

MEMOIRE

Présenté à

**L’Institut Supérieur d’Informatique
Et de Multimédia de Sfax**

En vue de l’obtention du diplôme de

LICENCE

***En BIG DATA ET ANALYSE DE DONNEES
intitulé***

**Mise en place d'un tableau de bord de suivi des
flux réseaux télécom de Tunisie Telecom**

Par

**Wala Ben Rhouma
Souad Achouri**

Soutenu le 30/05/2024, devant le jury composé de :

Mme Imen Tounsi
Mme Hana Mallek
Mme Mouna Ktari

Présidente
Membre
Encadrante

Dédicace

De plus profond de mon cœur, je dédie ce travail

À mon père Fethi

Celui qui m'a appris, à travers l'amour paternel, les conseils, les directives qu'il m'a prodigués tout le long de ma vie. Que ce travail soit le témoignage de l'amour que je lui avoue.

À ma mère Fatma

Pour son amour, son soutien et son aide précieuse. Je ferai toujours de mon mieux pour rester un sujet de fierté à ses yeux. Que Dieu lui procure santé, bonheur et longue vie et que ce travail soit le témoignage de mon éternelle reconnaissance pour son amour et ses sacrifices.

À mon frère Ayoub et ma sœur Wafa

Pour leur amour, leur soutien, et leur écoute en cas de besoin. Je vous souhaite un avenir radieux plein de succès et de bonheur.

À mes amis

Vous étiez à mes côtés tout au long de ce chemin. Je suis très reconnaissante de faire votre connaissance. C'est un vrai honneur pour moi. Merci pour vos aides et vos conseils aux moments de faiblesse. Que Dieu vous garde pour moi. Je vous adore.

Wala

Dédicace

Du plus profond de mon cœur, je dédie ce modeste travail

À ma famille

Qui m'a offert une éducation exemplaire et qui a façonné la personne que je suis aujourd'hui.

En particulier, à mon père ABDESSATTAR. Aucun hommage ne pourrait être à la hauteur de l'amour incommensurable dont il m'a toujours entouré.

À Mes frères et mes sœurs

Qui m'ont toujours soutenu et encouragé, je suis profondément reconnaissant. Leur guidance empreinte de sagesse a été une lumière dans les moments sombres. Le soutien constant de ma famille m'a apporté la confiance nécessaire pour surmonter les obstacles et saisir les opportunités qui se sont présentées sur mon chemin.

À mes précieux amis

Qui m'ont toujours encouragé, et à qui je souhaite plus de succès. Ils vont être mes compagnons de route tout au long de cette aventure académique. ces encouragements et ces moments de détente ont allégé le fardeau des études et ont rendu cette expérience plus joyeuse et mémorable.

Souad



REMERCIEMENT

Nous sommes reconnaissantes envers le Tout-Puissant pour notre santé et notre motivation à entreprendre etachever ce projet.

Nous voulons exprimer notre profonde gratitude envers notre enseignante et encadrante, Mme Mouna Ktari, pour son suivi, sa sympathie et ses encouragements qui ont grandement contribué au succès de notre travail. Nous tenons également à remercier M. Yessine Brahmi, Chef de Service Informatique à Tunisie Telecom, pour son assistance, sa disponibilité et ses précieux conseils tout au long de notre stage.

Nous souhaitons également remercier les membres du jury qui ont honoré notre travail en l'examinant et en l'évaluant. Enfin, nous sommes profondément reconnaissantes envers nos enseignants de l'Institut Supérieur d'Informatique et de Multimédia de Sfax pour la qualité exceptionnelle de leur enseignement.

■ TABLE DES MATIÈRES

LISTE DES FIGURES	ix
LISTE DES TABLEAUX	x
LISTE DES ABRÉVIATIONS	xi
INTRODUCTION GÉNÉRALE	xii
1 Chapitre 1 : Etude Préalable	2
1.1 Introduction	3
1.2 Présentation de l'organisme d'accueil	3
1.2.1 Organigramme de Tunisie Telecom	3
1.2.2 Direction Centrale des Finances	5
1.3 Définition de champ de l'étude	5
1.4 Étude de l'existant	6
1.4.1 Analyse de l'existant	6
1.4.2 Critique de l'existant	7
1.5 Solution proposée	8
1.6 Methodologie de travail	9
1.6.1 Méthodes Agile	9
1.6.2 Choix de la méthodologie de travail	10
1.7 Conclusion	12
2 Chapitre 2 : L'informatique décisionnelle	13
2.1 Introduction	14
2.2 Concepts généraux du BI	14
2.2.1 Le Business Intelligence	14
2.2.2 Les avantages du BI	14
2.2.3 Les limites du BI	15
2.3 Les principes des systèmes décisionnels	15
2.3.1 Sources de données	16
2.3.2 Entrepôt de données	16
2.3.3 Magasin de données	17
2.3.4 Extract - Transform - Load	17
2.4 Modélisation multidimensionnelle	17
2.4.1 Définition	17
2.4.2 Concepts de base	18
2.4.3 Schéma multidimensionnel	19
2.4.3.1 Schéma en étoile :	19

TABLE DES MATIÈRES

2.4.3.2	Schéma en flocon de neige :	19
2.4.3.3	Schéma en constellation :	19
2.4.4	L'objectif de la modélisation multidimensionnelle	19
2.5	Démarche de construction d'un entrepôt de données	20
2.5.1	Modélisation et conception de l'entrepôt	20
2.5.2	Alimentation de l'Entrepôt	21
2.5.2.1	Extraction des Données :	21
2.5.2.2	Transformation des données :	21
2.5.2.3	Chargement des données :	22
2.5.3	Administration et maintenance	23
2.6	Analyse et Fouille de Données (Data Mining)	23
2.6.1	Définition	23
2.6.2	L'utilité de l'analyse et fouille de données	24
2.6.3	Concept de Base	24
2.7	Conclusion	25
3	Chapitre 3 : Sprint 0 - Analyse et spécification des besoins	26
3.1	Introduction	27
3.2	Compréhension du domaine	27
3.2.1	Notion de fraude	27
3.2.2	Call Detail Record (CDR)	28
3.2.3	Flux télécoms et systèmes de taxation	29
3.2.3.1	Services Telecom	30
3.2.3.2	Systèmes de taxation	30
3.3	Analyse des besoins	32
3.3.1	Identification des Acteurs	32
3.3.2	Les besoins fonctionnels	33
3.3.3	Les besoins non fonctionnels	33
3.3.4	Diagramme de cas d'utilisation global	34
3.4	Pilotage du projet avec scrum	35
3.4.1	L'équipe SCRUM	35
3.4.2	Backlog du produit	36
3.4.3	Planification des sprints	37
3.5	Choix des outils de développement	38
3.6	Conclusion	39
4	Chapitre 4 : Sprint 1 - Construction de l'entrepôt de données	40
4.1	Introduction	41
4.2	Sprint backlog	41
4.3	Diagramme du cas d'utilisation de sprint 1	42
4.3.1	Diagramme du cas d'utilisation	42
4.3.2	Description textuelle du cas d'utilisation « Créer l'entrepôt de données »	43
4.4	Schéma conceptuel de la source de données	43
4.5	Modélisation conceptuelle de l'entrepôt de données	44
4.6	Construction de l'entrepôt de données	54
4.6.1	Modélisation dimensionnelle	54
4.6.2	Choix du modèle dimensionnel	55

TABLE DES MATIÈRES

4.6.3	Modèle physique de l'entrepôt de données	56
4.7	Conclusion	57
5	Chapitre 5 : Sprint 2 - Gestion du processus ETL	58
5.1	Introduction	59
5.2	Sprint Backlog « Gérer le processus ETL »	59
5.3	Diagramme du cas d'utilisation du sprint 2 « Gérer le processus ETL »	60
5.3.1	Diagramme du cas d'utilisation	60
5.3.2	Description textuelle du cas d'utilisation « Gérer le processus ETL »	61
5.4	Staging Area (SA)	61
5.5	Gestion de processus ETL : du Flux source vers SA	63
5.5.1	Extraction des données	63
5.5.2	Nettoyage et transformations :	63
5.5.3	Chargement des données	64
5.6	Gestion de processus ETL : du SA vers l'entrepot de données	64
5.6.1	Diagramme d'activités	64
5.6.2	Description de ce processus ETL : du SA vers l'ED	65
5.6.2.1	Phase d'extraction des données (Extract)	65
5.6.2.2	Phase de transformation des données (Transform)	67
5.6.2.3	Phase de chargement des données (Load)	69
5.7	Automatisation du chargement de l'ED	71
5.8	Conclusion	73
6	Chapitre 6 : Sprint 3 : Visualisation de données	74
6.1	Introduction	75
6.2	Sprint Backlog	75
6.3	Diagramme du cas d'utilisation du sprint 3 « Créer tableau de bord »	76
6.3.1	Diagramme du cas d'utilisation	76
6.3.2	Description textuelle du cas d'utilisation « Créer tableau de bord »	77
6.4	Création des tableaux de bord	77
6.4.1	Etablir la connexion entre Power BI Desktop et Oracle	78
6.4.2	Mesures spécifiques (DAX)	78
6.4.3	Choix des graphiques	80
6.4.4	Présentation des interfaces de réalisation	81
6.4.4.1	Page d'accueil	82
6.4.4.2	Rapport : Validation de système de taxation AIR / service de recharge électronique ETOPUP	82
6.4.4.3	Rapport : Validation de système de taxation AIR / service de recharge SOS solde USSD	83
6.4.4.4	Rapport : Validation de système de taxation AIR / service de recharge par carte Voucher	84
6.4.4.5	Rapport : Validation de système de taxation OCC / Service de transfert de données SASN	85
6.4.4.6	Rapport : Validation des flux de système de taxation CCN / service de SMS+ MMG	86
6.4.4.7	Raport : Validation des flux de taxation CCN et le service des appels et des sms MSC	87
6.5	Conclusion	88

TABLE DES MATIÈRES

7 Chapitre 7 : Sprint 4 : Analyse et Fouille de données	89
7.1 Introduction	90
7.2 Sprint Backlog	90
7.3 Diagramme du cas d'utilisation du sprint 4 « Gérer les structures de fouille de données »	91
7.3.1 Diagramme du cas d'utilisation	91
7.3.2 Description textuelle du cas d'utilisation « Gérer structure de fouille de données »	91
7.4 Compréhension des données et descriptions des variables	92
7.5 L'analyse descriptive des variables	92
7.5.1 Description statique	92
7.5.2 Distribution de la variable cible	93
7.5.3 Distribution de la variable statut	94
7.6 Choix du modèle	94
7.7 Entraînement du Modèle	95
7.8 Evaluation du Modèle	96
7.9 Interprétation des résultats	97
7.10 Conclusion	98
CONCLUSION GÉNÉRALE	99
BIBLIOGRAPHIE	100

LISTE DES FIGURES

1.1	Organigramme général de Tunisie Télécom	4
1.2	Fonctionnement de Scrum	11
2.1	Architecture générale d'un système décisionnel	16
2.2	Exemple de fait	18
2.3	Exemple de dimension	19
3.1	Les dimensions CDR MSC	29
3.2	Diagramme du cas d'utilisation global du Système décisionnel de TUNISIE TELECOM	34
3.3	Pilotage de projet par Scrum	37
3.4	Diagramme de Gantt de planification des sprints	37
4.1	Diagramme du cas d'utilisation de sprint 1	42
4.2	Diagramme de classe	44
4.3	Les étapes de l'approche descendante	45
4.4	Schéma multidimensionnel de la démarche descendante	53
4.5	Schéma en étoile	54
4.6	Schéma en flocons de neige	55
4.7	Schéma en constellation	55
4.8	Modèle physique de l'entrepôt de données	57
5.1	Diagramme du cas d'utilisation du sprint 2 « Gérer le processus ETL »	60
5.2	Schéma du flux décisionnel avec utilisation du Staging Area	62
5.3	La base de données du SA	64
5.4	Diagramme d'activités pour l'alimentation de l'ED	65
5.5	Création d'un job	66
5.6	Connexion talend et Oracle	66
5.7	Le composant TDBInput	67
5.8	Configuration de ces composants	67
5.9	Job aggregation par filename	68
5.10	Job aggregation par date	68
5.11	job de concatenation	68
5.12	Le composant tFilterRow	69
5.13	Le composant tAggregate	69
5.14	Le composant tMap	69
5.15	Creation de l'entrepôt de données	70
5.16	Decomposition de la date	70
5.17	Chargement des tables de fait et des tables des dimensions	71
5.18	Job d'automatisation de chargement de l'entrepot	71

LISTE DES FIGURES

5.19	Création de tâche d'automatisation du chargement de l'ED	72
5.20	Planification du job du chargement de l'entrepot	72
6.1	Diagramme du cas d'utilisation de sprint 3 « Créer tableau de bord »	76
6.2	Connexion du Power BI avec Oracle	78
6.3	Chargement de l'entrepot dans PowerBI	78
6.4	Creation d'une mesure de la difference entre les Event_Count de deux flux	79
6.5	Creation d'une mesure de la difference entre les Charge_Amount de deux flux	79
6.6	Creation d'une mesure de la difference entre les Bonus_Amount de deux flux	79
6.7	Histogramme groupé	80
6.8	Graphique en courbe	81
6.9	Graphique en secteurs	81
6.10	Graphique en segment	81
6.11	Première page du tableau de bord : Page d'accueil	82
6.12	Validation des flux AIR_ETOPUP	83
6.13	Validation des flux AIR_USSD	84
6.14	Validation des flux AIR_Voucher	85
6.15	Validation des flux OCC_SASN	86
6.16	Validation des flux CCN_MMG	87
6.17	Validation des flux MSC_CCN	88
7.1	Diagramme du cas d'utilisation du sprint 4	91
7.2	Statistiques descriptives	93
7.3	Distribution de la variable cible	93
7.4	Distribution de la variable status	94
7.5	Metriques d'évaluation	97
7.6	Les resultats obtenus	97
7.7	Courbes des Valeurs de Test et Prédites	98

LISTE DES TABLEAUX

3.1	Flux télécoms et systèmes de taxation	31
3.2	Backlog du produit	36
4.1	Backlog du sprint 1	41
4.2	Description textuelle du cas d'utilisation « Créer l'entrepôt de données »	43
5.1	Backlog du sprint 2 « Gérer le processus ETL »	59
5.2	Description textuelle du cas d'utilisation « Gérer le processus ETL »	61
6.1	Backlog du sprint 3	75
6.2	Description textuelle du cas d'utilisation « Créer tableau de bord »	77
7.1	Backlog du sprint 4	90
7.2	Description textuelle du cas d'utilisation « Gérer structures de fouille de données »	91



LISTE DES ABRÉVIATIONS

AIR Automatic Incident Reporting

BI Business Intelligence

CCN Charging Control Node

CDR Call Detail Record

DAX Data Analysis Expressions

DRAF Direction Revenue Assurance et Fraude

ETL Extract Transform Load

ETOPUP Electronic Top Up

KPI Key Performance Indicator

MMG Multimedia Management Gateway

MSC Mobile Switching Center

OCC Online Charging Control

USSD Unstructured Supplementary Service Data

RA Revenue Assurance

SASN Subscriber Authentication Service Node

INTRODUCTION GÉNÉRALE

Dans un monde de plus en plus connecté, les opérateurs de télécommunications jouent un rôle crucial dans la connectivité et la communication des individus et des entreprises. En tant que principal opérateur en Tunisie, Tunisie Telecom (TT) doit gérer efficacement ses flux de données et offrir un service fiable à ses clients, tout en optimisant la gestion de ses ressources, notamment pour le service SOS Solde.

Dans cette perspective le Business Intelligente (BI) vient résoudre les problèmes rencontrés par les entreprises en matière d'aide à la décision au moyen d'outils et de méthodes permettant de collecter, standardiser, modifier et restituer leurs données de productions afin de favoriser les meilleures prises de décision.

Dans ce cadre, Tunisie Telecom (TT) souhaite mettre en place un système d'aide à la décision permettant de répondre à ses besoins d'analyse. Ce système permettra de suivre, via des tableaux de bord, la validation des services TT en relation avec les systèmes de taxation des télécoms. En parallèle, un algorithme d'analyse et de fouille de données a été appliqué pour produire des prédictions précises, aidant ainsi à prendre des décisions informées concernant le remboursement des utilisateurs du service SOS en dépassement de leur forfait.

Le présent rapport montre clairement les différentes étapes de réalisations des travaux requis pour ce projet. Il est structuré en sept chapitres :

INTRODUCTION GÉNÉRALE

- **Chapitre 1** « Étude préalable » du projet : Ce chapitre est dédié pour la présentation de l’organisme d’accueil, le champ d’étude, l’étude de l’existant, les solutions proposées et la méthodologie choisie.
- **Chapitre 2** « Informatique décisionnelle » : Ce chapitre présente les notions de base de l’informatique décisionnelle.
- **Chapitre 3** « Sprint 0 : Analyse et spécification des besoins » : Ce chapitre s’intéresse à l’analyse des besoins, les outils décisionnels, le backlog du produit et le pilotage par Scrum.
- **Chapitre 4** « Sprint 1 : Construction de l’entrepôt de données » : Ce chapitre montre d’abord le schéma conceptuel des données source puis le schéma conceptuel et le modèle physique de l’entrepôt de données.
- **Chapitre 5** « Sprint 2 : Gestion du processus ETL » : Ce chapitre expose les transformations appliquées sur les données, les processus de chargement incrémental et de rafraîchissement des données.
- **Chapitre 6** « Sprint 3 : Visualisation des données » : Ce chapitre dévoile la construction du tableau de bord dynamique.
- **Chapitre 7** « Sprint 4 : Analyse et Fouille de données » : Ce chapitre décrit la partie machine learning concernant sos solde .

Chapitre 1 : Etude Préalable

Sommaire

1.1	Introduction	3
1.2	Présentation de l'organisme d'accueil	3
1.2.1	Organigramme de Tunisie Telecom	3
1.2.2	Direction Centrale des Finances	5
1.3	Définition de champ de l'étude	5
1.4	Étude de l'existant	6
1.4.1	Analyse de l'existant	6
1.4.2	Critique de l'existant	7
1.5	Solution proposée	8
1.6	Methodologie de travail	9
1.6.1	Méthodes Agile	9
1.6.2	Choix de la méthodologie de travail	10
1.7	Conclusion	12

1.1 Introduction

Le succès de toute étude dépend de la qualité de son départ. C'est pour cette raison que ce premier chapitre est dédié à l'étude du cadre de projet, ainsi, qu'à sa compréhension globale. Dans une première partie, nous présentons l'organisme d'accueil (Tunisie Telecom) ainsi que ses domaines d'activité. Puis, nous définissons le champ de l'étude et les objectifs à atteindre par notre projet. En second lieu, nous abordons l'étude de l'existant dans laquelle nous décrivons le système actuel de tunisie telecom. Puis, nous le critiquons en évoquant ses avantages et ses limites. Sur la base de ses critiques, nous proposons les solutions que nous comptons réaliser à travers notre projet. On clôture ce chapitre avec le choix de la méthodologie adéquate.

1.2 Présentation de l'organisme d'accueil

Tunisie Telecom, fondée en 1996, est le principal opérateur de télécommunications en Tunisie. Avec une présence étendue dans tout le pays, Tunisie Telecom offre une gamme complète de services de télécommunications, y compris la téléphonie fixe et mobile, l'accès à Internet, ainsi que des services de données et de transmission. Tunisie Telecom dispose d'une infrastructure robuste et moderne comprenant des réseaux de fibre optique, des stations de base mobiles, des centres de données, et des équipements de transmission haut débit. Cette infrastructure permet à Tunisie Telecom de fournir des services fiables et à haut débit à ses clients à travers tout le territoire tunisien. Tunisie Telecom s'engage à respecter des valeurs telles que la qualité et l'amélioration continue, et elle a plus de 6 millions d'abonnés dans les services de téléphonie fixe et mobile. La société est présente dans 24 régions et dispose de 140 Espaces TT ainsi que de plus de 13 mille points de vente privés, et emploie plus de 6 mille personnes[1].

1.2.1 Organigramme de Tunisie Telecom

L'organigramme est avant tout un outil de communication destiné à faciliter la compréhension des rapports et liens existants au sein de la société, il fait l'objet de plusieurs réformes et

CHAPITRE 1 : ETUDE PRÉALABLE

restructurations afin de s'adapter aux nouvelles exigences d'efficacité. En effet, l'organigramme de Tunisie Telecom se présente comme suit :

- **DGA** : Direction Générale Adjoint.
- **D.C.F** (Direction Centrale Financière) : est une unité dont le rôle est de faire la gestion financière ainsi que la comptabilité et l'administration.
- **D.C.C.M** (Direction Centrale Commerciale et Marketing) : est une unité ayant pour but de gérer les ventes, le chiffre d'affaires et le processus de marketing.
- **D.C.R.H** (Direction Centrale Ressources Humaines) : Comme indique son nom cette unité de travail se focalise sur la procédure de recrutement, d'intégration et de formation du personnel, la gestion administrative et la paie ainsi que la communication interne.
- **D.C.S.E** : Direction Centrale Solution d'Entreprise : cette unité est destinée à faire les études, l'installation et la maintenance des réseaux privés.
- **D.C.S.I** : Direction Centrale des systèmes d'informations : cette unité s'occupe de la mise en œuvre de l'infrastructure informatique et des réseaux sécurisés de l'entreprise ainsi que des solutions IT [2].

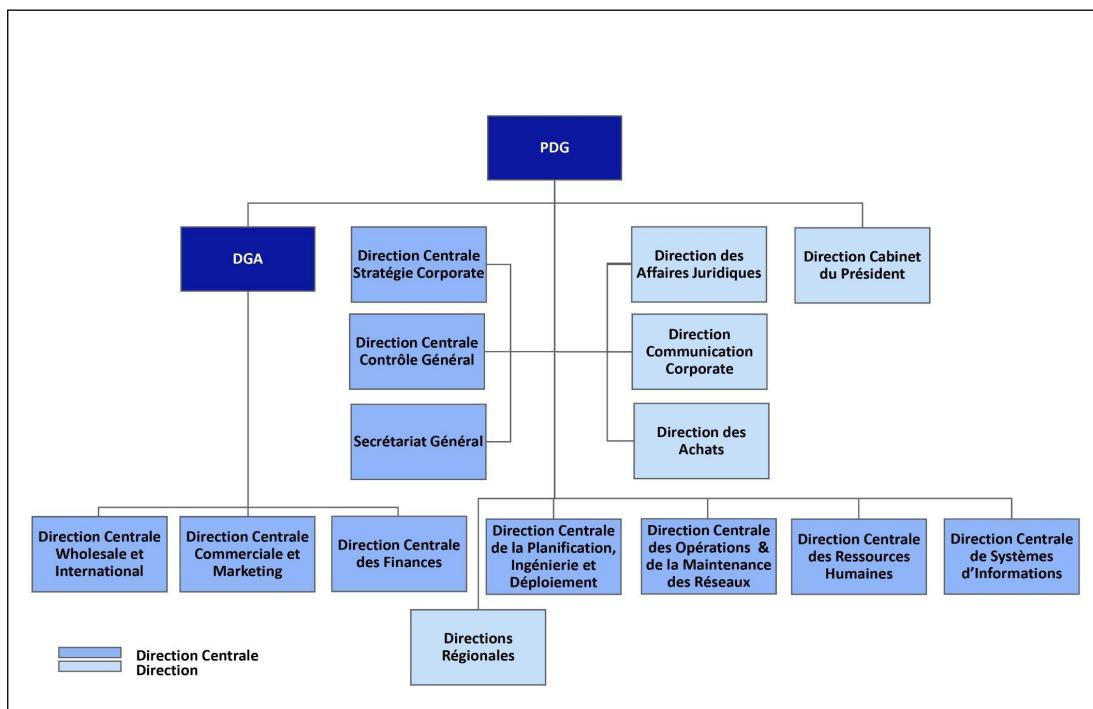


FIGURE 1.1 – Organigramme général de Tunisie Télécom

1.2.2 Direction Centrale des Finances

Nous avons effectué notre stage au sein de la Direction Centrale des Finances, qui est une direction opérationnelle dont la mission principale est de définir les politiques financières afin de garantir l'équilibre budgétaire et de produire des états financiers conformes à la réglementation en vigueur. Les activités de la DCF sont :

- L'activité comptable et fiscale.
- L'activité de trésorerie et du financement.
- L'activité de facturation et de recouvrement.
- L'activité de planification et contrôle financier.
- L'activité d'assurance.

Notre attention s'est portée spécifiquement sur les activités de la Direction Assurance Revenue et Fraude, relevant de la Direction Centrale des Finances. Cette entité est chargé de mettre en place des mécanismes visant à garantir la perception de tous les revenus. Son rôle consiste à détecter les fuites tout au long de la chaîne de revenus, à évaluer leur impact comptable, à récupérer les sommes manquantes et, surtout, à mettre en œuvre des actions correctives pour éviter leur récurrence.

1.3 Définition de champ de l'étude

Dans le cadre de notre projet de fin d'études, notre mission consiste à concevoir et mettre en place un tableau de bord permettant de suivre les flux réseaux télécom de Tunisie Telecom. Pour ce faire, nous allons construire un entrepôt de données à partir des informations provenant des différents services de télécommunications. Cette solution impliquera un processus de chargement comprenant la lecture, la transformation et le stockage des données, ainsi que la mise en place de différents contrôles de validation pour garantir la qualité des données. Une fois les données chargées dans l'entrepôt, nous mettrons en place des tableaux de bord permettant de comparer les données de chaque flux. Ces tableaux de bord fourniront une visualisation claire et intuitive des informations pertinentes, facilitant ainsi leur analyse et leur interprétation.

Par la suite, nous appliquerons des algorithmes d'analyse et de fouille de données afin de réaliser des prédictions. Par exemple, nous pourrions utiliser ces algorithmes pour prédire les personnes qui vont recharger leur solde SOS (service d'urgence). Ces prédictions pourront aider Tunisie Telecom à anticiper les besoins de ses clients et à prendre des décisions stratégiques plus éclairées.

1.4 Étude de l'existant

L'étude de l'existant constitue la première étape du processus de la conception. Elle vise à étudier l'existant et le critiquer pour déterminer les orientations et comprendre les fonctionnalités du futur système.

