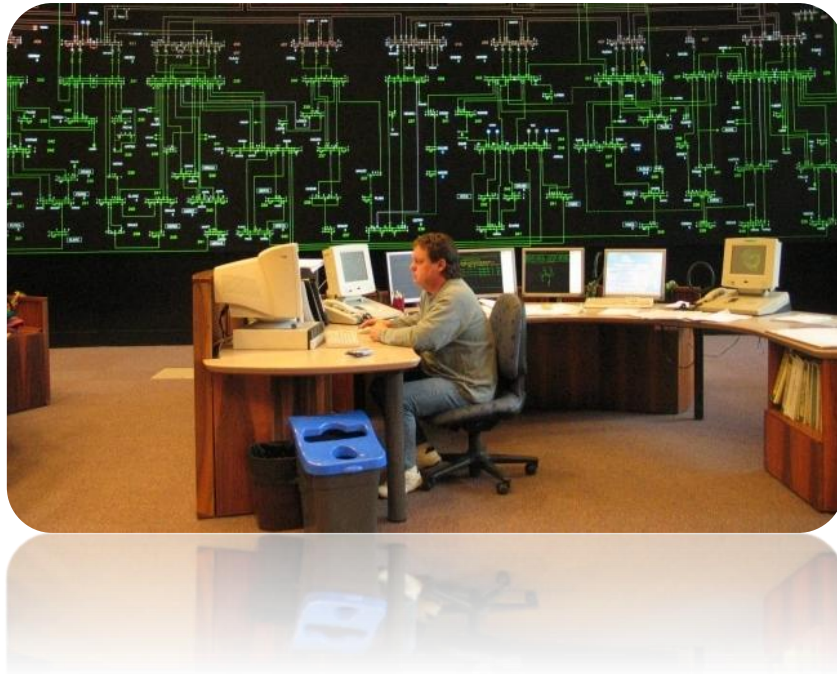


## SAE Intégration DW

-

« Un jour dans la vie... *d'un gestionnaire de réseau électrique* »



Terryl HASSEN  
William LEFEBVRE  
Makam TRAORE

## Table des matières

Présentation du projet .....	3
Contexte .....	3
Problématique .....	3
Objectifs .....	3
Modélisation du SID .....	4
Choix du processus métier à modéliser .....	4
Définition de la granularité.....	4
Choix des dimensions .....	4
Identification des faits (Mesures) .....	5
Intégration de données (Processus ETL).....	6
Description du processus d'extraction (Extract) .....	6
Transformation des données (Transform) .....	6
Transformations préalables .....	6
Chargement des données (Load) .....	9
Transformation des tables .....	9
Création des tables dimensions à partir des sources .....	11
Création et utilisation de tables de temps .....	11
Création des mesures (fait) avec DAX .....	12
<b>Description des modèles de données utilisés. ....</b>	<b>12</b>
<b>Restitution</b> .....	<b>15</b>
Tableaux de bord .....	16
Annexes.....	17

# Présentation du projet

<https://github.com/Lalmytox/SAE-SID>

## Contexte

Ce projet s'inscrit dans le cadre d'un enseignement lié au développement de systèmes d'information décisionnels (SID) pour analyser des données massives. Nous avons choisi de traiter des données concernant la consommation et la production énergétique en France, ainsi que les échanges commerciaux transfrontaliers d'électricité.

Les données utilisées couvrent plusieurs années et sont issues de sources variées : consommation régionale d'énergie, production annuelle d'électricité par filière, et import/export d'électricité entre la France et ses voisins. Ces données, publiées sous licences ouvertes, permettent de mener des analyses approfondies pour mieux comprendre les dynamiques énergétiques et commerciales sur le territoire.

## Problématique

Face aux enjeux énergétiques et climatiques, il est crucial de comprendre comment les données liées aux flux énergétiques peuvent être exploitées pour optimiser la gestion des ressources.

Pour ce faire, le gestionnaire du réseau doit avoir à disposition des bases de données lui permettant de réaliser simplement des analyses.

**Comment modéliser un Entrepôt de données à destination d'un gestionnaire de réseau électrique ?**

## Objectifs

Le projet vise à répondre à cette problématique en développant un système d'information décisionnel permettant :

- D'intégrer des données hétérogènes provenant de différentes sources, tout en assurant leur qualité et leur cohérence.
- D'élaborer une modélisation adaptée au processus métier étudié, en définissant la granularité et les dimensions pertinentes pour les analyses énergétiques et commerciales.
- De produire des indicateurs pertinents : grâce à des mesures dynamiques créées avec des requêtes DAX et d'autres outils d'analyse de Power BI
- De restituer les résultats sous forme de tableaux de bord interactifs et de rapports détaillés, permettant une prise de décision éclairée.
- De démontrer l'impact potentiel des données analysées dans des domaines tels que la gestion énergétique et les échanges commerciaux.

# Modélisation du SID

## Choix du processus métier à modéliser

Le processus métier sélectionné pour ce projet est **l'analyse des flux d'énergie électrique** en France. Ce processus pourra être divisé en sous-processus :

- La consommation d'électricité par région et secteur d'activité.
- Les échanges commerciaux d'électricité entre la France et ses pays voisins.
- La production électrique par filière (nucléaire, éolien, solaire, hydraulique, etc.).

