category	question	company	difficulty	type
SQL	Write a SQL query to find the time period when the most people were online, measured in seconds.	Amazon	Medium	Coding
SQL	Given a large table with 3 columns (datetime, employee, customer_response), find the top 10 employees with the most phone nun	Amazon	Hard	Coding
SQL	Calculate first day retention rate - players who log in the 2nd day immediately after first login.	Amazon	Medium	Coding
SQL	Write a SQL query to get total revenue generated by each subscriber in the year 2014.	Amazon	Medium	Coding
SQL	Calculate the median number of searches a person made last year using summary table.	Google	Hard	Coding
SQL	Calculate the sum of odd-numbered and even-numbered measurements separately for a particular day.	Google	Medium	Coding
SQL	Write a SQL query that returns the 2nd-highest salary in the engineering department.	Netflix	Easy	Coding
SQL	Write a query to find customers who placed more than three orders in the past year.	Microsoft	Medium	Coding
SQL	Write SQL query to find ranks of employee salaries, ordered by department.	Microsoft	Medium	Coding
SQL	Write a query to return all neighborhoods that have 0 users (LEFT JOIN with IS NULL).	Microsoft	Medium	Coding
SQL	Calculate average vacant days across AirBnbs in 2021 for active properties only.	Airbnb	Hard	Coding
SQL	Get average rating of each Airbnb property listing per month.	Airbnb	Medium	Coding
SQL	Find top 10 listings with most bookings by joining hosts and bookings tables.	Airbnb	Medium	Coding
SQL	Write a query to get list of customers with their last purchase details.	Google	Medium	Coding
SQL	How do you optimize a slow SQL query? Explain query tuning process.	Google	Hard	Theory
Python	Write a function to find non duplicate numbers preserving order: [1,1,3,2,5,6,5] → [1,3,2,5,6]	Amazon	Medium	Coding
Python	Given a json object with nested objects, write a function that flattens all objects to single key-value dictionary.	Amazon	Hard	Coding
Python	Write code to find maximum combinations of infinite coins {1,2,5} that add up to 20 rupees.	Amazon	Hard	Coding
Python	Extract unique values from a dictionary where values occur only once.	Uber	Medium	Coding
Python	What's the difference between list, tuple, and set in Python for data pipelines?	Multiple Compan	Easy	Theory
Python	Explain Python modules used in data engineering: Pandas, NumPy, Requests, SQLAlchemy.	Google	Easy	Theory
Python	How would you optimize a data pipeline written in Python using vectorized operations?	Multiple Compan	Medium	Theory
Python	Write Python code to rank hosts based on number of beds using pandas groupby.	Airbnb	Medium	Coding
Python	What are generators in Python and how are they useful in large data processing?	Netflix	Medium	Theory
Python	How do you handle large CSV files that don't fit in memory using Python?	Spotify	Medium	Theory
Python	Explain the Global Interpreter Lock (GIL) and its impact on data processing.	Multiple Compan	Hard	Theory
Python	Write a Python class for a data pipeline with error handling and logging.	Stripe	Hard	Coding
Python	How do you implement multithreading vs multiprocessing for data processing in Python?	Uber	Hard	Theory
Python	Write code to detect longest patterns in song plays from user listening data.	Spotify	Hard	Coding
Python	How do you handle missing data and data cleaning in pandas efficiently?	Meta	Medium	Theory
Data Modeling	Design a relational database schema for a ride-sharing app like Uber.	Amazon	Hard	Design
Data Modeling	Design a database schema for an e-commerce platform with products, orders, and customers.	Amazon	Medium	Design
Data Modeling	Present a design of a gaming company database with players, games, and achievements.	Meta	Medium	Design
Data Modeling	Design a fact table for retail sales database with appropriate measures and dimensions.	Multiple Compan	Medium	Design
Data Modeling	Explain difference between Star Schema and Snowflake Schema with examples.	Multiple Compan	Medium	Theory

Data Modeling	Design a dimension table for customer entity with at least five attributes.	Multiple Compan	Easy	Design
