УДК 004.65

РАЗРАБОТКА БЕНЧМАРКА ФРЕЙМВОРКОВ ДЛЯ РАБОТЫ С БОЛЬШИМ ОБЪЕМОМ ПРОСТРАНСТВЕННЫХ ДАННЫХ

Гараева А. А., Кабиров А. Д., Тихонова О. В.

Казанский Национальный Исследовательский Технический Университет им. Туполева

Сегодня обработка больших объемов пространственных данных в распределенных системах играет критическую роль во многих сферах нашей жизни. Большие данные часто неструктурированы и для их обработки требуются специальные алгоритмы. Одним из методов анализа больших данных является пространственный анализ. Источником больших данных в этом случае часто является географическая информационная система.

В данной статье разрабатывается бенчмарк для оценки фреймворков, работающих с такими данными. Также приводятся результаты оценки трех фреймворков, GeoSpark, STARK, SpecialSpark, с применением разработанного бенчмарка. В ходе данной работы мы рассматривали бенчмарк двух типов: макробенчмарк и микробенчмарк.

Основной целью нашей работы является тестирование топологических предикатов на различных топологических данных. Сравнение будет производится при помощи модели DE-9IM[1]. Это модель используется для определения типов топологических отношений (пересечение, равенство и др.).

Основной проблемой сравнения данных фреймворков является то, что они поддерживают не все операции выбранной модели. Это повлияло на формирования сценариев для микробенчмарка и макробенчмарка, так как невозможно было провести сравнение по всем пунктам DE-9IM.

Работа программы выполнялась на кластере, который состоит из 16 машин (CPU 2.90 ГГц, 4 ядра, оперативная память 16 Гб DDR3). Данные для тестирования программы состоят из набора точек и полигонов, имеющих нормальное распределение. Для измерения производительности мы использовали наборы данных, состоящих из 10000 (494KB), 50000 (2511KB), 100000 (5033KB), 250000 (13MB) точек, и состоящих из 10000 (19 MB), 50000 (96 MB), 100000 (192 MB), 250000 (480 MB) полигонов.

Фреймворки сравнивались по разным параметрам. На рис. 1 представлен результат выполнения макробенчмарка.