L'analyse de ce processus permet de répondre à des problématiques stratégiques telles que l'optimisation de la consommation énergétique, la prévision des besoins futurs et l'évaluation des échanges commerciaux avec d'autres pays.

## Définition de la granularité

La granularité des données détermine les différentes variables utilisées pour répondre aux objectifs du projet et le niveau de détail souhaité dans le modèle de données :

### Granularité temporelle :

- Grain annuel : Pour la consommation et la production énergétique, des analyses globales sur plusieurs années permettront de dégager des tendances.
- Grain horaire : Les échanges commerciaux d'électricité nous donnent la possibilité d'avoir une granularité plus fine. Ils sont donc analysés au pas horaire ce qui nous permettra d'identifier des variations temporelles précises.

### Granularité géographique :

- Grain régional : Les données de consommation et de production seront organisées par région.

Ces niveaux de granularité permettent d'adapter les analyses aux différentes dimensions temporelles et géographiques.

## Choix des dimensions

Les dimensions sélectionnées structurent le SID et offrent un cadre pour explorer les données :

- Dimension temporelle : Année, mois, jour, heure (selon les tables).
- Dimension géographique : Régions (codes et noms), pays (pour les échanges).
- Dimensions énergétique :

- Filière de production (éolien, solaire, hydraulique, nucléaire, bioénergies).
- Catégorie de consommation (secteur résidentiel, industriel, tertiaire).
- Type d'échange (import/export).

## Identification des faits (Mesures)

Les faits sont les données quantitatives qui alimentent les analyses.

La mesure que nous souhaitons analysées et qui sera au centre de notre modèle de données sera la **quantité d'énergie échangée** (en MWh) au sein du réseau électrique français.

Cette quantité d'énergie pourra être de différent type (consommation, production, import, export) et être analysée sous différentes dimensions. Mais quel que soit son type ou son axe d'analyse, cette quantité d'énergie continuera de décrire les flux électriques français.

## Intégration de données (Processus ETL)

Cette partie décrira l'ensemble du processus que nous avons effectué pour passer de nos sources de données à un modèle de données convenable.

Nous décrirons dans un premier temps l'Extraction des données en mentionnant la provenance et la qualité des données, puis la Transformation en détaillant le nettoyage et l'analyse de la qualité des données, pour finir avec Chargement des données vers une modèle de données créé à cet effet.

### Description du processus d'extraction (Extract)

Les données proviennent de 3 sources différentes :

- Imports et exports commerciaux (2005 à 2021) [\[lien\]](#)  
Cette base décrit les échanges commerciaux d'électricité entre la France et ses pays voisins de 2005 à 2021.
- Consommation annuelle d'électricité et gaz par région [\[lien\]](#)  
Cette base décrit la consommation d'électricité par région et secteur d'activité de 2011 à 2023.
- Production régionale annuelle par filière (2008 à 2023) [\[lien\]](#)  
Cette base la production électrique par filière de 2008 à 2023.

Pour plus d'information sur le contenu de ces tables de données, un document descriptif complet est disponible sur le GitHub.

Les 3 sources de données ont été téléchargées sur les différentes plateformes au format csv. A partir de ces fichiers obtenus (voir les fichiers sources sur GitHub), nous avons pu les importer sur PowerBi pour réaliser les étapes de transformations.

### Transformation des données (Transform)

Sur PowerBi, à l'aide de PowerQuery, nous avons pu réaliser différentes transformations. Dans un premier temps nous réalisons des transformations préalables pour s'assurer d'avoir des données de qualité. Puis nous réalisons d'autres transformations dans l'objectif de réaliser un modèle de données convenable seront réalisées dans la partie « Load ».

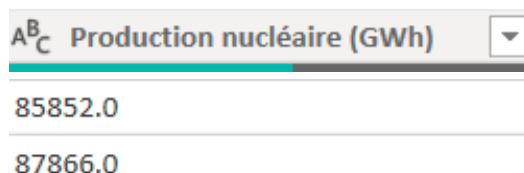
#### Transformations préalables

Cette première série de transformations nous permet de nous assurer d'avoir des données de qualité pour la suite du projet.

Afin de ne pas surcharger ce document, ces étapes seront décrites ici de manière assez succincte. Un document plus détaillé est disponible sur le GitHub.

### Modification du type des variables

Les fichiers plats étant des fichiers csv, les valeurs numériques sont indiquées dans le document avec des points plutôt que des virgules. Ces points sont remplacés et les colonnes sont passées en type décimal.



AB<sub>C</sub> Production nucléaire (GWh) ▼

85852.0

87866.0

Le . nécessite d'être remplacé par une , pour changer le type de la variable

### Gestion des erreurs et des données manquantes

La qualité des données est vérifiée grâce aux options *qualité de la colonne* et *profil de la colonne* disponible dans l'onglet transformation des données de PowerBi.

Aucune erreur n'apparaît lors de l'import des données cependant, des données manquantes apparaissent.

