Gouvernance et management des données clients dans un réseau de transport : exemple du Havre

Pierre Rolland

le cnam
intd

Mémoire

pour l'obtention du Master

Méga données et analyse sociale (Médas)

Gouvernance et management des données clients dans un réseau de transport : exemple du Havre

Pierre ROLLAND

Date et lieu de la soutenance

- 11/09/2019
- Saint-Denis

Membres du jury

- Béatrice ARRUABARRENA, Présidente du jury
- Ghislaine CHARTRON, Tutrice pédagogique
- Denis COUTROT, Maître d'apprentissage (entreprise : TRANSDEV)

Promotion 2 (2017-2019)

Remerciements

Je tiens à remercier l'ensemble de l'équipe du Département Data Science de Transdev et tout particulièrement mon maître d'apprentissage Denis Coutrot. J'ai passé deux années en apprentissage très enrichissantes sur les problématiques de transport et données.

Sur la partie académique, je remercie ma tutrice pédagogique du Master MEDAS, Madame Ghislaine Chartron pour nos échanges.

L'alternance pendant deux années, est une expérience plutôt intense, je salue l'ensemble de mes camarades de la promotion 2017-2019 du Master MEDAS.

Résumé

Ce mémoire de Master a pour objectif de présenter des éléments de gouvernance de la donnée dans le cadre de la consolidation d'une base clients dans le domaine des transports publics. Une première partie s'attachera à étudier ces éléments de gouvernance de la donnée de manière théorique tout en contextualisant les enjeux en termes de données clients (données de références) pour un réseau de transport en commun. Dans une seconde partie, un cas concret d'un réseau de transport public, le réseau du Havre, sera étudié. Plus précisément seront étudiées les différentes bases portant une information client-voyageur et les problématiques rencontrées lors de la consolidation de ces bases de données dans l'objectif de tendre vers une vision client unifié dans une logique de Customer Relationship Management (CRM).

Mots clés : Transports publics ; Données de référence ; Master Data Management ; Données clients ; Référentiel Client Unique ; Gouvernance de la donnée ; Connaissance Client

Abstract

This Master's thesis aims to present elements of data governance as part of the consolidation of a customer base in the field of public transport. A first part will focus on studying these elements of data governance in a theoretical way while contextualizing the issues in terms of customer data (reference data) for a public transport network. In the second part, a concrete case of a public transport network, Le Havre network, will be studied. More precisely, the different databases client information and the problems encountered when consolidating these databases in order to achieve a unified customer vision in a Customer Relationship Management (CRM) approach will be studied.

Keywords : Public transports ; Master Data Management ; Customer data ; Data management ; Customer knowledge

Table des matières

Introduction **5**

I- Contextualisation et enjeux de la consolidation d'une base clients dans le domaine des transports **7**

1.1. Clients, données clients et relation avec le réseau de transport 7

1.1.1. Les différents points de contact entre le client et le réseau de transport 7

1.1.2. Les données clients dans le domaine de la mobilité et des transports en commun 11

1.1.3. Les données clients comme données de référence pour l'entreprise 14

1.1.4. Une base clients pour alimenter le CRM (« Customer Relationship Management ») .. 16

1.2. Eléments de gouvernance de ces données clients 21

1.2.1. Construction d'un Référentiel Client Unique 21

1.2.2. Qualité des données 24

1.2.3. Protection des données personnelles 26

1.3. Le contexte de la donnée chez Transdev, une organisation décentralisée 29

1.3.1. Présentation de Transdev 29

1.3.2. La problématique « Data » liée à cette organisation décentralisée 30

1.3.3. A quelle échelle penser le Référentiel Client Unique (RCU) ? 31

1.3.4. Le Département Data Science dans cette organisation 31

II- Etude de cas : réseau de transport LiA au Havre **33**

2.1. Premières observations des recoupements des bases de données de LiA 33

2.1.1. Genèse du projet 33

2.1.2. « Data Camp » et récolte des différentes sources de données 34

2.1.3. Panorama des données collectées 35

2.1.4. Recoupement des différentes bases de données (vision unifiée des bases de données clients) 36

2.2. Application des recoupements au Transport A la Demande 40

2.2.1. Présentation du Transport à la Demande (TAD) 40

2.2.2. Les données du TAD 41

2.2.3. Premiers résultats des recoupements 42

2.3. Vers une consolidation d'une base clients 45

2.3.1. Qualité de la donnée 45

2.3.2. Champs structurants et problématique des données personnelles 46

2.3.3. Architecture orientée clients pour une approche « CRM 360° » 47

Conclusion **50**

Bibliographie **52**

Liste des graphiques **54**

Liste des tableaux **54**

Annexes **55**

Introduction

Une nouvelle approche s'est imposée aux entreprises concernant leur relation avec leurs clients nécessitant une relation plus personnalisée et plus directe. Cela implique une connaissance plus fine et plus globalisée des caractéristiques et comportements de leurs clients ainsi que de leurs habitudes (d'achats, d'utilisation de services…). On parle alors de « Connaissance Client » (Jallat et al., 2018) et de « Customer Relationship Management » (CRM) (Delers, 2015). Nous aborderons ces notions qui touchent également le domaine des transports publics, en insistant sur les spécificités de ce domaine, sous le prisme de la donnée.

Nous verrons, dans un premier temps, que cette nouvelle approche « orientée client » engendre une certaine complexité : comment centraliser, consolider toute l'information client portée par des données qui sont aujourd'hui très souvent stockées et administrées « en silo ».
Il s'agira avant tout de comprendre ce qu'est un client dans le domaine des transports publics. Celui-ci est à la fois acheteur de titres de transport (carte d'abonnement mensuel ou annuel, ou des titres unitaires) et voyageur (utilisant un service de mobilité à l'aide de son titre de transport). Comprendre ce qu'est un client mais également sa relation avec le réseau de transport et pour finir quelles données lui sont rattachées.
Toujours dans cette première partie, nous présenterons des éléments de gouvernance de la donnée centrés « client ». Nous verrons en quoi les données clients des entreprises peuvent être considérées comme stratégiques et être managées comme des données de références (Master Data Management). Nous nous intéresserons notamment, au-delà des aspects techniques de la consolidation d'une base clients, à la prise en compte d'un volet qualité de la donnée et d'un volet protection des données personnelles, éléments essentiels pour la réconciliation du patrimoine de données clients hétérogènes dans l'objectif d'acquérir une vue client unique.

Enfin, avant de présenter un cas concret d'une approche de consolidation d'une base clients dans un réseau de transport, nous analyserons la relation entre une base clients consolidée et l'approche CRM ainsi que certaines applications rendues possibles à la suite de cette consolidation.
Pour finir cette première partie de contextualisation et de présentation des enjeux de la consolidation d'une base clients dans le domaine des transports, nous présenterons le groupe Transdev dans lequel j'effectue mon apprentissage afin d'appréhender son organisation ainsi que le contexte dans lequel la donnée est créée et évolue.

Dans une deuxième partie, à partir des éléments de gouvernance de la données évoqués dans la première partie, nous présenterons une étude de cas que nous avons effectuée avec un réseau de transport en commun, celui de la Communauté Urbaine Le Havre Seine Métropole, aussi appelé LiA et dont une filiale de Transdev a la charge de l'exploitation.

Après avoir présenté la genèse du projet et le partenariat avec le réseau LiA, la récolte des données et leur recoupements possibles permettant d'avoir une certaine vision unifiée des bases de données clients, nous verrons un exemple concret pour le réseau de l'utilité de ces recoupements pour un mode de transport spécifique, le transport à la demande. Ces recoupements permettront d'apporter une certaine « Connaissance Client » pour un mode de transport où cette dernière est relativement faible, rendant également possible un ciblage marketing plus précis.

Pour finir, nous mettrons en avant, toujours en lien avec cette étude de cas, quelques éléments pratiques de gouvernance de la donnée et d'organisation dans une logique de pérennisation et d'automatisation d'une telle démarche de consolidation d'une base de données clients.

I- Contextualisation et enjeux de la consolidation d'une base clients dans le domaine des transports

Dans cette première partie, nous allons tenter de comprendre et mettre en avant les différents points de contact entre le client, qui est aussi un voyageur, et le réseau de transport à l'heure de la cross-canalité. Nous décrirons ces données clients et leur importance stratégique pour le réseau de transport ainsi que leurs potentielles utilisations et valorisations. Nous pourrons par la suite évoquer les problématiques de gouvernance de ces données clients et enfin, nous présenterons le contexte de l'organisation de Transdev notamment en lien avec ce sujet des données clients.

1.1. Clients, données clients et relation avec le réseau de transport

1.1.1. Les différents points de contact entre le client et le réseau de transport

A tous types de ventes de produits ou services correspondent des clients (ou voyageurs) et ces derniers entrent en contact avec les offrants par différents points de contact. Il existe en marketing une notion classique de « cycle de vie client » dont l'une des définitions désigne ce cycle de vie client comme « l'évolution de la relation entre un client et l'entreprise »[1] . A ce cycle de vie client, correspond différentes phases qui sont classiquement :

- L'état de prospect
- L'entrée en relation
- Le premier achat
- La relation continue (achats répétés)
- L'inactivité (« churn » ou période de vacance dans la relation)

Ces différents états sont communément admis dans la littérature marketing. En les adaptant au secteur de la mobilité, nous chercherons dans un premier temps à identifier les différents points de contacts entre un client et le réseau de transport.

Remarquons dès à présent que ces points de contacts ne sont **pas linéaires dans le temps**. Ils mixent par ailleurs un volet **physique et un aspect numérique**. Certaines actions peuvent aussi être **anonymes** (ticket à l'unité acheté avec de la monnaie).

Une des spécificités pour un client dans les transports en commun est que le service qu'il achète peut (et l'est souvent) être différé de sa « consommation », c'est-à-dire qu'il peut par exemple acheter un titre de transport (carnet de 10 voyages, un abonnement mensuel ou annuel) et le consommer en plusieurs fois, dans le temps… Ce sont tous ces différents actes (soit d'achat, soit de consommation) qui nous intéressent et qui constituent des points de contacts. Cette consommation est représentée généralement par les **validations** dans les différents modes de transport.

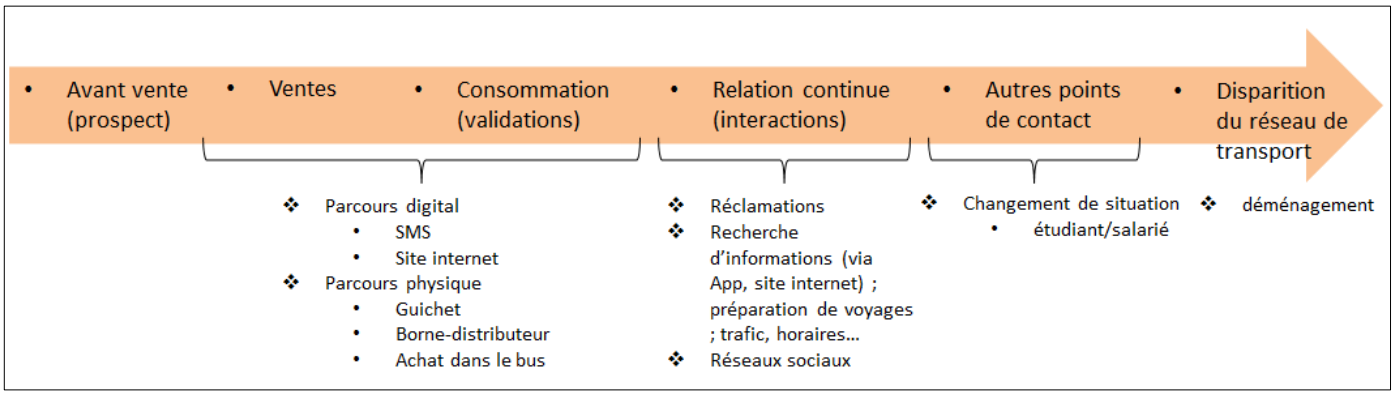
[1] Article « Définition : Cycle de vie client », Bertrand Bathelot, https://www.definitions-marketing.com/definition/cycle-de-vie-client/ , consulté le 17/04/2019.

De plus, est apparu ce que l'on nomme la « **crosscanalité** » (Belvaux et Notebaert, 2018) multipliant *de facto* les points d'entrée (achat par sms, par bornes, réservations par applications ou directement par téléphone, notation des différents services, avis sur les réseaux sociaux, consultation du site internet du réseau de transport ou de l'application smartphone pour chercher diverses informations…).

Du fait du développement de la **multi-modalité**, les offres de mobilité au sein d'un réseau ont évolué et se sont développées ces dernières années pour proposer des moyens de transports variés à leurs voyageurs. Avec le développement de ces différentes offres de transport les points de contact se sont multipliés.

A l'aide de ce qui vient d'être mentionné, nous avons tenté de construire une cartographie des différents points de contact avec les clients pour un réseau de transport :

Graphique 1 : Cartographie des points de contacts des clients avec un réseau de transport



Nous pouvons constater avec cette cartographie que l'information concernant le client, et non pas encore la donnée, est répartie sur l'ensemble du parcours client. La multiplicité de ces points de contact, ou points d'entrée est une source de complexité quant à la gestion des données clients, leur consolidation et recoupement.

En parallèle de cette cartographie des points de contact entre un client et un réseau de transport, nous pouvons établir une description de la « maturité client » du réseau de transport selon trois niveaux que l'on peut retrouver notamment chez Transdev :

- Niveau 1 : C'est le degré zéro de de l'information concernant le client (et donc le voyageur) tout au long de la chaîne des points de contact présentés plus haut. Cela correspond essentiellement à des voyageurs « non identifiés » ou occasionnels sur le réseau de transport qui n'ont pas de carte d'abonnement et qui ont acheté des **titres** (**à usage unique** ou à décompte) à un guichet, une borne ou directement dans le bus (qui sont les « points de contacts physiques » historiques et présents sur tous les réseaux). On ne peut étudier que statistiquement leurs validations. On est donc plus dans une approche d'étude sur un ensemble statistique de clients sans capacité d'interagir directement avec eux (seulement indirectement).
  C'est un cas de figure non négligeable dans le domaine des transports en commun. A titre d'exemple, en 2018 pour le réseau de transport de l'agglomération de Saint-Etienne, la STAS

gérée par Transdev, l'usage anonyme des titres à décompte représente un tiers des validations[2].

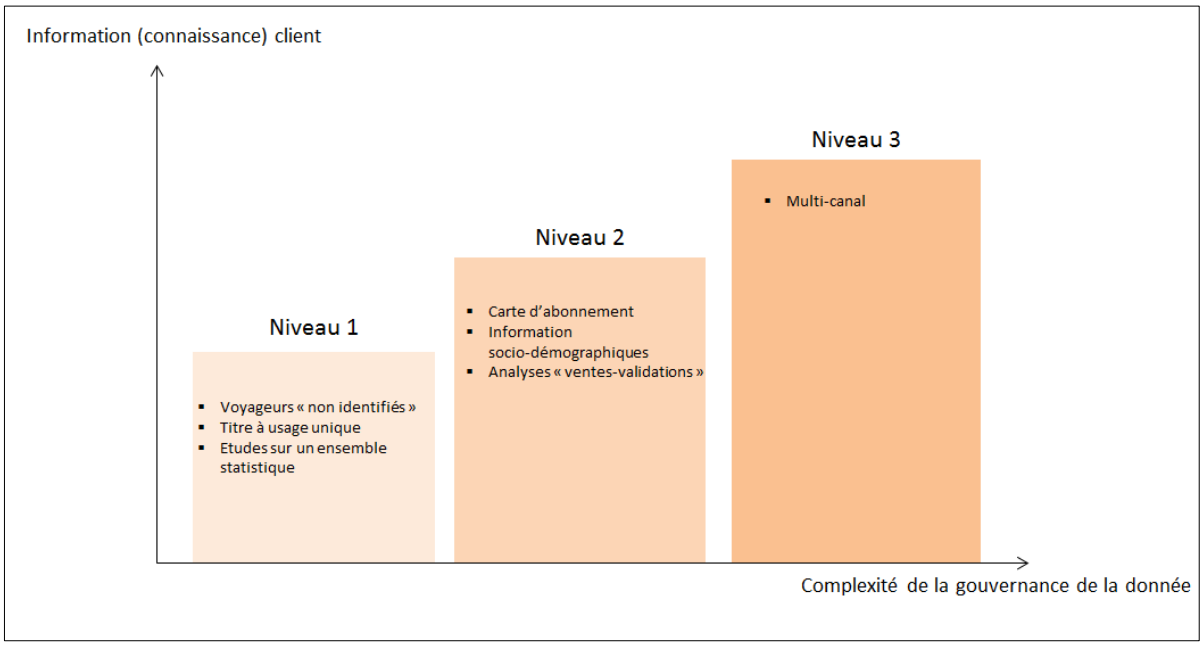
- Niveau 2 : Avec ce niveau, l'information est portée par une **carte d'abonnement** (le plus souvent mensuelle ou annuelle) comportant un identifiant du porteur de cette carte. Cette dernière permet l'accès à des **informations clients** que l'on peut qualifier de **« socio-démographiques** » (nous les présenterons dans la partie suivante) que l'on peut recouper avec les validations dans les différents modes de transport du réseau. Le point de contact de ce niveau reste exclusivement les « ventes-validations » de personnes abonnées permettant des études et analyses statistiques plus poussées (segmentation, profils voyageurs, RFM-PMG[3], habitudes de voyages…). Ces analyses se font sur des données anonymisées.
  La carte d'abonnement étant aujourd'hui adoptée dans quasiment l'ensemble des réseaux de transport, c'est essentiellement ce niveau d'information qui est traité dans les différents réseaux de transports de Transdev.

- Niveau 3 : ce niveau correspond à l'approche **multi-canal** qui engendre une multiplication des points de contact (notamment via le numérique et les approches services). Ce niveau englobe toute la chaîne des points de contacts des clients avec un réseau de transport (*cf.* Graphique 1) engendrant une augmentation de l'information sur le client. Nous le verrons par la suite, c'est avec ce niveau 3 que se pose notamment les problématiques de gouvernance et donc de réconciliation des différents systèmes d'information pas essence souvent incompatibles. C'est sur ce niveau 3 que repose essentiellement le sujet de ce mémoire.

A ces différents niveaux d'information client correspondent des leviers d'action différents (notamment en termes de marketing). La complexité de la gouvernance de la donnée augmente à mesure que l'information client augmentent. Au niveau 3, cette complexité est la plus forte. Nous avons schématisé cette « maturité client » dans le Graphique 2 :

[2] Chiffres internes à Transdev. A noter qu'un part très minoritaire d'utilisateurs de titres à décompte sont « embasés ».
[3] Segmentation RFM-PMG pour Récence (date de la dernière commande), Fréquence (des commandes) et Montant et Petit, Moyen, Grand qui permet d'établir des segments de clients homogènes.

Graphique 2 : Les trois niveaux de la « maturité client »



De manière générale, **c'est la valorisation en information de la donnée issue des contacts avec le client qui est recherchée et qui s'inscrit dans ce qu'on appelle la « Connaissance Client »**. Cette dernière peut être considérée dans un objectif stratégique visant à améliorer l'expérience vécue par les clients tout en permettant une mise en relation entre l'entreprise et les clients au bon moment et par le bon canal pour proposer une offre adaptée (Jallat et al., 2018).

