

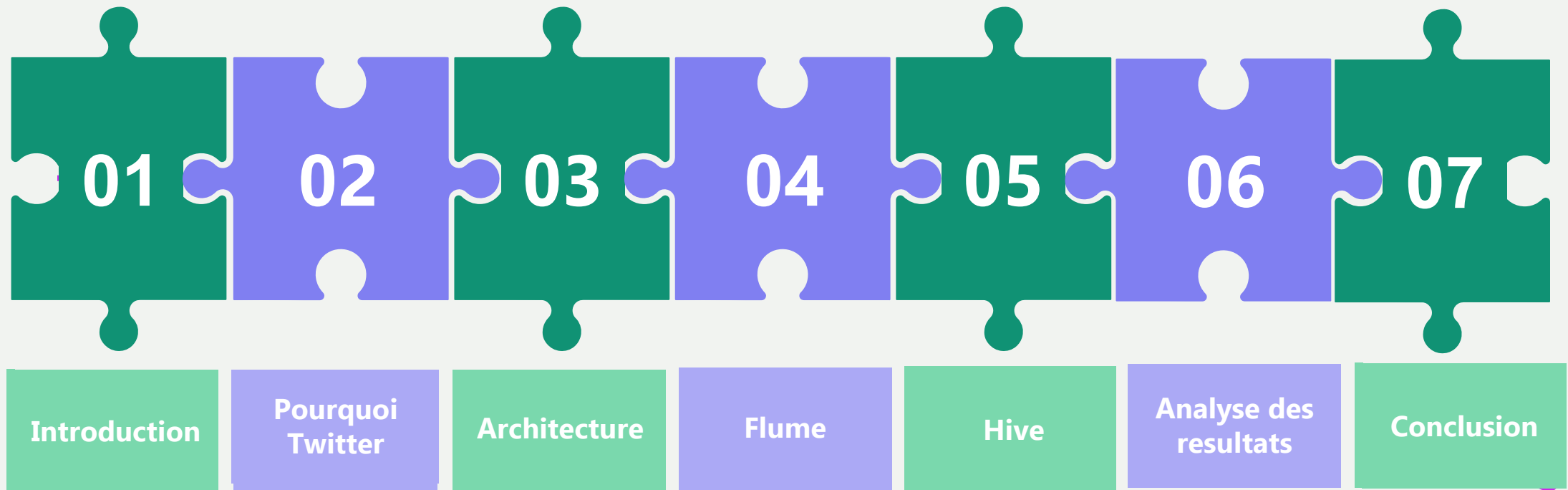
Projet Framework Big Data

Analyse des Tweets des Provox et des Antivax Covid-19

Lydia Ait Abdelkader
Yanis Hammoudi

Année universitaire: 2021/2022

Plan



Introduction

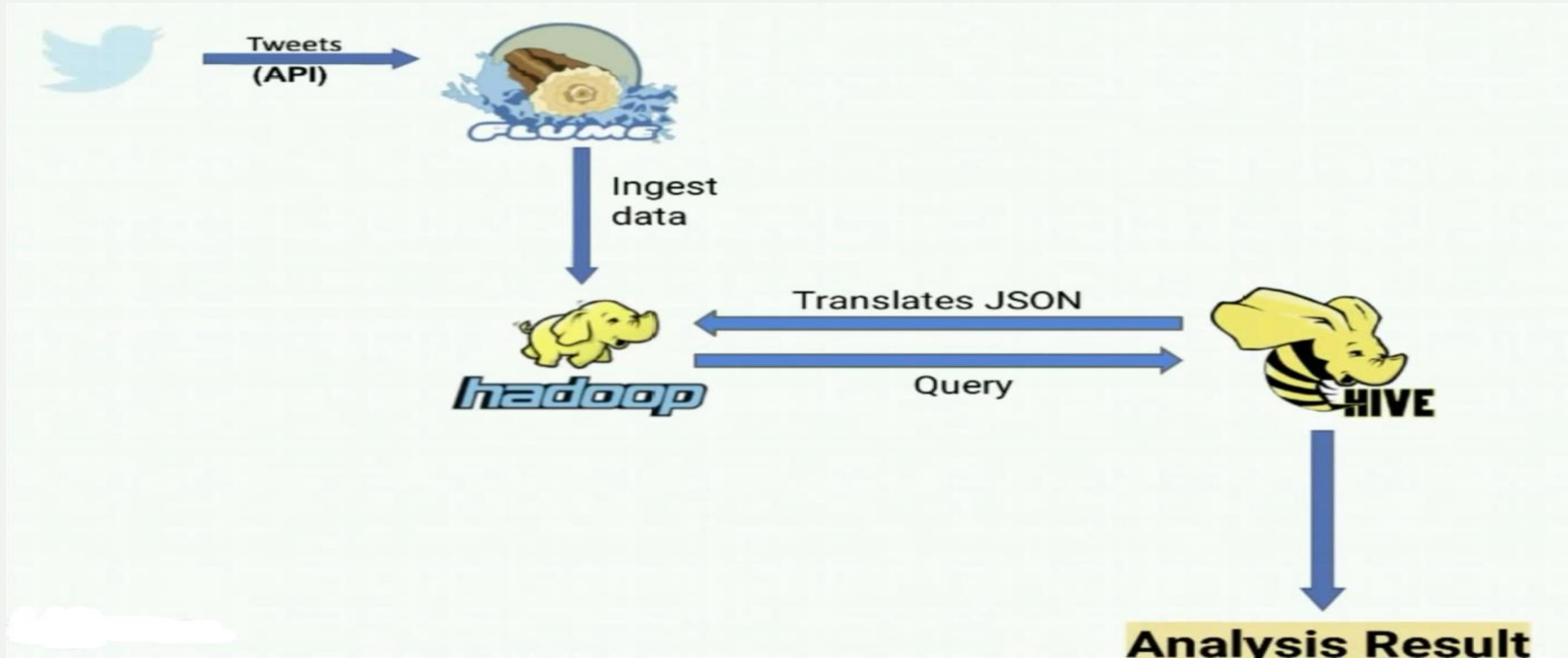
- Grand débat sur les réseaux sociaux depuis l'arrivée du COVID-19.
- Opinions partagées sur la vaccination.
- Raisons différentes.

Pourquoi Twitter ?

