Data Engineer Interview Questions

To view the live version of the page, click here.

Contents

Data Engineer Interview Questions for Freshers

Data Engineer Interview Questions for Freshers

(.....Continued)

21. What are the four Vs of Big Data?
22. Explain the Star Schema in Brief.
23. Explain the Snowflake Schema in Brief.
24. Name the XML configuration files present in Hadoop.
25. What is Hadoop Streaming?
26. What is the Replication factor?
27. What is the difference between HDFS block and InputSplit?
28. What is Apache Spark?
29. What is the difference between Spark and MapReduce?

Data Engineer Interview Questions for Experienced

30. What are Skewed tables in Hive?
31. What is SerDe in the hive?
32. What are the table creation functions in Hive?
33. What are *args and **kwargs used for?
34. What do you mean by spark execution plan?
35. What is executor memory in spark?
36. Explain how columnar storage increases query speed.
37. What is schema evolution?
38. What do you mean by data pipeline?

Data Engineer Interview Questions for Experienced

(.....Continued)

39. What is orchestration?
40. What are different data validation approaches?
41. What was the algorithm you used in a recent project?
42. Have you earned any certification related to this field?
43. Why are you applying for the Data Engineer role in our company?
44. What tools did you use in your recent projects?
45. What challenges did you face in your recent project and how did you overcome them?
46. Which Python libraries would you recommend for effective data processing?
47. How do you handle duplicate data points in a SQL query?
48. Have you ever worked with big data in a cloud computing environment?

Frequently Asked Questions

49. What are the roles and responsibilities of a data engineer?
50. How to become a Data Engineer?
51. Is Data Engineering a good career?
52. Are data engineers paid well?
53. What do Data Engineering interns do?

Let's get Started

The practice of developing and constructing large-scale data collection, storage, and analysis systems is known as **data engineering**. It is a vast field that has applications in almost every industry. It is a multidisciplinary subject that involves defining the data pipeline alongside **data scientists**, **data analysts**, and **software engineers**. Data engineers create systems that collect, process and turn raw data into information that data scientists and business analysts can understand. The future of data engineers looks prominent, given the ever-increasing reliance on massive amounts of data. Companies use the data collected to leverage their business, which means that there will always be a demand for skilled data engineers. Finding the appropriate person for data engineering roles is extremely challenging, and competition for that position can be fierce.

A considerable part of the questions you will be asked during an interview will be aimed at testing your understanding of how these critical systems operate and how you would respond to restraints and faults in their design and implementation. You can try to prepare for these types of questions by understanding quantitative and analytical approaches to data collection, preparation, and analysis, as well as some fundamental computer science principles. Domain knowledge is especially beneficial if you can discuss related projects or applications in your industry.

As much as we can do to assist you, we have compiled a list of 35+ **data engineering interview questions and answers** for your convenience. The questions have been chosen in such a way that they are suited for both freshers and experienced Data Engineers.

Data Engineer Interview Questions for Freshers

1. What is Data Engineering?

The application of data collecting and analysis is the emphasis of **data engineering**. The information gathered from numerous sources is merely raw information. Data engineering helps in the transformation of unusable data into useful information. It is the process of transforming, cleansing, profiling, and aggregating huge data sets in a nutshell.

2. What is Data Modeling?

Data Modeling is the act of creating a visual representation of an entire information system or parts of it in order to express linkages between data points and structures. The purpose is to show the many types of data that are used and stored in the system, as well as the relationships between them, how the data can be classified and arranged, and its formats and features. Data can be modeled according to the needs and requirements at various degrees of abstraction. The process begins with stakeholders and end-users providing information about business requirements. These business rules are then converted into data structures, which are used to create a concrete database design.

3. What are the design schemas available in data modeling?

There are two design schemas available in data modeling:

