans information about information. It refers to storing metadata in a fault-tolerant storage system such as HDFS. You can consider configurations, DStream actions, and unfinished batches as types of metadata.

Data checkpointing: Because some of the stateful operations demand it, we save the RDD to secure storage. The RDD for the next batch is defined by the RDDs from previous batches in this case.

26. In Spark, how would you calculate the total number of unique words?

1. Open the text file in RDD mode:

```
sc.textFile("hdfs://Hadoop/user/sample_file.txt");
```

2. A function that converts each line into words:

```
def toWords(line):

return line.split();
```

3. As a flatMap transformation, run the toWords function on each item of the RDD in Spark:

```
words = line.flatMap(toWords);
```

4. Create a (key,value) pair for each word:

```
def toTuple(word):

return (word, 1);

wordTuple = words.map(toTuple);
```

5. Run the reduceByKey() command:

```
def sum(x, y):


return x+y:


counts = wordsTuple.reduceByKey(sum)
```

6. Print:

```
counts.collect()
```

27. List some of the benefits of using PySpark.

PySpark is a specialized in-memory distributed processing engine that enables you to handle data in a distributed fashion effectively.

PySpark-based programs are 100 times quicker than traditional apps.

You can learn a lot by utilizing PySpark for data intake processes. PySpark can handle data from Hadoop HDFS, Amazon S3, and a variety of other file systems.

Through the use of Streaming and Kafka, PySpark is also utilized to process real-time data.

You can use PySpark streaming to swap data between the file system and the socket.

PySpark contains machine learning and graph libraries by chance.

Azure Interview Questions

1. What is meant by Microsoft Azure and Azure diagnostic?

Microsoft Azure is a cloud computing interface that is implemented by Microsoft so as to get benefited from cloud computing.

Azure diagnostics is an API based system that collects the data to diagnose the application which is constantly running. It tunes with the verbose monitoring by enabling roles of the cloud services.

2. What is meant by cloud computing?

Cloud Computing is the high-level abstraction procedure that focuses on business logic. This is a service delivered via the internet that aids you with the computing services without laying much importance on the infrastructural needs

3. What is the scalability of cloud computing?

Vertical scaling, where the configuration yields to increase the existing capacity of the machine. Just like expanding the size of the RAM from 4GB to 32GB.

Horizontal Scaling, an aspect where the physical aspect is increased like putting multiple machines at work instead of replacing the existing machine.

4. What are the advantages of cloud computing?

The versatility of the system

They are highly available.

The system is capable of fault tolerance.

The service allows you to pay as you go.

5. What is meant by PaaS, SaaS, and IaaS?

Platform as a Service that enables you to get a platform to deliver without directly giving authorization to the OS software.

Software as a Service is devoid of platform infrastructure software that can be used without direct purchase.

Infrastructure as a Service which enables you to get the hardware from the provider as the desired service which can be configured by the user.

6. Explain the different deployment model of the cloud?

Private Cloud Deployment Model

Public Cloud Deployment Model

Hybrid Cloud Deployment Model

7. What are the main functions of the Azure Cloud Service?

The main functions of the Azure Cloud Service are;

It is designed to host the running application and at the same time manage the background running application.

The application of web processing is termed as “web role” whereas the background processing is termed as the “worker role”.

8. State the purpose of the cloud configuration file?

There is a primary .csfg file available with each and every cloud service. The main purpose of this file is

They hold the main copy of certificates.

They have the storage of user-defined settings.

There are a number of instances in any service project.

9. Which services are used to manage the resources in Azure?

Azure resource manager is the infrastructure that is involved in managing deploys or deleting all the resources.

10. What do you mean by roles?

Roles in cloud management are often termed to be nothing servers that are linked to managing and balancing the platform

11. What are the different types of roles?

Web Role

VM Role

Worker Role

12. What do you mean by a domain?

The interconnected and interlinked nodes that are often a measure undertaken by the organization is known as the domain. These relations are carried by only one point of the organization.

13. Explain the fault domain.

it is a logical working domain in which the underlying hardware is sharing a common power source and switch network. This means that when VMs is created the Azure distributes the VM across the fault domain that limits the potential impact of hardware failure, power interruption or outages of the network.

14. What do you mean by a BLOB and what are their types?

BLOB is a Binary Large Object that is composed of any size and type of file. They are mainly of two types-the page and the block blob.

15. What is meant by the block blob and page BLOB?

Blob is a block that is having a specific block ID. Each block in this block BLOB comprises of the 4MB and maximum size of this BLOB limits to 200 GB. Whereas the Page blob contains pages in which data range is determined by the offsets. The maximum limit is 1TB where a single page is of the size 1TB.

16. What is meant by the Dead Letter queue?

Messages are transferred to the Dead Letter queue in the following situation;

When the delivery count has exceeded for a message that is on a queue.

When the expiry date of the message has crossed and the entire expired message is held in a queue.

When there is an evaluation exception set by default and the subscription is enabled with dead letter filter.

17. How is the price of the Azure subscription placed?

The Free Model

The BYOL Scheme

The Trial of the free software

Usage Based fee

Monthly Bills.

18. What are the sizes of the Azure VM?

The extra large computer has 8*1.6 GHz of Instance size, with instance storage of 2040 GB, CPU memory of 14 GB. The I/O performance is high.

The large computer has 4*1.6 GHz of Instance size, with instance storage of 1000 GB, CPU memory of 7 GB. The I/O performance is high.

The medium computer has 2*1.6 GHz of Instance size, with instance storage of 490 GB, CPU memory of 3.5 GB. The I/O performance is high.

Small computer has 1.6 GHz of Instance size, with instance storage of 225 GB, CPU memory of 1.75 GB. The I/O performance is moderate.

The extra small computer has 1.0 GHz of Instance size of 20 GB, with instance storage of 20 GB, CPU memory of 768MB. The I/O performance is low.

19. What is meant by table storage?

It is an interface that is capable of storing bulk amount of structured but non-relational data. It is a service of the NoSQL data store that takes authenticated calls from either outside or inside the Azure cloud.

20. Differentiate between the repository and the powerhouse server?

Repository servers are those which are in lieu of the integrity, consistency, and uniformity whereas the powerhouse server governs the integration of different aspects of the database repository.

21. What is meant by the enterprise warehousing?

It is the phenomenon where the data is developed by the organization having access at a single point throughout the globe. The warehousing enables the server to get linked to a single point with the assistance of periodic handling.

22. What do you mean by lookup transformation?

Lookup transformation aids to determine source qualifier. It can be active or passive lookup transformation. The process is yield to get the access the relevant information or the data.

23. What is the primary ETL service in Azure?

Ingest

Control Flow

Data flow

Schedule

Monitor

24. What are data masking features available in Azure?

Dynamic data masking plays various significant roles in data security. It restricts sensitive information to some specific set of users.

It is available for Azure SQL Database, Azure SQL Managed Instance and Azure Synapse Analytics.

It can be implemented as a security policy on all the SQL Databases across an Azure subscription.

Users can control the level of masking as per their requirements.

It only masks the query results for specific column values on which the data masking has been applied. It does not affect the actual stored data in the database.

25. What is Polybase?

Polybase optimises the data ingestion into PDW and supports T-SQL. It enables developers to query external data transparently from supported data stores, irrespective of the storage architecture of the external data store.

