Université Galatasaray
FIT
Département de Génie Informatique
INF438 - Bases de Données Avancées

2022 – 2023

Google BigQuery avec GDELT

Hazırlayanlar:

YUNUS SOLAK 17401773

BANU ÖZDEVECİ 18401806

DOĞA YAĞMUR YILMAZ 19401852

Contenu:

Qu'est-ce que BigQuery ? **2**
Stockage de BigQuery 3

GDELT **4**

Exemples de requêtes **5**
Exemples Supplémentaires: 8
Histogrammes Géographiques 9
Requête Supplémentaire: 11
Analyse de Réseau 12

Références **14**

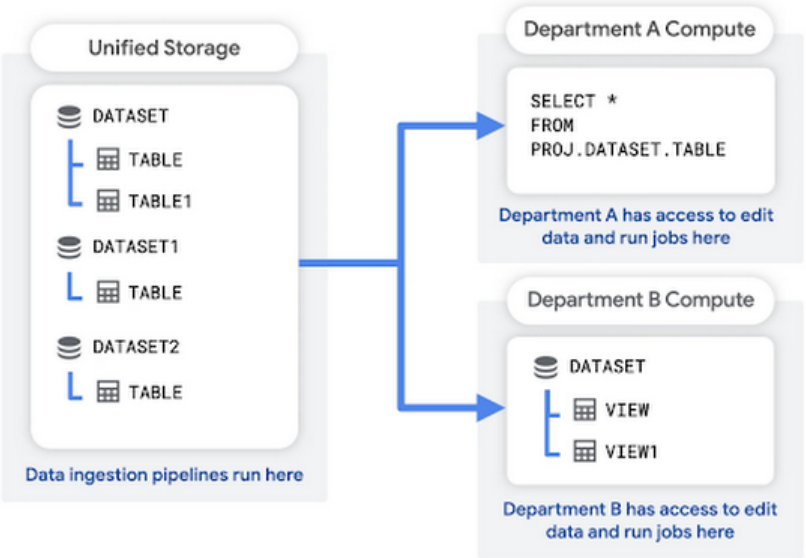
Qu'est-ce que BigQuery ?

BigQuery est le puissant service de base de données analytique en nuage de Google, conçu pour les plus grands ensembles de données de la planète. Il permet aux utilisateurs d'exécuter en quelques secondes des requêtes rapides, de type SQL, sur des ensembles de données de plusieurs téraoctets. Évolutif et facile à utiliser, BigQuery vous donne un aperçu en temps réel de vos données.
BigQuery est un entrepôt de données d'entreprise entièrement géré qui vous aide à gérer et à analyser vos données grâce à des fonctionnalités intégrées telles que l'apprentissage automatique, l'analyse géospatiale et la veille stratégique.
L'architecture sans serveur de BigQuery vous permet d'utiliser des requêtes SQL pour répondre aux plus grandes questions de votre entreprise sans aucune gestion de l'infrastructure.

BigQuery

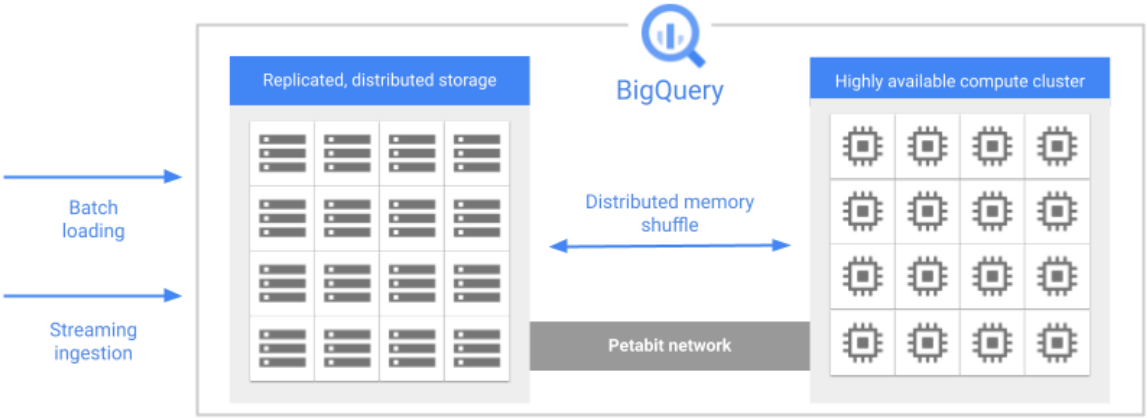
Le moteur d'analyse distribué et évolutif de BigQuery vous permet d'interroger des téraoctets en quelques secondes et des pétaoctets en quelques minutes.
BigQuery maximise la flexibilité en séparant le moteur de calcul qui analyse vos données de vos choix de stockage. Vous pouvez stocker et analyser vos données dans BigQuery ou utiliser BigQuery pour évaluer vos données là où elles se trouvent. Les requêtes fédérées vous permettent de lire des données provenant de sources externes, tandis que le streaming prend en charge les mises à jour continues des données. Des outils puissants comme BigQuery ML et BI Engine vous permettent d'analyser et de comprendre ces données.

