Кравець М. З., Пастернак І. І.

Національний університет "Львівська політехніка",

кафедра електронних обчислювальних машин

**Відображення результатів обробки великих наборів даних за допомогою технології OpenGL**

*© Кравець М. З., Пастернак І. І. 2019*

**Автором розглянуто проблему вибору методів обробки великих наборів даних, проаналізовано типові методи обробки великих наборів даних, розглянуто структурну схему програмного забезпечення, побудованого для прийому і аналізу даних з подальшим графічним відображенням, використовуючи технологію OpenGL.**

**Ключові слова: великі набори даних, аналіз даних, методи обробки, візуалізація, OpenGL.**

**The author deals with the problem of choosing the methods of processing BigData, analyzes the typical methods of processing large datasets, examines the block diagram of software built for receiving and analyzing data with subsequent graphical display using OpenGL technology.**

**Keywords: BigData, data analysis, processing methods, visualization, OpenGL.**

**Вступ.** Великі дані (big data) уже тривалу кількість часу є всюди - в продажі, медіа фінансовому секторі, телекомунікаціях, медицині, IT-індустрії. Переважна кількість даних, що збираються, зберігаються і передаються через нові технології, змінюють пріоритети для багатьох діючих підприємств. Тема обробки та пояснення результатів обробки великих даних є складною і об’ємною, оскільки на 2020 рік прогнозується, що світовий обсяг даних буде становити 40–44 зеттабайта. Багато людей працюють над тим, щоб вчасно обробляти інформацію, що постійно надходить та десь її зберігати для подальшого вивчення та оцінки.

**Стан проблеми.** Сучасні технології big data дають можливість перетворення маси неоднорідних та неструктурованих даних на інформацію, яку можна застосувати для підвищення ефективності сфери, до якої вони відносяться. Це пояснюється наступним: вважається, що коректний аналіз даних дозволяє виявляти приховані закономірності, що вислизають від обмеженого людського сприйняття. Це дає безпрецедентні можливості оптимізації багатьох сфер нашого життя. Але мало приймати та зберігати інформацію - адже тоді не отримати всіх переваг цієї інформації. Для того, щоб можна було передбачити щось, правильно оцінити стан певної області, потрібно навчитися правильно читати і трактувати результати обробки.

Для вирішення проблеми великих затрат часу на розуміння та вивчення результатів обробки вирішено створити застосунок, що буде відображати в найпростішому варіанті результати обробки даних. Також для отримання більш коректних результатів проводитиметься аналіз методів та вдосконалення, а також вибір того, що найкраще підходить до того чи іншого набору.

**Постановка задачі**. Розглянути можливості покращення сучасних методів аналізу великих обсягів даних та вибрати найоптимальніший спосіб їх відображення, використовуючи технологію OpenGL.

**Розв’язання задачі.** Для розв’язку поставленої задачі було вирішено спершу оглянути та оцінити принципи, які існують для обробки big data.

Якщо не враховувати фізичний об’єм, є інші характеристики, що підкреслюють складність задачі обробки і аналізу цих даних. Цими характеристиками є: , швидкість приросту даних і необхідність їх швидкої обробки, здатність обробляти дані різних типів, достовірність, цінність та інші. Також додано таку характеристику, як візуалізація. На основі вказаних характеристик можна визначити певні принципи роботи з даними:

* Горизонтальна масштабованість. Це — базовий принцип обробки великих даних. Як вже було зазначено, великих даних з кожним днем стає все більше. Відповідно, необхідно збільшувати кількість обчислювальних вузлів, за якими розподіляються ці дані, при чому обробка має відбуватись без погіршення продуктивності
* Відмовостійкість. Цей принцип витікає з попереднього. Оскільки обчислювальних вузлів у кластері може бути багато (іноді десятки тисяч) та їх кількість, не виключено, буде збільшуватись, зростає ймовірність виходу машин з ладу. Методи роботи з великими даними мають враховувати ймовірність таких ситуацій і передбачати превентивні заходи
* Локальність даних. Оскільки дані розподілені по великій кількості обчислювальних вузлів, то, якщо вони фізично знаходяться на одному сервері, а обробляються на іншому, витрати на передачу даних можуть бути невиправдано великими. Тому обробку даних бажано проводити на тій же машині, на якій вони зберігаються

Ці принципи відрізняються від тих, які характерні для традиційних, централізованих, вертикальних моделей зберігання добре структурованих даних. Власне, для роботи з великими даними розробляються підходи і технології.

