3D Smart Factory

ML TEMPS RÉEL SUR FLUX DE DONNÉES

Présenté par:

- EL BOUGRINI Nassim
- HAFSI Siham
- CHAHIDI Khadija
- RAHMOUN Hayat
- AGOUMI Achraf
- NACHOUR Ilham

Encadré par:

- M.BERTIN Thierry

PLAN

LES DONNÉES

INGESTION DES DONNÉES

PRÉTRAITEMENT DES DONNÉES

INGÉNIERIE DES CARACTÉRISTIQUES

ENTRAÎNEMENT DU MODÈLE

PRÉDICTIONS EN TEMPS RÉEL

Présentation des Données

ALPHA VANTAGE

Le site présente plusieurs avantages:

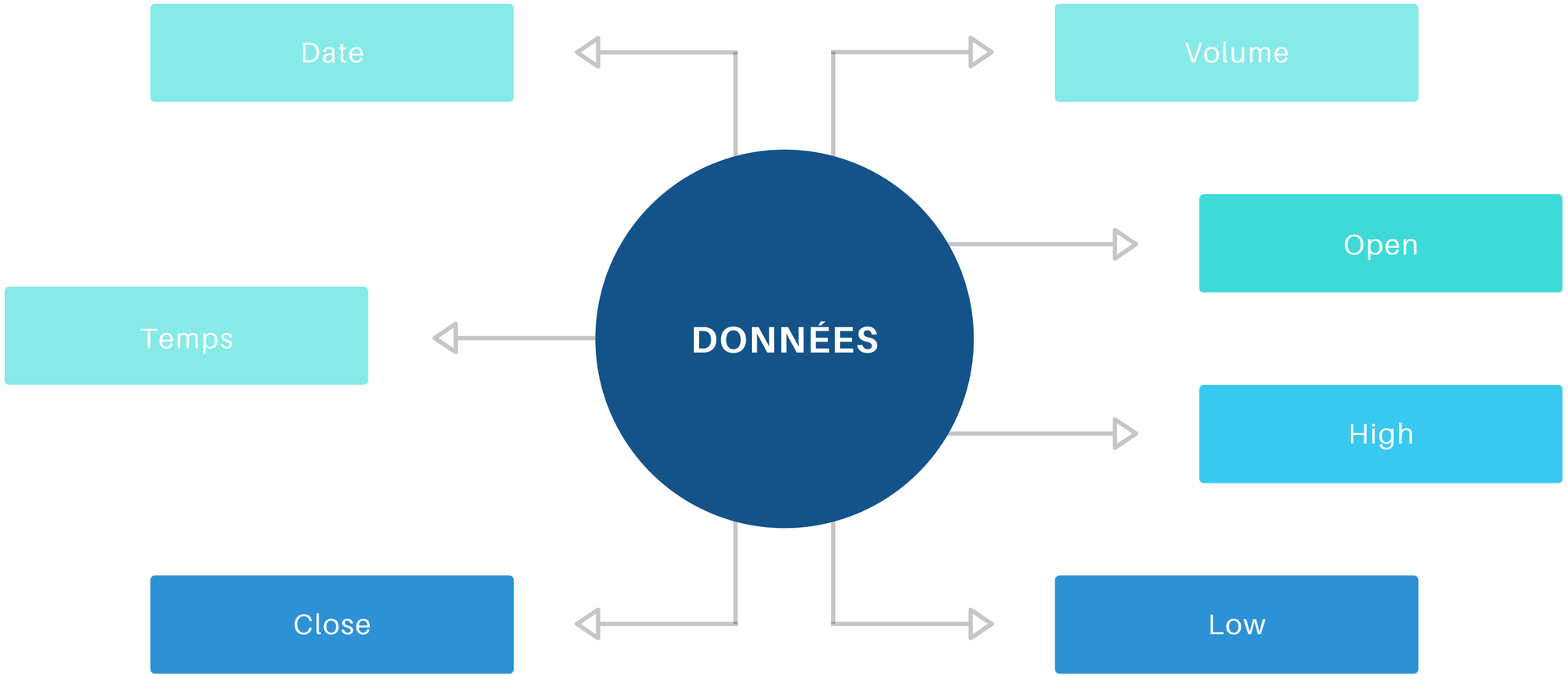
- Présentation d'un flux de données en temps réel.
- Choix du type de données suivant la société et les intervalles temporelles.
- Association de l'URL du site a notre model en utilisant les requêtes suivantes:

```
import requests

# replace the "demo" apikey below with your own key from https://www.alphavantage.co/support/
#api-key
url = 'https://www.alphavantage.co/query?function=TIME_SERIES_INTRADAY&symbol=IBM&interval=5m
in&apikey=demo'
r = requests.get(url)
data = r.json()

print(data)
```