Stockage de BigQuery



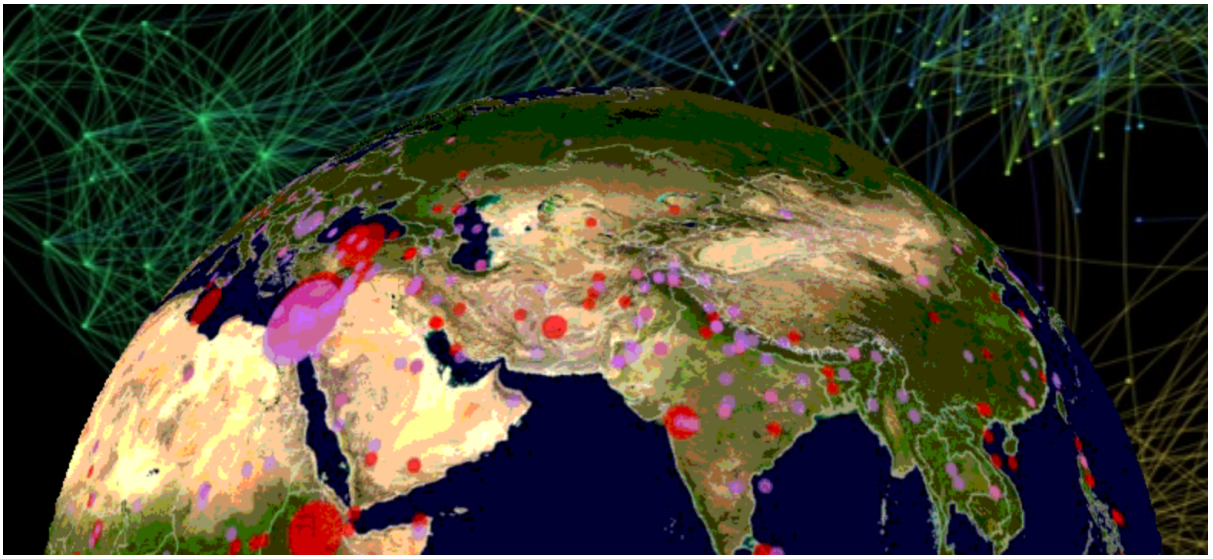
BigQuery stocke les données à l'aide d'un format de stockage en colonnes qui est optimisé pour les requêtes analytiques. BigQuery présente les données sous forme de tableaux, de lignes et de colonnes et prend entièrement en charge la sémantique des transactions de base de données (ACID). Le stockage BigQuery est automatiquement répliqué sur plusieurs sites pour assurer une haute disponibilité.

L'une des principales caractéristiques de l'architecture de BigQuery est la séparation du stockage et du calcul. Cela permet à BigQuery de faire évoluer le stockage et le calcul de manière indépendante, en fonction de la demande.
Lorsque vous exécutez une requête, le moteur de requête distribue le travail en parallèle entre plusieurs travailleurs, qui analysent les tables pertinentes dans le stockage, traitent la requête, puis rassemblent les résultats. BigQuery exécute les requêtes entièrement en mémoire, à l'aide d'un réseau pétabits pour garantir que les données se déplacent extrêmement rapidement vers les nœuds de travail.



