Une plateforme haute performance pour l'exploitation des données massives

Application aux données des réseaux sociaux

Annabelle Gillet & Éric Leclercq & Nadine Cullot

LIB - EA 7534 Université de Bourgogne Franche-Comté
Équipe Science des Données

DataBFC 2019
2e édition - Dijon - 20 au 22 novembre 2019


Plan

1. Entreposage et Analyse des Big Data
   - Problématique principale / verrous
   - Utilisation des données de Twitter
   - Spécificités des données de Twitter
2. Architectures haute performance
   - Objectifs et problématiques
   - Architectures existantes
3. Architecture et plateforme Hydre
   - Architecture générale
   - Performances
   - Utilisation
4. Bilan et conclusion

Les enjeux des Big Data

Les Big Data et les V caractéristiques comme Vélocité, Variété, Variabilité soulèvent de nouveaux enjeux pour les systèmes de gestion de données au niveau :

1. du stockage et de la restitution des données (interrogation)
2. du traitement en temps réel des flux de données
3. de la manipulation des données par différents algorithmes (différents modèles de représentation des données et différents algorithmes s'appuyant sur des paradigmes de programmation)
   - **modèles :** objets/attributs, séries temporelles, matrices, graphes, tenseurs, etc.
   - **paradigmes :** programmation GPU, cluster (map-reduce), concurrente, parallèle, fonctionnelle, etc.

Comment collecter les données et alimenter rapidement les différents algorithmes d'analyse ?

Contexte (1/3)

L'analyse des données de Twitter peut être appliquée dans différents domaines comme par exemple :

- le marketing, la santé, la politique, la gestion des crises (ex. sanitaires, environnementales)

Mais les objectifs sont différents (non exclusifs) :

- construire des modèles descriptifs, explicatifs, prédictifs
- élaborer des connaissances

Contexte (2/3)

Projets interdisciplinaires développés avec des chercheurs en sciences sociales et en sciences de la communication pour étudier la structure et la circulation des discours sur Twitter.

Plus précisément, il s'agit d'étudier comment Twitter reflète :

- Les questions de société dans la communication politique
- Les crises alimentaires, sanitaires, des habitudes de consommation
- La prise en compte des problèmes environnementaux

Collaboration scientifique établie depuis 2014, soutenue par des projets de recherche :

- Twitter aux Élections Européennes de 2014 et 2019 (TEE 2014, TEE 2019)
- Élections Présidentielles de 2017
- ISITE UBFC Cocktail 2019

Contexte (3/3)

Du point de vue de l'analyse de données les objectifs des différents projets s'articulent autour de :

- **Détection des sujets émergents** (par hashtags ou topics)
- **Détection et caractérisation des communautés** d'utilisateurs
- **Détection et caractérisation d'évènements**
- Étude le la structure de la communication, la circulation des discours
- Détections des **messages viraux et du rôle des robots** :
  - dans la circulation de l'information
  - dans la structuration des communautés

Spécificités des données de Twitter

Les tweets contiennent de nombreuses informations et révèlent un **sémantique complexe** (hashtags, mentions, retweets, etc).

Twitter propose des API pour collecter les tweets en respectant certaines limites :

- 1% du flux total (en moyenne 6 000 tweets/s) en temps réel
- récupération des tweets historiques datant au plus de 7 jours

