

Mini Projet en Data Warehouse

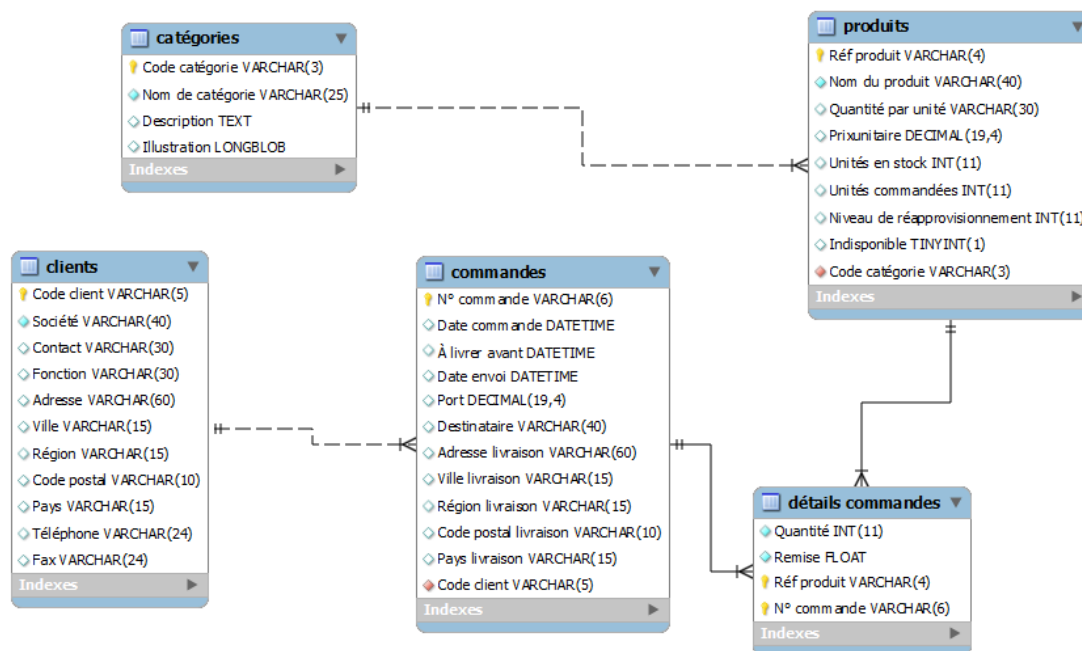
Objectif de ce projet est la Mise en œuvre d'un système d'aide à la décision à partir de la base de données comptoir (MySQL, Pentaho Data Integration,)

Modalités de réalisation du projet :

- Equipe de 2 personnes
- **Le travail doit être rendu sous la forme d'un rapport sur CD contenant la data et les résultats d'analyse OLAP avant le : Jeudi 22 décembre 2016**
- La date du contrôle oral individuel sera communiquée plus tard.

Data Warehouse

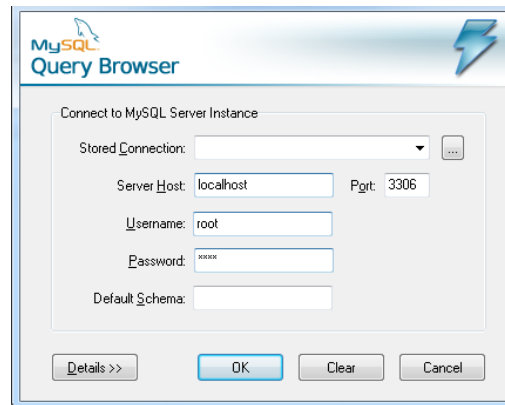
Soit le schéma de la base de données opérationnelle « comptoir » suivant :



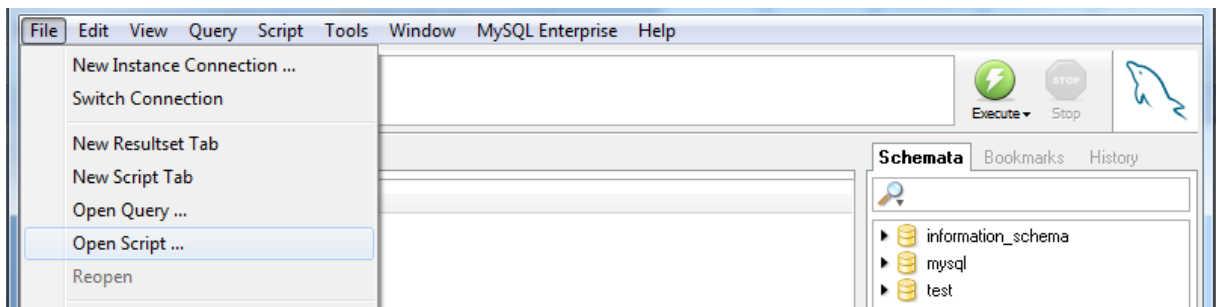
On veut à partir de cette base de données opérationnelle, produire un entrepôt de données pour étudier les performances de **Ventes** selon plusieurs dimensions : **Temps**, **Clients** et **Produits**. Pour ce faire nous allons suivre les étapes suivantes :

I. Préparation de la base de données en production

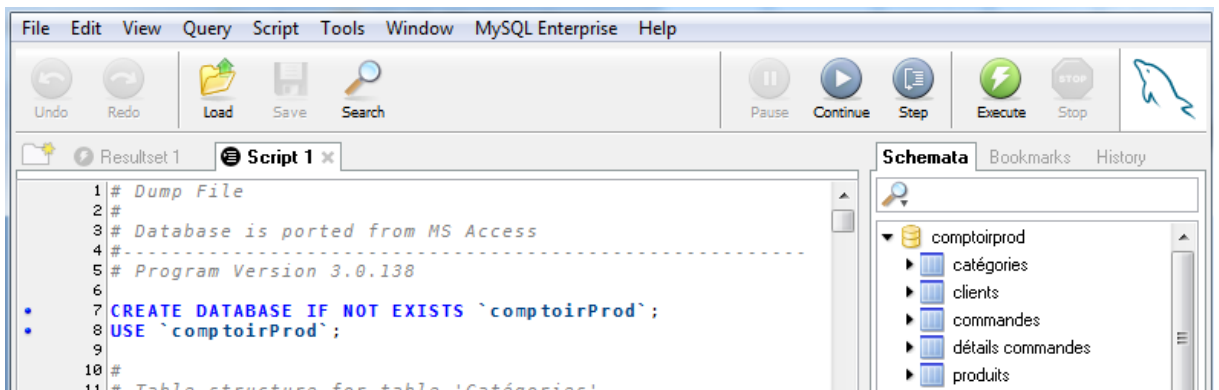
1. Démarrer l'outil MySQL Query Browser :