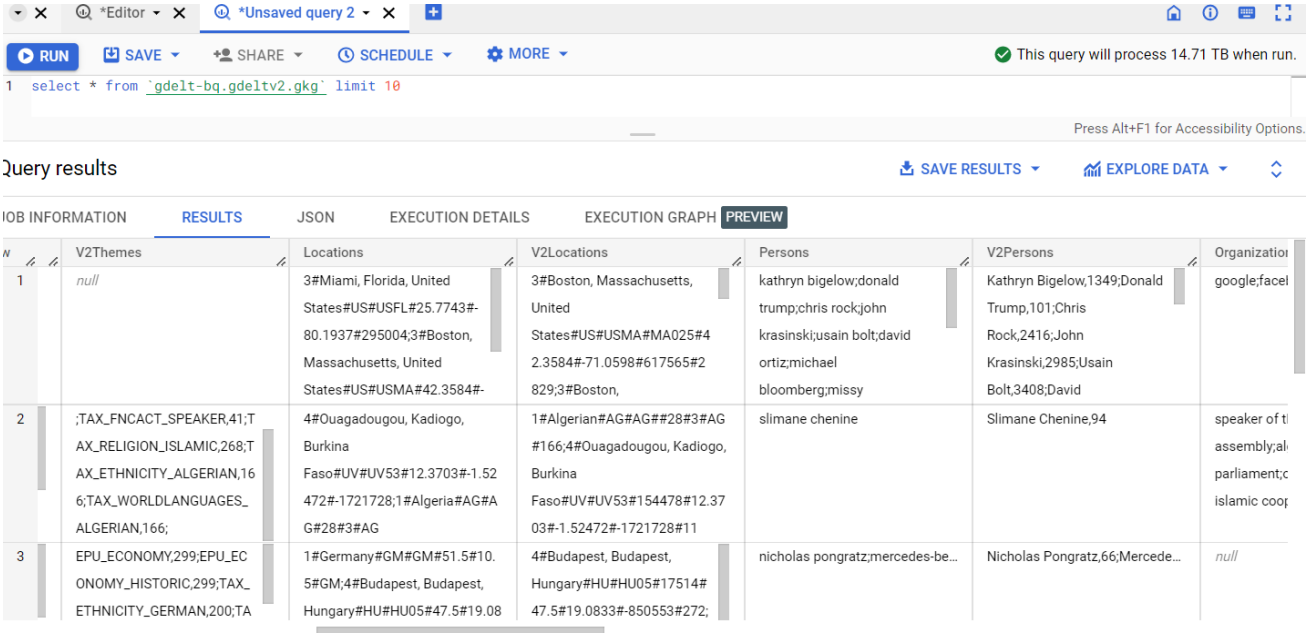
GDELT

La surveillance de la quasi-totalité des médias d'information du monde n'est qu'un début - même la plus grande équipe d'humains ne pourrait pas commencer à lire et à analyser les milliards et milliards de mots et d'images publiés chaque jour. GDELT utilise certains des algorithmes informatiques les plus sophistiqués au monde, conçus sur mesure pour les médias d'information mondiaux, fonctionnant sur "l'un des réseaux de serveurs les plus puissants de l'univers connu", ainsi que certains des algorithmes d'apprentissage profond les plus puissants au monde, pour créer un enregistrement informatique en temps réel de la société mondiale qui peut être visualisé, analysé, modélisé, examiné et même prévu.

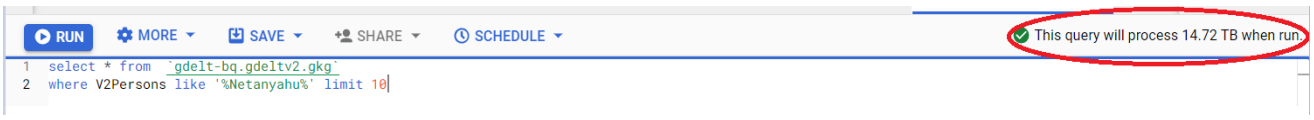


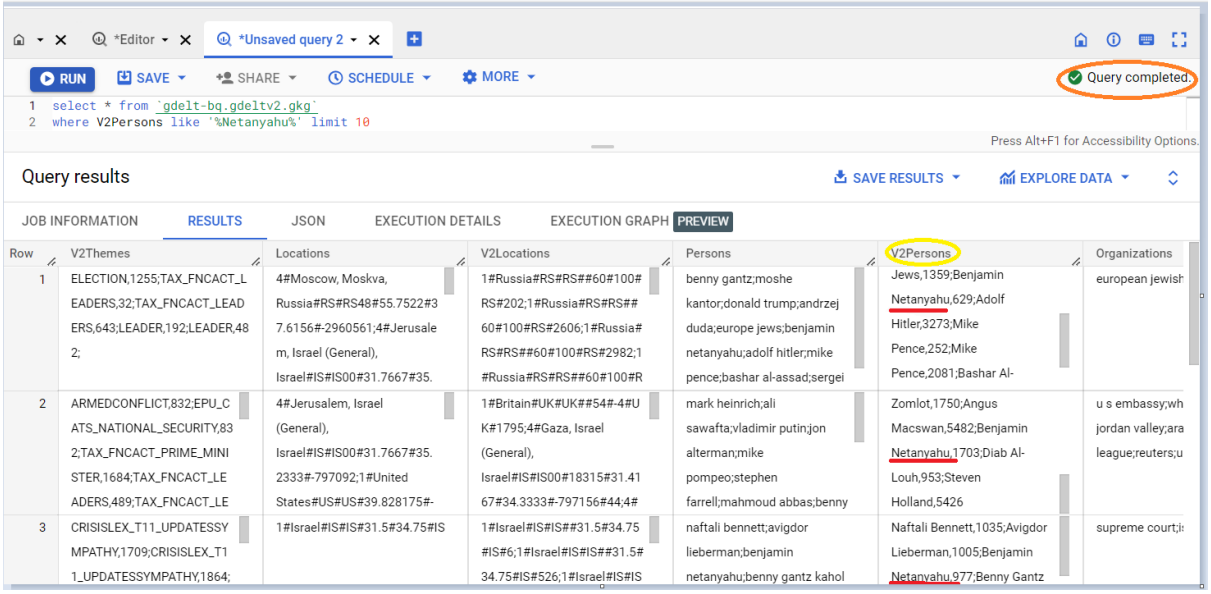
Un vaste éventail d'ensembles de données totalisant des trillions de points de données est disponible. Trois flux de données primaires sont créés, l'un codifiant les activités physiques à travers le monde dans plus de 300 catégories, l'autre enregistrant les personnes, les lieux, les organisations, les millions de thèmes et les milliers d'émotions qui sous-tendent ces événements et leurs interconnexions, et le dernier codifiant les récits visuels des images d'actualité dans le monde.

