

Projet IPBD

# Etude d'une BD d'Anime

Tessa DEPAOLI, DE WEERD Xavier, FORT Alexandre

# Introduction

Mettre en place les outils vu en cours



Quelles données peut-on extraire d'une BD sur les Anime ?

# Introduction - Quelles données



# Sommaire

I- Choix du Data Set

II- Analyse du Data Set (python : pandas, seaborn)

III- Construction de la plateforme de Big Data

- A- Présentation des outils

- B- Préparation de la VM

- C- Hadoop, Hive

- D- Schéma de la BD

IV- Apache Superset

V- Google Cloud

# I- Choix du DataSet



Anime-database :

- anime\_id
- name
- synopsis
- rank
- genre
- ...

Users :

- mal\_id
- location
- gender
- birthday
- ...

Score :

- mal\_id
- anime\_id
- score

—————▶ ~3Go

## II - Analyse du DataSet



## II - Analyse du DataSet

Extraire la proportion de genres d'anime chaque année

Cowboy Bebop	8.78	Action, Adventure, Comedy, Drama, Sci-Fi, Space
Cowboy Bebop: Tengoku no Tobira	8.39	Action, Drama, Mystery, Sci-Fi, Space
Trigun	8.24	Action, Sci-Fi, Adventure, Comedy, Drama, Shounen

Séparation

<b>△ Premiered</b>	
The season and year when the anime premiered.	
Unknown	70%
Spring 2017	0%
Other (4422)	30%
Spring 1998	
Unknown	
Spring 1998	

## II - Analyse du DataSet

Premiered	
The season and year when the anime premiered.	
Unknown	70%
Spring 2017	0%
Other (4422)	30%

remplacé par -1

Action, Sci-Fi,  
Adventure, Comedy,  
Drama, Shounen

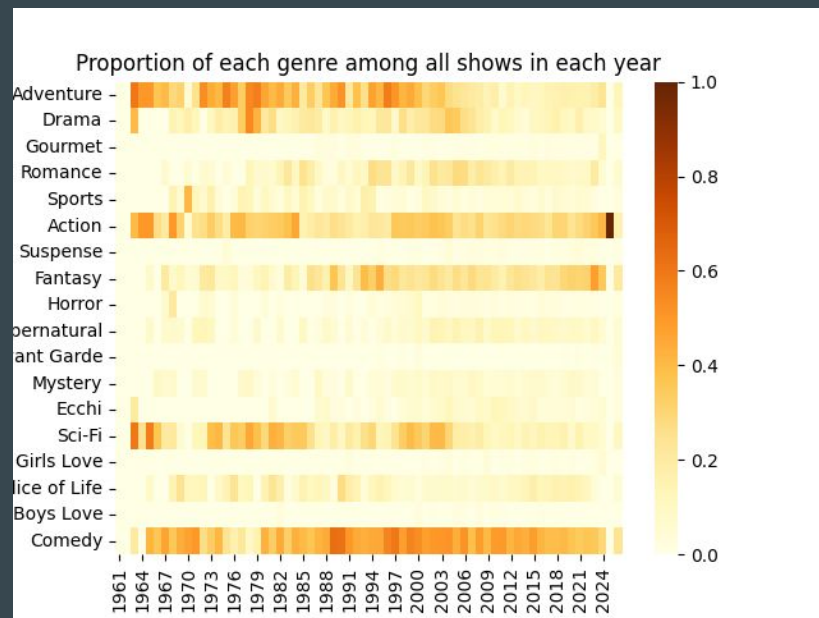
One-hot  
encoding :

Chaque genre =>  
nouvelle colonne  
(valeur 0 ou 1)

