VERSION 1.0 - MARS 2012

BIG DATA & OPEN SOURCE: *UNE CONVERGENCE INÉVITABLE?*

Stefane Fermigier

Table des matières

Introduction

Le “Big Data” recouvre de manière lâche les défis, les opportunités et les technologies impliquées par le **déluge des données** produites depuis quelques années par les entreprises. Par exemple, selon IDC, ce sont 1.8 Zettaoctets qui ont été produits en 2011 (l’équivalent d’un milliard de disque durs de grande capacité récents), un chiffre qui continue à augmenter de 50% chaque année.

Une définition plus précise, donnée par Wikipedia[1], indique qu’il s’agit d’une “expression anglophone utilisée pour désigner des **ensembles de données** qui deviennent tellement gros qu’ils en deviennent **difficiles à travailler** avec des **outils classiques** de gestion de base de données”. Ces derniers (bases de données relationnelles, principalement) ont en effet pour caractéristique de ne pouvoir monter en charge que de manière **verticale** (i.e. en augmentant la puissance d’un seul serveur) jusqu’à atteindre des prix prohibitifs. Par opposition, les outils utilisés dans le domaine des Big Data visent à atteindre une scalabilité **horizontale** (i.e. obtenue en rajoutant des serveurs à bas coût), au prix d’un renoncement au modèle de donnée relationnel et/ou au modèle transactionnel.

Les **enjeux économiques** sont considérables: c’est en étant les premiers à comprendre l’intérêt, et à maîtriser les difficultés techniques, du traitement des données issues des interactions de leurs utilisateurs avec leurs services, que des sociétés web comme Google, Amazon, Yahoo! ou Facebook ont réussi à provoquer une **disruption massive** de leur marché (“web 2.0” vs. “web 1.0”) et à s’imposer comme les leaders de leur catégorie. Dans le domaine scientifique, on voit émerger depuis quelques années des sous-disciplines (“data science”) entièrement fondées sur le traitement massif de données scientifique. Enfin, pour d’autres acteurs (ex: grande distribution), le Big Data ne représente pas une opportunité de disruption par un modèle nouveau, mais un moyen de plus en plus incontournable d’**optimiser leur efficience** et donc leur compétitivité.

[1] <http://fr.wikipedia.org/wiki/Big_data>.

Contexte économique et technologique

L'origine des données du Big Data

Les données traitées par le Big Data proviennent notamment[2] :

- du **Web**: journaux d'accès, réseaux sociaux, e-commerce, indexation, stockage de documents, de photos, de vidéos, *linked data*, etc. (ex: Google traitait 24 petaoctets de données par jour avec MapReduce en 2009).[3]
- plus généralement, de l'**internet** et des **objets communicants**: RFID, réseaux de capteurs, journaux des appels en téléphonie;
- des **sciences**: génomique, astronomie, physique subatomique (ex: le CERN annonce produire 15 petaoctets de données par an avec le LHC), climatologie (ex: le centre de recherche allemand sur le climat gère une base de données de 60 petaoctets), etc.;
- données **commerciales** (ex: historique des transactions dans une chaîne d'hypermarchés);
- données **personnelles** (ex: dossiers médicaux);
- données **publiques** (open data).

Les principaux acteurs

Parmi ces catégories, le monde du web a été le précurseur du mouvement (l'expression "web scale" a longtemps été synonyme de "big data"), et il n'est pas étonnant que les principales innovations du domaine trouvent leur origine chez les leaders du Web: Google (MapReduce et BigTable), Amazon (Dynamo, S3), Yahoo! (Hadoop, S4), Facebook (Cassandra, Hive), Twitter (Storm, FlockDB), LinkedIn (Kafka, SenseiDB, Voldemort), LiveJournal (Memcached), etc.

Compte-tenu de la culture et du modèle économique de ces sociétés, il n'est pas étonnant non plus que la plupart de ces projets soient open source, souvent développés de manière

---

[2] Source: wikipedia, op. cit.

[3] Cf. par exemple: *The Great Disk Drive in the Sky: How Web giants store big—and we mean big—data*, Ars Technica, janvier 2012. <http://arstechnica.com/business/news/2012/01/the-big-disk-drive-in-the-sky-how-the-giants-of-the-web-store-big-data.ars/>

collaborative après ouverture initiale de code développé en interne, et parfois confié à une entité extérieure.

La Fondation Apache est ainsi particulièrement active dans ce domaine, en lançant ou en recueillant plus d'une dizaine de projets, matures ou en incubation: Hadoop, Lucene/Solr, Hbase, Hive, Pig, Cassandra, Mahout, Zookeeper, S4, Storm, Kafka, Flume, Hama, Giraph, etc.

Outre les sociétés du Web, le secteur scientifique et plus récemment les promoteurs de l'Open Data (et de sa variante, l'Open Linked Data, issu du Web Sémantique), sont également historiquement très ouverts à l'open source, et ont logiquement effectué des contributions importantes dans le domaine du Big Data.

