Alain Clapaud / jeudi 5 avril 2018 / Thèmes: Big Data, Dossier, RGPD

Comment concilier Big Data et RGPD ?

Alors que le jour officiel de l'entrée en vigueur du règlement européen de la protection des données (RGPD) approche, le Big Data est au cœur des préoccupations de DSI et des DPO. Protections renforcées et nouvelles procédures sont en plein déploiement pour être prêt à la date du 25 mai 2018.

La mise en conformité RGPD d'une infrastructure Big Data passe nécessairement par une phase de découverte des données afin de localiser tous les référentiels qui contiennent des données personnelles.

Par leur nature même, les grands Data Lake mis en place par les entreprises sont les premiers à être touchés par l'entrée en application prochaine du règlement européen qui vise à renforcer la protection des données personnelles des citoyens de l'Union. Si les solutions qui arborent le logo « RGPD Ready » tiennent plus du marketing que du réel développement pour le nouveau règlement européen, un certain nombre de briques peuvent être déployées dans le système d'information pour aller vers la conformité.

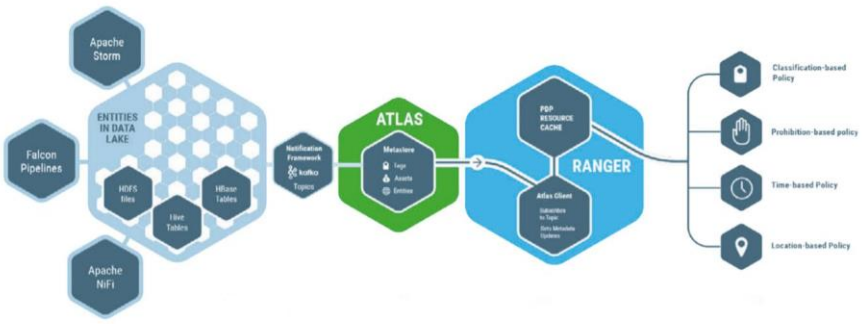
Première phase : partir à la découverte des données

Les DPO vont presser leurs DSI et les administrateurs de Data Lake de rapidement mettre en place des procédures pour faire face aux demandes d'accès aux données personnelles, au droit d'effacement de ces données ou encore de portabilité des données. Préalable indispensable à cette mise en conformité du Data Lake, il est nécessaire de répertorier où sont réellement stockées les données personnelles. Il faut le faire dans l'infrastructure Big Data, mais aussi dans tout le SI, notamment pour débusquer tous les exports de base de données réalisés pour les besoins du marketing, pour la RH, pour les commerciaux, des fichiers Excel, des bases Microsoft Access ou même des exports vers des outils analytiques type Qlikview ou Tableau qui échappent parfois au contrôle étroit de la DSI. Porté par le RGPD, le marché des outils de Data Discovery, littéralement découverte des données, se porte bien et s'il existe de nombreuses solutions pour inspecter un SI, il en existe certaines dédiées au Big Data. Outre les solutions des gros éditeurs comme SAS Institute, Oracle, de multiples éditeurs proposent des solutions capables de parcourir les données des grands Data Lake : Imperva, Prifender, Privacera. Alivia Smith, responsable marketing de l'éditeur Dataiku souligne : « *La première étape est de rechercher où l'on stocke de la donnée personnelle puis de documenter tous nos traitements de données en interne. Dans notre cas, c'était une tâche relativement simple car toutes nos données sont centralisées dans un même outil.* »

Même constat pour Abhas Ricky, à la tête de la stratégie d'Hortonworks, éditeur de l'une des distributions Hadoop les plus populaires dans les entreprises, notamment chez les Telcos, le secteur de la banque/ assurance en première ligne sur le RGPD : « *Beaucoup d'entreprises ne connaissent pas précisément l'ensemble des datasets où ils ont des données personnelles car ils collectent les données de manières très diverses, Il faut être capable d'identifier où se trouve cette donnée personnelle, la donnée sensible. La donnée peut être auto-classifiée via un jeu de règles qui vont permettre de taguer toutes les informations entrantes, via Atlas, un outil qui fait partie du stack Hortonworks, mais il est possible d'utiliser d'autres outils plus spécialisés sur notre stack open source.* »

