UPEC
UNIVERSITÉ PARIS-EST CRÉTEIL
iut
CRÉTEIL-VITRY
DÉPARTEMENT INFORMATIQUE

GROUP INFRASTRUCTURE PLATFORM

Rapport d'alternance BUT 3

BUT Informatique / IUT de Créteil – Vitry

2024/2025

Axelle Peenaert

Tuteur Entreprise : M. Rouinsard Antoine

Tuteur Pédagogique : M. Attal Ferhat

Remerciement

Avant d'entamer ce rapport, je tiens à exprimer ma profonde gratitude envers toutes les personnes qui ont rendu cette année d'alternance possible.

Je remercie tout d'abord mon tuteur de stage, M. Antoine Rouinsard, pour son accompagnement constant, ses conseils précieux et sa disponibilité tout au long de cette période.

Plus largement, je remercie chaleureusement l'ensemble de l'équipe « Plateformes et Référentiels » ainsi que l'équipe de la « Factory » de CA-GIP pour leur accueil, leur disponibilité, ainsi que pour les échanges enrichissants qui ont grandement contribué à mon apprentissage et au développement de mes compétences.

Je tiens également à remercier M. Ferhat Attal, mon tuteur pédagogique, pour son suivi durant cette année et pour ses conseils.

Je souhaite adresser mes sincères remerciements à toute l'équipe pédagogique du BUT Informatique de l'IUT de Créteil-Vitry. Grâce à la qualité de leur enseignement, j'ai pu acquérir des bases solides, tant théoriques que pratiques, qui m'ont permis d'aborder cette année d'alternance avec sérénité.

Enfin, je souhaite également remercier ma famille pour son soutien tout au long de cette période. Leur patience, leur bienveillance et leurs encouragements ont été d'une grande aide pour m'accompagner dans cette étape importante de mon parcours académique et professionnel.

Abstract

This report summarizes my third-year apprenticeship at Crédit Agricole Group Infrastructure Platform (CA-GIP). As a Data Engineer within the “Plateformes et Référentiels” team, I contributed to projects ensuring data quality, compliance, and accessibility.

CA-GIP, the IT subsidiary of Crédit Agricole, manages most of the group’s infrastructures. My team worked in an agile environment, focusing on data repositories, platform management, and internal IT services.

My missions included implementing a data quality control system to detect anomalies and automate alerts, building an HR reference system to centralize organizational information, and contributing to the “Dotation” project, which synchronized user and equipment data while detecting and justifying anomalies.

These experiences allowed me to strengthen my technical skills, while developing autonomy, problem-solving, and communication skills. This apprenticeship confirmed my interest in the data field.

Sommaire

I - Introduction

Dans le cadre de ma troisième année de BUT informatique, j'ai effectué une période d'apprentissage au sein de Crédit Agricole Group Infrastructure Platform (CA-GIP), où j'ai aussi eu l'opportunité de réaliser un stage de huit semaines, lors de la deuxième année du BUT.

CA-GIP est un acteur majeur de la transformation numérique du groupe Crédit Agricole. Durant cette année, j'ai intégré l'équipe Plateformes et Référentiels qui garantit la qualité, l'accessibilité et la sécurité des données en assurant leur gestion, leur intégration, leur conformité ainsi que le bon fonctionnement de leurs plateformes.

Ce rapport vise à présenter le contexte de mon année d'apprentissage, les missions réalisées, les compétences mobilisées et les enseignements tirés de cette expérience professionnelle.

Plusieurs raisons m'ont incité à choisir CA-GIP pour cette année d'apprentissage. Tout d'abord, intégrer un grand groupe comme le Crédit Agricole représente une opportunité exceptionnelle de développer mes compétences dans un environnement de travail dynamique et structuré. De plus, CA-GIP est reconnu pour son expertise en gestion des infrastructures informatiques et en développement de solutions innovantes, ce qui correspond parfaitement à mon ambition de me spécialiser dans les technologies de pointe et de participer à des projets à fort impact. Enfin, la diversité des projets menés par CA-GIP offre une occasion unique d'acquérir une expérience enrichissante.

Durant cette année, j'ai ainsi occupé le poste d'ingénieure Data au sein de l'équipe Plateformes et Référentiels rattachée à la direction du SI interne de l'entreprise. Mes missions principales ont consisté à :

- Réaliser les développements d'interfaces entre différents outils du SI Interne pour contribuer à l'urbanisation de celui-ci.

- Mettre en place et mettre à jour les référentiels d'entreprise en se basant sur les sources de données de confiance.

- Assister le Responsable de la plateforme data dans la mise en œuvre des évolutions et de la mise en conformité des outils (gestion de l'obsolescence notamment).

- Contribuer avec l'équipe sur des sujets innovants autour de la Data, de « l'APIsation » et de l'intelligence artificielle.

Intégrer une équipe spécialisée en data a été pour moi un véritable atout, car cela correspond pleinement à mes aspirations professionnelles. Cette expérience m'a ainsi permis de confirmer mon intérêt pour les métiers liés à la data science, ce qui m'a beaucoup aidé dans le choix de ma poursuite d'études.

Ce rapport se divise en deux grandes parties. Une première partie sera consacrée à la présentation de l'entreprise ainsi que de l'équipe où j'ai effectué mon apprentissage. Puis une seconde partie portera sur les différentes missions qui m'ont été confiées lors de cette année.

II - Présentation de l’entreprise et de l’équipe

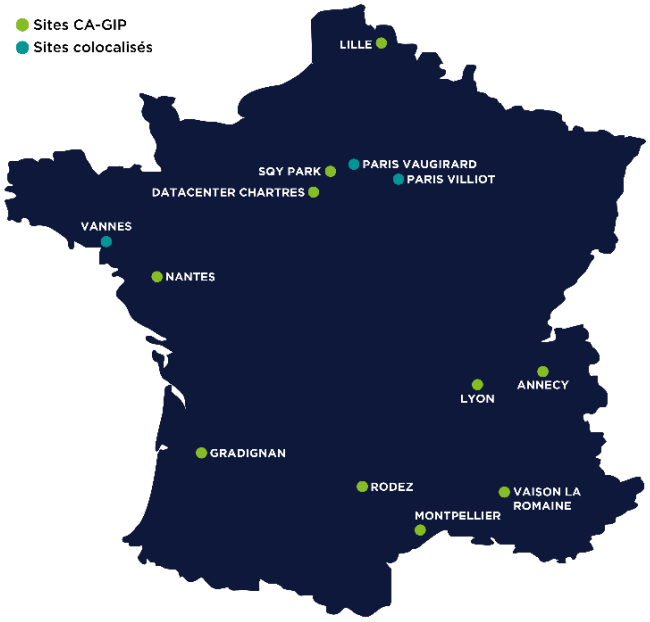
a) Présentation de l’entreprise

Crédit Agricole Group Infrastructure Platform (CA-GIP) est une filiale informatique du Crédit Agricole, ce qui signifie qu'elle est dirigée par sa société mère. Le Crédit Agricole est l'un des plus grands groupes bancaires et financiers en France et dans le monde. Fondé en 1894, il s'est initialement concentré sur le secteur agricole, mais a depuis diversifié ses services pour devenir un acteur majeur dans tous les domaines de la banque et de la finance.

CA-GIP regroupe depuis le 1er janvier 2019 toutes les activités de production informatique de toutes ses entités (Crédit Agricole Assurances, Crédit Agricole Corporate and Investment Bank, Crédit Agricole Technologies et Services ainsi que celui de l’entité SILCA). CA-GIP, avec ses 1,2 milliard d’euros de chiffres d’affaires lors de l’année 2023, regroupe 80% des productions informatiques et des infrastructures du groupe.

Son objectif est de répondre aux enjeux de la révolution digitale. En effet, CA-GIP a pour ambition de développer de nouvelles plateformes adaptées aux pratiques du digital, tout en garantissant un haut niveau de sécurité et de confidentialité. En bref, CA-GIP assure la gestion, la maintenance et l’évolution des systèmes d’information et des plateformes technologiques pour répondre aux besoins des différentes entités du groupe.

CA-GIP, c’est plus de 2000 collaborateurs répartis sur 10 sites en France, 6 Datacenters, 60 000 Serveurs Open (des serveurs dit « classique » comme Windows, linux…) et 6 Serveurs Mainframe (ordinateur puissant conçu pour traiter de grandes quantités de données et exécuter de nombreuses transactions simultanément).



(Figure 1 : Carte de France avec les différents sites)
Source : Document interne

Il existe deux types de sites. En vert, les sites CA-GIP, comme SQY Park qui se trouve à Saint-Quentin en Yvelines. Ces sites sont les sites classiques de l'entreprise où se trouve la majorité des employés. Les sites en bleu sont les sites colocalisés, c'est-à-dire qu'ils sont partagés avec les entités où les collaborateurs CA-GIP viennent travailler pour être au plus proches d'eux.

