Insee
Mesurer pour comprendre

Big Data et Statistique Publique

Lettre d'information
N°9 – décembre 2019

Data Lake : une destination singulière

C'est là une question vieille comme l'informatique, mais elle prend une actualité nouvelle avec le big data : comment stocker les données ? Comment optimiser leur rangement afin de pouvoir ensuite y accéder facilement ? La réponse a varié au cours du temps. Au commencement était le fichier : un stockage facile, mais pas de garantie de maîtrise de la structure des données. Puis sont apparues les bases de données relationnelles, et leurs systèmes de gestion (Système de Gestion de Base de Données ou SGBD) : ils ont joué un rôle majeur pour qu'on puisse multiplier les mises à jour des données (avec accès concurrents) tout en respectant des contraintes d'intégrité. Les SGBD étaient parfaits pour la gestion, avec des données évoluant fréquemment, mais insuffisants pour l'analyse, l'aide à la décision, sur données figées. Les entrepôts de données ont été inventés pour cela : ils imposent en amont un gros travail de transformation des données (axes d'analyse, nomenclatures communes…). En aval, la navigation dans les hypercubes devient très simple. Sauf que cette étape de transformation (le « T » du ETL, Extract-transform-load) est très coûteuse… trop, en cas de données massives. D'où l'idée des Data Lakes : on peut « poser » les données telles quelles, et on les transforme quand on en a besoin. Idée séduisante. Mais contrairement aux bases et entrepôts, il n'existe pas de littérature académique de référence : ainsi, chaque éditeur peut vendre sa propre vision du Data Lake. De plus, le concept reste encore à éprouver : selon une récente étude du Gartner group, 80 % des projets de lacs de données se soldent par un échec. Les difficultés sont multiples. Il n'en reste pas moins que mettre à disposition des données massives est un enjeu pour la statistique publique. Le chemin sera certes semé d'embûches, mais essayons, expérimentons, apprenons : c'est incontestablement une voie d'avenir.

Pascal Rivière, chef de l'Inspection Générale

Qu'est-ce qu'un lac de données ?

Est-ce une destination paradisiaque, un voyage enrichissant ou juste un terme inventé par un commercial bien avisé ?

Une définition peu précise

L'essor des Data Lakes est lié au développement relativement récent des entreprises du numérique qui tirent profit des données collectées massivement. Les secteurs économiques concernés par ces challengers sont tellement nombreux que tout le monde se met à revoir la place des données dans son activité. Le Data Lake, comme le Big Data ou la datascience, devient un nouveau *buzzword*.

Le principe de Data Lake fut conceptualisé pour la première fois par James Dixon, CTO de Pentaho (société informatique implantée aux États-Unis appartenant au groupe Hitachi et spécialisée dans les logiciels de traitement et exploitation de la donnée), pour établir un parallèle avec le Data Mart, entrepôt de données structurées. Selon Dixon : « si vous considérez le Data Mart comme un magasin d'eau en bouteille, où l'eau est empaquetée pour faciliter sa consommation – nettoyée, emballée et structurée – le lac de données est une grande étendue d'eau dans un état plus naturel. […] Divers utilisateurs du lac peuvent venir l'examiner, y plonger ou y prélever des échantillons. »

Le terme Data Lake ne désigne donc pas un concept précis. Il permet simplement de supposer un certain nombre de principes un peu flous :

- la centralisation des données pour éviter les silos ;
- une mise à disposition non dirigée des données, de manière à favoriser les idées nouvelles d'utilisation en complément de la réponse aux besoins déjà identifiés ;
- une masse importante de données à faible densité d'information ;
- des données peu structurées, très proches de leur état initial avant traitement, avec une logique inversée entre apurement et stockage (le schéma est défini au moment de sa lecture plutôt qu'au moment de son écriture).

Le consommateur du lac sera donc plutôt un spécialiste de la donnée comme le statisticien, le datascientist ou le développeur plutôt que le client final.

Des risques

Les deux premiers principes de centralisation et de mise à disposition non dirigée rentrent en contradiction avec le « principe de minimisation des données » du RGPD et avec les principes de la CNIL concernant le croisement de données et la conservation de données, qui devraient être précisément adaptés au juste besoin, considérés et décrits préalablement à la mise en œuvre d'un traitement de données personnelles.

Un second risque est la transformation du Data Lake en Data Swamp, un marécage de données, dès lors que les données sont empilées les unes à côté des autres, sans véritable effort de description de leur origine, de leur qualité ou de leur finalité. Or la problématique des métadonnées et la politique d'accès sont fondamentales dans un tel écosystème. Ce sont les piliers de ce qu'on appelle la gouvernance des données, élément essentiel à la survie et au bon usage des Data Lakes.

