L’inventaire des œuvres du musée d’art de Nantes

Les étudiants

Corentin Marionneau
Alexis Lamothe
Merlin Barzilai

Sont fiers de vous présenter
leur compte rendu de projet

Table of contents

Table des matières

Projet de Base de données OLAP

Introduction

Ce document présente un projet de master 1 d'informatique. Le sujet porte sur l'intégration d'une base de donnée de type OLAP, c'est à dire une base en lecture seule, structurée pour faciliter l'analyse.

L'objet de cette intégration sera l'inventaire des œuvres du musée d'art de Nantes. Il s'agit d'un dataset issu de l'ouverture de donnée de la ville de Nantes.

Présentation du dataset

Nous avons récupéré un dataset de l'ouverture de données de Nantes, qui porte sur l'inventaire des œuvres du musée d'art de Nantes.
Ce dataset consiste en une liste d'œuvres d'art pour lesquelles on indique

- Les noms et prénoms de l'auteur
- Les dates et lieux de naissance et de décès de l'auteur
- La date de réalisation de l'œuvre
- Le domaine de l'œuvre
- La technique
- Les dimensions
- L'année d'acquisition
- Le moyen d'acquisition
- Des précisions sur l'acquisition
- le n° d'inventaire
- le lien Navigart

Il s'agit d'un document assez complet, qui ouvre sur un certain nombre de possibilités d'exploitation ; comparaisons de dimensions, observation des tendances de types d'œuvre en fonction du temps, etc.

De plus il s'agit d'un ensemble de données que l'on peut aisément lier à d'autres par les informations relatives aux dates ou aux auteurs, comme par exemple des données sur l'identité, ou des données sur les noms de rues dans Nantes par exemple.

Nettoyage du dataset

Dans l'état dans lequel nous l'avons récupéré le dataset était assez chaotique. Un certain nombre de valeurs n'étaient pas standardisée, certaines informations étaient accumulées dans un unique champs alors qu'elles auraient dues être distinguées.

Nous avons utilisé le logiciel gratuit Talend data preparation, afin de préparer nos données avant leur intégration dans la base SQL.

Dimension

Les dimensions des œuvres étaient exprimée sous la forme «*x*x* cm » la plupart du temps. Une première étape simple de la préparation des données consistait en sa séparation en 3 colonnes ; x, y, et z (le cas échéant).

Talend à facilité le processus puisqu'il permet de faire cette séparation en choisissant le caractère de séparation.

Cela dit certaines données ne respectaient pas le format et il a fallu passer par une étape

préliminaire.

Nous avons commencé par spliter nos valeurs avec le délimiteur « cm ». D'abord ceci permettait de se débarrasser de la mention qui ne nous intéressait pas et qui n'était pas convertible en valeur numérique. Deuxièmement cela à permis d'extraire dans une colonne toutes les valeurs qui ne respectaient pas la norme (qui avait quelque chose après leur simple mention de dimensions).

En regardant ces données il en est apparu deux types :

- Les spécifications sur les dimensions hors marge.
- Les spécification sur le diamètre d'une œuvre

Dans le premier cas nous nous sommes contentés de décréter que les dimensions hors marge ne seraient pas prises en compte, car nous voulions les dimensions complètes des œuvres.

Dans le second cas avons envisagé de rectifier les données. Mais nous avons finalement décidé de retirer ces entrées de la table et ce pour trois bonnes raisons.

- La plupart du temps les données ne nous permettaient pas de corriger. Par exemple nous avions des œuvres avec des dimensions 15x20, sans profondeur avec un diamètre de 33. Dans ce genre de situation nous ne pouvions pas être sûrs de la manière d'interpréter ces valeurs
- Il y avait peu de données dans ce cas. Environ une vingtaine pour un dataset de quasiment six-mille entrées
- Nous pouvions le faire

Dates et lieux de naissance et de mort des artistes

Lorsque nous avons récupérer le dataset, ces informations étaient concaténées dans une unique colonne. Bien évidemment sous cette forme, la donnée était complètement inexploitable. Heureusement les valeurs semblaient suivre un schéma standard, de la forme « ddn, ldn – ddm, ldm ».

Grâce aux fonctions de Talend nous avons pu spliter deux fois ce document, une fois selon naissance ou mort, puis une deuxième selon lieu ou date. Nous avons ainsi obtenu quatre colonne parfaitement exploitables. À ceci près que l'observation à révélé la présence de lignes mal formatées dans le dataset. Le séparateur à pu être oublié ou confondu avec l'autre. Cela donne des dates de naissance comme « 1993Paris ». Ces problèmes tragiques ont pu être corrigés lorsque nous avons forcé le typage de nos dates en tant que nombres.

