Agrégateur de flux RSS

DSIA-4301B Data Engineering 2

Antonin Grimal
Joaquim Nicolas
Sébastien Luong
Antoine Susini
Adam Lafkih
Islaan Muhammad
Yanis Tissot

Projet disponible sur GitHub : https://github.com/antoningr/agregateur_de_flux_RSS

Sommaire

Présentation du projet

- Partie back-end d’un agrégateur de flux RSS

  → Informations sur des sites Web

  → Syndication de contenu au format RSS

  → Récupération d’informations
- Approche de microservices

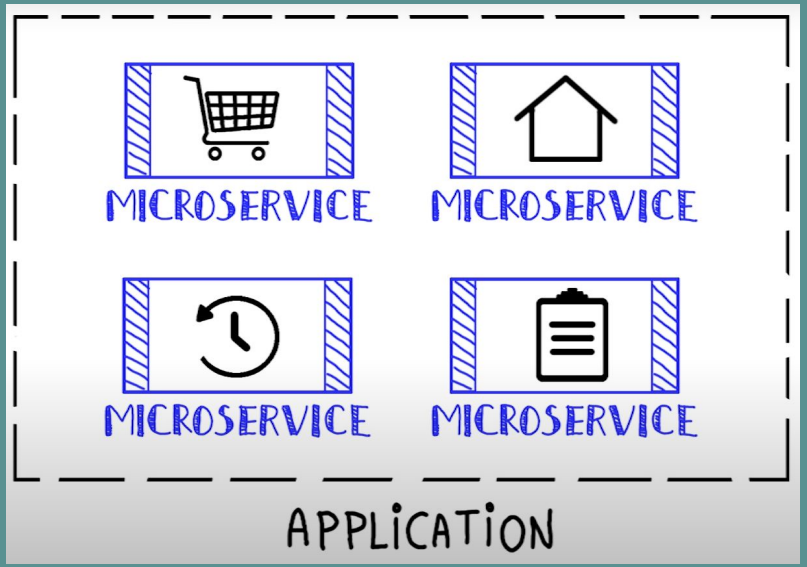
Présentation du projet

- Approche d’architecture logicielle

  → Découper en plusieurs modules fonctionnels

  → Partie spécifique et unique

  → Accessibles par le client via l’API du microservice



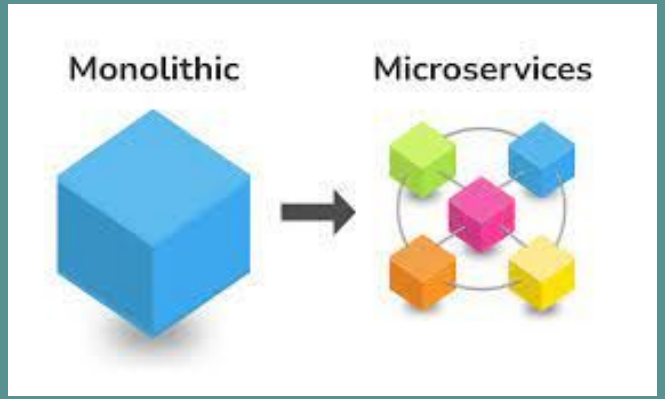
Présentation du projet

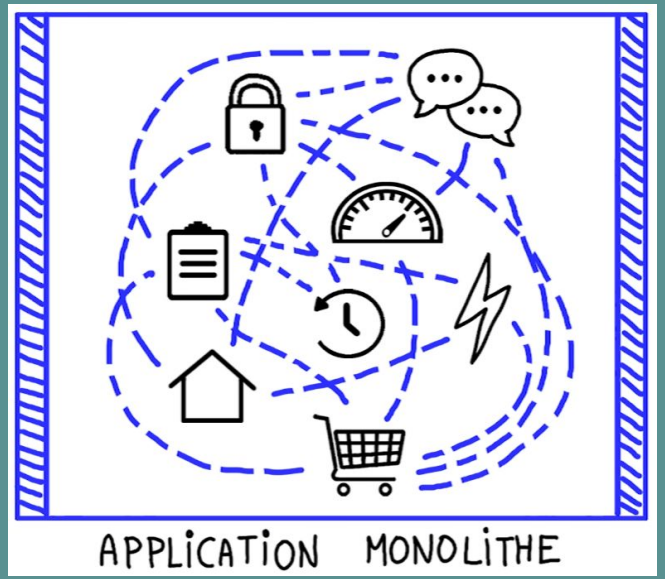
- Répondre aux problèmes de l'approche monolithe

