INTERVIEW D'EXPERT

KAMEL GADOUCHE

DIRECTEUR - CASD

Kamel Gadouche a suivi ses études à l'ENSAE (division CGSA), avant de travailler à l'Insee en tant que statisticien puis en tant que chef de projet réseaux et sécurité. Il a ensuite rejoint la direction informatique du Groupe des écoles nationales d'économie et statistique (Genes). Il a alors pris en charge un projet IT pour répondre au besoin d'ouvrir l'accès aux données de l'INSEE de façon plus large aux chercheurs, mais aussi celles de certains ministères ainsi que les données d'acteurs, notamment dans les domaines de la banque, l'assurance et l'énergie (avec RTE essentiellement).
Ce **projet d'ouverture et de partage des données à des fins de recherche, baptisé CASD, est né en 2010**. Parti d'une forte volonté de permettre l'accès aux données, le projet concerne de nombreuses données confidentielles, qui dès le départ ont appelé des conditions de sécurité élevées.

Kamel Gadouche a dû considérer les deux parties prenantes avec une approche différente : **les producteurs de données devaient être convaincus que les conditions d'accès étaient suffisamment sécurisées** pour accepter de les partager, tandis que les chercheurs qui en demandaient l'accès devaient être convaincus de l'ergonomie du dispositif. Il s'agit donc d'une part de rassurer le détenteur de la donnée et de lui fournir des garanties, et d'autre part de procurer aux chercheurs des outils simples et efficaces pour pouvoir les utiliser en toute sécurité.

Le cas s'est par exemple présenté avec les données de la compagnie d'assurance GENERALI, qui avait accepté de laisser certaines de ses données être étudiées par des chercheurs externes, mais nécessitait un cadre très sécurisé. Consciente de l'impossibilité de copier les données directement vers ses serveurs, l'équipe de recherche a alors proposé d'utiliser la structure du CASD, comme un intermédiaire de confiance. **La Banque Postale a rapidement adopté la même démarche**.
Concrètement, le CASD est une infrastructure informatique où est consignée la donnée, adossée à des outils de traitement type SAS, R, Matlab, Hadoop, Spark... Les utilisateurs y accèdent à distance, grâce à un boitier conçu par le CASD. Il s'agit là de terminaux lourds sécurisés, permettant de se connecter à l'infrastructure du CASD de façon sécurisée mais empêchant toute possibilité de téléchargement dans une structure externe. Les résultats, tableaux et analyses produits peuvent être récupérés selon une procédure particulière qui valide manuellement le téléchargement et le droit de publication. Ce sont les producteurs de données qui définissent les règles de ce qui sera publiable ou pas- en fonction de leurs obligations de confidentialité- et précisent les formes sous lesquelles les données peuvent être extraites. **La CNIL est également consultée lorsqu'il s'agit de données à caractère personnel**. Le processus d'anonymisation ne se limite d'ailleurs pas aux données personnelles, elle peut également concerner des données d'entreprises qui pour des raisons de confidentialité ne doivent pas pouvoir être identifiée.
A ce jour, les utilisateurs du CASD ont des profils variés : chercheurs, consultants, chargé d'étude, datascientist, médecins travaillant sur des données de santé, biostatisticiens... La diversité des secteurs et profils intéressés par la méthode est révélatrice des préoccupations en matière de sécurité des données.
Si au départ la technologie ne permettait pas d'effectuer des traitements Big Data, elle a intégré depuis 2013 des outils comme Hadoop ou Spark. **Le CASD peut alors s'apparenter à un datalab qui offre un accès sécurisé à des données**, pour une multitude de techniques de traitement. Le passage au Big Data complique souvent les aspects techniques d'accès aux données, d'où la pertinence d'une plateforme qui applique des conditions d'accès précises à chacun des partenaires. Un volume de données très important augmente les risques d'identification dans le cas d'un besoin d'anonymisation. Les traitements et les règles d'accès doivent être adaptés en conséquence, en gardant à l'esprit que **l'objectif est de donner l'accès à des données les plus brutes possibles, pour des études pointues et éclairées**.

