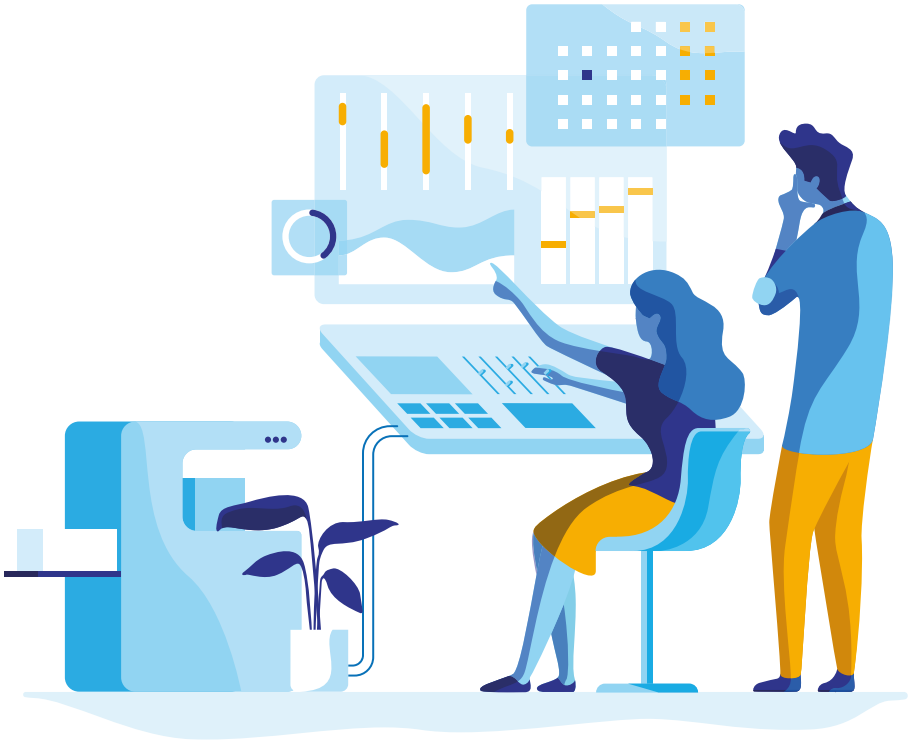
DATA ENGINEER

Interview Guide

simplilearn



Lead the Data Revolution

Whether you're new to the data and looking to break into a Data Engineering role, or you're an experienced Data Engineer looking for a new opportunity, preparing for an upcoming interview can be overwhelming. Given how competitive this market is right now, anticipating the questions that can be asked, and satisfactory answers to them can be the edge that you need in an interview. The following are some of the top data engineer interview questions you can likely expect at your interview, along with possible reasons why these questions are asked, plus the answers that interviewers are typically looking for.

Q: What is Data Engineering?

A: This may seem like a pretty basic question, but regardless of your skill level, this may come up during your interview. Your interviewer wants to see what your specific definition of data engineering is, which also makes it clear that you know what the work entails. So, what is it? In a nutshell, it is the act of transforming, cleansing, profiling, and aggregating large data sets. You can also take it a step further and discuss the daily duties of a data engineer, such as ad-hoc data query building and extracting, owning an organization's data stewardship, and so on.

Q. Why did you choose a career in Data Engineering?

A: An interviewer might ask this question to learn more about your motivation and interest behind choosing data engineering as a career. They want to employ individuals who are passionate about the field. You can start by sharing your story and insights you have gained to highlight what excites you most about being a data engineer.

Q. How Does a Data Warehouse Differ from an Operational Database?

A: This question may be more geared toward those on the intermediate level, but in some positions, it may also be considered an entry-level question. You'll want to answer by stating that databases using Delete SQL statements, Insert, and Update is standard operational databases that focus on speed and efficiency. As a result, analyzing data can be a little more complicated. With a data warehouse, on the other hand, aggregations, calculations, and select statements are the primary focus. These make data warehouses an ideal choice for data analysis.

