**                                                                                                             Logo de l'entreprise Orange**

**Lycée Parc de Vilgénis**                                                                                                   **Orange**

80 Rue de Versailles                                                                            111 Quai du Président Roosevelt

91305, Massy                                                                                                92130, Issy-les-Moulineaux

**Rapport de stage**

**Data Analyst**

**Yanis Parent**

Du 27 mai au 05 juillet 2024

Maître de stage : M. Thierry Bel, Head of Hub Data Value Reporting Factory

Année scolaire 2023 – 2024

**Présentation d’Orange :**

**Nom de l'entreprise :** Orange S.A.

**Secteur d'activité :** Télécommunications

**Date de création :** 1988 (sous le nom de France Télécom)

**Siège social :** 111, quai du Président Roosevelt, 92130 Issy-les-Moulineaux, France

**Président-directeur général :** Christel Heydemann

**Effectif :** Environ 137,000 employés

**Chiffre d'affaires :** 44.1 milliards d'euros en 2023

**Comité Exécutif :**

* **Christel Heydemann**– Directrice Générale du groupe Orange
* **Jean-François Fallacher** – Directeur Exécutif, CEO d’Orange France
* **Caroline Guillaumin** – Directrice Exécutive de la Communication du Groupe
* **Laurent Martinez** – Directeur Exécutif Finance, Performance et Développement du Groupe
* **Mari-Noëlle Jego-Laveissière** – Directrice Exécutive, CEO d’Orange Europe (hors France)
* **Michaël Trabbia** – Directeur Exécutif, CEO d’Orange Wholesale
* **Nicolas Guérin** – Secrétaire Général du Groupe
* **Elizabeth Tchoungui** – Directrice Exécutive en charge de la responsabilité sociétale d’entreprise du Groupe
* **Hugues Foulon** – Directeur Exécutif, CEO d’Orange Cyberdéfense
* **Jérôme Hénique** – Directeur Exécutif, CEO d’Orange Afrique et Moyen-Orient (OMEA)
* **Aliette Mousnier-Lompré** – Directrice Exécutive, CEO d’Orange Buisness
* **Vincent Lecerf** – Directeur Exécutif en charge des Ressources Humaines du Groupe
* **Bruno Zerbib** – Directeur Exécutif, Chief Technology & Innovation du Groupe

**Activités principales :**

* Téléphonie fixe et mobile
* Internet haut débit et fibre optique
* Services de télévision et de contenu numérique
* Solutions pour entreprises (Orange Business)
* Services cyberdéfense (Orange Cyberdéfense)

**Présence internationale :**

* Présent dans 26 pays
* Plus de 298 millions de clients dans le monde

**Historique :**

* 1988 : Création de France Télécom
* 2000 : Introduction en bourse
* 2013 : Changement de nom pour Orange S.A.

**Innovations récentes :**

* Déploiement de la 5G
* Développement de solutions IoT (Internet des Objets)
* Initiatives en matière de cybersécurité

**Engagements RSE (Responsabilité Sociétale des Entreprises) :**

* Réduction de l'empreinte carbone
* Risques climatiques
* Inclusion numérique
* Diversité et inclusion au travail

**Partenariats et acquisitions récentes :**

* Acquisition de Basefarm (2018)
* Partenariat avec Google Cloud (2019)
* Fusion avec MasMovil en Espagne (2024)

**Contact :**

* **Site web :** [www.orange.com](https://www.orange.com)

**Service client :** 3900 (depuis la France)

**Finance & Performance :**

La filière Finance & Performance d'Orange, dirigée par Laurent Martinez, compte plus de 6595 collaborateurs répartis à travers le monde. Elle est composée à 65% de cadres et à 35% de non-cadres, avec une répartition équilibrée entre la France (49,4% des effectifs) et l'international (50,6%).

La raison d'être de la filière Finance & Performance est de "renforcer la création de valeur durable pour Orange". Elle regroupe divers métiers, tels que le contrôle de gestion, la comptabilité, la gestion des risques, l'audit, la supply chain, les achats et la gestion administrative.