2. Utiliser le fichier dump « ComptoirDump.sql » pour créer la base de données comptoir.

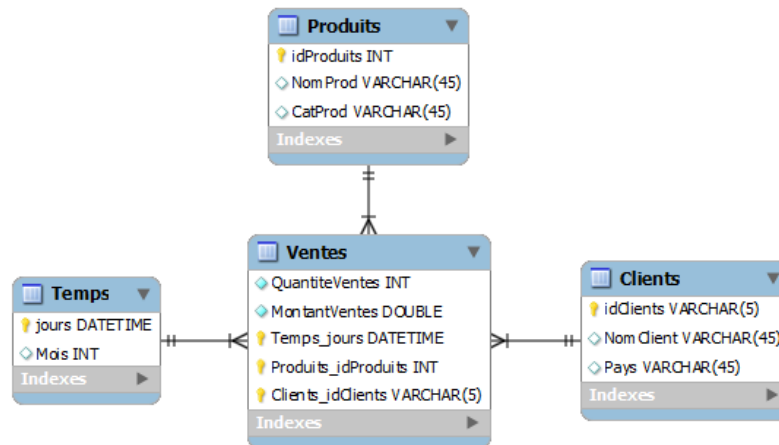


3. Exécuter le script SQL en utilisant la commande « Execute » pour avoir le schéma de la base de données suivant :



II. Conception du Datamart pour l'analyse du sujet ventes.

1. Proposer un schéma en étoile pour étudier les faits de ventes.
2. Implémenter le schéma, ainsi proposé, par l'outil MySQL Workbench et créer le script dump correspondant :
 - a. File >> New Model



b. Générer le script dump (Databases >>Forward engineer) :

```

SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0;
SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0;
SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='TRADITIONAL';
CREATE SCHEMA IF NOT EXISTS `ventes_datamart` DEFAULT CHARACTER SET latin1 COLLATE latin1_swedish_ci ;
USE `ventes_datamart` ;

-----
-- Table `ventes_datamart`.`Temps`
-----
CREATE TABLE IF NOT EXISTS `ventes_datamart`.`Temps` (
  `jours` DATETIME NOT NULL ,
  `Mois` INT NULL ,
  PRIMARY KEY (`jours`) )
ENGINE = InnoDB;

-----
-- Table `ventes_datamart`.`Produits`
-----
CREATE TABLE IF NOT EXISTS `ventes_datamart`.`Produits` (
  `idProduits` INT NOT NULL ,
  `NomProd` VARCHAR(45) NULL ,
  `CatProd` VARCHAR(45) NULL ,
  PRIMARY KEY (`idProduits`) )
ENGINE = InnoDB;

-----
-- Table `ventes_datamart`.`Clients`
-----
CREATE TABLE IF NOT EXISTS `ventes_datamart`.`Clients` (
  `idClients` VARCHAR(5) NOT NULL ,
  `NomClient` VARCHAR(45) NULL ,
  `Pays` VARCHAR(45) NULL ,
  PRIMARY KEY (`idClients`) )
ENGINE = InnoDB;
  
```

```

-----
-- Table `ventes_datamart`.`Ventes`
-----
CREATE TABLE IF NOT EXISTS `ventes_datamart`.`Ventes` (
  `QuantiteVentes` INT NOT NULL ,
  `MontantVentes` DOUBLE NOT NULL ,
  `Temps_jours` DATETIME NOT NULL ,
  `Produits_idProduits` INT NOT NULL ,
  `Clients_idClients` VARCHAR(5) NOT NULL ,
  PRIMARY KEY (`Temps_jours`, `Produits_idProduits`, `Clients_idClients`) ,
  INDEX `fk_Ventes_Produits1` (`Produits_idProduits` ASC) ,
  INDEX `fk_Ventes_Clients1` (`Clients_idClients` ASC) ,
  CONSTRAINT `fk_Ventes_Temps`
    FOREIGN KEY (`Temps_jours` )
    REFERENCES `ventes_datamart`.`Temps` (`jours` )
    ON DELETE NO ACTION
    ON UPDATE NO ACTION,
  CONSTRAINT `fk_Ventes_Produits1`
    FOREIGN KEY (`Produits_idProduits` )
    REFERENCES `ventes_datamart`.`Produits` (`idProduits` )
    ON DELETE NO ACTION
    ON UPDATE NO ACTION,
  CONSTRAINT `fk_Ventes_Clients1`
    FOREIGN KEY (`Clients_idClients` )
    REFERENCES `ventes_datamart`.`Clients` (`idClients` )
    ON DELETE NO ACTION
    ON UPDATE NO ACTION)
ENGINE = InnoDB;
SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;

```

3. Utiliser le fichier dump ainsi généré pour créer le schéma de l'entrepôt de donnée en MySQL.

III. Introduction à Pentaho Business Analytics.

1. Présentation de Pentaho

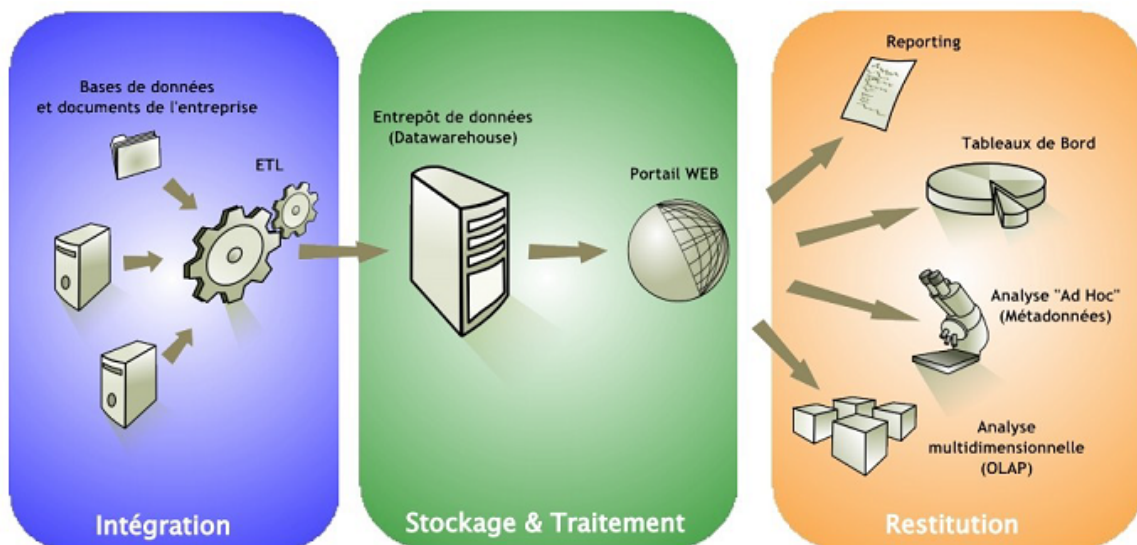
Pentaho est une plate-forme décisionnelle open source complète possédant les caractéristiques suivantes :