Les trois niveaux d'information client présentés précédemment dans le cadre d'un réseau de transport offrent finalement des caractéristiques pour profiler les clients et se retrouvent dans l'approche classique de la « Connaissance Client » que certains auteurs (Jallat et al., 2018) ont découpé en 12 dimensions :

Tableau 1 : Dimensions de la connaissance client

Dimensions	Variables
Identité	Nom ; prénom ; adresses ; téléphone…
Caractéristiques	Age ; sexe ; composition du ménage ; revenu ; catégorie socioprofessionnelle
Segments	Appartenance du client à des groupes constitués par l'entreprise (marketing)
Transactions	Date, type, produits, de la transaction ; canal utilisé ; récence, fréquence, montant
Services	Contrats en cours de validité ; dates d'expiration ; garanties…
Evénements	Evénements pouvant avoir un impact sur les comportements à venir
Valeur économique	Chiffre d'affaires ; part du portefeuille ; valeur à vie du client
Usages	Comment, avec qui, où et pourquoi le produit/service est utilisé ; satisfaction et expérience de l'usage du produit/service
Parcours	Persona ; score d'efforts

Comportement omnicanal	Canaux utilisés ; historiques des interactions entre le client et l'entreprise
Préférences relationnelles	Type de relation souhaitée ; préférence pour le type de communication…
Etat de la relation	Recommandations (NPS) ; Fidélité ; formes d'engagement ; réclamations…

Source : Extrait du tableau des 12 dimensions de la connaissance client de JALLAT, Frédéric., PEELEN, Ed., STEVENS, Eric., VOLLE, Pierre. Gestion de la relation client Expérience client, Performance relationnelle et Hub relationnel. 5ème éd. France : Pearson, 2018.

Ce tableau très exhaustif et théorique peut être **résumé** de la sorte :

« Historiquement la Connaissance Client était surtout constituée de données socio-démographiques (identité, âge, adresse, etc.) et de données transactionnelles (historique des achats) »[4]. Ce qui correspond aux dimensions **Identité** et **Caractéristiques** pour les premières et aux dimensions **Transactions** mais aussi **Usages** pour les secondes.

On trouve des dimensions de « renseignements » : **Services**, **Evénements**, **Préférences relationnelles.** D'autres dimensions (**Segments**, **Valeur économique**, **Etat de la relation**) font l'objet « d'opérations à posteriori » (enrichissement de données, analyses, calculs de score…) pour essayer de comprendre les comportements et d'établir des segmentations clients.

Cette « Connaissance Client » constituée de l'ensemble des données relatives aux clients qu'une organisation collecte constitue la matière première pour le marketing.

1.1.2. Les données clients dans le domaine de la mobilité et des transports en commun

Nous venons de le voir, à ces informations clients collectées par les différents points de contact correspondent des données de types différents.

Dans une approche « data marketing », on peut classer les données clients en quatre catégories (Hirth, 2017) :

- Les données sociodémographiques
- Les données transactionnelles
- Les données comportementales
- Les données contextuelles

Reprenons cette catégorisation pour présenter les données clients dans un réseau de transport en commun. Nous nous appuyons sur les différents cas que nous avons pu étudier dans le cadre de notre apprentissage.

[4] Article « Définition : Connaissance client », Bertrand Bathelot, https://www.definitions-marketing.com/definition/connaissance-client/ , consulté le 26/04/2019.

Pour les **données sociodémographiques**, nous l'avons dit, elles sont essentiellement accessibles lors de la souscription à une carte d'abonnement de transport (indépendamment du canal de vente : guichet, borne, internet, application mobile…) et son stockées dans une base de données relationnelle – **Base Clients**. On retrouve généralement les champs suivants (à quelques exceptions près selon les réseaux) :

- ID Client
- Titre
- Sexe
- Prénoms
- Nom
- Date de naissance
- Adresse
- Numéro de téléphone (mobile et fixe)

A cela s'ajoute des données liées à la carte d'abonnement (**Base Carte**) qui permettent de caractériser un client, notamment le Code Produit qui permet de catégoriser le client :

- ID Carte
- ID Client
- Code Produit Code Produit (étudiant, scolaire, - 26 ans, sénior, demi-tarif... / mensuel, annuel, 10 titres, titre unitaire …)
- Fin de validité de la carte d'abonnement (ou du titre)

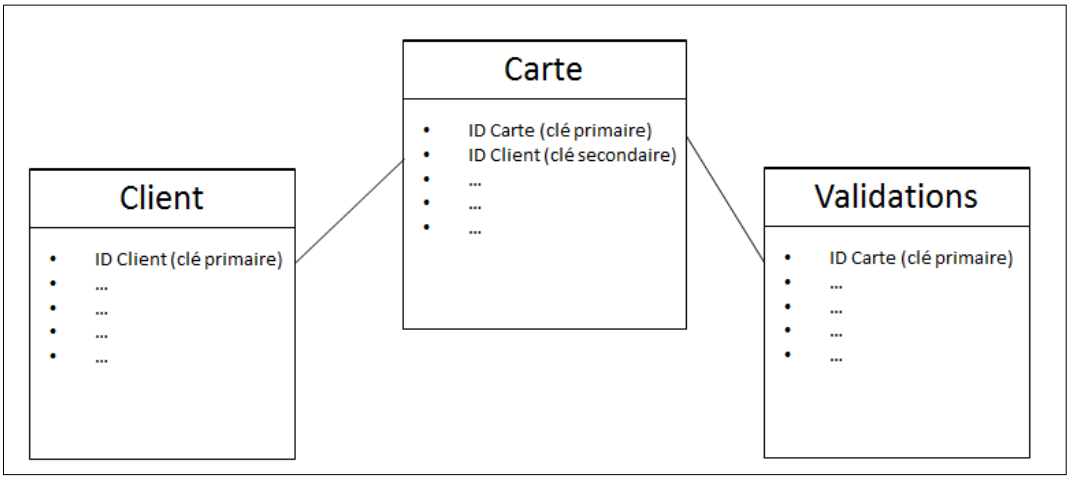
Les **données transactionnelles**, qui répondent à des questions telles que « combien ? », « quand ? », « quoi ? », « où ? », « comment ? » (Hirth, 2017), correspondent aux données de validations dans les différents modes de transport[5]. Ces dernières sont collectées dans leur ensemble par les appareils de validations. Elles sont soit collectées périodiquement pour être stockées dans une base de données relationnelle – **Base Validations**. On peut récupérer de cette base (à titre d'exemple) :

- ID Carte
- Date et Heure de la validation
- Code Produit (étudiant, scolaire, - 26 ans, sénior, demi-tarif... / mensuel, annuel, 10 titres, titre unitaire…)
- Mode utilisé (Tram, Bus…)
- Première montée ou correspondance
- Nombre de validations par passager (pour les accompagnants)
- Fin de validité de la carte d'abonnement (ou du titre)

Schématiquement, l'ensemble de ces bases de données relationnelles se croisent à l'aide de clés primaires et secondaires (les identifiants) :

[5] Précisons qu'il existe aussi la **base Ventes** qui référence l'ensemble des transactions de ventes de titres de transport mais n'offre pas d'information propre à caractériser un client si ce n'est ses actes d'achat.

Graphique 3 : Modélisation relationnelle entre les BDD Client, Carte et Validations



A ce schéma classique Client-Carte-Validations, s'ajoute les données provenant d'autres canaux. Il est par exemple possible dans certains réseaux de transport de créer un espace personnel (ou compte) sur le site internet du réseau de transport. Des informations, notamment sociodémographiques, caractérisant le client sont alors présentes (nom, prénom, numéro de téléphone, adresse e-mail).
De plus, de nouveaux modes de validations (autres que les tickets ou les cartes d'abonnement) tels que le ticket de bus par SMS[6] ou simplement le smartphone[7] comme titre de transport permettent de caractériser un client de manière plus partielle puisqu'on ne peut capter que le numéro de téléphone dans le cas du SMS. Tout l'enjeu étant d'articuler les données de ces nouveaux modes de validation aux bases de données habituelles Client-Carte-Validations.

Les **données comportementales** sont parfois assimilées aux données transactionnelles, elles permettent de caractériser notamment les comportements de navigation sur les sites internet et les applications et de prendre en compte l'activité du client sur les réseaux sociaux, les notations de type « likes » ou « étoiles »[8] ainsi que les réactions aux campagnes marketing (e-mails) ou enquêtes de satisfaction.

Ainsi, on peut déterminer quelles sont les lignes de transport et arrêts pour lesquels un client-voyageur effectue des recherches sur le site web du réseau (il est même souvent possible pour lui d'indiquer directement ses lignes et arrêts comme favoris pour recevoir des notifications spécifiques, de trafic par exemple) et capter ses habitudes d'achats de titres de transport (sur place, sur le site internet, via l'application…).

Certains services de mobilité tels que les voitures de transport avec chauffeur (VTC) comme Uber ou les lignes d'autocars longue distance (Isilines, Ouibus) offrent la possibilité de noter la course. Sont également inclus dans les données comportementales, les canaux privilégiés d'échange entre le client

[6] https://www.twisto.fr/titres-et-abonnements/titres/ticket-sms.html (les mobilités de Caen la mer), consulté le 15/05/2019.
https://www.optymo.fr/pass-tickets/ticket-sms/ (Territoire de Belfort), consulté le 15/05/2019.
[7] Navigo lab, l'application pour acheter et valider les titres de transport sur smartphone, publié le 25/09/2018, https://www.digital.sncf.com/actualites/navigo-lab-lapplication-pour-acheter-et-valider-les-titres-de-transport-sur-smartphone, consulté le 15/05/2019.
Le passe Navigo sur smartphone testé à l'automne 2018, publié le 01/10/2018, https://www.iledefrance.fr/le-passe-navigo-sur-smartphone-teste-lautomne-2018, consulté le 15/05/2019.
[8] De 1 à 5 étoiles par exemple.

et le réseau de transport. Chez Transdev, il existe par exemple l'application Listen qui permet aux usagers des réseaux de faire des réclamations via les canaux suivants : courrier, e-mail, site internet, téléphone.

Les **données contextuelles**, quant à elles, sont plus difficiles à cerner dans l'ensemble des données clients dans un réseau de transport. Elles sont rattachées moins directement au client et proviennent généralement d'enrichissements de données pour justement contextualiser, dans notre cas un trajet ou des habitudes de déplacement du client dans l'agglomération par exemple. Les données météo en sont un exemple de même que les données de géolocalisation des départs et des arrivées des différents trajets ou encore la période de la journée (heures de pointe, nuit…).

Nous le voyons, c'est l'ensemble de ces catégories de données qui permettent de caractériser un client le plus largement possible.
La multiplicité de l'information caractérisant un client, portée par des données, ainsi que la multiplication des canaux et des points de contacts entrainent une grande complexité de la gestion de ces données. Tout l'enjeu étant de réussir l'intégration de l'ensemble de ces données et de faire converger des sources différentes en respectant des règles différentes.

L'enjeu de ce mémoire est finalement d'apporter des éléments de réponse à la question : « Comment parvenir alors à réconcilier ce patrimoine de données clients hétérogènes pour acquérir une **vue client unique**, **une vue 360°**, une seule version de la réalité ? » (Berti-Equille, 2012). Une des réponses pouvant être apportée est la création d'un **référentiel client unique (RCU)** qui est un « véritable projet stratégique d'entreprise, et dont la garantie de succès passe indubitablement par la prise en compte des aspects de qualité et de gouvernance des données. » (Berti-Equille, 2012).

1.1.3. Les données clients comme données de référence pour l'entreprise

Après avoir observé plus en détail les données clients dans un réseau de transport et leurs caractéristiques, nous voulons dans cette partie présenter le concept de donnée de référence. Donnée de référence ou Master Data[9] en anglais renvoie notamment au Master Data Management (MDM). C'est une notion assez vaste qui peut se résumer comme étant une « démarche qui s'intéresse aux données de référence afin de garantir une gestion unifiée et sous la responsabilité des métiers » (Bonnet, 2009).
On retrouve notamment deux caractéristiques du MDM :

- C'est un entrepôt de données (ou Data Warehouse) permettant de gérer les versions et les droits d'accès. Les données de référence ainsi stockées et centralisées dans l'entrepôt de données sont redistribuées aux différentes applications permettant d'éviter les incohérences et les conflits. Cette centralisation des données de référence et de leurs règles de gestion permet de n'avoir qu'une seule version de la vérité (*a single version of the truth*) (Mariko, 2016).
- Il s'appuie sur un modèle, partagé dans l'ensemble de l'organisation, décrivant les données de référence en précisant les différentes relations, les significations et les règles de mise à jour.

[9] On parle aussi en français de données maîtres.

Nous allons présenter dans cette partie en quoi les données clients au sens large peuvent être qualifiées de données de référence dans une entreprise. Cette compréhension des données de référence nous permettra par la suite de poser les bases de la réflexion sur la gouvernance de la donnée pour la consolidation d'une base clients notamment à travers le Référentiel Client Unique (RCU).

On ne trouve pas dans la littérature de définition standard de la donnée de référence mais l'on retrouve dans plusieurs sources différentes caractéristiques et situations qui permettent de définir une donnée de référence.

Pour commencer, une donnée de référence est une **donnée métier** comportant une information de base, fondamentale pour l'activité de l'entreprise (Trigaux, 2009). Elle **s'oppose aux données de transactions**.

Classiquement, voici quelques exemples de données de référence que l'on peut retrouver comme dans une organisation (Loshin, 2010) :

- Customers
- Employees
- Vendors
- Suppliers
- Parts
- Products
- Locations
- Contact mechanisms
- Profiles
- Accounting items
- Contracts
- Policies

C'est à l'entreprise de prioriser ses « objets métiers » qu'elle souhaite voir considérer comme des données de références car la gestion des données de référence demande une certaine implication et peut entrainer des coûts supplémentaires en tant qu'investissement.

Une autre caractéristique d'une donnée de référence est son caractère « **dupliqué** » au sein de plusieurs systèmes (Bonnet, 2009). Dans l'exemple d'un client d'un réseau de transport, selon les points de contact nous pouvons retrouver une partie de l'information qui le caractérise soutenue par des données stockées à plusieurs endroits (par exemple, le numéro de téléphone est présent dans la base Réclamations ainsi que dans le système qui gère la carte de validation). Ce caractère de duplication génère de la complexité quant aux contraintes d'intégrité reliant les données entre elles dans les différents systèmes. En effet, la duplication d'une donnée a pour conséquence la duplication de ses règles de validation et il est nécessaire de synchroniser les mises à jour dans les différents systèmes puisque la donnée doit disposer de la même valeur dans tous les endroits où celle-ci est stockée (Bonnet, 2009).

Enfin, comme troisième caractéristique, une donnée de référence est une **donnée échangée avec des tiers**. Cette donnée qui se partage entre différents systèmes nécessite une traçabilité puisqu'il faut être en mesure de prouver la véracité de la donnée entre le système émetteur et le système récepteur. Un référentiel des échanges entre les différents systèmes est donc à prévoir.

La valeur des données de référence ne dépend donc pas de l'exécution des transactions soutenant l'activité de l'entreprise. Ces données n'en sont pas moins stratégiques puisqu'elles sous-tendent et permettent ces transactions. Les données de références sont finalement au cœur d'une chaîne de valeur et si elles sont erronées, c'est toute cette chaîne de valeur qui peut être mise en défaut (Bonnet, 2009). Nous avons pu voir dans les « analyses clients » réalisées aujourd'hui par les réseaux de transport l'importance de la billettique (données transactionnelles) permettant de rattacher un client à ses trajets par sa carte d'abonnement et son numéro d'identifiant. Si de plus, on cherche à augmenter la Connaissance Client en prenant en compte et en augmentant les points de contact entre le client et le réseau de transport (*cf*. Graphique 1), nous imaginons bien la nécessiter de traiter et de considérer l'ensemble des données clients comme des données de référence.

Aujourd'hui encore, dans beaucoup d'organisations, et nous le verrons par la suite avec l'exemple du réseau de transport du Havre, les données de références « sont traitées avec des outils et des processus de gestion hétérogènes, qui conduisent à des problèmes importants de qualité et de traçabilité » (Bonnet, 2009).
Cette notion de données de référence permet donc de mettre en avant dès à présent l'importance de la qualité des données, de leur gestion ainsi que l'organisation à mettre en place pour assurer leur intégrité et permettre de consolider au mieux une base clients.

Ajoutons également qu'étant considérée comme stratégique, la gestion des données de références ne concerne pas uniquement la direction informatique ou les systèmes d'information. Si ces derniers auront bien entendu en charge de gérer la partie technique (notamment l'architecture), les équipes métiers doivent participer notamment en amont à la cartographie du cycle de vie. Ce cycle de vie se définissant généralement autour des fonctions suivantes (Mariko, 2016) :

- La découverte et le profilage des données
- L'acquisition et l'intégration des données
- La maintenance des données
- L'usage des données
- L'archivage et la destruction des données

1.1.4. Une base clients pour alimenter le CRM (« Customer Relationship Management »)

Nous présentons dans cette partie l'importance de la consolidation d'une base clients notamment dans le cadre d'un réseau de transport. Dans quel but réaliser cette consolidation et quelles applications au sens large peut-on en tirer ? La principale application que nous voulons présenter est le CRM puisqu'il est le point de convergence entre l'entreprise et le client au moyen des données, puis nous évoquerons un ensemble de pratiques et d'analyses rendues possible et facilitées par la consolidation d'une base clients et qui alimenteront finalement le CRM.

Nous souhaitons évoquer plus largement le CRM dans la mesure où il est l'un des principaux aboutissements de tout le travail sur les données clients (gouvernance, management, point de départ de la relation client à l'heure du cross-canal)[10]. Et finalement, lorsqu'on parle de consolidation d'une base clients, celle-ci se comprend notamment chez Transdev comme alimentant le CRM (ce dernier faisant l'objet d'une réflexion à l'échelle du Groupe).

1.1.4.1.Customer Relationship Management (CRM)

L'ensemble de la « Connaissance Client » évoquée plus haut et ce profilage client (avec les enjeux techniques, marketings, de données et organisationnels, au sens large qui y sont liés) s'intègrent dans ce qu'on appelle l'approche « Customer Relationship Management » (CRM) qui « désigne l'ensemble des stratégies, outils et techniques qui permettent d'enregistrer, de gérer et d'enrichir les relations avec les clients – actuels, voire même les anciens à reconquérir – et les prospects » (Delers, 2015). Nous retrouvons cette notion marketing de cycle de vie client (clients prospects, clients actuels, clients à reconquérir) que nous avons présentée avec les différents points de contact entre le client et le réseau de transport.

De nombreuses entreprises ont intégré cette approche dans leur relation client. Et l'ensemble des principales entreprises de transport public en France ont développé ou sont en train de développer des solutions CRM[11].

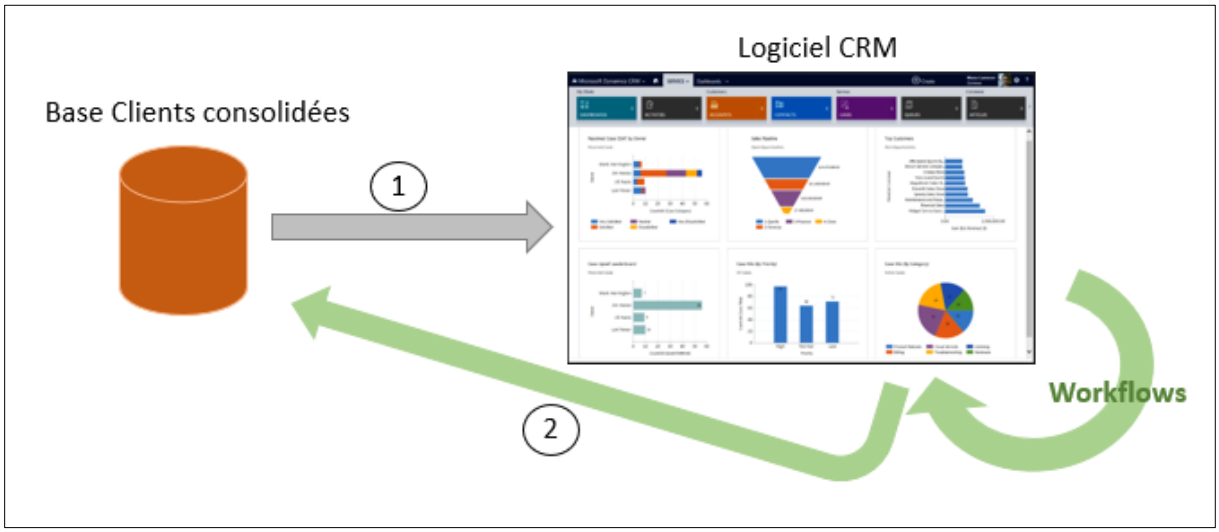
Le **CRM** est avant tout une **démarche** avant d'être un outil et finalement un **logiciel** où la place des données sur les différents clients est primordiale. De manière générale, on peut voir le CRM comme un outil qui « collecte, stocke et active les données personnelles dans une optique de fidélisation » (Hirth, 2017). On parle aujourd'hui, du fait de la multicanalité et de la crosscanalité (Belvaux et Notebaert, 2018), de CRM 360°, ce qui renvoi à l'idée d'avoir une vue unique et « panoramique » du client (ces habitudes, ces échanges…), ce qui nécessite la construction d'un Référentiel Client Unique (RCU).

