**Streaming POC - 2**

**Work Flow :**

**Phase 1:**

1. Fetching the API data from random API generator
2. Writes the API data into bronze table

**Phase 2:**

1. Apply flattening , joins on the data which is in bronze table and writes into silver table.

**Phase 3:**

1. Apply filter operations on the data which is in silver layer and writes into gold table.

**Phase 1:**

**Fetching the API data from random API generator and stored it in bronze:**

1.Five seconds once, we Fetch the data from random API generator using requests.get() method and stored as a json.

2.create a RDD using spark.context with Json . From that RDD , we can create a dataframe with schema and writes into bronze tables.

**Phase 2:**

**Apply flattening , joins on the data in bronze table:**

1.In the bronze table, create a dataframe by applying json function like explode to flatten the array type of data into columns.

2.In that dataframe, we can Apply withcolumn() to create a new columns (date, time ) from timestamp column.

3.Apply withcolumn() with regular expression to extract the username from ‘username’ column.

4.create a Another dataframe from Avro file.

5.create a dataframe by applying broadcast join on above two dataframe and writes into silver table.

**Phase 3:**

**Apply filters on the data which is in silver table:**

1.In the silver table, Apply the filter function on the nationality column with condition ‘not null’ as not\_Available customers.

2. Apply the filter function on the nationality column with condition ‘null’ as available\_customers.

3. Available customers and not\_available\_customers writes with partition by into gold table separately.

**REFERENCE: Work flow diagram link:**

<https://miro.com/welcomeonboard/VDg1QVpRekRHTGFKdjNGQnd6aTNJY21KMmd6WXFiRnlvVjlIQWdnNTlYUFo5TDRlWXR4VjBSRTRiYWZmRXJTVHwzNDU4NzY0NTc0Mjg0ODc1MjkwfDI=?share_link_id=367957530926>