**Finance Solutions & Data (FSD) :**

Au sein de la filière Finance & Performance, la direction Finance Solutions & Data (FSD) conçoit et déploie des solutions digitales et data pour maximiser la valeur, l'efficacité opérationnelle et les synergies au sein du groupe Orange. Dans FSD il y a d’une part des solutions digitales et de l’autre la data. La data et les services associés sont mutualisées au sein d’une organisation (IDS), tandis que les solutions sont distribuées dans plusieurs organisations selon leur adhérence à des processus métiers comme par exemple Oracle Fusion qui porte le Purchase to Pay (P2P) et sa comptabilisation. FSD s'engage à accompagner le développement des compétences et des carrières des experts en solutions et data, tout en poursuivant la mutualisation et la fédération des solutions Finance, Performance et RSE. Elle met également l'accent sur l'accélération de l'usage de la donnée via un écosystème enrichi, le Hub Data Value (HDV), pour améliorer l'efficacité opérationnelle à travers un panel de solutions data techniques ou fonctionnelles : Solutions d’échange (flux de données automatisés) ou de partage de données (patrimoine data mis en catalogue avec téléchargement sous contrôle d’accès), solutions d’exposition pour prise de décision (business intelligence dashboard), utilisation de techniques de data science ou d’IA (prédictions, mise en qualité).

**Innovation Data & Services (IDS) :**

Au sein de la direction Finance Solutions & Data, l'entité Innovative Data & Services (IDS) exploite et maximise la valeur des données de l'entreprise. IDS a pour mission de permettre des prises de décisions éclairées, d'optimiser les performances et de favoriser une innovation continue.

**Principaux enjeux & défis :**

* **Utiliser** la data pour intensifier l'agilité opérationnelle et répondre aux nouvelles réglementations et engagements RSE.
* **Promouvoir** une innovation continue de l'infrastructure technologique Data pour démocratiser l'usage de la data au sein des équipes métiers.
* **Consolider** les données dans le datalake corporate F&P pour faciliter l'accès et l'exploitation des données.
* **Intégrer** les technologies émergentes autour de l'IA et de l'IA générative pour améliorer les processus décisionnels.

**Principales missions :**

* **Industrialiser** la collecte, le stockage, l'analyse, la valorisation et la mise à disposition des données.
* **Accompagner** les métiers dans l'usage et la production de use cases et de reportings data.
* **Proposer** des solutions pour faciliter les initiatives data à forte valeur ajoutée.
* **Sécuriser** le patrimoine de données (sécurité, qualité, conformité réglementaire, éthique – data gouvernance).
* **Appliquer** les stratégies du groupe et assurer une veille technologique continue.

**Organisation et équipes :** IDS est structurée en plusieurs pôles, incluant des data product owners, des experts en infrastructure, des data engineers, des data scientists et des data analysts. Cette organisation agile permet de répondre rapidement aux besoins des métiers tout en garantissant la qualité et la sécurité des données.

**Présentation d'Hub Data Value (HDV) :**

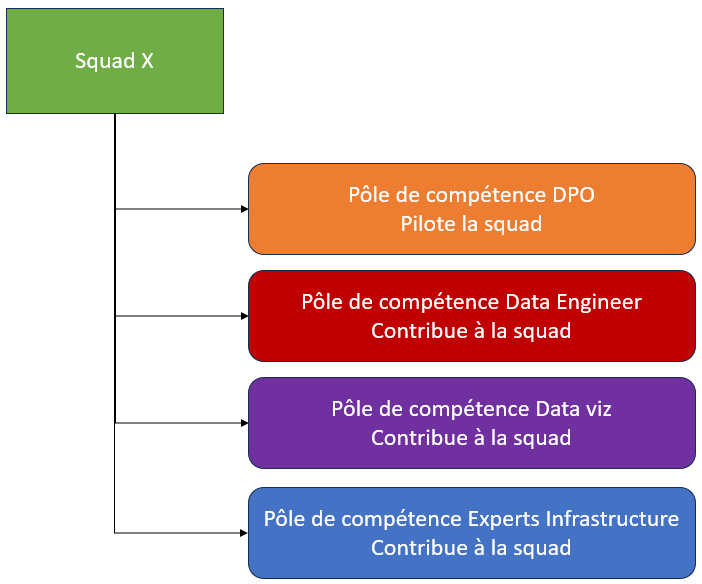
Le Hub Data Value (HDV) est un ensemble de services data comprenant des plateformes de stockage, des solutions d’aiguillage des données, des solutions de modélisation puis d’expositions, des solutions de gouvernances et de documentation.

Cet ensemble de service permet de répondre aux enjeux et défis de l’entité IDS sur les domaines clients comme Finance, Supply Chain, Achat, Immobilier ou RSE

**Organisation et Structure :** HDV est tout un écosystème managé par les personnes d’IDS et structuré en plusieurs services de data.