Exemples de requêtes

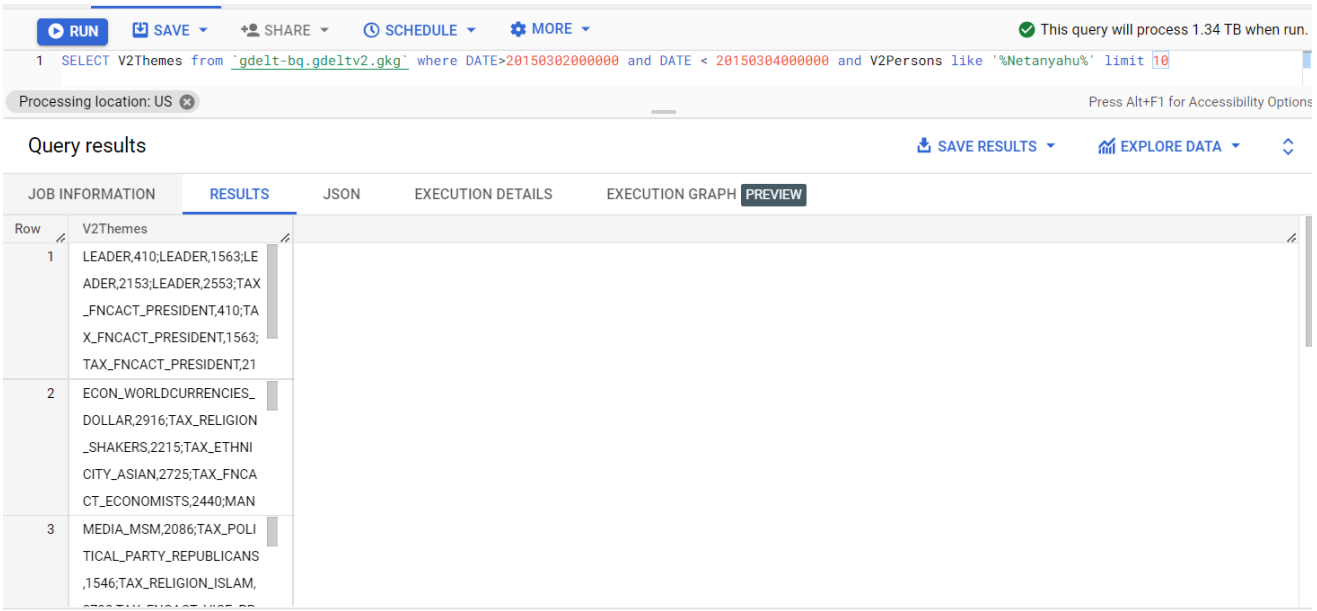


“gdelt-bq.gdeltv2.gkg” est une table de l'ensemble de données GDELT (Global Database of Events, Language, and Tone), qui est disponible sur Google BigQuery. L'ensemble de données GDELT contient un registre de la société humaine mondiale et de ses événements, tels qu'ils sont enregistrés dans les médias, avec plus d'un quart de milliard d'enregistrements ajoutés chaque jour.
La table GKG de l'ensemble de données GDELT contient des enregistrements de la couverture médiatique mondiale, avec un enregistrement pour chaque article ou rapport unique. Chaque enregistrement comprend des informations sur l'article, telles que la date de publication, la langue dans laquelle il a été rédigé, le ton de l'article et le lieu mentionné dans l'article. Il comprend également une liste d'entités nommées mentionnées dans l'article, telles que des personnes, des organisations et des lieux. La table GKG est mise à jour quotidiennement avec les dernières nouvelles du monde entier.





Par exemple, imaginons créer un histogramme des principaux thèmes associés au Premier ministre israélien Benjamin Netanyahu lors de sa visite au Congrès américain le 3 mars 2015. Demandons une liste des thèmes qui apparaissent dans chaque article dont le nom est mentionné. C'est facile dans BigQuery.



Le problème est que la colonne V2Themes utilise une délimitation imbriquée - chaque citation d'un thème reconnu dans un article est séparée par un point-virgule, et pour chaque citation, le thème et son décalage de caractères dans l'article sont séparés par une virgule.

Tout d'abord, nous utilisons la fonction SPLIT() pour demander à BigQuery de prendre le champ V2Themes, de le séparer par un point-virgule et de le renvoyer sous forme d'enregistrements multiples, un par citation. En d'autres termes, en utilisant "SPLIT(V2Themes,';')" BigQuery prendra l'exemple de l'enregistrement V2Thèmes.

RUN SAVE SHARE SCHEDULE MORE This query will process 1.34 TB when run.

```
SELECT SPLIT(V2Themes,';') from `gdelt-bq.gdeltv2.gkg` where DATE>20150302000000 and DATE < 20150304000000 and V2Persons like '%Netanyahu%' limit 10
```

Processing location: US Press Alt+F1 for Accessibility Options.