  → Traiter toutes les demandes possibles

  → Répondre au maximum de cas d'usage

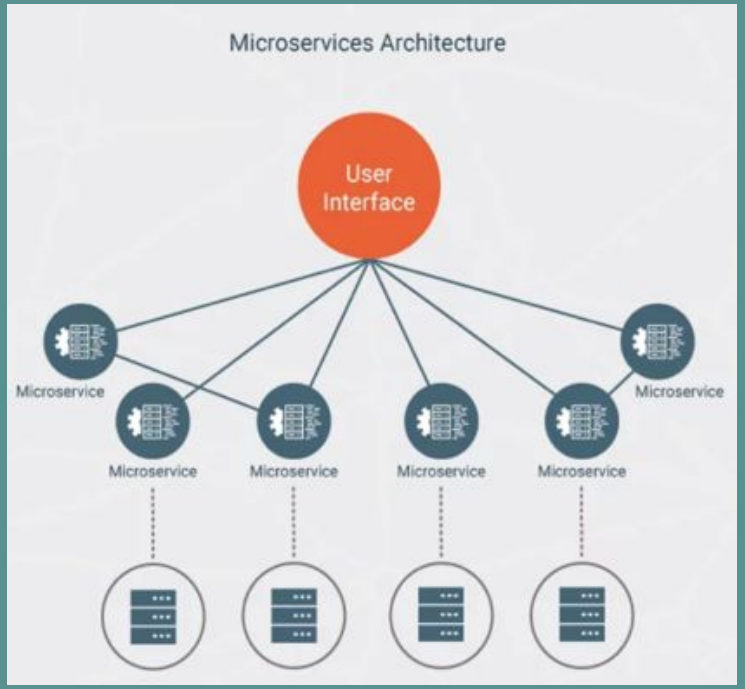
- Création d'interdépendances



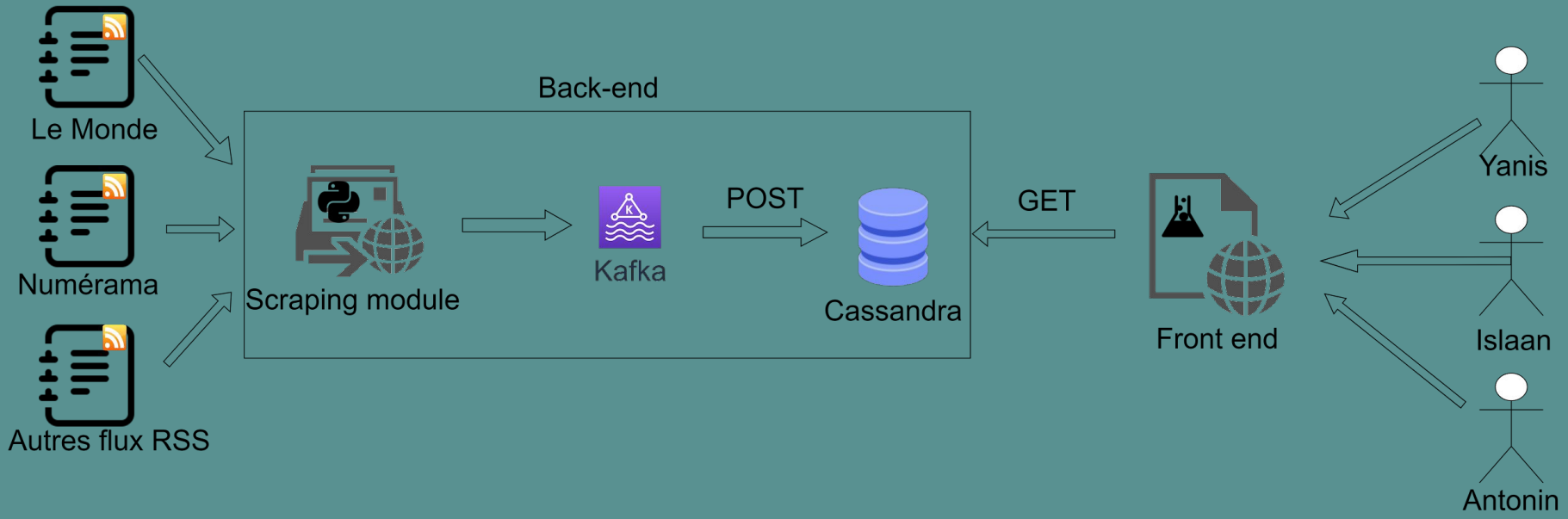


Présentation du projet

- Scalabilité
- Flexibilité
- Déploiement continu
- Sécurité



Architecture de l'agrégateur RSS



Description du modèle de données

- **Flux RSS**
  - Format de **données structuré**
  - Site web publie du contenu (articles de blog, actualités, ...)

- Deux entités : **Feed** et **Article**

- L'entité **Article** représente un article issu d'un flux RSS :
  - **feed_id**
  - **article_id**