1.4.1 Analyse de l'existant

Afin d'approfondir notre compréhension du sujet et d'avoir une idée plus claire sur notre projet et ses fonctionnalités attendues, nous avons mené une étude sur les limitations actuelles au sein de Tunisie Télécom.

* Au niveau de la collecte et du stockage des données :

Les différents systèmes de Tunisie Télécom génèrent, gèrent et stockent quotidiennement d'importantes volumes de données provenant de diverses sources telles que des tables de base de données, des fichiers d'entrée (csv, Txt, out, xlsx, etc.), des journaux (logs) ou des pièces jointes de gestion de cas. Cette diversité de sources requiert un traitement personnalisé de chaque source. Après l'envoi des données au serveur FTP, le stockage s'effectue manuellement dans la base de données via des requêtes SQL, PL/SQL et des scripts Shell. En raison de la grande quantité d'informations, le processus de stockage est trop lent. Tunisie Télécom utilise des requêtes exécutables directement sur SQLDeveloper et enregistre les résultats dans la base de données Oracle. Le temps d'attente pour le résultat de la requête est très long (plus de 500 millions de CDR par jour dans une seule table).

* Sur le plan analytique :

Le suivi des différents trafics (voix, SMS, etc.) au sein de la DRAF (Direction Revenue Assurance et Fraude) est actuellement assuré par la division RA (Revenue Assurance). Une fois les données enregistrées, l'étape suivante consiste à mettre en place le reporting, à analyser les différents trafics et à pérenniser le rapport journalier et le rapport hebdomadaire de l'équipe Revenue Assurance et fraude. Le rapport actuel est constitué d'une série de tableaux Excel récapitulant tous les flux. Ce reporting opérationnel est difficilement exploitable dans une approche très performante.

* Problème de gestion du SOS solde :

Un autre problème majeur auquel Tunisie Télécom est confronté concerne la gestion du SOS solde, où certains abonnés ne recharge pas leurs crédits et Tunisie Télécom les rembourse. Cette mauvaise gestion des ressources entraîne des pertes de revenus, car les fonds sont utilisés pour rembourser des abonnés inactifs au lieu de rembourser ceux qui sont actifs. Cela génère une inefficacité dans la gestion des ressources financières de l'entreprise.

Dans l'ensemble, l'existant montre un besoin critique d'amélioration dans la gestion des données et des processus analytiques au sein de Tunisie Télécom. Les méthodes manuelles actuelles sont inefficaces pour faire face au volume croissant de données et entravent la capacité de l'entreprise à tirer pleinement parti de ses données pour prendre des décisions éclairées.

1.4.2 Critique de l'existant

La critique du système existant constitue une étape importante permettant de porter un jugement objectif afin de déceler les insuffisances éventuelles rencontrées au cours de l'étude de l'existant. L'analyse de l'existant nous a permis d'extraire les défaillances suivantes :

- La collecte et le stockage des données sont principalement manuels, entraînant des retards dus au volume massif de données générées quotidiennement.
- Le processus actuel de stockage via des requêtes SQL, PL/SQL et des scripts Shell est inefficace et peu évolutif, limitant la gestion et l'utilisation efficace des données.

- Le suivi des différents trafics repose sur des rapports Excel, limitant la capacité des analystes à effectuer une analyse approfondie et en temps réel.
- L'absence d'analyse prédictive et de détection des fraudes constitue une lacune majeure, entravant la capacité à anticiper les tendances futures du trafic et à identifier les comportements frauduleux.
- Les méthodes manuelles et les processus rigides entravent l'efficacité opérationnelle et la réactivité aux changements du marché.

En résumé, l'existant présente des défis significatifs en matière de collecte, de stockage, d'analyse et d'utilisation des données, nécessitant des améliorations pour une gestion plus efficace et une prise de décision stratégique optimisée.

1.5 Solution proposée

Notre solution doit répondre aux besoins suivants :

- **Constituer une source de données souple et adaptable** (*staging area*) pour gérer les volumes importants de données générés quotidiennement par les différents systèmes de Tunisie Télécom.
- **Automatiser l'extraction de données et assurer la mise à jour de l'entrepôt de données** pour maintenir une base de données actuelle et fiable.
- **Créer des tableaux de bord clairs**, permettant de suivre les tendances d'utilisation des abonnés, de comparer l'activité réelle des clients avec les données enregistrées dans le système de Tunisie Télécom pour identifier les écarts et les anomalies, résoudre les incidents et limiter les pertes financières.
- **Developper des modèles prédictifs** pour estimer la probabilité de paiement des clients SOS, afin d'anticiper les flux de trésorerie et de prendre des décisions stratégiques concernant le remboursement des abonnés demandant un crédit après avoir dépassé leur limite d'accès autorisée.

1.6 Methodologie de travail

L'adoption d'une méthodologie de travail est impérative pour garantir la qualité et respecter les délais d'un projet complexe. Cette méthodologie définit les règles de conduite, les rôles des intervenants, l'ordonnancement des tâches et la séquence des actions. Dans le cadre de la préparation de ce système, nous essayons d'étudier les méthodes les plus populaires des méthodes agiles, pour pouvoir choisir la méthode la plus adéquate à ce projet.

1.6.1 Méthodes Agile

Agile est une approche itérative et collaborative qui prend en compte à la fois les besoins initiaux du client et les besoins liés aux changements. Le principe fondamental est de fournir une version minimale du logiciel puis, par un processus itératif, d'intégrer des fonctionnalités supplémentaires à cette base. Le processus itératif consiste en une série d'étapes, répétées autant de fois que nécessaire.

Parmi les méthodes agiles les plus couramment utilisées, on peut citer le Processus Unifié, la Programmation Extrême et le Scrum.

- **Processus Unifié (UP)** : Le Processus Unifié est une méthode de mise en œuvre et de développement principalement utilisée par les développeurs informatiques. C'est un processus itératif et incremental caractérisé par quatre aspects :

- Axé sur les cas d'utilisation.
- L'architecture est au cœur du processus.
- Utilisation maximale des modèles, en particulier des modèles UML.
- Résolution régulière des incertitudes grâce à sa nature itérative et cyclique.

- **Extreme Programming (XP)** : XP est une méthodologie agile de développement logiciel visant à assurer la qualité du code, la satisfaction client et la réactivité aux changements en privilégiant une communication étroite avec le client. Cela conduit à une amélioration de la qualité du code, une communication client plus efficace et une meilleure capacité à

s'adapter aux changements. Cependant, sa mise en œuvre nécessite une forte discipline et un engagement total de la part de l'équipe.

- **Scrum** : Scrum est un cadre de travail pour le développement de produits logiciels complexes. Scrum est considéré essentiel pour l'avancement du projet, la prise de décision et le partage d'informations. L'accent doit être mis sur la communication et la collaboration entre les membres de l'équipe, la qualité plutôt que la quantité , l'acceptation des changements, la forte implication du client et la mise en place des réunions.

1.6.2 Choix de la méthodologie de travail

Nous choisirons l'une des méthodes de développement Agile pour gérer le cycle de vie de notre projet. Ce choix repose sur les avantages offerts par ces méthodes, tels que leur capacité à s'adapter aux changements de l'environnement et à l'instabilité des spécifications qui sont souvent modifiées au cours du développement. Après avoir étudié les différentes approches et évalué leur pertinence pour notre projet, nous avons décidé d'adopter la méthode Scrum.

Scrum offre de nombreux avantages qui en font une méthode idéale pour notre travail. Parmi ces avantages, on trouve :

- * Une gestion plus souple et plus intelligente du travail, améliorant l'efficacité de l'équipe.
- * Une meilleure visibilité du projet et de son avancement.
- * Une communication interne renforcée, entraînant une meilleure cohésion de l'équipe.
- * Le partage des connaissances et la favorisation de l'entraide.
- * Un gain de temps et une meilleure réactivité grâce à des réunions fréquentes.
- * Sa gestion nécessite un minimum de flexibilité pour intégrer facilement des changements dans les plans initiaux.

Scrum définit un modèle d'équipe qui optimise la flexibilité, la créativité et la productivité.

L'équipe Scrum se compose de :

- **Product Owner** : il s'agit de la personne responsable de chaque phase du processus de développement ainsi que du produit final. Il joue un rôle primordial en inspectant et en évaluant l'avancement du produit à chaque itération.

CHAPITRE 1 : ETUDE PRÉALABLE

- **Scrum Master :** Cette personne agit en tant que coach et effectue les tâches suivantes :
 - * Assurer le bon déroulement et le respect de Scrum.
 - * Encourager l'équipe à apprendre et à progresser afin d'être plus productive et créative tout au long du projet.
 - * Éliminer les obstacles susceptibles de perturber l'avancement des travaux.

- **L'équipe du projet :** Cette équipe est généralement composée de 2 à 10 développeurs.

Elle réunit les rôles nécessaires à la réalisation d'un projet. Son rôle principal est de :

- * Transformer les besoins exprimés dans le Sprint Backlog en fonctionnalités utilisables.
- * Livrer régulièrement une version fonctionnelle du produit.

La méthode Scrum repose sur le concept de "Sprint", qui est un intervalle de temps pendant lequel l'équipe Scrum va réaliser un travail décrit dans le backlog du sprint au début de cet intervalle de temps, guidé par un objectif. Le Backlog du sprint est une liste de besoins classés par ordre de priorité par le Product Owner et évalués par l'équipe. Le Backlog du sprint est un extrait du Backlog produit. La mêlée quotidienne de quinze minutes chaque jour est d'une grande importance. Elle permet aux membres de l'équipe de se mettre à jour, de partager les problèmes rencontrés et de vérifier la progression du sprint. Le produit livrable est remis au Client une fois que tous les objectifs ont été atteints.

La Figure 1.2 illustre le fonctionnement de Scrum.

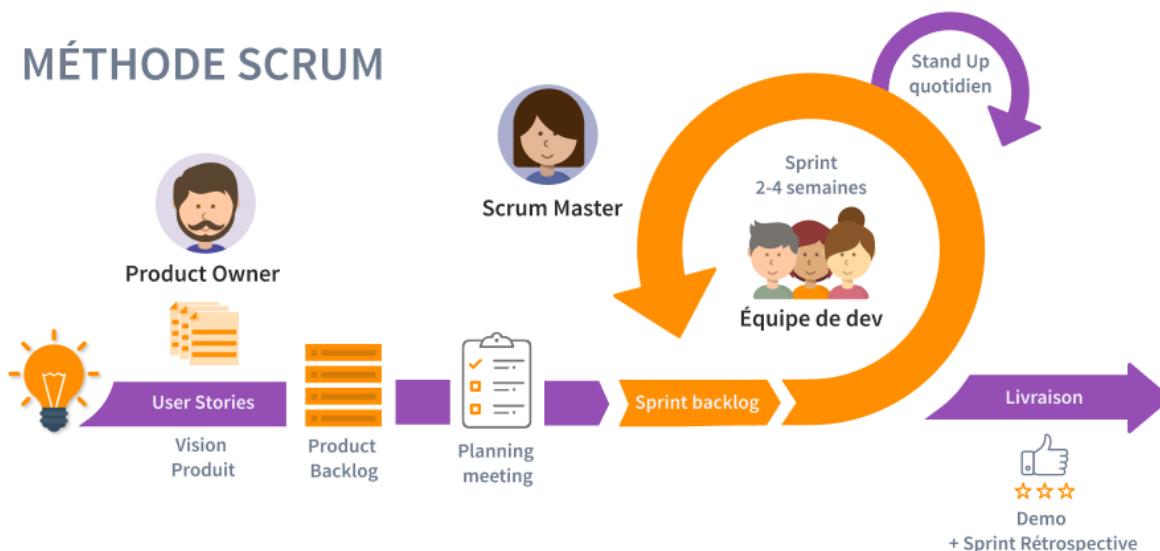


FIGURE 1.2 – Fonctionnement de Scrum

1.7 Conclusion

Dans ce chapitre, nous avons présenté le cadre général de notre projet, à savoir l'organisme d'accueil, la définition de la mission et les objectifs à atteindre. Par ailleurs, nous avons mené une étude de l'existant et sur la base des critiques dégagées de cette étude, nous avons défini la solution adéquate. De plus, nous avons précisé la méthodologie de gestion de projet que nous allons adopter.

Chapitre 2 : L'informatique décisionnelle

Sommaire

2.1	Introduction	14
2.2	Concepts généraux du BI	14
2.2.1	Le Business Intelligence	14
2.2.2	Les avantages du BI	14
2.2.3	Les limites du BI	15
2.3	Les principes des systèmes décisionnels	15
2.3.1	Sources de données	16
2.3.2	Entrepôt de données	16
2.3.3	Magasin de données	17
2.3.4	Extract - Transform - Load	17
2.4	Modélisation multidimensionnelle	17
2.4.1	Définition	17
2.4.2	Concepts de base	18
2.4.3	Schéma multidimensionnel	19
2.4.4	L'objectif de la modélisation multidimensionnelle	19
2.5	Démarche de construction d'un entrepôt de données	20
2.5.1	Modélisation et conception de l'entrepôt	20
2.5.2	Alimentation de l'Entrepôt	21
2.5.3	Administration et maintenance	23
2.6	Analyse et Fouille de Données (Data Mining)	23
2.6.1	Définition	23
2.6.2	L'utilité de l'analyse et fouille de données	24
2.6.3	Concept de Base	24
2.7	Conclusion	25

2.1 Introduction

Ce chapitre est consacré à la définition de la veille stratégique. Nous débuterons par examiner ses avantages et ses limites. Ensuite, nous approfondirons les termes et concepts clés, en proposant des explications détaillées sur l'ETL (Extraction, Transformation, Chargement) ainsi que sur l'entrepôt de données.

2.2 Concepts généraux du BI

2.2.1 Le Business Intelligence

L'informatique décisionnelle, également appelée BI, désigne les moyens, méthodes et outils qui offrent des solutions décisionnelles aux professionnels. Son objectif est de fournir une vue d'ensemble des activités de l'entreprise et de permettre une prise de décision éclairée grâce à l'utilisation de tableaux de bord de suivi et d'analyses.

2.2.2 Les avantages du BI

La mise en place d'une solution BI apporte de nombreux avantages :

- * Une meilleure visibilité des chiffres, des écarts et des anomalies.
- * La possibilité de combiner plusieurs sources de données (ERP, systèmes comptables, feuilles de calcul, budgets, etc.).
- * Une présentation cohérente d'informations fiables.
- * Une automatisation qui accélère la collecte et la diffusion des données.
- * Le calcul efficace d'agrégats pour de grands volumes de données.
- * Prise de décision facilitée par des indicateurs pertinents et une structure d'information cohérente.
- * Aide au nettoyage des données provenant de différents systèmes logiciels.
- * Anticipation des événements et projections futures.

2.2.3 Les limites du BI

Parmi les limites de la veille stratégique, on peut citer les suivantes :

- * La mise en œuvre d'une solution de BI prend beaucoup de temps, ce qui peut ne pas convenir à des entreprises évoluant dans des secteurs où tout va très vite.
- * La complexité est un autre inconvénient de la BI, notamment dans la mise en œuvre des données.
- * Des erreurs peuvent se produire dans les résultats produits par les systèmes décisionnels en raison de la complexité des conceptions informatiques et mathématiques. De plus, les résultats sont souvent statistiques et non déterministes, ce qui implique de prendre en compte la possibilité d'erreurs ou d'approximations inappropriées dans la prise de décision.

2.3 Les principes des systèmes décisionnels

Le système décisionnel est architecturé comme suit :

- * De multiples sources de données sont lues, et un entrepôt de données fusionne les données nécessaires.
- * Un processus ETL est utilisé pour alimenter l'entrepôt de données avec les données existantes.
- * Des marts de données sont utilisés pour simplifier l'entrepôt de données.
- * Des applications d'exploration de données sont utilisées pour présenter l'étude aux utilisateurs finaux et aux décideurs.

La figure 2.1 représente l'architecture générale d'un système décisionnel.

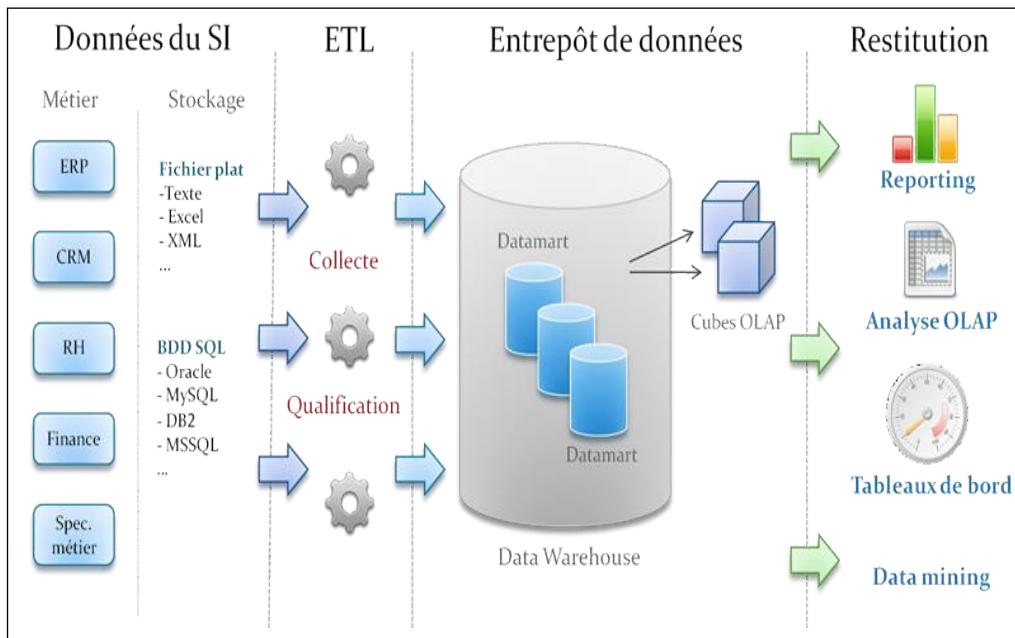


FIGURE 2.1 – Architecture générale d'un système décisionnel

2.3.1 Sources de données

Afin d'alimenter l'entrepôt, les informations doivent être identifiées et extraites de leurs emplacements d'origine. Il peut s'agir de sources de données hétérogènes comprenant des données internes à l'entreprise, stockées dans diverses bases de données de production. Il peut également s'agir de sources externes, récupérées par les services distants et des services Web, ou de sources sous forme de fichiers plats.

2.3.2 Entrepôt de données

Selon BILL Inmon : " *Un entrepôt de données est une collection de données thématiques, intégrées, non volatiles et historiques organisées pour la prise de décision* [3].

Sur la base de cette définition, nous pouvons distinguer les caractéristiques suivantes :

- * Données orientées vers un sujet : Les données de l'entrepôt sont organisées par sujet et triées par thème.

- * Données intégrées : Les données provenant de différentes sources doivent être intégrées avant d'être stockées dans l'entrepôt de données. Un nettoyage préalable des données est nécessaire pour assurer la cohérence et la normalisation des informations.
- * Données non volatiles : Contrairement aux données opérationnelles, les données de l'entrepôt sont permanentes et ne peuvent pas être modifiées. Le rafraîchissement de l'entrepôt consiste à ajouter de nouvelles données sans perdre les données existantes.
- * Données historiques : Les données doivent être datées.

2.3.3 Magasin de données

Chaque mart est un extrait de l'entrepôt. Personnalisé pour un groupe de décideurs ou une utilisation spécifique, un magasin contient uniquement les données d'un domaine d'activité d'une entreprise, tandis qu'un ED contient toutes les données décisionnelles de l'ensemble de l'entreprise, tous domaines confondus.

2.3.4 Extract - Transform - Load

ETL, qui signifie Extraction Transformation Loading, est un processus d'intégration de données. Il consiste à transférer des données brutes à partir d'un système source, à les préparer pour une utilisation en aval et à les envoyer à l'entrepôt de données. Ce système doit guider les données à travers divers processus pour les dénormaliser, les nettoyer, les contextualiser et les charger de manière appropriée. Cependant, la réalisation de l'ETL est une étape très importante et complexe, car elle représente en moyenne 70 % d'un projet de prise de décision [4].

2.4 Modélisation multidimensionnelle

2.4.1 Définition

La modélisation multidimensionnelle, introduite par Ralph Kimball, est une méthode de conception logique visant à présenter les données de manière normalisée et intuitive, permettant un accès très efficace. Elle adhère pleinement à la dimensionnalité et suit une approche disciplinée dans

l'utilisation du modèle. Cette méthode considère un sujet analysé comme un point dans un espace multidimensionnel, organisant les données pour mettre en évidence le sujet analysé et les différentes perspectives de l'analyse. Conceptuellement, la modélisation multidimensionnelle a donné naissance aux concepts de fait et de dimension [5].

2.4.2 Concepts de base

- * **Fait** : Il s'agit d'un point central d'intérêt pour la prise de décision, modélisant l'objet de l'analyse. Il englobe un ensemble d'attributs numériques représentant les mesures liées aux informations de l'activité analysée, ainsi que les identifiants associés aux dimensions [6].
- * **Mesure** : C'est un indicateur d'analyse numérique et agréable, accompagné d'un ensemble de fonctions d'agrégation permettant de l'agréger en fonction des axes d'analyse. Les mesures peuvent être additives, semi-additives ou non additives.

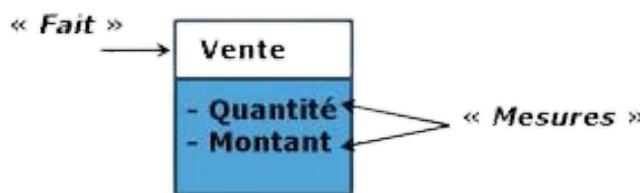


FIGURE 2.2 – Exemple de fait

- * **Dimensions** : représentent les différents axes d'analyse autour desquels les données sont organisées. Elles permettent de regrouper et de catégoriser les données en fonction de leurs attributs communs [6].
 - Les paramètres d'une dimension peuvent être accompagnés de descripteurs appelés attributs faibles qui ne sont pas utilisés dans les calculs de regroupement.
 - Une hiérarchie est une perspective d'analyse définie dans une dimension. Elle regroupe un ensemble de paramètres organisés de la granularité la plus fine vers la granularité la plus générale

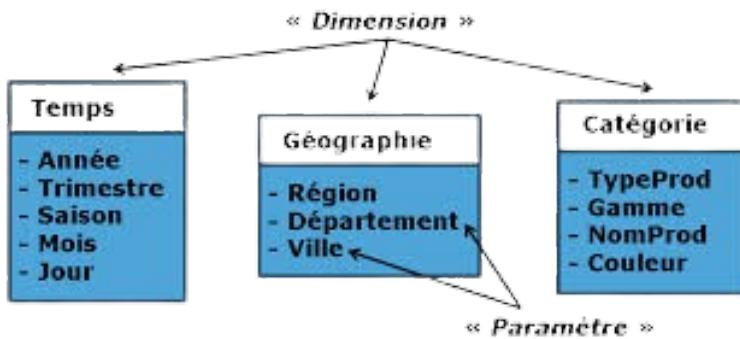


FIGURE 2.3 – Exemple de dimension

2.4.3 Schéma multidimensionnel

2.4.3.1 Schéma en étoile :

Il se caractérise par un seul sujet d'analyse (fait) contenant un ou plusieurs indicateurs (mesures) et plusieurs axes d'analyse (dimensions), y compris les descripteurs des dimensions. Chaque dimension est décrite par une seule table, où les attributs représentent les différentes granularités possibles.

2.4.3.2 Schéma en flocon de neige :

Une modélisation en flocon de neige consiste à décomposer les dimensions du modèle en étoile en sous-hierarchies. La modélisation en flocon de neige est donc une extension de la modélisation en étoile ; le fait est conservé et les dimensions sont étendues en fonction de la hiérarchie des paramètres.

2.4.3.3 Schéma en constellation :

Il s'agit de fusionner plusieurs modèles en étoile qui utilisent des dimensions communes. Un modèle en constellation comprend donc plusieurs faits et des dimensions communes ou non.

2.4.4 L'objectif de la modélisation multidimensionnelle

L'objectif principal de la modélisation multidimensionnelle est de répondre aux besoins d'analyse des utilisateurs en définissant les exigences qui déterminent les données nécessaires. Ce processus commence par la construction d'une matrice qui représente les processus commerciaux clés et

leurs dimensions. Ensuite, une analyse détaillée des données provenant des systèmes sources est effectuée.

Cette approche consiste à organiser les données en fonction de plusieurs dimensions afin de faciliter l'analyse. Elle est connue sous le nom de modélisation et de traitement des données multidimensionnelles.

Le résultat de cette approche est la création d'un modèle dimensionnel qui identifie la granularité des données factuelles, les dimensions associées avec leurs attributs et leurs hiérarchies. Enfin, ce processus se termine par l'établissement d'une correspondance entre les données sources et les données cibles dans les métadonnées.

2.5 Démarche de construction d'un entrepôt de données

L'entreposage de données se déroule en quatre phases principales :

1. Modélisation et conception de l'entrepôt.
2. Alimentation de l'entrepôt.
3. Mise en œuvre de l'entrepôt.
4. Administration et maintenance de l'entrepôt.

2.5.1 Modélisation et conception de l'entrepôt

Les approches les plus connues dans la conception des entrepôts sont :

- * **Approche Descendante** : Dans cette approche, le contenu de l'entrepôt est déterminé selon les besoins de l'utilisateur final. Les instructions sont données en amont et les objectifs du projet sont fixés par la direction.
- * **Approche Ascendante** : Cette approche détermine le contenu de l'entrepôt selon les sources de données. C'est un processus analytique qui examine des données de base pour en tirer un schéma multidimensionnel offrant une vision analytique des données.
- * **Approche Mixte** : C'est une approche hybride qui combine les approches ascendante et descendante. Elle prend en considération les sources de données et les besoins des utilisateurs.