Q. What Do *args and **kwargs Mean?

A: If you're interviewing for a more advanced role, you should be prepared to answer complex coding questions. This specific coding question is commonly asked in data engineering interviews, and you'll want to answer by telling your interviewer that *args defines an ordered function and that **kwargs represent unordered arguments used in a function. To impress your interviewer, you may want to write down this code in a visual example to demonstrate your expertise.

Q. As a Data Engineer, How Have You Handled a Job-Related Crisis?

A: Data engineers have a lot of responsibilities, and it's a genuine possibility that you'll face challenges while on the job, or even emergencies. Just be honest and let them know what you did to solve the problem. If you have yet to encounter an urgent issue while on the job or this is your first data engineering role, tell your interviewer what you would do in a hypothetical situation. For example, you can say that if data were to get lost or corrupted, you would work with IT to make sure data backups were ready to be loaded, and that other

team members have access to what they need.

Q. Do You Have Any Experience with Data Modeling?

A: Unless you are interviewing for an entry-level role, you will likely be asked this question at some point during your interview. Start with a simple yes or no. Even if you don't have experience with data modeling, you'll want to be at least able to define it: the act of transforming and processing fetched data and then sending it to the right individual(s). If you are experienced, you can go into detail about what you've done specifically. Perhaps you used tools like Talend, Pentaho, or Informatica. If so, say it. If not, simply being aware of the relevant industry tools and what they do would be helpful.

Q. Why are you interested in this job, and why should we hire you?

A: It is a fundamental question, but your answer can set you apart from the rest. To demonstrate your interest in the job, identify a few exciting features of the job, which makes it an excellent fit for you and then mention why you love the company.

For the second part of the question, link your skills, education, personality, and professional experience to the job and company culture. You can back your answers with examples from previous experience. As you justify your compatibility with the job and company, be sure to depict yourself as energetic, confident, motivated, and culturally fit for the company.

Q. What are the essential skills required to be a data engineer?

A: Every company can have its own definition of a data engineer, and they match your skills and qualifications with the company's assessment.

Here is a list of must-have skills and requirements if you are aiming to be a successful data engineer:

- Comprehensive knowledge about Data Modelling.
- Understanding about database design & database architecture. In-Depth Database Knowledge – SQL and NoSQL.
- Working experience of data stores and distributed systems like Hadoop (HDFS).
- Data Visualization Skills.
- Experience in Data Warehousing and ETL (Extract Transform Load) Tools.
- You should have robust computing and math skills.
- Outstanding communication, leadership, critical thinking, and problem-solving capabilities are an added advantage.

You can mention specific examples in which a data engineer would apply these skills.

Q. Can you name the essential frameworks and applications for data engineers?

A: This question is often asked to evaluate whether you understand the critical requirements for the position and have the desired technical skills. In your answer, accurately mention the names of frameworks along with your level of experience with each.

You can list all of the technical applications like SQL, Hadoop, Python, and more, along with your proficiency level in each. You can also state the frameworks which want to learn more about if given the opportunity.

Q. Are you experienced in Python, Java, Bash, or other scripting languages?

A: This question is asked to emphasize the importance of understanding scripting languages as a data engineer. It is essential to have a comprehensive knowledge of scripting languages, as it allows you to perform analytical tasks efficiently and automate data flow.

Q. Can you differentiate between a Data Engineer and Data Scientist?

A: With this question, the recruiter is trying to assess your understanding of different job roles within a data warehouse team. The skills and responsibilities of both positions often overlap, but they are distinct from each other.

Data Engineers develop, test, and maintain the complete architecture for data generation, whereas data scientists analyze and interpret complex data. They tend to focus on organization and translation of Big Data. Data scientists require data engineers to create the infrastructure for them to work.

Q. What, according to you, are the daily responsibilities of a data engineer?

A: This question assesses your understanding of the role of a data engineer role and job description.

You can explain some crucial tasks a data engineer like:

- Development, testing, and maintenance of architectures.
- Aligning the design with business requisites.
- Data acquisition and development of data set processes.
- Deploying machine learning and statistical models
- Developing pipelines for various ETL operations and data transformation
- Simplifying data cleansing and improving the de-duplication and building of data.
- Identifying ways to improve data reliability, flexibility, accuracy, and quality.