Існують компанії, що спеціалізуються на розв’язанні задач, пов’язаних зі стратегічним управлінням, використовуючи великі дані. Одна з них - McKinsey. Дана компанія виділяє 11 методів і технік аналізу, що застосовуються до великих даних:

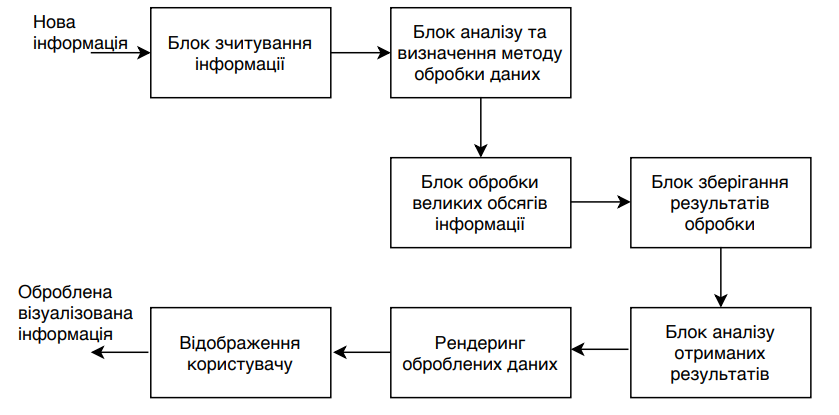
* Методи классу Data Mining — сукупність методів виявлення у даних раніше невідомих, нетривіальних, практично корисних знань, необхідних для прийняття рішень.
* Краудсорсинг — класифікація і збагачення даних силами широкого, неозначеного кола особистостей, що виконують цю роботу без вступу у трудові стосунки.
* Змішання та інтеграція даних (data fusion and integration) — набір технік, що дозволяють інтегрувати різнорідні дані з розмаїття джерел з метою проведення глибинного аналізу
* Машинне навчання, включаючи навчання з учителем і без учителя
* Штучні нейронні мережі, мережевий аналіз, оптимізація, у тому числі генетичні алгоритми
* Розпізнавання образів
* Прогнозна аналітика
* Імітаційне моделювання (simulation) — метод, що дозволяє будувати моделі, що описують процеси так, як вони би проходили у дійсності
* Просторовий аналіз (spatial analysis) — клас методів, що використовують топологічну, геометричну і географічну інформацію, що вилучається із даних.
* Статистичний аналіз — аналіз часових рядів, A/B-тестування
* Візуалізація аналітичних даних — подання інформації у вигляді малюнків, діаграм, з використанням інтерактивних можливостей і анімації, як для отримання результатів, так і для використання у якості вихідних даних для подальшого аналізу. Дуже важливий етап аналізу великих даних, що дозволяє показати найважливіші результати аналізу у найбільш зручному для сприйняття вигляді.

Завдяки візуалізації даних інформація краще сприймається і дозволяє швидко і ефективно донести до глядача власні думки та ідеї. Фізіологічно, сприйняття візуальної інформації є основною для людини. Успіх візуалізації безпосередньо залежить від правильності її застосування, а саме від вибору, наприклад, типу графіка, його вірного використання та оформлення. Для того, щоб правильно подати інформацію, необхідно вирішити, яким способом її краще буде представляти, та які інструменти застосовувати. Також необхідно визначати, які особливості даних необхідно виділяти, з чим порівнювати. Такі задачі теж мають бути вирішеними з використанням програмного забезпечення.

Інструментом графічного відображення оброблених даних буде відкрита графічна бібліотека OpenGL. Це специфікація, що визначає незалежний від мови програмування крос-платформний програмний інтерфейс (API) для написання застосунків, що використовують 2D та 3D комп'ютерну графіку. Цей інтерфейс містить понад 250 функцій, які можуть використовуватися для малювання складних тривимірних сцен з простих примітивів. Широко застосовується індустрією комп'ютерних ігор і віртуальної реальності, у графічних інтерфейсах (Compiz, Clutter), при візуалізації наукових даних, в системах автоматизованого проектування тощо.

