14 JAN. 2020

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Le Gouvernement

GRAND PLAN D'INVESTISSEMENT

FONDS POUR LA TRANSFORMATION DE L'ACTION PUBLIQUE

Contrat de transformation

« Datalake et API Management : valorisation et mise à disposition des données de la DGFiP »

Ministère de l'Action et des Comptes publics

Direction Générale des Finances Publiques

ACTION PUBLIQUE 2022

Ce contrat de transformation est conclu entre la Direction Générale des Finances Publiques (DGFiP) et le Secrétariat général des ministères économiques et financiers, d'une part, et la Direction interministérielle de la transformation publique (DITP) et la direction du budget, d'autre part. Il définit les modalités d'exécution du projet, qui conditionnent le versement des crédits au titre du fonds pour la transformation de l'action publique. Il engage également le porteur de projet sur des résultats mesurables.

Compte-tenu du coût du projet, le directeur de la Direction interministérielle du numérique (DINUM) sera saisi pour avis conforme sur le présent projet lors de la phase de cadrage du projet, dans les six mois de la signature du contrat, conformément à l'article 3 du Décret n° 2019-1088 du 25 octobre 2019 relatif au système d'information et de communication de l'Etat et à la direction interministérielle du numérique.

Sur les 8 295 500 € financés au titre du FTAP, le versement de la première année sera acquis à la signature du contrat. Le versement des années suivantes sera conditionné par l'avis conforme du directeur de la DINUM.

1. Présentation du projet de transformation

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.

La DGFiP s'est très tôt engagée dans cette orientation, par exemple sur les jeux de données cadastrales. Plus récemment, la mise à disposition des données sur les transactions immobilières (DVF) et l'utilisation des données publiques pour la détection anticipée des entreprises en difficultés témoignent de la démarche d'ouverture de la DGFiP. Pour aller plus loin et valoriser pleinement ces données présentes en quantité et diversité dans son système d'information, la DGFiP doit moderniser et décloisonner son architecture informatique.

Le projet présenté « Datalake et APIM : Valorisation et mise à disposition des données » consiste à mettre en œuvre cette démarche en s'appuyant sur deux technologies complémentaires et indispensables :

- la première, appelée « lac de données », offre un stockage global des données brutes de la DGFiP et des mécanismes de croisements et de présentation (data-visualisation) facilement adaptables aux demandes futures des utilisateurs. Le projet de lac de données a pour corollaire l'implémentation des modules indispensables de dictionnaire de données et de data management ;
- la seconde, « l'API management », est une technologie autorisant une gestion industrialisée et sécurisée des interfaces qui permettra d'exposer les données de la DGFiP à l'ensemble de ses partenaires ayant juridiquement la possibilité d'y accéder en cohérence avec les exigences du RGPD[1].

1 Règlement général sur la protection des données entré en vigueur le 25 mai 2018.

2. Besoin et modalités de financement du projet

Le projet DataLake et API Management présente un coût global de 14,876 M€ dont 11,270 M€ de coûts directs (dont assistance et matériel) et 3,606 M€ de coûts indirects. Les coûts sont décomposés dans le tableau ci-après :

Nature de dépenses	Catégorie de dépenses	Programme budgétaire	2018		2019		2020		2021		2022		Cumul	
			AE	CP	AE	CP	AE	CP	AE	CP	AE	CP	AE	CP
APIM – Agents DGFiP	**Titre 2**		**12,3**	**12,3**	**281,7**	**281,7**	**256,3**	**256,3**	**146,9**	**146,9**	**0,0**	**0,0**	**697,2**	**697,2**
..dont f hancement DGFiP		P156	12,3	12,3	281,7	281,7	256,3	256,3	146,9	146,9			697,2	697,2
..dont f hancement FTAP		P349											0,0	0,0
APIM – Assistance MOA	**Titre 5**		**45,4**	**0,0**	**472,5**	**119,4**	**355,0**	**540,5**	**0,0**	**177,5**	**0,0**	**35,5**	**872,9**	**872,9**
..dont f hancement DGFiP		P156	45,4		74,0	119,4							119,4	119,4
..dont f hancement FTAP		P349			398,5		355,0	540,5		177,5		35,5	753,5	753,5
APIM – Assistance MOE	**Titre 5**		**179,9**	**4,6**	**697,1**	**394,4**	**0,0**	**478,0**	**0,0**	**0,0**	**0,0**	**0,0**	**877,0**	**877,0**
..dont f hancement DGFiP		P156	179,9	4,6	295,1	394,4		76,0					475,0	475,0
..dont f hancement FTAP		P349			402,0			402,0					402,0	402,0
APIM – Matériel et travaux architecture	**Titre 5**		**0,0**	**0,0**	**190,0**	**76,0**	**0,0**	**114,0**	**0,0**	**0,0**	**0,0**	**0,0**	**190,0**	**190,0**
..dont f hancement DGFiP		P156											0,0	0,0
..dont f hancement FTAP		P349			190,0	76,0		114,0					190,0	190,0
DataLake – Agents DGFiP	**Titre 2**		**122,5**	**122,5**	**300,4**	**300,4**	**1 383,9**	**1 383,9**	**1 101,9**	**1 101,9**	**0,0**	**0,0**	**2 908,7**	**2 908,7**
..dont f hancement DGFiP		P156	122,5	122,5	300,4	300,4	1 383,9	1 383,9	1 101,9	1 101,9			2 908,7	2 908,7
..dont f hancement FTAP		P349											0,0	0,0
DataLake – Cadrage technique amont	**Titre 5**		**0,0**	**0,0**	**0,0**	**0,0**	**200,0**	**80,0**	**0,0**	**120,0**	**0,0**	**0,0**	**200,0**	**200,0**
..dont f hancement DGFiP		P156											0,0	0,0
..dont f hancement FTAP		P349					200,0	80,0		120,0			200,0	200,0
DataLake – Assistance MOA	**Titre 5**		**250,0**	**0,0**	**0,0**	**250,0**	**700,0**	**280,0**	**700,0**	**700,0**	**0,0**	**420,0**	**1 650,0**	**1 650,0**
..dont f hancement DGFiP		P156	250,0			250,0	200,0	80,0	200,0	200,0		120,0	650,0	650,0
..dont f hancement FTAP		P349					500,0	200,0	500,0	500,0		300,0	1 000,0	1 000,0
DataLake – Assistance MOE	**Titre 5**		**0,0**	**0,0**	**0,0**	**0,0**	**2 800,0**	**1 120,0**	**2 850,0**	**2 820,0**	**0,0**	**1 710,0**	**5 650,0**	**5 650,0**
..dont f hancement DGFiP		P156					800,0	320,0	800,0	800,0		480,0	1 600,0	1 600,0
..dont f hancement FTAP		P349					2 000,0	800,0	2 050,0	2 020,0		1 230,0	4 050,0	4 050,0
DataLake – Travaux d'architecture	**Titre 5**		**0,0**	**0,0**	**0,0**	**0,0**	**150,0**	**60,0**	**100,0**	**130,0**	**0,0**	**60,0**	**250,0**	**250,0**
..dont f hancement DGFiP		P156											0,0	0,0
..dont f hancement FTAP		P349					150,0	60,0	100,0	130,0		60,0	250,0	250,0
DataLake – Materiel, logiciel	**Titre 5**		**0,0**	**0,0**	**0,0**	**0,0**	**550,0**	**550,0**	**500,0**	**500,0**	**0,0**	**180,0**	**1 050,0**	**1 230,0**
..dont f hancement DGFiP		P156											0,0	0,0
..dont f hancement FTAP		P349					550,0	550,0	500,0	500,0		180,0	1 050,0	1 230,0
DataLake – Sécurisat on des accès au lac	**Titre 5**		**0,0**	**0,0**	**0,0**	**0,0**	**100,0**	**40,0**	**300,0**	**180,0**	**0,0**	**0,0**	**400,0**	**220,0**
...dont f hancement DGFiP		P156											0,0	0,0
...dont f hancement FTAP		P349					100,0	40,0	300,0	180,0			400,0	220,0
DataLake – Format on (MOA,MOE et exploitant)	**Titre 3**		**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**130,0**	**130,0**	**0,0**	**0,0**	**130,0**	**130,0**
...dont f hancement DGFiP		P156							130,0	130,0			130,0	130,0
...dont f hancement FTAP		P349											0,0	0,0
TOTAL			610,1	139,4	1 941,7	1 421,9	6 495,2	4 902,7	5 828,8	6 006,3	0,0	2 405,5	**14 875,8**	**14 875,8**

