RAPPORT DE PROJET DATA WAREHOUSING

FILIÈRE: E-MANAGEMENT AND BUSINESS INTELLIGENCE

Rapport de projet Data Warehousing

RÉALISÉ PAR:

Kassaoui Wissal

Sahlaoui Botaina

ENCADRANTS:

Pr. Lamia Benhiba

Année académique : 2021/2022

Table des matières

Table des figures

1 | Introduction

La gestion des données financières peut s'avérer complexe et fastidieuse. En effet, la collecte, le traitement et l'analyse de volumes d'informations financières nécessitent des processus précis et transparents, notamment l'utilisation de services de transfert de données fiables pour rationaliser vos flux de travail. C'est là que le Data Warehouse ou "l'entrepôt de données" s'avère utile.

Le data warehousing fournit une ressource de données critiques que vous pouvez facilement suivre dans le temps et analyser pour vous aider à gérer efficacement vos processus financiers et à prendre de meilleures décisions.

Le data warehouse permet ainsi d'améliorer la prise de décision et le déploiement de stratégies plus efficaces. Il constitue un avantage concurrentiel important pour une entreprise.

2 | Présentation du projet

2.1 Problématique

Les tableaux de bord et les rapports sont des outils de suivi et d'aide à la décision très utilisés par les dirigeants d'entreprise. Ainsi département financier de l'entreprise AdventureWorks a besoin de rapports/Tableaux de bord décrivant la situation financière de l'entreprise.

2.2 Objectifs

Afin de répondre au besoin de l'entreprise Adventure Works, nous aurons besoin de :

- Créer un Datamart Finance.
- Charger les données dans le Datamart
- Etablir des rapports et des tableaux de bord pour la restitution des données

2.3 Phases du projet

Pour réaliser nos objectifs, nous allons suivre les étapes suivantes :

- Exploration de la base de données Adventure Work.
- La création du schéma de l'entrepôt de données (DataMart)
- Implémenter les étapes de l'extraction, la transformation, et le chargement des données dans le DataMart
- Création d'un cube OLAP et des rapports

3 | Exploration de la bases de données

3.1 Adventure Works

Ce projet utilise utilise une entreprise fictive fournie par Microsoft, à savoir Adventure Works Bicycles, Inc. Adventure Works est n marchand en gros de bicyclettes fictif. Même si Adventure Works est une entreprise fictive, elle est conçue comme un cas réaliste, semblable à celui d'une entreprise réelle dans l'industrie. Adventure Works fournit une base de données et un entrepôt de données qui couvrent les processus d'affaires des ventes, de la gestion du matériel, de la production, des finances et de la gestion du capital humain.

3.2 La base de données Adventure Works

La base de données AdventureWorks est une base de données de traitement des transactions en ligne (OLTP), qui est riche en structure, contenu et variété.

La base de données OLTP se compose de 68 tables qui sont regroupées dans différentes classifications telles que Ventes, Achats, Production, Ressources humaines et Personnes. La base de données (dans son état brut) contient les données de près de 20 000 personnes (employés, clients, contacts de magasins, contacts de fournisseurs et contacts généraux). Elle contient également les données de plus de 31 000 transactions de vente aux clients et de plus de 4000 transactions d'achat aux fournisseurs.

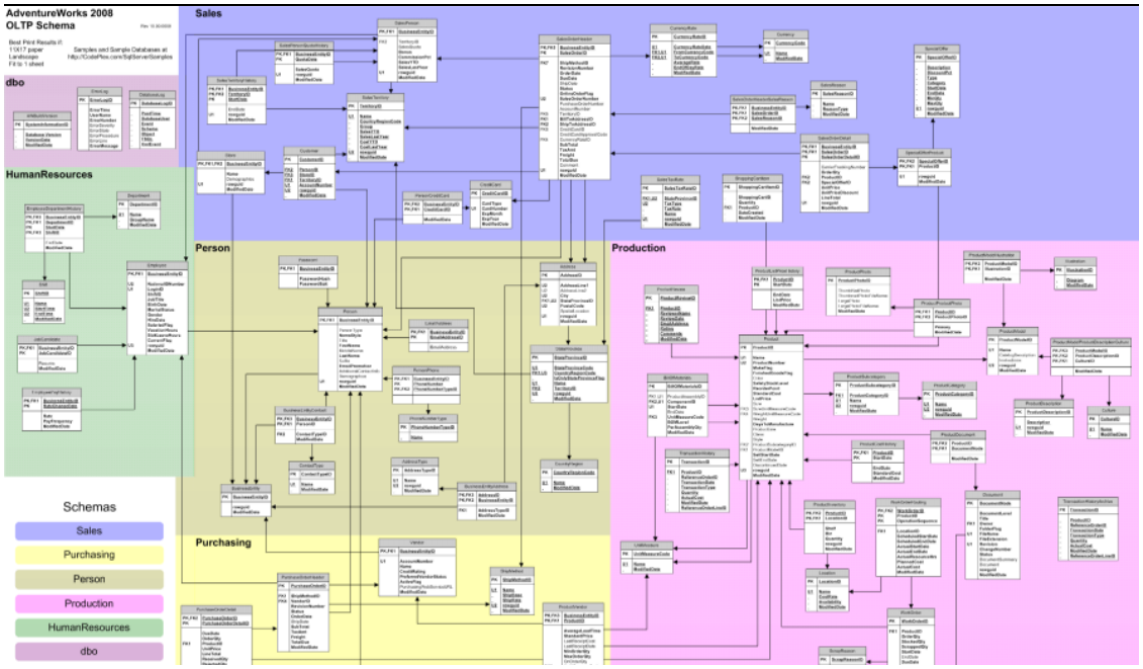


FIGURE 3.1 – Schéma de la base de données Adventure Works

3.3 Adventure Works Data Warehouse

Le DW d'Adventure Works est une architecture d'entrepôt centralisée composée de de tables de faits, de tables de dimensions, et contenant des données obtenues à partir de la base de données OLTP et de d'autres sources de données via un processus traditionnel d'extraction/transformation/chargement (ELT). Il y a au total 10 tables de faits, dont les sujets vont de la vente par Internet et par revendeur aux finances et l'inventaire des produits. Ces tables de faits sont entourées de 16 tables de dimensions, représentant les clients, les lignes de produits, les comptes, les employés, les départements, les régions géographiques, et le temps