présence dans  
chaque anime



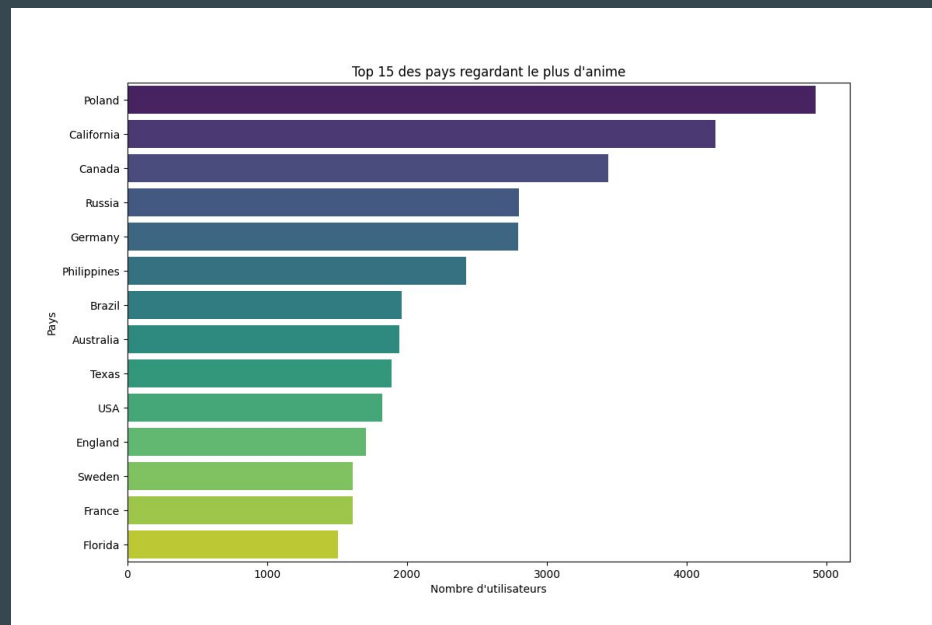
## II - Analyse du DataSet



## II - Analyse du DataSet

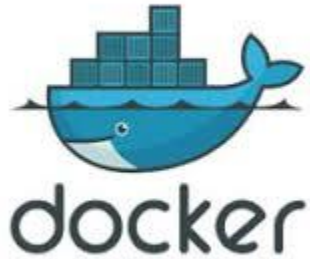
extraire les pays qui regardent le plus d'anime

the user.	
[null]	79%
Poland	0%
Other (150348)	21%
California	
Oslo, Norway	
Melbourne, Australia	



# III- Construction de la plateforme de Big Data

## A- Présentation des Outils



# III- Construction de la plateforme de Big Data

## B- Préparation de la VM

- Installation des paquets
- Mise en place du git
- Importation des données
- Préparation des conteneurs
- Préparation de Hadoop et de Hive

# III- Construction de la plateforme de Big Data

## C- Hadoop, Hive

- > HDFS : Import des données
- > Problème de connexion
- > Problème d'importation des données
  - Guillemets
  - Sauts de lignes

# III- Construction de la plateforme de Big Data

## D- Schéma de la DB

### Anime

- |                       |                    |                    |
|-----------------------|--------------------|--------------------|
| - anime_id INT        | - episodes FLOAT   | - duration STRING  |
| - name STRING         | - aired STRING     | - rating STRING    |
| - english_name STRING | - premiered STRING | - rank FLOAT       |
| - other_name STRING   | - status STRING    | - popularity INT   |
| - score FLOAT         | - producers STRING | - favorites INT    |
| - genres STRING       | - licensors STRING | - scored_by FLOAT  |
| - synopsis STRING     | - studios STRING   | - members INT      |
| - type STRING         | - source STRING    | - image_url STRING |

# III- Construction de la plateforme de Big Data

## D- Schéma de la DB

### Score

- anime\_id INT
- mal\_id INT
- score INT

# III- Construction de la plateforme de Big Data

## D- Schéma de la DB

- mal\_id INT
- username STRING
- gender STRING
- birthday TIMESTAMP
- location STRING
- joined TIMESTAMP
- days\_watched FLOAT
- mean\_score FLOAT

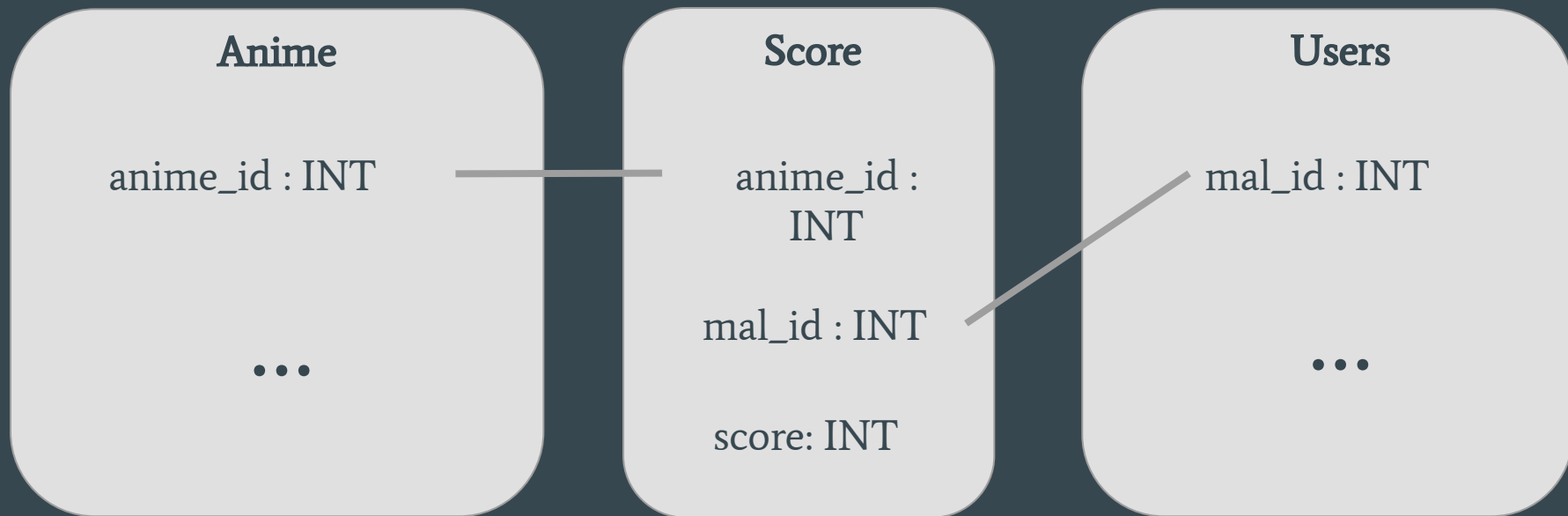
### Users

- watching FLOAT
- completed FLOAT
- on\_hold FLOAT
- dropped FLOAT
- plan\_to\_watch FLOAT
- total\_entries FLOAT
- rewatched FLOAT
- episodes\_watched FLOAT