TOTAL Financement DGFIP		***P156***	**610,1**	**139,4**	**951,2**	**1 345,9**	**2 640,2**	**2 116,2**	**2 378,8**	**2 378,8**	**0,0**	**600,0**	6 580,3	6 580,3
TOTAL Financement FTAP		***P349***	**0,0**	**0,0**	**990,5**	**76,0**	**3 855,0**	**2 786,5**	**3 450,0**	**3 627,5**	**0,0**	**1 805,5**	**8 295,5**	**8 295,5**

Financement du projet :

La contribution de la DGFiP est de 6,580 M€. Ainsi, la contribution demandée au Fonds pour la transformation de l'action publique (FTAP) s'élève à 8,295 M€.

Détail des dépenses financées par le Fonds :

L'apport du FTAP permet de financer le volet du projet correspondant à une partie des investissements dans l'infrastructure informatique et des prestations d'assistance aux bureaux MOA et MOE, afin de construire la solution.

3. Économies prévisionnelles engendrées par le projet

L'approche par la donnée offre une nouvelle vision de l'informatique d'entreprise. Les technologies de « Big Data » permettent d'exploiter des données massives structurées et non structurées avec des outils à l'état de l'art (Hadoop) de stockage et de traitement. Cette nouvelle approche permet de décloisonner les infocentres spécialisés existants.

Dans la mesure où elle sera déjà dupliquée dans le lac, une donnée pourra être réutilisée pour autant de traitements que nécessaire, sans impact supplémentaire sur l'application productrice de celle-ci.

Ces technologies ouvrent des possibilités d'application dans tous les domaines métier (fiscalité, contrôle, conseil, simulation de réformes, RH, ...). Ci-après figurent des cas d'usage identifiés à ce jour. Au fil de l'appropriation de ces outils, d'autres cas d'usage émergeront.

• Bénéfices attendus pour les partenaires de la DGFiP :

Les données DGFiP seront partagées avec les partenaires externes au travers du dispositif *d'API-management*. En effet, cet accès aux données est actuellement géré au cas par cas, avec beaucoup d'interventions humaines. Ce constat a été fait pour le déploiement de l'API « Impôt Particulier » au bénéfice de la Ville de Lyon (ouvert en janvier 2017) ou au ministère de l'Éducation nationale pour le dépôt des bourses des collégiens et lycéens (avril 2017).

Pour passer à l'échelle, et ouvrir largement les données de la DGFiP, il est nécessaire d'outiller ces démarches par la mise en place d'un composant mutualisé et industrialisé de gestion des API. Cet outil met à disposition un portail avec la documentation des différentes API et une capacité à tester leur utilisation ; le développement par les partenaires de nouveaux services appuyés sur les données DGFiP s'en trouve ainsi accéléré. Il offrira aussi un outillage de la contractualisation (lien avec l'outil Sign'Up de la DINUM).

Au-delà de la mise à disposition des données, l'outillage datalake pourra aussi mobiliser et valoriser des données publiques au bénéfice des partenaires. A titre d'exemple, la DGFIP a déjà évoqué les projets suivants :

- Ministère de la Justice : de prochains échanges seront engagés pour accélérer la fiabilisation et le traitement des dossiers d'aide juridictionnelle à l'aide des données DGFiP ;
- GiP Union Retraite : échange de données relatives à l'identification des personnes et, à terme, au revenu fiscal de référence ;
- Inspection Générale des Finances : ouverture du lac de données au pôle DataScience de l'IGF. Cela permettrait à l'IGF de réduire sa dépendance à l'outil payant CASD[2] pour l'accès

2 CASD : Centre d'accès sécurisé aux données.

aux données publiques et de pouvoir conduire ses études sur de très grands volumes de données grâce à la puissance de calcul du lac de données.