2.5.2 Alimentation de l'Entrepôt

Une fois l'entrepôt conçu, il faut l'alimenter et le charger en données. Cette alimentation s'effectue à travers le processus ETL (Extraction, Transformation, Chargement) et se déroule en trois phases :

2.5.2.1 Extraction des Données :

La première étape consiste à récupérer les informations dans l'environnement de l'entrepôt de données. Cette étape est cruciale car elle permet de collecter les données brutes provenant de diverses sources, telles que des bases de données transactionnelles, des fichiers plats, des API, etc. On distingue deux principaux types d'extraction :

- **L'extraction complète** : Dans ce cas, toutes les données sont extraites à partir des sources à chaque exécution du processus ETL. C'est utile lorsque les données sources sont relativement petites ou lorsque l'intégralité des données est nécessaire à chaque chargement de l'entrepôt.
- **L'extraction incrémentale** : Contrairement à l'extraction complète, l'extraction incrémentale ne récupère que les données qui ont été ajoutées ou modifiées depuis la dernière exécution du processus ETL. Cela permet d'économiser du temps et des ressources en ne traitant que les changements récents, ce qui est particulièrement avantageux pour les sources de données volumineuses ou qui évoluent fréquemment.

2.5.2.2 Transformation des données :

Une fois les données extraites, on applique plusieurs étapes de transformation pour les rendre homogènes et cohérentes. Ces transformations sont essentielles pour préparer les données à être stockées dans l'entrepôt de données et à être utilisées pour l'analyse ultérieure. Les étapes de transformation comprennent :

- * Résolution des cas d'informations manquantes : Identifier et gérer les valeurs manquantes dans les données en utilisant des techniques telles que l'imputation de données ou la suppression des enregistrements.

- * Combinaison des sources de données : Fusionner les données provenant de différentes sources pour créer un jeu de données unifié. Cela peut impliquer l'alignement des schémas, la normalisation des données et la déduplication.
- * Construction d'agrégats : Agréger les données pour résumer les informations sur une granularité supérieure. Par exemple, regrouper les ventes par mois ou par région pour obtenir des statistiques agrégées.
- * Application de filtres : Sélectionner les données pertinentes en fonction de critères spécifiques. Cela peut inclure l'exclusion de données obsolètes, la restriction aux données d'une certaine période de temps ou la sélection de catégories spécifiques.

2.5.2.3 Chargement des données :

C'est la dernière phase de l'alimentation de l'entrepôt. Une fois que les données ont été extraites, transformées et préparées, elles doivent être chargées dans l'entrepôt de données. Cette phase comprend trois types de chargement :

- **Chargement initial** : Dans ce type de chargement, l'intégralité des données est chargée pour la première fois dans l'entrepôt. C'est souvent utilisé lors de la création initiale de l'entrepôt ou lors d'une mise à jour majeure de la structure des données.
- **Chargement incrémental** : Contrairement au chargement initial, le chargement incrémental ne charge que les données qui ont été ajoutées ou modifiées depuis la dernière exécution du processus d'alimentation. Cela permet de maintenir l'entrepôt à jour avec les dernières informations sans avoir à recharger l'intégralité des données à chaque fois.
- **Chargement complet** : Ce type de chargement consiste à charger l'intégralité des données à chaque exécution du processus d'alimentation, qu'il y ait eu des modifications ou non. Bien que cela puisse être plus simple à mettre en œuvre, cela peut être coûteux en termes de temps et de ressources, surtout si les données sources sont volumineuses.

2.5.3 Administration et maintenance

Cette étape comprend plusieurs tâches essentielles pour garantir la qualité et la pérennité des données, la maintenance du système, la gestion de l'évolution, et la documentation. Parmi les principales tâches à accomplir :

- **Assurer la qualité des données** : Mettre en place des processus de contrôle de la qualité pour vérifier la précision, la cohérence et l'intégrité des données stockées dans l'entrepôt. Cela peut inclure la validation des données entrantes, la détection et la correction des erreurs, et la surveillance continue de la qualité des données.
- **Maintenir le système** : Veiller à ce que le système d'entrepôt de données fonctionne de manière optimale en effectuant des tâches de maintenance régulières telles que la sauvegarde des données, l'optimisation des performances, et la gestion des capacités.
- **Gérer l'évolution des besoins** : Adapter l'entrepôt de données aux évolutions des besoins métier en ajoutant de nouvelles sources de données, en modifiant les structures de données existantes, ou en mettant à jour les processus ETL pour répondre aux nouveaux cas d'utilisation.
- **Documenter le système** : Fournir une documentation complète et à jour sur la structure de l'entrepôt de données, les processus ETL, les schémas de données, et les règles métier associées. Cela permet de faciliter la compréhension du système par les utilisateurs et les développeurs, ainsi que la résolution des problèmes et la maintenance future.

2.6 Analyse et Fouille de Données (Data Mining)

2.6.1 Définition

L'analyse et Fouille de données est une pratique qui consiste à rechercher automatiquement un grand volume de données pour découvrir des comportements, des modèles et des tendances qui ne peuvent pas être trouvés par une simple analyse. Son but est de permettre aux entreprises de prendre des décisions proactives et fondées sur la connaissance qui leur donneront un avantage sur leurs concurrents [7].

Si les entrepôts de données sont utilisés pour l'analyse descriptive (ce qui s'est passé) et l'analyse diagnostique (pourquoi c'est arrivé), les entreprises doivent aller plus loin. Elles doivent utiliser l'analyse et l'exploration des données pour l'analyse prédictive (ce qui va se passer) et l'analyse prescriptive (comment pouvons-nous faire en sorte que cela se produise).

2.6.2 L'utilité de l'analyse et fouille de données

Aujourd'hui, le "data mining" est utilisé dans divers secteurs tels que la recherche, le marketing, le développement de produits, les soins de santé et l'éducation. Ce processus permet de résoudre rapidement des problèmes qui nécessitaient auparavant beaucoup de temps pour être traités manuellement. En utilisant différentes techniques statistiques pour analyser les données, les utilisateurs peuvent identifier des modèles, des tendances et des corrélations qui n'étaient pas clairs au départ. Grâce aux résultats de plusieurs analyses successives, ils peuvent prédire les résultats potentiels et prendre des mesures pour influencer et maximiser les résultats de l'entreprise. Lorsqu'il est utilisé efficacement, le data mining peut donner aux organisations un avantage significatif sur leurs concurrents. Il permet de mieux comprendre les clients, de développer des stratégies marketing efficaces, d'augmenter les revenus et de réduire les coûts.

2.6.3 Concept de Base

L'utilisation du "data mining" implique de comprendre les tâches utilisées pour résoudre les problèmes des entreprises. Ces tâches comprennent la classification, l'estimation, le regroupement, la prédiction, la séquence et l'association .

Classer : Catégoriser en fonction de différents attributs. Par exemple, catégoriser un client potentiel en fonction d'autres données telles que l'âge, le sexe, l'état civil, la profession, le niveau d'études, etc. Parmi les algorithmes de classification, on peut citer les arbres de décision, les règles de classification, les réseaux neuronaux.

Estimation : L'estimation se fait à l'aide de paramètres. Par exemple, les prix des maisons seront prédits en fonction de l'emplacement de la maison, de sa taille, etc.

Regroupement (clustering) : Également connu sous le nom de segmentation. Un regroupement naturel est effectué en fonction de différents attributs. La segmentation de la clientèle est un exemple commercial classique de regroupement. Parmi les algorithmes de regroupement, on peut citer les k-means, le regroupement hiérarchique, les essaims dynamiques, la classification pyramidale.

Prédire : Prédire des variables continues dans le temps. Prédire le volume des ventes pour les deux prochaines années est un scénario courant dans l'industrie. Certains algorithmes de prédiction comprennent la régression linéaire, la régression non linéaire, etc.

Associer : Recherche d'éléments ou de groupes communs dans une transaction.

Séquence : Prévoir la suite des événements.

2.7 Conclusion

Dans ce chapitre, nous avons détaillé toutes les notions relatives aux systèmes décisionnels (les avantages , les limites, l'architecture d'un système décisionnel, la modélisation multidimensionnelle, la démarche de construction d'un ED) et au Data Mining pour favoriser le bon déroulement du projet.

Chapitre 3 : Sprint 0 - Analyse et spécification des besoins

Sommaire

3.1	Introduction	27
3.2	Compréhension du domaine	27
3.2.1	Notion de fraude	27
3.2.2	Call Detail Record (CDR)	28
3.2.3	Flux télécoms et systèmes de taxation	29
3.3	Analyse des besoins	32
3.3.1	Identification des Acteurs	32
3.3.2	Les besoins fonctionnels	33
3.3.3	Les besoins non fonctionnels	33
3.3.4	Diagramme de cas d'utilisation global	34
3.4	Pilotage du projet avec scrum	35
3.4.1	L'équipe SCRUM	35
3.4.2	Backlog du produit	36
3.4.3	Planification des sprints	37
3.5	Choix des outils de développement	38
3.6	Conclusion	39

3.1 Introduction

Dans ce chapitre, nous approfondissons le concept du sprint 0, axé sur l'« Analyse et spécification des besoins ». Nous commençons par une exploration approfondie du domaine, en identifiant les acteurs clés et en déterminant les besoins fonctionnels et non fonctionnels du système décisionnel. Ensuite, nous présentons en détail le backlog du produit, suivi d'une discussion sur la planification des sprints à venir. Enfin, nous examinons de près les outils d'informatique décisionnelle que nous avons choisis pour la réalisation du projet.

3.2 Compréhension du domaine

3.2.1 Notion de fraude

La fraude est caractérisée par une action intentionnelle d'un individu, d'un groupe (par exemple, un syndicat) ou d'une entreprise (par exemple, un partenaire), visant à obtenir des produits, des services et des revenus auprès d'un prestataire de services cible par le biais de la tromperie, sans payer la valeur attendue pour ces produits ou services. Dans le domaine des télécommunications, la fraude se manifeste par une utilisation du réseau de télécommunication dans le but d'éviter tout paiement. Ceci peut entraîner des conséquences néfastes sur leur rentabilité et leur réputation.

Parmi les impacts les plus courants dans le domaine des télécommunications, on peut citer :

- Des pertes financières importantes dues à une diminution des revenus ou à des coûts supplémentaires liés à la prévention et à la lutte contre la fraude.
- Une détérioration de la confiance des clients et des partenaires commerciaux.
- Une charge de travail accrue pour les équipes chargées de la détection et de la prévention de la fraude.
- Des sanctions réglementaires et juridiques pouvant entraîner des amendes, des poursuites pénales ou des pertes de licences d'exploitation.

3.2.2 Call Detail Record (CDR)

Un enregistrement détaillé des appels (Call Detail Record ou CDR) est un fichier créé par un centre téléphonique, contenant les informations détaillées d'un appel établi via ce centre. Les CDR fournissent des données essentielles pour vérifier l'activité réseau et établir la facturation des appels.

En plus de son utilisation pour la facturation, les enregistrements détaillés peuvent être exploités par Tunisie Télécom pour identifier les appels erronés et fournir des données pertinentes pour soutenir leurs opérations. De plus, ils permettent d'estimer les niveaux de trafic et documente en détail les informations relatives à un appel téléphonique ayant transité par le périphérique concerné. Le CDR comprend plusieurs champs de données décrivant les différentes opérations de télécommunication, notamment :

- * Le numéro de téléphone de l'abonné origine de l'appel (appelant).
- * Le numéro de téléphone qui reçoit l'appel (appelé).
- * L'heure de début de l'appel (date).
- * La durée de l'appel (DURATION).
- * L'identification de l'équipement de commutation téléphonique (EQUIPEMENT ID).
- * Un numéro de séquence identifiant l'enregistrement (RECORD-ID).
- * La disposition ou les résultats de l'appel, en indiquant par exemple si l'appelé est occupé, ou l'appel a échoué.

CHAPITRE 3 : SPRINT 0 - ANALYSE ET SPÉCIFICATION DES BESOINS

Une description détaillée sur les champs du CDR est présentée par la figure 3.1.

Column name	Data type	Description
START_DATE	Date	La date du CDR
START_HOUR	Number(2)	L'heure du CDR
CALLING_NO_GRP	Varchar2(100 char)	CALLING operator telecom name
A_IMSI	Varchar2(50 char)	International Mobile Subscriber Identity
A_MSISDN	Varchar2(50 char)	Mobile Station International ISDN Number Calling
CALLED_NO_GRP	Varchar2(100 char)	CALLED operator telecom name
B_MSISDN	Varchar2(50 char)	Mobile Station International ISDN Number Called
CALL_REFERENCE	Varchar2(50 char)	IDENTIFENT UNIQUE D'APPEL
CALL_TYPE	Varchar2(50 char)	TYPE DE CDR
C_NUM	Varchar2(50 char)	Mobile Station International ISDN Number Call Forward
EVENT_DURATION	Number	durée d'appel
EVENT_TYPE	Varchar2(5 char)	Type d'événement du CDR
FILENAME	Varchar2(100 char)	Nom du fichier CDR généré par la plateforme
IMEI	Varchar2(50 char)	International Mobile Equipment Identity
ORIG_START_TIME	Varchar2(50 char)	timestamp
RECORD_TYPE	Varchar2(50 char)	
SUBSCRIBER_TYPE	Varchar2(50 char)	Prepaid or Postpaid or Hybrid (profil abonné)
PORTABILITY_FLAG	Varchar2(50 char)	
TRUNK_IN	Varchar2(50 char)	
TRUNK_OUT	Varchar2(50 char)	
PORTABILITY_FLAG	Varchar2(50 char)	
TRUNK_IN	Varchar2(50 char)	
TRUNK_OUT	Varchar2(50 char)	
LOAD_DATE	Date	la date d'insertion de la CDR dans la table détail

FIGURE 3.1 – Les dimensions CDR MSC

3.2.3 Flux télécoms et systèmes de taxation

Tunisie Télécom propose une variété de flux pour ses services, notamment les **flux de recharge**, les **flux de voix**, les **SMS**, les **SMS+** et les **flux de données**.

Pour garantir le bon fonctionnement et la validation de ces flux, Tunisie Télécom utilise plusieurs systèmes de taxation tels que le **système Air**, le **système CCN** et le **système OCC**.

Voici une description détaillée des services télécoms disponibles pour chacun de ces flux ainsi que des systèmes de taxation associés (voir tableau 3.1).

3.2.3.1 Services Telecom

a) Services du flux de recharge

* **ETOPUP (Electronic Top Up)** : Il s'agit d'une plateforme de recharge électronique qui permet de recharger les crédits téléphoniques via différents canaux (mobile, borne, caisse PC, etc.) de manière simple et centralisée pour tous les opérateurs.

* **VOUCHER** : Ce système contient les cartes de recharge et leur statut, permettant la vente de codes de recharge de manière sécurisée.

* **USSD (Unstructured Supplementary Service Data)** : C'est un système de communication en temps réel entre le téléphone mobile et le réseau de l'opérateur, utilisé par exemple pour vérifier le solde ou activer des services.

b) Services du flux voix, SMS et SMS+

* **MSC (Mobile Switching Center)** : C'est comme le centre de contrôle d'un réseau de téléphonie mobile. Il s'occupe des événements tels que les appels vocaux et les SMS. Il dirige le trafic des appels, assure la connexion avec d'autres réseaux et organise les appels téléphoniques et les SMS.

* **MMG (Multimedia Management Gateway)** : Ces systèmes gèrent les contenus multimédias comme les jeux HTML5 et les services à valeur ajoutée (VAS) accessibles via les téléphones mobiles.

c) Services du flux de données transférées

* **SASN (Service Access and Selection Network)** : Ce sont des systèmes de transfert de données pour les réseaux mobiles, comme la 3G et la 4G.

3.2.3.2 Systèmes de taxation

* **AIR (Automatic Incident Reporting)** : C'est un système de suivi des cartes de recharge et de leurs transactions associées. Il suit toutes les transactions effectuées sur le système de vouchers

CHAPITRE 3 : SPRINT 0 - ANALYSE ET SPÉCIFICATION DES BESOINS

(cartes de recharge), en enregistrant les numéros de série des cartes utilisées ainsi que les dates auxquelles elles ont été utilisées ou modifiées.

* **CCN (Charging Control Node)** : Il s'agit d'un point de contrôle de la tarification dans un réseau de communication, utilisé notamment dans les réseaux mobiles 3G pour suivre, évaluer et facturer l'utilisation des services de données.

* **OCC (Online Charging Control)** : Online Charging Control est un système de contrôle de tarification en temps réel dans un réseau de télécommunications. Ce système permet de facturer les services en fonction de leur utilisation en temps réel.

	Système de taxation	Services Telecom
Flux de recharge	Air	Etopup (recharge électronique)
		Voucher (recharge par carte)
		USSD (recharge SOS solde)
Flux voix, SMS et SMS+	CCN	MSC (flux voix et SMS)
		MMG (flux SMS+)
Flux données transférées	OCC	SASN

TABLE 3.1 – Flux télécoms et systèmes de taxation

Dans le cadre de notre analyse, nous avons entrepris plusieurs comparaisons entre les différents systèmes de taxation et les services de Tunisie Télécom en fonction de type de flux. Cette démarche, appelée validation, vise à évaluer la cohérence et la précision des mécanismes de taxation par rapport aux différents flux de services.

Nous avons débuté notre analyse en comparant le flux de taxation géré par le système **Automatic Incident Reporting (AIR)** avec différents services de recharge. Nous avons ainsi comparé AIR au flux de recharge physique via les cartes de recharge, représenté par **VOUCHER**, au flux de

recharge électronique fourni par le système **Electronic Top Up (ETOPUP)**, et au flux de SOS solde généré par **Unstructured Supplementary Service Data (USSD)**. Cette analyse nous a permis de mieux comprendre comment AIR assure la taxation des divers systèmes de recharge.

Ensuite, nous avons examiné la façon dont le système de taxation **Charging Control Node (CCN)** gère la tarification des services voix et SMS, en les comparant aux données enregistrées au niveau du **Mobile Switching Center (MSC)**, ainsi qu'à la tarification des services SMS+, gérée par le **Multimedia Management Gateway (MMG)**.

Enfin, nous avons étudié la comparaison entre le système **Online Charging Control (OCC)** et les systèmes de transfert de données tels que le **Service Access and Selection Network (SASN)**. Cette analyse nous a permis d'évaluer la performance du système OCC dans la tarification en temps réel des données par rapport au transfert de données dans les réseaux mobiles, comme géré par le SASN.

Ces comparaisons sont cruciales pour évaluer la fiabilité et l'efficacité des mécanismes de taxation utilisés par Tunisie Télécom, ce qui permettra d'optimiser les processus de tarification et de garantir une expérience utilisateur optimale.

3.3 Analyse des besoins

L'analyse des besoins est la phase initiale de l'élaboration d'un projet. La complexité de l'analyse des besoins tient aussi à ses multiples aspects. Il existe des besoins naturellement exprimés, des besoins implicites, des besoins non reconnus et de simples besoins dont les utilisateurs ne sont même pas conscients.

3.3.1 Identification des Acteurs

Dans un premier temps, nous présenterons les différentes personnes bénéficiant des divers accès aux données :

- * **Developpeur** : C'est l'acteur responsable de la mise en place du système décisionnel.
- * **Analyste** : C'est l'acteur responsable de l'analyse et de la visualisation des résultats obtenus par le système pour aider l'entreprise à prendre les décisions convenables.

3.3.2 Les besoins fonctionnels

Le projet nécessite de répondre à des exigences fonctionnelles qui imposent des contraintes et des conditions pour atteindre des objectifs prédéterminés. Ces exigences sont liées aux fonctions et aux caractéristiques que la solution doit offrir pour répondre aux besoins spécifiques du client.

- * Gérer l'entrepôt de données .
- * Gérer le processus ETL (Extract, Transform, Load) .
- * Mettre en place des tableaux de bord pour donner un sens aux données de l'entreprise.
Ces tableaux de bord offriront une représentation visuelle accessible et compréhensible des données.
- * Introduire des algorithmes d'analyse et de fouille de données pour réaliser des prévisions et dégager des déductions. Cela permettra d'optimiser l'activité d'approvisionnement en identifiant les tendances, les anomalies et les opportunités d'amélioration.

3.3.3 Les besoins non fonctionnels

Les exigences non fonctionnelles représentent des exigences implicites qui définissent les contraintes internes et externes caractérisant le système. Les besoins non fonctionnels du système décisionnel peuvent être décrits comme suit :

- * **Simplicité d'accès aux données** : Les analystes d'affaires demandent des interfaces conviviales offrant un accès simple à l'information, avec un affichage rapide permettant de visualiser les résultats facilement.
- * **Performance** : Le chargement des données (Big Data) nécessite une capacité de traitement et de stockage puissante.
- * **Fiabilité** : La solution doit être sécurisée, accessible et optimiste en terme de temps de réponse, ce qui entraînera une augmentation de l'efficacité de l'application.
- * **Disponibilité des données** : La redondance des matériels et la mise en cluster d'une architecture garantissent une haute disponibilité des données.

3.3.4 Diagramme de cas d'utilisation global

Un diagramme de cas d'utilisation (DCU) est un diagramme UML utilisé pour représenter le comportement fonctionnel d'un système logiciel.

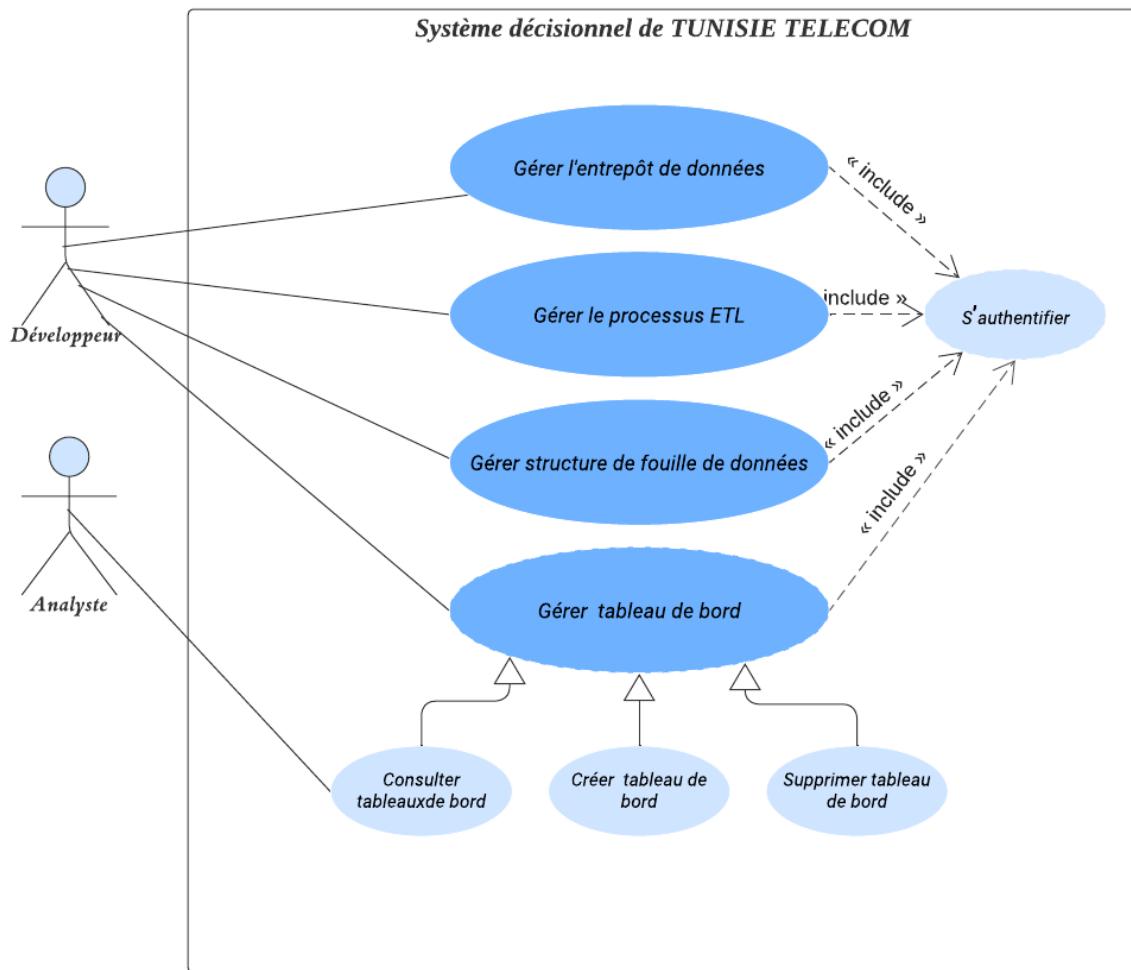


FIGURE 3.2 – Diagramme du cas d'utilisation global du Système décisionnel de TUNISIE TELECOM

3.4 Pilotage du projet avec scrum

Dans cette partie on va détailler les différents composants SCRUM tels que l'équipe, le backlog du produit ainsi que les sprints à développés.

3.4.1 L'équipe SCRUM

La répartition de notre projet se base sur la composition de l'équipe suivante :

Le SCRUM Master : Madame Mouna Ktari, Dr. en informatique à l'Institut Supérieur d'Informatique et Multimedia de Sfax.

Le Product Owner : Monsieur Yessine Brahmi , responsable chez Tunisie Telecom.

Les membres de l'équipe : Wala Ben Rhouma et Souad Achouri.

