Projet Big Data

Business Intelligence & Analytics

Real-time Stock Market Prediction Using Sentiment Analysis on Social Network

Réalisé par :

ABJAOU Oumaima
BENALI Amine
CHOUKHANTRI Ikram
GHAZOUAN Oumaima
SAADI Naoufal

Supervisé par :

M. Yasser EL Alami EL Madani

Année Universitaire 2024 - 2025

LISTE DE FIGURES

TABLE DES MATIÈRES

INTRODUCTION GÉNÉRALE

L'essor des technologies Big Data et l'omniprésence des réseaux sociaux ont profondément modifié les dynamiques économiques et sociales. Les plateformes comme Reddit, où les utilisateurs partagent leurs opinions et analyses sur des sujets variés, sont devenues une source précieuse de données exploitables pour des prédictions stratégiques. Dans ce contexte, notre projet vise à combiner l'analyse des sentiments exprimés sur les réseaux sociaux et les données financières pour prédire les variations des actions boursières, en particulier celles de Tesla, en temps réel. Ce rapport détaille les différentes étapes de conception et de réalisation de cette solution, depuis la collecte des données jusqu'à leur visualisation sur un tableau de bord interactif, en passant par l'entraînement de modèles prédictifs avancés. Ce travail met en lumière les défis et les opportunités associés à l'exploitation des flux massifs de données pour la prise de décision stratégique.

CHAPITRE 1

CONTEXTE GÉNÉRAL DU PROJET

1.1 Contexte général

L'essor des technologies numériques et l'explosion des réseaux sociaux ont transformé la manière dont les informations circulent et influencent les décisions économiques. Parmi ces réseaux, Reddit occupe une place particulière en tant que plateforme communautaire où les utilisateurs partagent des opinions, des analyses et des prédictions sur divers sujets, y compris les marchés financiers. Parallèlement, les actions comme TESLA suscitent un intérêt marqué, générant un flux constant de discussions et de données pouvant influencer les tendances boursières.

Dans ce contexte, la capacité à exploiter ces données massives en temps réel pour prédire l'évolution des marchés financiers est devenue un enjeu stratégique. Cependant, ces données présentent des défis uniques liés à leur volume, leur vélocité et leur variabilité.

1.2 Problématique

La principale problématique de ce projet réside dans l'intégration, le traitement et l'analyse en temps réel des données issues de Reddit et des marchés financiers pour prédire les variations des actions TESLA. Cela soulève plusieurs questions :

- Quels outils et technologies utiliser pour gérer le stockage, le traitement et l'apprentissage machine sur ces données ?
- Comment visualiser les résultats pour fournir des insights exploitables aux utilisateurs finaux ?

1.3 Planification du projet

Afin de mener ce projet à bien, une planification rigoureuse a été mise en place. Le projet a été divisé en plusieurs étapes majeures, chacune assortie de délais précis, comme illustré dans le diagramme de Gantt :

