

Projet DataEngineer

Réalisé par : Mezghani Mariem



Plan

Introduction

Objectif

Architecture du projet

Réalisation

Conclusion

Introduction

Ce projet appartient au domaine de la gestion logistique, de l'analyse de données en temps réel et du partage de vélos. Il s'inscrit également dans le cadre plus large de la mobilité urbaine durable, mettant en œuvre des technologies de Big Data, de streaming, et de visualisation pour améliorer l'efficacité opérationnelle des services de partage de vélos.

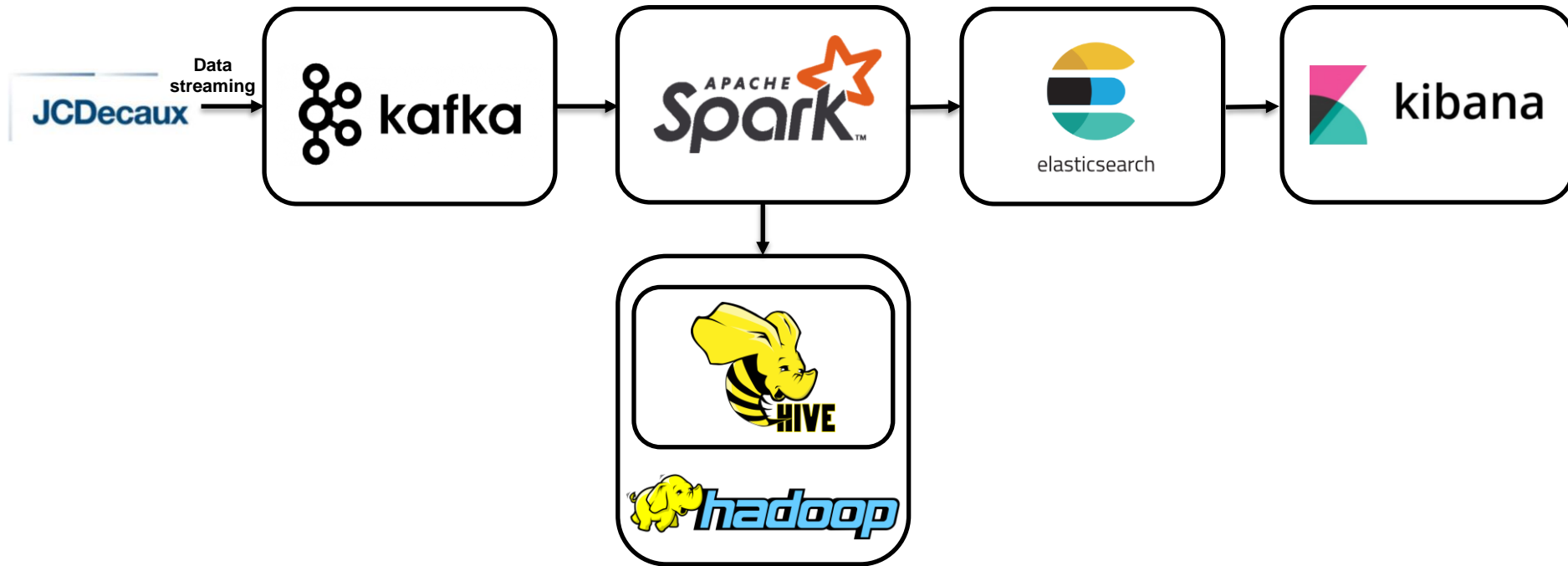


Objectif



Ce projet répond à un besoin crucial dans le domaine du partage de vélos, en comblant des lacunes opérationnelles et en relevant les défis spécifiques rencontrés par les utilisateurs et les gestionnaires de stations. Tout d'abord, il adresse le besoin fondamental des cyclistes qui cherchent à optimiser leurs déplacements. En fournissant une visibilité en temps réel sur la disponibilité des vélos et les emplacements des stations, le projet répond directement à la nécessité des utilisateurs de planifier efficacement leurs trajets et de trouver des vélos rapidement.

Architecture du projet



Réalisation

La source de données pour ce projet provient de **JCDecaux**, une entreprise internationale spécialisée dans le vélo en libre-service et la publicité extérieure. JCDecaux exploite des systèmes de vélos en libre-service dans de nombreuses villes à travers le monde, fournissant ainsi une source riche et dynamique d'informations sur la disponibilité des vélos et l'état des stations.



Réalisation

Le pipeline du projet se déploie en plusieurs étapes, mettant en œuvre un processus complet de collecte, traitement et visualisation des données.

- **Collecte des Données en Temps Réel avec Kafka :** Le processus débute par la collecte de données en temps réel à l'aide de Kafka, un outil puissant de streaming. Les informations sur la disponibilité des vélos et l'état des stations sont capturées de manière continue.
- **Analyse des Données avec Spark :** Les données collectées sont ensuite acheminées vers Spark, où un programme Python spécialement conçu analyse, formate et nettoie les données. Cette étape permet de préparer les informations pour une intégration ultérieure.



