POLYTECH LYON

Documentation technique du Data Warehouse

PROJET BUSINESS INTELLIGENCE

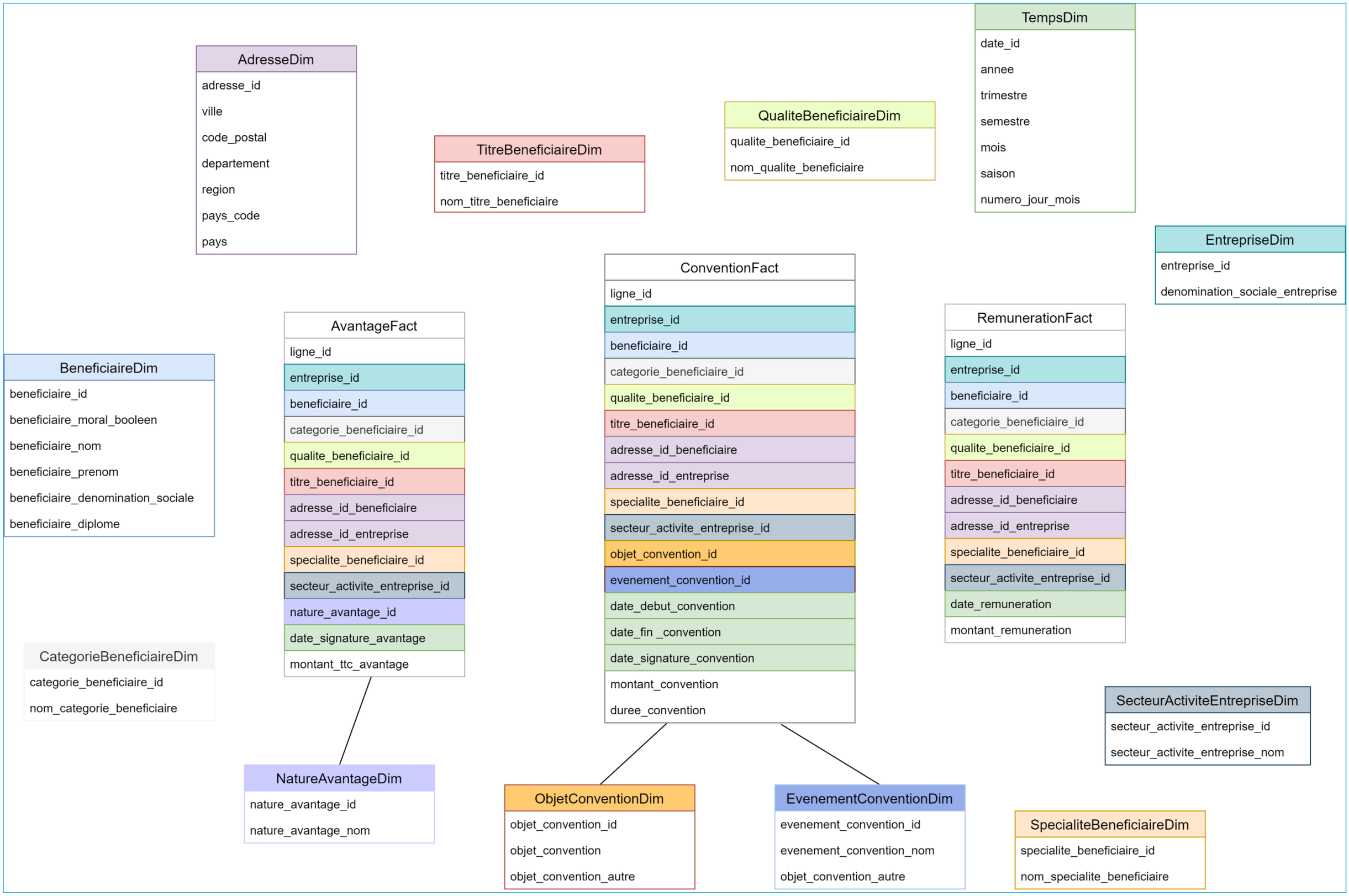
BENALI Myriam, BESSON Cécile, CARDOSO MORENO Ineida, NAAJI Dorian, VIRARAGAVANE Smaïline

POLYTECH LYON

5A INFO GROUPE 2

1. INTRODUCTION ET MODÉLISATION

Ce document constitue la documentation technique de la conception de l'entrepôt de données réalisé lors du projet de Business Intelligence. Voici le diagramme de l'entrepôt de données :



