

If you're an aspiring data engineer, don't let the pandemic hold your career prospects back any longer. Look for remote data engineering positions at companies like these. Chances are, you'll be invited for an interview (although it may just be nothing but Zoom calls). So, all that being said, how can you prepare for the data engineering interview in 2021?

4 Essential Data Engineering Skills to Practice

First and foremost, aspiring data engineers need to prepare by practicing (or learning) the necessary skills for any data engineering position. These are:

1. General Programming

As you can expect, one of the most crucial skills a data engineer needs to have (and prepare for) is coding. Be sure to study the basics of data structures and algorithms before your interviews. For instance, an aspiring data engineer should know exactly when a certain data structure or algorithm is best for a given situation. They should also be able to explain why this is the case.

To help prepare, check out the Springboard [Intro to Python for Data Science](#) course.

2. SQL

Being a position in the data science field, it should come as no surprise that SQL is another vital skill for data engineering. As a matter of fact, data engineering candidates may find that they need to complete two different technical interviews, one for SQL and another for other coding skills. Many different data science positions require competency with SQL. Data engineers, however, are expected to have some of the most advanced SQL skills, considering their critical role in building reliable and scalable data processing and modeling tools that are deeply consequential for their company.

To help prepare, check out the [Khan Academy SQL Course](#).

3. Database Design

Database and system design is another crucial skill for any data engineer. As such, most companies will ask their candidates to design a data warehouse given some real-life parameters or use cases. Be sure to use the whiteboard during these parts of the interview to illustrate your particular way of designing data systems.

To help prepare, check out this [Database Structure and Design Tutorial](#) by Lucidchart.

4. Data Architecture and Big Data Frameworks

Most companies will expect their candidates to be competent with specific big data frameworks like [Hadoop](#), [Kafka](#), [Spark](#), or [Hive](#). The best way to prepare is to become comfortable with as

many of these frameworks as possible. You can also find a lot of educational value in the official documentation for each framework.

To help prepare, check out the Springboard [Data Analysis With Python, SQL, and R](#) learning path.

What's a Typical Data Engineering Interview Structure?

1. Phone Screenings

Most candidates will need to complete some initial phone screenings before being invited for an in-person interview. For data engineering interviews, candidates will need to complete a screening with an HR rep or hiring manager along with another technical screening.

2. Take-Home Assignments

Some companies may have their candidates initially complete a take-home project to test their technical skills. Before inviting a candidate to an on-site interview, the hiring managers will need a good assessment of their data engineering skills. A take-home coding challenge is the best way to do that in the beginning stages.

3. On-Site Interview

Finally, if a candidate passes the previous interview stages, they will be invited to an on-site interview. Data engineering interviews have the potential to be a strenuous matter, with candidates sitting down with up to 10 people in an 8 hour day. The length and rigor of these interviews may come as a surprise to those not expecting them.

5 Data Engineering Interview Tips

While studying the necessary fundamental skills is the best way to prepare for a data engineering interview, there are other additional ways to give yourself an edge.

- **Complete coding challenges with LeetCode or HackerRank:** At some point, either during your phone screening or the on-site interview, or both, you will need to complete some programming assignments. The best way to prepare for this is by doing some beforehand using something like [LeetCode](#) or [HackerRank](#).
- **Practice with the Whiteboard:** During the technical interview questions, you will have the opportunity to use the Whiteboard. If candidates are not accustomed to using the Whiteboard in this way, they may not take advantage of it as much as they need to. For this reason, you should practice using the [Whiteboard](#) to answer data engineering questions.

- **Practice your soft skills:** Data engineering is indeed a primarily technical position, but that doesn't mean your soft skills don't matter! You'll definitely be asked some behavioral questions regarding your soft skills, so be sure to practice them as much as anything else.
- **Use the STAR method for behavioral questions:** When you are inevitably asked those behavioral questions, you can use the [STAR Method](#) to answer them sufficiently.
- **Review documentation and best practices:** Data engineering candidates can also find a foundation for their knowledge in the documentation or best practices of widely used frameworks or tools.
 - **Oracle Database SQL Tuning Guide**
 - **GitHub Best Practices**
 - **Google Python Style Guide**

Data Engineer Interview Questions & Answers

[Pixeltrue](#) by [Pixeltrue](#)

Technical Data Engineer Interview Questions

1. What is an example of an unanticipated problem you faced while trying to merge data together from many different places? What was the solution you found?

In this question, the interviewer will inquire about your capacity to handle unexpected problems along with the creativity you use while solving them. Ideally, candidates will come prepared with several experiences they can choose from to answer this question.

2. What ETL tools or frameworks do you have experience with? Are there any you prefer over others?

ETL is a fundamental procedure in SQL. As such, every hiring manager will ask some questions about your knowledge of the ETL process. Your interviewers will be especially interested in your experience with different ETL tools. Therefore, candidates should reflect and think about the ETL tools they have worked with before. When you are asked for your favorite, be sure to answer in a way that also demonstrates your knowledge about the ETL process more generally.

3. Do you have experience with designing data systems using the Hadoop framework or something like it?

Hadoop is a software framework that is often asked about during data engineering interviews. You can know which frameworks your interviewers will ask about beforehand by consulting the job posting. You should expect a question similar to this one during your interview. As such, you should be sure to do your homework and become familiar with the languages and frameworks the job requires. When giving your answer, provide a detailed account of the projects you

completed using the framework. Give your interviewer some tangible examples to highlight your experience and competency with the framework.

6. How much experience do you have with NoSQL? Give me an example of a situation where you decided to create a NoSQL database instead of a relational database. Why did you do so?

7. Do you have any experience with data modeling? If so, what data modeling tools did you use?

Many data engineers have some experience with data modeling, it may well be within the expected responsibilities of data engineers in some organizations. Some interviewers may ask a question like this. If so, be sure to catalog the modeling tools you worked with in the past. Don't forget to include details on the advantages and disadvantages of each. If you have knowledge or experience with data modeling, this question is your time to shine!

Behavioral Data Engineer Interview Questions

1. Tell me about a time you suggested a change to improve the reliability and quality of company data. Were those changes ever made? Why or why not?

Your interviewer will be most interested in the improvements you can bring to the table as a data engineering candidate. They may ask some variation of this question to see how you take the initiative in improving things in your role. If you are asked this question, be sure to point out how your previous experience demonstrates that you are a self-starter. However, if you do not yet have this experience, be sure to prepare some remarks on the improvements you would and could be making if offered the job. Ultimately, be sure to keep your answer focused on the actual methods you employ as a data engineer to improve the quality of data for your organization.

2. What are the non-technical or soft skills that are the most invaluable for data engineers?

Technical data skills, it goes without saying, are the foundation of a data engineering role. This does not mean, however, that data engineering candidates can have these skills and nothing else. Many non-technical skills are vital to successful data engineering. Be sure to be creative when delivering your answer. Try to tell your interviewer something that has not been heard before for this question.

3. What are the fundamental characteristics necessary for a data engineer?

This is, in part, a culture-fit question. The hiring managers will be interested in comparing your conception of a skilled data engineer with that of the company. If there is a significant disparity between the company and the candidate, there may not be a cultural fit. Be sure to explain the skills and capabilities you believe to be vital for any data engineer.

4. What is the most significant professional hurdle you have encountered working as a data engineer?

One of the primary goals of behavioral questions is to investigate how candidates handle conflicts in the workplace. Your interviewer will be less interested in the actual details of what the hurdle was. Instead, they will be interested in how you handled the conflict and how determined you acted in the face of a challenge. It is best to use the STAR method to ace these kinds of behavioral questions.

5. How would you begin the development of a new product working as a data engineer?

These kinds of questions investigate your level of understanding of the product development cycle, especially how data engineering fits into the puzzle. To ace this question, be sure to detail how your data engineering skills could simplify or improve product development at that particular organization. You could use examples from your previous experiences, but you should come prepared with sufficient knowledge of the company's products. For instance, if you were to answer this question by describing the ways you would improve the product development of that company's flagship product, your chances of nailing this question are high.

=====

Interview Questions for Data Engineers

1. Using the following SQL table definitions and data, how would you construct a query that shows...

A data engineer needs to be able to construct and execute queries in order to understand the existing data, and to verify data transformations that are part of the data pipeline.

You can ask a few questions covering SQL to ensure the data engineer candidate has a good handle on the query language.

Here are some examples:

- With a product table defined with a name, SKU, and price, how would you construct a query that shows the lowest priced item?
- With an order table defined with a date, a product SKU, price, quantity, tax rate, and shipping rate, how would you construct a query that shows the average order cost?

You can use the SQL below to setup the examples above:

```
CREATE TABLE products (  
    sku INT NOT NULL,  
    name VARCHAR(50) NOT NULL,  
    price DECIMAL(10, 2) NOT NULL,  
    PRIMARY KEY(sku)  
);
```

```
CREATE TABLE orders (  
    product_sku INT NOT NULL,
```

```

price DECIMAL(10, 2) NOT NULL,
quantity INT NOT NULL,
tax_rate DECIMAL(3, 2) NOT NULL,
shipping_rate DECIMAL(3, 2) NOT NULL,
FOREIGN KEY(product_sku) REFERENCES products(sku)
);

```

```

INSERT INTO products VALUES (1, 'shirt', 25.99);
INSERT INTO products VALUES (2, 'sweater', 34.99);
INSERT INTO orders VALUES (1, 25.99, 1, 15.0, 3.0);
INSERT INTO orders VALUES (2, 34.99, 3, 13.0, 2.0);

```

2. Which Python libraries would you use for efficient data processing?

Python is one of the most popular languages used for data engineering. This question lets you know if the candidate knows the basics of Python. The answer should include *NumPy* and *pandas*.

NumPy is used for efficient processing of arrays of numbers, and pandas is great for stats, which are the bread and butter of data science work. Pandas is also good for preparing data for machine learning work. You can ask a candidate why they would use those libraries, and also ask for examples of situations where they would not use them.

From here, you can ask specific Python coding questions, such as:

- How would you transpose this data?
- How would you filter out outliers?

Here are two examples based on the Pandas library documentation:

```

# Here is the data
df = pd.DataFrame(data={'col1': [-10, 1, 2, 8], 'col2': [-5, 0.2, 0.4, 7]})
# How to transpose it?
df_transposed = df.T
# How to filter out the outliers?
df_no_outliers = df.ge(-3).le(3)

```

3. In this example web app, what data points would you collect?

The example web app could be a calendar similar to Outlook or Google calendar. In that case, it would be worthwhile to collect data on which calendar views are being used.

This question asks the data engineer candidate to analyze and understand the domain they're working in. They *need* to understand the domain they're working with because collecting data from a calendar web app can differ vastly from collecting data from IoT (Internet of Things) devices.

While a data engineer does not need to implement the code that records data points in the web app, they should be able to understand the needs of developers who do need to implement that

code. You can guide the candidate to a more specific answer by asking additional questions such as: *what data would we need to collect to find out how users are using certain features?*

If you want to know if a data engineer candidate understands that problem domain, ask this question.

4. How would you deal with duplicate data points in an SQL query?

This is a good question to ask a candidate because it should get them to ask you questions in return. For instance, they should ask you what kind of data you are working with, and what columns or values would likely be duplicated.

They should also suggest using the SQL keywords, `UNIQUE` and `DISTINCT`, for reducing duplicate data points. After that they should also suggest other ways to deal with duplicate data points, such as grouping the data using `GROUP BY` and filtering it further.

If you want to know if a candidate has a good grasp of SQL, ask this question.

5. What is a memorable data pipeline performance issue that you solved?

This question will give you insight into the candidate's past experiences with data pipeline implementation and how they were able to improve performance. Performance issues in a data pipeline can not only slow down the gathering of data, but can disrupt and slow down data analysis. This can have a direct impact on business decisions.

Here are some examples of experiences candidates could discuss:

- how they improved the performance of a specific SQL query
- how they upgraded a database from one type to another
- how they reduced the time it took to run a set of queries
- how they improved performance of importing or exporting of data (as an example, importing CSV files or exporting JSON or XML or CSV)
- how they improved retrieval of data from a backup system (for instance, Amazon Glacier or moving data from S3 storage into a faster data storage system)

If you want to know if the candidate has ideas on how to improve the performance of your data pipeline, also ask this as a question!

You can also ask a candidate how they have solved issues with malformed data and incorrect taxonomies.

6. Given an expected increase in data volume, how would you plan to add more capacity to the data processing architecture?

When asking this question, you can come up with a simple scenario. For example, you could be collecting data from IoT devices and are planning a rollout of thousands more devices (which will send back sensor data to the data pipeline). The data would be processed in two ways and stored in three ways: a data warehouse for analysis, a database, and caching layer for the interaction between a control panel web app and a backup system. *How would they add more capacity in that situation?*

This question asks the candidate to formulate a plan for the data pipeline to handle more data. They should be able to tell you what would be needed for this, such as needing more database instances in the cloud on Amazon Web Services, Microsoft Azure, or Google Cloud Platform. Or they could suggest better data compression, or removing old sets of data, or redirecting subsets of data to other parts of the architecture.

The candidate should be able to point to the various components and give you ideas about preparing those pieces for an increase in data volume.

For startups this can be particularly important. As startups start rapidly gaining more customers, they need to handle a dramatic increase in data collection volume *and* need to be able to view business reports and analysis quickly to reorient their direction.

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

The data engineer candidate should be concerned with the validity of data and ensuring that no data is dropped. They should be able to explain how validation of data would happen. In some cases, a comparison between hashes or timestamps can be used; in other cases, a more thorough comparison of data is needed to be able to validate.

The candidate should be able to give you an idea of which type of validation is appropriate in different scenarios. For example, the validation could occur continuously as the data flows into both databases, or the validation could occur once after a complete data migration happens. There could also be other approaches that the candidate suggests. The validation could also be a simple comparison, or more involved (in the case of more complex data structures).

8. How would you prepare for the migration of a dataset that's 1GB from a NoSQL database to an SQL-based database?

This is an interesting question to ask a candidate. There are a few startups that have started out with MongoDB or Couch (or some other NoSQL database) and found that it didn't suit their needs as they grew. The NoSQL database may contain a lot of duplicate data and may not have validation of all data fields, and could be schema-less. This makes migration more difficult. A good candidate will ask for more information about this and inquire about the details around the NoSQL and SQL databases, and will also ask about performance requirements.

The data engineer candidate should be able to tell you what steps are needed for migrating from NoSQL to SQL. For instance they should recommend ways to understand the existing data

schema. The candidate should give ideas on designing the new database schema to accommodate that data.

If you want to know whether a candidate understands how to prepare database backups and how to design schemas, ask this question.

=====

Q #1) Why did you study data engineering?

Answer: This question aims to learn about your education, work experience, and background. It might have been a natural choice in the continuation of your Information Systems or Computer Science degree. Or, maybe you have worked in a similar field, or you might be transitioning from an entirely different work area.

Whatever your story may be, don't hold back or shy away. And while you are sharing, keep highlighting the skills that you have learned along the way and the excellent work you have done.

However, don't start storytelling. Start with your educational background a little and then reach to the part when you knew you wanted to be a data engineer. And then move on how you reach here.

Q #2) What is the toughest thing about being a data engineer according to you?

Answer: You must answer this question honestly. Not every aspect of all the jobs is easy and your interviewer knows that. The aim of this question is not to pinpoint your weakness but to know how you work through things you find difficult to deal with.

You can say something like, "As a data engineer I find it hard to complete the request of all the departments in a company where most of them often come up with conflicting demands. So, I often find it challenging to balance them accordingly.

But it has offered me a valuable insight into the workings of the departments and the role they play in the overall company's structure." And this is just one example. You can and should put your point of view.

Q #3) Tell us an incident where you were supposed to bring data together from various sources but faced unexpected issues and how did you resolve it?

Answer: This question is an opportunity for you to demonstrate your problem-solving skills and how you adapt to the sudden plan changes. The question could be addressed generally or specifically with context to data engineering. If you haven't been through such an experience you can deliver a hypothetical answer.

Here is a sample answer: “In my previous franchise company, I and my team were supposed to collect data from various locations and systems. But one of the franchises changed their system without giving us any prior notice. This resulted in a handful of issues for data collection and processing.

To resolve that, we had to come up with a quick short-term solution first for getting the essential data into the company’s system. And after that, we have developed a long-term solution to prevent such issues from happening again.”

Q #4) How is the job of a data engineer different from that of a data architect?

Answer: This question is meant to check if you understand that there are differences within the team of a data warehouse. You can’t go wrong with the answer. The responsibilities of both of them overlap or vary depending on what the database maintenance department or the company needs.

You can say that “according to my experience, the difference between the roles of a data engineer and a data architect varies from company to company. Although they work very closely together, there are differences in their general responsibilities.

Managing the servers and building the architecture of the data system of a company is the responsibility of a data architect. And the work of a data engineer is to test and maintain that architecture. Along with that, we, data engineers, make sure that the data that is made available to the analysts is of high quality and reliable.”

Technical Interview Questions

Q #5) What are Big Data’s four V’s?



[image [source](#)]

Answer:

The four V's of Big Data are:

- The first V is **Velocity** which is referred to the rate at which Big Data is being generated over time. So, it can be considered as analyzing the data.
- The second V is the **Variety** of various forms of Big Data, be it within images, log files, media files, and voice recordings.
- The third V is the **Volume** of the data. It could be in the number of users, the number of tables, size of data, or the number of records.
- The fourth V is **Veracity** related to the uncertainty or certainty of the data. In other terms, it decides how sure you can be about the accuracy of the data.

Q #6) How is structured data different from unstructured data?

Answer: The below table explain the differences:

Structured Data	Unstructured Data
It can be stored in MS Access, Oracle, SQL Server, and other similar traditional database systems.	It can't be stored in a traditional database system.
It can be stored within different columns and rows.	It can't be stored in rows and columns.
An example of structured data is online	Examples of unstructured data are Tweets,

Structured Data

application transactions.

- 4) It can be easily defined within the data model.
- 5) It comes with a fixed size and contents.

Unstructured Data

Google searches, Facebook likes, etc.

It can't be defined according to the data model.

It comes in various sizes and contents.

Q #7) Which ETL tools are you familiar with?

Answer: Name all the ETL tools you have worked with. You can say, “ I have worked with SAS Data management, IBM Infosphere, and SAP Data Services. But my preferred one is PowerCenter from Informatica. It is efficient, has an extremely high-performance rate, and is flexible. In short, it has all the important properties of a good ETL tool.

They smoothly run business data operations and guarantee data access even when there are changes taking place in business or its structure.” Make sure you only talk about the ones you have worked with and the ones you like working with. Or, it could tank your interview later.

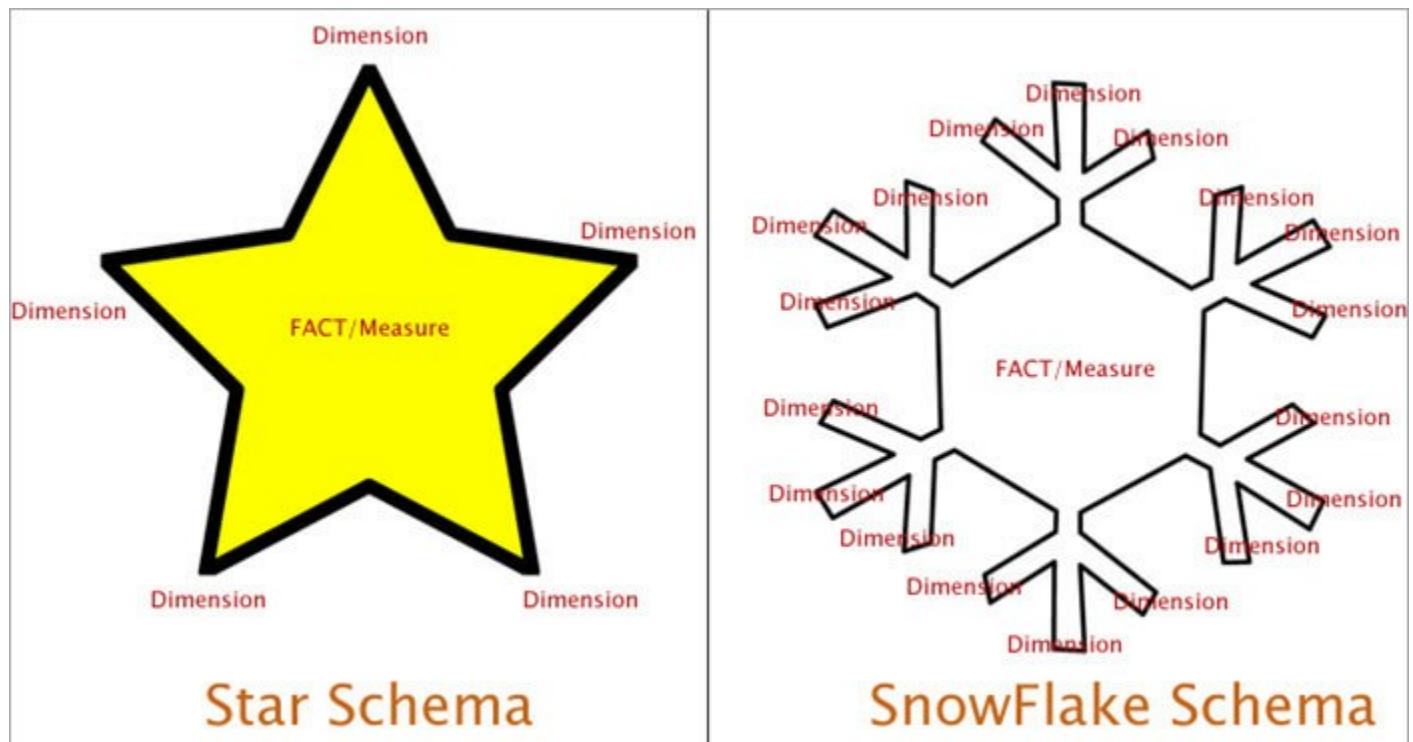
Q #8) Tell us about design schemas of data modeling.

Answer: Data modeling comes with two types of design schemas.

They are explained as follows:

- The first one is the **Star schema**, which is divided into two parts- the fact table and the dimension table. Here, both the tables are connected. Star schema is the simplest data mart schema style and is most widely approached as well. It is named so because its structure resembles a star.
- The second one is the **Snowflake schema** which is the extension of the star schema. It adds additional dimensions and is called a snowflake because its structure resembles that of a snowflake.

Q #9) What is the difference between Star schema and Snowflake schema?



