

## Part V

# 1001 Data Engineering Interview Questions

## 33 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's the difference between IaaS, PaaS and SaaS
- How do you move from the ingest layer to the Consumption layer? (In Serverless)
- Whats the difference between cloud and edge and on-premises
- What is edge computing

## 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 consumer groups
- What is a producer

## Coding

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

## Hadoop

- What File Formats can you use in Hadoop
- What's the difference between a name node 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's 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 Datalakes within single warehouse?
- What is a data maart?
- 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's an RDD
- What is a dataframe
- What is a dataset
- How is a dataset typesafe
- What is Parquet
- What's Avro
- Difference between Parquet and Avro
- Tumbling Windows Vs. Sliding Windows
- Difference between batch and stream processing
- What are microbatches

## MapReduce

- What's 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 s the easiest way to learn kubernetes fast

## Data Pipelines

- What is an example of a serverless pipeline
- What's 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?

## DataViszualization

- What's a BI tool

## Security/Privacy

- What is Kerberos
- What is a firewall
- Whats GDPR?
- What's 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