OpenGL використовується для таких цілей:

* Для того, щоб приховувати складнощі встановлення зв'язку комп'ютера з різними 3D акселераторами, надати програмістові один, загальноприйнятий API
* Для того, щоб приховати можливості базових інструментальних машин, які відрізняються своїм намаганням виконати підтримку повного набору особливостей OpenGL (використання програмної емуляції, якщо необхідно).



Структурна схема застосунку для обробки на графічного відображення великих обсягів даних

Згідно із вказаною структурною схемою, інформацію, що надходить до застосунку, спрямовують до блоку, у якому проводитиметься вибір методу обробки та перетворення інформації. На основі оцінки даних за їх характеристиками, вибраний метод та самі дані обробляються у блоці для обрахунків. Оскільки обсяги оброблених даних, нам необхідне їх зберігання і аналіз частинами. Тому запропоновано окремий блок для зберігання результатів обробки та блок аналізу результатів. Останній аналіз дозволяє виділяти певні набори даних, що, наприклад, потребують окремої уваги або мають критичні показники. Після оцінки результатів отримані дані візуалізуються та подаються користувачу для перегляду та вивчення.

Даний застосунок може працювати з великою кількістю даних за рахунок попереднього аналізу даних на розподіленість та подальшої багатопотокової обробки. Окрім того, забезпечення регулярного оновлення графічного вікна для користувача дозволить спостерігати за зміною стану даних, тобто проводити моніторинг.

Програмне забезпечення дозволятиме, за допомогою більш доступних засобів, аналізувати та оцінювати великі набори даних та отримувати простіший для розуміння та пояснення результат обробки даних.

**Висновки**. В роботі розглянуто серйозність проблеми вибору методу обробки великих наборів даних, проаналізовано типові методи, створено та пояснено структурну схему програмного забезпечення, описано технологію OpenGL.

**Література**

1. Beyer M.A., Laney D.The Importance of “Big Data”: A Defiition [Electronic resource] / MarkA. Beyer, Douglas Laney // Gartner Inc. — Electronic data. — [Stamford: Gartner, 2012]. — Mode of access: World Wide Web: http://www.gartner.com/id=2057415/ (viewed on October 11, 2017). — Title from the screen.

2. Черняк Л. Большие Данные — новая теория и практика [Электронний ресурс] / Л. Черняк // Открытые системы. СУБД. — Электронные данные. — [Москва: “Открытые системы”, 2011]. — № 10. — Режим доступа: https://www.osp.ru/os/2011/10/13010990/ (дата обращения 16.11.2017).

3. Аналитический обзор рынка Big Data [Электронний ресурс] / Хабрахабр. — Электронные данные. — [Москва: TechMedia, 2015]. — Режим доступа: https://habrahabr.ru/company/moex/blog/256747 (дата обращения 21.11.2017). — Название с экрана.

4. Юрасов С. Оцифровування статистики, або Перша їжа для Bigdata [Електронний ресурс] / Стас

Юрасов // Інтернет-видання Економічна правда. — Електронні дані. — [Київ, Економічна правда, 2015].

Режим доступу: https://www.epravda.com.ua/publications/2015/08/20/554624 (дата звернення 22.11.2017 р.).

Назва з екрану.

5. Глущенко Н. Большие данные большого города: как Big Data меняет жизнь Киева [Электронный ресурс] / Нина Глущенко // интернет-журнал AIN.UA. — Электронные данные. — [Киев: AIN.UA, 2017]. - Режим доступа: https://ain.ua/special/big-data-in-kyiv/ (дата обращения 24.11.2017). — Название с экрана.

6. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы

живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер; пер. с англ. Инны Гайдюк. —

Москва.: Манн, Иванов и Фербер, 2014 — 240с.

7. Elie Tahari combines fashion savvy with powerful analytics [Electronic resource] / IBM Business

Analytics. — Electronic data. — [NY: IBM Corporation, 2014]. — Mode of access: https://www01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=YTC03447USEN (viewed on October 14, 2017). — Title from the screen.

8. More Than 30 Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020

[Electronic resource] / ABI Research. — Electronic data. — [London: Allied Business Intelligence, Inc., 2013].

Mode of access: https://www.abiresearch.com/press/more-than-30-billion-devices-will-wirelessly-conne/ (viewed on October 19, 2017). — Title from the screen.