[image [source](#)]

Answer: The below table explain the differences:

Star Schema	Snowflake Schema
1) The dimension table contains the hierarchies for the dimensions.	There are separate tables for hierarchies.
2) Here dimension tables surround a fact table.	Dimension tables surround a fact table and then they are further surrounded by dimension tables.
3) A fact table and any dimension table are connected by just a single join.	To fetch the data, it requires many joins.
4) It comes with a simple DB design.	It has a complex DB design.
5) Works well even with denormalized queries and data structures.	Works only with the normalized data structure.
6) Data redundancy- high.	Data redundancy- very low.
7) Aggregated data is contained in a single dimension.	Data is split into different dimension tables.
8) Faster cube processing.	Complex join slows cube processing.

Q #10) What is the difference between Data warehouse and Operational database?

Answer: The below table explain the differences:

Data Warehouse

- 1) These are designed to support the analytical processing of high-volume.
- 2) Historical data affects a data warehouse.
- 3) New, non-volatile data is added regularly but remains rarely changed.
It is designed for analyzing business
- 4) measures by attributes, subject areas, and categories.
Optimized for heavy loads and complex
- 5) queries accessing many rows at every table.
It is full of valid and consistent
- 6) information and doesn't need any real-time validation.
- 7) Supports a handful of OLTP like concurrent clients.
- 8) Its systems are mainly subject-oriented.
- 9) Data out.
- 10) A huge number of data can be accessed.
- 11) Created for OLAP, on-line Analytical Processing.

Operational Database

- These support transaction processing of high-volume.
- Current data affects the operational database.
- Data is updated regularly as the need arises.
- It is designed for real-time processing and business-dealings.
- Optimized for a simple single set of transactions like retrieving and adding one row at a time for every table.
- Improved for validating incoming information and uses validation data tables.
- Supports many concurrent clients.
- Its systems are mainly process-oriented.
- Data In.
- A limited number of data can be accessed.
- Created for OLTP, on-line transaction Processing.

Q #11) Point out the difference between OLTP and OLAP.

Answer: The below table explain the differences:

OLTP

- 1) Used to manage operational data.
- 2) Clients, clerks and IT professionals use it.
- 3) It is customer-oriented.
It manages current data, the ones that are extremely detailed and are used for decision making.
- 4) It has a 100 MB-GB database size.
It uses an ER (entity-relationship) data model along
- 6) with a database design that is application-oriented.

OLAP

- Used to manage informational data.
- Managers, analysts, executives, and other knowledge workers use it.
- It is market-oriented.
- It manages a huge amount of historical data. It also provides facilities for aggregation and summarization along with managing and storing data at different levels of granularity. Hence the data becomes more comfortable to be used in decision making.
- It has a 100 GB-TB database size.
- OLAP uses either a snowflake or star model along with a database design that is subject-oriented.

OLTP

- 7) The volume of data is not very large.
- 8) Access mode- Read/Write.
- 9) Completely normalized.
- 10) Its processing speed is very fast.

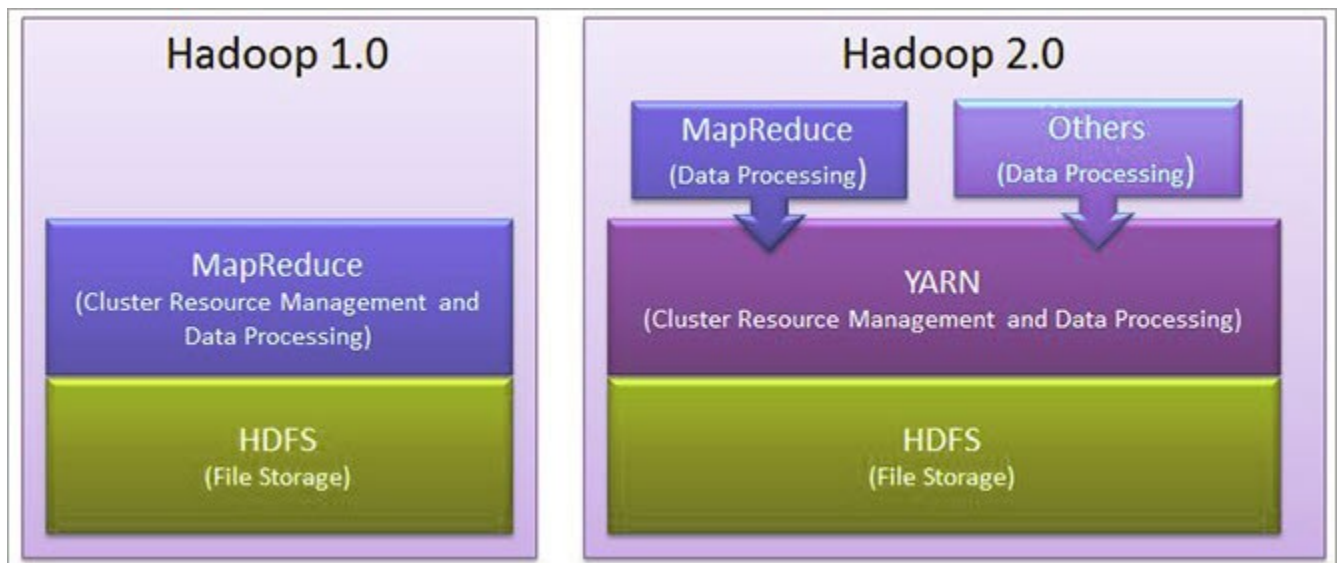
OLAP

- It has a large volume of data.
- The access mode is mostly write.
- Partially normalized.
- Its processing speed depends on the number of files it contains, complex queries, and batch data refresh

Q #12) Explain the main concept behind the Framework of Apache Hadoop.

Answer: It is based on the MapReduce algorithm. In this algorithm, to process a huge data set, Map and Reduce operations are used. Map, filters and sorts the data while Reduce, summarizes the data. Scalability and fault tolerance are the key points in this concept. We can achieve these features in Apache Hadoop by efficiently implementing MapReduce and Multi-threading.

Q #13) Have you ever worked with Hadoop Framework?



[image [source](#)]

Answer: Many hiring managers ask about the Hadoop tool in the interview to know if you are familiar with the tools and languages the company uses. If you have worked with the Hadoop Framework, tell them the details of your project to bring in light about your knowledge and skills with the tool and its capabilities. And if you haven't ever worked with it, some research to show some familiarity with its attributes will also work.

You can say, **for example**, "While working on a team project, I have had the chance to work with Hadoop. We were focused on increasing the efficiency of data processing, so, due to its ability to increase the speed of data processing without compromising the quality during its distributed processing, we decided to use Hadoop.

And as my previous company expected a considerable increase in data processing over the next few months, its scalability came in handy as well. Hadoop is also an open-source network based on Java, that makes it the best option for the projects with limited resources and an easy one to use without any additional training.”

Q #14) Mention some important features of Hadoop.

Answer: Features are as follows:

- Hadoop is a free open source framework where we can alter the source code as per our requirement.
- It supports the faster-distributed processing of data. HDFS Hadoop stores data in a distributed manner and uses MapReduce to parallel process the data.
- Hadoop is highly tolerant and by default, at different nodes, it allows the user to create three replicas of each block. So, if one of the nodes is unsuccessful, we can recover the data from another node.
- It is also scalable and is compatible with many hardware.
- Since Hadoop stored data in clusters, independent of all the other operations. Hence it is reliable. The stored data remains unaffected by the malfunctioning of the machines. And so, it is highly available as well.

Q #15) How can you increase the business revenue by analyzing Big Data?

Answer: Big data analysis is a vital part of the businesses since it helps them to differentiate from one another along with increasing the revenue. Big data analytics offers customized suggestions and recommendations to businesses through predictive analysis.

It also helps businesses in launching new products based on the preferences and needs of the customers. This helps the businesses earn significantly more, approximately 5-20% more. Companies like Bank of America, LinkedIn, Twitter, Walmart, Facebook, etc. use Big Data Analysis to increase their revenue.

Q #16) While deploying a Big Data solution, what steps you must follow?

Answer: There are three steps to be followed while deploying a Big Data solution:

- **Data Ingestion-** It is the first step in deploying a Big Data solution. It is the extraction of the data from various sources like SAP, MYSQL, Salesforce, log files, internal database, etc. Data ingestion can happen through real-time streaming or batch jobs.
- **Data Storage-** After the data is ingested, the extracted data should be stored somewhere. It is either stored in HDFS or NoSQL databases. HDFS works well for sequential access through HBase for random read or writes access.
- **Data Processing-** This is the third and the concluding step for deploying on a Big Data solution. After storage, the data is processed through one of the main frameworks like MapReduce or Pig.

Q #17) What is a block and block scanner in HDFS?

Answer: A block is the minimum amount of data that can be written or read in HDFS. 64MB is the default size of a block.

The block scanner is a program that tracks the number of blocks on a DataNode periodically along with verifying them for any possible checksum errors and data corruption.

Q #18) What are the challenges you have faced while introducing new data analytics applications if you have ever introduced one?

Answer: If you have never introduced new data analytics, you can simply say so. Because they are quite expensive and hence it is not often that companies do that. But if a company decides to invest in it, it can be an extremely ambitious project. It would need highly trained employees to install, connect, use, and maintain these tools.

So, if you have ever been through the process, tell them what obstacles you faced and how you overcame them. If you haven't, tell them in detail what you know about the process. This question determines if you have the basic know-how to get through the problems that might arise during the introduction of new data analytics applications.

Sample Answer; "I have been a part of introducing new data analytics in my previous company. The entire process is elaborate and needs a well-planned process for a smoothest possible transition.

However, even with immaculate planning, we can't always avoid unforeseen circumstances and issues. One such issue was an incredibly high demand for user licenses. It went over and beyond what we expected. For obtaining the additional licenses, the company had to reallocate the financial resources.

Also, training had to be planned in a way that it doesn't hamper the workflow. Also, we had to optimize the infrastructure to support the high number of users."

Q #19) What if NameNode crashes in the HDFS cluster?

Answer: The HDFS cluster only has one NameNode and it maintains DataNode's metadata. Having only one NameNode gives HDFS clusters a single point of failure.

So, if NameNode crashes, systems might become unavailable. To prevent that, we can specify a secondary NameNode that takes the periodic checkpoints in HDFS file systems but it is not a backup of the NameNode. But we can use it to recreate NameNode and restart.

Q #20) Difference between NAS and DAS in the Hadoop Cluster.

Answer: In NAS, storage and compute layers are separate, and then storage is distributed among various servers on the network. While in DAS, storage is usually attached to the computation node. Apache Hadoop is based on the principle of processing near a specific data location.

Hence, the storage disk should be local to computation. DAS helps you get performance on a Hadoop cluster and may be used on commodity hardware. In simple words, it is more cost-effective. NAS storage is preferred with high bandwidth of around 10 GbE.

Q #21) Is building a NoSQL database better than building a relational database?

Answer: In answer to this question, you must showcase your knowledge about both the databases. Also, you must back it up with an example of the situation demonstrating how you will or have applied the know-how in a real project.

Your answer could be something like this “ In some situations, it might be beneficial to build a NoSQL database. In my last company when the franchise system was exponentially increasing in size, we had to scale up quickly for making the most of all operational and sales data we had.

Scaling out is better than scaling up with bigger servers when handling the increased data processing load. It is cost-effective and easier to accomplish with NoSQL databases as it can easily deal with huge volumes of data. That comes in handy when you need to respond quickly to considerable data load shifts in the future.

Although relational databases come with better connectivity to any analytics tools. But NoSQL databases have a lot to offer.”

Q #22) What do you do when you encounter an unexpected problem with data maintenance? Have you tried any out-of-the-box solutions for that?

Answer: Inevitably, unexpected issues arise every once in a while in every routine task, even while data maintenance. This question aims to know if you can deal with high-pressure situations and how.

You can say something like “ data maintenance might be a routine task but it is vital to closely watch the specific tasks, including making sure of successful execution of the scripts.

Once while conducting the integrity check, I came across a corrupt index that could have caused serious issues in the future. That’s why I came up with a new maintenance task for preventing the addition of corrupt indexes into the database of the company.”

Q #23) Have you ever trained someone in your field? If yes, what have you found most challenging about it?

Answer: Usually data engineers are needed to train their coworkers on new systems or processes that you have created or train new employees on already existing systems and architecture. So, with this question, your interviewer wants to know if you can handle that. If you haven’t had the

chance to train someone yourself, talk about the challenges someone who trained or you know you faced.

A sample of the ideal answer will be something like this. “ Yes, I have had the chance to train small and large both groups of co-workers. Training new employees with significant experience in another company is the most challenging task I have come across. They are often so used to approaching data from one different perspective that they struggle to accept the way we do things.

Often, they are extremely opinionated and think they know everything right and that’s why it takes a lot of time for them to realize that a problem can have more than one solution. I try to encourage them to open their minds and accept alternate possibilities by emphasizing on how successful our architecture and processes have been.”

Q #24) What are the pros and cons of working in cloud computing?

Pros:

- No infrastructure cost.
- Minimum management.
- No hassles regarding management and administration.
- Easy to access.
- Pay for what you use.
- It is reliable.
- It offers data control, backup, and recovery.
- Huge storage.

Cons:

- It needs a good internet connection with equally good bandwidth to function well.
- It has its downtime.
- Your control of infrastructure will be limited.
- There is little flexibility.
- It has certain ongoing costs.
- There might be security and technical issues.

Q #25) The work of data engineers is usually ‘backstage’. Are you comfortable working away from the ‘spotlight’?

Answer: Your hiring manager wants to know if you love limelight or you can work well in both situations. Your answer should tell them that although you do like the limelight, you are comfortable working in the background as well.

“ What matters to me is that I should be an expert in my field and contribute to my company’s growth. If I have to work in the spotlight, I am comfortable doing that as well. If there is an issue

that executives need to address, I will not hesitate in raising my voice and bringing it to their attention.”

Q #26) What happens when the Block scanner detects a corrupt data block?

Answer: First of all DataNode reports to NameNode. Then NameNode starts creating a new replica through the replica of the corrupt block. Corrupted data block will not be deleted if the replication count of the right replicas matches the replication factor.

Q #27) Have you ever found a new innovative use for already existing data? Did it affect the company positively?

Answer: This question is meant for them to find out if you are self-motivated and eager enough to contribute to the success of the projects. If possible, answer the question with an example where you took the charge of a project or came up with an idea. And if you ever presented a novel solution to a problem, don't miss it either.

Example answer: “ In my last job, I took part in finding out why we have a high employee turnover rate. I observed the data closely from various departments where I found highly correlated data in key areas like finance, marketing, operations, etc. and the rate of employee turnover.

Collaborated with the department analysts for a better understanding of those correlations. With our understanding, we made some strategic changes that affected the employee turnover rate positively.”

Q #28) What non-technical skills do you think comes in most handy as a data engineer?

Answer: Try to avoid the most obvious answers like communicating or interpersonal skills. You can say, “prioritizing and multitasking have often come in handy in my job. We get various tasks in a day because we work with different departments. And hence, it becomes vital that we prioritize them. It makes our work easy and helps us efficiently finishing them all.”

Q #29) What are some common problems you have faced as a data engineer?

Answer: These are:

- Continuous and real-time integration.
- Storing huge amounts of data and information from those data.
- Constraints of resources.
- Considering which tools to use and which ones can deliver the best results.

Conclusion

Data engineering might sound like a routine boring job but there are many interesting facets to it. That is evident from the possible scenario questions interviewers might ask. You should be ready

to answer not just technical bookish questions but also situational questions like the above-listed ones. Only then you will be able to prove that you can do your job well and deserve it.

=====

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

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.

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

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.

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

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.

6. Do You Have Any Experience with Data Modeling?

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

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

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.

29. How can you deploy a big data solution?

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.

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

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.

31. Can you differentiate between list and tuples?

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.

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

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.

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

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

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

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.

=====

What Does a Data Engineer Do?

Given its varied skill set, a data engineering role can span many different job descriptions. A data engineer can be responsible for [database](#) design, schema design, and creating multiple database solutions. This work might also involve a Database Administrator.

As a **data engineer**, you might act as a bridge between the database and the [data science](#) teams. In that case, you'll be responsible for data cleaning and preparation, as well. If big data is involved, then it's your job to come up with an efficient solution for that data. This work can overlap with the DevOps role.

You'll also need to make efficient data queries for reporting and analysis. You might need to interact with multiple databases or write Stored Procedures. For many solutions like high-traffic websites or services, there may be more than one database present. In these cases, the data engineer is responsible for setting up the databases, maintaining them, and transferring data between them.

How Can Python Help Data Engineers?

Python is known for being the swiss army knife of programming languages. It's especially useful in data science, backend systems, and server-side scripting. That's because Python has strong typing, simple syntax, and an abundance of third-party libraries to use. [Pandas](#), [SciPy](#), [Tensorflow](#), [SQLAlchemy](#), and [NumPy](#) are some of the most widely used libraries in production across different industries.

Most importantly, Python decreases development time, which means fewer expenses for companies. For a data engineer, most code execution is database-bound, not CPU-bound. Because of this, it makes sense to capitalize on Python's simplicity, even at the cost of slower performance when compared to compiled languages such as C# and [Java](#).

Answering Data Engineer Interview Questions

Now that you know what your role might consist of, it's time to learn how to answer some data engineer interview questions! While there's a lot of ground to cover, you'll see practical Python examples throughout the tutorial to guide you along the way.

Questions on Relational Databases

Databases are one of the most crucial components in a system. Without them, there can be no state and no history. While you may not have considered database design to be a priority, know that it can have a significant impact on how quickly your page loads. In the past few years, several large corporations have introduced several new tools and techniques:

- NoSQL
- Cache databases
- Graph databases
- NoSQL support in SQL databases

These and other techniques were invented to try and increase the speed at which databases process requests. You'll likely need to talk about these concepts in your data engineer interview, so let's go over some questions!

Q1: Relational vs Non-Relational Databases

A **relational database** is one where data is stored in the form of a table. Each table has a **schema**, which is the columns and types a record is required to have. Each schema must have at

least one primary key that uniquely identifies that record. In other words, there are no duplicate rows in your database. Moreover, each table can be related to other tables using foreign keys.

One important aspect of relational databases is that a change in a schema must be applied to all records. This can sometimes cause breakages and big headaches during migrations. **Non-relational databases** tackle things in a different way. They are inherently schema-less, which means that records can be saved with different schemas and with a different, nested structure. Records can still have primary keys, but a change in the schema is done on an entry-by-entry basis.

You would need to perform a speed comparison test based on the type of function being performed. You can choose `INSERT`, `UPDATE`, `DELETE`, or another function. Schema design, indices, the number of aggregations, and the number of records will also affect this analysis, so you'll need to test thoroughly. You'll learn more about how to do this later on.

Databases also differ in **scalability**. A non-relational database may be less of a headache to distribute. That's because a collection of related records can be easily stored on a particular node. On the other hand, relational databases require more thought and usually make use of a master-slave system.

A SQLite Example

Now that you've answered what relational databases are, it's time to dig into some Python! [SQLite](#) is a convenient database that you can use on your local machine. The database is a single file, which makes it ideal for prototyping purposes. First, import the required Python library and create a new database:

```
import sqlite3

db = sqlite3.connect(':memory:') # Using an in-memory database
cur = db.cursor()
```

You're now connected to an in-memory database and have your cursor object ready to go.

Next, you'll create the following three tables:

1. **Customer:** This table will contain a primary key as well as the customer's first and last names.
2. **Items:** This table will contain a primary key, the item name, and the item price.
3. **Items Bought:** This table will contain an order number, date, and price. It will also connect to the primary keys in the Items and Customer tables.

Now that you have an idea of what your tables will look like, you can go ahead and create them:

```
cur.execute('''CREATE TABLE IF NOT EXISTS Customer (
                id integer PRIMARY KEY,
                firstname varchar(255),
                lastname varchar(255) )''')
```

```

cur.execute('''CREATE TABLE IF NOT EXISTS Item (
            id integer PRIMARY KEY,
            title varchar(255),
            price decimal )''')
cur.execute('''CREATE TABLE IF NOT EXISTS BoughtItem (
            ordernumber integer PRIMARY KEY,
            customerid integer,
            itemid integer,
            price decimal,
            CONSTRAINT customerid
                FOREIGN KEY (customerid) REFERENCES Customer(id),
            CONSTRAINT itemid
                FOREIGN KEY (itemid) REFERENCES Item(id) )''')

```

You've passed a query to `cur.execute()` to create your three tables.

The last step is to populate your tables with data:

```

cur.execute('''INSERT INTO Customer(firstname, lastname)
            VALUES ('Bob', 'Adams'),
            ('Amy', 'Smith'),
            ('Rob', 'Bennet');''')
cur.execute('''INSERT INTO Item(title, price)
            VALUES ('USB', 10.2),
            ('Mouse', 12.23),
            ('Monitor', 199.99);''')
cur.execute('''INSERT INTO BoughtItem(customerid, itemid, price)
            VALUES (1, 1, 10.2),
            (1, 2, 12.23),
            (1, 3, 199.99),
            (2, 3, 180.00),
            (3, 2, 11.23);''') # Discounted price

```

Now that there are a few records in each table, you can use this data to answer a few more data engineer interview questions.

A Python Best Practices Handbook

python-guide.org

[Remove ads](#)

Q2: SQL Aggregation Functions

Aggregation functions are those that perform a mathematical operation on a result set. Some examples include `AVG`, `COUNT`, `MIN`, `MAX`, and `SUM`. Often, you'll need `GROUP BY` and `HAVING` clauses to complement these aggregations. One useful aggregation function is `AVG`, which you can use to compute the mean of a given result set:

```
>>> cur.execute('''SELECT itemid, AVG(price) FROM BoughtItem GROUP BY
itemid''')
>>> print(cur.fetchall())
[(1, 10.2), (2, 11.73), (3, 189.995)]
```

Here, you've retrieved the average price for each of the items bought in your database. You can see that the item with an `itemid` of 1 has an average price of \$10.20.

To make the above output easier to understand, you can display the item name instead of the `itemid`:

```
>>> cur.execute('''SELECT item.title, AVG(boughtitem.price) FROM BoughtItem as
boughtitem
...             INNER JOIN Item as item on (item.id = boughtitem.itemid)
...             GROUP BY boughtitem.itemid''')
...
>>> print(cur.fetchall())
[('USB', 10.2), ('Mouse', 11.73), ('Monitor', 189.995)]
```

Now, you see more easily that the item with an average price of \$10.20 is the `USB`.

Another useful aggregation is `SUM`. You can use this function to display the total amount of money that each customer spent:

```
>>> cur.execute('''SELECT customer.firstname, SUM(boughtitem.price) FROM
BoughtItem as boughtitem
...             INNER JOIN Customer as customer on (customer.id =
boughtitem.customerid)
...             GROUP BY customer.firstname''')
...
>>> print(cur.fetchall())
[('Amy', 180), ('Bob', 222.42000000000002), ('Rob', 11.23)]
```

On average, the customer named Amy spent about \$180, while Rob only spent \$11.23!

If your interviewer likes databases, then you might want to brush up on nested queries, join types, and the steps a relational database takes to perform your query.

