

Projet Bigdata : Analyser l'année 2017 via la base de données GDELT

Intro

“ *The Global Database of Events, Language, and Tone (GDELT) (<https://www.gdeltproject.org/>), est une initiative pour construire un catalogue de comportements et de croyances sociales à travers le monde, reliant chaque personne, organisation, lieu, dénombrement, thème, source d'information, et événement à travers la planète en un seul réseau massif qui capture ce qui se passe dans le monde, le contexte, les implications ainsi que la perception des gens sur chaque jour*”.

Cette base de données a eu beaucoup d'utilisations, par exemple pour mieux comprendre l'évolution et l'impact de la crise financière du 2008 (Bayesian dynamic financial networks with time-varying predictors (<https://arxiv.org/pdf/1403.2272v1.pdf>)) ou analyser l'évolution des relations entre des pays impliquées dans des conflits (Massive Media Event Data Analysis to Assess World-Wide Political Conflict and Instability (http://www.gao.ece.ufl.edu/GXU/fun_reading/sbp_hurst.pdf)).

L'objectif du projet est de concevoir un système qui permet d'analyser les événements de l'année 2017 à travers leur récit dans les médias du monde collectés par GDELT.

Contexte

A. Jeu de données

GDELT est composé par trois jeux de fichiers CSV, avec un fichier compressé par tranche de 15 minutes:

- les events (schema (<https://bigquery.cloud.google.com/table/gdelt-bq:gdeltv2.events?tab=schema>), CAMEO Ontology (<http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf>), documentation (http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf))
- les mentions (schema (<https://bigquery.cloud.google.com/table/gdelt-bq:gdeltv2.eventmentions>), documentation (http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf))

- le graph des relations \Rightarrow GKG, Global Knowledge Graph (schema (<https://bigquery.cloud.google.com/table/gdelt-bq:gdeltv2.gkg>), documentation (http://data.gdeltproject.org/documentation/GDEL-Global_Knowledge_Graph_Codebook-V2.1.pdf))

Pour plus d'infos consulter la documentation.

(<https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>)

Motivation

Le jeu de données de GDEL- v2.0 est disponible sur Google BigQuery

(<https://www.gdeltproject.org/data.html#googlebigquery>) pour permettre d'analyser l'historique ou en temps réel. Cependant le stockage est assez basique (les trois tables sont stockées en format raw), ce qui fait que les requêtes nécessitent souvent le parcours de grandes quantités de données avec un coût très élevé.

Pour votre projet nous avons extrait un sous-ensemble du jeu de données GDEL- sur AWS S3 (dans le bucket [s3://telecom.gdelt]). Cet extrait correspond à la période du 1er janvier 2017 jusqu'au 15 décembre 2017 soit environ 3.5TB de données.

Objectif

L'objectif de ce projet est de proposer un système de stockage distribué performant sur AWS pour les données de GDEL-.

B. Cas d'utilisation

Exemple d'application minimale

1. A partir d'une *date* et un/*plusieurs critères* (theme, type d'action/relation, entité \Rightarrow à vous de proposer) le système affiche **les événements les plus importants** (avec le plus de mentions, le plus de mentions positives/negatives, les plus pertinents par rapport à un sujet particulier/critères) sur les dernières 24h et ceux de 30 derniers jours.
2. Suite au choix d'un événement, affichez une analyse de celui-ci, par exemple acteurs, description, timeline, mentions, timeline des mentions et leur localisation, analyse des sources qui ont relayé cet événement.
3. Vous êtes libre de proposer les fonctionnalités qui vous semblent les plus pertinentes.

C. Contraintes

1. Vous devez utiliser **au moins 1 technologie vue en cours** en expliquant les raisons de votre choix (SQL/Cassandra/MongoDB/Spark/Neo4j).

2. Vous devez concevoir **un système distribué et tolérant aux pannes** (le système doit pouvoir continuer après la perte d'un serveur).
3. Vous devez pre-charger au moins **un mois de données** dans votre cluster
4. Vous devez utiliser **AWS** pour déployer le cluster **impérativement dans la region us-east !**.

Budget AWS: 300E par groupe (à confirmer).

D. Les livrables

Vous devrez fournir:

- une archive avec votre code source (ou un lien sur github...)
- une courte description de votre modélisation, les avantages et inconvénients des choix de modélisation et d'architecture (slides de présentation)

IV. Organisation

Vous travaillerez par groupe de 6 personnes. La soutenance se déroulera de la manière suivante:

1. Présentation: 10 minutes
2. Démo: 10 minutes
3. Questions & Réponses : 10 minutes



Lors de la soutenance, les données devront être préalablement chargées dans votre cluster.

