1001 Data Engineering Interview Questions

Looking for a job or just want to know what people find important? In this chapter you can find a lot of interview questions we collect on the stream.

Ultimately this should reach at least one thousand and one questions.

But Andreas, where are the answers?? Answers are for losers. I have been thinking a lot about this and the best way for you to prepare and learn is to look into these questions yourself.

This cookbook or Google will help you a long way. Some questions we discuss directly on the live stream.

Live Streams

First live stream where we started to collect these questions.

Podcast Episode: #096 1001 Data Engineering Interview Questions
First live stream where we collect and try to answer as many interview questions as possible. If this helps people and is fun we do this regularly until we reach 1000 and one.
Watch on YouTube

All Interview Questions

The interview questions are roughly structured like the sections in the "Basic data engineering skills" part. This makes it easier to navigate this document. I still need to sort them accordingly.

SQL DBs

- What are windowing functions?
- What is a stored procedure?
- Why would you use them?
- What are atomic attributes?
- Explain ACID props of a database
- How to optimize queries?
- What are the different types of JOIN (CROSS, INNER, OUTER)?
- What is the difference between Clustered Index and Non-Clustered Index - with examples?

The Cloud

- What is serverless?
- What is the difference between IaaS, PaaS and SaaS?
- How do you move from the ingest layer to the Cosumption layer? (In Serverless)
- What is edge computing?
- What is the difference between cloud and edge and on-premise?

Linux

- What is crontab?

Big Data

- What are the 4 V's?
- Which one is most important?

Kafka

- What is a topic?
- How to ensure FIFO?
- How do you know if all messages in a topic have been fully consumed?
- What are brokers?
- What are consumergroups?

- What is a producer?

Coding

- What is the difference between an object and a class?
- Explain immutability
- What are AWS Lambda functions and why would you use them?
- Difference between library, framework and package
- How to reverse a linked list
- Difference between args and kwargs
- Difference between OOP and functional programming

NoSQL DBs

- What is a key-value (rowstore) store?
- What is a columnstore?
- Diff between Row and col.store
- What is a document store?
- Difference between Redshift and Snowflake

Hadoop

- What file formats can you use in Hadoop?
- What is the difference between a namenode and a datanode?
- What is HDFS?
- What is the purpose of YARN?

Lambda Architecture

- What is streaming and batching?
- What is the upside of streaming vs batching?
- What is the difference between lambda and kappa architecture?
- Can you sync the batch and streaming layer and if yes how?

Python

- Difference between list tuples and dictionary

Data Warehouse & Data Lake

- What is a data lake?
- What is a data warehouse?
- Are there data lake warehouses?
- Two data lakes within single warehouse?
- What is a data mart?
- What is a slow changing dimension (types)?
- What is a surrogate key and why use them?

APIs (REST)

- What does REST mean?
- What is idempotency?
- What are common REST API frameworks (Jersey and Spring)?

Apache Spark

- What is an RDD?
- What is a dataframe?
- What is a dataset?
- How is a dataset typesafe?
- What is Parquet?
- What is Avro?
- Difference between Parquet and Avro
- Tumbling Windows vs. Sliding Windows
- Difference between batch and stream processing
- What are microbatches?

MapReduce

- What is a use case of mapreduce?
- Write a pseudo code for wordcount
- What is a combiner?

Docker & Kubernetes

- What is a container?
- Difference between Docker Container and a Virtual PC
- What is the easiest way to learn kubernetes fast?

Data Pipelines

- What is an example of a serverless pipeline?
- What is the difference between at most once vs at least once vs exactly once?
- What systems provide transactions?
- What is a ETL pipeline?

Airflow

- What is a DAG (in context of airflow/luigi)?
- What are hooks/is a hook?
- What are operators?
- How to branch?

DataVisualization

- What is a BI tool?

Security/Privacy

- What is Kerberos?
- What is a firewall?
- What is GDPR?
- What is anonymization?

Distributed Systems

- How clusters reach consensus (the answer was using consensus protocols like Paxos or Raft). Good I didnt have to explain paxos
- What is the cap theorem / explain it (What factors should be considered when choosing a DB?)
- How to choose right storage for different data consumers? It's always a tricky question

Apache Flink

- What is Flink used for?
- Flink vs Spark?

GitHub

- What are branches?
- What are commits?
- What's a pull request?

Dev/Ops

- What is continuous integration?
- What is continuous deployment?
- Difference CI/CD

Development / Agile

- What is Scrum?
- What is OKR?
- What is Jira and what is it used for?