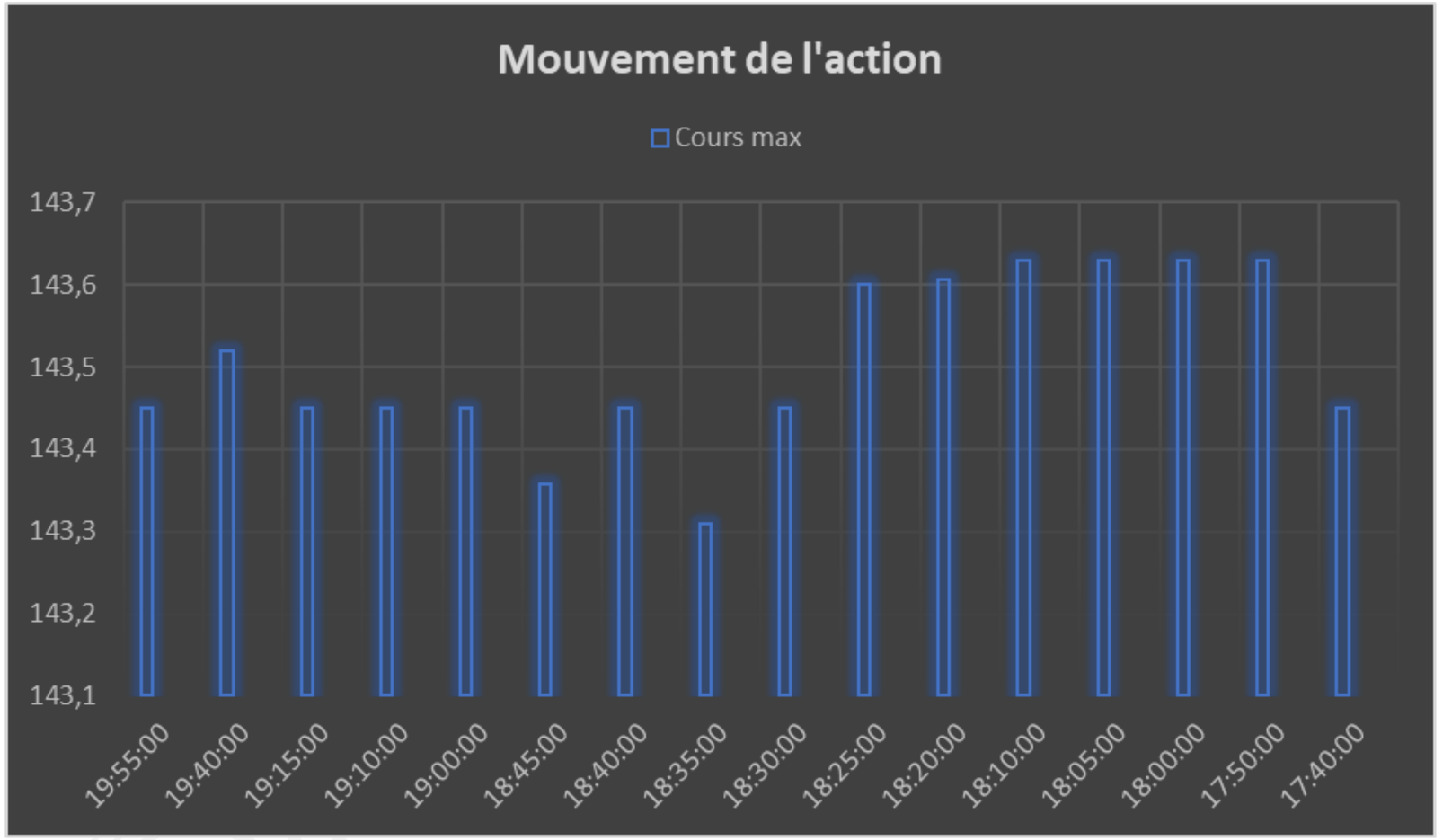
TEMPLATE DES DONNÉES

Date	Time	Open	High	Low	Close	Volume
28/07/2023	19:55:00	143.4200	143.4500	143.4200	143.4500	2
28/07/2023	19:40:00	143.5200	143.5200	143.5200	143.5200	10
28/07/2023	19:15:00	143.4500	143.4500	143.4500	143.4500	3
28/07/2023	19:10:00	143.4200	143.4500	143.4200	143.4500	2
28/07/2023	19:00:00	143.4500	143.4500	143.3100	143.3100	1262815
28/07/2023	18:45:00	143.3580	143.3580	143.3580	143.3580	50
28/07/2023	18:40:00	143.4500	143.4500	143.4500	143.4500	2
28/07/2023	18:35:00	143.3100	143.3100	143.3100	143.3100	100
28/07/2023	18:30:00	143.4500	143.4500	143.4500	143.4500	1262814
28/07/2023	18:25:00	143.6000	143.6000	143.6000	143.6000	10
28/07/2023	18:20:00	143.6060	143.6060	143.6000	143.6000	6
28/07/2023	18:10:00	143.6300	143.6300	143.6300	143.6300	1
28/07/2023	18:05:00	143.5500	143.6300	143.5500	143.6280	112
28/07/2023	18:00:00	143.5000	143.6300	143.5000	143.6130	26
28/07/2023	17:50:00	143.6300	143.6300	143.5000	143.5000	6
28/07/2023	17:40:00	143.4500	143.4500	143.4500	143.4500	16
28/07/2023	17:35:00	143.3100	143.3100	143.3100	143.3100	60
28/07/2023	17:30:00	143.5000	143.5000	143.4500	143.4500	71
28/07/2023	17:20:00	143.5000	143.5000	143.4500	143.4500	3
28/07/2023	17:15:00	143.5500	143.6200	143.5500	143.6200	2