Le boitier s'utilise avec une carte à puce biométrique dont les accès sont très fins, personnalisée pour chaque utilisateur. Son installation doit remplir des conditions strictes de sécurité et une infrastructure exigeante, engageant contractuellement chaque partie prenante. Un système de « bulle » crée une isolation totale du boiter et de son utilisateur, fonctionnant en circuit fermé, sans contact avec l'extérieur à partir du moment où l'utilisateur est entré sur la plateforme.

Kamel Gadouche nous détaille quelques projets avec lesquels le CASD a travaillé. L'un d'entre eux a été mené avec RTE, partant d'une problématique liée à des données issues de capteurs. **La question était de savoir comment valoriser les données des capteurs installés par RTE**. Topographies, géographie, ou météo par exemple devaient être pris en considération, dans un objectif de maintenance préventive et d'optimisation de l'allocation d'énergie. La complexité du projet a demandé l'aide de consultants externes, de startups et de chercheurs internes et externes, qui devaient chacun avoir un accès défini et contrôlé à chacune des données.

Un autre projet, mené avec l'Insee, a permis de traiter de gros volume de données dans le cadre de l'étude de l'indice des prix. Alors que jusqu'à présent, des enquêteurs devaient eux-mêmes relever les prix manuellement, une opération lente et couteuse, l'Insee peut désormais en grande partie récupérer les données de caisse des grandes enseignes françaises et établir directement un indice des prix.
Le domaine de la santé, où les données sont particulièrement sensibles, s'est lui aussi intéressé au dispositif proposé par le CASD. **La loi de 1978 qui régit les « donnée sensibles » et l'accompagnement apporté par la CNIL ont permis au CASD de vérifier que sa solution était adéquate au traitement de données de santé**.
Plusieurs cas de figure définissent le traitement des données de santé. Si l'objectif est de faire de l'open data, et donc d'ouvrir très largement les données, une anonymisation totale est alors indispensable. Lorsque les analyses nécessitent l'utilisation de données beaucoup plus précises, directement ou indirectement identifiantes, une infrastructure comme celle proposée par le CASD est alors indispensable.
Une cohorte gérée par l'INSERM – du même type que Constance, qui travaille sur les épidémiologies- a mené des études sur une population de 200 000 individus, avec un suivi dans le temps, pour observer les évolutions et déterminer les causes de certaines maladies. Basée sur le volontariat, cette étude a du apporter à ses participants d'importantes garanties de protection de leurs données, que la structure du CASD a su lui donner.

A l'avenir, **le CASD souhaite davantage faciliter la rencontre entre les producteurs des données, les startups et les consultants**. Une approche à l'image d'un de leur projet actuel mené avec ERDF, un concours d'innovation clôturé le 1er juillet 2015, où les données ont été mises à disposition de quatre startups, dans le but d'imaginer les nouveaux usages qui pourraient en découler.
A l'international, le CASD travaille déjà avec les instances européennes pour créer une infrastructure commune sécurisée. Le CASD représente la brique technologique du projet, avec un avantage non négligeable d'apporter une solution dont le coût à l'usage est faible. Il est important de noter ce point, car il a fait débat aux débuts de la création du CASD. Entre faible investissement de départ, et rentabilité sur le long terme, la question s'était posée. Originellement pensé sous forme d'un logiciel, le système était moins couteux à l'achat que le boitier proposé aujourd'hui. A l'usage, les coûts indirects, et notamment ceux de maintenance, s'ajoutaient à vitesse grand V. Sans contrôle des paramètres de l'utilisateur, de son outil de travail et ses potentiels virus, ou des accès donnés à des personnes tierces, garantir le niveau de sécurité promis était fastidieux et couteux.
Avec aujourd'hui 350 boitiers installés, plus de 1000 utilisateurs, le CASD parvient à se contenter de 4 techniciens, qui n'ont besoin d'intervenir que très rarement sur site.
Au-delà des considérations économiques, une notion d'ergonomie a aussi pesé dans le choix de la solution. Le CASD était soucieux de proposer un outil à l'usage simple et intuitif, malgré des règles strictes, qui auraient pu sembler rigides mais dont la mise en place a été simplifiée.
Les enjeux sont assez importants pour que les utilisateurs acceptent une dose de contrainte nécessaire, si l'outil les guide vers un usage efficace.