3.4.2 Backlog du produit

Le tableau 3.2 représente le Backlog du produit de notre projet

ID	Fonctionnalité	ID	User Stories	Priorité
1	Création de l'entrepôt de données	1.1	En tant que développeur, je peux créer l'entrepôt de données (ED).	1
2	Gérer le processus ETL	2.1	En tant que développeur, je peux transférer des données vers la zone de préparation des données (SA).	1
		2.2	En tant que développeur, je peux gérer le processus ETL du données sources vers SA.	2
		2.3	En tant que développeur, je peux gérer le processus ETL du SA vers l'entrepôt de données (ED).	3
		2.4	En tant que développeur, je peux rafraîchir et charger l'ED.	4
3	Gérer les tableaux de bord	3.1	En tant que développeur, je peux créer, modifier et consulter les tableaux de bord.	1
		3.2	En tant qu'analyste, je peux consulter des tableaux de bord de l'ED.	1
4	Analyse et fouille de données	4.1	En tant que développeur, je peux faire des prédictions et des classifications à partir des données de l'ED.	1

TABLE 3.2 – Backlog du produit

3.4.3 Planification des sprints

La planification de sprint constitue une étape essentielle de la méthodologie agile Scrum, où nous fragmentons notre projet en différentes tâches et estimons le temps requis pour les achever, tel que démontré dans la figure 3.3 :

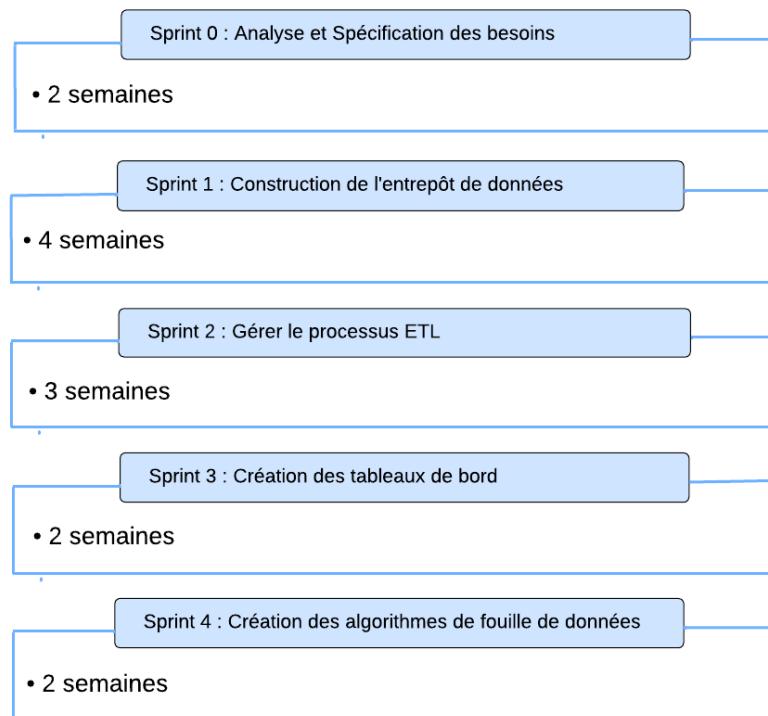


FIGURE 3.3 – Pilotage de projet par Scrum

Diagramme de Gantt : La figure 3.4 montre le diagramme de gantt de planification des sprints :

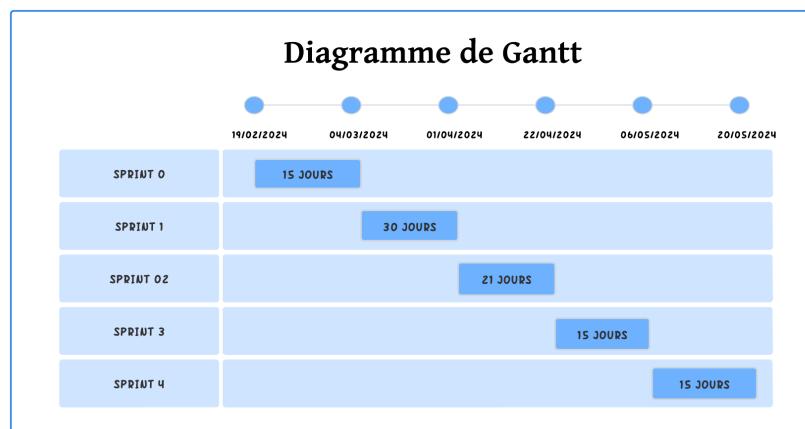


FIGURE 3.4 – Diagramme de Gantt de planification des sprints

3.5 Choix des outils de développement

Oracle SQL Developer



Oracle SQL Developer est un outil de développement SQL fourni par Oracle Corporation pour les développeurs SQL et PL/SQL travaillant avec des bases de données Oracle. C'est une application graphique multiplateforme qui offre une large gamme de fonctionnalités pour faciliter le développement, la gestion et la manipulation des bases de données Oracle [8].

Talend



Talend est une suite logicielle complète d'intégration de données et d'ETL (Extract, Transform, Load), développée par Talend SA. Cette suite offre un ensemble d'outils pour concevoir, développer, déployer et gérer des processus d'intégration de données complexes [9].

power bi



Power BI de Microsoft offre des fonctionnalités pour la création de rapports et de tableaux de bord interactifs basés sur des données provenant de différentes sources [10].

3.6 Conclusion

Dans ce chapitre, nous avons d'abord présenté le contexte du projet et identifié les besoins fonctionnels et non fonctionnels du système décisionnel à développer. Ensuite, nous avons illustré le diagramme de cas d'utilisation du projet, accompagné de descriptions détaillées de chaque cas. Nous avons également élaboré le backlog du produit et précisé la planification des sprints à venir. Enfin, nous avons présenté les outils techniques sélectionnés pour le développement de l'application.

Chapitre 4 : Sprint 1 - Construction de l'entrepôt de données

Sommaire

4.1	Introduction	41
4.2	Sprint backlog	41
4.3	Diagramme du cas d'utilisation de sprint 1	42
4.3.1	Diagramme du cas d'utilisation	42
4.3.2	Description textuelle du cas d'utilisation « Crée l'entrepôt de données »	43
4.4	Schéma conceptuel de la source de données	43
4.5	Modélisation conceptuelle de l'entrepôt de données	44
4.6	Construction de l'entrepôt de données	54
4.6.1	Modélisation dimensionnelle	54
4.6.2	Choix du modèle dimensionnel	55
4.6.3	Modèle physique de l'entrepôt de données	56
4.7	Conclusion	57

4.1 Introduction

Dans ce chapitre, nous nous concentrerons sur la description du sprint 1, intitulé « Construction de l'entrepôt de données ». Ainsi, nous abordons les éléments suivants :

- La conception de la source de données.
- La modélisation conceptuelle de l'entrepôt de données selon l'approche descendante.
- La modélisation physique de l'entrepôt de données.

4.2 Sprint backlog

Le tableau 4.1 représente le Backlog du sprint 1

ID	Fonctionnalité	User Stories	ID	Description des tâches
1	Création de l'entrepôt de données	En tant que développeur, je peux créer l'entrepôt de données	1.1.1 1.1.2 1.1.3	Définir le schéma conceptuel des sources de données. Établir le schéma multidimensionnel de l'entrepôt de données en utilisant l'approche descendante. Construire le modèle physique de l'entrepôt de données.

TABLE 4.1 – Backlog du sprint 1

4.3 Diagramme du cas d'utilisation de sprint 1

4.3.1 Diagramme du cas d'utilisation

La figure 4.1 représente le diagramme du cas d'utilisation de sprint 1

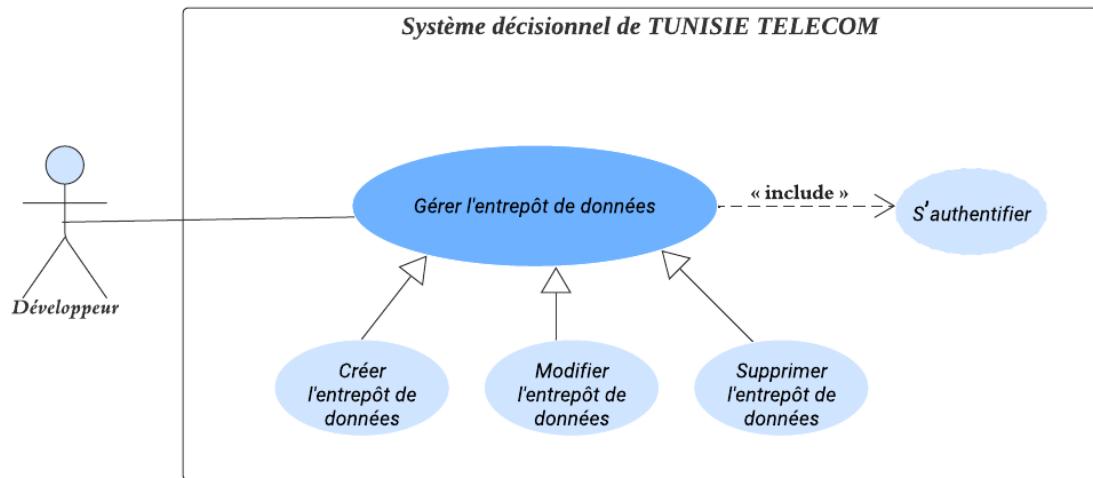


FIGURE 4.1 – Diagramme du cas d'utilisation de sprint 1

4.3.2 Description textuelle du cas d'utilisation « Créer l'entrepôt de données »

Le tableau 4.2 montre la description textuelle du cas d'utilisation « Créer l'entrepôt de données ».

Titre	Créer l'entrepôt de données.
Acteurs	Développeur.
Objectif	Permettre de construire un entrepôt de données.
Pré-condition	Le développeur doit être connecté au système.
Post-condition	Un nouveau entrepôt de données créé.
Scénario nominal	<ol style="list-style-type: none">1. Le développeur identifie le schéma conceptuel de la source de données.2. Le développeur choisit une démarche de conception de l'entrepôt.3. Le développeur identifie les tables de fait.4. Le développeur identifie les tables de dimensions.5. Le développeur identifie les hiérarchies.6. Le développeur construit le schéma multidimensionnel de l'ED.7. Le développeur construit la modélisation de l'ED.

TABLE 4.2 – Description textuelle du cas d'utilisation « Créer l'entrepôt de données »

4.4 Schéma conceptuel de la source de données

Avant d'entamer la conception de notre schéma multidimensionnel, nous avons procédé à l'analyse du diagramme de classe de nos sources de données. Dans notre contexte, ces sources sont constituées de multiples tables présentant une grande hétérogénéité. Cette diversité engendre plusieurs défis. Après avoir examiné attentivement ces tables, nous avons pu établir le diagramme de classe suivant, représenté dans la figure 4.2

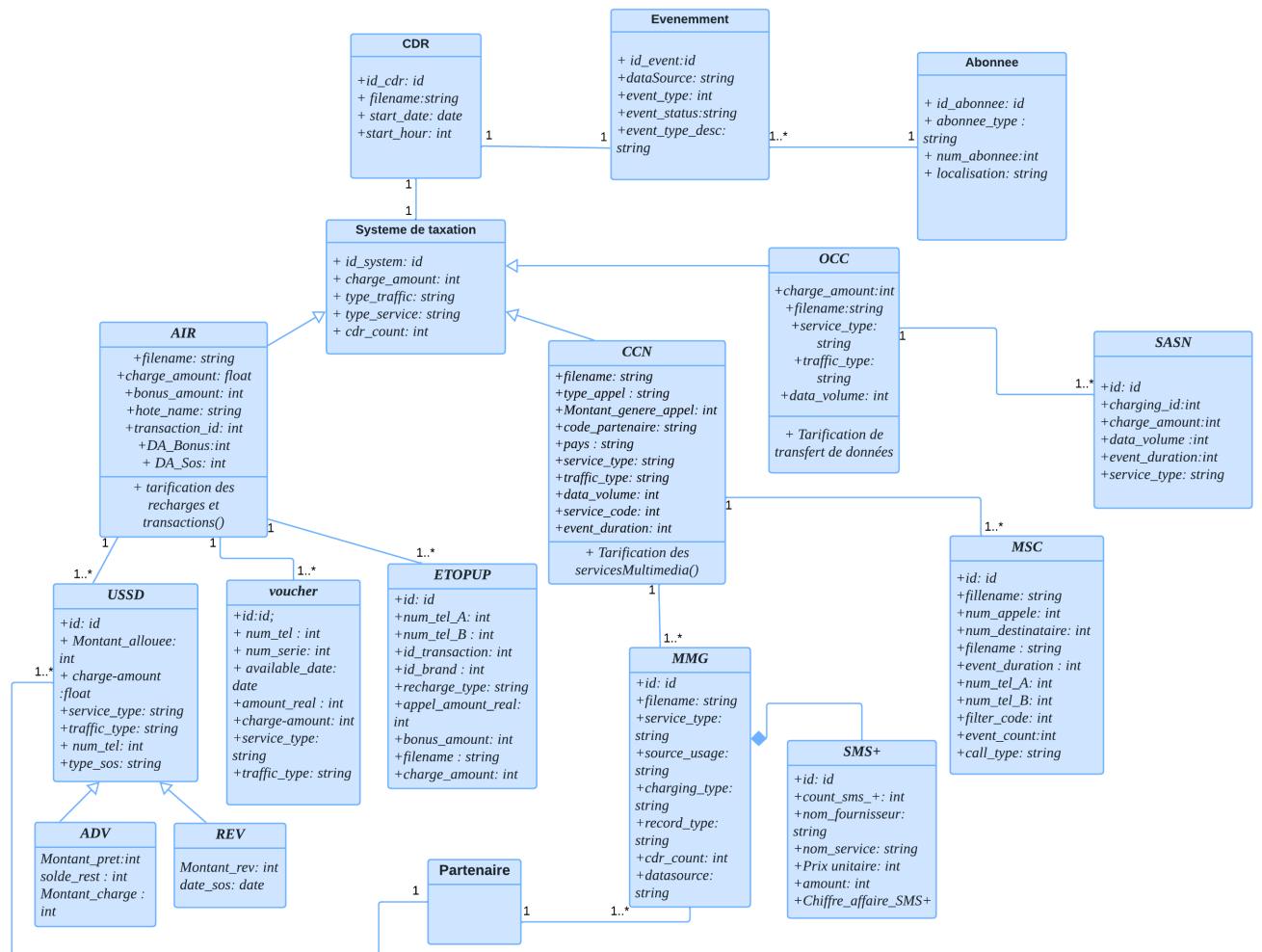


FIGURE 4.2 – Diagramme de classe

4.5 Modélisation conceptuelle de l'entrepôt de données

La modélisation des données est un élément fondamental dans la démarche de spécification de tout type de système d'information. L'entrepôt de données nécessite un type de modélisation spécifique connu sous le nom de modélisation multidimensionnelle (Ines, 2021), qui comprend trois méthodes.

La méthode ascendante se base sur les données sources pour proposer des étapes qui définissent les faits, les dimensions et les hiérarchies dans le but de créer un schéma multidimensionnel. Cependant, lorsque les tables ne sont pas interconnectées et que la base de données est mal organisée, la méthode descendante est plus appropriée.

La méthode descendante ne tient pas compte des données de la source et se concentre uniquement sur les spécifications des besoins des utilisateurs. Elle vise à créer un modèle de données qui répond précisément aux attentes des décideurs, sans être limité par la structure ou la qualité des données existantes.

La méthode mixte se base sur les données de la source ainsi que sur les besoins des décideurs. Cette méthode combine entre la démarche ascendante et descendante .

Dans notre cas, la méthode descendante est la plus adaptée pour la modélisation conceptuelle de notre ED. Nous partons des demandes des analystes et des questionnaires pour créer un modèle de données qui répond à leurs besoins spécifiques.

a) Démarche descendante :

La figure 4.3 illustre les différentes phases de cette méthode. Dans cette section nous présentons la modélisation de l'ED selon la démarche descendante .

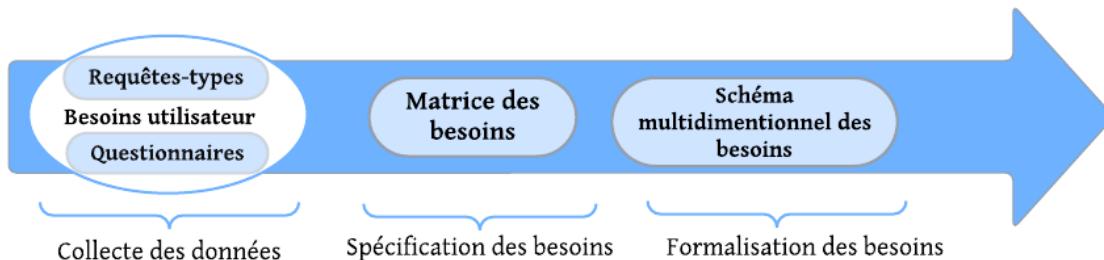


FIGURE 4.3 – Les étapes de l'approche descendante

1. Collecte de données :

Cette étape consiste à

- Interroger les décideurs afin de collecter des requêtes-types pertinentes.
- Identifier leurs besoins à travers des questionnaires. Ainsi, nous avons déterminé les requêtes-types suivantes :

R1 : **Analyser** Air_Event_Count en milliers

En fonction de heure, jour, mois, année, subscriberType **Pour** l'événement de type
Recharge électronique

R2 : **Analyser** ETOPUP_Event_Count en milliers

En fonction de heure, jour, mois, année, subscriberType

R3 : **Analyser** Air_out_charge_amount en milliers en

En fonction de jour, mois, année, subscriberType **Pour** l'évenement de type Recharge
électronique

R4 : **Analyser** ETOPUP_out_charge_amount en milliers

En fonction de jour, mois, année, subscriber_Type

R5 : **Analyser** Air_out_Bonus_amount en millions

En fonction de heure, jour, mois, année, subscriberType

R6 : **Analyser** ETOPUP_out_bonus_amount en millions

En fonction de heure, jour, mois, année, subscriberType

R7 : **Analyser** Voucher_Event_Count en milliers

En fonction de heure, jour, mois, année, subscriberType

R8 : **Analyser** Air_Event_Count en milliers

En fonction de jour, mois, année, subscriberType **Pour** l'événement de type recharge
manuelle

R9 : **Analyser** Voucher_Charge_amount en milliers

En fonction de heure, jour, mois, année, subscriberType

R10 : **Analyser** Air_Charge_Amount en milliers

En fonction de heure, jour, mois, année, subscriberType **Pour** l'événement de type
recharge manuelle

R11 : **Analyser** Air_Event_Count en milliers

En fonction de heure, jour, mois, année **Pour** l'événement de type SOS solde

R12 : **Analyser** USSD_Event_Count en milliers

En fonction de heure, jour, mois, année

R13 : **Analyser** Air_out_charge en milliers

En fonction de heure, jour, mois, année **Pour** l'événement de type SOS solde

R14 : **Analyser** USSD_out_charge_amount en milliers

En fonction de heure , jour, mois, année

R15 : **Analyser** OCC_out_Data_Volume en milliard

En fonction de heure,jour, mois, année

R16 : **Analyser** SASN_out_Data_volume en milliard

En fonction de heure , jour, mois, année

R17 : **Analyser** OCC_out_Event_Count en milliard

En fonction de heure , jour, mois, année

R18 : **Analyser** SANS_Event_Count en milliard

En fonction de heure , jour, mois, année

R19 : **Analyser** MMG_out_Event_Count en millions

En fonction de heure , jour, mois, année subscriber_Type **Pour** l'événement de type
SMS+

R20 : **Analyser** CCN_Event_Count en millions

En fonction de heure , jour, mois, année subscriber_Type **Pour** l'événement de type
SMS+

R21. **Analyser** SMS+_Event_Count en milliers

En fonction de nomFournisseur

R22 : **Analyser** SMS+_Event_Count en milliers

En fonction de nomService

R23 : **Analyser** le chiffre d'affaire SMS+

En fonction de heure,jour , mois, année

R24 : **Analyser** CCN_Event_Count en millions

En fonction de heure, jour, mois, année, CallType **Pour** l'événement de type call

R25 : **Analyser** MSC_Event_Count en millions

En fonction de heure, jour, mois, année, subscriber_Type, Call_Type **Pour** l'événement
de type call

R26 : **Analyser CCN_Event_duration en minute**

En fonction de heure, jour, mois, année, subscriber_Type, Call_Type

R27 : **Analyser MSC_Event_Duration en minute**

En fonction de heure, jour, mois, année, subscriber_Type, Call_Type

R28 : **Analyser la diff_Event_Count_Air_ETOPUP**

En fonction de heure, jour, mois, année, subscriberType

R29 : **Analyser la diff_Bonus_amount_Air_ETOPUP**

En fonction de l'heure, jour, mois, année, subscriberType

R30 : **Analyser la diff_charge_amount_Air_ETOPUP**

En fonction de heure, jour, mois, année, subscriberType

R31 : **Analyser la diff_charge_amount_Air_Voucher**

En fonction de heure, jour, mois, année, subscriberType

R32 : **Analyser la diff_Event_count_Air_Voucher**

En fonction de heure, jour, mois, année, subscriberType

R33 : **Analyser la diff_charge_amount_Air_USSD**

En fonction de heure, mois, année, subscriberType

R34 : **Analyser la diff_Event_count_Air_USSD**

En fonction de heure, jour, mois, année, subscriberType

R35 : **Analyser la diff_Event_count_SASN_OCC**

En fonction de heure, jour, mois, année

R36 : **Analyser la diff_Data_volume_SASN_OCC**

En fonction de heure, jour, mois, année

R37 : **Analyser la diff_Event_count_MMG_CCN**

En fonction de heure, jour, mois, année, subscriberType

R38 : **Analyser la diff_Event_Count_MSC_CCN**

En fonction de heure, jour, mois, année, subscriberType, CallType

R39 : **Analyser la diff_Event_Duration_MSC_CCN**

En fonction de heure, jour, mois, année, subscriberType, CallType

2. Matrice des besoins :

	Heure	Jour	Mois	Année	Subscriber_Type	Call_Type	NomFournisseur	NomService
Air_Event_Count	x	x	x	x	x			
ETOPUP_Event_Count	x	x	x	x	x			
Diff_Event_AIR_ETOPUP	x	x	x	x	x			
AIR_out_charge_amount	x	x	x	x	x			
ETOPUP_out_charge_amount	x	x	x	x	x			
Diff_charge_amount_AIR_ETOPUP	x	x	x	x	x			
AIR_out_Bonus_amount	x	x	x	x	x			
ETOPUP_out_Bonus_amount	x	x	x	x	x			
Diff_Bonus_amount_AIR_ETOPUP	x	x	x	x	x			
VOUCHER_Event_Count	x	x	x	x	x			
VOUCHER_charge_amount	x	x	x	x	x			
Diff_Event_Count_AIR_VOUCHER	x	x	x	x	x			
Diff_charge_amount_AIR_VOUCHER	x	x	x	x	x			
USSD_Event_Count		x	x	x	x			
USSD_out_charge_amount		x	x	x	x			
Diff_Event_Count_AIR_USSD	x	x	x	x	x			
Diff_charge_amount_AIR_USSD	x	x	x	x	x			
OCC_out_Event_Count	x	x	x	x				
SASN_out_Event_Count	x	x	x	x				
Diff_Event_Count_SASN_OCC	x	x	x	x				
OCC_out_Data_volume	x	x	x	x				
SASN_out_Data_volume	x	x	x	x				
Diff_Data_volume_SASN_OCC	x	x	x	x				
MMG_out_Event_Count	x	x	x	x	x			
CCN_Out_Event_Count	x	x	x	x	x			
Diff_Event_Count_MMG_CCN	x	x	x	x				
MSC_Event_Count	x	x	x	x	x	x		
CCN_Event_Count	x	x	x	x	x	x		
Diff_Event_Count_MSC_CCN	x	x	x	x	x	x		
MSC_Event_Duration	x	x	x	x	x			
CCN_Event_Duration	x	x	x	x	x			
Diff_Event_Duration_MSC_CCN	x	x	x	x	x	x		
SMS+_Event_Count							x	x
CA_SMS+	x	x	x	x				

3. Schéma multidimensionnel résultant de la méthode descendante :

***Définition des faits :**

Fact_Validation_MSC-CCN :

MSC_EVENT_count (sum, min, max, avg)

CCN_EVENT_count (sum, min, max, avg)

MSC_Event_duration (sum, min, max, avg)

CCN_Event_duration (sum, min, max, avg)

Fact_Validation_MMG-CCN :

MMG_EVENT_count (sum, min, max, avg)

CCN_EVENT_count (sum, min, max, avg)

chiffre_affaire_SMS+ (count)

SMS+_Event_Count (count)

Fact_Validation_OCC-SASN :

OCC_OUT_DATA (sum, min, max, avg)

SASN_OUT_DATA (sum, min, max, avg)

OCC_EVENT_count (sum, min, max, avg)

SASN_EVENT_count (sum, min, max, avg)

Fact_Validation_AIR-VOUCHER :

VOUCHER_EVENT_count (sum, min, max, avg)

AIR_EVENT_count (sum, min, max, avg)

VOUCHER_CHARGE_amount (sum, min, max, avg)

AIR_CHARGE_amount (sum, min, max, avg)

Fact_Validation_AIR-ETOPUP :

ETOPUP_EVENT_count (sum, min, max, avg)

AIR_EVENT_count (sum, min, max, avg)

ETOPUP_CHARGE_amount (sum, min, max, avg)

AIR_CHARGE_amount (sum, min, max, avg)

ETOPUP_BONUS_amount (sum, min, max, avg)

AIR_BONUS_amount (sum, min, max, avg)

Fact_Validation_AIR-USSD :

USSD_EVENT_count (sum, min, max, avg)

AIR_EVENT_count (sum, min, max, avg)

USSD_CHARGE_AMOUNT (sum, min, max, avg)

AIR_CHARGE_amount (sum, min, max, avg)

* **Définition des dimensions :**

DimDate(CleDate, heure, Jour, Mois, Annee)

DimSubscriber (IdSubscriber, subscriberType)

DimCall (IdCall, callType)

DimSms+ (IdSms+, nomFournisseur, nomService)

* **Définition des hiérarchies :**

H Date = cleDate → heure → jour → mois → année

H Subscriber = idSubscriber → subscriberType

H Call = idCall → CallType

H Sms+ = idSms+ → nomFournisseur

H1 SMS+ = IdSMS+ → nomService

* **Définition du schéma multidimensionnel :**

La figure 4.4 représente le schéma multidimensionnel en notation de Golfarelli en suivant la démarche descendante (Golfarelli, 1998)

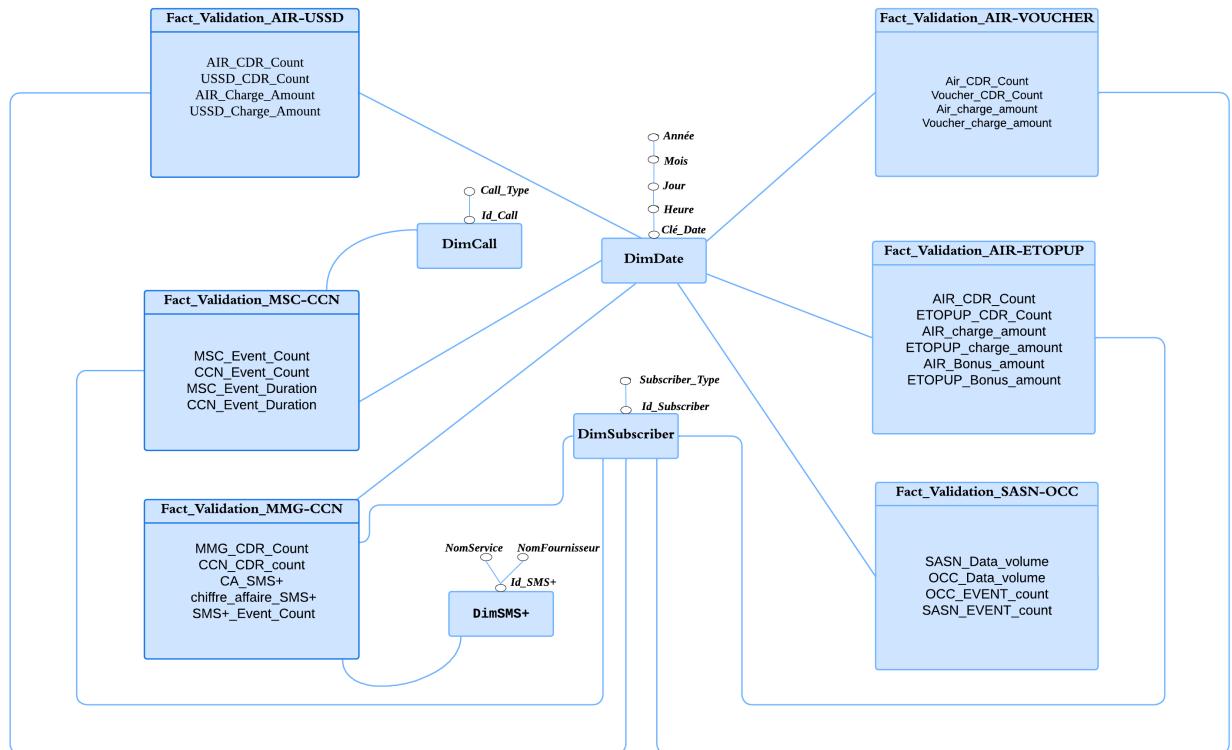


FIGURE 4.4 – Schéma multidimensionnel de la démarche descendante

4.6 Construction de l'entrepôt de données

4.6.1 Modélisation dimensionnelle

La modélisation dimensionnelle est une méthode de conception logique qui permet d'organiser les données. Elles deviennent divisées en faits et dimensions ce qui les rend plus intuitifs aux utilisateurs et plus performants aux requêtes. Il existe 3 méthodes de conception : * **Modèle en étoile :**

Le modèle de données "en étoile" est une structure multidimensionnelle utilisée pour stocker des données atomiques ou agrégées. Ce modèle, souvent qualifié de dénormalisé, est conçu pour optimiser les requêtes analytiques grâce à des jointures simplifiées entre la table de faits centrale et les tables de dimensions associées.