- Star Schema
- Snowflake Schema

4. What is the difference between a data engineer and a data scientist?

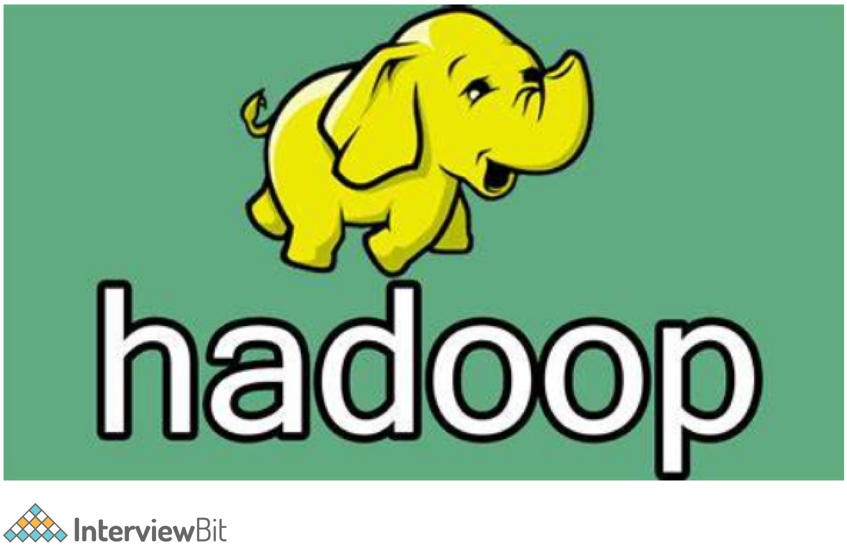
- Data science is a broad topic of research. It focuses on extracting data from extremely huge datasets (sometimes it is known as "big data"). Data scientists can operate in a variety of fields, including industry, government, and applied sciences. All data scientists have the same goal: to analyze data and derive insights from it that are relevant to their field of work.
- A data engineer's job is to develop or integrate many components of complex systems, taking into account the information needed, the company's goals, and the end requirements. This necessitates the creation of extremely complicated data pipelines. These data pipelines, like oil pipelines, take raw, unstructured data from a variety of sources. They then channel them into a single database (or larger structure) for storage.

5. What are the differences between structured and unstructured data?

On the basis of	Structured	Unstructured
Storage	Structured data is stored in DBMS.	It is stored in unmanaged file structures.
Flexibility	It is less flexible as it is dependent on the schema.	It is more flexible.
Scalability	Not easy to scale.	Easy to scale.
Performance	Since we can perform a structured query, the performance is high.	The performance of unstructured data is low.
Analysis factor	Easy to analyze.	Hard to analyze.

6. What are the features of Hadoop?

Hadoop has the following features:



- It is open-source and easy to use.
- Hadoop is extremely scalable. A significant volume of data is split across several devices in a cluster and processed in parallel. According to the needs of the hour, the number of these devices or nodes can be increased or decreased.
- Data in Hadoop is copied across multiple DataNodes in a Hadoop cluster, ensuring data availability even if one of your systems fails.
- Hadoop is built in such a way that it can efficiently handle any type of dataset, including structured (MySQL Data), semi-structured (XML, JSON), and unstructured (Images and Videos). This means it can analyze any type of data regardless of its form, making it extremely flexible.
- Hadoop provides faster data processing. **More Features**.

7. Which frameworks and applications are important for data engineers?

SQL, Amazon Web Services, Hadoop, and Python are all required skills for data engineers. Other tools critical for data engineers are PostgreSQL, MongoDB, Apache Spark, Apache Kafka, Amazon Redshift, Snowflake, and Amazon Athena.

8. What is HDFS?

HDFS is an acronym for Hadoop Distributed File System. It is a distributed file system that runs on commodity hardware and can handle massive data collections.

9. What is a NameNode?

The HDFS system is built on the foundation of NameNode. It keeps track of where the data file is kept by storing the directory tree of the files in a single file system.

10. What are the repercussions of the NameNode crash?

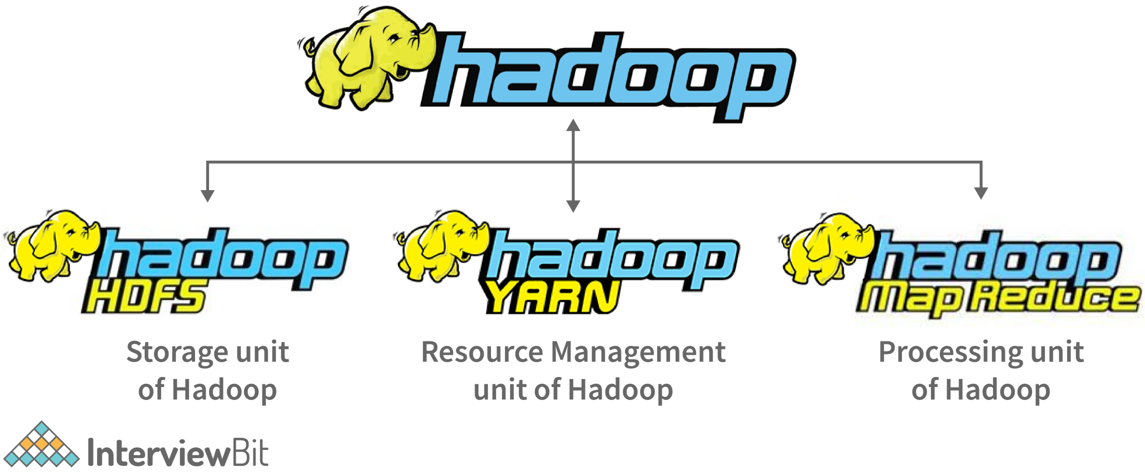
In an HDFS cluster, there is only one NameNode. This node keeps track of DataNode metadata. Because there is only one NameNode in an HDFS cluster, it is the single point of failure. The system may become inaccessible if NameNode crashes. In a high-availability system, a passive NameNode backs up the primary one and takes over if the primary one fails.