Plan

1 Entreposage et Analyse des Big Data

2 Architectures haute performance
- Objectifs et problématiques
- Architectures existantes

3 Architecture et plateforme Hydre

4 Bilan et conclusion

Objectifs et problématique

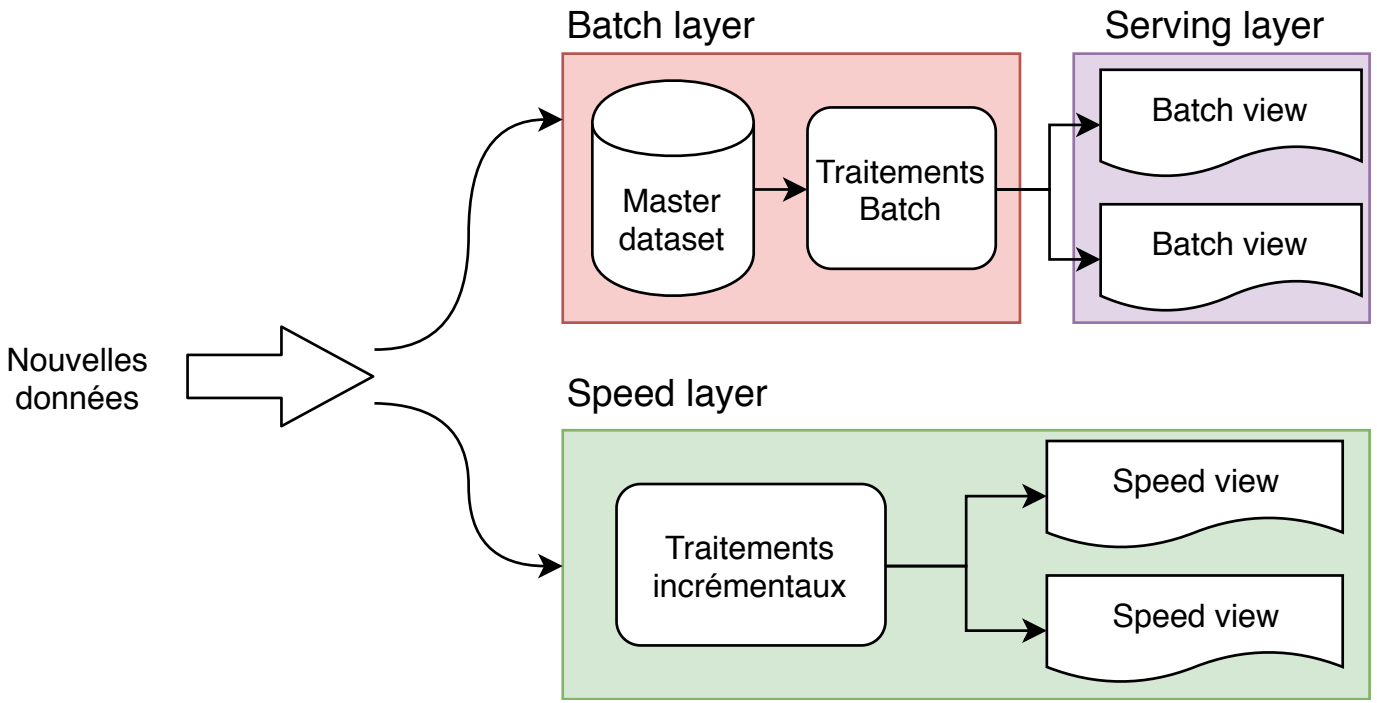
Concevoir et réaliser une **plateforme basée sur une architecture à haute performance** pour **collecter**, **stocker** et **supporter des analyses** des données des réseaux sociaux.

Problématiques :

- Collecter les données en étant capable d'absorber un flux important de plusieurs milliers de messages par seconde
- Extraire les informations des messages (ex. les composants d'un tweet)
- Effectuer des analyses avec une faible latence ("temps réel")
- Effectuer des analyses fines plus complexes en temps différé
  - Transformer les données rapidement pour s'adapter aux différents algorithmes (alg. lineaire, graphes, machine learning, etc.)