D'un point de vue pratique, il existe une multitude de solutions CRM sur le marché en tant qu'application logicielle qui cherchent à être le plus ergonomique possible et surtout utilisables et accessibles aux différents employés en charge de la relation client et du marketing.
L'utilisation « marketing » du CRM en bout de chaîne se fait donc indépendamment de l'architecture fonctionnelle. Cette dernière est étudiée et mise en place en amont. D'une manière générale, le CRM

[10] Un autre travail sur les données clients est l'ingénierie tarifaire.
[11] RATP Dev se lance dans le marketing relationnel, Florence Guernalec, 02/10/2015, http://www.mobilicites.com/011-4190-RATP-Dev-se-lance-dans-le-marketing-relationnel.htmlhttps://www.keolis.com/fr/nos-expertises/experience-client/relation-clients, consulté le 28/04/2019 ;
Transdev ou l'obsession du client, Laure Trehorel, 19/12/2016, https://www.relationclientmag.fr/Thematique/strategies-1255/Breves/Transdev-obsession-client-311883.htm, consulté le 28/04/2019 ;
Keolis : un outil commun de CRM pour les filiales, Aurélie Chandèze, 18/05/2015, Revue / Numéro : Best Practices SI n°151, https://www.bestpractices-si.fr/publications/retour-d-experience/keolis-un-outil-commun-de-crm-pour-les-filiales, consulté le 28/04/2019.

est le plus souvent lié à un ERP (Enterprise Ressource Planning[12]) de l'entreprise permettant de proposer une base de données unique et complète (Delers, 2015). Une autre approche, et c'est l'objet de ce mémoire, est de connecter le CRM avec une base clients consolidée (construite à l'aide d'un Référentiel Client Unique) en amont, comme présenté dans le Graphique 4 :

Graphique 4 : Flux de données entre la Base clients consolidées et le CRM



Dans ce schéma, nous avons ajouté la boucle **Workflows**. Celle-ci représente l'ensemble des actions réalisées par les salariés utilisant le logiciel CRM concernant l'ensemble des services qu'ils proposent aux clients (Blondeau et al., 2015), ainsi que l'ensemble des règles de déclenchement d'actions automatiques et de mises à jour de la page du client en réponse aux différentes actions (ouverture ou non de l'e-mail envoyé, message automatique de travaux sur une ligne de bus emprunté par le client…). Ce Workflows est un apport continu d'informations sur la relation client et alimente de manière continue le CRM. Certaines de ces informations portées par des données peuvent également alimenter la base clients (par exemple : la mise à jour d'une adresse par un salarié directement dans le logiciel CRM). Cela est représenté par la flèche numéro 2 sur le schéma précédent. Le CRM devient alors, au même titre que d'autres applications dans l'entreprise, une source de données alimentant la base clients.

1.1.4.2.Valoriser la relation client

Une fois la base clients consolidée et la solution CRM mise en place, il devient alors possible de pousser de véritables offres marketing personnalisées, de mieux comprendre et segmenter les différents clients du réseau et d'adopter une approche « expérience client ».

[12] ERP ou Progiciel de Gestion Intégré (PGI) en français, est un progiciel permettant de gérer l'ensemble des processus d'une organisation en intégrant l'ensemble de ces différentes fonctions (comme les ressources humaines, la comptabilité, la finance, les vente ou la distribution…).

L'expérience client

Avec le CRM, il s'agit bien de relation avec le client dans une optique crosscanal. Un exemple typique de cas d'usage du CRM, est l'envoi d'un e-mail personnalisé à la suite de différentes actions du client (recherche d'informations, d'horaires sur une ligne précise, comportement sur le site internet du réseau de transport). Finalement, le CRM, la « relation client » et la « Connaissance Client » au sens large s'intègrent dans ce qu'on appelle **l'expérience client**. Pour Kéolis, opérateur de transport public et filiale de la SNCF, cette expérience client se décline selon les axes suivants[13] :

- Connaissance client
- Conception des réseaux
- Tarification
- Parcours client
- Relation clients
- Promotion du réseau

Comme pour les différents points de contact entre le client et le réseau de transport, ces différents axes de l'expérience client génèrent des données notamment à l'échelle du client.
Les préférences (de prise de contact entre le réseau et le client), tout comme les mesures de satisfactions sont des informations qui s'intègrent pleinement au CRM qui permettent une relation personnalisée avec le client-voyageur.

D'une manière générale, l'expérience client cherche à mettre au cœur le client-voyageur dans les actions du réseau de transport en personnalisant son « expérience mobilité » physique (dans les modes de transport[14]) tout comme numérique (préparer son trajet en ligne, noter une course…). Le groupe Transdev s'inscrit également dans cette démarche en mettant notamment en place une direction Clients Voyageurs transversale au sein du Groupe, en promouvant le concept de MaaS (Mobility as a Service[15]) ou en continuant d'assembler et d'installer au sein des réseaux de transport des outils CRM[16].

Segmentation de la clientèle pour une information personnalisée et un marketing ciblé

Tout l'intérêt de la mise en place d'un CRM est de pouvoir segmenter, diviser sa clientèle en groupes distincts. Une fois les segmentations élaborées, il est alors possible par exemple de lancer des campagnes marketing ciblées selon les groupes de clients ou de donner des informations pertinentes sur le réseau de transport (offres, services…). Dans la partie II, avec l'exemple du réseau du Havre, nous verrons qu'avec l'aide de la consolidation des différentes bases de données, nous avons pu mettre en avant une typologie de population utilisant un service spécifique de transport (le transport à la demande la nuit), à savoir des personnes âgées de 18 à 25 ans.

[13] https://www.keolis.com/fr/nos-expertises/experience-client , consulté le 17/05/2019.
[14] Offre de services, accès au wifi, information voyageurs…
[15] La Mobility as a Service (MaaS) décrit une nouvelle approche des modes de transport et de la mobilité tendant à faire de ceux-ci des solutions de mobilité consommées en tant que service pour le voyageur.
[16] https://www.transdev.com/fr/innovations/experience-client/ , consulté le 17/05/2019.

Il existe de nombreuses façons de segmenter les clients. Nous ne les détaillons pas ici, mais insistons sur quatre contraintes permettant de réussir au mieux une segmentation dans une logique de marketing (Hirth, 2017) :

- coller à un objectif
- avoir un fort pouvoir discriminant
- avoir un sens métier
- pouvoir être exploité facilement par le marketing

Si certaines techniques de segmentation sont plus techniques et demandent certaines connaissances (en Data Science par exemple), il est tout de même toujours envisageable de réaliser des segmentations simples par une équipe marketing.

Voici trois exemples de segmentation selon leur ordre de complexité :

- Segmentation selon un seuil sur les variables. Observation statistique des variables caractérisant les voyageurs (âge, sexe, nombres de trajets…)
- Segmentation RFM très utilisée en marketing (Récence, Fréquence, Montant)
- Une autre approche de segmentation consiste à utiliser des algorithmes d'exploration de données de type « clustering » (regroupement automatique d'individus au sein d'un certain nombre de classes a priori inconnues) ou encore de type « classification » (attribution automatique d'un individu à une classe pré-existante) (Bellot et Espinasse, 2017).

Ajoutons pour conclure sur la segmentation client, qu'il est nécessaire de travailler sur une base clients (ou depuis le CRM) possédant une forte qualité de la donnée (au sens large).

Enrichissement de données : exemple du géomarketing

Il est aussi envisageable de faire l'enrichissement de données concernant la base clients. Nous avons beaucoup travaillé sur le géocodage des adresses des clients. Le géocodage d'adresses, effectuée à l'aide d'une API[17], consiste à déterminer la longitude et la latitude afin de pouvoir travailler de façon spatialisée et cartographiée. On peut ainsi rapprocher l'adresse d'un client d'un arrêt de bus ou de tramway et ainsi pouvoir pousser une offre de géomarketing ciblée (présenter l'arrivée d'un nouvel arrêt de bus, message personnalisé : « le centre-ville à 15 min avec la ligne de bus numéro 12 ») ou faire de « l'information voyageur » au plus proche du client (« l'arrêt République sera fermé pour cause de travaux du 01/08/2019 au 31/08/2019 »).

Nous travaillons avec l'API reliée à la Base Adresse Nationale[18] (BAN) qui est une base de données référençant l'intégralité des adresses du territoire français qui est constituée et gérée par la collaboration entre Etalab, La Poste, l'IGN, la DGFiP et OpenStreetMap France.

La précision et l'exactitude des adresses des clients sont essentielles pour ne pas « chaîner » les erreurs (adresse inexacte, mauvais calcul longitude/latitude, travail cartographique faussé, géomarketing inopérant).

[17] Utilisation de l'API (Application Programming Interface ou Interface de Programmation d'Application) : https://adresse.data.gouv.fr/api.
[18] https://www.data.gouv.fr/fr/datasets/base-adresse-nationale/

1.2. Eléments de gouvernance de ces données clients

Après avoir décrit comment un réseau de transport rentre en contact avec ses clients, mis en avant le caractère stratégique des données clients en tant que données de référence et présenté ce qui peut être fait avec la consolidation d'une base clients, nous voulons présenter dans cette partie les questions que soulèvent cette consolidation d'une base clients, notamment en termes de gouvernance de la donnée.

La gouvernance des données a été définit par l'administrateur général des données de l'Etat (SGMAP) comme « l'ensemble de principes et de pratiques qui visent à assurer la meilleure exploitation du potentiel des données » (Infolab-Fing, 2017). Ces principes et pratiques au sens large se doivent d'être partagés.
Tout l'enjeu est de développer une vision transverse du client en partant de l'existant. En effet, la dimension historique est très importante dans la mesure où il ne s'agit nullement de « faire table rase du passé » et des pratiques qui ont cours dans l'organisation mais de partir du contexte actuel. Chaque organisation (chaque réseau de transport) a déjà un certain savoir-faire et a déjà acquis une certaine connaissance client. Cependant, cette connaissance client est encore trop souvent « stockée » dans de nombreux silos (différents systèmes d'information) auxquels sont rattachés de nombreux référentiels qu'il est difficile de faire converger, rendant ainsi difficile les échanges entre applications (Cigref, 2016). C'est dans cette logique que nous avons présenté précédemment les données de références (Master Data). C'est le management de ces dernières[19] ou MDM (« Master Data Management ») qui vise à obtenir une **information unique** et **partagée** au sein de l'organisation.

Nous allons commencer par évoquer le Référentiel Client Unique (RCU) qui se rapproche de la notion de MDM est qui au cœur de notre problématique de consolidation d'une base clients. Dans un second temps, nous aborderons les enjeux de la qualité de la donnée. Et pour finir, nous évoquerons les questions juridiques que la consolidation et le stockage de données clients posent notamment en termes de protection des données personnelles.

1.2.1. Construction d'un Référentiel Client Unique

Que cela soit pour le CRM, ainsi que dans l'objectif d'avoir une vue 360° du client mais aussi pour les différentes « applications » rendues possibles à la suite de la consolidation d'une base clients, il est nécessaire de définir en amont quelles informations, c'est-à-dire quels champs dans les différentes bases de données permettront d'identifier clairement un client. C'est en cela que constitue la construction d'un **Référentiel Client Unique (RCU)** qui « correspond à la situation ou une entreprise réussie à alimenter une seule et unique BDD client, quels que soient les canaux de contact entrants ou sortants ayant permis de collecter une information relative au client »[20].

Cette base de données unique qui est modélisée à partir du RCU, est une réponse à l'organisation hétérogène des données dans les entreprises, le plus souvent en silos où les données sont

[19] Dans notre exemple, ce sont les données clients comme données de références.
[20] Article « Définition : Référentiel client unique », Bertrand Bathelot, https://www.definitions-marketing.com/definition/referentiel-client-unique/ , consulté le 20/05/2019.

fréquemment redondantes voir divergentes. Elle constitue un véritable socle de la gestion des données multi-sources permettant une bonne Connaissance Client, une vue unique, une seule version de la vérité (Berti-Equille, 2012).

A première vue, cette construction du RCU peut paraitre relativement simple dans la mesure où l'on pourrait penser qu'il suffise de remplir et de garder les champs tels que « nom », « prénom », « adresse », « numéro de téléphone » etc. comme information et faire en sorte de retrouver ces champs dans les différentes bases de données. Cependant, et nous le verrons concrètement dans l'exemple du Havre *(cf.* partie II), du fait de la multiplication des applications et des offres de service, il n'est pas si évident de retrouver l'ensemble de ces champs pouvant caractériser un client et finalement de l'identifier de façon « unique et vraie ».
Par exemple, avec le service d'achat d'un titre unitaire de transport par SMS, seule l'information « numéro de téléphone » est captée. Il n'y a donc, dans la base de données « SMS », en plus des informations sur le titre de transport (prix, durée…) qu'une seule information sur le client, son numéro de téléphone portable. Si cette information n'est pas présente dans la base générale Clients, il ne sera pas possible de faire une jointure et de constituer une vue d'ensemble des activités du client sur le réseau de transport.
La réflexion sur les **choix des champs** dans les différentes applications est donc une phase essentielle dans la constitution du RCU.

De plus, il ne s'agit pas seulement que l'information soit présente mais aussi que celle-ci **respecte des règles** et notamment en termes de qualité des données (*cf.* partie 1.2.2). Par exemple, si dans une application sont présentes les informations du « nom » et du « prénom », il est possible que dans une autre application ces informations existent également mais sous une autre forme (avec des accents, en majuscule, avec seulement la première lettre du prénom mais aussi avec une autre orthographe), rendant difficile le recoupement entre bases de données et le rapprochement des différentes sources de données pour permettre d'établir le RCU.

Il n'est pas toujours possible de respecter scrupuleusement ces règles entre différentes applications notamment en ce qui concerne les applications déjà existantes et pour lesquelles il serait trop lourd et trop risqué de modifier la structuration des champs. Des règles de type « ETL[21] » seront alors mises en place pour alimenter la base clients en respectant le RCU.

Cependant, à chaque nouvelle application ou lors du remplacement d'un logiciel par un autre (pour la même fonction), il sera nécessaire d'essayer de se rapprocher le plus possible du RCU qui aura été établi (en obligeant d'ajouter des champs, en permettant de via une BDD secondaire de rattacher à un client unique par le numéro de téléphone par exemple).

Dans les différentes règles rattachées au RCU, celles de **mise à jour** constituent un point important notamment dans le management des données de références (MDM), puisque la logique de ce dernier est d'assurer à tout moment la véracité des données. En effet, dans le RCU, il y a ce qui constitue les données fondamentales du client mais nous l'avons vu, il est possible de calculer des scores, de géocoder des adresses, de regrouper des clients (cluster) etc., tout cela apportant de l'information

[21] ETL pour Extract-Transform-Load.

supplémentaires « portées » par des données « secondaires » qui vont alimenter différentes applications (notamment le CRM). Selon les besoins métier, ces règles de mise à jour devront être explicitées dans les différents flux et processus et être intégrées à l'architecture.

La construction du RCU nécessite donc une réelle gouvernance et comme évoqué pour les données de références, elle doit se faire également conjointement entre les équipes métiers et les équipes informatiques pour permettre d'établir des règles (« humaines » et « automatiques » ou informatiques)

Ajoutons un dernier élément de complexité, celui de **l'échelle de partage de ce RCU** qui est lié à l'organisation de l'entreprise. En effet, une entreprise peut être considérée comme une seule et unique entité ou peut être constituée de plusieurs sites géographiques, de zones ou divisions mais aussi de filiales. Selon le degré d'intégration entre ces subdivisions, leur état de maturité en matière de données ainsi que leur intérêt dans la mise en œuvre du RCU, se posera la question de l'échelle de ce RCU, du partage d'un RCU commun. Nous présenterons plus loin l'organisation du Groupe Transdev.

Une fois le RCU établi, se posera alors pour les équipes informatiques (IT) les **questions d'architecture**. Nous ne développerons pas ici cette question de l'architecture qui est essentiellement de l'ordre technique. Cependant, il est nécessaire de comprendre quelques fondements d'architecture, tirés du management des données de références (MDM).

Pour la mise en œuvre d'un référentiel (et dans notre cas un Référentiel Client Unique), il est question en termes d'architecture de créer un « **point focal au sein du SI** ». Ce point permet de garantir la validité des informations portées par le référentiel, devenant ainsi un « **point de vérité** » disponible auquel viennent se greffer les différents processus consommateurs, les différentes applications (Régnier-Pécastaing et al., 2008).

Dans la « chaîne SI », ce point focal se situe entre un amont et un aval. L'amont constitue l'ensemble des processus transformant une donnée entrante en donnée valide, ce sont les « **points d'acquisition** » (Régnier-Pécastaing et al., 2008). Par exemple, saisir une adresse d'un client dans une application est un point d'acquisition qui touche à la fois les équipes métiers (comme le marketing) puisque ce sont elles qui saisissent la donnée (selon des règles définies) et la partie informatique (processus récoltant la donnée, transformation, stockage, vérifications de règles, synchronisation avec d'autres applications…).

L'aval quant à lui ne concerne plutôt que les équipes informatiques, puisque cela englobe l'ensemble des services de consultation et de diffusion des données pour l'ensemble des processus consommateurs dans l'organisation. Les équipes informatiques doivent mettre en place les processus pour la « **consommation** ». L'expression des besoins des équipes métiers est importante pour l'établissement de ces processus (fraicheur de la donnée et rafraichissement, synchronisation, déclenchement de processus à une temporalité précise…).

Dans l'ensemble, les équipes SI choisiront une architecture permettant d'articuler au mieux la **relation « amont – point de vérité – aval »** en prenant en compte le nombre de points d'acquisition[22] (l'amont) et le nombre d'applications et leurs interdépendances entre elles (l'aval) tout en s'appuyant sur l'architecture plus globale de l'entreprise déjà en place.

Ce stockage du « point de vérité », du Référentiel Client Unique s'effectue donc dans une base unique pouvant être un entrepôt de données (Data Warehouse) (Mariko, 2016) auquel s'intègrent les différentes applications de l'amont et l'aval.

1.2.2. Qualité des données

Comme nous avons commencé à le voir, les données peuvent être considérées comme une ressource, un actif portant une certaine valeur puisqu'elles jouent un rôle croissant dans les choix stratégiques des entreprises et dans les processus opérationnels. En ce qui concerne les données clients, nous avons évoqué l'importance de ces données pour l'étude des habitudes de voyages, le marketing ainsi que pour l'expérience client. Cette valorisation est étroitement liée à la qualité de la donnée.

De nombreuses dimensions concernant la qualité des données ont été recensées dans la littérature. Cinq d'entre-elles sont unanimement reconnues aujourd'hui (Berti-Equille, 2018) :

- La complétude des données
- La cohérence des données
- L'unicité des données
- L'exactitude et la véracité des données
- La fraicheur des données

La **complétude** des données fait référence à la quantité d'information renseignée dans les différents champs d'une base de données.

La **cohérence** des données révèle si les données respectent les règles métiers et les différentes contraintes prédéfinies (par exemple, dans le domaine des transports publics, le nombre de validation par jour ne peut pas dépasser un certain seuil ; 100 validations en une journée pour une seule personne indiquerait une donnée non-cohérente).

L'**unicité** des données renseigne sur l'absence de doublons dans une base de données.

