УДК 519.683.8

КОНЦЕПЦИЯ ПРИМЕНЕНИЯ MAPREDUCE В ИЕРАРХИЧЕСКОЙ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ

С.А. Ермоченко
Учреждение образования «Витебский государственный университет имени П.М. Машерова»

В статье рассматриваются особенности иерархической агломеративной кластеризации и проблемы, связанные с необходимостью распараллелить этот процесс при использовании разных мер близости объектов.

Цель работы – выработка концепции применения модели MapReduce для иерархического агломеративного кластерного анализа большого объема данных.

Материал и методы. Материалом являются объекты произвольной природы, имеющие набор числовых характеристик и требующие выполнения их иерархической кластеризации. Особенность набора объектов – большое их количество (более 10 000). Использованы описательно-аналитический метод и метод проектирования распределенных вычислительных систем.

Результаты и их обсуждение. Для применения модели MapReduce в рассматриваемой задаче выделены операции, выполнение которых предлагается осуществлять на стадии предварительной обработки (Map), и операции для стадии свертки (Reduce). Достоинством предложенной концепции является возможность обработки большого числа объектов, информация о которых хранится в распределенных хранилищах данных. Результаты могут применяться на практике при проектировании вычислительных систем, ориентированных на конкретные предметные области.

Заключение. Предложена концепция использования модели MapReduce для выполнения иерархической агломеративной кластеризации в распределенной вычислительной системе, позволяющей гибкое горизонтальное масштабирование.

Ключевые слова: иерархическая агломеративная кластеризация, распределенные вычисления, модель MapReduce, обработка большого объема данных, мера близости объектов.

CONCEPT OF USING MAPREDUCE IN HIERARCHICAL AGGLOMERATIVE CLUSTERING

S.A. Yermochenko
Educational Establishment «Vitebsk State P.M. Masherov University»

Features of hierarchical agglomerative clustering and problems connected with the necessity to parallel this process while using different measures of object proximity are considered in the article.

The purpose is to elaborate the concept of using MapReduce model for hierarchical agglomerative clustering analysis of a large amount of data.

Material and methods. The material is objects of arbitrary nature which have a set of numerical characteristics and require their hierarchical clustering. The peculiarity of the set of objects is their large amount (over 10 000). The descriptive and analytical method and the method of designing distributed computing systems were used.

Findings and their discussion. To apply MapReduce model in the considered problem operations are singled out which are suggested to be executed at the stage of preliminary processing (Map) as well as operations for the closing stage (Reduce).

The advantage of the offered concept is the possibility to process a large amount of objects, information about which is stored in the distributed data bases. The results can be used in practice in computer system design, which aim at definite object areas.

Conclusion. A concept of using MapReduce model for performing hierarchical agglomerative clustering in the distributed computer system, which allows flexible horizontal scaling, is offered.

Key words: hierarchical agglomerative clustering, distributed computing, MapReduce model, a large amount of data processing, measure of object proximation.