2. SPÉCIFICATION DES DONNÉES DU DW

2.1. AvantageFact

Nom colonne DW	Fichier initial	Description	Type
ligne_id	avantages (ligne_identifiant)	L'id de la ligne tel que trouvé dans les fichiers de base.	Texte
entreprise_id	entreprises (identifiant)	L'id de l'entreprise tel que trouvé dans les fichiers de base.	Texte
beneficiaire_id	avantages (benef_identifiant_valeur) ou généré	L'id du bénéficiaire qui est répertorié dans la dimension BeneficiaireDim. (RPPS ou numéro d'ordre des médecins pour une personne physique, numéro de SIREN pour une entreprise ou numéro généré à partir de différentes informations si pas de SIREN/RPPS)	Texte
categorie_beneficiaire_id	avantages (benef_titre_code)	Code de la catégorie du bénéficiaire qui est répertorié dans la dimension CategorieBeneficiaireDim.	Texte
qualite_beneficiaire_id	avantages (benef_qualite_code)	Code de la qualité du bénéficiaire qui est répertorié dans la dimension QualiteBeneficiaireDim.	Texte
titre_beneficiaire_id	avantages (benef_titre_code)	Code du titre du bénéficiaire se trouvant dans la dimension TitreBeneficiaireDim.	Texte
adresse_id_beneficiaire	n°insee ou généré	Id de l'adresse du bénéficiaire associé à cet avantage	Nombre ou Texte
adresse_id_entreprise	n°insee ou généré	Id de l'adresse de l'entreprise associé à cet avantage	Nombre ou Texte
specialite_beneficiaire_id	avantages (benef_specialite_code)	Id de la spécialité du bénéficiaire qui est répertorié dans la dimension SpecialiteBeneficiaireDim Code correspondant à la spécialité du bénéficiaire concerné par l'avantage.	Texte
secteur_activite_entreprise_id	entreprises (secteur_activite_code)	Id du secteur d'activité de l'entreprise rattaché à l'avantage en question. Il est répertorié dans la dimension SecteurActiviteEntrepriseDim	Texte
nature_avantage_id	généré	Id de la nature de l'avantage en question. Il est répertorié dans la dimension NatureAvantageDim	Texte
date_signature_avantage	avantages (avant_date_signature)	La date où l'avantage a été accordé	Date
montant_ttc_avantage	avantages (avant_montant_ttc)	Le montant en TTC pour payer l'avantage du bénéficiaire	Nombre

2.2. ConventionFact

Nom colonne DW	Fichier initial	Description	Type
ligne_id	conventions (ligne_identifiant)	L'id de la ligne tel que trouvé dans les fichiers de base.	Texte
entreprise_id	entreprises(identifiant)	L'id de l'entreprise tel que trouvé dans les fichiers de base.	Texte
beneficiaire_id	conventions (benef_identifiant_valeur) ou généré	L'id du bénéficiaire qui est répertorié dans la dimension BeneficiaireDim. (RPPS ou numéro d'ordre des médecins pour une personne physique, numéro de SIREN pour une entreprise ou numéro généré à partir de différentes informations si pas de SIREN/RPPS)	Texte
categorie_beneficiaire_id	conventions (benef_titre_code)	Code de la catégorie du bénéficiaire qui est répertorié dans la dimension CategoryBeneficiaireDim.	Texte
qualite_beneficiaire_id	conventions (benef_qualite_code)	Code de la qualité du bénéficiaire qui est répertorié dans la dimension QualiteBeneficiaireDim.	Texte
titre_beneficiaire_id	conventions (benef_titre_code)	Code du titre du bénéficiaire se trouvant dans la dimension TitreBeneficiaireDim.	Texte
adresse_id_beneficiaire	n°insee ou généré	Id de l'adresse du bénéficiaire associé à cette convention	Nombre ou Texte
adresse_id_entreprise	n°insee ou généré	Id de l'adresse de l'entreprise associée à cette convention	Nombre ou Texte
specialite_beneficiaire_id	conventions (benef_specialite_code)	Id de la spécialité du bénéficiaire qui est répertorié dans la dimension SpecialiteBeneficiaireDim Code correspondant à la spécialité du bénéficiaire concerné par la convention.	Texte
secteur_activite_entreprise_id	conventions (secteur_activite_code)	Id du secteur d'activité de l'entreprise rattaché à la convention en question. Il est répertorié dans la dimension SecteurActiviteEntrepriseDim	Texte
objet_convention_id	généré	Id de l'objet de la convention qui est répertorié dans la dimension ObjetConventionDim	Texte
evenement_convention_id	généré	Id de l'évènement lié à la convention qui est répertorié dans la dimension EvenementConventionDim	
date_debut_convention	conventions (conv_date_debut)	Date de début de la convention	

date_fin _convention	conventions conv(date_fin)	Date de fin de la convention	
date_signature_convention	conventions (conv_date_fin)	Date de la signature de la convention	
montant_convention	conventions (conv_montant_ttc)	Montant des dépenses liées à cette convention (ex : rémunération des acteurs de santé …)	
duree_convention	calculé	La durée de la convention (en jours). -1 si la convention n'est toujours pas terminée.	Nombre