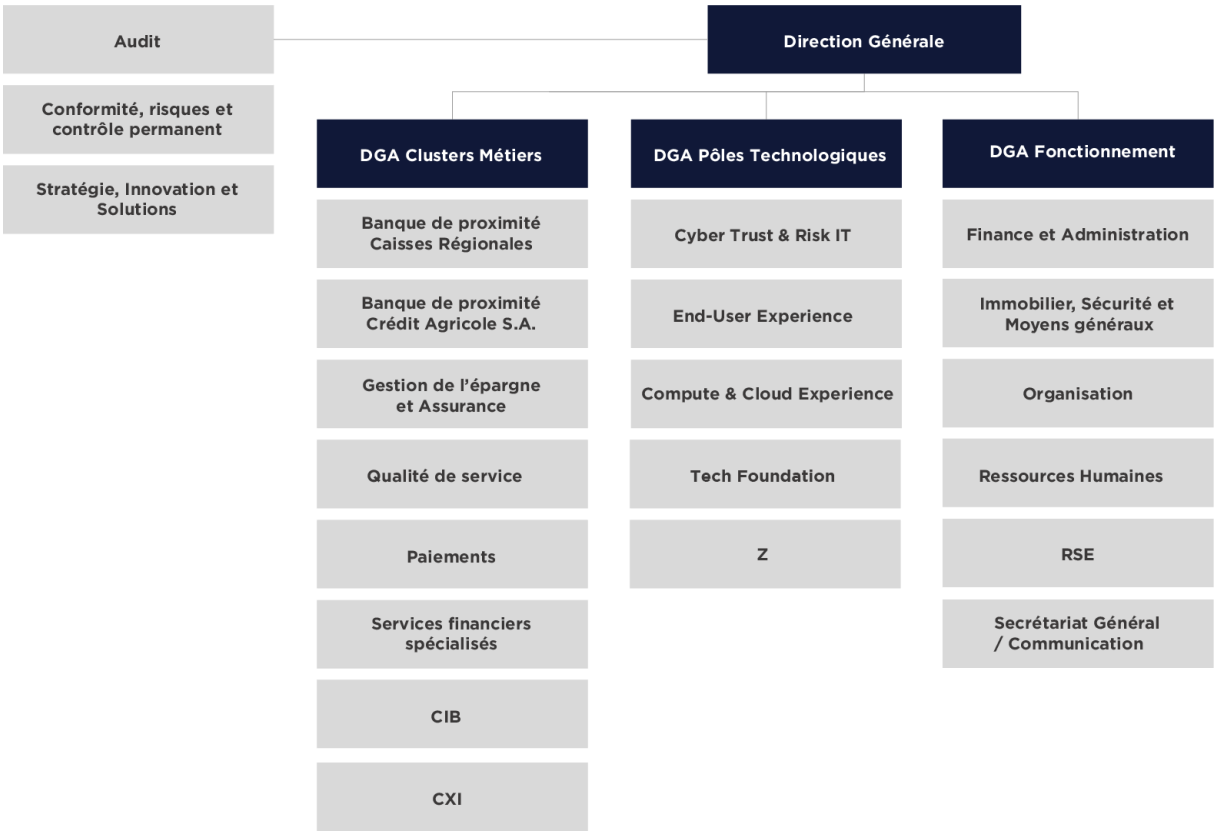
Le modèle opérationnel de CA-GIP est organisé en trois parties. Il y a les Clusters qui prennent en charge la production des applications métier en assurant une qualité de service et un maintien opérationnel au quotidien et qui accompagnent les Directeurs des Systèmes d'Information (DSI) et Entités dans l'implémentation des projets métier et dans leurs transformations.

On y retrouve ensuite les Pôles Technologiques qui portent les activités technologiques de CA-GIP, tant en fonctionnement qu'en construction et qui mutualisent les plateformes technologiques partagées à l'échelle du Groupe. On peut y retrouver 4 pôles :

- **Cyber Trust & Risk IT** qui joue un rôle transverse qui consiste à protéger les opérations et activités IT des Pôles Technologiques et à sécuriser les produits finaux.
- **End-User Expérience** qui vise principalement à fournir aux employés du Groupe un environnement de travail fiable, sécurisé, durable et compétitif.

- **Compute & Cloud Expérience** qui est dédié aux infrastructures supportant les applications et à leur cycle de vie.
- **Tech Foundation** qui fournit les fondations d'hébergement, de connectivité ainsi que les services transverses essentiels pour l'ensemble du Groupe.

Pour terminer, il y a la DGA Fonctionnement où l'on retrouve les fonctions transverses (la communication, les ressources humaines, le pôle juridique, la finance, le service client, les achats, l'immobilier et l'organisation). L'équipe que j'ai intégré faisait partie cette DGA, plus particulièrement du SI Interne au sein de l'Organisation.



(Figure 2 : Organigramme de CA-GIP)
Source : Document interne

CA-GIP n'oublie pas non plus d'affirmer et de démontrer leur fort engagement RSE. La RSE (responsabilité sociétale des entreprises) désigne la contribution des entreprises aux enjeux du développement durable mais aussi le fait d'avoir un impact positif sur la société tout en étant économiquement viable. Par exemple, entre 2023 et 2024, CA-GIP à fait baisser de 5,7 % son facteur

d'émissions carbone. CA-GIP souhaite devenir une entreprise durable de référence en plaçant le Green IT au cœur de son modèle et en intégrant une dimension sociétale dans l'ADN de tous ses projets. Même si l'IT génère des impacts négatifs, l'entreprise peut jouer un rôle moteur et facilitateur dans la décarbonation en intégrant la sobriété numérique.

C'est pourquoi l'entreprise est engagée avec détermination dans la sensibilisation aux enjeux du numérique responsable et dans l'écoconception by design de ses produits et services.

La stratégie RSE de l'entité se décline en 4 axes :

a. La mise en œuvre d'un dispositif global pour gérer et piloter l'emprunte environnementale sur lequel j'ai eu la chance de travailler durant cette année.

b. L'éco-conception des plateformes et des produits CA-GIP.

c. L'accompagnement des clients dans leur décarbonation (stimuler, encourager et guider).

d. La mise en place d'un environnement de travail respectueux.

En ce qui concerne sa politique sociale, CA-GIP est signataire de la Charte de la diversité en faveur des diversités et de l'inclusion depuis 2023. Cette politique vise à renforcer les diversités dans tous les métiers et à tous les niveaux de l'organisation en favorisant l'inclusion. Au-delà de la mixité, la politique des diversités de Crédit Agricole S.A. entend s'adresser à toutes les diversités (genres, âges, handicaps, orientations sexuelles, ethnies, origines socioculturelles etc.) par des initiatives à l'échelle du Groupe et/ ou des métiers en s'appuyant sur 5 engagements :

- Égalité des chances
- Ouverture et curiosité
- Représentativité
- Solidarité
- Responsabilité

Pour respecter son engagement, le groupe a mis en place diverses actions. Pour commencer, le groupe a mis en place un plan jeunesse pour favoriser l'insertion professionnelle des jeunes en y accueillant et en formant 50 000 jeunes d'ici 2025, notamment via l'alternance et les stages, que ce soit au sein des banques de proximité et des métiers du Groupe en France ou à l'international. Ils ont aussi développé un programme de Mentorat. Par ce dispositif, le Groupe renforce depuis 6 ans la diversité et favorise le

développement de carrières équitables et l'émergence de leaders responsables. Par la mise en relation avec des dirigeants mentors, le Groupe offre aux mentorés l'opportunité d'acquérir de nouvelles compétences, de développer leur réseau et de bénéficier de conseils sur leur carrière. Pour finir, ils ont créé un mécénat de compétences qui est un programme d'engagement permettant aux personnes travaillant chez CA-GIP de se mobiliser en externes en donnant de leur temps à des organisations à but non lucratif à l'instar d'associations qui œuvrent en faveur de l'égalité des chances comme Chemins d'avenirs et l'Institut de l'Engagement.

Crédit Agricole Group Infrastructure Platform a pour vocation de devenir un pôle technologique de 1er plan en Europe, attractif pour les collaborateurs, les nouveaux talents et les partenaires technologiques.

b) Présentation de l'équipe

Lors de mon apprentissage, j'ai été accueillie dans l'équipe IT Core au sein du SI Interne de CA-GIP. Cette équipe regroupe 20 collaborateurs qui sont répartis sur 2 sites (Saint Quentin en Yvelines et Annecy). Cette équipe comprend la Factory spécialisée dans le développement et le maintien en conditions opérationnelles d'applications et de solutions de visualisation des données. La Factory est donc un centre d'expertise constitué d'experts du développement et de la BI.

Enfin, il y a l'équipe Plateforme et Référentiels, que j'ai intégré. Ces missions sont de :

- **Consolider les référentiels de l'entreprise** : En assurant l'identification, la centralisation et la fiabilité des référentiels.
- **Cartographier des données** : En participant à la constitution d'une cartographie des données de référence de l'entreprise.
- **Gérer des plateformes** : En Développant, maintenant et optimisant les plateformes technologiques nécessaires aux opérations et aux services numériques pour la Factory et le reste du SI Interne.
- **Intégrer des données** : En Facilitant l'intégration et l'interopérabilité des données entre les différents systèmes d'information et leur mise à disposition aussi bien pour les besoins de reporting que de développement d'applications métier.

- **Sécuriser et conformiser** : En Garantissant la sécurisation des données et la conformité avec les réglementations en vigueur à chaque étape du traitement.
- **Assurer le support et la maintenance** : En Fournissant un support technique et une maintenance continue pour assurer la disponibilité et la performance des plateformes.

L'équipe réalise ces tâches pour un seul et même client, CA-GIP (socles, cluster et fonctions transverses).

Plateformes et Référentiels est composée d'une part de 4 personnes internes à l'entreprise qui ne travaillent que pour CA-GIP, et d'autre part de 4 externes qui travaillent pour une autre société et qui ont un contrat de prestation avec CA-GIP pour différentes missions.

Par ailleurs, l'équipe est composée d'une Responsable d'activité. La Responsable d'activité est chargée de gérer le backlog, qui est une liste priorisée de fonctionnalités, d'améliorations et de corrections de bugs à développer. Elle doit définir et communiquer clairement les éléments du backlog, s'assurer que l'équipe comprend bien ce qui est attendu, et prioriser ces éléments. De plus, la Responsable d'activité collabore étroitement avec l'équipe de développement pour affiner les exigences, répondre aux questions et fournir des retours rapides sur le travail en cours. Dans mon équipe, elle est aussi chargée d'organiser et d'animer les réunions clés, comme les réunions hebdomadaires, les revues de sprint, les rétrospectives et les planifications de sprint. Elle aide également l'équipe à surmonter les obstacles et les défis qui pourraient freiner l'avancement du travail, ce qui peut impliquer la résolution de conflits internes ou la coordination avec d'autres équipes.

Ensuite, on y retrouve un Lead Dev qui supervise et coordonne les activités de développement. Un Lead Dev ETL (Extract, Transform, Load) est un professionnel spécialisé dans la gestion des données et l'intégration de systèmes. Ce rôle implique la conception et le développement de processus ETL qui consistent à extraire les données de diverses sources, les transformer pour répondre aux besoins de l'entreprise, puis les charger dans des bases de données. Il supervise les projets ETL, coordonne les équipes de développeurs et s'assure que les projets sont livrés dans les délais et les budgets impartis.