- Réseau social populaire.
- 8500 tweets par seconde.
- Cibler facilement l'information recherchée.



Architecture: Installation en local from Scratch



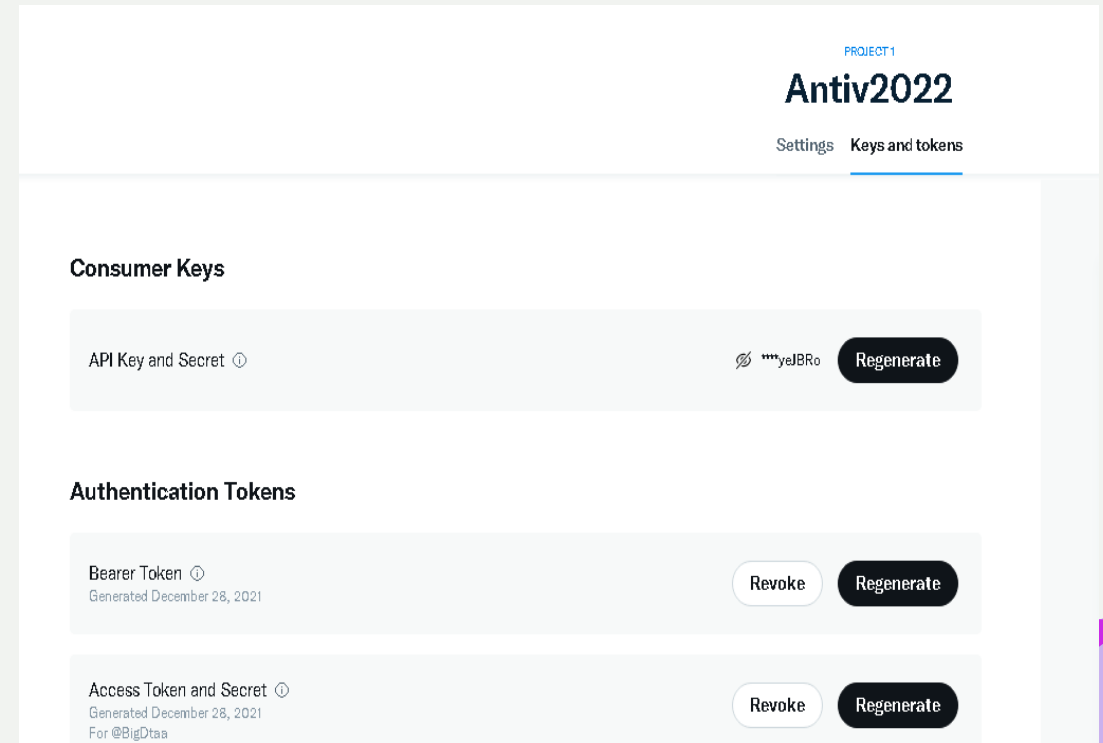
Flume

- Solution de collecte, agrégation et transfert de gros volumes de données.
- Conçu pour gérer des débits importants avec une fonctionnalité native d'écriture dans HDFS.
- Fait partie de l'écosystème Big Data open source Hadoop.



