TD ENTREPOT DONNEES ET BIG DATA

QUESTION 1

**Schéma d'Entrepôt de Données pour Flickr - TD1**

*Introduction :*

L'objectif principal de notre entrepôt de données est de permettre une analyse approfondie de l'utilisation des appareils photo à partir des photos publiées sur la plateforme Flickr. Nous visons à étudier les lieux, les périodes de l'année, les horaires de la journée, le lignage des appareils photo, et les créateurs de chaque modèle.

*Tables de Dimensions :*

1. **Utilisateur :**
   * IDUTIL (Clé primaire)
   * NOM
   * PRENOM
2. **Appareil :**
   * IDAPP (Clé primaire)
   * MARQUE
3. **Configuration :**
   * IDCONF (Clé primaire)
   * OUVERTURE\_FOCALE
   * TEMPS\_EXPO
   * FLASH
4. **Photographie :**
   * CODE (Clé primaire)
   * LIEU\_REALISATION
   * DATE\_REALISATION
   * LICENCE
   * *IDAPPAREIL* (Clé étrangère vers la dimension "Appareil")
   * *IDCONF* (Clé étrangère vers la dimension "Configuration")
   * *IDUTIL*\_*PUB* (Clé étrangère vers la dimension "Utilisateur")
   * *IDALB* (Clé étrangère vers la dimension "Album")
   * *IDGAL* (Clé étrangère vers la dimension "Galerie")
   * *IDDIS* (Clé étrangère vers la dimension "Discussion")
5. **Temps :**
   * DATE (Clé primaire)
   * Annee
   * Mois
   * Jour
6. **Lieu :**
   * LIEU (Clé primaire)
7. **Lignage :**
   * MODELE (Clé primaire)
   * *LIGNE\_PARENTE* (Clé étrangère vers la dimension "Lignage")
8. **Createur :**
   * CREATEUR (Clé primaire)
9. **Balise :**
   * TYPE\_BALISE (Clé primaire)
10. **Commentaire :**

* IDCOMMENTAIRE (Clé primaire)
* *ID*\_*UTILISATION* (Clé étrangère vers la dimension "Utilisation")
* TEXTECOMMENTAIRE
* *CODEPHOTOCOM* (Clé étrangère vers la dimension "Photographie")
* *IDUTIL*\_*COM* (Clé étrangère vers la dimension "Utilisateur")

1. **Utilisation :**

* IDDIS (Clé primaire)
* *CODEPHOTOPUB* (Clé étrangère vers la dimension "Photographie")
* *IDUTIL*\_*VIS* (Clé étrangère vers la dimension "Utilisateur")
* *IDALB* (Clé étrangère vers la dimension "Album")
* *IDGAL* (Clé étrangère vers la dimension "Galerie")
* *ID*\_*UTIL*\_*LIKE* (Clé étrangère vers la dimension "Utilisateur")
* *IDUTIL\_ABON\_1* (Clé étrangère vers la dimension "Utilisateur")
* *IDUTIL\_ABON\_2* (Clé étrangère vers la dimension "Utilisateur")

1. **Marquage :**

* BALISE\_MARQUER (Clé primaire)
* *CODEPHOTOMAR* (Clé étrangère vers la dimension "Photographie")
* *IDUTIL*\_*MAR* (Clé étrangère vers la dimension "Utilisateur")

*Table de Faits - Photographie :*

* CODE (Clé primaire)
* *IDUTIL\_PUB* (Clé étrangère vers la dimension "Utilisateur)
* *IDAPPAREIL* (Clé étrangère vers la dimension "Appareil")
* *IDCONF* (Clé étrangère vers la dimension "Configuration")
* *IDALB* (Clé étrangère vers la dimension "Album")
* *IDGAL* (Clé étrangère vers la dimension "Galerie")
* *IDDIS* (Clé étrangère vers la dimension "Discussion")