On peut y retrouver également des ingénieurs data. Le rôle des ingénieurs data est de concevoir, construire et maintenir les infrastructures nécessaires au traitement des données. Ils développent des pipelines permettant de collecter, nettoyer, structurer et stocker les données provenant de plusieurs sources. Ils

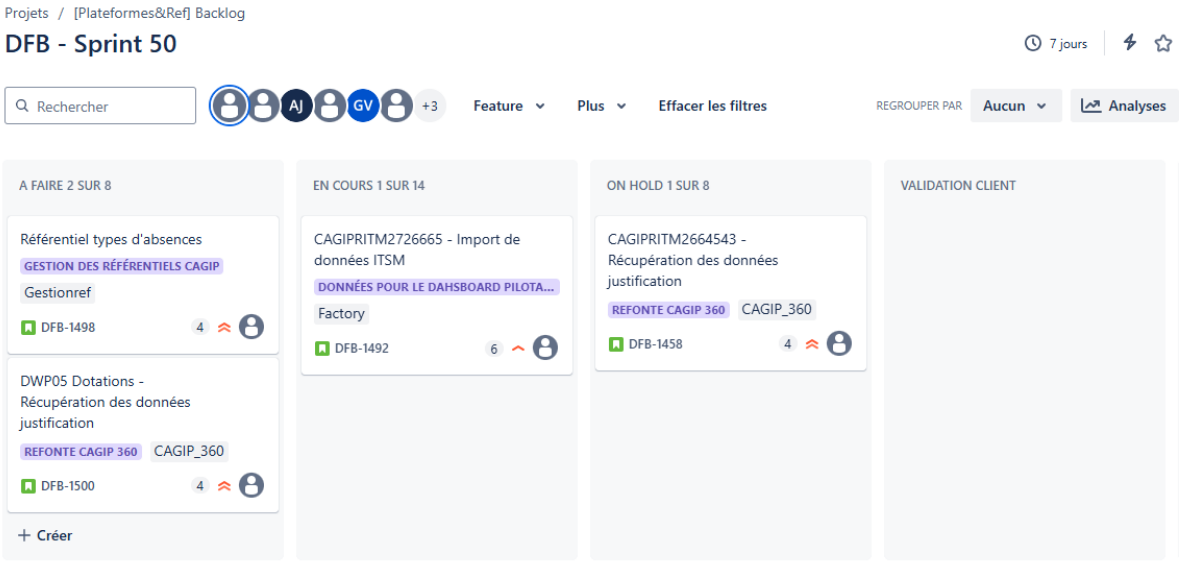
veillent à la qualité, à la performance et à la fiabilité des systèmes de données, tout en rendant les informations accessibles.

De plus, l'équipe compte un architecte data qui accompagne dans les choix des briques techniques de la data. Il accompagne aussi les lead Dev dans l'architecture et l'optimisation des flux de données, et donne des conseils sur les évolutions possibles de la data platform.

Pour finir, il y a dans l'équipe un AppOps. Il est responsable du bon fonctionnement, de la surveillance et de la maintenance des applications en production. Son rôle consiste notamment à déployer les nouvelles versions, détecter et résoudre les incidents, automatiser les opérations répétitives et collaborer avec les équipes de développement pour assurer une mise en production fluide.

L'équipe fonctionne avec la méthode Agile. La méthode Agile est une méthode de gestion de projet qui consiste à décomposer les projets en une suite de petits objectifs. Elle vise à fractionner les étapes de développement en cycles courts et à livrer des versions intermédiaires du produit. Lorsque la méthode Agile est utilisée, l'équipe travaille en petits cycles courts que l'on appelle sprint. Dans l'équipe dans laquelle je me trouvais, les sprints duraient 2 semaines. Pendant chaque sprint, il y avait deux réunions : une le premier jour du sprint (toujours un lundi) qui permettait de faire le point sur le sprint précédant, voir si les objectifs ont été atteints ou non. Cette réunion était aussi l'occasion de définir la répartition des tâches pour le prochain sprint. La deuxième réunion, quant à elle, se déroulait le lundi suivant et permettait de faire le point sur l'avancement de chacun, régler des problèmes s'il y en avait, aider une personne de l'équipe si besoin…

Pour attribuer les tâches, l'équipe utilise le logiciel Jira. Ce dernier est un outil de gestion de projet pour les équipes qui utilisent la méthode Agile. Il permet de planifier, suivre et gérer leurs projets. Il permet aussi de créer des tâches, de les attribuer aux membres de l'équipe et de suivre l'avancement du travail. Pour rendre cela plus parlant, voici une capture d'écran de mon tableau Jira lors du sprint 50.



(Figure 3 : Interface du logiciel Jira) Source : CdA

Dans chaque case, nous pouvons apercevoir différentes cartes que nous appelons ticket. Chaque ticket représente une tâche à effectuer. En fonction de notre progression, nous pouvons le déplacer dans les différents états, ce qui permet à tout le monde de voir notre avancement. Cela permet aussi d'avoir différentes informations sur le contexte de la tâche à effectuer (le temps que la carte est censée nous prendre, son niveau de priorité et de complexité…)

Il y avait également une autre réunion qui se déroulait chaque vendredi à laquelle j'étais conviée et qui réunissait toutes l'équipe IT Core (la Factory + Plateformes et Référentiels). Cette réunion était plutôt informative et portée davantage sur le côté « management » de l'équipe. En effet durant cette réunion, plusieurs sujets étaient abordés, notamment des sujets à propos de l'équipe. Tout le monde pouvait y prendre la parole et ainsi évoquer un problème qui a pu être rencontré. Ces réunions permettaient de réunir toute l'équipe, dont une partie est basée à Annecy, favorisant ainsi les échanges et renforçant la cohésion malgré la distance.

III – Les missions

Dans cette partie, je vais vous présenter quelques exemples des différentes missions qui ont pu m'être confiées lors de cette année. Durant mon alternance, j'ai eu l'opportunité d'être pleinement intégrée à mon équipe et de travailler sur des missions similaires à celles confiées aux ingénieurs data. Cette immersion m'a permis de développer de nombreuses compétences techniques, que je vais vous développer dans cette partie, mais aussi des compétences transversales telles que l'autonomie, la rigueur, la capacité d'adaptation, la gestion des priorités ainsi que l'esprit d'analyse. J'ai également renforcé mes qualités relationnelle grâce au travail d'équipe, à la communication régulière avec les équipes métier. J'ai participé à des projets stratégiques suivis de très près par la direction de l'entreprise.

L'ensemble des traitements de données que vous allez voir dans ce rapport ont été réalisés sur Dataiku. Dataiku est une plateforme de science des données qui met l'accent sur la préparation et le nettoyage des données, des étapes cruciales dans le processus d'analyse des données. Cette plateforme offre une gamme d'outils avancés pour faciliter la transformation des données brutes en informations exploitables. Grâce à Dataiku, les utilisateurs bénéficient de fonctionnalités puissantes pour nettoyer et préparer les données, en éliminant les valeurs non utiles, en gérant les données manquantes et en détectant les erreurs. Dataiku permet aux utilisateurs de manipuler facilement les données selon leurs besoins spécifiques, tout en offrant une visualisation en temps réel des transformations effectuées.

a) Data Quality

Cette mission avait pour objectif de mettre en place un système de contrôle de la qualité des données provenant de deux sources : le DP02 et Triskell.

Le DP02 est le référentiel central des prestataires. Il contient un ensemble d'informations administratives et contractuelles telles que le nom, le prénom, le type de contrat, la durée de présence ou encore la fonction occupée. D'autre part, Triskell est un outil de gestion des temps et des activités, utilisé

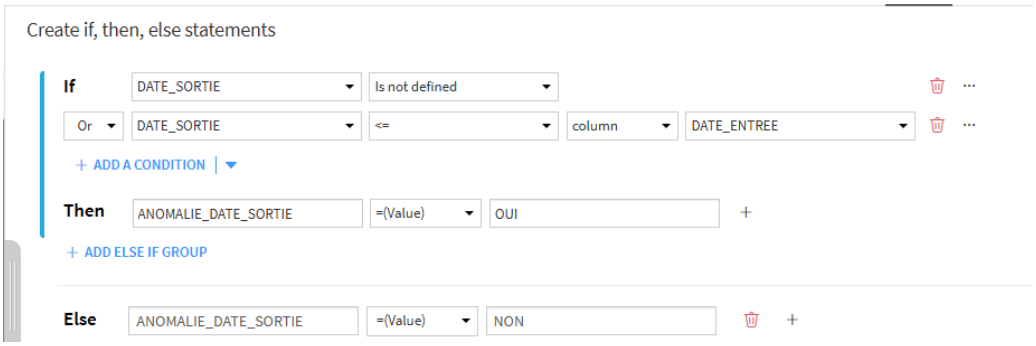
notamment pour la saisie des temps, le suivi des projets et la gestion des portefeuilles. Il contient également des informations sur les prestataires.

L'objectif était donc d'analyser les données issues de ces deux sources, d'y détecter des anomalies, et d'automatiser un processus d'alerte auprès des équipes concernées afin de passer d'un traitement manuel à un traitement automatique mais aussi d'améliorer le temps de résolution des anomalies.

J'ai commencé par vérifier la qualité des données du DP02. Pour ce faire, on m'a transmis un ensemble de règles métier à appliquer afin d'identifier les éventuelles incohérences :

- Si un IUG (Identifiant Unique Groupe : identifiant unique des internes) est renseigné, alors anomalie car il s'agit d'une information réservée aux internes.
- Le CGB (Centre de Gestion Budgétaire : identifiant unique des équipes) n'existe pas dans le référentiel de l'organisation, ou est inactif dans celui-ci, alors anomalie.
- Si un doublon d'identifiants IAM (outil de gestion des identités) est détecté, alors anomalie.
- Si l'adresse email ne finit pas par «-prestataire@ca-gip.fr », alors anomalie.
- Si la « DATE_SORTIE » est vide ou si elle est inférieure ou égale à la « DATE_ENTREE », alors anomalie.
- Si un collaborateur est comptabilisé comme étant manager alors qu'il ne fait pas partie de la liste des managers du référentiel des internes pour ce CGB, ou s'il n'y a pas de manager, alors anomalie.