Polybase can be used to:

Query Data

Import Data

Export data

26. What is reserved capacity in Azure?

Microsoft provides an option of reserved capacity on Azure storage to optimise the Azure Storage costs. For the reservation period on Azure cloud, the reserved storage provides a fixed amount of capacity to customers. It is available for Block Blobs and Azure Data Lake to Store Gen 2 data in a standard storage account.

27. Which service would you use to create Data Warehouse in Azure?

Azure Synapse is a limitless analytics service that brings together Big Data analytics and enterprise data warehousing. It gives users the freedom to query data on individual terms for using either serverless on-demand or provisioned resources at scale.

28. Explain the architecture of Azure Synapse Analytics

It is designed to process massive amounts of data with hundreds of millions of rows in a table. Azure Synapse Analytics processes complex queries and returns the query results within seconds, even with massive data, because Synapse SQL runs on a Massively Parallel Processing (MPP) architecture that distributes data processing across multiple nodes.

Applications connect to a control node that acts as a point of entry to the Synapse Analytics MPP engine. On receiving the Synapse SQL query, the control node breaks it down into MPP optimised format. Further, the individual operations are forwarded to the compute nodes that can perform the operations in parallel, resulting in much better query performance.

29. Difference between ADLS and Azure Synapse Analytics? Both Azure Data Lake Storage Gen2 and Azure Synapse Analytics are highly scalable and can ingest and process vast amounts of data (on a Peta Byte scale).

ADLS Gen2	Azure Synapse Analytics
Optimised for storing and processing structured and non-structured data	Optimised for processing structured data in a well-defined schema
Used for data exploration and analytics by data scientists and engineers	Used for Business Analytics or disseminating data to business users
Built to work with Hadoop	Built on SQL Server
No regulatory compliance	Compliant with regulatory standards such as HIPAA
USQL (combination of C# and TSQL) and Hadoop are used for accessing data	Synapse SQL (improved version of TSQL) is used for accessing data
Can handle data streaming using tools such as Azure Stream Analytics	Built-in data pipelines and data streaming capabilities

30. What are Dedicated SQL Pools?

Dedicated SQL Pool is a collection of features that enable the implementation of the more traditional Enterprise Data Warehousing platform using Azure Synapse Analytics. The resources are measured in Data Warehousing Units (DWU) that are provisioned using Synapse SQL. A dedicated SQL pool uses columnar storage and relational tables to store data, improving query performance and reducing the required amount of storage.

31. How do you capture streaming data in Azure?

Azure provides a dedicated analytics service called Azure Stream Analytics that provides a simple SQL-based language that is Stream Analytics Query Language. It allows developers to extend the ability of query language by defining additional ML (Machine Learning) functions. Azure Stream Analytics can process a huge amount of data on a scale of over a million events per second and also deliver the results with ultra-low latency.

32. What are the various windowing functions in Azure Stream Analytics?

A window in Azure Stream Analytics refers to a block of time-stamped event data that enables users to perform various statistical operations on the event data.

Tumbling Window

Hopping window

sliding window

session window

33.What are the different types of storage in Azure?

There are five types of storage in Azure:

Azure Blobs: Blob stands for a large binary object. It can support all kinds of files including, text files, videos, images, documents, binary data etc.

Azure Queues: Azure Queues is a cloud-based messaging store for establishing and brokering communication between various applications and components.

Azure Files: It is an organized way of storing data in the cloud. Azure Files has one main advantage over Azure Blobs, it allows organizing the data in a folder structure, and it is SMB compliant, i.e. it can be used as a file share.

Azure Disks: It is used as a storage solution for Azure VMs (Virtual Machines).

Azure Tables: A NoSQL storage solution for storing structured data which does not meet the standard relational database schema.

34. Explore Azure storage explorer and its uses?

It is a versatile standalone application available for Windows, Mac OS and Linux to manage Azure Storage from any platform. Azure Storage can be downloaded from Microsoft.

It provides access to multiple Azure data stores such as ADLS Gen2, Cosmos DB, Blobs, Queues, Tables, etc., with an easy to navigate GUI.

One of the key features of Azure Storage Explorer is that it allows users to work even when they are disconnected from the Azure cloud service by attaching local emulators.

35. What is Azure Databricks, and how is it different from regular data bricks?

It is the Azure implementation of Apache Spark that is an open-source big data processing platform. In the data lifecycle, Azure Databricks lies in the data preparation or processing stage. First of all, data is ingested in Azure using Data Factory and stored in permanent storage (such as ADLS Gen2 or Blob Storage). Further, data is processed using Machine Learning (ML) in Databricks and then extracted insights are loaded into the Analysis Services in Azure like Azure Synapse Analytics or Cosmos DB.

Finally, insights are visualized and presented to the end-users with the help of Analytical reporting tools like Power BI.

36. What is Azure table storage?

It is a storage service optimized for storing structured data. In structured data, table entities are the basic units of data equivalent to rows in a relational database table. Each entity represents a key-value pair, and the properties for table entities are as follows:

PartitionKey

RowKey

Timestamp

37. What is serverless database computing in Azure?

In a typical computing scenario, the program code resides either on the server or the client-side. But Serverless computing follows the stateless code nature, i.e. the code does not require any infrastructure.

Users have to pay for the compute resources used by the code during a short period while executing the code. It is very cost-effective, and users only need to pay for the resources used.

38. What Data security options are available in Azure SQL DB?

Azure SQL firewall Rules

Azure SQL always Encrypted

Azure SQL Transparent Data encryption

Azure SQL databaase auditing

39. What is data redundancy in Azure?

Azure constantly retains several copies of data to provide high levels of data availability. Some data redundancy solutions are accessible to clients in Azure, depending on the criticality and duration necessary to provide access to the replica.

Locally Redundant storage

zone Redundant storage

Geo Redundant storage

Read access geo Redundant storage

40. What are some ways to ingest data from on-premise storage to Azure?

While choosing a data transfer solution, the main factors to consider are:

Data Size

Data Transfer Frequency (One-time or Periodic)

Network Bandwidth

Offline transfer

network transfer: Over a network connection, data transfer can be performed in the following ways:

Graphical interface

programmic interface

onpremises devices

managed data factory pipeline.

41. What is the best way to migrate data from an on-premise database to Azure?

SQL server stretch database

azure sql database

sql server managed instance

sql server on a virtual machine.

42. What are multi-model databases?

Azure Cosmos DB is Microsoft's premier NoSQL service offering on Azure. It is the first globally distributed, multi-model database offered on the cloud by any vendor. It is used to store data in various data storage models such as Key-value pair, document-based, graph-based, column-family based, etc. Low latency, consistency, global distribution and automatic indexing features are the same no matter what data model the customer chooses.

43. What is the Azure Cosmos DB synthetic partition key?

It is crucial to select a good partition key that can distribute the data evenly across multiple partitions. We can create a Synthetic partition key when there is no right column with properly distributed values. The three ways to create a synthetic partition key are:

Concatenate properties

Random suffix

pre calculated suffix

44. What are various consistency models available in Cosmos DB?

strong

Bounded staleness

session

consistent prefix

eventual

45. How is data security implemented in ADLS Gen2?

ADLS Gen2 has a multi-layered security model. They are

Authentication

Access control

Network isolation

data protection

advanced threat protection

auditing

46. What are pipelines and activities in Azure?

ADF activities are grouped into three parts:

Data movement activities

data transformation activities

control activities

47. How do you manually execute the Data factory pipeline?

A pipeline can run with Manual or On-demand execution.

