Data Engineering Zoomcamp Project

Author: Anatolii Solovev

Problem statement

The goal of the project is to analyse Bandcamp's sales in respect to a particular country and date answering the next questions:

- Find top 5 countries by money paid in USD
- Compare money paid in these top 5 countries by date

The data set used is "1,000,000 Bandcamp sales" which can be found at: https://components.one/datasets/bandcamp-sales

Solution approach

Under the given project ELT process has been created. The data is retrieved, uploaded into S3 cloud storage as a parquet file, then transformed and uploaded into DWH for the further analysis and visualization.

Data Pipeline

It is assumed that pipeline can be run periodically depending on business needs. Hence, the ELT process assumes batch processing.

Technologies used

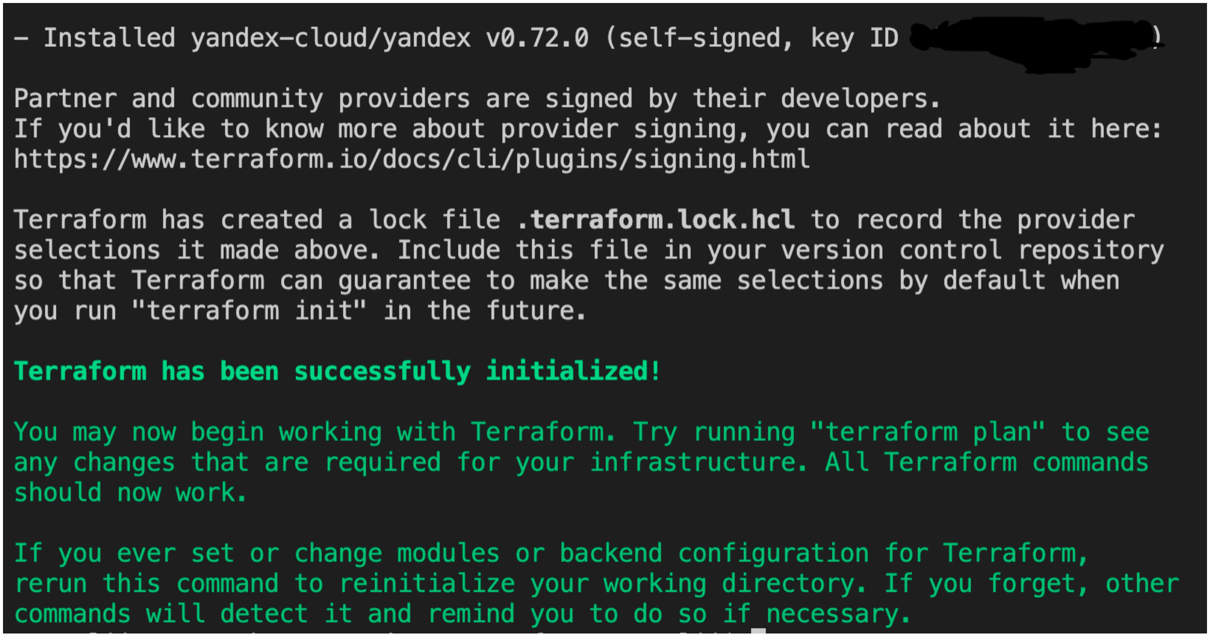
- Cloud: Yandex Cloud
- Infrastructure as code (IaC): Terraform
- Workflow orchestration: Airflow
- Data Warehouse: PostgreSQL
- Data transformation: Pandas

Part One: Infrastructure as Code

To create infrastructure for the project, I decided to use Yandex Cloud and Terraform.