Pour détecter ces anomalies j'ai utilisé la recette « prepared » de dataiku. Cette recette permet de nettoyer, transformer, enrichir les données de manière visuelle avec une succession d'étapes. Pour chaque règle, j'ai utilisé l'étape conditionnelle « if, then, else », qui permet de créer une colonne signalant si la condition est remplie ou non.



(Figure 4 : Visuel de l'étape if, then else du traitement) Source : CdA

Ce qui donne à la fin du traitement :

DATE_ENTREE string Date (unparsed)	DATE_SORTIE string Date (unparsed)	ANOMALIE_DATE_SORTIE string Text
2024-04-01 00:00:00.000		OUI
2024-07-01 00:00:00.000	2023-12-31 00:00:00.000	OUI
2024-08-19 00:00:00.000	2025-12-31 00:00:00.000	NON
2023-05-02 00:00:00.000	2025-12-31 00:00:00.000	NON
2023-12-07 00:00:00.000	2025-06-30 00:00:00.000	NON
2024-01-02 00:00:00.000	2025-12-31 00:00:00.000	NON
2023-10-02 00:00:00.000	2025-12-31 00:00:00.000	NON
2024-10-21 00:00:00.000	2025-12-31 00:00:00.000	NON

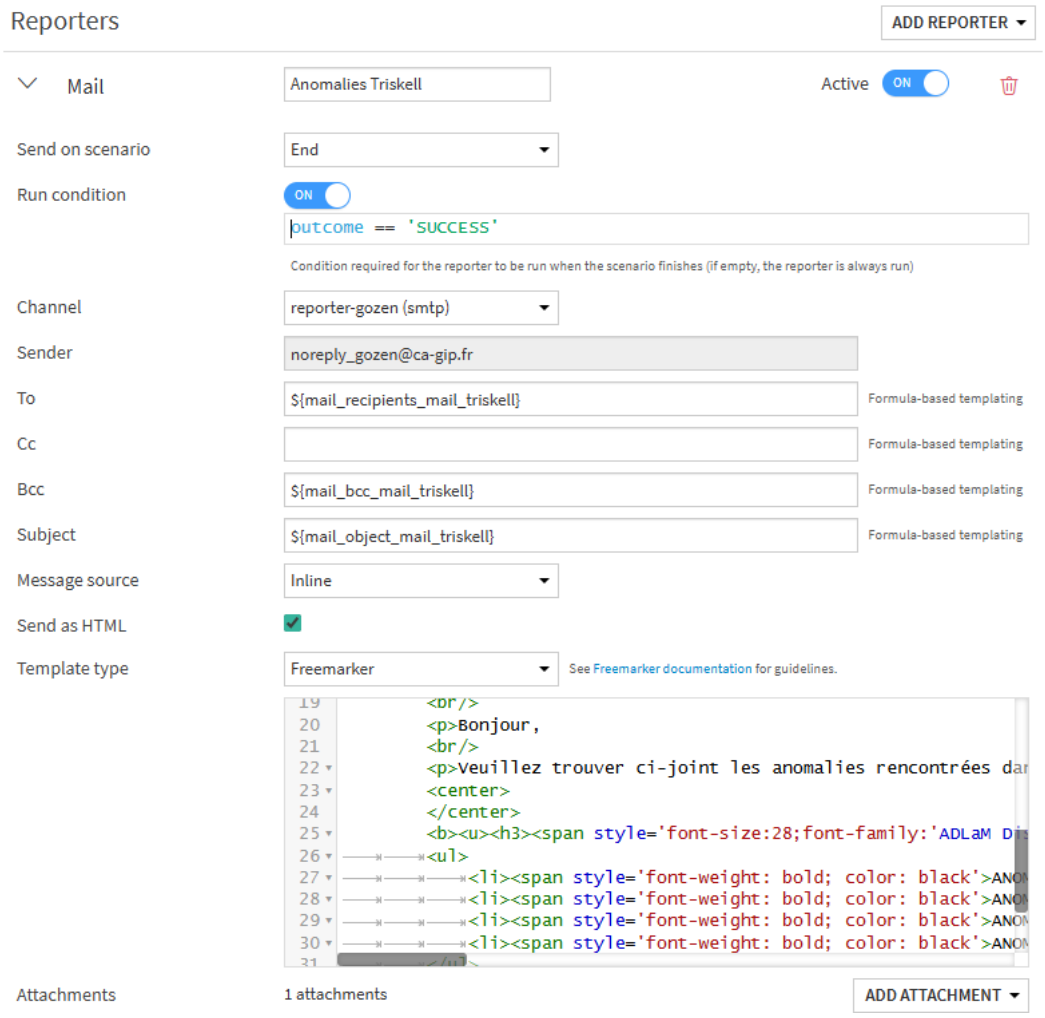
(Figure 5 : Résultat de l'étape if, then else du traitement) Source : CdA

Après avoir traité le DP02, j'ai effectué le même travail sur les données de Triskell. Les règles d'identification des anomalies étaient similaires dans leur logique, mais adaptées à l'outil :

- Une anomalie est identifiée si l'équipe des prestataires renseignés dans Triskell est vide ou n'est pas la même que l'équipe réelle déclarée dans le DP02.
- Une anomalie est identifiée si le matricule des prestataires dans Stan (outil de gestion des identités et des habilitations) n'est pas identique à celui dans Triskell.
- Une anomalie est identifiée si un prestataire en forfait (type de contrat) détient un compte de saisie des temps Triskell car un prestataire en forfait est payé pour un résultat global (objectif) et non en fonction du temps passé dans l'entreprise.
- Une anomalie est identifiée si la date de fin du contrat des prestataires n'est pas la même que la date de sortie dans Triskell.

L'ensemble de ces règles a été implémenté via une seconde recette de préparation, en suivant la même démarche que pour le DP02.

Une fois les anomalies détectées, il restait à définir une manière de prévenir les personnes concernées de façon automatisée. Nous avons décidé de mettre en place un système d'alerte par mail avec le descriptif et le nombre d'anomalie. Pour automatiser l'envoi de ces mails, nous avons utilisé une fonctionnalité sur Dataiku qui permet d'envoyer un mail quand le scénario (partie qui permet d'automatiser les projets) se lance. Voilà à quoi cela ressemble :



(Figure 6 : Création du mail dans Dataiku Source : CdA

J'ai rempli les différents paramètres. Les champs commençant par des $ sont des variables. Elles se trouvent dans l'onglet variable de Dataiku. Cela permet de déclarer les variables (global et local) afin qu'en cas de changement, cela soit plus simple à modifier.

Le mail est écrit en HTML afin de lui donner la forme que nous souhaitons. Voici le résultat final d'un des mails :

CONTRÔLE DE QUALITÉ DONNÉES
TRISKELL - Ressources Externes

Bonjour,

Veuillez trouver ci-joint les anomalies rencontrées dans le référentiel des externes. Nous comptabilisons **4** ressource(s) en anomalie(s) sur les **1742** ressource(s) présente(s) à date chez CA-GIP.

Les contrôles effectués sur le référentiel sont :

- **ANOMALIE_CGB** : Une anomalie est identifiée si l'équipe des prestataires renseignée dans Triskell est vide ou n'est pas la même que l'équipe réelle déclarée dans le DP02.
- **ANOMALIE_MATRICULE** : Une anomalie est identifiée si le matricule des prestataires dans Stan n'est pas identique à celui dans Triskell.
- **ANOMALIE_COMPTE_SAISIE** : Une anomalie est identifiée si un prestataire en forfait détient un compte de saisie des temps Triskell.
- **ANOMALIE_DATE_SORTIE** : Une anomalie est identifiée si la date de fin du contrat des prestataires n'est pas la même que la date de sortie dans Triskell

Si vous avez des questions ou des remarques, n'hésitez pas à contacter
CAGIP_DDO_Data_Factory@ca-gip.fr

Equipe IT-Core

(Figure 7 : Résultat du mail envoyé aux équipes) Source : CdA

Une fois l'outil mis en place en pré-production (espace de test qui reproduit à l'identique l'environnement de production afin de valider le flux avant sa mise en ligne), j'ai organisé des échanges avec les interlocuteurs métier responsables du DP02 et de Triskell. Ces entretiens m'ont permis de leur présenter le fonctionnement du dispositif, de recueillir leurs remarques et d'adapter la fréquence d'envoi des mails :

- Une équipe a demandé à recevoir un rapport hebdomadaire, en début de semaine ;
- L'autre a préféré un envoi quotidien, mais seulement si des anomalies sont détectées.

Pour ne pas envoyer de mail inutile dans le cas où aucune anomalie n'est détectée, j'ai ajouté un contrôle python dans le scénario, qui vérifie si le dataset d'anomalies est vide. Si c'est la cas, l'envoi du mail est automatiquement annulé :

```
import dataiku

dataset = dataiku.Dataset("Anomalies_Triskell") #On récupère le dataset
dataframe = dataset.get_dataframe() #On convertit le dataset en dataframe pandas

if(len(dataframe.index) == 0):
    project = dataiku.api_client().get_project(dataiku.default_project_key())
    scenario = project.list_scenarios(as_type="objects")[0]
    scenario.abort() # Si le dataset est vide, on récupère le scénario du projet et on l'arrête
```

Cette mission m'a permis d'approfondir mes compétences sur la gestion de la qualité des données et de l'automatisation dans Dataiku. Elle m'a également donné l'occasion de dialoguer avec des équipes métiers, de prendre en compte leurs besoins, et de livrer un outil concret, prêt à l'emploi.

b) Référentiel RH

Une autre mission majeure de mon alternance a été la participation à la construction de référentiels métiers, en particulier à destination de notre DGA. Ces référentiels ont pour objectif de centraliser des données fiables, claires et à jour, pour alimenter des tableaux de bord ou des applications internes, tout en rendant l'information accessible et compréhensible par les utilisateurs métiers.