# III- Construction de la plateforme de Big Data

## D- Schéma de la DB



# III- Construction de la plateforme de Big Data

## D- Schéma de la DB

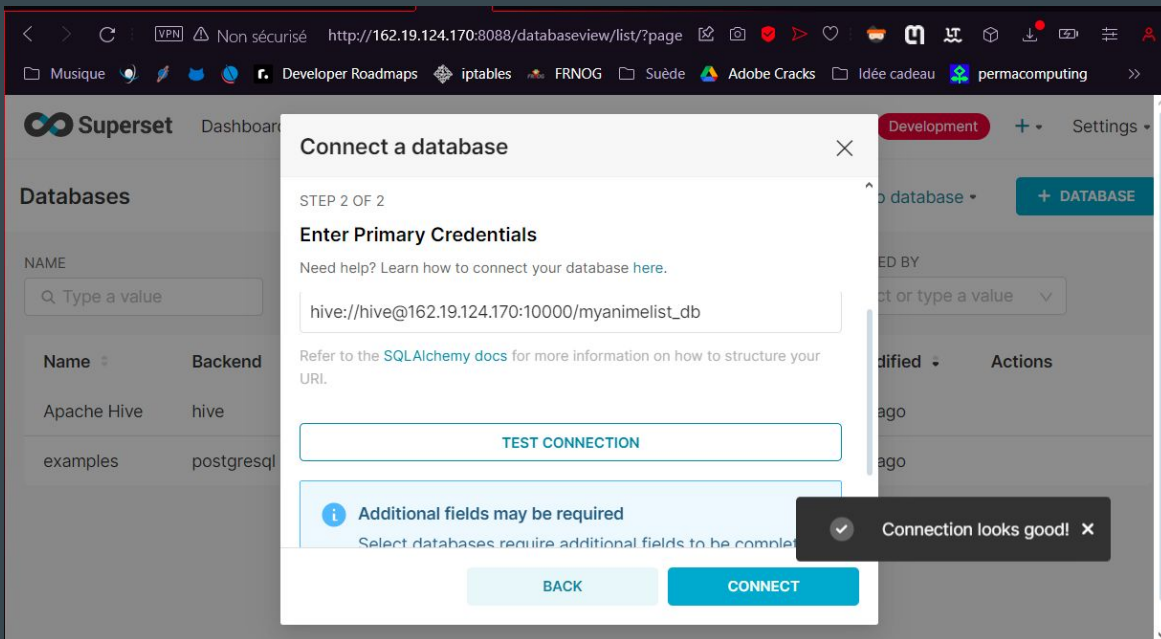
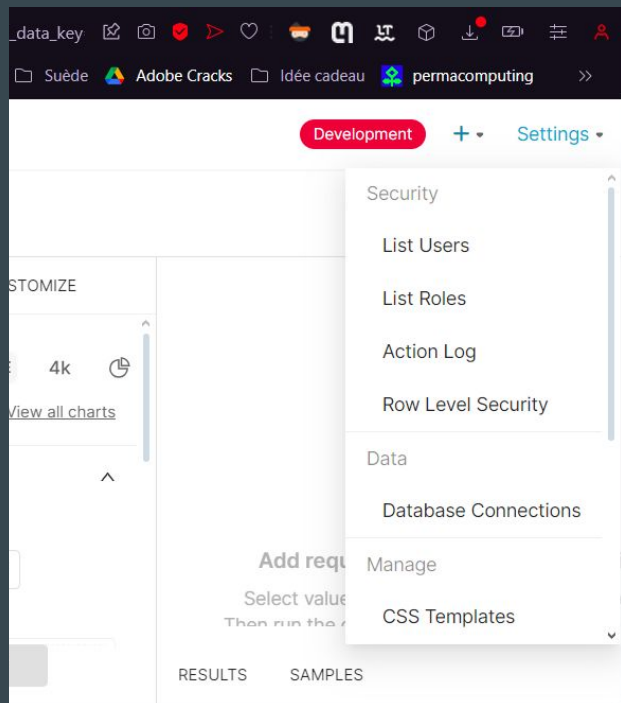
```
CREATE EXTERNAL TABLE IF NOT EXISTS users (  
    ...  
)  
ROW FORMAT SERDE  
'org.apache.hadoop.hive.serde2.OpenCSVSerde'  
WITH SERDEPROPERTIES (  
    "separatorChar" = ',',  
    "quoteChar"     = '"',  
    "escapeChar"    = '\\'  
)  
    STORED AS TEXTFILE  
    LOCATION '/dataset/users'  
    TBLPROPERTIES  
    ("skip.header.line.count"="1");
```