Ressources

[GDELT v2.0 dataset description](https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/) (<https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>)

[Une compilation des demos GDELT](https://blog.gdeltproject.org/a-compilation-of-gdelt-bigquery-demos/) (<https://blog.gdeltproject.org/a-compilation-of-gdelt-bigquery-demos/>)

[Article original sur la creation du dataset GDELT](http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf)

(<http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf>)

Annexe

Exploration des données GDELT via AWS EMR

Pour commencer à regarder le contenu des fichiers GDELT vous pouvez démarrer un cluster sur AWS et utiliser un notebook Spark (Apache Zeppelin (<https://zeppelin.apache.org/>)) pour une analyse exploratoire du jeu de données.

Démarrage d'un cluster sur AWS EMR

Creation/configuration du votre compte AWS

1. Si vous n'avez pas encore créé un compte sur AWS créez-en ici:
<https://aws.amazon.com/fr/premiumsupport/knowledge-center/create-and-activate-aws-account/>
2. Rendez-vous sur la console AWS et créez une paire de clefs *gdeltKeyPair*:
<https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#KeyPairs:sort=keyName>
3. la clé privée sera automatiquement sauvegardée après la création dans un fichier *gdeltKeyPair.pem*. Notez l'emplacement de ce fichier, vous en aurez besoin pour plus tard
4. Créez un utilisateur pour la gestion de vos clusters:
 - a. Allez sur la console IAM (<https://console.aws.amazon.com/iam/home?region=us-east-1#/users>) et cliquez sur *Add User*
 - b. Mettez comme nom d'utilisateur: *gdeltUser* et cochez la case **Programmatic access** pour créer un identifiant d'accès et une clé de sécurité (access key ID and secret access key)
 - c. A l'étape suivante mettez votre utilisateur dans le groupe administrateur et validez la création de l'utilisateur
 - d. Sur la page de confirmation de la création de votre utilisateur, cliquez sur *Download csv* pour sauvegarder dans un fichier *credentials.csv* l'ID et la clé de sécurité



Pour simplifier les procédures nous allons utiliser un utilisateur avec des droits d'admin. En général ce n'est pas recommandé, pour des raisons de sécurité d'utiliser des comptes avec trop de droits (si quelqu'un arrive à mettre la main sur votre identifiant d'accès et votre clé de sécurité il pourra démarrer des machines en votre nom). Nous vous conseillons de désactiver cet utilisateur à la fin du TP et créer un avec des droits plus spécifiques.

Configuration de votre machine

1. Sur votre machine installez le client AWS ([awscli](https://aws.amazon.com/cli/)
([https://aws.amazon.com/cli/?
sc_channel=PS&sc_campaign=acquisition_FR&sc_publisher=google&sc_medium=english_english_command_li](https://aws.amazon.com/cli/?sc_channel=PS&sc_campaign=acquisition_FR&sc_publisher=google&sc_medium=english_english_command_li)
)
2. Utilisez aws configure pour configurer votre installation (inserez votre *Access Key ID*, votre *Secret Access Key* (que vous avez sauvegardé dans credentials.csv), spécifiez la région par default à *us-east-1*, et le type de log par défaut à *text*:

```
[aar@wifibridge 2018]# aws configure
AWS Access Key ID [None]: *****JVBA
AWS Secret Access Key [None]: *****EiQv
Default region name [None]: us-east-1
Default output format [None]: text
```

Démarrage d'un cluster de 3 noeuds via AWS EMR

Nous allons utiliser la console *AWS EMR* pour démarrer notre cluster:

1. Allez dans la console *AWS EMR*: <https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#> et cliquer sur *Create cluster*
2. Modifiez les paramètres suivants puis validez:
 - a. *GdeltCluster* pour le nom du cluster
 - b. sélectionnez dans Applications \Rightarrow *Spark*
 - c. *instance_type* \Rightarrow *m1.large*
 - d. *key pair* \Rightarrow *gdeltKeyPair*

Create Cluster - Quick Options [Go to advanced options](#)

General Configuration

Cluster name

☒ Logging ⓘ

S3 folder

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

Software configuration

Release ⓘ

Applications

☐ Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.3.1, Hue 4.0.1, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4

☐ HBase: HBase 1.3.1 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 2.3.1, Hue 4.0.1, Phoenix 4.11.0, and ZooKeeper 3.4.10

☐ Presto: Presto 0.187 with Hadoop 2.7.3 HDFS and Hive 2.3.1 Metastore

☒ Spark: Spark 2.2.0 on Hadoop 2.7.3 YARN with Ganglia 3.7.2 and Zeppelin 0.7.3

☐ Use AWS Glue Data Catalog for table metadata ⓘ

Hardware configuration

Instance type

Number of instances (1 master and 2 core nodes)

Security and access

EC2 key pair ⓘ [Learn how to create an EC2 key pair.](#)