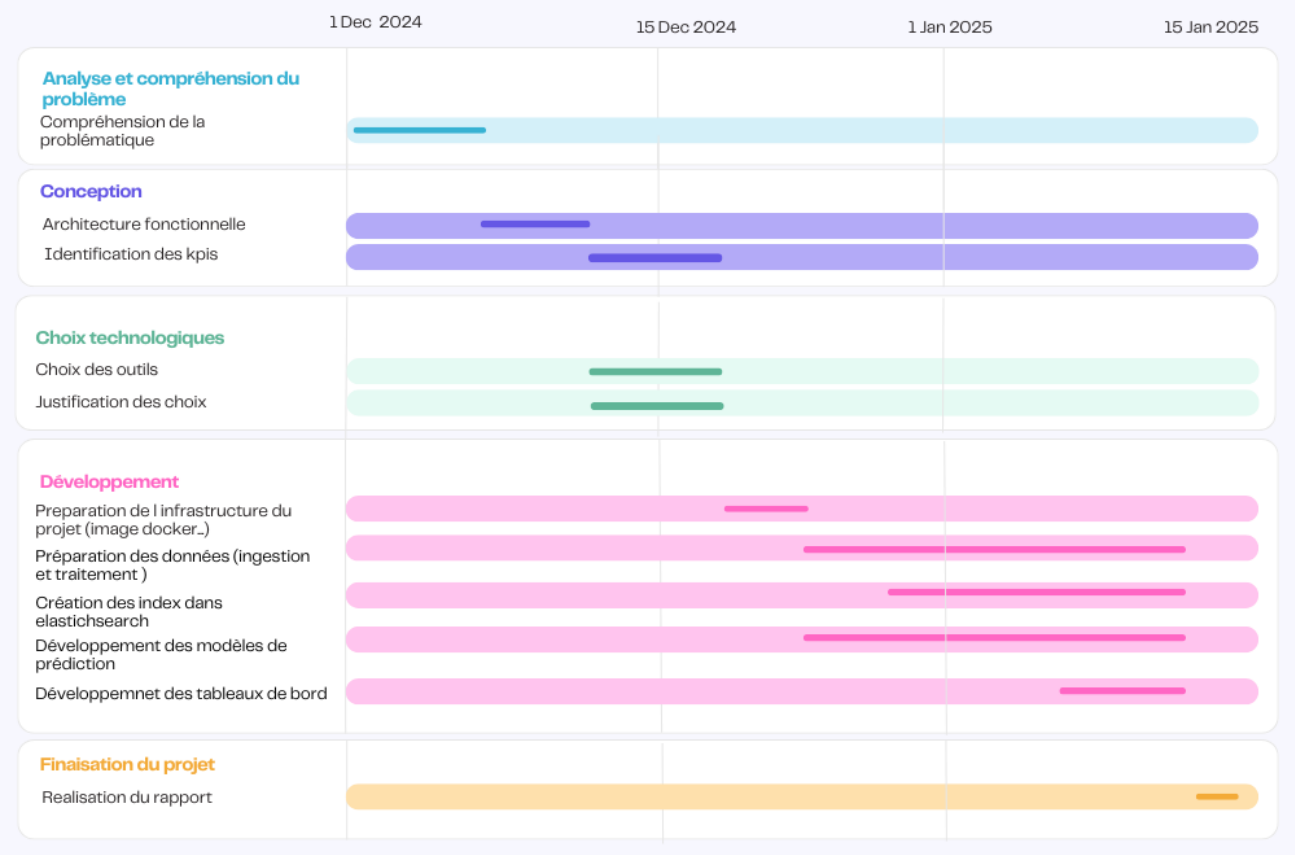

FIGURE 1.1 – Diagramme de Gantt

1.3.1 Étapes principales du projet

1. **Analyse et compréhension du problème (1er décembre 2024 - 7 décembre 2025)** :
   - Compréhension approfondie de la problématique.
   - Identification des objectifs clés et des besoins du projet.
2. **Conception (7 décembre 2024 - 20 décembre 2025)** :
   - Définition de l'architecture fonctionnelle.
   - Identification et sélection des KPIs pertinents.
3. **Choix technologiques (10 décembre 2025 - 20 décembre 2025)** :
   - Sélection des outils technologiques adaptés.
   - Justification des choix technologiques en fonction des besoins.
4. **Développement (10 décembre 2024 - 10 janvier 2025)** :
   - Préparation de l'infrastructure du projet (par exemple, configuration d'une image Docker).
   - Ingestion et traitement des données.
   - Création des index dans Elasticsearch.
   - Développement des modèles de prédiction.
   - Conception et intégration des tableaux de bord interactifs.
5. **Finalisation du projet (10 janvier 2025 - 13 janvier)** :

— Rédaction du rapport final.
— Préparation des livrables (rapport, présentation, code source).

Un suivi hebdomadaire des tâches a été réalisé à l'aide de Trello, permettant d'identifier rapidement les blocages et de les résoudre. L'intégration de DDS a renforcé la transparence et l'efficacité dans la gestion des cycles du projet.