Le lac des sigles : une connotation technique très marquée

Le terme Data Lake recouvre parfois un ensemble de technologies issues de l'écosystème Big Data sans en reprendre l'essence. En particulier, les Data Lakes sont souvent construits autour d'un cluster Hadoop (HDFS), un système de fichier distribué hautement élastique et opensource permettant la distribution des calculs sur un nombre arbitrairement grand de serveurs. Cette hégémonie est concurrencée par l'émergence du Cloud Computing dont certaines offres peuvent être considérées comme des Data Lakes as a Service (DLaaS).

Y a-t-il un Data Lake à l'Insee ?

Les Data Lakes peuvent aussi bien être à usage interne ou au contraire utilisés pour la diffusion. Faisons un tour de ce qui pourrait s'en rapprocher ou non à l'Insee.

L'entrepôt de données locales (EDL) : l'archétype du DataMart

Le Data Lake est souvent comparé ou opposé au *DataMart* comme l'est l'EDL. Il s'agit d'un entrepôt central mais dont les données sont denses en informations, structurées au moment de leur conception. Sans en renier la pertinence, un Data Lake est plutôt censé compléter ce type de stockage.

Données de caisse : un cluster BigData pur jus

L'infrastructure construite pour traiter les données de caisse des opérateurs de la grande distribution est d'un point de vue technologique ce qui s'apparente le plus à un Data Lake, sans pour

autant en avoir l'essence. En effet, il s'agit d'un cluster Hadoop sur lequel l'Insee fait du calcul distribué mais les données ne concernent qu'un domaine particulier et l'accès au cluster est limité du fait de la sensibilité des données.

AUS : l'essence d'un Data Lake

AUS est sans doute le service Insee le plus proche de l'essence du Data Lake puisqu'il s'agit d'une plateforme de stockage et de traitement de données centralisée de sources multiples ; même si les données sont plutôt structurées sous la forme de tables, la finalité statistique laisse un certain niveau d'autonomie. Au regard de cette offre, un Data Lake peut se voir comme une opportunité technique pour diversifier l'offre de stockage et de traitement des données en offrant des capacités nouvelles en particulier dans le domaine des données massives.

Le datalab de la plateforme innovation : un Data Swamp ?

La plateforme innovation a mis en œuvre des technologies proches des Data Lakes émergents des acteurs du Cloud Computing. Mais en l'absence de gouvernance des données (métadonnées, politique d'accès, règles d'usage, etc.) qui nécessiterait des moyens dédiés, son datalab s'apparente plus à un Data Swamp. Les capacités de stockage associées à un large choix de technologies de représentation et de traitement de données en font néanmoins un laboratoire de données intéressant.

Pour conclure ce voyage, il n'existe pas vraiment de Data Lake à proprement parler à l'Insee, des ambitions proches ont certes pu sous-tendre AUS, qui arrivait avant la mode et le *buzzword*. Une cure de jouvence pourrait apporter les dispositifs nécessaires à davantage de partage, à condition de l'accompagner d'un travail de gouvernance fort pour maîtriser les usages et faciliter l'explorabilité.

L'expérience du SDES : un Data Lake pour la diffusion

Le service statistique de l'écologie (SDES) diffuse une information riche, multi thématique et de qualité : 4 000 indicateurs statistiques sur EIDER, 600 indicateurs à la commune sur GéoIDD (outil de cartographie dynamique), des milliers de fichiers Excel sur son site web et une vingtaine de cubes sur Beyond.

Cependant, depuis plusieurs années, cette mise à disposition importante des données connaissait des faiblesses : des processus d'alimentation différents et chronophages, des incohérences entre les données publiées sur des outils différents, des données sous format Excel non exploitables pour des réutilisations… Cet écosystème d'outils et de procédures, résultat de la fusion en 2008 de trois services statistiques en un unique, s'avérait être un véritable labyrinthe pour les internautes et échouait *in fine* à proposer une offre de données publiques simples, cohérentes et valorisables.

En 2016, le SDES a lancé un projet pour revoir dans son ensemble son offre de diffusion. Une consultation des internautes a permis de définir les nouvelles bases de la diffusion : un site plus épuré et un lac de données pour explorer les sources de données produites ou utilisées dans les publications.