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS JSON EXECUTION DETAILS EXECUTION GRAPH PREVIEW

Row	f0_
	WB_2462_POLITICAL_VIOLENCE_AND_WAR,185
	WB_2473_DIPLOMACY_AND_NEGOTIATIONS,2331
	WB_2471_PEACEKEEPING,2331
	TAX_POLITICAL_PARTY_DEMOCRAT,1934
	TAX_FNCACT_AMBASSADOR,1388
	TAX_FNCACT_AMBASSADOR,2166
	MIL_WEAPONS_PROLIFERATION,540
	GENERAL_GOVERNMENT,1946
	ACT_MAKESTATEMENT,1641
	WB_2470_PEACE_OPERATIONS_AND_CONFLICT_MANAGEMENT,708
	WB_2503_WEAPONS_PROLIFERATION_AND_ARMS_CONTROL,708
	TAX_FNCACT_SENATOR,874

Le problème est que nous avons toujours le décalage de caractères indiqué à la fin de chaque citation de thème dont nous devons séparer. Pour y arriver, Nous disposons de la fonction utile REGEXP_REPLACE().
En SQL, les fonctions REGEXP_* vous permettent de rechercher et de manipuler des chaînes de caractères sur la base de modèles d'expressions régulières. Une expression régulière est une séquence de caractères qui définit un motif de recherche, principalement utilisé pour la correspondance de motifs avec des chaînes de caractères c'est-à-dire la recherche et le remplacement de motifs dans des chaînes de caractères.

- ^ correspond au début d'une chaîne de caractères.
- . correspond à n'importe quel caractère unique (sauf une nouvelle ligne).
- * correspond à zéro ou plusieurs occurrences du caractère ou du motif précédent.
- ? correspond à zéro ou une occurrence du caractère ou du motif précédent.

RUN SAVE SHARE SCHEDULE MORE This query will process 1.34 TB when run.

```
SELECT theme, COUNT(*) as count
FROM (
SELECT REGEXP_REPLACE(ARRAY_TO_STRING(SPLIT(V2Themes,';'), ","), r',.*', "") as theme
FROM `gdelt-bq.gdeltv2.gkg`
WHERE DATE > 20150302000000 AND DATE < 20150304000000 AND V2Persons LIKE '%Netanyahu%'
)
group by theme
ORDER BY 2 DESC
LIMIT 300
```

Processing location: US Press Alt+F1 for Accessibility Options.

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS JSON EXECUTION DETAILS EXECUTION GRAPH PREVIEW

Row	theme	count
1	TAX_POLITICAL_PARTY_REPUBLICANS	2088
2	LEADER	1217
3	MEDIA_MSM	989
4	TERROR	785
5	TAX_FNCACT_SENATORS	467
6	GENERAL_GOVERNMENT	452
7	TAX_FNCACT_DEPUTY	276

Results per page: 50 1 – 50 of 300

Comparons ces résultats à ceux du Premier ministre grec Alexis Tsipras durant la même période.

```
SELECT theme, COUNT(*) as count
FROM (
SELECT REGEXP_REPLACE(ARRAY_TO_STRING(SPLIT(V2Themes,';'), ","), r',.*', "") as theme
FROM `gdelt-bq.gdeltv2.gkg`
WHERE DATE > 20150302000000 AND DATE < 20150304000000 AND V2Persons LIKE '%Tsipras%'
)
group by theme
ORDER BY 2 DESC
LIMIT 300
```

Query results

Row	theme	count
1	GENERAL_GOVERNMENT	925
2	ECON_DEVELOPMENTORGS_I...	125
3	LEADER	64
4	PRIVATIZATION	43
5	ELECTION	43
6	SOVEREIGNTY	42
7	TAX_ETHNICITY_GERMAN	41

Comme prévu, nous observons un ensemble de thèmes très différents, qui reflètent fortement le discours sur l'économie et la crise de la Grèce.

Exemples Supplémentaires:

-Un exemple que nous pouvons comprendre la Turquie et les politiques gouvernementales en 2020. Dans cette requête, nous voyons les nombres "Erdogan" qui sont mentionnés dans différents thèmes.

```
SELECT theme, COUNT(*) as count
FROM (
SELECT REGEXP_REPLACE(ARRAY_TO_STRING(SPLIT(V2Themes,';'), ","), r',.*', "") as theme
FROM `gdelt-bq.gdeltv2.gkg`
WHERE DATE > 20200302000000 AND DATE < 20200304000000 AND V2Persons LIKE '%Erdogan%'
)
group by theme
ORDER BY 2 DESC
LIMIT 100
```

Query results

Row	theme	count
1	GENERAL_GOVERNMENT	859
2	LEADER	336
3	ARMEDCONFLICT	307
4	TERROR	211
5	TAX_ETHNICITY_TURKISH	186
6	KILL	123
7	EPU_POLICY_GOVERNMENT	111

-Une requête sur “Zelenski” et les thèmes des publications officielles, après la guerre russo-ukrainienne qui a débuté en février 2022.