To execute the pipeline manually or programmatically, we can use the PowerShell command.

48. Azure Data Factory: Control Flow vs Data Flow

Control Flow is an activity that affects the path of execution of the Data Factory pipeline. For example, an activity that creates a loop if conditions are met.

Data Flow Transformations are used when we need to transform the input data, for example, Join or Conditional Split.

Control Flow Activity	Data Flow Transformation
It affects the execution sequence or path of the pipeline	Transforms the ingested data
Can be recursive	Non-recursive
No source/sink	Source and sink are required
Implemented at the pipeline level	Implemented at the activity level

49. Name the data flow partitioning schemes in Azure

Partitioning Scheme is a way to optimize the performance of Data Flow. This partitioning scheme setting can be accessed on the Optimize tab of the configuration panel for the Data Flow Activity.

'Use current partitioning' is the default setting recommended by Microsoft in most cases that uses native partitioning schemes.

The 'Single Partition' option is used when users want to output to a single destination, for example, a single file in ADLS Gen2.

Round robin

Hash

Dynamic range

fixed range

key

50. What is the trigger execution in Azure Data Factory?

In Azure Data Factory, pipelines can be triggered or automated.

Schedule trigger

tumbling window trigger

event based trigger

51. What are mapping Dataflows?

Microsoft provides Mapping Data Flows that do not require writing code for a more straightforward data integration experience than Data Factory Pipelines. It is a visual way to design data transformation flows. The data flow becomes Azure Data Factory (ADF) activities and gets executed as a part of the ADF pipelines.

52. What is the role of Azure Data Engineer?

Azure Data Engineers are responsible for integrating, transforming, operating, and consolidating data from structured or unstructured data systems. They also build, implement and support Business Intelligence solutions by applying knowledge of technologies, methodologies, processes, tools and applications. In short, they handle all the data operations stored in the cloud, such as Azure.

53. What is Azure Data Factory?

Cloud-based integration service that allows creating data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.

Using Azure data factory, you can create and schedule the data-driven workflows(called pipelines) that can ingest data from disparate data stores.

54. What is the integration runtime?

The integration runtime is the compute infrastructure that Azure Data Factory uses to provide the following data integration capabilities across various network environments.

Azure integration runtime

self hosted integration runtime

Azure SSIS integration runtime

55. What is the limit on the number of integration runtimes?

There is no hard limit on the number of integration runtime instances you can have in a data factory.

56. What is the difference between Azure Data Lake and Azure Data Warehouse?

Data Warehouse is a traditional way of storing data that is still used widely. Data Lake is complementary to Data Warehouse i.e if you have your data at a data lake that can be stored in the data warehouse as well but there are certain rules that need to be followed.

DATA LAKE	DATA WAREHOUSE
Complementary to data warehouse	Maybe sourced to the data lake
Data is Detailed data or Raw data. It can be in any particular form. you just need to take the data and dump it into your data lake	Data is filtered, summarised, refined
Schema on read (not structured, you can define your schema in n number of ways)	Schema on write(data is written in Structured form or in a particular schema)
One language to process data of any format(USQL)	It uses SQL

57. What is blob storage in Azure?

Azure Blob Storage is a service for storing large amounts of unstructured object data, such as text or binary data. You can use Blob Storage to expose data publicly to the world or to store application data privately. Common uses of Blob Storage include:

Serving images or documents directly to a browser

Storing files for distributed access

Streaming video and audio

Storing data for backup and restore disaster recovery, and archiving

Storing data for analysis by an on-premises or Azure-hosted service

58. What are the steps for creating ETL process in Azure Data Factory?

While we are trying to extract some data from Azure SQL server database, if something has to be processed, then it will be processed and is stored in the Data Lake Store.

Steps for Creating ETL

Create a Linked Service for source data store which is SQL Server Database

Assume that we have a cars dataset

Create a Linked Service for destination data store which is Azure Data Lake Store

Create a dataset for Data Saving

Create the pipeline and add copy activity

Schedule the pipeline by adding a trigger

59. What are the top-level concepts of Azure Data Factory?

Pipeline

Activities

datasets

Linked services

60. How can I schedule a pipeline?

You can use the scheduler trigger or time window trigger to schedule a pipeline.

61. Explain the two levels of security in ADLS Gen2?

The two levels of security applicable to ADLS Gen2 were also in effect for ADLS Gen1. Even though this is not new, it is worth calling out the two levels of security because it's a very fundamental piece to getting started with the data lake and it is confusing for many people just getting started.

Role based access control(RBAC)

Access control Lists

62. Define reserved capacity in Azure.

Microsoft has included a reserved capacity option in Azure storage to optimize costs. The reserved storage gives its customers a fixed amount of capacity during the reservation period on the Azure cloud.

63. What is the linked service in the Azure data factory?

Linked service is one of the components in the Azure data factory which is used to make a connection hence to connect to any of the data sources you have to first create the linked service based upon the type of data source.

64. What is the dataset in the Azure Data factory?

Dataset needs to read-write data to any data source using the ADF. Dataset is the representation of the type of data holds by the data source.

65. What are the parameters in the ADF?

Linked service parameters

Dataset parameters

Pipeline parameters

Global parameters

66. How to check the history of pipeline execution run in ADF?

In the Azure Data factory, we can check the pipeline execution run by going to the monitor tab. There we can search all the pipelines run history. You can search the history based on various parameters like the name of the pipeline, time duration, the status of execution run (Pass/fail) and etc.

67. Why you will use the data flow from the Azure data factory

Data flow is used for a no-code transformation. For example when you are doing any ETL operation that you wanted to do a couple of transformations and put some logic on your input data. You may have not found it comfortable to type the query or when you are using the files as input then in that case you cannot write the query at all. Hence data flow will come as a Savior in this situation. Using the data flow you can just do drag and drop and write almost all your business logic without writing any code. Behind the scene, data flow get converted into the spark code and it will run on the cluster.

68. What is the foreach activity in the data factory?

In the Azure data factory whenever you have to do some of the work repetitively then probably you will be going to use the foreach activity. In the foreach activity, you pass an array and the foreach loop will run for all the items of this array. As of now nested foreach activity is not allow which means you cannot have one foreach activity into another for each activity.

69. What is the get metadata activity in the Azure data factory?

In the Azure data factory get metadata activity is used to get the information about the files.

70. What is the custom activity in the Azure data factory?

Custom activity is used in the Azure data factory to execute the Python or a PowerShell script. Assume that you have some code that is written in the Python or PowerShell script and you wanted to execute as a part of your pipeline. Then you can use a custom activity that will help you to execute the code.

71. How you can connect the Azure data factory with the Azure Databricks?

To connect to Azure databricks we have to create a linked service that will point to the Azure databricks account. Next in the pipeline, you will be going to use the notebook activity there you will provide the linked service created for Databricks. You will also be going to provide the notebook path available in the Azure Databricks workspace. That's how you can use the Databricks from the Data factory.

72. Is it possible to connect MongoDB DB from the Azure data factory?

Yes, it is possible to connect MongoDB from the Azure data factory. You have to provide the proper connection information about the MongoDB server. In case if this MongoDB server is residing outside the Azure workspace then probably you have to create a self-hosted integration runtime, and through which you can connect to the Mongo DB server.