## IV- Apache Superset



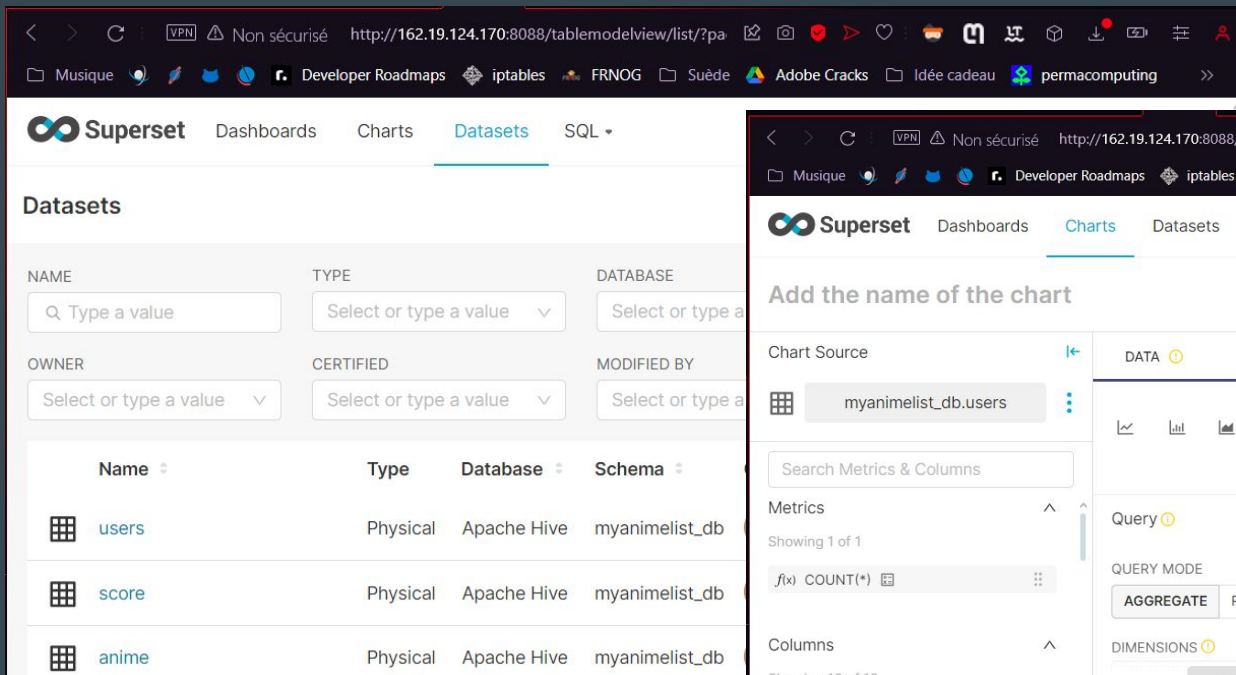
# IV- Apache Superset

## A- Connexion à Hive



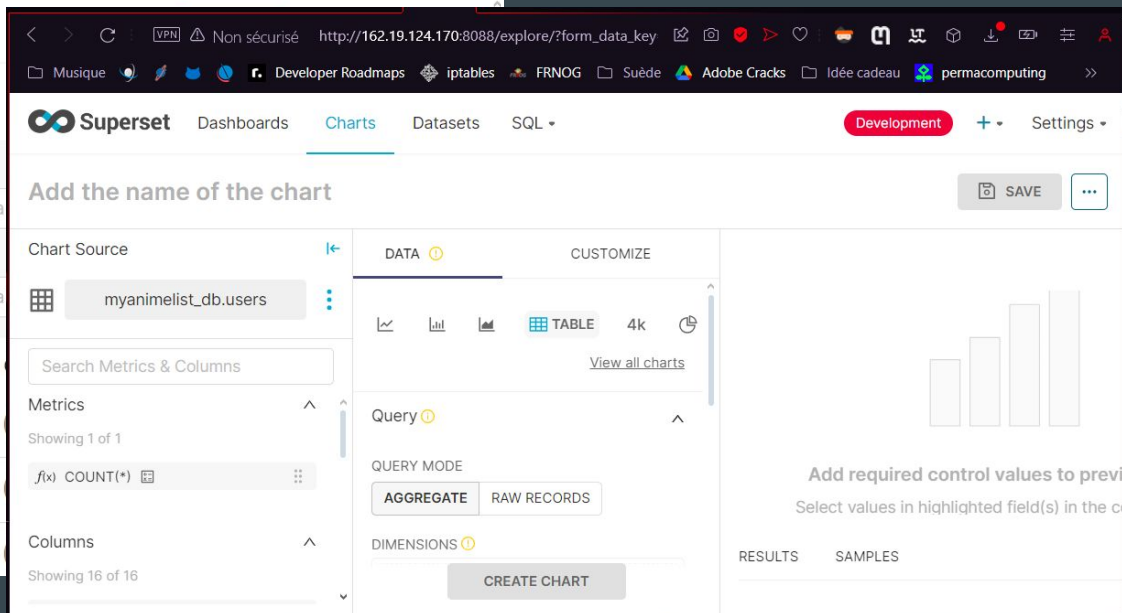
# IV- Apache Superset

## B- Ajout : datasets - chart - dashboards



The screenshot shows the Apache Superset web interface with the 'Datasets' tab selected. The page title is 'Datasets'. There are three input fields for filtering: 'NAME' (with a search icon and placeholder 'Type a value'), 'TYPE' (with a dropdown arrow and placeholder 'Select or type a value'), and 'DATABASE' (with a dropdown arrow and placeholder 'Select or type a value'). Below these are three more input fields for 'OWNER', 'CERTIFIED', and 'MODIFIED BY', each with a dropdown arrow and placeholder 'Select or type a value'. A table lists datasets with columns: Name, Type, Database, and Schema. The table contains three rows: 'users' (Physical, Apache Hive, myanimelist\_db), 'score' (Physical, Apache Hive, myanimelist\_db), and 'anime' (Physical, Apache Hive, myanimelist\_db).

