

L'écosystème régional du Big Data.

Retours d'expérience sur les technologies Big Data.

Robert VISEUR
Senior R&D Expert

Big Data Week 2015 : Bridging data events all over the world - Gosselies, 24/11/2015

FEDER



UNION EUROPEENNE



Wallonie



Centre d'Excellence en **Technologies** de
l'**Information** et de la **Communication**

www.cetic.be

Qui suis-je ?

- Dr Ir Robert VISEUR.
- Formation : Ingénieur Civil (FPMs), Mastère en Management de l'Innovation, Docteur en Sciences appliquées.
- Compétences : management de l'innovation, co-création, open source et traitement de l'information.
- Affiliations :
 - Assistant, UMONS Faculté Polytechnique (www.umons.ac.be), Service de Management de l'Innovation Technologique (Pr Vankerem).
 - Senior R&D Expert, CETIC (www.cetic.be).
- Photographe indépendant (www.derrierelevisueur.be).
- Plus d'informations : www.robertviseur.be.

Qu'est-ce que le Big Data ? #1

- Définitions ?
 - Mégadonnées (via “Journal officiel de la République française”) :

“Données structurées ou non dont le très grand volume requiert des outils d'analyse adaptés”.

- Trois V du Big Data :
 - Volume,
 - Vitesse,
 - Variété.
 - (+ Véracité, Valeur)

Qu'est-ce que le Big Data ? #2

- Limites du marché du Big Data ?
 - Données :
 - Web, smartphones, RFID, réseaux sociaux (Facebook, Twitter), capteurs (IoT : smart city, machines, énergie,...), images,...
 - Open data (secteur public).
 - Quadrant magique Gartner : Big Data comme tendance en Business Intelligence / Analytics.
- Limites du marché assez vagues.

Marché du Big Data

- Marché mondial (Transparency Market Research) : marché de 8,9 milliards USD (2014) et 24,6 milliards USD (2016, prévisions).
- Marché français.
 - Marché de 9 milliards EUR (2020).
 - Plan Big Data (centre de ressources technologiques, formations et accélérateurs de startups).
- Possibilité d'analyser les publications associées aux salons et aux syndicats professionnels.

Equipementiers Hardware Editeurs Software Editeurs BI
Editeurs Dataviz Consultants SSII Intégrateurs Institutionnels



Labels Annonceurs	Equipementier Hardware	Editeur Software	Editeur BI	Editeur Data	Consult / SSII Intégrateur	Institutionnel	Page
ACTUATE		*	*	*			44
A.I.D				*	*		46
AMAZON WEB SERVICES		*	*				48
BIME ANALYTICS				*			50
BITTLE		*	*				52
BLUESTONE					*		54
CAPGEMINI					*		56
CENTRALE FORMATION						*	58
CLOUDERA		*					60
CLUB DECISION DSI						*	62
CSC					*		64
DATASIX		*			*		66
EDIS CONSULTING					*		68
EKIMETRICS			*	*	*		70
EXALEAD DASSAULT SYSTEMES		*		*			72
OFI INFORMATIQUE					*		74
HP	*		*				76
IBM		*	*	*			78
KEYRUS					*		80
MICROPOLE					*		82
NEO TECHNOLOGY		*					84
PENTAHO		*	*	*			86
PIVOTAL		*	*	*			88
QLIK			*	*			90
SAP		*	*	*			92
SAS		*	*	*			94
SCINETIK			*	*	*		96
SEMDEE			*				98
SENTELIS					*		100
SINEQUA		*	*	*			102
SPLUNK		*	*	*			104
SYNCSORT		*	*	*			106
TABLEAU SOFTWARE		*	*	*			108
TALEND		*	*	*			110
TERADATA	*	*	*	*			112
TIBCO JASPERSOFT			*	*			114
VEISO				*	*		116

Fournisseurs

BIGDATA

Origine / Open source

- Open source.
 - Apache Hadoop (hadoop.apache.org).
 - Créé par Doug Cutting (www.linkedin.com/in/cutting).
 - Employé chez Yahoo!, puis Cloudera.
 - Travail sur Lucene (indexer plein texte), puis...
 - Travail sur Nutch (moteur de recherche open source).
 - Développement d'Hadoop (HDFS).
 - Base : publications de Google (GFS, MapReduce).
 - Hébergé par la Fondation Apache (2006).

- Apache Hadoop (hadoop.apache.org).
 - Services : Cloudera, Hortonworks & MapR.
 - Cloudera (2008) :
 - Cloudera Distribution for Hadoop (CDH) (open source) + Cloudera Management Suite (propriétaire).
 - Plus d'infos : www.cloudera.com.
 - Hortonworks (2011) :
 - Hortonworks Data Platform (HDP) (open source).
 - Plus d'infos : www.hortonworks.com.
 - MapR (2009) :
 - MapR Distribution (propriétaire).
 - Plus d'infos : www.mapr.com.

