Методология за подбор, извличане и зареждане на големи данни в Hadoop за целите на научни изследвания

Мариана Ковачева
Университет за Национално и Световно Стопанство(УНСС)
mkovacheva@unwe.bg

Конференция по проект EuroCC-България: Високопроизводителни пресмятания в полза на изследователите и обществото, 22.11.2022

Съдържание

- Какво представляват Големите данни?
- Типове Големи данни
- Подбор на Големи данни
- Извличане на Големи данни
- Зареждане на Големи данни
- Структура на първични данни в Hadoop за формиране на знания и за провеждане на научни изследвания
- Шаблон

Какво представляват Големите данни?

Дефиницията, която дава **Gartner** за **Големите данни (Big Data)** е, че това са информационни активи с голям обем, висока скорост и/или голямо разнообразие, които изискват рентабилни, иновативни форми на обработка на информация, които позволяват подобрена представа, вземане на решения и автоматизация на процеси.

Типове Големи данни

- Структурирани
- Полу-структурирани
- Неструктурирани

Структурирани данни

Структурираните данни са данните, които съответстват на модел на данни, имат добре дефинирана структура, следват последователен ред и могат да бъдат лесно достъпни и използвани от човек или компютърна програма. Обикновено се съхраняват в добре дефинирани схеми като бази данни, в табличен вид с колони и редове, които ясно дефинират неговите атрибути.

Полу-структурирани данни

Полу-структурираните данни са данни, които не съответстват на модел на данни, но имат известна структура. Липсва фиксирана или твърда схема. Това са данните, които не се намират в рационална база данни, но имат някои организационни свойства, които улесняват анализирането им. С някои процеси можем да ги съхраняваме в релационната база данни.

Неструктурирани данни

Неструктурираните данни са данните, които не съответстват на модел на данни и нямат лесно разпознаваема структура, така че да не могат лесно да се използват от компютърна програма. Неструктурираните данни не са организирани по предварително дефиниран начин или нямат предварително дефиниран модел на данни, поради което не са подходящи за основна релационна база данни.

Подбор на Големи данни

При подбор на Големи данни свързани с научни изследвания, има 2 варианта за подбор на данните и дефиниране на научно-изследователски въпрос.

- Дефиниране на проблем според вече наличните Големи данни в сферата
- Дефиниране на задача и търсене на Големи данни според зададената задача

Подбор на Големи данни

Данните, които биват подбирани и зареждани в Hadoop са сетове, предоставени свободно в интернет пространството. Поради тази причина, в някои случаи при предварително дефиниран научно-изследователски въпрос за сфери, които работят с чувствителна информация е изключително трудно да бъдат намерени подходящи данни, които да се използват.

Такива сфери са:

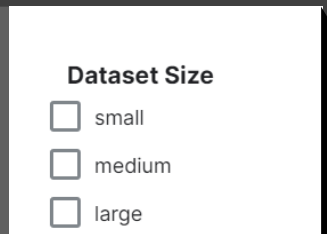
- Банки и финансови услуги
- Човешки ресурси
- Клиентска информация, т.н.

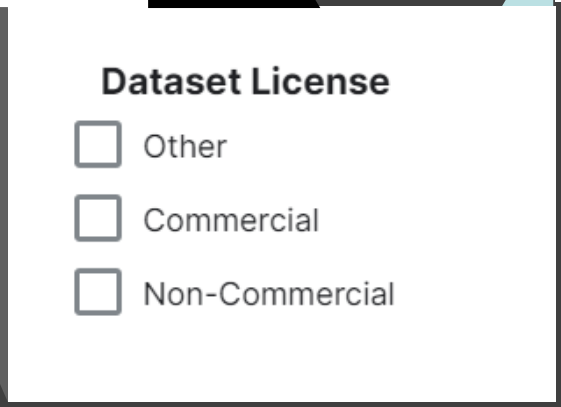
Извличане на Големи данни

Извличането на Големи данни се извършва от разнообразни източници онлайн, които имат вече налични сетове с данни. От официални правителствени сайтове със статистика, европейски и световни такива.

Примери:

- https://data.worldbank.org/
- https://ec.europa.eu/eurostat
- https://www.kaggle.com/





Извличане на Големи данни - пример

Сайтът https://www.kaggle.com/ предоставя изключително голям набор от свободни Големи данни от всички видове – структурирани, полуструктурирани или неструктурирани, като разполага с изключително подробни филтри. Те могат да се използват, за да се отсеят необходимите за нашата цел данни.

Зареждане на Големи данни - пример

След като приложим филтриране според критериите, които са ни необходими и сферата, в която искаме да намерим данни, ще демонстрираме практическо зареждане на такива.