Step 1

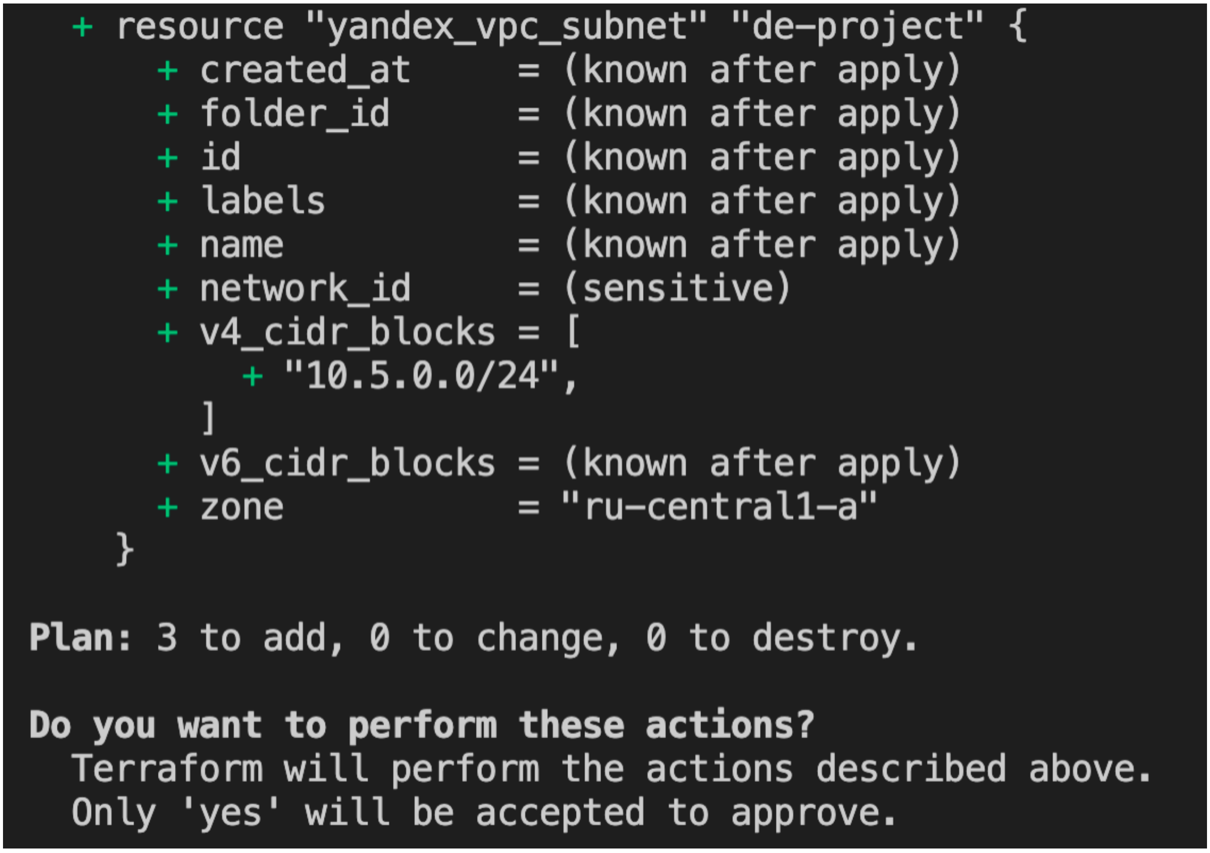
Init terraform with Yandex cloud prover running **terraform init**



```
- Installed yandex-cloud/yandex v0.72.0 (self-signed, key ID [redacted])

Partner and community providers are signed by their developers.
If you'd like to know more about provider signing, you can read about it here:
https://www.terraform.io/docs/cli/plugins/signing.html

Terraform has created a lock file .terraform.lock.hcl to record the provider
selections it made above. Include this file in your version control repository
so that Terraform can guarantee to make the same selections by default when
you run "terraform init" in the future.

Terraform has been successfully initialized!

You may now begin working with Terraform. Try running "terraform plan" to see
any changes that are required for your infrastructure. All Terraform commands
should now work.

If you ever set or change modules or backend configuration for Terraform,
rerun this command to reinitialize your working directory. If you forget, other
commands will detect it and remind you to do so if necessary.
```