Explication des colonnes



VISUALISATION GRAPHIQUE

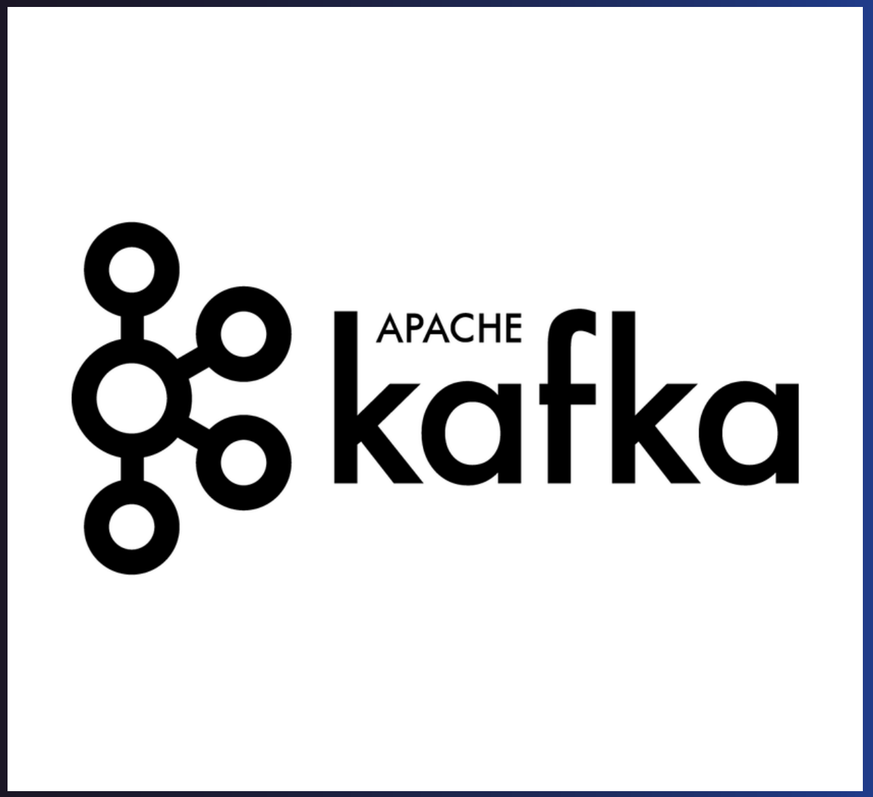


INGESTION DES DONNÉES

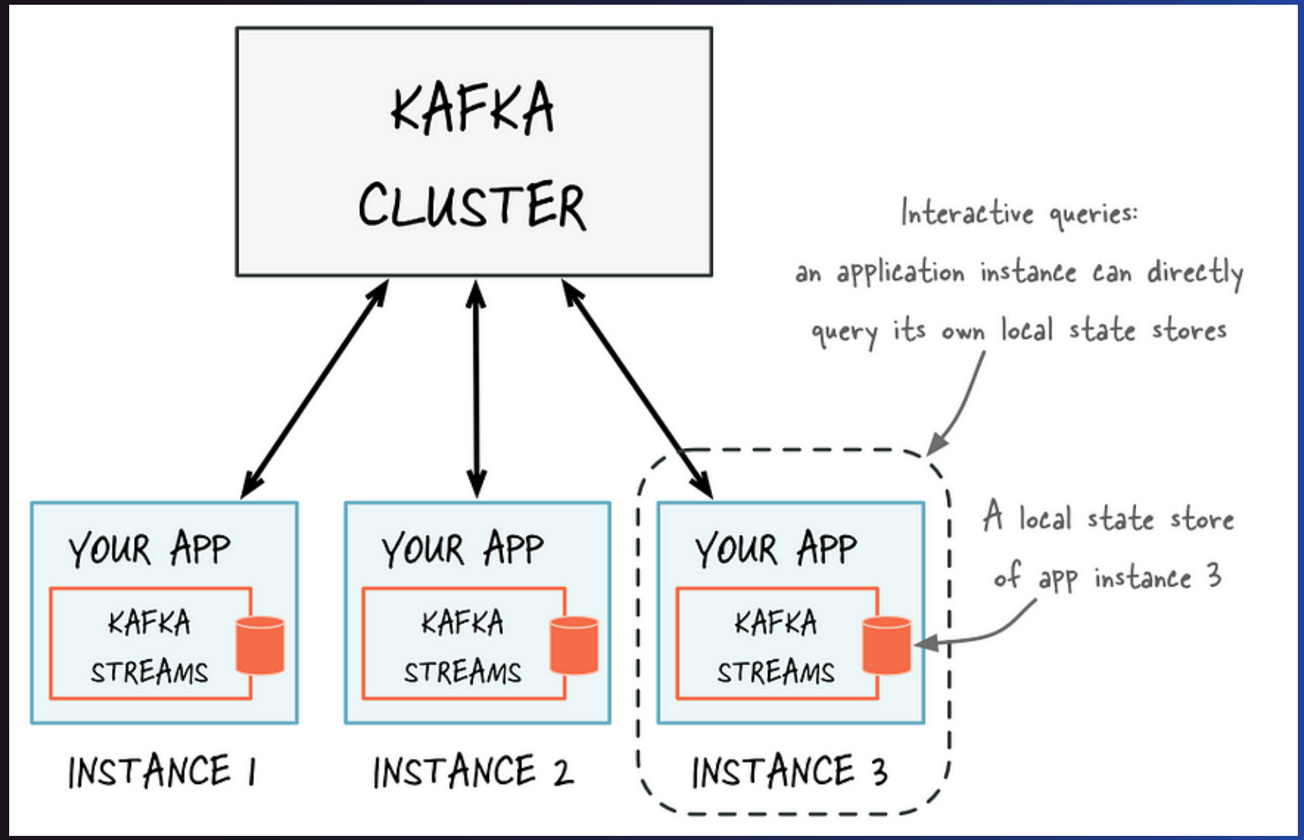
L'ingestion des données fait référence au processus de collecte, de réception et de stockage des données en temps réel provenant de la source (bourses de valeurs ou fournisseurs de données financières) dans un système centralisé où elles peuvent être traitées et analysées efficacement. Pour gérer l'ingestion des données financières en temps réel, nous utilisons un système de messagerie appelé Apache Kafka.