73. Can Azure data factory directly connect to the different Azure services?

Yes, the Azure data factory can connect with various other Azure services like Azure blob storage, Azure functions, logic app. However, for all of them, they have to provide the proper roles using the RBAC.

74. How to connect Azure data factory to GitHub?

Azure data factory can connect to GitHub using the GIT integration. You probably have using the Azure DevOps which has git repo. We can configure the GIT repository path into the Azure data factory. So that all the changes we do in the Azure data factory get automatically sync with the GitHub repository

75. How you can move your changes from one environment to another environment for the Azure data factory?

We can migrate the code from one environment to another environment for the Azure data factory using the ARM template. ARM template is the JSON representation of the pipeline that we have created.

76. What is Synapse SQL?

Synapse SQL is the ability to do T-SQL based analytics in the Synapse workspace. Synapse SQL has two consumption models: dedicated and serverless. For the dedicated model, use dedicated SQL pools. A workspace can have any number of these pools. To use the serverless model, use the serverless SQL pools. Every workspace has one of these pools.

Inside Synapse Studio, you can work with SQL pools by running SQL scripts.

77. How you can use Apache Spark through Azure Synapse Analytics?

In Azure analytics, you can run the spark code either using the notebook or you can create a job that will run the spark code. For running the Spark code you need a Spark pool which is nothing just a cluster of the nodes having a spark installed on it.

78. What are the different types of Synapse SQL pools available?

Azure Synapse Analytics is an analytics service that brings together enterprise data warehousing and Big Data analytics. Dedicated SQL pool (formerly SQL DW) refers to the enterprise data warehousing features that are available in Azure Synapse Analytics. There are two types of Synapse SQL pool

Serverless SQL pool

Dedicated SQL pool

79. What is Delta Lake?

Delta Lake is an open-source storage layer that brings ACID (atomicity, consistency, isolation, and durability) transactions to Apache Spark and big data workloads.

The current version of Delta Lake included with Azure Synapse has language support for Scala, PySpark, and .NET.

80. What is Azure Synapse Runtime?

Apache Spark pools in Azure Synapse use runtimes to tie together essential component versions, Azure Synapse optimizations, packages, and connectors with a specific Apache Spark version. These runtimes will be upgraded periodically to include new improvements, features, and patches.These runtimes have the following advantages:

Faster session startup times

Tested compatibility with specific Apache Spark versions

Access to popular, compatible connectors and open-source packages

81. Can you run machine learning algorithm using Azure Synapse Analytics?

Yes, it is possible to run the machine learning algorithm using Azure synapse Analytics. In the Azure synapse analytics, we have an Apache Spark and there we can write the machine learning code and which can be executed on the Spark cluster.

82. What is Azure Databricks?

Databricks is the organization that provides the Spark based Cluster over the cloud.

83. What are the two different types of execution modes provided by the Databricks?

You can run the Spark code in two modes that are interactive mode or job scheduled mode. In the interactive mode, you can run the code line by line and see the output. In the job mode, it will run all code together, and then you will see the output.

84. What are the two different types of the cluster provided by the Databricks?

Two different types of the cluster provided by the Databricks are the interactive cluster and the job cluster. To run the interactive notebook you will be going to use an interactive cluster and to run the job, we will use the job cluster.

85. What is azure data factory used for ?

Azure Data factory is the data orchestration service provided by the Microsoft Azure cloud. ADF is used for following use cases mainly :

Data migration from one data source to other

On Premise to cloud data migration

ETL purpose

Automated the data flow.

There is huge data laid out there and when you want to move the data from one location to another in automated way within the cloud or from on-premises to the azure cloud azure data factory is the best service we have available.

86. What are the main components of the azure data factory ?

Pipeline

Integration Runtime

Activities

DataSet

Linked Services

Triggers

87. What is the pipeline in the adf ?

Pipeline is the set of the activities specified to run in defined sequence. For achieving any task in the azure data factory we create a pipeline which contains the various types of activity as required for fulfilling the business purpose. Every pipeline must have a valid name and optional list of parameters.

88. What is the data source in the azure data factory ?

It is the source or destination system which contains the data to be used or operate upon. Data could be of anytype like text, binary, json, csv type files or may be audio, video, image files, or may be a proper database. Data source examples are : Azure blob storage, azure data lake storage, any database like azure sql database, mysql db, postgres and etc. There are 80+ different data source connector provided by the azure data factory to get in/out data from the data source.

89. What is the integration runtime in Azure data factory :

It is the powerhouse of the azure data pipeline. Integration runtime is also knows as IR, is the one who provides the computer resources for the data transfer activities and for dispatching the data transfer activities in azure data factory. Integration runtime is the heart of the azure data factory.

In Azure data factory the pipeline is made up of activities. An activity is represents some action that need to be performed. This action could be a data transfer which acquired some execution or it will be dispatch action. Integration runtime provides the area where this activity can execute.

90. What are the different types of integration runtime ?

There are 3 types of the integration runtime available in the Azure data factory.

Azure IR

self hosted

Azure SSIS

91. What is the main advantage of the AutoResolveIntegrationRuntime ?

Advantage of AutoResolveIntegrationRuntime is that it will automatically try to run the activities in the same region if possible or close to the region of the sink data source. This can improve the performance a lot.

92. What is Self Hosted Integration Runtimes in azure data factory?

Self hosted integration runtime as the name suggested, is the IR managed by you itself rather than azure. This will make you responsible for the installation, configuration, maintenance, installing updates and scaling. Now as you host the IR , it can access the on premises network as well.

93. What is the Azure-SSIS Integration Runtimes

As the name suggested the azure-SSIS integration runtimes are actually the set of vm running the SQL Server Integration Services (SSIS) engine, managed by Microsoft. Again the responsibility of the installation, maintenance, are of azure only. Azure Data Factory uses azure-SSIS integration runtime for executing SSIS packages.

94. How to install Self Hosted Integration Runtimes in azure data factory ?

1. Create self hosted integration runtime by simply giving general information like name description.

2. Create Azure VM (If u already have then you can skip this step)

3. Download the integration runtime software on azure virtual machine. and install it.

4. Copy the autogenerated key from step 1 and paste it newly installed integration runtime on azure vm.

95. What is use of lookup activity in azure data factory ?

Lookup activity used to pull the data from source dataset and keep it as the output of the activity. Output of the lookup activity generally used further in the pipeline for making some decision, configuration accordingly.

96. What do you mean by variables in the azure data factory ?

Variables is the adf pipeline provide the functionality to temporary hold the values. They are used for similar reason like we do use variables in the programming language. They are available inside the pipeline and there is set inside the pipeline. Set Variable and append variable are two type of activities used for the setting or manipulating the variables values. There are two types of the variable :

System variable

User Variables

97. There are two ways to create the Linked Service :

Using the Azure Portal

ARM template way

98. Can we debug the pipeline ?

Debugging is one of the key feature for any developer. To solve and test issue in the code developers uses the debug feature in general. Azure data factory also provide the debugging feature.

99. What is the breakpoint in the adf pipeline ?

Debug part of the pipeline using the break points : While doing if you want to check the pipeline up to certain activity you can do it by using the breakpoints.

For example you have 3 activities in the pipeline and you want to debug up to 2nd activity only. You can do this by putting the break point at the 2nd activity.