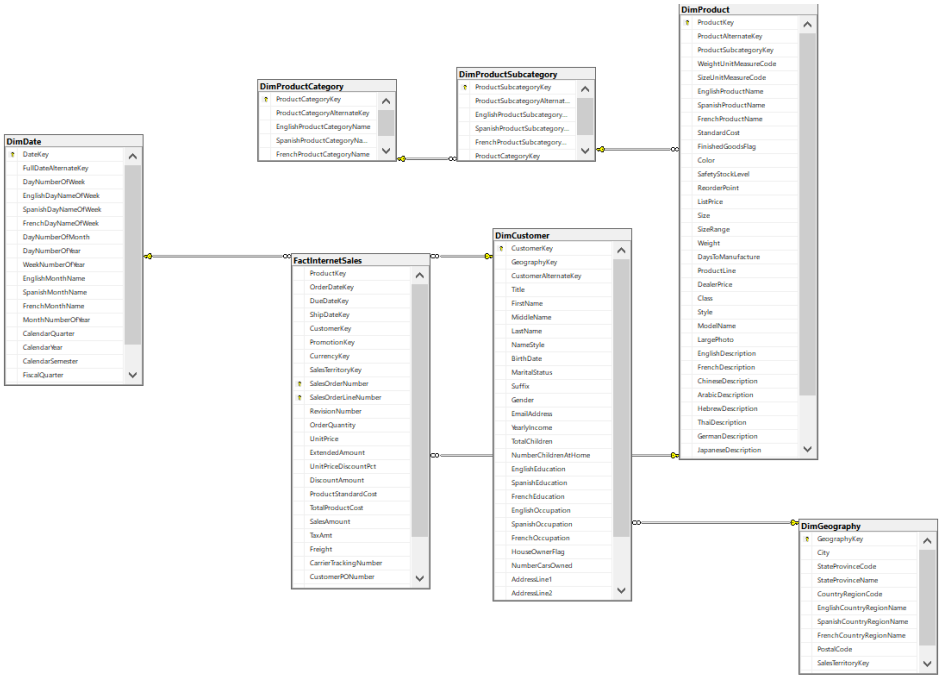


FIGURE 3.2 – Exemple de shéma du Data Warehouese Adventure Works

4 | Conception Du Datamart

4.1 Schéma de travail :

Notre Datamart contiendra 6 tables de dimensions : DimAccount, DimScenario, DimCurrency, DimOrganization, DimDepartmentGroup, DimDate et une table de fait Fact Finance.

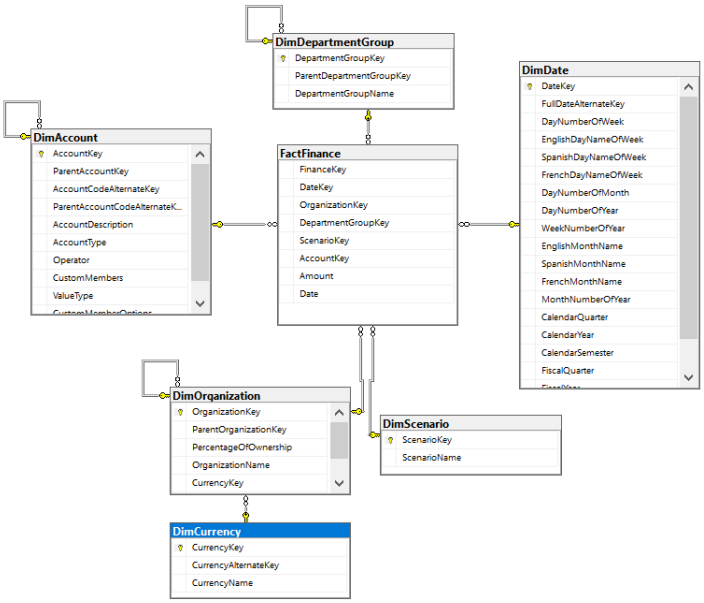


FIGURE 4.1 – Schéma conceptuel du Datamart

Les tables de dimension DimDate et DimScenario n'ont pas de tables Lookup supplémentaires qui lui sont associés et utilisent donc un schéma en étoile.

Les tables de dimension DimAccount, DimOrganization, DimDepartmentGroup, contiennent une une hiérarchie parent-enfant qui dépend d'une relation d'auto-référencement présente sur la table principale de la dimension. Par exemple : dans la dimension DimAccount la colonne ParentAccountKey a une relation de clé étrangère avec la colonne de clé primaire AccountKey.

De plus, la table de dimension DimOrganisation utilise un schéma hybride. Le premier niveau du schéma est en snowflake seulement à travers la table lookup DimCurrency.

4.2 Base de données :

On a choisit d'utiliser SQL server comme SGBD pour héberger notre Datawarehouse. Microsoft SQL Server est un système de gestion de base de données (SGBD) en langage SQL incorporant entre autres un SGBDR (SGBD relationnel ») développé et commercialisé par la société Microsoft.

5 | Intégration de données vers le Datamart

5.1 Source des données :

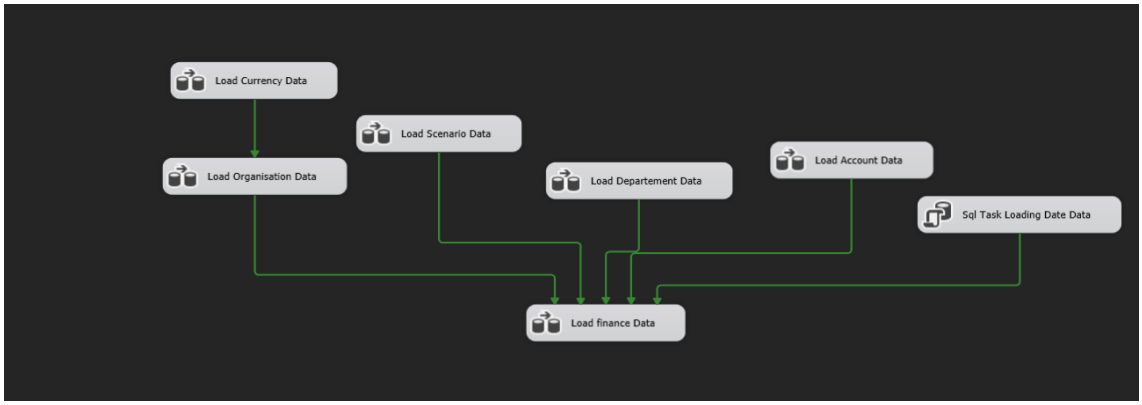
Notre entrepôt de données financières acquiert ses données de trois sources différentes (Entrepôt de données AdventureWorks, La base de données opérationnelle de l'entreprise et un fichier plat contenant des données à propos de la devise)

5.2 Outil ETL :

SQL Server Integration Services est un composant du logiciel de base de données Microsoft SQL Server qui peut être utilisé pour effectuer un large éventail de tâches de migration de données. SSIS est une plate-forme pour l'intégration des données et les applications de flux de travail. Il comporte un outil d'entreposage des données utilisé pour l'extraction, la transformation et le chargement des données.

5.3 Flux de contrôle des données :

Afin de respecter les différentes contraintes, on a définit un flux de contrôle pour contrôler l'ordre de chargement des dimensions.