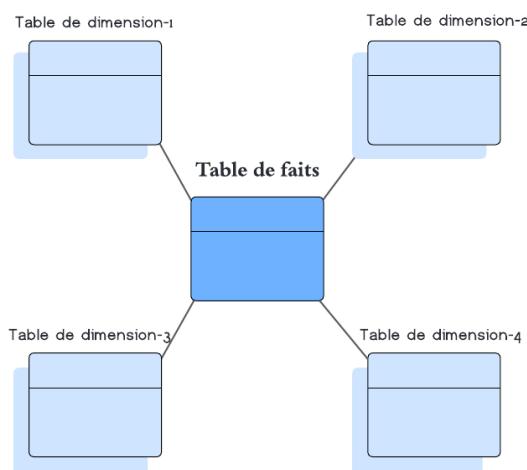


FIGURE 4.5 – Schéma en étoile

* **Modèle en flocon de neige :**

Le modèle de données en "flocon de neige" est une extension du schéma en étoile. Dans ce modèle, chaque table de dimension est normalisée pour représenter les hiérarchies sous-jacentes. Bien que cette normalisation ne soit pas indispensable, car les mises à jour et les suppressions ne se produisent pas directement sur l'entrepôt de données, elle permet une organisation plus détaillée des informations.

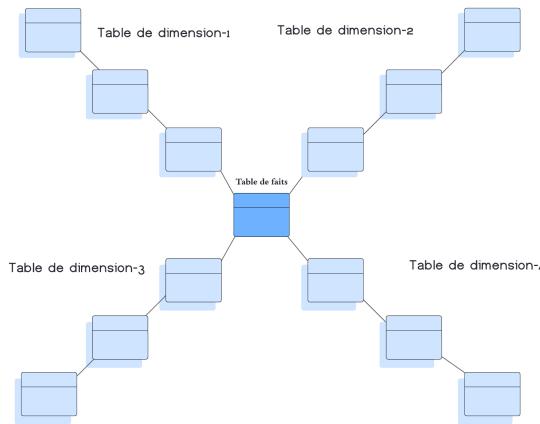


FIGURE 4.6 – Schéma en flocons de neige

*Modèle en constellation :

Le modèle de constellation est constitué de plusieurs tables de faits reliées à des tables de dimensions spécifiques. Un des principaux avantages de cette modélisation est l'absence de redondance pour les tables de dimensions partagées entre différentes tables de faits, ce qui permet de réduire l'espace de stockage nécessaire.

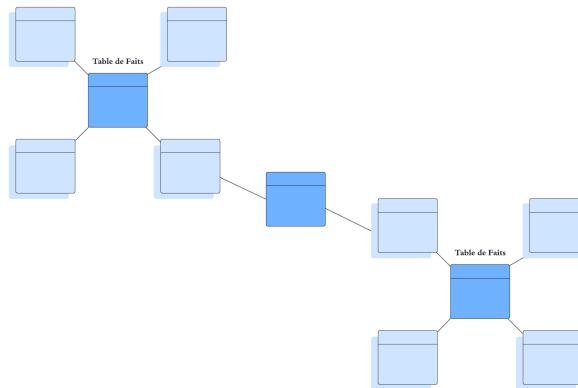


FIGURE 4.7 – Schéma en constellation

4.6.2 Choix du modèle dimensionnel

Suivant l'étude réalisée et en fonction de nos besoins, nous avons choisi le modèle en constellation comme modèle conceptuel de notre entrepôt de données pour les raisons suivantes :

- * **Flexibilité dans la modélisation** : Permet une représentation flexible des relations entre les données.
- * **Corrélation améliorée** : Facilite la corrélation entre différents sujets d'analyse.

* **Performance accrue :** Offre de meilleures performances, particulièrement avec des dimensions volumineuses.

* **Partage efficace des dimensions :** Permet le partage des dimensions entre plusieurs tables de faits, assurant la cohérence et réduisant les redondances.

4.6.3 Modèle physique de l'entrepôt de données

Nous passons maintenant au niveau physique de la modélisation, où nous rencontrons trois approches selon Ines (2021) :

- Approche relationnelle ROLAP : Cette approche implique le stockage de l'entrepôt de données dans un SGBD relationnel. Ensuite, le moteur OLAP émet des requêtes SQL vers l'entrepôt de données et fournit une vision multidimensionnelle permettant d'effectuer des calculs et des agrégations.
- Approche multidimensionnelle MOLAP : Cette approche repose sur un modèle multidimensionnel. En raison de la réduction du temps de réponse aux requêtes, l'accès aux données est très optimisé.
- Approche hybride HOLAP : Cette approche tente de combiner les avantages des approches ROLAP et MOLAP. Le système HOLAP stocke les données détaillées de l'entrepôt de données dans un SGBDR et les données agrégées dans un SGBD multidimensionnel.

Après une analyse approfondie des besoins, nous avons opté pour la solution ROLAP, qui répond parfaitement à toutes les exigences spécifiées par l'utilisateur.

La figure 4.8 représente la modélisation finale de l'entrepôt de données à construire.

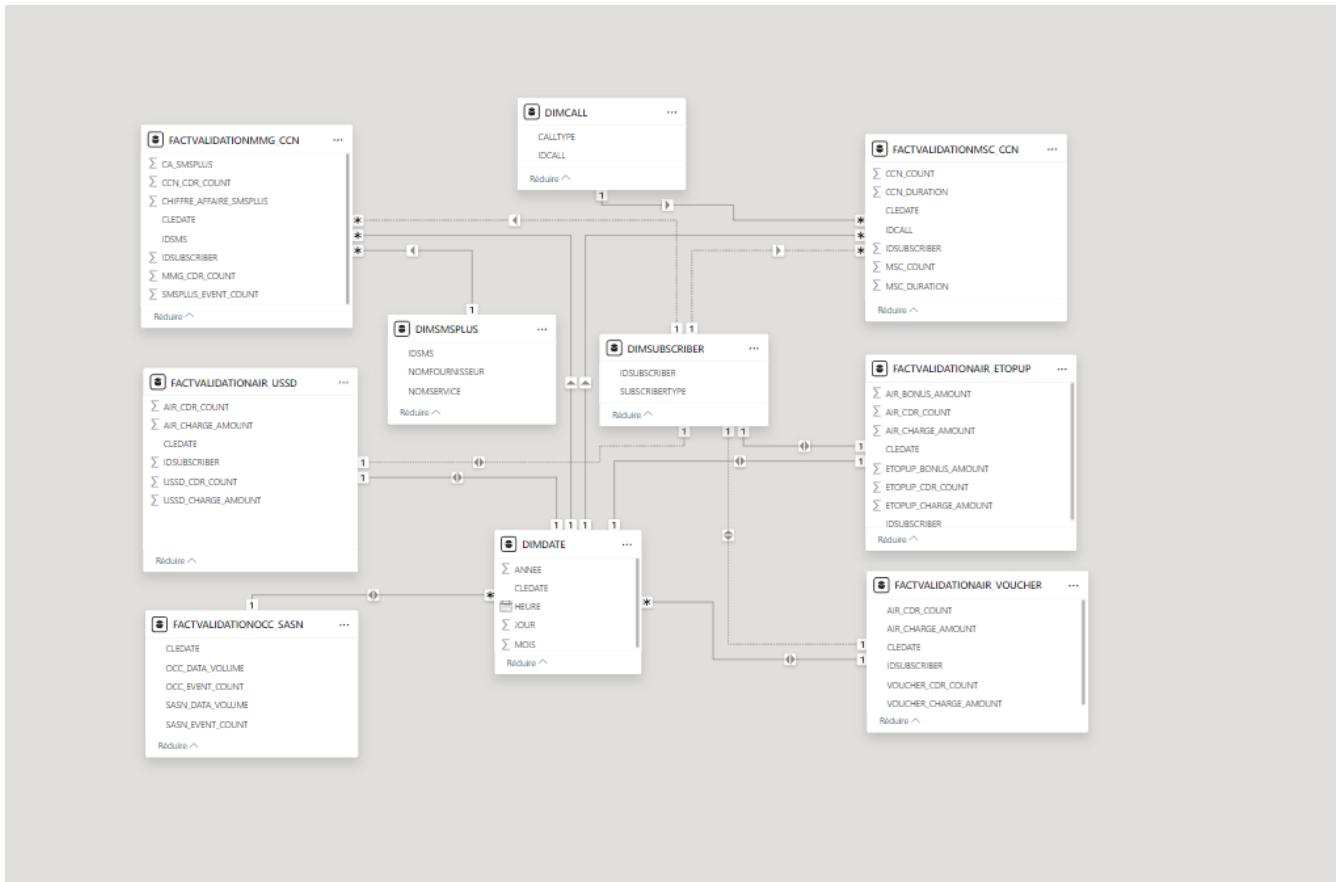


FIGURE 4.8 – Modèle physique de l'entrepôt de données

4.7 Conclusion

Dans ce chapitre nous avons précisé le schéma conceptuel des données source. Puis, nous avons présenté celui de l'entrepôt de données. Ensuite, nous avons introduit le schéma logique utilisé. Enfin, nous avons fixé le modèle physique de l'entrepôt de données.

Chapitre 5 : Sprint 2 - Gestion du processus ETL

Sommaire

5.1	Introduction	59
5.2	Sprint Backlog « Gérer le processus ETL »	59
5.3	Diagramme du cas d'utilisation du sprint 2 « Gérer le processus ETL »	60
5.3.1	Diagramme du cas d'utilisation	60
5.3.2	Description textuelle du cas d'utilisation « Gérer le processus ETL »	61
5.4	Staging Area (SA)	61
5.5	Gestion de processus ETL : du Flux source vers SA	63
5.5.1	Extraction des données	63
5.5.2	Nettoyage et transformations	63
5.5.3	Chargement des données	64
5.6	Gestion de processus ETL : du SA vers l'entrepot de données	64
5.6.1	Diagramme d'activités	64
5.6.2	Description de ce processus ETL : du SA vers l'ED	65
5.7	Automatisation du chargement de l'ED	71
5.8	Conclusion	73

5.1 Introduction

Dans ce chapitre, nous commençons par la conception de la zone de préparation de données (SA). Puis, nous montrons le déroulement du processus ETL de la source vers la SA et de la SA vers l'entrepôt de données. Enfin, nous précisons la méthode de rafraîchissement des données.

5.2 Sprint Backlog « Gérer le processus ETL »

le tableau 5.1 représente le backlog du sprint 2

ID	Fonctionnalité	ID	User Stories	ID	Description des tâches
2	Gérer le processus ETL	2.1	En tant que développeur, je peux créer le Staging Area (SA).	2.1.1	Créer les tables du « staging area »
		2.2	En tant que développeur, je peux gérer le processus ETL des données sources vers la SA.	2.2.1	Définition des sources de données
				2.2.2	Extraire les données des sources de données
				2.2.3	Transformer et nettoyer les données
				2.2.4	Charger les données dans le SA
		2.3	En tant que développeur, je peux gérer le processus ETL de la SA vers l'ED.	2.3.1	Extraire les données à partir du SA
				2.3.2	Transformer et nettoyer les données
				2.3.3	Charger les données dans l'ED
		2.4	En tant que développeur, je peux rafraîchir et charger l'entrepot de données l'entrepôt de données.	2.4.1	Créer un job pour charger les données.
				2.4.2	Planifier le job.

TABLE 5.1 – Backlog du sprint 2 « Gérer le processus ETL »

5.3 Diagramme du cas d'utilisation du sprint 2 « Gérer le processus ETL »

5.3.1 Diagramme du cas d'utilisation

La figure 5.1 représente le diagramme de cas d'utilisation du sprint 2

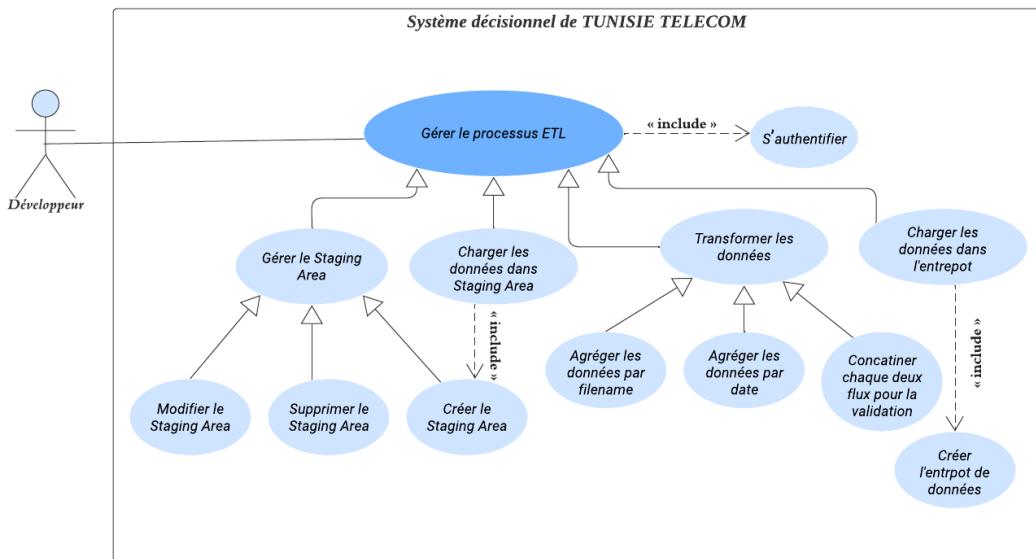


FIGURE 5.1 – Diagramme du cas d'utilisation du sprint 2 « Gérer le processus ETL »

5.3.2 Description textuelle du cas d'utilisation « Gérer le processus ETL »

Titre	Charger les données dans le Staging Area
Acteurs	Développeur
Objectif	Permettre de charger les données nettoyées dans le staging area.
Pré-condition	Existance d'un staging area créé.
Post-condition	Staging Area chargé.
Scénario nominal	<ol style="list-style-type: none">1. Le système vérifie la connexion à la base de données.2. Le développeur doit extraire les données à partir des fichiers sources3. Le développeur transforme et nettoie les données.4. Le système vérifie les transformations faites par le développeur.5. Le développeur doit charger les données nettoyées dans le staging area.
Scénario alternatif	<ol style="list-style-type: none">1. Échec de connexion à la base de données .<ol style="list-style-type: none">a) Le système affiche un message d'erreur.b) Le système reste à l'étape 2 du scénario nominal.4. a) Le développeur fait une erreur dans lors de la transformation des données.b) Le système affiche un message d'erreur.c) Le système reprend à l'étape 2 du scénario nominal.

TABLE 5.2 – Description textuelle du cas d'utilisation « Gérer le processus ETL »

5.4 Staging Area (SA)

La "Staging Area" est une étape importante dans le processus ETL (Extract, Transform, Load) qui consiste à préparer les données sources en vue de leur intégration et de leur exploitation dans le DW. Le Staging Area (SA) assume plusieurs rôles fondamentaux dans le processus ETL :

1. **Migration des données** : Le SA s'engage à collecter les informations issues de sources multiples, assurant ainsi l'intégrité et l'exhaustivité des données tout au long de ce processus.
2. **Zone de stockage temporaire** : Le SA agit comme une zone de stockage temporaire où les données sont stockées dans leur état brut à un moment donné. Ceci facilite la mise

en place d'un processus de reprise de données en cas de nécessité, garantissant ainsi la disponibilité des données même en cas d'incidents.

Nous répartirons les rôles de la manière suivante :

* Les flux entre les systèmes sources et le Staging Area (SA) consistent en des données transformées par des requêtes SQL, exécutées via SQL Developer. Le rôle du SA est crucial car il simplifie l'étape suivante.

* Les flux entre le Staging Area (SA) et l'entrepôt de données (ED) sont de véritables flux ETL. Dans ce processus, les données sont extraites, transformées et chargées dans l'ED pour une analyse ultérieure.

Le déroulement du flux décisionnel va donc se dérouler comme illustre la figure 5.2 :

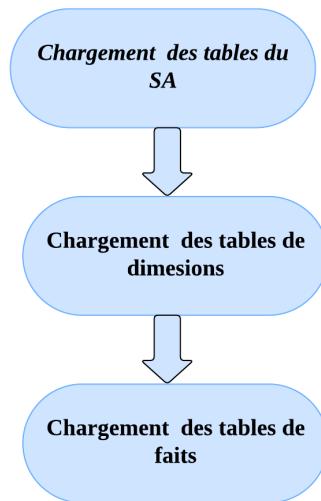


FIGURE 5.2 – Schéma du flux décisionnel avec utilisation du Staging Area

En décisionnel, il existe donc deux sortes de flux différents :

- les flux des données sources vers le SA.
- les flux de gestion et de mise à jour des dimensions et de chargement des tables de faits de l'ED.

5.5 Gestion de processus ETL : du Flux source vers SA

Tunisie Telecom nous a fourni des fichiers Excel et des fichiers plats contenant des enregistrements de CDR (Call Detail Records) pour chaque événement effectué par un abonné.

5.5.1 Extraction des données

Dans cette partie, nous allons examiner un processus visant à préparer et à intégrer des données provenant de différents fichiers sources . Pour réaliser cette tâche, nous optons pour l'exécution de requêtes SQL à travers l'application client SQL Developer en raison de sa souplesse et de sa capacité à gérer efficacement les transformations de données complexes qui seront archivées sous le système de gestion de base de données Oracle.

5.5.2 Nettoyage et transformations :

- **Conversion de données** : Cette fonctionnalité nous a permis de convertir les données d'un type à un autre, en les enregistrant dans des colonnes de sortie spécifiques.
- **Colonnes dérivées** : Grâce à cette fonction, nous avons créé de nouvelles valeurs de colonne en appliquant des expressions aux données d'entrée. Ces nouvelles valeurs ont été stockées dans des colonnes nouvellement créées ou dans des colonnes existantes.
- **Tri des données** : L'outil de tri nous a permis de réorganiser les données d'entrée dans un ordre croissant ou décroissant, tout en filtrant les doublons pour ne conserver que des valeurs uniques.
- **Fractionnement conditionnel** : Cette fonction nous a permis de diriger les lignes de données vers différentes sorties en fonction de conditions spécifiques, ce qui nous a permis de segmenter les données en fonction de leur contexte.

En plus de ces transformations, nous avons également ajouté des colonnes supplémentaires à la table et modifié les noms de certaines colonnes pour simplifier le processus d'extraction, de transformation et de chargement (ETL).

5.5.3 Chargement des données

Notre processus commence par l'extraction, la transformation et le chargement des tables relatives dans la base de données du SA.

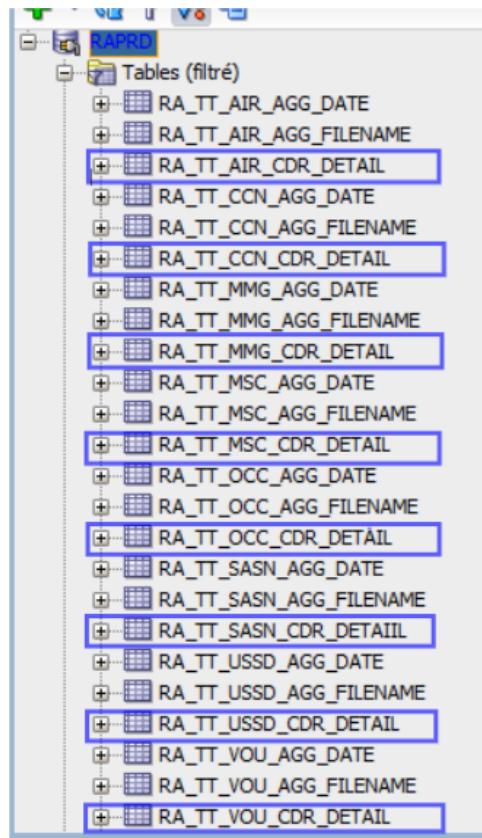


FIGURE 5.3 – La base de données du SA

5.6 Gestion de processus ETL : du SA vers l'entrepot de données

Lors de cette partie, nous allons étudier un flux permettant de nettoyer et transformer des données de SA vers la base de données de l'entrepôt. Les processus ETL représentent une étape cruciale de la gestion des données dans un projet décisionnel.

5.6.1 Diagramme d'activités

Dans notre démarche de développement du processus ETL, nous avons opté pour l'utilisation du diagramme d'activités en raison de sa capacité à modéliser efficacement le flux de travail. La figure 5.4 présente la phase d'intégration générale, qui implique la récupération des données

à partir du SA (Système d'Analyse), puis leur chargement dans la base de destination. Cette fonctionnalité sera assurée par les activités suivantes :

- Vérification de la connexion à la base de données du SA.
- Alimentation des dimensions.
- Alimentation des tables des faits.

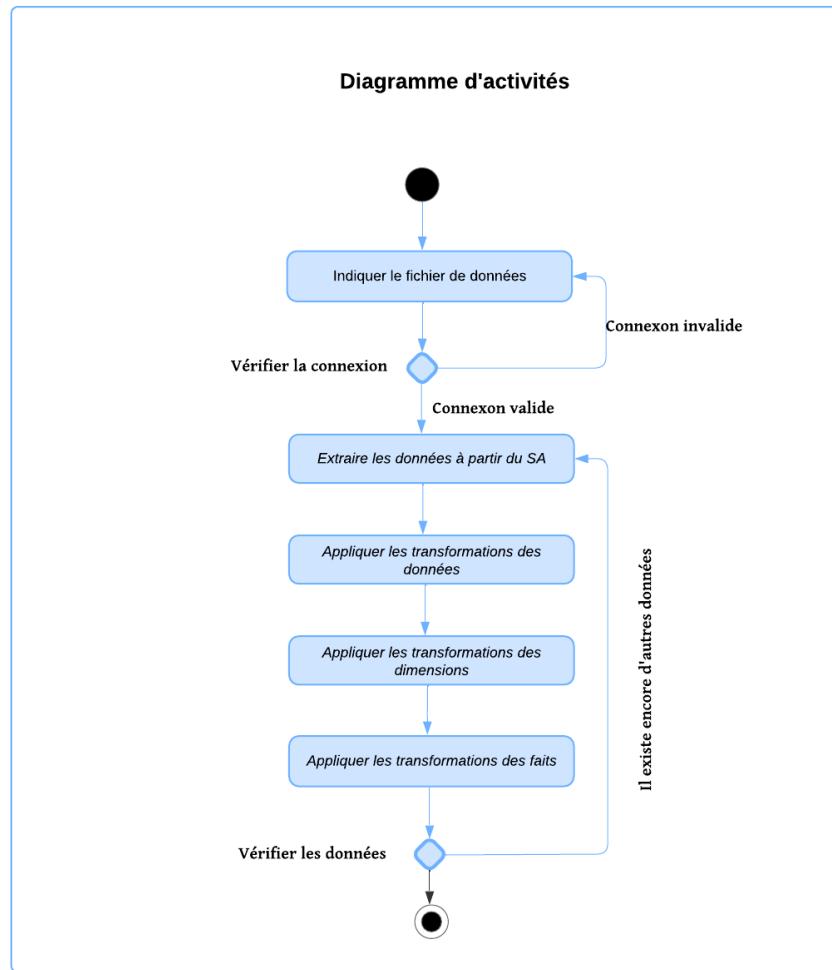


FIGURE 5.4 – Diagramme d'activités pour l'alimentation de l'ED

5.6.2 Description de ce processus ETL : du SA vers l'ED

5.6.2.1 Phase d'extraction des données (Extract)

Dans cette deuxième partie du processus ETL, nous allons utiliser Talend comme outil principal pour l'intégration des données dans l'entrepôt de données (ED). Talend Studio est une solution

open source d'intégration de données qui facilite le processus ETL grâce à une interface intuitive et graphique.

Le travail sous Talend est organisé à travers la création de "Jobs". Un Job est une représentation graphique et une implémentation technique de plusieurs composants connectés entre eux. Chaque Job comprend des tâches spécifiques, ou unités de traitement, qui permettent d'exécuter une quantité massive de données. La figure 5.5 montre la création d'un job sous Talend.