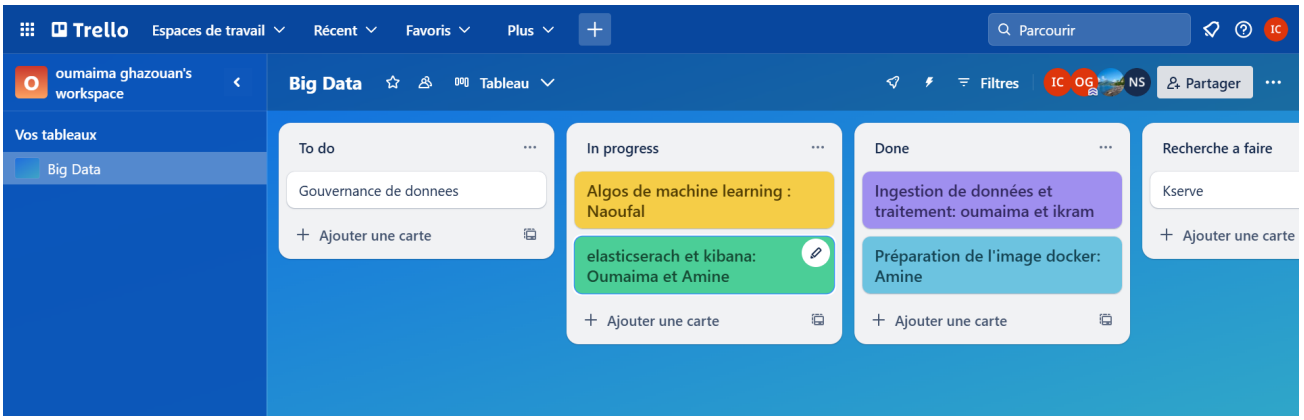

FIGURE 1.2 – Trello

1.3.2 Approche Data-Driven Scrum (DDS)

L'approche Data-Driven Scrum (DDS) combine les principes agiles avec une orientation basée sur les données pour structurer le développement du projet. Cette méthode garantit une adaptation continue aux exigences dynamiques des données massives et des analyses avancées. Les principales caractéristiques de cette approche sont les suivantes :

- **Travail par étapes** : Chaque sprint se concentre sur une tâche précise, comme le traitement des données, la création d'un modèle ou la visualisation.
- **Décisions basées sur les données** : Le progrès est évalué à chaque étape, et les priorités sont ajustées selon les besoins.
- **Travail d'équipe** : Les membres de l'équipe travaillent ensemble pour réaliser le projet.
- **Amélioration continue** : Après chaque sprint, le travail est évalué pour trouver des moyens de s'améliorer et avancer plus efficacement.

Cette approche permet de rester flexible et de progresser pas à pas pour atteindre les résultats attendus du projet tout en s'adaptant aux besoins.

CHAPITRE 2

CONCEPTION DU PROJET

2.1 Architecture Fonctionnelle

Le schéma ci-dessous illustre l'enchaînement des différentes étapes fonctionnelles, depuis la collecte des données jusqu'à leur restitution sous forme d'analyses et de visualisations exploitables :

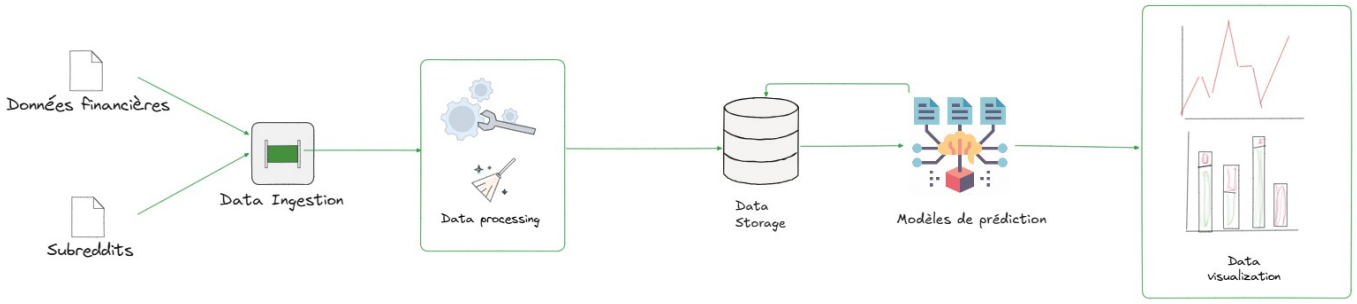

FIGURE 2.1 – Schéma de l'architecture fonctionnelle

Cette architecture assure une gestion fluide des données et une analyse efficace pour répondre aux objectifs du projet. L'architecture fonctionnelle mise en place repose sur une approche modulaire et scalable, adaptée à la gestion de flux de données massifs et au traitement analytique avancé. Elle se compose des éléments suivants :

- **Sources de données** :
  - *Données textuelles* : Provenant de plateformes de discussion Reddit, ces données incluent les messages des utilisateurs publiés dans des subreddits spécifiques, par exemple ceux relatifs à des entreprises ou secteurs ciblés (comme Tesla).
  - *Données financières* : Extraites de Yahoo Finance.
- **Pipeline d'ingestion** : Un mécanisme dédié au transfert des données brutes depuis les sources vers les modules de traitement.

- **Traitement des données** : Cette étape comprend le nettoyage et la transformation
- **Stockage** : Un système structuré et optimisé pour accueillir des données structurées, avec des capacités de recherche rapide.
- **Modélisation et analyse prédictive** :
  - **Apprentissage automatique pour les prévisions basées sur des séries temporelles** : Utilisation d'algorithmes avancés pour prédire les prix des actions en fonction des données passées et évaluer l'impact des sentiments extraits des tweets sur les prédictions des prix des actions.
- **Visualisation et restitution** : Un tableau de bord dynamique permettant de suivre en temps réel les indicateurs clés et les résultats des prédictions.