La plupart de ces technologies open source ont par ailleurs donné lieu à la création de startups, massivement financées pour certaines. Par exemple, autour de Hadoop, on peut citer: Cloudera (76M$ levés), Hortonworks (~20M$), Datameer (12M$), Zettaset, Drawntoscale, etc.

Les grands acteurs des logiciels et systèmes d'entreprises ne sont pas épargnés par cette vague du Big Data open source: Oracle a mis Hadoop au coeur de son "*big data appliance*" lancé en octobre 2011[4]; Microsoft a annoncé en novembre 2011 l'arrêt de son projet interne de MapReduce pour Azure (baptisé "Dryad") au profit d'Hadoop[5]; IBM, EMC et Netapp ont également intégré Hadoop dans leur offre de big data.

Les enjeux technologiques

Michael Stonebraker, "pape" de la base de données depuis 30 ans, déclarait récemment dans une interview au MagIT:

> *"Il y a beaucoup de bruit autour du Big Data. Ce concept a plusieurs significations en fonction du type de personnes. Selon moi, la meilleure façon de considérer le Big Data est de penser au concept de trois V. Big Data peut être synonyme de gros* ***volume****. Du teraoctet au petaoctet. Il peut également signifier la* ***rapidité*** *[Velocity, NDLR] de traitement de flux continus de données. Enfin, la troisième signification : vous avez à*

4 *Oracle Big Data Appliance stakes big claim*, GigaOM, 3 octobre 2011, et *Cloudera puts the Hadoop in Oracle's Big Data Appliance*, GigaOM, 10 janvier 2012.

5 "Dryad was intended to run big-data jobs across HPC, Microsoft's clustered server environment. But such a release would have presented a proprietary and competing alternative to Hadoop, which is rapidly emerging as the leading platform for distributed data processing." Source: Information Week, 17 novembre 2011.

*manipuler une grande **variété** de données, de sources hétérogènes. Vous avez à intégrer entre mille et deux mille sources de données différentes et l'opération est un calvaire. La vérité est que le Big Data a bien trois significations et que les éditeurs n'en abordent qu'une à la fois. Il est important de connaître leur positionnement pour leur poser les bonnes questions."*

Alex Popescu[6], suivant l'avis des analystes de Forrester Research, ajoute à cela un quatrième "V", celui de "variabilité", pour aboutir aux critères suivants:

- **Volume**: les données dépassent les limites de la scalabilité verticale des outils classiques, nécessitant des solutions de stockage distribués et des outils de traitement parallèles.
- **Variété**: les données sont hétérogènes ce qui rend leur intégration complexe et coûteuse.
- **Vélocité**: les données doivent être traitées et analysées rapidement eu égard à la vitesse de leur capture.
- **Variabilité**: le format et le sens des données peut varier au fil du temps.

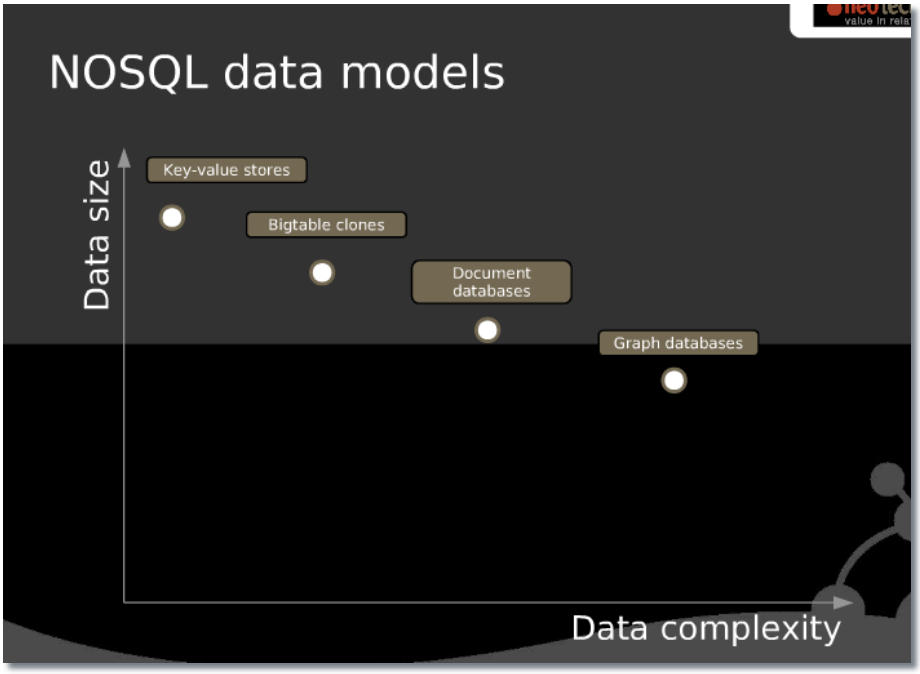
[6] Alex Popescu "Big Data Causes Concern and Big Confusion. A Big Data Definition to Help Clarify the Confusion", 27 février 2012 <http://nosql.mypopescu.com/post/18376617501/big-data-causes-concern-and-big-confusion-a-big-data>.