Origine / Fondation Apache

- Ecosystème Hadoop soutenu par la Fondation Apache.
 - Projets officiels (+ incubateur).
 - Exemples :
 - [PIG \(pig.apache.org\)](http://pig.apache.org).
 - [Hive \(hive.apache.org\)](http://hive.apache.org).
 - [HBase \(hbase.apache.org\)](http://hbase.apache.org).
 - [Zookeeper \(zookeeper.apache.org\)](http://zookeeper.apache.org).
 - [Apache Spark \(spark.apache.org\)](http://spark.apache.org).
 - [Apache Storm \(storm.apache.org\)](http://storm.apache.org).
 - ...

Origine / Propriétaires

- Propriétaires.
 - Univers BI : Qlik, SAS, SPSS,...
 - Remarque : BI et Analytics confondus avec Big Data (analystes).
- Base fréquente : Apache Hadoop.

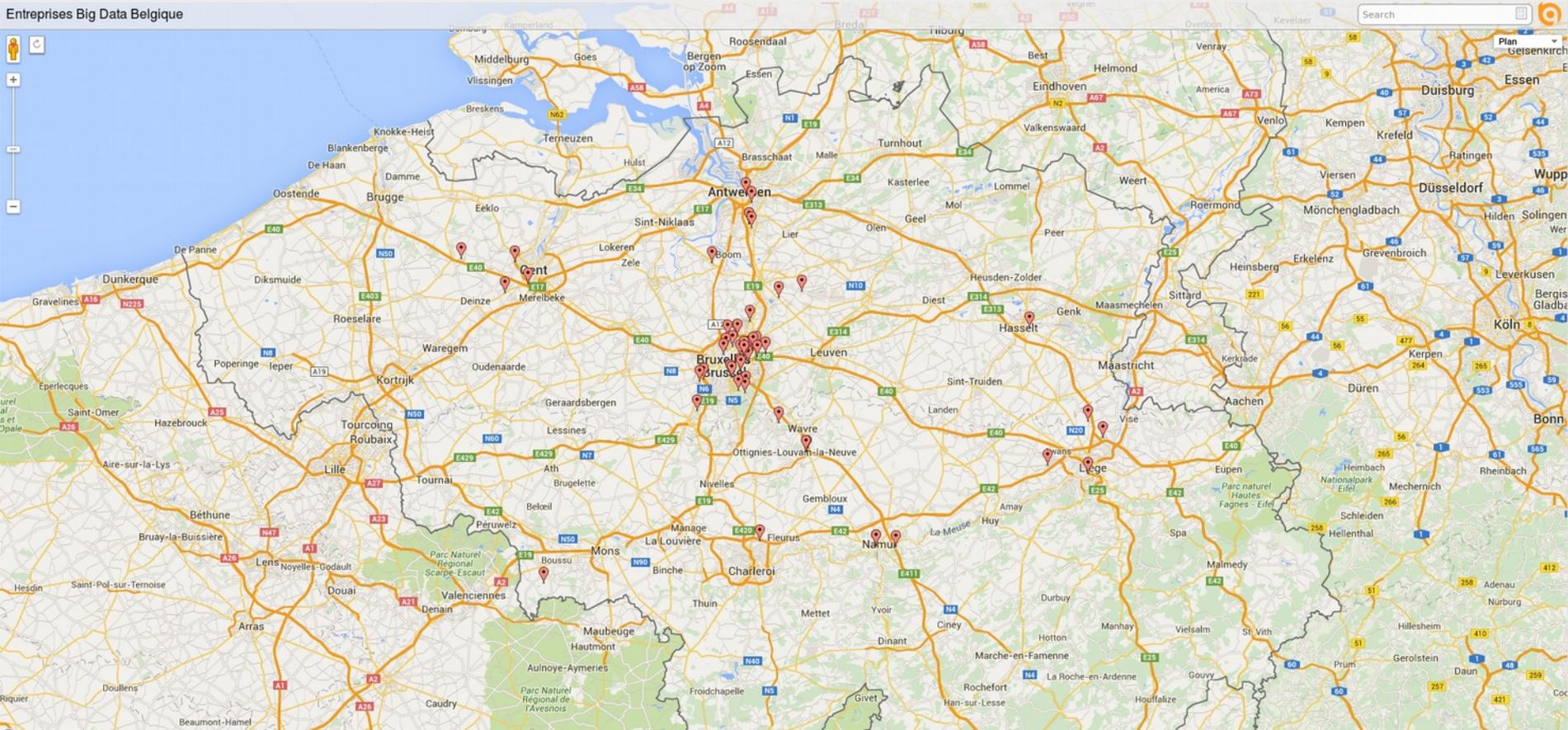
- Jeu de données :
 - Liste de noms tirée d'Agoria Data Alliance.
 - Ajout d'autres entreprises (connaissance marché CETIC).
 - Création d'un fichier complet.
- Création d'un index spécialisé.
 - requêtes + extraction des hyperliens.
- Annotation des compétences (base : site Web).
 - En manuel, puis automatique.
 - Création de tableaux d'activités (par types de produits, produits, technologies,...).
- Création d'un graphe de liens entre entreprises.
- Remarque : expérience préalable sur le marché OSS.
 - Ce qui suit = première analyse.