11. What is a block and block scanner in HDFS?

- **Block**: In HDFS, a "block" refers to the smallest amount of data that may be read or written.
- **Block Scanner**: Block Scanner keeps track of the list of blocks on a DataNode and checks them for checksum problems. To save disc bandwidth on the data node, Block Scanners use a throttling technique.

12. What are the components of Hadoop?

Hadoop has the following components:

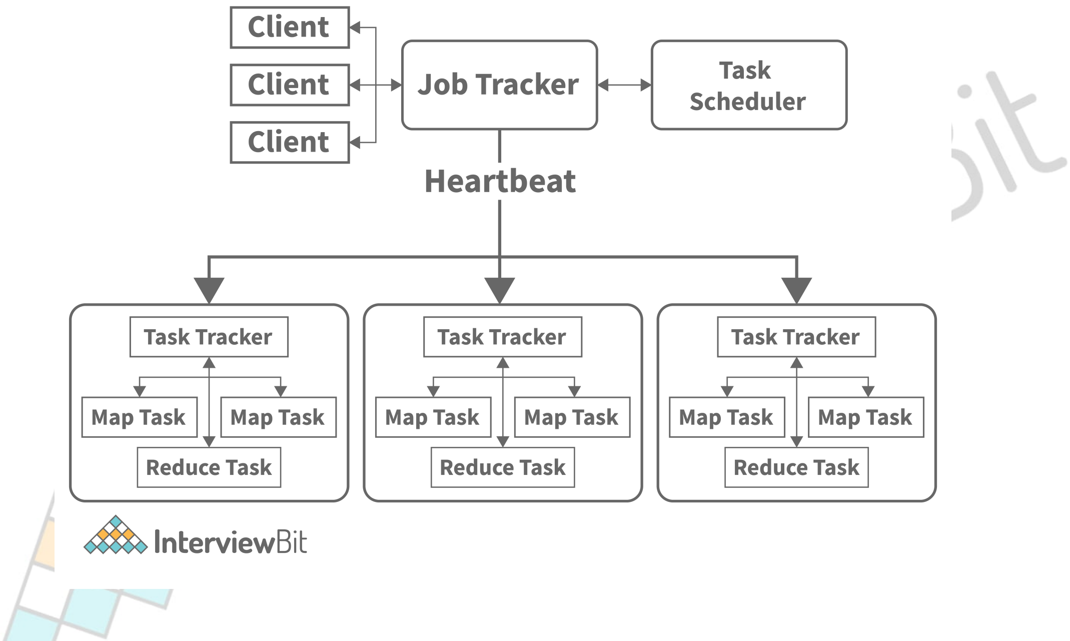


- **Hadoop Common:** A collection of Hadoop tools and libraries.
- **Hadoop HDFS:** Hadoop's storage unit is the Hadoop Distributed File System (HDFS). HDFS stores data in a distributed fashion. HDFS is made up of two parts: a name node and a data node. While there is only one name node, numerous data nodes are possible.
- **Hadoop MapReduce:** Hadoop's processing unit is MapReduce. The processing is done on the slave nodes in the MapReduce technique, and the final result is delivered to the master node.
- **Hadoop YARN:** Hadoop's YARN is an acronym for Yet Another Resource Negotiator. It is Hadoop's resource management unit, and it is included in Hadoop version 2 as a component. It's in charge of managing cluster resources to avoid overloading a single machine.