• Bénéfices attendus pour les agents et les services métiers de la DGFiP :

De nombreux besoins métier ont d'ores et déjà été exprimés par les services et les agents, par exemple :

- la mise à disposition de documents fiscaux (avis d'imposition notamment) reste fréquente, chronophage et sans valeur ajoutée pour la mission fiscale. L'accès à ces données par API dans le cadre de processus dématérialisés d'autres administrations permettra d'alléger la tâche des agents ;
- la valorisation des données RH des agents de la DGFiP tant dans l'approche GPEEC (meilleur profilage poste-agent ; détection des compétences) que dans l'évaluation de la masse salariale ;
- la réalisation du profil complet (360°) des redevables particuliers et professionnels par le recoupement de données fiscales, foncières, patrimoniales : à des fins d'optimisation du recouvrement ; pour la personnalisation de la relation à l'usager ; pour la mobilisation des données utiles au traitement du contentieux, comme l'a récemment mis en évidence le projet de réingénierie de la fonction contentieuse mené par la DITP et la DGFiP ;
- la sélection de populations « particuliers » ou « professionnels » qui répondent à des critères métiers ou qui peuvent être impactées par des mesures gouvernementales à des fins d'information ciblée de nos interlocuteurs ;
- la simulation de mesures fiscales ou sociales à tout ou partie d'un type de population ou de collectivités locales ;
- les analyses comportementales et prédictives afin d'orienter les stratégies métiers et de définir une offre de services adaptée aux usagers.

Les outils permettront de développer des API internes à la disposition des applications de gestion pour l'ouverture de services nouveaux non couverts par les technologies actuelles notamment : simulation, analyse de graphes, représentation graphique d'entités.

Ces outils faciliteront le travail des agents dans leur gestion avec une meilleure visualisation et une plus grande fraîcheur des données. La mise en production du projet évitera certaines sollicitations dans les SIP et permettra un gain annuel évalué à **276 ETP**.[3]

Hors SIP, ce gain annuel est estimé à **51 ETP**[4].

• Autres bénéfices métiers :

Le projet est également un facilitateur du déploiement du programme « Dites-le nous une fois » qui va connaître une accélération rapide avec notamment la mise en place par la DINUM dès le mois de septembre 2019 d'un indicateur de pré-remplissage de l'ensemble des formulaires administratifs avec les données déjà connues par l'administration (les données collectées par la DGFiP figurent parmi les principales et les plus fréquentes qui seront utilisées).

3 Durant la campagne de paiement des impôts des particuliers, à l'automne 2018, les accueils des SIP ont été sollicités par les usagers pour délivrer 8 600 000 documents et avis. La délivrance de 30 % de ces documents sera évitée par la mise en place du projet datalake et APIM. 8 600 000 x 30 % x 10' de traitement / 93160' par ETP = 276 ETP.
4 Le nombre annuel de sollicitations des directions par les organismes extérieurs est évalué à 234 000 avec un temps de traitement unitaire estimé à 20 minutes soit 234 000 x 20' / 93160' par etp = 51 ETP.

Ainsi de nouveaux usages ont été identifiés et sont listés ci-après (liste non exhaustive) :

- requêtes concernant les données budgétaires et comptables sur un périmètre ministériel, interministériel voire étendu aux collectivités locales, avec des usages tels que ceux identifiés lors des derniers hackathons DataFin (23 projets de réutilisations ont été identifiés) ;
- réalisation d'évaluations immobilières pour les collectivités ou les ministères dans le cadre des inventaires, voire pour des opérateurs publics, en rapprochant les évaluations immobilières des domaines, de données des notaires et des professionnels de l'immobilier par exemple ;
- consultation des données bancaires des personnes physiques par les acteurs institutionnels autorisés par la loi ;
- accélération des cas d'usage de l'API « impôt particulier » au bénéfice d'un nombre important de collectivités territoriales et aux organismes prestataires d'aides sociales en simplifiant notamment les démarches des usagers auprès des services municipaux (cantines scolaires, crèches, accès à des tarifs sociaux), et de l'ensemble des guichets d'aides sociales et d'accès aux droits (établissement des droits à pension, bénéfice de l'aide juridictionnelle, bourses et aides scolaires ...) ;
- transmission de données aux nombreux organismes bénéficiant de l'accès aux données d'origine fiscale via le droit de communication pour des missions de contrôle ou de recouvrement (officiers de police judiciaire, Tracfin, organismes sociaux, conseils départementaux, huissiers de justice...).

• Bénéfices techniques :

D'un point de vue du système d'information de la DGFiP, ce projet apporte aussi une réactivité accrue et une sécurisation de fonctionnement.

Sur le plan de la réactivité, en facilitant l'accès et le recoupement des données, les nouvelles technologies de traitement et de stockage des données permettront d'alimenter le lac avec des formats multiples : base de données Oracle ou Postgres, fichier texte ou CSV, fichiers XML, document PDF, image. Ainsi une seule requête permet d'obtenir toutes les informations détenues aujourd'hui par plusieurs applications en une seule interrogation par API. Ceci constitue un levier d'accélération de l'ouverture des données de systèmes reposant sur des technologies anciennes considérées comme la partie « legacy[5] » du SI.

De son côté, le management des API permettra de simplifier à la fois l'urbanisation du système d'information en standardisant les échanges inter-applicatifs (architecture REST), et la gestion des dépendances entre applications dans leur cycle de vie propre par une gestion du cycle de vie des API.

Sur le plan de la sécurité, l'API-management apportera des mécanismes de quota d'utilisation attribués aux applications internes et aux partenaires, qui protègent les sources de données, même en situation de forte sollicitation. Le datalake, au travers de son système d'habilitation, sécurise l'accès légitime aux données, en cohérence avec la réglementation RGPD.

5 Un système hérité, système patrimonial ou legacy system est un matériel et/ou logiciel continuant d'être utilisé dans une organisation alors qu'il est supplanté par des systèmes plus modernes.