Le stockage

Bases NoSQL

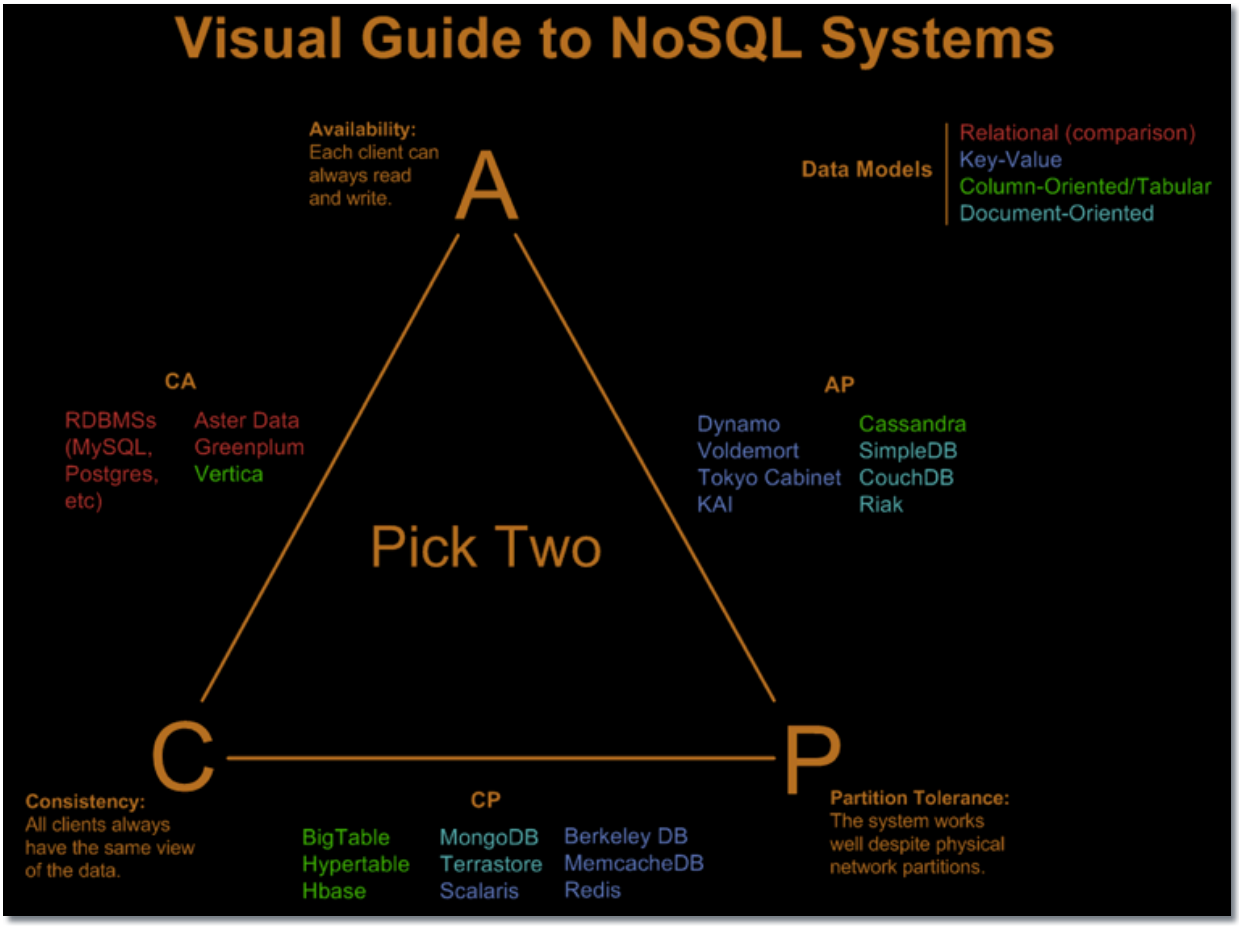
Les bases NoSQL visent à passer à l'échelle de manière horizontale en relâchant les conditions fortes de transactionnalité (ACID - atomiques, cohérentes, isolées et durables) attendues des bases traditionnelles, et en renonçant au modèle relationnel. On distingue actuellement 4 types de bases NoSQL:

- Clé-valeur (ex: Memcached)
- "Orientées colonne" ou "clones de BigTable" (ex: Cassandra)
- "Orientées document" (ex: CouchDB, MongoDB)
- Graphe (ex: Neo4j).



Chacune de ces catégories présente des caractéristiques différentes en termes de scalabilité horizontale (par exemple, les bases orientées graphes ne passent pas aussi facilement à l'échelle horizontalement, mais sont pourtant indispensables pour traiter efficacement les données issues des réseaux sociaux). De plus, au sein de chaque catégorie, différents compromis en termes de cohérence, disponibilité et résistance au morcellement (attendu qu'il est impossible,

selon le théorème de Brewer (aussi appelé “théorème CAP”), d’avoir ces trois caractéristiques simultanément dans un système distribué).