Q. What is your approach to developing a new analytical product as a data engineer?

A: The hiring managers want to know your role as a data engineer in developing a new product and evaluate your understanding of the product development cycle. As a data engineer, you control the outcome of the final product as you are responsible for building algorithms or metrics with the correct data.

Your first step would be to understand the outline of the entire product to comprehend the complete requirements and scope. Your second step would be looking into the details and reasons for each metric. Think about as many issues that could occur, and it helps you to create a more robust system with a suitable level of granularity.

Q. What was the algorithm you used on a recent project?

A: The interviewer might ask you to select an algorithm you have used in the past project and can ask some follow-up questions like:

- **Why did you choose this algorithm, and can you contrast this with other similar ones?**
- **What is the scalability of this algorithm with more data?**
- **Are you happy with the results? If you were given more time, what could you improve?**

These questions are a reflection of your thought process and technical knowledge. First, identify the project you might want to discuss. If you have an actual example within your area of expertise and an algorithm related to the company's work, then use it to pique the interest of your hiring manager. Secondly, make a list of all the models you worked with and your analysis. Start with simple models and do not overcomplicate things. The hiring managers want you to explain the results and their impact.

Q. What tools did you use in a recent project?

A: Interviewers want to assess your decision-making skills and knowledge about different tools. Therefore, use this question to explain your rationale for choosing specific tools over others.

- Walk the hiring managers through your thought process, explaining your reasons for considering the particular tool, its benefits, and the drawbacks of other technologies.
- If you find that the company works on the techniques you have previously worked on, then weave your experience with the similarities.

Q. What challenges came up during your recent project, and how did you overcome these challenges?

A: Any employer wants to evaluate how you react during difficulties and what you do to address and successfully handle the challenges.

When you talk about the problems you encountered, frame your answer using the STAR method:

- **Situation:** Brief them about the circumstances due to which problem occurred.
- **Task:** It is essential to elaborate on your role in overcoming the problem. For example, if you took a leadership role and provided a working solution, then showcasing it could be decisive if you were interviewing for a leadership position.
- **Action:** Walk the interviewer through the steps you took to fix the problem.
- **Result:** Always explain the consequences of your actions. Talk about the learnings and insights gained by you and other stakeholders.

Q. Have you ever transformed unstructured data into structured data?

A: It is an important question as your answer can demonstrate your understating of both the data types and your practical working experience. You can answer this question by briefly distinguishing between both categories. The unstructured data must be transformed into structured data for proper data analysis, and you can discuss the methods for transformation. You must share a real-world situation wherein you changed the unstructured data into structured data. If you are a fresh graduate and don't have professional experience, discuss information related to your academic projects.

Q. What is Data Modelling? Do you understand different Data Models?

A: Data Modelling is the initial step towards data analysis and database design phase. Interviewers want to understand your knowledge. You can explain that is the diagrammatic representation to show the relation between entities. First, the conceptual model is created, followed by the logical model and, finally, the physical model. The level of complexity also increases in this pattern.

Q. Can you list and explain the design schemas in Data Modelling?

A: Design schemas are the fundamentals of data engineering, and interviewers ask this question to test your data engineering knowledge. In your answer, try to be concise and accurate. Describe the two schemas, which are Star schema and Snowflake schema.

Explain that Star Schema is divided into a fact table referenced by multiple dimension tables, which are all linked to a fact table. In contrast, in Snowflake Schema, the fact table remains the same, and dimension tables are normalized into many layers looking like a snowflake.

Q. How would you validate a data migration from one database to another?

A: The validity of data and ensuring that no data is dropped should be of utmost priority for a data engineer. Hiring managers ask this question to understand your thought process on how validation of data would happen.

You should be able to speak about appropriate validation types in different scenarios. For instance, you could suggest that validation could be a simple comparison, or it can happen after the complete data migration.

