

M1 IASD EDBD
Rendu 1 Analyse Traitements et Schéma
Entrepôt
Rapport du projet

Tristan Tribes Manon Ramaroson Dorine Fontaine

Novembre 2021

Question 1

La vie d'un étudiant est bien remplie. Comment pouvons-nous allier à la fois excellence académique et vie sociale remplie ? Pour répondre à cette question, le groupe TMD a développé une application répondant à la fois aux étudiants et aux Bureaux des Étudiants de France. Cette application recense les fêtes et les notes de chaque étudiant afin de les aider à ne plus se soucier de jongler entre les cours et les soirées car nous leur proposons de le faire pour eux.

A l'air du numérique, les informations sont répandues de toute part. Les informations concernant les événements et les concerts deviennent beaucoup plus accessibles qu'avant. Une journée peut être bien vite remplie en quelques clics. Les étudiants sont aujourd'hui beaucoup plus susceptibles de sortir qu'avant. Ainsi l'application dédiée propose un large choix de soirées qui non seulement correspondent au profil de l'étudiant mais également ne se répercute pas sur ses notes. L'étudiant rentre les fêtes auxquelles il a participé, ainsi que ses notes scolaires et l'application se chargera automatiquement d'agréger les données. En plus de ça, il pourra se construire un profil pour lui indiquer les fêtes qui pourraient l'intéresser, et à minima, lui donner un indice déterminant si la participation à une fête serait bénéfique ou non sur ses résultats.

C'est également un excellent outil de mesure permettant d'analyser les facteurs de succès d'une fête. Elle pourra servir aux organisateurs de soirées (tels que les BDE) pour prospérer. Par exemple : planifier les fêtes ayant le meilleur rendement, ou le plus de monde présent en évaluant le nombre de participants à chaque soirée et les profils plus susceptibles d'être intéressés par telle ou telle soirée.

C'est une association qui veut créer une application afin de référencer des sorties plus adaptées à l'emploi du temps et au rythme de vie d'un étudiant. Le sens du mot adapter concerne les notes scolaires de l'étudiant, notre but est donc de maximiser les notes tout en continuant à vivre pleinement sa vie étudiante. Cette application est participative, personne n'est rémunéré. Bien sûr les dons sont acceptés afin de mieux développer l'application.

Pour avoir accès à nos services il faudra se connecter (et donc se créer un compte), un emploi du temps adapté sera créé après avoir rempli quelques champs (notes et soirées déjà passées, type de soirée préféré, heures de sortie). Ces fonctionnalités seront disponibles pour un élève ou bien une promo, concernant la promotion d'élèves les moyennes des notes seront alors prises en compte. Le service principal est donc des soirées recommandées à l'étudiant en tenant en compte de ses disponibilités, préférences et notes scolaires. Tenir compte des notes signifient de réserver plus de temps donc moins de soirées si un examen se prépare ou si l'élève a des difficultés dans une matière.

Les dons des utilisateurs sont acceptés afin de payer les éléments techniques tels que le domaine du site web, des moyens de publications (flyers,...), des entrepôt de données...

Des informations utiles et obligatoires pour faire fonctionner correctement le service principal tels que :

- horaires de disponibilités de l'étudiant
- notes
- estimation des cours s'il n'y a pas de note (coefficient de la note ou estimation de la difficulté du cours)
- heures de sortie (ex : j'aime sortir après manger donc après 20:00)
- type de sortie (journée/soir, à l'extérieur/intérieur...)

On se limitera donc à associer les fêtes des 14 derniers jours aux notes des étudiants. On estimera dans notre modèle qu'une fête plus ancienne que deux semaines n'a pas d'impact sur les notes de l'étudiant.

Question 2

Indiquez les actions / opérations (e.g., ventes, livraisons) à tracer pour récupérer ces informations.