- Mondrian - Open Source OLAP Server. It implements the MDX language, and the XML for Analysis (XMLA) and JOLAP specifications.
- JFreeReport - Open Source Reporting
- Kettle - Open Source Data Integration (ETL)

- Pentaho - Comprehensive Open Source BI Suite Weka - Open Source Data Mining

C'est une couverture globale des fonctionnalités de la Business Intelligence :

- ETL (intégration de données),
- reporting,
- tableaux de bord ("Dashboards"),
- analyse ad hoc (requêtes à la demande),
- analyse multidimensionnelle (OLAP) ;



- Pentaho permet d'adresser deux typologies d'utilisateurs :
 - les « one-clic users », utilisateurs de base, consommateurs d'indicateurs prédéfinis,
 - les utilisateurs avancés, qui ont besoin d'outils d'analyse et d'exploration avancés ;



➔ Reporting :

- Rapports & états préformatés (pdf, excel, html)
- Filtrage possible des données ("row level")

➔ Tableaux de bords :

- Présentation graphique et synthétique d'indicateurs
- Permettent l'obtention intuitive de rapports ("drill-down")

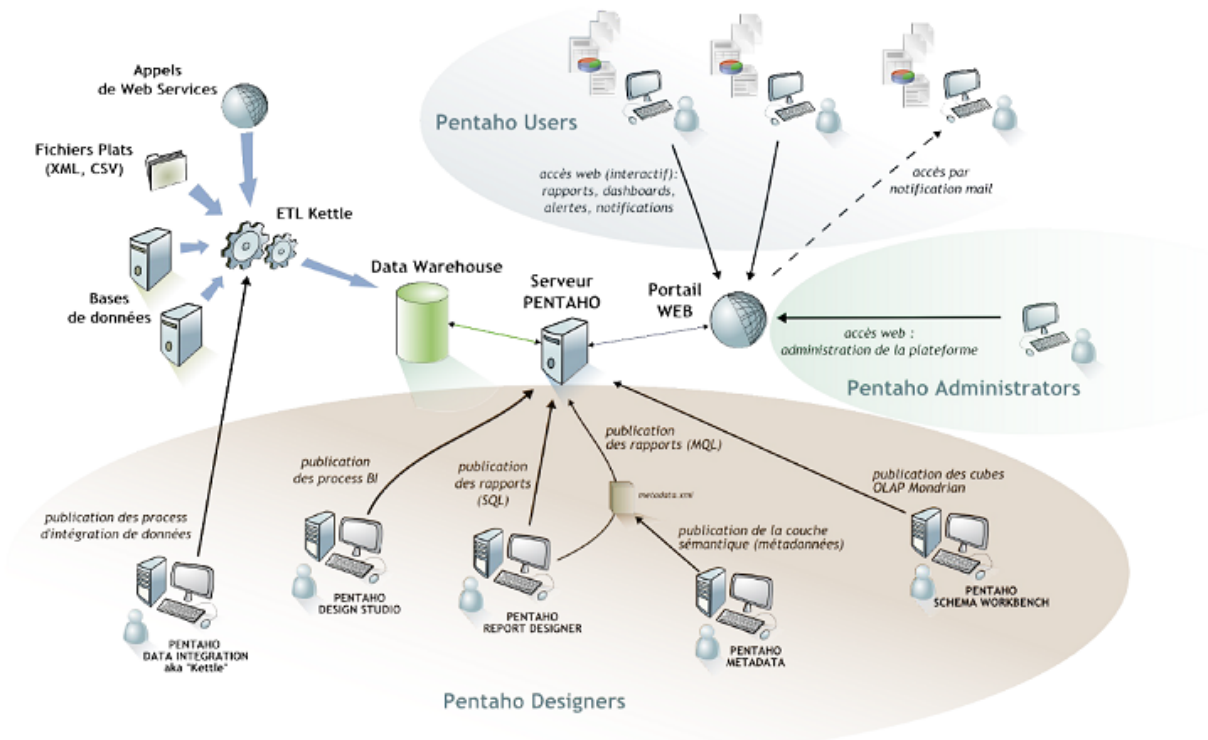
➔ Analyse Ad hoc :

- Permet de créer des rapports avancés (choix des colonnes, tableaux croisés) à partir de vues métiers ("business views"). Nécessite la mise en place d'une couche sémantique d'abstraction (métadonnées)

➔ Analyse multidimensionnelle (OLAP)

- Permet la manipulation de données selon plusieurs axes d'analyse. Nécessite une modélisation spécifique dans le SGBDR (ROLAP)

- une architecture Web 2.0 qui se compose :
 - d'un serveur Web J2EE permettant de mettre à disposition l'ensemble des ressources décisionnelles et ceci au travers d'URL Web uniques et standardisées. Le serveur est dénommé "Pentaho User Console" (PUC),
 - plusieurs clients riches permettant la conception et la publication des ressources. Ces derniers sont librement téléchargeables et peuvent être installés sous des environnements Windows, Linux ou Mac-OS (clients Java) ;



- le serveur Web Pentaho comporte également une plate-forme d'administration (Pentaho Administration Console) pour la gestion des droits d'accès, la planification d'évènements, la gestion centralisée des sources de données... ;
- Pentaho est reconnue pour être une solution d'une grande qualité conceptuelle et technique. La plate-forme est orientée « processus » : au travers de « séquences d'actions » on peut ainsi modéliser avec Pentaho des workflows BI avancés ;
- il n'est pas besoin de connaître JAVA pour travailler avec Pentaho : seule la maîtrise du langage SQL est nécessaire, ainsi que des connaissances de base en XML, HTML et JavaScript. Il faut bien sûr s'autoformer (ou être formé) aux clients de conception ;
- une communauté importante et très active s'anime autour de Pentaho. Celle-ci contribue au codage de nombreux plugins et de projets communautaires : plugins Kettle, Pentaho Analysis Tool, Pentaho Community Dashboard Framework, etc. ;
- Pentaho est une suite décisionnelle open source commerciale qui reste très « ouverte ». Les différences fonctionnelles entre la version libre (community edition) et la version payante (enterprise edition) restent limitées. La version libre de Pentaho permet d'installer une plate-forme décisionnelle complète !

