DOSSIER TECHNIQUE

Formation GEPP

CRISTIONA KADANGHA BARIKI

Membres du groupe :
Moussa
Ingrid

Jeudi, 7 décembre 2023

Table des matières

Introduction 3
1. Présentation du projet 3
1.1. Contexte 3
1.2. Objectifs du projet 3
1.3. Présentation du projet 4
1.3.1. Le sujet 4
1.3.2. Les exigences 4
1.3.3. Contraintes 6
2. Analyse du besoin 7
2.1. Contexte 7
2.2. Enjeux 7
2.3. Problématique et axes d'analyse 8
2.4. Stratégie Big Data 8
3. Déploiement et méthode de gestion de projet 9
3.1. Jalons Recommandés 9
3.2. Méthodologies Agile et DevOps 10
3.2.1. Gestion de projet Agile 10
3.2.2. Principes de déploiement DevOps 11
4. Réalisation du projet 12
4.1. Architecture 12
4.2. Collecte et Stockage des Données 13
4.2.1. Données sélectionnées 13
4.2.2. Politique de Données 16
3.4. Stockage des données 17
3.4.1. Système de stockage des données 17
3.4.2. Environnement de travail 18
3.5. Intégration des données 21
3.5.1. Technologies choisies 21
3.5.2. Processus d'intégration des données 22
3.5.3. Qualité de Données 24
3.6. Analyse et exposition des données 24
3.6.2. Analyse prédictive 25
3.6.2. Visualisation des données 32
Conclusion 37
Annexes 37
Sources 38

Introduction

L'obtention d'un master MIAGE (Méthodes informatiques appliquées à la gestion des entreprises) parcours IDA (Ingénierie des données et apprentissage) m'a permis d'intégrer Eviden en tant que Data Engineer. Comme l'indique le nom du parcours, je me suis spécialisé dans l'ingénierie des données. Les thèmes, méthodes et technologies proposés par cette formation me sont, pour la plupart connus, du moins théoriquement. Parfaitement alignée avec les problématiques actuelles d'exploitation des données, cette formation me permettra de consolider et d'enrichir les connaissances acquises tout au long de ces deux années d'études tout en découvrant de nouveaux outils. Elle m'est tout à fait utile pour exercer mon rôle de Data Engineer avec confiance.

1. Présentation du projet

1.1. Contexte

J'ai participé à la **formation Data Engineer GEPP (gestion des emplois et des parcours professionnels) du 02 octobre 2023 au 7 décembre 2023.** Il s'agit d'une formation organisée et orchestrée par **l'académie du numérique (ADN)**. Né d'un partenariat entre l'ESN (entreprise des services numériques) Eviden et le pôle européen des Deep Tech, Systematic, cette formation concilie formation théorique et pratique. C'est dans ce contexte que nous a été soumis un projet à livrer en fin de formation. L'objectif de ce dernier est de mettre en pratique les connaissances acquises tout au long de la formation théorique en les appliquant à une situation concrète. J'ai travaillé en collaboration avec Moussa Boutemine et Ingrid Tavares.

1.2. Objectifs du projet

L'objectif de la formation est de nous permettre de **développer les compétences nécessaires pour exercer le métier d'ingénieur des données**. Il est un acteur clé dans le monde du Big Data où les données à exploiter ne font que croître exponentiellement. Qualifiées par le terme « massives », les données sont au cœur de toutes les entreprises qui se veulent de plus en plus « Data Driven ». Pour être plus explicite, ces entreprises se basent sur une utilisation quantitative et qualitative des données tout le temps, par tous et partout au sein de leurs activités. Ainsi, l'ingénieur des données a pour mission la collecte, le stockage et le traitement des données caractérisées par un volume, une variété et une vélocité importantes. Il se charge également de développer et mettre en place l'infrastructure technique et les solutions IT adaptées aux modèles de traitement de la data et répondant aux besoins métiers.

Les tâches auxquelles sont confrontées un ingénieur des données sont les suivantes :

- L'analyse d'un besoin métier et la préparation et la mise en œuvre d'une stratégie Big Data
- La gestion des données pour des projets en question
- La manipulation et l'exploitation des données issues de diverses sources
- L'automatisation du traitement des données sur une infrastructure Big Data
- L'industrialisation des processus de traitement des données

Il convient alors de constater que tant qu'il y aura des données à traiter, les ingénieurs des données demeureront indispensables. Le projet proposé sert de moyen d'évaluation de la maîtrise des compétences et des savoir-faire acquis.

1.3. Présentation du projet

Le projet a été introduit dès le début de la formation pour nous donner l'opportunité de réfléchir dès le départ tout en assimilant de nouvelles notions.
Dans le cadre du projet, les participants de la formation endossent le rôle d'ingénieurs des données mandaté par un institut de sondage. Nous avons donc une mission dont l'intérêt est le traitement des données pour en tirer de la valeur en utilisant des outils.

1.3.1. Le sujet

Notre mission est d'analyser les tendances d'un secteur spécifique (au choix) en construisant, du début à la fin, une plateforme Big Data. La construction de cette plateforme implique le choix d'une architecture, d'outils, de méthodes de gestion de projet, techniques de traitement de données, etc…
Afin de réaliser ce projet, nous sommes chargés de récolter les données nécessaires provenant de plusieurs sources notamment **les sources open data**. La seule contrainte quant aux données est le volume : chaque ensemble de données doit contenir au moins 10 000 enregistrements. Bien sûr, il est important de corréler les données issues des différentes sources afin d'en tirer du sens. De plus, en fonction de la nature des données récoltées, nous avons la liberté de choisir les technologies les plus adaptés.

1.3.2. Les exigences

Nous sommes tenus de respecter plusieurs étapes dans la réalisation du projet, à savoir : la compréhension et l'analyse du besoin ; la collecte des données ; le stockage de données et l'architecture, les principes et méthodologies de déploiement DevOps et de gestion de projet Agile.

LA COMPREHENSION ET L'ANALYSE DU BESOIN
Cela implique d'appréhender le contexte global dans lequel se situe le projet, de saisir les enjeux impliqués et de définir la problématique à analyser en profondeur en fonction des **caractéristiques des données** à exploiter telles que :

- Le volume : la quantité totale de données générées, collectées et stockées
- La variété : la diversité des types et des formats de données disponibles (structurées, non structurées et semi-structurées). Cette diversité peut englober des données en provenance d'une multitude de sources (documents texte, images, vidéos, fichiers audios, des données de capteurs, bases de données relationnelles ou NOSQL, etc.)
- La vélocité : la vitesse à laquelle les données sont générées, collectées et traitées. On parle donc de la rapidité à laquelle les données sont produites et disponibles pour être analysées (temps réel ou en lots (batch)). Les flux de données à haute vélocité peuvent provenir de diverses sources telles que les capteurs IoT (Internet des objets), les réseaux sociaux, les transactions financières, les logs, etc.

- Du secteur d'origine des données (le métier) : Comprendre le domaine métier associé aux données pour contextualiser les informations (finance, énergie, vente, etc..). Les spécificités de chaque secteur peuvent influencer la manière dont les données sont collectées, traitées et analysées.

- **De la problématique et les critères d'analyse** : La problématique représente le cœur de la question à résoudre ou à explorer à travers l'analyse des données. Les critères d'analyse sont

les paramètres ou les indicateurs utilisés pour évaluer, mesurer ou examiner les données afin de répondre à la problématique. Ces critères sont généralement définis en fonction des objectifs de l'analyse et peuvent inclure des mesures de performance, des tendances, des corrélations ou d'autres paramètres spécifiques pertinents pour résoudre la problématique identifiée.

LA COLLECTE DE DONNEES

La collecte de données implique la mise en place d'une politique de données visant à définir les méthodes et les règles pour obtenir des données pertinentes et utiles pour l'analyse en tout sécurité et légalité. Cela comprend la politique de données. Il s'agit de l'établissement de directives et de procédures pour collecter les données, déterminer leur provenance, leur qualité, leur légalité et leur utilisation appropriée.

LE STOCKAGE DE DONNEES

Une fois les données collectées, l'étape logique est celle du stockage dans un espace dédié et convenable pour le type et le format de données correspondants. De ce fait, la création d'un schéma conceptuel décrivant la structure et les relations entre les différentes données à stocker s'avère primordial. Il s'agit d'une base pour la conception et la création de la base de données. Le choix du stockage des données influe également sur la sécurité des données. Il est donc important d'anticiper la mise en place de mesures de sécurité pour protéger les données stockées contre les accès non autorisés, les altérations ou les pertes éventuelles.

LE TRAITEMENT DES DONNEES

Le traitement des données est une étape importante dans tout projet Big Data. L'objectif principal d'une tâche de traitement des données est celui de préparer et rendre les données exploitables pour les utilisateurs tels que les Data Analysts et Scientists. L'objectif est la garantie de la qualité des données.

Le traitement des données comprend notamment l'intégration des données et la gestion de la qualité des données :

- Collecte et rassemblement de données provenant de différentes sources, par exemple, bases de données, fichiers CSV, API, etc.
- Jointure ou intégration de ces données pour former un ensemble de données cohérent et complet.
- Définition et mise en œuvre de règles pour assurer la qualité des données.
- Nettoyage des données en supprimant les valeurs aberrantes, les doublons, les valeurs manquantes ou incorrectes (s'assurer de l'unicité, la cohérence, l'interopérabilité, la disponibilité, la justesse et la complétude des données).

L'EXPOSITION DES DONNEES

L'exposition des données implique l'analyse approfondie et la restitution des résultats obtenus à partir de notre ensemble de données. Il est important de sélectionner l'outil le plus adapté à la nature des données. Parmi les outils de visualisation, l'on peut citer : Tableau, Power BI, Matplotlib (librairie Python), ggplot2, etc. Ainsi, l'exposition des données sert à :

- Résumer les principales découvertes, tendances et conclusions extraites des données
- Mettre en évidence les points clés répondant à la problématique de notre projet en utilisant des graphiques, des tableaux, des visualisations, des résumés textuels
- Argumenter les conclusions en s'appuyant sur des analyses statistiques, des modèles prédictifs, des corrélations

L'ARCHITECTURE, **LES PRINCIPES ET METHODOLOGIES DE DEPLOIEMENT DEVOPS ET DE GESTION DE PROJET AGILE**

La flexibilité et l'adaptation aux besoins du projet sont cruciales pour l'avancement du projet. L'ensemble des principes et méthodologies DevOps et gestion de projet Agile favorise la communication entre les membres du groupe tout en encourageant la collaboration étroite entre les membres de l'équipe pour résoudre les problèmes et trouver des solutions efficaces. Notons également que l'amélioration continue est un enjeu de ces principes et méthodes : Engagement à évaluer constamment les processus et à les améliorer pour une livraison plus rapide et de meilleure qualité.

Notre responsabilité est donc de mener à bien l'intégralité d'un projet Big Data en respectant à la fois les exigences techniques et organisationnelles. Néanmoins, nous avons détecté quelques contraintes quant à l'exécution du projet.

1.3.3. Contraintes

Les contraintes rencontrées sont nécessairement matérielles.

Nos postes de travail sont **inadaptés pour le traitement de volumes massifs de données.** En effet, le traitement et l'analyse de grandes quantités de données nécessitent souvent une puissance de calcul considérable. Des contraintes de ressources informatiques telles que la capacité de traitement du CPU, la mémoire RAM ou les capacités de calcul parallèle peuvent ralentir ou entraver les opérations de traitement des données.

Il convient de noter que cette limitation a été prise en considération lors de l'élaboration du projet, nous imposant ainsi une limite d'au moins 10 000 enregistrements par jeu de données.

En fin de compte, nous avons dû faire preuve de modération dans le choix de nos données et de notre architecture pour nous conformer aux contraintes matérielles.

En outre, nous manquons également d'espace de stockage. En raison de nos projets antérieurs, ainsi que des logiciels et des dossiers sur nos postes, nous ne disposons pas d'une capacité suffisante pour stocker non seulement les logiciels nécessaires, mais surtout pour gérer de grandes quantités de données.

2. Analyse du besoin

2.1. Contexte

Afin de mener à bien notre projet Big Data, nous avons choisis des données openData relatives au domaine de la télécommunication, plus précisément celui des réseaux mobiles en France et dans les DROM (Département et région d'outre-mer). Ce domaine connaît une transformation majeure avec l'adoption de la 5G.
La « 5G » est **la cinquième génération de réseaux mobiles,** qui succède aux technologies 2G, 3G et 4G. Cette avancée technologique offre des vitesses de transmission plus rapides et une réduction des délais, ouvrant ainsi de nouvelles opportunités pour les entreprises et les individus. En France, le déploiement de la 5G a été progressif et a débuté dans certaines régions en 2020. Les opérateurs de télécommunications tels que Orange, SFR, Bouygues Telecom et Free ont entrepris des efforts pour étendre la couverture de la 5G à travers le pays. De ce fait a été constatée une multiplication des antennes sur le territoire. Les fournisseurs de services de télécommunications réalisent d'importants investissements dans le cadre du déploiement de la 5G. Ces engagements financiers revêtent une grande importance car ils déterminent la capacité des opérateurs à répondre à la demande croissante de capacité de transmission. En outre, les régulateurs et les consommateurs sont concernés par cette nouvelle technologie.