RUN SAVE SHARE SCHEDULE MORE Query completed.

```
SELECT theme, COUNT(*) as count
FROM (
SELECT REGEXP_REPLACE(ARRAY_TO_STRING(SPLIT(V2Themes,';'), ","), r',.*', "") as theme
FROM `gdelt-bq.gdeltv2.gkg`
WHERE DATE > 20220202000000 AND DATE < 20220504000000 AND V2Persons LIKE '%Zelenski%'
)
group by theme
ORDER BY 2 DESC
LIMIT 100
```

Press Alt+F1 for Accessibility Options.

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS JSON EXECUTION DETAILS EXECUTION GRAPH PREVIEW

Row	theme	count
1	ARMEDCONFLICT	4194
2	GENERAL_GOVERNMENT	3046
3	TAX_FNCACT_DEPUTY	2782
4	LEADER	2744
5	EPU_CATS_NATIONAL_SECURI...	1516
6	TAX_FNCACT_CHIEF	1389
7	BAN	933
8	MILITARY	896
9	TAX_WORLDLANGUAGES_RUS...	809
10	KILL	791
11	WB_2433_CONFLICT_AND_VIO...	703
12	CRISISLEX_CRISISLEXREC	703

Histogrammes Géographiques

une entrée typique pourrait ressembler à :
"4#Berlin, Berlin, Germany#GM#GM16#16538#52.5167#13.4#-1746443#1340"

Si le champ V2Locations contient la valeur:
"4#Berlin, Berlin, Germany#GM#GM16#16538#52.5167#13.4#-1746443#1340",
cela indique que le document mentionne Berlin, Allemagne, qui est une ville ou un point de repère en dehors des Etats-Unis, avec un centroïde de latitude de 52.5167 et un centroïde de longitude de 13.4. Le champ Location FeatureID contient le GNS ou GNIS FeatureID numérique de Berlin, qui est -1746443. Le champ Character Offset indique que l'emplacement a été mentionné au décalage de caractère 1340 dans le document.

Voici un résumé de ce que représente quelques champ du champ V2Locations de la table “gdelt-bq.gdeltv2.gkg” :
Type d'emplacement : Il s'agit d'un nombre entier qui spécifie la résolution géographique de l'emplacement. Une valeur de 1 indique que l'emplacement est un pays, 2 indique qu'il s'agit d'un État américain, 3 indique qu'il s'agit d'une ville ou d'un point de repère américain, 4 indique qu'il s'agit d'une ville ou d'un point de repère hors des États-Unis et 5 indique qu'il s'agit d'une division administrative hors des États-Unis (à peu près l'équivalent d'un État américain).

Nom complet de l'emplacement : Il s'agit du nom complet lisible par l'homme de l'emplacement. Il peut être au format "État, nom du pays" pour les États américains et les pays du monde, ou "Ville/Point de repère, État, pays" pour tous les autres lieux.
Décalage des caractères : Il s'agit du décalage approximatif des caractères dans le document où le lieu est mentionné.

Une requête simple qui renvoie un histogramme des lieux mentionnés dans la couverture du Premier ministre grec Tsipras au cours de la même période que la requête thématique précédente :

RUN SAVE SHARE SCHEDULE MORE Query completed

```
SELECT location, COUNT(*)
FROM(
SELECT REGEXP_EXTRACT(ARRAY_TO_STRING(SPLIT(V2Locations,';'), ","),r'^.*?#(.*?)#') as location
FROM `gdelt-bq.gdeltv2.gkg`
WHERE DATE > 20150302000000 AND DATE < 20150304000000 AND V2Persons LIKE '%Tsipras%'
)
group by location
ORDER BY 2 DESC
LIMIT 100
```

Press Alt+F1 for Accessibility Option

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS JSON EXECUTION DETAILS EXECUTION GRAPH PREVIEW

Row	location	f0_
1	Madrid, Madrid, Spain	414
2	Greece	276
3	Spanish	140
4	Spain	103
5	Guindo, Castilla Y LeóP, Spain	83
6	null	75
7	Greek	55
8	Brussels, Bruxelles-Capitale, Be...	50
9	Russia	48
10	Greeks	41

Elapsed time	Slot time consumed	Bytes shuffled	Bytes spilled to disk
30 sec	13 hr 17 min	54.45 KB	0 B

SHOW AVERAGE TIME SHOW MAXIMUM TIME

Stages	Working timing				Rows
S00: Input	Wait: 12 sec	Read: 86 ms	Compute: 9 ms	Write: 3 ms	Records read: 1023815521 Records written: 1704
S01: Sort+	Wait: 165 ms	Read: 0 ms	Compute: 204 ms	Write: 1 ms	Records read: 1704 Records written: 102