AB <sub>C</sub> Column11	AB <sub>C</sub> Column12	AB <sub>C</sub> Column13
<div><div>● Valide</div><div>● Erreur</div><div>● Vide</div></div> <div>98 % 0 % 2 %</div>	<div><div>● Valide</div><div>● Erreur</div><div>● Vide</div></div> <div>100 % 0 % 0 %</div>	<div><div>● Valide</div><div>● Erreur</div><div>● Vide</div></div> <div>&lt; 1 % 0 % 99 %</div>
Conso moyenne (MWh)	Nombre de mailles secretisées	Part thermosensible (%)
3251.79933333333	0	
73.6005	0	
2225.343	0	
27.582	0	
103.115	0	

L'option *qualité de la colonne* nous permet de voir des colonnes vides.

La table sur la consommation contenait plusieurs colonnes vides ou possédant trop peu de données. Elles ont été supprimées pour une meilleure visibilité et compréhension. La table sur la consommation contenait des sommes de colonnes, elles seront supprimées.

### Création de colonnes personnalisées et conditionnelles

Peu de colonnes personnalisées et conditionnelles sont nécessaires, à ce niveau de la transformation.

Néanmoins, on peut mentionner la création de colonnes temporaires dans la table d'import/export. Certaines colonnes réalisent la somme d'autres colonnes alors nous créons, à des fins de vérifications, une colonne personnalisée additionnant les colonnes qui sont censées être additionnées et une colonne conditionnelle qui compare la valeur obtenue aux valeurs des colonnes. (Voir code ci-dessous)

Nouveau nom de colonne

verif\_export

Formule de colonne personnalisée ⓘ

```
= [#"FR vers GB (MWh)"+[#"FR vers CWE (MWh)"+[#"FR vers CH (MWh)"+[#"FR vers IT (MWh)"+[#"FR vers ES (MWh)"]]]]
```

Nouveau nom de colonne

is\_equal\_import

Formule de colonne personnalisée ⓘ

```
= if [verif_export]=["Export France (MWh)"] then 1 else 0
```

Formules utilisées pour les colonnes temporaires

Les résultats sont bien toujours égaux aux colonnes de bases, les données sont complètes et de bonne qualité.

### Analyse de la qualité des données

Les données étaient déjà de qualité, l'utilisation des options *qualité de la colonne* et *profil de la colonne* sur nos données transformées nous permettent de confirmer qu'après nos transformations préalables, nos données sont de qualités sur chacune des colonnes de chacune des tables.

A <sup>B</sup> <sub>C</sub> FILIERE	1	A <sup>B</sup> <sub>C</sub> Année	2	A <sup>B</sup> <sub>C</sub> Code Région		A <sup>B</sup> <sub>C</sub> Nom Région	3	A <sup>B</sup> <sub>C</sub> CODE CATEGORIE CONSOMM...	4
● Valide	100 %	● Valide	100 %	● Valide	100 %	● Valide	100 %	● Valide	100 %
● Erreur	0 %	● Erreur	0 %	● Erreur	0 %	● Erreur	0 %	● Erreur	0 %
● Vide	0 %	● Vide	0 %	● Vide	0 %	● Vide	0 %	● Vide	0 %

Les données sont de qualités, la mise en place du modèle peut commencer.

### Autres transformations

#### Filtre de données

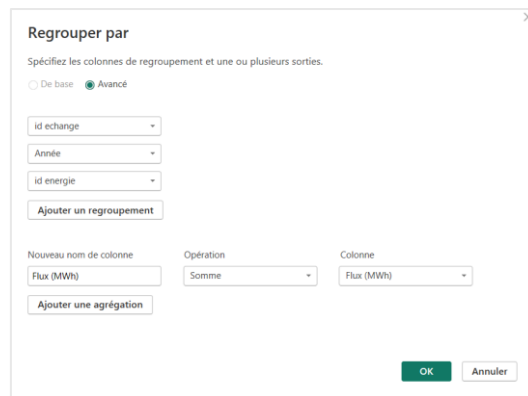
La table de consommation contenait des données sur le gaz. Ces données ne sont pas directement liées au processus étudié (flux d'électricité en France), elles ont donc été retirées.

Le filtrage étant été réalisé tard dans le projet, après la création du modèle, une table de dimension (dim\_type\_energie) reste disponible au cas où ces données sur le gaz venaient à être rajoutées à nouveau dans le modèle.

#### Création d'agrégats

Les tables sur l'import / export ont été agrégée à la dimension de l'année. Cela à été fait pour permettre une meilleure compréhension des données pour l'utilisateur final. Pour ce faire nous avons utilisé l'option « regrouper par » de l'onglet transformation des données.





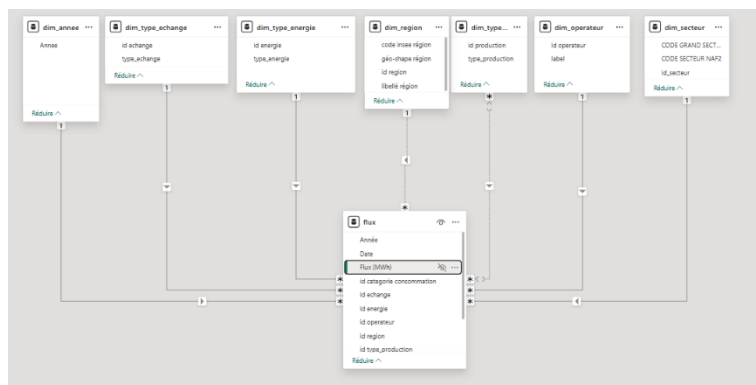
*Agrégation de la table flux-import. La mesure est sommée en fonction de l'année.*

## Chargement des données (Load)

Le processus de chargement des données nécessite de nouveau une transformation des données pour permettre la création du magasin de données souhaité.