2.2. Enjeux

L'évaluation des données concernant les déploiements des réseaux mobiles représente une préoccupation capitale pour les opérateurs, les organismes de régulation et les consommateurs :

- Pour les **opérateurs de téléphonie mobile**, cette analyse leur offre la possibilité de surveiller l'évolution de la compétition et d'optimiser la répartition de leurs investissements. De plus, le suivi de près des déploiements pour comprendre la couverture de leur réseau est important afin d’évaluer les zones à améliorer ou à étendre, et planifier les investissements pour offrir des services de qualité aux consommateurs.
- Pour les **organismes de régulation**, cette analyse est cruciale pour garantir la transparence du marché et favoriser une concurrence équitable pour tous les opérateurs sur le marché de la télécommunication. L’ANFR (agence nationale des fréquences) et l’ARCEP (Autorité de Régulation des Communications Électroniques, des Postes et de la Distribution de la Presse) sont les principaux régulateurs de ce type de technologie.
- Pour les **consommateurs**, cette analyse met la lumière sur la qualité de service qu'ils peuvent attendre de leur opérateur. Une évaluation précise des déploiements de réseaux mobiles leur permet de prendre des décisions éclairées lors du choix d'un fournisseur de services mobiles, en tenant compte de la couverture, de la vitesse et de la qualité du réseau dans leur région.

En prenant en considération l'auditoire visé, nous avons relevé une problématique et principalement déterminé les axes d'étude pour maximiser l'exploitation de la valeur des données.

2.3. Problématique et axes d'analyse

La problématique centrale de nos analyses réside dans la répartition des antennes du réseau de téléphonie, ainsi que dans la distribution des différentes antennes réseau à travers l'ensemble du territoire français (DROM inclus).

Afin d'apporter réponse à notre problématique, voici les axes d'analyses que nous proposons :

- Répartition des installations des antennes réseaux en France en fonction des régions, départements
- Evolution temporelle des installations des antennes (type de génération, type de système, de support d'antennes)
- Répartition des installations respectives par opérateurs et zone géographique
- Évaluation de la part de marché des principaux opérateurs.
- Prédiction du nombre futur d'antennes à installer, en utilisant des algorithmes de séries temporelles

Nous avons donc développé et coordonné un plan d'action pour répondre à la problématique mentionnée précédemment. Notre approche repose sur une sélection minutieuse d'outils, de techniques et de méthodes Big Data.

2.4. Stratégie Big Data

Nous disposons de trois ensembles de données distincts :

- Données concernant les **installations d'antennes de téléphonie mobile en France**
- Données concernant les **départements français**
- Données concernant les **régions françaises**.

Notre stratégie Big Data vise à exploiter ces données pour obtenir des informations pertinentes, ce que l'on appelle la valorisation des données. Cette stratégie comprend plusieurs étapes clés :

1. **Exploration et stockage des données** : L'étape d'exploration et de stockage des données brutes débutent par une analyse de la structure de nos différents ensembles de données. Dans un premier temps, l'élaboration d'un dictionnaire des données a été important car il nous permet de découvrir les données, ainsi que leurs relations.
Dans un second temps, afin de détecter tout problème de qualité éventuel au sein des données, nous effectuons du Data Profiling (avec la librairie Python ydata-profiling). Il s'agit d'une analyse approfondie des données pour identifier les incohérences, les erreurs, les valeurs manquantes et les doublons. Cette étape de diagnostic nous permet d'identifier les éventuels défauts ou les anomalies structurelles présentes dans nos données semi-structurées.

Une fois cette phase d'inspection complétée, nous entreprenons le stockage de ces données dans une base de données relationnelle. Ce choix stratégique de stockage est motivé par la nécessité de conserver ces données semi-structurées dans un environnement adapté à leur nature. La base de données relationnelle offre une structure optimale pour organiser et gérer ces données, facilitant ainsi leur manipulation.

2. **Intégration des données** : L'intégration des données est la phase la plus importante de notre stratégie. L'intégration des données, consiste à rassembler des données de plusieurs sources en les nettoyant, les transformant, puis les combinant pour en faire un ensemble de données unifié, et

prêt d'utilisation. Nous employons la méthode ETL (Extract, Transform, Load), une technique qui nous permet d'extraire les données, de les transformer et enfin de les charger dans une base de données cible. Cette méthode nous offre la possibilité de réaliser des opérations de nettoyage et de normalisation des données afin de garantir leur qualité et leur cohérence.

3. **Stockage des données traitées** : Nous choisissons d'utiliser un Datawarehouse (entrepôt de données) pour le stockage des données après leur traitement. Il s'agit d'une plateforme utilisée pour collecter et analyser des données provenant de multiples sources hétérogènes. Il permet le stockage d'un grand volume de données, mais aussi l'interrogation et l'analyse de ces dernières. C'est la raison pour laquelle il occupe une position centrale au sein d'un système de Business Intelligence (BI). En effet, les données brutes stockées sont transformées en informations utiles, et de les rendre disponibles et accessibles aux utilisateurs.
En optant pour une base de données relationnelle, nous nous orientons vers une approche qui favorise les analyses descriptives que nous prévoyons de réaliser par la suite.

4. **Analyse des données** : Une fois les données rassemblées, nettoyées et stockées dans notre entrepôt, nous procédons à l'analyse approfondie des informations afin de détecter des patterns, des constats et des observations.
En parallèle, nous employons des techniques d'analyse prédictive, avec des modèles de Machine Learning. Ces méthodes nous permettent d'anticiper et de prédire des comportements futurs à partir des données actuelles.

5. **Visualisation et présentation des résultats** : La phase de visualisation et de présentation des résultats marque la conclusion de notre stratégie. Après avoir effectué toutes les analyses et extrait les informations pertinentes, nous nous attelons à présenter ces résultats de manière accessible et aisément compréhensible pour les parties prenantes.
Pour ce faire, nous utilisons un outil spécialisé de visualisation de afin de transformer les données en représentations visuelles, telles que des graphiques, des tableaux de bord interactifs, des diagrammes ou des cartes. L'objectif principal est de traduire ces données complexes en visualisations claires et intuitives, permettant ainsi une compréhension immédiate des conclusions tirées.
Après avoir établi cette stratégie, nous nous sommes interrogés sur la méthode de gestion de projet.

3. Déploiement et méthode de gestion de projet

3.1. Jalons Recommandés

Pour garantir la réussite de notre projet dans les délais impartis, des jalons ont été établis, accompagnés de points de suivi par l'équipe ADN.

26/09 : Avoir posé le contexte, les enjeux et la problématique et choisi les données
08/11 : Avoir stocké les données brutes et les avoir traitées
22/11 : Avoir analysé les données et préparé les restitutions
01/12 - 07/12 : Finalisation du dossier technique

Nous avons réussi à respecter les deux premiers jalons initialement prévus. Néanmoins, l'analyse et la restitution ont nécessité deux jours supplémentaires, lesquels ont été achevés le 24 décembre. Par conséquent, le temps restant était très limité pour finaliser la rédaction du dossier technique.

3.2. Méthodologies Agile et DevOps

Au cours de cette formation, nous avons pris part à des modules qui se concentraient sur l'organisation, incluant des sujets tels que DevOps et la méthodologie de gestion de projet Agile.

3.2.1. Gestion de projet Agile

La gestion de projet Agile est une approche de travail itérative et flexible, axée sur la collaboration et la communication entre les membres d'une équipe. La méthode agile favorise : les individus et leurs interactions plus que les processus et les métiers ; des livrables opérationnels plutôt que qu'une documentation exhaustive ; la collaboration avec les clients ; et l'adaptation au changement plutôt que le suivi d'un plan.
Pour aller plus loin, nous avons plus ou moins utilisé le Framework Scrum qui prône la collaboration entre les membres d'une petite équipe de personnes (Max 10). L'équipe Scrum est composée de :

- **Scrum Master** : Il est responsable de l'efficacité de l'équipe Scrum en l'aidant à améliorer la façon dont l'équipe travaille ensemble pour créer de la valeur sur une base continue. Lucas et Olivier ont assumé ce rôle en prenant connaissance de l'évolution du projet tout en s'intéressant à nos interactions en tant qu'équipe.
- **Product Owner** : Il est responsable de définir les besoins, les fonctionnalités et les priorités du produit à développer. Lucas et Olivier ont assumé ce rôle en veillant à ce que les jalons prédéfinis soient respectés et en nous guidant vers les taches à prioriser.
- **Développeurs** : Les développeurs se concentrent sur la conception, la construction, les tests et la livraison du produit (logiciels ou d'autres types de produits). Ils sont responsables de la création du plan du Sprint et du Sprint Backlog. Moussa, Ingrid et moi avons joué le rôle des développeurs.

En tant que membres de l'équipe de développement, nous avons mis en pratique des méthodes organisationnelles tout en utilisant des outils facilitant la collaboration, tels que :

- **Réunions régulières** : Les réunions périodiques de l'équipe servent à évaluer notre progression collective et à assurer une communication constante. Plus spécifiquement, du 4 au 7 décembre, nous avons organisé des réunions quotidiennes (Daily) pour partager nos avancements et les défis rencontrés, dans le but d'harmoniser nos connaissances en vue de la rédaction du dossier technique.
- **Sprints** : Le travail est divisé en parties distinctes et délimitées dans le temps, appelées « sprints ». Chaque sprint a ses propres objectifs et tâches assignées et attribuées à chaque développeur, suivies de retours et d'échanges pour améliorer les processus.
- **Scrum board** : Nous avons utilisé OneNote comme un tableau Scrum pour suivre et organiser les progrès des sprints et des diverses tâches.

Dans la réalisation de nos livrables, nous avons adopté des méthodes agiles qui favorisent la souplesse dans la livraison des résultats :

- Approche incrémentale : Le livrable est construit progressivement par ajout d'éléments successifs, permettant des livraisons plus rapides de fonctionnalités.

- Aspect itératif : Le développement est mené par itérations successives, offrant l'opportunité de recueillir les retours des utilisateurs pour les intégrer au processus de développement.

- Collaboration : Les équipes travaillent en étroite collaboration, favorisant ainsi la réduction des délais et l'amélioration globale de la qualité du produit final.

Le DevOps est également une méthodologie permettant de favoriser la communication et la collaboration.

3.2.2. Principes de déploiement DevOps

Le DevOps combine le développement (Dev) et les opérations (Ops) pour augmenter l'efficacité, la rapidité et la sécurité du développement ainsi que de la livraison continue de logiciels. Il s'agit d'une extension des approches Agiles. Cependant, durant la formation, nous nous sommes concentrés sur le DataOps, Il s'agit d'une méthodologie de gestion des données dont l'objectif est d'améliorer la communication, l'intégration et l'automatisation des flux de données entre les gestionnaires et consommateurs de données au sein d'une organisation. Cette méthodologie est l'union entre l'Agilité, le DevOps et le Lean Management.
Nous avons abordé ce module vers la fin de notre formation, soit à la fin du mois de novembre. Par conséquent, nous n'avons pas eu l'occasion d'appliquer intentionnellement cette méthode lors de la réalisation de notre projet.

Le DataOps se base sur des acteurs, des rôles, des principes et des outils.

Les acteurs et les rôles : Les membres de l'équipe Data Ops étaient Moussa, Ingrid et moi-même. Bien que nous ayons tous occupé le poste d'ingénieur Data pour la construction du pipeline de données, nous avons réparti les tâches en fonction de nos préférences et compétences, compte tenu du temps limité pour le projet.

- Moussa, Analyste BI : Principalement impliqué dans l'analyse des données, la génération de rapports et de graphiques.
- Ingrid, Architecte Data : Responsable de l'évaluation du système de gestion des données, de la collecte des besoins métiers et de la planification des processus.
- Moi-même, Data Scientist : Chargé de la transformation et de la construction des pipelines de données, ainsi que de l'implémentation des modèles de Machine Learning.