Comme on peut s'y attendre, un grand nombre des premiers résultats sont des localisations au niveau national, comme "Grèce" et "Espagne", qui sont probablement moins intéressantes pour de nombreuses requêtes. Au lieu de cela, la requête suivante modifie légèrement le REGEXP_EXTRACT pour ajouter "[2-5]" au début de la requête au lieu du ".* ?" original (filtrant chaque mention de lieu pour exiger une valeur comprise entre 2 et 5 pour le champ "Type de lieu" (tout ce qui a une résolution plus élevée qu'une correspondance au niveau du pays).

```
SELECT location, COUNT(*)
FROM(
SELECT REGEXP_EXTRACT(ARRAY_TO_STRING(SPLIT(V2Locations,';'), ","),r'^[2-5]#(.*?)#') as location
FROM `gdelt-bq.gdeltv2.gkg`
WHERE DATE > 20150302000000 AND DATE < 20150304000000 AND V2Persons LIKE '%Tsipras%'
)
WHERE location is not null
group by location
ORDER BY 2 DESC
LIMIT 100
```

Row	location	f0_
1	Madrid, Madrid, Spain	414
2	Guindo, Castilla Y LeóP, Spain	83
3	Brussels, Bruxelles-Capitale, Belgium	50
4	Samara, Samarskaya Oblast', Russia	28
5	Dublin, Dublin, Ireland	21
6	Texas, United States	20
7	Athens, AttikíR, Greece	19
8	Bruxelles, Bruxelles-Capitale, Belgium	15
9	Mediterranean Sea, Oceans (General), Oceans	14
10	Toronto, Ontario, Canada	14
11	Lemnos, Perifereia Voreiou Aigaiou, Greece	13

Requête Supplémentaire:

```
SELECT location, COUNT(*)
FROM(
SELECT REGEXP_EXTRACT(ARRAY_TO_STRING(SPLIT(V2Locations,';'), ","),r'^[2-5]#(.*?)#') as location
FROM `gdelt-bq.gdeltv2.gkg`
WHERE DATE > 20150302000000 AND DATE < 20220304000000 AND V2Persons LIKE '%Ataturk%'
)
WHERE location is not null
group by location
ORDER BY 2 DESC
LIMIT 100
```

JOB INFORMATION | RESULTS | JSON | E

Row	location	f0_
1	Dublin, Dublin, Ireland	166
2	Izmir, Izmir, Turkey	41
3	Istanbul, Istanbul, Turkey	27
4	Edremit, Van, Turkey	12
5	Bingol, BingöU, Turkey	4
6	Balgownie, New South Wales, ...	4
7	Trabzon, Trabzon, Turkey	3
8	Silivri, Istanbul, Turkey	3
9	Sivas, Sivas, Turkey	3
10	Baghdad, Baghdad, Iraq	2
11	Kalymnos, Perifereia Notiou Ai...	2
12	Gursel, Kahramanmaras, Turkey	2

Intéressant que la Turquie ne soit pas au premier rang.

Collectons les villes d'un même pays dans un seul enregistrement en utilisant “Location CountryCode”.