2.2 Identification des KPIs

Dans le cadre de ce projet, plusieurs KPIs ont été identifiés pour mesurer les performances et fournir des insights basés sur les données collectées et analysées. Les principaux KPIs sont les suivants :

- **Prix moyen des actions (*Average of Stock Price*)** :
  - Cet indicateur donne une vue d'ensemble de la tendance générale des prix des actions sur une période donnée.
  - Utilité : Identifier les variations moyennes et les tendances à long terme.
- **Prix maximum des actions (*Maximum Stock Price*) par date** :
  - Permet de déterminer les périodes où les actions ont atteint leur pic.
  - Utilité : Suivre les opportunités de marché.
- **Prix prédit des actions (*Predicted Stock Price*)** :
  - Ce KPI est basé sur les modèles prédictifs et fournit une estimation future des prix des actions.
  - Utilité : Aider à anticiper les fluctuations du marché pour la prise de décision.
- **Distribution des sentiments (*Sentiment Analysis*)** :
  - La proportion des tweets à tonalité négative, neutre ou positive.
  - Utilité : Mesurer l'humeur globale des utilisateurs concernant une entreprise ou un sujet donné.
- **Nombre total de followers (*Maximum Followers*)** :
  - Cet indicateur met en évidence le potentiel de portée des tweets.
  - Utilité : Identifier les influenceurs clés.
- **Activité des utilisateurs (*Count of Username*)** :
  - Le nombre total d'utilisateurs qui ont publié des tweets.
  - Utilité : Analyser l'activité sociale et l'engagement sur les plateformes.
- **Nombre total de tweets par utilisateur (*Total Tweets by User*)** :
  - Cet indicateur permet de suivre l'activité des utilisateurs clés.
  - Utilité : Identifier les contributeurs les plus actifs et leur impact.

Ces KPIs permettent de fournir une vision claire des performances financières et sociales, ainsi que des relations entre les sentiments exprimés sur les réseaux sociaux et les variations des prix des actions.

CHAPITRE 3

RÉALISATION DU PROJET

3.1 Architecture technique

L'architecture utilise **Kafka** pour l'ingestion de données provenant de *Yahoo Finance* et *Reddit*. Les données sont traitées par **Apache Spark**, puis stockées dans **Elasticsearch**. Les modèles de prédiction sont entraînés et suivis avec **MLflow**, et les visualisations sont effectuées avec **Kibana**. Les tâches sont orchestrées par **Apache Airflow**, et les données volumineuses sont gérées avec **Hadoop** dans un environnement conteneurisé.