Architectures existantes : Lambda Architecture

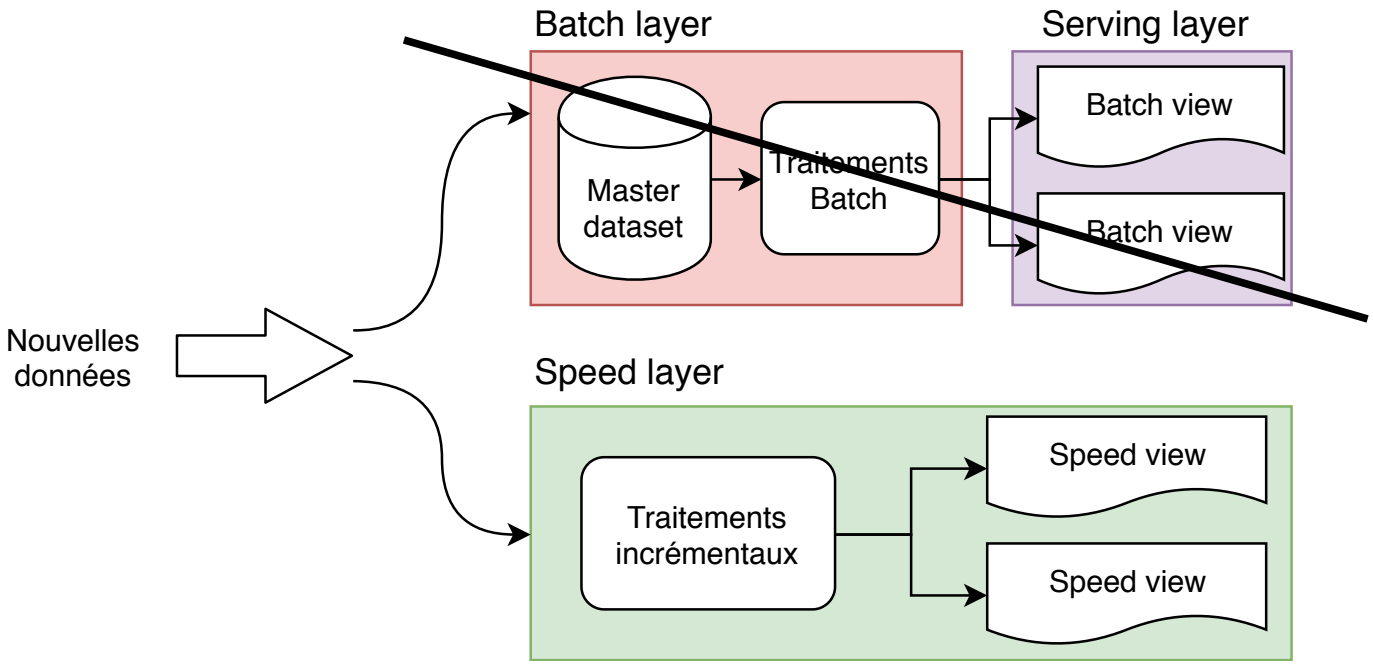
Patron d'architecture spécifié en 2011 par Marz.
Traitement des **données massives** avec une **faible latence** et par lot de manière simultanée, avec une **forte résistance aux pannes**.



Architectures existantes : Kappa Architecture

Définie par Kreps :

- *everything is a stream*
- Peu de flexibilité, les algorithmes doivent s'adapter à une structure de flux
- Pas de conservation des données brutes



Synthèse

Du point de vue des données

- La Lambda Architecture (LA) offre une conservation des données brutes pour une possibilité de retraitement
- LA et Kappa Architecture (KA) sont orientées vers la production de résultats pour des requêtes pré-définies
- LA et KA ne sont pas centrées données et schéma
- LA : agrégation des vues *Batch* et *Speed* floue

Du point de vue de l'organisation des traitements

- LA et KA respectent le principe de faible latence
- KA : toutes les analyses ne sont pas adaptées au stream processing
- LA : traitements en double avec objectif d'arriver au même résultat
- LA difficultés techniques : différentes technologies avec des concepts différents à utiliser