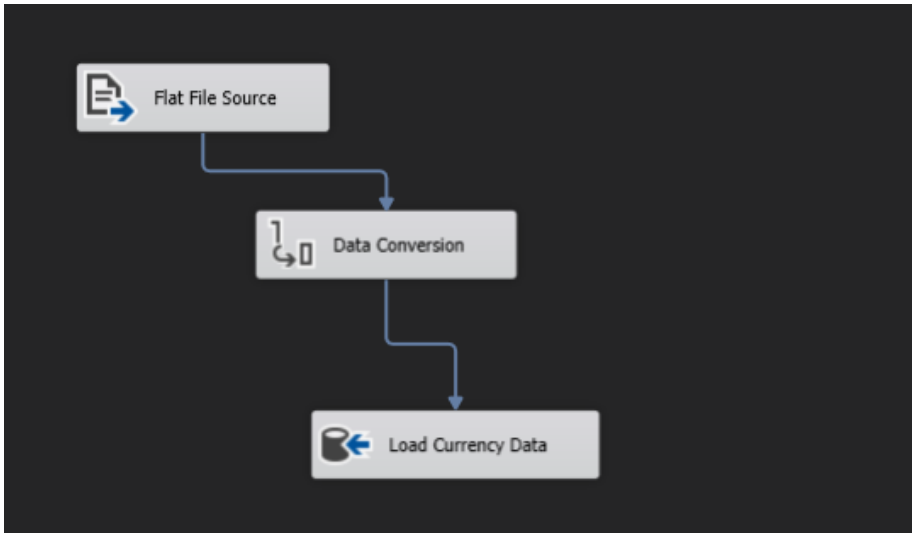
On définit pour chaque tache de flux de données des sous taches d'extraction et d'alimentation

5.3.1 Dim Date :

Le chargement de données pour cette partie est réalisée par un script code T-SQL. En choisissant une date début, la requête SQL va charger une ligne pour chaque jours qui suit jusqu'à atteindre le jour présent. (Annexe 1)

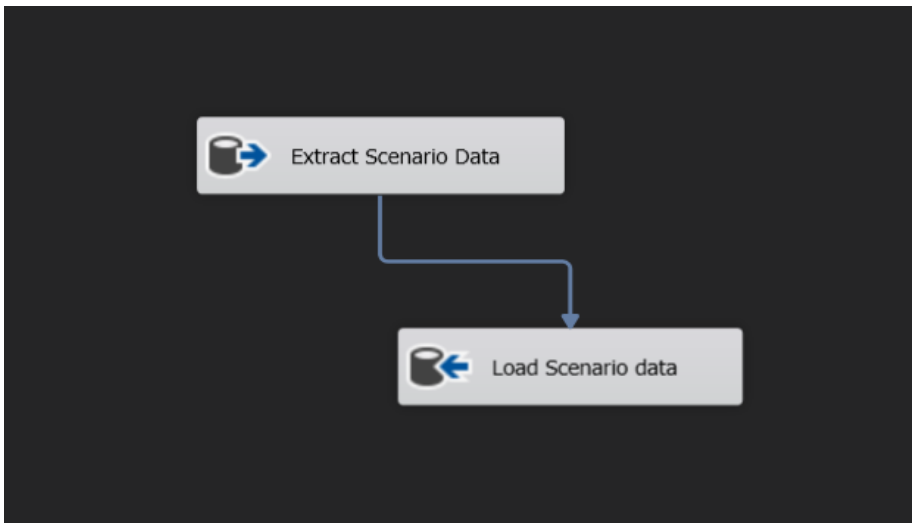
5.3.2 Dim Currency :

La source de données ici est un fichier plat, il contient les détails concernant chaque devise. Quelques transformations ont été nécessaires avant d'intégrer ces données au sein de l'entrepôt pour s'assurer que les données ont le même type que la source sortie.



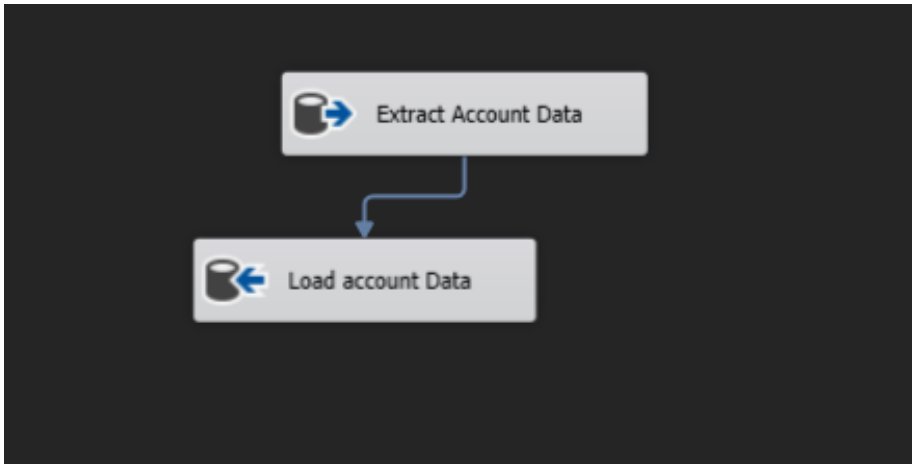
5.3.3 Dim Scénario :

Le chargement de cette dimension est intuitif. La seule source de données présente est l'entrepôt de données AdventureWorksDW2019 de la table *DimScenario*.



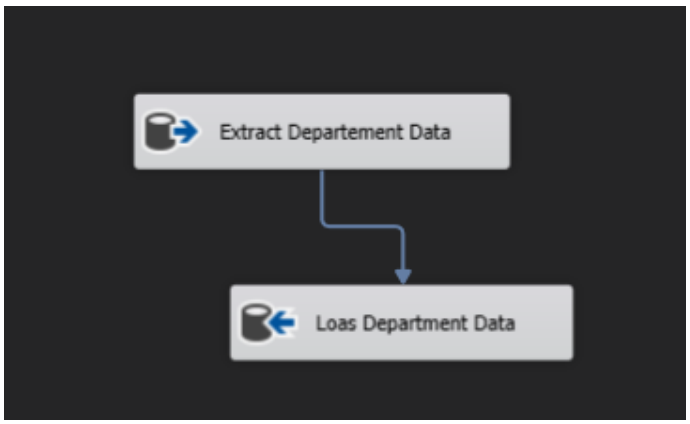
5.3.4 Dim Account :

Le chargement de cette dimension est à partir de la table *DimAccount* de l'entrepôt de données AdventureWorksDW2019.



5.3.5 Dim Departement :

Le chargement de la dimension Department est à partir de la table *DimDepartment* de l'entrepôt de données AdventureWorksDW2019.