100. What are the different pricing tier of the azure databricks available ?

There are basically two tier provided by the azure for databricks service :

Standard Tier

Premium Tier

101. How many different types of cluster mode available in the Azure Databricks?

Standard cluster

High concurrency cluster

Single node cluster

102. How can you connect your ADB cluster to your favorite IDE (Eclipse, IntelliJ, PyCharm, RStudio, Visual Studio)?

Databricks connect is the way to connect the databricks cluster to local IDE on your local machine. You need to install the databricks-connect client and then needed the configuration details like ADB url, token etc. Using all these you can configure the local IDE to run and debug the code on the cluster.

103. How typical Azure Databricks CI/CD pipeline consist of?

continuous integratoin

continuous delivery

104. Explain azure cloud services?

Azure Cloud Services is a Paas (platform-as-a-service) product that intends to provide robust, efficient, and cost-effective applications. Azure Cloud Services are hosted on virtual machines. By launching a cloud service instance, Azure cloud services can be utilized to implement multi-tier web-based apps in Azure.

There are two types azure cloud services….

1. Web role
2. Worker role

Azure Data Factory

1. Is azure data factory ETL or ELT tool?

   It is a cloud-based Microsoft tool that provides a cloud-based integration service for data analytics at scale and supports ETL and ELT paradigms.

2. Why is ADF needed?

   With an increasing amount of big data, there is a need for a service like ADF that can orchestrate and operationalize processes to refine the enormous stores of raw business data into actionable business insights.

3. What sets azure data factory apart from conventional ETL tools?
   Azure Data Factory stands out from other ETL tools as it provides: -

   i) Enterprise Readiness: Data integration at Cloud Scale for big data analytics!

   ii) Enterprise Data Readiness: There are 90+ connectors supported to get your data from any disparate sources to the Azure cloud!

   iii) Code-Free Transformation: UI-driven mapping dataflows.

   iv) Ability to run Code on Any Azure Compute: Hands-on data transformations

   v) Ability to rehost on-prem services on Azure Cloud in 3 Steps: Many SSIS packages run on Azure cloud.

vi) Making DataOps seamless: with Source control, automated deploy & simple templates.

vii) Secure Data Integration: Managed virtual networks protect against data exfiltration, which, in turn, simplifies your networking.

4. What are Major components of azure data factory?

Pipelines: Data Factory can contain one or more pipelines, which is a logical grouping of tasks/activities to perform a task. e.g., An activity can read data from Azure blob storage and load it into Cosmos DB or Synapse DB for analytics while transforming the data according to business logic.

Activities: Activities represent a processing step in a pipeline. For example, you might use a copy activity to copy data between data stores. Data Factory supports data movement, transformations, and control activities.

Datasets: Datasets represent data structures within the data stores, which simply point to or reference the data you want to use in your activities as inputs or outputs.

Linked service: This is more like a connection string, which will hold the information that Data Factory can connect to various sources. In the case of reading from Azure Blob storage, the storage-linked service will specify the connection string to connect to the blob, and the Azure blob dataset will select the container and folder containing the data.

Integration Runtime: Integration runtime instances provide the bridge between the activity and linked Service. It is referenced by the linked service or activity and provides the computing environment where the activity either runs on or gets dispatched. This way, the activity can be performed in the region closest to the target data stores or compute service in the most performant way while meeting security (no exposing of data publicly) and compliance needs.

Data Flows: These are objects you build visually in Data Factory, which transform data at scale on backend Spark services. You do not need to understand programming or Spark internals. Just design your data transformation intent using graphs (Mapping) or spreadsheets (Power query activity)

5. What are the different ways to execute pipelines in azure data factory?

Debug mode can be helpful when trying out pipeline code and acts as a tool to test and troubleshoot our code.

ii) Manual Execution is what we do by clicking on the 'Trigger now' option in a pipeline. This is useful if you want to run your pipelines on an ad-hoc basis.

iii) We can schedule our pipelines at predefined times and intervals via a Trigger. As we will see later in this article, there are three types of triggers available in Data Factory.

6. What is the purpose of Linked services in azure data factory?

Linked services are used majorly for two purposes in Data Factory:

1. For a Data Store representation, i.e., any storage system like Azure Blob storage account, a file share, or an Oracle DB/ SQL Server instance.
2. For Compute representation, i.e., the underlying VM will execute the activity defined in the pipeline.

7. Can you elaborate more on data factory integration runtime?

   The Integration Runtime or IR is the compute infrastructure for Azure Data Factory pipelines. It is the bridge between activities and linked services. It's referenced by the linked service or activity and provides the compute environment where the activity is run directly or dispatched. This allows the activity to be performed in the closest region to the target data stores or compute service.

   There are three types of integration runtime supported by Azure Data Factory.

   1, azure integration runtime

   Self hosted integration runtime

   Azure SSIS integration runtime

8. What is the limit on the number of integration runtimes, if any?

   Within a Data Factory, the default limit on any entities is set to *5000*, including pipelines, data sets, triggers, linked services, Private Endpoints, and integration runtimes. One can create an online support ticket to raise the limit to a higher number if required.

9. What are ARM templates in azure data factory? What are they used for?

   An ARM template is a JSON (JavaScript Object Notation) file that defines the infrastructure and configuration for the data factory pipeline, including pipeline activities, linked services, datasets, etc. The template will contain essentially the same code as our pipeline.

   ARM templates are helpful when we want to migrate our pipeline code to higher environments, say Production or Staging from Development, after we are convinced that the code is working correctly.

10. How can we deploy code to higher environments in data factory?

    Create a feature branch that will store our code base.

    Create a pull request to merge the code after we’re sure to the Dev branch.

    Publish the code from dev to generate ARM templates.

    This can trigger an automated CI/CD DevOps pipeline to promote code to higher environments like Staging or Production.

11. Which three activities can you run in azure data factory?

    There are three types.

Data movement activities

Data transformation activities

Control flow activities

12. What are two types of compute environments supported by data factory to execute the transform activities?

    On-demand computing environment

    Bring your own environment

13. What are the steps involved in ETL process?

    The ETL process follow four main steps…

    Connect and collect

    Data transformation using computing services such as HDInsight, Hadoop, Spark etc.

    Publish

    Monitor

14. If you want to use the output by executing a query, which activity shall use?

    Look-up activity can return the result of executing a query or stored procedure.

    The output can be a singleton value or an array of attributes, which can be consumed in subsequent copy data activity, or any transformation or control flow activity like ForEach activity.

15. Have you used execute notebook activity in data factory? How to pass parameters to a notebook activity?

    We can use execute notebook activity to pass code to our databricks cluster. We can pass parameters to a notebook activity using *baseParameters* property. If the parameters are not defined/ specified in the activity, default values from the notebook are executed.

16. What are some useful constructs available in data factory?

    Parameter

    Coalesce

    Activity

17. Is it possible to push code and have CI/CD in ADF?

Data Factory offers full support for CI/CD of your data pipelines using Azure DevOps and GitHub. This allows you to develop and deliver your ETL processes incrementally before publishing the finished product. After the raw data has been refined into a business-ready consumable form, load the data into Azure Data Warehouse or Azure SQL Azure Data Lake, Azure Cosmos DB, or whichever analytics engine your business uses can point to from their business intelligence tools.

18. What do you mean by variables in azure data factory?

Variables in the Azure Data Factory pipeline provide the functionality to hold the values. They are used for a similar reason as we use variables in any programming language and are available inside the pipeline.