**Explications sur le Schéma d'Entrepôt de Données pour Flickr :**

1. **Objectif Principal :**
   * L'entrepôt de données vise à analyser en profondeur l'utilisation des appareils photo à partir des photos publiées sur Flickr.
2. **Table de Faits - Photographie :**
   * La table "Photographie" est la table de faits centrale, représentant chaque photo avec des clés étrangères vers des dimensions telles que "Utilisateur," "Appareil," "Configuration," etc.
3. **Dimensions Clés :**
   * Les tables de dimensions comme "Utilisateur," "Appareil," "Configuration," "Temps," "Lieu," "Lignage," et "Createur" fournissent des informations détaillées pour enrichir l'analyse.
4. **Interactions Utilisateur :**
   * La table "Utilisation" capture les interactions des utilisateurs avec les photos, y compris la publication, les likes, les abonnements, etc.
5. **Métadonnées et Interaction :**
   * Les tables "Commentaire" et "Marquage" ajoutent des métadonnées et des informations sur l'interaction des utilisateurs avec les photos.
6. **Hiérarchie des Appareils :**
   * La table "Lignage" permet de modéliser la hiérarchie des modèles d'appareils photo, important pour comprendre l'évolution des produits.
7. **Flexibilité pour l'Analyse :**

* La structure permet des analyses diverses telles que l'étude des lieux, des dates, des créateurs, et des interactions, offrant une vision complète de l’environnement Flickr.

2/

Voir le Fichier creation.sql

3/ Requête chez Nabil

A REECRIRE LEXO 2 !!

1/

Pour concevoir un entrepôt de données pour l'analyse des statistiques d'appel de Bouygues Télécom, vous pouvez commencer par élaborer un schéma logique et physique. Voici une proposition de schéma logique pour votre entrepôt de données, suivi d'une explication des choix faits :

**Schéma Logique de l'Entrepôt de Données**

1. **Table "Appels" :**
   * **ID\_Appel (Clé primaire)**
   * Emetteur
   * Destinataire
   * Date\_heure\_debut
   * Duree\_appel
   * Type\_appel
   * Compte\_facture
   * Code\_erreur
2. **Table "Clients" :**
   * **ID\_Client (Clé primaire)**
   * Nom
   * Prénom
   * Numéro\_de\_téléphone

**Explications et Justifications :**

1. **Table "Appels" :**
   * **ID\_Appel (Clé primaire) :** Un identifiant unique pour chaque appel, utilisé comme clé primaire.
   * **Emetteur et Destinataire :** Numéro ou identifiant de l'appelant et de l'appelé.
   * **Date\_heure\_debut :** Pour enregistrer la date et l'heure du début de chaque appel.
   * **Duree\_appel :** Pour stocker la durée de chaque appel.
   * **Type\_appel :** Pour spécifier le type d'appel (Mobile, SMS, Fax, Fixe, etc.).
   * **Compte\_facture :** Le compte facturé pour l'appel (peut être différent de l'émetteur).
   * **Code\_erreur :** Pour enregistrer tout dysfonctionnement ou erreur pendant l'appel.
2. **Table "Clients" :**
   * **ID\_Client (Clé primaire) :** Identifiant unique pour chaque client.
   * **Nom, Prénom :** Informations sur le client.
   * **Numéro\_de\_téléphone :** Pour établir la relation avec la table "Appels".

Le schéma physique de l'entrepôt de données pour Bouygues Télécom conserve la structure fondamentale du schéma logique, avec deux tables principales, "Appels" et "Clients". Cependant, une adaptation a été apportée pour améliorer les performances des requêtes temporelles. Concrètement, un index a été ajouté à la colonne "Date\_heure\_debut" de la table "Appels". Cet index vise à accélérer les opérations de filtrage et de tri basées sur la date et l'heure du début de l'appel, fournissant ainsi une optimisation spécifique pour les requêtes temporelles fréquentes. L'intégrité relationnelle entre les tables et les contraintes clés étrangères demeurent pour assurer la cohérence des données, tout en introduisant une amélioration ciblée des performances.