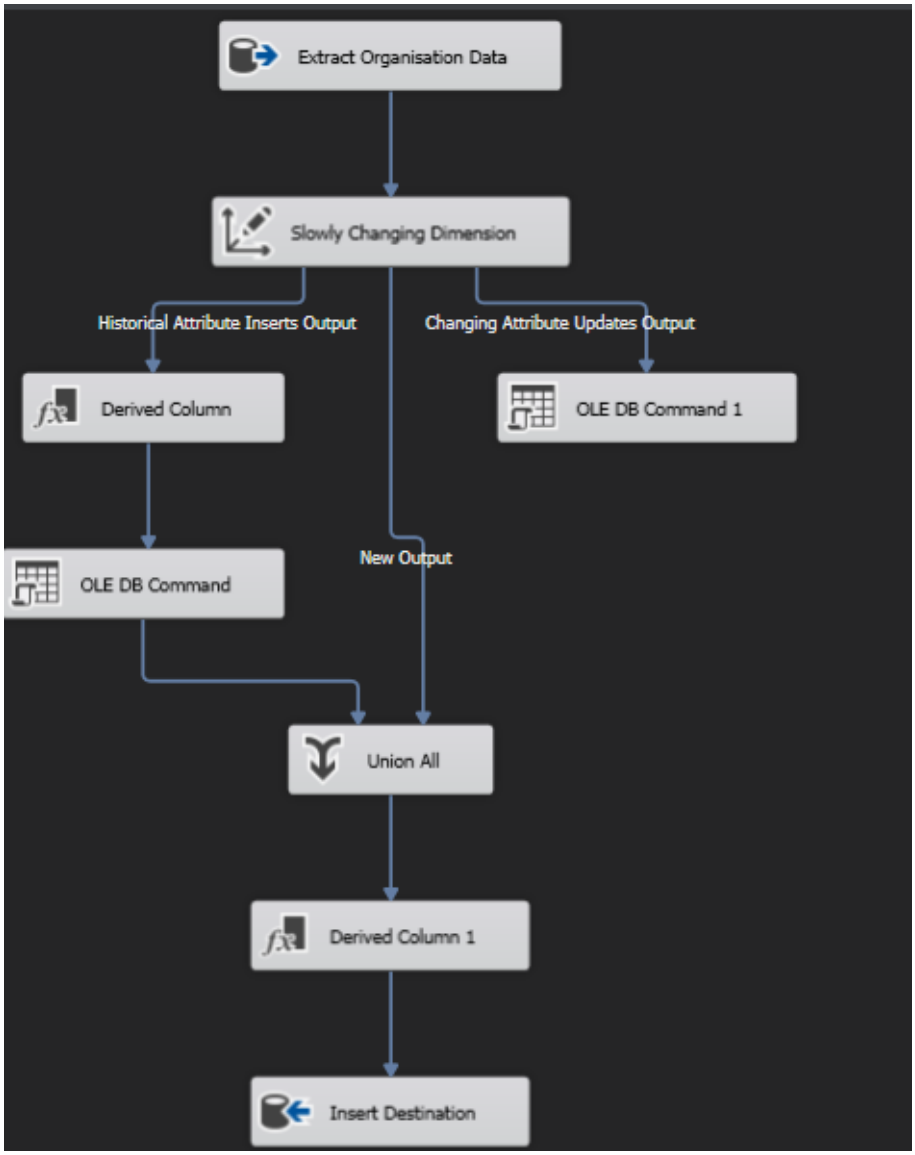


5.3.6 Dim Organisation :

La source initiale des données est la dimension Organization de l'entrepôt général. L'insertion de données est réalisée par historisation complète des modifications réalisées au pourcentage de propriété.

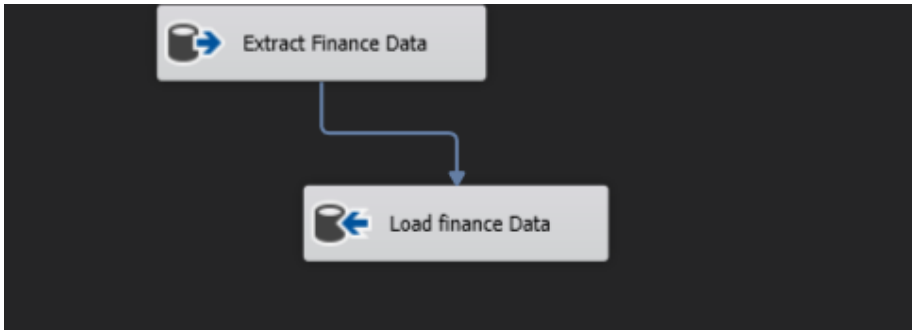
L'historisation de cette information est essentielle car elle révèle le taux de profit que la société AdventureWorks réalise pour chaque période de temps lors de l'attribution des dividendes.

Toutes les clés de la source sont stockées dans de nouvelles colonnes pour faire références aux clés d'entreprise. Un deuxième flux de données est ajouté pour insérer les informations manquantes à propos des clés père de chaque organisation insérée



5.3.7 Fact Finance

Après avoir assuré la présence de toutes les données de l'entrepôt, on insère les lignes de la table de faits par l'option de vérification ligne par ligne pour prévenir l'insertion de lignes qui ne satisfont pas les contrainte de stockage.



Lorsqu'il y a présence de lignes corrompues, on les visualise pour comprendre la nature de l'erreur puis on les ignore afin d'insérer les lignes sûres. Lorsque l'erreur est fixée une réexécution du paquet sera nécessaire et elle ne concernerait que les lignes non insérés. Une telle option ralentit le processus de l'insertion mais nous aide à détecter l'erreur de manière rapide. Toutefois, la deuxième option arrive à minimiser le temps gaspillé lors du premier échec.

6 | Génération du cube Finance Datamart :

6.1 Modes de stockage de partition

Un cube OLAP est un tableau multidimensionnel de données. Le traitement analytique en ligne (OLAP) permet d'analyser des données pour en tirer des enseignements. Cependant, il existe plusieurs approches pour le stockage du Cube Olap. Nous choisissons le mode MOLAP. MOLAP est l'acronyme de Multidimensional Online Analytical Processing. Le MOLAP utilise un cube multidimensionnel qui accède aux données stockées par diverses combinaisons. Les données sont pré-calculées, pré-sommées et stockées.

6.2 Outil de génération du Cube

Microsoft SQL Server Analysis Services est un outil de traitement analytique et d'exploration de données en ligne dans Microsoft SQL Server. SSAS est utilisé comme un outil par les organisations pour analyser et donner du sens aux informations éventuellement réparties dans plusieurs bases de données, ou dans des tables ou des fichiers disparates.

6.3 Vue de source de donnée :

Après avoir configuré la connexion du service SSAS OLAP et attribué les droits nécessaires pour la lecture du Finance DW, on crée notre première et unique vue de source de données.