2. Liste des clients de conception Pentaho

Outil	Code	Fonction
Pentaho Report Designer	PRD	Client de conception de rapports avancés. Il s'agit d'un outil de mise en page similaire à iReport, Eclipse BIRT, Crystal Reports... Permet de se connecter à de nombreuses sources de données : SGBD, XML, Excel, CSV, flux de données venant de Kettle, MDX (OLAP)...
Pentaho Design Studio	PDS	Client Eclipse de modélisation de workflows BI (Xactions) propre à Pentaho. Design Studio permet de mettre en œuvre de nombreuses ressources BI en minimisant l'écriture de code (envoi de mails automatisé par ex.).
Pentaho Metadata	PME	Client riche permettant la mise en place d'une couche sémantique d'abstraction (métadonnées) sur la couche physique (tables et colonnes d'une base de données). Le but est de rendre les objets d'un SGBD compréhensibles et manipulables par un utilisateur final afin de lui permettre d'effectuer ses propres requêtes et ceci sans connaître le langage SQL. La couche de métadonnées peut-être utilisée dans le requêteur Web ad hoc, dans Pentaho Report Designer et dans Pentaho Design Studio.
Pentaho Schema Workbench	PSW	Client riche permettant la définition des schémas Mondrian à partir d'un modèle en étoile ou flocon de l'entrepôt de données. Un autre outil, Pentaho Aggregation Designer (PAD), permet de construire et de charger automatiquement des tables d'agrégation en vue d'améliorer les performances lors du requêtage des cubes Mondrian. Téléchargement spécifique ici : http://sourceforge.net/projects/mondrian/files/

Pentaho Data Integration (Kettle)	PDI	<p>Outil ETL (Extract Transform Load) complet, pouvant être utilisé indépendamment de la plate-forme Pentaho. Kettle est comparable à Talend Open Studio en termes de fonctionnalités. Pour consulter les différences techniques et fonctionnelles, un livre blanc est disponible à cette adresse :</p> <p>http://www.atolcd.com/actualites/detail-actualite/actualite/2/comparatif-etopen-source-1.html</p>
-----------------------------------	-----	--

3. Démarrer & arrêter les serveurs Pentaho

a. Pentaho User Console

Les commandes suivantes permettent de lancer et stopper la console d'utilisation Web

Action	Commande
Démarrage (Windows)	{PENTAHO-HOME}\biserver-ce\start-pentaho.bat
Arrêt (Windows)	{PENTAHO-HOME}\biserver-ce\stop-pentaho.bat
Démarrage (Linux)	{PENTAHO-HOME}\biserver-ce\start-pentaho.sh
Arrêt (Linux)	{PENTAHO-HOME}\biserver-ce\stop-pentaho.sh

On accède à la console d'utilisation Pentaho en saisissant l'URL suivante dans un navigateur Web :

<http://localhost:8080/pentaho>

Puis en saisissant l'identifiant et mot de passe ci-dessous :

- login : joe ;
- password : password.

b. Pentaho Administration Console

Les commandes suivantes permettent de lancer et stopper la console d'utilisation Web

Action	Commande
Démarrage (Windows)	{PENTAHO-HOME}\administration-console\start-pac.bat
Arrêt (Windows)	{PENTAHO-HOME}\administration-console\stop-pac.bat
Démarrage (Linux)	{PENTAHO-HOME}\administration-console\start-pac.sh
Arrêt (Linux)	{PENTAHO-HOME}\administration-console\stop-pac.sh

On accède à la console d'administration Pentaho en saisissant l'URL suivante dans un navigateur Web :

http://localhost:8099

Puis en saisissant l'identifiant et mot de passe ci-dessous :

- login : admin ;
- password : password.

IV. Intégration des données à travers la couche ETL.

1. **Après avoir installer le logiciel Pentaho, copier le fichier « mysql-connector-java-5.1.28-bin.jar » dans les répertoires suivants :**

- a. C:\Program Files\pentaho\design-tools\data-integration\lib
- b. C:\Program Files\pentaho\design-tools\aggregation-designer\lib
- c. C:\Program Files\pentaho\design-tools\metadata-editor\lib
- d. C:\Program Files\pentaho\design-tools\report-designer\lib
- e. C:\Program Files\pentaho\design-tools\schema-workbench\lib

2. **Lancer le serveur Pentaho Data Integration PDI**

3. **Utiliser Pentaho Data Integration pour alimenter l'entrepôt en utilisant les données provenant de la base de données en production.**

- a. Créer une nouvelle transformation : Fichier >> Nouveau >> Transformation.
- b. Ajouter une fonction d'extraction : Extraction depuis Table.
- c. Renommer la fonction d'extraction : Extraction produits.
- d. Créer une nouvelle connexion à la base de données comptoirprod:
cn_prod

General
Advanced
Options
Pooling
Clustering

Connection Name:
cn_prod

Connection Type:
MonetDB
MySQL
Native Mondrian
Neoview
Netezza
OpenERP Server
Oracle
Oracle RDB
Palo MOLAP Server
PostgreSQL
Remedy Action Request System

Access:
Native (JDBC)
ODBC
JNDI

Settings
Host Name: localhost
Database Name: comptoirprod
Port Number: 3306
User Name: root
Password: ****
☒ Use Result Streaming Cursor

Tester Spécification Explorer

OK Cancel

e. Tester la connexion :



f. Intégrer la requête sql permettant d'extraire les granularités de la dimension produit :

```
SELECT DISTINCT p.`Réf produit` , p.`Nom du produit` , c.`Nom de catégorie` FROM
comptoirprod.produits p, comptoirprod.catégories c where p.`Code catégorie`=c.`Code
catégorie`;
```

- g. Ajouter une fonction d'alimentation : Insertion dans table
- h. Renommer la fonction d'alimentation: Alimentation dimension produit
- i. Créer une nouvelle connexion à l'entrepôt ventes_datamart:
cn_datamart

General
Advanced
Options
Pooling
Clustering

Connection Name:
cn_datamat

Connection Type:
MySQL
Native Mondrian
Neoview
Netezza
OpenERP Server
Oracle
Oracle RDB
Palo MOLAP Server
PostgreSQL
Remedy Action Request System
SAP FRP System

Access:
Native (JDBC)
ODBC
JNDI

Settings
Host Name:
localhost
Database Name:
ventes_datamat
Port Number:
3306
User Name:
root
Password:

☒ Use Result Streaming Cursor

Tester Spécification Explorer

OK Cancel

j. Configurer correctement le mapping des champs :

Nom étape Alimentation dimension produits

Connexion cn_datamat [Modifier...] [Nouvelle...] [Assistant...]

Schéma cible [Parcourir...]

Table cible produits [Parcourir...]

