Projet Entrepôt de données

**Talend Open Studio**

# Fourniture

Vous recevez une base de données SQLite contenant des statistiques sur le football. Elle contient des données sur les joueurs, les équipes, les matchs…

Vous pouvez télécharger cette base de données à l’URL suivante : http://shorturl.at/ivxY2

# Scénario

Vous devez être en capacité d’analyser les caractéristiques des joueurs et des équipes présentées dans le jeu de données.

# Concept

Vous devez construire un modèle de stockage des données reçues dans votre 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…).

Enfin, vous devrez répondre obligatoirement aux questions suivantes :

* Que pensez-vous de la qualité des données ?
* Quels problèmes avez-vous rencontrées ?
* Quelle est la composition des équipes au fil des saisons ?
* Quelle est l’évolution des caractéristiques des joueurs au fil des saisons ?
  + À minima : âge, poids et taille.
* Nous désirons des courbes d'évolution par équipe des buts/nombre de matchs gagnés, par rapport aux saisons / leagues / teams…
* Nous désirons pouvoir visualiser le « radar » par saison des compétences des joueurs par rapport à leur équipe et par rapport à tous les autres joueurs.

MAITCH : id, league\_id, match\_api\_id, season\_id, home\_team\_api\_id, away\_team\_api\_id, home\_team\_goal, away\_team\_goal (Nous désirons des courbes d'évolution par équipe des buts),away\_team\_win (1 si away\_team\_goal > home\_team\_goal sinon 0), home\_team\_win (1/0), home/away\_player\_1\_11

# Modalités sur la soutenance et le rapport

Le rapport devra contenir vos propositions d’états de reporting, et répondre aux questions posées.

Vous pouvez compléter, en proposant de nouveaux rapports ayant un intérêt (ex : géo-visualisation, dataviz sur smartphone, …), un diagramme de la chaîne de chargement, des indicateurs supplémentaires, .... Plus globalement, tous axes d'améliorations sont bienvenus.

**Remise du rapport :** Le Vendredi 15 décembre 2023 13h59 dernier délai.

Une version électronique du rapport devra être envoyée à l’adresse benoit.delplace@dauphine.fr

Préfixer le sujet du mail par [Dauphine - RAPPORT\_EDT Groupe n° XX].

**Soutenance le 18/12/2023**, par groupe de 2 ou 3, mais notation individuelle.

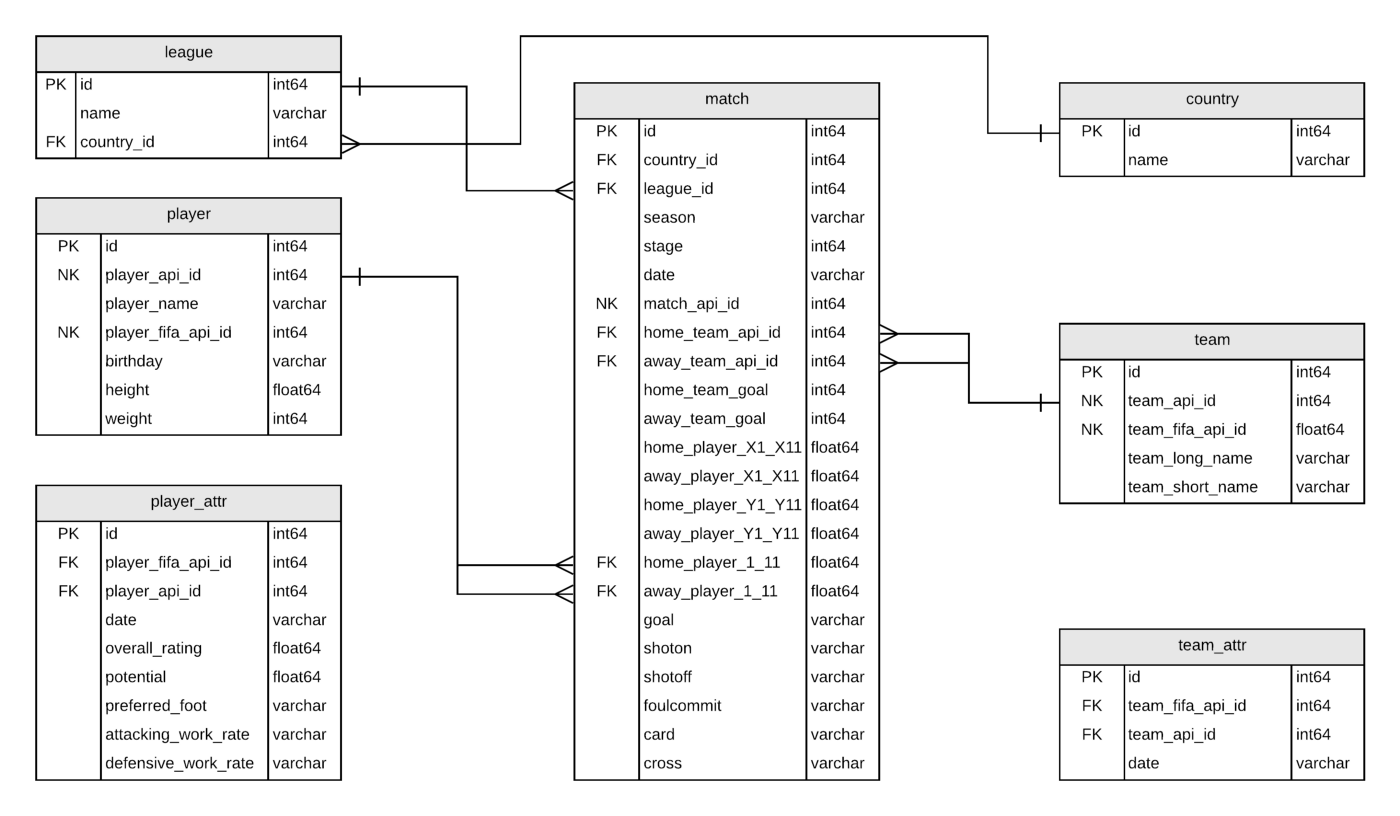
Les soutenances se décomposent en trois parties :