Data Modeling	How do you handle slowly changing dimensions (SCDs) Type 1, 2, and 3?	Microsoft	Hard	Theory
Data Modeling	Design a schema to keep track of customer address changes over time.	Amazon	Medium	Design
Data Modeling	What are factless fact tables and when would you use them?	Multiple Compan	Medium	Theory
Data Modeling	Explain normalization vs denormalization trade-offs in data modeling.	Google	Medium	Theory
Data Modeling	Design a data model for tracking user behavior on a streaming platform like Netflix.	Netflix	Hard	Design
Data Modeling	How do you model many-to-many relationships in dimensional modeling?	Multiple Compan	Medium	Theory
Data Modeling	Design ERD for a social media platform with users, posts, and interactions.	Meta	Hard	Design
Data Modeling	What is dimensional modeling according to Ralph Kimball methodology?	Multiple Compan	Medium	Theory
Data Modeling	Design a data model for IoT sensor data with time-series considerations.	Multiple Compan	Hard	Design
ETL/Pipelines	Design an end-to-end ETL pipeline that ingests payment data into analytics warehouse.	Stripe	Hard	Design
ETL/Pipelines	How would you design a data pipeline for ingesting data with increasing volume?	Amazon	Medium	Design
ETL/Pipelines	Design data pipeline for hourly user analytics with batch and streaming options.	Uber	Hard	Design
ETL/Pipelines	How do you handle data quality issues and validation in data pipelines?	Netflix	Medium	Theory
ETL/Pipelines	How would you manage incremental data loads in a data warehouse?	Multiple Compan	Medium	Theory
ETL/Pipelines	When should you consider streaming over batch processing?	Netflix	Medium	Theory
ETL/Pipelines	Explain the key stages in ETL process: Extract, Transform, Load.	Multiple Compan	Easy	Theory
ETL/Pipelines	How do you implement error handling and recovery mechanisms in ETL processes?	Multiple Compan	Medium	Theory
ETL/Pipelines	What are the differences between ETL and ELT approaches?	Multiple Compan	Medium	Theory
ETL/Pipelines	How do you handle schema changes in source systems during ETL?	Microsoft	Medium	Theory
ETL/Pipelines	Design a real-time data pipeline to process payment transactions and update accounts.	Stripe	Hard	Design
ETL/Pipelines	How do you optimize slow-running ETL jobs for better performance?	Amazon	Medium	Theory
ETL/Pipelines	How do you implement data lineage tracking in ETL pipelines?	Multiple Compan	Hard	Theory
ETL/Pipelines	Design ETL process to collect and process tickets in real-time manner.	Amazon	Hard	Design
ETL/Pipelines	How do you handle late-arriving data in streaming ETL pipelines?	Netflix	Hard	Theory
System Design	Design a scalable data processing system for Netflix's recommendation engine.	Netflix	Hard	Design
System Design	Design a system similar to BookMyShow handling concurrency and fault tolerance.	Uber	Hard	Design
System Design	Design a data warehouse to capture and analyze sales data.	Amazon	Medium	Design
System Design	Design a real-time analytics platform for a ride-sharing company like Uber.	Uber	Hard	Design
System Design	Design distributed storage system for petabyte-scale data.	Google	Hard	Design
System Design	Design monitoring and alerting system for data pipelines.	Multiple Compan	Medium	Design
System Design	Design data lake architecture for storing multi-structured data.	Amazon	Hard	Design
System Design	Design API for high-frequency trading data with low latency requirements.	Multiple Compan	Hard	Design
System Design	Design caching strategy for frequently accessed data in distributed system.	Multiple Compan	Medium	Design
System Design	Design backup and disaster recovery strategy for critical data systems.	Multiple Compan	Medium	Design

System Design	Design A/B testing platform for data-driven product decisions.	Meta	Hard	Design
System Design	Design real-time fraud detection system for payment processing.	Stripe	Hard	Design
System Design	Design data platform for processing IoT sensor data at scale.	Multiple Compan	Hard	Design
System Design	Design search and recommendation system for e-commerce platform.	Amazon	Hard	Design