• Calculs des économies de personnel :

	Catégorie A	Catégorie B	Catégorie C
Coût moyen par emploi (en €)	42 835	35 069	30 317

	2020	2021	2022	2023
Emplois supprimés	0	0	51	276
dont catégorie A			5	
dont catégorie B			21	
dont catégorie C			25	276
Economie en €	0	0	854 275	5 892 295
dont catégorie A	0	0	107 088	214 175
dont catégorie B	0	0	368 225	736 449
dont catégorie C	0	0	378 963	4 941 671

Les économies sur les coûts de fonctionnement liés à l'agent sont évaluées à 2 490 € par agent.

	Catégorie de dépenses	Programme budgétaire	2022		2023		Économies annuelles pérennes post-déploiement du projet	
			AE	CP	AE	CP	AE	CP
Délivrance automatique de documents (avis) en SIP et hors SIP	T2	P156	854 275	854 275	5 892 295	5 892 295	10 076 041	10 076 041
Economies sur les coûts de fonctionnement liés à l'agent	HT2	P156	63 495	63 495	470 610	470 610	814 230	814 230
TOTAL			917 770	917 770	6 362 905	6 362 905	10 890 271	10 890 271

4. Calendrier, gouvernance et modalités de réalisation des projets

4.1. Calendrier prévisionnel

Les principales réalisations envisagées sont listées ci-après :

– un Datalake :

1. mise en œuvre du dictionnaire des données de la DGFiP ;
2. installation de l'infrastructure permettant d'accueillir le lac de données ;
3. développement des composants d'extraction et de transformation des données de gestion pour alimenter le lac de données ;
4. développement et intégration dans le plan de production DGFiP des traitements batchs de chargement des flux de données dans le lac de données ;
5. sécurisation des couloirs d'accès aux données du lac ;
6. développement des traitements d'extraction des données dans le lac ;
7. exposition de ces données via des API par la mise à disposition.

– un portail à destination des développeurs internes, partenaires et externes doté d'un bac à sable et de la documentation technique pour chaque API ;

– un portail de gestion d'API sur lequel sont publiées les API, déposées les documentations et gérés les autorisations des partenaires et les quotas d'utilisation.

4.1.1 Calendrier prévisionnel datalake :

Après une étude de cadrage et la mise en place d'un projet transitoire « datalab » préfigurateur de la cible « datalake », réalisé sur la base d'une structure d'innovation et de transformation à gouvernance simplifiée conduit en 2019 qui a permis de qualifier l'outillage technique « lac de données », l'ensemble des technologies nécessaires à la construction d'un datalake est maintenant maîtrisé.

La DGFiP est techniquement prête à passer à l'échelle dés 2020. Dès le dernier trimestre 2019, les travaux suivants ont été menés : stratégie de basculement du projet pilote datalab vers le projet datalake et sélection de l'outil « dictionnaire des données ».

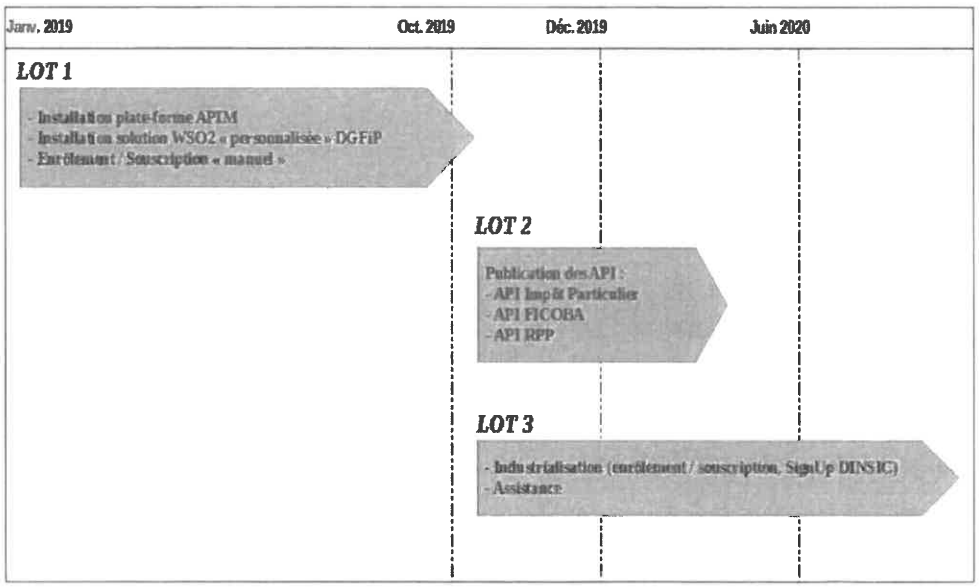
Au 1er trimestre 2020, les nouveaux chantiers décisionnels seront planifiés en lien avec les orientations prises dans le cadre de l'exposition des données et de leur ouverture pilotées par l'administrateur des données de la DGFIP.

La composition des différents lots du projet datalake se présente comme suit :

- Lot 1 : **statistiques EAI V2** (mise en production en septembre 2020) et **reprise de l'infocentre SIRIUS PART** (mise en production en mars 2021) ;
- Lot 2 : **reprise de l'infocentre SIRIUS PRO** (mise en production en 2021) et **infocentre SCL** et prise en compte de nouveaux besoins dont ROCSP (mise en production en 2021) **;**
- Lot 3 : **reprise des données des derniers infocentres réalisés avec les technologies COGNOS/SAS** (mise en production en 2022) **et prise en compte de besoins nouveaux.**

4.1.2 Calendrier prévisionnel « API Management »

L'API Management (APIM), plate-forme ouverte aux partenaires de la DGFiP, facilitera le partage des données ou de services de la DGFiP en outillant chaque étape : expérimentation itérative, contractualisation, contrôle d'accès, mise en service sécurisée. La plate-forme constitue le pré-requis nécessaire et indispensable à l'industrialisation du service des API exposées par la DGFiP.