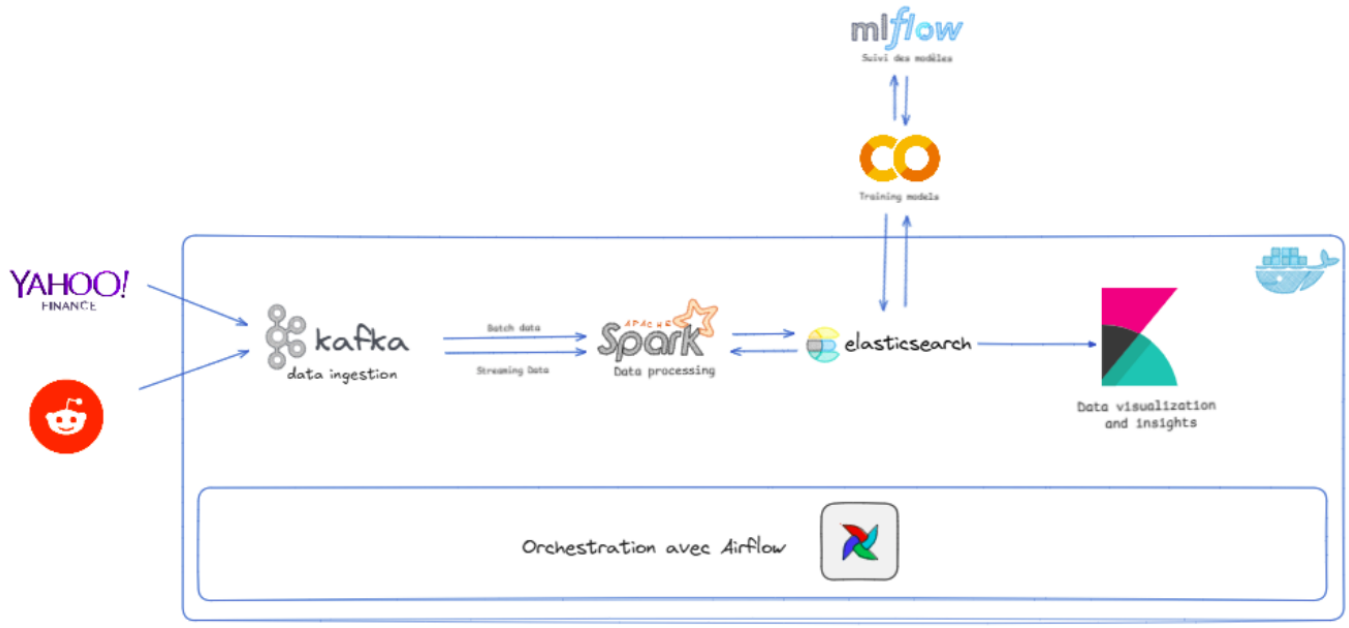

FIGURE 3.1 – Architecture technique

3.2 Collecte et ingestion des données

La collecte des données s'est déroulée en deux volets principaux :

- **Données Reddit** : Extraction des discussions relatives à Tesla à partir de Reddit en utilisant une technique de scraping. Cette collecte permet d'analyser les échanges sur divers aspects liés à Tesla, incluant les événements marquants, les réactions du public, ainsi que les opinions et débats concernant les produits et les initiatives de la société.

Date	Subreddit	Followers	Title	Upvotes	Comments
01-01-2025	r/AbruptChaos	2.4M	Tesla suddenly explodes outside Trump Hotel in Las Vegas in in	13000	785
01-01-2025	r/unusual_whales	140K	Video shows Tesla, $TSLA, Cybertruck explosion at the Trump H	11000	1400
04-01-2025	r/worldnews	44M	Forbes: Elon Musk Pushes For Britain's King Charles To Dissolve	31000	3000
05-01-2025	r/mildlyinfuriating	9M	The line to this Tesla charging station in Sweden.	27000	3200
02-01-2025	r/news	29M	Driver of Tesla Cybertruck in Las Vegas blast identified as US arn	27000	3200
01-01-2025	r/technology	18M	Tesla replaced laid off US workers with foreign workers using H-	37000	1500
02-01-2025	r/technology	18M	Tesla reports 1.1% sales drop for 2024, first annual decline in at l	20000	1700
02-01-2025	r/news	29M	Man who died in Tesla in Las Vegas 'suffered gun shot wound' b	16000	1900
20-12-2024	r/technology	18M	Tesla recalls 700,000 vehicles over tire pressure warning failure	31000	1600
21-12-2024	r/MarkMyWords	146K	MMW: Elon Musk and DJT will have a major fallout closer to the	13000	2600
10-01-2025	r/RealTesla	106K	25% of Americans Avoiding Tesla Tech Because of Elon Musk	16000	956
10-01-2025	r/FluentInFinance	534K	I used to respect Musk for being an innovator...	7900	1800
01-01-2025	r/technology	18M	Tesla Is Secretly Recalling Cybertruck Batteries	19000	858
05-01-2025	r/CyberStuck	196K	It's Impossible to Get Rid of a Tesla Cybertruck. I Want to Cut My	12000	1000
03-01-2025	r/UFOs	3.1M	Drones in the U.S. are from China and have gravitational propuls	6600	2500
27-12-2024	r/csMajors	321K	Elon laid off Tesla employees and requested H1B workers	15000	903
02-01-2025	r/WhitePeopleTwitter	3.1M	The guy who died in the Tesla Cybertruck explosion was a Trump	11000	977
12-12-2024	r/technology	18M	Trump transition wants to scrap crash reporting requirement op	15000	849
27-12-2024	r/clevercomebacks	2.5M	Now he's trying to justify it with Tesla's history	8200	1200
26-12-2024	r/politics	8.7M	"Dire shortage": Elon Musk sparks MAGA backlash after calling f	9700	868
01-01-2025	r/vegas	231K	Security footage of Tesla Cybertruck Exploding in front of Trump	5600	1200
10-01-2025	r/NoShitSherlock	81K	25% of Americans Avoiding Tesla Tech Because of Elon Musk	8300	639
01-01-2025	r/technology	18M	The Tesla Cybertruck that exploded and the New Orleans attack	7700	658
10-01-2025	r/mildlyinteresting	24M	This Tesla with a sticker distancing the owner from Elon Musk.	7100	614
01-01-2025	r/therewasanattempt	7.2M	to own a tesla cybertruck	7800	620
28-12-2024	r/Autobody	85K	Is my Tesla totaled?	4000	1800
02-01-2025	r/news	29M	Musk donated $108 million in Tesla shares to unnamed charities	7000	527
20-12-2024	r/CyberStuck	196K	Tesla is recalling 700k vehicles, including all Cybertrucks...	9000	485
05-01-2025	r/RealTesla	106K	Tesla as a rental... I couldn't make a worse car if I tried.	3700	1200
20-12-2024	r/europe	8.2M	Tesla Sales Are Tanking In Europe	4300	1100
12-12-2024	r/wallstreetbets	17M	What it feels like shorting Tesla now...	5300	859
12-12-2024	r/energy	205K	Musk and Trump: A modern-day Teapot Dome scandal waiting t	4500	1000
26-12-2024	r/CyberStuck	196K	Tesla rolled out a software update for Christmas and it bricked s	9900	407