  - **title**
  - **pubDate**
  - **description**
  - **link**

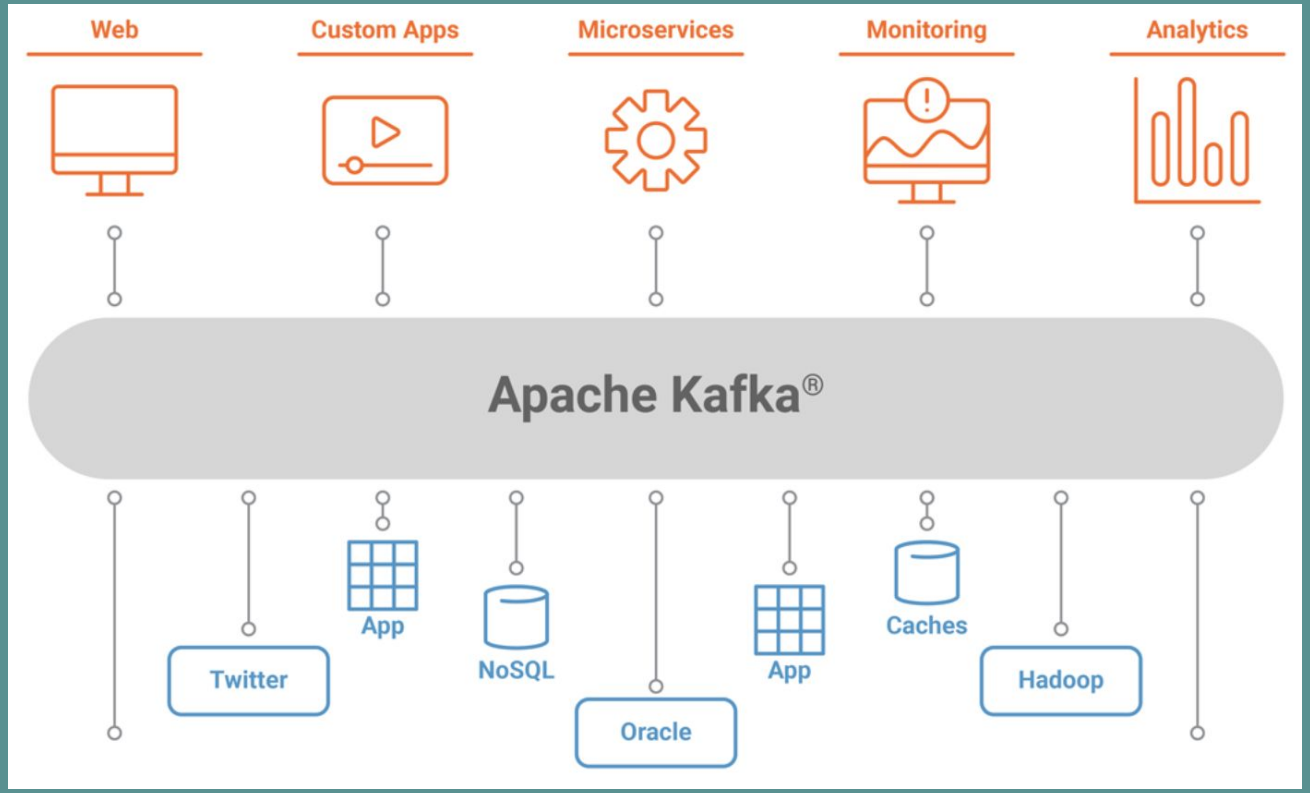
Exemple d'article


```
[
 {
  "feed_id":     "2ms1vWAxwtalfO9i1MxD3Zn2_P7xR3xp",
  "article_id":  "F-boNQaGhvWum01IWwo9A4ySnAB-12i4",
  "title":       "Xfce 4.18 Coming Soon and Offers Subtle Improvements",
  "pubDate":     "Wed, 08 Feb 2023 15:43:44 +0000",
  "description": "The Xfce team has announced the release date of the next iteration of
  "link":        "http://www.linux-magazine.co/News/Xfce-4.18-Coming-Soon"
 },
 {
  "feed_id": "8ZzejRTaQ5xI_nqQXWiI3SEVObXpC_5K",
  …
 },
 …
]
```

