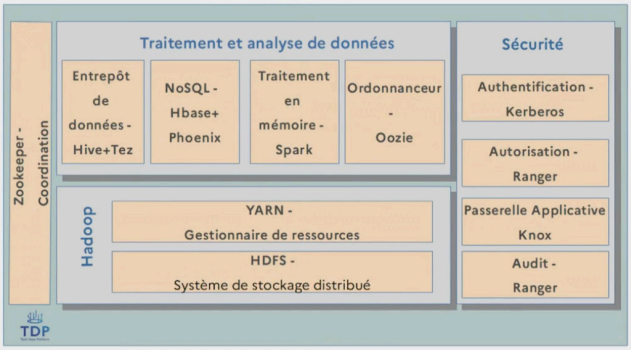
Le contexte

La DGFiP a mis en production depuis mars 2021, une infrastructure Big Data dite « Lac de données », open-source, basée sur la distribution Hortonworks. Elle en a défini l'architecture technique et logicielle, et réalisé l'intégralité de sa mise en œuvre.

- Faciliter et centraliser l'accès aux données pour optimiser leur valorisation.
- Favoriser la réutilisation et le partage des données entre services de la DGFiP (données internes, données des partenaires) et accélérer notamment l'Open Data.
- Développer les mécanismes de croisements de données (décloisonner les infocentres spécialisés).
- Valoriser les données issues des applications de gestion en veillant notamment à son décloisonnement ou provenant des partenaires de la DGFiP.
- Promouvoir en facilitant les usages de la datascience et de la datavisualisation.

TOSiT Data Platform



TOSit Data Platform

Les versions de composants sont alignées avec celles de HDP 3.1.5.

Composant	Version	Apache Git branch
Apache ZooKeeper	3.4.6	release-3.4.6
Apache Hadoop	3.1.1	branch-3.1.1
Apache Hive	3.1.3	branch-3.1
Apache Hive 1	1.2.3	branch-1.2
Apache Tez	0.9.1	branch-0.9.1
Apache Spark	2.3.5	branch-2.3
Apache Ranger	2.0.1	ranger-2
Apache Oozie	5.3.0	master

L'écosystème de Hadoop est très riche. Le projet commence par les composants les plus essentiels des cas d'usage chez EDF et DGFiP.

D'autres composants pourraient être ajoutés suivant l'évolution des besoins techniques et fonctionnels.

Équipes

Côté DGFiP :

- L'équipe a été constituée en mars-avril 2021 avec 2 ETP et avec une cible en 2022 de 5 ETP.
- Les travaux se font sur l'environnement Cloud interministériel de la DGFiP.
- La Mise En Production du DATALAKE cible de la DGFiP à base de TDP est prévue fin 2022 - début 2023

Côté EDF :

- Les travaux techniques ont démarré fin 2020 avec 2 ETP et une augmentation de ressources et de moyens financier est prévue en 2022.
- Les travaux de développement se font sur l'environnement virtuel des containers.
- **Le déploiement d'un POC sur l'environnement bare metal est prévu au 4ème trimestre de 2021.**

Membres

Contributeurs aux codes et binaires TDP :

- EDF depuis fin 2020
- DGFiP/Dtnum depuis mars 2021Les travaux se font sur l'environnement Cloud interministériel de la DGFiP.

Intéressés :

- NATIXIS : Souhaiterait contribuer aux travaux prochainement
- CNAM : Réflexions en cours sur les moyens financiers et ressources humaines
- DOUANE : Réflexions en cours sur les moyens financiers et ressources humaines
- ORANGE : Suit attentivement l'avancement de nos travaux
- OCDE : En cours d'adhésion au TOSiT afin de participer aux travaux
- SFR : Très intéressé par nos travaux et des discussions en cours au niveau de la direction
- Pôle Emploi I : Très intéressé par nos travaux et des discussions en cours au niveau de la direction