Name	Type	Database	Schema
users	Physical	Apache Hive	myanimelist_db
score	Physical	Apache Hive	myanimelist_db
anime	Physical	Apache Hive	myanimelist_db



The screenshot shows the Apache Superset web interface with the 'Charts' tab selected. The page title is 'Add the name of the chart'. There are two tabs: 'DATA' and 'CUSTOMIZE'. The 'DATA' tab is active, showing a 'Chart Source' section with a dropdown menu set to 'myanimelist\_db.users'. Below this is a 'Search Metrics & Columns' input field. The 'Metrics' section shows 'Showing 1 of 1' with a metric 'COUNT(\*)'. The 'Columns' section shows 'Showing 16 of 16'. The 'QUERY MODE' section has two buttons: 'AGGREGATE' and 'RAW RECORDS'. The 'DIMENSIONS' section has a 'CREATE CHART' button. The 'CUSTOMIZE' tab is partially visible on the right, showing a bar chart and a 'View all charts' link.

# IV- Apache Superset

## B- Ajout : datasets - chart - dashboards

The screenshot displays the Apache Superset web interface. The top navigation bar includes the Superset logo, tabs for Dashboards, Charts (active), Datasets, and SQL, a 'Development' status indicator, and a 'Settings' dropdown. The main heading is 'Add the name of the chart'. Below this, the 'Chart Source' is set to 'myanimelist\_db.anime'. The 'Metrics' section shows 'COUNT(\*)'. The 'Columns' section shows 'Showing 24 of 24'. The 'Query' section is in 'AGGREGATE' mode. The 'CUSTOMIZE' tab is active, showing a 'TABLE' view with '4k' rows. An 'UPDATE CHART' button is at the bottom. On the right, an 'Unexpected error' message is displayed, indicating a 'SemanticException [Error 10025]: Expression not in GROUP BY key 'na''.

Superset Dashboards Charts Datasets SQL Development + Settings

Add the name of the chart SAVE ...

Chart Source myanimelist\_db.anime

Search Metrics & Columns

Metrics Showing 1 of 1

$f(x)$  COUNT(\*)

Columns Showing 24 of 24

DATA CUSTOMIZE

TABLE 4k View all charts

Query AGGREGATE RAW RECORDS

QUERY MODE

DIMENSIONS

UPDATE CHART

0 rows

**Unexpected error**

Error:  
TExecuteStatementResp(status=TStat  
de=3, infoMessages=  
["\*org.apache.hive.service.cli.Hiv  
on:Error while compiling statement  
SemanticException [Error 10025]: U  
Expression not in GROUP BY key 'na  
'org.apache.hive.service.cli.oper  
ion:toSQLException:Operation.java  
'org.apache.hive.service.cli.oper

RESULTS SAMPLES

## V- Google Cloud



Google  
Big Query



Looker

# V- Google Cloud

Google Cloud

My First Project

Tapez / pour rechercher des ressources, des document...

Recherche

1

Explorateur

AJOUTER

Saisissez un terme à rechercher

Afficher les ressources

N'AFFICHER QUE LES FAVORIS

vernal-reality-424000-s8

Requêtes

Notebooks

Canevas de données

Canevas de données part...

Canevas sans titre

Connexions externes

RÉSUMÉ

ACTIVITÉ

21 mai 2024

Enregistré - Tessa DePaoli 17:36

Enregistré - Tessa DePaoli 16:51

Enregistré - Tessa DePaoli 16:50

Enregistré - Tessa DePaoli 16:49

Canevas sans titre

ENREGISTRER

PARTAGER

EXPORTER EN TANT QUE NOTEBOOK

10 requêtes d'analyse les plus nombreuses

10 requêtes d'analyse les plus nombreuses

10 requêtes d'analyse les plus nombreuses

10 requêtes d'analyse les plus nombreuses

10 requêtes d'analyse les plus nombreuses

Visualisation

Visualisation

10 requêtes d'analyse les plus nombreuses

Requête	Nombre de requêtes
1. Analyse des données de base	10
2. Analyse des données de base	10
3. Analyse des données de base	10
4. Analyse des données de base	10
5. Analyse des données de base	10
6. Analyse des données de base	10
7. Analyse des données de base	10
8. Analyse des données de base	10
9. Analyse des données de base	10
10. Analyse des données de base	10

10 requêtes d'analyse les plus nombreuses

Requête	Nombre de requêtes
1. Analyse des données de base	10
2. Analyse des données de base	10
3. Analyse des données de base	10
4. Analyse des données de base	10
5. Analyse des données de base	10
6. Analyse des données de base	10
7. Analyse des données de base	10
8. Analyse des données de base	10
9. Analyse des données de base	10
10. Analyse des données de base	10