Le référentiel que j'ai développé permet de représenter l'organisation RH, en identifiant les Responsables RH (RRH), Gestionnaires RH (GRH), Assistantes RH (ARH) et Gestionnaires de Paie (GP) responsables de chaque périmètre. Il est destiné principalement aux équipes RH, qui pourront ainsi retrouver facilement les interlocuteurs dédiés à chaque équipe ou service.

La première étape a été la construction du flux sur Dataiku pour récupérer les informations demandées depuis plusieurs sources :

- Référentiel GRH orga : est un fichier déposé régulièrement sur un partage réseau, contenant les correspondances entre les équipes et les RRH, GRH, ARH et GP associés. Le fait qu'il soit mis à jour par le service RH permet de garantir l'exactitude des données utilisées. On le récupère avec un script Python
- DP07 : est une table regroupant les informations personnelles et professionnelles de l'ensemble des collaborateurs internes (nom, prénom, âge, équipe, métier…).
- IN03 : référentiel de l'organisation de l'entreprise (liste des équipes avec leurs directions, départements, services…).

Une fois les sources de données récupérées, j'ai pu faire une première jointure entre le fichier référentiel grh orga et le dp07, ce qui m'a permis de récupérer des informations supplémentaires sur les RRH, GRH, ARH et GP (Nom, Prénom et Adresse mail) car ces informations ne sont pas présentes sur le fichier issu du partage réseaux (nous n'avons que l'IUG qui est un identifiant unique décerné à tous les internes).

Une seconde jointure avec la table IN03 a ensuite permis d'enrichir les données avec les informations organisationnelles telles que la direction, le service, le département, ou encore l'identifiant de l'équipe.

J'ai ensuite fait différentes recettes « prepared » afin de nettoyer les données et les remettre dans la forme attendue.

Pour garantir une mise à jour quotidienne du référentiel, j'ai mis en place un scénario Dataiku permettant de lancer automatiquement le flux chaque jour à heure fixe. Cela permet d'avoir une donnée toujours à jours, sans intervention manuelle.

Afin de valider la conformité du référentiel aux besoins métier, j'ai organisé un échange avec l'interlocutrice RH pour ce projet. Cet entretien a permis d'expliquer en détail la construction du référentiel, de recueillir ses retours, et de prendre en compte certaines demandes de modifications telles que :

- Changer le nom de certains champs.
- Supprimer certaines colonnes devenues inutiles pour ce référentiel.
- Réorganiser les colonnes pour respecter un ordre logique.

Une fois les ajustements effectués, le fichier lui a été renvoyé pour validation, avant la mise en production.

La mise en production du projet a ensuite été réalisée. Une dernière étape a été ajoutée, codée en python, permettant d'enregistrer automatiquement le fichier final sur le partage réseau, afin qu'il soit accessible ailleurs que dans la base de données et plus facilement accessible par le service RH. Les jours suivant la MEP, il a fallu surveiller l'exécution du projet pour assurer qu'aucune erreur ne survenait, et que le fichier se mettait bien à jour comme prévu.

Voici un extrait de référentiel CH00 :

Equipe_Code	Equipe_Label	Date_debut_de_validité	Date_fin_de_validité	Date_derniere_modificati	GRH_IUG	GRH_NOM	GRH_PRENOM	GRH_EMAIL
C3093	Achats	2020-01-01	2099-01-01	2025-01-13				
C3105	Juridique & Corporate	2020-01-01	2099-01-01	2025-01-13				
R3001	DGA Fonctionnement	2022-01-01	2099-01-01	2024-01-09				
A3004	Secrétariat Général et Communication	2023-05-01	2099-01-01	2024-01-09				
T3012	Communication	2020-01-01	2099-01-01	2024-01-09				
T3007	RSE	2022-01-01	2099-01-01	2024-01-09				
A3003	Développement et projets RH	2023-05-01	2099-01-01	2024-01-09				
A3005	Capital Humain	2023-05-01	2099-01-01	2024-01-09				
R3016	Frais déplacements élus	2024-01-09	2099-01-01	2024-01-09				
R3020	Frais de déplacement pour les formations	2023-12-21	2099-01-01	2024-01-09				
R3021	CSE CA-GIP	2024-01-09	2099-01-01	2024-01-09				
R3022	JOCA2024	2024-04-03	2099-01-01	2024-04-03				
T3006	Ressources Humaines	2022-01-01	2099-01-01	2024-01-09				
T3008	Dialogue Social et QVT	2022-01-01	2099-01-01	2024-01-09				
T3014	Performance Sociale et Data RH	2020-01-01	2099-01-01	2024-12-05				
C3001	DGA Clusters	2020-01-01	2099-01-01	2020-01-01				
C3211	Programme Transfo Cluster	2024-02-22	2099-01-01	2024-02-22				
C3021	Cluster Banque de Proximité CASA	2020-01-01	2099-01-01	2025-03-31				
C3168	Pôle WorkPlace et Services	2023-02-01	2099-01-01	2023-02-01				
C3169	Chapitre des Tribus et Relations Entités	2023-02-01	2099-01-01	2025-05-12				
C3170	Tribu Crédits et Paiement	2023-02-01	2099-01-01	2025-05-12				
C3173	Tribu Data	2023-02-01	2099-01-01	2025-05-12				

(Figure 8 : Extrait du référentiel CH00) Source : CdA

Enfin, une dernière évolution du projet a consisté à intégrer une historisation du référentiel, afin de garder une trace des différents RH rattachés aux équipes au fil du temps.

Pour cela, j'ai :

1. Ajouté une colonne PERIODE dans la dernière recette « Prepared », pour dater chaque ligne à la date du traitement ;
2. Ajouté un traitement post-écriture en SQL permettant d'éliminer les doublons, tout en conservant uniquement la version la plus récente pour chaque combinaison équipe + période.

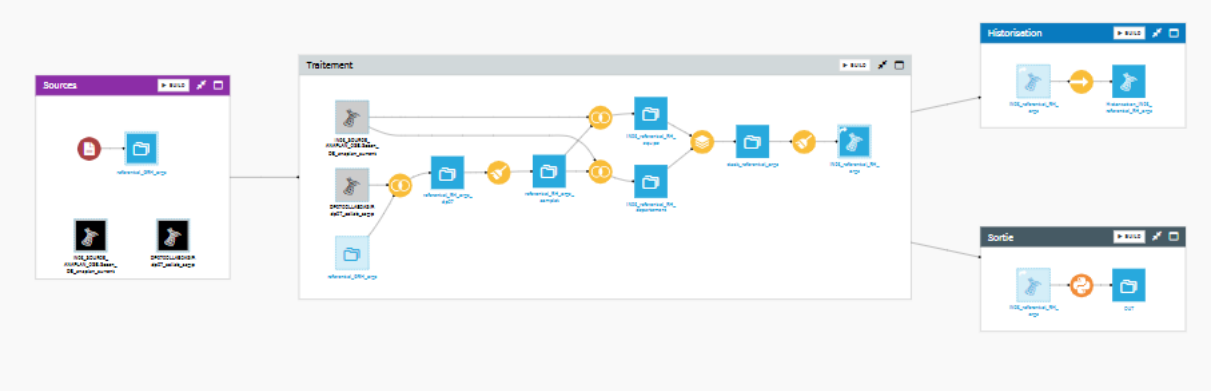
Voici la requête utilisée :

```
DELETE s1 FROM "Historisation_IN03_referentiel_RH_orga" s1
WHERE TIMEST < (SELECT MAX(TIMEST) FROM "Historisation_IN03_referentiel_RH_orga" s2 WHERE
s1.Equipe_Code = s2.Equipe_Code AND s1."PERIODE" = s2."PERIODE") ;
```

Une fois ce traitement terminé, j'ai synchronisé les données vers la base de données où est stockée la table.
Pour permettre une historisation incrémentale, j'ai coché l'option « Append instead of overwrite », ce qui permet d'ajouter uniquement les nouvelles lignes

sans écraser les anciennes. Ainsi, il est possible de suivre l'évolution de l'organisation RH et les changements de rattachement d'une équipe à une autre dans le temps.

Pour faciliter la compréhension du processus, vous pouvez vous référer à cette image qui représente le traitement des données du début à la fin :



(Figure 9 : Flow dataiku du projet CH00) Source : CdA

Zone voilette : Récupération des données

- Cette zone contient toutes les sources de données utilisées.

Zone grise : Traitement

- Pastille jaune avec un pinceau : action de nettoyage (suppression de colonnes, concaténation de colonnes, renommer une colonne…).
- Pastille jaune avec deux ronds : jointure.
- Pastille jaune avec plusieurs carrés superposés : fusionner différentes tables.

Zone bleue : Historisation

- Pastille jaune avec une flèche : synchronisation avec une base de données.

Zone grise : Sortie

- Recette python : permet de transférer la table en fichier Excel sur le partage réseaux.

Cette mission m'a permis d'aborder plusieurs aspects essentiels de la gestion de référentiels : récupération de données, transformation, automatisation, validation métier, historisation et mise en production.

Elle m'a également permis de mieux comprendre les besoins concrets des équipes RH, et de produire un outil durable, fiable et directement exploitable par les utilisateurs finaux.

c) Dotation

Cette dernière mission que je vais vous présenter est sans doute celle qui m'a prise le plus de temps durant cette année. Le projet « DWP05 – Outil de dotations » a pour objectif de synchroniser mensuellement les informations sur les utilisateurs et leurs dotations (postes de travail et mobiles), en détectant les anomalies telles que des dotations attribuées à des utilisateurs inactifs ou partis.

J'ai d'abord travaillé sur la partie du flux dédiée à la récupération des données utilisateurs. L'objectif était de ne conserver que les collaborateurs actuellement présents ou ceux ayant quitté l'entreprise depuis moins d'un mois. J'ai donc commencé par faire une jointure pour récupérer l'ensemble des informations puis fait une recette « prepared » pour mettre en ordre les données.

Dans un second temps, j'ai intégré la partie concernant les données sur les postes de travail. Les spécifications des « devices » avaient été définies en amont dans un fichier Excel, ce qui m'a permis de rédiger une requête SQL pour extraire les informations nécessaires. Après la validation d'un membre de mon équipe, cette requête a été intégrée au flux dans Dataiku.