Action/Opérations :

1. Notre but est d'augmenter le rendement du BDE. On veut donc connaître toutes les soirées auxquelles les étudiants participent. On voudra aussi connaître certaines informations sur ces soirées : si l'événement est organisé par un BDE, le prix du ticket (même si c'est gratuit), le type de sortie, ainsi que la date. Enfin on voudra connaître le lieu de vente du ticket s'il existe, par exemple en ligne ou devant un établissement.
2. On veut aussi stocker et mettre en lien les contrôles et les sorties. Pour cela on va donc conserver d'abord l'étudiant, les contrôles auxquels il doit participer, la matière et l'horaire, ainsi que les sorties liées. Pour ces sorties on voudra connaître l'heure de début et de fin pour l'étudiant (à ne pas confondre avec l'heure officielle de début et de fin de la sortie). On notera aussi l'heure de début du contrôle. On aura aussi des informations sur les contrôles, le coefficient du contrôle et de la matière (selon les parcours), ainsi que le type pour pouvoir filtrer entre devoirs maison ou contrôle terminal. Enfin on gardera la note du contrôle pour quantifier le succès et les verres d'alcools bus durant la sortie.

3. On va stocker tous les cours auxquels un étudiant est censé participer. Pour chacun de ces cours on notera si l'étudiant y a participé, ainsi que les horaires du cours (heure, jour). On pourra ainsi avoir des informations sur la participation à des cours selon le type de cours, le profil de l'étudiant ou encore les horaires pour ce même cours.

Question 3

Pour chaque action / opération, proposez au moins trois traitements possibles (i.e., requêtes analytiques) permettant d'aider à la prise de décision sur le sujet.

1. Vente d'un ticket du BDE pour une soirée :
 - (a) Je veux connaître le nombre de ventes par type de soirées et par filières.
 - (b) Je souhaite connaître les ventes par rapport à la période et au coût total de la fête.
 - (c) Je veux obtenir le pourcentage de participation par rapport au prix des soirées.
2. Note par rapport aux sorties :
 - (a) Je veux connaître la moyenne des notes par rapport à une soirée selon l'alcoolémie.
 - (b) Je veux comparer l'heure de fin d'une sortie sur les notes liées.
 - (c) Je veux connaître la moyenne des notes sur des contrôles par rapport aux nombre de sorties
3. Présence en cours :
 - (a) Je veux comparer la présence dans un cours des étudiants par rapport à la distance du domicile
 - (b) Je veux comparer la présence dans les cours selon l'existence d'un contrôle dans la semaine à venir
 - (c) Je veux comparer la présence dans un cours par rapport à la durée du cours totale.

Question 4

Ordonnez les actions par ordre d'importance / rentabilité potentielle (e.g., augmentation des ventes vs. utilisation optimale de l'espace de stockage dans le magasin).

Tout d'abord un rappel de l'objectif de l'application est de maximiser les notes par rapport à la vie personnelle. Pouvoir vivre pleinement sa vie étudiante tout en gardant une assiduité au niveau des notes scolaires.

L'action la plus importante est donc naturellement celle des notes par rapport aux sorties. Les deux choix restants sont assez durs à classer, mais nous trouvons que la vente de billets seraient plus intéressantes.

Question 5

Identifiez les deux actions / opérations les plus importantes à analyser.

Les deux actions les plus importantes à analyser sont la corrélation note/sorties et le rendement des soirées.

Question 6&7

Pour chaque actions / opérations, concevez un data-mart indépendant (c'est à dire, un modèle en étoile)

Pour le premier Datamart, on veut les notes par rapport aux sorties. On va donc avoir une table de fait de la forme (e, s, hds, hfs, hdc, d, m, a, c, n). e pour étudiant, s pour la sortie associée, hds et hfs pour respectivement l'heure de début et de fin de la sortie, hdc pour l'heure de début du contrôle, d pour la date du contrôle, m pour la matière du contrôle, a pour l'alcool bu dans la soirée, c pour les informations sur le contrôle et enfin la mesure n qui sera la note.