On peut également noter que nous avons soustraites du dataset toutes les lignes dont les dates et lieux de naissance de l'auteur manquaient et ce pour trois bonnes raisons :

- Lorsque de tels cas se révélaient, nous n'avions pas d'information non plus sur le nom de l'auteur, sa date et son lieu de mort, et d'une manière générale, peu d'information sur l'œuvre
- Ces cas étaient rares, au même titre que les précisions sur le diamètre des œuvres
- Nous pouvions le faire

Noms et prénoms des auteurs

La dernière donnée a préparer concernait les noms et prénoms de auteurs. Comme pour les dates et lieux de naissance et de mort, les noms et prénoms étaient regroupés dans une unique colonne. En soi ce n'est pas un problème tant que l'on fait des requêtes en interne (quoi que ça peut être pertinent pour chercher des filiations, même si c'est improbable).

L'intérêt de séparer noms et prénoms, c'est d'une part pour augmenter la lisibilité. En effet cela rend l'ensemble plus cohérent, les gens ont un nom et un prénom. Pas un nomprénom. Et parfois ils ont d'autres choses qui ne sont ni nom, ni prénom, et qui devraient être dans une autre catégorie. « titre » par exemple.

D'autre part, cela est plus pratique si il faut combiner nos données avec celles d'un autre groupe. Il est beaucoup plus facile de matcher l'un ou l'autre du nom ou du prénom, que les deux en même

temps.

En définitive nous avons trié les noms et prénoms en prenant avantage de leur formatage. À savoir que les noms étaient en majuscules, et les prénoms en minuscules. Il y a cependant eu quelques données qui ne sont pas passées correctement au filtre, comme « Charles MELLIN dit Charles de LORRAINE ». Selon le cas nous avons pu :

- Corriger la donnée (erreur de typo)
- Laisser tel quel (personne dont le titre est le nom. Ex : « Cedric de la TOURBIERE »)
- Retirer la ligne si elle nous posait vraiment problème

Structuration

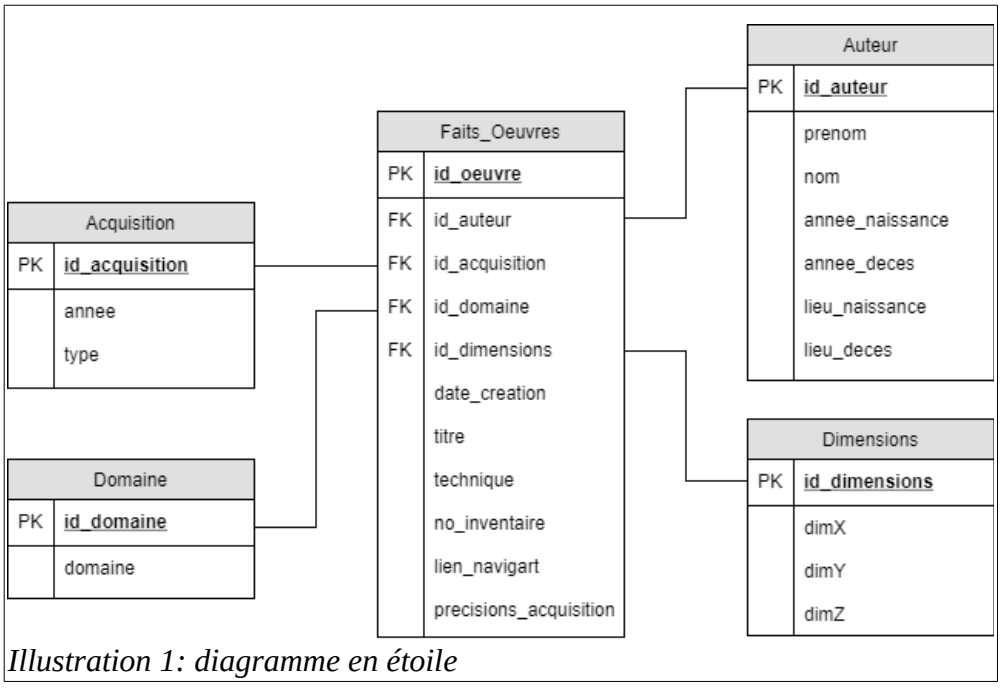

Illustration 1: diagramme en étoile

Nous avons décidé de centrer notre modèle de donnée autour des œuvres, qui constituent ainsi notre table de faits.

