ДОДАТОК А

Слайди презентації

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

АТЕСТАЦІЙНА РОБОТА МАГІСТРА

Дослідження методів аналізу даних в data set для прийняття рішень з урахуванням вподобань

Науковий керівник:
к.т.н., проф. Дудар З.В.

Виконав:
студент групи ІПЗм-18-4 Хілюк Є.В.

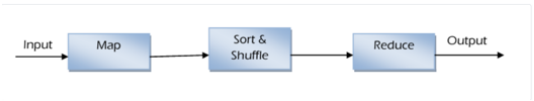
Мета роботи

- Дослідження основних методів, технологій та алгоритмів для аналізу великих обсягів даних з метою ефективного застосування в системах обробки даних
- Проектування та реалізація програмної системи з використанням технологій обробки і зберігання великих обсягів даних

Постановка задачі

- провести аналіз предметної галузі
- провести аналіз з існуючих методів та підходів до обробки великих обсягів даних
- вибрати оптимальний фреймворк для обробки та генерації великих обсягів даних (Hadoop MapReduce або Apache Spark)
- спроектувати архітектуру програмної системи
- виконати програмну реалізацію серверної та клієнтської частини
- зробити висновки про виконану роботу

Hadoop. Структура.



MapReduce. Як працює ?

Hadoop MapReduce. Переваги.

- Лінійна обробка величезних наборів даних
- Економічне рішення, якщо не очікується негайних результатів

- Hadoop зберігає дані на жорсткому диску разом з кожним кроком алгоритму MapReduce. У той час як Spark здійснює всі операції, використовуючи пам'ять з випадковим доступом.

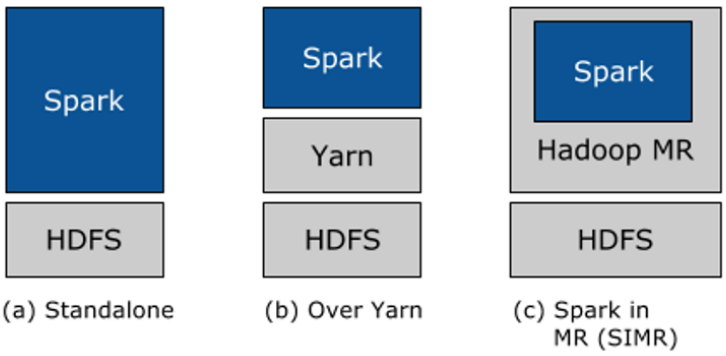
Apache Spark

7

Apache Spark. Переваги.

- швидка обробка даних
- ітеративна обробка
- обробка в режимі реального часу
- обробка графіків
- приєднання до наборів даних