La composition des différents lots du projet d'API Management se présente de la manière suivante :

- Lot 1 : **Primo-installation API Management** (22 octobre 2019) :
  - Installation de la plate-forme APIM et de la solution WSO2 « personnalisée » DGFiP ;
  - Enrôlement[6] / Souscription[7] « manuel ».
- Lot 2 : **Publication des premières API DGFiP** (octobre 2019 - avril 2020) :
  - API Impôt Particulier (données de la fiscalité des particuliers) ;
  - API FICOBA (données bancaires) ;
  - API RPP (données d'état civil et d'adresse des personnes physiques).
- Lot 3 : **Industrialisation** (dernier trimestre 2020) :
  - Automatisation des processus d'enrôlement et de souscription (échanges avec les briques SSO et SignUp de la DINUM) ;
  - Assistance interne (projets API DGFiP) et externe (partenaires DGFiP).

4.2. Gouvernance du projet

- Maîtrise des risques :

La maîtrise des risques du projet est assurée par plusieurs actions dont certaines sont déjà réalisées et d'autres sont en cours.

Avant 2019, des premières expérimentations ont été menées :

- mise en place d'une plate-forme « bac à sable » sur Hadoop, qui a permis l'expérimentation et la qualification technique de l'outillage « lac de données ». Les tests suivant ont été conduits :
  - tests de performance des traitements Spark vis-à-vis des traitements faits sur Oracle ;
  - mise en place expérimentale des règles de sécurité d'accès aux données via Apache Ranger ;
  - réalisation en mode agile et en moins d'un mois d'une application R Shiny qui permet à un artisan de choisir dans quelle commune s'implanter compte tenu de la présence d'autres artisans et de la population présente ;
  - qualification de Hadoop 3 vis-à-vis de Hadoop 2 ;
- test de l'organisation projet avec les bureaux MOA et MOE, montée en compétence des équipes grâce à cette même plate-forme « bac à sable » ;
- réalisation de deux POC sur l'API management, l'un à visée organisationnelle pour préciser l'organisation projet à mettre en place dans ce cadre, et l'autre à visée technique. Dans un premier temps, une étude à dire d'expert a permis d'identifier la solution open source de la Société WSO2 (parmi cinq autres solutions). La réalisation d'une preuve opérationnelle de concept avec WSO2 a confirmé la couverture fonctionnelle de la solution, sa capacité à fonctionner dans notre environnement et avec notre annuaire LDAP ainsi qu'en lien avec France Connect. Il a pu être vérifié que les performances

6 Enrôlement : création d'un compte permettant à un partenaire d'accéder au portail des API de la DGFiP (Store)
7 Souscription : abonnement à une API permettant à un partenaire d'accéder à l'API

annoncées par l'éditeur sont reproductibles sur les environnements de la DGFiP. Cela a nécessité des réglages approfondis sur l'ensemble des composants de la plate-forme – Socle Linux, JDK et base de données PostgreSQL. Les résultats collectés dans un document de près de 200 pages ont été déterminants pour le lancement du projet d'APIM de la DGFiP ;

- études d'architecture (impacts sur le SI notamment) ;
- mise en place de formations et de monitorat.

En 2019, un projet transitoire « datalab » préfigurateur de la cible « datalake » a été mis en place sur la base d'une structure d'innovation et de transformation à gouvernance simplifiée dont les trois objectifs sont :

- une bonne maîtrise par la DGFIP d'architectures innovantes et de technologies émergentes liées à la filière Hadoop/Hortonworks ;
- la réalisation de projets orientés data et numérique à valeur ajoutée pour la DGFiP :
  - rénovation de l'outil de recoupement et de contrôle fiscal dans le domaine des impôts des particuliers, en liaison avec le projet Pilat financé par le FTAP ;
  - production des statistiques sur les échanges internationaux de données fiscales dans le cadre des accords internationaux (CRS, FATCA) ;
  - développement de métiers autour de la datascience, etc ;
- la mise en place d'un centre de compétences autour de l'exploitation d'une Infrastructure Hortonworks / Hadoop mutualisée devant éviter une dispersion des compétences techniques « data engineer » dans le SSI.

Le pilotage et la cartographie des risques s'inspirent des processus établis par la DINUM.

La sécurisation de la phase de généralisation va se traduire par les actions ci-dessous :

- instancier une cartographie des risques en début de projet, répertoriant l'ensemble des risques projets identifiés. Pour chaque risque identifié, seront définis les éléments suivants : le niveau de criticité, le niveau d'impact, un plan d'action de couverture ainsi qu'un responsable de la mise en œuvre des actions identifiées. La cartographie s'inspire de celle établie par la DINUM ;
- inscrire le pilotage des risques dans les instances actuelles de gouvernance, afin de veiller à ce que les risques identifiés soient notamment couverts à travers la mise en œuvre du plan d'action défini.

● Gouvernance et responsabilité opérationnelle

Dans le cadre de ce projet, la DGFiP va affecter des ressources internes compétentes dans le domaine des infocentres et qui s'approprieront les technologies de valorisation de la donnée.

La gouvernance du projet est d'ores et déjà organisée autour de parties prenantes clairement identifiées en administration centrale :

➢ MOA confiée au SCN Cap Numérique - bureau Cap Usagers : son expérience sur les référentiels fondamentaux, et son investissement sur les problématiques France Connect et de management des API font de ce Bureau une structure experte dans le domaine de la

valorisation et de l'exposition des données. En interne, il bénéficiera aussi de l'expertise et du soutien du Bureau Cap Particuliers qui a porté la création en 2016 de l'API « Impôt Particulier ». ;

- MOE du lac de données : bureau SI-1D : réalisation de projets majeurs (comptes bancaires ; archivage électronique ; système SI RH...), et conduite de projets en méthode agile / devops (Bofip), travaux sur l'API management et sur les technologies Hadoop ;
- MOE de l'*API management* : bureau SI-1F : réalisation de projets majeurs (Portail impots.gouv.fr, Hélios), il est également spécialisé dans les annuaires et les portails d'accès.

  Référent technologique : bureau SI-1A : qualification technique des composants Hadoop et WSO2, structuration des API REST à la DGFiP.