Cette mesure sera semi-additive, il existe une fonction avec laquelle on peut associer un nouveau résultat issu des anciennes données. On pourra faire une moyenne des notes en faisant bien attention aux tuples dupliqués à cause des soirées. On fera une projection en omettant toutes les informations liés aux sorties (s, hds, hfs, a).

En ce qui concerne les dimensions :

Etudiants : On va stocker sa filière, son niveau d'étude (Bac+ x), son année de première inscription dans le supérieur, son établissement actuel, son âge, son genre, son état matrimonial, la distance de son domicile par rapport à son établissement ainsi que son budget moyen pour les activités "extrascolaires".

Matière : La dimension qui concerne une matière, une discipline ou une UE. On notera son titre, le nombre d'heures de Travaux Dirigés, Travaux Pratiques, ainsi que le nombre d'heures de cours magistraux. On stockera enfin le niveau de la matière (Bac+x).

Date : La dimension date est assez classique, on ajoutera cependant des informations pertinentes dans la vie étudiante, c'est-à-dire des indicateurs de week-end, de vacances et de jours fériés. On notera aussi l'ajout de période de l'année (par exemple Noël, l'automne ou la Rentrée...).

Heure : La table heure est elle aussi assez classique. On stockera en supplément la période de la journée, en particulier pour avoir les différences entre après-midi, soirée et nuit.

Sortie : La table Sortie doit nous renseigner au maximum sur une sortie effectuée par un étudiant. Cette sortie ne se limite pas uniquement aux sorties en bars ou en club, on peut y mettre les concerts, les dîners entre amis, des sorties en parcs d'attractions... On stockera donc, le nombre de personnes présentes, si elle est payante (ou pas), l'établissement de la soirée, si la soirée était publique ou privée, une petite description, le public ciblé (étudiants, ou une filière particulière par exemple), le coût de la soirée (0 si gratuit), le nombre d'invités maximum et la présence ou non d'alcool.

Alcool : La dimension Alcool est une mini-dimension qui renseigne uniquement sur la consommation d'alcool à une soirée. Ces informations seront données soit en verres (mesure normalisées, 1 verre = 0.2 g/l dans le sang) soit en grammes d'alcool ingéré.

Contrôle : La dimension contrôle donne des informations basiques sur un contrôle. Le type du contrôle, par exemple partiel ou devoir maison. Le coefficient du contrôle en particulier et de sa matière (qui seront indirectement liés à l'étudiant), ainsi que la moyenne historique de la matière pour ce type d'étudiant, c'est-à-dire la moyenne des moyennes de la matière depuis sa création.

Enfin la mesure note sera simplement la note, mise sur 20 que l'étudiant aura obtenu au contrôle. On pourra faire une moyenne sur cette note via une fonction, ou filtrer selon les notes (plus de 15 par exemple).