System variable

User variables

19. what are mapping data flows?

Mapping data flows are visually designed data transformations in Azure Data Factory. Data flows allow data engineers to develop data transformation logic without writing code. The resulting data flows are executed as activities within Azure Data Factory pipelines that use scaled-out Apache Spark clusters. Data flow activities can be operationalized using existing Azure Data Factory scheduling, control flow, and monitoring capabilities.

Mapping data flows provide an entirely visual experience with no coding required. Data flows run on ADF-managed execution clusters for scaled-out data processing. Azure Data Factory manages all the code translation, path optimization, and execution of the data flow jobs.

20. What is copy activity in azure data factory?

Copy activity is one of the most popular and universally used activity in the Azure data factory. It is used for ETL or Lift and Shift, where you want to move the data from one data source to another. While you copy the data, you can also do the transformation.

21. Can you elaborate more on copy activity?

The copy activity performs the following steps at high-level:

i) Read data from the source data store. (e.g., blob storage)
ii) Perform the following tasks on the data:

- Serialization/deserialization
- Compression/decompression
- Column mapping

iii) Write data to the destination data store or sink. (e.g., azure data lake)

22. What are different activities you performed in azure data factory?

Here you can share some of the major activities if you have used them in your career be it your work or college project. Here are a few of the most used activities :

1. Copy Data Activity to copy the data between datasets.
2. ForEach Activity for looping.
3. Get Metadata Activity which can provide metadata about any data source.
4. Set Variable Activity to define and initiate variables within pipelines.
5. Lookup Activity to do a lookup to get some values from a table/file.
6. Wait Activity to wait for a specified amount of time before/in between the pipeline run.
7. Validation Activity will validate the presence of files within the dataset.
8. Web Activity to call a custom REST endpoint from an ADF pipeline.

23. How can I schedule pipeline?

You can use the time window trigger or scheduler trigger to schedule a pipeline.

Currently, the service supports three types of triggers:

- Tumbling window trigger: A trigger that operates on a periodic interval while retaining a state.
- Schedule Trigger: A trigger that invokes a pipeline on a wall-clock schedule.
- Event-Based Trigger: A trigger that responds to an event. e.g., a file getting placed inside a blob.
  Pipelines and triggers have a many-to-many relationship (except for the tumbling window trigger).

24. When should you choose azure data factory?

One should consider using Data Factory-

i) When working with big data, there is a need for a data warehouse to be implemented; you might require a cloud-based integration solution like ADF for the same.

ii) Not all the team members are experienced in coding and may prefer graphical tools to work with data.

iii) When raw business data is stored at diverse data sources, which can be on-prem and on the cloud, we would like to have one analytics solution like ADF to integrate them all in one place.

iv) We would like to use readily available data movement and processing solutions and like to be light in terms of infrastructure management. So, a managed solution like ADF makes more sense in this case.

25. How can you access data using the other 90 dataset types in data factory?

Use copy activity. The mapping data flow feature allows Azure SQL Database, Azure Synapse Analytics, delimited text files from azure storage account or Azure Data Lake Storage Gen2, and Parquet files from blob storage or Data Lake Storage Gen2 natively for source and sink data source.

26. What is the difference between mapping and wrangling data flow?

Mapping data flows transform data at scale without requiring coding. You can design a data transformation job in the data flow canvas by constructing a series of transformations. Start with any number of source transformations followed by data transformation steps. Complete your data flow with a sink to land your results in a destination. It is excellent at mapping and transforming data with known and unknown schemas in the sinks and sources\.

Power Query Data Wrangling allows you to do agile data preparation and exploration using the Power Query Online mashup editor at scale via spark execution. With the rise of data lakes, sometimes you just need to explore a data set or create a dataset in the lake.

27. Is it possible to calculate a value for a new column from existing column from mapping in ADF?

We can derive transformations in the mapping data flow to generate a new column based on our desired logic. We can create a new derived column or update an existing one when generating a derived one. Enter the name of the column you're creating in the Column textbox.

You can use the column dropdown to override an existing column in your schema. Click the Enter expression textbox to start creating the derived column's expression. You can input or use the expression builder to build your logic.

28. How is lookup activity useful in the data factory?

In the ADF pipeline, the Lookup activity is commonly used for configuration lookup purposes, and the source dataset is available. Moreover, it is used to retrieve the data from the source dataset and then send it as the output of the activity. Generally, the output of the lookup activity is further used in the pipeline for taking some decisions or presenting any configuration as a result.

29. What is get metadata activity in data factory?

The Get Metadata activity is used to retrieve the metadata of any data in the Azure Data Factory or a Synapse pipeline. We can use the output from the Get Metadata activity in conditional expressions to perform validation or consume the metadata in subsequent activities.

30. How to debug ADF pipeline?

Debugging is one of the crucial aspects of any coding-related activity needed to test the code for any issues it might have. It also provides an option to debug the pipeline without executing it.

31. What does it mean by breakpoint in ADF pipeline?

To understand better, for example, you are using three activities in the pipeline, and now you want to debug up to the second activity only. You can do this by placing the breakpoint at the second activity. To add a breakpoint, you can click the circle present at the top of the activity.

32. What is the use of ADF service?

ADF is primarily used to organize the data copying between various relational and non-relational data sources hosted locally in data centers or the cloud. Moreover, you can use ADF Service to transform the ingested data to fulfill business requirements. In most Big Data solutions, ADF Service is used as an ETL or ELT tool for data ingestion.

33. Explain the data source in azure data factory?

The data source is the source or destination system that comprises the data intended to be utilized or executed.

34. Can you share any difficulties you faced while getting data from on premises to azure data factory?

One of the significant challenges we face while migrating from on-prem to cloud is throughput and speed. When we try to copy the data using Copy activity from on-prem, the speed of the process is relatively slow, and hence we don't get the desired throughput.

35. How to copy multiple sheet data from an excel file?

When we use an excel connector within a data factory, we must provide a sheet name from which we have to load data. This approach is nuanced when we have to deal with a single or a handful of sheets' data, but when we have lots of sheets (say 10+), this may become a tedious task as we have to change the hard-coded sheet name every time!

36. Is it possible to have nested looping in data factory?

There is no direct support for nested looping in the data factory for any looping activity (for each / until). However, we can use one for each/until loop activity which will contain an execute pipeline activity that can have a loop activity.

37. How to copy multiple tables from one datastore to another data store?

An efficient approach to complete this task would be:

i) Maintain a lookup table/ file which will contain the list of tables and their source, which needs to be copied.
ii) ii) Then, we can use the lookup activity and each loop activity to scan through the list.
iii) iii) Inside the for each loop activity, we can use a copy activity or a mapping dataflow to accomplish the task of copying multiple tables to the destination datastore.

38. What are some limitations of ADF?

Azure Data Factory provides great functionalities for data movement and transformations. However, there are some limitations as well.

i) We can't have nested looping activities in the data factory, and we must use some workaround if we have that sort of structure in our pipeline. All the looping activities come under this: If, Foreach, switch, and until activities.
ii) ii) The lookup activity can retrieve only 5000 rows at a time and not more than that. Again, we need to use some other loop activity along with SQL with the limit to achieve this sort of structure in the pipeline.
iii) iii) We can have a maximum of 40 activities in a single pipeline, including everything: inner activity, containers, etc. To overcome this, we should try to modularize the pipelines regarding the number of datasets, activities, etc.

