УДК 004.62

Федоренчік М.О., магістрант, Пазюк Ю.М., доцент, науковий керівник

ДОСЛІДЖЕННЯ ТА ОБРОБКА ВЕЛИКИХ ДАНИХ З МЕТОЮ ВИЯВЛЕННЯ ЗАКОНОМІРНОСТЕЙ

Запорізька державна інженерна академія, кафедра ПЗАС

Початком історії розвитку великих об`ємів даних та їх обробки вважають XIX століття. Зміни в знаннях сприяли не тільки приплив великої кількості нової інформації з усього світу, а й зрушення у виробництві, обробці та аналізі цієї інформації. Минуло два сторіччя після першої революції великих даних, але багато з проблем і шляхи їх вирішення зберігаються аж до сьогоднішнього дня [1].

Проблема «великих даних» існувала протягом всієї історії розвитку IT. Завжди виникало бажання обробити великі масиви даних за мінімальний час і завжди для цього виявлялося недостатньо потужності існуючої IT-інфраструктури. Під терміном Big Data в різному контексті можуть матися на увазі дані великого обсягу, технологія їх обробки, проекти, ринок і навіть компанії, що активно використовують цю технологію [2].

Дані більше не розглядаються як якась статична або застаріла величина, яка стає марною по досягненні певної мети, наприклад, приземлення літака. Швидше, вони стали сировинним матеріалом бізнесу, життєво важливим економічним внеском, використовуваним для створення нової економічної вигоди. Виявилося, що при правильному підході їх можна спритно використовувати повторно, як джерело інновацій і нових послуг. Великі дані призначені для прогнозування.

В сучасній Україні питання вибору професії в учнів старших класів ставиться все складніше і складніше. У свою чергу вузи повинні мати статистику в якій вони будуть бачити які предмети здаються на зно більшу кількість разів і які предмети з якими вибирають одинадцятикласники. Так як обсяг цих даних досить великий, на допомогу нам приходять сучасні технології обробки великих даних.

Важливим етапом є вибір інструментів розробки, їх порівняння та вибір найефективніших, найгнучкіших та найзручніших для використання. Вибір полягав між середовищами розробки, допоміжними бібліотеками та фреймворками для роботи з великими даними. Для аналізу даних було обрано декілька фреймворків. Вибір було зроблено на основі наступних особливостей:

- Hadoop складається з пакету Hadoop Common, який надає абстракції операційної та файлової системи, рушій MapReduce (або MapReduce/MR1 або YARN/MR2) та Hadoop Distributed File System (HDFS). Пакет Hadoop Common містить файли JAR та скрипти потрібні для запуску Hadoop.
- Застосунок Spark складається з процесу-драйвера (англ. driver process) та багатьох процесів виконавців (англ. executor processes). Драйвер є серцем застосунку Spark, і виконує наступні функції: зберігає та обробляє інформацію про стан застосунку, відповідає на запити користувацьких програм, аналізує та розподіляє завдання між виконавцями та порядок їх виконання. Виконавці натомість виконують завдання та звітують про їх виконання і свій стан драйверу.

Для розробки застосунку обрано середовище PyCharm, оскільки воно має найкращу підтримку Python при написанні коду, безліч можливостей по рефакторінгу та аналізу коду.

Розроблений програмний комплекс складається з двох модулів:.

- **Модуль обробки даних**

Це досить простий модуль, який відповідає за обробку вхідних даних ЗНО.

- **Модуль інтерфейсу користувача**

Доступ до модулю генерування користувачу надається через REST API. Ця задача вирішувалась стандартними методами оптимізації швидкодії веб-застосунків на зразок кешування, асинхронних запитів до бази даних, горизонтальне масштабування серверів і т.д.

Після дослідження предметної області та детального аналізу поставленої проблеми було встановлено актуальність аналізу даних з ціллю виявлення закономірностей. Комплекс обраних методів обробки даних дозволить досягнути якісного результату, дасть уявлення про

основні проблеми, що виникають при рішенні задач цього класу, та методи їх вирішення. Аналіз використаних алгоритмів дозволить краще зрозуміти принципи їх побудови і використати отримані знання для реалізації власних алгоритмів обробки даних.

Висновки:

- Була досліджена методологія обробки даних та маніпуляцій з ними.
- Досліджена архітектура та принципи розробки Big data моделей.
- Досліджені методи побудови моделей глибокого аналізу даних з використанням фреймворку Hadoop.

Література

1. Хэмиш Робертсон , Джоанн Травалья , История Big Data восходит к практикам общественного порядка XIX века, Веб-ресурс: https://22century.ru/popular-science-publications/big-data-problems
2. Андрей Найдич, Big Data: проблема, технология, рынок, Веб-ресурс: https://compress.ru/article.aspx?id=22725