La majorité des technologies NoSQL sont open source (cf. annexe de ce document). Celles qui ne le sont pas sont le plus souvent des outils utilisés en interne par des sociétés internet (ex: Google).

Bases “NewSQL”

En réponse à la menace que représentent les technologies NoSQL, les éditeurs de bases relationnelles s’appliquent à présent à développer la capacité de leurs technologies à passer à l’échelle de manière horizontale, sans pour autant renoncer totalement au modèle relationnel.

Un exemple d’écosystème open source dynamique sur ce sujet est l’écosystème MySQL.

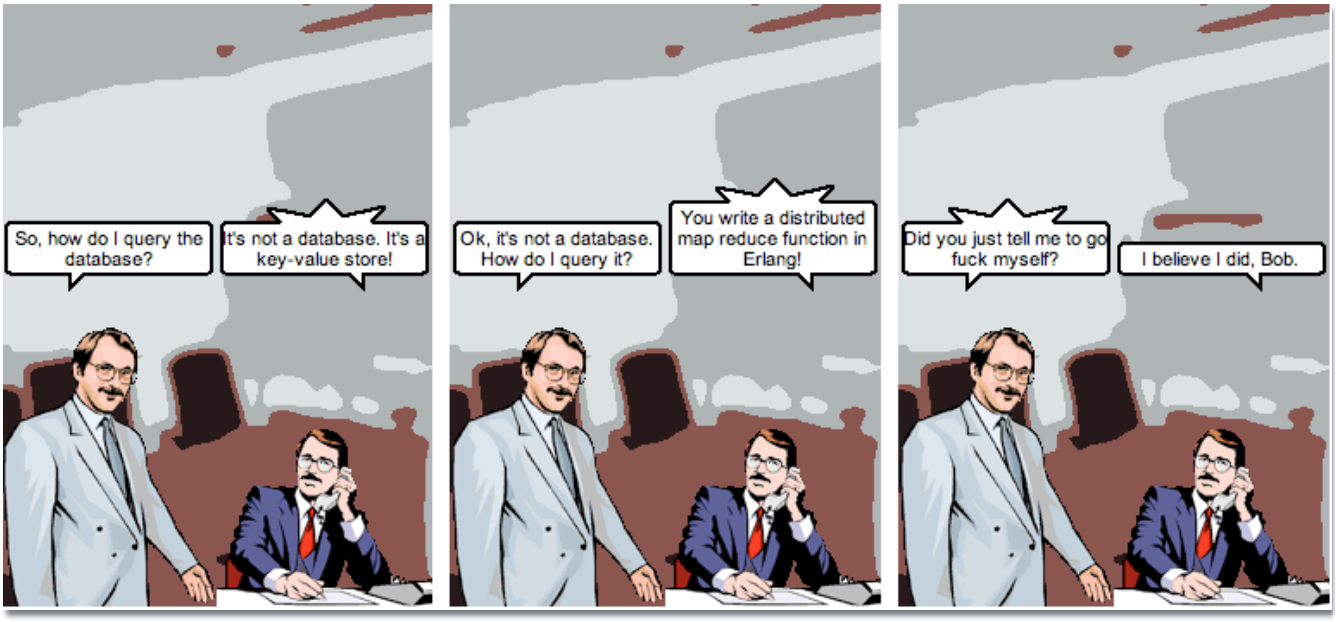
Le traitement et l'analyse

MapReduce

MapReduce est à l'origine une technique de programmation connue de longue date en programmation fonctionnelle, mais surtout un framework développé par Google en 2004[7].

Depuis cet article séminal, de nombreuses implémentations open source du principe général ont vu le jour: Hadoop (Yahoo! puis Fondation Apache), mais aussi Disco (Nokia), MrJob (Yelp!), etc.

Citons aussi les implémentations de MapReduce intégrées dans les bases de données NoSQL: CouchDB, MongoDB, Riak, etc.



Enfin, d'autres paradigmes de calcul massivement parallèles, plus orientés temps-réel ou aux calculs sur les graphes, ont commencé à être utilisés par les acteurs du Web (Google, Twitter, Yahoo!, etc.), et ont pour certains donné lieu à des projets Open Source: Pregel (technologie interne à Google, qui a à son tour inspiré les projets open source suivants: Apache Hama, GoldenOrb, Apache Giraph, Phoebus, Signal/Collect, HipG), S4, Storm, etc.

---

7 Jeffrey Dean et Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters. <http://research.google.com/archive/mapreduce.html>.