**Écosystème et Services :** HDV se distingue par une architecture technique robuste qui inclut plusieurs couches :

* **Front End :** Interface utilisateur pour l'accès aux données et aux reportings.
* **Middle End :** Traitement et transformation des données, avec notamment la plateforme Dataiku pour la préparation et l'enrichissement des données.
* **Backend :** Stockage et gestion des données dans le datalake corporate.

**Pôles de Compétences et Squads :**

* **Les pôles de compétences** regroupent des spécialistes autour de compétences particulières comme les data engineer pour l’acquisition et la modélisation des données ou les dataviz coder pour la construction des dashboards par exemple.
* **Les squads** regroupent pour un domaine métier donné des compétences fournies par les pôles de compétences. Chaque squad est pilotée par un DPO (Data Product Owner). L’ensemble des travaux réalisés par les squads sont articulés en mode agile avec l’aide d’un scrum master.

**Focus sur la Squad RSE/DS et HDV Reporting Factory :** La Squad RSE/DS se concentre sur la migration et la transformation des datasets entre Oracle et BigQuery, en utilisant Dataiku pour automatiser et optimiser les pipelines de données.

**Activités principales de la Squad RSE/DS :**

* **Migration de Datasets :** Transfert des données d'Oracle vers BigQuery pour améliorer l'efficacité et la performance des analyses.
* **Utilisation de Dataiku :** Plateforme utilisée pour nettoyer, préparer, enrichir et transformer les données avant leur exposition finale.
* **Automatisation des Pipelines :** Mise en place de processus automatisés pour assurer la qualité et la cohérence des données.
* **Documentation et Contrôles de Qualité :** Assurer la documentation des processus et la mise en place de contrôles de qualité pour garantir l'intégrité des données.

**Focus sur la partie middle end :** Dans le cadre de mes travaux, je me concentre sur le projet ? Dim Fact avec Dataiku. Il s’agit d’effectuer des migrations de dataset d'Oracle vers BigQuery tout en vérifiant les données et en s’assurant de son intégrité.

* **Dataiku** est une plateforme unique qui permet de nettoyer, préparer, enrichir, et transformer les données. Elle facilite également l'expérimentation rapide et la création de modèles performants.
* **Dim Fact** se réfère aux dimensions et aux faits dans le modèle de données, essentiels pour structurer les données de manière à ce qu'elles soient facilement exploitables pour les analyses et les reportings.

L'utilisation de Dataiku dans le cadre de Dim Fact permet d'automatiser les pipelines de données, d'assurer la qualité des données et de garantir que les bonnes pratiques sont appliquées tout au long du processus de transformation. Cela inclut la documentation des processus de données, la mise en place de contrôles de qualité, pour ainsi faciliter le partage communautaire.

**Mission réalisée**

Au cours de ce stage, j’ai réalisé une transformation et une migration de données depuis Dataiku FCA (connexion sous Oracle) vers Dataiku FE (connexion sous Bigquery) en utilisant dataiku.

**Mission précise**

Au niveau du domaine Sourcing, la chaine de données middle end avait été développée sur la couche normalement dédiée au front end. Plusieurs raisons justifiaient ce choix :

* Les codeurs dataviz étaient très proches des business et leur compréhension des business-modèles leur permettaient de modéliser les données de manière adaptée et rapide (inutile d’expliquer à un autre acteur)
* L’outil qui leur était très familier (Qlik) permettait de scripter des traitements dans un langage qu’ils maitrisaient
* L’outil était extrêmement performant en matière de rapidité dans la manipulation de millions de lignes de données (travail en full mémoire).

Cependant, cette solution présentait plusieurs inconvénients majeurs :

* **Adhérence** de la chaine middle end à un outil dataviz dont le cycle de vie est potentiellement plus volatil que des solutions backend.
* **Impossibilité** de partager autrement que via les dashboards le patrimoine des données

À la suite de cela, un projet de redescente de cette chaine middle end vers une couche plus basse a donc été lancé mi-2022. Dataiku a été choisi afin de réaliser cette opération pour :

* Sa capacité à garantir la reprise des traitements scripts Qlik
* Sa capacité à être pris en main de manière professionnelle par les équipes en place (historiquement dataviz).

Ce choix évitait de lancer un projet de transfert technique (d’une plateforme vers une autre) en même temps qu’un transfert de compétences (d’une équipe vers une autre).

Malgré tout, compte tenu du volume et de la complexité du projet, ce dernier aura nécessité plus de 1500 jours.