13. Explain MapReduce in Hadoop.

MapReduce is a programming model and software framework for processing large volumes of data. Map and Reduce are the two phases of MapReduce. The map turns a set of data into another set of data by breaking down individual elements into tuples (key/value pairs). Second, there's the reduction job, which takes the result of a map as an input and condenses the data tuples into a smaller set. The reduction work is always executed after the map job, as the name MapReduce suggests.

14. What is the Heartbeat in Hadoop?



The heartbeat is a communication link that runs between the Namenode and the Datanode. It's the signal that the Datanode sends to the Namenode at regular intervals. If a Datanode in HDFS fails to send a heartbeat to Namenode after 10 minutes, Namenode assumes the Datanode is unavailable.

15. How does the NameNode communicate with the DataNode?

The NameNode and the DataNode communicate via these messages:

- Block reports
- Heartbeats

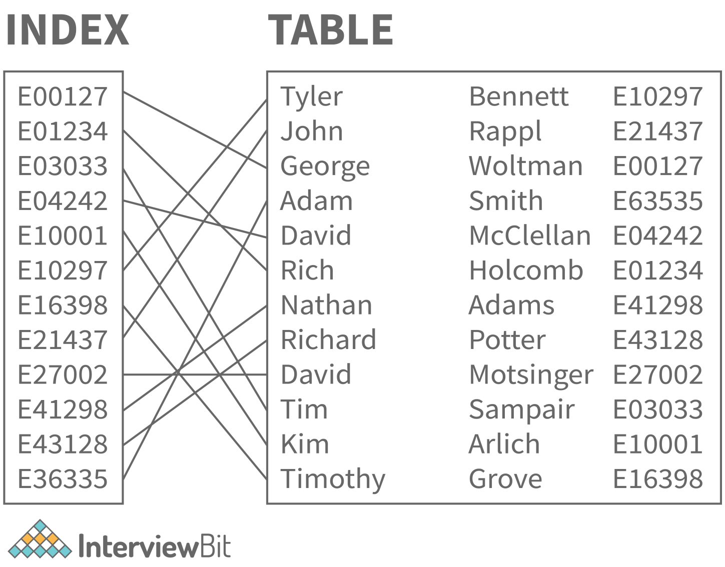
16. What happens when the block scanner detects a corrupt data block?

The following steps occur when the block scanner detects a corrupt data block:

- First and foremost, when the Block Scanner detects a corrupted data block, DataNode notifies NameNode.
- NameNode begins the process of constructing a new replica from a corrupted block replica.
- The replication factor is compared to the replication count of the right replicas. The faulty data block will not be removed if a match is detected.

17. Explain indexing.

Indexing is a technique for improving database performance by reducing the number of disc accesses necessary when a query is run. It's a data structure strategy for finding and accessing data in a database rapidly.



Tyler	Bennett	E10297
John	Rappl	E21437
George	Woltman	E00127
Adam	Smith	E63535
David	McClellan	E04242
Rich	Holcomb	E01234
Nathan	Adams	E41298
Richard	Potter	E43128
David	Motsinger	E27002
Tim	Sampair	E03033
Kim	Arlich	E10001
Timothy	Grove	E16398

18. Explain the main methods of reducer.

These are the main methods of reducer:

- **setup()**: This command is used to specify parameters such as the size of input data and the distributed cache.
- **cleaning()**: is a function for deleting temporary files.
- **reduce()**: it's called once per key with the corresponding reduced task.

19. What is COSHH?

Classification and Optimization-based Scheduling for Heterogeneous Hadoop Systems (COSHH), as the name implies, enables scheduling at both the cluster and application levels to have a direct positive impact on task completion time.

20. What is the relevance of Apache Hadoop's Distributed Cache?

Hadoop Distributed Cache is a Hadoop MapReduce Framework technique that provides a service for copying read-only files, archives, or jar files to worker nodes before any job tasks are executed on that node. To minimize network bandwidth, files are usually copied only once per job. Distributed Cache is a program that distributes read-only data/text files, archives, jars, and other files.