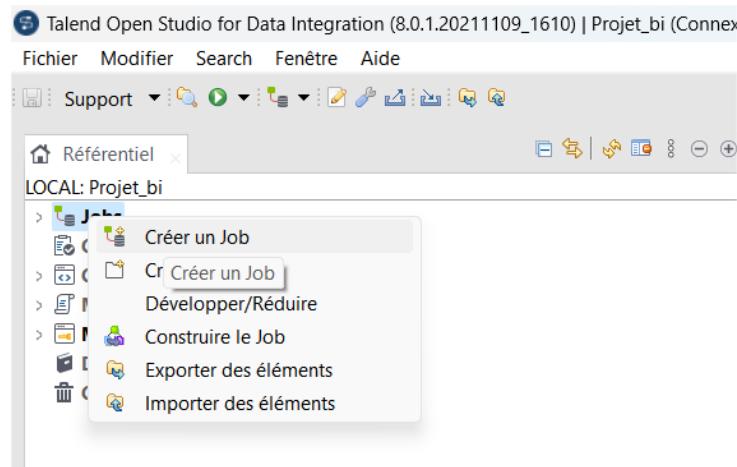


FIGURE 5.5 – Crédit d'un job

La figure ci-dessous 5.6 montre les paramètres de connexion entre Talend et la base de données Oracle.

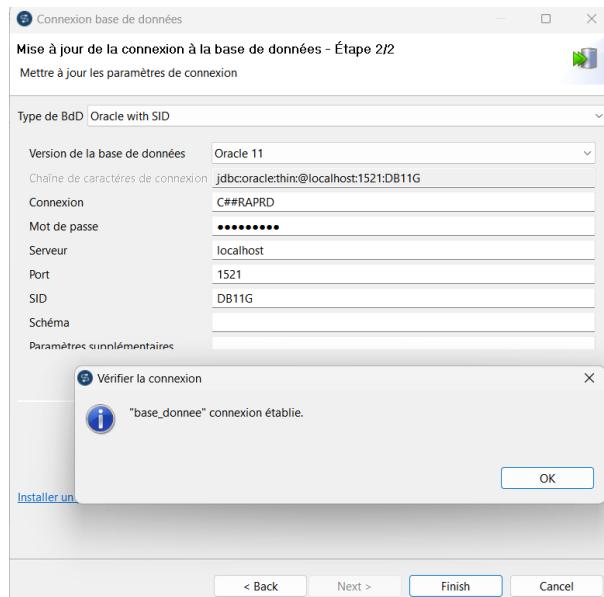


FIGURE 5.6 – Connexion talend et Oracle

* **Composants et connecteurs Talend :**

- **TDBInput** : Ce composant lit une table dans la base de données ligne par ligne et envoie les champs au composant suivant (voir figure 5.7).



FIGURE 5.7 – Le composant TDBInput

Ce composant doit être configuré selon les paramètres indiqués dans la figure 5.8. La liste des champs sera transmise au composant suivant via une connexion Main Row.

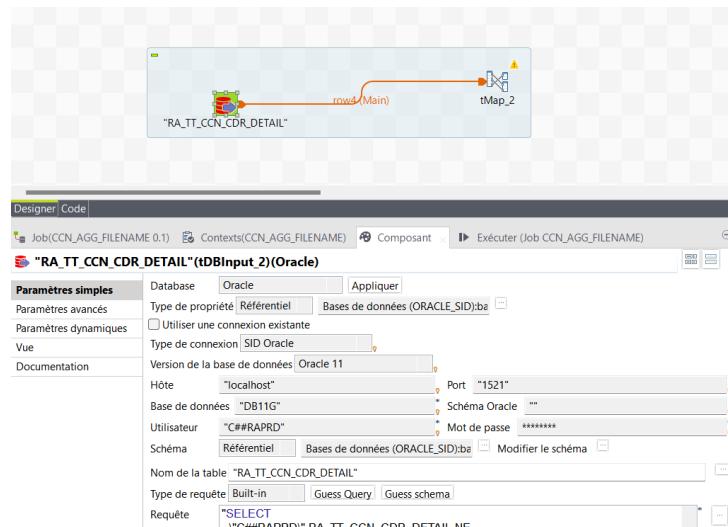


FIGURE 5.8 – Configuration de ces composants

5.6.2.2 Phase de transformation des données (Transform)

Dans cette partie, nous allons expliquer les transformations appliquées sur les tables.

* **Agrégation par "filename"**

Chaque flux doit être agrégé par "filename" en priorité pour vérifier la réception correcte des fichiers et leur contenu, notamment pour déterminer le nombre de CDR par "filename". Après avoir filtré les lignes en fonction de conditions spécifiques, nous avons agrégé les données

pour regrouper les tables et effectuer les opérations nécessaires. Enfin, nous avons procédé au mappage des résultats agrégés et au chargement des données dans la sortie "agg_filename".

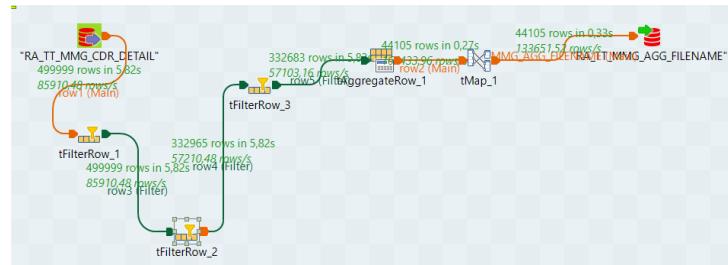


FIGURE 5.9 – Job aggregation par filename

* Agrégation par date

Chaque flux, agrégé par "filename", doit ensuite être agrégé par date pour surveiller le trafic et effectuer une validation avec le système de taxation basée sur la date. Après avoir agrégé les données pour regrouper les tables et effectué les opérations nécessaires, nous avons procédé au mappage des résultats agrégés et au chargement des données dans la sortie "agg_date".

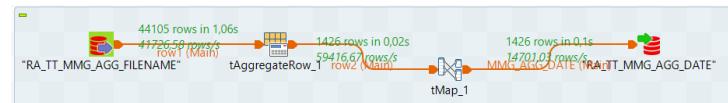


FIGURE 5.10 – Job aggregation par date

* Concatination de chaque deux flux

Enfin, pour valider la conformité entre les deux flux, nous avons procédé à la concaténation de chaque paire de flux. Cela consiste à unir les données de deux tables distinctes que nous comparerons ensuite avec la table de sortie.

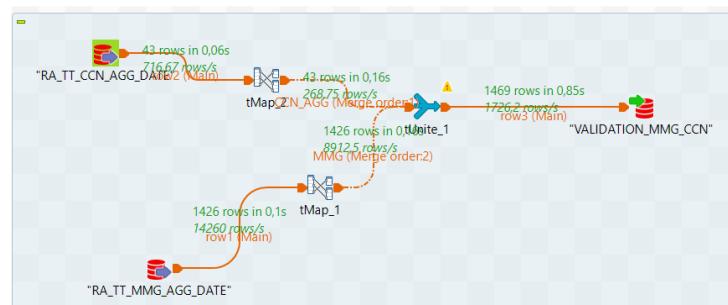


FIGURE 5.11 – job de concatenation

* **Les composants Talend utilisés sont les suivants :** Lors de cette transformation on a utilisé les composants suivants :

tFilterRow : est utilisé pour filtrer des lignes de données en fonction de certaines conditions.

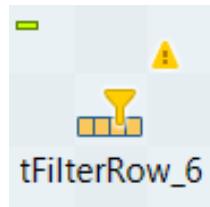


FIGURE 5.12 – Le composant tFilterRow

tAggregateRow : Ce composant se manifeste par la réception d'un flux de données et exerce une agrégation qui se base sur une ou plusieurs colonnes. En fait, il sert à établir des statistiques fondées sur des calculs et des valeurs.

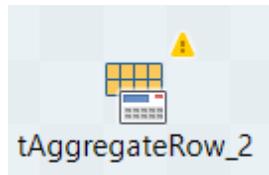


FIGURE 5.13 – Le composant tAggregate

tMap : Un composant très important dans Talend, permet d'effectuer des opérations de jointures, de filtrages, suppression, ect...



FIGURE 5.14 – Le composant tMap

5.6.2.3 Phase de chargement des données (Load)

* Crédit de l'entrepôt de données

Pour commencer, il est crucial de mettre en place un job dans lequel nous pouvons alimenter notre entrepôt de données avec les faits et les dimensions. Dans ce job, nous devons nous assurer que toutes les dimensions nécessaires sont correctement liées aux faits, créant ainsi une structure solide pour l'analyse ultérieure.

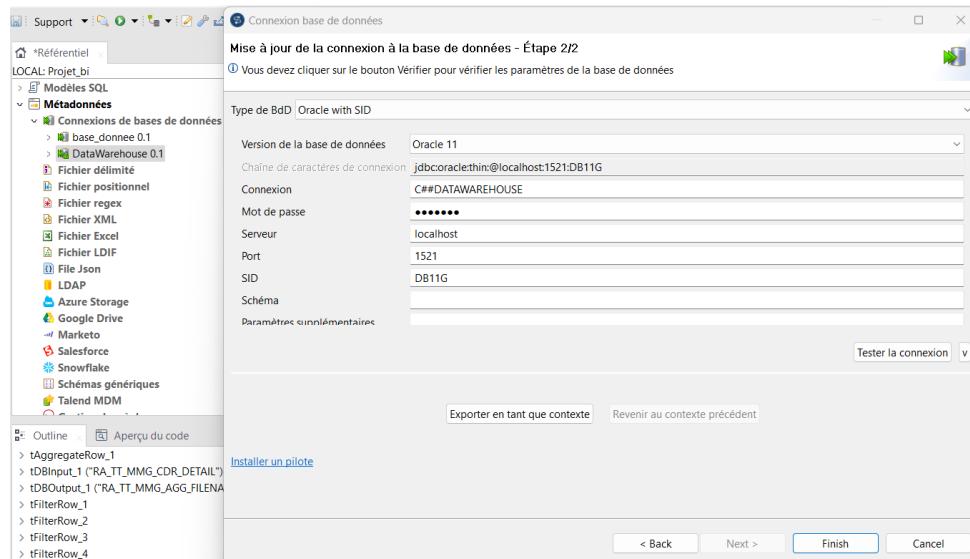


FIGURE 5.15 – Creation de l’entrepôt de données

* Alimentation de l’entrepôt de données

Une fois l’entrepôt de données établi, nous avons entamée le processus d’alimentation des dimensions et des faits à partir des données agrégées disponibles. Pendant cette étape, nous avons appliqué un traitement spécifique à la dimension temporelle, consistant à décomposer la date en ses composants de jour, mois et année. Cette action a pour avantage d’offrir une granularité plus fine lors de l’analyse des données temporelles, ce qui facilite la segmentation et l’agrégation des données en fonction de ces unités temporelles.

date	Column
Expression	CLEDATE
TalendDate.formatDate("yyyy", row1.START_DATE) + "/" + ... Tal...	HEURE
row1.START_HOUR	JOUR
new BigDecimal(Integer.parseInt(TalendDate.formatDate("dd", r...)	MOIS
new BigDecimal(Integer.parseInt(TalendDate.formatDate("MM", ...	ANNEE
new BigDecimal(Integer.parseInt(TalendDate.formatDate("yyyy"...	

FIGURE 5.16 – Décomposition de la date

Par la suite, nous avons chargé les tables des dimensions et les tables de fait dans un job nommé “Chargementfaidimension” comme l’indique la figure 5.17

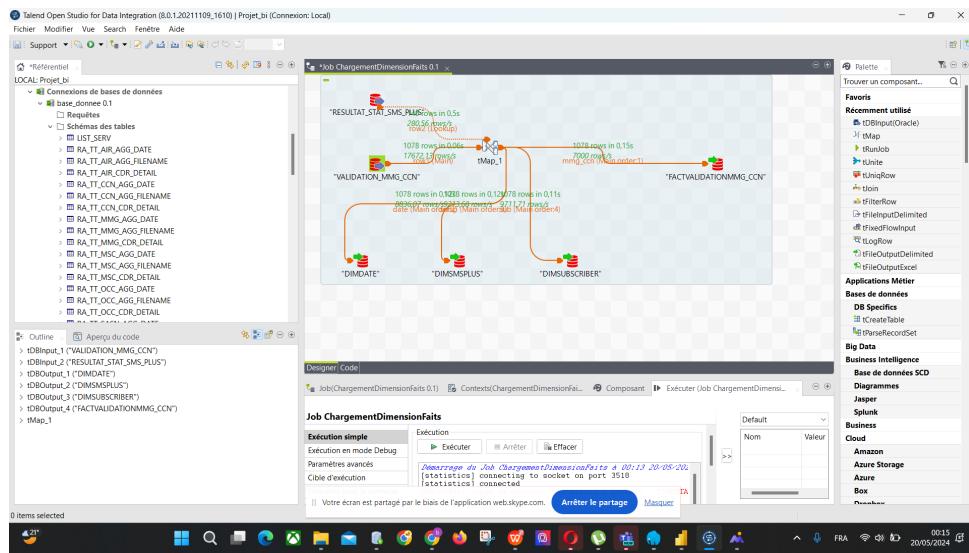


FIGURE 5.17 – Chargement des tables de fait et des tables des dimensions

5.7 Automatisation du chargement de l’ED

L’automatisation des mises à jour dans l’entrepôt de données peut être effectuée grâce à divers outils et techniques. Pour ce faire, nous avons opté pour un "**planificateur de tâches**".

Nous avons implémenté un job Talend pour automatiser les tâches de rafraîchissement ETL. L’utilisation d’un planificateur de tâches nous permet de programmer les différents processus ETL et les mises à jour dans l’entrepôt de données. Ces outils nous permettent de définir des horaires réguliers pour l’exécution des tâches, assurant ainsi une gestion efficace et continue des flux de données. Cela garantit que les données sont toujours à jour et prêtes pour l’analyse.

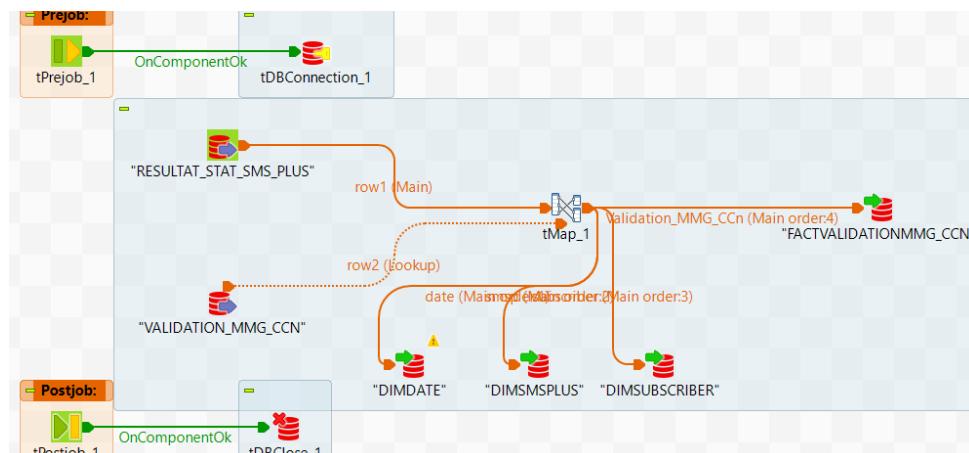


FIGURE 5.18 – Job d’automatisation de chargement de l’entrepot

Sous Windows, nous avons utilisé le "Planificateur de tâches" pour créer une nouvelle tâche automatisée dédiée au chargement de l'entrepôt de données (ED).

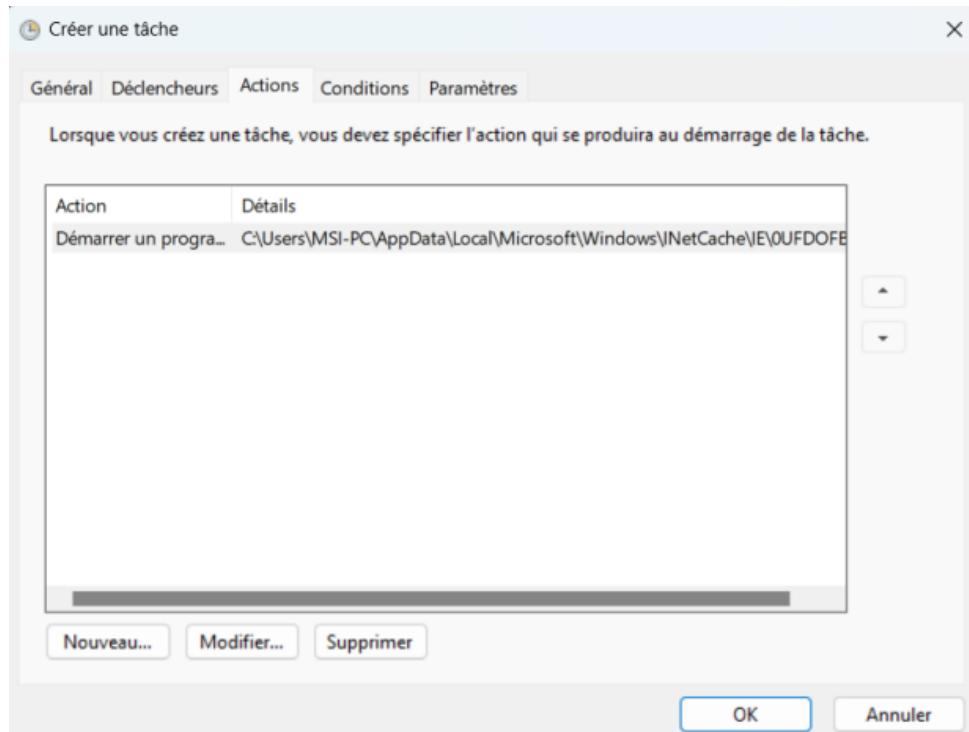


FIGURE 5.19 – Création de tâche d'automatisation du chargement de l'ED

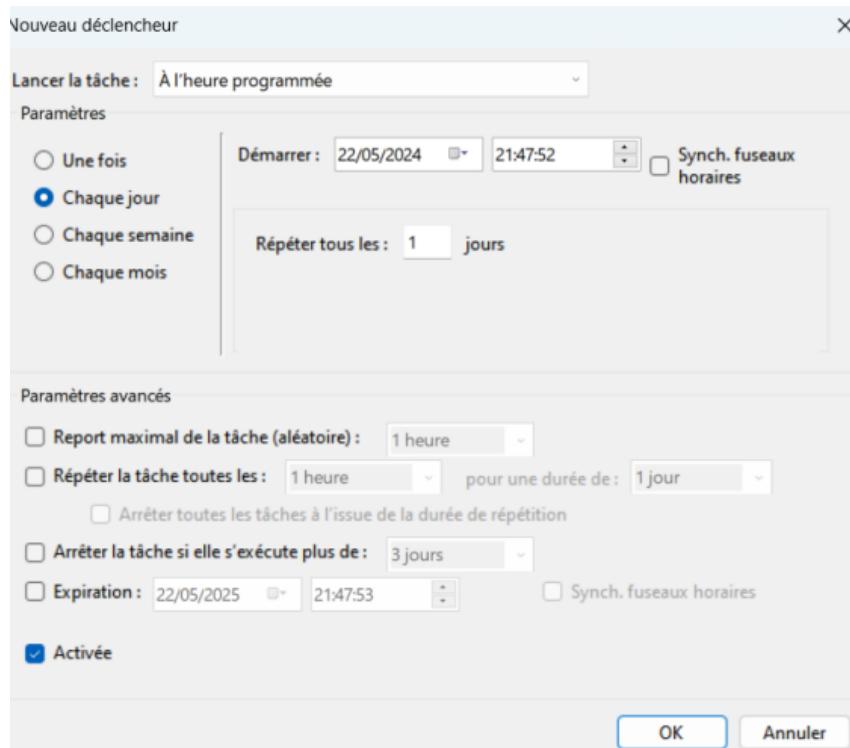


FIGURE 5.20 – Planification du job du chargement de l'entrepot

5.8 Conclusion

Dans ce chapitre, nous avons d'abord présenté une conception détaillée de la zone de préparation des données (Staging Area). Ensuite, nous avons expliqué en détail le déroulement du processus ETL (Extraction, Transformation, Chargement). Pour finir, nous avons décrit la méthode de planification efficace de ce processus ETL.

Chapitre 6 : Sprint 3 : Visualisation de données

Sommaire

6.1	Introduction	75
6.2	Sprint Backlog	75
6.3	Diagramme du cas d'utilisation du sprint 3 « Crée tableau de bord »	76
6.3.1	Diagramme du cas d'utilisation	76
6.3.2	Description textuelle du cas d'utilisation « Crée tableau de bord »	77
6.4	Création des tableaux de bord	77
6.4.1	Etablir la connexion entre Power BI Desktop et Oracle	78
6.4.2	Mesures spécifiques (DAX)	78
6.4.3	Choix des graphiques	80
6.4.4	Présentation des interfaces de réalisation	81
6.5	Conclusion	88

6.1 Introduction

D'une façon similaire aux sprints précédents, ce chapitre est consacré au troisième sprint, qui se focalise sur la restitution des données de l'entrepôt et la présentation des tableaux de bord. Cela permettra aux décideurs de prendre des décisions plus efficaces et performantes en termes de temps et de gestion des données.

6.2 Sprint Backlog

Le tableau 6.1 met en évidence le backlog du sprint 3, illustrant l'étape de gestion du tableau de bord par les deux acteurs : développeur et analyste.

ID	Fonctionnalité	ID	User Stories	ID	Description des tâches
3	Gérer tableau de bord	3.1	En tant que développeur, je peux créer et modifier des tableaux de bord à partir de l'entrepôt de données.	3.1.1	Construire les graphiques.
			En tant qu'analyste business, je peux consulter des tableaux de bord de l'entrepôt de données.	3.1.2	Visualiser les graphiques.

TABLE 6.1 – Backlog du sprint 3

6.3 Diagramme du cas d'utilisation du sprint 3 « Crée tableau de bord »

6.3.1 Diagramme du cas d'utilisation

La figure 6.1 représente le diagramme de cas d'utilisation du sprint 3

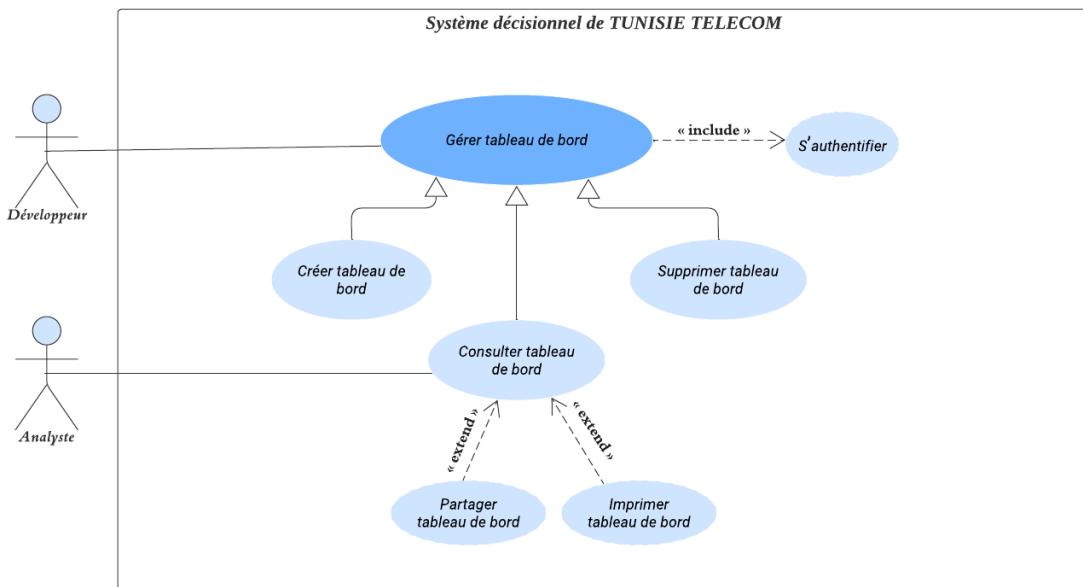


FIGURE 6.1 – Diagramme du cas d'utilisation de sprint 3 « Crée tableau de bord »

6.3.2 Description textuelle du cas d'utilisation « Créer tableau de bord »

Le tableau 6.2 montre la description textuelle du cas d'utilisation « Créer tableau de bord »

Titre	Créer tableau de bord.
Acteurs	Développeur.
Objectif	Permettre à l'utilisateur de créer des tableaux de bord.
Pré-condition	Le développeur doit être connecté au logiciel.
Post-condition	Tableau de bord créé.
Scénario nominal	<ol style="list-style-type: none">1. Le développeur s'authentifie à la base de données.2. L'application affiche une interface permettant la création des tableaux de bord.3. Le développeur choisit les dimensions et les faits.4. Le développeur choisit les modèles disponibles dans l'application pour construire des graphiques afin de mieux visualiser les données.5. Le développeur enregistre le tableau de bord créé.
Scénario alternatif	<p>5.a) Le développeur fait une erreur dans la manipulation des données.</p> <ol style="list-style-type: none">a) 1. Le système affiche un message d'erreur.a) 2. Le système reprend à l'étape 2 du scénario nominal.

TABLE 6.2 – Description textuelle du cas d'utilisation « Créer tableau de bord »

6.4 Crédit des tableaux de bord

Une fois l'ED est créé et chargé, nous passons à la phase de visualisation de données. Cette phase permet aux utilisateurs finaux d'exploiter les données afin de savoir l'état de l'entreprise et de prendre les décisions stratégiques.

6.4.1 Etablir la connexion entre Power BI Desktop et Oracle

Au niveau de Power BI, on a besoin d'importer les dimensions et les tables de faits pour effectuer la visualisation. Mais avant tout, on doit connecter Power BI et Oracle pour charger les données. Les figures suivantes 6.2 , 6.3 montrent respectivement cette connexion.