Indexation et recherche

A l'aube du Web, les moteurs de recherche (Inktomi, Infoseek, AltaVista, etc.) ont été les premiers à devoir développer des technologies innovantes pour indexer le Web de manière horizontale dans de multiples serveurs. Les mêmes enjeux se retrouvent dans les moteurs de recherche d'entreprises, qui doivent non seulement indexer efficacement (souvent, sans l'apport de l'algorithme *PageRank* de Google) les données des intranets et des applications métiers des entreprises, mais aussi à présent servir de base à des applications métiers basées sur la recherche, aussi appelées SBA (*Search Based Applications*).

De nombreux éditeurs propriétaires existent dans ce secteur (ex: en France, Exalead, Sinequa, Antidot), mais de plus en plus d'éditeurs se tournent vers moteurs d'indexation open source (principalement Apache Lucene / Solr) comme base de leurs produits (ex: en France, Open Search Server, Polyspot).

Machine learning et statistiques

L'enjeu majeur du Big Data n'est pas dans la collecte et le stockage, problème difficile mais principalement technique, mais dans la valorisation de ces données, qui touche à la technique mais principalement au business de chaque organisation.

Parmi les techniques utilisées pour extraire de la connaissance actionnable par le business à partir des données brutes, les techniques de *machine learning* tiennent une place de choix.

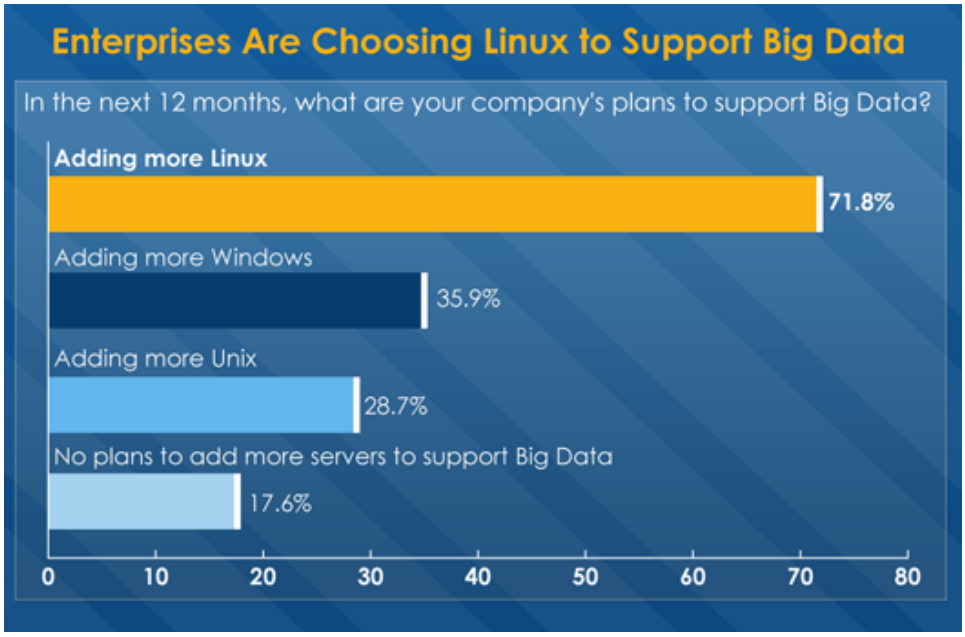
Parmi les projets open source aboutis dans le domaine, citons par exemple le projet Apache Mahout, boite à outils d'algorithmes de *machine learning* en Java, ou le projet Scikit-Learn, initié en France et dont une grande partie des développeurs sont français.

Dans le domaine des statistiques, la référence a été pendant des années le logiciel propriétaire SAS. Mais depuis quelques années, on constate la montée en puissance de projets open source comme R, soutenu à présent par la société commerciale Revolution Analytics (17.6M$ de levés), ou encore le projet Pandas.

Infrastructure

Lorsqu'une organisation gère plusieurs dizaines de milliers de serveurs, à plus forte raison plus d'un million comme Google, il est naturel qu'elle se tourne vers un système d'exploitation libre comme Linux afin d'une part d'avoir la plus grande maîtrise possible sur la pile logicielle qu'elle exploite, d'autre part ne pas dépenser exagérément en coûts de licences.

Ainsi, il n'est pas étonnant que 71.8% des grandes entreprises interrogées récemment envisagent de faire appel à des serveurs Linux pour faire face à leur besoins de Big Data, contre seulement 35.9% pour Windows.



De manière similaire, on voit se développer des besoins importants, et des projets matures, dans les domaines de la gestion de parc, du déploiement automatique, du monitoring, etc. Parmi les projets significatifs, citons: Chef, Puppet, Fabric, Zookeeper, etc.

Dans tous les cas, on retrouve dans le domaine de l'infrastructure des problématiques très semblables, et les mêmes outils, que dans le *cloud computing*.