Qu'est-ce que Kafka ?

Kafka est un logiciel open source qui fournit un cadre pour le traitement, la lecture et l’analyse de données en continu.
Être open source garantit qu’il est essentiellement gratuit à utiliser et dispose d’un vaste réseau d’utilisateurs et de développeurs qui contribuent avec des mises à jour, de nouvelles fonctionnalités et un nouveau support utilisateur.



Kafka est conçu pour fonctionner dans un environnement « distribué », ce qui signifie qu'il fonctionne sur plusieurs serveurs au lieu de rester assis sur l'ordinateur d'un utilisateur, exploitant la puissance de calcul et la capacité de stockage supplémentaires que cela fournit.

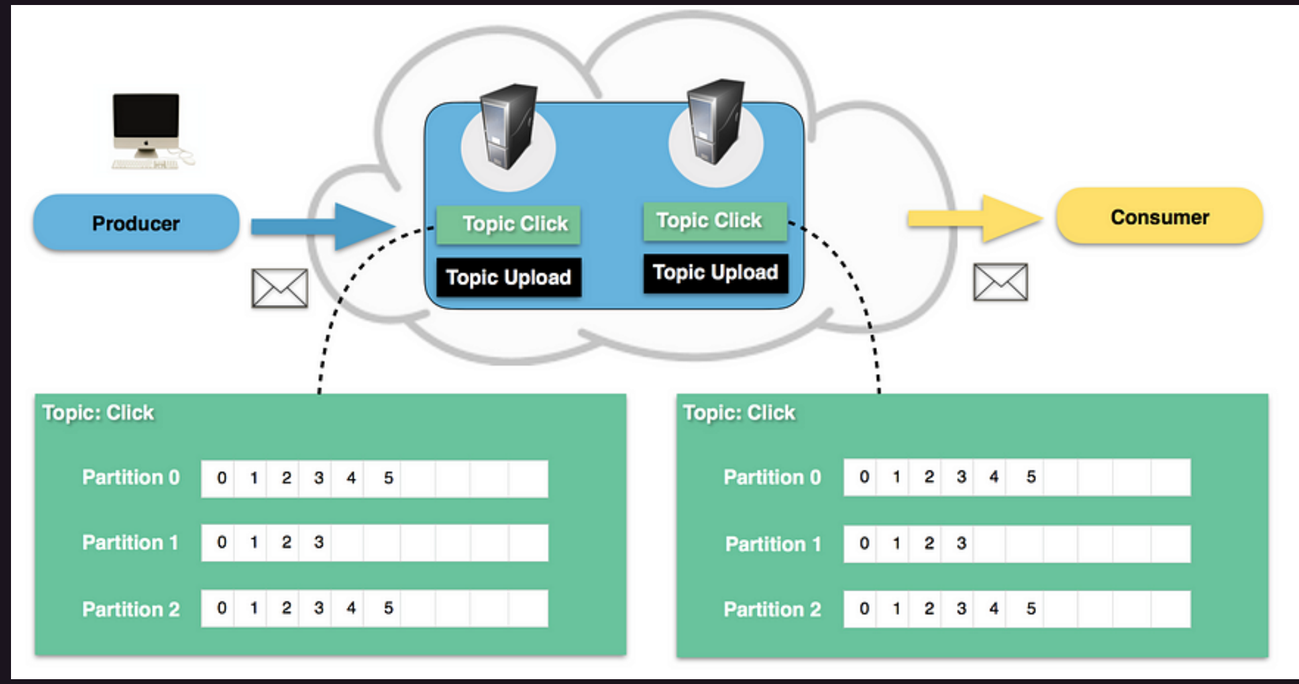


Quelle est l’utilité de Kafka ?