C'est une décision qui nous semble logique et assez directe puisque ce c'est clairement l'aspect qui est au centre de nos données brutes :

« Inventaire des œuvres du musée d'art de Nantes »

C'est dans le titre.

Y sont reliées nos dimensions :

Auteur

C'est une dimension qui vient assez naturellement pour quelques raisons :

- C'est facile de se le représenter : il n'y a aucun mal à imaginer qu'un auteur soit une entité extérieure à son œuvre
- Cela synthétise l'information : il y a des œuvres qui disposent du même auteur, c'est un situation fréquente. Dès lors, l'auteur, bien qu'il soit une propriété de l'œuvre, il n'est plus une propriété unique
- C'est un axe d'analyse pertinent : il y a moyen de mettre en évidence des liens et des tendances spécifiques à des artiste, et ce, d'une manière intéressante

Dimensions

La dimension dimensions regroupe les différentes mesure dans les deux ou les trois dimensions. Bien qu'elle ne semble pas forcément indispensable elle présente quelques intérêts :

- Cela synthétise l'information : certaines œuvres partagent des dimensions, notamment quand il s'agit d'ensembles d'œuvres
- C'est une donnée sur laquelle il est facile de faire de la comparaison, y compris avec des données extérieures.
- Cela fait une dimension dimensions, ce qui est assez rare pour être souligné

Acquisition

Toutes les œuvres ont une origine et sont arrivés dans le musée par un biais. C'est ce biais qui est explicité plus ou moins informativement dans la dimension acquisition. Nous avons choisi d'en faire une dimension car :

- C'est facile de se le représenter : le mode d'acquisition ne fait pas partie intégrante de l'œuvre. l'acquisition est un processus externe et il est justifié de l'extraire.
- Cela synthétise l'information : Il y a au final peu de modes d'acquisitions différentes, même si il peut y avoir beaucoup de précisions et de description sur le processus.

Domaine

Le domaine de l'œuvre représente son type ; sculpture, peinture, etc. On en fait une dimension pour des raisons simple :

- Il y a peu de valeurs possible, ce qui signifie que l'on peut indexer les valeurs en question
- Il est pertinent de vouloir classer les œuvres en fonction de leur domaine, car cela permet de créer un contexte de comparaison pertinent (il est n'est pas pertinent de comparer une sculpture à une peinture, en terme de dimensions par exemple)

Intégration

Une fois notre nettoyage effectué nous avons alors un fichier csv possédant une vingtaine de colonnes, certaines non exploitées car ayant été décomposées en plusieurs autres colonne exploitables (Exemple : Dates_naissance_deces_artiste dans la table originelle laissera sa place dans l'entrepôt aux 4 colonnes Lieu_naissance, Lieu_deces, Annee_naissance et Annee_deces qui iront dans la table Dim_Auteur).

Avant d'utiliser ce fichier, il faut tout d'abord créer les tables de l'entrepôt. Pour cela, nous avons utilisé Oracle pour créer 5 tables : Faits_Oeuvres, Dim_Auteur, Dim_Domaine, Dim_Dimensions et Dim_Acquisition, tel que spécifié dans notre modélisation en étoile vu précédement.

Table de faits

Le gros du travail était évidement de créer la table des faits. Elle est donc composée de 11 colonnes :

- id_oeuvre : un nombre incrément automatique identifiant l'entrée, clé primaire
- id_auteur : identificateur de l'auteur (du texte, son nom complet), clé étrangère vers la table Dim_Auteur
- id_acquisition : identificateur de l'acquisition de l'œuvre, sous la forme d'une chaîne de caractères « (*annee*,*type*) » (Exemple : (1854,Don)), clé étrangère vers la table Dim_Acquisition
- id_domaine : identificateur du domaine artistique dont l'œuvre fait parti (seuls 4 domaines différents apparaissent donc l'identificateur est un nombre entre 1 et 4), clé étrangère vers la table Dim_Domaine
- id_dimensions : identificateur des dimensions de l'œuvre sous la forme d'un texte tel que