RUN SAVE SHARE SCHEDULE MORE

```
SELECT location, COUNT(*)
FROM(
SELECT REGEXP_EXTRACT(ARRAY_TO_STRING(SPLIT(V2Locations,';'), ","),r'^[2-5]#(.*?)#(?:GR|GM|SP)#') as location
FROM `gdelt-bq.gdeltv2.gkg`
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

Row	location	f0_
1	Madrid, Madrid, Spain	414
2	Guindo, Castilla Y LeóP, Spain	83
3	Athens, AttikíR, Greece	19
4	Lemnos, Perifereia Voreiou Aigaiou, Greece	13
5	Brussels, Bruxelles-Capitale, Belgium#BE#BE11#5850#50.8333#4.33333#-1955538#13,1#Greek	6

Analyse de Réseau

Comprenons le réseau de connexions qui entoure le Premier ministre grec Alexis Tsipras, en créant un diagramme de réseau montrant toutes les personnes avec lesquelles il est le plus étroitement lié dans la presse pendant une période donnée. La requête suivante crée un diagramme de réseau de cooccurrence des personnes apparaissant dans la couverture du Premier ministre Tsipras pendant la période du 2 mars 2015 au 4 mars 2015 .

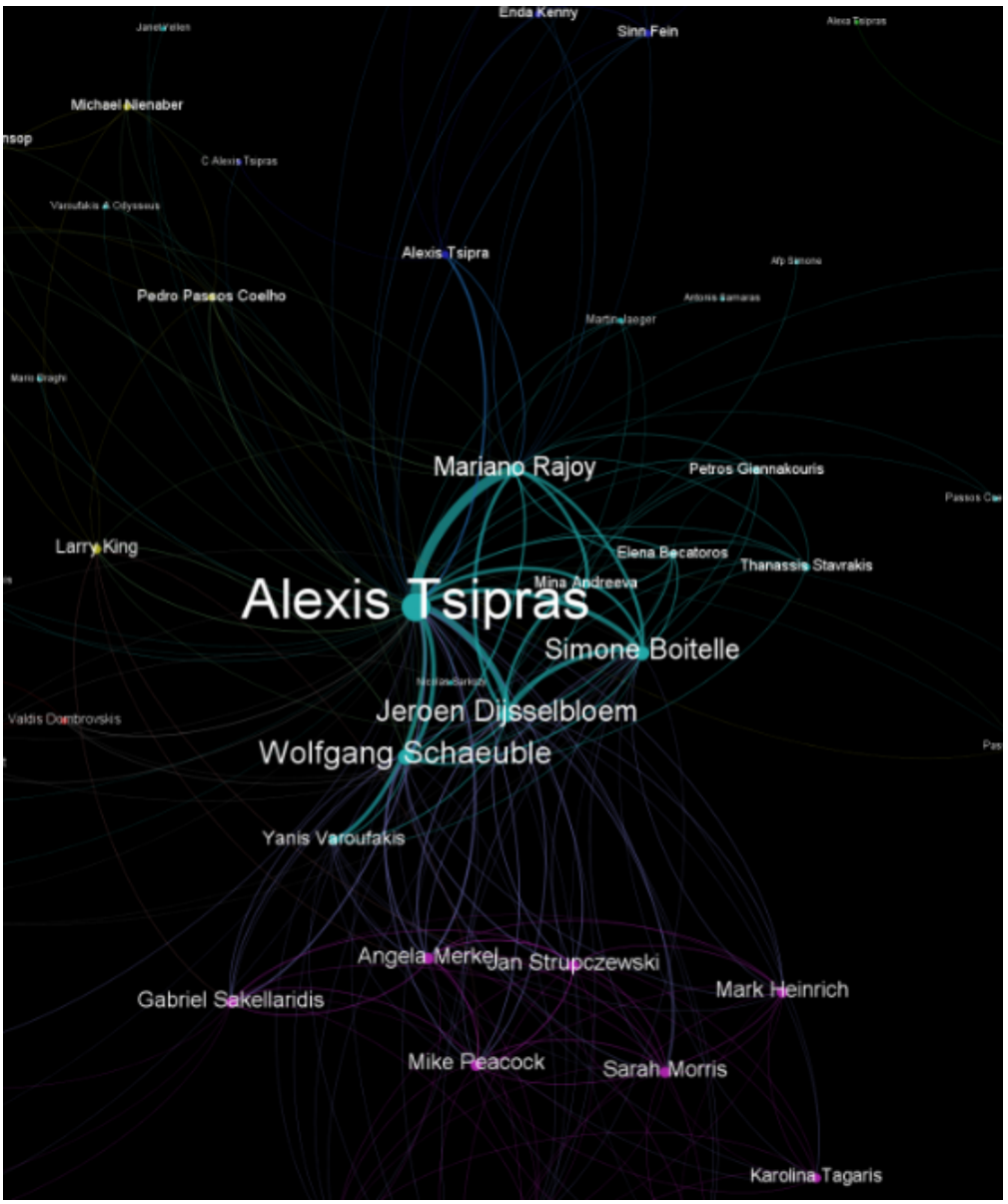
RUN SAVE SHARE SCHEDULE MORE Query completed.

```
SELECT person1,person2, COUNT(*) as count

FROM(
SELECT GKGRECORDID, REGEXP_REPLACE(ARRAY_TO_STRING(SPLIT(V2Persons,';'), ","), r',.*', "") person1
FROM `gdelt-bq.gdeltv2.gkg`
where DATE>20150302000000 and DATE < 20150304000000 and V2Persons like '%Tchipras%'
)a

INNER JOIN

(
SELECT GKGRECORDID, REGEXP_REPLACE(ARRAY_TO_STRING(SPLIT(V2Persons,';'), ","), r',.*', "") person2
FROM `gdelt-bq.gdeltv2.gkg`
where DATE>20150302000000 and DATE < 20150304000000 and V2Persons like '%Tchipras%'
)b

ON a.GKGRECORDID = b.GKGRECORDID
WHERE a.person1<b.person2
GROUP BY 1,2
ORDER BY 3 DESC
LIMIT 250
```

Press Alt+F1 for Accessibility Option

Source	Target	count	Type	Weight
Alexis Tsipras	Mariano Rajoy	496	Undirected	0.056998
Alexis Tsipras	Jeroen Dijsselbloem	334	Undirected	0.038382
Alexis Tsipras	Simone Boitelle	229	Undirected	0.026316
Jeroen Dijsselbloem	Simone Boitelle	229	Undirected	0.026316
Alexis Tsipras	Wolfgang Schaeuble	221	Undirected	0.025396

Nous avons visualisé le réseau de cooccurrence des personnes cooccurrentes dans la couverture du Premier ministre de la Grèce sur une période de deux jours grâce à open source Gephi logiciel. En regardant le réseau ci-dessous, les noms les plus étroitement associés au Premier ministre sont ceux qui sont les plus impliqués dans les négociations de la Grèce avec l'UE ou dans les discussions sur la restructuration de la crise de la dette grecque et l'austérité.



Références

https://cloud.google.com/bigquery/docs/introduction

https://blog.gdeltproject.org/google-bigquery-gkg-2-0-sample-queries/

https://www.gdeltproject.org/