Marché belge / wallon #1

Bruxelles	19
Flandre	24
Brabant flamand	13
Anvers	6
Limbourg	1
Flandre-Orientale	4
Wallonie	14
Brabant wallon	4
Namur	3
Hainaut	2
Liège	5

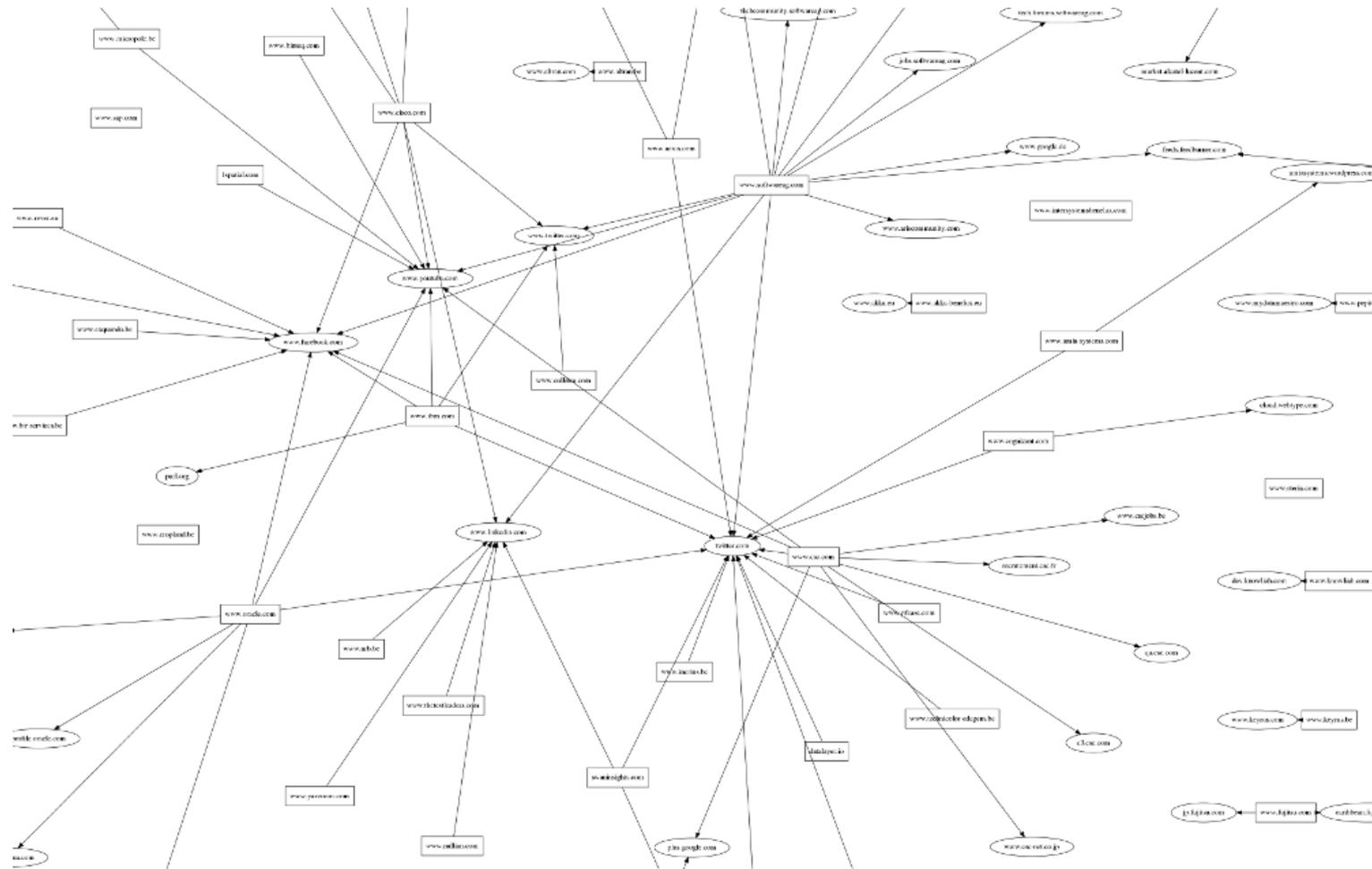
(nombre d'entreprises Big Data par province / région)

