УДК 519.876

О. М. Москвін
М. В. Плис

ОСОБЛИВОСТІ РОЗРОБКИ РОЗПОДІЛЕНОЇ СИСТЕМИ ОБРОБКИ ГЕНЕТИЧНОЇ ІНФОРМАЦІЇ

Вінницький національний технічний університет

Анотація

Описано підходи для обробки великих масивів генетичних даних. Розроблено розподілену систему для обробки генетичної інформації.

Ключові слова: кластер, геном, генетична інформація, біоінформатика, розподіленні обчислення, Hadoop, Spark, Cassandra.

Abstract

A set of approaches for big genetic data analysis considered. Developed a distributed system for genetic information processing.

Keywords: cluster, genome, genetic information, bioinformatics, distribution calculations, Hadoop, Spark, Cassandra.

Вступ

В зв’язку з стрімким розвитком багатоядерних та багатопроцесорних обчислювальних систем набувають популярності розподілені та хмарні обчислення. За умови наявності ефективного алгоритмічного забезпечення, підвищення щільності апаратних ядер дозволяє більш ефективно вирішувати задачі аналізу даних, що є основою сучасних SaaS, PaaS та IaaS систем. Актуальність даної теми підтверджується тим фактом, що кількість даних, які були вироблені людством до 2003 року становили 5 мільярдів гігабайт, така ж кількість даних вироблялася кожні два дні в 2011 році, і кожні 10 хвилин в 2013. Отже має місце експоненційне зростання обсягів інформації. Аналогічним чином, але у бік зменшення змінюються вимоги щодо часу на обробку даної інформації. Очевидно, що для обробки таких масивів даних і отримання результату за певний кінцевий час, потужностей однієї обчислювальної системи недостатньо. Тому проблема побудови ефективних архітектур програмно-апаратних систем та алгоритмічного забезпечення для задач аналізу великих масивів інформації є актуальною.

Однією з областей науки, що оперує надвеликими масивами даних є біоінформатика. В середньому один оцифрований геном людини складається з 150 млн. записів і займає 120 гігабайт даних. Задачі сучасної біоінформатики потребують швидкого аналізу множини геномів з метою пошуку в них певних закономірностей. Вимоги до часу обробки – до 5 секунд. Для обробки одного такого геному дана задача з вказаним обмеженням за часом може бути вирішена засобами однієї обчислювальної системи, але не у випадку наявності сотень, тисяч геномів, що підлягають аналізу одночасно, і потребує розробки нових підходів в основі яких лежить використання розподілених багатопроцесорних систем. Вирішення даної задачі представляє науковий інтерес і є предметом розгляду даної роботи.

Результати дослідження

Основними вимогами до кінцевої системи обробки генетичної інформації є наступні:

- основний робочий формат геному – VCF (Variant Call Format) версій 41 та 42;
- використання розподіленого підходу до збереження генетичної інформації;
- можливість здійснення повнотекстового пошуку по множині файлів;
- можливість здійснення виконання аналітичних запитів з використанням агрегатних функцій;
- граничний час виконання запитів обробки інформації – 5 секунд.

Аналіз існуючих систем обробки генетичної інформації з відкритим кодом, серед яких Gemini, LuceGene, BigQ, TABIX, BCF2, показав, що вони характеризуються достатньо потужними засобами до обробки геномів, але здебільшого орієнтовані на обробку поодиноких файлів у не розподіленому