2.3. RemunerationFact

Nom colonne DW	Fichier initial	Description	Type
ligne_id	remunerations (ligne_identifiant)	L'id de la ligne tel que trouvé dans les fichiers de base.	Texte
entreprise_id	entreprises(identifiant)	L'id de l'entreprise tel que trouvé dans les fichiers de base.	Texte
beneficiaire_id	remunerations (benef_identifiant_valeur) ou généré	L'id du bénéficiaire qui est répertorié dans la dimension BeneficiaireDim. (RPPS ou numéro d'ordre des médecins pour une personne physique, numéro de SIREN pour une entreprise ou numéro généré à partir de différentes informations si pas de SIREN/RPPS)	Texte
categorie_beneficiaire_id	remunerations (benef_titre_code)	Code catégorie du bénéficiaire qui peut être trouvé dans dimension CategorieBeneficiaireDim	Texte
qualite_beneficiaire_id	remunerations (benef_qualite_code)	Code qualité du bénéficiaire se trouvant dans la dimension QualiteBeneficiaireQualiteDim	Texte
titre_beneficiaire_id	remunerations (benef_titre_code)	Code titre du bénéficiaire se trouvant dans la dimension TitreBeneficiaireDim	Texte
adresse_id_beneficiaire	n°insee ou généré	Id de l'adresse du bénéficiaire correspondant à la convention en question	Nombre ou Texte
adresse_id_entreprise	n°insee ou généré	Id de l'adresse de l'entreprise correspondant à la convention en question	Nombre ou Texte
specialite_beneficiaire_id	remunerations (benef_specialite_code)	Id de la spécialité du bénéficiaire qui est répertorié dans la dimension SpecialiteBeneficiaireDim	Texte
secteur_activite_entreprise_id	remunerations (ligne_identifiant)	Id du secteur d'activité de l'entreprise rattaché à la convention en question. Il est répertorié dans la dimension SecteurActiviteEntrepriseDim	Texte
date_remuneration	remunerations (remu_date)	La date de la rémunération	Date
montant_remuneration	remunerations (remu_montant_ttc)	Le montant de la rémunération	Nombre

2.4. BeneficaireDim

Nom colonne DW	Fichier initial	Description	Type
beneficiaire_id	benef_identifiant_valeur	Cette colonne correspond à l'id du bénéficiaire qui est un acteur de santé. Si c'est une personne morale alors cet id correspond au numéro de SIREN de l'entreprise. Si c'est une personne physique, il correspond au numéro RPPS/Ordre de Médecin ou Autre. Généré si « Autre ».	Texte
beneficiaire_moral_booleen	calculé	Ce booléen permettra de savoir si l'acteur de santé en question est une personne morale.	Booléen
beneficiaire_nom	benef_nom	Nom du bénéficiaire en question (N/A si c'est une entreprise (personne morale))	Texte
beneficiaire_prenom	benef_prenom	Le prénom du bénéficiaire en question (N/A si c'est une entreprise (personnes morale))	Texte
beneficiaire_denomination_sociale	benef_denomination_sociale	La dénomination sociale du bénéficiaire si c'est une entreprise (N/A si c'est une personne physique)	Texte
beneficiaire_diplome	rpps_diplome.	L'intitulé du diplôme du bénéficiaire (N/A si c'est une entreprise, « Inconnu » si numéro RRPS introuvable ou si pas de numéro).	Texte

2.5. TitreBeneficiaireDim

Nom colonne DW	Colonnes fichier initial	Description	Type
titre_beneficiaire_id	benef_titre_code	Cette colonne correspond à l'id du titre du bénéficiaire, soit un code de plusieurs lettres entre crochets représentants le titre du bénéficiaire.	Texte
nom_titre_beneficiaire	benef_titre_libelle	Titre du bénéficiaire en question Ex : Docteur, Pharmacien, Professeur, etc.	Texte

2.6. QualiteBeneficaireDim

Nom colonne DW	Colonnes fichier initial	Description	Type
qualite_beneficiaire_id	benef_qualite_code	Cette colonne correspond à l'id de la qualité du bénéficiaire, soit un code de deux chiffres entre crochets représentants la qualité du bénéficiaire.	Texte
nom_qualite_beneficiaire	qualite	Qualité du bénéficiaire Ex: Médecin, infirmier, chirurgien-dentiste, etc.	Texte