L'**exactitude** et la **véracité** des données indique la validité et l'absence d'erreur par rapport à la vérité.

La **fraîcheur** des données renseigne si les données sont correctement à jour et si la validité temporelle est vérifiée.

Le domaine de la qualité de la donnée est un domaine très vaste, puisque cela concerne à la fois des enjeux techniques, d'organisation et de traitement. C'est un domaine transverse à l'organisation.

[22] Processus où l'on rentre de la donnée.

L'amélioration de la qualité des données est donc à considérer comme une démarche en continue. Il existe plusieurs approches pour améliorer la qualité des données. Retenons deux approches en lien notamment avec notre problématique de la consolidation d'une base clients : l'approche « processus » et l'approche « nettoyage » (ou data cleansing) (Régnier-Pécastaing et al., 2008).

La première **approche, « processus »**, a pour but de prémunir les systèmes d'information de l'introduction de données erronées. On parle de processus car il est question de l'ensemble de la chaîne de traitement des données. L'objectif est donc de mettre en place des procédés pour empêcher de faire circuler le long de cette chaîne des données erronées. Par exemple, des formats de saisi de champs (code postal, ville, date…) dans des logiciels et applications peuvent être proposé dans une liste pour éviter les erreurs textuelles et numériques notamment lors de la saisie (lors de la création d'une carte d'abonnement dans le réseau de transport au guichet, ou lors de la création d'un compte par le client sur le site internet). Cette approche a donc un caractère préventif à la différence de la seconde.

En effet, **l'approche « nettoyage »** (cleansing) arrive plutôt en bout de chaîne. La donnée est déjà saisie ou présente. Ce « nettoyage » consiste essentiellement en une correction des données à l'aide de différents outils (code à la main, règles automatiques pré-codées, logiciels...). Ce « nettoyage » fait partie intégrante notamment des processus ETL[23] avant le chargement des données (Load) par exemple dans un Data Warehouse (Régnier-Pécastaing et al., 2008). Ce « nettoyage » peut être programmé de façon périodique ou être mis en œuvre avant la migration de données existantes de qualité médiocres vers un nouveau système.

Enfin, pour notre sujet de la consolidation d'une base clients, se pose particulièrement la problématique de la « **résolution d'identité** » pour les personnes physiques, c'est-à-dire la mise en place de démarches de qualité pour le traitement des données d'individus, de clients.
Des rapprochements, des recoupements erronés entre deux individus peuvent avoir des conséquences négatives aussi bien pour les personnes concernées que pour l'organisation. Pour la personne car elle peut se voir affecter de fausses activités et pour l'organisation, cela peut engendrer des dysfonctionnements et entrainer des coûts pour chercher et corriger les erreurs.

Une des difficultés pour la résolution et le rapprochement d'identité provient de la multitude d'anomalies qui peuvent affecter les données saisies concernant les personnes physiques (plusieurs possibilités de saisie d'une même information, abréviations, codifications (code postal, pays…), ordre des mots (nom prénom, prénom nom, adresse…) (Berti-Equille, 2012). Le tableau suivant résume différents types d'anomalies pouvant affecter les données :

[23] Extract Transform Load.

Tableau 2 : Anomalies liées aux identités

Type d'anomalie	Exemple
variantes de saisie	
erreurs de saisie ou de transcription	
utilisation d'abréviations	avenue ou av. ou ave
utilisation des initiales	Henri Martin ou H. Martin
utilisation de surnoms	Bill pour William, Bob pour Robert
permutation de mots	rue Henri Martin ou rue Martin Henri
permutation de lettres	rue Herni Martin
fautes d'orthographe	
données non-nettoyées	
différences de structure, de format et de localisation des données	
omissions : données non renseignées ou imparfaitement	
doublons	
des différences de langues	Genf et Genève ou jj/mm/aaaa et mm/jj/aaaa
des différences syntaxiques	ordre des mots
des codifications différentes	codification pays sur 3 car ou sur 2 car par exemple Afghanistan AFG ou AF, Afrique du sud ZAF ou ZA

Source : tableau "Les anomalies" (Chapitre 2, Les critères pour la résolution d'identité appliqués aux personnes physiques, page 62) de BERTI-EQUILLE, Laure. La qualité et la gouvernance des données : au service de la performance des entreprises. Paris : Hermes/Lavoisier, 2012.

Ajoutons que pour la résolution d'identité, un autre enjeu de la qualité des données est à prendre en compte : la problématique des homonymes. Des personnes pouvant avoir le même nom et prénom, ces deux informations se sont donc pas suffisantes pour identifier de façon exacte.

1.2.3. Protection des données personnelles

Pour toute gestion d'une base clients et dans notre cas lors de la consolidation d'une telle base dans l'optique par exemple d'alimenter un CRM, se posent des questions de conformité en ce qui concerne notamment la protection des données personnelles. Cette problématique de la protection des données personnelles était déjà prise en compte dans la loi informatique et liberté du 6 janvier 1978. Le Règlement européen dit Règlement Général sur la protection des données personnelles (RGPD)[24] adopté le 25 mai 2016 constitue un nouveau cadre pour la protection des données personnelles en Europe et en France. Ce texte a été transposé en France via une modification de la loi informatique et liberté.

Le RGPD comprend trois points principaux :

- Les personnes concernées par la collecte de données personnelles les concernant doivent être auparavant informé de la finalité de cette collecte
- Pour cette collecte, nécessité de recueillir le consentement (démarche volontaire et active : opt-in)

[24] Règlement 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.

- Les personnes concernées doivent pouvoir exprimer un droit d'accès, de rectification, d'opposition, de portabilité et de suppression des données les concernant.

Rappelons ce qu'on entend par « **donnée à caractère personnel** » et par **« traitement des données à caractère personnel** ».

- Une **donnée à caractère personnel** selon l'article 4 du RGPD est définie comme « toute information se rapportant à une personne physique identifiée ou identifiable ; une personne physique identifiable est celle qui peut être identifiée directement ou indirectement, notamment par référence à un numéro d'identification, à des données de localisation, à un identifiant en ligne ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale ».
  Il s'agit concrètement de l'ensemble des données qui permettent de réidentifier une personne grâce à l'ensemble des moyens disponibles (comme le croisement de plusieurs sources). Les données personnelles visées par le RGPD concernent donc d'une part des données personnelles qui permettent une identification directe (données nominatives) mais aussi des données autres qui permettent une identification indirecte, i.e. les données qui sont croisées avec d'autres, parfois mêmes externes à l'entreprise, permettent l'identification d'une personne[25].
  Sans toutes les lister, on retrouve les données à caractère personnel suivante dans le domaine de la mobilité (Arroyo et al., 2018) :

Tableau 3 : Données à caractère personnel dans le domaine de la mobilité

Etat civil	Nom, prénom, genre, date de naissance, adresse physique
Vie personnelle	Habitude de mobilité sur le territoire (adresses récurrentes)
Comportements	Réseaux sociaux (modes de transport), notations et « likes », navigation web (recherche d'itinéraire…), consommation en ligne (achats de titres de transport, réservation Transport à la Demande (TAD)[26]
Localisation	Localisation des arrêts de transport utilisés, GPS, itinéraire
Identifiants numériques	Login, adresse IP, adresse e-mail

- Pour le RGPD, un **traitement de donnée** est « toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données ou des

[25] J. P. Arroyo, S. Brunet et R. Sage dans leur ouvrage « RGPD et Marketing. De la contrainte à l'opportunité » (E-théque, 2018), donnent les exemples suivants :
- une plaque d'immatriculation permet de trouver le nom et l'adresse d'un citoyen, par le biais du fichier des cartes grises ;
- une adresse IP permet de remonter jusqu'à l'adresse de l'utilisateur en passant par le fichier de l'opérateur.

[26] Exemple inspiré de (Arroyo et al., 2018) citant des situations permettant d'identifier des personnes : « Comportements : activité messagerie, réseaux sociaux, « likes », recherches web, consommation en ligne, pétitions… » (page 34).

ensembles de données à caractère personnel, telles que la collecte, l'enregistrement, l'organisation, la structuration, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, la diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, la limitation, l'effacement ou la destruction[27] ».

La problématique du consentement a été beaucoup médiatisée mais celle-ci n'est pas la seule base légale sur laquelle doit reposer tout traitement de données personnelles.
Dans la pratique, trois bases légales sont envisageables pour les entreprises du commerce, dans le cadre de leurs activités commerciales (Arroyo et al., 2018) :

- **l'exécution d'un contrat** souscrit entre le responsable de traitement et la personne concernée ;
- l'intérêt légitime du responsable de traitement ;
- le **consentement** de la personne concernée.

Dans le domaine de la mobilité et des transports en commun, la livraison d'une carte d'abonnement (mensuelle ou annuelle) à une personne physique est typiquement une finalité pour laquelle les traitements de données reposent sur **l'exécution d'un contrat[28]**. Peuvent alors être conservées des données comme le Nom, Prénom, Adresse, Numéro de téléphone, l'adresse mail (Arroyo et al., 2018).

En ce qui concerne **la prospection commerciale**, c'est le principe du **consentement** qui prévaut. Il faut que les personnes concernées, au moment de la collecte de leurs coordonnées (téléphone, e-mail, adresse…), aient explicitement donné leur accord pour être démarchées.
Deux exceptions existent à ce principe[29] :

- Si la personne prospectée est déjà cliente de l'entreprise et si la prospection concerne des produits ou services analogues à ceux déjà fournis par l'entreprise ;
- Si la prospection n'est pas de nature commerciale.

Ainsi, pour un réseau de transport qui propose à des personnes qui ont une carte d'abonnement une nouvelle offre de transport (location de vélos) ou une adaptation de l'offre (nouvelles lignes, fermetures d'arrêts…), c'est le cadre des exceptions qui s'applique ; ainsi que dans le cadre de l'exécution d'un contrat. De même pour de l'information concernant le réseau de transport (changements d'horaires, travaux, déviations d'itinéraire). Dans ces cas de figure, les différents canaux de contact peuvent être utilisé.

Cependant, en ce qui concerne la vente et l'utilisation de titres de transport unitaires (à décompte) qui ne font pas l'objet d'un abonnement, l'exécution du contrat prend fin une fois le trajet effectué. Il n'y a pas en soi d'enjeu de protection des données personnelles pour des tickets papier acheté

[27] Associé au traitement des données, le RGPD définit un responsable de traitement : « la personne physique ou morale, l'autorité publique, le service ou un autre organisme qui, seul ou conjointement avec d'autres, détermine les finalités et les moyens du traitement ». C'est lui qui peut être sanctionné en cas de manquement à ses obligations. Le RGPD définit aussi le sous-traitant comme « la personne physique ou morale, l'autorité publique, le service ou un autre organisme qui traite des données à caractère personnel pour le compte du responsable du traitement ».
[28] L'offre d'un service de mobilité tout le mois (abonnement mensuel) ou toute l'année (abonnement annuel) sur l'ensemble du réseau de transport.
[29] https://www.cnil.fr/fr/la-prospection-commerciale-par-courrier-electronique

anonymement (au conducteur, à un distributeur) mais apparaissent de plus en plus des titres de transport innovants « digitalisés ».
Au Havre par exemple, il est possible d'acheter des titres de transport à l'unité par SMS. Dans ce cas de figure, rentre en jeu une donnée à caractère personnel, le numéro de téléphone portable. Celui-ci ne peut pas être utilisé pour de la prospection commerciale (par exemple proposer des offres de mobilité du réseau de transport), c'est le **consentement de la personne concernée** qui prévaut.

Ce « consentement » préalable et explicite (opt-in) doit être stocké pour garder un enregistrement de la façon dont il a été obtenu en incluant les données de l'individu, la date d'obtention, et le but précis de l'utilisation des données personnelles.

Cette gestion des « consentements » possède un cycle de vie qui doit prendre en compte un certain nombre de droits attribués aux individus quant à ses données personnelles, dans le cadre du RGPD (Mattatia, 2018) :

- Le droit d'accès aux données : selon l'article 15 du RGPD, toutes personnes peut demander si ses données ont fait l'objet d'un traitement, obtenir des informations sur les opérations réalisées sur celles-ci et une copie des données traitées ;
- Le droit à la rectification : selon l'article 16, toutes personnes à droit à demander à compléter ses données ou qu'elles soient mises à jour ;
- Le droit à l'effacement : selon l'article 17, dans certains cas précis, toute personne peut demander à l'effacement de ses données. Ce droit peut s'exercer par exemple, si le traitement est illicite, si la personne retire son consentement nécessaire au traitement ou encore si elle s'oppose à un traitement de ses données.
- Le droit d'opposition : selon l'article 21 du RGPD, toutes personnes peut s'opposer au traitement de ses données par exemple à des fins de prospections commerciales. Mais ce droit n'est pas total dans la mesure où il ne peut s'appliquer de manière automatique.

1.3. Le contexte de la donnée chez Transdev, une organisation décentralisée

Nous venons d'évoquer les différentes problématiques autour des données clients-voyageurs et leur gouvernance dans une logique de consolidation d'une base clients. Nous étudierons dans la partie II un exemple concret dans un réseau de transport, le réseau LiA du Havre. Intéressons-nous maintenant au groupe Transdev, dans lequel j'effectue mon apprentissage afin d'appréhender son organisation ainsi que le contexte dans lequel la donnée évolue.

1.3.1. Présentation de Transdev

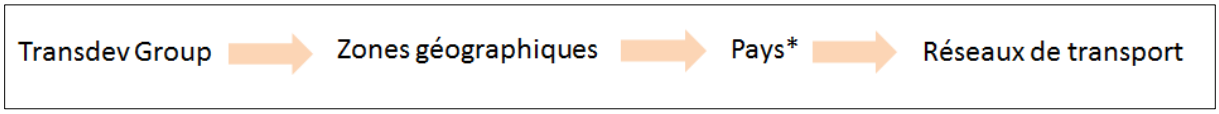
La société Transdev a été créée en 2001 à la suite de la fusion de Transdev (groupe Veolia) et de Veolia Transport (groupe Caisse des Dépôts et Consignations). C'est un opérateur international de transport présent dans une vingtaine de pays exploitant de nombreux modes de mobilité tels que les bus, les tramways, les trains, les lignes d'autocar, le transport à la demande mais aussi les ferries, les navettes, les services médicaux, les services scolaires, l'autopartage, les véhicules autonomes, les vélos. L'ensemble de ces modes de transport peuvent être classés entre modes lourds, modes actifs et nouvelles mobilités. L'une des principales missions du groupe Transdev étant l'accompagnement des

collectivités territoriales ainsi que des entreprises dans la recherche et la mise en œuvre de solutions de mobilité. A ce jour, Transdev « transporte » quotidiennement 11 millions de passagers via les différents modes de transport, et emploie 83 000 personnes (dont 56 390 conducteurs)[30].

La grande majorité de l'exploitation se fait sur des réseaux de transport de villes ou d'agglomérations en réponse à des appels d'offres. Chaque ville ou agglomération a sa propre organisation et ses propres modes de transport. Majoritairement, ce sont des filiales « locales » qui ont la charge de l'exploitation de ces différents réseaux.

S'il existe une expertise « groupe » regroupant des fonctions transverses ou mutualisées, chaque réseau de transport est porté par une filiale qui possède sa propre marge de manœuvre (notamment en termes financiers et d'organisation) et ses propres problématiques et contraintes locales. De plus, chaque réseau est différent en termes d'offre de transport. Il y a donc une **forte décentralisation**. A cela s'ajoute la problématique de la géographie, puisque le groupe Transdev est présent dans de nombreux pays[31]. On retrouve donc également une organisation et des pratiques selon les pays ainsi que selon les zones géographiques.

Graphique 5 : Les différentes échelles de l'organisation de Transdev (également la subdivision « Région » pour la France)*



1.3.2. La problématique « Data » liée à cette organisation décentralisée

Une telle organisation décentralisée et « diffuse » géographiquement entraine une certaine complexité dans l'organisation de l'IT, des systèmes d'information (SI) et donc de la gestion des données.
Si certaines applications sont communes à l'ensemble des filiales-réseaux de transport, d'autres sont plutôt partagées à l'échelle d'une zone géographique ou d'un pays. Dans certains domaines de l'entreprise et selon des directives stratégiques (c'est notamment le cas pour le CRM) il peut y avoir une volonté d'uniformiser et homogénéiser les différentes applications à l'échelle la plus vaste possible.
Cependant plusieurs facteurs peuvent rendre difficile cette uniformisation et participer à la fragmentation de l'IT du fait du contexte local.

Tout d'abord, il peut y avoir une gouvernance partagée de l'IT entre la filiale prestataire et le client (en France les clients sont dans l'ensemble les Autorités Organisatrices de Mobilité, AOM), c'est-à-dire que des applications et des systèmes informatiques sont déjà présents (héritage de l'IT) et restent en place même lorsqu'il y a un changement de prestataire lors d'une nouvelle Délégation de Service Publique. Dans certain réseau, on récupère donc des IT antérieures lorsque dans un autre réseau, à la même période, il a été possible d'intégrer une nouvelle « application groupe » (faisant référence au sein du

[30] Le Groupe Transdev – Carte d'identité 2017 (document interne)
[31] Voir annexe A.

groupe Transdev) dans un domaine particulier (par exemple le Transport à la Demande)[32] et qui sera « débranchée » si le contrat n'est pas renouvelé.

Un autre facteur, est celui du cycle de vie des applications. La mise en place, la maitrise par les équipes (métiers et informatiques), le fait que « cela fonctionne » et les moyens à mettre en œuvre pour d'éventuels changements font que dans la pratique, le cycle de vie des applications est plutôt long.
De plus, aux différents systèmes et applications s'ajoutent les différentes versions (logiciels, solutions…) qui peuvent être différentes entre les différents réseaux de transport.

1.3.3. A quelle échelle penser le Référentiel Client Unique (RCU) ?

Nous avons présenté dans la partie précédente les enjeux de la construction du RCU. L'un des points importants de ce RCU est son échelle de partage (au sein de l'organisation). Dans ce mémoire nous évoquons la problématique générale de la consolidation d'une base clients dans le cas d'un réseau de transport ; mais si la consolidation d'une base clients peut s'effectuer à l'échelle d'un réseau de transport, travailler à une échelle plus large (zone géographique, pays…) peut être dans une « logique groupe » plus intéressant (reporting groupe, vision stratégique à l'échelle la plus large, réseau d'expert et d'entraide…). Alors que le groupe Transdev a engagé le programme de relation client « CRM 360° » visant à la mise en place d'un « CRM groupe » partagé à l'échelle du groupe, poser le questionnement du partage du RCU en articulation avec ce programme arrive à un moment opportun.

1.3.4. Le Département Data Science dans cette organisation

Le Département Data Science, dans lequel j'effectue mon apprentissage, a été constitué en 2016. A l'origine de ce département, les principales missions étaient la prévision de fréquentation et de recettes lors des appels d'offres ainsi qu'un appui aux différentes filiales pour ces appels d'offres, demandant des connaissances économétriques et statistiques poussées. Sur la base de ce socle de connaissance, s'est ajouté le développement d'une « expertise Data » consistant à faire émerger au sein du groupe et des filiales de Transdev une « culture Data » en faisant la preuve notamment de l'apport de la Data Science dans différents domaines traités dans le groupe.

Cette preuve de l'apport de la Data Science se fait notamment par autosaisie de sujets en lien avec différents services du groupe ou de filiales et fait l'objet sur des jeux de données de « proof of concept (POC)[33] ». Ces POC peuvent par la suite se généraliser à d'autres réseaux de transport ou « s'industrialiser » à l'échelle du groupe. Cette deuxième étape (« l'après POC ») ne concerne moins le Département Data Science qui n'a pas vocation à industrialiser les différentes solutions ; il est bien souvent accompagnateur, assistant à maitrise d'ouvrage (AMO) de cette industrialisation.

En 2018, le Département en lien avec la Stratégie du groupe, a élaboré une feuille de route sur les données (Data Roadmap) confirmant ce rôle de « fer de lance » sur les projets Data dans le groupe tout en proposant une méthodologie propre.