Récupérer les Clés/Tokens d'accès Twitter

- Création d'un compte développeur sur twitter.
- Récupération des clés/ tokens pour utiliser l'API twitter.



- ```
Naming the components of the current agent.
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = 6mmHdG3TcykLWtNjy308u
TwitterAgent.sources.Twitter.consumerSecret = 12WY4nEPT3XW4AL4V7Z0u673657L411T7t0uN126Y
TwitterAgent.sources.Twitter.accessToken = 147F844EAC62E6276274-147W42M074L4G03P50003300012W
TwitterAgent.sources.Twitter.accessTokenSecret = 1P0G7M1E20670ym4mL0rj0T4700G18000000000
TwitterAgent.sources.Twitter.language = fr
TwitterAgent.sources.Twitter.keywords = vaccin , covid , je , suis , pour , pass

Describing/Configuring the sink
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:50070/user/Hadoop/twitter_data/ProVax

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 100
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 100

Describing/Configuring the channel
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 1000

Binding the source and sink to the channel
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```



# Configurer l'Agent Flume

- Mettre à jour le fichier de configuration twitter en y ajoutant les identifiants récupérés précédemment.
- Définir les mots clés.
- Définir le chemin dans HDFS dans lequel les résultats seront enregistrés.

```
Naming the components on the current agent.
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = [REDACTED]
TwitterAgent.sources.Twitter.consumerSecret = [REDACTED]
TwitterAgent.sources.Twitter.accessToken = [REDACTED]
TwitterAgent.sources.Twitter.accessTokenSecret = [REDACTED]
TwitterAgent.sources.Twitter.language = fr
TwitterAgent.sources.Twitter.keywords = vaccin , covid , je , suis , contre , anti, pass

Describing/Configuring the sink

TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:50070/user/Hadoop/twitter_data/AntiVax

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 100
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 100

Describing/Configuring the channel
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 1000

Binding the source and sink to the channel
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```