Q3: Speeding Up SQL Queries

Speed depends on various factors, but is mostly affected by how many of each of the following are present:

- Joins

- Aggregations
- Traversals
- Records

The greater the number of joins, the higher the complexity and the larger the number of traversals in tables. Multiple joins are quite expensive to perform on several thousands of records involving several tables because the database also needs to cache the intermediate result! At this point, you might start to think about how to increase your memory size.

Speed is also affected by whether or not there are **indices** present in the database. Indices are extremely important and allow you to quickly search through a table and find a match for some column specified in the query.

Indices sort the records at the cost of higher insert time, as well as some storage. Multiple columns can be combined to create a single index. For example, the columns `date` and `price` might be combined because your query depends on both conditions.

Q4: Debugging SQL Queries

Most databases include an `EXPLAIN QUERY PLAN` that describes the steps the database takes to execute the query. For SQLite, you can enable this functionality by adding `EXPLAIN QUERY PLAN` in front of a `SELECT` statement:

```
>>> cur.execute('''EXPLAIN QUERY PLAN SELECT customer.firstname, item.title,
...               item.price, boughtitem.price FROM BoughtItem as boughtitem
...               INNER JOIN Customer as customer on (customer.id =
boughtitem.customerid)
...               INNER JOIN Item as item on (item.id =
boughtitem.itemid)''')
...
>>> print(cur.fetchall())
[(4, 0, 0, 'SCAN TABLE BoughtItem AS boughtitem'),
 (6, 0, 0, 'SEARCH TABLE Customer AS customer USING INTEGER PRIMARY KEY
(rowid=?)'),
 (9, 0, 0, 'SEARCH TABLE Item AS item USING INTEGER PRIMARY KEY (rowid=?)')]
```

This query tries to list the first name, item title, original price, and bought price for all the bought items.

Here's what the query plan itself looks like:

```
SCAN TABLE BoughtItem AS boughtitem
SEARCH TABLE Customer AS customer USING INTEGER PRIMARY KEY (rowid=?)
SEARCH TABLE Item AS item USING INTEGER PRIMARY KEY (rowid=?)
```

Note that fetch statement in your Python code only returns the explanation, but not the results. That's because `EXPLAIN QUERY PLAN` is not intended to be used in production.

Your Guide to the Python Programming Language and a Best Practices Handbook

python-guide.org

[Remove ads](#)

Questions on Non-Relational Databases

In the previous section, you laid out the differences between relational and non-relational databases and used SQLite with Python. Now you're going to focus on NoSQL. Your goal is to highlight its strengths, differences, and use cases.

A MongoDB Example

You'll use the same data as before, but this time your database will be [MongoDB](#). This NoSQL database is document-based and scales very well. First things first, you'll need to install the required Python library:

```
$ pip install pymongo
```

You also might want to install the [MongoDB Compass Community](#). It includes a local [IDE](#) that's perfect for visualizing the database. With it, you can see the created records, create triggers, and act as visual admin for the database.

Note: To run the code in this section, you'll need a running database server. To learn more about how to set it up, check out [Introduction to MongoDB and Python](#).

Here's how you create the database and insert some data:

```
import pymongo

client = pymongo.MongoClient("mongodb://localhost:27017/")

# Note: This database is not created until it is populated by some data
db = client["example_database"]

customers = db["customers"]
items = db["items"]

customers_data = [{ "firstname": "Bob", "lastname": "Adams" },
                  { "firstname": "Amy", "lastname": "Smith" },
                  { "firstname": "Rob", "lastname": "Bennet" },]
items_data = [{ "title": "USB", "price": 10.2 },
```

```

        { "title": "Mouse", "price": 12.23 },
        { "title": "Monitor", "price": 199.99 },]

customers.insert_many(customers_data)
items.insert_many(items_data)

```

As you might have noticed, MongoDB stores data records in **collections**, which are the equivalent to a list of dictionaries in Python. In practice, MongoDB stores [BSON documents](#).

Q5: Querying Data With MongoDB

Let's try to replicate the `BoughtItem` table first, as you did in SQL. To do this, you must append a new field to a customer. MongoDB's [documentation](#) specifies that the keyword operator `set` can be used to update a record without having to write all the existing fields:

```

# Just add "boughtitems" to the customer where the firstname is Bob
bob = customers.update_many(
    {"firstname": "Bob"},
    {
        "$set": {
            "boughtitems": [
                {
                    "title": "USB",
                    "price": 10.2,
                    "currency": "EUR",
                    "notes": "Customer wants it delivered via FedEx",
                    "original_item_id": 1
                }
            ]
        }
    },
)

```

Notice how you added additional fields to the `customer` without explicitly defining the schema beforehand. Nifty!

In fact, you can update another customer with a slightly altered schema:

```

amy = customers.update_many(
    {"firstname": "Amy"},
    {
        "$set": {
            "boughtitems": [
                {
                    "title": "Monitor",
                    "price": 199.99,
                    "original_item_id": 3,
                    "discounted": False
                }
            ]
        }
    },
)

```

```
print(type(amy))    # pymongo.results.UpdateResult
```

Similar to SQL, document-based databases also allow queries and aggregations to be executed. However, the functionality can differ both syntactically and in the underlying execution. In fact, you might have noticed that MongoDB reserves the \$ character to specify some command or aggregation on the records, such as \$group. You can learn more about this behavior in the [official docs](#).

You can perform queries just like you did in SQL. To start, you can create an index:

```
>>> customers.create_index([("name", pymongo.DESCENDING)])
```

This is optional, but it speeds up queries that require name lookups.

Then, you can retrieve the customer names sorted in ascending order:

```
>>> items = customers.find().sort("name", pymongo.ASCENDING)
```

You can also iterate through and print the bought items:

```
>>> for item in items:
...     print(item.get('boughtitems'))
...
None
[{'title': 'Monitor', 'price': 199.99, 'original_item_id': 3, 'discounted':
False}]
[{'title': 'USB', 'price': 10.2, 'currency': 'EUR', 'notes': 'Customer wants
it delivered via FedEx', 'original_item_id': 1}]
```

You can even retrieve a list of unique names in the database:

```
>>> customers.distinct("firstname")
['Bob', 'Amy', 'Rob']
```

Now that you know the names of the customers in your database, you can create a query to retrieve information about them:

```
>>> for i in customers.find({"$or": [{'firstname': 'Bob'},
{'firstname': 'Amy'}]},
...                         {'firstname': 1, 'boughtitems': 1,
'_id': 0}):
...     print(i)
...
{'firstname': 'Bob', 'boughtitems': [{'title': 'USB', 'price': 10.2,
'currency': 'EUR', 'notes': 'Customer wants it delivered via FedEx',
'original_item_id': 1}]}
{'firstname': 'Amy', 'boughtitems': [{'title': 'Monitor', 'price': 199.99,
'original_item_id': 3, 'discounted': False}]}
```

Here's the equivalent SQL query:

```
SELECT firstname, boughtitems FROM customers WHERE firstname LIKE ('Bob',  
'Amy')
```

Note that even though the syntax may differ only slightly, there's a drastic difference in the way queries are executed underneath the hood. This is to be expected because of the different query structures and use cases between SQL and NoSQL databases.

Find Your Dream Python Job

pythonjobshq.com

[Remove ads](#)

Q6: NoSQL vs SQL

If you have a constantly changing schema, such as financial regulatory information, then NoSQL can modify the records and nest related information. Imagine the number of joins you'd have to do in SQL if you had eight orders of nesting! However, this situation is more common than you would think.

Now, what if you want to run reports, extract information on that financial data, and infer conclusions? In this case, you need to run complex queries, and SQL tends to be faster in this respect.

Note: SQL databases, particularly [PostgreSQL](#), have also released a feature that allows queryable [JSON](#) data to be inserted as part of a record. While this can combine the best of both worlds, speed may be of concern.

It's faster to query unstructured data from a NoSQL database than it is to query JSON fields from a JSON-type column in PostgreSQL. You can always do a [speed comparison](#) test for a definitive answer.

Nonetheless, this feature might reduce the need for an additional database. Sometimes, pickled or serialized objects are stored in records in the form of binary types, and then de-serialized on read.

Speed isn't the only metric, though. You'll also want to take into account things like transactions, atomicity, durability, and scalability. **Transactions** are important in financial applications, and such features take precedence.

Since there's a wide range of databases, each with its own features, it's the data engineer's job to make an informed decision on which database to use in each application. For more information, you can read up on [ACID](#) properties relating to database transactions.

You may also be asked what other databases you know of in your data engineer interview. There are several other relevant databases that are used by many companies:

- [Elastic Search](#) is highly efficient in text search. It leverages its document-based database to create a powerful search tool.
- [Newt DB](#) combines [ZODB](#) and the PostgreSQL JSONB feature to create a Python-friendly NoSQL database.
- [InfluxDB](#) is used in time-series applications to store events.

The list goes on, but this illustrates how a wide variety of available databases all cater to their niche industry.

Questions on Cache Databases

Cache databases hold frequently accessed data. They live alongside the main SQL and NoSQL databases. Their aim is to alleviate load and serve requests faster.

A Redis Example

You've covered SQL and NoSQL databases for long-term storage solutions, but what about faster, more immediate storage? How can a data engineer change how fast data is retrieved from a database?

Typical web-applications retrieve commonly-used data, like a user's profile or name, very often. If all of the data is contained in one database, then the number of **hits** the database server gets is going to be over the top and unnecessary. As such, a faster, more immediate storage solution is needed.

While this reduces server load, it also creates two headaches for the data engineer, backend team, and DevOps team. First, you'll now need some database that has a faster read time than your main SQL or NoSQL database. However, the contents of both databases must eventually match. (Welcome to the problem of **state consistency** between databases! Enjoy.)

The second headache is that DevOps now needs to worry about scalability, redundancy, and so on for the new cache database. In the next section, you'll dive into issues like these with the help of [Redis](#).

Q7: How to Use Cache Databases

You may have gotten enough information from the introduction to answer this question! A **cache database** is a fast storage solution used to store short-lived, structured, or unstructured data. It

can be partitioned and scaled according to your needs, but it's typically much smaller in size than your main database. Because of this, your cache database can reside in memory, allowing you to bypass the need to read from a disk.

Note: If you've ever used [dictionaries](#) in Python, then Redis follows the same structure. It's a key-value store, where you can `SET` and `GET` data just like a Python `dict`.

When a request comes in, you first check the cache database, then the main database. This way, you can prevent any unnecessary and repetitive requests from reaching the main database's server. Since a cache database has a lower read time, you also benefit from a performance increase!

You can use [pip](#) to install the required library:

```
$ pip install redis
```

Now, consider a request to get the user's name from their ID:

```
import redis
from datetime import timedelta

# In a real web application, configuration is obtained from settings or utils
r = redis.Redis()

# Assume this is a getter handling a request
def get_name(request, *args, **kwargs):
    id = request.get('id')
    if id in r:
        return r.get(id) # Assume that we have an {id: name} store
    else:
        # Get data from the main DB here, assume we already did it
        name = 'Bob'
        # Set the value in the cache database, with an expiration time
        r.setex(id, timedelta(minutes=60), value=name)
        return name
```

This code checks if the name is in Redis using the `id` key. If not, then the name is set with an expiration time, which you use because the cache is short-lived.

Now, what if your interviewer asks you what's wrong with this code? Your response should be that there's no [exception handling](#)! Databases can have many problems, like dropped connections, so it's always a good idea to try and catch those exceptions.

Your Weekly Dose of All Things Py

pycoders.com

[Remove ads](#)

Questions on Design Patterns and ETL Concepts

In large applications, you'll often use more than one type of database. In fact, it's possible to use PostgreSQL, MongoDB, and Redis all within just one application! One challenging problem is dealing with state changes between databases, which exposes the developer to issues of consistency. Consider the following scenario:

1. **A value** in Database #1 is updated.
2. **That same value** in Database #2 is kept the same (not updated).
3. **A query** is run on Database #2.

Now, you've got yourself an inconsistent and outdated result! The results returned from the second database won't reflect the updated value in the first one. This can happen with any two databases, but it's especially common when the main database is a NoSQL database, and information is transformed into SQL for query purposes.

Databases may have background workers to tackle such problems. These workers **extract** data from one database, **transform** it in some way, and **load** it into the target database. When you're converting from a NoSQL database to a SQL one, the Extract, transform, load (ETL) process takes the following steps:

1. **Extract:** There is a MongoDB trigger whenever a record is created, updated, and so on. A callback function is called [asynchronously](#) on a separate thread.
2. **Transform:** Parts of the record are extracted, normalized, and put into the correct data structure (or row) to be inserted into SQL.
3. **Load:** The SQL database is updated in batches, or as a single record for high volume writes.

This workflow is quite common in financial, gaming, and reporting applications. In these cases, the constantly-changing schema requires a NoSQL database, but reporting, analysis, and aggregations require a SQL database.

Q8: ETL Challenges

There are several challenging concepts in ETL, including the following:

- Big data
- Stateful problems
- Asynchronous workers
- Type-matching

The list goes on! However, since the steps in the ETL process are well-defined and logical, the data and backend engineers will typically worry more about performance and availability rather than implementation.

If your application is writing thousands of records per second to MongoDB, then your ETL worker needs to keep up with transforming, loading, and delivering the data to the user in the requested form. Speed and latency can become an issue, so these workers are typically written in fast languages. You can use compiled code for the transform step to speed things up, as this part is usually CPU-bound.

Note: Multi-processing and separation of workers are other solutions that you might want to consider.

If you're dealing with a lot of CPU-intensive functions, then you might want to check out [Numba](#). This library compiles functions to make them faster on execution. Best of all, this is easily implemented in Python, though there are some limitations on what functions can be used in these compiled functions.

Q9: Design Patterns in Big Data

Imagine Amazon needs to create a [recommender system](#) to suggest suitable products to users. The data science team needs data and lots of it! They go to you, the data engineer, and ask you to create a separate staging database warehouse. That's where they'll clean up and transform the data.

You might be shocked to receive such a request. When you have terabytes of data, you'll need multiple machines to handle all of that information. A database aggregation function can be a very complex operation. How can you query, aggregate, and make use of relatively big data in an efficient way?

Apache had initially introduced [MapReduce](#), which follows the **map, shuffle, reduce** workflow. The idea is to map different data on separate machines, also called clusters. Then, you can perform work on the data, grouped by a key, and finally, aggregate the data in the final stage.

This workflow is still used today, but it's been fading recently in favor of [Spark](#). The design pattern, however, forms the basis of most big data workflows and is a highly intriguing concept. You can read more on MapReduce at [IBM Analytics](#).

Q10: Common Aspects of the ETL Process and Big Data Workflows

You might think this a rather odd question, but it's simply a check of your computer science knowledge, as well as your overall design knowledge and experience.

Both workflows follow the **Producer-Consumer** pattern. A worker (the Producer) produces data of some kind and outputs it to a pipeline. This pipeline can take many forms, including network messages and triggers. After the Producer outputs the data, the Consumer consumes and makes use of it. These workers typically work in an asynchronous manner and are executed in separate processes.

You can liken the Producer to the extract and transform steps of the ETL process. Similarly, in big data, the **mapper** can be seen as the Producer, while the **reducer** is effectively the Consumer. This separation of concerns is extremely important and effective in the development and architecture design of applications.

=====

50 Data Engineer Interview Questions to Help You Prepare

April 2, 2021

By: Indeed Editorial Team

Receiving an interview request for a data engineer job is a key step toward obtaining the career you want. A job interview gives you a chance to impress your potential employer and encourage them to view you as an excellent candidate. To get ready for your job interview, think over your answers to both general questions and in-depth inquiries into your experience and background. In this article, we discuss 50 common data engineer interview questions and share some sample answers to help you prepare.

General questions

The following general interview questions allow the interviewers to learn about you and your interest in both the position and the company:

- Tell us about yourself.
- How would you describe yourself?
- What is your greatest strength?
- What is your biggest weakness?
- What motivates you to work hard?
- How do you handle pressure at work?
- What causes you stress at work and how do you manage it?
- Why are you leaving your current role?

- What are some positive things your last manager would say about you?
- What do you find interesting about this position?
- Why do you want to work for our organization?
- Why should we hire you to work here?
- What makes you unique?
- What makes you the best candidate for this position?
- What salary range are you expecting?
- Do you have questions for us?

Related: [125 Common Interview Questions and Answers \(with Tips\)](#)

Questions about experience and background

The following background and experience questions help the hiring team evaluate your qualifications and assess whether your goals are in line with the organization's values and objectives:

- What would you bring to our organization?
- What do you like most about your current job?
- What do you like least about your current position?
- Tell us about your data engineering work experience.
- What do you appreciate most about data engineering?
- What do you enjoy least about data engineering?
- Can you describe your biggest accomplishment?
- Share one of your biggest challenges and how you overcame it.
- What is your preferred work environment?
- Are you comfortable with reporting to superiors younger than you?
- Do you consider yourself a leader?
- What is your definition of professional success?
- How do you envision your career path?
- Where do you see yourself in five years?

Related: [12 Tough Interview Questions and Answers](#)

In-depth questions

The following interview questions enable the hiring manager to gain a comprehensive understanding of your competencies and assess how you would respond to issues that may arise at work:

- What are the most important skills for a data engineer to have?
- What data engineering platforms and software are you familiar with?
- Which computer languages can you use fluently?
- Do you tend to focus on pipelines, databases or both?
- How do you create reliable data pipelines?
- Tell us about a distributed system you've built. How did you engineer it?

- Tell us about a time you found a new use case for an existing database. How did your discovery impact the company positively?
- Do you have any experience with data modeling?
- What common data engineering maxim do you disagree with?
- Do you have a data engineering philosophy?
- What is a data-first mindset?
- How do you handle conflict with coworkers? Can you give us an example?
- Can you recall a time when you disagreed with your supervisor? How did you handle it?

Data engineer interview questions with sample answers

Below are seven of the most common job interview questions for data engineers. Review the explanation and sample responses as you think of your own answers to prepare for your interview.

1. What is data engineering?
2. What are the essential qualities of a data engineer?
3. Which frameworks and applications are critical for data engineers?
4. Can you explain the design schemas relevant to data modeling?
5. Do you consider yourself database- or pipeline-centric?
6. What is the biggest professional challenge you have overcome as a data engineer?
7. As a data engineer, how would you prepare to develop a new product?

1. What is data engineering?

Interviewers frequently bring this question up to assess whether you can discuss your field in an understandable and competent way. When you answer, try to include a general summary as well as a brief discussion of how data engineers collaborate with colleagues.

Example: *“Data engineering powers the collection and processing of information through a combination of desktop software, mobile applications, cloud-based servers and physical infrastructure. Effective data engineering requires careful construction, strong pipelines and smart collaborators. Data engineers are essential partners of data scientists, who analyze and use the information we collect.”*

2. What are the essential qualities of a data engineer?

An interviewer might ask this question to determine whether your idea of a skilled professional matches the company's assessment. In your answer, discuss the skills and abilities that you think are essential for data engineers. Try to mention specific instances in which a data engineer would apply these skills.

Example: *“A successful data engineer needs to know how to architect distributed systems and data stores, create dependable pipelines and combine data sources effectively. Data engineers also need to be able to collaborate with team members and colleagues from other departments.”*

To accomplish all of these tasks, a data engineer needs strong math and computing skills, critical thinking and problem-solving skills and communication and leadership capabilities.”

3. Which frameworks and applications are critical for data engineers?

Hiring managers often ask this question to assess your understanding of the key requirements for the job and to find out whether you have essential technical skills. When you answer, be as specific as possible and mention the names of frameworks and applications. Consider mentioning your experience level with each as well.

Example: *“Data engineers have to be proficient in SQL, Amazon Web Services, Hadoop and Python. I am fluent with all of these frameworks and I am also familiar with Tableau, Java, Hive and Apache Spark. I embrace every opportunity to learn new frameworks.”*

4. Can you explain the design schemas relevant to data modeling?

Hiring teams may question you about design schemas as a way to test your knowledge of the fundamentals of data engineering. When you respond, do your best to explain the concept clearly and concisely.

Example: *“Data modeling involves two schemas, star and snowflake. Star schema includes dimension tables that are connected to a fact table. Snowflake schema includes a similar fact table and dimension tables with snowflake-like layers.”*

5. Do you consider yourself database- or pipeline-centric?

Interviewers frequently bring this question up to determine whether you have a focus area and if it matches with the needs of the company. In your response, provide an honest assessment of your specialization. Try your best to reflect the whole scope of your technical knowledge and experience, especially if you have the skills for both data engineering specialties.

Example: *“Because I usually opt to work for smaller companies, I am a generalist who is equally comfortable with a database or pipeline focus. Since I specialize in both components, I have comprehensive knowledge of distributed systems and data warehouses.”*

6. What is the biggest professional challenge you have overcome as a data engineer?

Hiring managers often ask this question to learn how you address difficulties at work. Rather than learning about the details of these difficulties, they typically want to determine how resilient you are and how you process what you learn from challenging situations. When you answer, try using the STAR method, which involves stating the situation, task, action and result of the circumstances.

Example: *“Last year, I served as the lead data engineer for a project that had insufficient internal support. As a result, my portion of the project fell behind schedule and I risked*

disciplinary measures. After my team missed the first deadline, I took the initiative to meet with the project manager and proposed possible solutions. Based on my suggestions, the company assigned additional personnel to my team and we were able to complete the project successfully within the original timeline.”

Related: [How to Use the STAR Interview Response Technique](#)

7. As a data engineer, how would you prepare to develop a new product?

Hiring teams may question you about product development to determine how much you know about the product cycle and the data engineer's role in it. When you respond, mention some of the ways your knowledge could streamline the development process and some of the questions you would consider to make the best possible product.

Example: *“As a lead data engineer, I would request an outline of the entire project so I can understand the complete scope and the particular requirements. Once I know what the stakeholders want and why, I would sketch some scenarios that might arise. Then I would use my understanding to begin developing data tables with the appropriate level of granularity.”*

=====

Top 50 Data Engineer Interview Questions And Answers

Data Engineering is one of the best and most sought fields, which is the fastest-growing job option globally. Companies always want professional data engineers for their team, so they interview every candidate thoroughly. They look for specific knowledge and skills; candidates have to be prepared accordingly to meet interviewer expectations.

Being prepared for a **Data Engineer** interview with a question type that the interviewee might ask is an excellent way to ace the interview. No matter how experienced the candidate is because the interviewer's questions usually target multiple areas such as compatibility of the interviewee, leadership skills, project management skills, knowledge of technical tools, and so on.

Here [Wissenhive](#) has collected the **top 50 data engineer interview questions** to help you in preparing for the data engineer field role.

1. What do you understand by the term Data Engineer?

Data Engineering refers to a term that is used while working on **Big Data**. Data Engineer focuses on performing data collections and in-depth research on the application by generating raw data from various sources. Through data engineering, raw entity data can be converted into useful and beneficial information.