Afin de compléter les données sur les postes de travail, j'ai dû rajouter les données liées à la télécollecte (informations sur les postes remontant automatiquement vers des bases de données. Par exemple : dernière date de connexion, l'adresse IP, le modèle, la RAM, la capacité totale du disque…). J'ai ensuite procédé à la jointure entre ces données et celles déjà existantes.

Au cours de cette étape plusieurs anomalies sont apparues : champs manquants, données incomplètes ou incohérentes (par exemple des lignes sans date de rafraîchissement, ou avec deux colonnes email différentes sans indication claire sur celle à utiliser). Ces constats ont été remontés aux équipes qui nous ont fournies les données afin d'avoir une solution. Après validation des choix, j'ai réalisé les corrections nécessaires.

Pour finir, j'ai ensuite récupéré les données sur les mobiles depuis un partage réseaux, puis j'ai modifié les données dans une recette « prepared » afin de ne garder que les données intéressantes et de renommer les colonnes pour qu'elles aient un nom plus parlant.

J'ai ensuite exporté tous ces fichiers vers un SharePoint afin que l'équipe de développement puisse y accéder, pour ce faire j'ai utilisé un script python :

```
from gozenlib import sharepoint_api as spa
import dataiku
import pandas as pd
from datetime import datetime
import json
dtk_vars = dataiku.get_custom_variables()


spm = spa.SharePointManager(
    tenant_id=dtk_vars['sharepoint_tenant_id'],
    client_id=dtk_vars['sharepoint_client_id'],
    client_secret=dtk_vars['sharepoint_client_secret'],
    server_relative_url=dtk_vars['server_relative_url'],
    proxies={'https': dtk_vars['proxy_cagip']}
)

file_path = json.loads(dtk_vars['sp_file_path'])
file_type = dtk_vars['sp_file_type']


for dataset_name, file in file_path.items():
    logging.info(f"{dataset_name=}")
    ds = dataiku.Dataset(dataset_name)
    df = ds.get_dataframe(infer_with_pandas=False, keep_default_na=True)
    if len(df) > 0:
        retour = spm.upload_df_to_file(file, df, file_type=file_type,
sheet_name='Tableau1')
    else:
        logging.info("Le référentiel est vide.")
```

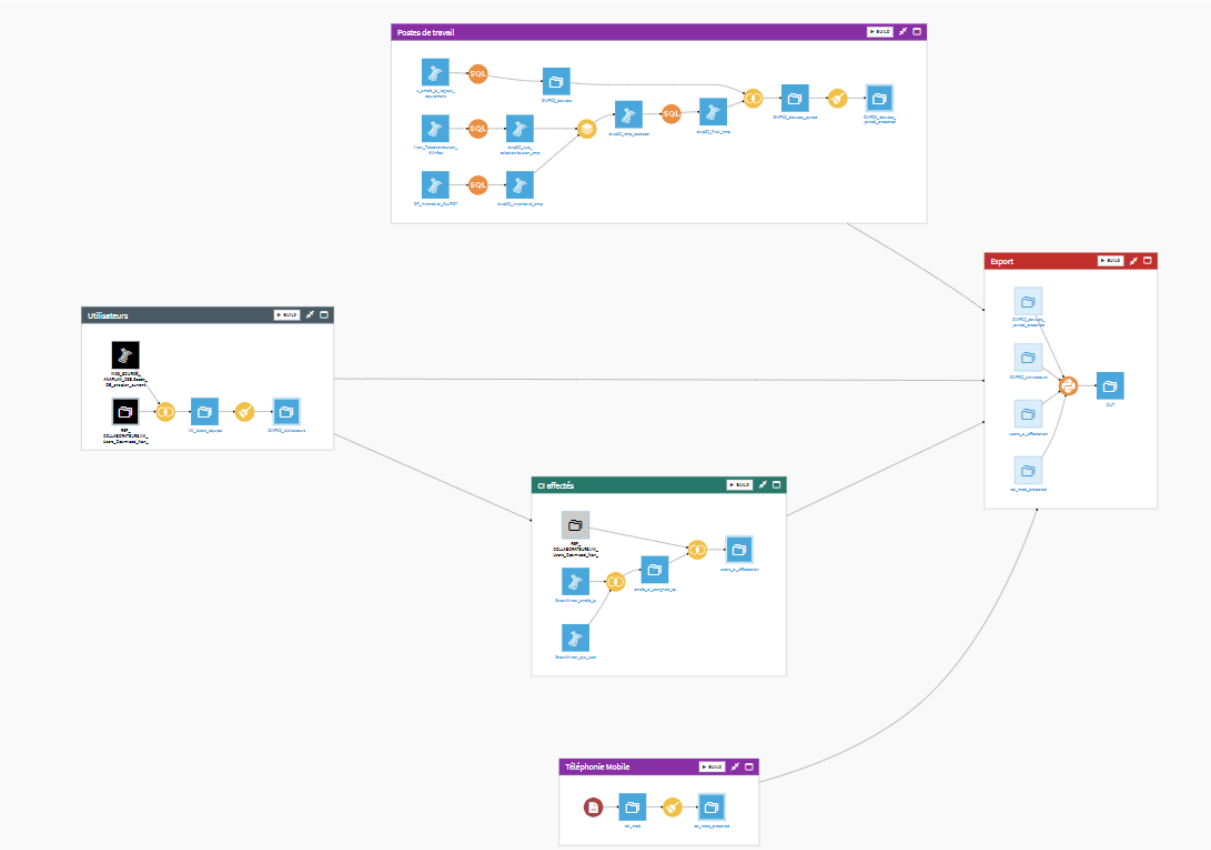
Ce code permet de :

- Prendre les informations de connexion et les chemins depuis les variables personnalisées de Dataiku.

- Parcourir les Datasets configurés et de les convertir en DataFrame Pandas.
- Exporter les fichiers vers SharePoint
- Ne pas exporter le Dataset s'il est vide.

La fonction SharePointManager est une fonction issue d'une bibliothèque développée par l'équipe, utilisant les API Microsoft, qui permet d'écrire sur du SharePoint Online.

Après cela, j'ai pu faire la mise en production du projet.



(Figure 9 : Flow dataiku du projet DWP05 - Dotation) Source : CdA

Plus tard dans l'année, le responsable du projet nous a présenté son projet d'évolution concernant la Dotation. Aujourd'hui, les Dashboard existants concernant les dotations affichent déjà les données, mais les règles métier permettant de détecter les anomalies sont encore calculées directement sur les Dashboard. L'objectif est désormais d'intégrer ces règles directement dans Dataiku, afin d'automatiser le processus. Voici la liste des règles à mettre en place :

Anomalies concernant les devices (poste de travail) :

- Poste de test rattaché à une personne
- Poste sans utilisation depuis 90 J

Anomalies concernant les mobiles :

- Plus de 3 mois sans consommation
- Coût du forfait n
- Coût du forfait n-1
- Coût du forfait n-2

Anomalies concernant les users :

- Possède uniquement un poste admin
- Possède uniquement un poste test
- Possède plus d'un poste bureautique
- Possède plusieurs postes admin

De plus, le responsable du projet voudrait y intégrer un système de justification des anomalies. L'idée est que l'utilisateur puisse être notifié lorsqu'une anomalie le concerne. Par exemple si le collaborateur possède seulement un poste admin, il pourra alors justifier sa situation. Une fois justifié, l'anomalie sera neutralisée et n'apparaitra plus ni dans les données ni dans l'application.

Afin de savoir exactement ce qui était attendu de notre part, l'équipe en charge du développement de l'application nous a fourni un fichier avec l'ensemble des champs que devait contenir chaque fichier. L'idée étant d'avoir à la fin du traitement, 3 fichiers (un sur les mobiles, un sur les users et un sur les devices) contenant les anomalies détectées ainsi que le système de justification.

J'ai donc commencé par mettre en place les différentes règles pour la détection des anomalies pour les dotations en commençant par les mobiles.

Ces règles concernaient principalement les prix des forfaits ainsi que la consommation. Un tableau récapitulatif des règles m'a été transmis. Dans un premier temps, j'ai transformé ce document en un tableau Excel.

ID	NB_MOIS_SANS_CONSO	OPERATEUR	PROFIL	COUTS FORFAIT	ENVELOPPE	ANOMALIES
1	3	SFR	Profil 1	9	60 Go	OUI
2	3	SFR	Profil 2	4,75	20 Go	OUI
3	3	SFR	Profil 3	4,75	20 Go	OUI
4	3	SFR	Profil 4	0,33	N/A	OUI
5	3	SFR	Profil i	6	20 Go	OUI
6	3	SFR	Profil o	6	20 Go	OUI
7	3	ORANGE	Profil 1	9,7	40 Go	OUI
8	3	ORANGE	Profil 2	6,7	20 Go	OUI
9	3	ORANGE	Profil 3	6,7	20 Go	OUI
10	3	ORANGE	Profil 4	0,8	N/A	OUI
11	3	ORANGE	Profil i	6,7	20 Go	OUI
12	3	ORANGE	Profil o	6,7	20 Go	OUI
13	3	ORANGE	Profil a	9,7	N/A	OUI
14	3	ORANGE	Profil u	13,7	N/A	OUI
15	3	BOUYGUES	Profil 1	12,83	50 Go	OUI
16	3	BOUYGUES	Profil 2	7,88	25 Go	OUI
17	3	BOUYGUES	Profil 3	6,08	10 Go	OUI
18	3	BOUYGUES	Profil 4	2,3	N/A	OUI
19	3	BOUYGUES	Profil i	6,7	25 Go	OUI
20	3	BOUYGUES	Profil o	5,9	10 Go	OUI

(Figure 10 : tableau anomalies mobiles) Source : CdA

J'ai ensuite importé ce tableau dans Dataiku et fait une jointure avec le fichier contenant les informations sur les mobiles. J'ai ensuite fait une recette « prepared » ou j'ai utilisé l'étape conditionnelle « if, then, else », qui permet de créer une colonne signalant si la condition est remplie ou non afin de comparer les prix des forfaits en fonction de l'opérateur et du profil. Par exemple, pour un téléphone avec l'opérateur Orange et le profil 3, le coût du forfait ne doit pas dépasser 6,7€. Si c'est le cas, alors la colonne « Anomalies_COUT_FORFAIT_N » sera égale à True sinon, à False. Cette même logique a été appliquée pour la détection des périodes sans consommation.