10 requêtes d'analyse les plus nombreuses

Requête	Nombre de requêtes
1. Analyse des données de base	10
2. Analyse des données de base	10
3. Analyse des données de base	10
4. Analyse des données de base	10
5. Analyse des données de base	10
6. Analyse des données de base	10
7. Analyse des données de base	10
8. Analyse des données de base	10
9. Analyse des données de base	10
10. Analyse des données de base	10

10 requêtes d'analyse les plus nombreuses

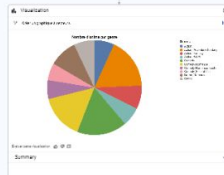
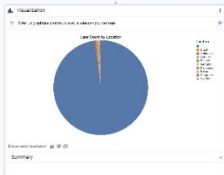
Requête	Nombre de requêtes
1. Analyse des données de base	10
2. Analyse des données de base	10
3. Analyse des données de base	10
4. Analyse des données de base	10
5. Analyse des données de base	10
6. Analyse des données de base	10
7. Analyse des données de base	10
8. Analyse des données de base	10
9. Analyse des données de base	10
10. Analyse des données de base	10

10 requêtes d'analyse les plus nombreuses

Requête	Nombre de requêtes
1. Analyse des données de base	10
2. Analyse des données de base	10
3. Analyse des données de base	10
4. Analyse des données de base	10
5. Analyse des données de base	10
6. Analyse des données de base	10
7. Analyse des données de base	10
8. Analyse des données de base	10
9. Analyse des données de base	10
10. Analyse des données de base	10

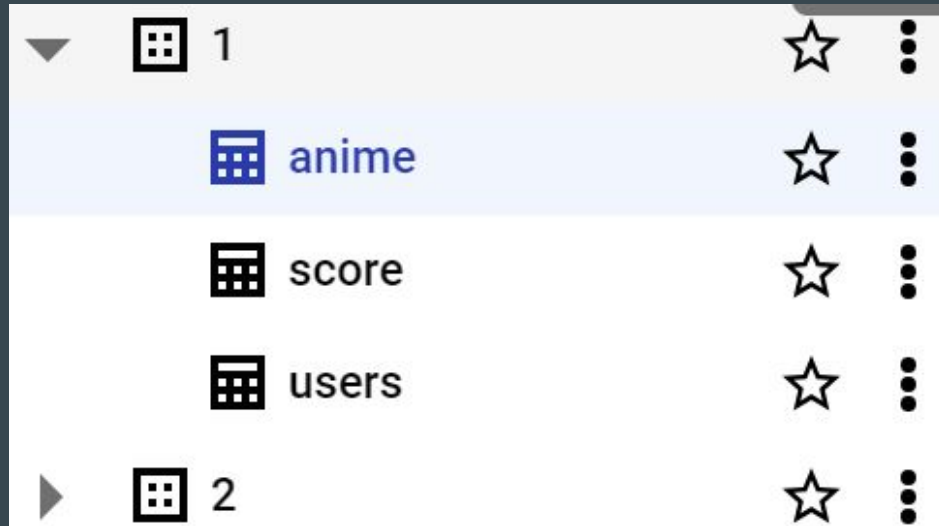
10 requêtes d'analyse les plus nombreuses

Requête	Nombre de requêtes
1. Analyse des données de base	10
2. Analyse des données de base	10
3. Analyse des données de base	10
4. Analyse des données de base	10
5. Analyse des données de base	10
6. Analyse des données de base	10
7. Analyse des données de base	10
8. Analyse des données de base	10
9. Analyse des données de base	10
10. Analyse des données de base	10





# V- Google Cloud



# V- Google Cloud

anime

REQUÊTE

PARTAGER

COPIER

INSTANTANÉ

SUPPRIMER

EXPORTER

SCHÉMA

DÉTAILS

APERÇU

TRAÇABILITÉ

PROFIL DE DONNÉES

QUALITÉ DES DONNÉES

Filtre

Saisissez le nom ou la valeur de la propriété

?

	Nom du champ	Type	Mode	Clé	Classement	Valeur par défaut	Tags avec stratégie	Description
<input type="checkbox"/>	anime_id	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Name	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	English name	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Other name	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Score	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Genres	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Synopsis	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Type	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Episodes	STRING	NULLABLE	-	-	-	-	-

MODIFIER LE SCHÉMA

AFFICHER LES RÈGLES D'ACCÈS AUX LIGNES

# V- Google Cloud

```
✓ SELECT
  Location,
  COUNT(`Mal ID`) AS user_count
✓ FROM
  `vernal-reality-424000-s8.1.users`
✓ GROUP BY
  Location
✓ ORDER BY
  user_count DESC
LIMIT 10;
```

Résultats de la requête



Ligne	Location	user_count
1	<i>null</i>	578484
2	Poland	2458
3	Germany	1706
4	Canada	1693
5	California	1506
6	Brazil	1056
7	Sweden	992
8	Singapore	942
9	Philippines	939
10	Finland	927