8



Приклади аналогічних систем

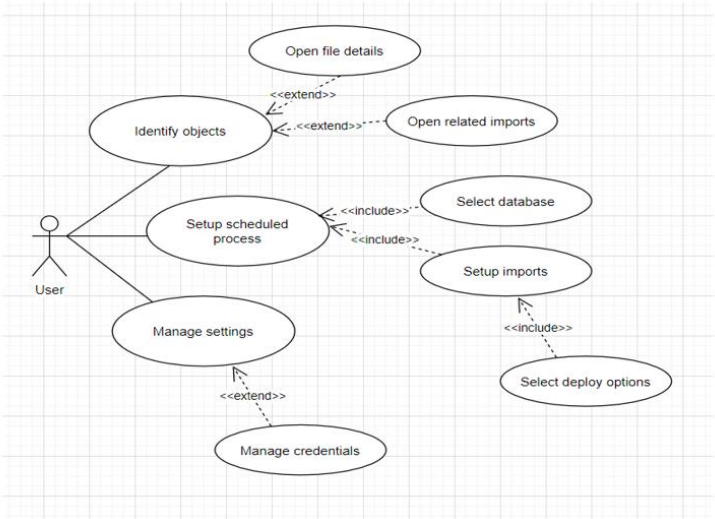
- OpenRefine
- Orange
- Weka

Weka

eSoftner

orange
DATA MINING
FRUITFUL&FUN

UML-моделювання. Use Case діаграма



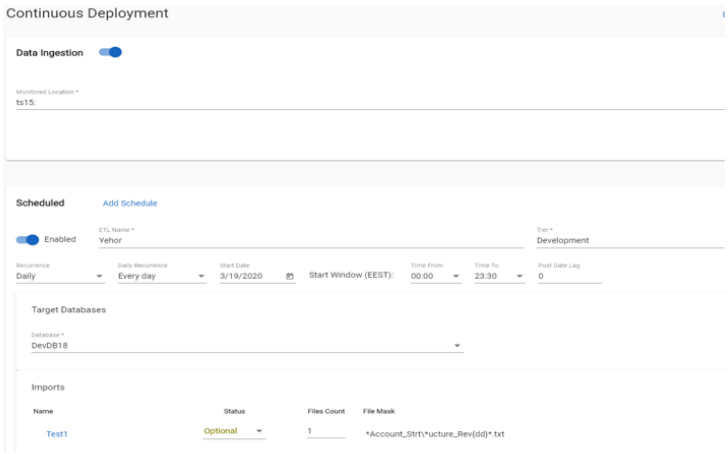
11

Засоби реалізації

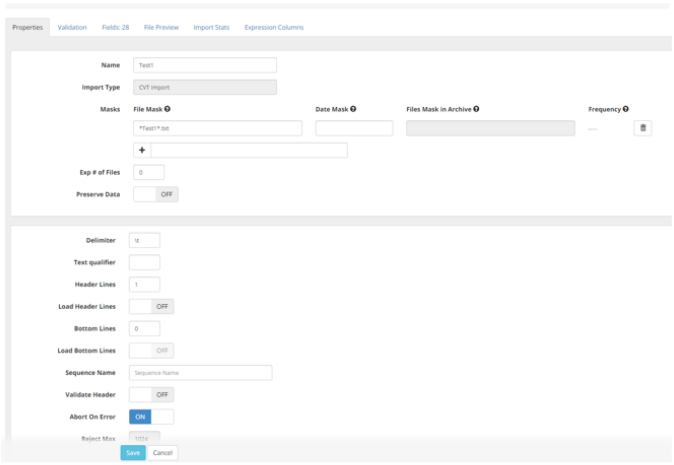
- C# (ASP.NET Web API)
- Angular + Material UI
- Apache Spark
- Hadoop (HDFS)
- MongoDb (MongoDriver)

12

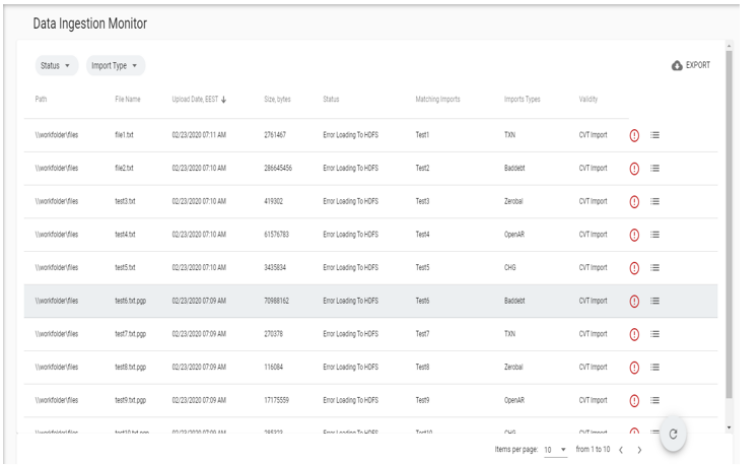
Програмний інтерфейс. Головна сторінка.



Програмний інтерфейс. Сторінка деталей конкретного імпорту



Програмний інтерфейс. Сторінка моніторингу.



15

Висновки

- У результаті роботи були проаналізовані та досліджені методи і підходи з аналізу і обробки великих обсягів даних.
- Був розроблений прототип програмної системи з урахування недоліків та переваг існуючих систем, що дозволяє оброблювати файли різних форматів із різної періодичністю, яку задає сам користувач.
- Програмна система має зрозумілий та зручний інтерфейс і демонструє можливості повної реалізації продукту

16

Апробація результатів роботи

- Хілюк Є.В., к.т.н., проф. Дудар З.В «Методи та техніки аналізу великих обсягів даних (BIG DATA)» / Конференція «Інформаційні інтелектуальні системи». Секція 3. Програмна інженерія. Інформаційні технології в світі

ДОДАТОК Б

Апробація результатів роботи

В ході науково-дослідної роботи було зроблено тезу для участі у наукових конференціях:

– Хілюк Є.В., к.т.н., проф. Дудар З.В «Методи та техніки аналізу великих обсягів даних (BIG DATA)» / Конференція «Інформаційні інтелектуальні системи». Секція 3. Програмна інженерія. Інформаційні технології в світі.