OPERATEUR	PROFIL	MONTANT_HT	MONTANT_HT_M1	MONTANT_HT_M2	COUTS_FORFAIT	ANOMALIE_COUT_FORFAIT_N_MOINS_2	ANOMALIE_COUT_FORFAIT_N_MOINS_1
SFR	Profil 3	4,75	4,75	4,75	4,75	false	false
ORANGE	Profil 3	21,64	28,15	8,26	6,7	true	true
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false
ORANGE	Profil 3	6,7	7,05		6,7	false	true
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false
SFR	Profil o	6	6	6	6	false	false
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false
SFR	Profil 3	4,75	4,75	4,75	4,75	false	false
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false
ORANGE	Profil 3	6,7	6,7	6,7	6,7	false	false

(Figure 11 : extrait du fichier anomalie Mobile) Source : CdA

Je suis ensuite passée à la détection des anomalies concernant les utilisateurs. J'ai réalisé plusieurs recettes « group », ce qui m'a permis de compter le nombre de poste bureautique, poste admin, poste de test et poste de production par personne.

En filtrant sur mon nom, on peut voir que je n'ai qu'un poste dit classique (bureautique).

EMAIL_METIS	CATEGORIE_POSTE_calculated	count_poste_classique
string	string	bigint
E-mail address	Text	Integer
axelle.peenaert@ca-gip.fr	MOWE	1

(Figure 12 : extrait de la recette group) Source : CdA

Par la suite, la recette « prepared » m'a permis de calculer les anomalies. Par exemple, pour « Possède uniquement un poste test », on regarde si le compteur de poste de production est égal à 0 ET que le compteur de poste de test est supérieur ou égal à 1. Si ces deux conditions sont réunies, alors la colonne est égale à True, si non, elle est égale à False. Pour l'anomalie « Possède plus d'un poste bureautique », je regarde si la colonne « count_poste_classique » est supérieure à 1. Si oui, alors la colonne « ANOMALIE_PLUSIEURS_POSTES_BUREAUTIQUES » sera alors égale à True, si non à False. On répète ce processus pour l'ensemble des anomalies concernant les utilisateurs.

EMAIL	COMPTEUR_POSTE_ADMIN	ANOMALIE_NB_POSTE_ADMIN	COMPTEUR_POSTE_PROD	COMPTEUR_POSTE_TEST	ANOMALIE_UNIQUEMENT_POSTE_TEST
	0	FALSE	0	0	FALSE
	0	FALSE	0	0	FALSE
	1	FALSE	2	0	FALSE
	0	FALSE	1	0	FALSE
	1	FALSE	2	0	FALSE
	1	FALSE	2	0	FALSE
	0	FALSE	0	0	FALSE
	0	FALSE	1	0	FALSE
	0	FALSE	1	0	FALSE
	0	FALSE	1	0	FALSE
	0	FALSE	1	0	FALSE
	0	FALSE	1	0	FALSE
	0	FALSE	0	0	FALSE
	0	FALSE	1	0	FALSE
	1	FALSE	2	0	FALSE
	0	FALSE	0	0	FALSE
	1	FALSE	2	0	FALSE

(Figure 13 : extrait du fichier anomalie User) Source : CdA

Pour finir, je calcule les anomalies pour les postes de travail. Pour l'anomalie « Poste sans utilisation depuis 90 J », on regarde si le nombre de jour d'inactivité est supérieur à 90 jours : si oui, la colonne est égale à True, si non, elle est égale à False. Pour ce qui est de l'anomalie « Poste de test rattaché à une personne », je regarde d'abord si le poste est un poste de test et ensuite, je fais une regex (Regular expression). Une regex est une séquence de caractères qui forme un motif de recherche utilisé pour identifier, rechercher ou manipuler des ensembles caractères en fonction de motifs définis. Ici, la regex permet d'identifier les adresses mails de service (présence de chiffres

avant le « @ »). Si c'est le cas, la colonne « Poste de test rattaché à une personne » est égale à False, si non, elle est égale à True.

USAGE_POSTE	CATEGORIE_POSTI	USAGE	EMAIL_METIS	ANOMALIE_DEVICE_POSTE_TEST_PERSONNEL	ANOMALIE_SANS_UTILISATION_DEPUIS_90J
PTF	MOWE	Test / Exploitat	0000517679@ca-gip.fr	FALSE	TRUE
PTF	MOWE	Test / Exploitat	66767@ca-gip.fr	FALSE	TRUE
PTF	MOWE	Test / Exploitat	0000517679@ca-gip.fr	FALSE	TRUE
PTF	MOWE	Test / Exploitat	0000517679@ca-gip.fr	FALSE	TRUE
PTF	MOWE	Test / Exploitat	0000517679@ca-gip.fr	FALSE	TRUE
PTF	MOWE	Test / Exploitat	0000517679@ca-gip.fr	FALSE	TRUE
PROD	PDTA	Administratior	hicham.mimouni-prestatair	FALSE	TRUE
PTF	MOWE	Test / Exploitat	0000517679@ca-gip.fr	FALSE	TRUE
PROD	PDTA	Administratior	sandie.corsellas@ca-gip.fr	FALSE	TRUE
PROD	PDTA			FALSE	TRUE
PROD	PDTA			FALSE	TRUE
PROD	PDTA			FALSE	TRUE
PROD	PDTA			FALSE	TRUE
PROD	PDTA			FALSE	TRUE
PTF	MOWE			FALSE	TRUE
PTF	MOWE			FALSE	TRUE
PROD	PDTA			FALSE	TRUE
PROD	PDTA			FALSE	TRUE

(Figure 14 : extrait du fichier anomalie Device) Source : CdA

Une fois les anomalies identifiées, j'ai pu intégrer la partie justification des anomalies. Pour ce faire, l'équipe en charge du projet nous envoie un fichier avec les anomalies à justifier depuis un SharePoint. J'ai donc commencé par l'exporter vers le projet Dataiku grâce à un script python. L'étape suivante a consisté en un pivot, ce qui m'a permis de transformer les lignes en colonnes. Cela m'a donc permis de compter le nombre d'anomalies justifiées par personne pour chaque anomalie.

EMAIL_METIS	ANOMALIE_COUT_FORFAIT_N_count	ANOMALIE_COUT_FORFAIT_N_1_count	ANOMALIE_COUT_FORFAIT_N_MOINS_2_count	ANOMALIE_DEVICE_POSTE_TEST_PERSONNEL_count
string	bigint	bigint	bigint	bigint
E-mail address	Integer	Integer	Integer	Integer
	1	0	0	0
	0	1	0	0
	0	0	0	0
	0	0	0	0
	0	0	1	1

(Figure 15 : extrait du pivot) Source : CdA

Ensuite, il a fallu faire 3 jointures différentes qui permettent à la fois de séparer les anomalies en fonction de leurs types (mobiles, users et devices), mais aussi d'enrichir chaque fichier avec les données de dotation récupérées en début de projet.

L'étape d'après consiste à neutraliser les anomalies. Par exemple, prenons le cas de l'anomalie « ANOMALIE_COUT_FORFAIT_N_1 ». Si le compteur (qui compte les anomalies justifiées) de cette anomalie est supérieur à 1 (donc qu'il y a au moins une anomalie sur le coût du forfait remonté pour cette personne), alors la colonne « JUSTIFICATION_ANOMALIE_COUT_FORFAIT_N » sera égale à True, si non à False.

Pour avoir un aperçu, voici le contenu de la table sur les mobiles (contenant les anomalies et la justification) :

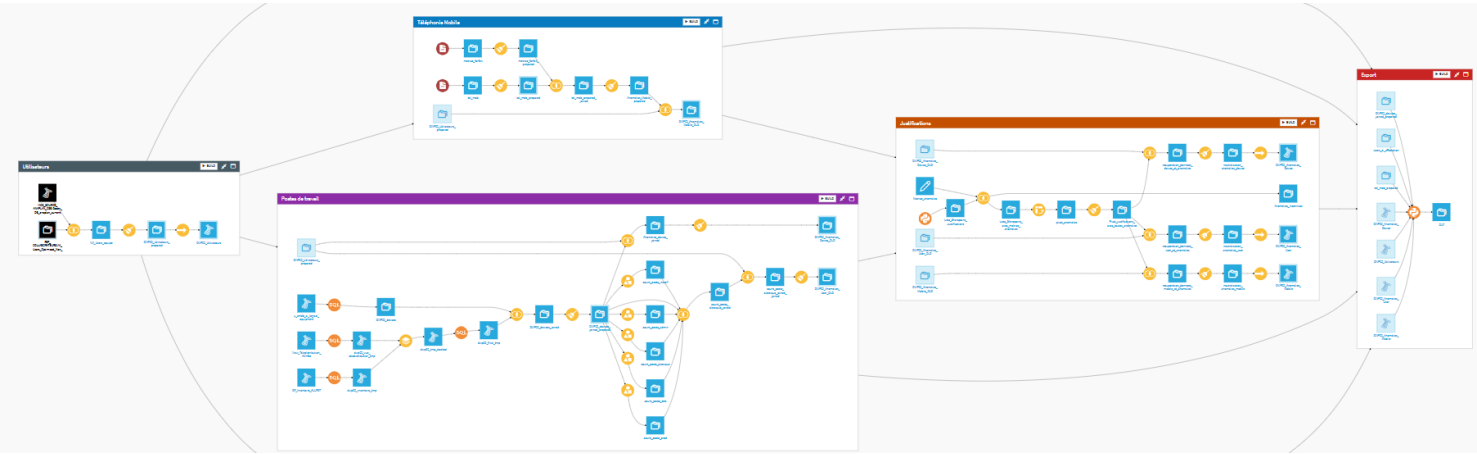
Nom du champs	Signification
Matricule	Identifiant de la personne
Nom	Le nom de la personne
Prénom	Le prénom de la personne
Email	L'adresse mail de la personne
Modèle	Le Modèle du téléphone
Numéro_Ligne	Le numéro de téléphone de la personne
Opérateur	L'opérateur attribué au téléphone
Profil	Le profil du forfait
Envelloppe	Enveloppe Internet du forfait
NB_Mois_Sans_Consommation	Le nombre de mois sans consommation
Date_Facture	Date de la dernière facture
Montant_ht	Montant de la facture du mois
Montant_ht_M1	Montant de la facture du mois précédent
Montant_ht_M2	Montant de la facture d'il y a 2 mois
NB_Mois_Sans_Conso	Le nombre de mois possible sans consommation avant de passer en anomalie
Anomalie_Plus_De_3_Mois_Sans_Conso	Si une anomalie est constatée, alors la colonne est égale à True, si non, False
Cout_Forfaits	Montant de forfait qu'il ne faut pas dépasser, sinon une anomalie sera remontée
Anomalie_Cout_Forfait_N_Moins_2	Si une anomalie est constatée, alors la colonne est égale à True, si non, False
Anomalie_Cout_Forfait_N_Moins_1	Si une anomalie est constatée, alors la colonne est égale à True, si non, False
Anomalie_Cout_Forfait_N	Si une anomalie est constatée, alors la colonne est égale à True, si non, False
Libelle_Equipe	Nom de l'équipe de la personne
Libelle_Service	Nom du service de la personne
Libelle_Département	Nom du département de la personne
Libelle_Direction	Nom de la direction de la personne
INT_EXT	Interne ou Externe
Nom_Manager	Nom du manager
Prénom_Manager	Prénom du manager