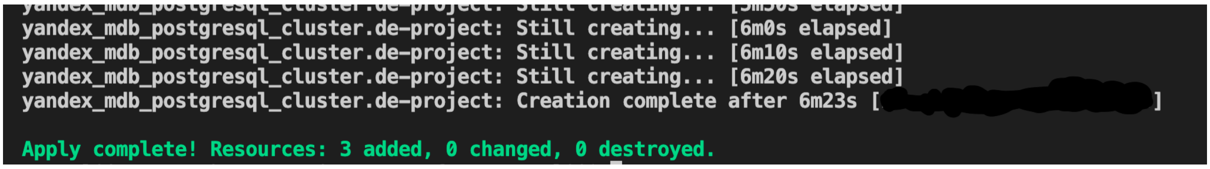
Step 2

Then, create infrastructure running **terraform apply -var-file="secret.tfvars"**



```
  + resource "yandex_vpc_subnet" "de-project" {
      + created_at     = (known after apply)
      + folder_id      = (known after apply)
      + id             = (known after apply)
      + labels         = (known after apply)
      + name           = (known after apply)
      + network_id     = (sensitive)
      + v4_cidr_blocks = [
          + "10.5.0.0/24",
        ]
      + v6_cidr_blocks = (known after apply)
      + zone           = "ru-central1-a"
    }

Plan: 3 to add, 0 to change, 0 to destroy.

Do you want to perform these actions?
  Terraform will perform the actions described above.
  Only 'yes' will be accepted to approve.
```

The structure has been created.



```
yandex_mdb_postgresql_cluster.de-project: Still creating... [5m50s elapsed]
yandex_mdb_postgresql_cluster.de-project: Still creating... [6m0s elapsed]
yandex_mdb_postgresql_cluster.de-project: Still creating... [6m10s elapsed]
yandex_mdb_postgresql_cluster.de-project: Still creating... [6m20s elapsed]
yandex_mdb_postgresql_cluster.de-project: Creation complete after 6m23s [[redacted]]

Apply complete! Resources: 3 added, 0 changed, 0 destroyed.
```