Scraping de données

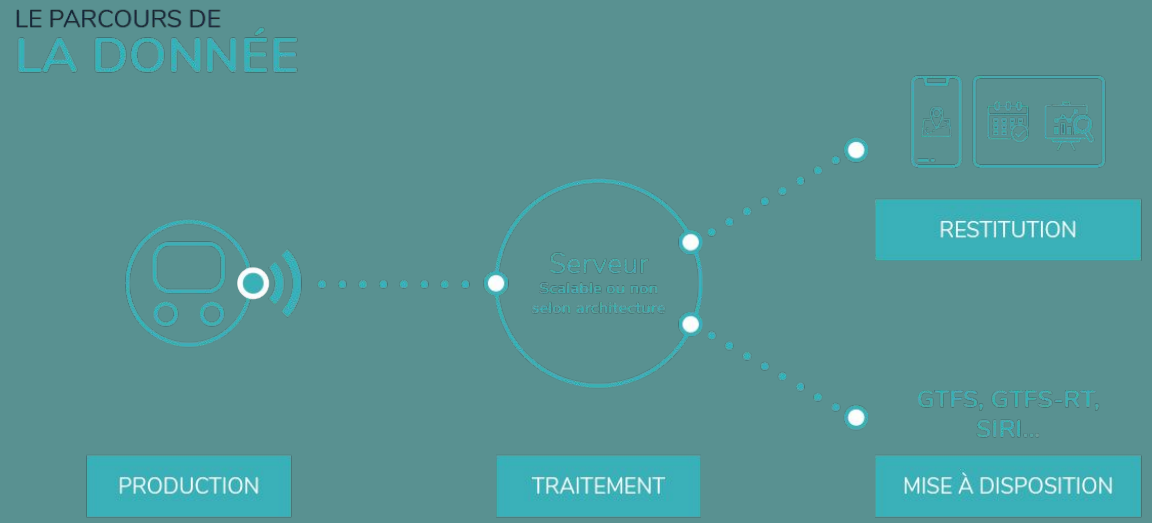
- Objectif : récupérer les articles de sites web d'actualité
- **Web Scraping** « technique qui permet l'extraction des données de pages web de façon automatique. »
  - Via la librairie Python **requests_html**

Streaming et gestion de flux (Kafka)



