Projet Entrepôt de données
Talend Open Studio

Fourniture

Vous recevez un jeu de données sous la forme d'une archive contenant des fichiers csv présentant des statistiques sur les Jeux Olympiques de Paris 2024. Elle contient des données sur les athlètes, les épreuves, les médailles …

L'archive contant les fichiers est : **TP_Donnees_JO_Paris_2024.zip**

Cette archive contient les fichiers suivants :

athletes.csv
coaches.csv
events.csv
medallists.csv
medals.csv
nocs.csv
schedules.csv
schedules_preliminary.csv
teams.csv
technical_officials.csv
torch_route.csv
venues.csv

Concept

Vous devez construire un modèle de stockage des données reçues dans une base de données. De plus, vous devez modéliser et construire le datamart permettant les analyses souhaitées.

Les technologies que vous utiliserez sont à votre discrétion, à l'exception de l'ETL : vous devez utiliser Talend Open Studio. La qualité des développements dans cet outil sera évaluée.

L'outil de restitution est libre (Power BI, Qlik, Tableau, autres…).

Modalités sur la soutenance et le rapport

Le rapport devra contenir vos propositions d'états de reporting, un diagramme de la chaîne de chargement et répondre aux questions posées. Vous pouvez compléter, en proposant de nouveaux rapports ayant un intérêt (ex : géo-visualisation, datavisualisation sur smartphone, …), des indicateurs supplémentaires… Plus globalement, les axes d'améliorations sont bienvenus.

Remise du rapport : Le **12/01/2025 à 16h00** dernier délai.
Une version électronique du rapport doit être envoyée à l'adresse benoit.delplace@dauphine.fr
Préfixer le sujet du mail par [RAPPORT_EDT Groupe n° XX].

Soutenance le15/01/2025, par groupe de 3, mais notation individuelle. Les soutenances se décomposent en trois parties :

- Présentation et démonstration : 10 minutes chacune
- Questions : 5 minutes

L'évaluation se décomposera de la façon suivante :

- Rapport et soutenance : chacun 40 % de la note
- Implication et maîtrise individuelle du sujet : 20 % de la note

Réalisations demandées

Vous devez ajouter la hiérarchie des sports suivante à votre modèle :

- Power Sports : Weightlifting, Boxing, Judo, Karate, Taekwondo, Wrestling,
- Endurance Sports : Cycling, Rowing, Triathlon,
- Speed Sports : Athletics, Swimming, Basketball, Handball, Hockey, Football, Rugby,
- Skill Sports : Gymnastics, Fencing, Golf, Shooting, Archery, Table Tennis, Badminton, Tennis, Baseball/Softball,
- Water Sports : Aquatics, Canoeing, Sailing, Surfing,
- Board Sports : Skateboarding, Surfing,
- Combination Sports : Modern Pentathlon,
- Team Sports : Basketball, Volleyball, Handball, Hockey, Football, Rugby, Baseball/Softball.

Vous devrez répondre aux questions suivantes :

- Que pensez-vous de la qualité des données ?
- Quels problèmes avez-vous rencontrées ?
- Fournissez le modèle de votre base de données.
    - S'agit-il d'un modèle Meurise, en étoile, en flocon…
- Trouvez-vous une corrélation entre une politique nationale et la hiérarchie des sports ?

Enfin vous devrez fournir les représentations suivantes :

- Construisez la pyramide des âges des athlètes par sexe.
    - Prévoyez les axes : tous les participants, uniquement les médaillés.
- Construisez le rapport entre le nombre de médaillés et le nombre de participants par pays.
- Construisez le tableau des médailles.
    - Prévoyez les axes pays, hiérarchie des sports, sport, sexe, type de médaille...
- Construire une représentation chronologique du nombre de médaille sur la période.
    - Prévoyez les axes pays, hiérarchie des sports, sexe, type de médaille...

Voici une représentation possible pour le rapport entre le nombre de médaillés et le nombre de participants par pays :