Pour un développement du Big Data open source en Ile-de-France

Quelques acteurs industriels de l'écosystème francilien

Les **sociétés de service** (ex: OpenWide, Smile, Zenika...) ont commencé à proposer à leurs clients des prestations autour de technologies open source liées au Big Data :

- **Smile** est l'auteur d'un livre blanc sur les bases de données NoSQL (http://www.smile.fr/Livres-blancs/Culture-du-web/NoSQL).
- **Zenika** a travaillé dans deux catégories différentes de Big Data, en utilisant des technologies open source comme MongoDB, CouchDB, ElasticSearch, Apache Hadoop et Redis :
  - les données bancaires, de risque notamment, qui nécessitent des grilles de données distribuées (*data-grids*) du fait de la taille des données à manipuler et de la vitesse d'exécution requise ;
  - les données de type utilisateur, dans le secteur du "Web 2.0" c'est-à-dire des startups françaises financées par des VC qui nous ont confiées le traitement de leur passage à l'échelle (*scalability*) sur les données nominatives stockées.

Parmi les **éditeurs**, on peut noter :

- **Core-Techs**, éditeur et société de conseil dans la GED et l'e-commerce, s'y intéresse à plus d'un titre, notamment avec le projet de R&D collaboratif GEO+ qui est une plateforme de collecte et de représentation de données ouvertes (ou non) mêlant des dimensions BI, sémantique et cartographie. La société à également travaillé avec différents acteurs publics autour de deux enjeux :
  - ouvrir quelles données et avec quels moyens et pour quels bénéfices ?
  - quel format / normalisation / ontologies mettre en œuvre ?
- **DataPublica**, qui développe un catalogue et une plateforme d'accès aux données publiques françaises (Open Data), utilise des technologies NoSQL (MongoDB) pour faire face à la fluidité des données traitées et pour passer à l'échelle.

- **Nexedi**, éditeur d'un ERP et d'une plateforme de cloud distribués open source, a développé, dans le cadre du projet FEDER NEOPPOD, une base de données objets NoSQL transactionnelle distribuée.
- **Nuxeo**, éditeur d'une plateforme de GED open source, utilise des technologies d'indexation et de *machine learning* pour apporter de la valeur aux contenus manipulés par sa plateforme.[8]
- **Open Search Server** (Jaeksoft) développe une plateforme d'indexation et de recherche open source, avec un accent sur la scalabilité horizontale de sa solution.

Place du big data dans l'agenda de la recherche publique

L'ANR (Agence Nationale de la Recherche) indique dans son document de programmation 2012[9] ses attentes en termes de Big Data:

> *Cet axe thématique regroupe une classe de problèmes où le volume et la complexité des données manipulées et traitées constituent un verrou majeur. Ces données sont caractérisées par leur nature différente (temporelle, spatiale, hybride, etc.), leur forme (signaux, déstructurées, semi–structurées, etc.), leur représentation matérielle et logicielle, leur gestion à grande échelle (transport, stockage, volatilité, acuité, pérennité, etc.).*
>
> *Concernant la simulation, tous les aspects de la gestion des données impliquées dans les cycles de simulation sont concernés. Les données du processus de simulation doivent être modélisées, stockées, traitées et manipulées par des algorithmes robustes, performants, et adaptés aux supports répartis.*

Elle précise également les sous-thèmes qui l'intéressent:

> *Les sous–thèmes importants sont, de façon non exhaustive, le stockage, la gestion et le traitement de BigData, i.e. très grands volumes de données (web, smart grids, wireless sensor networks) avec notamment le stream computing (traitement en flux tendu des données) dans lequel le stockage classique est irréalisable voire non souhaitable (p.ex. caméras de vidéo–protection), les techniques innovantes de modélisation par les données,*

---

[8] Cf. <http://www.fiercecontentmanagement.com/story/big-data-and-smart-content-new-challenges-content-management-applications/2011-12-19>.

[9] <http://www.agence-nationale-recherche.fr/fileadmin/user_upload/documents/2011/Programmation-ANR-2012.pdf>

de pré et post traitement, de fouille des données, d'interprétation... provenant notamment de dispositifs ubiquitaires de collecte d'informations fixes et mobiles qui sont « enfouis » et « omniprésents » toujours en plus grand nombre dans le monde réel (assistants personnels, téléphones cellulaires, traceurs GPS, caméras de vidéo–protection, réseaux RFID, etc.).

