ROMÂNIA
MINISTERUL APĂRĂRII NAŢIONALE
ACADEMIA TEHNICĂ MILITARĂ „FERDINAND I”
FACULTATEA DE SISTEME INFORMATICE ȘI SECURITATE CIBERNETICĂ

Specialize: Calculatoare și sisteme informatice pentru apărare și securitate națională

Dezvoltarea unui motor de indexare, clasificare și căutare pentru știri

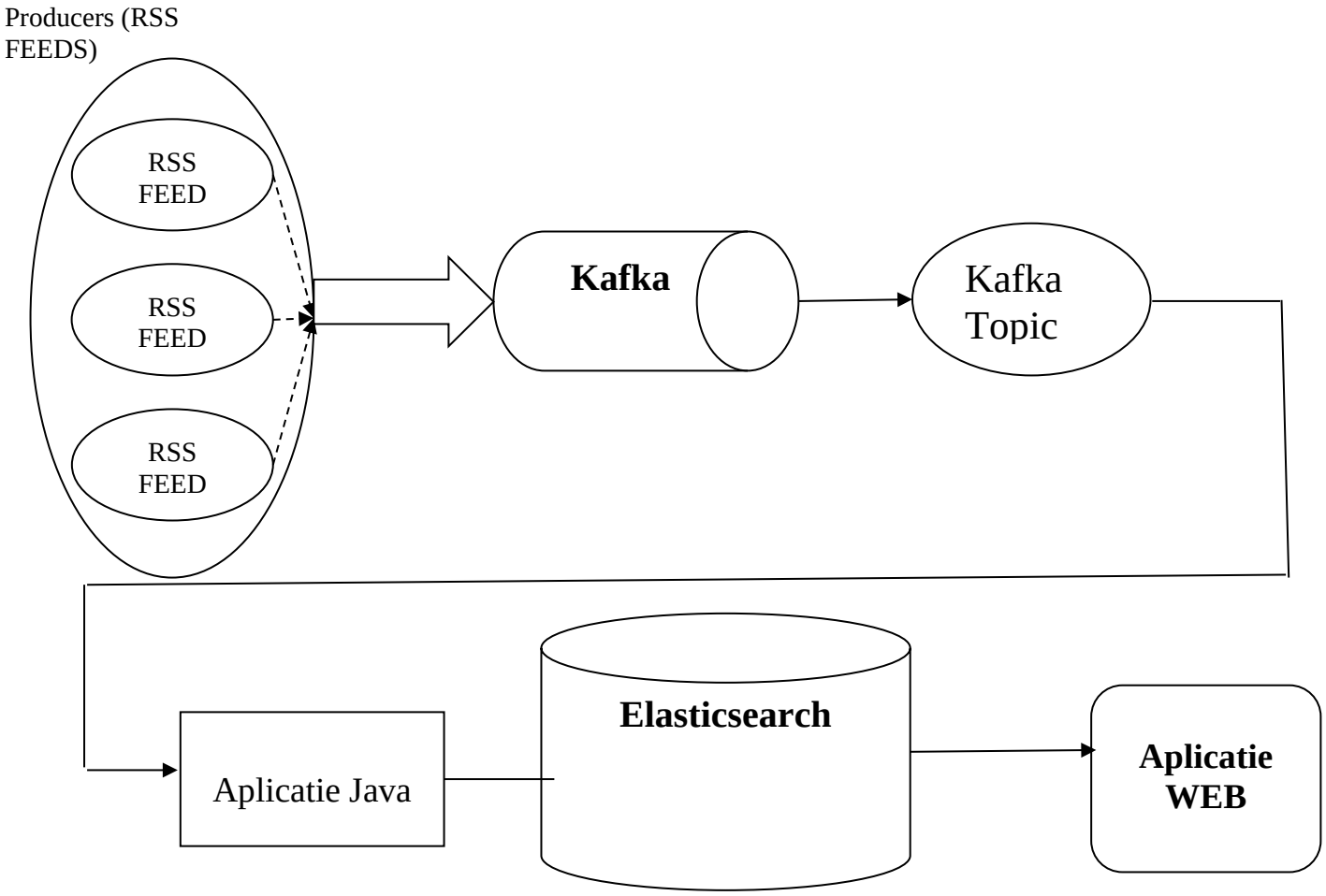
CONDUCĂTOR ŞTIINŢIFIC:
Ș.L. dr. ing. Cristian CHILIPIREA

STUDENT:
Sg. Ciprian-George Pesu

BUCUREŞTI
2021

Contents

Prezentare generala a arhitecturii proiectului



RSS FEEDS:

RSS este o familie de formate de fluxuri web, realizate în format XML și folosite pentru Websyndication. RSS este folosit (printre altele) pentru știri, weblog-uri și podcasting.