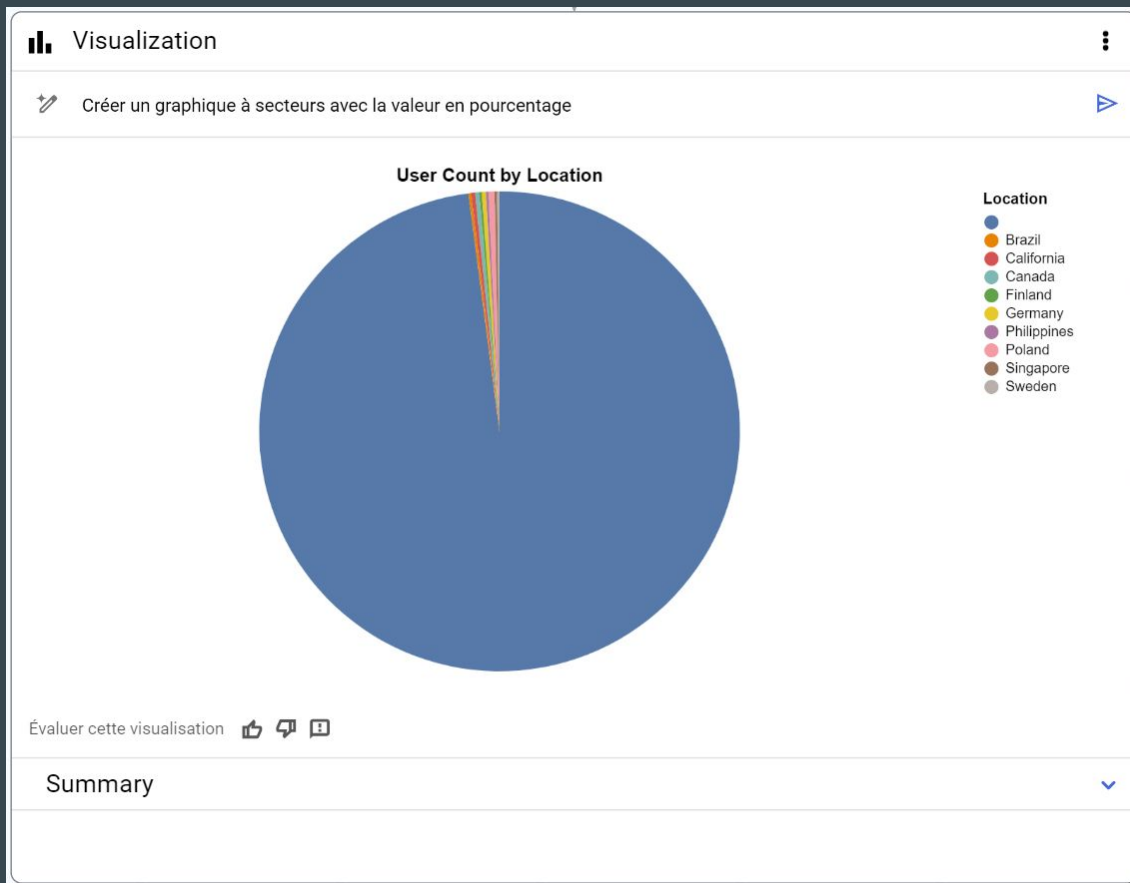
Créer un autre nœud de branche

INTERROGER CES RÉSULTATS

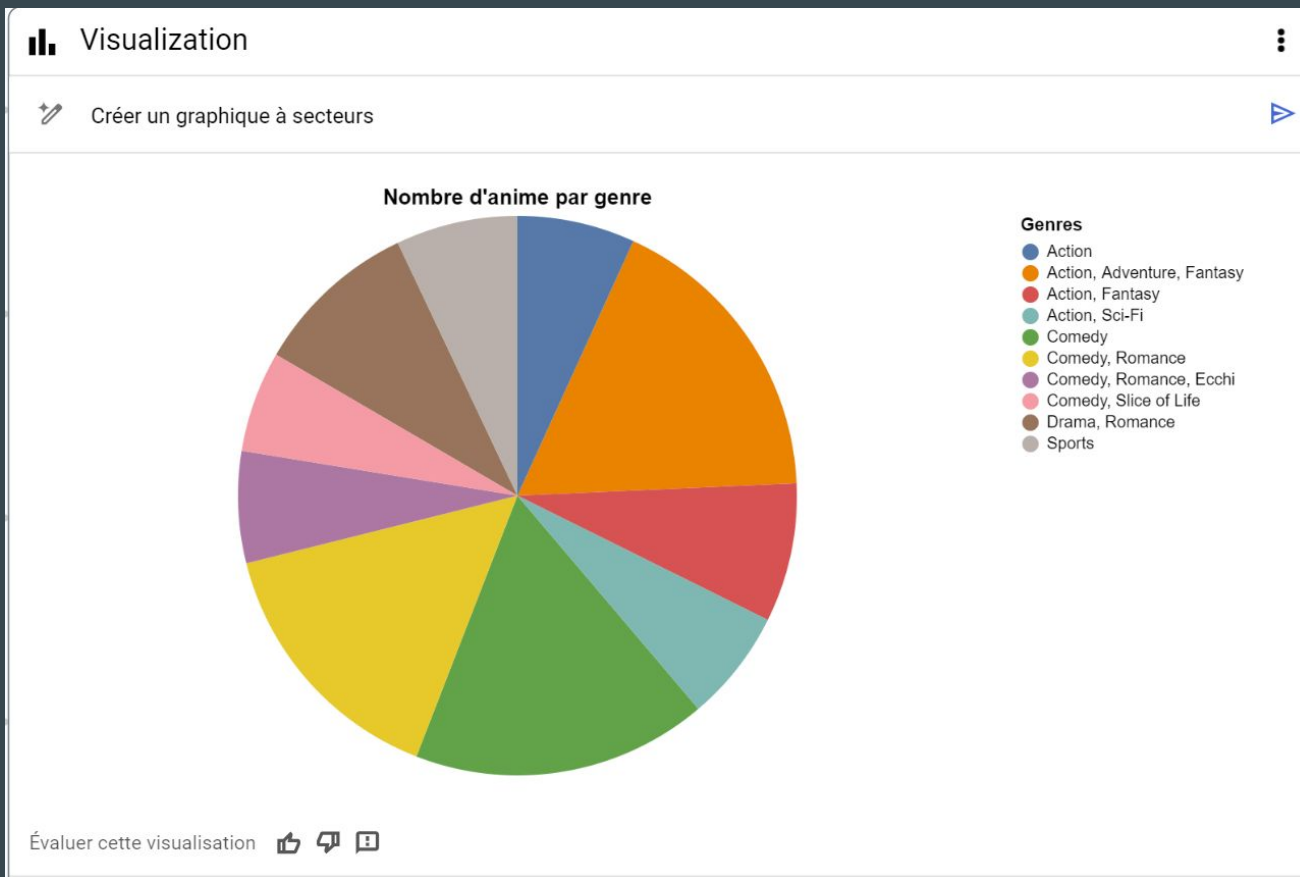
VISUALISER

JOIN

# V- Google Cloud



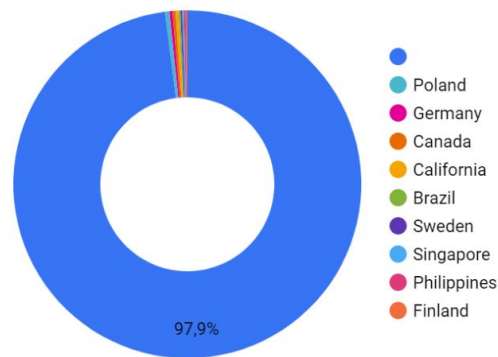
# V- Google Cloud



# V- Google Cloud

## Location/user

	Location	user_count ▾
1.	null	578484
2.	Poland	2458
3.	Germany	1706
4.	Canada	1693
5.	California	1506
6.	Brazil	1056
7.	Sweden	992
8.	Singapore	942
9.	Philippines	939
10.	Finland	927



# V- Google Cloud

**Reporting Looker Studio - 22/05/2024 07:46**

Fichier Modification Vue Insertion Page Organisation Ressource Aide

Enregistrer et partager

Rechercher

Ajouter une page Ajouter des données Ajouter un graphique Ajouter un sélecteur

Suspendre les mises à jour

+ Ajouter un filtre rapide

Réinitialiser

**anime**

**Table 1: anime\_id**

Name	anime_id
1. Awakening	110276
2. Azur Lane	93679
3. Utopia	76685
4. Sousseki	59083
5. Shijakuku Nichi	55735
6. Bokura no Seshuo Sengou	55734
7. Di Yi Xue	55733
8. Bu Xing Si: Yuan Qi	55732
9. Wu Nao Monu	55731
10. Energy	55730