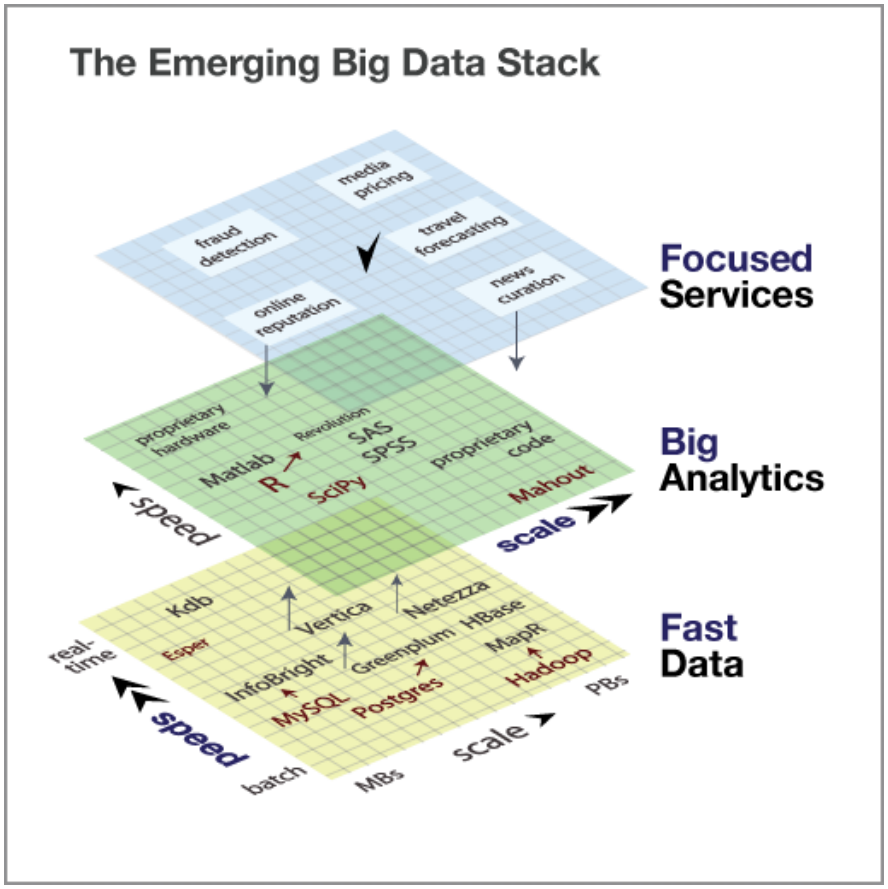
Concernant l'open source, sans exclure de financer le développement de logiciels propriétaires, l'ANR affiche clairement des arguments en faveur de l'open source:

[Le programme] s"intéresse à la production et la fourniture de logiciels propriétaires et libres (i.e., « open source »). Ces deux modes vont économiquement cohabiter dans le futur mais le logiciel libre a vocation à faciliter l"accès, la connaissance et l"utilisation à coût modéré de résultats de R&D accessibles directement par l'Internet et donc partout sur la planète.

Conclusion

Le logiciel libre est particulièrement actif dans le domaine du Big Data, avec plusieurs dizaines de projets de grande valeur, mais dont le centre de gravité se trouve en général aux USA, du fait de l'origine de ces projets, et plus généralement dans les pays qui ont des sites web de plusieurs dizaines de millions d'utilisateurs.

Néanmoins les compétences existent en Ile-de-France, que ce soit au sein de la recherche publique (ex: INRIA) ou des entreprises. L'émergence d'actions collectives, comme un ou des projets de R&D collaborative fédérateurs, ou des conférences sur le sujet, paraissent utiles pour catalyser ce potentiel et permettre à notre écosystème d'être présent sur ce marché à la fois stratégique et à très fort potentiel de croissance.



(Source: Michael Driscoll, "Building data startups: Fast, big, and focused", 9 août 2011 <http://radar.oreilly.com/2011/08/building-data-startups.html>).

Annexe: quelques projets open source

NB: liste probablement encore incomplète, n'hésitez pas à me contacter pour la compléter.

Bases NoSQL

Clé-valeur

- Tokyo Cabinet <http://fallabs.com/tokyocabinet/> et Tokyo Tyrant <http://fallabs.com/tokyotyrant/>: stockage clé-valeur local et distribué.
- Kyoto Cabinet <http://fallabs.com/kyotocabinet/> et Kyoto Tycoon <http://fallabs.com/kyototycoon/>: successeurs de Tokyo Cabinet et Tokyo Tyrant.
- Riak <http://wiki.basho.com/>: base clé-valeur répartie avec support de MapReduce.
- Voldemort <http://project-voldemort.com/>: stockage clé-valeur distribué développé par LinkedIn.
- Redis <http://redis.io/>: stockage clé-valeur de "structures de données" (listes, ensembles, dictionnaires), en mémoire, ultra-rapide.
- KumoFS <http://kumofs.sourceforge.net/>: stockage clé-valeur distribué développé par Nico Nico Douga, le "Youtube japonais".
- Memcached <http://memcached.org/>: base clé-valeur distribuée optimisée pour être utilisée comme cache de données provenant de systèmes de stockages plus traditionnelles (ex: SGBRD).

Orientées documents

- MongoDB <http://mongodb.org/>: base distribuée orientée documents développée par la start-up new yorkaise 10gen.
- Apache CouchDB <http://couchdb.apache.org/>: base distribuée orientée documents avec une API REST, développée en Erlang.

Orientées graphes

- Neo4j <http://neo4j.org/>: base de données orientée graphes développée en Java.
- Infinitegraph <http://www.infinitegraph.com>: base de données distribuée orientée graphes.
- OrientDB <http://www.orientechnologies.com/>: base de données orientée graphes et documents.