Les principes : Le DataOps se base plusieurs valeurs, à savoir : la fourniture continue des données aux utilisateurs finaux, l'innovation, l'évaluation constante de la qualité des données, l'amélioration continue de l'intégration et du déploiement, ainsi que la collaboration. Ces valeurs sont mises en avant à travers ces quatre principes (cf Figure 1) :

Principes DataOps	Orchestration des données et automatisation des taches	Qualité des données	Surveillance des données	Gouvernance des données
Application au sein du projet	Pas d'automatisation : Nous avions prévu une optimisation en créant des environnements via Docker Compose ainsi qu'un processus automatisé d'extraction et stockage hebdomadaire des données ANFR	La gestion de la qualité des données est assurée par nos jobs ETL (Talend). Une fois stockées dans la base de données cible, les données sont garanties d'être de qualité et prêtes à être utilisées.	Surveillance manuelle à l'aide du data profiling et de la détection d'anomalies sur nos tableaux de bord pour garantir la qualité et l'intégrité de nos données. Dans l'idéal, la mise en place d'un système d'alerte et de suivi (rapports, tableaux de bords en temps réel)	La gouvernance des données inclut la définition de rôles clairs au sein de l'équipe, favorisant ainsi une collaboration rapide et fluide

Tableau 1: Les principes DataOps et leur application au sein du projet

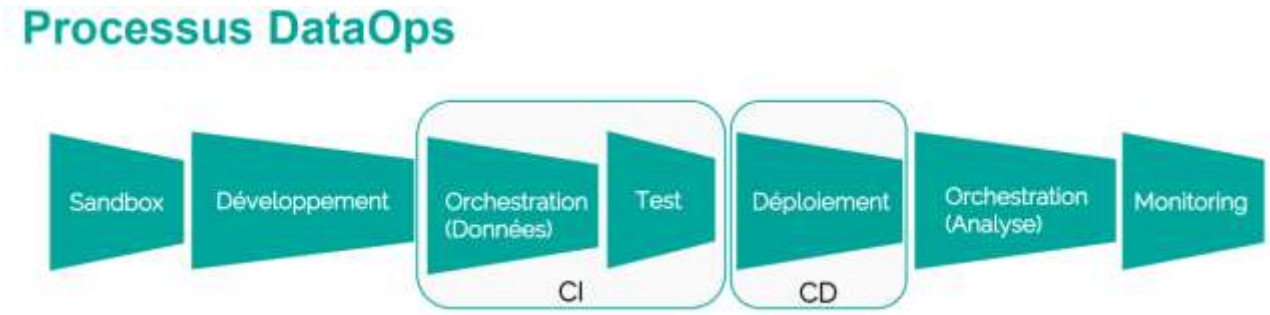


Figure 1: Processus DataOps (Source : Module DevOps (ADN))

Cela concerne les outils, nous avons mis en place un pipeline complet qui gère les données depuis la collecte jusqu'à la restitution.

Collecte des données	Stockage des données	Intégration des données	Gouvernance des données	Analyse des données
Manuelle puis stockage dans MariaDB	PostgreSQL (Données cibles)	Talend Open Studio for Data Integration	N/A	Power BI & Jupyter

Tableau 2: Outils utilisés dans le cadre du DataOps

Nous n'avons pas utilisé **d'outils de déploiement** dans ce projet. Cependant, nous sommes conscients de l'existence d'outils pour les projets futurs, tels que Bash, git et Kubernetes.
Il n'y a pas eu l'utilisation **d'outils spécifiques pour l'orchestration ou les tests**. Cependant, nous avons exploité Teams et OneDrive pour la **collaboration** au sein de l'équipe.

Après avoir établi notre organisation, abordons maintenant la phase de mise en œuvre du projet.

4. Réalisation du projet

Pour mener à bien le projet, nous avons sélectionné l'architecture Big Data appropriée, puis nous avons traité les données en accord avec cette architecture choisie.

4.1. Architecture

L'architecture de big data fait référence à la structure logique et physique qui dicte la manière dont les données massives sont ingérées, traitées, stockées et gérées de façon optimale.
Ci-dessous se trouve notre architecture. Elle compte quatre grandes zones :

- Sources : l'ensemble de nos ensembles de nos données
- Ingestion : le stockage et le traitement de nos données
- Analyses : Les analyses prédictives avec des modèles de Machine Learning
- Visualisation : La Business Intelligence pour valoriser les données

Dans les parties suivantes, les choix d'outils et de technologies seront détaillés.

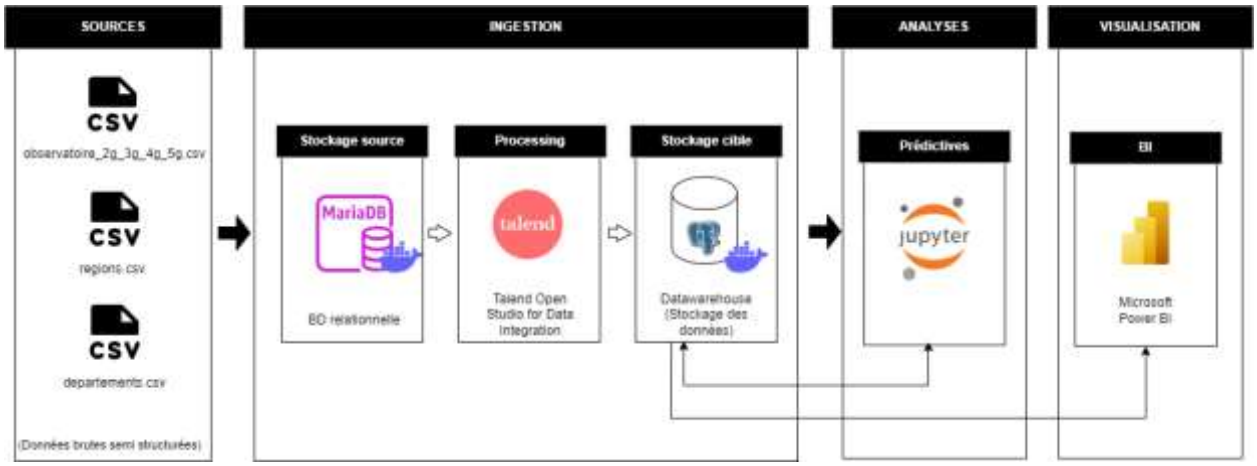


Figure 2: Architecture Big Data

4.2. Collecte et Stockage des Données

4.2.1. Données sélectionnées

Afin de mener à bien notre projet, nous avons récolté des données qualifiées de « open data ». Pour être plus explicite, il s'agit de données ouvertement accessibles, exploitables, modifiables et partagées par quiconque et à n'importe quelle fin.

ANFR : Données sur les installations des antennes des réseaux mobiles en France et DROM
Notre première source de données provient du portail Open data de l'ANFR. Ce portail a pour objectif de répertorier toutes les données rendues publiques par l'ANFR et disponibles sous licence libre. Il offre à l'ensemble de la communauté des utilisateurs (individus, radio-amateurs, jeunes entreprises, collectivités publiques, etc.) l'accès aux informations de base concernant les fréquences en France.
Description : Le jeu de données sélectionné porte sur les réseaux mobiles en France et DROM. Pour être plus explicite, il s'agit de données sur les installations des réseaux mobiles télécoms mises à jour hebdomadairement.
Format : Les données récoltées sont stockées dans un fichier de format csv
Type de données : Les données sont de type semi-structurées. Il s'agit d'un type de données flexible, entre les données non structurées et les données structurées. Étant stockées dans un fichier csv, elles ne suivent pas le format d'un modèle de données tabulaires ou de bases de données relationnelles (structurées). Autrement dit, elles n'ont pas de schéma fixe. Elles sont séparées par une virgule. Néanmoins, les données ne sont pas non structurées car elles contiennent certains éléments structurels tels que des métadonnées organisationnelles (nom d'attributs, etc.) qui facilitent leur analyse et compréhension.
Nombre d'enregistrements : 187442
Nombre d'attributs : 22
Dernière date de mise à jour : 24 novembre 2023
Flux de données : Pour mener à bien ce projet, nous avons opté pour l'utilisation d'un fichier CSV téléchargé et stocké sur notre espace de travail. Bien que l'idée d'établir un flux de données batch hebdomadaire ait été envisagée, le temps nécessaire pour sa mise en place était insuffisant. Ce flux de données faisait partie de notre liste d'optimisations à implémenter une fois le projet lancé. Nous nous sommes appuyés sur ce fichier en attendant d'intégrer ultérieurement un flux de données régulier pour enrichir notre processus.

Afin d'explorer les données, nous avons dressé un dictionnaire de données (simplifié). Il s'agit d'un répertoire de métadonnées qui offre des informations clés permettant d'interpréter le contenu d'un ensemble de données. Le dictionnaire contient souvent les informations concernant la longueur (nombre de caractères), les valeurs acceptées dans le champ, les relations et dépendances, l'emplacement des données, le propriétaire ou responsable du fichier, règles de validation de la qualité des données ou contraintes propres au champ (ex : > Date du jour, etc..).

	Attributs	Libellé	Type
1	id	Numéro d'identification unique d'une installation d'antenne de téléphonie mobile	Numérique
2	adm_lb_nom	Nom d'un opérateur de réseau de téléphonie mobile	Texte
3	sup_id	Numéro d'identification unique d'un support d'une antenne de téléphonie mobile	Numérique
4	emr_lb_systeme	Bandes de fréquences utilisées dans les réseaux de télécommunications mobiles	Texte
5	emr_dt	Date d'installation d'une antenne de téléphonie mobile	Date
6	sta_nm_dpt	Numéro du département dans lequel est installée une antenne de téléphonie mobile	Numérique
7	Code_insee	Code Insee de la zone géographique (commune, canton, etc..) dans laquelle est installé l'antenne réseau	Texte
8	generation	Génération de technologie de communication mobile	Texte
9	date_maj	Date d'extraction des données	Date
10	sta_nm_anfr	Numéro d'identification d'une antenne de téléphonie mobile (anfr)	Texte
11	nat_id	Numéro d'identification national d'une antenne de téléphonie mobile(anfr)	Numérique
12	sup_nm_haut	Hauteur de l'antenne sur son support par rapport au sol	Numérique
13	tpo_id	Numéro d'identification TPO	Numérique
14	adr_lb_lieu	Nom du lieu d'installation de l'antenne réseau	Texte
15	adr_lb_add1	Adresse du lieu d'installation de l'antenne réseau (pt 1)	Texte
16	adr_lb_add2	Adresse du lieu d'installation de l'antenne réseau (pt 3)	Texte
17	adr_lb_add3	Adresse du lieu d'installation de l'antenne réseau (pt 3)	Texte
18	adr_nm_cp	Code postal du lieu d'installation de l'antenne de téléphonie mobile	Numérique
19	com_cd_insee	Code Insee de la zone géographique dans laquelle est installé l'antenne de téléphonie mobile	Texte
20	Coordonnees	Coordonnées géographiques du lieu d'installation de l'antenne réseau (format "degrés décimaux")	Texte
21	Coord	Coordonnées géographiques du lieu d'installation de l'antenne réseau (format "degré, minutes et secondes")	Texte
22	statut	Statut du support réseau	Texte

Tableau 3: Dictionnaire de données : observatoire5G (BD source)

INSEE : Données relatives aux départements de France (Métropole et DROM)

L'INSEE (Institut National des Statistiques et des Études Économiques) offre un accès libre à ses données sur son site web officiel. Afin de visualiser géographiquement l'emplacement des antennes de téléphonie mobile, nous avons cherché à associer des données géographiques (départements et régions) à celles dont nous disposions déjà. Pour ce faire, nous avons utilisé les données officielles de l'INSEE

Description : Ce jeu de données contient des informations concernant les populations légales des départements en vigueur au 1er janvier 2017 en France. Ces données ont été mises à jour en mars 2017.

Les départements recensés sont ceux de la France métropolitaine, départements d'outre-mer de la Guadeloupe, de la Guyane, de la Martinique et de La Réunion, limites territoriales en vigueur au 1er janvier 2016.

Le jeu de données sélectionné contient à la fois les informations relatives aux départements et aux régions. Elles ont été enregistrées dans un fichier contenant plusieurs feuilles distinctes

Format : Les données récoltées sont stockées dans un fichier de format xls. Nous avons entrepris une démarche de segmentation du fichier en sélectionnant la feuille dédiée aux départements pour créer un fichier csv spécifique à ces derniers. De même, nous avons procédé de manière similaire pour les régions en extrayant les données pertinentes vers un fichier csv distinct.

Type de données : Les données sont de type semi-structurées

Nombre d'enregistrements : 101

Nombre d'attributs : 9

Dernière date de mise à jour : mars 2017

Flux de données : Les données géographiques sont souvent statiques et peu actualisées à moins qu'il y ait des événements significatifs ou des modifications légales majeures ayant un impact sur ces données spécifiques. Effectivement, les informations les plus récentes disponibles remontent à l'année 2017, ces données font suite à la loi de novembre 2015, portant sur la Nouvelle organisation territoriale de la République qui « confie de nouvelles compétences aux régions et redéfinit clairement les compétences attribuées à chaque collectivité territoriale »[1].

Ainsi, nous travaillons avec des données récupérées et chargées sur notre environnement de travail. Ci-contre se trouve le dictionnaire des données :

	Attributs	Libellé	Type
1	id_deprtmt	Numéro d'identification unique d'un département	Numérique
2	code_region	Numéro d'identification de la région à laquelle appartient un département	Numérique
3	nom_region	Nom d'une région à laquelle appartient un département	Texte
4	code_deprtmt	Numérotation standardisée d'un département	Texte
5	nom_deprtmt	Nom d'un département	Texte
6	num_arrndssmt	Nombre d'arrondissements d'un département	Numérique
7	num_cantons	Nombre de cantons d'un département	Numérique
8	num_communes	Nombre de communes d'un département	Numérique
9	ppl_mncpl	Population totale des municipalités d'un département	Numérique
10	ppl_ttl	Population totale d'un département	Numérique

Tableau 4: Dictionnaire de données - départements (BD source)

[1] Gouvernement.fr. « La réforme territoriale ». Consulté le 6 décembre 2023. https://www.gouvernement.fr/action/la-reforme-territoriale.