Réalisation

Une fois les données préparées, ces informations sont stockées soit dans Elasticsearch, soit dans Hive Hadoop.

- **Injection des Données dans Elasticsearch:** Spark injecte les informations dans le moteur Elasticsearch. Cela offre une solution de stockage efficace et flexible pour les données traitées, prêtes à être exploitées pour des analyses ultérieures.
- **Stockage Historique dans Hive Hadoop :** Pour assurer la rétention à long terme des données historiques, le pipeline intègre le stockage dans Hive Hadoop. Hive fournit une interface SQL pour interagir avec Hadoop, permettant ainsi le stockage structuré des données historiques.



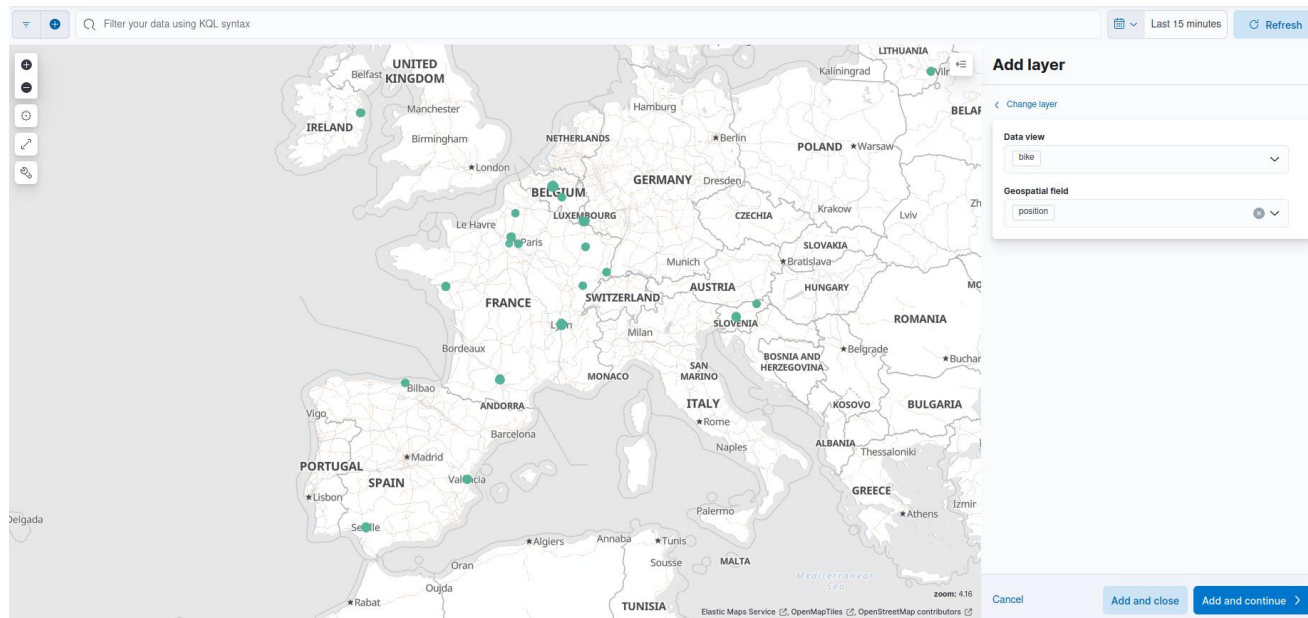
Réalisation

- **Visualisation avec Kibana :** Kibana entre en scène pour créer des tableaux de bord interactifs. Les données stockées dans Elasticsearch sont utilisées pour générer des visualisations dynamiques, offrant une vue en temps réel de l'emplacement des stations, du nombre de vélos disponibles, et d'autres métriques pertinentes.



Résultats

Cette carte générée par Kibana affiche géographiquement les vélos disponibles :











Résultats

Voici la table générée sur Elasticsearch :

Columns

Sort fields

	address	available_bikes	position	available_bike_sta...	banking	bike_stands	contract_name	numbers	status	timestamps
 <input type="checkbox"/>	26 RUE DE LA THIBAUDIERE	4	POINT (4.845048 45.750344)	24	true	28	lyon	7,053	OPEN	2023-11-23 15:04:45
 <input type="checkbox"/>	57 RUE DU PRINTEMPS	1	POINT (1.443571 43.614574)	16	true	17	toulouse	93	OPEN	2023-11-23 15:01:17
 <input type="checkbox"/>	39 BIS AV DE LOMBEZ	9	POINT (1.418503 43.595503)	16	true	25	toulouse	139	OPEN	2023-11-23 15:05:02
 <input type="checkbox"/>	Pio XII - Campanar	0	POINT (-0.39343615667867 39.4814301267876)	20	true	20	valence	172	OPEN	2023-11-23 15:02:49
 <input type="checkbox"/>	PISCINE / ZWEMBAD - RUE DE LOMBARTZYDE / ...	5	POINT (4.376931 50.893158)	9	false	14	bruxelles	283	OPEN	2023-11-23 15:00:23
 <input type="checkbox"/>	Vestbygata 61, Lillestrom	0	POINT (11.040774 59.964446)	20	true	20	lillestrom	3	CLOSED	2023-11-23 15:00:43
 <input type="checkbox"/>	ALLEE CHARLES DE FITTE / FACE JARDIN RAYMOND VI	28	POINT (1.429521 43.599797)	5	true	33	toulouse	81	OPEN	2023-11-23 15:03:27
 <input type="checkbox"/>	A côté de l'église	1	POINT (4.887311 45.749565)	24	true	25	lyon	3,012	OPEN	2023-11-23 15:04:43