Plan

1 Entreposage et Analyse des Big Data

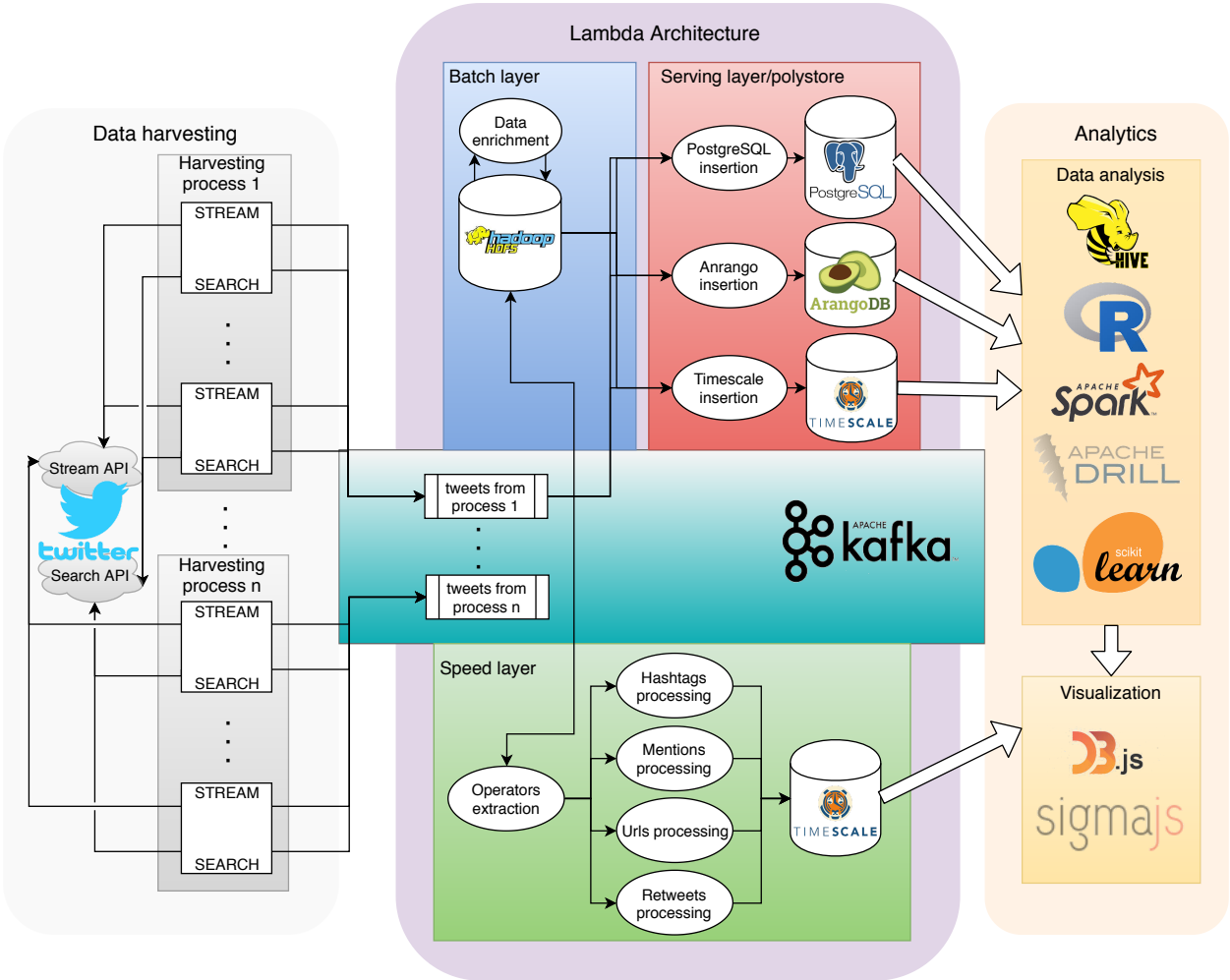
2 Architectures haute performance

3 Architecture et plateforme Hydre
- Architecture générale
- Performances
- Utilisation

4 Bilan et conclusion

Architecture générale

Hydre est une évolution de SNFreezer que nous avons réalisée en 2014 pour le projet TEE 2014 (Licence GPL), maintenue depuis 2019 par la MSH de Dijon.