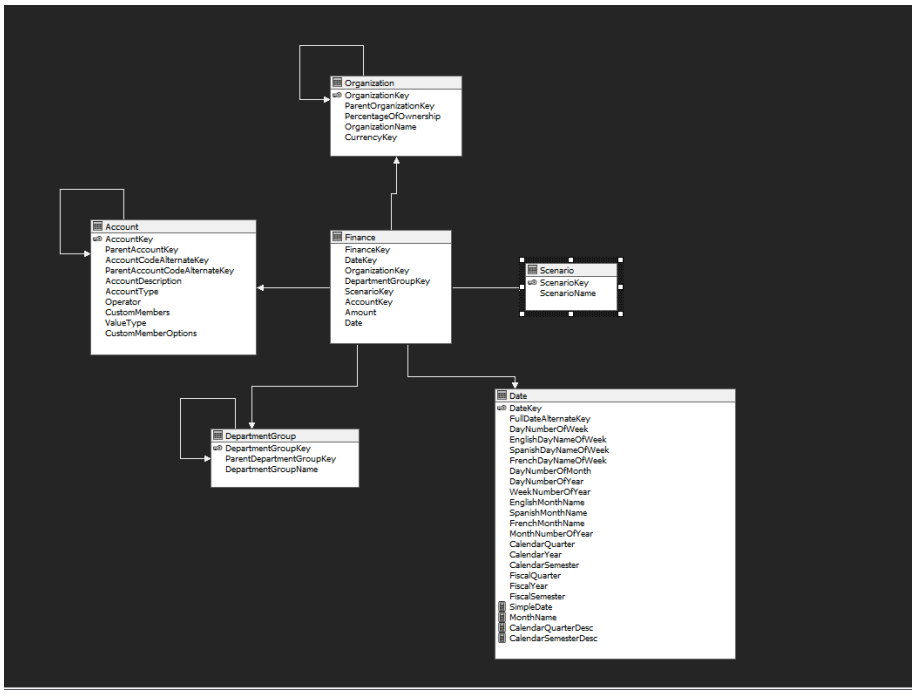


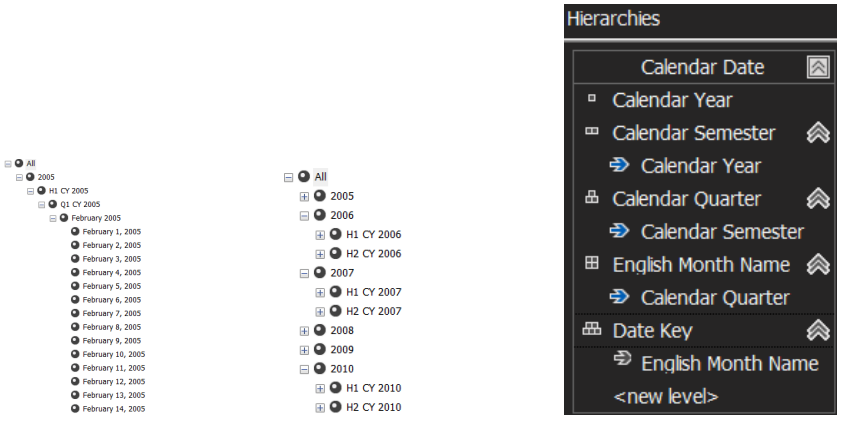
FIGURE 6.1 – Vue de la source de donnée

Nous avons sélectionné le schéma en étoile du Datamart

6.4 Définition des dimensions :

6.4.1 Dim Date

La dimension date est l'une des dimensions les plus importantes dans un cube Analysis Server (SSAS). Les cubes ont besoin d'une dimension date afin d'analyser les informations historiques.



(a) (b) (c)

FIGURE 6.2 – (a) Hiéarchies de la dimension Date

6.4.2 Dim Account

Cette dimension présente une relation Parent-Child définissant la hiérarchie des comptes. On associe la description du compte à la clé AccountKey pour assurer une meilleure visualisation des données, puis nous avons configuré le paramètre ParentAccountKey pour indiquer la présence d'une relation père fils sans faire dupliquer la feuille du père.

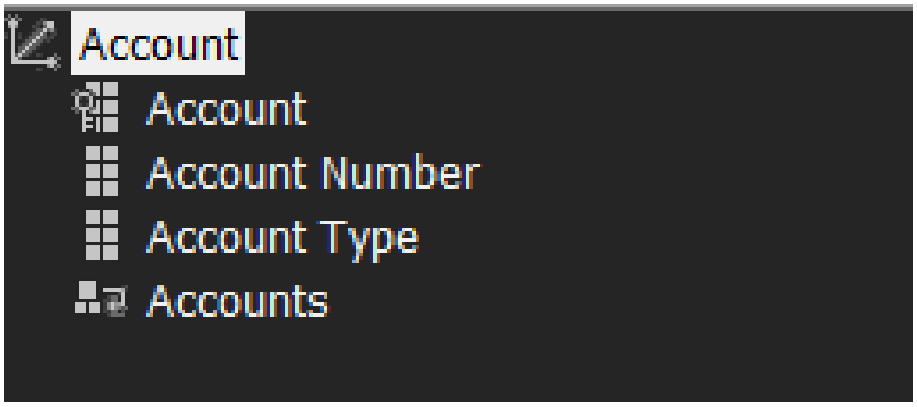


FIGURE 6.3 – Attributs Choisis

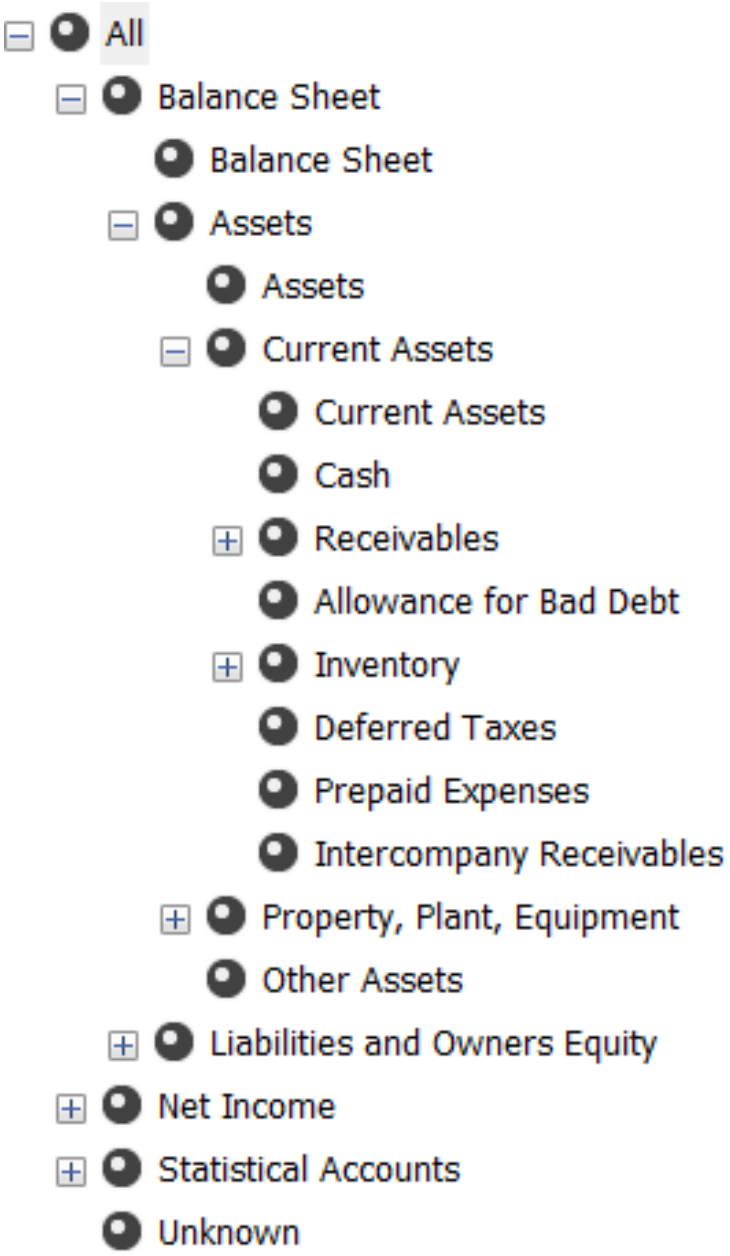
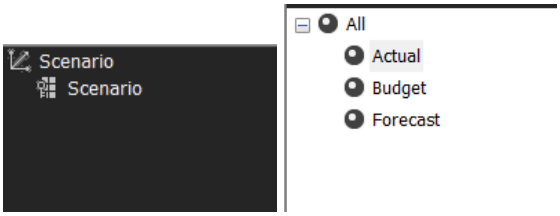