RSS înseamnă Really Simple Syndication si se referă la fișierele XML care se actualizează automat . Aceste informații sunt preluate de un cititor de fluxuri RSS al unui utilizator care convertește fișierele într-un format ușor de citit. Un flux RSS preia titlurile, rezumatele și notificările de actualizare și apoi face legătura cu articolele de pe pagina site-ului web preferat.

Acest conținut este distribuit în timp real, astfel încât rezultatele de top ale fluxului RSS să fie întotdeauna cele mai recente conținut publicat pentru un site web

Principalele fluxuri RSS indentificate pentru folosire in proiect :

- BBC : BBC News - Home XML Feed (bbci.co.uk)
- CNN : RSS (Really Simple Syndication) - CNN.com
- FOX : FOXNews.com RSS Feeds | Fox News

Din aceste fluxuri RSS putem identifica titlurile , rezumatele si linkurile catre stirile prezente in flux.

Exemplu de date extrase din fluxuri RSS:

{
"link": "https://www.cnn.com/style/article/oldest-whiskey-auction-style-trnd/index.html",
"description": "A historic whiskey, which could date back over 250 years, smashed auction estimates to sell for $110,000 on Wednesday.",
"source": "CNN.com - RSS Channel - App International Edition",
"title": "The world's oldest known whiskey sells for $110K",
"pubDate": "Thu, 01 Jul 2021 05:17:27 GMT"
}
{
"link": "https://www.cnn.com/2021/07/01/business/gap-store-closures-uk-ireland/index.html",
"description": "Gap will close all 81 of its stores in the United Kingdom and Ireland by the end of September and go fully online, as the brand adjusts to changes in shopping habits following the pandemic.",
"source": "CNN.com - RSS Channel - App International Edition",
"title": "Gap to close all stores in UK and Ireland ",
"pubDate": "Thu, 01 Jul 2021 09:27:27 GMT"
}

Extragerea cuvintelor cheie din titlu:

Titlul știrilor este necesar pentru recunoașterea conținutului articolului din pagina de știri. Dacă localizăm corect poziția titlului într-o pagină de știri, poziția articolului ar fi găsită cu ușurință deoarece pagina web este o listă de paragrafele precedate de titlu.

În plus, pentru o știre, rezumatul extras fluxul RSS descrie în detaliu subiectul știrii de unde se pot scoate mai multe cuvinte cheie pentru a identfica corect articolul in pagina .

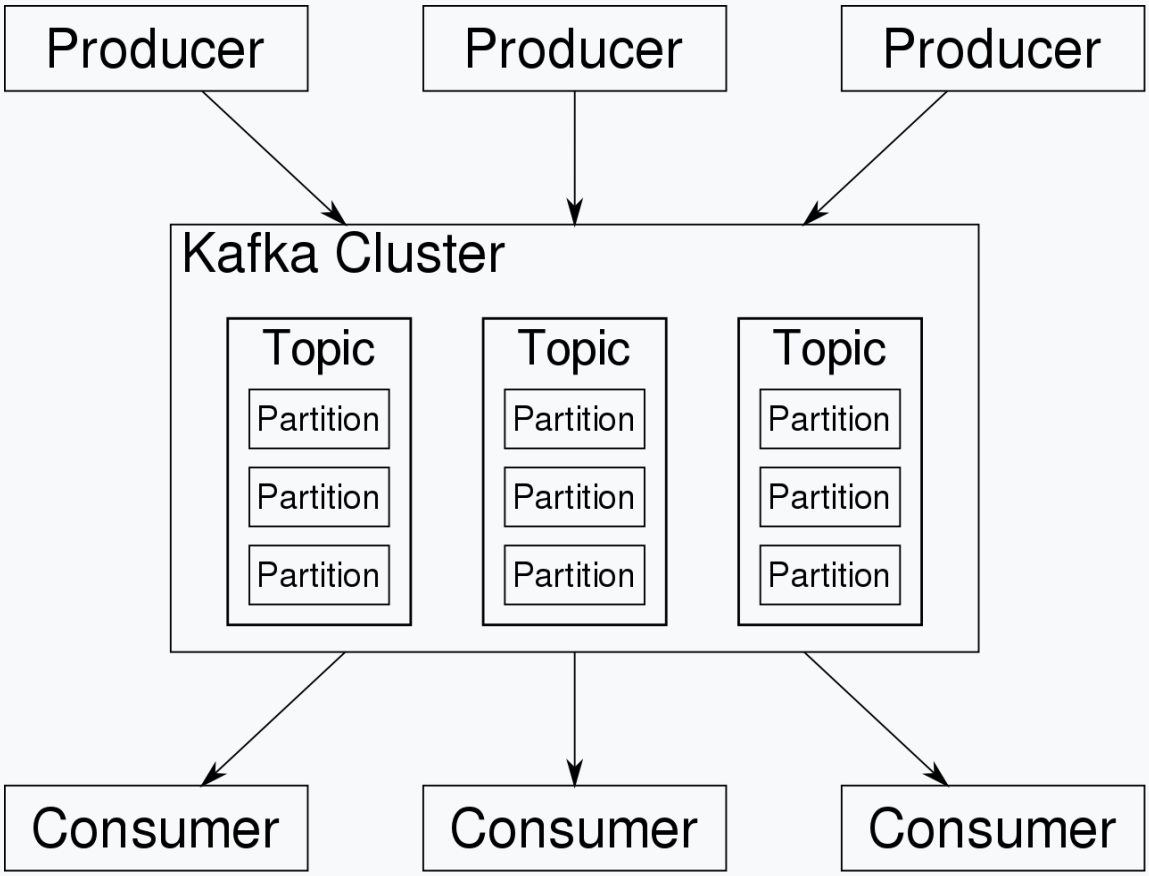
Extragerea continutului articolului din documentul HTML:

Un document HTML poate fi reprezentat ca o structură structura arborescenta. Cu ajutorul cuvintelor cheie extrase din titlul si rezumatul primit de la fluxul RSS putem identifica titlul si paragrafele ce alcatuiesc articolul in cadrul documentului HTM .

După extragerea continutului articolului, primim un paragraf cu din întregul al conținutului știrilor. De obicei, întregul conținut text al știrilor este o listă de paragrafe continue. Cu toate acestea, există reclame si imagini publicitare printre paragrafele articolului , acestea trebuiesc identificate si extrase .

Apache Kafka:

Apache Kafka este un framework pentru procesarea fluxului. Este o platformă software open-source dezvoltată de Apache Software Foundation, scrisă în Scala și Java. Proiectul își propune să ofere o platformă unificată, cu randament ridicat și cu latență redusă pentru gestionarea fluxurilor de date în timp real.



Principii de functionare Kafka :

- Topicele Kafka sunt întotdeauna multi-producătoare și multi-consumatoare: un topic poate avea zero, unul sau mulți producători care îi scriu evenimente, precum și zero, unul sau mulți consumatori care se abonează la aceste evenimente. Evenimentele dintr-un subiect pot fi citite ori de câte ori este necesar - spre deosebire de sistemele tradiționale de mesagerie, evenimentele nu sunt șterse după consum. În schimb, se poate defii cât timp Kafka trebuie să păstreze evenimentele printr-o setare de configurare pe topic, după care evenimentele vechi vor fi eliminate.

- Producatorii (in cazul nostru fluxurile RSS) in pun datele (stirile) in unul sau mai multe Topice (depinde de modul de implementare).

- Consumatorii (aplicatia care proceseaza stitile) extrage date din Topice si le trimite catre **Elasticsearch**

Există multe limbaje de programare care oferă biblioteci pentru Kafka (C , C++ , Java , C# , Pyhon ...).
Pentru acest proiect am ales **Java** deoarece este varinata oficiala de Kafka si nu necesita librarii neoficiale

Zookeeper:

Zookeeper este un software dezvoltat de Apache care acționează ca un serviciu centralizat și este utilizat pentru a menține datele de denumire și configurare și pentru a oferi sincronizare flexibilă și robustă în cadrul sistemelor distribuite. Zookeeper ține evidența stării nodurilor clusterului Kafka și, de asemenea, ține evidența topicelor, partițiilor etc.

Zookeeper permite mai multor clienți să efectueze citiri și scrieri simultane și acționează ca un serviciu de configurare partajat în cadrul sistemului. Protocolul

Zookeeper atomic broadcast (ZAB) este creierul întregului sistem, făcând posibil ca Zookeeper să acționeze ca un sistem de difuzare și să emită actualizări .

Elasticsearch:

Elasticsearch este un motor de căutare bazat pe biblioteca Lucene. Acesta oferă un motor de căutare full-text distribuit, capabil de multitenant, cu o interfață web HTTP și documente JSON fără schemă. Elasticsearch este dezvoltat în Java . Clienții oficiali sunt disponibili în Java, .NET (C #), PHP, Python, Apache Groovy, Ruby și multe alte limbi. Conform clasamentului DB-Engines, Elasticsearch este cel mai popular motor de căutare pentru întreprinderi.

Pentru acest proiect am ales limbajul **Java**

Kibana :

Kibana este o interfață de utilizator gratuită , care permite vizualizarea datelelor din Elasticsearch și navigarea în Elastic Stack.

Kibana o sa fie folosit in general ca un dashboard pentru statisitici si vizualizare a datelor din Elasticsearch

Bibliografie:

- Apache Kafka
- Kibana Guide [master] | Elastic
- Elastic Stack and Product Documentation | Elastic
- An Automatic Web News Article Contents Extraction