2. What are the roles and responsibilities of a Data Engineer?

The roles and responsibility of Data Engineer covers a wide array of things, but some of the important are

- Processing pipelines and handling data inflow
- Managing data staging areas
- ETL data transformation exercises responsibility
- Elimination of redundancies and performing data cleaning
- Creating operations, building ad-hoc, and extraction of native data methods.

3. Difference between Data Scientist and Data Analytics?

Scope	Data Scientist	Data Analytics
Background	It deals with various data operations.	It is related to data cleansing, transforming, and generating inferences from data.
Scope	Involves several underlying data procedures	Involves limited small data and static inferences
Data Type	Manages structured and unstructured data	Manages structured data only
Skills	Processes knowledge of statistics, mathematics, and learning algorithms	Has problems solving skills and knowledge of basic statistics.
Tools	Proficient in Python, SAS, TensorFlow, R, Spark, and Hadoop.	Knows SQL, Excel, and R, and Tableau

4. What do you mean by Data Modeling?

Data modeling refers to the simplification method for documenting complex designs of software in the form of diagrams for easy understanding without any prerequisites. It offers numerous benefits, such as simple conceptual and visualized representation of data objects, associated between data objects and their rules.

5. What are the various types of design schemas performed in Data Modelling?

There are two different types of design schemas performed in Data Modelling, and those are

1. Star Schema

2. Snowflake Schema

6. Differentiate between structured data and unstructured data?

Parameters	Structure Data	Unstructured Data
Storage Strategies	DBMS	Unmanageable structure of the file
Protocol Standards	SQL, ADO.net, and ODBC	CSV, SMSM, XML, and SMTP
Integration Tool	ELT	Batch processing or Manual data entry that includes codes
Scaling Level	Difficult	Easy
Example	Ordered dataset text file	Photos, videos, etc.

7. What do you understand about the Hadoop Application?

Hadoop refers to a set of open-source software framework utilities that facilities by using many computers' networks to solve a massive amount of computation and data-related problems. It provides the software framework for Big data processing and distributing storage by using the MapReduce programming model.

8. What are the main components of the Hadoop Application?

There are several components required while working on the Hadoop application, but some of the popular components are

- Hadoop Common: Consist of a standard set of libraries and utilities
- HDFS: Used for storing data and providing a distributed file system having high bandwidth.
- Hadoop MapReduce: Used for managing resources and scheduling task
- Hadoop YARN: Provides access to users for large-scale data processing.

9. What is Heartbeat in Hadoop?

In the Hadoop application, DataNode, and NameNode establish communication together. It refers to a signal which DataNode usually sends to NameNode to show its presence regularly.

10. What do you understand by the term Hadoop streaming?

Hadoop streaming refers to widely used Hadoop utilities for creating maps, performing various reduction operations, and submitting specific clusters.

11. What are some features available in Hadoop?

- It is an open-source framework.
- Works on distributed computing basis
- Uses parallel computing for processing data faster
- Store data in separate clusters
- It gives data redundancy to ensure no loss of data.

12. What are the three different Hadoop usage modes?

There are three different modes used in Hadoop, and those are

- Standalone mode
- Fully distributed mode
- Pseudo distributed mode

13. How is data security assured in the Hadoop application?

There are some steps that help in securing data in Hadoop Application.

- Begin with securing genuine and authentic channels, which helps in connecting clients to the platform.
- Consumers are making the usage of stamps received to request a service ticket.
- Clients make the usage of service tickets for correcting the corresponding server authentically.

14. What is a NameNode?

NameNode is one of the centerpieces or vital parts of HDFS. It helps in storing HDFS data and tracking multiple files from the clusters. However, the actual data does not store the information in NameNodes but is stored in DataNodes.

15. What are the main functions of secondary NameNode?

- **Fsimage** - It stores a copy of the FsImage file and EditLog
- **NameNode Crash** - If the original NameNode crashes, then secondary Fsimage of NameNode can be used to create the NameNode again.
- **Update** - It helps in automatically updating the Fsimage and EditLog file to keep the Fsimage file updated on secondary NameNode.
- **Checkpoint** - Secondary NameNode uses checkpoint to check if data is secured or not in HDFS.

16. How does the DataNode communicate with the NameNode?

There are two different methods that help DataNode and NameNode in communicating via messages, and those are

- Block reports
- Heartbeats

17. What are the 4V's of Big Data?

- Variety
- Velocity
- Volume
- Veracity

18. Define HDFS Block and Block Scanner?

HDFS Block refers to a single data entity, which is considered the smallest factor. Blocks automatically divide files into a smaller section when Hadoop encounters a large file.

HDFS Block Scanner verifies whether Hadoop created loss-of-blocks is updated on the DataNode successfully or not.

19. How does Block Scanner handle corrupted files?

- DataNode reports NameNode about a particular file when the block scanner finds any corrupted file in the system.
- Then NameNode processes files by using original corrupted files to create replicas.
- If created file replicas matches and replications get blocked, then it saves the corrupted data block instead of removing it.

20. What do you understand by COSHH?

The full form of COSHH is Classification and Optimization-based Scheduling for Heterogeneous Hadoop systems. The COSHH provides detailed scheduling at both the application and the cluster levels to have a direct positive impact on the job completion times.

21. Differentiate between Snowflake schema and Star schema?

Parameters	Snowflake Schema	Star Schema
Data stored in	Individual tables	Dimension tables
Redundancy	Low data	High data
Processor	Slower cube processor	Fast cube processor
Data presentation	Complex data-handling storage	Simple database designs

22. What are some of the Hadoop XML configuration files?

There are four different types of Hadoop XML configuration files available or presented in Hadoop, and those are:

- Mapred-site
- HDFS-site
- YARN-site
- Core-site

23. What is Combiner in Hadoop Application?

Combiner refers to an optional step between Reduce and Map. It takes all the necessary output from the Map function, building the key-value pairs, and submitting to the Hadoop reducer. Combiners' main task is to summarize the final result into summary records from Map with an identical key.

24. What are the three main methods of Reducer?

The three main methods of Reducer are

- Setup () - For configuring parameters
- Cleanup () - For clearing temporary files
- Reduce () - For reducing associated task

25. What do you mean by Star Join Schema?

Star Join Schema or Star Schema is one of the simplest type schemas for the Data Warehousing concept. This schema's structure is like a star, consisting of multiple associated dimension tables and fact tables. It is widely used while working with large data sets.

26. What do you mean by Snowflake Schema?

The snowflake schema refers to a star schema's primary extension with the presence of more dimensions. As the name suggests, the structure of this schema looks like a snowflake. It structures the data and, after normalization, split it into multiple tables.

27. What do you understand by FSCK?

FSCK or file system check refers to one of the most important commands that HDFS uses. It is mostly used to check for file discrepancies, inconsistencies, and problems.

28. What are some of the important languages or fields used by data engineers?

Data engineers use few languages or fields, and those are

- Mathematics (Linear algebra and probability)
- Trend regression and analysis
- Machine Learning
- Summary statistics
- Python
- SQL and Hive QL databases

- SAS and R programming languages

29. What are the objects created by the CREATE statement in MySQL?

- Database
- Table
- Trigger
- Function
- Index
- Event
- User
- View
- Procedure

30. How to check the structure of the database in MySQL?

To check the structure of the database in MySQL, Data Engineers can use the DESCRIBE command.

31. What do you understand by Rack Awareness?

Rack Awareness refers to a process where NameNode takes access from DataNode to enhance the network traffic while writing or reading any document or file in the Hadoop cluster, which is a nearby rack to Write or Read request. NameNode manages the id of the rack of every DataNode to achieve information from the rack.

32. What are the default port numbers for Task Tracker, Port Tracker, and NameNode in Hadoop Application?

- The default port of Task Tracker: 50060
- The default port of Port Tracker: 50060
- The default port of NameNode Tracker: 50060

33. What are distributed file systems in the Hadoop application?

Hadoop application works with a distributed scalable file system such as

- HFTP FS
- HDFS
- S3
- FS

The distributed file system of the Hadoop application is made on Google File System, which is designed to run with a large cluster on the computer system.

34. What is Big Data?

Big Data refers to a large amount of structured data and unstructured data, which is difficult to process by using traditional data storage methods. Data engineers prefer using the Hadoop application for managing large amounts of data.

35. How can Big Data Analytics improve a company's revenue?

Big Data Analytics helps many organizations in multiple ways, and those fundamental strategies are

- Effective usage of data to relate with structured growth
- Forecasting manpower and improving staffing strategies
- Increasing customer value and analyzing retention
- Strategies of bringing down the cost of production majorly

36. What is the difference between NAS and DAS in Hadoop Application?

	NAS	DAS
Storage	10(9) to 101(2) in byte	10(9) in byte
Management cost	GB is moderate	GB is high
Transmit data	Ethernet or IP/TCP	IDE/ SCSI

37. What do you understand by FIFO scheduling?

FIFO (First In First Out) scheduling is one of the Job scheduling algorithms, which is also known as FCFS (first come, first served). In this scheduling, the reporter chooses the job from the queue of work, the oldest job first.

38. What are the complex data collection/types supported by Hive?

- Struct
- Union
- Map
- Array

39. What is the usage of context objects in the Hadoop Application?

A context object refers to a means of communication that is used in the Hadoop application with mapper class. It presents system configuration jobs and details in the constructor obtained easily by using context objects. Context objects are used in three methods to send information, and those methods are map(), setup(), and cleanup().

40. What is Data Locality in Hadoop Application?

In Big Data, the data size is huge, which makes it difficult to move huge data across the different networks. Here, Hadoop helps in moving computation closer to data, and in that way, the data remains stored to the location.

41. What is the main use of hive?

Hive in the Hadoop ecosystem is used to build the user interface to manage all the Hadoop stored data. HBase's data mapped tables work when needed. Hive queries are very similar to SQL queries that are executed and converted into MapReduce jobs, which can be done to keep to manage complexity when multiple executing jobs at once.

42. What are the three components available in the Hive model of data?

There are three different types of components available in Hive data mode, and those are

- Tables
- Buckets
- Partitions

43. What do you understand by Hive's Metastore?

Metastore in Hive is used to store locations for the Hive tables and Schema. Data such as mappings, definitions, and other metadata are stored in the Hive Metastore. Later, it started storing data in an RDMS. It uses metadata records to enable the MapR FS and Hadoop FS destinations to write parquet data or drifting Avro.

44. What are the functions available in Hive to create the table?

The important functions available in the Hive to creation table are

- Explode (Map)
- Explode (Array)
- Stack ()
- JSON-tuple ()

45. Explain the role of the .hiverc file?

The role and responsibility of the .hiverc file is initialization. When individuals want to create or write code for the hive, they can open up the command-line interface. .hiverc is the first to load when CLI opened while containing the parameters that the user initially set.

46. What are the uses of **kwargs and *args?

The **kwargs function is used for donating argument sets that are in line to be input and unordered to function. The *args function allows users to define an ordered function for the usage in the command line.

47. What do you understand about Skewed tables in Hive?

Hive's skewed table refers to a special type of table that contains values column more often. The skewed table is a specific table in a hive, usually split into various separate files, and the remaining values go to the other files.

48. What do you understand about the Hive's SerDe?

In Hive, SerDe's full form is Serialization and Deserialization. It refers to an operation that involves passing records through the tables of the Hive. The Deserializer takes a record and changes it into an object of Java that the Hive understands. And then the serializer takes that object of Java and changes it into a processable format by HDFS. Later, HDFS take over the functions of storage.

49. What are the different types of SerDe implementations included in Hive?

There are several SerDe implementations available in Hive; you even can create your own custom. Some of the popular SerDe implementations are

- RegexSerDe
- ByteStreamTypedSerDe
- DelimitedJSONSerDe
- OpenCSVSerde

50. Differentiate between Database and Data Warehouse?

Scope	Database	Data Warehouse
Suited solution	OLTP Solutions	OLAP solutions
No. of users	It can handle thousands of users	It can handle a smaller number
Use	Recording data	Analysis data
Downtime	Always available	Scheduled downtime
Optimization	CRUD operations	Complex analysis
Data Type	Real-time detailed data	Summarized historical data

Here, this blog brings us to the end of the top 50 most asked questions by interviewers in a Data Engineering interview.

=====

The role of an Amazon Data Engineer

Much like several other top companies, Amazon has data engineering roles as one of their most critical hires in which they're expanding roles dramatically. Data engineering involves the collection and validation of data that can help the company meet its objectives. Data engineers face very unique challenges as to what kind of data must be selected, processed, and shaped, and to do this with competence makes it one of the most challenging jobs out there.

Data Engineers work alongside product managers, designers, data scientists, software engineers, and are an integral part of the team. They are responsible for extracting most of the data and transforming it into pipelines, which the rest of the team works upon. In simple words, a data engineer manages data and the data scientists explore it.

Some qualities of a Data Engineer at Amazon:

- Good command of database systems such as SQL and other programming languages is essential to be able to work on complex datasets.
- In-depth knowledge and experience with big data processing frameworks such as Hadoop or ApacheSpark, and analytical environments.

Additionally, Amazon has and operates by its [14 leadership skills](#), and hence looks for people who live these principles every day.

Interview Guide

The Amazon Data Engineering interview can be broadly divided into 3 rounds.

After applying for the job, a screening round is conducted, which is a telephonic interview with a recruiter. Candidates then proceed to the second round, which is a technical phone interview, with questions focusing on SQL and Data Modeling. This is followed by the third (and final) round i.e. the onsite round, which typically consists of 3-4 interviews, focussing on SQL, Database Management, Data Warehousing, and behavioral topics.

Recruiter Phone Screen

Overview

This is the first telephonic screening interview.

What the interviewer will assess

The recruiter will assess your knowledge of SQL and Data Modelling. They may also ask you to solve basic coding problems in Python.

Tips

- You must be prepared for the “Tell me about yourself” starter.
- Have a solid understanding of SQL and Python.
- Keep your answers short and crisp.

Interview Questions

1. Tell me about yourself, why do you want to join this company.
2. Give the basics of the window function in SQL.
3. Write the required SQL queries for a given order table.
4. Write the code for the Travelling Salesman problem and explain.
5. Write a code for finding any two (or three) numbers in the given array whose sum is equal to x.

Note: Questions related to applications of Pandas in Python may also be asked.

Technical Phone Screen

Overview

In this round, there is usually just only one interview, and you’ll be talking to a Data Engineer. Your knowledge on the following topics will be tested -

- Data Warehousing, ETL
- Data Modeling
- SQL
- Data Structures

You can expect some scenario-based questions on these topics. They may also inquire about previous experiences in projects.

What the interviewer will assess

- Fundamentals of SQL such as joins, subqueries, aggregations, filters, case statements, etc to solve the scenario-based questions.
- Your speed and efficiency in solving coding problems, and most importantly, your approach towards the problem.

Tips

- Don’t rush into coding, clarify all doubts beforehand.
- Think out loud so that the interviewer understands your approach. Sometimes they may give you hints indirectly, so you should use the opportunity to either reconsider the approach or explain why you tackled the problem that way.
- The interviewer wants to assess how you handle a problem, so it’s fine to check in with them on small syntax issues.

Interview Questions

1. Given two binary trees, determine whether they have the same inorder traversal.
2. Given an order table, write the SQL queries for the desired output. For example - find the maximum frequency of a name in a group.

Note: Questions related to linked lists, stacks, queues, doubly linked lists, performance tuning of a program may also be asked.

Want to prepare for your technical phone screen with an Amazon Data Engineer?

[→ Book a Mock Interview now!](#)

Onsite Round

Overview

This is the hardest of the three rounds, as it focuses on problem-solving skills through scenario-based questions. The following interviews take place -

- 2 Technical interviews; 45-60 minutes each
- Bar Raiser Round
- HR Round

The technical interviews sometimes have a “mixed” aspect, with questions ranging from Data Warehousing definitions to real-life problems solved in past projects.

3 broad types of interviews happen in the onsite round:

1. Technical Interviews:

The topics tested during the technical interviews are -

- Data Warehousing
- Data Modeling
- Complex SQL
- Big Data Technology

What the interviewer will assess:

- Problem analysis: How well do you understand the problem, what use of the data you make, and the use cases that you solve.
- SQL: Fundamentals, as well as complex SQL, will be tested; how well you handle SQL queries
- Data Modeling: How well you understand the need for the data and how it supports the use cases. Further, they will assess how you execute this in the form of SQL queries.

Tips:

- Data Modeling: Incorporate only the relevant details in the data model. Don't include unnecessary details and complicate the model with tangents that aren't core to the problem you're solving.
- SQL: You must have a good command of SQL. Practice joins, aggregate functions, analytical functions, correlated subqueries, etc.

Sample interview questions:

1. Explain the design schemas - star schema and snowflake schema (Data Warehousing)
2. Write SQL queries for a given order table (*this one is frequently asked*)
3. Design a data model to track products from vendors to the warehouse and then ultimately to customers.
4. Create a data model for a multinational company like Amazon.
5. Given a list of edges and nodes in a graph, write code to find the minimal canopy count.
6. What is the difference between a correlated query and a nested query?
7. What is a chasm trap?
8. What is an index? Give the different types of indexes and explain and differentiate between them.
9. Create required tables for eBay, define necessary relations, primary keys, and foreign keys.
10. Design a simple OLTP architecture for RedBus.
11. Write SQL queries to find groups having exactly three different tags.

Note: Practice questions related to designing a data pipeline given a specific reporting need.

2. Bar Raiser Round

For the bar raiser round, indirect questions will be asked regarding the [14 leadership skills](#).

This interview examines your ability to influence peers, collaborators, and stakeholders in cross-functional roles, as well as your previous experiences in doing so.

Sample interview questions:

- Tell us about a time when you worked in strict timelines?
- Have you ever had some kind of a conflict with your manager? How did you resolve it?
- Tell me about a time you made a mistake. How did you communicate your mistakes to the team?
- Tell me about a time you faced a crisis at work, how did you handle it?

Note: Don't forget to incorporate the [14 leadership skills](#) in your answers!

3. HR Round

For the HR round, questions may revolve around your previous projects and the use cases you have worked on in the past. Multiple HR rounds may be conducted to test different abilities. Group discussions may also take place.

Tips:

- They may ask about projects you've previously worked on, so be prepared for that. Don't forget to relate it to concepts of data engineering.
- Technical questions related to your projects, such as issues faced during the creation of pipelines, how you managed your databases, tables in the projects, etc. may be asked.

=====

Data engineer interviews at Facebook are really challenging. The questions are difficult, specific to Facebook, and cover a wide range of topics.

The good news is that the right preparation can make a big difference, and can help you land a data engineer job at Facebook. We have put together the ultimate guide to help you maximize your chances of success.

1. Interview process and timeline

What's the Facebook data engineer interview process and timeline? It usually takes around two months and follows these steps:

1.1 What interviews to expect

1. Application and referrals
2. Recruiter phone screen
3. Technical screening
4. Onsite interviews

Let's take a look at each of those steps in a bit more detail.

1.1.1 Application and referrals

Step one is getting a Facebook interview in the first place. In this guide we're focusing primarily on the interviews, so we'll keep this portion brief. You can apply to Facebook directly or a recruiter may reach out to you. In either case, it helps to have a quality (and up-to-date) resume that is tailored to front end positions, and to Facebook more specifically.

If you do have a connection to someone in Facebook, it can be really helpful to get an employee referral to the internal recruiting team, as it may increase your chances of getting into the interview process.

1.1.2 Recruiter phone screen

In most cases, you'll start your interview process with Facebook by talking to an HR recruiter on the phone for 30-45 minutes. They are looking to confirm that you've got a chance of getting the job at all, so be prepared to explain your background and why you're a good fit at Facebook. You should expect typical behavioral and resume questions like, "Tell me about yourself", "[Why Facebook?](#)", as well as some SQL and data structure questions.

If you get past this first HR screen, the recruiter will then help schedule a technical screen with a Facebook engineer. One great thing about Facebook is that they are very transparent about their recruiting process. Your HR contact will therefore walk you through the remaining steps in the hiring process, and will also share with you a helpful email listing resources you can use to prepare.

1.1.3 Technical screening

If you make it past the HR screen you'll have an hour-long technical screening where you'll spend half the time on 5 SQL questions and the other half on 5 coding questions.

To provide your answers you'll use a simple online code editor without syntax highlighting or auto-completion (e.g. CoderPad) and it's a good idea to get used to using one of them beforehand. It's up to you which programming language you use.

1.1.4 Onsite interviews

If you make it past the technical screening you'll be invited to the real test: a full day of "[onsite](#)" interviews at a Facebook office. As you'll see in this really useful [Facebook guide](#), you can expect four interviews in total: three hour-long technical interviews and one 30-minute "ownership" interview. The questions you'll face can be grouped into 5 categories:

1. [Coding questions](#)
2. [SQL questions](#)
3. [Data modeling questions](#)
4. [Product sense questions](#)
5. [Ownership questions](#)

In addition to these interviews, you'll also have lunch with a fellow engineer while you are onsite. The lunch interview is meant to be your time to ask questions about what it's like to work at Facebook. The company won't be evaluating you during this time, but we recommend that you behave as if they were.

[COVID note] It's likely that your onsite interviews will be held [virtually](#) instead of in-person, given the COVID-19 pandemic. However, your recruiter should be able to provide you with the most up-to-date information on Facebook's onsite interview procedures. Feel free to ask your Facebook recruiter for details after you've been officially invited to participate in the onsite interviews.

1.2 Individual contributors vs managers

Facebook has two career tracks. You can either grow into a [manager](#) where you end up leading teams of engineers (management track). Or you can stay very hands-on technically and specialize as you become more senior (individual contributor track).

If you're interviewing as an individual contributor then, as we've discussed regarding a data engineer role, you can mainly expect technical interviews and will typically only have a single behavioral (ownership) interview.

However, if you are interviewing as a manager, director, or above, then you should expect at least two interviews that are not technical where you need to answer behavioral questions about how you develop people, work with cross functional teams, execute on projects, grow an organization, etc.

1.3 What happens behind the scenes

Throughout the interview process at Facebook, the recruiter usually plays the role of "facilitator" and moves the process from one stage to the next. Here's an overview of what typically happens behind the scenes:

- **After the technical screen**, the interviewer you've talked to will have 24h to submit their ratings and notes to the internal system. Your recruiter then reviews the feedback, and decides to move you to the onsite interview or not depending on how well you've done.
- **After the onsite**, your interviewers will make a recommendation on hiring you or not and the recruiter compiles your "packet" (interview feedback, resume, referrals, etc.). If they think you can get the job, they will present your case at the next candidate review meeting.
- **Candidate review meetings** are used to assess all candidates who have recently finished their interview loops and are close to getting an offer. Your packet will be analyzed and possible concerns will be discussed. Your interviewers are invited to join your candidate review meeting, but will usually only attend if there's a strong disagreement in the grades you received (e.g. 2 no hires, 2 hires). If after discussion the team still can't agree whether you should get an offer or not, you might be asked to do a follow up interview to settle the debate. At the end of the candidate review meeting, a hire / no hire recommendation is made for consideration by the hiring committee.
- **The hiring committee** includes senior leaders from across Facebook. This step is usually a formality and the committee follows the recommendation of the candidate review meeting. The main focus is on fine-tuning the exact level and therefore the compensation you will be offered.