39. how do you send email notifications on pipeline failure?

There are multiple ways to do this:

1. Using Logic Apps with Web/Web hook activity.
2. Using Alerts and Metrics from pipeline options.

40. What is azure SQL database? How to integrate with data factory?

Part of the Azure SQL family, Azure SQL Database is an always up-to-date, fully managed relational database service built for the cloud for storing data. We can easily design data pipelines to read and write to SQL DB using the Azure data factory.

41. Can you host sql server instance on azure?

Azure SQL Managed Instance is the intelligent, scalable cloud database service that combines the broadest SQL Server instance.

42. What is azure data lake analytics?

Azure Data Lake Analytics is an on-demand analytics job service that simplifies storing data and processing big data.

Azure Databricks

1. How does pyspark data frames work?

The distributed collection of structured data is called a PySpark DataFrame. They are stored in named columns and are equivalent to relational database tables. Additionally, PySpark DataFrames are more effectively optimized than Python or R code. Various sources, including Structured Data Files, Hive Tables, external databases, existing RDDs, etc., can be used to create them.

2. Define PySpark partition? What's the maximum number of partitions pyspark allows?

The PySpark Partition method divides a large dataset into smaller datasets using one or more partition keys. When transformations on partitioned data run more quickly, execution performance is improved. This is due to the concurrent operation of each partition's transformations. PySpark supports two partitioning methods: partitioning in memory (DataFrame) and partitioning on a disc (File system). partitionBy (self, *cols) is the syntax for it. Adding 4x as many partitions are accessible to the cluster application core count is advisable.

3. How do you import data into delta lake?

Loading data into Delta Lake is quite simple. Using Databricks Auto Loader or the COPY INTO command with SQL, you can automatically intake new data files into Delta Lake as they arrive in your data lake (for example, on S3 or ADLS). Additionally, you can batch-read your data using Apache SparkTM, make any necessary changes, and then store the outcome in Delta Lake format.

4. Does delta lake offers access controls for security and governance?

You can leverage access control lists (ACLs) to set permissions for workspace objects (folders, notebooks, experiments, and models), clusters, pools, tasks, data schemas, tables, views, etc., using Delta Lake on Databricks. Both administrators and users with delegated access control list management rights have this access.

5. What does an index mean in the context of delta lake?

An index is a data structure that supports Delta Lake in improving query performance by enabling fast data lookups. It is possible to speed up queries that filter on a particular column or columns by using an index on a Delta table.

6. Is PySpark DataFrames' implementation entirely different from other Python DataFrames like Pandas, or are there some similarities?

While Pandas are the source of inspiration for Spark DataFrames and behave similarly, they are not the same. You should use DataFrames in Spark rather than Pandas. To lessen the performance impact of switching between the two frameworks, users of Pandas and Spark DataFrames should consider implementing Apache Arrow.

7. Is it possible to utilize multiple languages in one notebook, or are there substantial restrictions? Is it usable in later stages if you build a DataFrame in your Python notebook using a % Scala magic?

A Scala DataFrame can be produced and then used as a Python reference. Write your program in Scala or Python if possible.

8. Which programming languages can be used to integrate with azure databricks?

Python, Scala, and R are a few examples of languages that you can use with the Apache Spark framework. Azure Databricks also supports SQL as a database language.

9. Suppose you have just begun your job at XYZ Company. Your manager has instructed you to develop business analytics logic in the Azure notebook leveraging some of the general functionality code written by other team members. How would you proceed?

You must import the databricks code into your notebook to reuse it. There are two additional options to import the code. If the code is present in the same workspace, you can import it immediately. If the code is outside the workspace, you should build a jar or module out of it and import it into the databricks cluster.

10. How do you handle the Databricks code while using TFS or Git in a collaborative environment?

The initial problem is the lack of support for Team Foundation Server (TFS). You can only use Git or a repository system built on the distributed format of Git. While it would be ideal to link Databricks to your Git directory of notebooks, you can still consider Databricks as a replica of your project, even though this is not presently feasible. Your first step is to start a notebook, which you then update before submitting to version control.

11. 1

If you type "%sql" at the beginning of any block in the Databricks notebook, the block will be converted from a Python/Scala notebook to a Simple SQL Notebook.

12. What is method for creating a databricks private access token?

Go to the "user profile" icon and choose "User setting" to create a private access token. You must choose the "Access Tokens" tab to access the "Generate New Token" button. To create the token, press the right button.

13. Define a databricks secret?

A secret is a key-value pair that can be used to safeguard confidential data. it consists of a special key name enclosed in a secure environment. There are only 1000 secrets allowed per scope. There is a 128 KB maximum size limit.

14. What possible challenges might you come across using databricks?

If you don't have enough credits to construct more clusters, you might experience cluster creation failures. If your code is incompatible with the Databricks runtime, Spark errors will occur. Network issues may occur if your network is not set up correctly or if you try to access Databricks from an unsupported location.

Azure Synapse Analytics

1. How does azure synapse differ from databricks?

Azure Synapse is a data integration service with some amazing transformation capabilities while Azure Databricks is data analytics focussed platform build on top of Spark. Azure Synapse integrates big data analytics and enterprise data warehouse into a single platform. While Databricks allows customers to develop complex machine learning algorithms and perform big data analytics. However, both both Synapse Analytics and Azure Databricks can be used together when building a data pipeline.

2. Define a linked service in azure synapse analytics?

The external sources outside the Azure Synapse Analytics workspace are connected via Linked Services. A linked service is required in a data pipeline to read or write data to source or destination. For instance, you can use it to establish a connection to the Azure Data Lake storage account to execute SQL queries on the files. Suppose you want to pull data from an on-premises server into the cloud or connect to the Azure data lake storage account to perform SQL queries on the files. In these cases, you must build a linked service by providing data such as a username, password, and server address to establish a connection.

3. What do you understand by sql pool in azure synapse analytics?

The default SQL pool in Azure Synapse Analytics helps with query optimization, data distribution, and data compression.

4. What does the azure synapse analytics OPENROWSET function do?

The OPENROWSET function lets data engineers read data from diverse data sources including flat files, RDBM's, and other OLE DB sources. In Azure Synapse, one uses the OPENROWSET function to read the file as a table. For instance, you want to perform the queries on a file saved in the ADLS account. Using the rowset function, which converts each row of a file into a row of a table, you can read this file as a table.

5. What makes synapse analytics different from azure blob storage?

Large volumes of unstructured data, such as video, audio, and images, can be stored in the cloud using Azure Blob storage. You can gather, store, and analyze huge amounts of data with the help of Azure Synapse Analytics, a cloud-based data warehouse service.In a nutshell, Azure Blob Storage majorly focusses on storing and retrieving data while Azure Synapse Analytics focusses on preparing, managing, and analyzing data.

6. Assume you are the lead data engineer at your company, and a few spark notebooks have been run in the past few hours. You want to examine each notebook execution. Where will you find the past hour few hours' histories and retrieve the event log for them in Azure Synapse Analytics?

You must access the monitor tab in Azure Synapse Analytics. You can find the activities area in the monitor tab on the left. The Apache Spark application has a separate tab below it. The Spark history server will appear on the following page when you click on it. Simply click on that to view the whole history and to download the event log for a particular application run.