2/

La granularité des faits dans le contexte de l'entrepôt de données pour les statistiques d'appel de Bouygues Télécom est définie au niveau de chaque appel individuel. Chaque enregistrement dans la table "Appels" représente un appel spécifique, avec des détails tels que l'émetteur, le destinataire, la date et l'heure du début de l'appel, la durée, le type d'appel, le compte facturé, et le code d'erreur. Cette granularité fine permet une analyse détaillée des comportements d'appel et des motifs associés.

En ce qui concerne l'additivité des mesures, examinons quelques exemples spécifiques dans le contexte de Bouygues Télécom :

1. **Additivité :** La durée totale des appels peut être considérée comme additive. En agrégeant les durées individuelles des appels, il est possible d'obtenir la durée totale des appels sur une période donnée, offrant ainsi une vue consolidée du temps d'utilisation du réseau.
2. **Semi-additivité :** Le coût total des appels peut être considéré comme semi-additif. Si l'on agrège les coûts individuels des appels par client, cela a du sens. Cependant, l'agrégation des coûts pour tous les clients peut ne pas être pertinente, car certains coûts peuvent être partagés entre plusieurs clients, introduisant une nuance de semi-additivité.
3. **Non-additivité :** Le type d'appel peut être considéré comme non-additif. Aggréger les types d'appels différents n'aurait pas de signification concrète, car cela ne fournirait pas d'informations utiles sur la répartition des types d'appels.

Ainsi, la granularité fine des faits, à savoir chaque appel individuel, permet une analyse approfondie des comportements d'appel, tandis que l'additivité, la semi-additivité et la non-additivité des mesures dépendent de la nature spécifique de chaque mesure enregistrée, répondant ainsi aux besoins analytiques particuliers de Bouygues Télécom.

### 3/

### Requête 1 : Analyse des tendances d'utilisation par type d'appel

**Description :** Cette requête permet d'analyser comment l'utilisation des différents types d'appels évolue au fil du temps.

**SQL :**

sql

SELECT

Type\_appel,

YEAR(Date\_heure\_debut) AS Annee,

MONTH(Date\_heure\_debut) AS Mois,

COUNT(\*) AS Nombre\_appels,

SUM(Duree\_appel) AS Duree\_totale

FROM

Appels

GROUP BY

Type\_appel, Annee, Mois

ORDER BY

Annee, Mois, Type\_appel;

**Description :** Cette requête vise à identifier les clients qui ont effectué le plus grand nombre d'appels.

**SQL :**

sql

SELECT

Emetteur AS Numero\_de\_telephone,

COUNT(\*) AS Nombre\_appels,

SUM(Duree\_appel) AS Duree\_totale

FROM

Appels

GROUP BY

Emetteur

ORDER BY

Nombre\_appels DESC;

### Requête 3 : Analyse des erreurs les plus fréquentes

**Description :** Cette requête permet d'identifier les types d'erreurs les plus fréquemment rencontrés pendant les appels.

**SQL :**

sql

SELECT

Code\_erreur,

COUNT(\*) AS Nombre\_occurrences

FROM

Appels

WHERE

Code\_erreur IS NOT NULL

GROUP BY

Code\_erreur

ORDER BY

Nombre\_occurrences DESC;

**Question 2**

### 1/

### Schéma Logique :

1. **Table "Clients" :**
   * **ID\_Client (Clé primaire)**
   * **Nom**
   * **Prenom**
   * **Numero\_de\_compte (Unique)**
2. **Table "Factures" :**
   * **ID\_Facture (Clé primaire)**
   * **ID\_Client (Clé étrangère vers la table Clients)**
   * **Periode\_Reference (DATE)**
   * **Cout\_Abonnements (DECIMAL)**
   * **Cout\_Communications\_Incluses (DECIMAL)**
   * **Cout\_Communications\_Hors\_Forfait (DECIMAL)**

**Schéma Physique :**

sql

CREATE TABLE Clients (

ID\_Client INT PRIMARY KEY,

Nom VARCHAR(50),

Prenom VARCHAR(50),

Numero\_de\_compte VARCHAR(20) UNIQUE

);