System Design	Design data governance framework for enterprise data management.	Multiple Compan	Medium	Design
Big Data/Spark	What is PySpark and how does it differ from Apache Spark?	Multiple Compan	Easy	Theory
Big Data/Spark	Explain concept of RDD in PySpark and how to create one.	Multiple Compan	Medium	Theory
Big Data/Spark	What is SparkSession and how do you create one in PySpark?	Multiple Compan	Easy	Theory
Big Data/Spark	Explain concept of lazy evaluation in PySpark with examples.	Multiple Compan	Medium	Theory
Big Data/Spark	What are broadcast variables in PySpark and when to use them?	Multiple Compan	Medium	Theory
Big Data/Spark	How do you optimize PySpark jobs for better performance?	Netflix	Hard	Theory
Big Data/Spark	What is role of Catalyst optimizer in PySpark?	Multiple Compan	Medium	Theory
Big Data/Spark	Write PySpark code to perform groupBy operation and calculate average of column.	Multiple Compan	Medium	Coding
Big Data/Spark	What are UDFs in PySpark? Write example and demonstrate usage.	Multiple Compan	Medium	Coding
Big Data/Spark	Explain difference between DataFrame and RDD in PySpark.	Multiple Compan	Medium	Theory
Big Data/Spark	Write PySpark code to join two DataFrames and explain different join types.	Multiple Compan	Medium	Coding
Big Data/Spark	What is PySpark Streaming and how to stream data using TCP/IP protocol?	Multiple Compan	Hard	Theory
Big Data/Spark	Explain the architecture of Hadoop and its core components (HDFS, MapReduce, YARN).	Netflix	Medium	Theory
Big Data/Spark	How do you handle data skew in Spark applications?	Uber	Hard	Theory
Big Data/Spark	Write PySpark code to perform window functions for calculating running totals.	Multiple Compan	Hard	Coding
Data Warehousir	What are OLTP vs OLAP systems and their impact on database design?	Multiple Compan	Medium	Theory
Data Warehousir	What is materialized view in BigQuery and when to use it?	Google	Medium	Theory
Data Warehousir	Explain concept of data lake vs data warehouse differences.	Multiple Compan	Medium	Theory
Data Warehousir	What are components of data warehouse architecture (staging, integration, access)?	Multiple Compan	Medium	Theory
Data Warehousir	How do you handle schema evolution in data warehouse environment?	Microsoft	Medium	Theory
Data Warehousir	What are best practices for data warehouse performance optimization?	Multiple Compan	Medium	Theory
Data Warehousir	Explain concept of data mart vs data warehouse and when to use each.	Multiple Compan	Easy	Theory
Data Warehousir	What are different types of indexes in data warehouse and their use cases?	Multiple Compan	Medium	Theory
Data Warehousir	How do you implement data retention policies in warehouse environment?	Multiple Compan	Medium	Theory
Data Warehousir	What are partitioning strategies for large tables in data warehouse?	Multiple Compan	Hard	Theory
Behavioral	Tell me about the most challenging data engineering project you worked on.	Amazon	Medium	Behavioral
Behavioral	Describe a time you had to optimize a data pipeline for performance. What was your approach?	Netflix	Medium	Behavioral
Behavioral	How do you prioritize multiple competing data engineering tasks and deadlines?	Amazon	Medium	Behavioral
Behavioral	Tell me about a time you applied judgment to a data decision when complete data was not available.	Amazon	Medium	Behavioral
Behavioral	How would you convey complex data insights to non-technical stakeholders?	Uber	Medium	Behavioral

Behavioral	Describe a situation where you had to collaborate with cross-functional teams on a data project.	Meta	Medium	Behavioral
Behavioral	Tell me about a time you made a mistake in a data project and how you handled it.	Amazon	Medium	Behavioral
Behavioral	How do you stay updated with the latest data engineering technologies and trends?	Multiple Compan	Easy	Behavioral
Behavioral	Describe a time you had to learn a new technology quickly for a project.	Spotify	Medium	Behavioral
Behavioral	How do you ensure data quality and reliability in your engineering projects?	Stripe	Medium	Behavioral