Le casse-tête de l'anonymisation des données

Dès lors que l'on dispose d'une cartographie à jour du Data Lake, le bon sens veut que pour limiter les risques de fuite et d'usages non souhaités de la donnée personnelle, il faille anonymiser au maximum les données et les faire ainsi sortir du champ d'application du RGPD. Outre des algorithmes open source que les administrateurs peuvent exécuter sur leurs données, il existe de multiples solutions logicielles pour anonymiser ou masquer les données stockées ou à la volée comme ce que réalise la solution DataRespect du Bordelais Magush : « *Il s'agit d'un proxy d'anonymisation qui filtre toutes les données qui entrent ou qui sortent du Data Lake* », explique Philippe Michel, directeur général de Magush. « *L'idée, c'est que pour être en conformité avec le RGPD, on ne stocke que des données non identifiables sur le serveur d'entreprise. Il s'agit d'une anonymisation temps réel sans impact sur la base de données elle-même.* » Attention !, croire qu'il suffit de remplacer les noms et le prénom par des « * » ou réaliser un hashage de ces champs pour ne pas tomber sous le coup du règlement européen serait une lourde erreur. L'avis du G29 – le groupement de toutes les « Cnil » européennes – est très différent car il est assez simple de remonter à un individu en analysant et en faisant de la corrélation de données. Ainsi, l'adresse IP est considérée comme personnelle par la Cnil. De même qu'avec le genre, l'adresse d'une personne et un diplôme par exemple, il est facile de retrouver son nom, mais plus on efface des données, moins les analyses seront pertinentes. L'article 26 du règlement européen est très clair sur ce point : « *Les données à caractère personnel qui ont fait l'objet d'une pseudonymisation et qui pourraient être attribuées à une personne physique par le recours à des informations supplémentaires devraient être considérées comme des informations concernant une personne physique identifiable.* » L'anonymisation demande donc une analyse des données soigneuse et décider de quelles données il faudra se passer et celles qui seront réellement utiles aux Data Scientists. À titre d'exemple, le CASD qui cumule nombre de données sensibles sur le français à souhaité conserver absolument toutes les données, mais il le fait au prix d'un niveau de sécurité ultra-élevé car la moindre fuite de données serait catastrophique.



Les distributions Hadoop avancées intègrent des dispositifs sophistiqués de gestion d'accès aux données et de traçabilité des opérations réalisées sur les données. Ici, l'architecture Hortonworks qui met en œuvre le référentiel Atlas et l'outil de gestion d'accès Ranger.

La cybersécurité du Data Lake doit s'élever

utre technique souvent évoquée pour protéger les données personnelles, le chiffrement. La technique est bien connue, ses inconvénients aussi, notamment du fait de la charge de calcul imposée par le chiffrement et le déchiffrement des données. En Big Data, s'il est bien évidemment possible de chiffrer intégralement un Data Lake, en pratique le temps de traitement des algorithmes inspire les administrateurs à plus de modération : « *Plusieurs niveaux de chiffrement sont réalisables en ce qui concerne un Data Lake* », explique Thibault Storai, expert Big Data chez Teradata. « *Un chiffrement au niveau des disques durs n'est pas pénalisant en termes de performances, mais ne permet que de se prémunir du vol d'un disque dur dans le data center. Au niveau de la couche logicielle, Hortonworks et Cloudera supportent le chiffrement des données, néanmoins ce chiffrement demande beaucoup de ressources et il faut absolument le limiter aux données sensibles, comme les numéros de carte bancaire par exemple.* »

Outre l'anonymisation, le chiffrement, l'accès aux données est un point crucial dans la sécurisation d'un Data Lake. L'humain reste le maillon faible de la cybersécurité d'un SI et c'est tout particulièrement le cas pour les comptes ayant accès au Data Lake. De l'avis général des experts, la sécurisation façon Unix d'Apache Hadoop ne suffit pas à une mise en conformité. Il faut désormais appuyer les accès aux Data Lake sur un système de gestion des droits qui permet une plus grande granularité et qui est surtout capable de tracer absolument toutes les manipulations réalisées par chaque Data Scientist, chaque Data Engineer ou chaque administrateur. « *Les éditeurs de distributions commerciales d'Hadoop ont étendu la plateforme Big Data et lui ont donné des capacités qui permettent de répondre à 100 % au RGPD* », estime Thibault Storai. « *Nous travaillons avec les deux partenaires les plus présents sur le marché, Cloudera et Hortonworks, et leurs distributions donnent la*

capacité d'identifier et de garantir l'authentification d'un utilisateur. Leur gestion des droits d'accès aux données permet d'avoir un bon niveau de finesse et on dispose d'un reporting complet des accès. »

Jérémy Greze, Data Analyst chez Dataiku ajoute : « *Il faut absolument cloisonner le Data Lake afin qu'il réponde aux besoins de chaque équipe, de chaque métier. Sur notre plateforme, nous traçons l'ensemble des actions de chaque utilisateur et non pas celles réalisés par un profil. En cas de fuite de données, c'est préférable pour retracer ce qui s'est passé.* »