Dans un premier temps, le modèle envisagé était un modèle en étoile centré autour d'une seule table flux.

Cette table de Flux est liée aux différentes dimensions. On citera notamment la dimension dim\_type\_echange qui décrit si la quantité d'énergie mesurée dans la table le Flux est de la production, de la consommation, de l'import ou de l'export.



*L'ancien modèle envisagé. Flux est la table de fait centrale entourée de ses dimensions.*

On verra [plus loin dans ce document](#) que ce modèle sera modifié pour devenir un modèle en constellation avec plusieurs tables de faits décrivant chacune un « sous-processus ».

## Transformation des tables

Les transformations appliquées ont été parfois assez lourdes. Les tables sources de la production et des échanges internationaux par exemple avaient la particularité d'avoir pour chacune des lignes plusieurs mesures différentes.

Pour la table de production originale, chaque année, région possédait plusieurs productions en fonction du type d'énergie. De la même manière, la table d'échange internationaux avait pour chaque ligne la quantité d'énergie exportée vers un pays X dans une colonne, puis celle exportée vers un pays Y dans une autre colonne.

Année	id region	Production nucléaire (GWh)	Production thermique (GWh)
2021	75	36682	906
2021	76	14815	322

Plusieurs colonnes contenant notre les mesures dans la table de production.

(ici, la ligne contient la production nucléaire et la production thermique mais elles devraient être dans une même colonne)

Date	Tranche horaire	FR vers GB (MWh)	GB vers FR (MWh)	FR vers CWE (MWh)
lundi 27 mai 2013	10	1405	-857	2901
lundi 27 mai 2013	13	1500	-20	3570

De la même manière, plusieurs colonnes contiennent la mesure dans la table d'import/export.

Cette disposition de colonne pose un problème pour la réalisation de notre modèle et un simple pivot nous ne permet pas de corriger cette situation. Il nous faudra donc « **empiler** les colonnes » de ces tables.

Pour ce faire, des dimensions sont créées pour distinguer le type de production (dim\_type\_production) et les type d'échanges (dim\_type\_echange). Les tables sont dupliquées à l'aide de référence puis les colonnes nécessaires sont conservées pour créer des tables à empiler par la suite.

Tranche horaire	Date	Flux (MWh)	type_echange
22	vendredi 1 juillet 2005	2414	Export IT
9	samedi 2 juillet 2005	2414	Export IT

Une table « exports-commerciaux IT » va par exemple conserver les colonnes nécessaires pour l'export d'électricité vers l'Italie.

Année	id region	Flux (GWh)	type_echange	type_energie
2014	11	1100	Production	Bioénergie
2014	44	526	Production	Bioénergie

Une table « production-region-annuelle-filiere BIOENER » va par exemple être créé pour la production d'électricité via la bioénergie.

Ces tables sont ensuite « empilées » à l'aide de l'option *ajouter des requêtes* afin d'avoir une seule table pour toutes les productions d'électricité et une seule table pour tous les imports / exports d'électricité. Cette unique table possède une unique colonne pour la mesure des flux.

A <sup>B</sup> <sub>C</sub> Année	A <sup>B</sup> <sub>C</sub> id region	1.2 Flux (GWh)	A <sup>B</sup> <sub>C</sub> type_echange	A <sup>B</sup> <sub>C</sub> type_energie
2013	24	68317	Production	Nucléaire
2013	76	19099	Production	Nucléaire
2014	11	2063	Production	Thermique
2014	44	6115	Production	Thermique

Après l'ajout de requêtes nous obtenons une table de production regroupant les différents types de production énergétique.

(Ici on voit que la production nucléaire et thermique d'électricité sont dans la même table)

Une fois ces transformations réalisées, nous pouvons extraire les dimensions de ces tables en créant des tables de dimensions. Ces tables deviendront donc nos tables de flux (tables de fait).

## Création des tables dimensions à partir des sources

Les tables dimensions sont diverses, certaines sont liées à une unique table source, comme `dim_type_production`. D'autres sont partagées entre toutes les tables sources, comme `dim_annee`.

Le déroulement de la création de dimension est la même pour la plupart des dimensions, on duplique la table contenant les valeurs originales, on conserve les colonnes nécessaires puis on *supprime les doublons*.

Une fois les doublons supprimés on crée une colonne d'index. Puis on *fusionne la requête* pour ramener la colonne d'index dans la table souhaitée. Nous pouvons ainsi supprimer les colonnes qui ont été mise dans la table de dimension.