Rows per page: 100

<

1

2

3

4

5

>

Résultats

Voici la table générée sur Hive :

```
marlen@marlen:~$ hive
Logging initialized using configuration in jar:file:/home/marlen/Downloads/apache-hive-2.3.9-bin/lib/hive-common-2.3.9.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> SELECT * FROM bikes_stations.bikes_stations LIMIT 20;
OK
100 valence false 20 3 17 Blasco Ibañez - Mestre Ripoll OPEN {"lat":39.4716021114468,"lng":-0.33824098596173} 2023-12-19 19:43:08
21 vilnius true 12 0 0 J. Tumo-Vaižganto g. - Lukiškių g. sankryža CLOSED {"lat":54.69207484,"lng":25.2711658} 2023-11-21 10:56:11
50015 cergy-pontoise false 18 8 10 PLACE DES LINANDES - 95000 CERGY OPEN {"lat":49.043284,"lng":2.068934} 2023-12-19 19:41:03
130 luxembourg false 20 5 14 Rue Gaston Thorn 4 OPEN {"lat":49.622397,"lng":6.036232} 2023-12-19 19:40:39
79 nantes false 15 6 9 15, quai de Malakoff OPEN {"lat":47.213395,"lng":-1.536954} 2023-12-19 19:41:09
7 santander true 30 15 15 Facultad de Derecho (Aprox. Facultad de Ciencias) OPEN {"lat":43.4701661089386,"lng":-3.80649947753026} 2023-12-19 19:44:33
191 toulouse true 18 13 5 395 ROUTE DE SAINT SIMON OPEN {"lat":43.590046,"lng":1.415246} 2023-12-19 19:38:50
97 toulouse true 21 2 19 FACE AU 66 RUE RAYMOND IV OPEN {"lat":43.612475,"lng":1.451558} 2023-12-19 19:36:11
205 seville false 20 19 1 AVENIDA KANSAS CITY - Aprox. Pza el Tato OPEN {"lat":37.3999693,"lng":-5.961848969} 2023-12-19 19:38:11
8009 lyon true 20 17 3 Boulevard Jean XXIII - Angle de la rue Mermoz OPEN {"lat":45.73648,"lng":4.869789} 2023-12-19 19:44:01
3094 lyon true 38 38 0 74 RUE DE LA VILLETTE OPEN {"lat":45.757808,"lng":4.861512} 2023-12-19 19:44:09
104 toulouse true 18 9 9 219 AV DE MURET OPEN {"lat":43.589126,"lng":1.431449} 2023-12-19 19:42:42
6012 lyon false 10 2 8 Angle rue Vendôme OPEN {"lat":45.771794,"lng":4.84459} 2023-12-19 19:44:41
15 amiens true 20 13 7 Rue des Jacobins OPEN {"lat":49.8909435451942,"lng":2.300843419403305} 2023-12-19 19:35:25
13 toulouse true 15 15 0 1 RUE DE L ESQUITE OPEN {"lat":43.606146,"lng":1.44135} 2023-12-19 19:42:41
129 toulouse true 16 12 4 2 BD DEODAT DE SEVERAC OPEN {"lat":43.585732,"lng":1.427853} 2023-12-19 19:39:03
52 toulouse true 21 6 15 39 BD ARMAND DUPORTAL OPEN {"lat":43.608571,"lng":1.434762} 2023-12-19 19:41:53
26 namur true 15 8 7 026 - ARSENAL - BOULEVARD FRERE ORBAN OPEN {"lat":50.464457,"lng":4.85644} 2023-12-19 19:40:03
249 bruxelles false 25 11 14 SAINT-JOB/SINT JOB - PLACE SAINT JOB/SINT-JOB PLEIN OPEN {"lat":50.794186,"lng":4.366336} 2023-12-19 19:43:56
49 toulouse true 24 5 19 DEVANT 12 PLACE DU PONT NEUF OPEN {"lat":43.599598,"lng":1.440942} 2023-12-19 19:39:39
Time taken: 1.738 seconds, Fetched: 20 row(s)
hive>
```

Conclusion

En conclusion, ce projet de Logistique Vélo et Analyse en Temps Réel offre une solution complète et innovante pour répondre aux défis de la gestion du partage de vélos. En unifiant des technologies de pointe telles que Kafka, Spark, Elasticsearch, et Kibana, le pipeline garantit une gestion efficace des données en temps réel, permettant aux utilisateurs de localiser rapidement des vélos disponibles et aux gestionnaires de stations de maintenir une distribution optimale des ressources.