It's also important to note that hiring managers and people who refer you have little influence on the overall process. They can help you get an interview at the beginning, but that's about it.

2. Example questions

As we mentioned above, for the position of Facebook data engineer you'll face questions across five different topics; four technical and one behavioral.

1. [Coding questions](#)
2. [SQL questions](#)
3. [Data modeling questions](#)
4. [Product sense questions](#)
5. [Ownership questions](#)

Now let's dig a little deeper into those topics. We've analyzed scores of candidates' interview reports on [Glassdoor](#) to provide you with real interview questions that have been previously asked by Facebook. We've categorized them and we've changed the grammar and phrasing in some places to make the questions easier to understand.

Let's get into it.

2.1 Coding questions

Data engineers at Facebook solve some of the company's biggest challenges with code. Facebook will want to make sure you've got the necessary problem-solving skills and that you can think in a structured way when it comes to code.

You'll need to know how to manipulate data structures, how to use dictionaries, loops and lists, and show a good understanding of string, set operations, etc, in your coding language of choice. You'll need to provide accurate, bug-free, efficient, and well-thought-out code, sharing your thoughts out loud while you work through it on a whiteboard (or online equivalent).

Let's take a look at some real examples of coding questions that we found in the Glassdoor data.

Example coding questions asked by Facebook in data engineer interviews

- Given a two dimensional list, for example [[2,3],[3,4],[5]] person 2 is friends with 3 etc, find how many friends each person has. Note, one person has no friends
- Can you do the following without using subquery?: {1,None,1,2,None} --> [1,1,1,2,2] Ensure you take care of case input[None] which means None object
- Complete a function that returns a list containing all the mismatched words (case sensitive) between two given input strings # For example: # - string 1 : "Firstly this is the first string" # - string 2 : "Next is the second string" # # - output : ['Firstly', 'this', 'first', 'Next', 'second']
- Complete a function that returns the number of times a given character occurs in the given string.
For example:
- input string = "mississippi"
- char = "s"

- output : 4
- Given an array containing None values fill in the None values with most recent non None value in the array. For example:input array: [1,None,2,3,None,None,5,None] # - output array: [1,1,2,3,3,3,5,5]

- Complete a function that returns a list containing all the mismatched words (case sensitive) between two given input strings # For example: string 1 : "Firstly this is the first string" # - string 2 : "Next is the second string" # # - output : ['Firstly', 'this', 'first', 'Next', 'second']
- Given an array of integers, we would like to determine whether the array is monotonic (non-decreasing/non-increasing) or not. Examples:

```
// 1 2 5 5 8
```

```
// true
```

```
// 9 4 4 2 2
```

```
// true
```

```
// 1 4 6 3
```

```
// false
```

```
// 1 1 1 1 1 1
```

```
// true
```

- Given two sentences, construct an array that has the words that appear in one sentence and not the other
- Given an ip address as an input string, validate it and return True/False
- Count the neighbors of each node in a graph. Input graph is a multi dimensional list
- Given a dictionary, print the key for nth highest value present in the dict. If there are more than 1 record present for nth highest value then sort the key and print the first one
- Flatten a nested dictionary
- You have a 2-D array of friends like [[A,B],[A,C],[B,D],[B,C],[R,M], [S],[P], [A]]. Write a function that creates a dictionary of how many friends each person has. People can have 0 to many friends. However, there won't be repeat relationships like [A,B] and [B,A] and neither will there be more than 2 people in a relationship
- What is a loop that goes on forever?
- Recursively parse a string for a pattern that can be either 1 or 2 characters long
- Write a simple spell-checking engine
- Given two sentences, you have to print the words those are not present in either of the sentences.(If one word is present twice in 1st sentence but not present in 2nd sentence then you have to print that word too)

To see more example coding problems asked by Facebook in engineering interviews, check out our [Facebook software engineer interview guide](#). You'll see that those examples come with links to Leetcode solutions.

2.2 SQL questions

With Facebook's backend handling billions of data fetch operations each day, knowledge of SQL is very important for the company's data engineers. That's why they include SQL questions both in the technical screening and the onsite interviews.

The interviewer will test your knowledge of basic SQL constructs. You'll need to know how to use joins, aggregate functions, analytical functions, set operators, and subqueries, while thinking about efficiency and scalability.

Let's take a look at some real examples of SQL questions that we found in the Glassdoor data.

Example SQL questions asked by Facebook in data engineer interviews

- How do you join two tables with all the information on the left one unchanged?
- Does database view occupy the disk space?
- Given full authority to ""make it work"", import a large data set with duplicates into a warehouse while meeting the requirements of a business intelligence designer for query speed
- Find the top 10 colleges/companies that an averagely social person interacts with
- The ORDER BY command in SQL is automatically set in what format if you didn't set it: Ascending or Descending?
- When you want to delete or add a column of a table in a database, what command will you use?
- What command would you want to use if you want to keep all the info of the left table?
- You want to combine two columns after removing two duplicates, would you use UNION or UNION ALL?
- What is the term used to select non-duplicates in SQL?
- Given a raw data table, how would you write the SQL to perform the ETL to get data into a desired format?
- Perform a merge-sort with SQL only
- A table has two data entries every day for # of apples and oranges sold. Write a query to get the difference between the apples and oranges sold on a given day
- Given a database schema showing product sales: calculate what percent of our sales transactions had a valid promotion applied? And what % of sales happened on the first and last day of the promotion?
- Display the most common name in a table

2.3 Data modeling questions

Facebook is an organization built on collecting huge amounts of data. Modeling that data in the best way possible is therefore very important.

You can expect to be asked to brainstorm the data needs of one of Facebook's products. You'll then be asked to design a data mart to support analytics use cases and to write select SQL statements to produce specific results.

Let's take a look at some real examples of data modeling questions that we found in the Glassdoor data.

Example data modeling questions asked by Facebook in data engineer interviews

- Present a design of a gaming company data
- Design a database for an app
- Design a relational database for a ride-sharing app

- Given data, design a table schema for this data to be used by a data scientist to query metrics such as process with max average elapsed time, and so they can plot each process
- Create DDL (table and foreign keys) for several tables in a provided ERD. The ERD contains at least one many-to-many relationship

2.4 Product sense questions

Data engineers play an important role in Facebook's product development strategy. They therefore need to have a strong product awareness and the ability to hold strategic conversations about a product and its possibilities.

All three technical interviews will be case studies of typical product challenges that Facebook normally solves with data. You'll need to show that you can think critically about the needs of a product and give solid technical solutions.

Let's take a look at some real examples of product sense questions that we found in the Glassdoor data.

Example product sense questions asked by Facebook in data engineer interviews

- Design a dashboard to highlight a certain aspect of the user behaviour
- How do you calculate unique logins by a user on facebook.com?
- How would you rate the popularity of a video posted online?
- How would you check if Facebook should change something in the newsfeed? How would you define the KPI in this case?
- Design an experiment to test whether a certain feature generates conversation

2.5 Ownership questions

Facebook data engineers need more than just strong technical skills - they need the soft skills that will enable them to leverage this technical knowledge by taking initiative and influencing their fellow engineers and cross-functional partners.

The ownership questions that you'll be asked are what we call "behavioral" questions. They will aim to assess you on your past behavior, looking to see if you have demonstrated good leadership, communication and teamwork skills. To dig deeper in behavioral questions, check out our guide on [how to answer them](#), as well as our take on the ["Why Facebook?"](#) question.

Below we've listed some frequent ownership questions that Facebook tends to ask, according to data from Glassdoor.

Example ownership questions asked by Facebook in data engineer interviews

- Tell me about yourself
- Tell me about a challenge you faced and how you overcame it?
- Why data engineering?

- [Why Facebook?](#)
- Tell me about a project you have worked on
- Describe a situation where you did not agree with the stakeholders. How did you handle it?
- What product would you want to most work on and why? What would you do if you worked on that project?

3. How to prepare

Now that you know what questions to expect, let's focus on how to prepare. It's no secret that the performance bar at Facebook is high. Some people even go as far as [quitting their job](#) to prepare for interviews full time.

This is obviously extreme and not what we recommend doing, but it shows how much effort some candidates are ready to put in. We've listed the four steps we recommend taking to prepare as efficiently as possible below.

3.1 Learn about Facebook's culture

Most candidates fail to do this. But before investing tens of hours preparing for an interview at Facebook, you should take some time to make sure it's actually the right company for you.

Facebook is prestigious and it's therefore tempting to ignore that step completely. But in our experience, the prestige in itself won't make you happy day-to-day. It's the type of work and the people you work with that will.

If you know engineers who work at Facebook (or used to), it's a good idea to talk to them to understand what the culture is like. In addition, we would recommend reading about Facebook's [5 core values](#) and [hacker culture](#).

3.2 Practice by yourself

To help you further organise your preparation, make sure to read [Facebook's own guide](#) to the data engineer onsite interview. It's a great overview of what to expect and contains some invaluable tips. Then you'll want to start digging a bit deeper into each of the five types of questions you'll be facing (coding, SQL, data modeling, product sense, and ownership) and start to practice answering them by yourself.

For coding questions, we recommend reading the following [article](#) written by an ex-Facebook interviewer to understand more about the step-by-step approach you should use to solve coding questions in an interview.

And to practice, we recommend using [Leetcode](#), where you can get a lot done with the free tier, and you can also access Facebook-specific questions using the premium tier. If you haven't already, you should also take a look at our [Facebook software engineer interview](#) article for useful example coding questions and additional information.

For SQL questions, you can also find a lot of practice questions on [Leetcode](#) or on [this website](#) recommended by Facebook. We also recommend reading this article on the [3 types of SQL questions](#).

For data modeling questions, [learndatamodeling.com](#) has some useful videos that explain some of the core concepts. To practice, take a main product from a top tech company and try and work out how you would model each of its functions. Create logging designs and design data models. This kind of preparation can go a long way in helping you make a strong impression.

For product sense questions, [this video](#) offers an easy-to-understand overview and tackles some example questions. It's aimed at data scientists but is also relevant to data engineers. In addition, we recommend having a close look at Facebook products and thinking hard about their metrics.

For ownership questions, we recommend learning [our step-by-step method](#) for answering behavioral questions. In addition, you'll want to write down your answers to the example questions we gave you in the previous section.

Finally, a great way to practice answering interview questions is to interview yourself out loud. This may sound strange, but it will significantly improve the way you communicate your answers during an interview. Play the role of both the candidate and the interviewer, asking questions and answering them, just like two people would in an interview. Trust us, it really helps!

3.3 Practice with peers

Practicing by yourself will only take you so far. One of the main challenges of coding interviews is to have to communicate what you are doing as you are doing it. As a result, we strongly recommend practicing live interviews with a peer interviewing you.

A great place to start is to practice with friends if you can. You can also sign up to the software engineer waitlist on our [free mock interview platform](#). We'll let you know as soon as we've activated the software engineer category (which includes data engineering) so that you can start practicing with other candidates.

3.4 Practice with ex-interviewers

The main benefit of practicing with peers is that it's free. But at some point you'll start noticing that the feedback you are getting from peers isn't helping you that much anymore. Once you reach that stage, we recommend practicing with ex-interviewers from top tech companies.

If you know a data engineer who has experience running interviews at Facebook or another big tech company, then that's fantastic. But for most of us, it's tough to find the right connections to make this happen. And it might also be difficult to practice multiple hours with that person unless you know them really well.

Here's the good news. We've already made the connections for you. We've created a coaching service where you can practice 1-on-1 with ex-interviewers from leading tech companies like Facebook. [Learn more and start scheduling sessions today.](#)

Data engineer interview questions are a major component of your interview preparation process. However, if you want to maximize your chances of [landing a data engineer job](#), you must also be aware of how the data engineer interview process is going to unfold.

This article is designed to help you navigate the data engineer interview landscape with confidence. Here's what you will learn:

- the most important skills required for a data engineer position;
- a list of real data engineer questions and answers (practice makes perfect, right?);
- how the data engineer interview process goes down in 3 leading companies.

As a bonus, we'll reveal 3 common mistakes you should avoid at all costs during your data engineer interview questions preparation.

But first things first...

What skills do you need to become a data engineer?

Skills and qualifications are the most crucial part of your preparation for a data engineer position. Here are the top 5 must-have skills for anyone aiming for a data engineer career:

- Knowledge of data modeling for both data warehousing and Big Data;
- Experience in ETLs;
- Experience in the Big Data space (Hadoop Stack like M/R, HDFS, Pig, Hive, etc.);
- SQL and Python;
- Mathematics;
- Data visualization skills (e.g., Tableau or PowerBI).

If you need to improve your skillset to launch a successful career as a data engineer, you can [register for the complete 365 Data Science Program](#) today. Start with the fundamentals with our Statistics, Maths, and Excel courses, and build up step-by-step experience with SQL, Python, R, Power BI and Tableau.

What are the most common data engineer interview questions you should be familiar with?

General Data Engineer Interview Questions

Usually, interviewers start the conversation with a few more general questions. Their aim is to take the edge off and prepare you for the more complex data engineering questions ahead. Here are a few that will help you get off to a flying start.

1. How did you choose a career in data engineering?

How to answer

The answer to this question helps the interviewer learn more about your education, background and work experience. You might have chosen the data engineering field as a natural continuation of your degree in Computer Science or Information Systems. Maybe you've had similar jobs before, or you're transitioning from an entirely different career field. In any case, don't shy away from sharing your story and highlighting the skills you've gained throughout your studies and professional path.

Answer Example

"Ever since I was a child, I have always had a keen interest in computers. When I reached senior year in high school, I already knew I wanted to pursue a degree in Information Systems. While in college, I took some math and statistics courses which helped me land my first job as a Data Analyst for a large healthcare company. However, as much as I liked applying my math and statistical knowledge, I wanted to develop more of my programming and data management skills. That's when I started looking into data engineering. I talked to experts in the field and took online courses to learn more about it. I discovered it was the ideal career path for my combination of interests and skills. Luckily, within a couple of months, a data engineering position opened up in my company and I had the chance to transfer without a problem."

2. What do you think is the hardest aspect of being a data engineer?

How to answer

Smart hiring managers know not all aspects of a job are easy. So, don't hesitate to answer this question honestly. You might think its goal is to make you pinpoint a weakness. But, in fact, what the interviewer wants to know is how you managed to resolve something you struggled with.

Answer Example

"As a data engineer, I've mostly struggled with fulfilling the needs of all the departments within the company. Different departments often have conflicting demands. So, balancing them with the capabilities of the company's infrastructure has been quite challenging. Nevertheless, this has been a valuable learning experience for me, as it's given me the chance to learn how these departments work and their role in the overall structure of the company."

3. Can you think of a time where you experienced an unexpected problem with bringing together data from different sources? How did you eventually solve it?

How to answer

This question gives you the perfect opportunity to demonstrate your problem-solving skills and how you respond to sudden changes of the plan. The question could be data-engineer specific, or a more general one about handling challenges. Even if you don't have particular experience, you can still give a satisfactory hypothetical answer.

Answer Example

“In my previous work experience, my team and I have always tried to be ready for any issues that may arise during the ETL process. Nevertheless, every once in a while, a problem will occur completely out of the blue. I remember when that happened while I was working for a franchise company. Its system required for data to be collected from various systems and locations. So, when one of the franchises changed their system without prior notification, this created quite a few loading issues for their store's data. To deal with this issue, first I came up with a short-term solution to get the essential data into the company's corporate wide-reporting system. Once I took care of that, I started developing a long-term solution to prevent such complications from happening again.”

4. Data engineers collaborate with data architects on a daily basis. What makes your job as a data engineer different?

How to Answer

With this question, the interviewer is most probably trying to see if you understand how job roles differ within a data warehouse team. However, there is no “right” or “wrong” answer to this question. The responsibilities of both data engineer and data architects vary (or overlap) depending on the requirements of the company/database maintenance department you work for.

Answer Example

“Based on my work experience, the differences between the two job roles vary from company to company. Yes, it's true that data engineers and data architects work closely together. Still, their general responsibilities differ. Data architects are in charge of building the data architecture of the company's data systems and managing the servers. They see the full picture when it comes to the dissemination of data throughout the company. In contrast, data engineers focus on testing and maintaining of the architecture, rather than on building it. Plus, they make sure that the data available to analysts within the organization is reliable and of the necessary high quality.”

5. Can you tell us a bit more about the data engineer certifications you have earned?

How to Answer

Certifications prove to your future employer that you've invested time and effort to get formal training for a skill, rather than just pick it up on the job. The number of certificates under your belt also shows how dedicated you are to expanding your knowledge and skillset. Recency is also important, as technology in this field is rapidly evolving, and upgrading your skills on a regular basis is vital. However, if you haven't completed any courses or online certificate programs, you can mention the trainings provided by past employers or the current company you work for. This will indicate that you're up-to-date with the latest advancements in the data engineering sphere.

Answer Example

"Over the past couple of years, I've become a certified Google Professional Data Engineer, and I've also earned a Cloudera Certified Professional credential as a Data Engineer. I'm always keeping up-to-date with new trainings in the field. I believe that's the only way to constantly increase my knowledge and upgrade my skillset. Right now, I'm preparing for the IBM Big Data Engineer Certificate Exam. In the meantime, I try to attend big data conferences with recognized speakers, whenever I have the chance."

Technical Data Engineer Interview Questions

The technical data engineer questions help the interviewer assess 2 things: whether you have the skills necessary for the role; and if you're experienced with (or willing to advance in) the systems and programs utilized in the company. So, here's a list of technical questions you can practice with.

6. Which ETL tools have you worked with? Do you have a favorite one? If so, why?

How to Answer

The hiring manager needs to know that you're no stranger to the ETL process and you have some experience with different ETL tools. So, once you enumerate the tools you've worked with and point out the one you favor, make sure to substantiate your preference in a way that demonstrates your expertise in the ETL process.

Answer Example

"I have experience with various ETL tools, such as IBM Infosphere, SAS Data Management, and SAP Data Services. However, if I have to pick one as my favorite, that would be Informatica's PowerCenter. In my opinion, what makes it the best out there is its efficiency. PowerCenter has a very top performance rate and high flexibility which, I believe, are the most important properties of an ETL tool. They guarantee access to the data and smoothly running business data operations at all times, even if changes in the business or its structure take place."

7. Have you built data systems using the Hadoop framework? If so, please describe a particular project you've worked on.

How to Answer

Hadoop is a tool that many hiring managers ask about during interviews. You should know that whenever there's a specific question like that, it's highly likely that you'll be required to use this particular tool on the job. So, to prepare, do your homework and make sure you're familiar with the languages and tools the company uses. More often than not, you can find that information in the job description. If you're experienced with the tool, give a detailed explanation of your project to highlight your skills and knowledge of the tool's capabilities. In case you haven't worked with this tool, the least you could do is do some research to demonstrate some basic familiarity with the tool's attributes.

Answer Example

"I've used the Hadoop framework while working on a team project focused on increasing data processing efficiency. We chose to implement it because of its ability to increase data processing speeds while, at the same time, preserving quality through its distributed processing. We also decided to implement Hadoop because of its scalability, as the company I worked for expected a considerable increase in its data processing needs over the next few months. In addition, Hadoop is an open-source network which made it the best option, keeping in mind the limited resources for the project. Not to mention that it's Java-based, so it was easy to use by everyone on the team and no additional training was required."

8. Do you have experience with a cloud computing environment? What are the pros and cons of working in one?

How to Answer

Data engineers are well aware that there are pros and cons to cloud computing. That said, even if you lack prior experience working in cloud computing, you must be able to demonstrate a certain level of understanding of its advantages and shortcomings. This will show the hiring manager that you're aware of the present technological issues in the industry. Plus, if the position you're interviewing for requires using a cloud computing environment, the hiring manager will know that you've got a basic idea of the possible challenges you might face.

Answer Example

"I haven't had the chance to work in a cloud computing environment yet. However, I have a good overall idea of its pros and cons. On the plus side, cloud computing is more cost-effective and reliable. Most providers sign agreements that guarantee a high level of service availability which should decrease downtimes to a minimum. On the negative side, the cloud computing environment may compromise data security and privacy, as the data is kept outside the company. Moreover, your control would be limited, as the infrastructure is managed by the service

provider. All things considered, cloud computing could be both right or wrong choice for a company, depending on its IT department structure and the resources at hand.”

9. In your line of work, have you introduced new data analytics applications? If so, what challenges did you face while introducing and implementing them?

How to Answer

New data applications are high-priced, so introducing such within a company doesn’t happen that often. Nevertheless, when a company decides to invest in new data analytics tools, this could turn into quite an ambitious project. The new tools must be connected to the current systems in the company, and the employers who are going to use them should be formally trained. Additionally, maintenance of the tools should be administered and carried out on a regular basis. So, if you have prior experience, point out the obstacles you’ve overcome or list some scenarios of what could have gone wrong. In case you lack relevant experience, describe what you know about the process in detail. This will let the hiring manager know that, if a problem arises, you have the basic know-how that would help you through.

Answer Example

“As a data engineer, I’ve taken part in the introduction of a brand-new data analytics application in the last company I’ve worked for. The whole process requires a well-thought-out plan to ensure the smoothest transition possible. However, even the most careful planning can’t rule out unforeseen issues. One of them was the high demand for user licenses which went beyond our expectations. The company had to reallocate financial resources to obtain additional licenses. Furthermore, training schedules had to be set up in a way that doesn’t interrupt the workflow in different departments. In addition, we had to optimize our infrastructure, so that it could support the considerably higher number of users.”

10. What is your experience level with NoSQL databases? Tell me about a situation where building a NoSQL database was a better solution than building a relational database.

How to Answer

There are certain pros and cons of using one type of database compared to another. To give the best possible answer, try to showcase your knowledge about each and back it up with an example situation that demonstrates how you have applied (or would apply) your know-how to a real-world project.

Answer Example

“Building a NoSQL database can be beneficial in some situations. Here’s a situation from my experience that first comes to my mind. When the franchise system in the company I worked for was increasing in size exponentially, we had to be able to scale up quickly in order to make the most of all the sales and operational data we had on hand.

