Aarte GCP Data Engineer Interview Guide – Experienced 1.5+

Technical Round 1:

1. **Tell us about your technical experience?**
2. **SQL and Functions**:
    - What is the difference between ROW_NUMBER(), RANK(), and DENSE_RANK() functions?
    - What is the difference between UNION and UNION ALL? Which one is faster and why?
    - List the different types of joins in SQL.
    - Given a table with 10 records and another with 4 records, how many records result from a cross join?
3. **Python and CSV Files**:
    - What are the different ways to read a CSV file in Python?
    - How do you interact with Google BigQuery using Python?
    - How can you automate data insertion into BigQuery using Python?
4. **Data Manipulation**:
    - How to merge two tables with identical structures into one?
5. **BigQuery**:
    - What types of columns support PARTITION_BY in BigQuery?
    - Where is the PARTITION_BY option in the BigQuery UI?
    - Can you modify a partitioned table into a non-partitioned one and vice-versa? How?
    - Explain how to flatten a multi-level nested JSON file while loading it into BigQuery.
6. **SQL Queries with Tables**:
    - Customer table (customer_id, name, dob, address, contact_number)
    - Sales table (product_id, customer_id, product_name, quantity, price, purchase_date)
    - Query examples:
        - Calculate the total sales amount for customers born between 1998-01-15 and 2000-01-15.
        - Identify the top 5 customers with the highest purchases in the last quarter.

7. **BigQuery Casting and Indexes**:
    - How to cast an integral column to a string in BigQuery and vice-versa?
    - What is the difference between SAFE_CAST() and CAST()?
    - Does BigQuery support indexes? If not, why?
8. **Data Warehousing**:
    - Have you worked on Data Warehousing projects?

Technical Round 2: Focusing on GCP Services.

Dataflow

1. Explain the key components of Apache Beam in the context of Google Dataflow.
2. How do you optimize resource allocation in a Dataflow job to reduce costs?
3. Describe the use of side inputs in Dataflow.
4. Explain the purpose of windowing and triggering in streaming data pipelines.
5. How would you handle a large-scale data shuffle in a Dataflow pipeline?

Dataproc

6. What are the advantages of using Dataproc over a traditional Hadoop setup?
7. Explain the concept of preemptible VMs in Dataproc and their cost implications.
8. How do you configure autoscaling for a Dataproc cluster?
9. Describe how Dataproc integrates with BigQuery for processing large datasets.
10. How would you debug a failing Spark job running on Dataproc?

Data Fusion

11. What is the purpose of Data Fusion in building ETL pipelines?
12. Describe how to configure a custom plugin in Data Fusion.
13. How would you schedule a recurring pipeline in Data Fusion?
14. Explain the difference between batch and streaming data processing in Data Fusion.
15. How do you monitor and troubleshoot data pipeline failures in Data Fusion?

Cloud Composer

16. Explain the role of Airflow DAGs in Cloud Composer.
17. How do you manage dependencies between tasks in a Cloud Composer DAG?
18. Describe how to set up retries and timeout for tasks in Cloud Composer.
19. Explain the concept of XComs in Airflow and their use in Cloud Composer.
20. How would you secure sensitive credentials in Cloud Composer workflows?

By thoroughly preparing for these technical and system-level questions, you'll be well-equipped to succeed in the Aarete Data Engineer interview process. Good luck!

Glassdoor Aarate Review –

https://www.glassdoor.co.in/Reviews/AArete-Reviews-E406227.htm

Aarate Careers –

https://areteir.com/careers/

Subscribe to my YouTube Channel for Free Data Engineering Content –

https://www.youtube.com/@shubhamwadekar27

Connect with me here –

https://bento.me/shubhamwadekar

Checkout more Interview Preparation Material on –

https://topmate.io/shubham_wadekar