CODE GRAND SECTEUR	CODE SECTEUR NAF2	id_secteur
AGRICULTURE	1	6
INDUSTRIE	10	7

*dim\_secteur sera liée à notre table de fait via l'id\_secteur. Dans la table de fait, seule la colonne id\_secteur sera présente.*

## Création et utilisation de tables de temps

Les dimensions temporelles ont fait l'objet de profondes réflexions en raison de la différence au niveau de la granularité temporelle des différentes tables. Il a un temps été décidé de créer une dimension date (comprenant année, jour, mois) pour la table d'échange commerciaux et de créer une dimension année qui sera commune entre toutes les tables de fait.

La création de ces deux dimensions a été réalisée grâce au langage DAX.

```
1 dim_annee = GENERATESERIES(2005, 2023, 1)
```

*Code utilisé pour la création de dim\_annee.*

```
1 dim_date =  
2 VAR DebutDate = DATE(2005, 1, 1)  
3 VAR FinDate = DATE(2023, 1, 1)  
4 RETURN  
5 ADDCOLUMNS(CALENDAR(DebutDate, FinDate),  
6     "Annee", FORMAT([Date], "YYYY"),  
7     "Trimestre", "T" & FORMAT([Date], "Q"),  
8     "Mois", FORMAT([Date], "MM"),  
9     "Nom du mois", FORMAT([Date], "MMMM YY")  
10 )
```

*Code utilisé pour la création de dim\_date*

Si la table `dim_date` est encore présente dans le fichier `pbix`, elle n'est plus utilisée suite au choix d'agréger les tables d'imports/exports à l'année.

## Création des mesures (fait) avec DAX

Les mesures implicites ont été remplacées par des mesures explicites que nous avons créées avec DAX.

```
1 Import annuel (MWh) = sum('flux-import-aggr'[Flux (MWh)])
```

*Mesure créée avec DAX.*

Les mesures restent simples, seules des sommes ont été nécessaires pour créer le rapport.

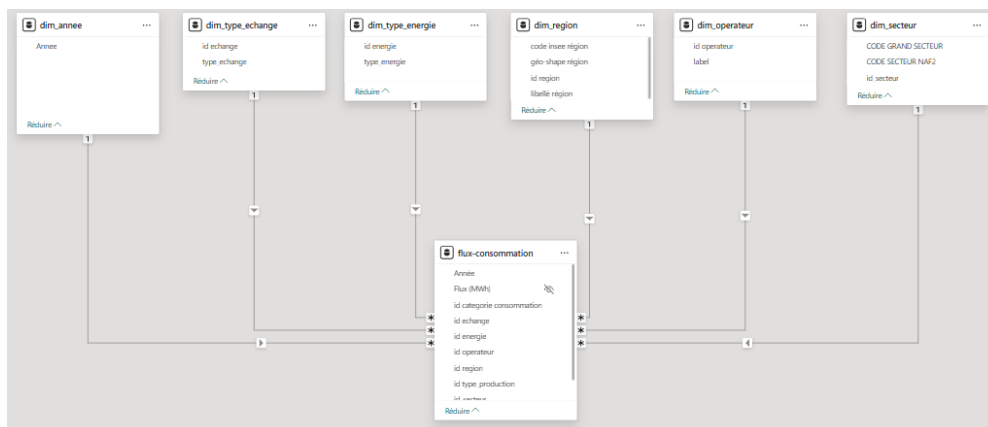
## Description des modèles de données utilisés.

Comme indiqué précédemment, le premier modèle visé était un modèle en étoile avec une unique table de fait nommée « Flux ». Cette modélisation avait comme avantage la bonne compréhension du processus étudié, les échanges sur le réseau électrique, les flux d'électricité.

Cependant, afin d'améliorer la lisibilité du modèle une autre solution a été adoptée. La table de fait « Flux » a été divisé en fonction des types d'échange. On retrouvera donc une table de fait pour chacun des sous-processus (production / consommation / importation / exportation).

Chacun des sous processus possèdera son propre datamart et son propre schéma en étoile. Cependant toutes les tables resteront reliées entre elles pour continuer de pouvoir réaliser des analyses sur l'entièreté du processus, les flux électriques.