INSEE : Données relatives aux régions de France (Métropole et DROM)

Description : Ce jeu de données fournit des informations relatives aux populations légales des régions en vigueur au 1er janvier 2017. Ces données ont été mises à jour en mars 2017. Les régions concernées sont celles de la France métropolitaine, des départements d'outre-mer de la Guadeloupe, de la Guyane, de la Martinique et de La Réunion, limites territoriales en vigueur au 1er janvier 2016.

Format : Les données récoltées sont stockées dans un fichier de format xls. Nous avons entrepris une démarche de segmentation du fichier en sélectionnant la feuille dédiée aux régions vers un fichier csv distinct tel que cela a été fait pour les départements.

Type de données : Les données sont de type semi-structurées

Nombre d'enregistrements : 13

Nombre d'attributs : 7

Dernière date de mise à jour : mars 2017

Flux de données : Ainsi, nous travaillons avec des données récupérées et chargées sur notre environnement de travail.

Ci-contre se trouve le dictionnaire des données :

	Attributs	Libellé	Type
1	id_region	Numéro d'identification unique d'une région	Numérique
2	code_region	Numérotation standardisée d'une région	Numérique
3	nom_region	Nom d'une région	Texte
4	num_arrndssmt	Nombre d'arrondissements d'une région	Numérique
5	num_cantons	Nombre de cantons d'une région	Numérique
6	num_communes	Nombre de communes d'une région	Numérique
7	ppl_mncpl	Population totale des municipalités d'une région	Numérique
8	ppl_ttl	Population totale d'une région	Numérique

Tableau 5: Dictionnaire des données - régions (BD source)

Avant de commencer toute manipulation ou traitement des données, il est important de s'informer sur la gestion responsable et conforme de ces données, conformément à la législation visant à assurer leur protection.

4.2.2. Politique de Données

Pour être brève, la politique des données représente un ensemble de règles et de directives décidées par une entreprise ou une organisation. Elle explique comment les données sont manipulées :

Il existe à ce jour une panoplie de réglementation régissant la manipulation des données. La CNIL (Commission nationale de l'informatique et des libertés) garantit le respect du RGPD (Règlement général sur la protection des données). L'un des règlements les plus connus dans le monde du Big Data. Il assure la protection des données personnelles, c'est à dire, « toute information se rapportant à une personne physique identifiée ou identifiable[2] ».

Dans notre projet, nous n'avons affaire à aucune donnée personnelle, ce qui nous exclut du champ d'application du RGPD.

[2] « Donnée personnelle ». Consulté le 6 décembre 2023. https://www.cnil.fr/fr/definition/donnee-personnelle.

Les données sélectionnées proviennent de sources officielles qui ont ouvert librement l'accès à leurs données. Ces données sont diffusées selon une licence ouverte. Par conséquent, elles sont disponibles et utilisables par tous sans aucune entrave. Nous avons donc la possibilité de manipuler et d'analyser ces données sélectionnées sans aucune restriction dans le cadre de notre projet.

Une fois la politique de données vérifiée et analysée, nous passons au stockage des données. Le choix de notre solution repose sur notre volume de données, ainsi que sur nos futures analyses et manipulations.

3.4. Stockage des données

Le choix de notre outil de stockage de données est fortement corrélé à la structure, l'évolutivité de nos données sources et le traitement des données cibles. Nous avons le choix entre un système de gestion des données relationnelles (SGBDR) ou une base de données NoSQL.

- En ce qui concerne la **structure des données** sources, elles sont stockées dans des fichiers CSV. Avec un script d'extraction, il est facile de transférer les données dans un SGBDR. Ce système de stockage garantit une structure et un schéma aux données. Cela est en phase avec les traitements finaux que nous souhaitons faire. Nous avions envisagé d'ajouter de nouvelles sources de données non-structurées en guise d'optimisation du projet. Dans l'idéal, il aurait été judicieux d'opter pour une BD NoSQL.
- En ce qui concerne **l'évolutivité des données** sources, nous nous sommes limités à environ 200 000 lignes avec des données dont le schéma ne varie pas. Nos données sont stockées sur nos postes de travail dans des fichiers. Un SGBDR est optimal pour ce type de données. Notons cependant que les données peuvent évoluer, dans le cas où nous choisissons de traiter un plus grand volume de données. Il est important de faire le choix d'une BD répondant positivement à l'évolutivité. Pour les données à grande échelle ou lorsque le schéma n'est pas fixe, les bases de données NoSQL peuvent être plus appropriées en raison de leur évolutivité et de leur flexibilité.
- En ce qui concerne les traitements, notre objectif final est l'analyse des données. Cela nécessite des données structurées. Les solutions d'entreposage de données (Data Warehouse) pour les données cibles sont bénéfiques en raison de leurs performances optimisées en matière d'interrogation.

3.4.1. Système de stockage des données

Nous faisons donc le constat que nous avons besoin d'une base de données relationnelle pour stocker nos données sources et cibles. **Nous avons étudié deux solutions principales : MariaDB pour les données sources et PostgreSQL, pour les données cibles**.

Le choix de MariaDB et Postgres pour notre projet trouve sa justification dans plusieurs raisons fondamentales :

- La **performance et accessibilité** : MariaDB et Postgres, des bases de données open source, sont réputées pour leurs performances optimales et leur aptitude à traiter efficacement de vastes ensembles de données.
- L'**évolutivité** : Ces deux systèmes offrent une grande capacité d'évolutivité, ce qui facilite leur adaptation aux besoins en constante évolution de l'institut de sondage, surtout en matière d'expansion des données.

- La **fiabilité** : MariaDB et Postgres sont reconnues pour leur fiabilité, conçues pour maintenir la cohérence des données et résister aux pannes potentielles.
- La **sécurité des données** : Ces bases de données intègrent des fonctionnalités de sécurité robustes pour protéger la confidentialité des données, prévenir tout accès non autorisé et garantir la sécurité des informations stockées.

D'autres alternatives, telles que MongoDB, Apache Cassandra et Oracle SQL Developer, ont été évaluées mais écartées :

- MongoDB : Cette BD document est reconnue pour sa flexibilité et sa capacité à gérer des données non structurées. Bien que flexible, cette option n'était pas adaptée car les données à traiter dans le cadre de ce projet sont structurées.
- Apache Cassandra : Réputée pour sa scalabilité et sa résilience en tant que base de données NoSQL open source. Bien qu'évolutive, elle ne répond pas à notre besoin, similairement à MongoDB. Elle ne garantit pas les mêmes performances et fiabilité que MariaDB et Postgres.
- Oracle SQL Developer : Connu pour sa robustesse et sa sécurité en tant que base de données relationnelle propriétaire. Cependant, elle n'est pas adaptée à cause de sn cout.

3.4.2. Environnement de travail

Pour installer et exécuter nos bases de données, nous avons exploré plusieurs environnements : **machines virtuelles, containers Docker et notre machine physique**. Dans notre analyse, nous avons privilégié deux options principales : Docker et les machines virtuelles.

Initialement, nous avons **exclu l'utilisation de nos machines physiques**, notamment en raison de notre désir de mettre en pratique les connaissances acquises lors de la formation, comme l'utilisation des environnements Docker. De plus, les contraintes matérielles, le manque de flexibilité et l'absence d'isolation des environnements de travail étaient des facteurs contraignants. Les exécutions de bases de données pouvaient potentiellement entrer en conflit avec d'autres logiciels ou configurations existants sur la machine.

L'option d'installer MariaDB et PostgreSQL dans des containers Docker offre de multiples avantages pour notre projet de plateforme Big Data :

- **La simplicité et facilité de configuration** : L'installation et la configuration de MariaDB et PostgreSQL sous Docker sont simples. Télécharger les images Docker adaptées et les lancer permet de se concentrer davantage sur le développement de l'application que sur la gestion des bases de données.

- **La portabilité** : Les images Docker peuvent être déployées sur différents environnements, qu'il s'agisse du cloud, du on-premise ou même sur des appareils mobiles. Cette flexibilité est cruciale pour le projet, offrant la possibilité de déployer la plateforme sur divers environnements.

- **L'optimisation des ressources** : Docker utilise efficacement les ressources par rapport aux machines virtuelles, réduisant ainsi les coûts opérationnels de la plateforme Big Data.

- **La sécurité** : Docker offre une isolation des environnements pour MariaDB et PostgreSQL, limitant ainsi les risques de compromission des données.

En revanche, la manipulation et la configuration des machines virtuelles sont plus complexes. Elles sont également moins portables, économiques et sécurisées.

En conclusion, nous avons opté **pour l'installation de nos bases de données, MariaDB et PostgreSQL**, dans des containers Docker pour bénéficier des avantages cités ci-dessus.

BASE DE DONNEES SOURCE

Une fois les bases de données installées, il est nécessaire d'effectuer l'insertion des données pour les stocker en vue de leur traitement ultérieur.

Pour MariaDB, la procédure requiert un schéma pour l'injection des données dans la base de données. Cela justifie la nécessité de dresser un modèle conceptuel de données (MCD). Ce modèle graphique fait partie de la méthodologie Merise et représente les entités, leurs attributs et les relations qui les lient. Son objectif principal est l'analyse structurelle de la base de données. Ci-dessous, vous trouverez le modèle conceptuel de notre base de données MariaDB.

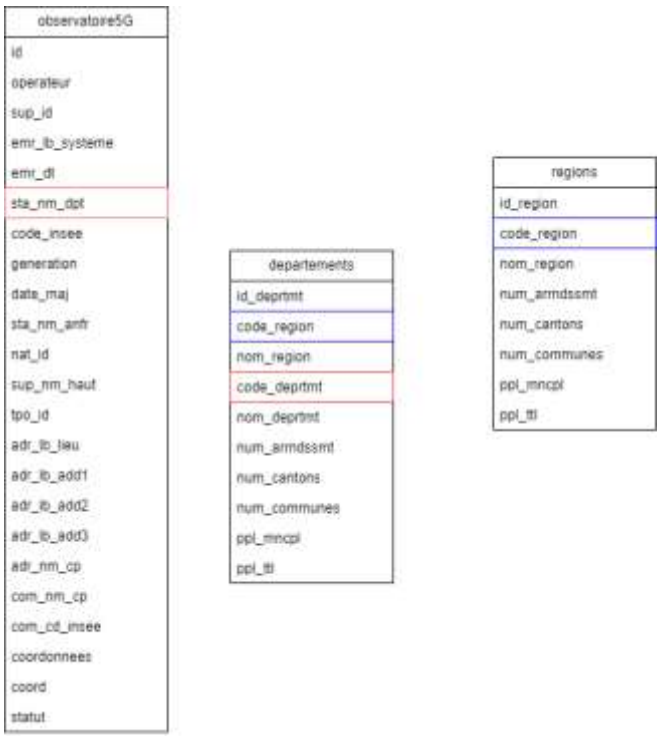


Figure 3: MDC DB MariaDB

Comme illustré sur la figure 3, l'on constate l'absence de relations entre les entités dans le modèle. Cette décision a été délibérée car nous avons opté pour l'utilisation exclusive de MariaDB en tant que solution de stockage des données.

Idéalement, notre modèle devrait plutôt ressembler à celui-ci de la figure 4, avec des relations entre l'entité "Observatoire5G" (représentant les antennes de téléphonie mobile) et l'entité "Départements", puis entre "Départements" et "Régions".

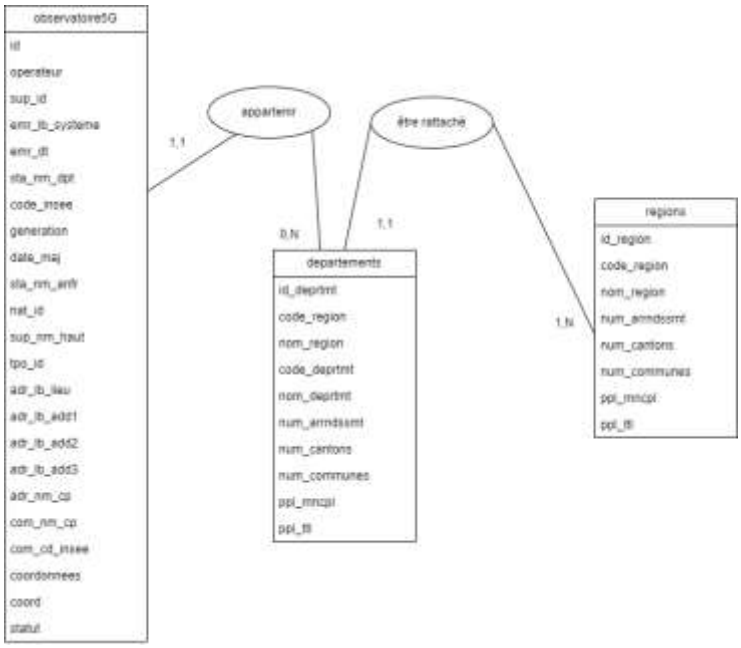


Figure 4: MCD idéale BD MariaDB

BASE DE DONNEES CIBLE

Les données traitées à la suite des opérations ETL sont stockées dans notre base de données cible. Voici le schéma représentant la structure de la table "Observatoire5G" située dans la base de données PostgreSQL.