As I can be seen, all secrets are stored in a file called **secret.tfvars**

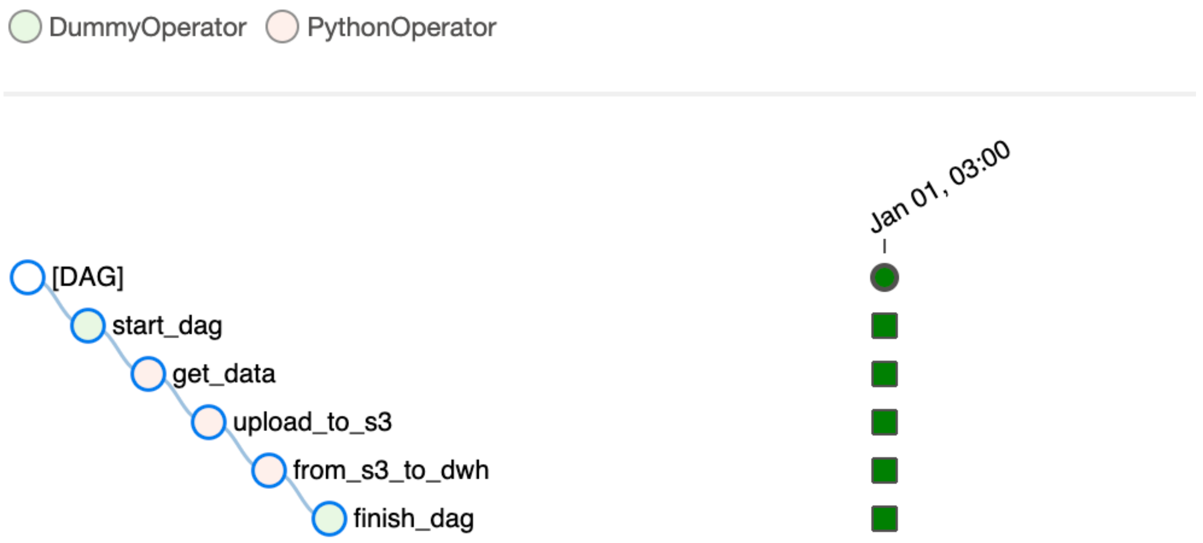
Part Two: Create data mart in DWH

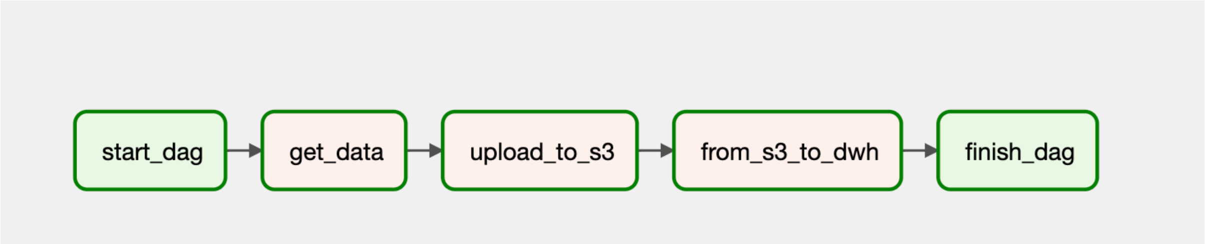
Since we suppose that data will be used for analytical purposes, I decided to create schema called “**marts**” with a table “**bandcamp_sales**”.
The table is partitioned by month assuming a lot of analytical queries will be made based on a particular month of sales. Partitioning by date helps make queries faster and effective if questions considering filtering or grouping by date will be asked.
Also, two indexes have been created. Going by the interest of making analysis filtered by countries and a particular slug type, two indexes on columns **country** and **slug_type** have been made respectively. Although questions related to slug types are not answered in the given project, they could be covered in the further analysis.

Part Three: Orchestration of data processing

The tool of orchestration is Apache Airflow. It has a DAG called project_dag.
The DAG has next steps:

- start_dag (a dummy DAG just to indicate DAG run)
- get_data (this DAG retrieves data stored locally and transform it into a temporarily parquet file stored locally as well)
- upload_to_s3 (this DAG uploads a parquet file into a S3 storage and deletes the parquet file)
- from_s3_to_dwh (this DAG reads data from S3 bucket, transforms it and inserts into DWH)
- finish_dag (a simple dummy DAG to indicate the pipeline is done)