* Présentation (10 minutes)
* Démonstration (10 minutes)
* Questions (5 minutes)

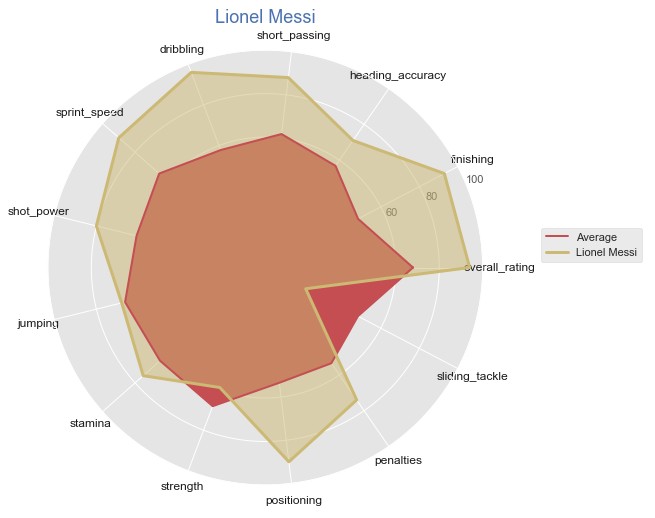
L’évaluation se décomposera de la façon suivante :

* Rapport : 40 % de la note
* Soutenance : 40 % de la note
* Implication et maîtrise individuelle du sujet : 20 % de la note