But here's the thing. Scaling out is the better option, compared to scaling up with bigger servers, when it comes to handling increases data processing loads. Scaling out is also more cost-effective and it's easier to accomplish through NoSQL databases. The latter can deal with larger volumes of data. And that can be crucial when you need to respond quickly to considerable shifts in data loads in the future. Yes, it's true that relational databases have better connectivity to various analytics tools. However, as more of those are being developed, there's definitely a lot more coming from NoSQL databases in the future. That said, the additional training some developers might need is certainly worth it."

By the way, if you're finding this answer useful, consider sharing this article, so others can benefit from it, too. Helping fellow aspiring data engineers reach their goals is one of the things that make the data science community special.

11. What's your experience with data modeling? What data modeling tools have you used in your work experience?

How to Answer

As a data engineer, you probably have some experience with data modeling. In your answer, try not only to list the relevant tools you have worked with, but also mention their pros and cons. This question also gives you a chance to highlight your knowledge of data modeling in general.

Answer Example

"I've always done my best to be familiar with the data models in the companies I've worked for, regardless of my involvement with the data modeling process. This is one of the ways I gain a deeper understanding of the whole system. In my work experience, I've utilized Oracle SQL Developer Data Modeler to develop two types of models. Conceptual models for our work with stakeholders, and logical data models which make it possible to define data models, structures and relationships within the database."

Behavioral Data Engineer Questions

Behavioral data engineer interview questions give the interviewer a chance to see how you have handled unforeseen data engineering issues or teamwork challenges in your experience. The answers you provide should reassure your future employer that you can deal with high-pressure situations and a variety of challenges. Here are a few examples to consider in your preparation.

12. Data maintenance is one of the routine responsibilities of a data engineer. Describe a time when you encountered an unexpected data maintenance problem that made you search for an out-of-the-box solution".

How to Answer

Usually, data maintenance is scheduled and covers a particular task list. Therefore, when everything is operating according to plan, the tasks don't change as often. However, it's

inevitable that an unexpected issue arises every once in a while. As this might cause uncertainty on your end, the hiring manager would like to know how you would deal with such high-pressure situations.

Answer Example

“It’s true that data maintenance may come off as routine. But, in my opinion, it’s always a good idea to closely monitor the specified tasks. And that includes making sure the scripts are executed successfully. Once, while I was conducting an integrity check, I located a corrupt index that could have caused some serious problems in the future. This prompted me to come up with a new maintenance task that prevents corrupt indexes from being added to the company’s databases.”

13. Data engineers generally work “backstage”. Do you feel comfortable with that or do you prefer being in the “spotlight”?

How to Answer

The reason why data engineers mostly work “backstage” is that making data available comes much earlier in the data analysis project timeline. That said, c-level executives in the company are usually more interested in the later stages of the work process. More specifically, their goal is to understand the insights that data scientists extract from the data via statistical and machine learning models. So, your answer to this question will tell the hiring manager if you’re only able to work in the spotlight, or if you thrive in both situations.

Answer Example

“As a data engineer, I realize that I do most of my work away from the spotlight. But that has never been that important to me. I believe what matters is my expertise in the field and how it helps the company reach its goals. However, I’m pretty comfortable being in the spotlight whenever I need to be. For example, if there’s a problem in my department which needs to be addressed by the company executives, I won’t hesitate to bring their attention to it. I think that’s how I can further improve my team’s work and reach better results for the company.”

14. Do you have experience as a trainer in software, applications, processes or architecture? If so, what do you consider as the most challenging part?

How to Answer

As a data engineer, you may often be required to train your co-workers on the new processes or systems you’ve created. Or you may have to train new teammates on the already existing architectures and pipelines. As technology is constantly evolving, you might even have to perform recurring trainings to keep everyone on track. That said, when you talk about a challenge you’ve faced, make sure you let the interviewer know how you handled it.

Answer Example

“Yes, I have experience training both small and large groups of co-workers. I think the most challenging part is to train new employees who already have significant experience in another company. Usually, they’re used to approaching data from an entirely different perspective. And that’s a problem because they struggle to accept the way we handle projects in our company. They’re often very opinionated and it takes time for them to realize there’s more than one solution to a certain problem. However, what usually helps is emphasizing how successful our processes and architecture have proven to be so far. That encourages them to open their minds to the alternative possibilities out there.”

15. Have you ever proposed changes to improve data reliability and quality? Were they eventually implemented? If not, why not?

How to Answer

One of the things hiring managers value most is constant improvements of the existing environment, especially if you initiate those improvements yourself, as opposed to being assigned to do it. So, if you’re a self-starter, definitely point this out. This will showcase your ability to think creatively and the importance you place on the overall company’s success. If you lack such experience, explain what changes you would propose as a data engineer. In case your ideas were not implemented for reasons such as lack of financial resources, you can mention that. However, try to focus on your continuous efforts to find novel ways to improve data quality.

Answer Example

“Data quality and reliability have always been a top priority in my work. While working on a specific project, I discovered some discrepancies and outliers in the data stored in the company’s database. Once I’ve identified several of those, I proposed to develop and implement a data quality process in my department’s routine. This included bi-weekly meetups with coworkers from different departments where we would identify and troubleshoot data issues. At first, everyone was worried that this would take too much time off their current projects. However, in time, it turned out it was worth it. The new process prevented the occurrence of larger (and more costly) issues in the future.”

16. Have you ever played an active role in solving a business problem through the innovative use of existing data?

How to Answer

Hiring managers are looking for self-motivated people who are eager to contribute to the success of a project. Try to give an example where you came up with a project idea or you took charge of a project. It’s best if you point out what novel solution you proposed, instead of focusing on a detailed description of the problem you had to deal with.

Answer Example

“In the last company I worked for, I took active part in a project that aimed to identify the reason’s for the high employee turnover rate. I started by closely observing data from other areas of the company, such as Marketing, Finance, and Operations. This helped me find some high correlations of data in these key areas with employee turnover rates. Then, I collaborated with the analysts in those departments to gain a better understanding of the correlations in question. Ultimately, our efforts resulted in strategic changes that had a positive influence over the employee turnover rates.”

17. Which non-technical skills do you find most valuable in your role as a data engineer?

How to Answer

Although technical skills are of major importance if you want to advance your data engineer career, there are many non-engineering skills that could aid your success. In your answer, try to avoid the most obvious examples, such as communication or interpersonal skills.

Answer Example

“I’d say the most useful skills I’ve developed over the years are multitasking and prioritizing. As a data engineer, I have to prioritize or balance between various tasks daily. I work with many departments in the company, so I receive tons of different requests from my coworkers. To cope with those efficiently, I need to put fulfilling the most urgent company needs first without neglecting all the other requests. And strengthening the skills I mentioned has really helped me out.”

Brainteasers

Interviewers use brainteasers to test both your logical and creative thinking. These questions also help them assess how quickly you can resolve a task that requires an out-of-the-box approach.

18. You have eight balls of the same size. Seven of them weigh the same, and one of them weighs slightly more. How can you find the ball that is heavier by using a balance and only two attempts at weighing?

You can put six of the balls on the balance. If one of the sides is heavier you will know that the heavier ball is on that side. If not, the heavier ball is among the two that you did not measure and it will be really easy to determine precisely which ball is heavier with your second weighing.

After you determine which side is heavier, you will have 3 balls left to choose from. You have another attempt at weighing left. You can put two of the balls on the balance and see if one of them is heavier. If it is, then you have found the heavier ball. If it is not, then the third ball is the one that is heavier.

19. A windowless room has three light bulbs. You are outside the room with 3 switches, each of them controlling one of the light bulbs. If you were told that you can enter the room only once, how are you going to tell which switch controls which light bulb?

You have to be creative in order to solve this one. You switch on two of the light bulbs and then wait for 30 minutes. Then you switch off one of them and enter the room. You will know which switch controls the light bulb that is on. Here is the tough part. How are you going to be able to determine which switch corresponds to the other two light bulbs? You will have to touch them. Yes. That's right. Touch them and feel which one is warm. That will be the other bulb that you had turned on for 30 minutes.

You will be in serious trouble if the interviewer says that the light bulbs are LED (given that they don't emit heat).

Guesstimate

Although guesstimates aren't an obligatory part of the data engineer interview process, many interviewers would ask such a question to assess your quantitative reasoning and approach to solving complex problems. Here's a good example.

20. How many gallons of white house paint are sold in the US every year?

Find the number of homes in the US: Assuming that there are 300 million people in the US and the average household contains 2.5 people then we can conclude that there are 120 million homes in the US.

Number of houses: Many people live in apartments and other types of buildings different than houses. Let's assume that the percentage of people living in houses is 50%. Hence, there are 60 million houses.

Houses that are painted in white: Although white is the most popular color, many people choose different paint colors for their houses or do not need to paint them (using other types of techniques in order to cover the external surface of the house). Let's hypothesize that 30% of all houses are painted in white, which makes 18 million houses that are painted in white.

Repainting: People need to repaint their houses after a given amount of years. For the purposes of this exercise, let's hypothesize that people repaint their houses once every 9 years, which means that every year 2 million houses are repainted in white.

I have never painted a house, but let's assume that in order to repaint a house you need 30 gallons of white paint. This means the total US market for white house paint is 60 million gallons.

What is the data engineer interview process like?

A phone screen with a recruiter or a team member? How many onsite interviews you should be ready for? Will there be one or multiple interviewers?

Short answer: It depends on the company, its hiring policy and interviewing approach.

That said, here is what you can expect from a data engineer job interview at three top companies – Yahoo, Facebook, and Walmart. We believe these overviews will give you a good initial idea of what happens behind the scenes.

Yahoo

Generally, Yahoo recruit candidates from the top 10-20 schools. However, you can still get a data engineer interview through large job search platforms, such as Indeed.com and Glassdoor. Or, if you are lucky enough – with an internal referral. Anyhow, once you make the cut, you can expect a phone screen with a manager or a team lead. What about the onsite interviews? Usually, you'll interview with 6-7 data engineer team members for about 45 minutes each. Each interview will focus on a different area, but all of them have a similar structure. A short general talk (5 minutes), followed by a coding question (20 minutes) and a data engineering question (20 minutes). The latter will often tap into your previous experience to solve a current data engineering issue the company is experiencing.

In the end, you'll have a more general talk with a senior employee. At the same time, the interviewers will gather to share their feedback on your performance and check in with the hiring manager. If you've passed the data engineer interview with flying colors, you could get a decision on the day of the interview! However, if a few days have passed and you haven't received an answer, don't be shy to send HR a polite update request.

Facebook

Usually, the data engineering interviewing process starts with an email or a phone call with a recruiter, followed by a phone screen or an in-person interview. The screening interview is conducted by a coworker and takes about 1 hour. It consists of SQL questions and online test coding tasks that you have to solve through a collaborative editor (CoderPad) in a programming language of your choice. Also, prepare to answer questions related to your resume, skills, interests, and motivation. If those go well, they'll invite you to a longer series of interviews at the Facebook office - 5 hours of in-person interviews, including a 1-hour lunch interview.

Three of the onsite interviews are focused on problem-solving. You'll be questioned about data engineering issues that the company is facing and how you can help them solve them, for example, how to identify the metrics for performance for this specific feature) and you will be expected to write SQL and actual code for the context of the problem itself. There is also a behavioral interview portion, asking you about your work experience, and how you deal with interpersonal problems. Finally, there is an informal lunch conversation where you can ask about the work culture and other day-to-day questions.

What's typical of Facebook interviews is that many data engineer interview questions focus on a deep understanding of their product, so make sure you demonstrate both knowledge and genuine interest in the data engineer job.

Once the interviews are over, everyone you've interviewed with compare notes to decide if you'll be successful in the data engineer role. Then all left to do is wait for your recruiter to contact you with feedback from the interview. Or, if you haven't heard from a company rep within a week or so, take matters into your own hands and send a kind follow-up email.

Walmart

The data engineer interview process will usually start with a phone screen, followed by 4 technical interviews (expect some coding, big data, data modeling, and mathematics) and 1 lunch interview. More often than not, there is one more data engineer technical interview with a hiring manager (and guess what - it involves some more coding!). Anything specific to remember? Yes. Walmart has been utilizing huge amounts of big data, even before it was coined as "big". MapReduce, Hive, HDFS, and Spark are all used internally by their data science and data engineering teams. That said, a little bit of practice every day goes a long way. And, if you diligently prepare for some coding and big data questions, you have every chance of becoming a data engineer in the world's biggest retail corporation.

What common mistakes to avoid in your data engineer interview questions preparation?

We know that sometimes the devil's in the details. And we wouldn't want you to miss a single detail that could cost you your success! So, here are 3 common mistakes you should definitely refrain from making:

Not practicing behavioral data engineer interview questions

Even if you have the technical part covered, that doesn't necessarily mean smooth sailing! Behavioral questions are becoming increasingly important, as they tell the interviewer more about your personality, how you handle conflicts and problematic work situations. So, remember to prepare for those by rehearsing some relevant stories from your past experience and getting familiar with the behavioral data engineer interview questions we've listed.

Skiping the mock interview

Are you so deep into your interview preparation process that you've cut all ties with the outside world? Big mistake! Snap out of it now, call a fellow data engineer and ask them to do a mock interview with you. Every interview has a performance side to it, and just imagining how you're going to act or sound wouldn't give you a realistic idea. So, while you're doing the mock interview, pay special attention to your body language and mannerisms, as well as to your tone of voice and pace of speech. You'll be amazed by the insight you're going to get!

Getting discouraged

There's one more thing you should remember about interviews. Once you pass the easier problems, you're bound to get to the harder data engineer interview questions. But no matter how difficult they seem, don't give up. Stay cool, calm, and collected, and don't hesitate to ask for guidance or additional explanations. If anything, this will prove two things: that you're not afraid of challenging situations; and you're willing to collaborate to find an efficient solution.

In Conclusion

Overview of SQL Interview Questions

The first step of analytics for most workflows involves quick slicing and dicing of data in SQL. That's why being able to write basic queries efficiently is a very important skill. Although many may think that SQL simply involves SELECTs and JOINs, there are many other operators and details involved for powerful SQL workflows. For example, utilizing subqueries is important and allows you to manipulate subsets of data by which later operations can be performed, while window functions allow you to cut data without combining rows explicitly using a GROUP BY. The questions asked within SQL are usually quite practical to the company at hand - a company like Facebook might ask about various user or app analytics question, whereas a company like Amazon will ask about products and transactions.

Overview of Databases Design Questions

Although it isn't explicitly necessary to know the inner workings of databases (which is typically more data engineering oriented), it helps to have a high level understanding of basic concepts in Databases and Systems. Databases refers not to specific ones but more so how they operate at a high level and what design decisions and trade-offs are made during construction and querying. "Systems" is a broad term but refers to any set of frameworks or tools by which analysis of large volumes of data relies on. For example, a common interview topic is the MapReduce framework which is heavily utilized at many companies for parallel processing of large datasets.

20 SQL Data Science Interview Questions

1. **[Robinhood - Easy]** Assume you are given the below tables for trades and users. Write a query to list the top 3 cities which had the highest number of completed orders.

users

<u>Aa</u> column_name	≡ type
<u>user_id</u>	integer
<u>city</u>	string
<u>email</u>	string
<u>signup_date</u>	datetime

2. **[Facebook - Easy]** Assume you have the below events table on app analytics. Write a query to get the click-through rate per app in 2019.

events

<u>Aa</u> column_name	≡ type
<u>app_id</u>	integer
<u>event_id</u>	string ("impression", "click")
<u>timestamp</u>	datetime

3. **[Uber - Easy]** Assume you are given the below table for spending activity by product type. Write a query to calculate the cumulative spend for each product over time in chronological order.

total_trans

<u>Aa</u> column_name	≡ type
<u>order_id</u>	integer
<u>user_id</u>	integer
<u>product_id</u>	string
<u>spend</u>	float
<u>date</u>	datetime

4. **[Snapchat - Easy]** Assume you have the below tables on sessions that users have, and a users table. Write a query to get the active user count of daily cohorts.

sessions

<u>Aa</u> column_name	≡ type
<u>user_id</u>	integer
<u>session_id</u>	integer
<u>date</u>	datetime

users + Add a view

<u>Aa</u> column_name	≡ type
<u>user_id</u>	integer
<u>email</u>	string
<u>date</u>	datetime

5. **[Facebook - Easy]** Assume you are given the below tables on users and user posts. Write a query to get the distribution of the number of posts per user.

users


<u>Aa</u> column_name	≡ type
<u>user_id</u>	integer
<u>date</u>	datetime

posts

<u>Aa</u> column_name	≡ type
<u>post_id</u>	integer
<u>user_id</u>	string
<u>body</u>	string
<u>date</u>	datetime


6. **[Amazon - Easy]** Assume you are given the below table on purchases from users. Write a query to get the number of people that purchased at least one product on multiple days.

purchases

<u>Aa</u> column_name	 type
<u>purchase_id</u>	integer
<u>user_id</u>	integer
<u>product_id</u>	integer
<u>quantity</u>	integer
<u>price</u>	float
<u>purchase_time</u>	datetime

7. **[Opendoor - Easy]** Assume you are given the below table on house prices from various zip codes that have been listed. Write a query to get the top 5 zip codes by market share of house prices for any zip code with at least 10000 houses.

housing

<u>Aa</u> column_name	 type
<u>house_id</u>	integer
<u>zip_code</u>	integer
<u>price</u>	float
<u>listing_date</u>	datetime

8. **[Etsy - Easy]** Assume you are given the below table on transactions from users for purchases. Write a query to get the list of customers where their earliest purchase was at least \$50.

user_transactions

<u>Aa</u> column_name	≡ type
transaction_id	integer
product_id	integer
user_id	integer
spend	float
transaction_date	datetime

9. **[Disney - Easy]** Assume you are given the below table on watch times (in minutes) for all users, where each user is based in a given city. Write a query to return all pairs of cities that have total watch times within 10000 minutes of one another.

watch_activity

<u>Aa</u> column_name	≡ type
user_id	integer
session_id	integer
watch_time	float
city_name	string
date	datetime

10. **[Twitter - Easy]** Assume you are given the below table on tweets by each user over a period of time. Calculate the 7-day rolling average of tweets by each user for every date.

tweets + Add a view

<u>Aa</u> column_name	≡ type
tweet_id	integer
msg	string
user_id	integer
tweet_date	datetime

11. **[Stitch Fix - Easy]** Assume you are given the below table on transactions from users. Write a query to get the number of users and total products bought per latest transaction date where each user is bucketed into their latest transaction date.

user_transactions

Aa column_name	≡ type
transaction_id	integer
product_id	integer
user_id	integer
spend	float
transaction_date	datetime


12. **[Amazon - Easy]** Assume you are given the below table on customer spend amounts on products in various categories. Calculate the top three most bought item within each category in 2020.

product_spend

Aa column_name	≡ type
transaction_id	integer
category_id	integer
product_id	integer
user_id	integer
spend	float
transaction_date	datetime

13. **[DoorDash - Easy]** Assume you are given the below table on transactions on delivery locations and times for meals - a start location, an end location, and timestamp for a given meal_id. Certain locations are aggregation locations - where meals get sent to, and where meals then go to their final destination. Calculate the delivery time per meal to each final destination from a particular aggregation location, loc_id = 4.

delivery_times

<u>Aa</u> column_name	 type
<u>meal_id</u>	integer
<u>start_loc_id</u>	integer
<u>end_loc_id</u>	integer
<u>cost</u>	float
<u>timestamp</u>	timestamp


14. **[Facebook - Medium]** Assume you have the below tables on user actions. Write a query to get the active user retention by month.

user_actions

<u>Aa</u> column_name	 type
<u>user_id</u>	integer
<u>event_id</u>	string ("sign-in", "like", "comment")
<u>timestamp</u>	datetime


15. **[Twitter - Medium]** Assume you are given the below tables for the session activity of users. Write a query to assign ranks to users by the total session duration for the different session types they have had between a start date (2020-01-01) and an end date (2020-02-01).

sessions

<u>Aa</u> column_name	 type
<u>session_id</u>	integer
<u>user_id</u>	integer
<u>session_type</u>	string
<u>duration</u>	integer
<u>start_time</u>	datetime

16. **[Snapchat - Medium]** Assume you are given the below tables on users and their time spent on sending and opening Snaps. Write a query to get the breakdown for each age breakdown of the percentage of time spent on sending versus opening snaps.

activities

<u>Aa</u> column_name	 type
<u>activity_id</u>	integer
<u>user_id</u>	integer
<u>type</u>	string ('send', 'open')
<u>time_spent</u>	float
<u>activity_date</u>	datetime

age_breakdown

<u>Aa</u> column_name	 type
<u>user_id</u>	integer
<u>age_bucket</u>	string

17. **[Google - Medium]** Assume you are given the below table on sessions from users, with a given start and end time. A session is concurrent with another session if they overlap in the start and end times. Write a query to output the session that is concurrent with the largest number of other sessions.

sessions

<u>Aa</u> column_name	 type
<u>session_id</u>	integer
<u>start_time</u>	datetime
<u>end_time</u>	datetime

18. **[Yelp - Medium]** Assume you are given the below table on reviews from users. Define a top-rated place as a business whose reviews only consist of 4 or 5 stars. Write a query to get the number and percentage of businesses that are top-rated places.

reviews

Aa column_name	≡ type
<u>business_id</u>	integer
<u>user_id</u>	integer
<u>review_text</u>	string
<u>review_stars</u>	integer
<u>review_date</u>	datetime

19. **[Google - Medium]** Assume you are given the below table of measurement values from a sensor for several days. Each measurement can happen several times in a given day. Write a query to output the sum of values for every odd measurement and the sum of values for every even measurement by date.

measurements

Aa column_name	≡ type
<u>measurement_id</u>	integer
<u>measurement_value</u>	float
<u>measurement_time</u>	datetime

20. **[Etsy - Medium]** Assume you are given the below table on transactions from various product search results from users on Etsy. For every given product keyword, there are multiple positions that being A/B tested, and user feedback is collected on the relevance of results (from 1-5). There are many displays for each position of every product, each of which is captured by a display_id. Define a highly relevant display as one whereby the corresponding relevance score is at least 4. Write a query to get all products having at least one position with > 80% highly relevant displays.

product_searches

Aa column_name	≡ type
product	string
position	integer
display_id	integer
relevance	integer
submit_time	datetime

10 Database And Systems Design Interview Questions

21. **[MongoDB - Easy]** For each of the ACID properties, give a one-sentence description of each property and why are these properties important?
22. **[VMWare - Easy]** What are the three major steps of database design? Describe each step.
23. **[Microsoft - Easy]** What are the requirements for a primary key?
24. **[DataStax - Easy]** A B+ tree can contain a maximum of 5 pointers in a node. What is the minimum number of keys in leaves?
25. **[Databricks - Easy]** Describe MapReduce and the operations involved.
26. **[Microsoft - Easy]** Name one major similarity and difference between a WHERE clause and a HAVING clause in SQL.
27. **[Facebook - Easy]** How does a trigger allow you to build business logic into a database?
28. **[DataStax - Easy]** What are the six levels of database security and briefly explain what each one entails?
29. **[Databricks - Easy]** Say you have a shuffle operator, whereby the input is a dataset and the output is simply a randomly ordered version of that dataset. Describe the algorithm steps in English.
30. **[Rubrik - Medium]** Describe what a cache is and what a block is. Say you have a fixed amount of total data storage - what are the some trade-offs in varying block size?