Методи та техніки аналізу великих обсягів даних (BIG DATA)

There are a lot of methods and approaches to big data analysis. This article reviews the solution of the problem of big data analysis. The main goal is to describe main models and approaches of big data analysis techniques using existing resolutions and approaches. It describes main principles and techniques of big data analysis and their modifications. Comparison of different data analysis techniques. Finally, we will make a conclusion about the use and application of these methods and techniques on practice.

Великі дані (англ. Big Data) в інформаційних технологіях — набори інформації (як структурованої, так і неструктурованої) настільки великих розмірів, що традиційні способи та підходи (здебільшого засновані на рішеннях класу бізнесової аналітики та системах управління базами даних) не можуть бути застосовані до них [1].

Термін Big Data з'явився порівняно нещодавно, приблизно в 2011 році. З кожним роком означення цього терміну змінюється через те, що обсяги даних зростають, а людям, що працюють з ними потрібно покращувати та вигадувати нові методи та алгоритми роботи з великими об’ємами даних.

Розглянемо основні методи та техніки роботи з big data:

- A/B тестування;
- злиття та інтеграція даних;
- data mining;
- машинне навчання;

- natural language processing (NLP);
- статистика;

A/B тестування включає в себе порівняння контрольної групи з різними тестовими групами, щоб визначити, які способи покращення або зміни поліпшать задану цільову змінну. Прикладом аналізу може служити копія, текст, зображення чи макет, що невідмінно покращать коефіцієнти конверсій на веб-сайті електронної комерції. Великі дані вписуються в цю модель, оскільки вона може перевірити величезну кількість записів, однак досягти цього можна лише в тому випадку, якщо групи мають достатньо великий розмір, щоб отримати значущі відмінності.

Злиття та інтеграція даних поєднує набір методів, які аналізують та інтегрують дані з різних джерел та рішень. Через це розуміння предметної області є більш ефективними та потенційно точнішими, ніж якщо б вона розроблялася через єдине джерело даних.

Загальний інструмент, що використовується в аналітиці великих даних, видобуток даних або “data mining” витягує зразки з великих наборів даних, поєднуючи методи статистики та машинного навчання, в рамках управління базами даних. Прикладом може виступати ситуація, коли дані клієнтів видобуваються, щоб визначити, які сегменти ринку найбільш ймовірно реагують на пропозицію.

Добре відоме в галузі штучного інтелекту, машинне навчання також використовується для аналізу даних. Виходячи з інформатики, вона працює з комп’ютерними алгоритмами для створення припущень, заснованих на даних. Вона дає прогнози, які неможливі для людських аналітиків.

“Natural language processing” або обробка природної мови, відома як галузь інформатики, штучного інтелекту та лінгвістики. Вона використовує алгоритми для аналізу людської (природної) мови.

Статистика працює для збору, організації та інтерпретації даних у межах опитувань та експериментів.

Інші методи аналізу даних включають просторовий аналіз, прогнозне моделювання, навчання правилам асоціацій, аналіз мереж та багато іншого.

Технології, які обробляють, керують та аналізують ці дані, є зовсім іншими, які аналогічно розвиваються з часом.

Після аналізу існуючих інструментів роботи з великими обсягами даних (big data), можна зробити висновок, що кожен з методів чи технік є вкрай важливим для окремих галузей та чітких задач, які перед ними ставлять. Також можна зазначити, що з дуже стрімким розвитком інформаційного кола, ці методи будуть тільки покращуватися та модифікуватися, а також будуть розроблятися нові, але вже сьогодні зрозуміло, що для того, щоб отримати найбільш оптимальний результат, необхідно намагатися поєднувати декілька методів або технік по роботі з великими обсягами даних.

Перелік використаних джерел:

1. Великі дані. URL: https://uk.wikipedia.org/wiki/Великі_дані (дата звернення: 25.02.2020).

2. Big Data Analysis Techniques. URL: https://www.getsmarter.com/blog/career-advice/big-data-analysis-techniques/ (дата звернення: 25.02.2020).

3. Big Data. URL: https://www.it.ua/ru/knowledge-base/technology-innovation/big-data-bolshie-dannye (дата звернення: 25.02.2020).

4. Класифікація методів аналізу великих даних. URL: http://science.lpnu.ua/sites/default/files/journal-paper/2018 (дата звернення: 25.02.2020).