FIGURE 6.4 – Visualisation de la hiéarchie

6.4.3 Dim Scénario

Cette dimension contient les scénarios possibles. Pour une meilleure visualisation nous avons associé le Nom du scénario à la clé primaire de la dimension



(a) (b)

FIGURE 6.5 – Dimension Scénario

6.4.4 Dim Organisation

Cette dimension présente aussi une relation Parent-Child définissant la hiérarchie des organisations de l'entreprise. On associe la description du compte à la clé OrganisationKey (rennomé Organisation) pour assurer une meilleure visualisation des données.

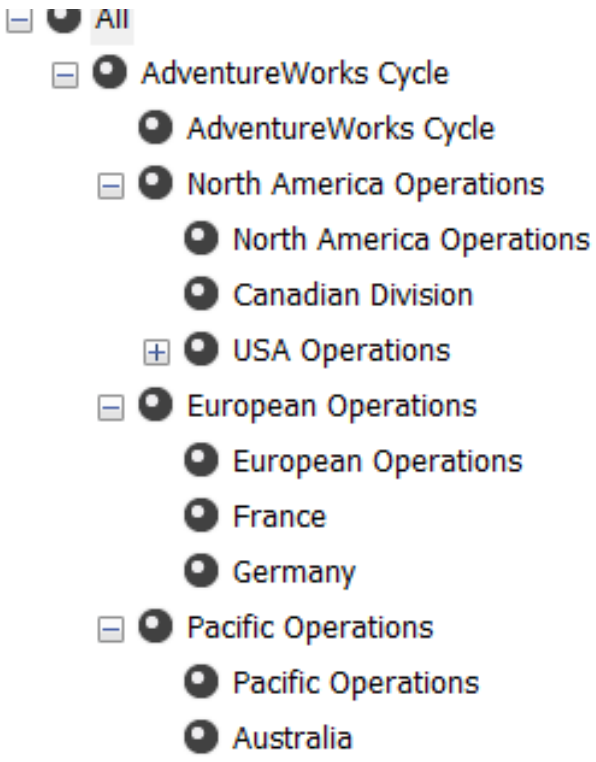


FIGURE 6.6 – Attributs Choisis

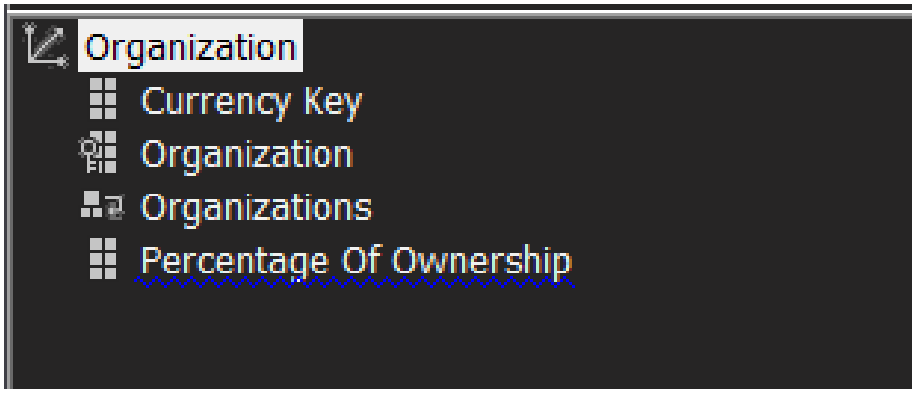


FIGURE 6.7 – Visualisation de la hiéarchie

6.4.5 Dim Departement Group

Cette dimension présente de même une relation Parent-Child définissant la hiérarchie des départements de l'entreprise. On associe la description du compte à la clé DepartementKey (rennomé Departement) pour assurer une meilleure visualisation des données.

FIGURE 6.8 – Visualisation de la hiéarchie

6.5 Génération du Cube :

Après le traitement et déploiement de chacune des dimensions, on a pu créer et générer le Cube.

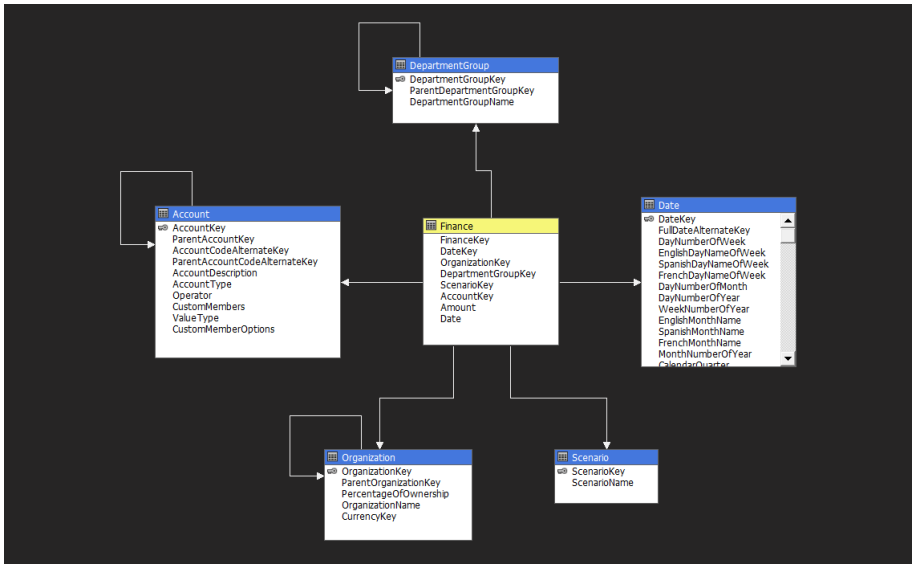


FIGURE 6.9 – Cube Finance

7 | Restitution des données et génération des rapport

7.1 Restitution des données avec Excel :

Les Tableaux Excel permettent d'analyser les différentes transactions effectuées par les différentes organisations, appartenant à Adventures Work et aussi de façon plus détaillée, comme par exemple par département ou par scénario…