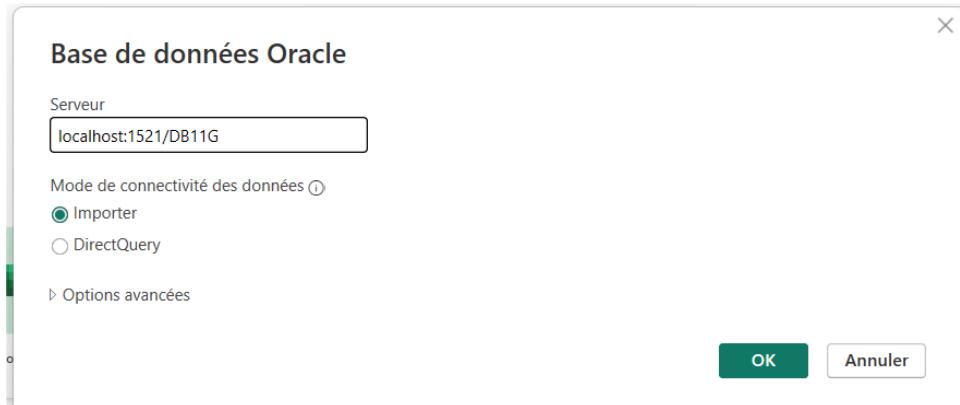


FIGURE 6.2 – Connexion du Power BI avec Oracle

MSC_COUNT	CCN_COUNT	MSC_DURATION	CCN_DURATION	C#
3385	3356	21400	23021	
127638	127391	793126	913933	
238019	288050	22070942	30472239	
376913	448432	29901528	40969802	
37988	37985	0	0	
1342	1342	0	0	
12373	12372	0	0	
2451	2416	16007	16811	
5225	5976	759498	607475	
82363	95823	5440523	6883779	
22548	20955	1158150	999718	
259005	258202	1628331	1884727	
11499	11499	0	0	
8594	8591	0	0	
17941	17860	102121	118897	
597681	540574	54213398	49035571	
11104	11099	0	0	
4472	4472	0	0	
31918	31454	203856	215658	
2895	2856	18876	19816	

FIGURE 6.3 – Chargement de l'entrepôt dans PowerBI

6.4.2 Mesures spécifiques (DAX)

DAX est un langage puissant utilisé pour créer des mesures spécifiques dans le cadre de l'analyse de données. Ces mesures sont conçues pour calculer et exprimer des indicateurs personnalisés afin de répondre à des problématiques spécifiques en exploitant les données disponibles dans l'entrepôt. Les formules écrites en DAX permettent de manipuler les données de manière flexible et de créer des indicateurs précis et pertinents pour l'analyse [12].

On peut utiliser DAX pour calculer des indicateurs de performance clés (KPI) adaptés à des besoins spécifiques. Par exemple, la différence entre le système de taxation et les services peut être illustrée par l'utilisation de DAX. Le système de taxation fournit des données brutes sur les transactions financières, tandis que les services offrent des détails sur l'utilisation des clients. En utilisant DAX, on peut créer des mesures pour analyser la différence entre eux. Des exemples de formules écrites avec DAX peuvent être décrits par les figures 6.4, 6.5 et 6.6.

```
Différence_Event_count =
DIVIDE(
    ABS(
        CALCULATE(SUM('FACTVALIDATIONAIR_ETOPUP'[AIR_OUT_EVENT_COUNT])) - SUM(
            'FACTVALIDATIONAIR_ETOPUP'[ETOPUP_OUT_EVENT_COUNT])
    ),
    MAX(SUM('FACTVALIDATIONAIR_ETOPUP'[AIR_OUT_EVENT_COUNT]), SUM('FACTVALIDATIONAIR_ETOPUP'[ETOPUP_OUT_EVENT_COUNT]))
)
```

FIGURE 6.4 – Creation d'une mesure de la difference entre les Event_Count de deux flux

```
Différence_Charge_amount =
DIVIDE(
    ABS(
        CALCULATE(SUM(FACTVALIDATIONAIR_ETOPUP[AIR_OUT_CHARGE_AMOUNT])) - CALCULATE(SUM(
            FACTVALIDATIONAIR_ETOPUP[ETOPUP_OUT_CHARGE_AMOUNT]))
    ),
    MAX(SUM(FACTVALIDATIONAIR_ETOPUP[AIR_OUT_CHARGE_AMOUNT]), SUM(FACTVALIDATIONAIR_ETOPUP[ETOPUP_OUT_CHARGE_AMOUNT]))
)
```

FIGURE 6.5 – Creation d'une mesure de la difference entre les Charge_Amount de deux flux

```
Différence_Bonus_Amount =
DIVIDE(
    ABS(
        CALCULATE(SUM(FACTVALIDATIONAIR_ETOPUP[AIR_OUT_BONUS_AMOUNT])) - CALCULATE(SUM(
            FACTVALIDATIONAIR_ETOPUP[ETOPUP_OUT_BONUS_AMOUNT]))
    ),
    MAX(SUM(FACTVALIDATIONAIR_ETOPUP[AIR_OUT_BONUS_AMOUNT]), SUM(FACTVALIDATIONAIR_ETOPUP[ETOPUP_OUT_BONUS_AMOUNT]))
)
```

FIGURE 6.6 – Creation d'une mesure de la difference entre les Bonus_Amount de deux flux

Analyse des résultats des différences :

Dans des circonstances normales, les pourcentages devraient être nuls, indiquant une conformité totale entre l'utilisation par le client et les enregistrements dans le système de taxation. Toute déviation de cette norme constitue un signal d'alarme important.

Un écart de 10% dépasse largement la marge d'erreur acceptable et suggère une fraude potentielle. Cette différence entre les montants taxés et utilisés peut être due à des fichiers manquants. Si

le montant taxé est supérieur à celui utilisé, cela signale un problème de qualité de fichier. À l'inverse, si le montant utilisé est plus élevé, cela représente une perte de revenus pour Tunisie Telecom.

Il est crucial d'alerter immédiatement l'équipe concernée de ces divergences pour résoudre le problème. Une intervention rapide et efficace est nécessaire pour garantir l'intégrité des données fiscales et éviter toute perte financière pour l'entreprise.

6.4.3 Choix des graphiques

Power BI propose une multitude de graphiques ainsi que différentes zones de données. Il est donc essentiel de choisir les graphiques appropriés pour obtenir une visualisation claire. Dans notre visualisation, nous avons sélectionné plusieurs types de graphiques, notamment :

L'histogramme groupé : Ce graphique permet de comparer visuellement la répartition des données entre différents ensembles en fonction du temps. Offrant une représentation graphique claire , il permet ainsi une analyse comparative rapide et précise.

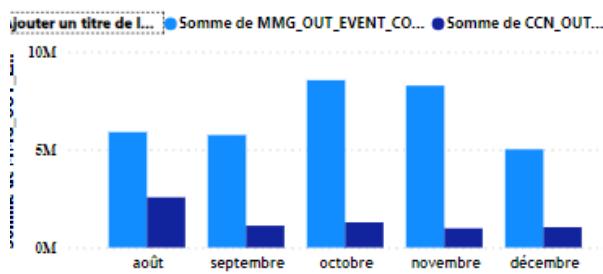


FIGURE 6.7 – Histogramme groupé

Graphique en courbe : Ce graphique est utilisé pour observer la tendance des données sur une échelle continue, telle que celle du temps, permettant ainsi de visualiser les variations et les tendances au fil du temps de manière fluide et continue, figure6.8



FIGURE 6.8 – Graphique en courbe

Graphique en secteurs : Il est choisi pour représenter la répartition des données catégoriques, fournissant une vue visuelle des proportions relatives de différentes catégories dans un ensemble de données, figure6.9

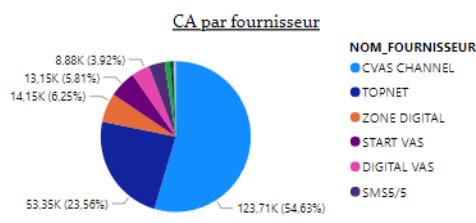


FIGURE 6.9 – Graphique en secteurs

Graphique en segment : Comme illustré dans la figure 6.10, ce type de graphique est utilisé pour générer des filtres dynamiques dans les tableaux de bord, figure 6.10



FIGURE 6.10 – Graphique en segment

6.4.4 Présentation des interfaces de réalisation

Cette étape, également appelée Reporting, se charge de présenter les informations à valeur ajoutée de telle sorte qu’elles apparaissent de la façon la plus lisible possible dans le cadre de l’aide à la décision. Les données sont principalement modélisées par des représentations à base de requêtes afin de construire des tableaux de bord ou des rapports via des outils d’analyse décisionnelle.

6.4.4.1 Page d'accueil

La page d'accueil peut être utile pour donner à l'analyste une vue d'ensemble rapide de ce qui est disponible dans notre projet décisionnel. Elle peut servir également de point d'entrée central, permettant à l'analyste de naviguer facilement vers les tableaux de bord spécifiques qui les intéressent, comme indiqué dans la figure 6.11.



FIGURE 6.11 – Première page du tableau de bord : Page d'accueil

6.4.4.2 Rapport : Validation de système de taxation AIR / service de recharge électronique ETOPUP

Dans cette section, nous analysons et validons les flux de type recharge entre le système de taxation Air et le service de recharge électronique ETOPUP. Les éléments suivants sont visualisés :

- L'évolution de l'AIR_Bonus_Amount et de l'ETOPUP_Bonus_Amount au fil du temps.
- L'évolution de l'AIR_Charge_Amount et de l'ETOPUP_Charge_Amount au fil du temps.
- L'évolution de l'AIR_Event_Count et de l'ETOPUP_Event_Count au fil du temps.
- La somme de l'ETOPUP_Event_Count par type d'abonné (Subscriber_type).
- La différence entre les Event_Count d'Air et d'ETOPUP.
- La différence entre les Charge_Amount d'AIR et d'ETOPUP.

- La différence entre les Bonus_Amount d'AIR et d'ETOPUP.

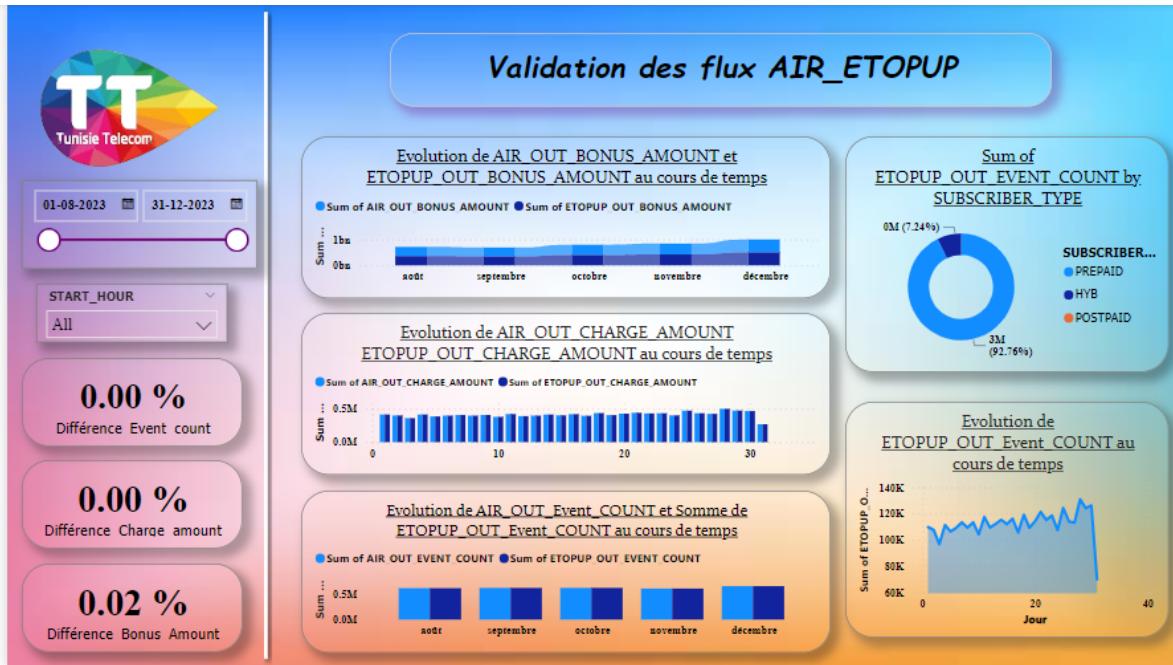


FIGURE 6.12 – Validation des flux AIR_ETOPUP

6.4.4.3 Rapport : Validation de système de taxation AIR / service de recharge SOS solde

USSD

Dans cette section, nous nous concentrons sur la validation des flux de taxation AIR et le flux d'utilisation de SOS solde fournis par l'USSD. Les visualisations incluent :

- La différence entre les Event_Count d'AIR et d'USSD.
- La différence entre les Charge_Amount d'AIR et d'USSD.
- L'évolution de l'AIR_Event_Count et de l'USSD_Event_Count au fil du temps.
- L'évolution de l'AIR_Charge_Amount et de l'USSD_Charge_Amount au fil du temps.

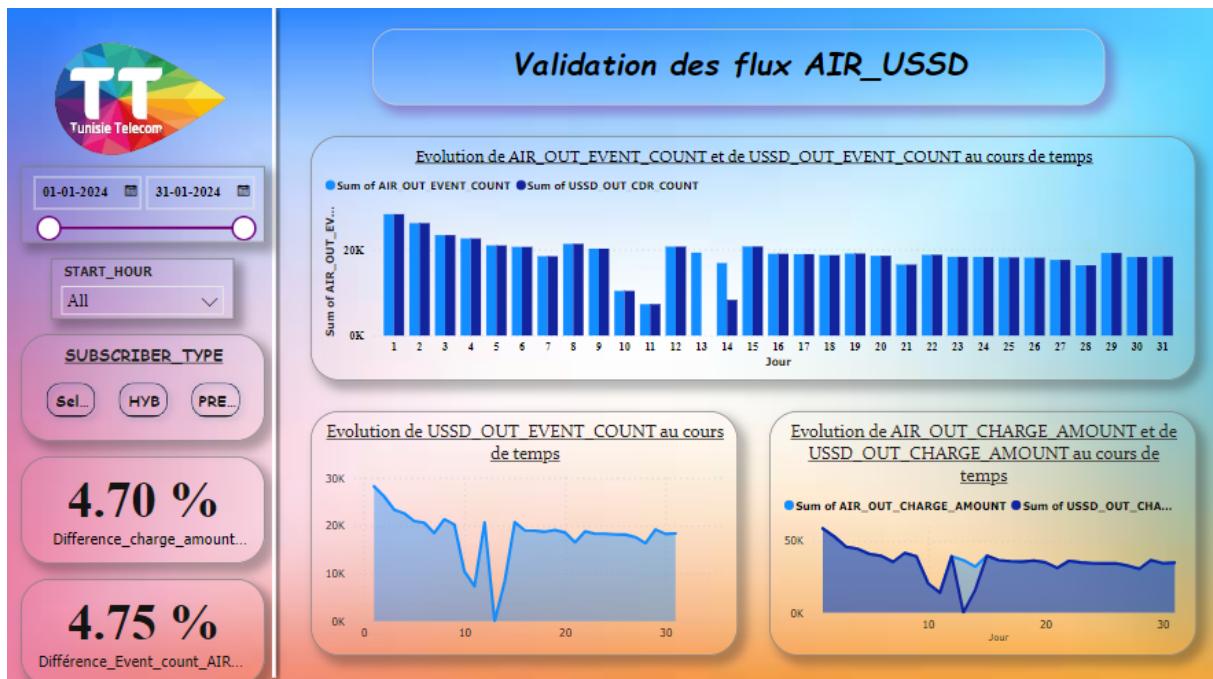


FIGURE 6.13 – Validation des flux AIR_USSD

6.4.4.4 Rapport : Validation de système de taxation AIR / service de recharge par carte Voucher

Cette section porte sur la validation des flux de taxation Air et le flux de recharge par carte Voucher. Les visualisations sont :

- La différence entre les Event_Count d'AIR et de VOUCHER.
- La différence entre les Charge_Amount d'AIR et de VOUCHER.

Pour la comparaison :

- L'évolution de l'AIR_Event_Count et du VOUCHER_Event_Count au fil du temps.
- L'évolution de l'AIR_Charge_Amount et du VOUCHER_Charge_Amount au fil du temps.

Pour le suivi de la tendance du trafic :

- L'évolution de l'AIR_Event_Count au fil du temps.
- L'évolution du VOUCHER_Event_Count au fil du temps.

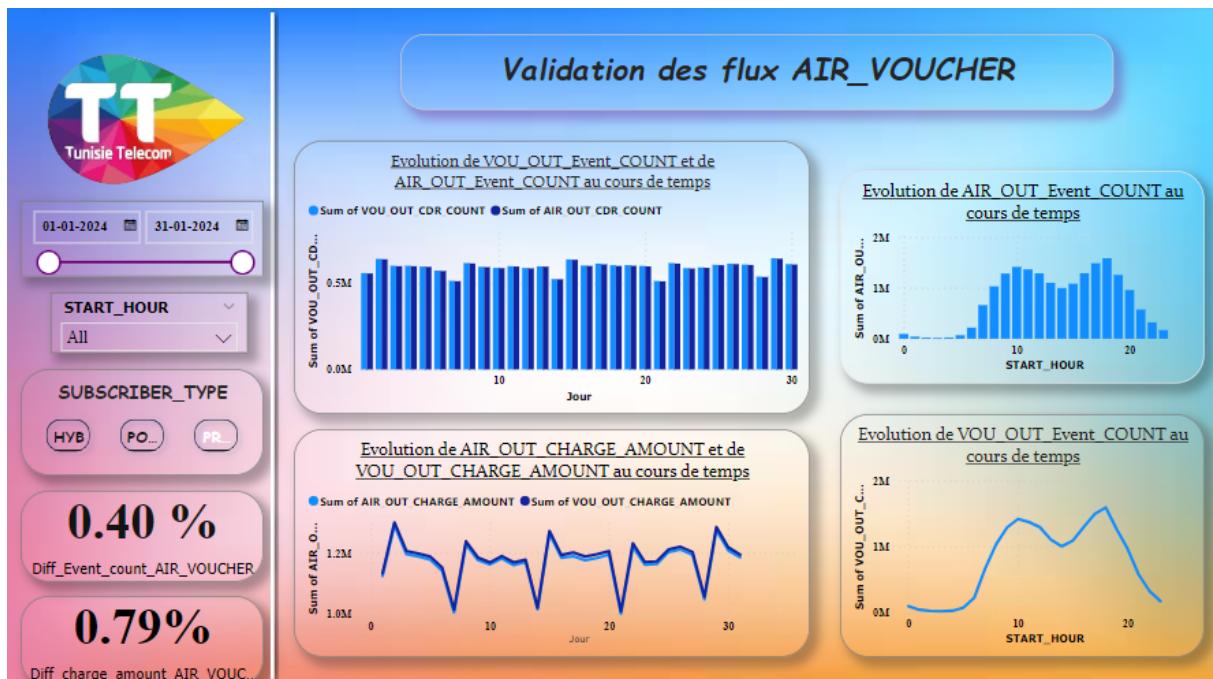


FIGURE 6.14 – Validation des flux AIR_Voucher

6.4.4.5 Rapport : Validation de système de taxation OCC / Service de transfert de données SASN

Nous analysons les flux entre le système de taxation OCC et le flux de transfert de données SASN dans cette section. Les visualisations sont :

- La différence de Data_Volume entre OCC et SASN.
- La différence d'Event_Count entre OCC et SASN.

*Pour la comparaison :

- L'évolution de l'OCC_Event_Count et du SASN_Event_Count au fil du temps.
- L'évolution de l'OCC_Charge_Amount et du SASN_Charge_Amount au fil du temps.

*Pour le suivi de la tendance du trafic :

- L'évolution de l'OCC_Event_Count au fil du temps.
- L'évolution du SASN_Event_Count au fil du temps.

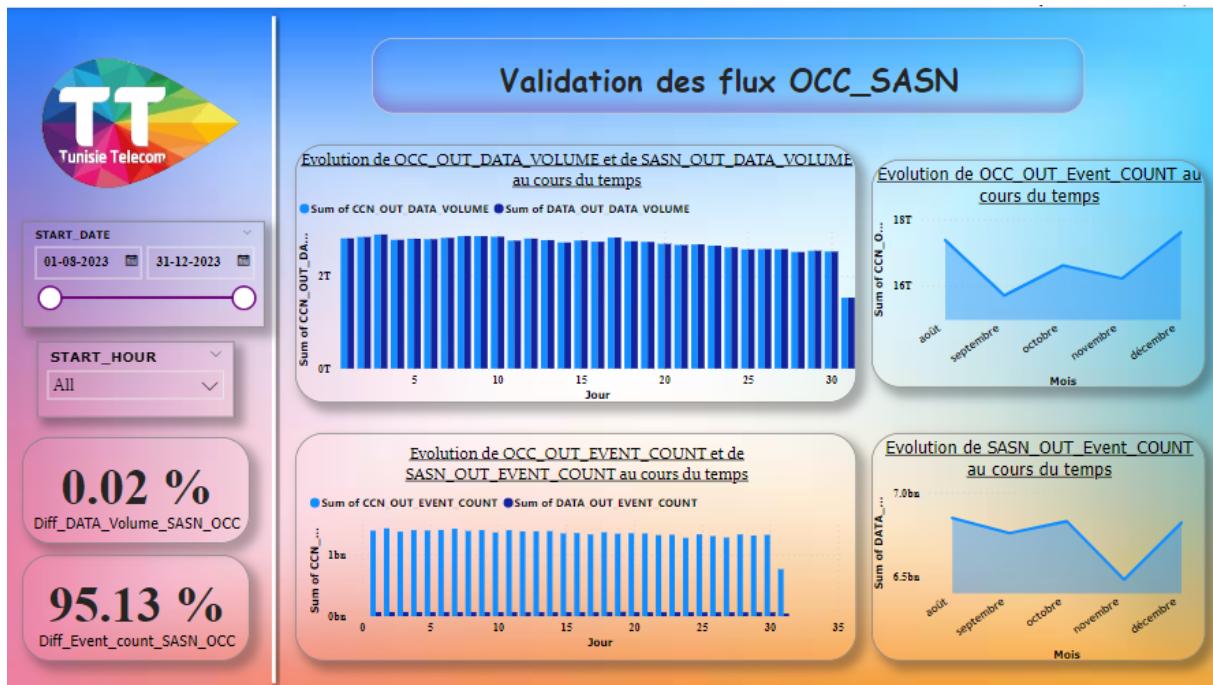


FIGURE 6.15 – Validation des flux OCC_SASN

6.4.4.6 Rapport : Validation des flux de système de taxation CCN / service de SMS+ MMG

Dans cette section, nous nous intéressons aux flux entre MMG et CCN. Les visualisations sont :

— La différence d'Event_Count entre MMG et CCN.

* Pour la comparaison :

— L'évolution du MMG_Event_Count et du CCN_Event_Count au fil du temps.

*Pour le suivi de la tendance du trafic :

— L'évolution du MMG_Event_Count au fil du temps.

— La répartition du chiffre d'affaires par nom_fournisseur.

— Le chiffre d'affaires par nom_service.

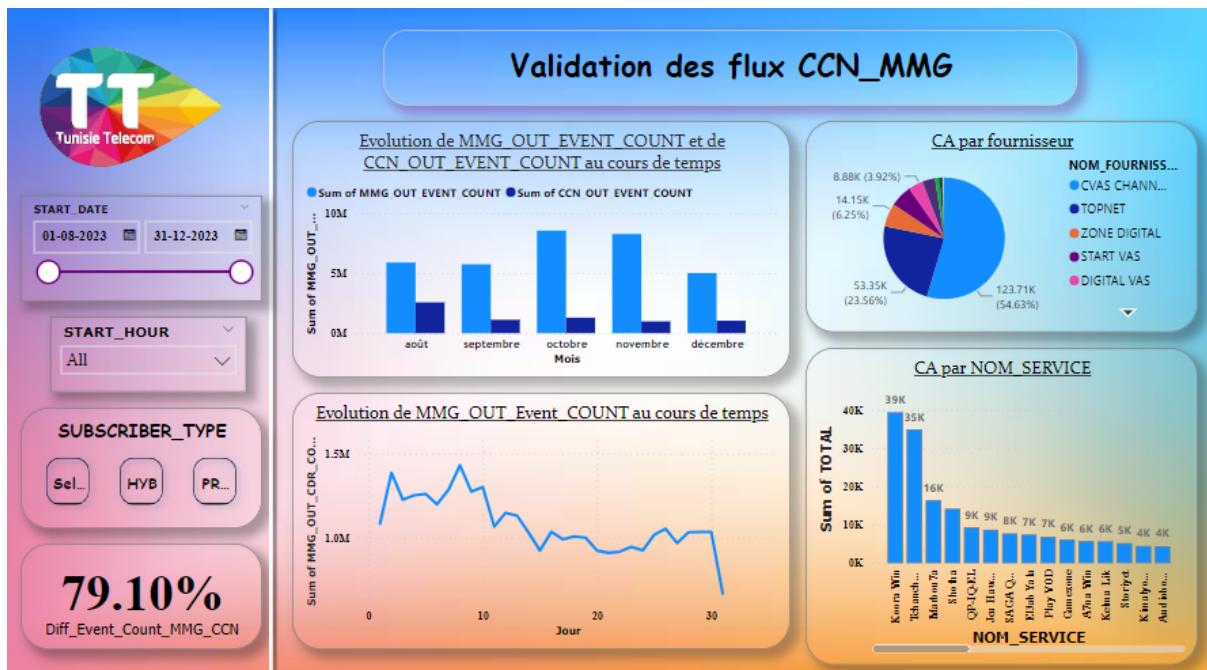


FIGURE 6.16 – Validation des flux CCN_MMG

6.4.4.7 Rapport : Validation des flux de taxation CCN et le service des appels et des sms MSC

Enfin, nous analysons les flux de taxation CCN et service des appels et des sms MSC. Les éléments visualisés incluent :

- La différence d'Event_Count entre MSC et CCN.
- La différence d'Event_Duration entre MSC et CCN.
- * Pour la comparaison :
 - L'évolution du CCN_Event_Count et du MSC_Event_Count au fil du temps.
 - L'évolution du CCN_Event_Duration et du MSC_Event_Duration au fil du temps.
- * Pour le suivi de la tendance du trafic :
 - L'évolution du MSC_Event_Count au fil du temps.
 - La répartition du MSC_Event_Count par type d'appel (call_type).