Colonne	Clé	Type	Nullable	Modèle de date (C...	Longueur
id	☑	int	☐		10
operateur	☐	String	☐		255
sup_id	☐	int	☐		10
emr_lb_systeme	☐	String	☐		255
emr_dt	☐	Date	☐	"dd/MM/yyyy"	255
code_deprtmt	☐	Integer	☐		10
nom_deprtmt	☐	String	☐		255
code_region	☐	Integer	☐		10
nom_region	☐	String	☐		255
generation	☐	String	☐		255
date_maj	☐	Date	☐	"dd/MM/yyyy"	255
sta_nm_anfr	☐	String	☐		255
nat_id	☐	int	☐		10
sup_nm_haut	☐	int	☐		10
tpo_id	☐	int	☐		10
com_cd_insee	☐	int	☐		10
coordonnees	☐	String	☐		255
coord	☐	String	☐		255
statut	☐	String	☐		255

Figure 5: Schéma de la table Observatoire5G

Les données prévisionnelles mensuelles concernant le nombre d'installations d'antennes de décembre 2023 à novembre 2024 sont également enregistrées dans la base de données cible. Ces données sont générées à partir des modèles de Machine Learning, se présentant sous la forme de cinq tables (cf Figure 6 pour la liste des 5 tables), toutes suivant une structure similaire. Voici le dictionnaire de ces données.

	Attributs	Libellé	Type
1	id	Le numéro d'identification unique d'une prévision d'installation d'antennes	Numérique
2	dt	La date (MM-YYYY) d'une prévision d'installation d'antennes	Date
3	count_antennas	Le nombre prédit d'antennes installées à une date donnée	Numérique

Tableau 6: Dictionnaire des données - données prévisonnelles

Figure 6: Liste des 5 tables prévisionnelles

3.5. Intégration des données

3.5.1. Technologies choisies

Au cours de notre formation, nous avons exploré divers outils pour intégrer les données, tels que **dbt, Hadoop Spark, Spark Streaming et Apache Nifi**. Cependant, notre choix s'est finalement porté **sur Talend Open Source for Data Integration pour nos traitements**. Ce choix initial a été fait au début de la formation, lorsque nos options d'outils pour l'intégration des données étaient limitées.

Après avoir évalué les opérations clés nécessaires telles que la connexion aux sources de données, l'application de filtres, la réalisation de jointures et le stockage dans une base de données cible, nous avons conclu que Talend Open Studio for Data Integration serait la solution la plus adéquate. La familiarité de la majorité de notre équipe avec cet outil en a fait un choix pratique pour nos besoins.

Il est important de noter que Talend offre plusieurs avantages :

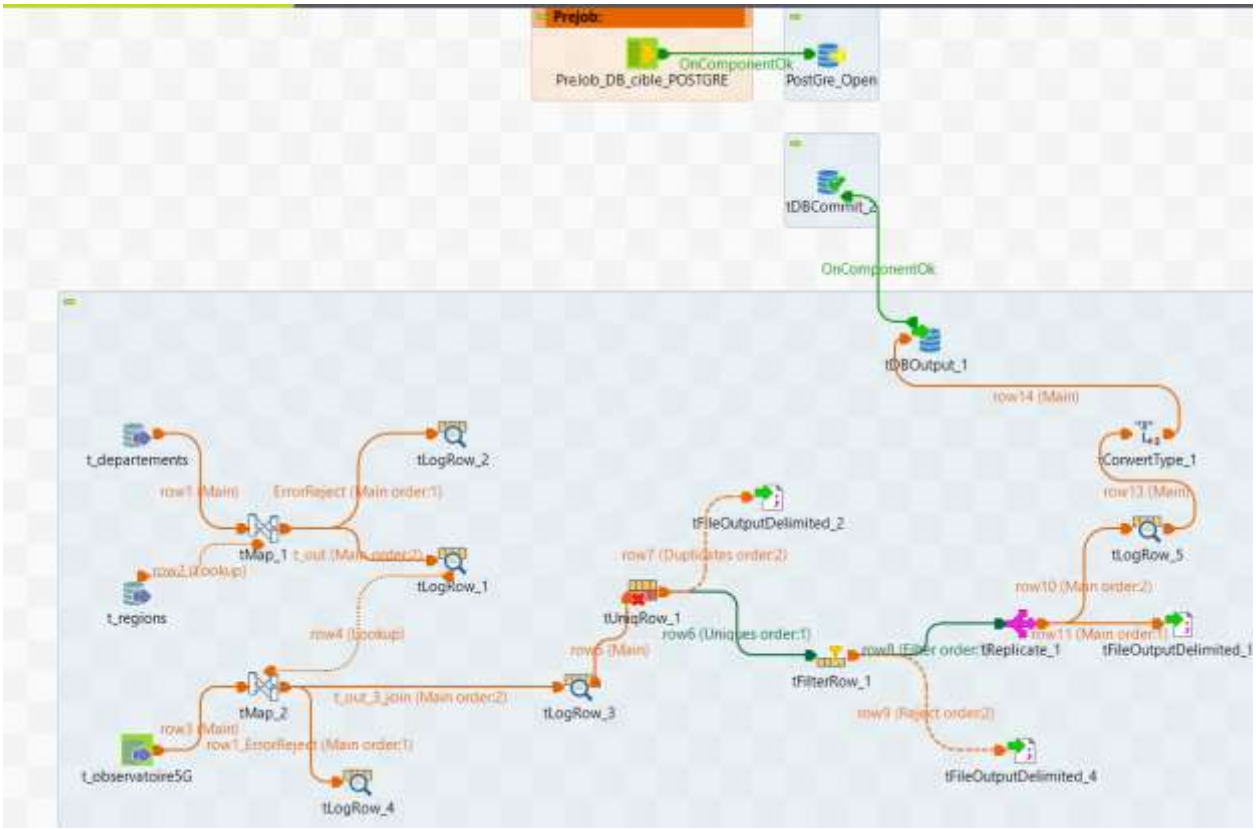
- **Une interface graphique intuitive**, équipée de composants prédéfinis pour les opérations courantes d'ETL, simplifiant ainsi la conception et l'assemblage des flux de données sans nécessiter de compétences avancées en programmation.
- La présence **de connecteurs multiples et prédéfinis** pour se connecter à diverses sources de données (bases de données relationnelles, fichiers plats, API, etc.), facilitant ainsi l'intégration avec différentes sources.

- Une **communauté active** ainsi qu'une documentation complète et riche.

Maintenant, passons à la coordination des composants pour atteindre notre objectif : obtenir des données de qualité pour nos analyses.

3.5.2. Processus d'intégration des données

Voici une vue globale de notre job d'intégration sur Talend

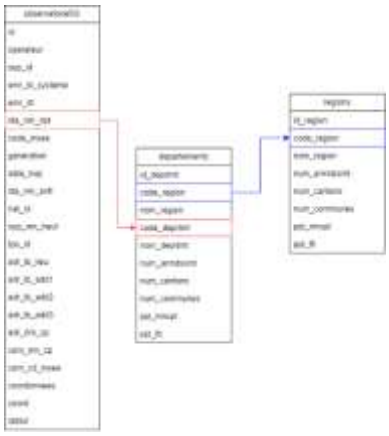


Phase 1: Connexion à la BD source
La première phase de notre processus ETL est l'extraction. Nous nous connectons donc à notre BD de données MariaDB pour extraire les données des tables Observatoire5G, departements, regions en utilisant un composant tDBInput.

Phase 2: Jointures
Les jointures entre les tables ont été effectuées avec Talend for Data Integration et non dans la BD source pour maintenir une certaine flexibilité et maintenabilité de nos données. Ce choix présente plusieurs avantages:

- Flexibilité : La jointure des tables dans Talend permet une plus grande flexibilité des analyses. Cela est dû au fait que la jointure peut être effectuée à tout moment, y compris après l'importation des données
- Maintenabilité: La jointure des tables dans Talend rend le code plus facile à maintenir. Cela est dû au fait que la jointure est effectuée dans un composant Talend, qui peut être facilement modifié ou remplacé.

Cependant, il est vrai que Talend n'est pas un moteur de base de données spécialisé dans les jointures et que la jointure doit être effectuée à chaque fois que le job talend est lancé. données sont consultées.



Une jointure est faite entre la table departements et regions avec le code INSEE des régions qui se trouvent respectivement dans chacune des tables.

Une seconde jointure est effectuée entre la table departements et la table observatoire5G, en utilisant le code départements INSEE que contiennent chacune des tables.

Les jointures ont été réalisées à l'aide du composant tMap pour combiner les tables tout en sélectionnant les attributs les plus pertinents à conserver dans les tables fusionnées. Le composant tFilterRow offre une alternative à celui précédemment utilisé.

Phase 3: Filtres

- Extraction des données uniques: En utilisant le composant tUniqRow, nous avons effectué une extraction pour vérifier la présence de doublons dans notre table de données. Les doublons ont été rejetés dans un fichier Excel pour une visualisation et une analyse ultérieures. Seules les données uniques ont été acheminées vers l'étape suivante du processus.
- Extraction des données non nulles: Avec le composant tFilterRow, nous avons réalisé une extraction des données non nulles.

Phase 4: Conversion

Lors de l'extraction de la base de données source MariaDB, les données temporelles, notamment les dates, ont été extraites sous forme de chaînes de caractères (string). Afin de normaliser ces données et de les rendre compatibles pour l'insertion dans notre base de données cible, nous avons prévu convertir leur type en format de date "dd/MM/yyyy". Cette opération est effectuée à l'aide du composant tConvertType.

Phase 5: Chargement dans la BD cible

Les données ont été préparées pour répondre aux exigences de qualité et correspondent aux critères requis pour les analyses. La dernière étape consiste à stocker ces données dans notre entrepôt de données, PostgreSQL. Pour ce faire, nous utilisons le composant tDBOutput, qui permet d'insérer les données dans la table "observatoirecible5G" tout en créant la table si elle n'existe pas. Enfin, l'ensemble du processus est validé par un commit pour enregistrer définitivement les insertions effectuées.

3.5.3. Qualité de Données

La qualité des données est incontournable afin de garantir la confiance dans les données consultées. L'on compte six critères majeurs qui en sont constitutifs :

Critère de qualité	Définition	Application
Fraicheur	accès à la version la plus récente des informations. Cela implique que les données sont constamment mises à jour et disponibles au moment opportun	Ce critère n'a pas nécessité d'attention particulière car les données ont été sélectionnées et stockées dans notre espace de travail.
Cohérence	respect d'un pattern : il s'agit d'évaluer si, attribut par attribut (ou champ par champ), les valeurs d'un jeu de donnée précis sont toujours saisies selon un même format (ou pattern)	Conversion des données temporelles au format date
Complétude	les attributs obligatoires pour identifier de manière unique un enregistrement à sa création, ne comptent aucune information manquante et/ou nulle	Filtre et rejet des données dont la longueur < 1
Exactitude	Considérer qu'une information est exacte, c'est se poser la question de savoir si elle reflète la réalité de ce qu'elle est vouée à représenter	Vérification des données géogrpahiques (nombre et noms des departements et régions et code INSEE sur le sites officiels)
Unicité	Il s'agit de pouvoir confirmer ou à minima évaluer le nombre d'enregistrements qui représente la même information	Filtre et rejet des doublons
Interopérabilité	Le critère d'interopérabilité couvre cette définition : il s'agit d'évaluer à quel point une donnée peut être (ou est déjà) échangée, partagée et partageable, dans un format clair, préférentiellement non propriétaire, entre plusieurs systèmes	L'utilisation de technologies d'intégration de données tel l'ETL faciliter l'échange et le partage des données entre divers systèmes (ex: BD source et cible); il est meme possible d'exporter les données dans un ficher CSV, JSON, etc..

3.6. Analyse et exposition des données

La dernière phase de notre stratégie est celle de l'analyse des données traitées et prêtes à être exploitées. Il s'agit d'une des phases les plus importantes pour répondre à notre problématique.
Nous avons procédé à une phase d'analyse prédictive et une phase de visualisation.