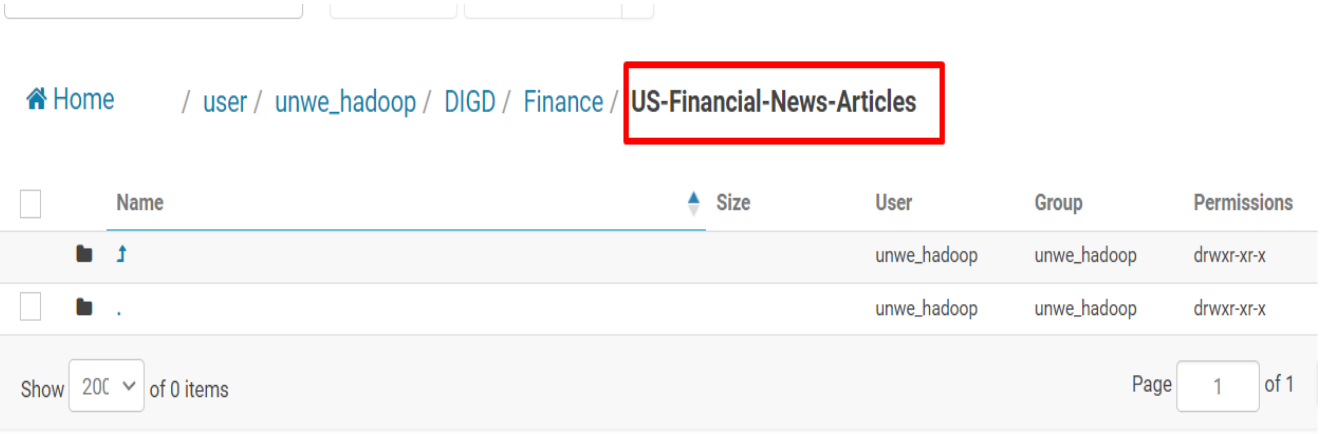
Пример: https://www.kaggle.com/datasets/jeet2016/us-financial-news-articles?resource=download

Сетът с данни се нарича: **US Financial News Articles**

Зареждане на Големи данни - пример

Създаваме папка в HDFS на Hadoop, чрез Hue.

Зареждаме данните, които предварително сме свалили локално – може да се разархивират предварително или да се качат, директно като архив.