FIGURE 3.2 – Extrait des données Reddit

Chaque entrée est caractérisée par plusieurs colonnes :

— **Date** : Date de publication du subreddit.
— **Subreddit** : Indique le nom de la communauté Reddit (ex. : r/technology) où la publication a été postée.
— **Followers** : Le nombre d'abonnés au subreddit correspondant.
— **Title** : Titre ou contenu principal du post.
— **Upvotes** : Nombre de votes positifs (upvotes) obtenus par la publication.
— **Comments** : Nombre de commentaires associés à la publication.

- **Données financières** : Utilisation de l'API Yahoo Finance pour récupérer les séries temporelles des actions TESLA, incluant le prix d'ouverture, de fermeture, le volume et les variations journalières.

Date	Close	High	Low	Open	Volume	Stock Name
12/2/2024	357.089996	360	351.149994	352.380005	77986500	TSLA
12/3/2024	351.420013	355.690002	348.200012	351.799988	58267200	TSLA
12/4/2024	357.929993	358.100006	348.600006	353	50810900	TSLA
12/5/2024	369.48999	375.429993	359.5	359.869995	81403600	TSLA
12/6/2024	389.220001	389.48999	370.799988	377.420013	81455800	TSLA
12/9/2024	389.790009	404.799988	378.01001	397.609985	96359200	TSLA
12/10/2024	400.98999	409.730011	390.850006	392.679993	97563600	TSLA
12/11/2024	424.769989	424.880005	402.380005	409.700012	104287600	TSLA
12/12/2024	418.100006	429.299988	415	424.839996	87752200	TSLA
12/13/2024	436.230011	436.299988	415.709991	420	89000200	TSLA
12/16/2024	463.019989	463.190002	436.149994	441.089996	114083800	TSLA
12/17/2024	479.859985	483.98999	457.51001	475.899994	131223000	TSLA
12/18/2024	440.130005	488.540009	427.01001	466.5	149340800	TSLA
12/19/2024	436.170013	456.359985	420.019989	451.880005	118566100	TSLA
12/20/2024	421.059998	447.079987	417.640015	425.51001	132216200	TSLA
12/23/2024	430.600006	434.51001	415.410004	431	72698100	TSLA
12/24/2024	462.279999	462.779999	435.140015	435.899994	59551800	TSLA
12/26/2024	454.130005	465.329987	451.019989	465.160004	76366400	TSLA

FIGURE 3.3 – Extrait des données financières de TSLA

Chaque entrée du dataset est caractérisée par plusieurs colonnes :

— **Date** : Indique la date de l'enregistrement des données concernant l'action Tesla (TSLA).
— **Open** : Prix d'ouverture de l'action Tesla à la date indiquée.
— **Close** : Prix de clôture de l'action Tesla à la fin de la journée.
— **High** : Prix le plus élevé atteint par l'action Tesla au cours de la journée.
— **Low** : Prix le plus bas atteint par l'action Tesla au cours de la journée.
— **Volume** : Volume total des transactions sur l'action Tesla pour la journée concernée.
— **Stock Name** : Nom de l'action, ici *Tesla (TSLA)*, pour identifier l'actif analysé.

3.3 Pipeline de traitement des données Reddit et Yahoo Finance

Ce projet met en œuvre un pipeline de traitement des flux de données provenant de **Reddit** et de **Yahoo Finance**, en utilisant **Kafka** pour la gestion des flux. Voici les différentes étapes et transformations effectuées :

3.3.1 Construction du producteur et du consommateur Kafka

- Les données brutes sont publiées dans des **topics Kafka** spécifiques à chaque type de source :
  — `topic tesla_twitter_data` : pour les publications Reddit.
  — `topic tesla_stock_prices` : pour les données financières.
- Un producteur Kafka injecte les données dans ces topics, et un consommateur les récupère pour les étapes de traitement.