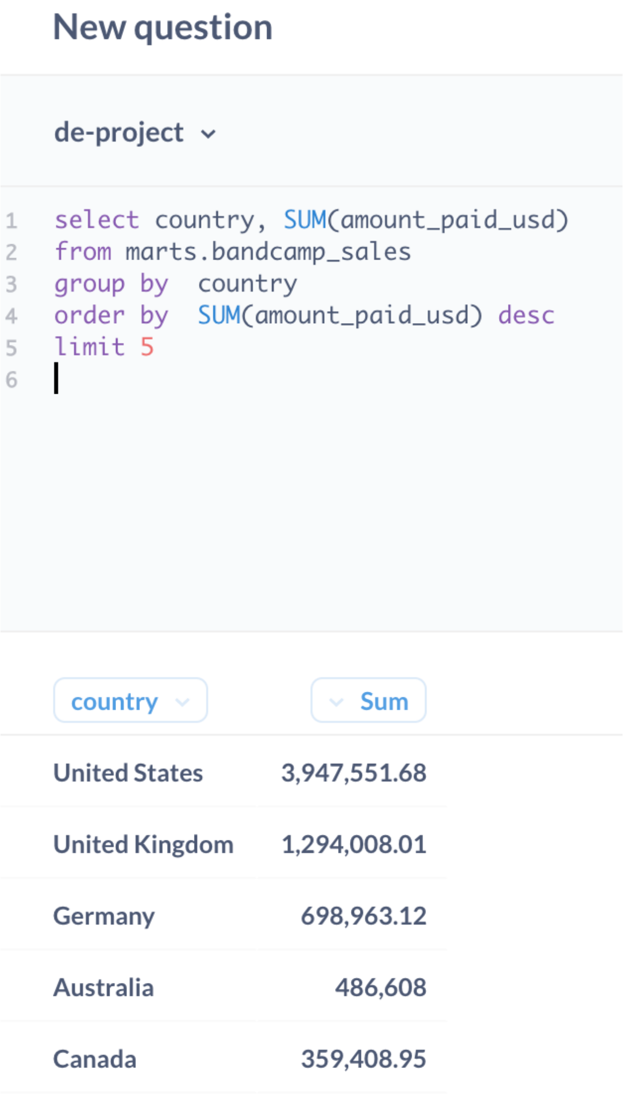




For data transformation python library pandas is used. The usage is based on the assumption a file will not be more than several gigabytes. Hence, it can be pre-processed in RAM directly. In this case, pandas can be a good solution. A function data_transformation transforms date from unix format to timestamp. Also, it creates an additional field date which retrieves date from timestamp. Moreover, it make trimming on string columns in a dataframe.

Part Four: Make visualisations

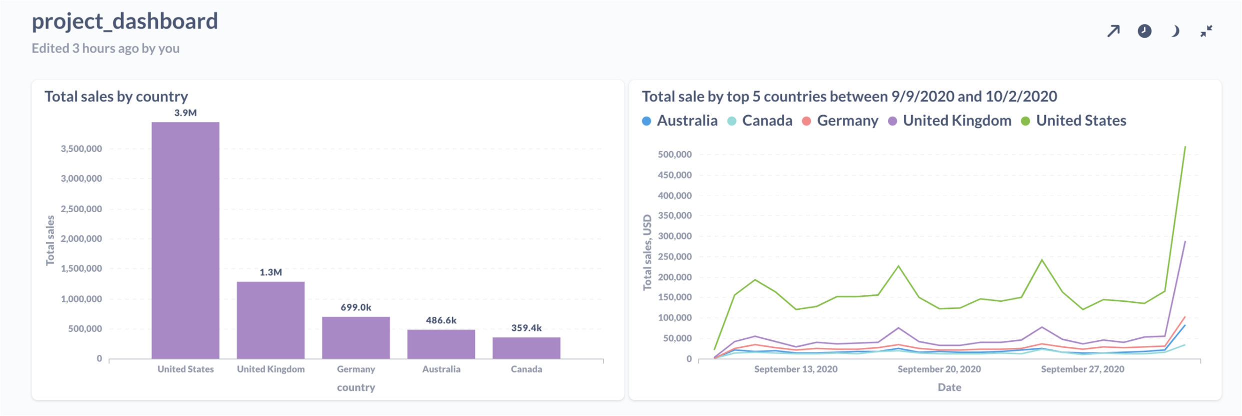
Answer for the first question can be found by an SQL query.



```
select country, SUM(amount_paid_usd)
from marts.bandcamp_sales
group by  country
order by  SUM(amount_paid_usd) desc
limit 5
```

country	Sum
United States	3,947,551.68
United Kingdom	1,294,008.01
Germany	698,963.12
Australia	486,608
Canada	359,408.95

Final dashboard looks like:



project_dashboard
Edited 3 hours ago by you
Total sales by country
3.9M
1.3M
699.0k
486.6k
359.4k
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Total sales
United States
United Kingdom
Germany
Australia
Canada
country
Total sale by top 5 countries between 9/9/2020 and 10/2/2020
● Australia ● Canada ● Germany ● United Kingdom ● United States
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Total sales, USD
September 13, 2020
September 20, 2020
September 27, 2020
Date