2.7. TempsDim

Nom colonne DW	Colonnes fichier initial	Description	Type
date_id	généré	La date en question	Date sous le format : JJ/MM/YYYY
annee	généré	L'année de la date concernée	Nombre
trimestre	calculé	Le trimestre auquel correspond la date concernée	Texte
semestre	calculé	Le semestre auquel correspond la date concernée	Texte
mois	généré	Le mois de la date concernée	Texte
saison	calculé	La saison de la date concernée	Texte
numero_jour_mois	généré	Le numéro du jour de la date concernée	Nombre

2.8. AdresseDim

Nom colonne DW	Colonnes fichier initial	Description	Type
adresse_id	communes-departement-region ou généré	L'id représente un entier qui est associé à une adresse. On récupérera le numéro INSEE dans communes-departement-region.csv ou alors on générera un id.	Nombre
ville	nom_commune (fichier communes-departement-region)	La ville de l'adresse	Texte
code_postal	code_postal (fichier communes-departement-region)	Le code postal de l'adresse	Nombre
departement	departement (fichier dep_region)	Le département de l'adresse	Texte
region	regionname (fichier dep_region)	La région de l'adresse	Texte
pays_code	benef_pays_code	Code correspondant au pays de l'adresse Ex : [FR] pour France	Texte
pays	pays	Pays de l'adresse	Texte

2.9. EntrepriseDim

Nom colonne DW	Colonnes fichier initial	Description	Type
entreprise_id	entreprise_identifiant	Numéro fourni par le fichier entreprises.csv	Nombre
denomination_sociale_entreprise	denomination_sociale	Dénomination sociale de l'entreprise	Texte

2.10.SecteurActiviteEntrepriseDim

Nom colonne DW	Colonnes fichier initial	Description	Type
secteur_activite_entreprise_id	secteur_activite_code (fichier entreprises)	Le code du secteur d'activité tel que fourni par le fichier	Texte
secteur_activite_entreprise_nom	secteur (fichier entreprises)	Le libellé du secteur d'activité tel que fourni par le fichier	Texte

2.11.SpecialiteBeneficiaireDim

Nom colonne DW	Colonnes fichier initial	Description	Type
specialite_beneficiaire_id	benef_speicalite_code	Le code de la spécialité tel que fourni par le fichier. (attention faute de frappe dans le nom de colonne du fichier initial) Ex : [SMR7], [SCD03]	Texte
nom_specialite_beneficiaire	benef_speicalite_libelle	Le libellé de la spécialité tel que fourni par le fichier. (attention faute de frappe dans le nom de colonne du	Texte

		fichier initial) Ex : Médecine interne, Médecine Bucco-Dentaire	

2.12.EvenementConventionDim

Nom colonne DW	Colonnes fichier initial	Description	Type
evenement_convention_id	généré	Un id généré en auto_increment ou à partir du texte de evenement_convention_nom	Nombre / Texte
evenement_convention_nom	conv_manifestation_nom (fichier conventions)	Le type de manifestation encadrée par la convention Ex : Congrès, formation, etc.	Texte

2.13.ObjetConventionDim

Nom colonne DW	Colonnes fichier initial	Description	Type
objet_convention_id	généré	Un id généré en auto_increment ou à partir du texte de l'objet_convention	Nombre / Texte
objet_convention	conv_objet (fichier conventions)	La raison pour laquelle on a réalisé la convention. Ex : Hospitalisation, Formation, etc.	Texte
objet_convention_autre	conv_objet_autre (fichier conventions)	Le champ est rempli si conv_objet = Autre. Il correspond au détail de l'objet de la convention. N/A si conv_objet est différent de : Autre.	Texte

2.14.NatureAvantageDim

Nom colonne DW	Colonnes fichier initial	Description	Type
nature_avantage_id	généré	Un id généré en auto_increment ou à partir du texte de avant_nature	Nombre / Texte
nature_avantage_nom	avant_nature (fichier avantages)	Libellé des avantages accordés des professionnels de santé participant dans les conventions. Ex : Hébergement, Repas, etc.	Texte

2.15. CategorieBeneficiaireDim

Nom colonne DW	Colonnes fichier initial	Description	Type
categorie_beneficiaire_id	benef_categorie_code	Le code de la qualité du bénéficiaire tel que fourni par le fichier. Ex : [PRS] pour Professionnel de santé, etc.	Texte
nom_categorie_beneficiaire	categorie	Libellé des catégories des professionnels de santé participant dans les conventions. Ex : Professionnel de santé, Étudiant, etc.	Texte