3.6.2. Analyse prédictive

L'analyse prédictive implique l'utilisation de données afin de prédire des données futures. Cette méthode exploite l'analyse des données et l'intelligence artificielle pour détecter des tendances qui permettent de projeter des comportements futurs. Pour ce faire, nous avons opté pour des modèles de Machine Learning (ML).
Le Machine Learning, apprentissage automatique en français, est une branche de l'intelligence artificielle. Il se concentre sur la capacité des machines à apprendre à partir de données en utilisant des modèles mathématiques. La machine apprend donc à détecter des tendances et des schémas. Cela revient donc à tirer des informations importantes à partir d'un ensemble de données d'entraînement.

L'objectif principal est de configurer les paramètres d'un modèle pour qu'il puisse prédire au mieux à partir des données sur lesquelles il a fait son apprentissage. Une fois le processus d'apprentissage terminé, le modèle peut être utilisé dans des situations réelles.
Pour ce projet, notre objectif est la prédiction du nombre probable d'antennes susceptibles d'être installées dans l'année à venir.

3.6.2.1. JUSTIFICATION DU TYPE DE MODELES DE ML

Pour atteindre notre objectif, nous avons opté pour l'utilisation de modèles spécifiquement conçus pour la prédiction des données temporelles, ce que l'on appelle communément les "séries temporelles", contrairement aux modèles de ML classiques. Ces modèles sont adaptés pour analyser et anticiper les tendances ou les variations dans un ensemble de données qui évoluent au fil du temps, telles que les installations d'antennes de téléphonie mobile.
Les modèles classiques de machine learning peuvent être utilisés pour traiter une grande variété de données. Les principales différences entre les deux approches sont les suivantes :

- **La nature des données** : Les modèles de prédiction sur des séries temporelles sont conçus pour traiter des données temporelles, qui sont des données qui évoluent au fil du temps. Les modèles classiques de machine learning peuvent être utilisés pour traiter une grande variété de données, y compris les données temporelles.
- **Les objectifs de modélisation** : Les modèles de prédiction sur des séries temporelles sont généralement utilisés pour prédire des valeurs futures à partir de données historiques. Les modèles classiques de machine learning peuvent être utilisés pour une grande variété d'objectifs de modélisation, y compris la classification, la régression et l'apprentissage par renforcement.
- **Les méthodes de modélisation** : Les modèles de gestion des séries temporelles utilisent des méthodes spécifiques qui sont adaptées aux données temporelles. Les modèles classiques de machine learning utilisent une variété de méthodes, y compris les méthodes d'apprentissage automatique supervisé et non supervisé.

En résumé, les modèles classiques de machine learning sont généralement utilisés pour des tâches comme la classification d'images, les recommandations de produits, ou encore la recherche d'informations sans prendre en compte la dimension temporelle. En revanche, les modèles

spécialement conçus pour les séries temporelles sont capables de prédire des valeurs, car ils prennent en compte le facteur temporel. Ces types de modèles excellent dans des tâches telles que la prédiction des ventes, des prix, des conditions météorologiques ou encore du nombre d'installations d'antennes dans l'avenir, ce qui correspond à notre objectif.

3.6.2.2. MODELES SELECTIONNES

Parmi la variété de modèles disponibles et dédiés aux séries temporelles, nous en avons écarté quelques uns:

Modèles de régression : Ce modèle utilise des techniques de régression classiques pour prédire les valeurs futures en se basant sur des variables explicatives et des caractéristiques temporelles. Cependant, dans le cadre de ce projet, nous avons écarté l'utilisation de ce modèle car notre intention est d'utiliser un modèle spécifique et adapté aux prévisions avec des séries temporelles.

Moyenne mobile (Moving Average, MA) : Ce modèle se fonde sur la moyenne des observations passées dans une fenêtre temporelle pour prédire les valeurs futures. Actuellement, les modèles de prédiction ne se limitent plus à la simple utilisation de la moyenne mobile. La plupart de ces modèles sont intégrés à des modèles d'autorégression comme ARMA ou ARIMA. C'est pourquoi ce modèle a été écarté de notre sélection.

Autoregressive (AR) : Ce modèle repose sur les valeurs passées d'une série temporelle pour prédire les valeurs futures en utilisant une combinaison linéaire de ces observations. Actuellement, les modèles AR sont incorporés dans les modèles ARIMA et sont associés à des modèles de moyenne mobile (MA). Par conséquent, il est plus pertinent d'exploiter ce type de modèle car il offre une approche combinée qui tient compte des variations et tendances temporelles pour des prévisions plus précises.

Réseaux de neurones récurrents (Recurrent Neural Networks, RNN) : Ces modèles de deep learning sont conçus pour traiter les données séquentielles en prenant en compte les dépendances temporelles. Pour des séquences très longues, les RNN peuvent rencontrer des difficultés à conserver les informations antérieures, ce qui peut impacter leurs performances sur certaines prédictions ou analyses de séries temporelles.

Nous avons choisi d'utiliser ces trois modèles:

LSTM **(Long Short-Term Memory, LSTM) :** Une architecture de RNN (Recurrent Neural Network) améliorée, capable de conserver des informations à long terme.

ARIMA (Autoregressive Integrated Moving Average): Un modèle qui combine les composants de moyenne mobile, de moyenne mobile intégrée et de modèle autorégressif pour la prédiction des séries temporelles.

Prophet: Prophet est un modèle de séries temporelles développé par Facebook, conçu pour faciliter la prévision des données temporelles avec des tendances saisonnières. Ce modèle est conçu pour être simple à utiliser tout en étant efficace pour la prévision des données temporelle

Une fois les modèles choisis, il convient de les utiliser sur un environnement adapté.

3.6.2.3. ENVIRONNEMENT CHOISI

Les modèles de machine learning sont implémentés en utilisant le langage de programmation Python. Pour travailler avec ces modèles, un environnement de développement intégré (IDE) dédié à Python est nécessaire pour écrire, exécuter et déboguer le code.
Nous avons donc fait une étude sur deux environnements couramment utilisés par les Data Scientists pour effectuer des prédictions et prévisions:

- Google Collaboratory: **Google Collaboratory:** Google Colab est une plateforme destinée au développement et à l'exécution du code Python facilement et rapidement sans aucune installation à travers un navigateur web.Cette plateforme est notamment utilisée pour effectuer des traitements de Machine Learning, et l'analyse des données
- Jupyter Notebook: Jupyter est une application interactive basée sur le web. Jupyter permet de créer et partager des documents (notebooks) contenant du code, des visualisations et du texte. Sur cette application, il est possible d'exécuter des programmes écrits avec plus de 40 langages de programmation, dont Python, R et Julia.

	Google Colaboratory	Jupyter Notebook
Avantages	• Gratuit: La seule condition est de disposer d'un compte Google • Plateforme Cloud : Tout le traitement est effectué sur les serveurs de Google, ce qui facilite l'utilisation sur n'importe quel appareil et permet d'économiser la mémoire du PC. • Accès à des GPU et TPU gratuits (mais limité): Utilisation d' un GPU gratuitement pendant 12 heures maximum, • Collaboration: en temps réel & partage du notebook • Accès à une variété de librairies Python: Bibliothèques intégrées, telles que NumPy, pandas, matplotlib, et bien d'autres encore, qui doivent être importée pour être utilisées)	• Open source et gratuit • Application de bureau: Installation sur la machine locale (on premise) • Accès à une variété de librairies Python (NumPy, pandas, matplotlib, et bien d'autres encore, qui doivent être importée pour être utilisées)

Inconviénients	• Fichiers stockés sur Google Drive (peu sécurisé) • Installation des paquets utiles à nos traitements à chaque session	• Utilisation du GPU de la machine locale: Utilisation des ressources informatiques de la machine locale (limiter la taille et la complexité des tâches à effectuer si la machine n'est pas assez performante) • Collaboration not by design: utilisateurs partagent manuellement leurs notebooks, soit en envoyant des fichiers, soit en les hébergeant sur un serveur partagé

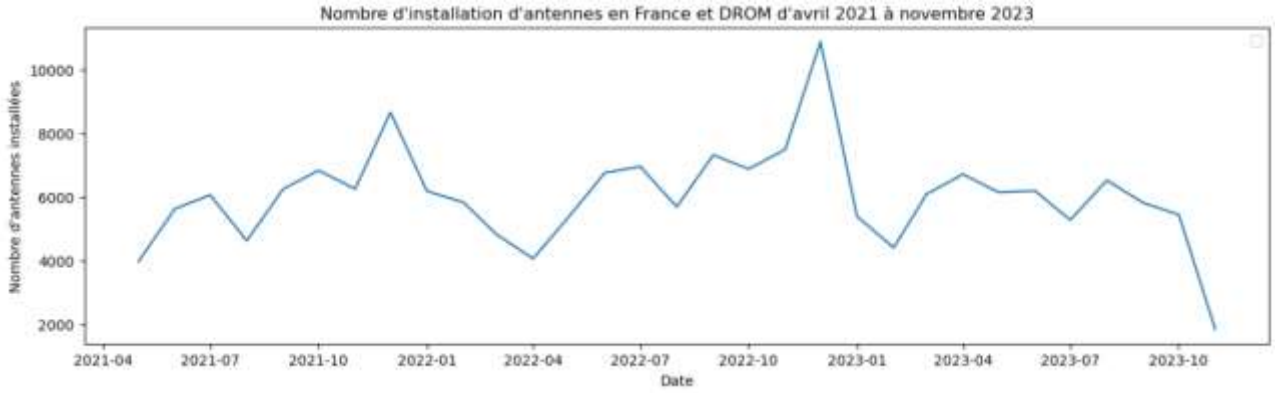
En raison de nos limitations en termes de capacité de RAM et de GPU sur nos machines, ainsi que de notre travail en équipe, nous avons recherché une plateforme permettant un partage efficace et une visualisation en temps réel de l'avancée du projet. Bien que Google Collaboratory réponde à la plupart de nos besoins, nous avons opté pour Jupyter Notebook en raison de sa connectivité directe et de son accès à notre base de données PostgreSQL en local. Contrairement à Colab, qui est un service cloud, Jupyter fonctionne en local et peut également se connecter à la base de données PostgreSQL, ce qui correspond davantage à nos besoins spécifiques.
C'est ainsi que nous nous sommes connecté à notre BD PostgreSQL pour en extraire les données et commencer les prédictions.

Pour mener à bien les prédictions, nous nous sommes servis de plusieurs librairies dont: Pandas, Numpy, Tensorflow, Keras, Psycopg2, slalchemy, matplotlib, et sklearn.

3.6.2.4.INTERPRETATION DES PREVISIONS

Pour rappel, nos données portant sur les installations d'antennes étaient journalières. Afin de faciliter l'apprentissage et les prévisions, nous avons compté par mois, le nombre d'antennes installées. Cela constituait notre feature de base, en plus des dates en tant qu'index dans nos dataframes.
Notons également que nous avons travaillé avec quatre dataframes: l'un contenant les données générales de tous les opérateurs, et les autres filtrés et dédiés respectivement aux 4 opérateurs principaux en France: SFR, Orange, Bouygues et Free Mobile sur la période d'avril 2021 à novembre 2023. Dans ce dossier technique, je me contenterai de commenter les tendances sur les données générales avec tous les opérateurs confondus.

