
pwc

Introduction

My Full name is Anurag Srivastava, i have been working as Data Engineer from past 3 years, i started my career in 2021, where i got mapped to a project of Cloud Data Warehouse that was basically a data migration project, where i got opportunities to work in 2 different teams. Both the team i worked was core Data Team.

Then explain about the project description in brief, quantify the data example worked on loading of 975 PB of Data.

Roles and Responsibilities

As mentioned i worked in 2 Teams -:

DDL Team-: where we were responsible for conversion of the objects like table,views, procedures, macros, triggers to GCP compatible and also support the cross functional teams if there were any issues with the testing and ETL jobs, or object definitions.

DMU Team-: Here we were responsible to load the history and incremental data, there were tables divided as Full load(<=100GB) and Full Large Loa(>100GB), and also help the teams for incremental load of data in batch. Also helped cross functional teams for data validation and data analysis.

Follow-up Questions on projects

1. What was the volume of data you were loading daily?
2. Number of Objects you were responsible to load and convert?
3. How you were extracting the data?
4. How were your transforming the data?
5. What all Steps you followed to load the data?
6. How was structure present in BigQuery?

SQL Questions

Write a SQL Query to find the latest and the oldest record from SCD Table.

sql Copy code

```sql
SELECT * FROM (
    -- Latest record
    SELECT *
    FROM SCD_Table
    ORDER BY effective_date DESC
    LIMIT 1

    UNION ALL

    -- Oldest record
    SELECT *
    FROM SCD_Table
    ORDER BY effective_date ASC
    LIMIT 1
) AS Latest_Oldest_Record;
```

SQL Questions

What is SCD and how it is used?

SCD stands for "Slowly Changing Dimension." It's a concept used in data warehousing to manage changes to dimensional data over time. Dimensional data typically represents descriptive attributes of business entities, such as customers, products, or locations.

1. TYPE 1 SCD (Overwrite)
2. TYPE 2 SCD (Add New Row)
3. TYPE 3 SCD (Add New Column)

Python Questions

Write a Python Code to find the frequency of each word in String

python Copy code

```python
def word_frequency(string):
    # Convert the string to lowercase to ensure case-insensitive counting
    string = string.lower()

    # Split the string into words
    words = string.split()

    # Create an empty dictionary to store word frequencies
    word_freq = {}

    # Iterate through each word in the list
    for word in words:
        # Update the frequency of the word in the dictionary
        word_freq[word] = word_freq.get(word, 0) + 1

    return word_freq

# Example usage
input_string = "This is a sample string. It contains some words. This string is used
result = word_frequency(input_string)
print(result)
```

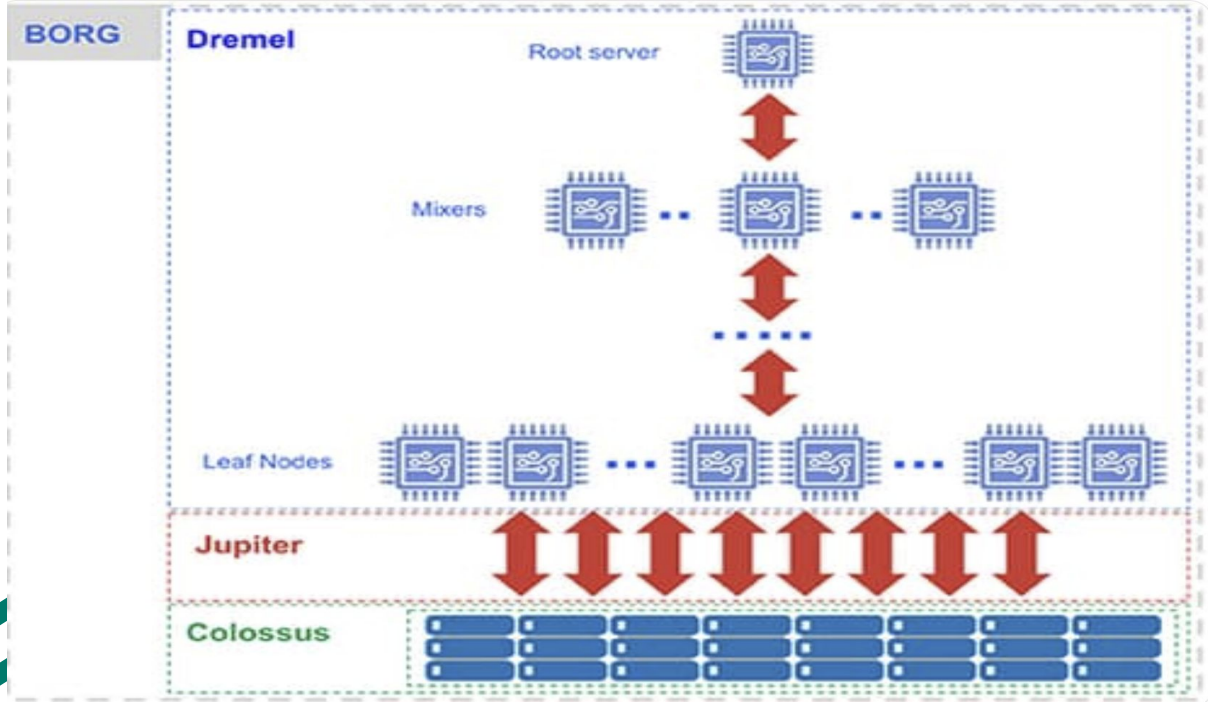
SQL Question

Query to write the second highest salary

sql Copy code

```sql
SELECT DISTINCT salary
FROM employees
ORDER BY salary DESC
LIMIT 1 OFFSET 1;
```

BigQuery Architecture



Knowledge about DE Tools

1. Apache Airflow
2. Apache Spark
3. DAGs
4. RDD

Pyspark Question

Write a Pyspark code to read a csv and get top 5 salary

python Copy code

```python
from pyspark.sql import SparkSession

# Create a SparkSession
spark = SparkSession.builder \
    .appName("Top 5 Employee Salaries") \
    .getOrCreate()

# Read the CSV file into a DataFrame
df = spark.read.csv("path/to/your/csv/file.csv", header=True, inferSchema=True)

# Sort the DataFrame by the 'salary' column in descending order and get the top 5 row
top_5_salaries = df.orderBy(df["salary"].desc()).limit(5)

# Show the result
top_5_salaries.show()
```