### Datamart-consommation

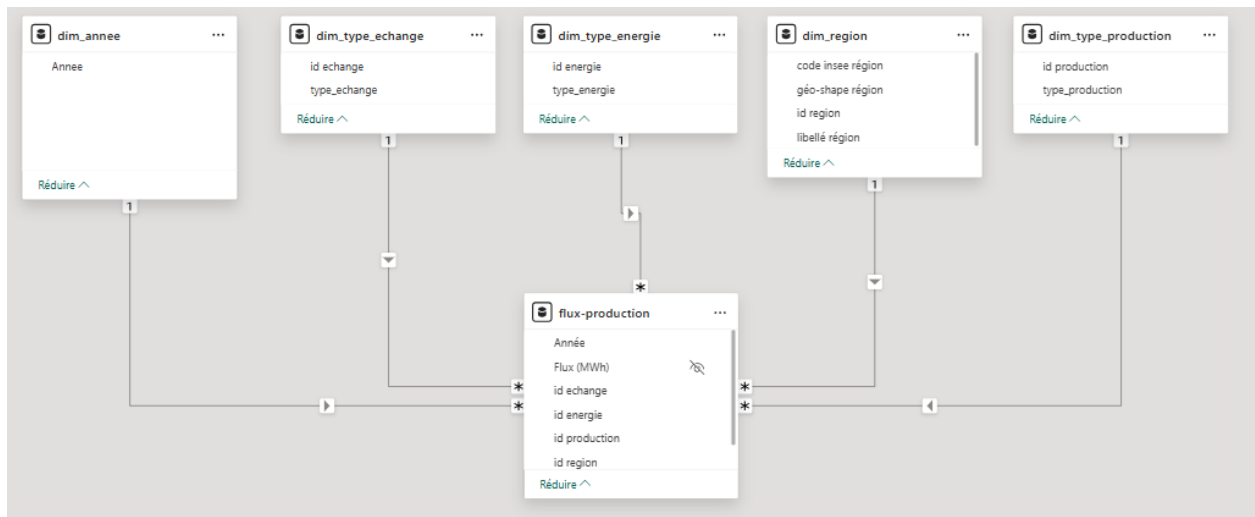


*Modèle en étoile du datamart-consommation.*

Autour de la table de fait flux-consommation se trouve les différentes dimensions que l'on voit le sur schéma ci-dessus. Dans cette table de fait, se trouve les flux d'électricité liés à la consommation d'électricité.

Certaines dimensions sont uniquement liées à cette table comme la dimension des secteurs de consommateurs (dim\_secteur) ou la dimension sur les opérateurs (dim\_operateur).

### *Datamart-production*

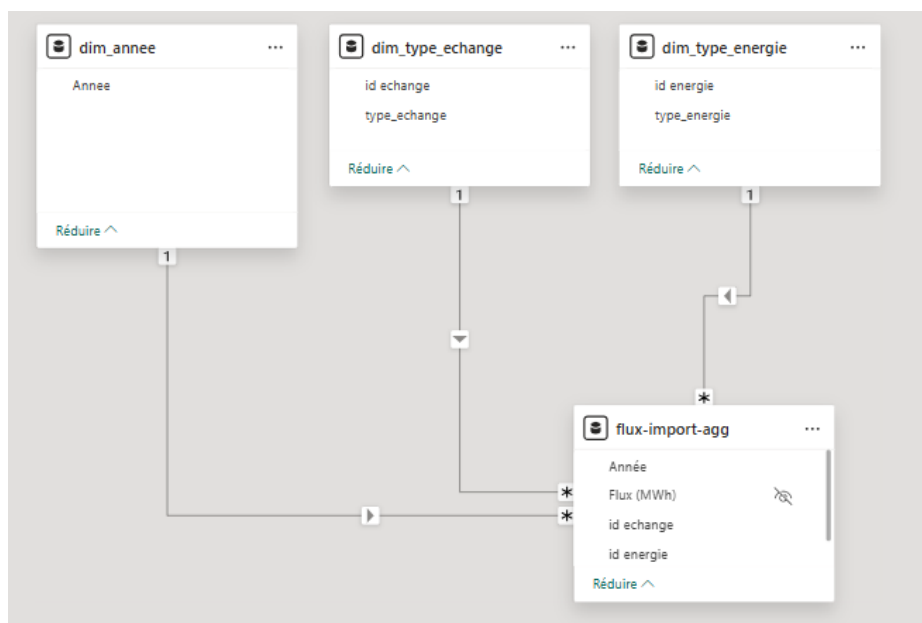


*Modèle en étoile du datamart-production.*

Autour de la table de fait flux-production se trouve les différentes dimensions que l'on voit le sur schéma ci-dessus. Dans cette table de fait, se trouve les flux d'électricité liés à la consommation d'électricité.

Une dimension est uniquement liée à cette table, la dimension sur le type de production (dim\_type\_production).

## Datamart-importation



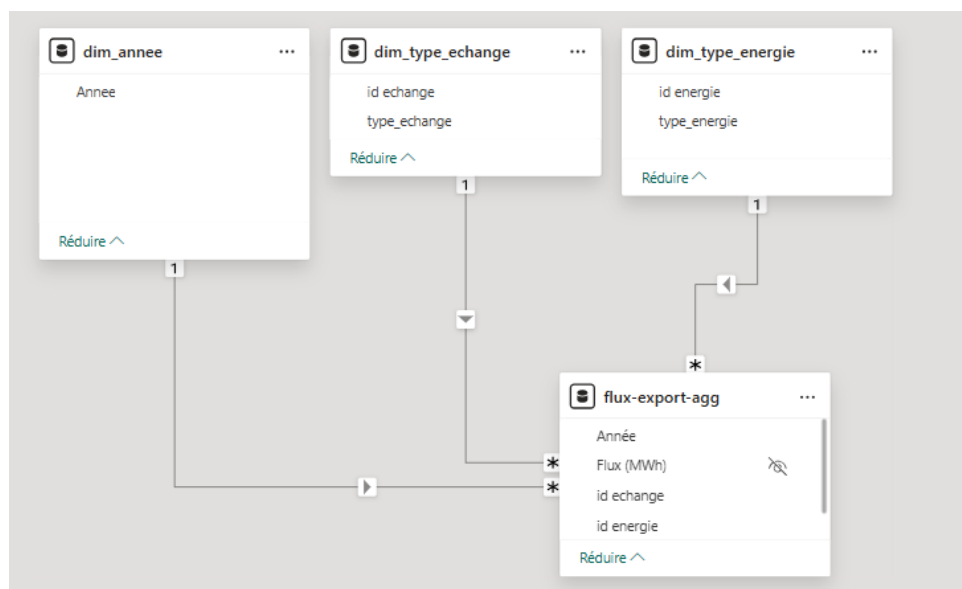
Modèle en étoile du datamart-importation.