Clones de BigTable

- Apache Cassandra <http://cassandra.apache.org/>: clone de BigTable développé à l'origine par Facebook, en Java.
- Apache HBase <http://hbase.apache.org/>: clone de BigTable basé sur Hadoop, développé en Java.
- Hypertable <http://www.hypertable.com>: clone de BigTable développé par une startup dédiée en C++.

Systèmes de fichiers distribués et stockages de BLOBs

- Lustre <http://wiki.whamcloud.com/>: système de fichier distribué utilisé par plus de 70% des supercalculateurs actuels.
- GlusterFS <http://www.gluster.org/>: système de fichiers distribué horizontalement scalable, développé par Red Hat.
- Ceph <http://ceph.newdream.net/wiki/>: système de fichier distribué, qui propose également une interface de stockage d'objets compatible S3.
- OpenStack Open Storage "Swift" <http://openstack.org/projects/storage/>: stockage d'objets distribué développé dans le cadre du projet cloud OpenStack.
- dCache <http://www.dcache.org/>: système de fichier distribué développé par le CERN et des institutions similaires.
- HDFS <http://hadoop.apache.org/hdfs/>: système de fichiers distribué dédié au stockage de données pour le framework MapReduce Apache Hadoop.
- DDFS <http://discoproject.org/doc/howto/ddfs.html>: système de fichier distribué dédié au stockage de données pour le framework MapReduce Disco (cf. infra).

Bases NewSQL

- Cubrid <http://www.cubrid.org/>: base de données distribuée compatible avec MySQL, développée par le principal portail coréen.
- InfiniDB <http://infinidb.org/>: base de données analytique.
- VoltDB <http://voltdb.com/>: base de données en mémoire compatible MySQL.

MapReduce

- Apache Hadoop <http://hadoop.apache.org/>: implémentation de l'algorithme MapReduce développée à l'origine par Yahoo! en Java et placée ensuite sous l'égide de la Fondation Apache.
- Disco <http://discoproject.org/>: implémentation de MapReduce en Erlang et Python développée par Nokia.

Moteurs d'indexation et de recherche

- Apache Lucene <http://lucene.apache.org/core/>: bibliothèque Java pour l'indexation et la recherche.
- Apache Solr <http://lucene.apache.org/solr/>: plateforme d'*enterprise search* basée sur Lucene.
- Katta: <http://katta.sourceforge.net/>: plateforme d'*enterprise search* basée sur Lucene.
- ElasticSearch <http://www.elasticsearch.org/>: plateforme d'*enterprise search* basée sur Lucene.
- OpenSearchServer <http://www.open-search-server.com/>: plateforme d'*enterprise search* basée sur Lucene.

Statistiques

- R <http://www.r-project.org/>: langage dédié aux statistiques et à l'analyse de données. Projet mature, largement utilisé dans les milieux universitaires mais aussi dans la finance.
- Pandas <http://pandas.pydata.org/>: bibliothèque Python pour les statistiques.
- Rapid Miner <http://rapid-i.com>: outil d'analyse de données en Java.

Machine learning

- Apache Mahout <http://mahout.apache.org/>: bibliothèque Java de *machine learning* et de *data mining* qui utilise Hadoop. Développée activement dans le cadre d'un projet Apache.
- Scikit Learn <http://scikit-learn.org>: bibliothèque Python de *machine learning*. Développée activement par un groupe de chercheurs et de *data hackers*.
- WEKA <http://www.cs.waikato.ac.nz/ml/weka/>: bibliothèque Java de *machine learning*. Développée activement par un groupe de recherche à l'université de Waikato.

A propos / crédits

Auteur

Ce texte a été rédigé par Stefane Fermigier <http://fermigier.com/>.

Stefane Fermigier est un entrepreneur du logiciel libre. Il a fondé Nuxeo, pionnier de l'ECM open source, en 2000. Il a également été cofondateur de l'AFUL <http://www.aful.org/>, et du Groupe Thématique Logiciel Libre <http://www.gt-logiciel-libre.org/> (au sein du Pôle de Compétitivité Systematic-Paris-Region), qu'il préside depuis 3 ans.

Il travaille actuellement comme consultant dans les domaines de l'intelligence des contenus et des données, de l'entreprise 2.0, du développement collaboratif et des business models de l'open source.

Si vous avez des commentaires sur ce document, n'hésitez pas à me contacter: sf@fermigier.com.

Contributeurs

Ont contribué à ce document: Jean-Paul Smets (Nexedi), Patrice Bertrand (Smile), Marine Soroko (Core Techs), Patrick Moreau (INRIA), Patrick Bénichou (OpenWide), Pierre Queinnec (Zenika), Raphael Perez (Jaeksoft), Vincent Heuschling (Affini-Tech).

Photo de couverture: The Planet.

Illustrations: Neo4j, Nathan Hurst, Linux Foundation, @jrecursive, Michael Driscoll.

Licence