3. REMARQUES DIVERSES CONCERNANT LA PHASE D'ETL

- Les données sont de manière générale très hétérogènes. Il faudra porter un soin particulier au remplissage des dimensions et les moyens de réaliser par la suite les jointures pour le remplissage de la table des faits.
- Entre la date de signature de la convention et la date des paiements / attributions des avantages, il peut y avoir des changements d'adresse, pour une même personne (qu'elle soit physique ou morale). Il faudrait dans ce cas pour chaque fichier repasser sur l'ensemble des lignes et mettre à jour les lignes en gardant l'adresse la plus récente.
- Porter attention aux erreurs de saisie, traits d'unions, accents, ponctuation, etc.
- Pour les bénéficiaires qui n'ont pas d'id (RPPS/SIREN/ORDRE), il faut générer un id. Une hypothèse serait de générer un ID en concaténant plusieurs champs, comme par exemple NOM+PRENOM+SPECIALITE+CP et voir si des doublons peuvent apparaître ou non. Les bénéficiaires n'ayant pas de numéro RPPS/SIREN/ORDRE restent des cas isolés dans l'ensemble. On peut utiliser BENEF_DENOMINATION_SOCIALE+CP si c'est une personne morale.
- Les rémunérations versées avant le 31 décembre 2016 sont situées dans le fichier des avantages, sous le nom (avant_nature) « HONORAIRES » ou « AUTRES : [COUTS PAR CENTRE] ». Il faudra veiller à cela.
- Disparité dans la saisie des villes (ponctuation, CEDEX, etc.)

De manière générale, les données utilisées sont très hétérogènes. Il faudra donc veiller à normaliser, nettoyer l'ensemble des données afin d'obtenir un *data warehouse* le plus propre possible. L'analyse des valeurs au sein des fichiers est indispensable pour la phase d'ETL. L'utilisation de fichier minimisés pour des filtres et tests est fortement recommandée. Liens utiles :

- https://www.childs.be/blog/post/split-and-extract-the-first-1000-n-rows-from-a-text-csv-data-file-in-windows : split and extract first n rows
- https://gofile.io/d/loEEoP : classeur excel : doc + top 10 000 rows.

4. INDICATEURS CLÉS DE PERFORMANCE

4.1. Avantages

- Quelles sont les villes/pays où les entreprises font le plus appel à des acteurs du secteur de la santé en échange d'avantages ?
- Quelles sont les 5 premières professions de santé les plus demandées par les entreprises ?
- Quelles sont les 5 premières spécialités les plus recherchées par les entreprises ?
- Quel est en moyenne le montant des frais engendrés pour payer l'avantage au bénéficiaire ?
- Ce montant est-il plus ou moins élevé selon la ville, le poste, la spécialité du bénéficiaire ?
- Quels sont les principaux avantages accordés par les entreprises ? Certaines entreprises font-elles plus appel à des prestataires que d'autre en échange d'avantages ?
- Quelle est le montant moyen des avantages en fonction des régions/départements/villes ?

4.2. Conventions

- Quelle est la durée moyenne d'une convention ?
- Certaines entreprises établissent-elle significativement plus de conventions que d'autres ?
- Y-t-il un lien entre le nombre de convention établies entre une entreprise et une prestation selon la ville ou la spécialité ?
- Quel est le montant moyen d'une convention ? En fonction des villes, régions, départements ?
- Le moment de l'année influence-t-il la signature des conventions ?
- De quel secteur sont issues les entreprises signant le plus de conventions ? Nombre de conventions signées en fonction du secteur d'entreprise.
- Quels sont les objets de convention les plus récurrents ? Les plus chers ?
- Zone géographique des bénéficiaires/entreprises en fonction du montant moyen de la convention

4.3. Pour les rémunérations :

- Quel est le montant moyen d'une rémunération ?
- Certains professionnels sont-ils mieux rémunérés que d'autres selon la ville, leur poste, leur spécialité ?
- Certaines entreprises font-elles plus appel à des prestataires que d'autre en échange d'une rémunération ?
- Quelle est la rémunération moyenne en fonction des régions/départements/villes ?
- Quel secteur d'entreprise fournit les plus grosses rémunérations ? les plus faibles ? montant moyen des rémunérations en fonction du secteur ? en fonction de la qualité du bénéficiaire ?
- La date / moment de l'année / saison influence-t-elle le montant des rémunérations ?

Ces KPIs sont donnés à titre informatif.