# Modèle de la base SQLite fournie



# Exemple de visualisation « radar »

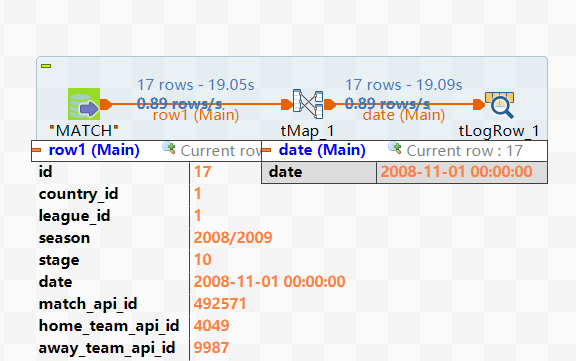


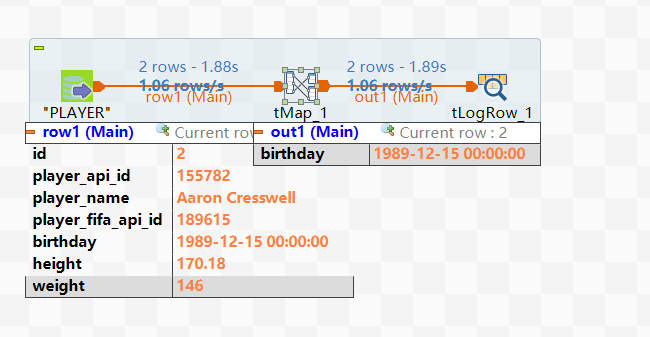
**Fait :**

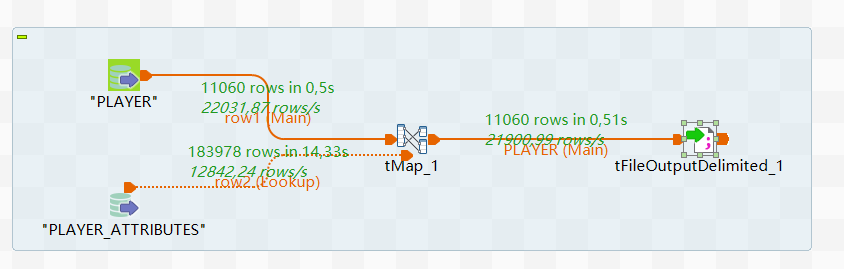
Match ( ?) : idMatch,

**Dimesion :,**

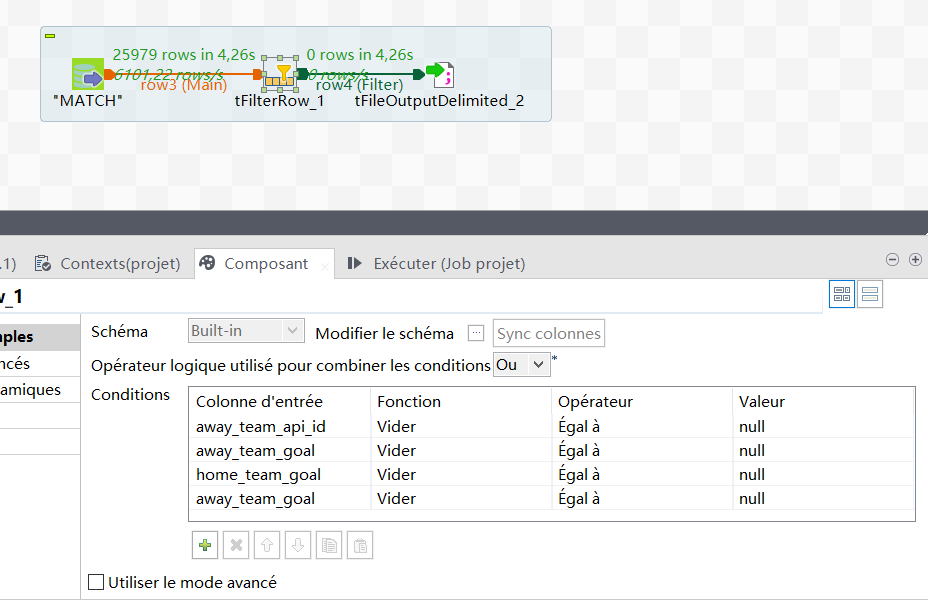
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Temps** | idTms | saison |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

****

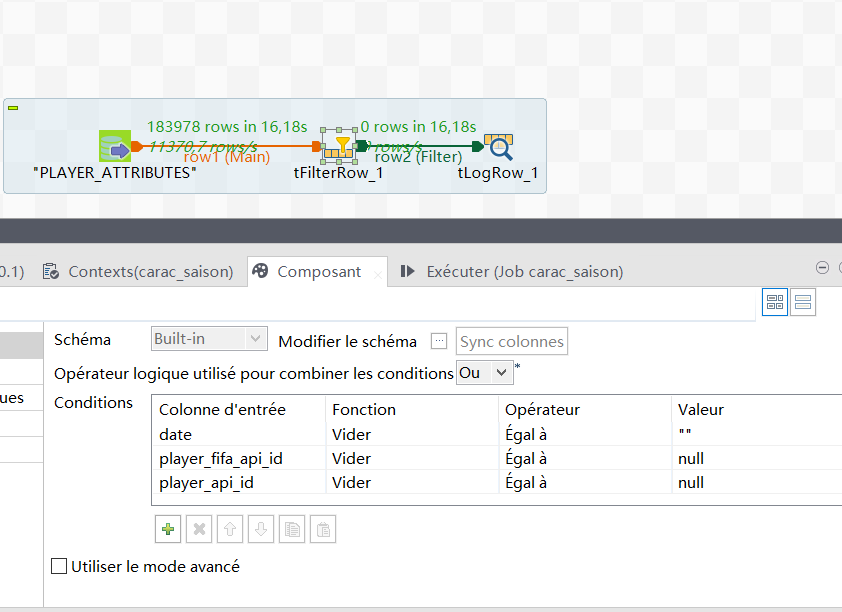
****

****

****

**证明所有match记录都有两个队伍、每个队伍的进球数也都有记录：**

**Player\_attribute的关键信息都没有缺失：**

****

**经过转换player\_attribut的date到saison，我们发现其中的saison要多于dimension包含的Saison，所以可以filter掉无关的saison**

**数据质量问题：球员的player\_api\_id不变，但fifa变了，比如api\_id = 110189**

Integer.parseInt(Var.saison.substring(0,4)) >= 2008 && Integer.parseInt(Var.saison.substring(0,4)) <= 2015

队员某个赛季的年龄：在赛季开始时的年龄