[32] Si cela est rendu possible lors de l'Appel d'Offre ou lorsqu'est proposée une nouvelle offre de service.
[33] Preuve de concept en français.

En termes d'organisation pour les futurs projets de Data Science, la Data Roadmap a intégré, entre autres, la mise en place de trois phases successives :

- Deep Dive
- Data Camp
- Use Case Definition and Deployment

Au sein d'un réseau de transport ou d'un service du groupe (RH, Comptabilité, Marketing…) partenaires, l'objectif du **Deep Dive** est d'établir un état des lieux du potentiel de valorisation des données d'un réseau de transport et de répondre à différentes questions telles que[34] :

- Quelles sont les données utilisées et comment sont-elles utilisées ?
- Existe-t-il des besoins qui permettraient d'améliorer la performance opérationnelle ou la stratégie commerciale ;
- Quelles sont les compétences mobilisables sur les données ?

Concrètement, le Deep Dive consiste, en une série d'entretiens ainsi que des extraits de bases de données. C'est une approche exploratoire qui vise à mieux définir les enjeux du groupe vis-à-vis de la question de la valorisation de la donnée en partant notamment d'une échelle plus « locale » (un service ou un réseau de transport).

La suite logique du Deep Dive, à la suite de ce premier diagnostic et si différentes problématiques « Data » ont été soulevées, est le **Data Camp** qui consiste notamment en un premier appui technique dans un temps court et dans un objectif concret de valorisation de données. Nous présenterons plus en détail cette phase dans la Partie II, puisque nous avons effectué un Data Camp dans le réseau LiA (au Havre).

Enfin, une fois la priorisation des premières solutions et POC à la suite du Data Camp, l'enjeu est de réaliser le déploiement des solutions retenues (phase **Use Case Deployment**). C'est cette organisation en trois phases que nous avons suivi dans l'exemple suivant que nous allons présenter.

[34] Questions extraites de la Note de Projet « Data Deep Dive », 31/08/2018 (document interne).

II- Etude de cas : réseau de transport LiA au Havre

Après avoir explicité les données clients-voyageurs et présenté des éléments de gouvernance de la donnée ainsi que les enjeux de la consolidation d'une base clients dans le domaine des transports publics, nous allons présenter dans cette partie un exemple concret, celui du réseau de transport en commun de la Communauté Urbaine Le Havre Seine Métropole[35]. Ce réseau de transport en commun, appelé LiA, est constitué d'un réseau de 21 lignes de bus, 2 lignes de tramway ainsi qu'un funiculaire et des services de transport à la demande (Fil'Bus, Mobi'Fil, LiA de Nuit, FlexiLiA)[36]. C'est la CTPO (Compagnie des Transports de la Porte Océane), filiale du groupe Transdev, qui a la charge de l'exploitation du réseau LiA par convention de délégation de service public (DSP) pour la période 2018-2023.

Ce travail sur la réflexion des recoupements possibles entre les différentes bases de données du réseau LiA a été réalisé par le Direction Data Science, en relation avec les équipes du réseau LiA. Nous évoquerons dans un premier temps la genèse de ce projet, l'état des lieux des bases de données existantes et leurs recoupements potentiels. Dans un second temps, nous présenterons le cas du Transport à la Demande (TAD), une offre de service proposée par LiA ; et nous verrons comment à partir des premiers recoupements entre les différentes bases de données nous avons pu qualifier les utilisateurs de ce service dans un but de ciblage marketing. Enfin, nous aborderons plus en détail la problématique de la consolidation d'une base clients en proposant quelques recommandations et propositions en matière de gouvernance et de management de la donnée.

2.1. Premières observations des recoupements des bases de données de LiA

2.1.1. Genèse du projet

Ce projet est donc le fruit d'un partenariat entre un réseau-filiale de Transdev (LiA) et le Direction Data Science du groupe (au siège de Transdev). Il fait suite à des discussions entre les parties prenantes du projet autour de la problématique suivante exprimée au sein du réseau LiA : comment mettre en place et pérenniser une vue « 360° » des voyageurs afin de tendre vers une gestion personnalisée de la relation avec les voyageurs du réseau LiA. Plus particulièrement, cette problématique a soulevé des questions telles que :

- Comment « pousser » l'offre de Transport à la Demande (TAD) aux clients abonnés du réseau ?
- Comment détecter le « churn[37] » chez les clients abonnés (notamment à partir du parcours client) ?
- Comment cibler géographiquement des zones à fort potentiel pour pousser des offres ?
- Quels sont les usages sur le site internet du réseau LiA ?

Les objectifs du projet, prenant la forme d'une étude réalisée par la direction Data science de Transdev Group en appui au réseau LiA, étaient donc de :

- Comprendre les usages du site internet par les utilisateurs du réseau ;

[35] Anciennement la Communauté de l'Agglomération Havraise (CODAH), changement le 1er janvier 2019.
[36] Voir Annexe B pour une liste plus détaillée des offres de transport du réseau LiA et quelques chiffres clés
[37] Ou attrition.

- Actualiser et augmenter la base client afin de « pousser » des offres personnalisées ;
- Proposer une stratégie de gestion dynamique de la base client.

Ce sont essentiellement les deux derniers points qui nous intéressent pour ce mémoire, même si bien entendu le comportement des clients sur le site internet du réseau LiA s'intègre totalement dans l'approche « vue client 360° ».

2.1.2. « Data Camp » et récolte des différentes sources de données

Ce projet a débuté par un « Data Camp ». Celui-ci a pour objectif d'apporter un appui technique aux entités opérationnelles du groupe afin de leur fournir des éléments tangibles de valorisation de leurs données. Cela constitue donc une démarche projet qui consiste pour l'« équipe Data » du groupe[38] à se déplacer trois jours dans un réseau de transport pour apporter cet appui technique suite à des problématiques discutées en amont (notamment lors d'un « Data Deep Dive »). Il est important de préciser qu'un « Data Camp » fait suite à une démarche volontaire du réseau de transport et nécessite notamment une disponibilité des responsables de la partie opérationnelle de l'IT (notamment pour répondre aux demandes d'extractions des données ainsi que pour permettre les échanges sur le « concret » de l'architecture SI du réseau de transport) et un engagement fort de la Direction du réseau.
L'un des enjeux d'un « Data Camp » est de réussir au bout des trois jours à produire des premiers résultats et livrables. Ces résultats reflètent les utilisations potentielles et futures des données disponibles localement et peuvent donner suite à d'autres travaux soit par les équipes du réseau elles-mêmes soit par l'« équipe Data » du groupe en investiguant davantage. Ces livrables peuvent prendre la forme de productions analytiques sur des données cibles (graphiques, rapport, métriques...), de scripts d'extraction, de propositions de mise en forme des données ou d'une proposition d'architecture dans l'objectif de construire un MVP[39].
La réussite d'un « Data Camp » comprend notamment une bonne identification des problématiques métiers pour lesquelles la valorisation de la donnée pourrait apporter des solutions, une bonne identification et qualification des sources de données utilisées ainsi que la capacité de l'« équipe Data » produire de la valeur lors de ces trois jours et la restituer.

Le « Data Camp » est donc une offre de projet du groupe Transdev à destination des différents réseaux de transports et en partenariat avec eux dans une logique de valorisation des données de ce même réseau.

En ce qui concerne notre projet, un « Data Camp » a donc été réalisé dans les locaux du réseau LiA entre le 20 et le 22 novembre 2018. Cette étape a permis, avec l'appui des équipes locales, de mettre en évidence l'ensemble des données disponibles touchant la relation client et d'en faire un premier état des lieux.

Listons ici l'ensemble des sources de données qui ont pu être récoltées au sein du réseau LiA, nous permettant ainsi d'avoir un panorama assez vaste des données que l'on retrouve dans un réseau de transport en commun :

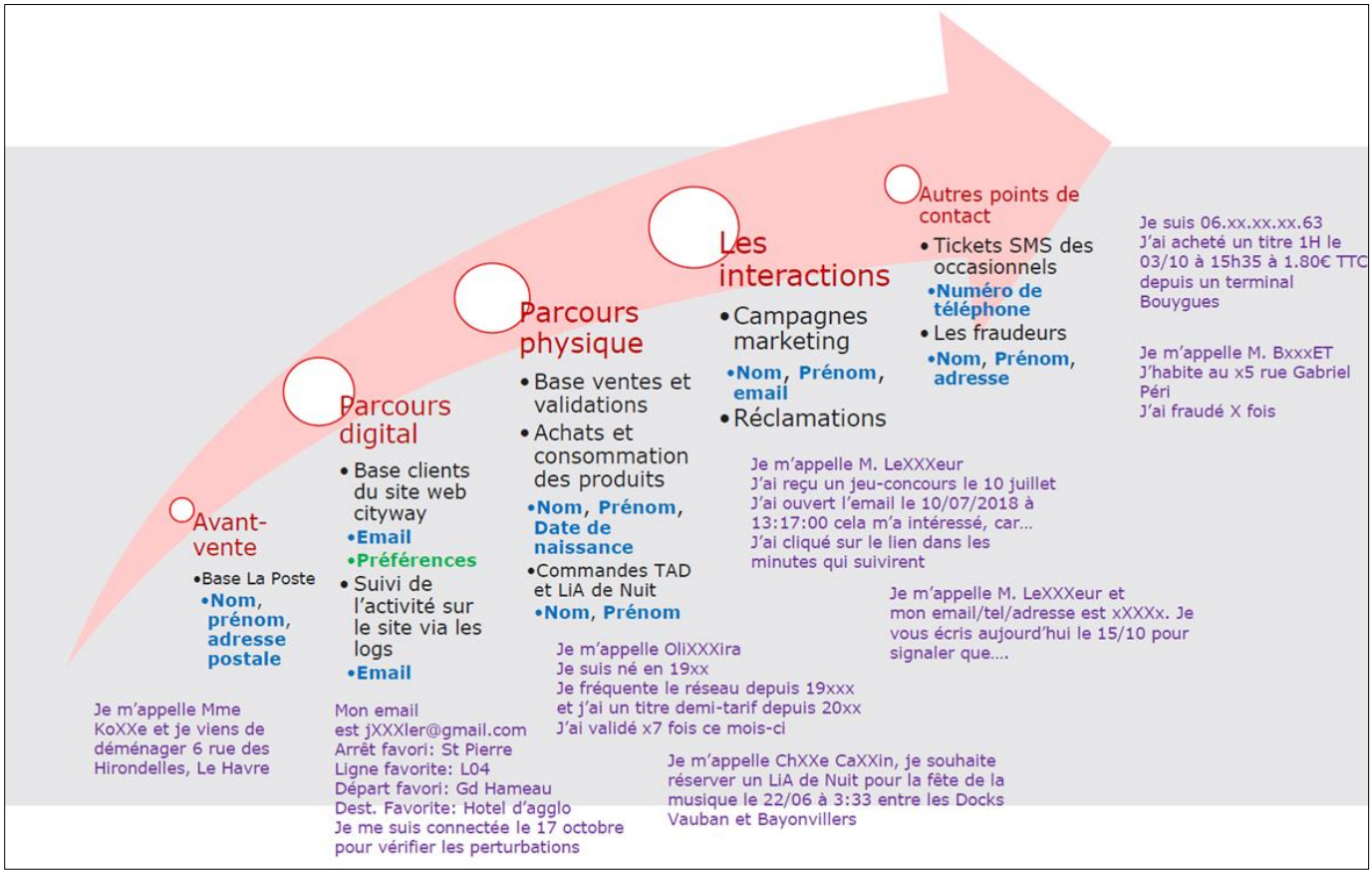
[38] Equipe Data pluridisciplinaire resserrée au sein de Transdev Corp.
[39] MVP pour minimum viable product, en français : produit minimum viable.

- Base La Poste des nouveaux arrivants en novembre 2018 au Havre
- Base Selligent des résultats d'une campagne marketing
- Base des achats de tickets par SMS, depuis le lancement du service au 1er août 2018
- Base des ventes de titres, issue de la billettique
- Base des validations mensuelles par personne et par titre, issue de la billettique
- Base clients issue de la billettique
- Base des clients ayant un compte en ligne
- Un mois de log sur le site internet / appli mobile
- Base des réclamations issue de Listen
- Base des commandes FlexiLiA (TAD) et LiA de Nuit depuis le lancement du service au 01/06/2018
- Liste des missions TAD Filbus et MobiFil, entre le 01/09/2018 et le 18/11/2018
- Base des fraudeurs
- Liste des tweets liés au compte @LiA_LeHavre entre le 18/11/2018 et le 28/11/2018

2.1.3. Panorama des données collectées

Avec l'ensemble de ces bases collectées sur le réseau LiA, et sans avoir commencé les recoupements entre elles, nous pouvons schématiquement reconstruire un parcours client à l'aide des différents points de contact entre le client et le réseau de transport LiA. Nous sommes encore à l'état de parcours théorique en lisant base par base l'information qui peut caractériser un client-voyageur.

Graphique 6 : Les contacts avec le client sur le réseau LiA (extrait du livrable de présentation de la fin du Data Camp par l'« équipe Data » ; document interne)



En **Avant-vente**, le réseau LiA achète tous les mois à l'entreprise La Poste une base de données des nouveaux arrivants dans la Communauté Urbaine Le Havre Seine Métropole pour pouvoir leur faire une offre marketing en leur décrivant le réseau et en espérant capter de nouveaux voyageurs. Le client est à l'état de prospect.

Pour le **Parcours digital**, la base des logs du site internet et de l'application mobile contient pour chaque utilisateur un fichier dont on peut extraire l'ensemble du parcours (connexion, pages consultées, recherches effectuées, déconnexion) sur le site ou l'application.

Le **Parcours physique**, plus classique pour un réseau de transport correspond aux achats et à la consommation des titres de transport sur le réseau ainsi que l'utilisation du service de transport à la demande (TAD).

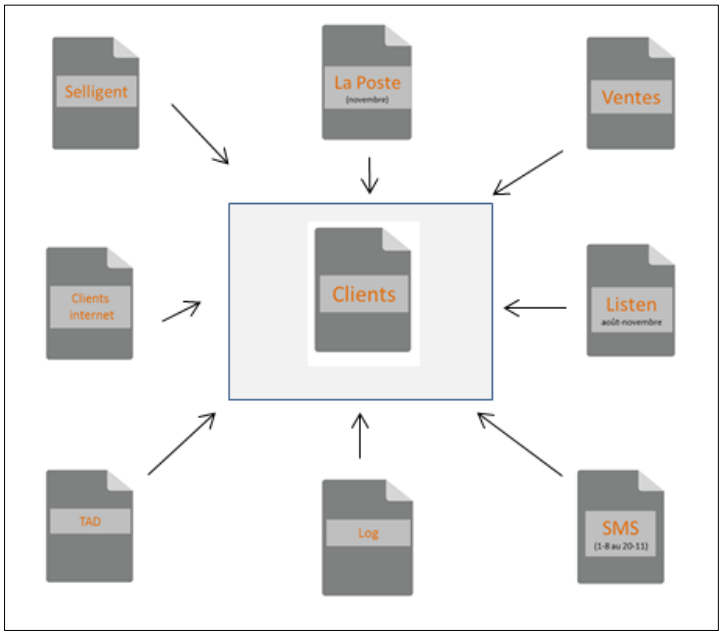
Les **Interactions** sont constituées des réclamations que peut faire un client ainsi que la « relation marketing » entre ce dernier et le réseau LiA.

Enfin, dans les **Autres points de contacts** on retrouve le service de ticket de transport par SMS (celui-ci est en fait un mix entre le parcours digital et le parcours physique) et la base des fraudeurs.

2.1.4. Recoupement des différentes bases de données (vision unifiée des bases de données clients)

Avant de proposer une consolidation d'une base clients permettant une gestion pérenne de la relation client et un suivit 360° du client, un premier travail a été effectué pour quantifier les recoupements entre les différentes bases de données.
Nous avons considéré comme base centrale pour réaliser ces recoupements la base clients issue de la billettique, déjà constituée par le réseau LiA.

Graphique 7 : Représentation des recoupements à partir de la base clients issue de la billettique



Cette quantification des recoupements entre ces différentes bases a été effectuée dans un premier temps de façon non-automatisée avec le logiciel R.

2.1.4.1.Création des « clés-identifiants »

La première étape a été de **créer des « clés-identifiants »** (qui permettront les jointures entre les bases) à partir des différents champs des bases de données. La création de ces clés a été réalisée à partir des champs suivants :

- Numéro du client
- Nom
- Prénom
- Adresse
- Date de naissance
- E-mail
- Numéro du téléphone portable

Tout un travail de codage en R a été nécessaire sur l'ensemble des champs lorsqu'ils étaient présents pour les formater et les homogénéiser entre les différentes bases. Ce travail de codage directement sur les données a consisté en du nettoyage (« cleansing ») et du formatage de texte (suppression des accents, des caractères spéciaux et des espaces pour les champs textuels tels que le Nom, le Prénom, l'Adresse) et en de la vérification-correction (taille du numéro de téléphone et remplacement des patterns « +33 » en « 06 » ; présence des caractères « @ » et « . » dans l'adresse e-mail ; taille du code postal…).

Enfin, nous générons les clés par concaténation de ces différents champs. Voici un exemple de code commenté pour la génération de clés dans la base clients d'origine :

Graphique 8 : Extrait du code R pour la génération de clés dans les différentes bases (exemple de la base Clients d'origine)

```
#### traitement de la base clients ####
clients$PrenomBis <- tolower(clients$Prenom)   # met en minuscule
clients$NomBis <- tolower(clients$Nom)         # met en minuscule

clients$PrenomBis <- unaccent(clients$PrenomBis)  # fonction qui enlève les accents et caractères spéciaux
clients$NomBis <- unaccent(clients$NomBis)        # fonction qui enlève les accents et caractères spéciaux

# suppression des espaces
clients$PrenomBis <- str_replace_all(clients$PrenomBis," ","")
clients$NomBis <- str_replace_all(clients$NomBis," ","")

# adresse complète
clients$adresseComplete <- paste(as.character(clients$Adresse),       # on colle les différents champs portant
                                 as.character(clients$Code.postal),   # une information sur l'adresse
                                 as.character(clients$Ville),
                                 sep=" ")

clients$adresseComplete <- tolower(clients$adresseComplete)                  # met en minuscule
clients$adresseComplete <- unaccent(clients$adresseComplete)                 # suppression accents et caractères spéciaux
clients$adresseComplete <- str_replace_all(clients$adresseComplete," ","")   # supression des espaces

## génération de clés

# clé 1 nomPrenom
clients$nomPrenom <- paste(clients$PrenomBis,
                           clients$NomBis,sep="")     #concaténation des champs PrenomBis et NomBis

# clé 2 nomPrenomAdresse
clients$nomPrenomAdresse <- paste(clients$PrenomBis,  #concaténation des champs PrenomBis, NomBis et adresseComplete
                                  clients$NomBis,
                                  clients$adresseComplete,sep="")
```

2.1.4.2. Observation/Etat des lieux des recoupements

La génération des différentes clés a donc été effectuée sur l'ensemble des bases. Voici un résumé des clés pouvant être générées selon la présence des différents champs structurants :

Tableau 4 : Clés qui ont pu être crées (en vert) dans les différentes bases

	les différentes clés crées dans les bases					
	nomPrenom	**nomPrenomAdresse**	**email**	**nomPrenomDateNaissance**	**numeroTelPortable**	**numeroClient**
Clients (via billétique)						
Selligent (campagnes marketings)						
Clients internet						
TAD (transport à la demande)						
Log (parcours site internet)						
SMS (titres de transport par SMS)						
Listen (réclamations)						
Ventes (titres de transport)						
La Poste (nouveaux arrivants)						

Nous pouvons voir à l'aide de ce tableau que l'on ne retrouve pas l'ensemble des différentes clés dans toutes les bases. Cela est logique dans la mesure où tous les champs mentionnés ne se retrouvent pas tous dans les différentes bases de données du fait de leur constitution et des applications auxquelles elles sont rattachées. Par exemple, on ne retrouve pas l'adresse e-mail dans l'application de vente de tickets de transport par SMS ou pour l'application de réclamations Listen, il est un peu délicat de demander à l'utilisateur de saisir sa date de naissance.