Row Labels	2018	2019	2020	2021	Grand Total
Balance Sheet	**7929449.36**	**251070222.3**	**391114398.7**	**456896848.2**	**1107010919**
Assets	**3964724.68**	**125535111.1**	**195557199.4**	**228448424.1**	**553505459.3**
Current Assets	3085240.82	114627595.2	179780757.7	208012742.6	505506336.3
Property, Plant, Equipment	765481.93	9486660.56	13778136.44	17888838.48	41919117.41
Other Assets	114001.93	1420855.41	1998305.21	2546843.03	6080005.58
Liabilities and Owners Equity	**3964724.68**	**125535111.1**	**195557199.4**	**228448424.1**	**553505459.3**
Liabilities	2652102.27	44138324.73	66421151.09	84641645.61	197853223.7
Owners Equity	1312622.41	81396786.4	129136048.3	143806778.5	355652235.6
Net Income	**3234990.15**	**86709470.53**	**68573306.18**	**80739402.32**	**239257169.2**
Operating Profit	**3187162.62**	**84727171.9**	**65490135.45**	**78106523.56**	**231510993.5**
Gross Margin	2001333.37	68860202.09	50739295.95	59855136.92	181455968.3
Operating Expenses	1185829.25	15866969.81	14750839.5	18251386.64	50055025.2
Other Income and Expense	**8178.39**	**118536.45**	**161684.86**	**213984.94**	**502384.64**
Interest Income	2433.67	32405.6	47995.92	64167.09	147002.28
Interest Expense	3739.56	49755.42	73685.36	98543.25	225723.59
Gain/Loss on Sales of Asset	-2984.2	-39720.65	-58726.23	-78577.6	-180008.68
Other Income	4989.36	-21256.92	94086.81	145364.2	223183.45
Curr Xchg Gain/(Loss)		97353	4643	-15512	86484
Taxes	**39649.14**	**1863762.18**	**2921485.87**	**2418893.82**	**7243791.01**
Statistical Accounts	**220445**	**3382867**	**4426284**	**4342729**	**12372325**
Headcount	**125**	**1566**	**2302**	**3009**	**7002**
Units	**820**	**28301**	**76982**	**95220**	**201323**
Square Footage	**219500**	**3353000**	**4347000**	**4244500**	**12164000**
Grand Total	**11384884.51**	**341162559.8**	**464113988.9**	**541978979.5**	**1358640413**

FIGURE 7.1 – Evolution des différents type de comptes dans le temps

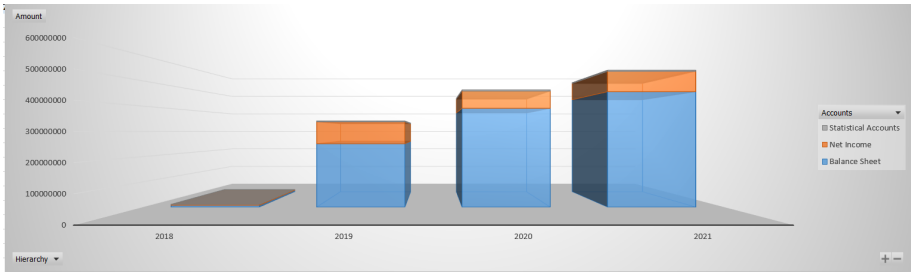


FIGURE 7.2 – Visualisation avec excel d l'évolution des différents type de comptes dans le temps

On remarque que les profit de l'entreprise évolue énormément de 2018 vers 2021 et que les comptes de type Balance Sheets représente la majorité des comptes

Amount	Column Labels				
Row Labels	2018	2019	2020	2021	Grand Total
Corporate	**11384884.51**	**341162560**	**464113989**	**541978980**	**1358640413**
Corporate	**526536.65**	**6674298.92**	**6355457.54**	**7497631.5**	**21053924.6**
AdventureWorks Cycle	526536.65	6674298.92	6355457.54	7497631.5	21053924.6
North America Operations	526536.65	6645717.65	5977998.2	6263520.11	19413772.6
European Operations		28581.27	352090.49	896036.81	1276708.57
Pacific Operations			25368.85	338074.58	363443.43
Executive General and Administration	**287661.02**	**9021344.63**	**13091638.9**	**15836808.9**	**38237453.5**
AdventureWorks Cycle	287661.02	9021344.63	13091638.9	15836808.9	38237453.5
North America Operations	287661.02	8966565.29	12169209.9	11886998	33310434.3
European Operations		54779.34	812494.64	2351978.32	3219252.3
Pacific Operations			109934.35	1597832.54	1707766.89
Inventory Management	**548846**	**19510038.5**	**29859920.1**	**34623729.4**	**84542533.9**
AdventureWorks Cycle	548846	19510038.5	29859920.1	34623729.4	84542533.9
North America Operations	548846	19492108	29695256	34318355	84054565
European Operations		17930.45	164664.1	305374.37	487968.92
Manufacturing	**46000**	**560586.83**	**472169.44**	**548683.98**	**1627440.25**
AdventureWorks Cycle	46000	560586.83	472169.44	548683.98	1627440.25
North America Operations	46000	546267	274402	260016	1126685
European Operations		14319.83	197767.44	288667.98	500755.25
Quality Assurance	**27488**	**323492**	**255480.97**	**264281.2**	**870742.17**
AdventureWorks Cycle	27488	323492	255480.97	264281.2	870742.17
North America Operations	27488	323492	198526	199758	749264
European Operations			56954.97	64523.2	121478.17
Research and Development	**2702102.36**	**80483260.1**	**64006851**	**74122673.9**	**221314887**
AdventureWorks Cycle	2702102.36	80483260.1	64006851	74122673.9	221314887
North America Operations	2702102.36	80204251.8	59471194.4	56993650.4	199371199
European Operations		279008.34	4182893.02	11145494	15607395.4
Pacific Operations			352763.57	5983529.49	6336293.06
Sales and Marketing	**7246250.48**	**224589539**	**350072471**	**409085171**	**990993431**
AdventureWorks Cycle	7246250.48	224589539	350072471	409085171	990993431
North America Operations	7246250.48	223171639	328859798	328480126	887757814
European Operations		1417900.21	20308233.4	54363911.8	76090045.4
Pacific Operations			904439.11	26241132.5	27145571.6
Grand Total	**11384884.51**	**341162560**	**464113989**	**541978980**	**1358640413**

FIGURE 7.3 – Evolution du finance amount par département et par organisation dans le temps

