Projet ISCRI

Voici ce que j’ai fait cette semaine :

* L’import des fichiers automatique depuis GDELT fonctionne avec reprise sur panne, dézippage auto et vérification de la signature MD5
* J’ai créé la BD et optimisé les indexes pour des accès rapides
* J’ai importé entièrement 5 années pour les tests dont 2015
* Le typage des données notamment des dates est amélioré pour des requêtes plus faciles
* J’ai créé une vingtaine de tests automatiques
* J’ai codé le calcul d’un risque journalier (Riskd = ), les tests sont en cours, ce n’est pas encore automatisé
* J’ai prototypé le calcul de l’ISCRI à partir des Riskd dans un fichier Excel pour 2015 pour bien comprendre

Quelques remarques :

* L’ensemble des fichiers fait 272Go
* La base de données fera entre 500Go et 800Go
* 4 fichiers sont absents sur gdelt : 20140123.export.CSV.zip, 20140124.export.CSV.zip, 20140125.export.CSV.zip et 20140319.export.CSV.zip, ça ne me semble pas très grave
* Le GoldsteinScale est parfois absent, c’est assez rare pour les GOV
* **Le AvgTone a l’air d’être systématiquement positif avant 2015-02** : j’ai vérifié ce point aléatoirement sur une dizaine de fichiers entre 2014 et 2010. Rien n’apparait sur ce point dans la doc GDELT
* Certains CountryCode ne sont pas dans la norme ISO3 comme les continents (EUR) ou des sous-continents, pour l’instant ce n’est pas grave mais cela risque de poser un problème sur les GeoMap, j’ai bien vu que GDELT fournissait la liste des codes qui est heureusement souvent compatible ISO3
* L’import d’un fichier prend 50s en moyenne, soit 5 heures pour 1 an, les contraintes d’intégrités sont vérifiées pour chaque ligne afin d’éviter tout problème
* Le GlobalEventId n’est pas une clé unique ce qui m’a posé beaucoup de problème. Il peut se retrouver sur plusieurs fichiers, je pense qu’il s’agit d’une mise à jour de l’event donc le nouveau remplace l’ancien
* Le calcul des Riskd pour toutes les pairs et sa sauvegarde prend < 1s soit moins de 30s pour 1 mois
* J’ai bon espoir que le calcul de l’ISCRI prenne < 2s en live et < 0.1s si précalculé
* Certains events peuvent être ajoutés dans le BD jusqu’à 10 ans après. Par exemple l’export 20241010 possède un event de la date 20141010, du coup le calcul du risque effectué au mois M ne sera peut-être plus valable le mois M+1. En général passé 2 mois se phénomène est rare.

Calcul de l’ISCRI :

J’ai du mal à bien comprendre ton code stata donc je me fie à ta notice Word et à ton mail (% par jour d’événements hostiles dans événements totaux et moyenne mensuelle de ces % journaliers)**, peux-tu valider ma démarche** ?

* **Pour une paire de CountryCode le risque journalier est Risk = NbEvents / NbNegativeEvents** ()
  + Un event = **(Actor1Type1Code == "GOV" | Actor2Type1Code == "GOV") & IsRootEvent == True & Actor1CountryCode != Actor2CountryCode**
    - Ceci représente ton
    - Comme validé par mail c’est bien un OR sur les GOV
  + Un NegativeEvent = Event où **QuadClass >= 3 & GoldsteinScale < 0 & AvgTone < 0**
    - Ceci représente ton ton
    - Je fais de même pour les Events de QuadClass == 3 et QuadClass == 4
  + **L’AvgTone ne semble pas fiable avant 2015-02, es-tu au courant ?**
  + Je stocke tous les calculs intermédiaires dans la BD pour faciliter le débogage
  + **Pourquoi dans ton code stata tu utilises NumArticles** ? Est-il utile dans ce calcul ? Car rien n’apparait sur ce point dans la notice Word
* Puis je calcule le risque mensuel que je nomme **Riskm =**  pour une paire de CountryCode en faisant la somme des Riskd divisé par le nombre de jour du mois, il s’agit donc de la moyenne mensuelle des Riskd.
  + Je fais de même pour QuadClass=3 ou 4
  + Il sera facile + tard de faire une convolution sur une fenêtre de 30j
* J’ai prototypé le calcul du Riskm dans un fichier Excel en partant de mes Riskd. Pour être indépendant de la graine utilisée par l’ISCRI, j’ai utilisé le calcul suivant Riskm = ISCRIm – 0.9 \* ISCRIm-1
  + Les chiffres que tu exposes dans ta notice Word pour l’année 2015 sont-ils fiables ? Car les valeurs que je trouve dans mon proto sont certains mois identiques (2015-07, 2015-10, …) et parfois assez différentes (2015-02, …)

La semaine prochaine j’implémente l’ISCRI et commence à automatiser les imports SQL et les calculs.