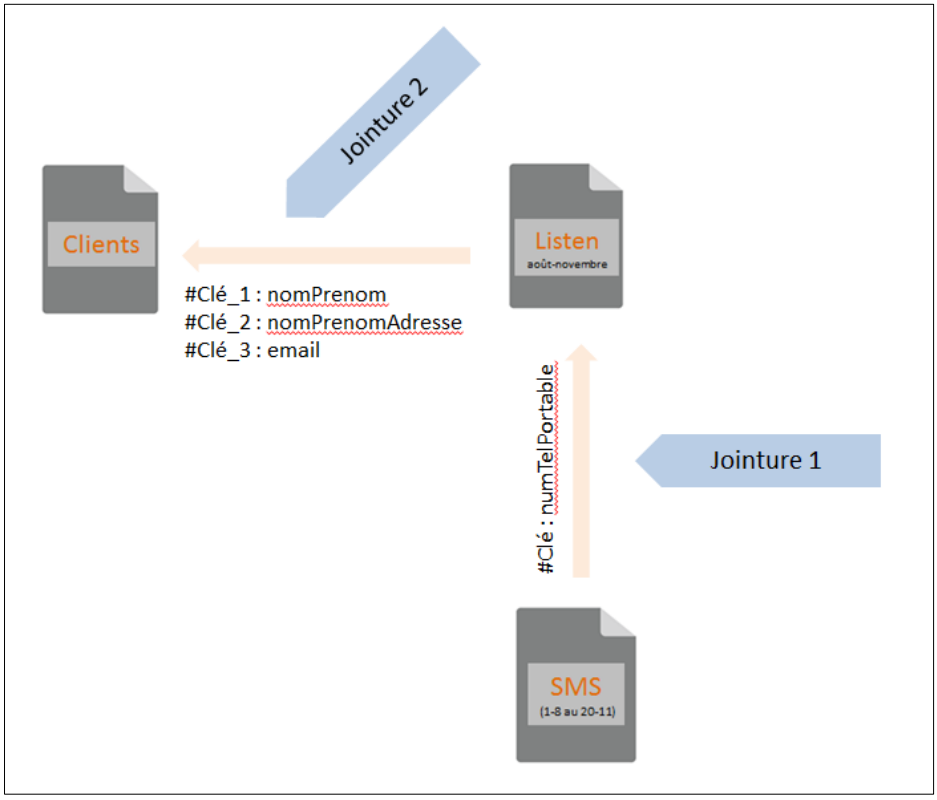
A l'aide de ces différentes clés, nous avons réalisé des **jointures** entre la base clients et les autres bases et calculé le pourcentage de recoupement ou **taux de correspondance entre ces bases**[40]. Ce pourcentage de recoupement correspond au pourcentage de clients d'une base dans la base Clients. Nous avons voulu montrer à l'aide de ces recoupements comment les différentes bases après jointures peuvent préfigurer d'un début de vue unique en mettant en avant les quelques difficultés rencontrées qu'il faudra prendre en compte pour la consolidation de la base clients unique.

Entre la base Ventes et la base Clients, nous avons un taux de correspondance de 99%, ce qui est logique et rassurant étant donné que la jointure se fait à l'aide de la clé *numeroClient*. Cette clé est la plus sûre étant donné que c'est un identifiant unique généré dans la base Clients. La correspondance base Ventes – base Clients permet de rattacher les différents trajets d'un client. On se retrouve dans le niveau 2 de la « maturité client » (*cf.* Graphique 2) que nous avions présenté au début de ce mémoire.

Pour la base SMS, nous pouvons voir à l'aide du tableau 4 précédent que l'unique clé qu'elle contient (*numeroTelPortable*) n'est pas présente dans la base Clients. Nous avons donc dû faire une première jointure entre la base SMS et la base Listen (qui contient la clé *numeroTelPortable)* avant de réaliser la jointure avec la base Clients :

[40] Voir Annexe C reprenant les pourcentages de correspondances entre les différentes bases.

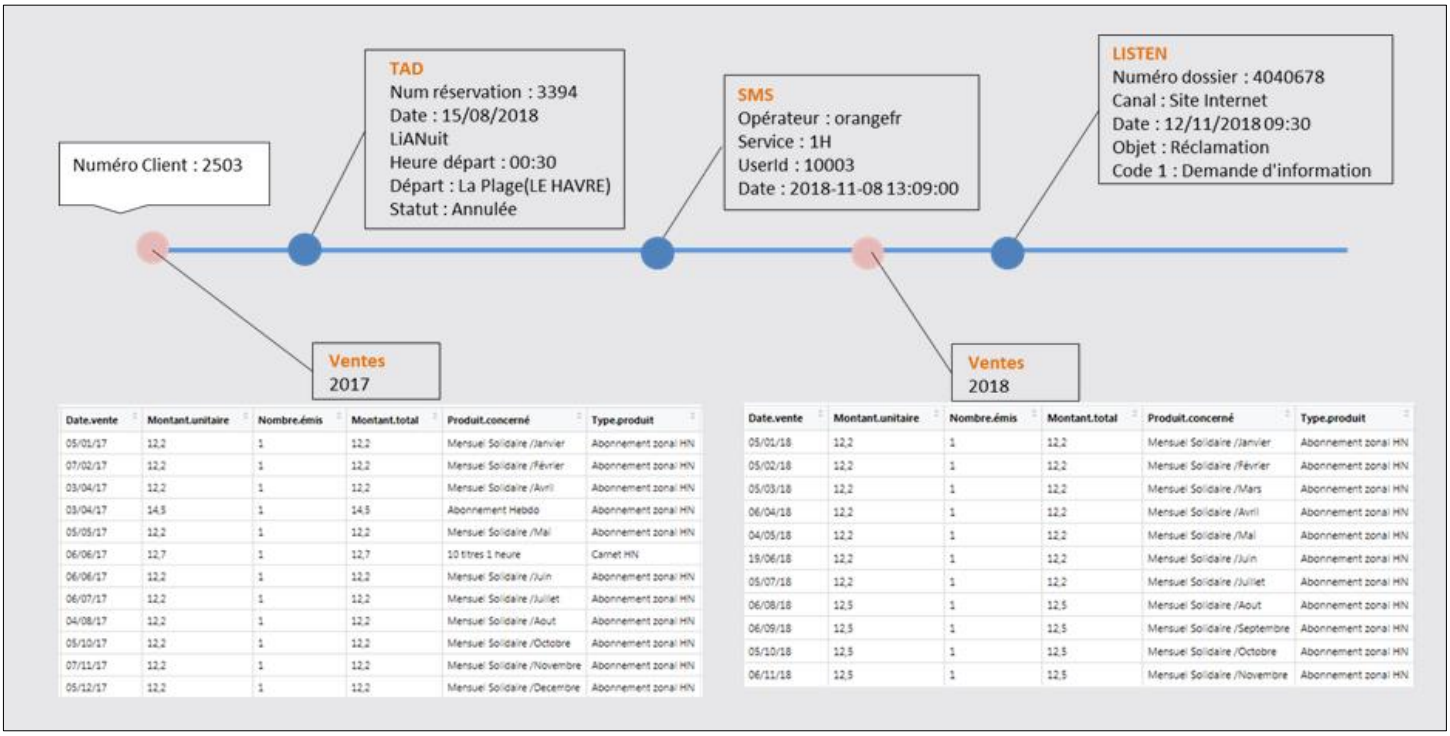
Graphique 9 : Les différentes étapes de recoupement (par deux jointures) entre la base SMS et la base clients



Précisons également qu'il a été nécessaire de prendre en compte la problématique des homonymes (nom + prénom) des différents clients dans la création des clés et pour les jointures. Nous évoquerons ce sujet dans la partie suivante.

Nous avons donc dû jongler entre les différentes clés pouvant être créées dans les bases pour permettre les différentes jointures et réaliser un début de vue « 360° » des clients (non automatisée) nous permettant notamment d'être en mesure de décrire le parcours temporel complet de quelques clients au travers de l'ensemble de ses contacts avec le réseau LiA :

Graphique 10 : Suivi personnalisé du parcours du client n°2503 à travers l'ensemble du réseau LiA



Date.vente	Montant.unitaire	Nombre.émis	Montant.total	Produit.concerné	Type.produit
05/01/17	12,2	1	12,2	Mensuel Solidaire /Janvier	Abonnement zonal HN
07/02/17	12,2	1	12,2	Mensuel Solidaire /Février	Abonnement zonal HN
03/04/17	12,2	1	12,2	Mensuel Solidaire /Avril	Abonnement zonal HN
03/04/17	14,5	1	14,5	Abonnement Hebdo	Abonnement zonal HN
05/05/17	12,2	1	12,2	Mensuel Solidaire /Mai	Abonnement zonal HN
06/06/17	12,7	1	12,7	10 titres 1 heure	Carnet HN
06/06/17	12,2	1	12,2	Mensuel Solidaire /Juin	Abonnement zonal HN
06/07/17	12,2	1	12,2	Mensuel Solidaire /Juillet	Abonnement zonal HN
04/08/17	12,2	1	12,2	Mensuel Solidaire /Aout	Abonnement zonal HN
05/10/17	12,2	1	12,2	Mensuel Solidaire /Octobre	Abonnement zonal HN
07/11/17	12,2	1	12,2	Mensuel Solidaire /Novembre	Abonnement zonal HN
05/12/17	12,2	1	12,2	Mensuel Solidaire /Decembre	Abonnement zonal HN

Date.vente	Montant.unitaire	Nombre.émis	Montant.total	Produit.concerné	Type.produit
05/01/18	12,2	1	12,2	Mensuel Solidaire /Janvier	Abonnement zonal HN
05/02/18	12,2	1	12,2	Mensuel Solidaire /Février	Abonnement zonal HN
05/03/18	12,2	1	12,2	Mensuel Solidaire /Mars	Abonnement zonal HN
06/04/18	12,2	1	12,2	Mensuel Solidaire /Avril	Abonnement zonal HN
04/05/18	12,2	1	12,2	Mensuel Solidaire /Mai	Abonnement zonal HN
19/06/18	12,2	1	12,2	Mensuel Solidaire /Juin	Abonnement zonal HN
05/07/18	12,2	1	12,2	Mensuel Solidaire /Juillet	Abonnement zonal HN
06/08/18	12,5	1	12,5	Mensuel Solidaire /Aout	Abonnement zonal HN
06/09/18	12,5	1	12,5	Mensuel Solidaire /Septembre	Abonnement zonal HN
05/10/18	12,5	1	12,5	Mensuel Solidaire /Octobre	Abonnement zonal HN
06/11/18	12,5	1	12,5	Mensuel Solidaire /Novembre	Abonnement zonal HN

Le Graphique 10 nous permet de suivre le client n°2503. Nous observons par année ses achats en titres de transport. Ce client avait réservé le 15/08/2018 un transport à la demande (TAD) qu'il a finalement annulé. Il a voyagé à l'aide d'un titre de transport par SMS le 08/11/2018. Enfin, via le canal du site internet du réseau LiA, il a fait une réclamation (demande d'information) le 12/11/2018.

2.2. Application des recoupements au Transport A la Demande

Nous allons présenter ici un exemple d'application concrète en termes de « Connaissance Client » et de ciblage marketing rendue possible à la suite des premiers recoupements que nous avons effectués entre les différentes bases de données du réseau LiA. Ce dernier cas concerne le service de Transport à la Demande (TAD), offre de service proposée par LiA.

2.2.1. Présentation du Transport à la Demande (TAD)

Il existe plusieurs définitions et pratiques du Transport à la demande (TAD) selon les pays. En Europe, le TAD est souvent présenté comme une forme intermédiaire entre le taxi et le bus et plus particulièrement en France l'accent est mis sur la réservation qui consiste à faire savoir par différents canaux (historiquement par téléphone) son souhait d'utiliser ce service de transport ; **service** de transport de fait **individualisé** (suite à une demande individuelle) mais aussi **collectif** puisque le gestionnaire du service cherche le plus souvent le regroupement de différents passagers en optimisant les trajets (Castex, 2007).

Très souvent, mais pas exclusivement, le TAD est une offre de service de transport utilisée pour desservir les zones peu denses, plutôt à la périphérie des agglomérations ou les zones rurales et périurbaines peu denses.

Il existe différents modes de TAD, que l'on retrouve également dans les différents TAD dont les filiales de Transdev ont la charge. Les deux principaux types sont :

- Le TAD « porte-à-porte » ou TAD « adresse à adresse » : Prise en charge (dans les zones définies du TAD) à une adresse donnée par le voyageur (domicile, travail) qui est transporté à une autre adresse. Solution ressemblant au service proposé par un taxi ;
- Arrêts TAD : Des arrêts répartis dans différents localités où les voyageurs doivent se rendre pour emprunter le service TAD. On parle aussi de « lignes virtuelles TAD ».

A ces différents modes de TAD s'ajoute des différences d'horaires dans la prise en charge des voyageurs : horaires libres, horaires fixes, plages horaire définies.

Dans leur ensemble, ces services TAD sont organisés par les collectivités locales (communes, conseils départementaux, conseils régionaux, Etablissement Public de Coopération Intercommunale) et les Autorités Organisatrice de Transports (AOT). Certaines associations (d'assistance aux personnes en difficulté, assimilables à des centrales de mobilité…) peuvent également organiser des TAD (Délégation à l'aménagement du territoire et à l'action régionale, 2004).

Notons qu'il existe également un cas particulier du TAD, c'est le Transport des Personnes de Mobilité Réduite (TPMR) avec un véhicule adapté notamment pour les fauteuils roulants.

Quatre offres de TAD sont proposées comme offre de service au sein du réseau LiA[41] :

- LiAdeNuit (type arrêts desservis selon la demande) de de 00h30 à 5h00 ;
- FiLBus pour la desserte des zones peu denses permettant de se déplacer entre des arrêts dans les communes de la CODAH ;
- FlexiLiA, quatre lignes de bus desservant les entreprises de la Zone Industrielle et Portuaire avec réservation ;
- MobiFil service de transport collectif à la demande réservé aux personnes à mobilité réduite ne pouvant prendre les transports en commun classiques.

2.2.2. Les données du TAD

L'ensemble des données TAD pour le réseau LiA sont stockées dans la base « Optycall » et comprend les champs principaux suivants :

- Date
- **Nom**
- **Prénom**
- Type de TAD
- Nombre de personnes
- Lieu de départ
- Heure de départ
- Lieu d'arrivée
- Heure d'arrivée

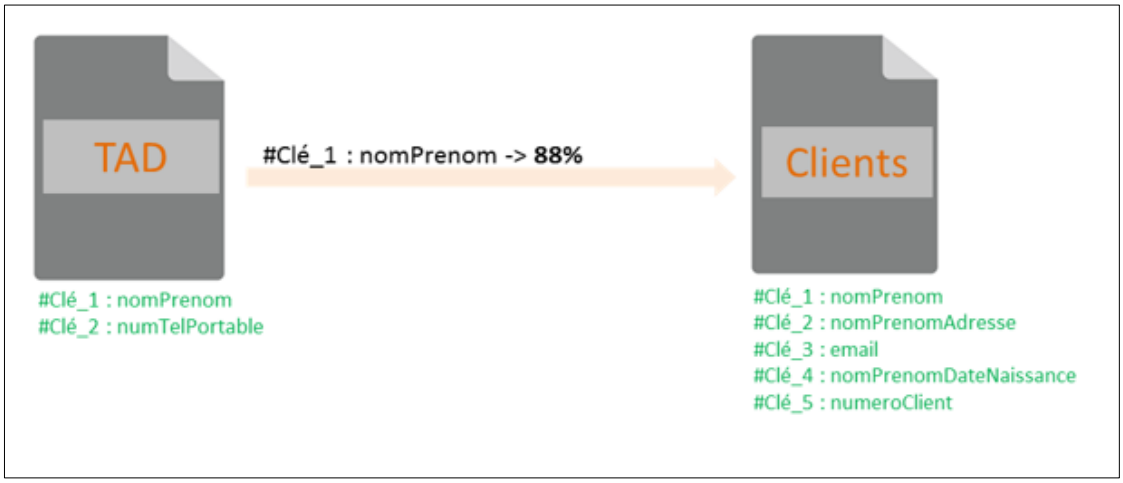
[41] https://www.transports-lia.fr/

- Distance de la course
- **Numéro de téléphone**

Nous avons pu avoir accès aux données des réservations pour les services LiAdeNuit et FlexiLiA sur la période du 01/06/2018 au 20/11/2018 soit l'équivalent de 11 362 lignes (une ligne correspondant à une réservation[42]).

Seuls trois champs structurants permettent de qualifier un client-voyageur (Nom, Prénom, Numéro de téléphone). Avec la clé *nomPrenom*, nous avons pu faire les recoupements avec la base clients d'origine :

Graphique 11 : Taux de recoupement entre la base TAD et la base clients



Nous retrouvons 88% des clients-voyageurs du TAD dans la base clients d'origine à l'aide de la clé *nomPrenom*.

Etant donné que la réservation des services LiAdeNuit et Flexilia peut se faire sur le site internet du réseau LiA[43] via un compte utilisateur, dans le but d'avoir le plus de champs structurants (qualifiants un client) par bases de données, il serait pertinent de retrouver dans cette base TAD « Optycall » les données liées au compte du site web tel que l'adresse mail et/ou l'identifiant du compte.

De même, le champ téléphone n'étant pas présent dans la base clients d'origine[44], nous n'avons pas utilisé ce champ pour la consolidation avec la base TAD.

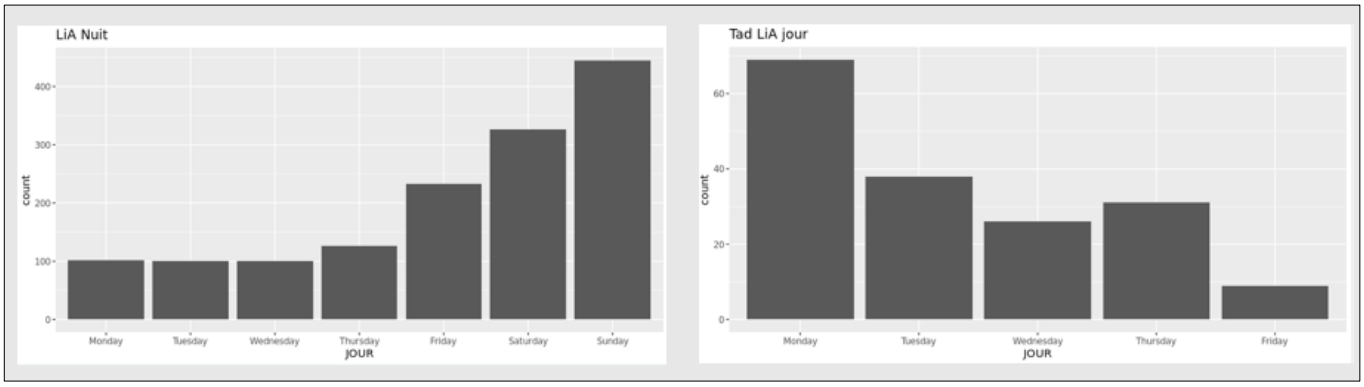
2.2.3. Premiers résultats des recoupements

Cette première consolidation TAD/base clients non-automatisée effectuée, nous avons pu identifier et qualifier les clients intéressés par les offres de mobilité à la demande. Avant cela nous avons déterminé les typologies d'usage du TAD.

[42] Certaines réservations ont le statut « annulé ».
[43] https://www.transports-lia.fr/
[44] Numéros de téléphone non présent dans nos premières extractions ; cependant dans le document (interne) de modélisation des bases de données du réseau LiA, ce champ est présent. Ajouter ce champ pour les prochains travaux.