CREATE TABLE Factures (

ID\_Facture INT PRIMARY KEY,

ID\_Client INT,

Periode\_Reference DATE,

Cout\_Abonnements DECIMAL(10,2),

Cout\_Communications\_Incluses DECIMAL(10,2),

Cout\_Communications\_Hors\_Forfait DECIMAL(10,2),

FOREIGN KEY (ID\_Client) REFERENCES Clients(ID\_Client)

);

2/

**Description de la Granularité des Faits :**

La granularité des faits dans le contexte de l'entrepôt de données pour les factures clients de la compagnie de télécommunications est définie par chaque enregistrement individuel de facturation. Chaque ligne dans la table des factures représente une unité indivisible d'information, incluant des détails tels que le compte client, la période de référence, les coûts mensuels des abonnements, les coûts des communications incluses dans le forfait, les coûts des communications hors forfait, et un identifiant unique pour chaque facture. Cette granularité fine permet une analyse détaillée des coûts associés à chaque client pour chaque période de facturation.

**Justification de l'Additivité :**

1. **Additivité du Coût des Abonnements :**
   * *Justification :* Les coûts des abonnements sont généralement des coûts fixes mensuels et peuvent être agrégés de manière significative. La somme des coûts des abonnements pour chaque facture donne le coût total des abonnements sur la période.
2. **Semi-Additivité des Coûts des Communications :**
   * *Justification :* Les coûts des communications incluses dans le forfait peuvent être agrégés au niveau de l'ensemble des clients, car ils représentent des coûts groupés. Cependant, les coûts des communications hors forfait ne peuvent être agrégés de la même manière, car ils dépendent des activités individuelles de chaque client.
3. **Non-Additivité de l'Identifiant Unique de la Facture :**
   * *Justification :* L'identifiant unique de chaque facture ne peut pas être agrégé, car il représente une caractéristique unique et non additive. Il identifie de manière distinctive chaque enregistrement dans la table des factures.

La semi-additivité des coûts des communications reflète le fait que certaines catégories de coûts peuvent être agrégées dans certains contextes (par exemple, les communications incluses dans le forfait par type de communication), tandis que d'autres catégories (par exemple, les coûts hors forfait) peuvent ne pas être agrégées de manière significative sans perdre de détails pertinents. La non-additivité de l'identifiant unique souligne le caractère distinct de chaque facture.

3/

### Requête 1 : Analyse des Coûts Moyens Mensuels par Client

**Description :** Cette requête permet d'analyser les coûts moyens mensuels, en agrégeant les coûts des abonnements, des communications incluses et des communications hors forfait pour chaque client.

SELECT

ID\_Client,

AVG(Cout\_Abonnements) AS Cout\_Moyen\_Abonnements,

AVG(Cout\_Communications\_Incluses) AS Cout\_Moyen\_Communications\_Incluses,

AVG(Cout\_Communications\_Hors\_Forfait) AS Cout\_Moyen\_Communications\_Hors\_Forfait

FROM

Factures

GROUP BY

ID\_Client;

### Requête 2 : Analyse des Coûts Mensuels par Type de Communication

**Description :** Cette requête permet d'analyser la répartition des coûts mensuels entre les différentes catégories de communication (abonnements, communications incluses, communications hors forfait).

SELECT

Periode\_Reference,

SUM(Cout\_Abonnements) AS Cout\_Total\_Abonnements,

SUM(Cout\_Communications\_Incluses) AS Cout\_Total\_Communications\_Incluses,

SUM(Cout\_Communications\_Hors\_Forfait) AS Cout\_Total\_Communications\_Hors\_Forfait

FROM

Factures

GROUP BY

Periode\_Reference;

### Requête 3 : Analyse des Clients les Plus Rentables

**Description :** Cette requête identifie les clients ayant le coût total le plus élevé, en agrégeant les coûts des abonnements, des communications incluses et des communications hors forfait.

SELECT

ID\_Client,

SUM(Cout\_Abonnements + Cout\_Communications\_Incluses + Cout\_Communications\_Hors\_Forfait) AS Cout\_Total

FROM

Factures

GROUP BY

ID\_Client

ORDER BY

Cout\_Total DESC

LIMIT 5;