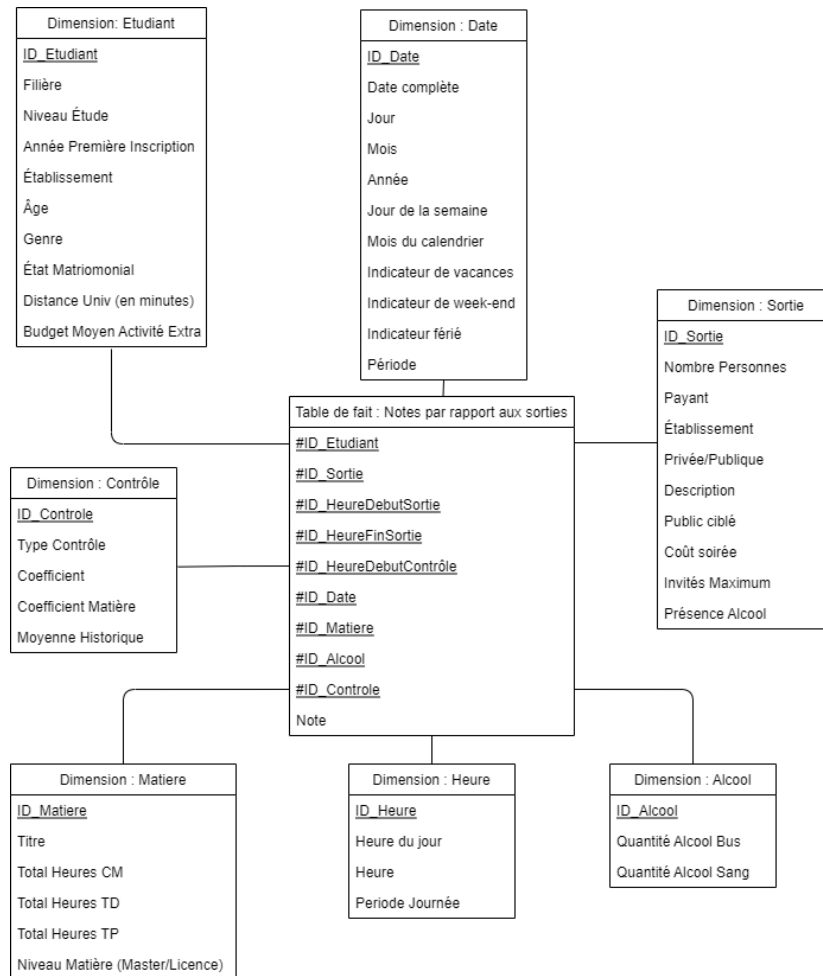


Figure 1: Data-mart pour les notes par rapport aux sorties

Pour la table de fait Vente de tickets qui sera une table snapshot, on a le schéma suivant (s, d, h, b, lvts, p, nb) où s correspond à la sortie, d à la date de la vente, b au BDE qui a vendu le ticket (avec une valeur spéciale si ce n'est pas un BDE qui a vendu le ticket), lvts une junk dimension qui regroupe le type de sortie et le lieu de vente. Pour les mesures on aura p pour le prix de vente total des tickets et nb pour le nombre de tickets vendus. Les deux mesures sont additives.

Cette table doit nous renseigner sur les tendances de ventes d'un ticket afin de pouvoir améliorer le profit des BDE. On a donc 5 dimensions dont trois en commun avec le datamart précédent (Heure, Date et Sortie).

Pour les deux nouvelles dimensions on a :

BDE : Regroupe toutes les informations concernant un BDE. Notamment, son nom, son budget annuel en euros, la nécessité de s'inscrire pour y participer ou non, son statut, privé ou public (n'importe qui peut y rentrer ou non), le type du BDE (c'est à dire le corps étudiant visé : médecine, sciences, école d'ingénieur...), la portée du BDE (seulement pour une filière, un établissement complet, ...), le nombre d'inscrits ou de personnes se revendiquant de ce BDE, le prix pour adhérer (0 ou +), ainsi que l'existence d'une carte donnant accès à des promotions pour les établissements affiliés.

Lieu Vente / Type Sortie : Cette junk dimension, regroupant à la fois le lieu de vente d'un ticket et le type de la sortie doit permettre d'accélérer les requêtes d'une part, d'autre part de donner des informations sur les meilleurs endroits de ventes/les soirées les plus populaires par type. On aura donc trois attributs : le lieu de vente (internet, devant l'établissement, au BDE...) puis le type de soirée (Halloween, déguisée, ou aucun thème en particulier) puis le type de l'événement (Concert, Dîner, Barathon...).