Amount	Column Labels				
Row Labels	2018	2019	2020	2021	Grand Total
AdventureWorks Cycle	11384884.51	341162559.8	464113988.9	541978979.5	1358640413
North America Operations	11384884.51	339350040.4	436646384.9	438402423.9	1225783734
Canadian Division	2922350.51	78654132.35	107976739.1	102621560.7	292174782.7
USA Operations	8462534	260695908	328669645.8	335780863.2	933608951
European Operations		1812519.44	26075098.06	69415986.49	97303603.99
Pacific Operations			1392505.88	34160569.11	35553074.99
Grand Total	11384884.51	341162559.8	464113988.9	541978979.5	1358640413

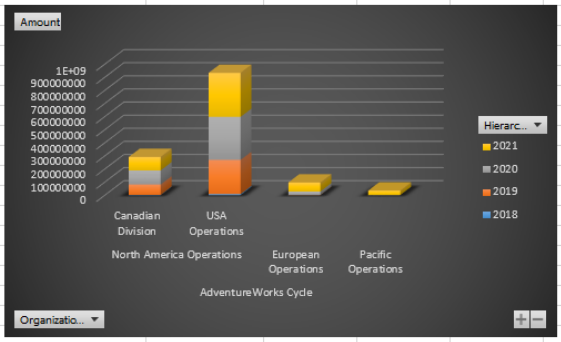


FIGURE 7.4 – Visualisation avec excel d l'évolution du finance amount par organisation dans le temps

Amount	Column Labels				
Row Labels	2018	2019	2020	2021	Grand Total
Actual	9766694.51	300368899.8	464113988.9	541978979.5	1316228563
Budget	1618190	40793660			42411850
Grand Total	**11384884.51**	**341162559.8**	**464113988.9**	**541978979.5**	**1358640413**

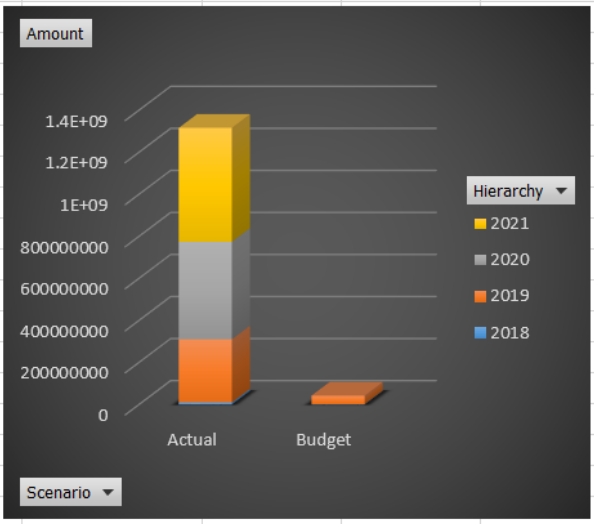


FIGURE 7.5 – Visualisation avec excel d l'évolution du finance amount par scénario

7.2 Restitution des données avec PowerBI :

Power BI est un logiciel interactif de visualisation de données développé par Microsoft, principalement axé sur la veille économique et faisant partie de la plateforme Microsoft Power. Power BI est une collection de services logiciels, d'applications et de connecteurs qui fonctionnent ensemble pour transformer des sources de données non liées en informations cohérentes, visuellement immersives et interactives.

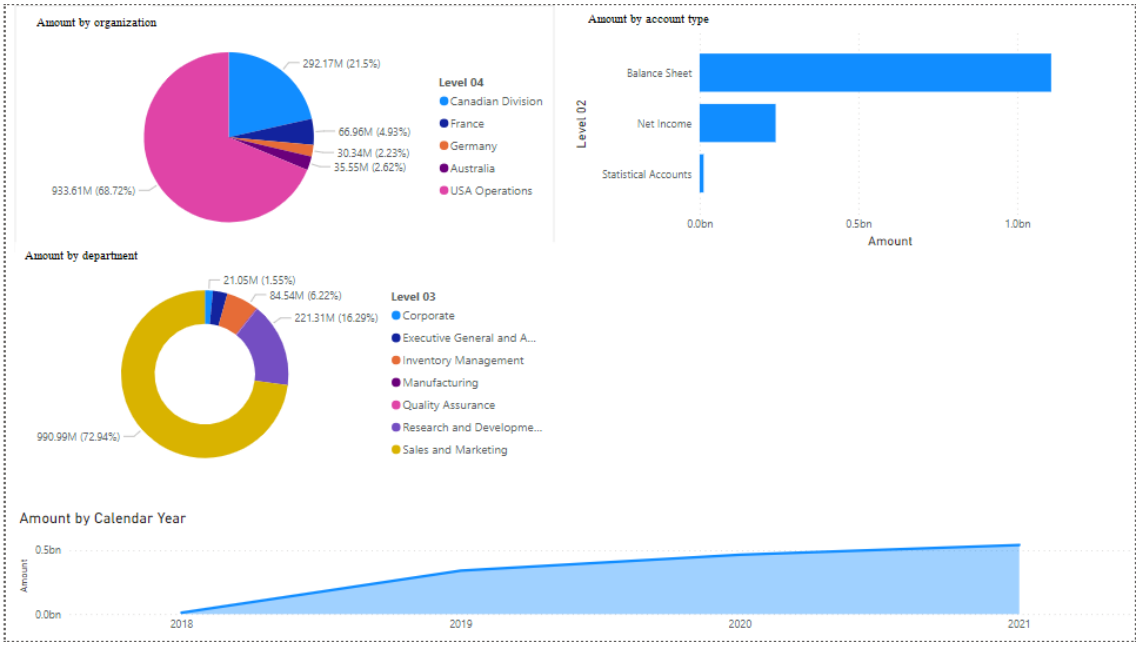


FIGURE 7.6 – Dashboard avec PowerBI

8 | Conclusion

Le projet DataWarehouse consiste à réaliser une Data Mart finance à partir d'AdventureWorkDB qui est une base de données open source. Ce cube facilitera la rédaction des rapports et des Dashboards permettant de garder le suivi de l'indicateur en fonction de plusieurs autres dimensions.

Pour répondre à nos objectifs, nous avons commencé par créer la base de données LightFinanceDW qui permettra d'extraire et stocker les données depuis AdventureWorkDW vers cette dernière. Ensuite, on a intégré les données vers LightFinanceDW ou on a créé l'architecture de notre flux de donnée, pour après générer notre cube contenant nos différentes dimensions et notre table de fait. Enfin, nous avons abordé la phase de reporting qui consiste à la mise en œuvre de la solution proposée. La réalisation a porté sur le choix de certains outils , en l'occurence SQL Server, Visual studio Data Tools, Power BI. La concrétisation de ce projet nous a permis de raffiner notre capacité de conception et de renforcer nos compétences en matière de technologies telles que Data Warehousing, ETL et le Reporting.

Toutefois, le projet reste toujours ouvert à des éventuelles améliorations.