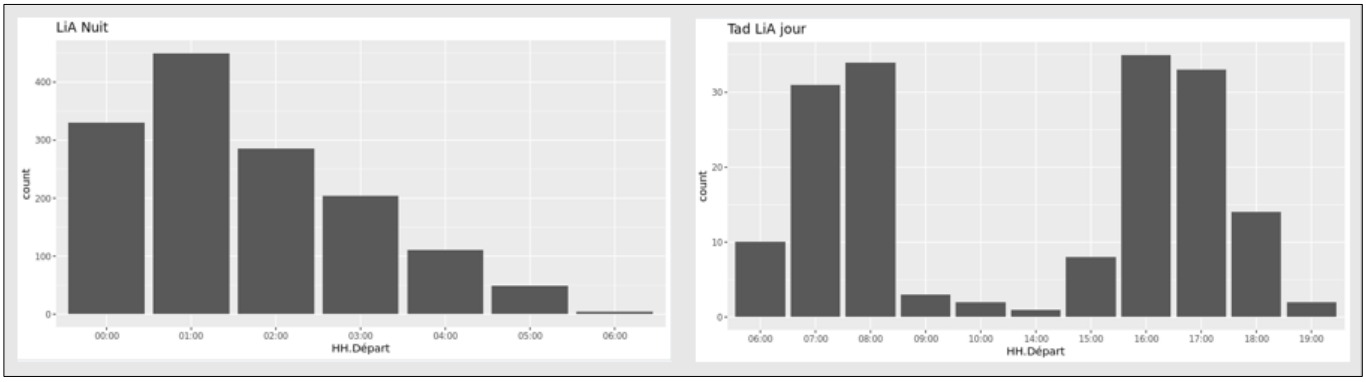
a) Typologie d'usage du TAD

Par jour de la semaine :



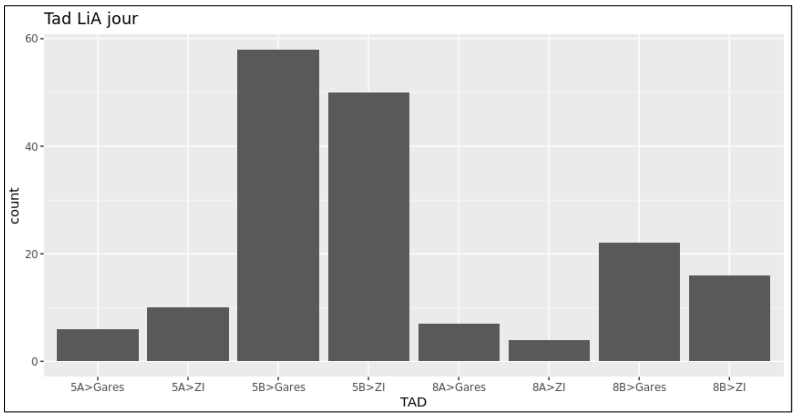
- Forte utilisation du TAD nuit le weekend
- Nombre important de TAD jour le lundi

Par horaires de départ :



- TAD nuit principalement entre 00:00 et 03:00
- TAD jour en début de matinée et fin d'après-midi

Par « zones TAD » :

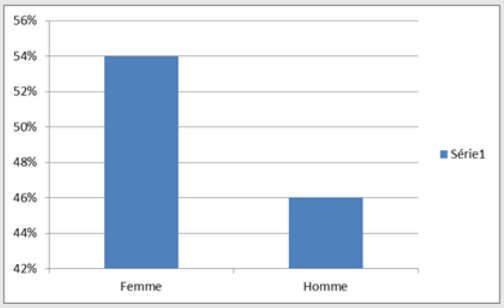


- La ligne 5B qui dessert notamment la zone industrielle plus utilisée

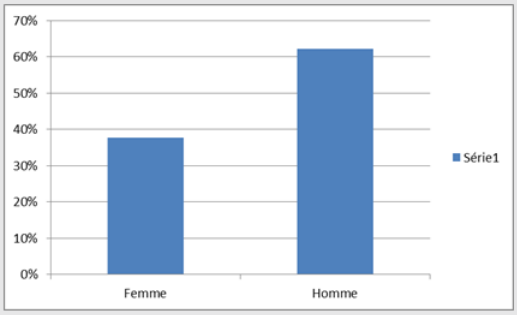
b) Typologie des utilisateurs du TAD

A l'aide du prénom des utilisateurs du TAD, nous avons pu déterminer le sexe de la personne[45] :

TAD Nuit		
Femme	Homme	Total
719	613	1332
54%	46%	100%

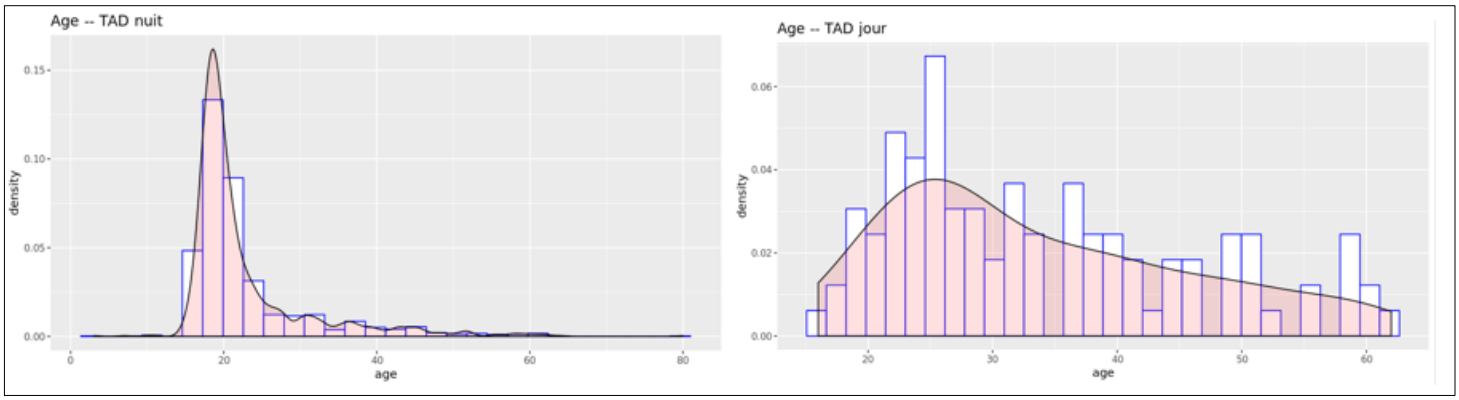


TAD Jour		
Femme	Homme	Total
54	89	143
38%	62%	100%



Les femmes sont plus nombreuses à utiliser le service de TAD la nuit.

Nous avons également observé la répartition de l'âge des utilisateurs à l'aide de la date de naissance :



On constate une forte utilisation du TAD nuit par une population jeune 18-25 ans (soit 67% des clients TAD nuit présents dans la base clients). Cette population des 18-25 ans représente 25% de la base clients.

Nous le voyons, avec seulement deux informations socio-démographiques (le sexe et l'âge), nous avons pu mieux caractériser les utilisateurs des services TAD du réseau LiA. Cette caractérisation des utilisateurs peut permettre de pousser des offres marketing ciblées et tout simplement mieux appréhender les utilisateurs de ces services.

[45] Nous n'avions pas le champ « Sexe » dans les bases de données, nous avons donc réalisé un « matching » à partir du fichier des prénoms en France depuis 1900 de l'INSEE. La problématique des prénoms unisexes n'a pas été gérée (observations à titre indicatif : importance de d'avoir l'information sexe dans la base clients générale).

2.3. Vers une consolidation d'une base clients

Nous voulons à présent, au regard des éléments théoriques de la première partie et des premiers travaux effectués sur les données du réseau LiA, présenter quelques éléments de gouvernance de la donnée dans une logique d'ouverture et de pérennisation d'une telle démarche de recoupements de bases de données clients. En effet, les recoupements que nous avons effectués ont été réalisés dans un cadre théorique, non-automatisé et dans un objectif de mettre en avant la faisabilité d'une telle démarche. Une deuxième étape sera de penser l'architecture de données qui permettra une vision 360° du client en prenant notamment en compte, dans sa mise en place, les problématiques de la qualité de la donnée et la protection des données personnelles.

2.3.1. Qualité de la donnée

A la suite du « Data Camp » et dans notre logique de recoupement entre les différentes bases de données, nous avons utilisé l'approche « nettoyage » pour améliorer la qualité des données. A cette étape de notre étude, nous avons donc travaillé sur les dimensions **« cohérence des données »** et **« l'exactitude des données »**. Nous avons effectué cette étape de « nettoyage » à l'aide d'instructions également avec le logicielle R.

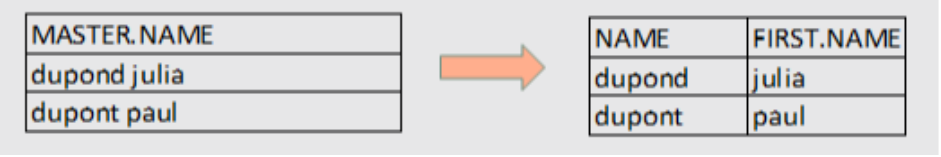
Nous avons notamment rencontré ces différents cas de nettoyage :

- Différents formats du champ téléphone :
  - **06** 88 89…
  - **+33** 6 88 89…
  - **6** 88 89…

- Également sur la date de naissance :
  - 10/09/1992
  - 01 09 1992
  - 01-09-1992
  - 1/9/1992

Pour chacun de ces cas, des routines ont été codées afin d'homogénéiser ces différents champs.
Nous avons proposé de mettre en place, lorsque cela était possible, dans les différentes applications la **complétion automatique** (aide à la saisie d'informations dans un champ de formulaire lié à une source de données) et des **formats de champs de saisie spécifiques** (et non manuel, par exemple choisir une date dans un calendrier et non l'écrire à la main).
Nous n'avons pas trouvé d'anomalies en ce qui concernait les champs « code postal » et « e-mail ».

Dans la base Marketing (Selligent), il est plus pertinent de garder deux colonnes pour NOM et PRENOM plutôt qu'une seule pour faciliter les recoupements :

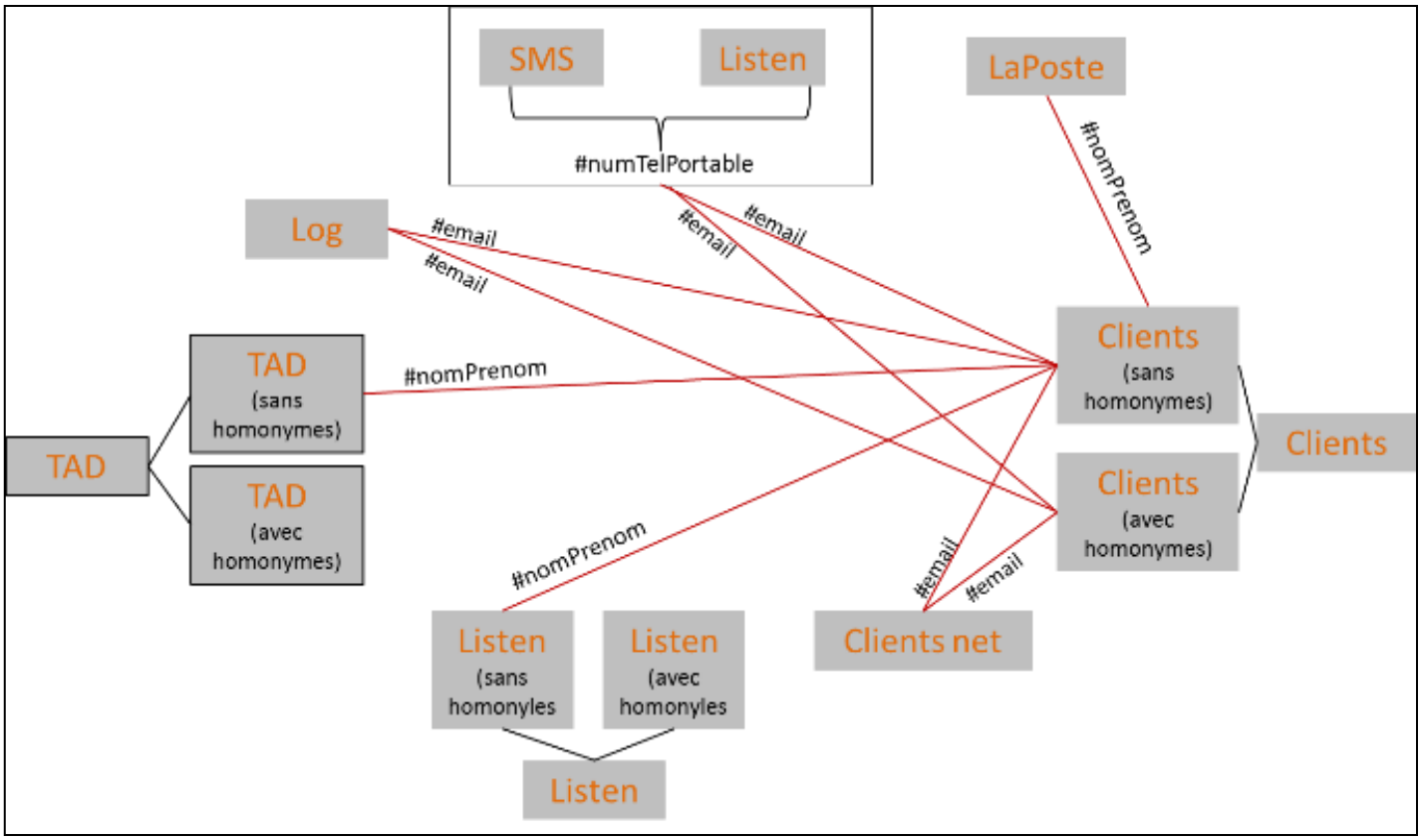


2.3.2. Champs structurants et problématique des données personnelles

L'ensemble des champs qui nous ont permis de réaliser des clés pour les recoupements entre les bases de données sont des **champs structurants**. La qualité de ceux-ci (qualité de la donnée) est primordiale pour assurer notamment la « résolution des identités ». Dans une logique d'automatisation, ce travail de « qualité de la donnée » s'effectue à l'aide ETL avant d'alimenter la base centrale clients. De plus, les données de références clients (MDM) s'appuient également sur ces champs structurants. Il est donc primordial d'essayer d'avoir le plus de champs structurants dans les différentes applications et bases de données qui y sont rattachées.

En effet, une des principales difficultés que nous avons rencontrée dans la « résolution des identités » a été la gestion des homonymes (nom+prénom), notamment lorsqu'il n'était pas possible de concaténer un autre champ structurant pour créer une clé. Nous avons dû passer par un script (en R) pour séparer les identités « homonymes nomPrenom » (c'est-à-dire des identités pour lesquelles on retrouve plusieurs fois l'information nomPrenom dans la même base de données) des identités « uniques ». Les homonymes ont été mis à l'écart et n'ont pas été repris dans les recoupements avec la base clients. Le Graphique 12 montre l'architecture des recoupements entre les différentes bases de données et la base clients avec cette mise à l'écart des homonymes :

Graphique 12 : Architecture des recoupements entre les différentes bases de données et la base clients avec la mise à l'écart des homonymes



La gestion des champs structurants implique de prendre en compte la problématique des données personnelles. Reprenons l'ensemble des bases de données que nous avons récolté au sein du réseau LiA, en évoquant les différents « points RGPD » correspondant :

Tableau 5 : Bases clients et cadre juridique de l'exécution du contrat

Base clients (d'origine)	Cadre juridique de **l'exécution du contrat**
Base Listen (les réclamations)	Le voyageur fait une réclamation via un de ces canaux : Formulaire web, Email, Téléphone, Courier, Agence commerciale. ➔ Réponse dans un délai contractuel ➔ Consentement (opt-in) pas nécessaire, **les coordonnées du client ne doivent servir <u>uniquement</u> dans le cadre d'une réponse à sa demande**
Base clients internet (compte sur le site internet)	**Mention de l'utilisation (finalité) des données** lors de la création d'un compte sur le site internet (pour approbation)
Base Logs (connexions internet)	**Mention de l'utilisation (finalité) des données** lors de la création d'un compte sur le site internet (pour approbation)
Base TAD (transport à la demande)	**Mention de l'utilisation (finalité) des données Exécution du contrat**
Base SMS (ticket de transport par SMS)	La donnée personnelle (téléphone portable) peut être conservé (dans les limites légales de durée) ➔ **Exécution du contrat** **Prospection commerciale interdite** à partir de cette donnée **sans consentement** préalable
Base Vente (titres de transport)	Cadre juridique de **l'exécution du contrat**
Base La Poste	Base achetée au groupe La Poste. Cette base doit déjà respecter les contraintes RGPD (mention « vos données personnelles soient utilisées à des fins de prospection commerciale »)

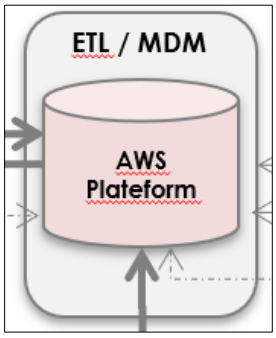
2.3.3. Architecture orientée clients pour une approche « CRM 360° »

Un des enjeux de ce travail a été de mettre en évidence le potentiel des différents recoupements entre l'ensemble des différentes bases de données portant une information sur les clients ainsi que leurs comportements et leur relation avec le réseau de transport. La suite logique (qui ne sera pas traitée en détail ici) est de tendre vers une architecture orientée client permettant d'automatiser l'ensemble de ces recoupements pour assurer une relation client personnalisée. Cette seconde phase fait l'objet d'un travail en cours piloté par la Direction Client Groupe (DCG) de Transdev sur le programme « CRM 360° ».

Cette architecture se doit d'intégrer les différents éléments de gouvernance de la donnée, que nous avons notamment mentionnés dans ce mémoire (qualité de la donnée, résolution d'identités, protection des données personnelles, référentiel clients, …). Et notamment d'intégrer les obligations légales dues au RGPD (stockage des opt-in) car une « vision 360° » et une relation client personnalisée ne signifient pas *tout recouper à priori*. Par exemple, les données personnelles en lien avec une réclamation au sein du réseau de transport ne doivent pas être à priori en lien avec l'application de marketing de campagne de mail. Cette centralisation, en tant qu'architecture, est cependant pertinente en ce qui concerne les données de références des clients (MDM) : pour avoir la vérité en un seul point relié aux différentes applications clients (constituant le CRM).