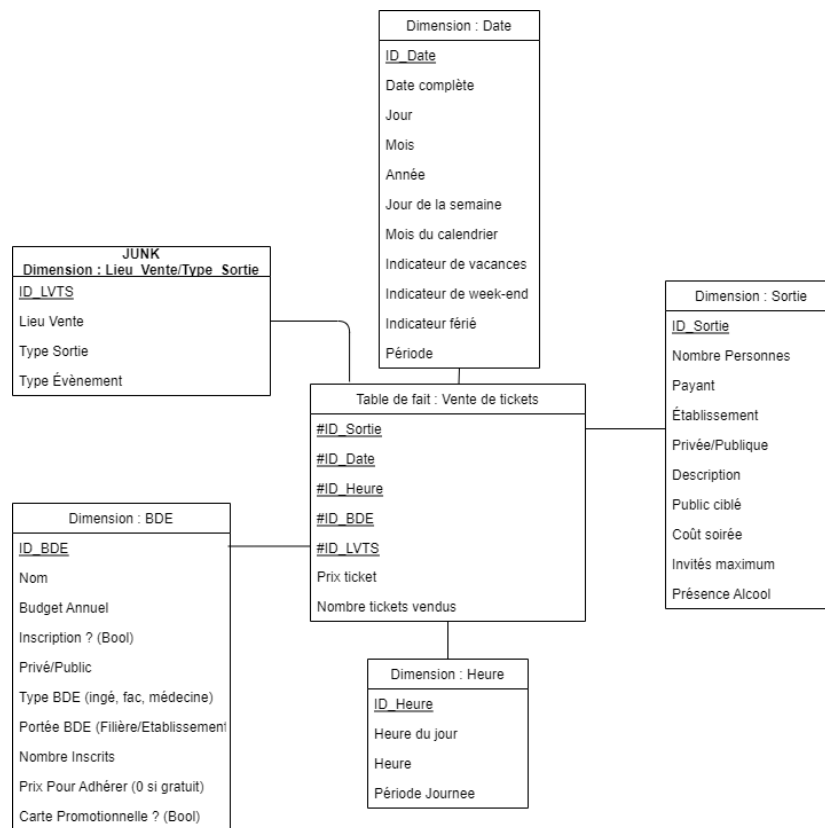


Figure 2: Data-mart pour vente ticket

Question 8

Question : Est-il possible de répondre aux traitements que vous avez indiqué avec le modèle que vous avez mis en place ? Expliquez pourquoi et comment.

Pour le datamart sur les notes et sorties :

Prenons par exemple la première requête de la question 3. "Je veux connaître la moyenne des notes par rapport à une soirée selon l'alcoolémie."

On partitionne les étudiants présents dans la soirée selon leur alcoolémie, par exemple on peut les diviser en 3 groupes (0-0.5g/L, 0.5g/L-1.5g/L et 1.5g/L-). On prend maintenant les notes des étudiants pour chaque catégorie afin de faire une moyenne des notes par catégorie.

Pour le datamart sur la vente de tickets :

Nous allons étudier la deuxième requête que nous avons proposé. "Je souhaite connaître les ventes par rapport à la période et au coût total de la fête."

Les périodes sont déjà fixées grâce au paramètre dans la dimension Date.

Pour chaque période nous faisons donc la moyenne du nombre de tickets vendus en regardant le coût du billet et le coût de la fête. Après plusieurs opérations nous pouvons comparer les résultats obtenus pour les différentes périodes et voir lequel est le plus important.

Question 9

1. Datamart notes par rapport aux sorties :

Instanciation d' Etudiant :

Etudiant	ID_Etudiant	Filiere	NiveauEtude	AnneePremiereInscription	Etablissement
1	51806744	Cinema	Master	2021	Paul Valery
2	21098771	Informatique	Master	2018	Universite de Montpellier
3	00044432	Lettres Modernes	Licence	2021	Paul Valery

Etudiant	...	Age	Genre	Etat_Matrimonial	Distance_Univ (en m)	Budget_Moyen_Activite_Extra
1	...	20	Femme	Celibataire	2000	100
2	...	22	Homme	Celibataire	500	50
3	...	18	Autre	Celibataire	10000	100

Instanciation de la table Sortie :