SQL And Database Interview Solutions

Problem #4 Solution:

By definition, daily cohorts are active users from a particular day. First, we can use a subquery to get the sessions of new users by day using an inner join with users. This is to filter for only active users by a particular join date for the cohort. Then we can get a distinct count to return the active user count:

```
WITH new_users_by_date AS (
```

```

SELECT sessions.*
FROM sessions
JOIN users on
  sessions.user_id = users.user_id
  sessions.date = users.date
)
SELECT date, COUNT(DISTINCT user_id) as active_user_count
FROM new_users_by_date
GROUP BY 1 ORDER BY 1 ASC

```

Problem #8 Solution:

Although we could use a self join on `transaction_date = MIN(transaction_date)` for each user, we can also use the `RANK()` window function to get the ordering of purchase by customer, and then use that subquery to filter on customers where the first purchase (rank one) is at least 50 dollars. Note that this requires the subquery to include `spend` as well

```

WITH purchase_rank AS (
  SELECT user_id, spend,
    RANK() OVER
      (PARTITION BY user_id ORDER BY transaction_date ASC) as rank
  FROM user_transactions u
)
SELECT
  user_id
FROM
  purchase_rank
WHERE rank = 1 AND spend >= 50.00

```

Problem #11 Solution:

First, we need to get the latest transaction date for each user, along with the number of products they have purchased. This can be done in a subquery where we `GROUP BY user_id` and take a `COUNT(DISTINCT product_id)` to get the number of products they have purchased, and a `MAX(transaction_date)` to get the latest transaction date (while casting to a date). Then, using this subquery, we can simply do an aggregation by the transaction date column in the previous subquery, while doing a `COUNT()` on the number of users, and a `SUM()` on the number of products:

```

WITH latest_date AS (
  SELECT user_id,
    COUNT(DISTINCT product_id) AS num_products,
    MAX(transaction_date::DATE) AS curr_date
  FROM user_transactions
  GROUP BY )
SELECT curr_date,
  COUNT(user_id) AS num_users,
  SUM(num_products) AS total_products

```

```
FROM
latest_date
GROUP BY 1
```

Problem #16 Solution:

We can get the breakdown of total time spent on each activity by each user by filtering out for the activity_type and taking the sum of time spent. In doing this, we want to do an outer join with the age bucket to get the total time by age bucket for both activity types. This results in the below two subqueries. Then, we can use these two subqueries to sum them by joining on the appropriate age bucket and take the proportion for send time and the proportion for open time per age bucket:

```
WITH send_timespent AS (
    SELECT age_breakdown.age_bucket, SUM(activities.time_spent) AS
send_timespent
    FROM age_breakdown
    LEFT JOIN activities on age_breakdown.user_id = activities.user_id
    WHERE activities.type = 'send'
    GROUP BY 1
),
open_timespent AS (
    SELECT age_breakdown.age_bucket, SUM(activities.time_spent) AS
open_timespent
    FROM age_breakdown
    LEFT JOIN activities on age_breakdown.user_id = activities.user_id
    WHERE activities.type = 'open'
    GROUP BY 1
),

SELECT a.age_bucket,
    s.send_timespent / (s.send_timespent + o.open_timespent) AS pct_send,
    o.open_timespent / (s.send_timespent + o.open_timespent) AS pct_open,
FROM age_breakdown a
LEFT JOIN send_timespent s ON a.age_bucket = s.age_bucket
LEFT JOIN open_timespent o ON a.age_bucket = o.age_bucket
GROUP BY 1
```

Problem #18 Solution:

First, we need to get the places where the reviews are all 4 or 5 stars. We can do this using a HAVING clause, instead of a WHERE clause since the reviews need to all be 4 stars or above. For the HAVING condition, we can use a CASE statement that filters for 4 or 5 stars and then take a SUM over them. This can then be compared with the total row count of the particular business_id reviews to ensure that the count of top reviews matches with the total review count. With the relevant businesses, we can then do an outer join with the original table on business_id to get a COUNT of distinct business_id matches, and then the percentage by comparing the COUNT from the top places with the overall COUNT of business_id:

```
WITH top_places AS (
```

```

SELECT business_id
FROM user_transactions
GROUP BY 1
HAVING
    SUM(CASE WHEN rating >= 4 THEN 1 ELSE 0 END) = COUNT(*)
)

SELECT
    COUNT(DISTINCT t.business_id) AS top_places,
    COUNT(DISTINCT t.business_id)/COUNT(DISTINCT r.business_id) AS
top_places_pct
FROM reviews r
LEFT JOIN top_places t
    ON r.business_id = t.business_id

```

Problem #21 Solution:

ACID is a set of properties that ensures that even in the event of errors, power outages, and other unforeseen circumstances, a database will still hold up. It is an important framework for studying database systems.

A: Atomicity, meaning that an entire transaction happens as a whole or it doesn't happen at all (no partial transactions). This prevents partial updates which can be problematic. Therefore, transactions cannot be "in progress" to any user.

C: Consistency, meaning that there are integrity constraints such that the database is consistent before and after a given transaction. Essentially, if I search for what is in Row 3 and then do so again without any modifications to the database (no deletes or inserts), I should get the same result. Any referential integrity is handled by appropriate checks for the primary and foreign keys.

I: Isolation, meaning that transactions happen in isolation and thus multiple transactions can occur independently without interference. This maintains concurrency properly.

D: Durability, meaning that once a transaction is complete, that information is now updated in the database even in the event of a system failure.

Problem #25 Solution:

MapReduce is a framework that is heavily used in processing large datasets across a large number of clusters (many machines). Within the groups of machines, there are worker nodes (which carry out the computations) and master nodes (which delegate the tasks for each worker node). The three steps are generally as follows.

1. Map step: each worker node applies a specific map operations on input data (which the master node ensures will not be duplicated) and writes the output to a memory buffer.
2. Data is re-distributed based on the output keys from the prior step's map function, such that for any given key, it is located on the same worker node.

3. Each worker node processes each key in parallel using specific reduce operations to get the output result.

Since the mapping and reducing functions can all be done in parallel, the amount of processing done is only limited by the amount of compute and data available. Note that there are edge cases if there are failures from worker nodes - in those cases, the desired operations can be re-scheduled by the master node.

Problem #27 Solution:

A trigger is like a CHECK condition, but every time there is an update to the database, the trigger condition will be checked to see if it has been violated. This allows you to implement some level of control and assurance that all your data entries meet a certain condition. For instance, a trigger that states that all ID values must be > 0 will ensure that you get no null values or negative values. When someone tries to enter such a value, the entry will not go through.

That being said, there are reasons why to not include business logic within database triggers. For example: 1) introduction of side effects which lead to bugs or other unintended consequences, or 2) performance problems in which case there is a cascading effect on triggers that leads to locking and other issues.

=====

Introduction

SQL is the good old friend that has always worked. It's something you always come back to, even as Pandas, Julia, Spark, Hadoop, and NoSql attempt to dethrone and replace SQL as the new de-facto data tool.

Eventually though, they all fail in the face of the consistently reliable SQL. And that's why SQL continues to get asked in interviews.

A note before we start...

This guide should be used for anyone who is preparing for an interview in which they know SQL will show up. This guide **is not** a search engine optimized listicle (top 50 sql questions for 2021...really?).

Rather, this is **real advice and REAL interview questions and exercises** gathered from hundreds of data scientists, engineers, and analysts. We sprinkle exercises throughout this post after learning concepts. Be sure to try attempting the questions first before we walk through solving them.

Lastly, if you enjoy this article, **please give us a share** and check out our [SQL course](#) that goes a little deeper with more exercises and problems.

1. Why does SQL show up on the interview?

SQL allows data scientists and engineers to do a couple of important things.

One is to **effectively store and retrieve information at scale for analytics**. Even though Google Sheets allows users to easily manipulate and visualize data, it cannot store and scale like a SQL database can. Other popular programs –namely Hadoop and Spark– can scale much further than SQL, but still don't have a clean and easy-to-use language like SQL to retrieve data efficiently.

Another great thing about SQL is that understanding the fundamentals bridges the gap between **engineering and data science**. Knowing SQL well gives you a competitive edge over any other candidate, whether you're competing for a position as a product manager, software engineer, or even as a business analyst. Having the skillset to write and pull your own queries is like being a magician that can come up with analyses out of thin air.

And at the end of the day, you could just be really good at SQL if you wanted to and make tons of money creating ETL jobs or pulling dashboards with efficiency. That's how valued SQL is.

How often does SQL show up in interviews?

One prevailing question is how often SQL shows up in interviews.

At Interview Query, we analyzed a dataset of Glassdoor data science interview experiences and responses submitted by our users. The analysis came back that SQL was asked:

- [70% of the time during Facebook's data science interviews](#)
- [94% of the time during Amazon's business intelligence and analyst interviews](#)

[Skills tested in Facebook's Data Science Interview as of 2021](#)

Due to its nature in being able to get and manipulate your own data, this is by far the most important skill now towards nabbing a data science position. And while pandas and other languages are useful, note that [SQL will always be the one that matters](#).

If you're looking for a job as a data engineer, SQL is just the beginning. Check out our [Data Engineering Interview Questions & Guide](#) today.

2. Strategies for the live SQL interview

Let's go over the common strategies when tackling SQL interview questions.

1.Repeat the problem statement

When presented with a SQL question, listen carefully to the problem description and repeat back what you think the crux of the problem is. The interviewer can then help verify if your understanding is correct.

2. Understand the edge cases

If time permits, write out a base case and an edge case to show that you understand the problem. For example: if the interviewer asks you to pull the average number of events per user per day, write out an example scenario where you're verifying this metric.

Do duplicate events matter? Are we looking at distinct users? These are questions we need to clarify.

3. Try working backwards if the problem is tricky

Sketching out what the output of the SQL question will look like is a great strategy towards solving the problem. Usually, if I know what the end output table is supposed to look like, I can work backwards from there on what functions need to be applied before.

For example, if the output looks like this:

date	average events per user
2021-12-01	3.5
2021-12-02	4.0

I know that the table before this aggregation would have to look something like this.

date	event	user_id
2021-12-01	click	1
2021-12-01	view	1
.....		

And then, I can figure out what functions I should use to get to my desired output!

4. Pattern match to different functions

As you practice more and more SQL exercises, what you'll find is that many SQL problems follow similar patterns. There are techniques we can use in SQL, like utilizing `HAVING` on aggregations, self-joins and cross-joins, and applying window functions. But, additionally, we'll see problems that run in a similar vein.

For example, writing a query to get the second highest salary or writing a query to isolate every fifth purchase by a user utilizes the same `RANK` function in SQL.

Understanding the commonalities between questions will help you understand the first step to solving SQL questions faster because you can re-use similar code and stitch together techniques on top of each other.

5. Start writing SQL

Finally, it's important to just start writing SQL. It's better to start writing an imperfect solution vs trying to perfectly understand the problem or trying to perfect the solution on the first try.

Verbalize your assumptions and what you're doing as you write SQL and your interviewer can then be put on the same page as you.

3. The 7 different SQL interview questions

SQL questions asked during interviews can vary widely across companies, but even more so across positions. You won't see data scientists asked the same SQL questions as software engineers, and that's because data scientists have to write different types of queries compared to software engineers.

Generally, each SQL interview question can be bucketed into these categories:

- Definition based SQL questions
- Basic SQL questions
- Reporting and metrics SQL questions
- Analytics SQL questions
- ETL SQL questions
- Database design questions
- Logic based SQL questions

In this next section, we'll go over which types of SQL questions are expected for different roles and what those different kinds of SQL questions are in detail.

4. SQL questions for data scientists and analysts

SQL interview questions for data scientists and data analysts will likely show up in three parts of the interview process: the technical round, the take-home challenge, and the onsite interview.

The technical round and take-home challenge will usually consist of SQL questions **designed to filter out candidates**. Since SQL is commonly used as a filter mechanism for data scientists, it's important to perform well on this part of the interview in order to demonstrate competence.

Depending on what type of data science role you're interviewing for, you'll find that most SQL questions will be split into these three types:

- Basic SQL Interview Questions
- Reporting and Metrics SQL Interview Questions
- Analytics SQL Interview Questions

Basic SQL Interview Questions

Basic SQL questions are what they sound like. These questions will be generally easy and focus on assessing if you know the basics.

Definition based SQL questions are grouped into this category because they're super easy to learn. All you have to do is study a list of definitions of SQL terms and applications. These questions will include understanding the differences between joins, what kinds of aggregations exist, and knowing the basic functions like `CASE WHEN` or `HAVING`.

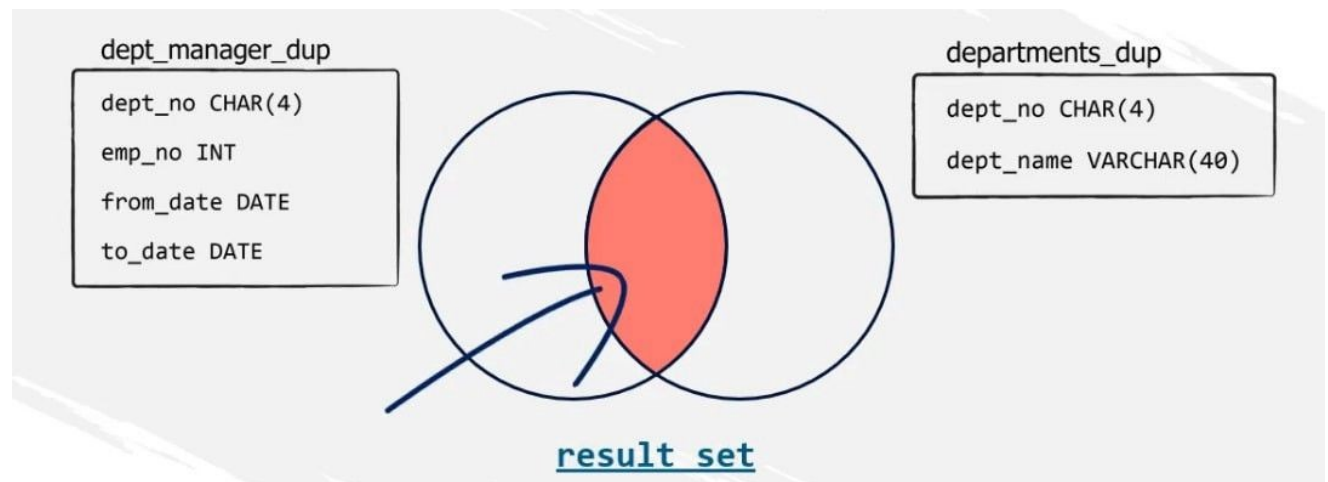
Basic SQL interview questions that involve a user actually writing a query are slightly different. These will involve getting the `COUNT` of a table, knowing what the `HAVING` clause does, and figuring out how to utilize `LEFT JOIN` versus `INNER JOIN` to give you the values that you need.

Read more on the the [basic concepts you need to know to pass your data science interview](#) here.

[Three SQL Concepts for your Data Scientist Interview](#)

[I've interviewed a lot of data scientist candidates and have found there are a lot of SQL interview questions for data science that eventually boil down to three generalized types of conceptual understandings.](#)

[Interview Query BlogJay Feng](#)



Basic SQL Concepts to Review

- What's the difference between a `LEFT JOIN` and an `INNER JOIN`?
- When would you use `UNION` vs `UNION ALL`? What if there were no duplicates?
- What's the difference between `COUNT` and `COUNT DISTINCT`?
- When would you use a `HAVING` clause versus a `WHERE` clause?

Basic SQL Question Example:

users table

columns	type
id	int
name	varchar
neighborhood_id	int
created_at	datetime

neighborhoods table

columns	type
id	int
name	varchar
city_id	int

We're given two tables, a *users* table with demographic information and the neighborhood they live in and a *neighborhoods* table.

Write a query that returns all of the neighborhoods that have 0 users.

Try answering this question with our [interactive SQL editor](#).

Here's a hint:

*Our predicament is to find all the neighborhoods without users that live in them. This means we have to introduce a **concept of existence of a field in one table, while not existing in another**.*

For example, let's say we generate some fake data of user's and the neighborhoods they live in. We would expect it to look something like this.

neighborhoods.name	users.id
castro	1
castro	2

cole valley	null
castro heights	3
sunset heights	4

We see each user from one to four is appropriately placed in their respective neighborhood except for the neighborhood of Cole Valley. That's the neighborhood we're targeting for returning in our query.

Strategies: whenever the question asks about finding values with 0 something (users, employees, posts, etc..), immediately think of the concept of **LEFT JOIN**! An inner join finds any values that are in both tables, a left join keeps only the values in the left table.

Our predicament is to find all the neighborhoods without users. To do this, we must do a left join from the neighborhoods table to the users table.

If we then add in a where condition of **WHERE users.id IS NULL**, we will get every single neighborhood without a singular user as shown above.

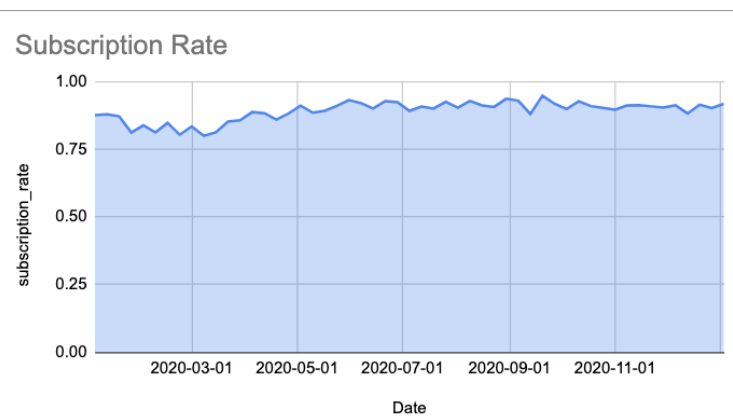
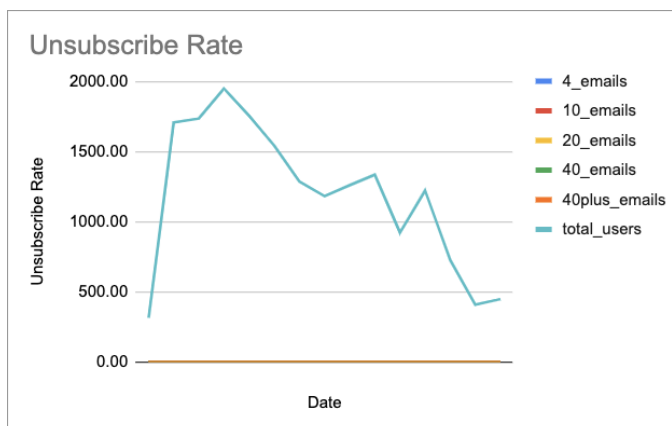
```
SELECT n.name
FROM neighborhoods AS n
LEFT JOIN users AS u
  ON n.id = u.neighborhood_id
WHERE u.id IS NULL
```

[Empty Neighborhoods — Interview Query sql problem](#)

[users table columnstype idint namevarchar neighborhood_idint created_atdatetime](#)
[neighborhoods table columnstype idint namevarchar city_idint We're](#)

Check out the full problem and different solutions we can apply

Reporting and Metrics SQL Interview Questions



Example of Reporting SQL queries to Dashboard

Reporting and metrics SQL questions are probably the most common type of SQL question to show up in interviews.

Reporting SQL interview questions replicate the work that many business and reporting analysts do on a day-to-day basis. This means writing queries that end up in dashboards and key metrics.

For example, a typical B2C software company would likely want to understand how many new users signed up. Better yet, they'd want to understand the daily, weekly, and monthly active users that are on their platform. What about the number of week-over-week new users that signed up as well?

These questions are generally some of the most common problems that analysts and data scientists will run into when pulling metrics for their day-to-day job. The output of what they need to pull is very clearly defined and the queries themselves require complex joins, sub-queries, self-joins, window functions, and more.

Interested in doing this sort of work in the field of business intelligence? [Check out our article on Business Intelligence Interview Questions on Interview Query.](#)

Easy Metrics-Based SQL Questions

- What's the total distance traveled for all riders on Lyft in the month of March?
- How many bookings did Airbnb get in the month of December?
- How many existing users booked at least one place in Airbnb in the month of December?

Advanced Metrics-Based SQL Questions:

- Write a query to get the month-over-month change of new users in January.
- Write a query to calculate the monthly retention of subscribers.
- Write a query to get a histogram of the number of posts per user in 2020.

Let's take a look at tackling an example problem posed by LinkedIn in their data science interview.

``job_postings`` table

column	type
id	integer
job_id	integer
user_id	integer
date_posted	datetime

[Rep](#)

[eat Job Postings](#)

Given a table of job postings, write a query to breakdown the number of users that have posted their jobs once versus the number of users that have posted at least one job multiple times.

[Repeat Job Postings — Interview Query sql problem](#)

``job_postings`` table column type id integer job_id integer user_id integer date_posted datetime Given a table of job postings, write a query

Try the problem on Interview Query

Here's a hint on tackling the problem.

First, let's visualize what the output would look like and clarify the question.

We want the value of two different metrics, the **number of users that have posted their jobs once** versus the **number of users that have posted at least one job multiple times**. What does that mean exactly?

Well, if a user has 5 jobs but only posted them once, then they are part of the first statement. But if they have a 5 jobs and posted a total of 7 times, that means that they had to **at least posted one job multiple times**.

We can visualize it the following way with an example. Let's say this is our end output:

```
Users posted once | Posted multiple times
-----
1                | 1
```

To get to that point, we need a table with the count of user-job pairings and the number of times each job gets posted.

user_id	job_id	number of times posted
1	1	2
1	2	1
2	3	1

We can pretty easily get to that point with just a simple GROUP BY on two variables, user_id and job_id.

```
WITH user_job AS (  
    SELECT user_id, job_id, COUNT(DISTINCT date_posted) AS num_posted  
    FROM job_postings  
    GROUP BY user_id, job_id  
)
```

```
SELECT * FROM user_job
```

Watch how I solve the rest of the solution below.

Check out this video where I tackle a harder reporting SQL question featuring multiple self-joins and year over year metrics.

Analytics SQL Interview Questions

Analytics SQL interview questions are some of the trickiest interview questions that you will face. This is because they test two concepts.

1. Understanding what metrics we need to answer the question.
2. Writing the correct SQL query that will output these metrics.