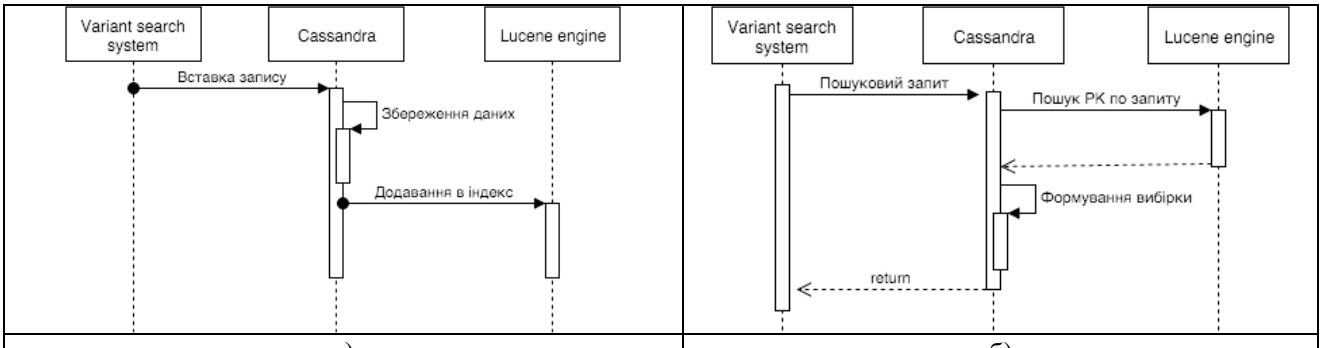
середовищі. Система ADAM є розподіленою, але використовує формат даних, що за умови аналізу множини файлів не задовольняє зазначене вище часове обмеження на виконання при аналізі більш ніж 11 геномів. Затрати на трансформацію перелічених систем для приведення їх до відповідності вимогам та обмеженням за часом сумірні з розробкою нової системи з архітектурою, що дозволить задовольнити зазначені вимоги.

З метою побудови архітектури нової системи обробки генетичної інформації було проведено аналіз існуючих підходів до обробки великих масивів даних в результаті чого було виявлено наступні:

- ETL (Extract, Transform, Load) – один із основних процесів в управлінні сховищами даних, який включає в себе витяг даних із зовнішніх джерел, їх трансформацію, перетворення для потреб бізнес-моделі, та завантаження в сховище даних OLAP. ETL+OLAP є досить потужним механізмом обробки та аналізу даних. Серед недоліків – погано піддається масштабуванню, не підтримує концепцію горизонтальної сегментації даних (sharding).
- HPC (High performance computing) – спеціалізована технологія для розробки програмних систем для виконання на багатопроцесорних обчислювальних машинах. Найбільш поширеною реалізацією є OpenMP. Проте для реалізації такої обчислювальної потужності необхідні великі затрати на апаратне забезпечення, а саме на швидкісні носії для забезпечення даних та мережеві обладнання, оскільки саме обмін великими обсягами даних та їх синхронізація є основними проблемами в даній технології.
- MapReduce – найбільш відома технологія паралельної обробки даних великої розмірності. Особливістю є розділення логіки програми на підзадачі розбиття даних та об'єднання результатів з можливістю необмеженого горизонтального масштабування шляхом додавання обчислювальних вузлів до кластеру. Дана технологія дозволяє ефективно вирішувати різноманітні задачі аналізу даних, але не виправдана для задач пошуку даних за заданим шаблоном.

MapReduce найбільше задовольняє зазначеним вимогам для забезпечення аналітичних обрахунків над генетичною інформацією.

Наступним кроком є вибір сховища даних. Для цього було проаналізовано ряд реляційних баз (SQL) даних Oracle, PostgreSQL, MySQL, MariaDB так і нереляційних (NoSQL) Cassandra, Redis, HBase, MongoDB, Couchbase та ін. Реляційна модель не підходить для вирішення задачі, що розглядається в даній роботі із міркувані швидкодії та можливості до масштабування. З цих же міркувань, для забезпечення швидкісного розподіленого сховища, було обрано NoSQL систему управління базами даних Apache Cassandra. Її перевагами є висока масштабованість, пропускна здатність для операцій запису і зчитування, SQL-подібна мова, підтримка реплікації, можливість розробки власних індексів для даних та надійність. З недоліків Apache Cassandra – суттєві обмеження пов'язані з використанням декількох індексів одночасно при пошуку даних. Для вирішення задачі пошуку було розроблено модуль інтеграції Apache Cassandra з бібліотекою Lucene, що дозволило реалізувати високошвидкісний повнотекстовий пошук необмеженої складності.



а) б)

Рис. 1 – UML діаграма а) додавання даних, б) пошуку даних з використанням бібліотеки Lucene у Cassandra

Особливістю розробленого модуля пошуку є те, що він повністю сумісний і інтегрований з інфраструктурою СУБД Apache Cassandra і дозволяє об'єднувати результати пошуку Cassandra та Lucene в одне логічне ціле. Висока швидкість пошуку за індексом Lucene досягається за рахунок

використання B+tree індексів. Відповідно до проведених в роботі досліджень, індекс Lucene призводить лише до 30% збільшення використання дискового простору, що є прийнятним, оскільки не призводить до дублювання наявних даних.