21. What are the four Vs of Big Data?

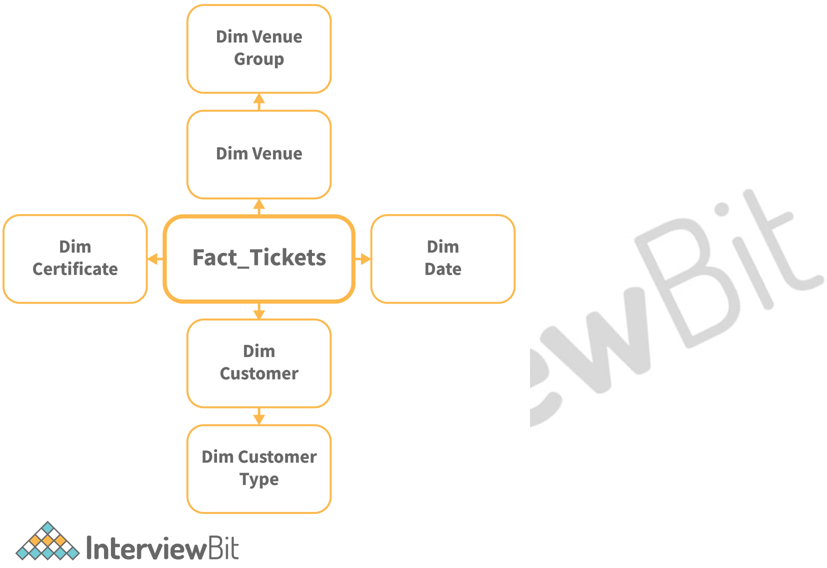
The four characteristics or four Vs of Big data are:

- Volume
- Veracity
- Velocity
- Variety

22. Explain the Star Schema in Brief.

In a data warehouse, a star schema can include one fact table and a number of associated dimension tables in the center. It's called a star schema because its structure resembles that of a star. The simplest sort of Data Warehouse schema is the Star Schema data model. It is also known as the Star Join Schema, and it is designed for massive data sets.

23. Explain the Snowflake Schema in Brief.



A snowflake schema is a logical arrangement of tables in a multidimensional database that matches the snowflake shape (in the ER diagram). A Snowflake Schema is an enlarged Star Schema with additional dimensions. After the dimension tables have been normalized, the data is separated into new tables.

Snowflaking has the potential to improve the performance of certain queries. The schema is organized so that each fact is surrounded by its related dimensions, and those dimensions are linked to other dimensions, forming a snowflake pattern.

24. Name the XML configuration files present in Hadoop.

XML configuration files available in Hadoop are:

- Core-site
- Mapred-site
- Yarn-site
- HDFS-site

25. What is Hadoop Streaming?

It is a utility or feature included with a Hadoop distribution that allows developers or programmers to construct Map-Reduce programs in many programming languages such as Python, C++, Ruby, Pearl, and others. We can use any language that can read from standard input (STDIN), such as keyboard input, and write using standard output (STDOUT).

26. What is the Replication factor?

The replication factor is the number of times the Hadoop framework replicates each Data Block. Fault tolerance is provided by replicating the block. The replication factor is set to 3 by default, however, it can be modified to 2 (less than 3) or raised to meet your needs (more than 3.)

27. What is the difference between HDFS block and InputSplit?

Block	InputSplit
In Hadoop, a block is the physical representation of data.	InputSplit is the logical representation of data in a block. It is primarily used in the MapReduce program or other data processing techniques.
The HDFS block size is set to 128MB by default, but you can modify it to suit your needs. Except for the last block, which can be the same size or less, all HDFS blocks are the same size.	By default, the InputSplit size is nearly equal to the block size.

28. What is Apache Spark?

Apache Spark is an open-source distributed processing solution for big data workloads. For rapid queries against any size of data, it uses in-memory caching and efficient query execution. Simply put, Spark is a general-purpose data processing engine that is quick and scalable.

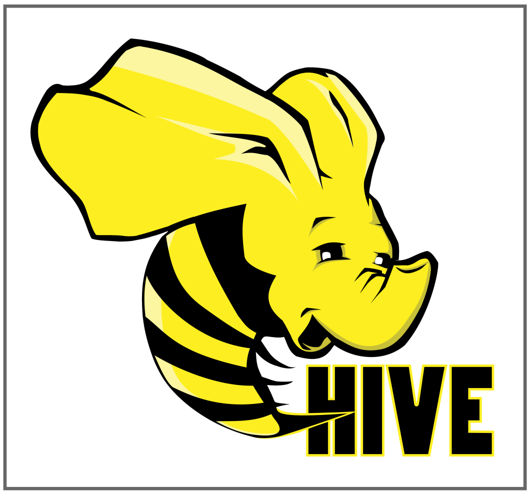
29. What is the difference between Spark and MapReduce?