Kafka est capable de fonctionner très rapidement en raison de sa nature distribuée et de la manière automatisée dont il gère les données entrantes – les grands clusters peuvent suivre et répondre à des millions de changements dans un ensemble de données chaque seconde.

Comment fonctionne Kafka ?

Apache Kafka est une plateforme de traitement de flux de données qui organise les informations provenant de diverses sources en "sujets". Ces sujets représentent des thèmes spécifiques, tels que le nombre de ventes d'un produit ou une plage horaire donnée. Contrairement aux bases de données traditionnelles, Kafka gère efficacement d'énormes volumes de données en temps réel. Son "processeur" agit comme une interface entre les applications et les sujets de Kafka. Les données générées peuvent alimenter des bases de données distribuées et des pipelines d'analyse Big Data. Un composant supplémentaire, le "consommateur", transfère les informations aux applications concernées. Ainsi, Kafka joue un rôle central dans la circulation et l'analyse des flux de données pour les systèmes Big Data.



PRÉTRAITEMENT DES DONNÉES

- Le prétraitement des données en temps réel consiste à nettoyer, normaliser et transformer les données en continu, juste après leur réception, avant d'être consommées par le modèle pour garantir que les données sont bien préparées.

- Le prétraitement commence par la consommation des données en temps réel à partir des sujets Kafka par le framework de traitement de flux, tel qu'Apache Flink ou Apache Spark.

Apache Spark et Apache Flink:
qu’est-ce que c’est et à quoi ça sert ?

Spark est un moteur d'analyse rapide et unifié adapté au Big Data et au Machine Learning, tandis que **Flink** est une plateforme de traitement distribué conçue pour les applications Big Data avec un accent sur l'analyse de données stockées dans des clusters Hadoop.

APACHE SPARK OU APACHE FLINK?

- **Apache Spark** : Traitement en temps réel et par lots, écosystème plus mature, latence légèrement supérieure en flux.

- **Apache Flink** : Traitement de flux en temps réel, latence faible, modèle de fenêtres avancé.

Processus de prétraitement des données en temps réel

1 **Consommation des données en temps réel depuis Kafka**

Les données sont produites en continu par les producteurs et envoyées à des sujets (topics) Kafka. Le framework de traitement de flux (par exemple, Apache Flink ou Apache Spark) se connecte à Kafka pour consommer ces données en temps réel.

2 EXTRACTION DES DONNÉES

Le framework de traitement de flux extrait les données des messages Kafka, qui sont souvent au format JSON, CSV, Avro, ou autre.

3

Nettoyage des données

Le prétraitement en streaming peut impliquer le filtrage des données pour supprimer les enregistrements inutiles ou corrompus, remplir les valeurs manquantes ou supprimer les valeurs aberrantes.

4

NORMALISATION

La normalisation est le processus de mise à l'échelle des caractéristiques (features) des données pour les ramener dans une plage spécifique.

5

TRANSFORMATION DES DONNÉES

Dans cette étape, les données sont remodelées ou structurées de manière à être adaptées à la consommation par le modèle.

6

Enrichissemen des données (optionnel)

il peut être nécessaire d'enrichir les données avec des informations supplémentaires provenant d'autres sources, pour améliorer la performance du modèle ou augmenter sa pertinence.

7

TRANSMISSION DES DONNÉES PRÉTRAITÉES

FEATURE ENGINEERING

FEATURE ENGINEERING

DATA RAW

FEATURE SELECTION

MODEL

FEATURE ENGINEERING

Le processus de sélection, de manipulation et de transformation des données brutes en caractéristiques pertinentes qui peuvent être utilisées dans les modèles d'apprentissage automatique. L'objectif est d'améliorer la performance et l'efficacité des modèles prédictifs en fournissant des informations utiles pour la tâche d'apprentissage supervisé.