L’étalement de ce projet dans le temps a inéluctablement conduit à des adaptations à faire lier aux évolutions de l’écosystème. Oracle database sur lequel ce projet avait été réalisé étant remplacé par GCP (Google Cloud Plateform) Big Query.

Mon arrivé sur ce projet coïncidait donc avec deux migrations :

* **Migration** des données d’Oracle vers Big Query
* **Migration / adaptation** des composants Dataiku vers une nouvelle plateforme Dataiku et une connexion Big Query (vs Oracle)

Pour réaliser ma mission Orange a mis à ma disposition :

* Un PC de l’entreprise
* Un accès à HDV
* Une licence Dataiku

L’objectif de ma mission était de migrer différents datasets entre deux environnements spécifiques de dataiku : **Dataiku FCA** et **Dataiku FE**.

Pendant mon stage, j’ai donc réalisé la migration de différents datasets entre deux environnements spécifiques de Dataiku : Dataiku FCA et Dataiku FE

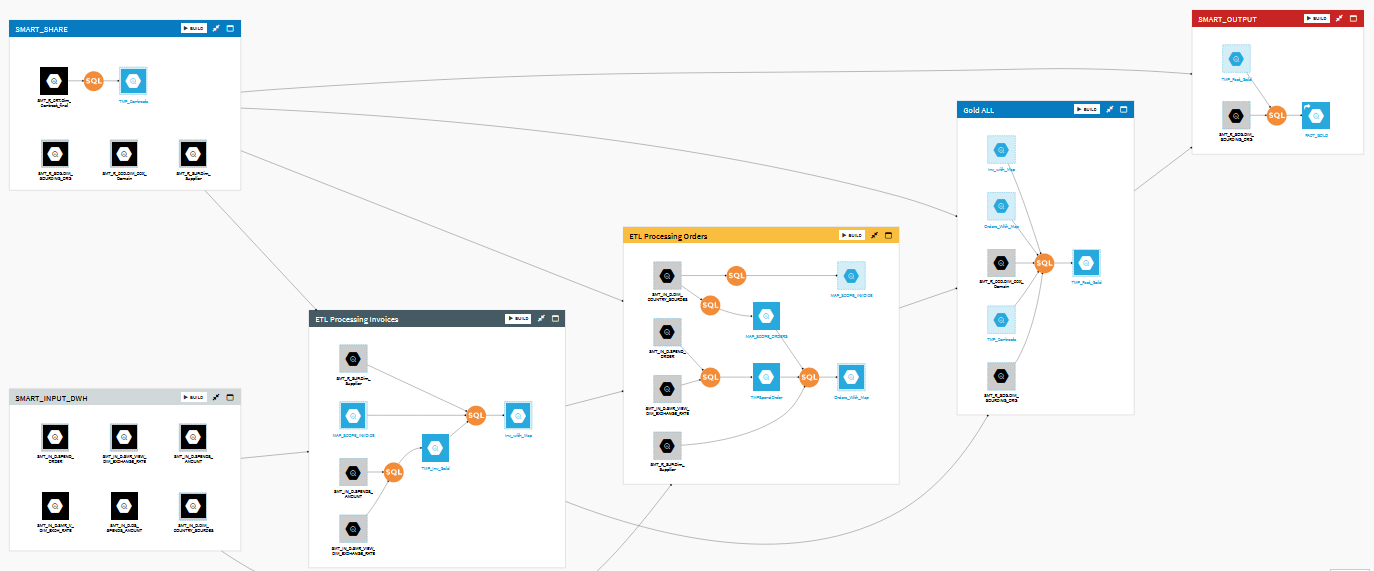
**Dataiku FCA (Flexible Computing Advanced) :**

* **Utilise** différents formats de texte (Majuscule, Minuscule)
* **Utilise** des connexions Oracle pour les datasets

**Dataiku FE (Flexible Engine) :**

* **Utilise** des normes de formatage de texte (minuscule)
* **Utilise** des connexions GCP Big Query pour les datasets

La mission se décompose en plusieurs étapes :

* **Migration** des projets entre deux environnements
* **Modification** de la connexion de chaque dataset, de passer d’Oracle à BigQuery utilisé pour FE
* **Modification** des traitements sur les données pour changer le nom de chaque colonne afin de respecter la norme de formatage de texte
* **Modification** des différents scripts SQL qui n’étaient pas adaptés à l’environnement et s’assurer de leur bon fonctionnement
* **Réalisation** de tests entre les données de FCA et les données de FE pour s’assurer de n’avoir aucune perte ou modification de la donnée.