Spark is a MapReduce improvement in Hadoop. The difference between Spark and MapReduce is that Spark processes and retains data in memory for later steps, whereas MapReduce processes data on the disc. As a result, Spark's data processing speed is up to 100 times quicker than MapReduce for lesser workloads. Spark also constructs a Directed Acyclic Graph (DAG) to schedule tasks and orchestrate nodes throughout the Hadoop cluster, as opposed to MapReduce's two-stage execution procedure.

Data Engineer Interview Questions for Experienced

30. What are Skewed tables in Hive?

Skewed tables are a type of table in which some values in a column appear more frequently than others. The distribution is skewed as a result of this. When a table is created in Hive with the SKEWED option, the skewed values are written to separate files, while the remaining data are written to another file.



31. What is SerDe in the hive?

Serializer/Deserializer is popularly known as SerDe. For IO, Hive employs the SerDe protocol. Serialization and deserialization are handled by the interface, which also interprets serialization results as separate fields for processing.

The Deserializer turns a record into a Hive-compatible Java object. The Serializer now turns this Java object into an HDFS-compatible format. The storage role is then taken over by HDFS. Anyone can create their own SerDe for their own data format.

32. What are the table creation functions in Hive?

The following are some of Hive's table creation functions:

- Explode(array)
- Explode(map)
- JSON_tuple()
- Stack()

33. What are *args and **kwargs used for?

The *args function allows users to specify an ordered function for use in the command line, whereas the **kwargs function is used to express a group of unordered and in-line arguments to be passed to a function.

34. What do you mean by spark execution plan?

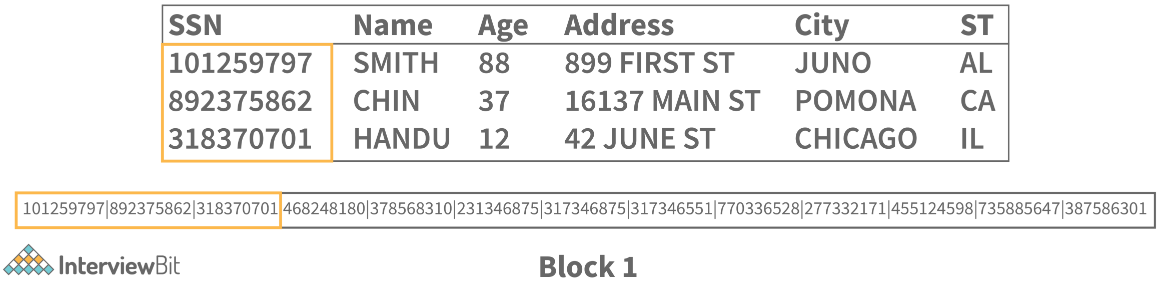
A query language statement (SQL, Spark SQL, Dataframe operations, etc.) is translated into a set of optimized logical and physical operations by an execution plan. It is a series of actions that will be carried out from the SQL (or Spark SQL) statement to the DAG(Directed Acyclic Graph), which will then be sent to Spark Executors.

35. What is executor memory in spark?

For a spark executor, every spark application has the same fixed heap size and fixed number of cores. The heap size is regulated by the spark.executor.memory attribute of the –executor-memory flag, which is also known as the Spark executor memory. Each worker node will have one executor for each Spark application. The executor memory is a measure of how much memory the application will use from the worker node.

36. Explain how columnar storage increases query speed.

Since it dramatically reduces total disc I/O requirements and the quantity of data you need to load from the disc, columnar storage for database tables is a critical factor in increasing analytic query speed. Each data block stores values of a single column in multiple rows using columnar storage.



SSN	Name	Age	Address	City	ST
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL

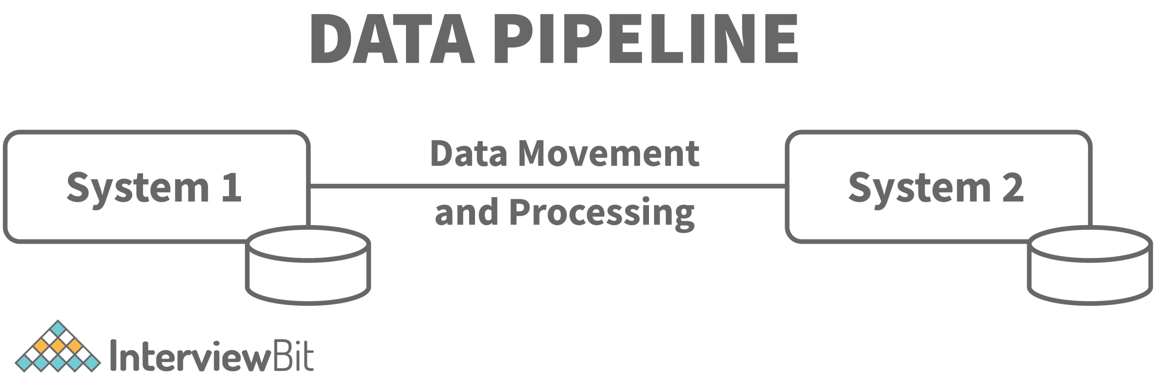
101259797|892375862|318370701|468248180|378568310|231346875|317346875|317346551|770336528|277332171|455124598|735885647|387586301