Q. Have you worked with ETL? If yes, please state, which one do you prefer the most and why?

A: With this question, the recruiter needs to know your understanding and experience regarding the ETL (Extract Transform Load) tools and process. You should list all the tools in which you have expertise and pick one as your favourite. Point out the vital properties which make that tool stand out and validate your preference to demonstrate your knowledge in the ETL process.

Q. What is Hadoop? How is it related to Big data? Can you describe its different components?

A: This question is most commonly asked by hiring managers to verify your knowledge and experience in data engineering. You should tell them that Big data and Hadoop are related to each other as Hadoop is the most common tool for processing Big data, and you should be familiar with the framework.

With the escalation of big data, Hadoop has also become popular. It is an open-source software framework that utilizes various components to process big data. The developer of Hadoop is the Apache foundation, and its utilities increase the efficiency of many data applications.

Hadoop comprises of mainly four components:

1. HDFS stands for Hadoop Distributed File System and stores all of the data of Hadoop. Being a distributed file system, it has a high bandwidth and preserves the quality of data.
2. MapReduce processes large volumes of data.
3. Hadoop Common is a group of libraries and functions you can utilize in Hadoop.
4. YARN (Yet Another Resource Negotiator)deals with the allocation and management of resources in Hadoop.

Q. Do you have any experience in building data systems using the Hadoop framework?

A: If you have experience with Hadoop, state your answer with a detailed explanation of the work you did to focus on your skills and tool's expertise. You can explain all the essential features of Hadoop. For example, you can tell them you utilized the Hadoop framework because of its scalability and ability to increase the data processing speed while preserving the quality.

Some features of Hadoop include:

- It is Java-Based. Hence, there may be no additional training required for team members. Also, it is easy to use.
- As the data is stored within Hadoop, it is accessible in the case of hardware

failure from other paths, which makes it the best choice for handling big data.

- In Hadoop, data is stored in a cluster, making it independent of all the other operations.

  In case you have no experience with this tool, learn the necessary information about the tool's properties and attributes.

Q. Can you tell me about NameNode? What happens if NameNode crashes or comes to an end?

A: It is the centre-piece or central node of the Hadoop Distributed File System(HDFS), and it does not store actual data. It stores metadata. For example, the data being stored in DataNodes on which rack and which DataNode the information is stored. It tracks the different files present in clusters. Generally, there is one NameNode, so when it crashes, the system may not be available.

Q. Are you familiar with the concepts of Block and Block Scanner in HDFS?

A: You'll want to answer by describing that Blocks are the smallest unit of a data file. Hadoop automatically divides huge data files into blocks for secure storage. Block Scanner validates the list of blocks presented on a DataNode.

Q. What happens when Block Scanner detects a corrupted data block?

A: It is one of the most typical and popular interview questions for data engineers. You should answer this by stating all steps followed by a Block scanner when it finds a corrupted block of data.

Firstly, DataNode reports the corrupted block to NameNode.NameNode makes a replica using an existing model. If the system does not delete the corrupted data block, NameNode creates replicas as per the replication factor.

Q. What are the two messages that NameNode gets from DataNode?

A: NameNodes gets information about the data from DataNodes in the form of messages or signals.

The two signs are:

1. Block report signals which are the list of data blocks stored on DataNode and its functioning.
2. Heartbeat signals that the DataNode is alive and functional. It is a periodic report to establish whether to use NameNode or not. If this signal is not sent, it implies DataNode has stopped working.

Q. Can you elaborate on Reducer in Hadoop MapReduce? Explain the core methods of Reducer?

A: Reducer is the second stage of data processing in the Hadoop Framework. The Reducer processes the data output of the mapper and produces a final output that is stored in HDFS.

The Reducer has 3 phases:

1. **Shuffle:** The output from the mappers is shuffled and acts as the input for Reducer.
2. **Sorting** is done simultaneously with shuffling, and the output from different mappers is sorted.
3. **Reduce:** in this step, Reduces aggregates the key-value pair and gives the required output, which is stored on HDFS and is not further sorted.

There are three core methods in Reducer:

1. **Setup:** it configures various parameters like input data size.
2. **Reduce:** It is the main operation of Reducer. In this method, a task is defined for the associated key.
3. **Cleanup:** This method cleans temporary files at the end of the task.