Для реалізації технології MapReduce разом із розробленою підсистемою збереження, розподілу та пошуку даних було обрано Apache Spark, що використовує технологію In-Memory, яка дозволяє значно зменшити час необхідний на проведення розрахунків. Для цього було здійснено інтеграцію Apache Spark та Apache Cassandra з метою мінімізації дискового простору.

В кожному вузлі кластера міститься Apache Cassandra з проінтегрованим модулем для повнотекстового пошуку, на основі Apache Lucene, та Apache Spark для проведення аналітичних розрахунків. Схематично структура розподіленої системи показана на рисунку 2.

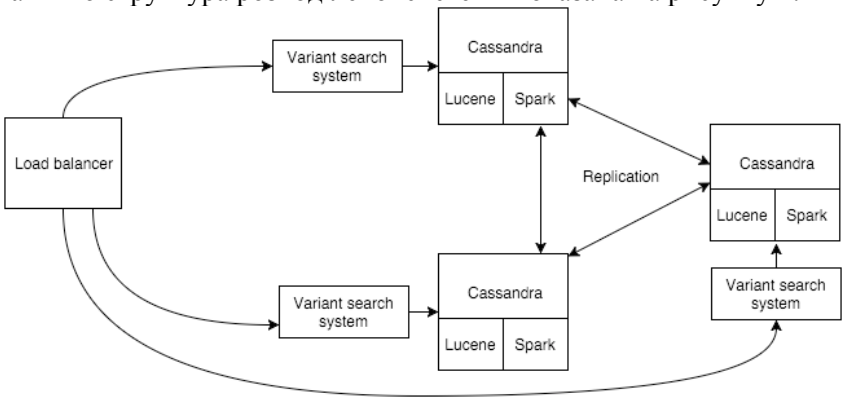


Рис. 2 – Схема структури розробленої розподіленої системи

Для мінімізації розміру бази даних була проаналізована структура VCF формату та використана концепція референсних (базових) геномів. Ідея полягає в тому, що замість використання повного геному зберігається лише одна копія певного базового геному, відносно якого створюється похідні файли, що містять лише ті алельні гени, що відрізняються від генів описаних у базовому файлі. Такий підхід дозволяє зменшити розмір геному у 30-50 раз і суттєво підвищити швидкість пошуку та загальну ефективність системи.

Результати тестування розробленої розподіленої системи, що використовує Apache Cassandra, реалізовані засоби пошуку та аналітики з концепцією зберігання даних на основі референсних геномів показані в таблиці 1. Отримані результати задовольняють вимогам та обмеженням визначеним у роботі.

Таблиця 1 – Результати тестування розробленої системи

Записів бази даних	**Операція пошуку, с**	**Агрегатна операція, с**
$10 \cdot 10^6$	0,3	0,4
$25 \cdot 10^6$	0,7	0,75
$50 \cdot 10^6$	1,5	1,65
$100 \cdot 10^6$	1,8	1,8

Висновки

В роботі запропонована архітектура системи для вирішення задачі обробки генетичної інформації, що дозволяє здійснювати швидкий пошук та аналітичних підрахунок інформації про гени із заданими характеристиками. Відповідно до отриманих результатів тестування система входить в обмеження по часу, допустиме для поставленої задачі.

Москвін Олексій Михайлович — к.т.н., ст. викладач кафедри комп'ютерних систем управління, Вінницький національний технічний університет, Вінниця.

Плис Максим Валентинович — студент кафедри комп'ютерних систем управління, Вінницький національний технічний університет, Вінниця, e-mail: maksm.plis1995@gmail.com;

Maksim V. Plys – Department of Computer Control Systems, Vinnytsya National Technical University, Vinnytsya, e-mail: maksm.plys1995@gmail.com;

Oleksiy M. Moskvin – Ph.D., lecturer in Computer Control Systems, Vinnytsia National Technical University, Vinnitsa.