УДК 004.6

Исследование применения распределенных систем для полнотекстового поиска библиотечных данных

Ф.М. Бикмуратов, К.А. Косолапов

Московский государственный университет им М.В. Ломоносова

Современные технологии все глубже проникают во все сферы деятельности так или иначе связанные с хранением информации. Там, где несколько десятков лет назад все обходилось бумажным архивом, сейчас хранится в электронном виде. Библиотеки в данном вопросе не являются исключением, большинство хранят полные тексты в своих базах данных. Это увеличивает производительность персонала библиотеки, а также открывает пространство для создания более удобных и продвинутых средств работы с самой библиотекой. Одной из самых актуальных задач является создание систем полнотекстового поиска среди гигантского набора текстов книг, содержащихся в библиотеке.

Задача полнотекстового поиска является хорошо изученной и существует масса средств и программных реализаций для ее решения. Все современные средства управления базами данных в той или иной степени имеют собственные решения по реализации полнотекстового поиска. Однако существуют и отдельные системы, которые поддерживают работу с удаленными базами данных.

Поддержка полнотекстового поиска обычно разбивается на следующие подзадачи: создание *полнотекстового индекса* – специального словаря, в котором для каждого слова указано в каких текстах и в каких местах оно встречается; *средства поддержки создания запросов* поиска – эти средства нормализуют текст запроса для более релаксированного поиска, в качестве нормализации могут быть использованы методы по удалению суффиксов слова для выделения главной части, а также отсеивание стоп слов.

В данной работе приводится сравнительный анализ имеющихся в мире средств для осуществления полнотекстового поиска на основе классических реляционных СУБД, а также и для нереляционных NoSQL баз данных. Одним из выводов этого анализа является преимущество использования NoSQL за счет высокой горизонтальной масштабируемости, позволяющей увеличивать производительность с помощью использования распределенных вычислений.

Большинство электронных библиотек использует реляционные СУБД и хранит данные о книгах - библиографические описания, в определенном формате. Таким образом,

для использования потенциала нереляционных баз данных требуется создания модели отображения данных, уже имеющихся в определенном формате в библиотеке, в данные типичные для NoSQL. В работе предоставляется модель отображения на примере конкретной библиотеки, а также приводится теоретическая оценка роста производительности системы поиска при росте числа узлов вычислительной системы.

В результате работы были исследованы технологии хранения данных, а также системы полнотекстового поиска, была реализована модель отображения библиографических данных из хранилища SQL в NoSQL. Предложена теоретическая оценка роста производительности. Дальнейшие направления работы связаны с практической реализацией системы полнотекстового поиска на нереляционных хранилищах и сравнение производительности с теоретической оценкой.

Литература

- *Adam Lith, Jakob Mattson* Investigating storage solutions for large data: A comparison of well performing and scalable data storage solutions for real time extraction and batch insertion of data. – Göteborg: Department of Computer Science and Engineering, Chalmers University of Technology. 2010. – 70 с.
- *B. Yuwono, D.L. Lee* Search and ranking algorithms for locating resources on the World Wide Web // 12th International Conference on Data Engineering (ICDE'96). – 1996. – С. 164-171.