Marché belge / wallon #2



(géolocalisation des prestataires ; visualisation : batchgeo.com)

Liens entre entreprises #1



Liens entre entreprises #2

- Peu de liens apparents (via le graphe d'hyperliens).
 - Excepté pour les fournisseurs technologiques (via annotation).
- Indice de la faible maturité et de la faible structuration du marché ?
- Par contraste : marché open source en Belgique.
 - Ex. : importance des centres de compétence (formation).

- Possibilités via les circuits classiques ?
- Possibilités via les MOOCs.
 - Exemples :
 - “Bioinformatics: Genomes and Algorithms”,
 - “Fondamentaux pour le Big Data”,
 - “Introduction à la statistique avec R”,
 - “Comprendre le Big Data à travers les films de cinéma”,
 - ...

Organisations Big Data #1

- Services.
 - Consultance (conseil, intégration,...).
 - Ex. : Altran, Hermes, Steria,...
 - BI, analytics, reporting (dashboard), warehousing.
 - Ex. : Pfease.
 - Pure / new players.
 - Ex. : Datalayer, Nextlab / Data Fellas,...
 - Spécialisation.
 - Ex. : 1Spatial ou PROjections / GIM (données géographiques), iCare (maintenance industrielle), Qualifio (réseaux sociaux), Real Impact Analytics (données télécoms), Rever (qualité des données),...
 - Remarque : trajectoire vers le Big Data plus ou moins planifiée suivant l'entreprise.

Organisations Big Data #2

- Développeurs de produits.
 - Développement.
 - Ex. : IBM, Oracle,...
 - Base / partenariat open source.
 - Ex. : InfoSphere BigInsights (Hadoop), Oracle avec Cloudera,...
- Revendeurs.
- Ex. : NRB.

Organisations Big Data #3

- Plate-formes / DaaS / hosting.
 - Généralistes.
 - Ex. : Amazon AWS, IBM Softlayer,...
 - Généralistes outillés
 - Ex. : Microsoft Azure.
 - Remarque : Microsoft Azure HDInsight (base Hadoop).
 - Spécialisés.
 - Ex. : Swan Insights (Swanalytics), Virdata (IoT / data processing),...

Organisations Big Data #4

- Matériel / Appliance.
 - Ex. : Oracle (ex. : Big Data Appliance), Cisco (ex. : Cisco Tidal Enterprise Scheduler),...
- Secteur public.
 - Open Data.
 - Ex. : UrbIS / OSM (cartographie).
- Utilisateurs finaux.
 - Ex. : Data.be (open data).

Remarque : certaines organisations citées sont localisées à Bruxelles et en Flandre.

Activités #1

- Business Intelligence / Analytics : ~1/4 du marché.
- Détails :

Business Intelligence

Cognos	10	19.23%
Qlik	12	23.08%
SAS	18	34.62%
Spotfire	4	7.69%
SPSS	7	13.46%
Pentaho	1	1.92%

Total: 52 prestataires. Source: CETIC (2015).

PaaS / IaaS

Amazon AWS	11	21.15%
Microsoft Azure	7	13.46%
IBM Softlayer	2	3.85%
Virdata	1	1.92%

Total: 52 prestataires. Source: CETIC (2015).

Activités #2

Open source

Apache (hors Hadoop)	16	30.77%
Hadoop	17	32.69%
NoSQL	11	21.15%
Pentaho	1	1.92%
Talend	5	9.62%

Total: 52 prestataires. Source: CETIC (2015).

Hadoop

Cloudera	10	19.23%
Hortonworks	7	13.46%
MapR	6	11.54%

Total: 52 prestataires. Source: CETIC (2015).

NoSQL

Cassandra	3	5.77%
CouchDB	1	1.92%
HBase	3	5.77%
MongoDB	10	19.23%
Neo4J	1	1.92%

Total: 52 prestataires. Source: CETIC (2015).

- Résultats :
 - Première cartographie du marché belge / wallon du Big Data.
 - Localisation des entreprises.
 - Analyse des liens entre sites Web.
 - Cartographie des compétences (semi-automatique).
- Améliorations :
 - Mise à jour de la base de données d'entreprises (→ contact).
 - Par exemple : veille sur les sites d'annonces d'événements Big Data (→ identification de startups).
 - Finalisation du travail de cartographie des compétences.
 - Infographies à améliorer, nouvelles questions à aborder (?).

>> Questions ? Remarques ? Suggestions ? <<