  En outre, un référent technique au sein du SSI est d'ores et déjà en poste et plus particulièrement positionné sur la mise en œuvre de l'infrastructure (bureau SI-2B).

Enfin, l'administrateur des données de la DGFiP appuiera l'ensemble des équipes de travail compte tenu des enjeux contenus dans ces travaux au regard de l'exposition des données et de leur ouverture dont il est le garant pour la DGFiP.

En termes de gouvernance, le projet sera suivi au sein des instances classiques et éprouvées par la DGFiP : comité de suivi projet (CSP) au niveau des chefs de projet, comité de suivi opérationnel (COMOP) au niveau des chefs de bureau et comité de pilotage (COPIL) présidé par le Directeur général adjoint (DGA). Un reporting de l'état d'avancement du projet et le suivi des risques inhérents (cf. infra) seront systématiquement portés à l'ordre du jour du comité trimestriel où sont spécifiquement suivis les projets liés à la valorisation des données, présidé par le DGA.

4.3. Modalités de réalisation du projet – respect des principes de l'Etat plateforme

4.3.1 Description du projet de datalake

Dans un premier temps, un premier module datalab, ouvert à tous les domaines métiers (fiscalité, contrôle, conseil, simulation de réformes, RH...), est conçu pour :

- la transformation et /ou restitution de données en différé ;
- l'exploitation de données massives structurées et non structurées avec des outils à l'état de l'art (Hadoop) de stockage et de traitement.

Ce projet est préfigurateur de la cible « datalake » et orienté offre de services techniques :

- une technologie HADOOP et des services BigData transparents pour les « clients » ;
- une allocation de ressources (stockage, CPU, mémoire) ;
- des accès sécurisés aux données et des traitements pour les couloirs de valorisation ;
- des outils pour des nouveaux usages métiers : datascience, Self-service, Datavisualisation ...

Il permet :

- l'accélération de l'exposition des données ;
- la rénovation rapide de l'architecture des infocentres spécialisés actuels ;

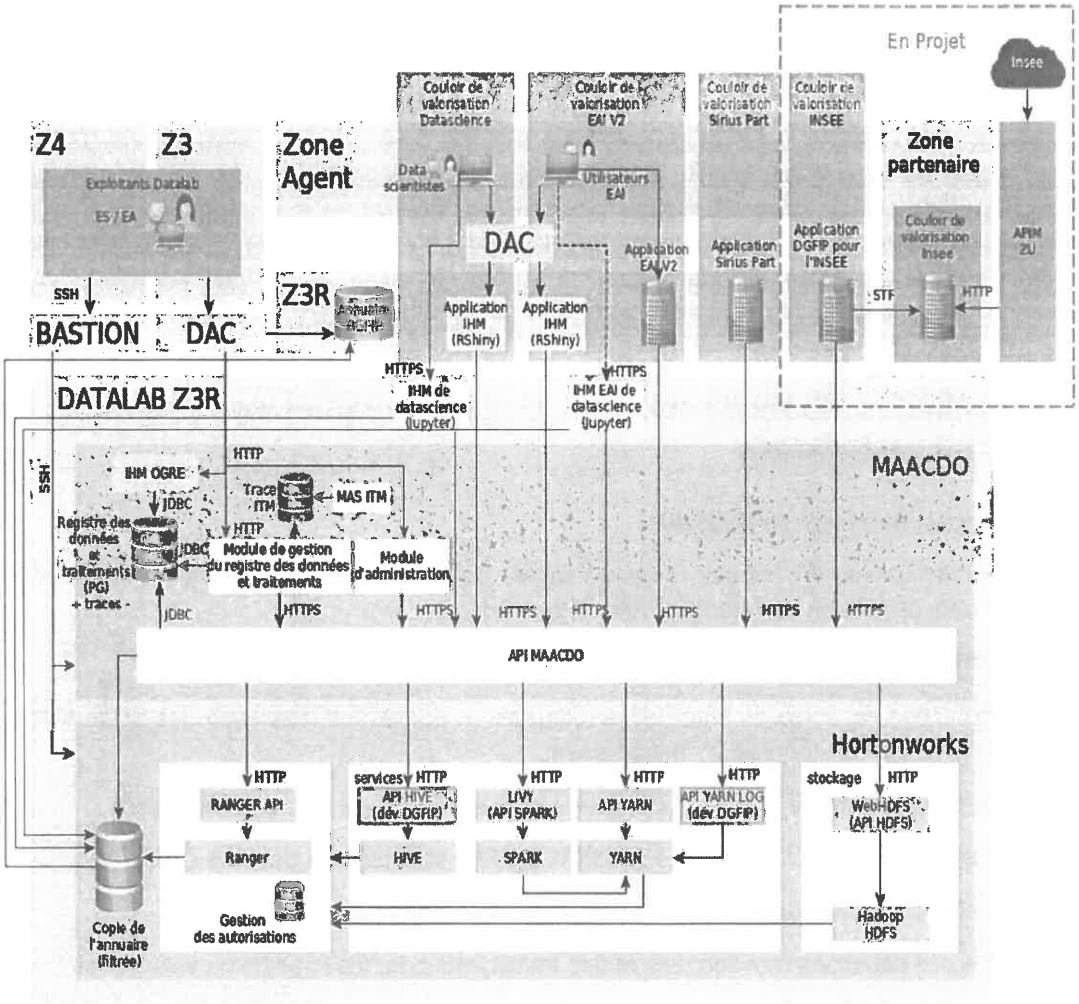
- le développement des API internes à destination des applications de gestion, pour de nouveaux services non couverts par les technologies actuelles, notamment : statistiques, prototypage rapide, représentation graphique, datavisualisation...

Le datalake permettra ensuite :

- la récupération de données unitaires (le Big Data est orienté pour le traitement de batch sur de fortes volumétries) ;
- l'hébergement de données vivantes (i.e non modifiables par les actes métiers des agents).