Email_Manager	Adresse mail du manager
CGB	Identifiant unique de l'équipe
Justification_Anomalie_Plus_De_3_Mois_Sans_Conso	Si l'anomalie est justifiée, alors la colonne est égale à True, sinon False
Justification_Anomalie_Cout_Forfait_N	Si l'anomalie est justifiée, alors la colonne est égale à True, sinon False
Justification_Anomalie_Cout_Forfait_N_Moins_1	Si l'anomalie est justifiée, alors la colonne est égale à True, sinon False
Justification_Anomalie_Cout_Forfait_N_Moins_2	Si l'anomalie est justifiée, alors la colonne est égale à True, sinon False
TIMESTAMP	Date de la dernière mise à jour de la donnée

Les deux autres fichiers (pour les devices et les users) sont construits de la même manière que celui-ci mais avec les informations les concernant.

Une fois l'ensemble des fichiers finalisés, je les ai synchronisés à la fois dans une base de données, mais aussi dans le SharePoint pour que les données soient accessibles pour les personnes en charge du développement de l'application.

Voici le rendu final du flux Dataiku, une fois le projet terminé dans sa globalité :



(Figure 16 : Flux final du projet dotation) Source : CdA

- Zone grise : Partie sur les utilisateurs
- Zone bleue : Partie sur les mobiles
- Zone violette : Partie sur les postes de travail (devices)
- Zone orange : Partie sur la justification
- Zone rouge : Partie sur la synchronisation des fichiers vers le SharePoint

Pour conclure, l'ensemble des missions qui m'ont été confiées au cours de cette année m'ont permis d'élargir mes compétences techniques, en particulier dans le traitement, l'analyse et la valorisation de données, ainsi que dans l'utilisation d'outils spécialisés comme Dataiku. J'ai également développé des aptitudes organisationnelles et relationnelles, en apprenant à gérer des projets de bout en bout, à collaborer avec différents interlocuteurs et à m'adapter à des contextes variés.

IV) Conclusion

Au terme de cette année d'alternance, l'ensemble des missions qui m'ont été confiées ont constitué une expérience à la fois riche, formatrice et exigeante. Les projets sur lesquels j'ai travaillé m'ont permis d'acquérir une vision globale du traitement et de la valorisation des données au sein d'une grande entreprise, tout en développant des compétences techniques solides.

Sur le plan technique, j'ai renforcé de manière significative ma maîtrise de l'outil Dataiku, qui a été au cœur de la plupart des projets. J'ai appris à exploiter pleinement ses fonctionnalités pour la préparation, la transformation et l'analyse des données, mais aussi pour la mise en place de flux automatisés grâce aux scénarios. En parallèle, j'ai consolidé mes compétences en SQL pour interroger et structurer les données issues de différentes sources, et en Python pour automatiser certaines étapes, interagir avec des API ou encore déposer des fichiers sur des espaces de partage tels que SharePoint.

Malgré les difficultés rencontrées, j'ai toujours réussi à trouver une solution, soit en m'appuyant sur l'aide de mes collègues, soit en surmontant les obstacles par moi-même. Cette expérience m'a ainsi permis de développer non seulement mes compétences techniques, mais aussi ma capacité à résoudre des problèmes de manière autonome et collaborative. Elle m'a appris à analyser les situations, à tester plusieurs approches et à persévérer jusqu'à obtenir un résultat fiable. Les échanges réguliers avec les différents interlocuteurs ont été l'occasion de mieux comprendre les attentes des utilisateurs finaux, d'adapter mes livrables et de développer des compétences en communication technique et vulgarisation.

Sur le plan personnel, l'intégration dans l'équipe Plateformes et Référentiels a été particulièrement enrichissante. J'ai eu la chance de collaborer avec des professionnels, de partager des connaissances, de participer à des résolutions collectives de problèmes et de bénéficier de conseils précieux. Ces interactions m'ont permis de progresser en communication, de gagner en assurance et de m'adapter à des environnements de travail variés et dynamiques.

En conclusion, cette alternance a été une étape essentielle dans mon parcours, me permettant de consolider mes connaissances dans le domaine de la data tout en développant des qualités indispensables au monde professionnel : rigueur, adaptabilité, esprit d'équipe et sens de l'initiative. Elle m'a donné une meilleure compréhension du fonctionnement interne d'une grande entreprise et m'a préparé à relever de nouveaux défis, avec confiance et efficacité.

De plus, j'ai eu l'opportunité de signer un contrat pour deux années supplémentaires au sein de la même structure, dans le cadre du Master MIAGE Informatique Décisionnelle à l'Université Paris-Saclay que j'intégrerai à la rentrée. Cette nouvelle étape me permettra d'approfondir mes compétences, de m'investir dans des projets plus ambitieux et de continuer à évoluer dans un environnement stimulant, en lien direct avec mes objectifs professionnels.

V) Bibliographie

Je souhaite préciser que je me suis aussi beaucoup aidée de documents fournis directement par l'entreprise (organigramme, schéma, présentation, PowerPoint, document écrit...) qui ne peuvent pas être partagés car ils sont internes à l'entreprise et donc confidentiels.

Histoire du Crédit Agricole :

Histoire du groupe Crédit Agricole | Crédit Agricole 1re banque des particuliers

Présentation de CA-GIP :

Crédit Agricole Group Infrastructure Platform CA-GIP | Crédit Agricole

Mieux connaître Crédit Agricole Group Infrastructure Platform | Crédit Agricole Carrières

Informations sur Dataiku :

Dataiku | The Universal AI Platform™

About Dataiku | Dataiku

Dataiku — Wikipédia

Définition RSE :

Qu'est-ce que la responsabilité sociétale des entreprises (RSE) ? | Ministère de l'Économie des Finances et de la Souveraineté industrielle et numérique

Définition DSI :

C'est quoi un DSI ? Définition et explications de son rôle

VI) Glossaire

- **Backlog** : Liste de tâches non terminées.

- **CA-GIP** (Crédit Agricole Group Infrastructure Platform) : maison de production informatique du Groupe Crédit Agricole, regroupant 80% de la production informatique, des infrastructures.

- **Cluster** : Les Clusters portent la responsabilité de la production informatique de bout-en-bout vis-à-vis des Entités du Groupe, à travers l'intégration, le support aux projets et l'exploitation applicative. Ils travaillent en proximité avec eux.

- **Dashboard** : Interface visuelle interactive qui permet de visualiser et de surveiller un ensemble de données de manière consolidée et intuitive.

- **Directeur des Systèmes d'Information** (DSI) : Pilote l'ensemble de l'infrastructure technologique et informatique d'une entreprise. Sa mission englobe la gestion, la mise à jour, et la sécurisation de tous les équipements et logiciels informatiques. Il s'occupe également de la supervision des réseaux et systèmes de communication. Le DSI veille à l'alignement de la stratégie informatique avec les objectifs globaux de l'entreprise, garantissant ainsi une intégration technologique efficace et sécurisée.

- **Ecoconception** : Conception d'un produit, d'un bien ou d'un service, qui prend en compte ses effets négatifs sur l'environnement tout au long de son cycle de vie, afin de les réduire tout en s'efforçant de préserver ses qualités ou ses performances.

- **Entités** : Clients de CA-GIP faisant partie du groupe Crédit Agricole.

- **ETL** (Extract, Transform, Load) : Processus utilisé pour intégrer des données provenant de différentes sources dans un entrepôt de données. L'extraction consiste à collecter les données brutes depuis diverses sources. La transformation implique de nettoyer, structurer et convertir ces données en un format approprié. Enfin, le chargement dépose les données transformées dans une base de données ou dans un entrepôt de données pour une analyse ultérieure.

- **MEP** (Mise en Production) : Processus par lequel une application, un système ou une fonctionnalité nouvellement développée est transférée de

l'environnement de développement, où elle sera utilisée par les utilisateurs finaux.

- **Pôles Technologiques** : Les Pôles Technologiques portent les activités technologiques de CA-GIP. Ils mutualisent les plateformes technologiques partagées à l'échelle du Groupe.

- **Pré production** : Espace de test qui reproduit à l'identique l'environnement de production afin de valider le flux avant sa mise en ligne.

- **Regex** : Séquence de caractères qui forme un motif de recherche utilisé pour identifier, rechercher ou manipuler des ensembles caractères en fonction de motifs définis.

- **Serveur Mainframe** : Ordinateur puissant conçu pour traiter de grandes quantités de données et exécuter de nombreuses transactions simultanément.

- **Sprint** : Période de travail définie dans la méthode agile, généralement d'une à quatre semaines, durant laquelle une équipe de développement s'engage à compléter un ensemble spécifique de tâches ou de fonctionnalités.