11. Thailand	55729

1 - 50 / 24901

**Donut Chart 1: Genres**

Genre	Proportion
UNKNOWN	27%
Comedy	42%
Fantasy	5.7%
Avant Garde	1.9%
Mental	1.9%
Slice of Life	1.9%
Drama	1.9%
Adventure, Fantasy	1.9%
Comedy, Fantasy	1.9%
Autres	1.9%

**Table 2: Popularity**

Name	Popularity
1. 空 (Hot Sauce) (MINIMONSTER Rem...	0
2. エンジンゼロ	0
3. ○	10011
4. #OLIVE	2982
5. Ōka Uka	17749

1 - 100 / 24905

**Donut Chart 2: Popularity**

Popularity	Proportion
1.0	48.1%
12.0	15.7%
26.0	3.2%
52.0	3.2%
13.0	3.2%
3.0	3.2%
4.0	3.2%
Autres	3.2%

**Graphique**

CONFIGURER STYLE

Source de données

anime - 22/05/2024 07:46

COMBINER LES DONN...

Dimension associée à la plage de dates

Ajouter une dimension

Dimension

Genres

Ajouter une dimension

Afficher le détail

Genres

Métrique

anime\_id

**Données**

Rechercher

anime - 22/05/2024 07:46

Aired

anime\_id

Duration

English name

Episodes

Favorites

Genres

Image URL

Licensors

Members

Name

Other name

Popularity

Premiered

Ajouter un champ

Ajouter un paramètre

Ajouter des données

# V- Google Cloud