3.3.2 Nettoyage et prétraitement des données issues de Reddit et Yahoo Finance

- Suppression des doublons et des données manquantes pour garantir la qualité des flux.
- Nettoyage et transformation des données de Reddit :

  - Suppression des caractères indésirables.
  - Normalisation des noms de **subreddits** en minuscules.
  - Conversion des champs de **followers** en valeurs numériques réelles (ex. : `2M` → `2 000 000`).
- Transformation des données financières :
  - Conversion des données JSON en colonnes structurées avec **Spark Streaming**.
  - Utilisation de **fenêtres temporelles** pour des agrégations comme les **prix moyens des actions** (ex. : moyennes par minute).

3.4 Analyse des sentiments

L'analyse des sentiments a été réalisée sur les données textuelles de Reddit à l'aide de **VADER**, qui calcule la polarité des subreddits grâce à son lexique intégré.

Un **indice d'influence sociale** a été obtenu en multipliant la polarité des subreddits par le nombre de followers des auteurs, fournissant une mesure pondérée de l'impact.

3.5 Fusion et préparation des données pour l'entrainement

Les données provenant de Reddit et des prix des actions de Tesla ont été fusionnées à l'aide du champ temporel. La dataset d'entrainement incluent : les prix des actions et les indices d'influence sociale. Cette fusion a permis de relier les sentiments exprimés sur les réseaux sociaux aux variations des prix des actions.

3.6 Entraînement du modèle et Prédiction

Pour la structure du modèle, nous avons opté pour un réseau LSTM afin de capturer la dépendance à long terme entre les données. De plus, nous avons ajouté un mécanisme d'attention pour permettre au modèle de se concentrer sur les éléments les plus importants de la séquence.

Chaque semaine, à l'aide de la technologie d'orchestration Airflow, l'entraînement du modèle sera effectué et les résultats seront ajoutés à MLflow, permettant de suivre l'évolution des entraînements. Une fois le modèle entraîné et enregistré dans MLflow, un script de prédiction sera déclenché pour estimer les prix des actions des sept jours suivants en utilisant ce dernier modèle entraîné.

3.7 Visualisation et suivi en temps réel

Un tableau de bord interactif a été développé pour suivre les tendances des actions Tesla, visualiser les analyses de sentiments de Reddit et présenter les prédictions des modèles. Alimenté par Kafka et Elasticsearch, il assure une mise à jour continue des données et des prédictions.