Permissions ☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role ⓘ

EC2 instance profile ⓘ

[Cancel](#) [Create cluster](#)

3. Votre cluster est en train de démarrer:

Amazon EMR

Clusters

Security configurations

VPC subnets

Events

Help

Clone

Terminate

AWS CLI export

Cluster: GdeltCluster Starting

Summary

Application history

Monitoring

Hardware

Events

Steps

Configurations

Bootstrap actions

Connections: --

Master public DNS: --

Tags: -- [View All / Edit](#)

Summary

ID: j-DC001QE66DYD

Creation date: 2017-12-17 23:39 (UTC+1)

Elapsed time: 0 seconds

Auto-terminate: No

Termination protection: Off [Change](#)

Security and access

Key name: gdeltKeyPair

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master:

Security groups for Core & Task:

Configuration details

Release label: emr-5.10.0

Hadoop distribution: Amazon 2.7.3

Applications: Ganglia 3.7.2, Spark 2.2.0, Zeppelin 0.7.3

Log URI: s3://aws-logs-486467272538-us-east-1/elasticmapreduce/

EMRFS consistent view: Disabled

Custom AMI ID: --

Network and hardware

Availability zone: --

Subnet ID: subnet-9a19d5c3

Master: Provisioning 1 m1.large

Core: Provisioning 2 m1.large

Task: --

4. Dans quelques minutes votre cluster aura démarré, notez l'adresse de votre master Spark (en rouge):


Cluster: GdeltCluster **Waiting** Cluster ready after last step completed.

[Summary](#) [Application history](#) [Monitoring](#) [Hardware](#) [Events](#) [Steps](#) [Configurations](#) [Bootstrap actions](#)

Connections: [Enable Web Connection](#) – Zeppelin, Spark History Server, Ganglia, Resource Manager ... (View All)

Master public DNS: [ec2-54-208-16-31.compute-1.amazonaws.com](#) [SSH](#)

Tags: – [View All / Edit](#)

Summary	Configuration details	Network and hardware
<p>ID: j-DC001QE66DYD</p> <p>Creation date: 2017-12-17 23:39 (UTC+1)</p> <p>Elapsed time: 12 minutes</p> <p>Auto-terminate: No</p> <p>Termination protection: Off Change</p>	<p>Release label: emr-5.10.0</p> <p>Hadoop distribution: Amazon 2.7.3</p> <p>Applications: Ganglia 3.7.2, Spark 2.2.0, Zeppelin 0.7.3</p> <p>Log URI: s3://aws-logs-486467272538-us-east-1/elasticmapreduce/ </p> <p>EMRFS consistent view: Disabled</p> <p>Custom AMI ID: --</p>	<p>Availability zone: us-east-1a</p> <p>Subnet ID: subnet-9a19d5c3</p> <p>Master: Running 1 m1.large</p> <p>Core: Running 2 m1.large</p> <p>Task: --</p>

Security and access

Key name: gdeltKeyPair

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for [sg-1307c177](#) (ElasticMapReduce-Master: master)

Security groups for [sg-1c07c178](#) (ElasticMapReduce-Core & Task: slave)