Autour de la table de fait flux-import se trouve les différentes dimensions que l'on voit le sur schéma ci-dessus. Dans cette table de fait, se trouve les flux d'électricité liés à l'import depuis les autres pays.

Toutes les données des flux sont positives.

La table de fait est agrégée à l'année.

## Datamart-exportation



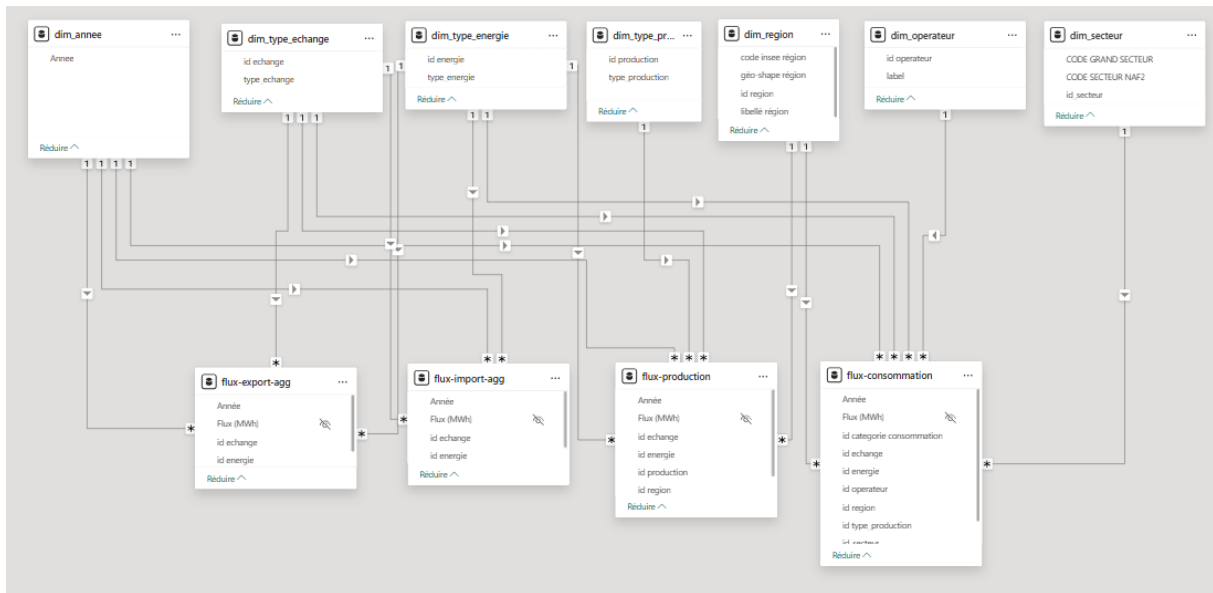
Modèle en étoile du datamart-exportation.

Autour de la table de fait flux-export se trouve les différentes dimensions que l'on voit le sur schéma ci-dessus. Dans cette table de fait, se trouve les flux d'électricité liés à l'export vers les autres pays.

Toutes les données des flux sont positives.

La table de fait est agrégée à l'année.

## Datawarehouse



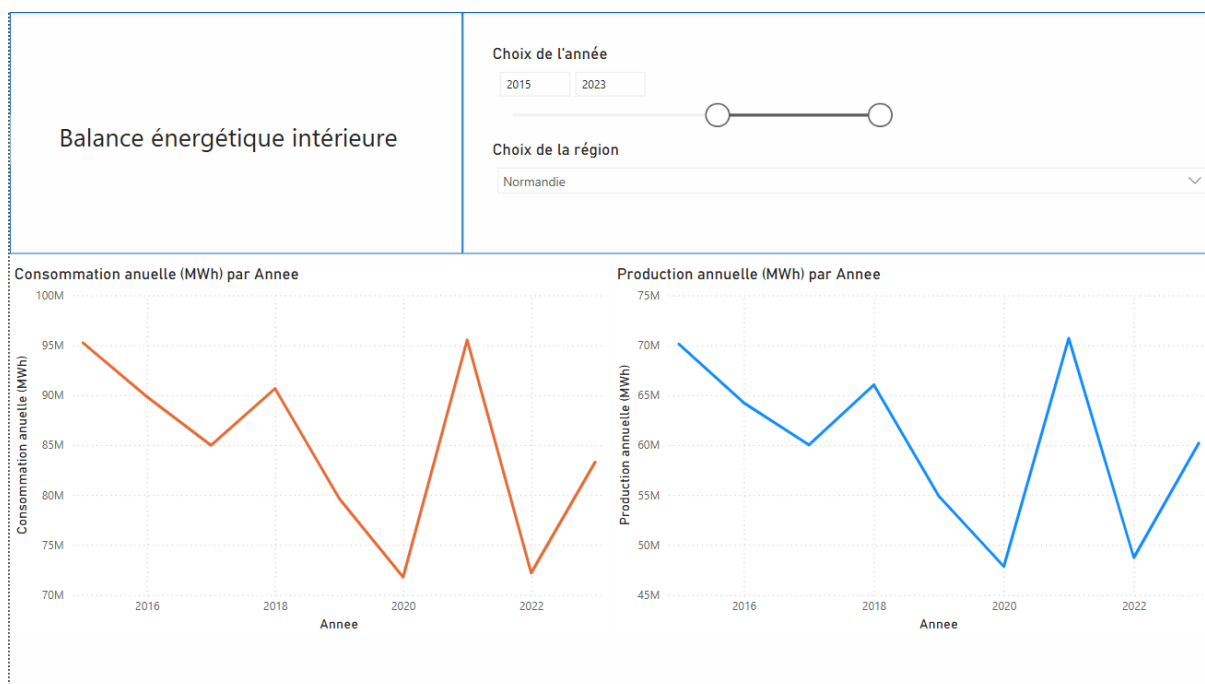
Modèle en constellation de l'ensemble des tables de faits.