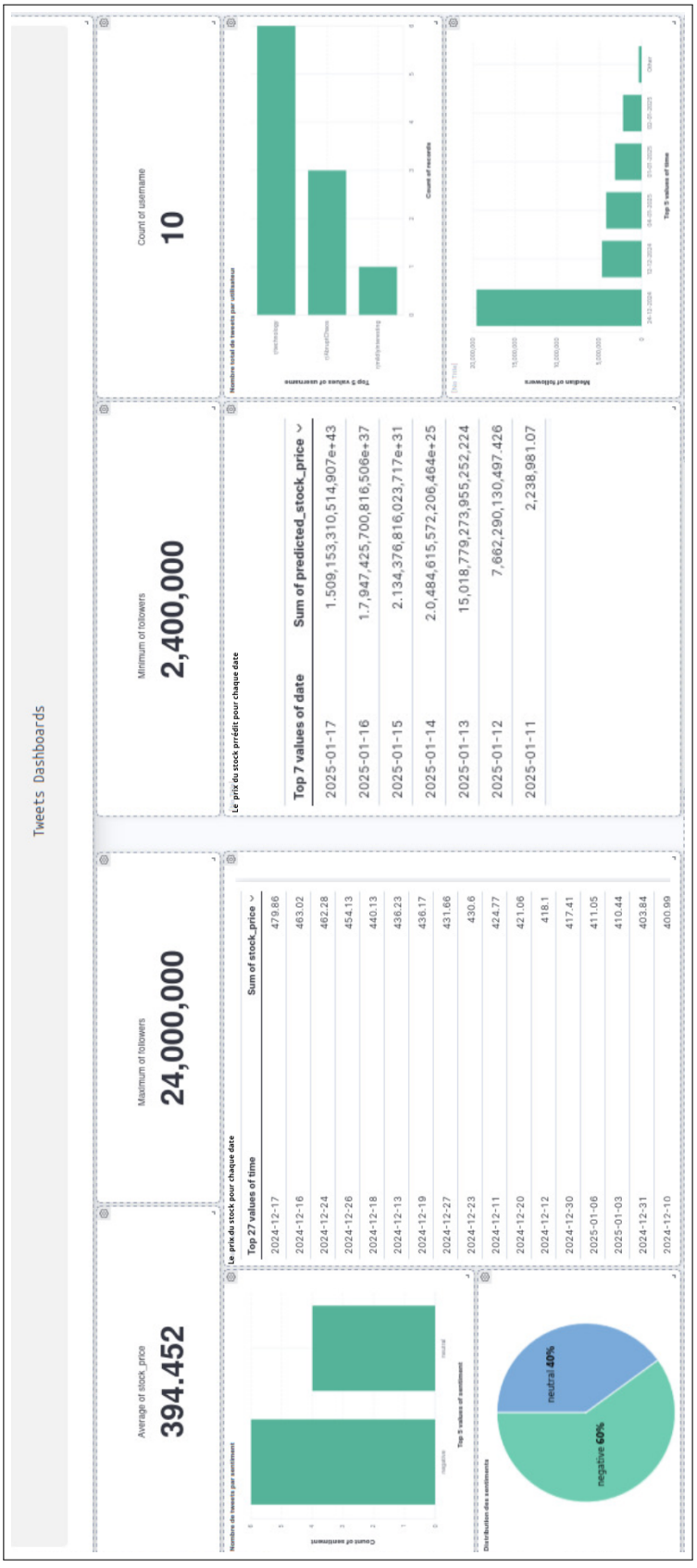

FIGURE 3.4 – Vue globale du dashboard

Métriques Globales

- **Moyenne des prix des actions** : La moyenne des prix des actions est de 394,452.
- **Maximum des abonnés** : Le nombre maximum d'abonnés pour un utilisateur est de 24 millions.
- **Minimum des abonnés** : Le nombre minimum d'abonnés pour un utilisateur est de 2,4 millions.
- **Nombre d'utilisateurs uniques** : Le nombre total d'utilisateurs pris en compte est de 10.

Graphiques et Analyses

- **Nombre de tweets par sentiment** : Un graphique en barres affiche la répartition des tweets en catégories "négatifs" ou "neutres", montrant une majorité de tweets négatifs.
- **Distribution des sentiments** : Un diagramme circulaire montre que 60% des tweets sont négatifs et 40% sont neutres.
- **Nombre total de tweets par utilisateur** : Un graphique en barres présente les utilisateurs les plus actifs (ex. : "technology", "otAngelChess") et le nombre de tweets qu'ils ont publiés.

Analyse des Prix des Actions

- **Prix des actions pour chaque date** : Un tableau présente les valeurs des prix des actions pour différentes dates, classées par ordre décroissant.
- **Prix des actions prédites pour chaque date** : Un tableau affiche les valeurs de prix prédites pour différentes dates. Certaines valeurs sont extrêmement élevées, probablement dues à des erreurs ou anomalies dans les données.

Analyse Temporelle des Abonnés

Graphique des abonnés par date : Un graphique en barres montre le nombre maximal d'abonnés pour les dates clés, permettant de visualiser les variations au fil du temps.

Analyse Globale

Ce tableau de bord offre une vue d'ensemble complète et structurée des tweets et des données de stock. Il inclut les sentiments des tweets, la répartition des utilisateurs, et l'évolution des prix des actions. Cela permet de détecter les tendances, les anomalies (comme les prix prédites anormalement élevés), et d'extraire des informations pertinentes à partir des données disponibles.

CONCLUSION

Ce projet a démontré le potentiel des technologies Big Data et de l'analyse des sentiments pour fournir des insights stratégiques en temps réel dans le domaine financier. En combinant des outils comme Apache Kafka, Spark, et Elasticsearch avec des modèles prédictifs, il a permis de lier les données sociales et financières pour anticiper les variations des actions de Tesla.

Kafka a facilité l'ingestion en temps réel, Spark le traitement des données volumineuses, et Elasticsearch la recherche et la visualisation des résultats. Les modèles LSTM permet de capturer les dynamiques complexes des marchés financiers influencés par les sentiments.

Le projet souligne cependant des pistes d'amélioration : élargir les sources de données, intégrer des facteurs externes (politiques monétaires, événements géopolitiques), et optimiser davantage les modèles via des approches hybrides. Ces évolutions pourraient enrichir les prédictions et réduire les erreurs.

Ces avancées ouvrent la voie à des applications plus poussées dans la gestion financière, comme l'optimisation des portefeuilles, la prévision des risques, et l'ajustement des stratégies d'investissement en temps réel.