Connexion à l'interface du Zeppelin

1. Démarrez un client ssh avec la redirection du port 8890 vers le maître de votre cluster:

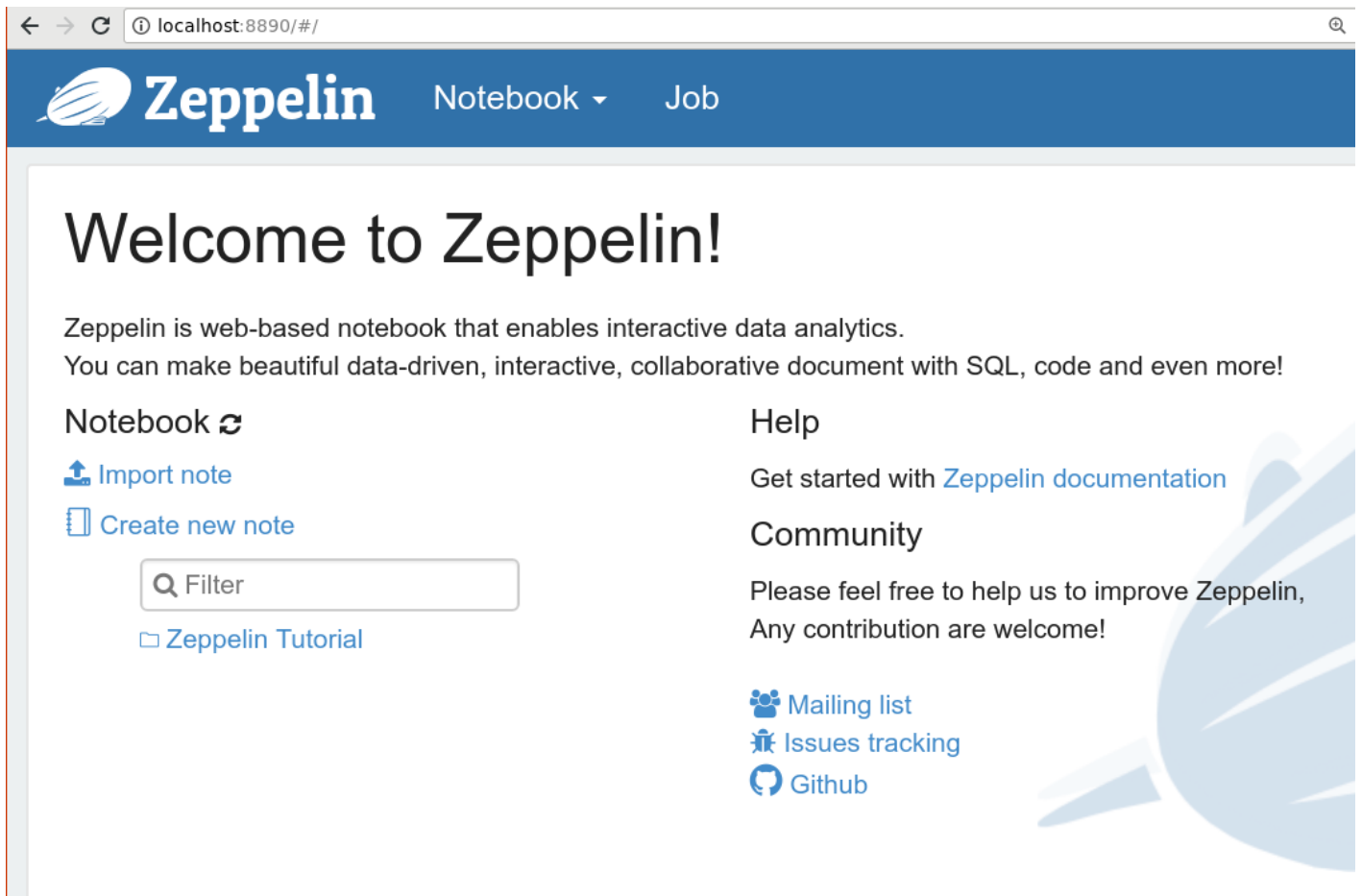
```
[aar@wifibridge 2018]# ssh -i /home/aar/Downloads/gdeltKeyPair.pem\
-L 8890:127.0.0.1:8890 hadoop@ec2-54-208-16-31.compute-1.amazonaws.com 1
bind: Cannot assign requested address
Last login: Sun Dec 17 22:52:44 2017
```

```
  _|  _|_  )
 _| (    /   Amazon Linux AMI
_| \_|_|_|
```

<https://aws.amazon.com/amazon-linux-ami/2017.09-release-notes/>
 9 package(s) needed **for** security, out of 15 available
 Run "**sudo yum update**" to apply all updates.

```
EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRRRRRRRRRRR
E::::::::::::::::::::E M:::::::::M          M:::::::::M R::::::::::::R
EE::::::::EEEEEEEE::::E M:::::::::M          M:::::::::M R::::RRRRRR:::::R
  E::::E          EEEEE M:::::::::M          M:::::::::M RR::::R          R::::R
  E::::E          M:::::::::M:M          M:::M:::::M          R:::R          R::::R
  E::::EEEEEEEEEEE M:::::M M:::M M:::M M:::::M          R::RRRRRR:::::R
  E:::::::::::::E M:::::M M:::M:M::M M:::::M          R::::::::::::RR
  E::::EEEEEEEEEEE M:::::M          M:::::M          R::RRRRRR:::::R
  E::::E          M:::::M          M:::M          M:::::M          R:::R          R::::R
  E::::E          EEEEE M:::::M          MMM          M:::::M          R:::R          R::::R
EE::::::::EEEEEEEE::::E M:::::M          M:::::M          R:::R          R::::R
E:::::::::::::E M:::::M          M:::::M RR::::R          R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRR          RRRRRR
```

2. Ouvrez un navigateur vers <http://localhost:8890> (<http://localhost:8890>) et vous aurez accès à l'interface du Zeppelin



3. Importer le notebook suivant [gdeltExploration.json](http://andreiarion.github.com/gdeltExploration.json)

(<http://andreiarion.github.com/gdeltExploration.json>) et explorer les données

Éteindre votre cluster

Allez a <https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1>, sélectionner votre cluster et cliquez sur Terminate

Last updated 2017-12-19 07:07:25 CET