Documentation sur les Transformations Appliquees aux Donnees

Introduction

Ce pipeline de streaming Spark lit des donnees en temps reel depuis Kafka, les nettoie en appliquant des transformations specifiques, et les stocke dans la zone cleansed de HDFS. Ce document decrit les etapes principales et les transformations appliquees a chaque type de donnees.

1. Types de Donnees Geres

a. Transactions

Contient les donnees des transactions realisees par les utilisateurs.

- Schema : transaction_id, user_id, product_id, date, quantity, total_amount, payment_method

b. Publicites

Donnees relatives aux campagnes publicitaires.

- Schema : ad_name, platform, objective, impressions, clicks, conversions, ctr, cpc, conversion_rate

c. Logs

Logs techniques contenant des informations textuelles.

- Schema : content

2. Transformations Appliquees

a. Transactions

Nettoyages effectues :

1. Remplacement des valeurs manquantes :

- user_id -> -1

- total_amount -> 0.0

- payment_method -> Unknown

2. Suppression des transactions sans date :

- Les lignes ou date est null sont eliminees.

3. Marquage des transactions incompletes :

- Une nouvelle colonne is_incomplete est ajoutee pour indiquer si product_id ou quantity est null.

4. Suppression des doublons :

- Les transactions ayant le meme transaction_id sont dedupliquees.

b. Publicites

Nettoyages effectues :

1. Remplacement des valeurs manquantes :

- ad_name -> Unknown Ad

- platform -> Unknown

- objective -> Unknown

- impressions, clicks, conversions -> 0

- ctr, cpc, conversion_rate -> 0.0

2. Suppression des doublons :

- Les lignes ayant la meme combinaison ad_name, platform, et objective sont dedupliquees.

c. Logs

Nettoyages effectues :

Documentation sur les Transformations Appliquees aux Donnees

1. Extraction des champs pertinents depuis la colonne content :

- response_time : temps de reponse, extrait sous forme float

- user_id : identifiant utilisateur, extrait sous forme float

2. Suppression des doublons :

- Les logs ayant le meme contenu sont dedupliquees.