37. What is schema evolution?

One set of data can be kept in several files with various yet compatible schemas with schema evolution. The Parquet data source in Spark can automatically recognize and merge the schema of those files. Without automatic schema merging, the most common method of dealing with schema evolution is to reload past data, which is time-consuming.

38. What do you mean by data pipeline?

A data pipeline is a system for transporting data from one location (the source) to another (the destination) (such as a data warehouse). Data is converted and optimized along the journey, and it eventually reaches a state that can be evaluated and used to produce business insights. The procedures involved in aggregating, organizing, and transporting data are referred to as a data pipeline. Many of the manual tasks needed in processing and improving continuous data loads are automated by modern data pipelines.



39. What is orchestration?

IT departments must maintain many servers and apps, but doing it manually isn't scalable. The more complicated an IT system is, the more difficult it is to keep track of all the moving elements. As the requirement to combine numerous automated jobs and their configurations across groups of systems or machines grows, so does the demand to combine multiple automated tasks and their configurations across groups of systems or machines. This is where orchestration comes in handy.

The automated configuration, management, and coordination of computer systems, applications, and services are known as orchestration. IT can manage complicated processes and workflows more easily with orchestration. There are many container orchestration platforms available such as Kubernetes and OpenShift.

40. What are different data validation approaches?

The process of confirming the accuracy and quality of data is known as data validation. It is implemented by incorporating various checks into a system or report to ensure that input and stored data are logically consistent. Common types of data validation approaches are

- **Data type check:** It confirms that the data entered is of the correct data type.
- **Code check**: A code check verifies that a field is chosen from a legitimate list of options or that it corresponds to specific formatting constraints. Checking a postal code against a list of valid codes, for example, makes it easier to verify if it is valid.
- **Range check**: It ensures that input falls in a predefined range.
- **Format check**: Many data types follow a predefined format. Format check confirms that. For example, a date has formats like DD-MM-YY or MM-DD-YY.
- **Consistency check:** It confirms that the data entered is logically correct.
- **Uniqueness check:** It ensures that the same data is not entered multiple times.

41. What was the algorithm you used in a recent project?

First, decide which project you'd want to talk about. If you have a real-world example in your field of expertise and an algorithm relevant to the company's work, utilize it to capture the hiring manager's attention. Maintain a list of all the models and analyses you deployed. Begin with simple models and avoid overcomplicating things. The hiring supervisors want you to describe the outcomes and their significance. There could be follow-up questions like:

- Why did you choose this algorithm?
- What is the scalability of your model?
- If you were given more time, what could you improve?

42. Have you earned any certification related to this field?

The interviewer wants to how much you have invested in this field and whether you are an interested candidate. Mention all your certifications related to the field in chronological order and briefly explained what you learned to earn that certificate.

43. Why are you applying for the Data Engineer role in our company?

You must expect this question. The interviewer wants to know how much you have researched before applying to this role. While answering this question, keep your explanation concise on how you would create a plan that works with the company set-up and how you would implement the plan, ensuring that it works by first understanding the company's data infrastructure setup. Reading job descriptions and researching the company will help you to tackle the question easily.

44. What tools did you use in your recent projects?

Interviewers seek to analyze your decision-making abilities as well as your understanding of various tools. As a result, utilize this question to describe why you chose certain tools over others. Tell the interviewer about the tools you used and why you used them. You can also mention the features and drawbacks of the tool you used. Also, try to use this opportunity to tell the interviewer how you can use the tool for the company’s benefit.

45. What challenges did you face in your recent project and how did you overcome them?

With this question, the panel generally wants to know your problem-solving ability and how well you perform under pressure. To answer the question, first, brief them about the situations that lead to the problem. You should tell them about your role in that situation. For example, if you played a leading role in solving that problem, that would tell the interviewer about competency as a leader. After that tell them about the action you took to solve the problem. To end the answer on a positive note, you should tell them about the consequences of the challenge and the learning you took out of it.