- « *a* x *b* » où *a* représente la longueur en cm et *b* la largeur en cm si l'oeuvre est en deux dimensions, sinon c'est un texte sous la forme « *a* x *b* x *c* » où *c* représente la profondeur en cm (Exemple : « 24,7 x 35 x 10 »), clé étrangère vers la table Dim_Dimensions
- date_creation : texte exprimant approximativement la date de création de l'œuvre
- titre : texte, le titre de l'œuvre
- technique : texte, description éventuelle de l'œuvre et de la technique utilisée
- no_inventaire : texte (car c'est plus un code qu'un numéro, il y a parfois des lettres),le numéro d'inventaire de l'œuvre dans le musée
- lien_navigart : texte, le lien vers l'œuvre sur le site navigart
- precisions_acquisition : texte, description des conditions d'acquisition de l'œuvre

Dimensions

Ensuite, il ne restait plus qu'à créer les tables des dimensions, en prenant comme clé primaire la valeur clé étrangère correspondante dans la table des faits. Ainsi, voici les colonnes des différentes tables.

Dim_Auteur

- id : texte, le nom complet de l'auteur comme dans la table des faits, clé primaire
- prenom : texte, prénom de l'auteur
- nom : texte, nom de famille de l'auteur
- annee_naissance : nombre, année de naissance de l'auteur
- annee_deces : nombre, année de décès de l'auteur
- lieu_naissance : texte, ville de naissance de l'auteur
- lieu_deces: texte, ville de décès de l'auteur

Dim_Acquisition

- id : texte tel qu'expliqué dans id_acquisition de la table des faits, clé primaire
- annee : nombre, année d'acquisition
- type : texte, type d'acquisition, l'une des valeurs suivantes : Inconnu, Don, Achat, Dépôt ou Legs

Dim_Domaine

- id : nombre entre 1 et 4 comme expliqué pour id_domaine de la table des faits
- domaine : texte, le domaine artistique correspondant : Sculpture, Peinture, Dessin ou Estampe

Dim_Dimensions

- id : texte formaté comme précisé pour id_dimensions de la table des faits
- x : nombre, dimension x en cm
- y : nombre, dimension y en cm
- z : nombre, dimension z en cm si l'œuvre est en 3D

Une fois ces tables créées, nous avons décidé de créer un index Bitmap afin d'accélérer l'accès aux valeurs d'id_domaine de la table des faits car id_domaine ne prend que 4 valeurs différentes (chiffre de 1 à 4).

Ensuite, l'intégration des valeurs depuis le CSV s'est effectuée très facilement car Oracle Developer permet l'import de données dans une table depuis un fichier CSV directement via l'interface. Pour cela, il a suffi de dire à Oracle Developer quelle colonne du CSV correspond à quelle colonne de la

table des faits et il importe les données dans la table des faits. Pour ce qui est des dimensions, c'est exactement la même procédure mais la clé primaire permet d'empêcher les duplicats et ainsi importer un CSV de 5500 lignes ne fait que 4 entrées dans la table de dimension Domaine par exemple. En somme, l'intégration des données s'est effectuée très aisément, il a suffi de créer les tables tels qu'on les avait auparavant modélisées et d'importer les données du CSV dans chaque table via l'outil d'Oracle Developer.

Requêtes intéressantes

Afin d'analyser les données de l'entrepôt, nous avons rédigé un total de 8 différentes requêtes SQL permettant d'analyser l'entrepôt sous différents angles.

Requête 1

```
  Select f.titre, f.id_auteur AS Auteur, dim.x * dim.y AS Aire, rank()
over (order by dim.x * dim.y desc) AS Rang
  from faits_oeuvre f, dim_dimensions dim
  where f.id_dimensions = dim.id AND dim.z IS NULL;
```

Cette requête permet d'afficher les différentes œuvres (titre et auteur) en 2 dimensions (dont les dimensions ont un Z nul) classées de l'aire la plus grande à la plus petite.

Requête 2

```
  Select * from
  (Select f.titre, f.id_auteur AS Auteur, dim.x * dim.y * dim.z AS
Volume, rank() over (order by dim.x * dim.y * dim.z desc) AS Rang
  from faits_oeuvre f, dim_dimensions dim
  where f.id_dimensions = dim.id AND dim.z IS NOT NULL)
  where Rownum <= 10;
```

Cette requête permet d'afficher le top 10 des différentes œuvres (titre et auteur) en 3 dimensions (dimension Z non nul) classées de la plus volumineuse à la moins volumineuse.

Requête 3

```
  Select acq.annee AS Annee_Acquisition,acq.type AS Type_Acquisition,
count(f.id_oeuvre) AS Oeuvres
  from faits_oeuvre f, dim_acquisition acq
  where f.id_acquisition = acq.id AND acq.annee != 0
  group by ROLLUP (acq.annee,acq.type)
  order by acq.annee;
```

Cette requête affiche le nombre d'œuvres (dont l'année d'acquisition est connue, donc différente de 0) et pour chaque année d'acquisition affiche combien d'œuvre ont été acquise pour chaque type d'acquisition (achat, don, leg…) cette année-ci.