Valider transaction toutes les 1000

Tronquer la table ☐

Ignorer les erreur d'insertion ☐

Sélectionner champs ☒

Général Champs table

Champs:

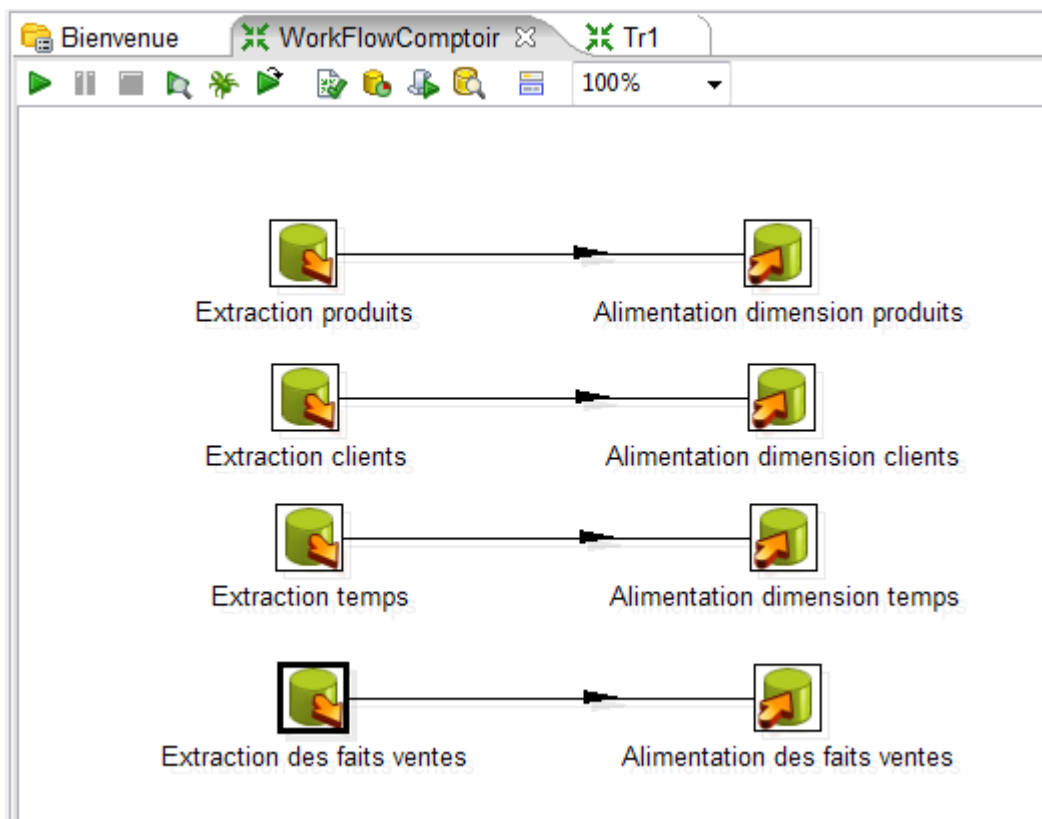
#	Champ table	Champ flux
1	idProduits	Réf produit
2	NomProd	Nom du prod...
3	CatProd	Nom de caté...

[Copier champs] [Editer schéma]

Help OK Annuler SQL

k. Refaire la même chose pour les deux autres dimensions et la table de faits ventes. Ci-après les requêtes d'agrégation :

- i. `SELECT DISTINCT c.`Code client`, c.`Société`, c.`Pays`
FROM comptoirprod.clients c;`
- ii. `SELECT DISTINCT Date(c.`Date commande`), month(c.`Date
commande`) FROM comptoirprod.commandes c;`
- iii. `SELECT
sum(d.`Quantité`),sum(d.`Quantité`*p.prixunitaire),
date(c.`Date commande`),p.`Réf produit`, clt.`Code
client` FROM comptoirprod.`détails commandes` d ,
comptoirprod.`produits` p , comptoirprod.`commandes` c ,
comptoirprod.`clients` clt where clt.`Code client` =
c.`Code client`and c. `N° commande`=d.`N° commande` and
p.`Réf produit`=d.`Réf produit` group by month(c.`Date
commande`),p.`Réf produit`, clt.`Code client`;`

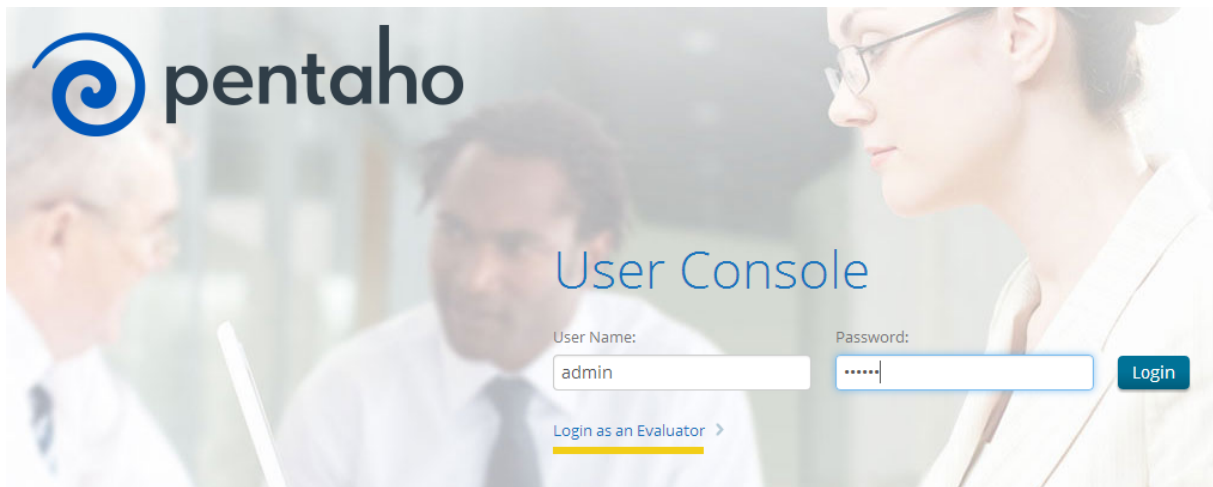


- l. Démarrer le workflow ETL.
- m. Vérifier que votre datamart est bien rempli.