Au plan technique, le projet datalab DGFIP repose sur :

- une offre de services Big DATA (via la distribution Hortonworks - version 3.1) permettant le stockage des données et les traitements distribués sur la plate-forme HADOOP ;
- des technologies BigData/ Hadoop ;



- des règles de sécurité d'accès aux données fortes

4.3.2 l'API Management

L'API Management vise à promouvoir le catalogue des API de la DGFiP auprès de ses partenaires. Il offre également des fonctionnalités permettant de simplifier, d'automatiser et de sécuriser l'accès aux API.

Dans cette perspective, et dans une logique de mutualisation inter-ministérielle, « l'API Management » va s'appuyer sur 3 briques logicielles majeures de la DINUM :

- le portail des API de l'État « api.gouv.fr », afin de promouvoir les API de la DGFiP ;
- la brique d'authentification « SSO », afin d'automatiser le processus d'enrôlement des partenaires ;
- la brique de contractualisation « SignUp », afin d'automatiser les processus de contractualisation et de souscription aux API ;

En termes de conduite du changement, une documentation pédagogique sera mise à disposition des partenaires pour les accompagner dans leur démarche de raccordement aux API de la DGFiP. Elle sera accompagnée d'un dispositif d'assistance dédié afin de répondre aux questions des partenaires (FAQ, formuel, assistance technique). En interne, un plan de formation sera mis en place afin que les agents de cette assistance puissent acquérir et consolider les connaissances et compétences nécessaires à l'exercice de leur mission.

À ce stade, il est prévu que les API suivantes soient exposées sur le portail des API de la DGFiP :

API	***Données exposées***	***Partenaires API Oct. 2019 – 1er semestre 2020***	***Partenaires API cible***
API Impôt Particulier	*Données de la fiscalité des particuliers*	*20 collectivités/organismes*	*~700 collectivités/organismes*
API FICOBA	*Données bancaires*	*CHORUS*	*Banques (~150)*
API RPP (Recherche Par Personne)	*Données d'état civil et d'adresse des personnes physiques*	*10 collectivités/organismes*	*50 000 collectivités/organismes*

Socle technologique :

Les composants majeurs de l'API Manager sont :

– WSO2 Identity Server 5.7.0, pré-packagé pour être Key Manager de l'APIM 2.6.0 ;

– Composant WSO2 Business Process Server profile de WSO2 Enterprise Integrator 6.4.0 pour gérer les Workflows avec validation humaine ;

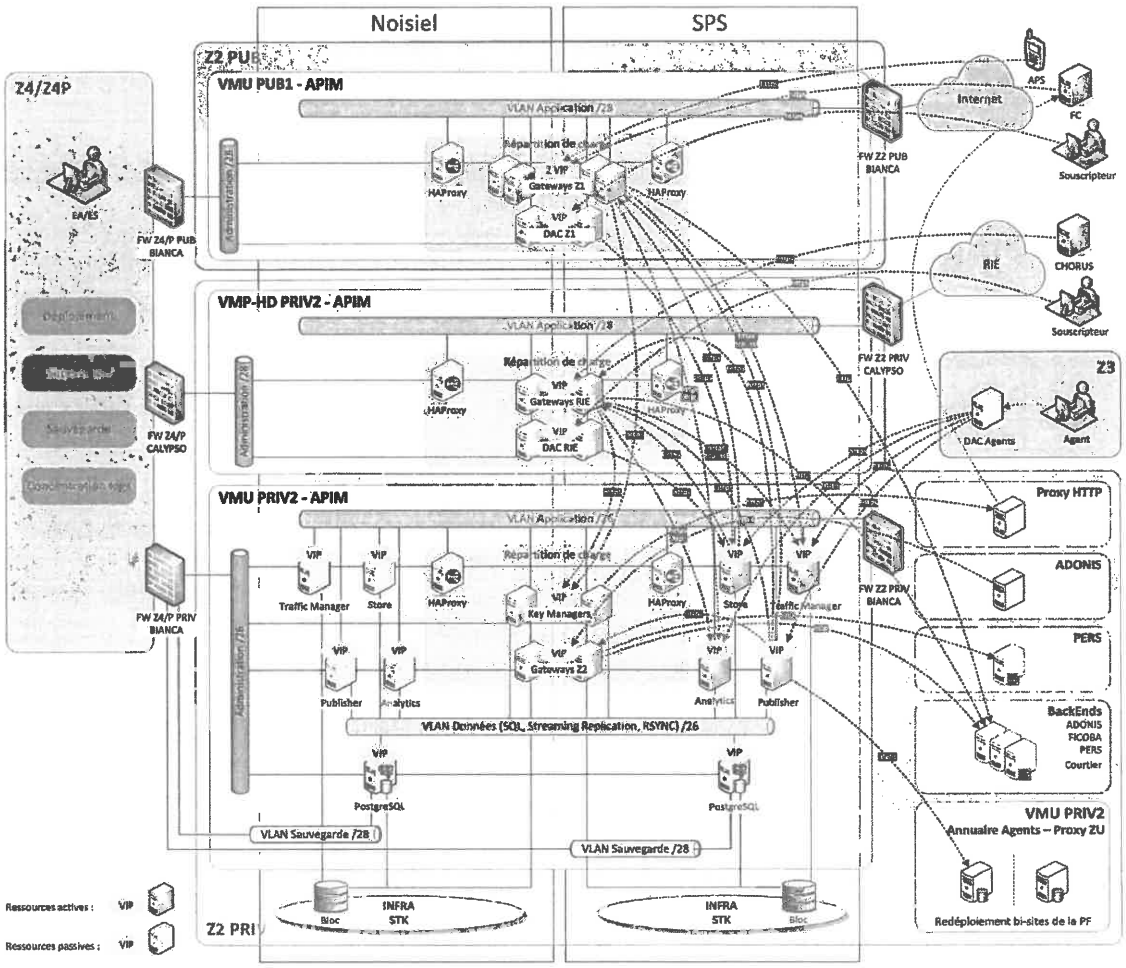
– WSO2 API-M Analytics 2.6.0 : composant de gestion des statistiques adapté à l'APIM ;

[remarque : ces 4 composants sont hors MATECH. Ils nécessitent une dérogation. Ils sont supportés sur l'OpenJDK 8 (APIM est certifié avec ce dernier)]

– Linux socle 2016 DGFIP, OpenJDK 8, PosgreSQL 10.4 ;

– LOAD BALANCER : HAProxy 1.8.