Mesures de performance

Nombre de processus	Kafka ingestion	Stockage HDFS	Schéma non-normalisé	Schéma normalisé	Insertion Neo4j	Calcul d'indicateurs
1	3 082	7 181	1 840	460	10	1 091
2	5 615 (1.86)	10 194 (1.48)	3 508 (1.95)	855 (1.89)	19 (1.94)	2 149 (2.04)
3	7 227 (2.29)	11 364 (1.69)	4 977 (2.63)	1 210 (2.70)	26 (2.72)	3 010 (2.81)
4	9 128 (2.98)	12 017 (1.79)	6 089 (3.35)	1 563 (3.49)	**30** (3.45)	3 459 (3.47)

Nombre de tweets/s traités par chaque composant (et facteur d'accélération).

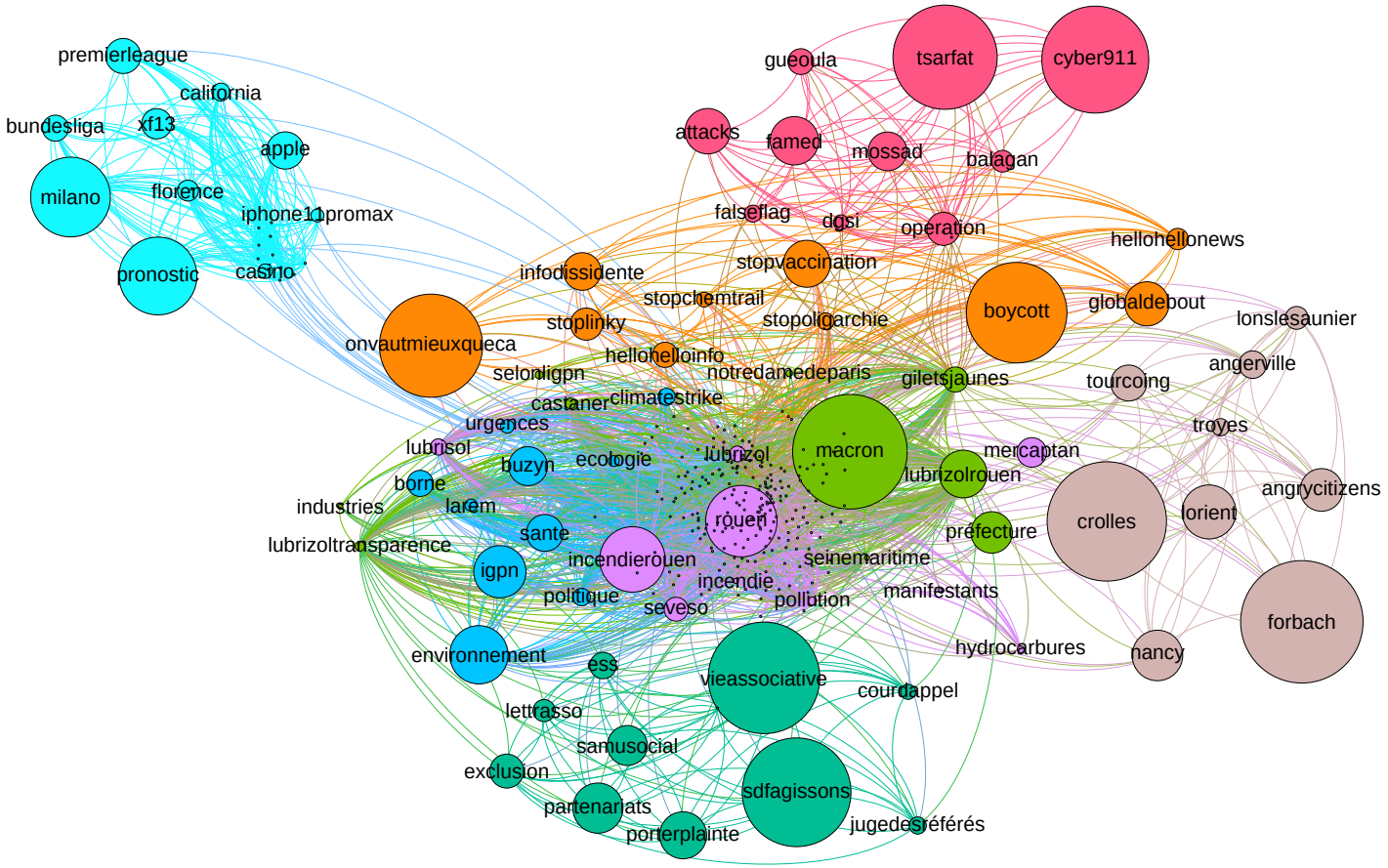
La plateforme a été et est utilisée pour :