ID_Sortie	NbPersonnes	Payant	InvitéMax	Alcool	Etablissement	Privée/Publicue	Desc	Public ciblé	Coût (€)
135	1500	Non	500	Oui	HeatClub	Privé	...	Médecine	15
1823	800	Oui	200	Non	GameTaverne	Privé	...	M1	6
1023	300	Non	90	Oui	Titi Twister	Privé	...	L3	4.50

Instanciation de la table Controle :

ID_Control	TypeControl	Coefficient	Coefficient_Matiere	Moyenne_Historique
135	CC	5	5	12
1823	CT	2.5	2	15
1023	DM	2	1	11

Instanciation de la table Heure :

ID_Heure	Heure du jour	Heure	PeriodeJournée
123	12:00	12	Midi
124	13:15	13	Après-midi
125	15:45	15	Après-midi

Instanciation de la table Date :

ID_Date	DateComplète	Jour	Mois	Année	JourSem	MoisCal	Vacance	Weekend	Férié	Période
1	01-01-2021	1	1	2021	Vendredi	Janvier	True	False	True	Nouvel An
2	24-12-2000	24	12	2000	Dimanche	Décembre	True	True	False	Noël
2	18-06-2000	18	6	2000	Dimanche	Juin	False	False	False	Été

Instanciation de la table Matières :

ID_Matiere	Titre	TotalheuresTD	TotalHeuresCM	TotalHeuresTP	NiveauMat
1	"Introduction Bases de Données"	30	60	10	Licence
2	"Mathématiques compliquées"	90	0	12	Master
3	"Musique classique"	15	60	20	Licence

Instanciation de la table BDE :

ID_BDE	Nom	BudgetAnnuel	Inscription	Privé/Public
1	"Lyons"	24400	True	"Privé"
2	"La Glacière"	4000	False	"Public"
3	"Les Ingélos"	12500	True	"Public"

ID_BDE	...	TypeBDE	PortéeBDE	NbInscr	PrixAdherer	CartePromo
1	...	"Faculté"	"Sciences"	230	20	True
2	...	"IUT"	"Gestion"	40	0	False
3	...	"Ingénieur"	"Lettre"	60	5	True

Instanciation de la table Alcool :

ID_Alcool	QuantitéAlcoolBus	Quantité Alcool Sang (en g/l)
1	1	0.2
2	5	1
3	3.5	0.70

Instanciation de la JUNK table Lieux/Ventes/Types Sortie:

ID_LVTS	LieuVente	TypeSortie	TypeEvenement
1	"Stand"	"Soirée Halloween"	"Boite"
2	"En ligne"	"Soirée Film"	"Cinéma"
3	"Bureau BDE"	"AfterWork"	"Barathon"

Fait de Notes par rapport aux sorties :

ID_Etudiant	ID_Sortie	ID_HeureDebut	ID_HeureFin	ID_Date
1	1	1	2	8
2	4	8	3	7
4	1	5	7	4

ID_Matiere	ID_Controle	ID_Alcool	ID_HeureDebutControle
17	5	14	1
2	2	17	18
2	87	9	7

Fait de ventes de tickets :

ID_Sortie	ID_Date	ID_BDE	ID_Heure	LVTS	PrixTicket	NbTicketsVendus
1	4	4	12	9	1	105
2	3	6	1	13	10	50
3	1	6	4	13	32	88

Question 10

Table de faits "Notes par rapport aux sorties" :

Il y a 2,73 millions étudiants inscrits dans l'enseignement supérieur en France.
On sait que 60% des étudiants sortent.
Donnons une moyenne de 2 soirées par semaine.

On donne également 3 contrôles par matière et 6 matières par semestre.
Chaque contrôle a une portée de 2 semaines, c'est-à-dire que l'on associera à chaque note les sorties des 14 derniers jours. Donc chaque contrôle aura en moyenne 4 sorties associées.