Q. How can you deploy a big data solution?

A: While asking this question, the recruiter is interested in knowing the steps you would follow to deploy a big data solution. You should answer by emphasizing on the three significant steps which are:

1. **Data Integration/Ingestion:** In this step, the extraction of data using data sources like RDBMS, Salesforce, SAP, MySQL is done.
2. **Data storage:** The extracted data would be stored in an HDFS or NoSQL database.
3. **Data processing:** The last step should be deploying the solution using processing frameworks like MapReduce, Pig, and Spark.

Q. Which Python libraries would you utilize for proficient data processing?

A: This question lets the hiring manager evaluate whether the candidate knows the basics of Python as it is the most popular language used by data engineers.

Your answer should include NumPy as it is utilized for efficient processing of arrays of numbers and pandas, which is great for statistics and data preparation for machine learning work. The interviewer can ask you questions like why would you use these libraries and list some examples where you would not use them.

Q. Can you differentiate between list and tuples?

A: Again, this question assesses your in-depth knowledge of Python. In Python, List and Tuple are the classes of data structure where Lists are mutable and can be edited, but Tuples are immutable and cannot be modified. Support your points with the help of examples.

Q. How can you deal with duplicate data points in an SQL query?

A: Interviewers can ask this question to test your SQL knowledge and how invested you are in this interview process as they would expect you to ask questions in return. You can ask them what kind of data they are working with and what values would likely be duplicated?

You can suggest the use of SQL keywords DISTINCT & UNIQUE to reduce duplicate data points. You should also state other ways like using GROUP BY to deal with duplicate data points.

Q. Did you ever work with big data in a cloud computing environment?

A: Nowadays, most companies are moving their services to the cloud. Therefore, hiring managers would like to understand your cloud computing capabilities, knowledge of industry trends, and the future of the company's data.

You must answer it stating that you are prepared for the possibility of working in a virtual workspace as it offers many advantages like:

- Flexibility to scale up the environment as required,
- Secure access to data from anywhere
- Having backups in case of an emergency

Q. How can data analytics help the business grow and boost revenue?

A: Ultimately, it all comes down to business growth and revenue generation, and Big Data analysis has become crucial for businesses. All companies want to hire candidates who understand how to help the business grow, achieve their goals, and result in higher ROI.

You can answer this question by illustrating the advantages of data analytics to boost revenue, improve customer satisfaction, and increase profit. Data analytics helps in setting realistic goals and supports decision making. By implementing Big Data analytics, businesses may encounter a 5-20% significant increase in revenue. Walmart, Facebook, LinkedIn are some of the companies using big data analytics to boost their income.

GET YOUR DATA CAREER GOING AND SUCCEED AS A DATA ENGINEER

One of the best ways to crush your next job interview is to get formal training and earn your certification. If you're an aspiring data engineer, enroll in our Data Engineer Course today and get started by learning the skills that can help you land your dream job.

Our Data Engineer Master's Program is co-developed with IBM and includes hands-on industry training in Hadoop, PySpark, database management, Apache Spark, and countless other data engineering techniques, skills, and tools. Upon completion, you will receive certifications from both IBM and Simplilearn, showcasing your knowledge in the field of data engineering.

With the job market being so competitive nowadays, earning the relevant credentials has never been more critical. The technology industry is booming, and while more opportunities seem to open up as technology continues to advance, it also means more competition. A Data Engineering certificate can not only help you to land that job interview, but it can help prepare you for any questions that you may be asked during your interview. From fundamentals to advanced techniques, learn the ins and outs of this exciting industry, and get started on your career.

simplilearn

INDIA

Simplilearn Solutions Pvt Ltd.
53/1 C, Manoj Arcade, 24th Main,
Harlkunte
2nd Sector, HSR Layout
Bangalore: 560102

Call us at: 1800-212-7688

USA

Simplilearn Americas, Inc.
201 Spear Street, Suite 1100,
San Francisco, CA 94105
United States

Phone No: +1-844-532-7688