Schéma d"architecture APIM :



Pour permettre à l'ensemble de nos différents partenaires d'accéder aux API de la DGFiP de façon sécurisée, l'APIM disposera de 4 couples de GateWay :

– Internet ;

– Internet France Connecté ;

– RIE ;

– Interne (VMU PRIV 2) ;

L'ensemble des composants est redondé, il pourra à terme être en bi-site.

5. Modalités de suivi et critères d'évaluation du projet

Le présent contrat donne lieu à un suivi du projet financé. Des indicateurs d'avancement et de résultats sont suivis dans le cadre du financement du projet. Ces indicateurs sont communiqués, à sa demande et au moins une fois par an, au secrétariat du fonds. Des réunions de suivi pourront être organisées à la demande d'une des parties lors de la communication de ces indicateurs.

5.1. Indicateurs d'avancement

Les indicateurs d'avancement permettent de sécuriser la mise en œuvre du projet. Ils sont définis comme suit :

- Montant des crédits consommés en AE et CP et respect des enveloppes allouées à chacune des tranches, par nature de dépense, par rapport aux besoins de financement présentés dans le point 2 du présent contrat ;
- Respect du calendrier prévisionnel de déploiement du projet, par rapport au calendrier présenté dans le point 4.1 du présent contrat.

5.2. Indicateurs de résultat et d'impact

Les indicateurs de résultat permettent d'évaluer l'atteinte des objectifs du projet :

- Montant d'économies générées (€) et répartition par nature de dépenses. Les économies réalisées seront comparées aux économies prévisionnelles présentées dans le point 3 du présent contrat ;
- Les indicateurs d'impacts permettent d'évaluer la réalisation des objectifs du projet :
    - **Taux d'avancement de l'adhésion des partenaires aux API via l'APIM :**
        - l'industrialisation des échanges de données par la DGFiP via des API devra croître et aura pour conséquence, en parallèle, la réduction de projets destinés à mettre en place des échanges de données spécifiques, les partenaires/clients de ces données devant s'abonner à APIM pour en bénéficier. A titre d'exemple, la transmission des données au bénéfice des collectivités territoriales ou des administrations sociales donnera lieu à une mesure spécifique à caractère semestriel ;
        - donc, pour un périmètre d'API de référence (API Impôt Particulier, API FICOBA et API RPP), il s'agira de calculer le nombre de partenaires ayant adhéré aux API/nombre de partenaires cible.
    - **Taux de réduction des sollicitations annuelles des usagers (particuliers et professionnels) pour la délivrance de documents et avis :**

      Il s'agit de mesurer la baisse de la sollicitation annuelle des demandes de documents et avis par rapport aux 8 600 000 demandes constatées en 2018.

Indicateur	Valeur actuelle	Cible fin 2020	Cible fin 2021	Cible fin 2022	Cible fin 2023
Taux d'avancement de l'adhésion des partenaires aux API via l'APIM	**< 1 %**	**so**	**5 %**	**20 %**	**50 %**
Taux de réduction des sollicitations des usagers (particuliers et professionnels) pour la délivrance de documents et avis				**7%**	**30 %**

6. Modalités et calendrier de versement des aides

Les crédits sont mis à la disposition **de la secrétaire générale du ministère de l'économie et des finances**. La secrétaire générale procède aux diligences nécessaires pour permettre l'ordonnancement des crédits du FTAP par les directions concernées.

La mise à disposition des crédits s'effectue par tranche. Le montant de chaque tranche sera définitivement arrêté par le secrétariat du fonds, les dépenses annuelles détaillées dans la présente convention étant prévisionnelles, à l'exception de la première année de financement (2020). À partir de 2021, le secrétariat du fonds décide à échéance régulière, *a minima* au 1er trimestre de chaque année, du montant des nouvelles tranches de financement au regard de l'avancement du projet, du suivi des indicateurs et de l'avis rendu par le DINUM sur le projet.

S'agissant le cas échéant des opérations d'investissement (titre 5 majoritaire), le montant des AE nécessaires pour le financement d'une phase fonctionnelle du projet[8] devra faire l'objet d'une affectation au sens de la comptabilité budgétaire, en cohérence avec les phases du projet décrites au 4.1.

Les crédits sont mis à disposition dans le cadre de gestion BOP-UO décrit en annexe. La consommation des crédits (AE et CP) sur le programme 349 est opérée en référençant la nomenclature budgétaire d'activités annexée au présent contrat.

7. Matérialisation des économies réalisées

La matérialisation des économies liées au projet est suivie annuellement, conformément aux indicateurs définis au paragraphe 5.2. Le porteur de projet communique au secrétariat du fonds les économies effectivement réalisées et explicite les raisons des éventuels écarts avec les prévisions exposées dans le présent contrat.

8 Au sens de l'article 8 de la LOLF

8. Modifications du contrat de transformation

Le présent contrat peut être modifié par voie d'avenant.

Toute difficulté majeure dans la réalisation du projet sera portée à la connaissance du comité de pilotage qui pourra suspendre ou interrompre les financements initialement définis.

9. Communication liée au projet

Dans toute communication relative à son projet, le porteur est invité à préciser qu'il a reçu le soutien financier du Fonds pour la transformation de l'action publique.

14 JAN. 2020

Le directeur général des finances publiques

Jérôme FOURNEL

La secrétaire générale des ministères économiques et financiers,

Marie-Anne BARBAT-LAYANI

Le délégué interministériel à la transformation publique

Thierry LAMBERT

La directrice du budget

Amélie VERDIER

ANNEXE RELATIVE AUX NOMENCLATURES BUDGETAIRES D'EXECUTION

Cadre de gestion BOP-UO : 0349-CDBU-CEFI

Action – Domaine fonctionnel : 0349-01

Référentiel de programmation :

Code Chorus	Désignation Chorus	Commentaires
034901014201	MACP - DataLake	Concerne toutes les dépenses HT2 relatives au projet et imputées sur le programme 0349