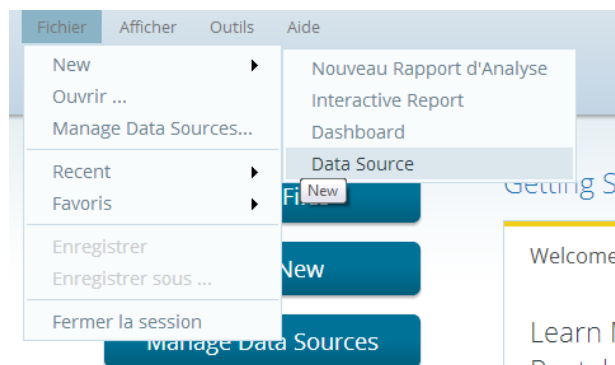
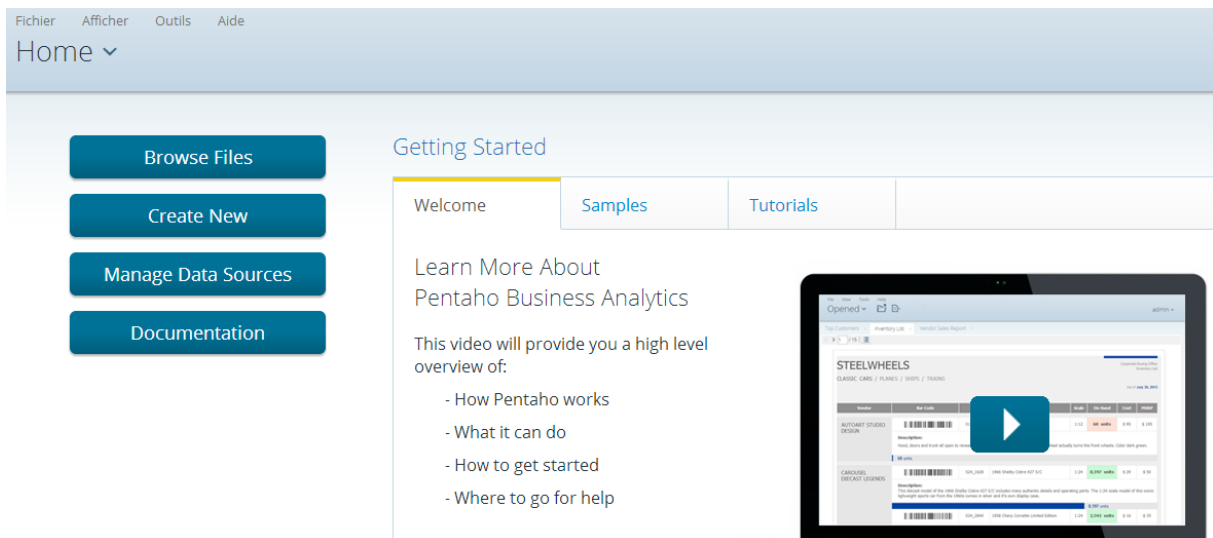
V. Analyse OLAP et Reporting. (La version du Pentaho BA utilisée ici n'est pas la plus récente, plusieurs changements peuvent être constatés)

N.B. : Cette partie peut être réalisée par n'importe quel outil de ton choix.

1. Démarrer le serveur Business Analytics
2. Lancer le « User console » et connectez-vous avec le compte admin



3. Créer une nouvelle source de données à partir de l'entrepôt de données :



> Selection type




Selection tables

Jointures

Nom source de données:
MyDataMart

Type de Source:
Base de donnees

Sélectionner une base de données et cliquer sur Suivant pour choisir les tables disponibles.

Connexion:   

SampleData
cn_datamart
pentaho_operations_mart

Créer source de données pour:

☐ Uniquement rapport

☒ Rapport et Analyse(Nécessite un schéma en étoile)


4. Vous devez ajouter la connexion « cn_datamart » :

Database Connection

General
Advanced
Options
Pooling

Connection Name:
cn_datamart

Database Type:
MySQL
Hadoop Hive
Hadoop Hive 2
Hypersonic
Impala
MS SQL Server
PostgreSQL
H2
MonetDB

Adding Databases 

Access:
Native (JDBC)
ODBC
JNDI

Settings

Host Name:
localhost

Database Name:
ventes_datamart

Port Number:
3306

User Name:
root

Password:
....

Test

5. Sélectionner les tables de dimensions et la table de faits :

Tables disponibles:

Tables sélectionnées:

>

<

Table de fait:

< Retour

Suivant >

Fin

Annuler

6. Créer les jointures du schéma en étoile :

Table gauche:

Table droite:

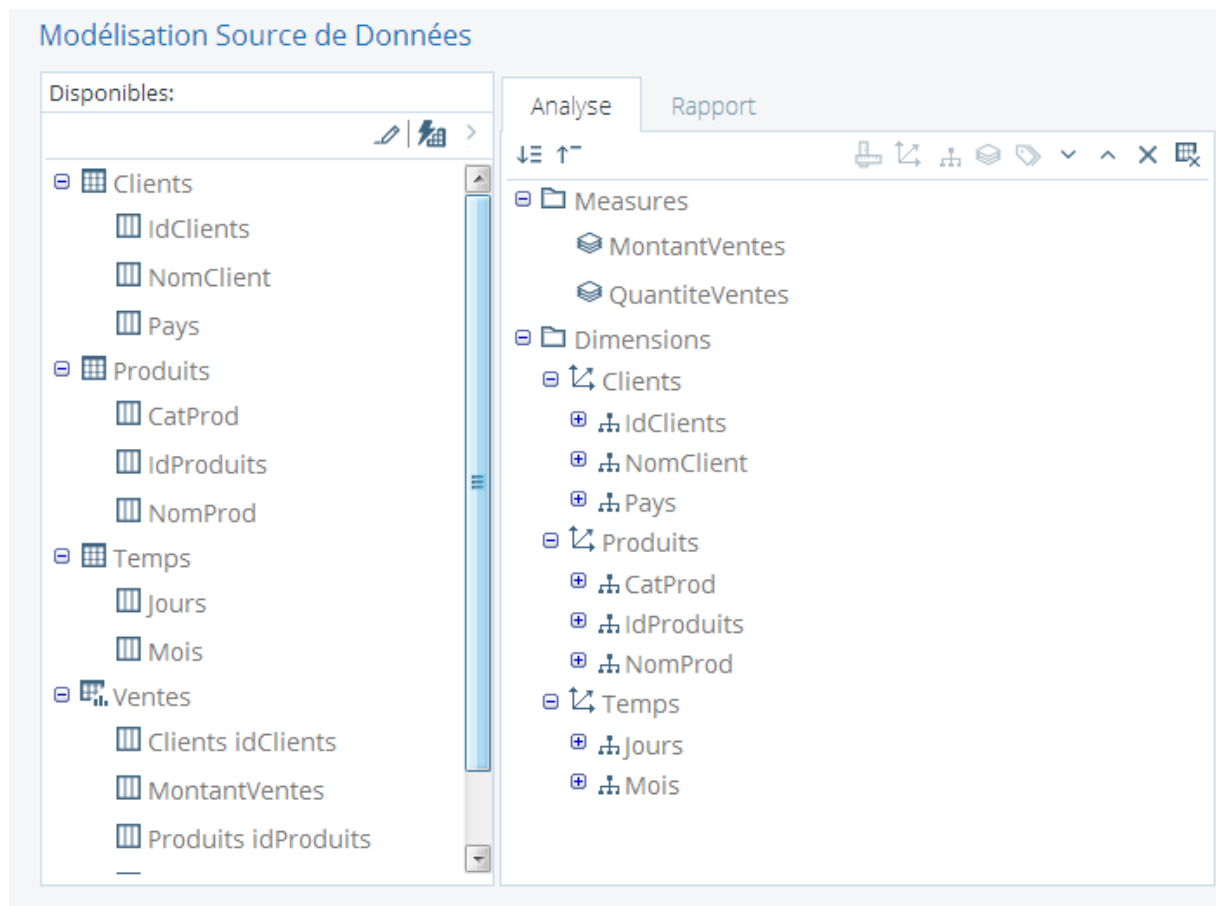
Champ clef:

Champ clef:

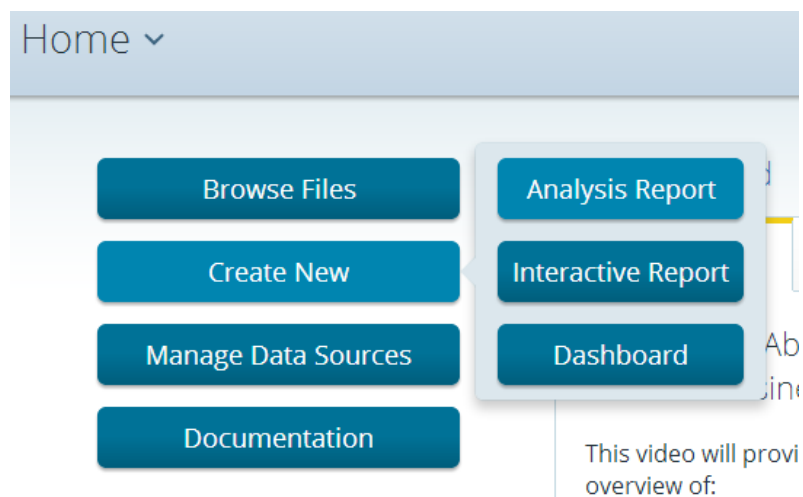
Créer jointure

Joitures:

7. Vérifier la modélisation et les méta-données du datamart



8. Création d'un rapport d'analyse



Analysis Report x

Champs disponibles (10)
pour:
MyDataMart

Chercher
Visualiser

CatProd
IdClients
IdProduits
Jours
Mesures
MontantVentes
QuantiteVentes
Mois

Mise en page

Lignes
Pays
Deplacer le niveau ici

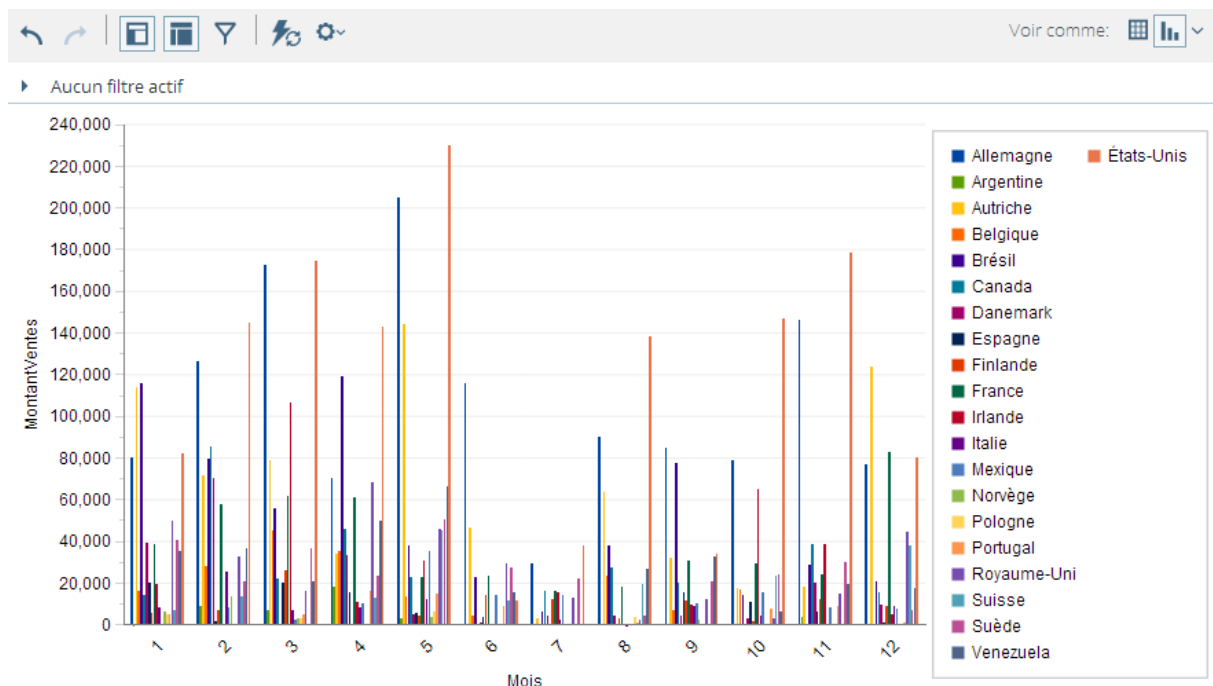
Columns
Mois
Deplacer le niveau ici

Mesures
MontantVentes
Deplacer la mesure ici

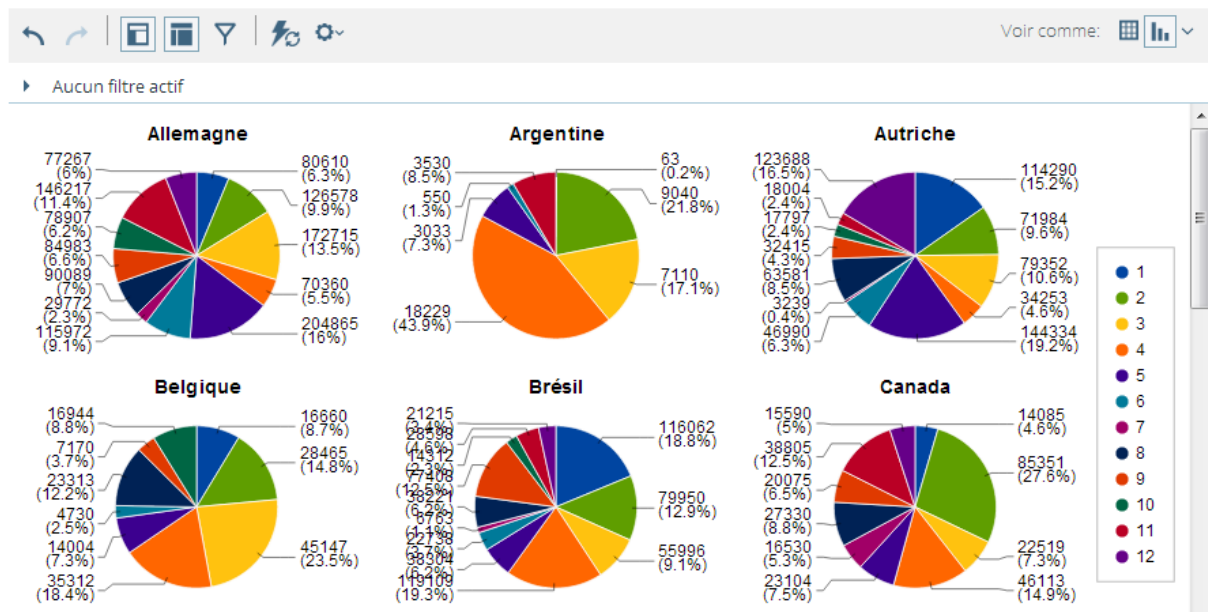
Aucun filtre actif

	Mois
	1
Pays	MontantVentes
Allemagne	80610
Argentine	63
Autriche	114290
Belgique	16660
Brésil	116062
Canada	14085
Danemark	39174
Espagne	20179
Finlande	5662
France	38750

9. Donner le mois et le pays où les ventes ont été les meilleurs :

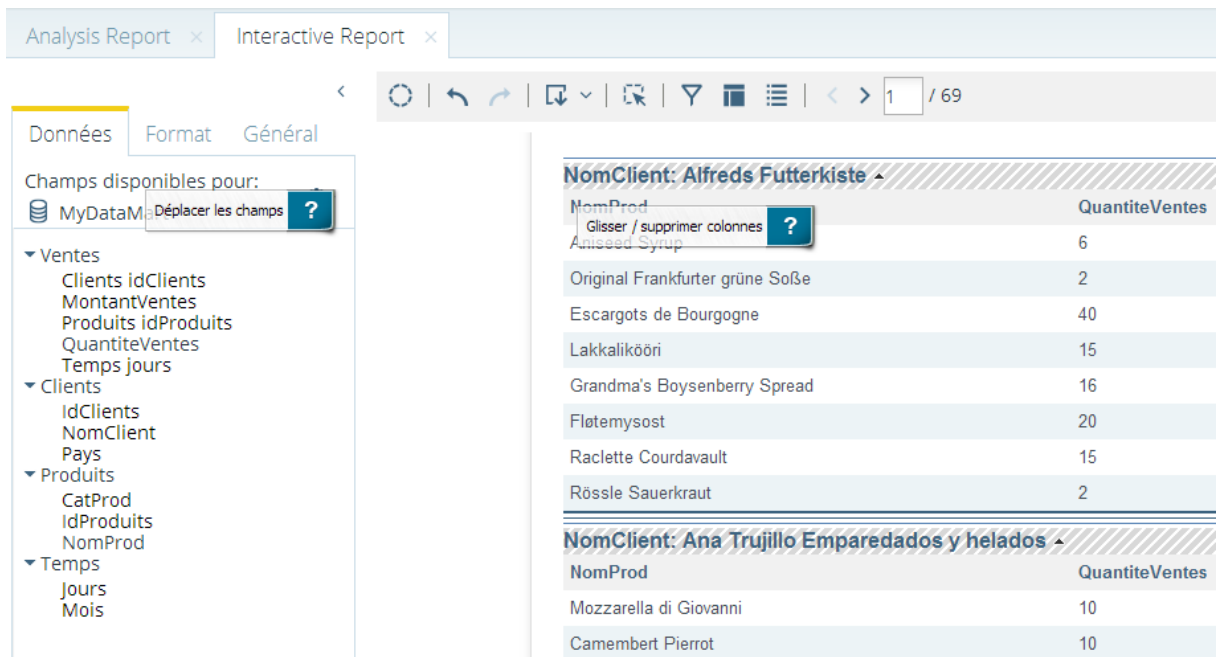


10. Donner le pays où les ventes sont réparties sur le nombre le plus réduit des mois :



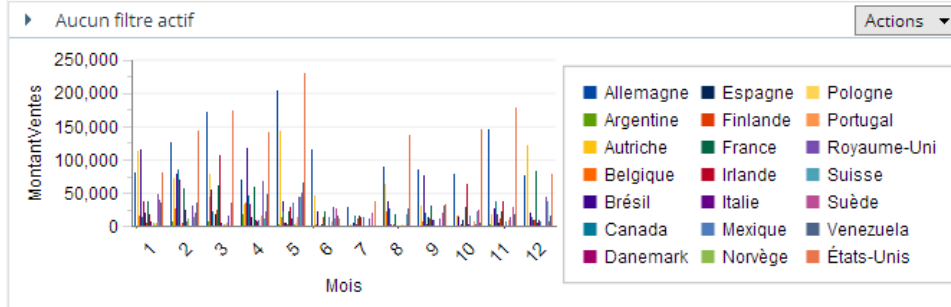
11. Création d'un rapport interactif

- a. Dans un rapport interactif, afficher pour chaque client le nom du produit et la quantité achetée :

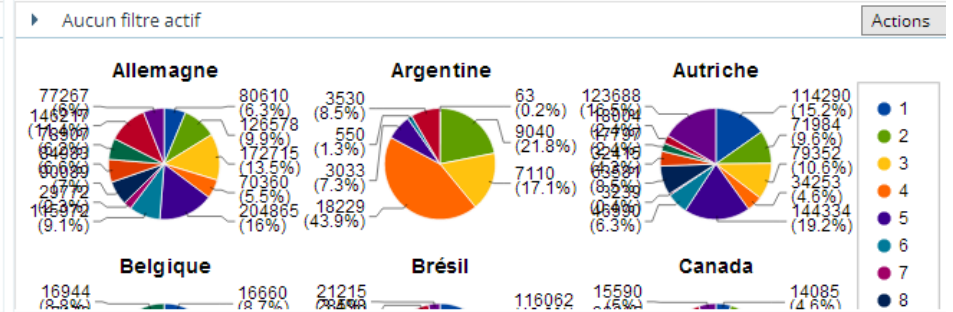


12. Création des Dashboards (les tableaux de bord dynamiques) :

Sans titre 1



Sans titre 2



Sans titre 3

< > 1 / 69

Aniseed Syrup	6
Original Frankfurter grüne Soße	2
Escargots de Bourgogne	40
Lakkalikööri	15
Grandma's Boysenberry Spread	16
Fløtemysost	20