Voici la tendance générale des installations des antennes de téléphonie mobile en France:



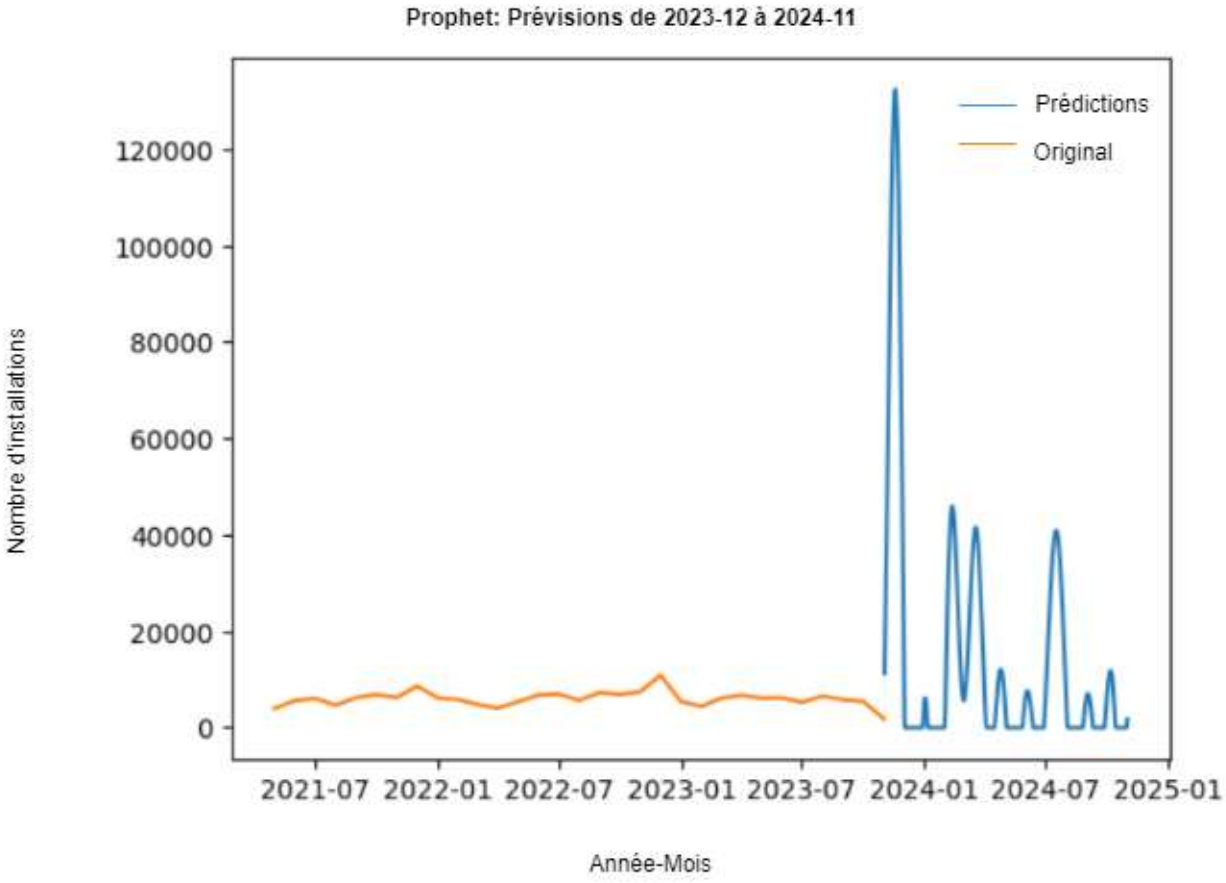
En avril 2021, le nombre d'installations s'élève à 4000 environ. Il augmente ensuite progressivement pour atteindre 9000 en décembre 2023. La croissance est particulièrement forte entre juin et décembre 2022. L'on constate un pic d'installations en fin d'année avec plus de 10000 installations. Cette tendance devrait se poursuivre dans les années à venir, avec le développement des réseaux 5G.
On remarque une tendance saisonnière à la fin de chaque année, avec un pic marqué du nombre d'installations. Cela est probablement dû à plusieurs facteurs, à savoir:

- Le calendrier des vacances scolaires. Les vacances scolaires de Noël sont une période propice aux travaux de construction et d'installation, car les écoles et les entreprises sont fermées. Cela permet aux opérateurs de téléphonie mobile de travailler sans interruption et d'installer plus d'antennes en moins de temps.
- Les objectifs de déploiement. Les opérateurs de téléphonie mobile ont souvent des objectifs de déploiement à atteindre à la fin de l'année. Cela peut les inciter à accélérer le rythme des installations en décembre, afin de pouvoir atteindre leurs objectifs.

Vérifions les prévisions proposées par chacun des modèles de ML un an plus tard.

Prophet

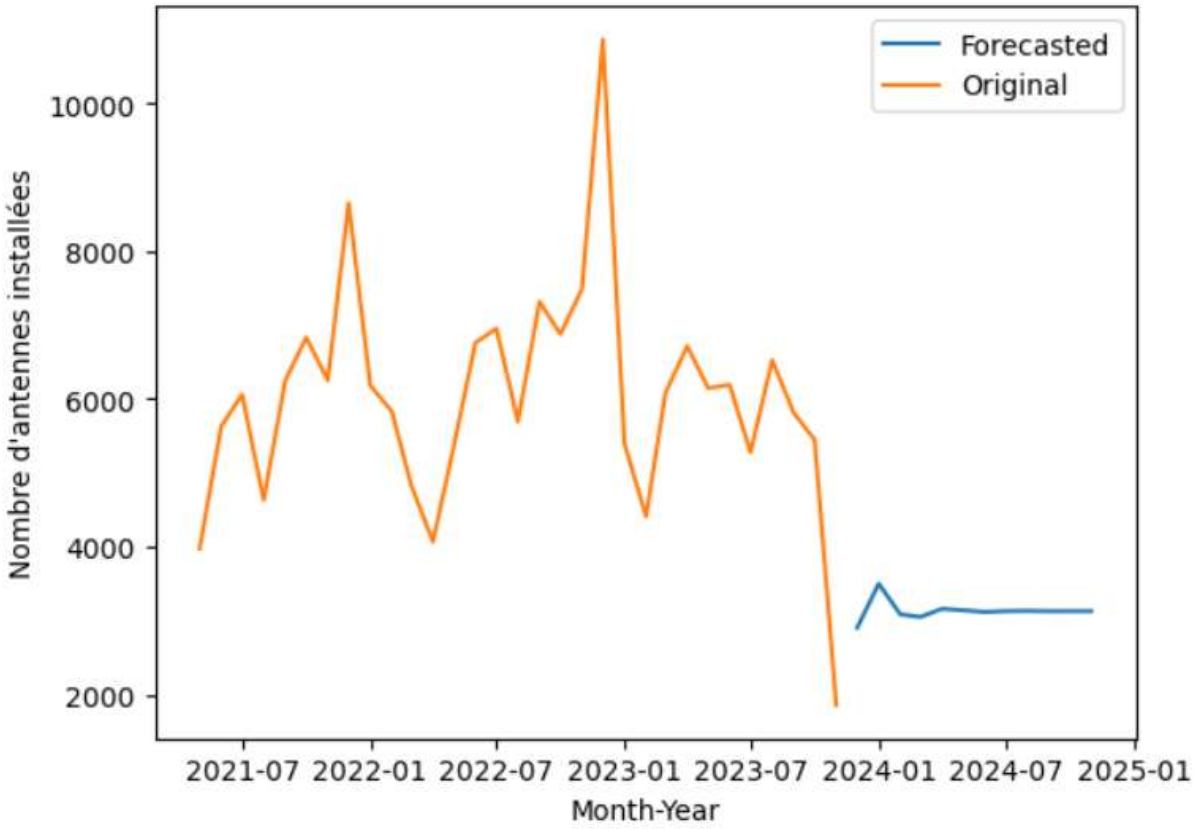
Le modèle Prophet prend en paramètres le DataFrame contenant l'ensemble des données puis fait de l'apprentissage en faisant un "fit". Ensuite, les prédictions sont effectuées grâce à la méthode "predict" qui prend en paramètres des dates futures. Dans ce cas, l'on passe 365 jours pour que le modèle effectue une prédiction sur 12 mois. Voici les résultats qui en sortent.



En décembre 2023, Prophet prévoit plus d'installations sur l'ensemble de l'année comparativement aux tendances constatées sur les données réelles. 120 000 installations sur le territoire français, tout opérateur confondu, est ce que préconise Prophet en décembre 2023. Il existe des différences notables entre les données réelles et les données prédites. Selon la tendance générale, le nombre d'installation d'antennes devrait être inférieur aux peaks proposés par le modèle. Cependant, il est important de garder à l'esprit que ces données sont des prédictions, et qu'il existe une certaine marge d'erreur.
On peut également parler de l'underfitting. Il me semble que le modèle n'ait pas eu assez de données pour apprendre, d'où des prévisions aberrantes.

ARIMA

Au modèle ARIMA, nous avons donné en paramètres, le DataFrame contenant nos données puis (2,1,0), l'ordre (p,d,q) du modèle pour les composantes autorégressives, différences et moyennes mobiles. d est toujours un entier, tandis que p et q peuvent être des entiers ou des listes d'entiers.



En général, les données prédites sont très écartées des données réelles. La courbe prédite suit peu la tendance générale des données réelles, et les écarts sont importants.
Néanmoins, la courbe prédite prévoit bien un pic d'installations en décembre 2022, telles que les données réelles.
La courbe devient linéaire sur le temps. Je suspecte encore une fois un sous-apprentissage lié à un faible taux de données pour l'apprentissage. Notons que les prédictions d'ARIMA, ces dernières sont plus cohérentes que celles du modèle précédent. Faute de temps, nous ne pouvons pas chercher les hyperparamètres permettant de faire les meilleurs apprentissages et prédictions.

LSTM

En ce qui concerne le modèle LSTM, nous ne sommes pas allés jusqu'aux prévisions, faute de temps. Nous n'avons fait que l'apprentissage et les tests avec un seul réseau de neurones. Notons cependant que les données ont été standardisées avec MinMaxScaler. Cet estimateur met à l'échelle et traduit chaque champ individuellement de manière à ce qu'elle se situe dans l'intervalle donné sur l'ensemble d'apprentissage, par exemple entre zéro et un. Il préserve la forme originale de la distribution tout en transformant les données, empêchant ainsi certaines variables de dominer en raison de leur plus grande échelle. Le MinMaxScaler permet de maintenir les relations entre les variables tout en s'assurant qu'elles se situent toutes dans une fourchette spécifiée, ce qui peut améliorer les performances de certains modèles d'apprentissage automatique. Tout compte fait, l'apprentissage est sur la bonne voie, les pics sont bien détectés. Avec un peu plus d'apprentissage et de stockage des patterns, LSTM pourra etre capable de faire de bonnes prédictions.

Notre décision est celle d'enregistrer les prévisions issues d'ARIMA dans notre base de données finale afin de les utiliser pour des représentations graphiques. Il est important de souligner que ces prévisions demeurent des estimations et peuvent différer des données réelles.

3.6.2.5. AMELIORATIONS DE L'APPRENTISSAGE ET DES PREDICTIONS

Dans le cadre de l'amélioration des modèles ARIMA, LSTM et Prophet, plusieurs approches peuvent être explorées :

1. GridSearch Cette méthode consiste à effectuer une recherche systématique des meilleurs hyperparamètres pour les modèles. En ajustant différents ensembles de paramètres prédéfinis, on sélectionne ceux qui produisent les meilleures performances pour chaque modèle.

2. Jeu de paramètres : Explorer une plus grande variété de paramètres ou d'hyper paramètres pour les modèles. Cela implique de tester et d'optimiser diverses combinaisons de paramètres pour affiner les prédictions.

3.Volume plus important de données : L'extension du jeu de données peut améliorer la capacité des modèles à capturer des tendances plus complexes. Collecter davantage de données peut être bénéfique pour améliorer la précision des prédictions, surtout lorsqu'elles sont disponibles et qu'elles peuvent améliorer la représentation des tendances et des variations temporelles.

3.6.2. Visualisation des données

La Business Intelligence (BI) est un processus technologique qui consiste à analyser les données et à présenter des informations, dans le but d'aider les décideurs à prendre des décisions stratégiques. Il s'agit d'un outil d'aide à la décision.
Une fois les données collectées et traitées et préparées pour l'analyse, afin de générer des rapports, des tableaux de bord et d'autres éléments visuels de visualisation des données, rendant ainsi les résultats analytiques accessibles aux profils respectifs.

3.6.2.1. OUTIL D'EXPOSITION CHOISI

Il existe plusieurs outils de visualisation de données sur le marché dont: Tableau; Power BI, Dash, et Grafana qui sont parmi les plus connus. Nous avons étudié chacun des outils et leur pertinence par rapport au besoin de notre projet.
Dès le début, Dash et Grafana ont été écartés de la sélection. Bien que Grafana soit réputé pour ses visualisations de séries temporelles, nos besoins en visualisations vont au-delà de ce type spécifique de graphiques. En outre, Dash est souvent utilisé pour créer des visualisations interactives en Python, mais pour notre projet, nous recherchons des solutions plus généralistes pour répondre à une variété de besoins visuels.
Le choix doit se faire donc entre Power BI et Tableau. Cela dépend de plusieurs facteurs, notamment les besoins spécifiques de notre projet, des compétences de notre équipe et les coûts. Voici les points pris en compte lors de la décision entre Power BI et Tableau :

1. Facilité d'utilisation et apprentissage :
    - Power BI est souvent réputé pour sa convivialité, surtout pour les utilisateurs habitués à d'autres produits de la suite Microsoft. De plus, notre équipe est plus familière avec cet outil, ce qui pourrait accélérer le processus de création des graphiques.
2. Coût :
    - Power BI offre une version test plus longue et plus accessible de 60 jours.
    - Tableau est plus contraignant avec une version d'essai de seulement 14 jours.

3. Performance :
    - Les performances, considérant l'architecture et le volume de données, semblent assez similaires pour les deux outils (tests de performance avec des jeux de données représentatifs).
4. Communauté et support :
    - Les deux outils bénéficient de grandes communautés d'utilisateurs actives, ce qui signifie qu'il est généralement facile de trouver des ressources en ligne, des tutoriels et des forums de discussion.
5. Sécurité et gouvernance des données :
    - Les 2 outils offrent des fonctionnalités de sécurité et de gouvernance des données en conformité avec notre organisation.
6. Intégration avec d'autres outils ou bases de données : La capacité des outils à se connecter facilement à d'autres sources de données ou à être intégrés à des systèmes existants est un critère important. Power BI s'intègrent mieux avec les bases de données offrant une large gamme de connecteurs.