Requête 4

```
  Select acq.annee AS Annee_Acquisition,dom.domaine AS Domaine,
count(f.id_oeuvre) AS Oeuvres
  from faits_oeuvre f, dim_acquisition acq, dim_domaine dom
  where f.id_acquisition = acq.id AND f.id_domaine = dom.id
  group by CUBE (acq.annee,dom.domaine)
  order by acq.annee;
```

Cette requête affiche le nombre d'œuvres acquises par année d'acquisition et par domaine

artistique.

Requête 5

```
  Select f.id_auteur AS Auteur,dom.domaine AS Domaine,
count(f.id_oeuvre) AS Oeuvres, GROUPING_ID(f.id_auteur,dom.domaine) AS
GroupingID
  from faits_oeuvre f, dim_domaine dom
  where f.id_domaine = dom.id
  group by GROUPING SETS ((f.id_auteur,dom.domaine),(f.id_auteur),());
```

Cette requête affiche pour chaque auteur le nombre d'oeuvres qu'il a réalisé pour chaque domaine artistique.

Elle affiche également un Grouping_ID permettant de savoir quelle lignes du résultat sont des totaux.

Requête 6

```
  Select acq.type AS Type_Acquisition,dom.domaine AS Domaine,
count(f.id_oeuvre) AS Oeuvres, GROUPING(acq.type) AS Acqu,
GROUPING(dom.domaine) AS Dom
  from faits_oeuvre f, dim_acquisition acq, dim_domaine dom
  where f.id_acquisition = acq.id AND f.id_domaine = dom.id
  group by CUBE (acq.type,dom.domaine)
  order by acq.type,dom.domaine;
```

Cette requête affiche le nombre d'oeuvres acquise par type d'acquisition et par domaine artistique.

Elle affiche également des grouping afin de définir quelles lignes de résultat sont des totaux pour chaque dimension du cube.

Requête 7

```
  Select au.lieu_naissance AS Lieu_Naissance_Auteur, count(f.id_oeuvre)
AS Oeuvres, rank() over (order by count(f.id_oeuvre) desc) AS Rang
  from faits_oeuvre f, dim_auteur au
  where f.id_auteur = au.id
  group by au.lieu_naissance;
```

Cette requête classe les villes de naissance d'artistes selon combien il y a d'œuvres d'artistes nés dans cette ville dans le musée. Elle affiche la ville, le nombre d'œuvres correspondant et le rang de la ville.

Requête 8

```
  Select f.id_auteur AS Auteur, acq.type AS Type_Acquisition,
count(f.id_oeuvre) AS Oeuvres
  from faits_oeuvre f, dim_acquisition acq
  where f.id_acquisition = acq.id
  group by ROLLUP (f.id_auteur,acq.type)
  order by f.id_auteur, acq.type;
```

Cette requête affiche pour chaque artiste le nombre d'œuvres qu'il a réalisé en fonction de leur type d'acquisition.

Ces 8 requêtes permettent ainsi d'analyser les données de l'entrepôt sous différents angles de vue,

que ce soit pour des analyses plus axés sur l'acquisition de l'œuvre par le musée, ou bien faire des statistique sur les artistes par exemple.

Améliorations éventuelles

Notre base de données permet une abstraction pertinente de nos données, apte à l'analyse et à la production de statistiques. Cela dit il y a encore beaucoup de choses qui pourraient être accomplies pour porter ce projet plus loin. De plus certaines données importantes et facilement exploitables ne sont pas incluses dans le dataset.

La base de donnée mériterait d'être liée à d'autres datasets afin de permettre une analyse plus large et de mettre en valeur des relations plus subtiles. En effet on pourrait lier les informations relatives aux auteur avec d'autres données sur l'inventaire des bibliothèques pour voir si il y a des auteurs en commun, pour trouver des livres sur la démarche des artistes. On pourrait observer les dimensions pour faire des comparaisons amusantes avec d'autres objets dimensionnés. On pourrait chercher des relations entre dates et événements historiques pour mettre en valeur des variations dans les tendances artistiques.

Il y a encore plein de possibilité, quand il s'agit de l'évolution d'un dataset, et on peut encore imaginer bien d'autres manière de lier nos données.