УДК 614.844

ІНФОРМАЦІЙНЕ ТА ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ПОБУДОВИ РОЗПОДІЛЕНОЇ СИСТЕМИ ЗБЕРІГАННЯ ДАНИХ НА ОСНОВІ MAPREDUCE

А.А. Марюха; А.А. Підкуйко; С.О. Петров
Сумський державний університет
e-mail: tohahak540@gmail.com

Значне збільшення обсягів даних змушує розробників інформаційних систем більше уваги приділяється питанням ефективної обробки даних. Традиційні реляційні СУБД не здатні обробляти обсяги даних, що не можуть бути розміщені в рамках однієї файлової системи, або не структуровані дані. Саме тому виникає необхідність створення нових моделей розподіленої обробки інформації таких як MapReduce [1]. Фактично подібні моделі реалізуються як програмний каркас, що розроблений для проведення розподіленої паралельної обробки великих масивів даних з використанням кластерів. Програми, написані для цієї системи, автоматично являються розпаралелюваними і виконуються на легкомасштабуємому кластері [2]. Система також обробляє збої машин в кластері, координує повідомлення між комп'ютерами усередині кластера, мінімізуючи навантаження на мережу і сховище даних.

Використання даного підходу дозволяє розглянути наступну задачу: необхідно реалізувати інформаційну систему, яка буде перевіряти орфографію написання слів та у випадку знаходження помилки – виправляти її. Під виправленням помилок розуміється знаходження та

запропонування користувачу кількох найбільш схожих слів зі словнику. «Схожість» двох слів буде визначатися за допомогою триграм [3]. Триграмами слова називаються всі його підрядки довжини 3 (рис. 1), кількість співпавших триграм і є «схожість». Для роботи системи, необхідно обробити та завантажити досить великий (приблизно 20 мегабайт) орфографічний словник, по якому і буде шукатись найбільш схоже слово для даного.

Рисунок 1 – Розбиття слова на триграми

Оскільки рішення даної задачі передбачає обробку та формування великої кількості інформації, класичні методи обробки даних не є досить ефективними. Тому очікується що застосування методу MapReduce до цієї задачі дасть високий приріст обчислюваної продуктивності. Також слід зазначити що дана задача створить гарні умови для дослідження особливості роботи методу MapReduce та створить умови для розширення сфер застосування даної технології.

1. Stonebraker M. MapReduce and parallel DBMSs: friends or foes? / M. Stonebraker, D. Abadi, D. J. DeWitt. – USA: Commun. ACM. – 2010. – Vol. 53. – №1. – P. 64-71.

2. Jimmy Lin Data-intensive Text Processing with MapReduce / Jimmy Lin, Chris Dyer –USA: Morgan & Claypool Publishers, – 2010. – P. 8-13.

3. Gonzalo Navarro. A Practical q-Gram Index for Text Retrieval Allowing Errors / Gonzalo Navarro, Ricardo Baeza-Yates. – Chile: CLEI Electronic Journal. – 1998. – Vol. 1 – P. 3-5.