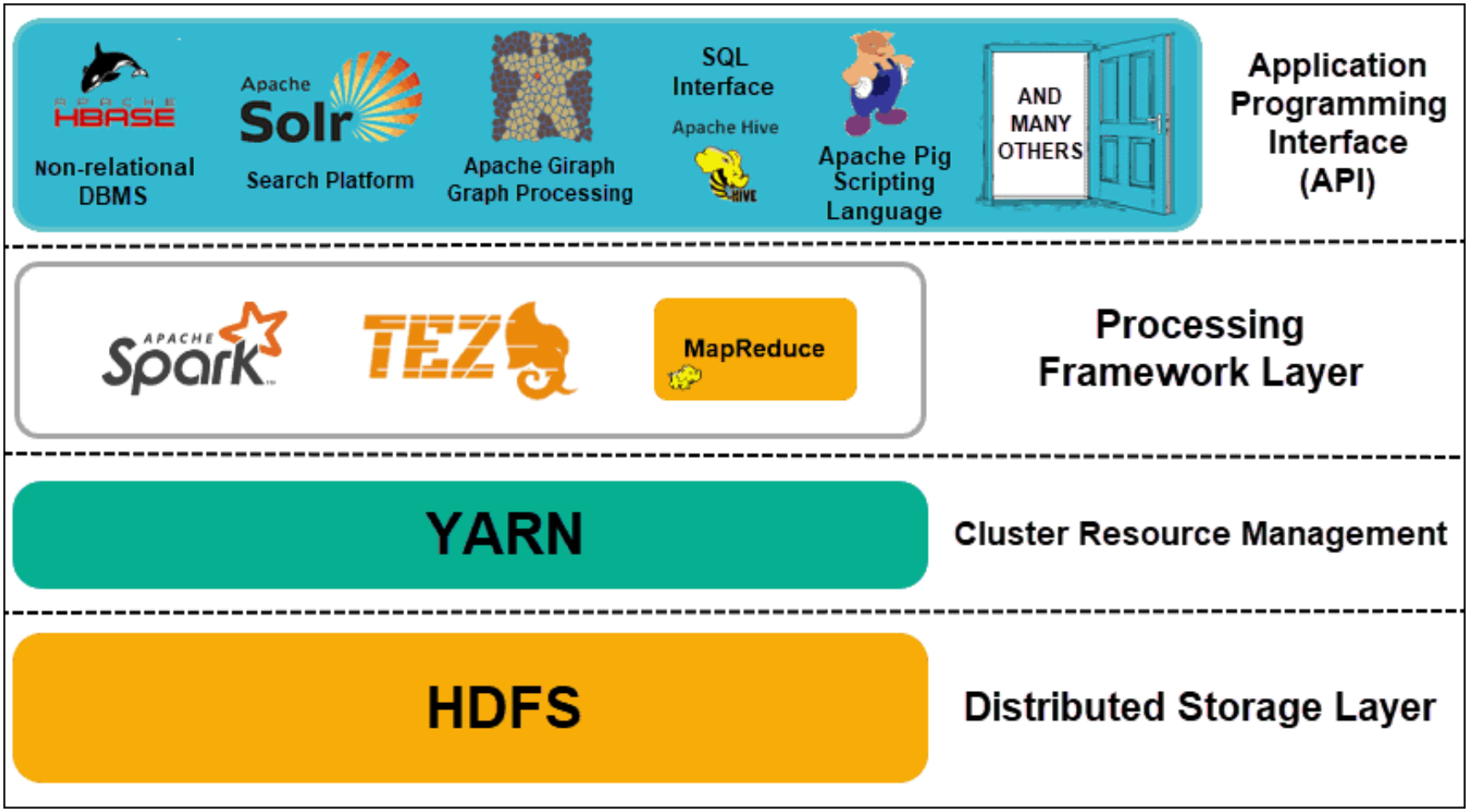
Зареждане на Големи данни - пример

Home / user / unwe_hadoop / DIGD / Finance / **US-Financial-News-Articles**

Name	Size	User	Group	Permissions
↑		unwe_hadoop	unwe_hadoop	drwxr-xr-x
.		unwe_hadoop	unwe_hadoop	drwxr-xr-x
news_0063117.json	10.2 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063118.json	5.5 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063119.json	4.0 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063120.json	1.7 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063121.json	11.9 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063122.json	2.2 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063123.json	2.1 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063124.json	1.8 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063125.json	1.5 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063126.json	4.0 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--
news_0063127.json	3.1 KB	unwe_hadoop	unwe_hadoop	-rw-r--r--

Зареждане на Големи данни - пример

Веднъж заредени данните в HDFS на Hadoop, те стават достъпни за използване от всички компоненти на инфраструктурата – Hive, Impala, Apache Spark и т.н.


Non-relational DBMS
Search Platform
Apache Giraph Graph Processing
SQL Interface
Apache Hive
Apache Pig Scripting Language
AND MANY OTHERS
Application Programming Interface (API)
MapReduce
Processing Framework Layer
YARN
Cluster Resource Management
HDFS
Distributed Storage Layer

След зареждане на данните в Hadoop, чрез интерфейса Hue в HDFS на Hadoop, бе създаден файл със структура и описание на наличните данни, който се актуализира своевременно при добавянето на нови данни.

Файлът се нарича „***Структура на първични данни в Hadoop за формиране на знания и за провеждане на научни изследвания*** “, като в него могат да се открият данни от различни сфери и различни типове.

Структура на първични данни в Hadoop за формиране на знания и за провеждане на научни изследвания

Наличните данни са разделени в следните категории:

- Транспорт
- Accommodation/Настаняване/
- Комуникация и съобщения
- Образование
- Медии
- Животни
- Финанси
- Агрикултура
- Икономика
- Медицина
- Социални медии

Шаблон

Създаден е шаблон с параметри, който се попълва за всеки един сет с данни. Параметрите са следните:

- Сфера
- Тип данни
- Период на събиране на заредените данни
- Колони / Полета
- Кратко описание
- Източник на данни
- Допълнителни данни

Пример - Emails

Сфера: Комуникация

Тип данни: неструктурирани

Период на събиране на заредените данни: не е упоменат

Колони / полета: Сетът е организиран в папки, като всяка папка е разделена на различни видове папки със съобщения в текстови формат – входящи, изходящи, чернови и т.н.

Кратко описание: Този набор от данни е събран и подготвен от проекта CALO (Когнитивен асистент, който учи и организира). Той съдържа данни от около 150 потребители, предимно висше ръководство на Enron, организирани в папки. Корпусът съдържа общо около 0,5 милиона съобщения.

Източник на данни: https://www.cs.cmu.edu/~./enron/

Допълнителни данни: Да. В линка по-горе могат да бъдат открити още допълнителни данни, които не са заредени в Hadoop.

Пример – Emails/изглед в Hue/