Analytics SQL interview questions are designed to test how you would think about solving a problem, and are purposely left more ambiguous than other types of problems. The tough part is that you not only have to think critically about what the SQL output has to look like, you also need to **understand EXACTLY the right data points to pull**.

For example, an interviewer might ask you to write a SQL query (given a few tables) to understand which AB test variant won. But there might not even be any understanding of **what winning actually means**.

Here's another example:

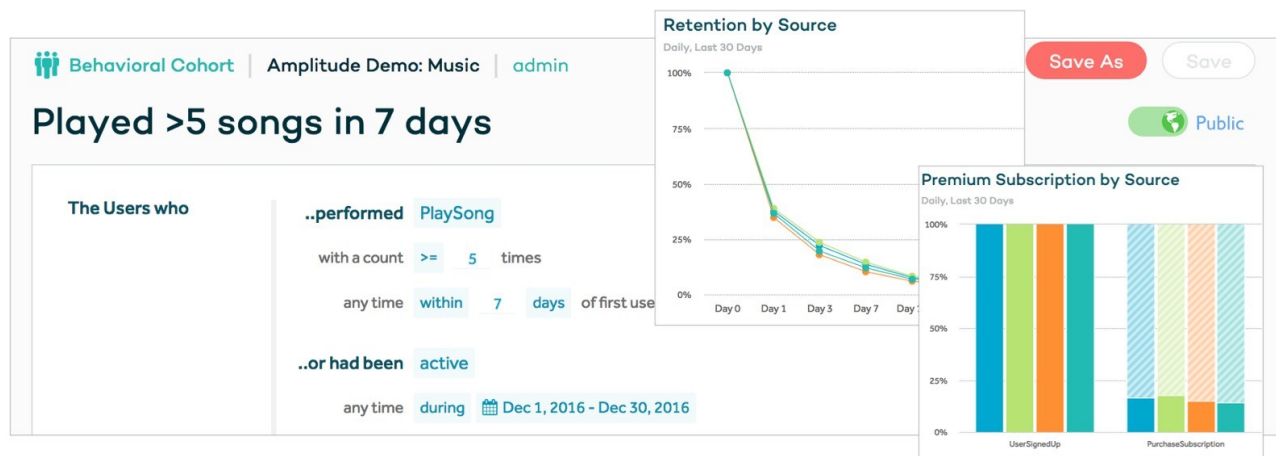
Let's say that we're given a table of users and sessions. A product manager wants to know if the new sign-up flow launched last week improved the sign up conversion rate at all.

Write a query to pull a metric to determine if this is the case.

We can see that there isn't a clearly defined output like in the previous question. The interviewer isn't defining a metric that the candidate needs to write a query to find, such as the number of new users signed up last week or the conversion rate.

Rather, the interviewer is asking the candidate to first *define a metric* to solve a problem *and then* write a query to get that metric. The reason why these problems are so difficult is because getting the first part right is critical to getting the second part as well.

It also makes for a great way to test product analysts and data scientists that are especially analytically focused. The SQL question is testing if a candidate solve an ambiguous question and prove their technical chops.



Example of UI analytics from CXL

Analytics SQL Question Concepts

- We ran an A/B test on two different sign up funnels. Write a query to see which variant "won".
- We're looking to understand the effect of a new Uber driver incentive promotion released in the past month on driver behavior. Write a query to figure out if the incentive worked as indicated.
- Amazon released a new recommendation widget on their landing page. Write a query to determine the impact the recommendation widget made on user behavior for one metric.

Analytics SQL Question Example:

`friends` table

column	type
user_id	integer
friend_id	integer

`page_likes` table

column	type
user_id	integer
page_id	integer

[Liked Pages](#)

Let's say we want to build a naive recommender.

We're given two tables: one table called `friends` with a `user_id` and `friend_id` columns representing each user's friends, and another table called `page_likes` with a `user_id` and a `page_id` representing the page each user liked.

Write an SQL query to create a metric to recommend pages for each user based on recommendations from their friends' liked pages.

Note: It shouldn't recommend pages that the user already likes.

[Liked Pages — Interview Query sql problem](#)

[`friends` table column type user_id integer friend_id integer `page_likes` table column type user_id integer page_id integer Let's](#)

Try the question in the SQL editor

Here's a hint.

Let's solve this problem by visualizing what kind of output we want from the query.

*Given that we have to create a metric for each user to recommend pages, we know we want something with a `user_id` and a `page_id` along with some sort of **recommendation score**.*

*Let's try to think of an easy way to represent the scores of each `user_id` and `page_id` combo. One naive method would be to create a score by **summing up the total likes by friends on each page that the user hasn't currently liked**. The max value of total likes will be the most recommendable page.*

Then, the first thing we have to do is write a query to associate users with their friends' liked pages. We can easily do that with an initial join between the two tables.

```
WITH t1 AS (  
    SELECT  
        f.user_id  
        , f.friend_id  
        , pl.page_id  
    FROM friends AS f  
    INNER JOIN page_likes AS pl  
        ON f.friend_id = pl.user_id  
)
```

Now, we have every single `user_id` associated with the friends' liked pages. Can't we just do a `GROUP BY` on `user_id` and `page_id` fields and get the `DISTINCT COUNT` of the `friend_id` field?

Not exactly. We still have to filter out all of the pages that the original users also liked. How do we do that?

We can do that by joining the original `page_likes` table back to the `t1` CTE. We can filter out all the pages that the original users liked by doing a `LEFT JOIN` on `page_likes` and then selecting all the rows where the `JOIN` on `user_id` and `page_id` are `NULL`.

```
LEFT JOIN page_likes AS pl  
    ON t1.page_id = pl.page_id  
    AND t1.user_id = pl.user_id  
WHERE pl.user_id IS NULL # filter out existing user likes
```

More study practice? Check out three tricky analytics interview questions that I solved with my friend Andrew.

ETL SQL Interview Questions

ETL stands for "Extract, Transfer, Load" and describes the process for which data flows between different data warehousing systems.

Extract does the process of reading data from a database. Transform converts data into a format that could be appropriate for reporting, analysis, machine learning, etc., and Load writes the transformed data into another database, table, or any other data storage service that can be then used by another data scientist or engineer for reporting.

Many times, ETLs are stacked on top of each other, creating a system of complex data-flows that eventually need to be managed by scheduling systems, such as Airflow or MLflow.

In the interview, ETL concepts are important to know for virtually all roles. The more difficult interview questions, however, will likely be focused and asked in data engineering, business intelligence, and related interviews.

Basic ETL SQL Concepts

- What's the difference between `TRUNCATE` and `DROP`?
- What is a `PRIMARY KEY` in SQL syntax?
- What is a `FOREIGN KEY`?

Advanced ETL SQL Concepts

- List an example of when you would add an `INDEX` to a table?
- What's the difference between a `PARTITION` and an `INDEX`?
- Does creating a view require storage in a database?
- Let's say that we have two ETL jobs that feed into a single production table each day. Can you think of any problems this might bring up?

Example ETL Interview Questions

Let's say you have a table with a billion rows.

How would you add a column inserting data from the original source without affecting the user experience?

[Modifying a billion rows — Interview Query system design problem](#)

[Let's say you have a table with a billion rows. How would you add a column inserting data from the original source without affecting the user experience?](#)

Here's a hint.

In a general database, writing a column would lock up the whole table. However, we can potentially do the update in steps.

One strategy is taking an exact replica of the existing table and updating the results offline. So, we could create a new table by copying the old table, update the new column, and then drop the old existing table and renaming it to the new table.

*However, this does produce a problem in that we may have a potential mismatch of data. From the time that we take a copy of the new table to then switching the tables over, **we may have lost data.***

View the solution in the mock interview with my friend Scott who works as a data and machine learning engineer.

[Employee Salaries - ETL Error](#)

employees table

column	type
id	integer
first_name	string
last_name	string
salary	integer
department_id	integer

[Table for](#)

[Interview Query problem](#)

Let's say we have a table representing a company payroll schema.

Due to an ETL error, the employees table, instead of updating the salaries every year when doing compensation adjustments, did an insert instead. The head of HR still needs the current salary of each employee.

Write a query to get the current salary for each employee.

Assume no duplicate combination of first and last names. (I.E. No two John Smiths)

Here's a hint.

The first step would be to remove duplicates and retain the current salary for each user.

*Given that we know there aren't any duplicate first and last name combinations, we can remove duplicates from the employees table by running a `GROUP BY` on two fields, the **first and last name**. This allows us to then get a unique combinational value between the two fields.*

Run the SQL query below [in the editor](#). What does this `max_id` value get us?

```
SELECT first_name, last_name, MAX(id) AS max_id
FROM employees
```

GROUP BY 1,2

[Employee Salaries \(ETL Error\) — Interview Query sql problem](#)

[employees table columntype idinteger first_namestring last_namestring salaryinteger department_idinteger Let's say we have a table](#)

5. SQL questions for engineers

A little while ago, I talked to a hiring manager at a prominent tech company in Silicon Valley, and he mentioned that he always tested SQL first for all members on his team - even the engineers.

I asked him to elaborate and he emphasized the importance of SQL.

"No matter what, the candidate needs to know SQL. I don't care if they're the best machine learning expert in the world— if you can't pull your own data, you can't work on my team. **No one is going to pull your data for you.**"

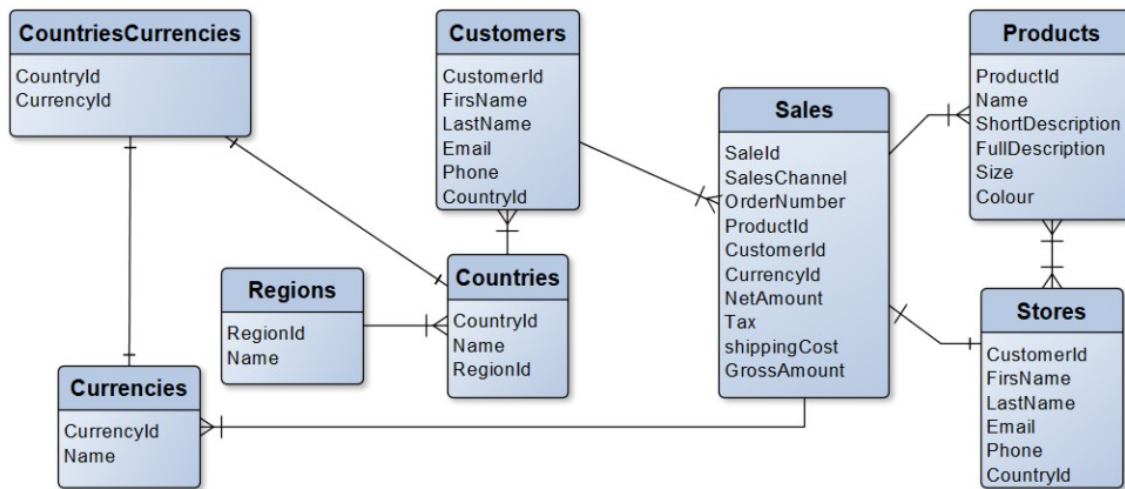
If you were the world's best machine learning researcher in the world, how would it look if you had to go to an analyst and say:

"Hey, can you get me this dataset for me, clean it, and then also feature engineer it a bit? Thx."

This doesn't exactly make for great teamwork. It also wouldn't make sense that you would be the best machine learning expert in the world and still not know how to master something as simple as SQL.

Each engineering position that works with SQL tests it differently, but let's start with data engineering, because data engineers definitely have to be masters at SQL.

Data Engineering SQL Interview Questions



Exa

mple of Data Architecture

Data engineers specialize in designing large scale data systems. This work can span from building out production and backup databases and data warehouses to writing code to transfer the data between different states.

One key understanding is that data engineers always require SQL knowledge. It's the fundamental building block to their position. Most of the required interview questions we have already gone over before, except for database design.

- Reporting SQL Interview Questions
- ETL SQL Interview Questions
- Database Design SQL Interview Questions

Database Design SQL Interview Questions

Database design SQL questions test your knowledge of data architecture and design. Most importantly, it tests whether you know how to design a database from scratch, given a business idea, application, or any other software that needs to interact with a database.

Many times, when databases need to scale for performance or breadth size of tables, we need to realize how to modify our database to fit the new requirements. Starting from a solid initial design is always important when building out databases.

Example Database Design Questions

- Let's say we're working at Spotify. In the users table we have columns such as `name` and `location` but also columns like `total_liked_songs` and `favorite_artist`. Should we do something about this table?
- How would you design a database for a fast food restaurant?

- Let's say we are productionizing a machine learning model. How would you set up the data architecture to ensure that the model serves all of our users in real time? How would you ensure that the model can be re-trained each day with new data coming in?

Example Database Design Interview Question

How would you create a schema to represent client click data on the web?

[Click Data Schema — Interview Query system design problem](#)

[How would you create a schema to represent client click data on the web?](#)

Here's a hint.

What exactly does click data on the web mean?

Any form of button clicks, scrolls, or action at all is an interaction with the client interface—in this case desktop—and would somehow be represented into a schema form for the end user to query. This does not include client views.

*A simple but effective design schema would be to first represent **each action with a specific label**. In this case, assigning each click event a name or label describing its specific action.*

For example, let's say the product is Dropbox on web and we want to track each folder click on the UI of an individual person's Dropbox. We can label the clicking on a folder as an action name called `folder_click`. When the user clicks on the side-panel to login or logout and we need to specify the action, we can call it `login_click` and `logout_click`.

What other fields do we need?

[Check out the rest of the solution here](#)

Machine Learning Engineering SQL Interview Questions

Machine learning engineers need to pull data in a similar fashion to data scientists. Depending on the role, however, machine learning engineers might not be asked SQL interview questions. At more established companies like Facebook and Google, ML engineers have processes to get their data without using SQL.

The most common types of SQL interview questions that will show up are:

- Reporting and Metrics SQL Interview Questions
- ETL SQL Interview Questions

Software Engineering SQL Interview Questions

Software engineers will be asked SQL questions mostly a precursor to test that they know the very basics. Software engineers interact with data when building APIs and endpoints and have to query the database for data.

The syntax that software engineers usually use for SQL is generally different from data scientists and analysts. This is because of the rise in interfacing with a SQL relational mapper and toolkit, such as [SQLAlchemy](#) for Python. The software engineer isn't exactly writing SQL syntax, but rather writing Python, Go, Java, etc. that wraps around SQL.

The syntax can also be more like this:

```
SELECT * From Emp, Dept
```

Or like this:

```
SELECT *  
FROM Emp, Dept  
WHERE Emp.dept_id = Dep.id
```

The most common SQL interview questions that software engineers would receive would be:

- Basic SQL Interview Questions
- ETL SQL Interview Questions
- Database Design Interview Questions
- Logic Based SQL Interview Questions

Logic based SQL Interview Questions

Logic based SQL interview questions are very tricky. They aren't really based on real life examples so much as putting the trickiness of algorithms and data structure interviews into SQL questions. This is exemplified on sites such as LeetCode, where you'll see a lot of interview questions that aren't very practical for real life scenarios.

Here's an example of a logic based SQL interview question.

`flights` table

column	type
id	integer
source_location	string
destination_location	string

Write a query to create a new table, named flight routes, that displays unique pairs of two locations.

Duplicate pairs from the flights table, such as Dallas to Seattle and Seattle to Dallas, should have one entry in the flight routes table.

[Flight Records — Interview Query sql problem](#)

[`flights` table columntype idinteger source_locationstring destination_locationstring Write a query to create a new table, named flight routes.](#)

6. Quick SQL concepts & review

Here, I want to go over an extensive review list of all of the concepts we just went through. In the second part, I've linked actual real SQL exercises to try.

SQL Concepts and Questions

What is a join?

A `JOIN` is a keyword used to merge together two or more tables on the same key.

Which SQL command is used to add rows to a table?

The `INSERT` command.

When might you denormalize your data?

Denormalize when its OLAP operations and normalize when OLTP.

What is OLAP and OLTP?

OLTP are databases intended for online transaction processing and OLAP are databases intended for online analytical processing.

What's the difference between WHERE and HAVING?

The main difference is that a `WHERE` clause is used to filter rows before grouping and `HAVING` is used to exclude records after grouping.

When do you use the CASE WHEN function?

`CASE WHEN` lets you write complex conditional statements on the `SELECT` clause and also allows you to pivot data from wide to long formats.

When would you use a SELF-JOIN?

A self-join joins a table with itself. It's most commonly used when needing to perform aggregation functions when data is stored in one large table rather than smaller ones.

7. SQL interview questions and exercises

Try more SQL interview questions with our interactive editor!

[Employee Salaries — Interview Query sql problem](#)

[employees table columnstypes idint first_namevarchar last_namevarchar salaryint department_idint departments table columnstypes idint namevarchar Given](#)

[Comments Histogram — Interview Query sql problem](#)

[users table columnstype idinteger namestring created_atdatetime neighborhood_idinteger mailstring comments table columnstype user_idinteger bodytext created_atdatetime Write](#)

[Upsell Transactions — Interview Query sql problem](#)

[`transactions` table columntype idinteger user_idinteger created_atdatetime product_idinteger quantityinteger We're given](#)

[Random SQL Sample — Interview Query sql problem](#)

[`big_table` columntype idint namevarchar Let's say we have a table with an id and name field. The table holds over 100 million rows and we want to sample](#)

8. SQL Interview Questions asked by Facebook, Amazon, and Google (FAANG)

What's the difference between SQL interview questions asked by the big tech companies like Facebook, Amazon, Microsoft, etc. vs other types of companies? Generally, FAANG companies ask questions that are more product and analytically facing that are conceptualized with a case study.

For example, Facebook will likely ask questions surrounding practical queries they have to run on a day-to-day with their platform. That means pulling different ad bids, looking at daily active users, and understanding product performance with dashboards.

Generally, it's helpful to tailor your SQL interview practice according to industry of the company that you'll be interviewing for. If you're interviewing for a bio-tech company, think of the type of data that an data analyst or data scientist will be querying on a day-to-day basis, and practice questions related to that.

Check out a SQL mock interview that I did with my friend Ben replicating the Facebook data science interview.

9. Last tips and notes

There's a couple of important details we need to consider in preparing for SQL technical interviews.

SQL interviews are almost always “white-boarding”.

SQL interviews are rarely conducted in a live console where you can run queries in a playground database. Even CoderPad, which is the de-facto interviewing tool for engineers and most tech companies, has only around four example tables you can use. And I've never heard of any interviewers using questions from their tables for interview problems.

Why is this the case?

First and foremost, it is difficult to set up a shared environment to test SQL with real data. I also suspect that many companies don't think it's a good test if a candidate does have a live SQL playground to work in. For example, Facebook chooses not to create a SQL playground for interviewees to use.

The main reason, however, is that testing SQL by white-boarding SQL syntax accomplishes the task of assessing real mastery of the language that is also practical for speaking towards your efficiency as an individual contributor.

The worst-case scenario is where you **write a wrong query**, put it into an ETL job, and then be comfortably oblivious as your company ingests wrong metrics for who knows how long. Learning how to write correct SQL is very necessary and the expectation is that you can whiteboard it without too much failure.

SQL can be quickly learned.

One can improve their SQL ability by simply practicing. The steps towards mastering SQL just include repeatedly working on problems. SQL interview questions are very similar to the tasks that you do on the job as a data scientist, which, on the plus side, means that mastering SQL can sometimes consist of you doing your actual day job.

However, the downside is that if your role is specialized and you're restricted to only using SQL for something like time-series forecasting, or you don't use SQL on a day-to-day basis, you can't practice it effectively across the board for interviews.

Therefore, the unique problem sets that we surface in Interview Query allow you to **hack the learning curve of SQL experience** and master the querying language.

Our job within this course is to expose you to the most common types of SQL questions that data scientists have to work on. Examples range from practicing applied analytics to dash-boarding metrics to queries for model and dataset generation.

Optimize the Performance of your SQL Query

Always try optimizing the performance of your SQL queries during the interview. If you end up writing a SQL query that is horribly inefficient, you will likely be rejected from the interview.

If you're working at a company with large swaths of data, your queries are likely to be slow. My co-workers and I on the data science team would write a query, go to the snack bar across the office and walk back, and it still would not be done. There was just so much data we had to query.

Running queries generally takes a long time, which is why **getting it right on the first try is a good demonstration of general competence**.

Imagine running an expensive join and waiting an hour until it times out just from non-optimized code. You likely just got blocked on a task for a whole hour during the workday. Then, imagine again an hour-long SQL query that **comes back with wrong data** because you messed up a simple join.

Take this interview question for example:

Let's say we have a table with an `id` and `name` field. The table holds over 100 million rows and we want to sample a random row in the table without throttling the database.

Write a query to randomly sample a row from this table.

The most obvious solution for this question would be to just order the rows by `RAND()` and select one row.

```
SELECT id, name
FROM big_table
ORDER BY RAND()
LIMIT 1
```

But it's noted that the table has 100 million rows. Ordering by `RAND()` would never finish the query! So even though it's technically a correct solution, it's not practical for the question being asked.

[Random SQL Sample — Interview Query sql problem](#)

[`big_table` columntype idint namevarchar Let's say we have a table with an id and name field. The table holds over 100 million rows and we want to sample](#)

Try the rest of the problem here

Use Correct Formatting for your SQL Queries.

I won't tell you how many times I have marked people down because their formatting is so atrocious I can't tell if the actual solution is correct or not.

Always make sure that you're adding line breaks and spacing to make your SQL code easy to read and digestible. Otherwise, you will find yourself writing SQL code that your interviewer cannot even verify as being correct, ruining the entire point of the interview itself.

Does it matter which SQL Engine I use for my interview?

No, not really. The only thing to watch out for is if you use a very uncommon SQL engine that goes against common SQL technique and syntax.

If you start using window functions such as `PARTITION BY` or `RANK` and your interviewer doesn't use that kind of SQL engine or doesn't understand window functions, you'll likely have to re-write your query in another way that is intelligible.

10. SQL Study plan for your next interview

SQL interview questions on Interview Query are bucketed into **easy, medium, and hard** and go over every SQL concept that we went through in this blog post.

Use Interview Query to level yourself at how good you are at SQL. Here's how you should approach each problem.

1. Try the SQL question without running any SQL code in the engine. Give yourself 15 to 20 minutes to work on each problem by just reading the problem statement and using the editor.
2. Work out all the bugs and make sure that when you click run or submit, you feel like this is the definitive solution that will work without syntax errors.
3. If you get stuck, try the problem hint. The hints on certain questions will push you into learning how to think about solving the problem at hand.
4. Try submitting your code to check if it passes our test cases. If it misses a few test cases, figure out why by debugging your code.
5. If you need to debug your code, you can hit the `RUN SQL` button to try it on our test database. Run your code and check for syntax errors. If you get any, re-read your query to fix the necessary errors.
6. Check your solution against ours. Take note of whether there are any optimizations you can use to write a more efficient query.