# Lancer Apache Flume

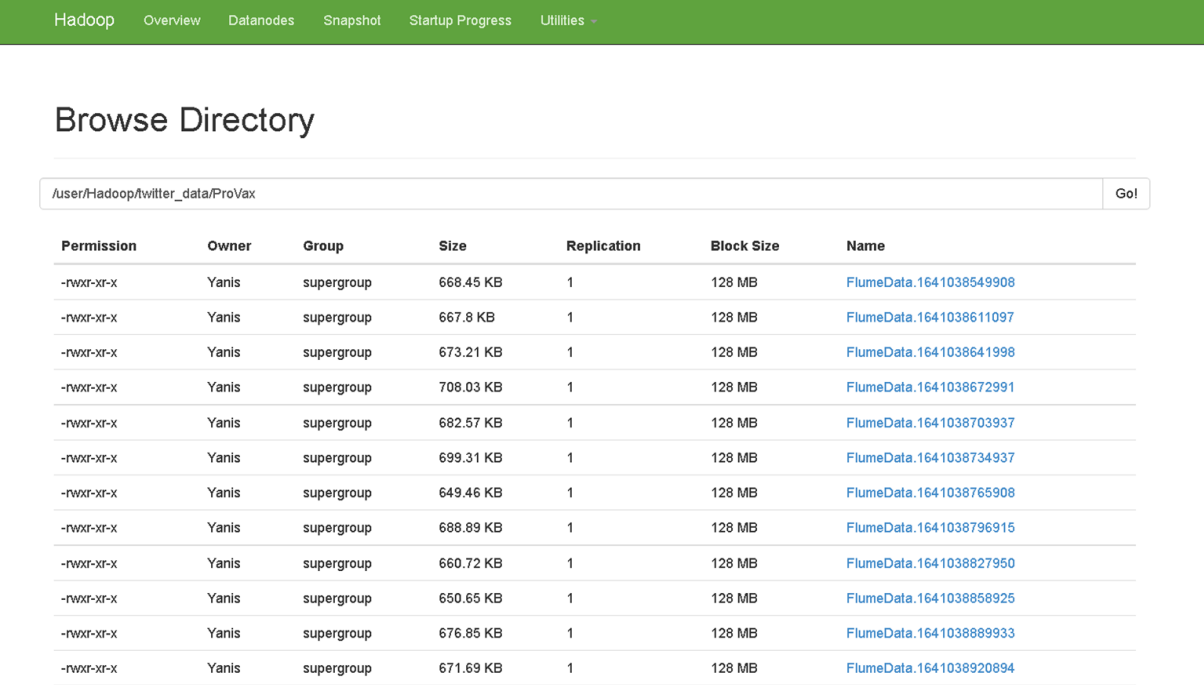
---

Lancement d'Apache flume via la commande :

```
flume-ng agent -n TwitterAgent -c conf -f C:/apache-flume-1.9.0-
bin/conf/twitter.conf -property flume.root.logger=DEBUG,console
```

# Vérifier les résultats dans HDFS

- Vérification des résultats dans HDFS via localhost:50070.
- Fichiers au format JSON.



The screenshot shows the Hadoop web interface for browsing the directory `/user/hadoop/twitter_data/ProVax`. The interface includes a navigation bar with links to Overview, Datanodes, Snapshot, Startup Progress, and Utilities. Below the navigation bar, the title "Browse Directory" is displayed. A search bar contains the path `/user/hadoop/twitter_data/ProVax` and a "Go!" button. The main content area displays a table of files and directories.

| Permission | Owner | Group      | Size      | Replication | Block Size | Name                                    |
|------------|-------|------------|-----------|-------------|------------|-----------------------------------------|
| -rwxr-xr-x | Yanis | supergroup | 668.45 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038549908</a> |
| -rwxr-xr-x | Yanis | supergroup | 667.8 KB  | 1           | 128 MB     | <a href="#">FlumeData.1641038611097</a> |
| -rwxr-xr-x | Yanis | supergroup | 673.21 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038641998</a> |
| -rwxr-xr-x | Yanis | supergroup | 708.03 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038672991</a> |
| -rwxr-xr-x | Yanis | supergroup | 682.57 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038703937</a> |
| -rwxr-xr-x | Yanis | supergroup | 699.31 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038734937</a> |
| -rwxr-xr-x | Yanis | supergroup | 649.46 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038765908</a> |
| -rwxr-xr-x | Yanis | supergroup | 688.89 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038796915</a> |
| -rwxr-xr-x | Yanis | supergroup | 660.72 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038827950</a> |
| -rwxr-xr-x | Yanis | supergroup | 650.65 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038858925</a> |
| -rwxr-xr-x | Yanis | supergroup | 676.85 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038889933</a> |
| -rwxr-xr-x | Yanis | supergroup | 671.69 KB | 1           | 128 MB     | <a href="#">FlumeData.1641038920894</a> |

# Hive

---

- Infrastructure d'entrepôt de données.
- Permet de faire des requêtes via un langage proche syntaxiquement de SQL



# Décryptage des données à travers HIVE

---

- Format des données twitter: JSON.
- Récupération d'une librairie JAVA afin de pouvoir décrypter les données.

# Décryptage des données à travers HIVE

---

- Création de deux tables externes pour la lecture des données.
- Chargement des données depuis HDFS vers les tables externes.

# Analyse des données dans HIVE

---

- Utilisation de requêtes pour l'analyse des résultats.
- Requetes faites sur la base de deux catégories.

# Analyse des résultats : PROVAX

---

- Se protéger contre les différents variants surtout ceux qui sont sujets à risque.
- Éviter de se faire tester à chaque fois grâce au pass sanitaire.
- Croire à la science et à l'efficacité du vaccin.
- Minimiser la propagation du virus.
- Voyager/ partir en vacances.
- Etre libre et retrouver la vie d'avant.





# Analyse des résultats : ANTIVAX

---

- Ne pas avoir confiance en la substance qui sera injectée dans les corps.
- Avoir peur des effets secondaires que pourraient engendrer le vaccin.
- Ne pas avoir assez de recul sur le vaccin.
- Avoir l'impression d'être un cobaye.
- Avoir peur de mourir.
- Ne pas se sentir libre.



# Conclusion

---

- Analyse des données non structurées.
- Mieux comprendre les différents points de vu.

**Merci pour votre attention**