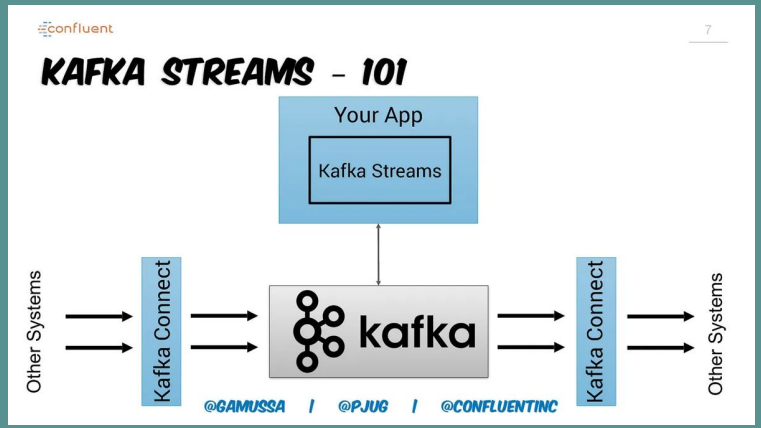
Gestion des flux en temps réels

- Collecter les flux RSS
- Stocker et traiter les flux
- Diffuser les flux agrégés



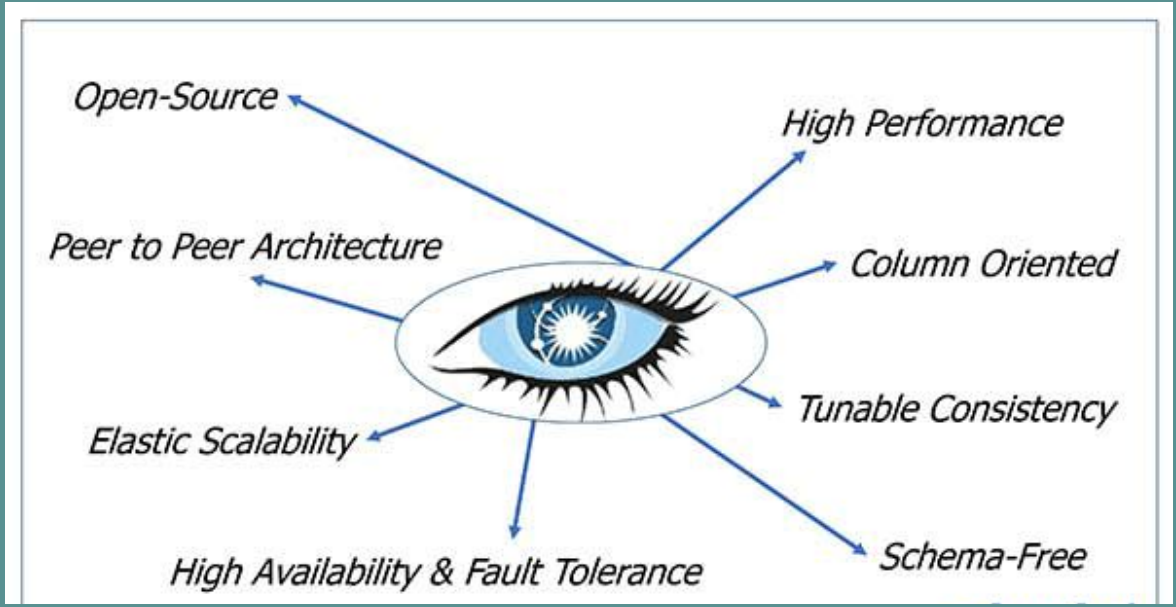
Interaction entre Kafka et les autres composants

- Utilisé en tant que middleware
- Flux RSS envoyé à Kafka par un scrapper de flux RSS
- Kafka stocke les flux dans des partitions puis les transmet à l'agrégateur
- L'agrégateur traite ensuite les flux reçus à partir de Kafka
- Interface utilisateur



Base de données (Cassandra)

Base de données (Cassandra)