46. Which Python libraries would you recommend for effective data processing?

This question allows the hiring manager to determine whether the candidate understands the fundamentals of Python, which is the most commonly used language among data engineers. NumPy, which is used for efficient processing of arrays of numbers, and pandas, which is useful for statistics and data preparation for machine learning work, should be included in your solution.

47. How do you handle duplicate data points in a SQL query?

This is a question that interviewers may ask to test your SQL expertise. To reduce duplicate data points, you can advise using the SQL keywords DISTINCT & UNIQUE. You should also provide additional approaches, such as utilizing GROUP BY to deal with duplicate data items.

48. Have you ever worked with big data in a cloud computing environment?

Since most companies are now shifting to cloud-based environments, this question lets the interviewer know how prepared you are to work in a cloud-based environment. You should show your preparedness and familiarity with the cloud-based environment along with the pros of cloud computing such as:

- Its flexibility and scalability.
- Security and mobility.
- Risk-free data access from anywhere.

Conclusion

Data Engineering is a demanding career and it takes a lot of effort to become one. As a data engineer, you must be prepared for data science challenges that may arise during an interview. Many problems have multi-step solutions, and having them planned ahead of time allows you to map out solutions as you go through the interview process. Here, you will not only get information about commonly asked interview questions on data engineering, but you will also ace the interview with your responses.

Useful Resources:

- Big Data Interview Questions
- Python Interview Questions
- Azure Interview Questions
- AWS Interview Questions
- Additional Technical Interview Resources

Frequently Asked Questions

49. What are the roles and responsibilities of a data engineer?

Roles and responsibilities of a data engineer include:

- **Work on data Architecture:** Plan create and maintain data architecture.
- **Collect data:** Obtain data from reliable sources.
- **Research:** Look for any underlying issues.
- **Upgrade skills:** Remain updated with the latest algorithms and tools.
- **Create models:** Create predictive models to forecast future patterns and demands.
- **Automate tasks**.

50. How to become a Data Engineer?

To become a data engineer you have to:

- Learn computer-science fundamentals
- Master a programming language
- Understand concepts of software testing
- Learn database concepts, try to learn both relational and nonrelational database concepts.
- Learn how to design and construct a data warehouse as it is crucial.
- Understand the basics of cloud computing
- Learn about frameworks for batch, hybrid, and streaming data processing. Apache Pig, Apache Spark, and Apache Kafka are just a few examples.
- Scheduling your workflow- you can use tools like Apache Airflow for this
- Understand the basics of networking
- Learn how to use machine-readable configuration files to manage and provision your data centre. Learn how to use containers with tools like Docker, Kubernetes, and AWS CloudFormation.
- The final step of learning– learn about cybersecurity to protect your data.

51. Is Data Engineering a good career?

Data engineering is a trending and well-paid career. It's one of the world's fastest-growing positions in one of the world's fastest-growing industries, with one of the highest average earnings. The growth of Big Data attests to the fact that data engineers will always be in high demand.

52. Are data engineers paid well?

Yes, due to the shortage of talent in the field, companies are willing to pay a huge amount to mid-level as well as fresher data engineers. According to Glassdoor, the average salary of a data engineer in India is Rs. 8,56,643 LPA. **Learn More**.

53. What do Data Engineering interns do?

As a data engineering intern, you'll collaborate with business leads, analysts, and data scientists to gain a better understanding of the business domain and work with other fellow engineers to create data products.

Links to More Interview Questions

C Interview Questions

Web Api Interview Questions

Cpp Interview Questions

Machine Learning Interview Questions

Css Interview Questions

Django Interview Questions

Operating System Interview Questions

Git Interview Questions

Dbms Interview Questions

Pl Sql Interview Questions

Ansible Interview Questions

Php Interview Questions

Hibernate Interview Questions

Oops Interview Questions

Docker Interview Questions

Laravel Interview Questions

Dot Net Interview Questions

React Native Interview Questions

Java 8 Interview Questions

Spring Boot Interview Questions

Tableau Interview Questions

Java Interview Questions

C Sharp Interview Questions

Node Js Interview Questions

Devops Interview Questions

Mysql Interview Questions

Asp Net Interview Questions

Kubernetes Interview Questions

Aws Interview Questions

Mongodb Interview Questions

Power Bi Interview Questions

Linux Interview Questions

Jenkins Interview Questions