Groupe de travail Transformation Numérique
19 octobre 2020

Fiche n° 3

L'enjeu des données comme levier de transformation
Le projet de lac de données de la DGFiP

La loi pour une République numérique du 16 octobre 2016 a créé l'obligation pour les organisations publiques de publier et d'échanger leurs bases de données, sous réserve notamment d'anonymisation quand il s'agit de données personnelles, de protection de la propriété intellectuelle, ou du secret industriel et commercial. Ces données doivent ainsi pouvoir être exploitées et réutilisées facilement notamment par les particuliers, les entreprises et les acteurs du secteur public.

Le projet lac des données s'inscrit dans cette orientation gouvernementale d'ouverture, de réutilisation et de valorisation des données publiques, auquel vient en complément le projet API management.

En capitalisant sur les acquis du socle technique développé par le SSI, le lac des données permet de répondre à un besoin triple :

- **décloisonner** les applications de gestion et les infocentres qui utilisent les données issues des applications de gestion,
- contribuer à la définition d'une **gouvernance** des données qui permettrait par la suite d'améliorer la qualité de la donnée et ainsi une meilleure valorisation des données de la DGFiP,
- proposer de nouvelles plateformes et de nouveaux services adaptés aux besoins du **Big Data**, c'est-à-dire du traitement d'une très grande volumétrie de données.

En d'autres termes, le lac des données qui est une des composantes du dossier lauréat du FTAP 2019-2, a pour vocation d'amener au fur et à mesure, et en fonction des projets pilotes, une bonne partie des données disponibles de la DGFiP dans une infrastructure dédiée aux traitements des données en grand volume. Le lac des données constitue donc un outil de traitement des données massives, en ce sens ce n'est pas un outil de gestion. Les premiers bénéfices sont assez immédiats :

- temps de traitement plus rapide (de l'ordre de 10 à 20 fois plus rapide qu'une infrastructure classique type Hadoop sur une volumétrie de quelques centaines de millions de cellules),
- des gains en coûts liés à la licence (de l'ordre de 50 à 100 000€ par infocentre),
- des gains en coûts de maintenance (entre 30 % et 50%),
- des gains en temps de développement (plus de 40 %).

Cette première typologie d'utilisation du lac des données, c'est-à-dire **un entrepôt des données avec une infrastructure adaptée aux calculs en grand volume** s'accompagne de services et d'outils supplémentaires qui permettraient à des utilisateurs de mieux valoriser les données disponibles dans le lac des données.

L'un des premiers outils, et le plus important, est **le dictionnaire des données**. Ce dernier constitue un outil d'administration et de consultation des résultats de cartographie, de modélisation et du lignage des données. La navigation du dictionnaire permettra de répondre à des questions relatives à la localisation, au sens (fonctionnel tout comme métier) et à la qualité des données. C'est également l'outil indispensable pour mettre en place une réelle gouvernance des données. Cet outil participe activement à la stratégie Open Data de la DGFiP en offrant une vue concise des données de la DGFiP mais aussi via une meilleure connaissance de nos propres données et donc de savoir lesquelles sont de qualité suffisante en vue d'une ouverture de ces données.

D'autres services et outils sont à prévoir autour du lac des données, comme des **outils dédiés à la datascience** (pour la supervision des modèles ou encore du self-service analytics pour donner les outils aux métiers pour la valorisation des données), des **outils d'intelligence économique** comme des **outils de visualisation** (dont la veille est lancée avec en objectif de choix du logiciel en fin d'année 2020) ou encore des **outils de requêtage** comme les bases orientées colonnes (Inscrit au Plan Annuel d'Activité 2021).

Le lac des données est également un lieu où s'exprime la volonté de la DGFiP de soutenir l'écosystème Open Source en proposant de développer, en partenariat avec d'autres administrations et grands comptes, une version Open Source de composantes pour applications distribuées suite au rachat de la seule alternative Open Source viable (Hortonworks) par une entreprise américaine (Cloudera).

Plus concrètement, le cas d'usage pilote du lac des données, SIRIUS PART, va voir son application portée dans l'environnement de production dès février 2021 avec son exploitation par les métiers dès la campagne (en fin d'année 2021.

Le dictionnaire des données de SIRIUS PART ainsi que celui des référentiels des entreprises utilisés dans le cadre des travaux de la MRV ont déjà été construits manuellement. Des POC sont en cours pour l'intégrer dans un dictionnaire sous format logiciel avec une date de production qui doit coïncider avec celle de SIRIUS PART, donc premier trimestre 2021. Les travaux du dictionnaire se poursuivent avec les données d'EAI -V2.

Plusieurs projets sont à apprécier à la fois pour utiliser le lac des données de manière native comme infrastructure et plateforme et pour être candidat à intégrer les données concernées dans le dictionnaire des données. C'est par exemple le cas pour le projet Foncier Innovant (FTAP 2019), les travaux du Service des collectivités locales notamment en lien avec l'application Hélios et l'infocentre Delphes ou encore les travaux en lien avec le projet ROC SP (FTAP 2020).

Au-delà des données et services de la DGFiP, ces outils peuvent être requis pour éventuellement porter, valoriser et mettre à disposition des données publiques ou des services opérés par d'autres administrations et services publics, conformément à la stratégie numérique de l'État.