- collecter des tweets sur le Brexit en Écosse : 45 millions
- collecter des tweets sur les Élections Européennes 2019 : 20 millions
- collecter des tweets autour de l'alimentation et la santé : 30 millions depuis 2 mois
  - dont tweets autour de Lubrizol : 400 000

Monitoring

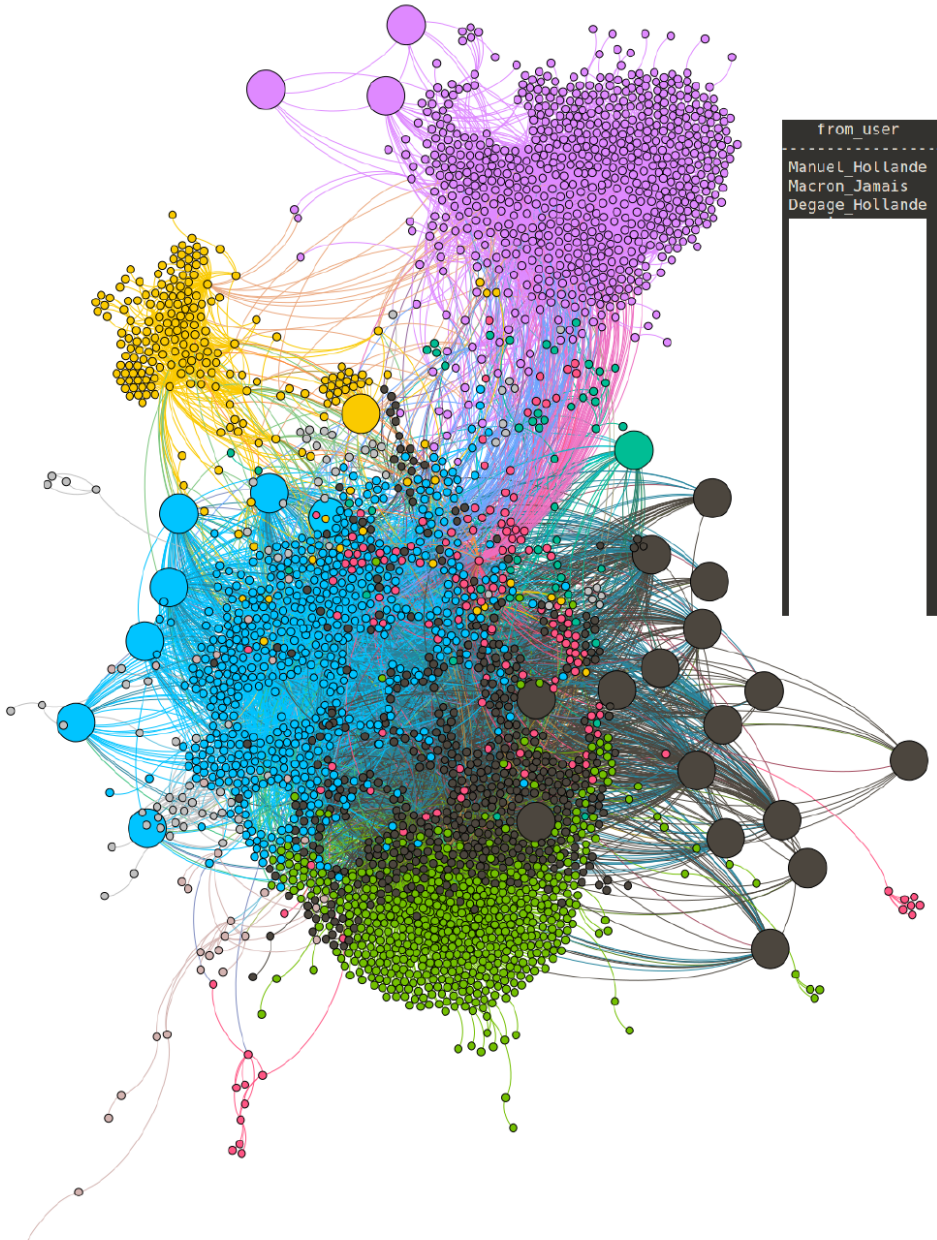
Écran de monitoring pour surveiller le bon déroulement des collectes.

