

Cahier des Charges : Projet Data Football

Introduction

- **Contexte:** Mise en place de workflows Batch et Stream pour l'analyse des données football.
- **Objectifs:** Collecte, transformation, stockage, et visualisation des données de joueurs, équipes, et événements de match.

Description du Projet

Batch Processing:

1. Collecte de données (CSV, JSON, API) de différentes ligues sur plusieurs saisons.
2. Ingestion dans le Data Lake (HDFS), transformation, et création d'un Data Warehouse en étoile.
3. Data Marts distincts pour les équipes et les joueurs.
4. Orchestration avec Apache Airflow.
5. Visualisation des Data Marts avec Power BI.

Stream Processing:

1. Collecte de données d'événements de match via une API en continu avec Kafka et Spark.
2. Stockage dans MongoDB.
3. Visualisation en temps réel avec des graphiques interactifs.

Besoins et Exigences

- **Besoins Métier:** Analyse approfondie des performances des équipes et des joueurs.
- **Exigences Techniques:** Intégration avec diverses sources, respect des délais, résilience en cas de problèmes de connexion.

Architecture et Conception

Batch Processing:

- Sources: CSV, JSON, API.
- Data Lake (HDFS), Data Warehouse en étoile.
- Outils: Hadoop, SQL Server, Apache Airflow, Power BI, Docker.

Stream Processing:

- Source: API (événements de match).
- Stockage: MongoDB.
- Outils: Kafka, Spark, Dash (Python), Docker.

Livrables

- Data Marts pour équipes et joueurs.
- Visualisations interactives avec Power BI et Dash.

Contraintes

Techniques:

- Gestion des problèmes de connexion.
- Optimisation des performances sur l'ordinateur local.

Temporelles:

- Livraison du projet en un mois.

Risques et Mitigation

- Suivi hebdomadaire du projet avec le formateur.
- Allocation de ressources supplémentaires en cas de besoin.