LES TECHNIQUES DE FEATURE ENGINEERING

Création de nouvelles caractéristiques

consiste à générer de nouvelles informations à partir des caractéristiques existantes dans les données. Cette approche permet d'enrichir la représentation des données, d'introduire des relations non linéaires et d'apporter des connaissances spécifiques au domaine

Transformation

fait référence au processus de modification des données existantes pour les rendre plus appropriées pour l'analyse . normalisation,one-hot-encoder,handling outliers,...

feature selection

L'objectif est de ne conserver que les caractéristiques les plus pertinentes et informatives pour améliorer la performance du modèle, accélérer l'entraînement, et éviter le surapprentissage.

Entraînement du modèle

Entraînement du modèle

CHOIX DU MODÈLE

SÉLECTION DES CARACTÉRISTIQUES

PARTITIONNEMENT DES DONNÉES

ENTRAÎNEMENT DU MODÈLE

RÉGLAGE DES HYPERPARAMÈTRES

Entraînement du modèle

CHOIX DU MODÈLE

Pour notre analyse en temps réel des flux de données financières, nous avons opté pour l'utilisation de modèles de séries temporelles.

Entraînement du modèle

CHOIX DU MODÈLE

SÉLECTION DES CARACTÉRISTIQUES

Pour nos modèles de séries temporelles, nous avons sélectionné des caractéristiques telles que les moyennes mobiles sur différentes périodes, les ratios financiers clés et les scores de sentiment provenant de l'analyse de textes financiers.

Entraînement du modèle

CHOIX DU MODÈLE — SÉLECTION DES CARACTÉRISTIQUES — PARTITIONNEMENT DES DONNÉES

Nous avons divisé nos données historiques en deux ensembles : 80% pour l'entraînement du modèle et 20% pour le test.

Entraînement du modèle

CHOIX DU MODÈLE — SÉLECTION DES CARACTÉRISTIQUES — PARTITIONNEMENT DES DONNÉES — ENTRAÎNEMENT DU MODÈLE

Notre modèle est entraîné à l'aide de Python et la bibliothèque scikit-learn en utilisant l'algorithme des moindres carrés.

Entraînement du modèle

CHOIX DU MODÈLE

SÉLECTION DES CARACTÉRISTIQUES

PARTITIONNEMENT DES DONNÉES

ENTRAÎNEMENT DU MODÈLE

RÉGLAGE DES HYPERPARAMÈTRES

Nous avons effectué une recherche systématique des hyperparamètres pour trouver les valeurs optimales qui améliorent les performances du modèle.

Predictions en Temps Réel

Intégration du modèle dans le code de traitement de flux

Une fois que le modèle d'apprentissage automatique est entraîné, il est intégré au code de traitement de flux de données. Une connexion est établie avec la plateforme de streaming Kafka pour recevoir les nouvelles données en temps réel. Grâce à un mécanisme asynchrone, le flux continu de données est géré, ce qui permet au modèle d'effectuer des prédictions en temps réel sans retard.

Prédictions en temps réel

Une fois que les nouvelles données sont reçues depuis Kafka, le modèle d'apprentissage automatique est activé pour effectuer des prédictions en temps réel sur l'évolution des actions en bourse. Le modèle utilise les caractéristiques spécifiques de ces données en temps réel pour générer des prédictions actualisées.

Adaptation continue

Les modèles d'apprentissage automatique peuvent nécessiter des mises à jour régulières pour maintenir leur précision dans un environnement en constante évolution comme le marché boursier.En fonction des performances du modèle et des changements dans les données de marché, il peut être nécessaire de ré-entraîner périodiquement le modèle ou de l'adapter en utilisant des techniques telles que le "transfer learning"

3D Smart Factory

MERCI POUR VOTRE ATTENTION