On aura donc deux semestres, avec six matières par semestres et trois contrôles par matière. Chaque contrôle sera lié en moyenne à quatre sorties et on a 60% d'étudiants qui sortent.

$$6 \text{ matières} * 3 \text{ contrôles} * 2 \text{ semestres} * 4 \text{ sorties_par_controle} * \text{nb_étudiants_sortant}$$

$$\text{Pour une année : } 6*3*2*4*(2\,730\,000*0.6) = 235\,872\,000$$

Au final on aura environ 235 millions de lignes par an dans notre table de fait.

Table de faits "Vente de tickets" :

On estime qu'un BDE organise au maximum une soirée payante par semaine, sauf en période de vacances d'été. On aura donc 52 - 9 semaines, donc 43 fêtes payantes par an et par BDE. En 2010 il y avait environ 13 500 BDE.

On aura donc le calcul suivant :
nombre fêtes par an * nombre de BDE = 43 * 13 500 = 580 500 tuples par an.

Les dimensions :

Etudiants : Il y a 2.73 millions d'étudiants en France, on peut supposer que ce nombre augmente chaque année. On a donc environ 3 millions de lignes par an.

Date : Il y a 366 jours par an, l'indicateur de vacances qui peut varier selon le jour (multiplication par deux) et on supposera au minimum quatre périodes pour les saisons, et d'autres périodes comme les fêtes, les vacances etc... On supposera donc une vingtaine de périodes. On aura donc $366*2*20 = 14640$ lignes par an.

Heure : Cette dimension ne s'intéresse qu'à l'heure et pas aux minutes ou aux secondes. On aura donc 24 heures et les périodes correspondent à une heure, donc 24 tuples par an, qui ne changeront jamais d'une année à l'autre.

Alcool : On donnera 10 verres d'alcool au maximum (après on ne compte plus sur l'étudiant pour se souvenir exactement des données). On aura donc 10 lignes par an, qui ne changeront pas d'une année à l'autre. On peut aussi augmenter la finesse avec des incréments de 0.1g dans le sang, jusqu'à 4g/L. On aura dans ce cas une quarantaine de lignes.

Matière : Il est difficile de donner une estimation. Il y aurait environ 3500 établissements d'enseignement supérieur publics ou privés en France. On suppose pour chacun possède une cinquantaine de formations, pour chaque formation il y a 12 matières sur une année. On a donc au final $3500 \times 50 \times 12 = 2\,100\,000$ lignes avec une fourchette très haute.

Contrôles : Avec environ 2 000 000 de matières, on estime dans chaque matière 4 contrôles (de tous types) et disons jusqu'à 4 coefficients différents (nombre de formations différentes qui participe à une matière).

On aura donc $2\,000\,000 \times 4 \times 4 = 32\,000\,000$ de contrôles par an.

La taille est assez élevée pour une table de dimension mais on peut espérer qu'en réalité il y ait moins de contrôles, étant donné que l'on dépend de la table matière qui est elle aussi incertaine.

Sorties : Il est encore une fois difficile de trouver des données sur les événements en France, d'autant plus que le recensement de sorties pourrait faire l'objet d'un datamart. On pourrait donc supposer que chaque jour il y a une vingtaine de fêtes par grande ville étudiante (moyenne de villes à faible et haute densité) et il y a une trentaine de villes étudiantes en France. A cela il faudrait aussi ajouter les sorties privées, on en donnera une par semaine et par étudiant qui sort (60%).

Finalement on obtient :

$$52 \times 20 \times 30 + 2\,730\,000 \times 0.67 = 31\,200 + 234\,000 = 265\,200 \text{ sorties par an.}$$

BDE : Il y a 13 500 BDE en France donc 13 500 lignes par an au maximum.

LVTS : Il y a 5 lieux de ventes principaux (Internet, Bureau BDE, Etablissement scolaire, Etablissement soirée, Bouche à oreille) et on en trouvera certainement d'autres, disons 10. On va donner 50 types d'événements et 30 types de sorties.

On fait un produit cartésien pour le tout et on aura : $10 \times 50 \times 30 = 15\,000$ par an, le résultat ne change pas vraiment entre chaque année.