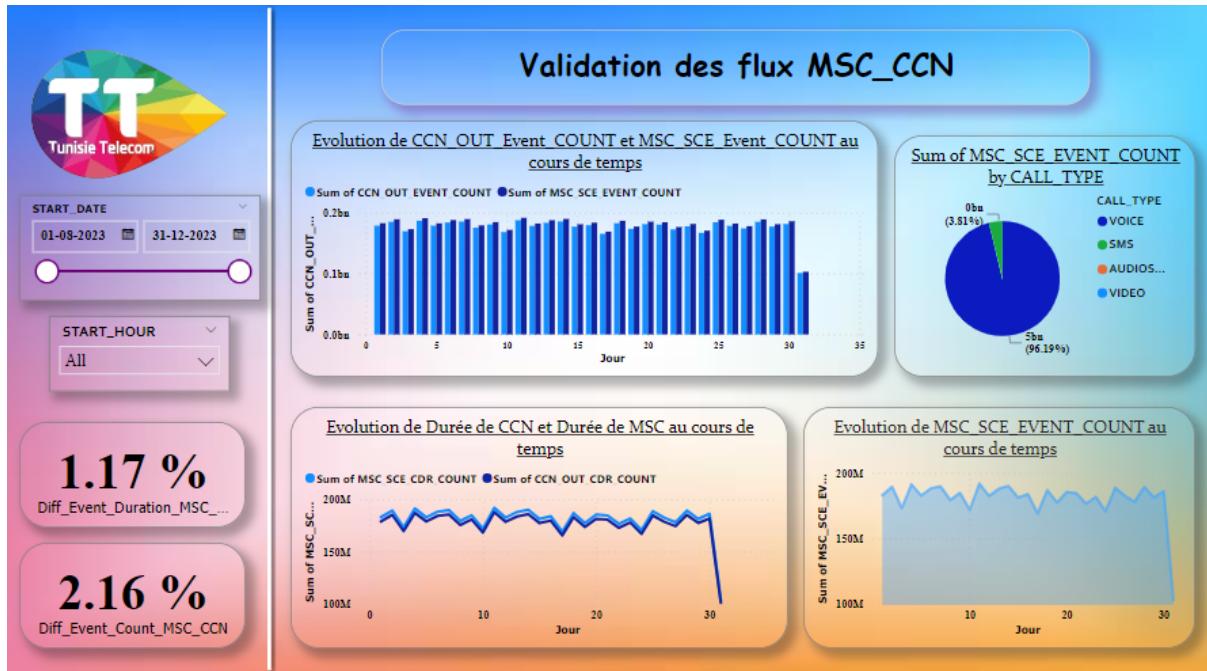


FIGURE 6.17 – Validation des flux MSC_CCN

6.5 Conclusion

Dans ce chapitre, nous avons créé et configuré un tableau de bord dynamique. Ensuite, Nous avons détaillé les pages de ce tableau en indiquant les thèmes d'analyse de données et en justifiant le choix des graphiques utilisés.

Chapitre 7 : Sprint 4 : Analyse et Fouille de données

Sommaire

7.1	Introduction	90
7.2	Sprint Backlog	90
7.3	Diagramme du cas d'utilisation du sprint 4 « Gérer les structures de fouille de données »	91
7.3.1	Diagramme du cas d'utilisation	91
7.3.2	Description textuelle du cas d'utilisation « Gérer structure de fouille de données »	91
7.4	Compréhension des données et descriptions des variables	92
7.5	L'analyse descriptive des variables	92
7.5.1	Description statique	92
7.5.2	Distribution de la variable cible	93
7.5.3	Distribution de la variable statut	94
7.6	Choix du modèle	94
7.7	Entraînement du Modèle	95
7.8	Evaluation du Modèle	96
7.9	Interpretation des résultats	97
7.10	Conclusion	98

7.1 Introduction

Après avoir mis en place divers tableaux de bord pour surveiller l'utilisation des services et comparer l'activité réelle des abonnés aux systèmes de taxation de Tunisie Telecom afin d'identifier les écarts et anomalies, résoudre les incidents et réduire les pertes financières, en remboursant uniquement les abonnés actifs plutôt que de privilégier ceux qui ont récemment rechargé. Nous continuerons à améliorer nos services pour nos abonnés. En intégrant l'intelligence artificielle, nous allons développer des modèles prédictifs pour estimer la probabilité de paiement des clients SOS. Cela permettra d'anticiper les flux de trésorerie et de prendre des décisions stratégiques concernant le remboursement des abonnés demandant un crédit après avoir dépassé leur limite d'accès autorisée.

Nous débutons ce chapitre par la description du sprint 4 "Analyse et fouille de données", puis nous mettons en place des structures d'analyse et de fouille de données en utilisant plusieurs algorithmes.

7.2 Sprint Backlog

Le tableau 7.1 représente le Backlog du sprint 4

ID	Fonctionnalité	ID	User Stories	ID	Description des tâches
4	Analyse et fouille de données	4.1	En tant que développeur, je peux faire des prédictions à partir des données de l'entrepôt de données	4.1.1	Importer les données de l'entrepôt de données (ED).
				4.1.2	Gérer les structures de fouille de données.

TABLE 7.1 – Backlog du sprint 4

7.3 Diagramme du cas d'utilisation du sprint 4 « Gérer les structures de fouille de données »

7.3.1 Diagramme du cas d'utilisation

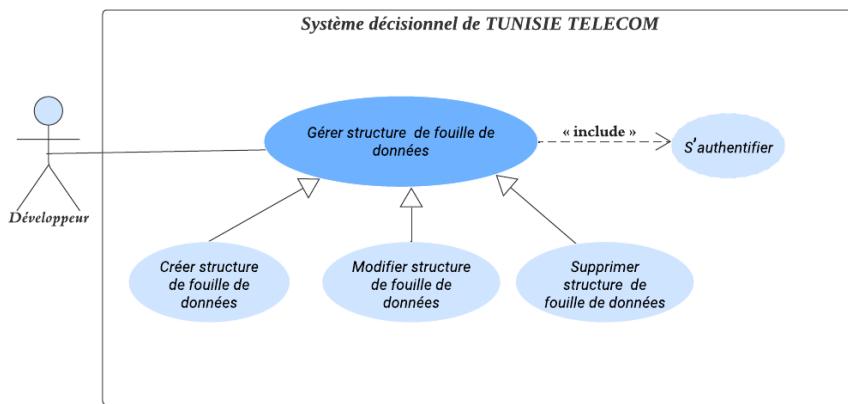


FIGURE 7.1 – Diagramme du cas d'utilisation du sprint 4

7.3.2 Description textuelle du cas d'utilisation « Gérer structure de fouille de données »

Le tableau 7.2 montre la description textuelle du cas d'utilisation « Gérer structure de fouille de données»

Titre	Gérer structure de fouille de données.
Acteurs	Développeur.
Objectif	Permet de créer des algorithmes de prédictions.
Pré-condition	Les données nécessaire pour la validation des algorithmes sont disponibles.
Post-condition	Les performances de l'algorithme sont évaluées et vérifiées pour assurer sa fiabilité et son efficacité..
Scénario nominal	<ol style="list-style-type: none"> 1. Le développeur accède à l'environnement de développement du système. 2. le développeur doit identifier la source des données. 3. Le développeur met en œuvre l'algorithme de prédition en utilisant un langage de programmation et les bibliothèques. 4. Le développeur procède à l'entraînement de l'algorithme en utilisant les données disponibles, en ajustant les paramètres et en évaluant les performances. 5. Le développeur effectue des tests et des validations pour s'assurer de la qualité et de la précision de l'algorithme.

TABLE 7.2 – Description textuelle du cas d'utilisation « Gérer structures de fouille de données »

7.4 Compréhension des données et descriptions des variables

Les données fournies pour cette étude de prédiction comprennent un ensemble exhaustif de variables relatives aux comportements des abonnés dans le réseau de télécommunications durant les trois derniers mois. Voici un aperçu des principales caractéristiques :

- **Numéro de téléphone (MSISDN)** : Identifie de manière unique l'abonné associé à la transaction.
- **Statut (STATUS)** : Décrit l'état de la transaction ou de l'enregistrement.
- **Type d'abonné (SUBSCRIBER_TYPE)** : Catégorise le type d'abonné, permettant une segmentation précise.
- **Montants divers (AMOUNT_*)** : Incluent des données financières telles que les montants associés à des recharges, des transferts, des services spécifiques (voix, données, SMS, etc.), ainsi que des crédits et des débits SOS.
- **Classe de service (SERVICE_CLASS_*)** : Identifie la classe de service à laquelle l'abonné est associé, avec une description correspondante.

Le score de remboursement (**REPAYMENT_SCORE**) est une variable clé dans nos données, car elle fournit une indication du comportement de remboursement des abonnés. Ce score sera utilisé comme variable cible pour prédire la propension des abonnés à rembourser leur solde SOS. Cette prédiction revêt une importance capitale dans la gestion du risque financier, car elle permet d'anticiper si un abonné remboursera ses dettes ou non. Le score peut prendre la valeur 0 si l'abonné n'a jamais rechargé son solde SOS, et 1 s'il l'a rechargé au moins une fois.

7.5 L'analyse descriptive des variables

7.5.1 Description statique

Cette image présente les statistiques descriptives des montants associés à différentes services offrant ainsi un aperçu détaillé des comportements financiers des abonnés (un aperçu rapide des données du fichier source).

	AMOUNT_VOUCHER	AMOUNT_ETOPUP	AMOUNT_TRANSFERT_IN	AMOUNT_SOS_CREDIT	\	
count	29068.000000	29068.000000	29068.000000	29068.000000		
mean	0.087106	0.000845	0.003939	0.018784		
std	0.849728	0.049103	0.117132	0.239999		
min	0.000000	0.000000	0.000000	0.000000		
25%	0.000000	0.000000	0.000000	0.000000		
50%	0.000000	0.000000	0.000000	0.000000		
75%	0.000000	0.000000	0.000000	0.000000		
max	40.000000	5.000000	6.000000	10.000000		
	AMOUNT_SOS_DEBIT	AMOUNT_TRANSFERT_OUT	AMOUNT_INJECTION	AMOUNT_DATA	\	
count	29068.000000	29068.000000	29068.0	29068.000000		
mean	0.055199	0.000069	0.0	0.009206		
std	0.570768	0.008295	0.0	0.117754		
min	0.000000	0.000000	0.0	0.000000		
25%	0.000000	0.000000	0.0	0.000000		
50%	0.000000	0.000000	0.0	0.000000		
75%	0.000000	0.000000	0.0	0.000000		
max	35.830000	1.000000	0.0	4.850000		
	AMOUNT_VOICE	AMOUNT_SMS	AMOUNT_VAS	AMOUNT_RBT	AMOUNT_MMS	\
count	29068.000000	29068.000000	29068.000000	29068.000000	29068.0	
mean	0.011110	0.000449	0.000002	0.000165	0.0	
std	0.127452	0.012837	0.000411	0.009009	0.0	
min	0.000000	0.000000	0.000000	0.000000	0.0	
25%	0.000000	0.000000	0.000000	0.000000	0.0	
50%	0.000000	0.000000	0.000000	0.000000	0.0	
75%	0.000000	0.000000	0.000000	0.000000	0.0	
max	4.370000	1.130000	0.070000	0.800000	0.0	

FIGURE 7.2 – Statistiques descriptives

7.5.2 Distribution de la variable cible

Le graphique montre clairement la répartition des scores de remboursement (**REPAYMENT_SCORE**) dans l'échantillon, ce qui est essentiel pour comprendre la propension globale des abonnés à rembourser leurs SOS solde.

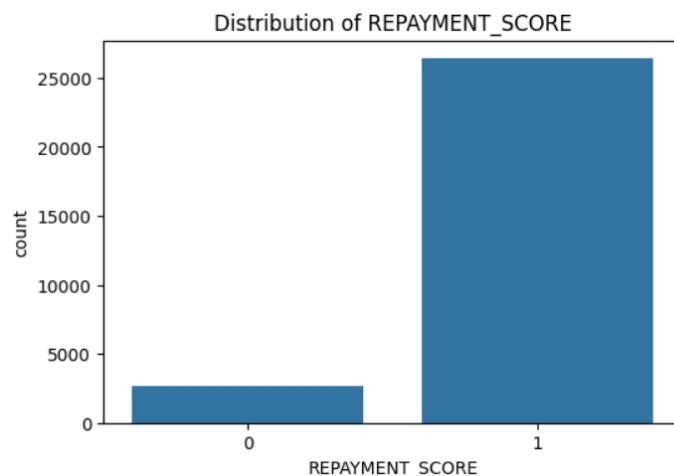


FIGURE 7.3 – Distribution de la variable cible

7.5.3 Distribution de la variable statut

Cette représentation graphique facilite une compréhension rapide de la distribution des différents statuts de transaction et des enregistrements dans le système, notamment les statuts "Actif" (Active), "En attente" (On hold), et "Suspendu" (Suspended).

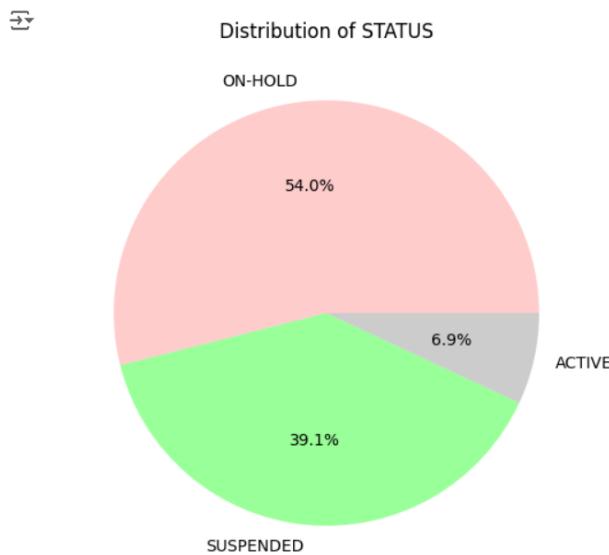


FIGURE 7.4 – Distribution de la variable status

7.6 Choix du modèle

Dans notre contexte, nous distinguons deux types de modèles : les modèles supervisés et les modèles non supervisés.

- * **Les modèles supervisés** sont utilisés lorsqu'il existe des données étiquetées et que l'objectif est de prédire une variable cible à partir des caractéristiques.
- * **Les modèles non supervisés** sont employés pour découvrir des structures dans les données en l'absence d'étiquettes.

Après avoir analysé notre problématique de prédiction du score de remboursement des abonnés en utilisant des variables explicatives, il est évident que nous sommes face à un problème de classification supervisée. Notre objectif est de prédire la catégorie d'un échantillon basé sur ses variables explicatives. Les données sont étiquetées pour chaque abonné, ce qui permet d'apprendre à partir de ces étiquettes.

Parmi les modèles de classification supervisée, nous avons plusieurs options :

- **Régression Logistique** : Adaptée aux problèmes de classification binaire ou multiclasse, elle est également facile à interpréter.
- **Arbres de Décision** : Permettent de modéliser des relations complexes entre les caractéristiques et la variable cible.
- **Forêts Aléatoires** : Une combinaison d'arbres de décision qui peut améliorer les performances prédictives.
- **Réseaux de Neurones** : Des modèles complexes capables de capturer des relations non linéaires dans les données.

Pour notre problématique de prédiction du score de remboursement des abonnés, nous avons choisi d'utiliser la régression logistique pour plusieurs raisons. Premièrement, elle offre une excellente interprétabilité grâce à ses coefficients, permettant de comprendre l'impact relatif de chaque caractéristique sur la prédiction du score de remboursement. De plus, elle est généralement rapide à entraîner, même sur des ensembles de données volumineux. Enfin, elle est particulièrement adaptée aux problèmes binaires, puisque notre objectif est de déterminer si un abonné remboursera ou non.

7.7 Entraînement du Modèle

* Séparation des Données

Nous avons commencé par diviser les données en ensembles "d'entraînement (training)" et de "test" à l'aide de la fonction `train_test_split`, en attribuant 80 % des données à l'entraînement et 20% aux tests. Cette étape est essentielle pour évaluer la capacité de généralisation du modèle.

* Entraînement du Modèle

Nous avons ensuite procédé à l'entraînement du modèle en utilisant une pipeline complète qui intègre à la fois le prétraitement des données et l'algorithme de régression logistique. Cette approche structurée permet de garantir que toutes les étapes nécessaires sont correctement appliquées avant le processus d'entraînement du modèle.

* Prédiction avec le Modèle

Une fois le modèle entraîné, nous avons utilisé la méthode `predict` pour réaliser des prédictions sur l'ensemble de test. Cette étape est très importante car elle permet de mesurer la performance du modèle sur des données qu'il n'a jamais vues auparavant.

7.8 Evaluation du Modèle

Pour évaluer les performances de notre modèle de classification, nous utilisons plusieurs métriques et outils. Voici une analyse détaillée de ces résultats :

- * **Exactitude (Accuracy)** L'exactitude mesure la proportion de prédictions correctes parmi toutes les prédictions. Pour notre modèle, l'exactitude est de 97.9%, ce qui signifie que près de 98% des prédictions sont correctes.
- * **Précision (Precision)** La précision est la proportion des vrais positifs parmi toutes les instances prédites comme positives. Pour la classe 1 (remboursement effectué), notre modèle a une précision de 98%. Cela signifie que parmi toutes les prédictions positives, 98% sont réellement positives.
- * **F1-Score** Le F1-score est la moyenne harmonique de la précision et du rappel. Il fournit une mesure équilibrée entre la précision et le rappel. Pour la classe 1, notre modèle obtient un F1-score de 99%.
- * **Matrice de Confusion** La matrice de confusion fournit une vue détaillée des performances du modèle. Elle montre le nombre de vrais positifs, de faux positifs, de vrais négatifs et de faux négatifs. Dans notre cas, la matrice de confusion montre que notre modèle a correctement classé la plupart des instances, avec seulement quelques faux positifs.

```
Accuracy: 0.979016167870657
Classification Report:
precision    recall    f1-score   support
          0       1.00      0.78      0.88      567
          1       0.98      1.00      0.99     5247

accuracy                           0.98      5814
macro avg                           0.99      0.89      0.93      5814
weighted avg                          0.98      0.98      0.98      5814

Confusion Matrix:
[[ 445  122]
 [  0 5247]]
```

FIGURE 7.5 – Metriques d’evaluation

7.9 Interpretation des résultats

Les scores obtenus indiquent que notre modèle de classification performe très bien à la fois sur le jeu d’entraînement et sur le jeu de test. Avec un score de 97.87% sur le jeu d’entraînement et 97.90% sur le jeu de test, témoignent d’une bonne capacité de généralisation du modèle, ce qui est essentiel pour garantir sa performance dans des situations réelles et pour éviter le surajustement aux données d’entraînement. En résumé, ces résultats indiquent que le modèle est bien adapté à la tâche de classification sur laquelle il a été formé.

```
Training set score: 0.9787
Test set score: 0.9790
```

FIGURE 7.6 – Les résultats obtenus

On a recours à un graphique pour illustrer clairement les résultats du modèle. Les valeurs sont tracées en fonction de l’indice de l’échantillon, avec les valeurs de test en bleu et les valeurs prédictes en orange. L’axe des x représente l’indice de l’échantillon, tandis que l’axe des y représente les valeurs réelles ou prédictes. Cette visualisation permet de comparer visuellement les valeurs réelles et les prédictions du modèle sur un ensemble de données.

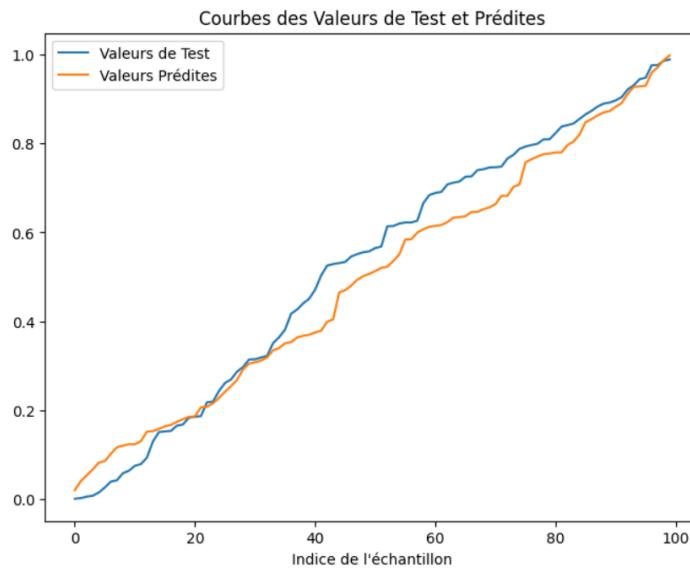


FIGURE 7.7 – Courbes des Valeurs de Test et Prédites

7.10 Conclusion

Dans ce chapitre, nous avons abordé la prédiction du score de remboursement des abonnés en utilisant des caractéristiques explicatives variées. Pour accomplir cela, nous avons mis en place des structures de fouille de données en recourant à l'algorithme de régression logistique.

CONCLUSION GÉNÉRALE

Au terme de ce rapport, nous présentons une évaluation détaillée de notre projet de fin d'études, qui visait à créer un tableau de bord pour surveiller et analyser les flux de données télécom. Ce projet nous a permis de renforcer nos compétences en informatique décisionnelle tout en acquérant une précieuse expérience professionnelle dans ce domaine.

Après une analyse approfondie des données opérationnelles et plusieurs discussions avec les décideurs pour identifier leurs besoins, nous avons d'abord procédé à la transformation, au nettoyage et au chargement des données sources dans une zone de transit appelée Staging Area (SA). La mise en place de cette zone était cruciale pour garantir le bon déroulement des processus ETL (Extraction, Transformation, Chargement). Ensuite, nous avons développé les flux de données entre la SA et l'entrepôt de données (ED), mettant en œuvre les processus ETL après avoir conçu notre entrepôt de données.

Dans un deuxième temps, nous avons instauré un processus automatisé pour alimenter l'entrepôt de données à partir de la SA via les processus ETL. Cette automatisation a assuré l'exécution régulière et fiable des mises à jour des données. Sur la base de l'entrepôt de données alimenté, nous avons pu générer des rapports composés de tableaux de bord, en analysant diverses mesures issues des tables de faits ainsi que celles calculées à l'aide de requêtes DAX.

Enfin, nous avons appliqué un algorithme d'analyse et de fouille de données pour produire des prédictions précises. Ces prédictions ont permis de prendre des décisions éclairées concernant

CONCLUSION GÉNÉRALE

le remboursement des utilisateurs du service SOS lorsqu'ils dépassent leur forfait.

Comme perspectives, toujours dans le cadre de la facilitation de la prise de décision, nous visons à renforcer le processus de fouille de données et à améliorer notre modèle de prédition. Pour ce faire, nous prévoyons d'implémenter une gamme plus large d'algorithmes de machine learning afin d'identifier ceux qui offrent la plus grande précision et performance.



BIBLIOGRAPHIE

- [1] **Tunisie Telecom** <https://www.tunisetelecom.tn/particulier/a-propos-de-tt/>
- [2] **Organigramme Tunisie Telecom** <https://myspace.tunisetelecom.tn/Fr/Documents/tt-organisation+structurelle.pdf>
- [3] **BILL Inmon** <https://theses.hal.science/tel04443377v1/file/theseinternet/benhissen/r.pdf>
- [4] **ETL** " [https://www.talend.com/fr/resources/guide-etl/"](https://www.talend.com/fr/resources/guide-etl/)
- [5] **Ralph Kimball** " <https://www.lamsade.dauphine.fr/negre/fichierjoints/BI-classique.pdf> "
- [6] **Ghozzi,F. (2004)** [https://fr.scribd.com/document/601031703/these-Ghozzi-faiza.](https://fr.scribd.com/document/601031703/these-Ghozzi-faiza)
- [7] **Asanka, 2019** <https://www.uv.es/nemiche/cursos/polycopies/120Data20Mining.pdf>
- [8] **Oracle sql developper** <https://en.wikipedia.org/wiki/Oracle/SQL/Developer>
- [9] **talend** <https://www.talend.com/fr/>
- [10] **power bi** <https://www.microsoft.com/fr-fr/power-platform/products/power-bi>
- [11] **Staging Area (SA)** <https://datascientest.com/staging-area-tout-savoir>
- [12] **Dax** <https://www.questionpro.com/blog/fr/quest-ce-que-lanalyse-de-donnees/>

Mise en place d'un tableau de bord de suivi des flux réseaux télécom de Tunisie Telecom

WALA BEN RHOUMA & SOUAD ACHOURI

الخلاصة:

هذا العمل يندرج ضمن إطار مشروع نهاية الدراسات الذي نسعى من خلاله للحصول على درجة البكالوريوس في تحليل البيانات الكبيرة والبيانات. الهدف من هذا المشروع هو إنشاء لوحة معلومات لمراقبة وتحليل تدفقات البيانات الهاتفية. لتحقيق هذا الهدف، قمنا بتنفيذ مستودع بيانات واستعادة البيانات المعالجة في لوحة معلومات ديناميكية، وفقاً لمواصفات مشروعنا. وأخيراً، قمنا بتطبيق نموذج للتنقيب في البيانات.

Résumé :

Ce travail s'inscrit dans le cadre de notre projet de fin d'études en vue de l'obtention de la licence en big data et analyse de données. L'objectif de ce projet est de mettre en place un tableau de bord pour surveiller et analyser les flux de données télécom. Pour atteindre cet objectif, nous avons mis en œuvre un entrepôt de données et restitué les données traitées dans un tableau de bord dynamique, conformément aux spécifications de notre projet. Enfin, nous avons appliqué un modèle de fouille de données .

Abstract:

This work is part of our final year project aimed at obtaining a Bachelor's degree in Big Data and Data Analytics. The objective of this project is to implement a dashboard to monitor and analyze telecom data flows. To achieve this goal, we implemented a data warehouse and presented the processed data in a dynamic dashboard, in accordance with the specifications of our project. Finally, we applied a data mining model.

المفاتيح : نظام صنع قرار ، مستودع البيانات ، التحليل متعدد الأبعاد ، لوحة معلومات ، تحليل وتنقيب البيانات .

Mots clés : Système décisionnel, entrepôt de données, analyse multidimensionnelle, tableau de bord , analyse et fouille de données .

Key-words: Decision-making system, data warehouse, multidimensional analysis, dashboard, data mining and analysis.