Modèle des tables (Cassandra)

```
session.execute("CREATE KEYSPACE IF NOT EXISTS your_keyspace WITH REPLICATION = {'class' : 'SimpleStrategy', 'replication_factor' : 1}")
session.execute("USE your_keyspace")
session.execute("CREATE TABLE IF NOT EXISTS RSS (title text, pubdate text, link text, description text, PRIMARY KEY(feed_id, pubdate)) WITH CLUSTERING ORDER BY (pubdate DESC)")
```

PRIMARY KEY(feed_id, pubdate)

Partition key

Clustering key

Interaction avec les composants (Cassandra)

Stockage des messages de Kafka dans Cassandra

```
for _, row in data.iterrows():
    session.execute(f"INSERT INTO RSS (title, pubdate, link, description) VALUES ('{row['title']}', '{row['pubDate']}', '{row['link']}', '{row['description']}')")
```

Affichage des données avec Flask

Titre	Date de publication	Lien	Description
Entre l'Etat et les sociétés d'autoroutes, un absurde contentieux fiscal	2023-04-07 10:19:02+02:00	https://www.lemonde.fr/economie/article/2023/04/07/entre-l-etat-et-les-societes-d-autoroutes-un-lourd-contentieux-fiscal_6168636_3234.html	Quand les pouvoirs publics manquent d'argent pour les infrastructures, ils taxent les sociétés d'autoroutes, lesquelles négocient systématiquement des contreparties. Le point de litige : l'indexation de la taxe d'aménagement du territoire.
Space: What lies above the Cloud – Part II	2023-03-01 14:45:00+00:00	https://blog.ovhcloud.com/?p=24812	As we mentioned in our previous article on the matter, space has made a welcome comeback in the news. Long gone are the days of reduced funding and the shutting down of NASA projects. Once again, we can dream of conquering new gallactical frontiers. While in our first article we introduced several crucial points about … Space: What lies above the Cloud – Part II Read More » The post Space: What lies above the Cloud – Part II appeared first on OVHcloud Blog.
Ransomware targeting VMware ESXi	2023-02-03 16:10:29+00:00	https://blog.ovhcloud.com/?p=24513	A wave of attacks is currently targetting ESXi servers. No OVHcloud managed service are impacted by this attack however, since a lot of customers are using this operating system on their own servers, we provide this post as a reference in support to help them in their remediation. These attacks are detected globally. According to … Ransomware targeting VMware ESXi Read More » The post Ransomware targeting VMware ESXi appeared first on OVHcloud Blog.
Picasso : des expositions, des documentaires, un livre et un podcast pour le redécouvrir	2023-04-07 04:00:00+02:00	https://www.lemonde.fr/culture/article/2023/04/07/des-expositions-des-documentaires-un-livre-et-un-podcast-pour-redecouvrir-picasso_6168588_3246.html	Le cinquantième anniversaire de la disparition de l'artiste s'accompagne d'une riche offre culturelle. Le service Culture du « Monde » vous propose sa sélection.
Transformer les réseaux en catalyseurs d'activité avec l'hyper-résilience	2023-02-13 17:04:21+00:00	https://blog.ovhcloud.com/?p=24672	Pourquoi la confiance dans les réseaux est essentielle pour toutes les organisations Les pannes de réseau sont une source constante d'inquiétude pour les entreprises. Les temps d'arret causés par des pannes non planifiées se traduisent par des pertes financières et peuvent ainsi coûter très cher. Selon Gartner, leur coût total moyen par minute s'élève à … Transformer les réseaux en catalyseurs d'activité avec l'hyper-résilience Read More » The post Transformer les réseaux en catalyseurs d'activité avec l'hyper-résilience appeared first on OVHcloud Blog.

Démonstration

Conclusion

Notre outil :

- Architecture microservice
- Apache Kafka
- Cassandra

Avantages :

- Flexibilité
- Efficacité

Bibliographie

- https://cassandra.apache.org/_/index.html
- https://kafka.apache.org/

Projet disponible sur GitHub :

https://github.com/antoningr/agregateur_de_flux_RSS