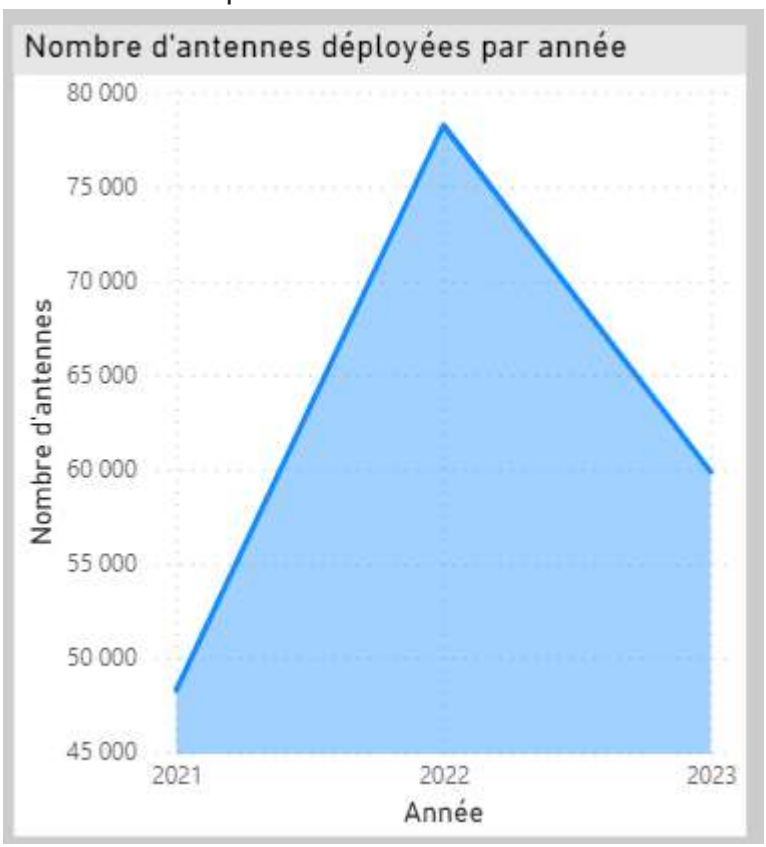
Finalement, nous optons pour Power BI en tenant compte des exigences particulières de notre projet, notamment les contraintes liées aux coûts et à la familiarité de l'équipe avec l'outil. Après avoir testé les

deux solutions pour déterminer celle qui convient le mieux à nos besoins, la question des licences est rapidement devenue cruciale et a joué un rôle déterminant dans notre choix.

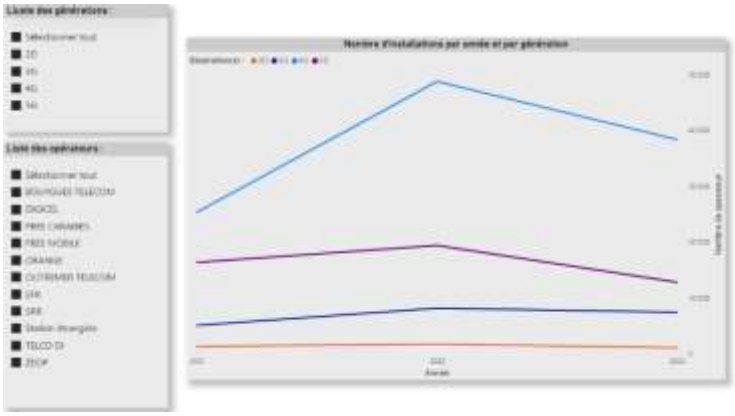
3.6.2.2.INTERPRETATION DES RESULTATS

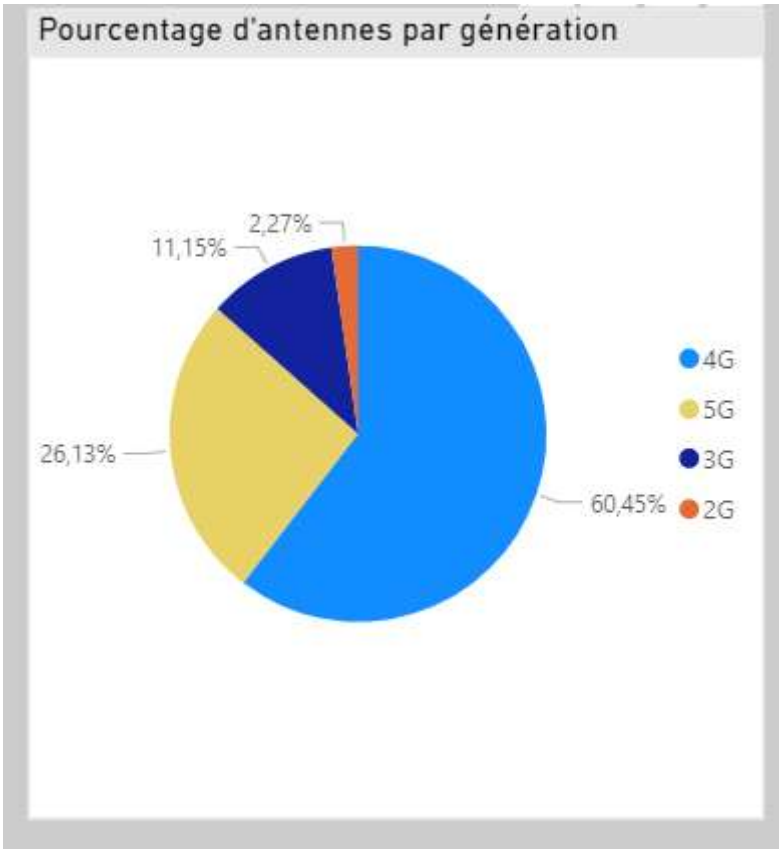
Avant de commencer les interprétations, il est important de noter que ces analyses se basent sur un échantillon de données. Elles peuvent ne pas refléter la réalité des faits sur le marché de la téléphonie mobile.

La téléphonie mobile est devenue un élément indispensable de notre vie quotidienne. Pour garantir une bonne couverture et une qualité de service optimale, les opérateurs de téléphonie mobile doivent déployer un réseau d'antennes dense et performant.
Selon des données de l'ANFR, 186 419 antennes ont été installées en France entre avril 2021 et novembre 2023 avec un pic d'environ 80 000 en 2022. Cette augmentation est due à plusieurs facteurs, notamment le développement des réseaux 4G et 5G, l'augmentation de la demande de services de téléphonie mobile et la simplification des procédures administratives pour l'installation d'antennes.



Les antennes 4G représentent la grande majorité des nouvelles installations, avec plus de 45 000 antennes installées. Les antennes 5G, plus récentes, sont encore en phase de déploiement, avec moins de 16 000 antennes installées.



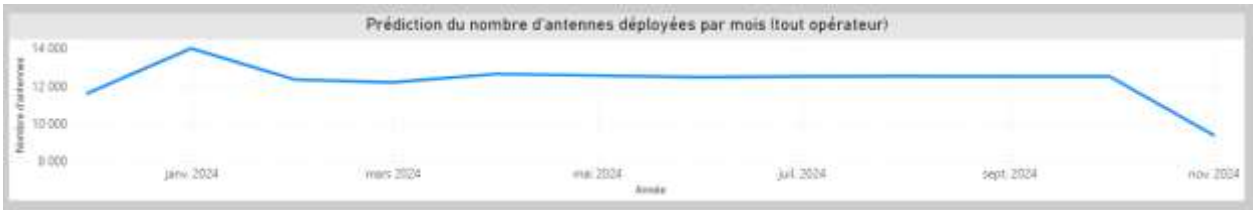


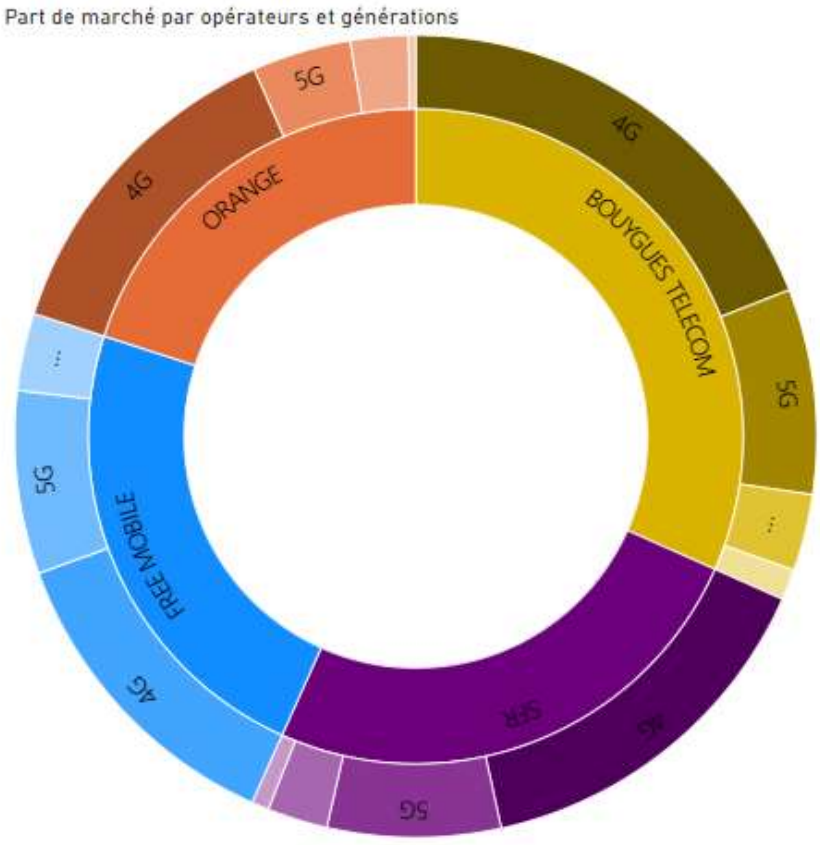
La majorité de ces antennes se trouvent dans la région d'Auvergne-Rhône-Alpes, avec 24 477 antennes installées. Viennent ensuite les régions d'Île-de-France (20 951 antennes), d'Occitanie (20 252 antennes) et de Nouvelle-Aquitaine (20 077 antennes).

Nombre d'installations par région

Région	Nbr d'installations
Auvergne-Rhône-Alpes	24477
Île-de-France	20951
Occitanie	20252
Nouvelle-Aquitaine	20077
Grand-Est	16804
Provence-Alpes-Côte d'Azur	14153
Hauts-de-France	12822
Bourgogne-Franche-Comté	10938
Pays de la Loire	10712
Bretagne	10035
Normandie	9554
Centre-Val de Loire	7632
Corse	3388
Total d'installations	**186419**

Le déploiement des antennes de téléphonie mobile en France est un phénomène majeur qui a un impact significatif sur notre quotidien. Cette croissance exponentielle devrait se poursuivre dans les années à venir indépendamment des opérateurs, avec le développement des réseaux 5G.





Avec Bouygues Telecom 31,66
4G = 19,06%
5G = 8,27%

SFR 24,05%
4g = 14.92%
5g = 6,98%

Free Mobile 23,25%
4g = 12%

5g=7%
Orange 20,05%
4g=13%
5g = 3%

Les opérateurs de téléphonie mobile doivent surveiller de près la compétition et ajuster leurs investissements, particulièrement en ce qui concerne le déploiement des antennes 5G. Par exemple, Orange devrait envisager d'accroître ses investissements dans la 5G afin de maintenir sa position concurrentielle. L'étroite surveillance des déploiements est également cruciale pour évaluer la couverture du réseau. Cela permet à chaque opérateur d'identifier les zones nécessitant des améliorations ou des extensions, et de planifier les investissements nécessaires pour fournir des services de qualité à leurs abonnés.
Les organismes de régulation tels que l'ANFR et l'ARCEP peuvent utiliser ces données pour favoriser une concurrence équitable entre tous les opérateurs du marché des télécommunications. En veillant à ce que chaque opérateur ait un accès équitable à des ressources régulées, ils peuvent éviter les inégalités sur le marché.

Pour les consommateurs, ces données leur permettent désormais d'évaluer la couverture, la vitesse et la qualité du réseau proposé par différents opérateurs dans leur région. Ceci leur permettra de faire des choix informés lorsqu'ils sélectionnent un fournisseur de services mobiles. En conclusion, il est indéniable que les consommateurs des régions telles que l'Auvergne-Rhône-Alpes, l'Île-de-France, l'Occitanie et la Nouvelle-Aquitaine bénéficient de meilleures infrastructures en matière de téléphonie mobile.

Conclusion

- Récapitulation des points clés.
- Mise en évidence de la maîtrise des compétences visées :

1. Analyse des besoins métier et mise en œuvre d'une stratégie Big Data
2. Collecte et stockage des données
3. Traitement et exposition des données
4. Déploiement d'une plateforme big data et maintenance des développements

Annexes

- Graphiques, schémas, captures d'écran.
- Code source (si pertinent).
- Références bibliographiques.

Sources