Ce graphique rassemble toutes les tables de faits. L'intérêt de relier la totalité des tables est de permettre de réaliser des analyses sur l'ensemble du processus étudié.

## Restitution

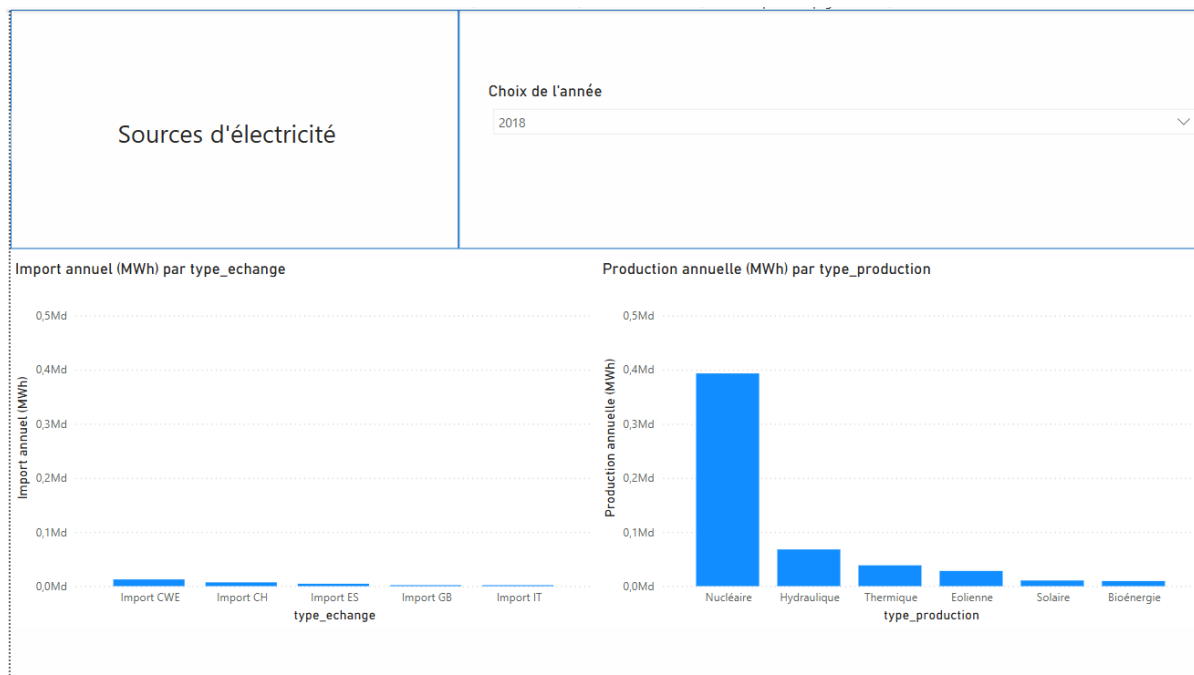
L'entrepôt de données ainsi mis en forme nous permet de réaliser des reporting et tableaux de bords.

## Tableaux de bord



Balance énergétique intérieure entre 2015 et 2023

Ce premier tableau de bord va permettre au gestionnaire de réseau de voir l'évolution de la consommation et de la production d'énergie sur le réseau. Ici, sur la capture d'écran on peut observer l'évolution de 2015 à 2023 sur la région Normandie.



Provenance de l'électricité sur le réseau en 2018.



Ce second tableau de bord permet au gestionnaire de croiser différents « sous-processus ». Ici, l'import et la production d'électricité illustrent les flux entrant sur le réseau électrique français en 2018.

On voit que l'énergie électrique française provient principalement du nucléaire.

## Annexes

### Notes de fin

\*\* Certaines étapes aurait pu être réalisées plus en amont dans la transformation des données. Néanmoins ce projet étant notre premier réel projet sur PowerBi nous n'avons pas immédiatement pensé à optimiser le code.

D'autres optimisations ont cependant été réalisées au sein des requêtes ce qui nous a permis d'améliorer la fluidité de l'importation des données. On peut notamment citer le fait de déplacer la suppression de colonnes au début des requêtes et la désactivation de l'actualisation lors de l'actualisation du rapport.