7. Suppose you are a data engineer for the XYZ company. The company wants to transfer its data from an on-site server to the Azure cloud. How will you achieve this using Azure synapse Analytics?

You must develop an integration runtime to transfer data from an on-premises server to a cloud server. This integration runtime will act as the self-hosted IR since the on-premise servers cannot be connected using the auto-resolve integration runtime, After creating the self-hosted IR, you can use the copy function to build a pipeline that will transfer data from the on-premise server to the cloud server.

8. Suppose you are the lead data engineer at XYZ company, which is transitioning from on-premise to cloud. On one of their on-site servers, your organization holds some mission-critical data. As a result, you must ensure that the team's self-hosted IR is only accessible to specified team members. What is the best way to meet this requirement using Azure Synapse Analytics?

Open the Azure synapse analytics workspace studio. Navigate to the access control section under the monitor tab. You can grant access at the workspace item level there. Choose the IR item and the IR's name, role, and member data.

9. Assume that a different application will access the data created by your stored procedure in the SQL pool database. Daily updates of this data are necessary. What approach would you apply to address this issue?

To begin, open the Azure Synapse Analytics workspace and create the pipeline in the integration tab of the resource. The activity for the SQL pool stored procedure must be added to this pipeline. Choose the specified stored procedure from the SQL pool stored procedure activity. Schedule this pipeline by including the trigger that will run it daily.

10. Let's say you want to know how many SQL requests your team has already executed in the SQL pool. Where can you see a list of every historical query run in the Azure Synapse Analytics workspace?

Any time you perform a query in Azure Synapse Analytics, a history log is generated. You must select the manage tab to see the SQL request. You can sort by time to see the executions and their respective results.

11. What does synapse analytics control node do?

The primary component of a Synapse SQL architecture is the control node. A distributed query engine is executed on the Control node to optimize and coordinate parallel queries. The Control node converts a T-SQL query into parallel queries executed against each distribution when submitted to a specific SQL pool. The DQP engine, used in serverless SQL pools, runs on the Control node and divides user queries into smaller ones that Compute nodes will process to optimize and coordinate distributed execution.

12. What is the method to create pipeline in azure synapse analytics?

You can create a pipeline in Azure Synapse with the following steps:

- Open the Integrate hub in Synapse Studio.
- Click on 'Pipeline' to build a new pipeline. Open the Pipeline designer by clicking on the newly created pipeline object.
- You can start dragging the activities from the ‘Activities’ panel on the left into the blank space in the middle.
- You must publish your pipeline once it has been created. After that, you can test it to ensure there are no errors and verify that it works.

13. In azure synapse analytics, how much data can you keep in a single column simultaneously?

The maximum size of a single column in Azure Synapse Analytics depends on the storage technology used and also on the datatype of the column. You can store up to 1 billion 2-byte Unicode characters using nvarchar [(n | max)].

14. Briefly discuss about setting spark job in azure synapse analytics?

i) Create a Spark Pool
ii) Prepare the Data
iii) Write the Spark Job
iv) Submit the Spark Job
v) Monitor the Spark Job using Synapse Analytics Studio
vi) Schedule the Job

15. Mention some of data security features offered by synapse analytics?

Synapse Analytics provides certain data security features for dedicated SQL pools, including Data Discovery & Classification, Dynamic Data Masking, Vulnerability Assessment, Advanced Threat Protection, and Transparent Data Encryption.Synapse Analytics supports encryption of data both

in transit and at rest. Data can be encrypted using Azure Key Vault, which provides a secure way to store and manage encryption keys.

ETL Pipeline questions

1. How do you analyze tables in ETL?

You can validate the structures of system objects using the ANALYZE statement. This statement generates statistics, which are then used by a cost-based optimizer to determine the most effective strategy for data retrieval. ESTIMATE, DELETE, and COMPUTER are some additional operations.

2. What SQL commands allow you to validate data completion?

You can validate the completeness of the data using the intersect and minus statements. When you run source minus target and target minus source, the minus query returns a value indicating rows that don't match. A duplicate row is present when the count intersects is less than the source count, and the minus query returns the value.

3. What roles does impact analysis play in ETL system?

Impact analysis analyzes the metadata relating to an object to determine what is impacted by a change in its structure or content. A data warehouse's proper loading can be affected by changing data-staging objects in processes. You must conduct an impact analysis before modifying a table once it has been created in the staging area.

4. What are the various Phases of data mining?

A phase in data mining is a logical procedure for sifting through vast amounts of data to identify crucial data.

- Exploration: The exploration phase aims to identify significant variables and determine their characteristics.
- Pattern Identification: In this phase, the main task is looking for patterns and selecting the best prediction.
- Deployment stage: This stage can only be attained once a reliable, highly predictive pattern is identified in stage 2.

5. Explain the Use of ETL in data migration projects?

The usage of ETL tools is popular in data migration projects. For instance, if a company previously managed its data in Oracle 10g and now wants to switch to a SQL Server cloud database, it will need to be moved from Source to Target. ETL tools are quite beneficial when performing this kind of data migration. ETL code writing will take a lot of the user's time. As a result, the ETL tools are helpful because they make coding easier than P-SQL or T-SQL. Therefore, ETL is a beneficial process for projects involving data migration.

6. What performs better, joining data first, then filtering it, or filtering data first, then joining it with other resources?

Filtering data and then joining it with other data sources is better.

Filtering unnecessary data as early as possible in the process is an excellent technique to enhance the efficiency of the ETL process. It cuts down on time necessary for data transfer, I/O, and/or memory processing.

The general idea is to minimize the number of processed rows and avoid altering data that is never utilized.

7. Explain how ETL and OLAP tools differ?

ETL tools: ETL tools allow you to extract, transform, and load the data in the data warehouse or data mart.

OLAP (Online Analytical Processing) tools: OLAP tools help generate reports from data marts and warehouses for business analysis.

8. Briefly explain ETL mapping sheets?

ETL mapping sheets usually offer complete details about a source and a destination table, including each column and how to look them up in reference tables.ETL testers may have to generate big queries with several joins at any stage of the testing process to check data. When using ETL mapping sheets, writing data verification queries is much easier.

9. Find every user who was active for three or days in a row?

(DataFrame: sf_events)

sf_events

date: datetime64[ns]

account_id: object

user_id: object

```
import pandas as pd

df = sf_events.drop_duplicates()

df = df[['user_id', 'date']].sort_values(['user_id', 'date'])

df['3_days'] = df['date'] + pd.DateOffset(days=2)

df['shift_3'] = df.groupby('user_id')['date'].shift(-2)

df[df['shift_3'] == df['3_days']]['user_id']
```

10. How would you modify a big table with more than 10 million rows?

Using batches is the most typical approach: Divide a large query into smaller ones, for example, ten thousand rows in one batch.

```
DECLARE @id_control INT = 0 --current batch

    ,@batchSize INT = 10000 --size of the batch

    ,@results INT = 1 --row count after batch

-- if 0 rows are returned, exit the loop

WHILE (@results > 0)

BEGIN

  UPDATE [table]

  SET [column] = [value]

  WHERE [PrimaryKey column] > @id_control

   AND [PrimaryKey column] <= @id_control + @batchSize

  -- the latest row count

  SET @results = @@ROWCOUNT

  -- start the next batch

  SET @id_control = @id_control + @batchSize

END
```

If the table is too big, it might be preferable to make a new one, add the updated data, and then switch tables.