Voici un exemple de flow dataiku d’un projet sous Dataiku FE. On voit ici que le projet est réparti en plusieurs zones qui nous permettent de se repérer dans les projets et de retrouver plus facilement les datasets. Les datasets sont les carrés bleus et noirs. Les bleus sont ceux qui sont créés ou modifiés dans ce projet tandis que les noirs sont simplement des copies d’un dataset créé dans un autre projet. On voit également des ronds orange avec écrit SQL qui représentent donc un « code recipes », c’est un script SQL qui prend en entrée plusieurs datasets afin d’en créer un dernier, ce qui nous permet de filtrer la donnée et d’obtenir des résultats voulus. Les traits gris nous montrent les utilisations des différents dataset dans les différentes zones.

Le résultat de cette mission est plutôt mitigé. En effet, les scripts SQL sont fonctionnels et la donnée est bien transformée entre les datasets quand il le faut mais malheureusement les phases de tests montrent une différence entre les données. Les données ne sont pas erronées mais elles sont manquantes. Il est difficile d’effectuer les tests à la suite de problèmes de migrations d’Oracle vers GCP bigquery ou de tests sur d’autres projets non terminés.

Lors de ce stage j’ai renforcé mes compétences en SQL mais également en anglais car je travaillais avec une équipe en Roumanie.

Durant ce stage j’ai été accueilli par une équipe très gentille, professionnelle et à l’écoute. Ils étaient tous là pour m’aider et répondre à mes questions lorsque j’en avais besoin. J’ai participé à plusieurs réunions qui m’ont permis de mieux comprendre mon travail et le monde professionnel. J’ai effectué une mission enrichissante et intéressante. Tout s’est bien passé et je suis très content d’avoir pu effectuer ce stage chez Orange, dans l’équipe de Thierry Bel

**Lexique**

**Datalake Corporate** : Un datalake corporate est un référentiel centralisé qui permet de stocker de grandes quantités de données brutes dans leur format natif jusqu'à ce qu'elles soient utilisées. Il est conçu pour consolider les données de toute l'entreprise, facilitant ainsi l'accès et l'analyse des données par différents départements.

**Reportings** : Les reportings sont des rapports générés à partir de données collectées et analysées. Ils servent à fournir des informations pertinentes et structurées pour aider à la prise de décision au sein de l'entreprise.

**Data Product Owners** : Les data product owners sont des responsables de produits de données. Ils gèrent le cycle de vie des produits de données, de la conception à la mise en œuvre, en passant par la maintenance et l'amélioration continue. Ils travaillent en étroite collaboration avec les équipes techniques et les utilisateurs finaux pour s'assurer que les produits de données répondent aux besoins de l'entreprise.

**Squad** : Une squad est une petite équipe autonome et multidisciplinaire qui travaille ensemble pour atteindre un objectif commun. Dans le contexte d'Orange, une squad peut inclure des data engineers, des data scientists, des data analysts, et d'autres rôles nécessaires pour mener à bien des projets spécifiques.

**Digital Sourcing** : Le digital sourcing fait référence à l'utilisation de technologies numériques pour optimiser et automatiser les processus d'approvisionnement et de gestion des fournisseurs. Cela inclut l'utilisation de plateformes et d'outils numériques pour améliorer l'efficacité, la transparence et la collaboration avec les fournisseurs.

**Datasets** : Les datasets sont des ensembles de données structurées ou non structurées qui sont collectées et utilisées pour l'analyse. Ils peuvent inclure des données provenant de diverses sources, telles que des bases de données, des fichiers CSV, des API, etc.

**Dataiku** : Dataiku est une plateforme de science des données et de machine learning qui permet aux équipes de collaborer, de préparer, d'analyser et de visualiser des données. Elle offre des outils pour automatiser les pipelines de données et créer des modèles prédictifs.

**BigQuery** : BigQuery est un entrepôt de données cloud entièrement géré par Google Cloud. Il permet d'exécuter des requêtes SQL sur de grandes quantités de données en temps réel, offrant des capacités d'analyse rapide et évolutive.

**Supply Chain** : La supply chain (chaîne d'approvisionnement) est l'ensemble des processus impliqués dans la production et la distribution de biens et services, depuis l'approvisionnement en matières premières jusqu'à la livraison des produits finis aux clients. Elle inclut la gestion des fournisseurs, la logistique, la production, et la distribution.