Rapport : Analyse de Données Footballistiques

Résumé Exécutif

Ce projet approfondi porte sur l'analyse de données liées au football à travers deux flux de travail distincts : le traitement par lots (batch processing) et le traitement en continu (stream processing). Les objectifs principaux sont d'extraire des informations significatives sur les performances des équipes et des joueurs dans diverses ligues de football et saisons. En combinant la collecte, le traitement et la visualisation des données, le projet vise à fournir des informations exploitables pour la prise de décisions éclairées.

Introduction

Le projet englobe deux flux de travail principaux, chacun traitant de différents aspects des données du football :

- Batch processing:

Implique la collecte de données à partir de sources diverses telles que CSV, JSON et API, couvrant plusieurs ligues de football (Bundesliga, EPL, La Liga, Ligue 1, Serie A) et différentes saisons. Les données sont ingérées dans un lac de données (HDFS) et subissent des transformations pour créer un entrepôt de données selon le modèle en étoile. Deux data marts sont créés, se concentrant sur les équipes et les joueurs. L'orchestration de ce flux de travail est gérée à l'aide d'Apache Airflow. Les données résultantes sont ensuite visualisées à l'aide de Power BI.

- Stream processing:

Englobe la collecte en temps réel de données sur les événements de matchs à partir d'une source API. Ces données sont diffusées à l'aide de Kafka et Spark, puis stockées dans MongoDB pour une récupération efficace. Dash, un framework Python, est utilisé pour créer une application web permettant des visualisations interactives des événements de matchs.

Détails du Batch processing

<u>Étape 1 : Collecte de Données</u>

Collecte exhaustive de données à partir de sources CSV, JSON et API, couvrant une variété de ligues de football.

<u>Étape 2 : Ingestion de Données</u>

Ingestion des données dans HDFS en tant que lac de données central pour un traitement ultérieur.

<u>Étape 3 : Transformation des Données</u>

3.1 Nettoyage :

Identification et suppression des lignes avec des entrées problématiques, améliorant la qualité des données.

3.2 Suppression de Colonnes :

Élimination des colonnes redondantes ('xGChain', 'xGBuildup') pour simplifier l'ensemble de données.

3.3 Colonnes en Pourcentage :

Introduction de nouvelles colonnes en pourcentage ('xG', 'xA', 'npxG') pour fournir des informations sur la contribution relative des métriques.

3.4 Gestion des NaN :

Imputation des valeurs NaN dans certaines colonnes avec 0 pour assurer un ensemble de données cohérent.

3.5 Calcul des Tirs Manqués :

Création d'une nouvelle colonne ('missed_shots') en soustrayant les 'buts' des 'tirs', offrant une vue nuancée de la performance du joueur.

3.6 Réorganisation des Colonnes :

Optimisation de l'ordre des colonnes pour une meilleure lisibilité et facilité d'utilisation.

Étape 4 : Création de Data Marts

Développement de deux data marts selon les principes du modèle en étoile, permettant une analyse nuancée des équipes et des joueurs.

Étape 5 : Orchestration

Utilisation d'Apache Airflow pour une orchestration sans heurts, assurant l'exécution coordonnée des tâches.

Étape 6 : Visualisation

Utilisation de Power BI pour la création de visualisations informatives, offrant une interface conviviale pour explorer les statistiques des équipes et des joueurs.

- Tableau de Bord des Statistiques des Équipes

6.1.1 Top 10 des Équipes les Plus Prolifiques :

Mise en avant des équipes avec les meilleurs scores, facilitant les comparaisons rapides.

6.1.2 Top 5 des Équipes avec le Plus de Cartons Rouges et Jaunes :

Identification des équipes avec un nombre notable d'actions disciplinaires, offrant des insights sur leur style de jeu.

6.1.3 Top 10 des Équipes Créant le Plus d'Opportunités de Marquer :

Mise en avant des équipes excellent dans la création d'opportunités de marquer, éclairant sur les stratégies offensives.

6.1.4 Buts par Ligue :

Représentation visuelle de la répartition des buts dans différentes ligues de football, facilitant l'analyse des performances par ligue.

- Tableau de Bord des Performances des Joueurs

6.1.5 Buts et Tirs Manqués par Nom de Joueur :

Offre une vue détaillée de la performance individuelle des joueurs, mettant l'accent sur les buts et les tirs manqués.

6.1.6 Cartons Rouges et Jaunes par Nom de Joueur :

Illustration des antécédents disciplinaires des joueurs, aidant à dresser le profil des joueurs.

6.1.7 Top 5 des Joueurs avec le Plus de Temps de Jeu :

Identification des joueurs avec le temps de jeu le plus élevé, mettant en lumière leur contribution à l'équipe.

Détails du Stream processing

Étape 1 : Collecte de Données

Collecte en temps réel de données sur les événements de matchs à partir d'une source API dédiée.

Étape 2 : Stream Processing

Intégration de Kafka et Spark pour un traitement en continu efficace, assurant l'ingestion et l'analyse des données en temps réel.

Étape 3 : Stockage des Données

Stockage des données traitées dans MongoDB pour une récupération efficace et une analyse future.

Étape 4 : Visualisation

Implémentation d'une application web Dash pour des visualisations interactives des événements de matchs.

Détails Techniques

Bibliothèques et Outils Utilisés

- Traitement par Lots :

  Apache Airflow

  Power BI

  SQL Server

  Python

  Hadoop

  Docker

- Traitement en Continu :

  Kafka

  Spark

  Dash

  MongoDB

Conclusion

Ce projet combine avec succès la puissance du traitement par lots et en continu pour effectuer une analyse approfondie des données liées au football. Les flux de travail détaillés, les étapes de transformation et les tableaux de bord de visualisation fournissent aux parties prenantes une compréhension complète des performances des équipes et des joueurs. L'intégration de divers outils et technologies démontre une approche robuste et évolutive de l'analyse des données sportives. Les insights issus de ce projet peuvent orienter les décisions stratégiques pour les équipes, les entraîneurs et les passionnés de football.