Exemples d'analyses : Lubrizol



Identification des communautés thématiques en utilisant le graphe des cooccurrences des hashtags.

Exemples d'analyses : détection des robots dans les communautés durant l'élection présidentielle 2017



from_user	user_id	nbrt	nbrtjour	nbtweetp2	proba	retrycount	max
Manuel_Hollande	639	1073	76	1078	0.66	1	
Macron_Jamais	622	1481	105	1706	0.61	1	
Degage_Hollande	277	1394	99	1464	0.66	0	
	724	1594	113	5688	0.5	0	
	933	3405	243	3727	0.31	1	
	477	2037	145	2370	0.39	0	
	761	2254	161	2263	0.3	1	
	1071	2289	163	2538	0.39	2	
	197	1988	142	2415	0.37	0	
	932	1967	140	2041	0.35	0	
	158	1932	138	1951	0.38	0	
	1070	1814	129	1820	0.54	0	
	979	1472	105	2396		9	
	712	1365	97	1428	0.48	0	
	968	4513	322	4558	0.39	1	
	248	6466	461	6672	0.5	0	
	345	6134	438	6549	0.34	0	
	446	5228	373	5336	0.34	0	
	112	1291	92	1395	0.69	0	
	834	7001	500	7105		8	
	877	1223	87	1243	0.42	4	
	324	1108	79	1111	0.4	0	
	347	3525	251	3908	0.49	0	
	896	1509	107	1509	0.62	0	

Conclusion et perspectives

Extensions au niveau des données :

- Traiter d'autres sources de données Web (vidéo, images, données géolocalisées, traces d'interactions, blogs, etc).
- Enrichir les données avec des Linked Open Data et des ontologies

Extensions autour des outils de traitement :

- Enrichir les algorithmes disponibles dans les outils interfacés avec la plateforme (Jupyter, R, Scala/Spark)
- Abstraire les analyses pour construire des workflows réutilisables
- Compléter le modèle de données tensoriel du polystore pour utiliser d'autres frameworks d'analyse : p. ex. Tensor Flow et des algorithmes de deep learning

Merci pour votre attention.

Ce travail est soutenu par le programme « Investissements d'Avenir », projet ISITE-BFC (contrat ANR-15-IDEX-0003). Le projet Cocktail est piloté scientifiquement par Gilles Brachotte, laboratoire CIMEOS EA-4177, Université de Bourgogne.