Au moment du lancement du projet, la construction du lac de données était avant tout un défi technique : définition des orientations, priorisation des fonctionnalités, choix des composants techniques, tests et recette, préparation de la mise en exploitation, installation, etc.

Au fil des mois, il s'avère que le défi le plus complexe n'est pas technique, mais humain et organisationnel. Car, le lac de données implique des changements importants dans l'organisation du SDES et dans les méthodes de travail. Qui doit documenter ? Qui doit préparer le fichier de données ? Est-ce bien utile de mettre à disposition cette donnée ? Qui doit être au contact des utilisateurs ? Pour lever ces difficultés, la mise en place du Data Lake au SDES s'est appuyée sur un modèle de conduite du changement socio-dynamique, qui consiste à repérer et valoriser au maximum des « alliés » du projet. Par une dynamique de ruissellement, ces alliés au projet – deux à trois agents dans chacune des sous-directions – ont pour mission d'encourager les hésitants et les passifs, et ainsi ne laisser personne sur le côté du chemin. Aujourd'hui, malgré le départ de la quasi-totalité de l'équipe initiale, le projet est dans sa dernière ligne droite : la mise en production du Data Lake doit intervenir dans les prochaines semaines. Encore un peu de patience…

Comment éviter le drame du lac ?

Le point de vue de l'administrateur ministériel des données, Stéphane Trainel : « Il ne fait aucun doute aujourd'hui que la gouvernance de la donnée qui englobe l'identification, le pilotage, le classement et le partage des données grâce à des processus, des acteurs et des instances clairement définis dans une entité publique ou privée, impose la création de lacs de données. C'est en fait la première étape nécessaire (mais non suffisante) pour établir un nouveau cadre d'usage des données détenues, et *in fine* proposer des outils d'aides à la décision qui associent les données de l'entité et les « nouveaux outils » de la science des données (traitements des données massives, machine learning, deep learning, etc.).

Pour réussir un lac de données, à l'instar des projets de transformation, il faut s'assurer d'une part que l'organisation du projet est viable. L'équipe projet intègre les métiers, les utilisateurs finaux et dispose des compétences techniques et humaines pour mener le projet à son terme. Le mandat est clairement établi et le sponsor s'implique dans les phases cruciales du projet. L'accompagnement des métiers et l'évolution des processus sont également traités dès le début du projet.

D'autre part, la gouvernance du lac de données doit être claire et partagée avec tous les acteurs (producteurs et utilisateurs) pour éviter le syndrome du marécage. Pour répondre aux différents besoins du métier, le lac doit s'organiser en plusieurs zones : un stockage temporaire, un stockage des données brutes, un bac à sable exploratoire et un stockage des données enrichies prêtes à l'emploi. Sans oublier le lignage (le suivi du cycle de vie) et la sécurité des données (droits d'en connaître, anonymisation, pseudonymisation, agrégation) qui sont des exigences fortes pour le bon fonctionnement du lac. L'administration du lac nécessite donc des moyens humains conséquents. Pour terminer, le lac de données n'est probablement pas un projet comme les autres. Il est profondément transformateur. Sa réussite dépendra de sa capacité à répondre aux enjeux et à la stratégie de l'entité qu'il sert. »

Quelques références

Sur Data Lakes : "How to Avoid Data Lake Failures", 2018 chez Gartner.

Sur les bases de données relationnelles : E. F. Codd. 1970. A relational model of data for large shared data banks. Commun. ACM 13, 6 (June 1970), 377-387

Sur les technologies d'entrepôts de données : Surajit Chaudhuri and Umeshwar Dayal. 1997. An overview of data warehousing and OLAP technology. SIGMOD Rec. 26, 1 (March 1997), 65-74.

Actualités / Brèves

Save the date ! Le prochain séminaire Big data « Nouvelles approches pour coder dans une nomenclature : machine learning et autocomplétion » se tiendra mardi **14 janvier 2020**, de 14h00 à 17h00 en Salle Closon Malinvaud.

Consulter la page « Parlons Innovation statistique » pour en savoir plus sur le 62e congrès mondial de statistiques de l'International Statistical Institute (ISI) 2019

Merci à Frédéric Comte, Juliette Fourcot, Pascal Rivière et Stéphane Trainel pour leur contribution.

Cette lettre est une occasion d'informer largement et d'échanger.
N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à mathilde.poulhes@insee.fr**.**

Les archives de cette lettre sont disponibles sur :
https://www.agora.insee.fr/cms/sites/dmcsi/home/ssp-lab/BigData.html
Demande d'inscription individuelle à la lettre : dg75-l001@insee.fr