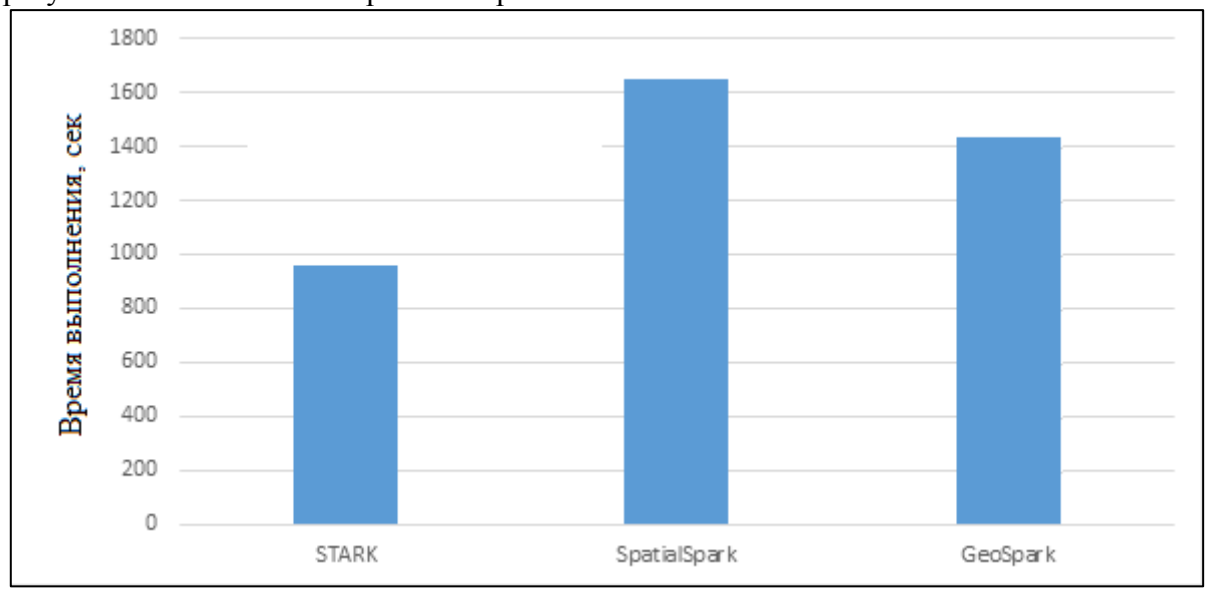

Рис.1. Результат макробенчмарка

На Рис. 2 приведён сравнение выполнения одной и той же модели отношений из DE-9IM на наборе данных разной размерности для фреймворка STARK.

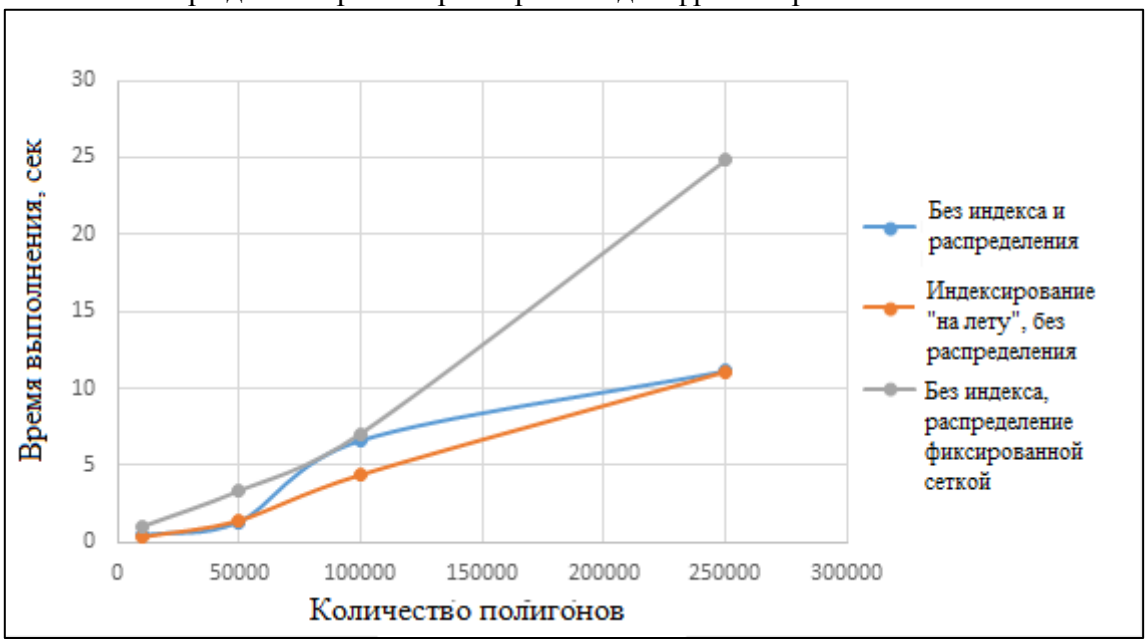

Рис.2. Результат выполнения операции фильтр для фреймворка STARK

По результатам нашей работы, STARK является наиболее подходящим фреймворком для работы с большими пространственными данными. SpatialSpark не имеет возможности быстрого индексирования, поэтому возникают проблемы с разработкой. У GeoSpark есть несколько видов распределения и индексирования , но количество пространственных отношений ограничено.

Библиографический список

1. Egenhofer, M.; Sharma, J.; Mark, D. A critical comparison of the 4-intersection and 9-intersection models for spatial relations: Formal analysis. In Proceedings of the AutoCarto Conference, Minneapolis, MN, USA, 30 October–1 November 1993; pp. 1–12.
2. Stefan Hagedorn, Philipp Goetze, Kai-Uwe Sattler; Big Spatial Data Processing Frameworks: Feature and Performance Evaluation EDBT, To appear, 2017.