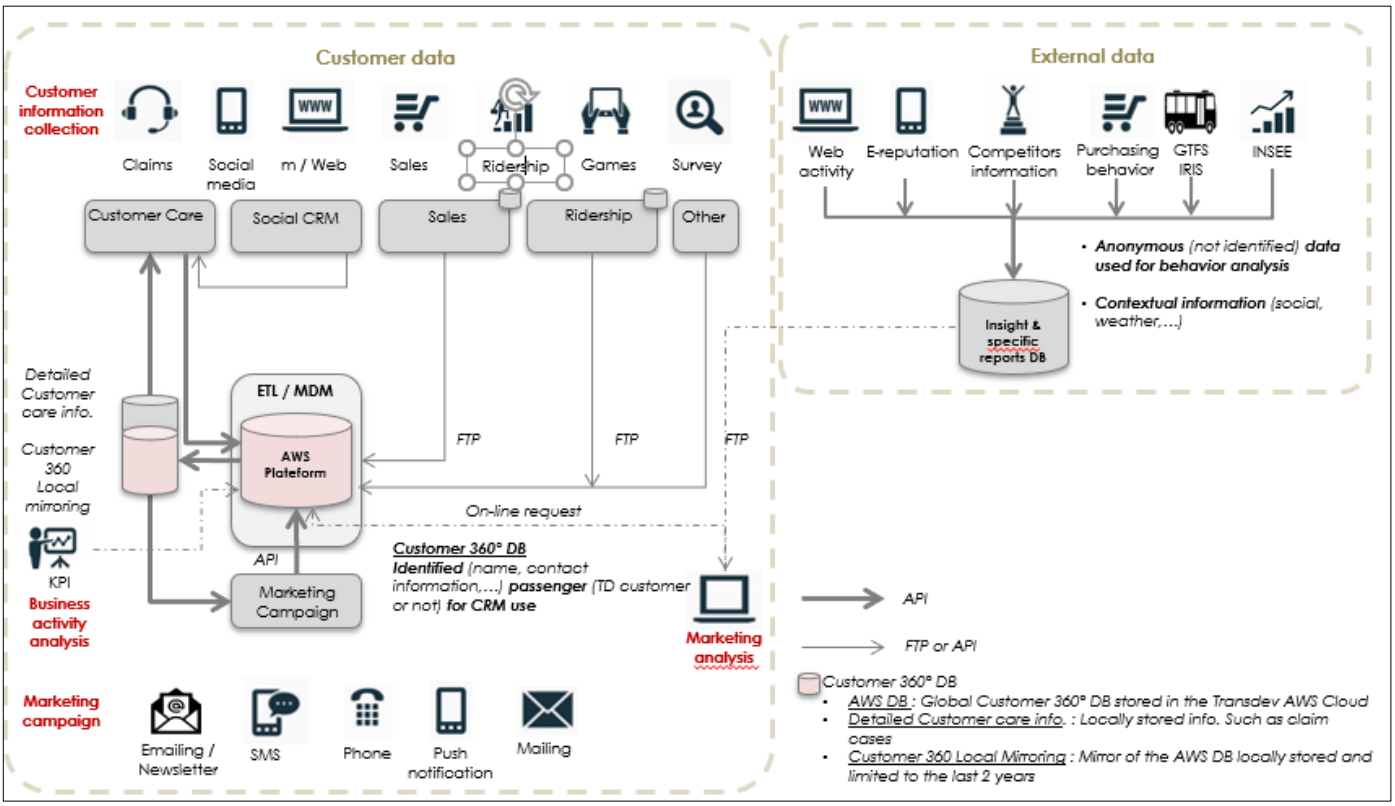
Les équipes de la DCG ont retenu une solution « Cloud » (Amazon Web Service) comme solution centralisée pour gérer les flux et stocker les données.
L'ensemble des recoupements pertinents ainsi que la gestion des Master Data qui devront être définies seront automatisés et intégrés au Data Warehouse sur le Cloud :

Graphique 13 : Data Warehouse sur le Cloud relié à l'ETL et au MDM



Ce Data Warehouse sera lui-même relié à deux solutions (logiciels), une d'ETL et l'autre de MDM. L'ensemble des différentes sources de données que nous avons décrites dans l'ensemble de ce mémoire concernant les clients, alimenteront ce Data Warehouse qui sera directement relié à la solution CRM permettant l'activation des campagnes marketings et la gestion de la relation client de manière plus automatisé et personnalisé :

Graphique 14 : Présentation de l'architecture centralisée « CRM 360° » (document interne)



Nous pouvons voir dans le schéma précédent que l'ensemble des flux à partir des bases de données passent par l'ETL et le MDM avant d'aller vers la plateforme AWS. Cette dernière alimente le CRM qui est lui-même relié au « Marketing Campaign ». Ces dernières alimentent également la base centrale.
A cela s'ajoute également les données extérieures que nous avons peu traitées dans ce mémoire mais qui s'intègrent également dans cette architecture orientée client, plus dans une logique de contextualisation (données contextuelles).

Conclusion

La consolidation d'une base clients, nous avons pu le voir, est avant tout un processus. Ce processus est technique (mettre en place des solutions techniques de réconciliation de données) tout en donnant une place primordiale à la gouvernance de la donnée. L'enjeu de ce mémoire a été d'apporter des éléments de gouvernance de la donnée dans l'optique de répondre aux problématiques que soulève la réconciliation du patrimoine de données clients hétérogènes pour acquérir une vue client unique, une vue « 360° », une seule version de la réalité.

Avoir une vision « métier » et pas uniquement technique ou informatique, nous a semblé primordial. C'est dans cette logique que nous avons décrit les données clients comme des données de références (Master Data). Rappelons que ces dernières sont justement des « objets métiers »[46] qui ont pour caractéristiques d'être « dupliquées » au sein de plusieurs systèmes et d'être échangées avec des tiers au sein de l'organisation.
C'est d'ailleurs au travers de la relation « métiers/systèmes d'information » que nous pensons au sens large le terme de Gouvernance de la donnée.
Nous avons également insisté sur cette « relation métiers/systèmes d'information » dans la constitution du Référentiel Client Unique qui par sa modélisation permet la constitution d'une unique base de données clients, quels que soient les canaux d'entrée et de sortie reliés à une information relative au client. Rappelons que la problématique de l'échelle de partage de ce Référentiel Client Unique est un point important à discuter en amont de sa création en lien avec l'organisation de l'entreprise (selon la subdivision et/ou l'organisation géographique de l'entreprise).

Avant même de s'intéresser à ces éléments propres de gouvernance de la donnée, nous nous sommes intéressés à l'objet métier « client » en lui-même : comprendre ce qu'est un client dans le domaine des transports publics ; présenter les points de contact entre le client et le réseau de transport ; décrire les données clients-voyageurs.
Nous l'avons vu, à l'heure du multi-canal se multiplient les différents points de contact, complexifiant la gestion et la gouvernance des données clients. Se pose ainsi la question de comment consolider et centraliser l'information client portée par des données pour beaucoup encore stockées et administrées « en silo ».

A travers la relation entre une base clients consolidée et le CRM, nous avons porté un regard essentiellement marketing (en tant qu'application de la consolidation d'une base clients étant donné que le marketing est l'un des principaux « métiers » qui traite le plus de l'objet « client »). Cependant, on peut noter d'autres applications d'une base clients consolidée comme l'expérience client (faciliter l'expérience et le parcours dans les transports, étudier le ressenti dans les transports), l'information voyageur personnalisée, le design du réseau de transport, la tarification…

Avec l'étude de cas du réseau du Havre, LiA, nous avons pu concrètement voir et décrire l'ensemble des bases de données qui contiennent de l'information sur les clients. Nous avons, dans un premier temps, établi le lien entre ces bases et la base clients d'origine (issue de la billétique) en schématisant leurs relations à l'aide des *champs structurants*. Puis, dans un second temps, nous avons créé, à l'aide

[46] Comportant une information de base, fondamentale pour l'activité de l'entreprise.

de scripts, des clés-identifiants. A l'aide de ces différentes clés, nous avons réalisé des jointures entre la base clients et les autres bases et calculé le pourcentage de recoupement ou taux de correspondance entre ces bases. Ce pourcentage de recoupement correspond au pourcentage de clients d'une base dans la base Clients.
Nous avons voulu montrer à l'aide de ces recoupements comment les différentes bases après jointures peuvent préfigurer d'un début de vue unique (ou vue « 360° ») en mettant en avant les quelques difficultés rencontrées qu'il faudra prendre en compte lors de consolidation d'une base clients unique :

- La difficulté de non non-correspondance des champs entre les différentes bases de données ;
- La qualité de la donnée.

Avec ce travail, nous n'étions pas encore dans une démarche d'automatisation mais plutôt dans une logique de préfiguration, avant l'instauration d'un véritable processus de consolidation et d'une architecture data appropriés. Cette deuxième phase pourra s'appuyer sur l'ensemble des éléments de gouvernance de la donnée que nous avons présentés dans ce mémoire.

Nous avons pu montrer en effet l'intérêt d'une telle démarche de consolidation d'une base clients notamment avec la reconstitution du parcours temporel complet de quelques clients au travers de l'ensemble de ses contacts avec le réseau LiA (*cf.* Graphique 10) préfigurant une vision « 360° » dans une logique de CRM.

Nous avons aussi décrit plus en détail l'utilité de la consolidation d'une base clients avec l'exemple du transport à la demande. Alors qu'actuellement quasiment aucune information sur les utilisateurs de ce service de mobilité n'est remontée, nous avons pu via les recoupements avec la base clients réconciliée (manuellement) obtenir des informations socio-économiques de base (âge, sexe) permettant de mieux cibler les clients utilisant ce service.
Ce genre de démarche pouvant tout à fait se réaliser pour d'autres modes de mobilité.

Pour finir, ajoutons également que deux points particuliers de la gouvernance de la donnée sont essentiels tout au long de ces processus, de ces démarches de consolidation de bases clients : la qualité de la donnée, tout particulièrement la « résolution des identités », et la problématique de la protection des données personnelles. Nous avons donc présenté ces deux enjeux de manières théoriques au début de ce mémoire pour ensuite les évoquer dans le cas concret du réseau de transport du Havre.

Bibliographie

Ouvrages :

ARROYO, Jean-Philippe, BRUNET, Sylvie, SAGE, Roselyne. *RGPD et Marketing, De la contrainte à l'opportunité*. Paris : e-theque, 2018, 267 p.

BELVAUX, Bertrand, NOTEBAERT, Jean-François. *Crosscanal et Omnicanal : La digitalisation de la relation client*. Paris : Dunod, 2018, 192 p.

BERTI-EQUILLE, Laure. *La qualité et la gouvernance des données : au service de la performance des entreprises*. Paris : Hermes/Lavoisier, 2012, 388 p.

BLONDEAU, Géraldine, DIGOUT, Jacques, ROUALDES, Emmanuelle. *Relation client / CRM*. Paris : Vuibert, 2015, 201 p.

BONNET, Pierre. *Management des données de l'entreprise : Master Data Management et modélisation sémantique*. Paris : Hermes/Lavoisier, 2009, 286 p.

DELERS, Antoine. *CRM : La gestion de la relation client*. Paris : 50 Minutes, 2015, 70 p.

HIRTH, Judith. Le Data marketing. Paris : Eyrolles, 2017, 295 p.

JALLAT, Frédéric, PEELEN, Ed, STEVENS, Eric, VOLLE, Pierre. *Gestion de la relation client Expérience client, Performance relationnelle et Hub relationnel.* 5ème éd. France : Pearson, 2018, 547 p.

LOSHIN, David. *Master Data Management*. Burlington : Morgan Kaufman, 2010, 304 p.

MATTATIA, Fabrice. RGPD et droit des données personnelles : enfin un manuel complet sur le nouveau cadre juridique issu du RGPD et de la loi Informatique et Libertés de 2018. 3ème éd. Paris : Eyrolles, 2018, 235 p.

REGNIER-PECASTAING, Franck, FINET, Jacques, GABASSI, Michel. MDM, *Enjeux et méthodes de la gestion des données*. Paris : Dunod, 2008, 336 p.

Articles

BELLOT, Patrice, ESPINASSE, Bernard. Introduction au Big Data - Opportunités, stockage et analyse des mégadonnées. *Techniques de l'Ingénieur*, 10 février 2017

BERTI-ÉQUILLE, Laure. Qualité des données. *Techniques de l'Ingénieur*, 10 octobre 2018

Thèses

Elodie Castex. Le Transport A la Demande (TAD) en France : de l'état des lieux à l'anticipation. Modélisation des caractéristiques fonctionnelles des TAD pour développer les modes flexibles de demain. Géographie. Université d'Avignon, 2007. Français. fftel-00199865v2
Disponible sur : https://tel.archives-ouvertes.fr/tel-00199865v2

Rapports

CIGREF. *Valorisation des données dans les grandes entreprises. Maturité, pratiques et modèles*. Novembre 2016
Consultable sur : https://www.cigref.fr/wp/wp-content/uploads/2016/11/CIGREF-Valorisation-des-donnees-Pratiques-Modele-2016.pdf

Délégation à l'aménagement du territoire et à l'action régionale ; FRANCE. Ministère de l'équipement, des transports, de l'aménagement du territoire, du tourisme et de la mer. Direction des transports terrestres. *Services à la demande et transports innovants en milieu rural : de l'inventaire à la valorisation des expériences*. Délégation à l'aménagement du territoire et à l'action régionale, Novembre 2004, 311 p.
Consultable sur : https://www.ladocumentationfrancaise.fr/rapports-publics/054000165/index.shtml

INFOLAB-FING. *Nouvelles efficacités et création de valeur : les projets de gouvernance de la donnée*. Cahier n°3, Janvier 2017
Consultable sur : https://infolabs.io/gouv16

MARIKO, Dominique. Le Master Data Management (MDM) et la qualité des données de l'entreprise : synergies digitales et collaboratives. *INTD-CNAM,* 5 avril 2016

TRIGAUX, Jean-Christophe. Master Data Management-Mise en place d'un référentiel de données. 2009/TRIM4/01. *Smals Research*, Décembre 2009

Sites d'internet

BATHELOT, Bertrand. Définition : Cycle de vie client. **In** https://www.definitions-marketing.com/definition/cycle-de-vie-client/

BATHELOT, Bertrand. Définition : Connaissance client. **In** https://www.definitions-marketing.com/definition/connaissance-client/

BATHELOT, Bertrand. Définition : Référentiel client unique. **In** https://www.definitions-marketing.com/definition/referentiel-client-unique/

Commission Nationale Informatique et Liberté. La prospection commerciale par courrier électronique **In** https://www.cnil.fr/fr/la-prospection-commerciale-par-courrier-electronique

Liste des graphiques

Graphique 1 : Cartographie des points de contacts des clients avec un réseau de transport p.8

Graphique 2 : Les trois niveaux de la « maturité client » p.10

Graphique 3 : Modélisation relationnelle entre les BDD Client, Carte et Validations p.13

Graphique 4 : Flux de données entre la Base clients consolidées et le CRM p.18

Graphique 5 : Les différentes échelles de l'organisation de Transdev p.30

Graphique 6 : Les contacts avec le client sur le réseau LiA p. 35

Graphique 7 : Représentation des recoupements à partir de la base clients issue de la billettique p. 36

Graphique 8 : Extrait du code R pour la génération de clés dans les différentes bases (exemple de la base Clients d'origine) p. 37

Graphique 9 : Les différentes étapes de recoupement (par deux jointures) entre la base SMS et la base clients p. 39

Graphique 10 : Suivi personnalisé du parcours du client n°2503 à travers l'ensemble du réseau LiA p. 40

Graphique 11 : Taux de recoupement entre la base TAD et la base clients p. 42

Graphique 12 : Architecture des recoupements entre les différentes bases de données et la base clients avec la mise à l'écart des homonymes p. 46

Graphique 13 : Data Warehouse sur le Cloud relié à l'ETL et au MDM p.48

Graphique 14 : Présentation de l'architecture centralisée « CRM 360° » p.49

Liste des tableaux

Tableau 1 : Dimensions de la connaissance client p.10

Tableau 2 : Anomalies liées aux identités p.26

Tableau 3 : Données à caractère personnel dans le domaine de la mobilité p.27

Tableau 4 : Clés qui ont pu être crées (en vert) dans les différentes bases p. 38

Tableau 5 : Bases clients et cadre juridique de l'exécution du contrat p. 47

Annexes

Annexe A : Implantation de Transdev dans le monde

Carte de l'ensemble des pays où Transdev est présent

Annexe B : Quelques chiffres sur le réseau LiA

Le réseau LiA :

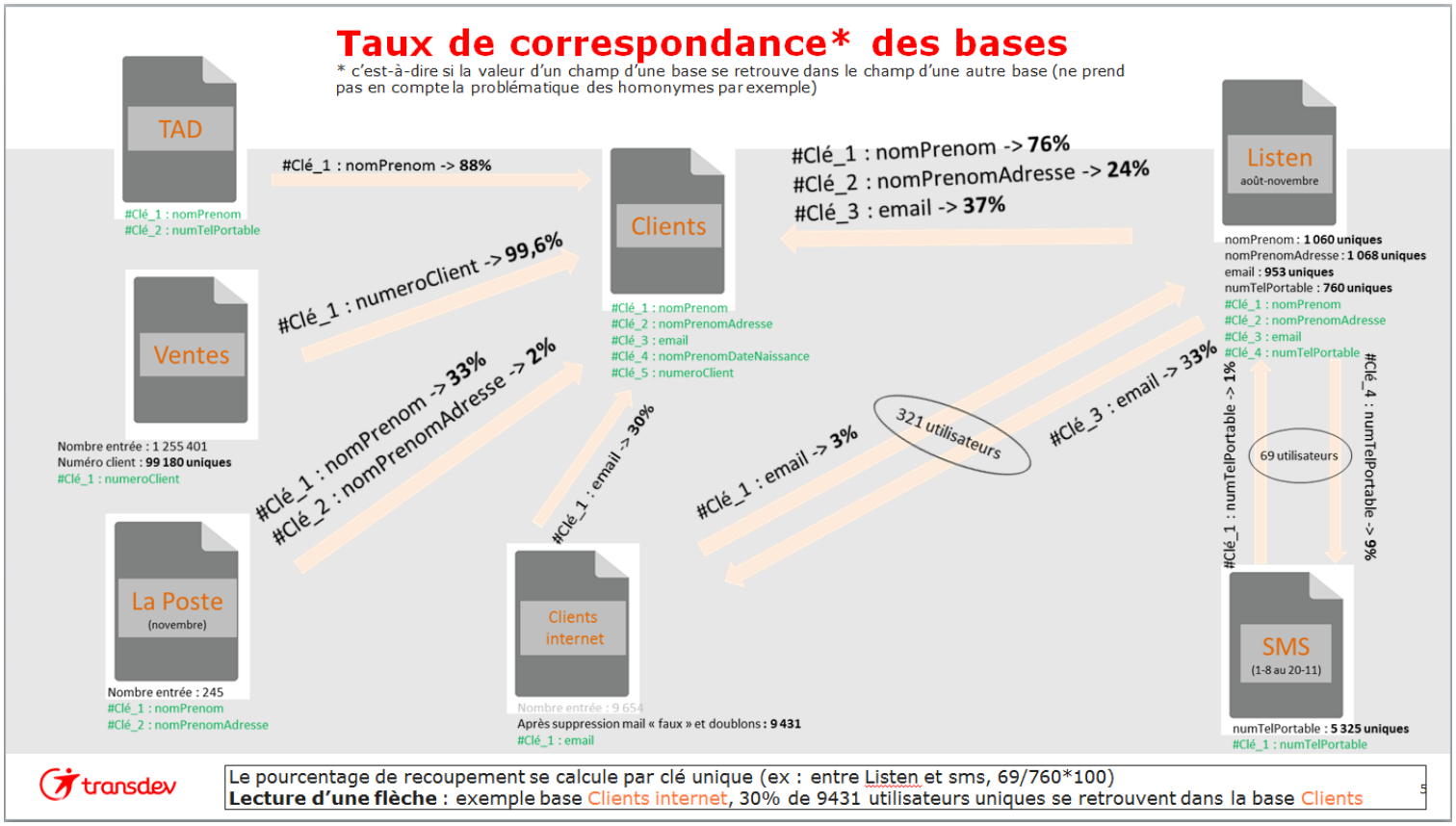
- 2 lignes de Tramway A et B
- 21 lignes de bus
- des lignes complémentaires : 30, 31, 40, 41, 50, 60, 70, 71, 80, 90, 91
- le funiculaire
- le service de transport à la demande Fil'Bus
- le service de transport de personnes à mobilité réduite Mobi'Fil
- Le service de nuit à la demande toute l'année : LiA de Nuit
- La desserte innovante de la Zone Industrielle et Portuaire : FlexiLiA
- La ligne de train à tarification urbaine, exploitée par la SNCF : la LER
- Le service de location de vélos : LiAvélos
- 2 parkings relais (P+R)
- 12 parcs à vélos (P+V)

Les chiffres clefs :

- 611 personnes dont 419 conducteurs-receveurs
- 138 bus et 22 rames de tramway
- plus de 79 000 voyages par jour
- 28 millions de voyages
- 9,26 millions de kilomètres parcourus par an
- 970 points d'arrêt (dont 278 abribus et 23 stations tramway)

Source : https://www.lehavreseinemetropole.fr/article/r%C3%A9seau-lia

Annexe C : Taux de correspondance entre les bases



Etat des lieux des recoupements entre les différentes bases de données et la base Clients d'origine (extrait de l'étude réalisée par la Direction Data Science, 2018)