Le comportemental garde un œil sur les Data Scientists

Pouvoir définir finement les droits de chacun et tracer les actions est rendu nécessaire par le RGPD mais cela restera inefficace si un pirate se connecte avec les login/password valides d'un administrateur. De nombreux logiciels de sécurité sont dédiés à la gestion des comptes à privilèges. Wallix, Bomgar, Balabit se sont spécialisés dans ce type d'outils, notamment en couplant ces accès à un système d'authentification forte pour les comptes les plus critiques. Autre acteur présent sur ce type de solution IBM qui évoque une tendance force dans la surveillance de ces comptes, l'utilisation du Machine Learning pour réaliser un contrôle comportemental de ces comptes. « *Notre solution Security Guardium protège aussi bien les bases de données relationnelles que les environnements plus typés Big Data ou les bases de données mainframe* », résume David Batut, directeur commercial chez IBM Security. « *Elle va générer des alertes en cas d'accès suspect ou même de blocage de certains accès, réaliser un " Dynamic Data Masking " afin de cacher certaines données à certains profils d'utilisateurs.* » Ces solutions, souvent mises en place pour cadrer le comportement des DBA des bases relationnelles les plus sensibles, sont désormais déployées sur les Data Lake devenus ultrasensibles vis-à-vis du RGPD. Le Data Scientist ou l'administrateur qui, soudainement, fait des exports de données sur un disque local ou une clef USB va éveiller l'attention du moteur comportemental et déclencher une contre-mesure et remonter cet incident vers le SOC où les analystes en cybersécurité vont pouvoir enquêter.

Beaucoup reste encore à faire pour les entreprises afin de mettre leur Data Lake en conformité avec le RGPD d'ici au 25 mai 2018, mais attention, il ne s'agit pas d'un projet ponctuel. Il va falloir auditer régulièrement le SI de l'entreprise et tout particulièrement son Data Lake afin de rester en conformité alors que l'on commence déjà à évoquer le droit futur de l'internaute à demander des explications sur la décision d'un algorithme, mais c'est une autre histoire...

Les dispositifs hardware d'authentification forte ou les boîtiers de sécurité tels que la SD-Box du CASD sont un moyen d'élever le niveau de sécurité d'un Data Lake.

« Qui dit grand volume de données, dit grande responsabilité ! »
Mathias Lemaire, expert sécurité, membre du pôle OcSSImore digital & sécurité

« Il n'y a pas d'incompatibilité entre Big Data et RGPD à partir du moment où l'on respecte quelques grands principes qui sont directement issus de la Loi informatique et liberté et de la LCEN (Loi pour la confiance dans l'économie numérique). Si une entreprise estime qu'elle a besoin de détenir de l'information personnelle en grand nombre, il n'y a pas antinomie avec le RGPD, pourvu qu'elle ait les accords éclairés des personnes et qu'elle protège la donnée. Qui dit grands volumes de données, dit grandes responsabilités ! » « L'une des principales difficultés est liée aux usages en entreprises où les ressources humaines font des extractions de données pour tel ou tel usage, ce sont des copies de fichiers client dans tel ou tel service, etc. Il faut mettre en place des outils qui vont chercher dans le SI les endroits où ces données personnelles ont été dupliquées, ainsi que là où elles ne sont pas protégées à l'état de l'art attendu. L'humain est un élément clé de la cybersécurité. Il faut l'aider à agir de manière plus responsable, notamment lui donner les moyens de travailler sans devoir dupliquer la donnée. »

« Il faut créer un bunker autour de la donnée »
Kamel Gadouche, directeur du Centre d'accès sécurisé aux données (CASD)

« *La vocation du CASD est de faciliter l'accès aux données pour les chercheurs. Nous hébergeons des données issues des recensements et des enquêtes de l'Insee, les déclarations sociales des entreprises, les déclarations d'impôts, les données de santé liées aux séjours hospitaliers, etc. Le chercheur, sous réserve évidemment qu'il ait obtenu les autorisations nécessaires – producteur, comité du secret, Cnil, etc. – dispose ainsi d'un excellent environnement de travail. Nous avons développé une technologie spécifique qui repose sur le couple : confinement des données dans une bulle sécurisée et authentification forte des utilisateurs via un boîtier biométrique, la SD-Box. Ce dispositif ultra-sécurisé est gage de confiance entre producteurs de données et chercheurs ou Data Scientists, qui disposent de tous les moyens pour mener leurs analyses. J'estime que cette approche est largement transposable dans les entreprises. Nous avons déjà eu des demandes de la part d'entreprises telles que Generali, BNP Paribas ou encore RTE viennent exploiter nos infrastructures.* »