MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
TERNOPIL IVAN PULUJ NATIONAL TECHNICAL UNIVERSITY
FOREIGN STUDENT DEPARTMENT
COMPUTER SCIENCE DEPARTMENT

AKOKHIA EMMANUEL OSHOKE

UDC 681.518

NETLIX FILM DATABASE CLUSTERING USING MAPREDUCE TECHNOLOGY

8.05010101 "Information Control System and Technologies "

The work at the Department of Manufacturing Engineering Ternopil Ivan Puluj National Technical University Ministry of Education and Science of Ukraine

Supervisor: Ph.D., assistant professor of mechanical engineering technology
Sitkar Taras,
Ternopil Ivan Puluj National Technical University,

Reviewer: Ph.D., assistant professor of design tools, instruments and machines
Ihor konovalenko,
Ternopil Ivan Puluj National Technical University,

Defence will be held in February 26, 2017 at 10.00 am at the meeting of the examination board №32 of Ternopil Ivan Puluj National Technical University at 46001, Ternopil, st. Ruska 56, educational building № 2, Aud. 702

GENERAL DESCRIPTION OF WORK

Actuality of work: Case details are intended to provide relative position with the required precision for details. Multithreading is one of the popular way of doing parallel programming, but major complexity of multi-thread programming is to co-ordinate the access of each thread to the shared data. We need things like semaphores, locks, and also use them with great care, otherwise dead locks will result. This is the fundamental concept of functional programming. Data is explicitly passed between functions as parameters or return values which can only be changed by the active function at that moment. Imagine functions are connected to each other via a directed acyclic graph. Since there is no hidden dependency (via shared state), functions in the DAG can run anywhere in parallel as long as one is not an ancestor of the other. In other words, analyze the parallelism is much easier when there is no hidden dependency from shared state. . Therefore, development of technological processes and design of body parts on their base production site is an actual scientific and practical problem that defined the direction of research thesis.

The purpose of the work: Hadoop got two important components namely hadoop distribured file system which is used for storing data and map reduce used for processing data . map reduce consists of mapping of related data using filtering and sorting algorithms (eg: filtering data based on likes in a post) and shuffling is an intermediate process which works according to the key vlaues based on mapping stage and then finally reduce stage does the actual working where summary operation takes places using merging algorithms (eg: counting the number of likes in each queue) . map reduce is influenced by functional programming.

Object, methods and sources of research. The main object of study is given technological process for processing data using MapReduce. Methods of work: economic and statistical, graphical, comparative, mathematical modeling; theoretical and empirical.

Scientific novelty of the results:

- The research features of the method of genetic algorithms to optimize data clustering;
- design and analysis service purpose facility production, the analysis of adaptability;
- chosen and designed the necessary technological equipment;
- completed a feasibility study of the decisions;
- The question of the use of information technology, safety, safety in emergencies is ecology;

The practical significance of the results.

Developed in process that can be implemented in a real production. The method of optimizing the layout for data clustering and mining.

Approbation. The results have been reported at V International scientific conference of young scientists and students << Current Issues in Modern Technologies >> Ternopil, November 17-18, 2016.

The structure of the work. The work consists of a cash-explanatory note and graphical part. Cash-explanatory note consists of an introduction, 5 parts, conclusions, list of

references and applications. Scope of work: settlement and explanatory note-__99_ pages. A4, graphic part- __7_ sheets A1

MAIN CONTENTS

In the introduction, a review of what data clustering is and gave a brief introduction on what MapReduce is and its application.

In the first part analysis of the issue according to the literature and other sources, the urgency of work done on the formulation of the problem thesis.

The research features of the method of genetic algorithms to optimize data processing.

As part of the project held designing manufacturing site for the implementation of the developed technological process implemented refine program production at the site, the calculation complexity and verstatomistkosti manufacturing products based on the developed processes, determine annual needs for technological equipment, preparation of summary information equipment, identification of quantitative structure of workers in the mechanical department, determination the size of the main and auxiliary space station shop and identifying key sizes and a choice of type and design of the building layout plan designed shop equipment layout, choice of load carried and vehicles.

As a special part The research capabilities of Hadoop MapReduce as a software framework is an open-source software framework used for distributed storage and processing of big data sets using the MapReduce programming model. It consists of computer clusters built from commodity hardware

As part of the "Substantiation of economic efficiency" The question of production and calculations conducted technical and economic efficiency of design solutions.

As part of "Ecology" analyzes the current status of Ukraine, the issues of pollution arising from the implementation process and proposed measures to reduce environmental pollution.

CONCLUSIONS

Adopted thesis work in analytic solutions demonstrated how MapReduce can be used to implement various data clustering algorithms. We have also shown how this clustered data may be used to provide movie predictions and recommendations for the Netflix problem. In the process we also demonstrated how MapReduce frameworks can collaborate with Database Management Systems allowing for interesting possibilities.

Economic efficiency calculations confirmed the correctness of the design decisions and showed that implementation of the new process reduced the cost of parts, improved loading details, decreased capital investment and improved a number of other technical and economic indicators.

List of published author of THE TOPIC OF WORK

1. AKOKHIA E.O, B.B. Mlynko. Mapreduce and its application in data clustering using netflix movie data /. Of V International scientific and technical conference of young researchers and students << Current Issues in Modern Technologies >>, Ternopil, November 17-18, 2016, vol.2. – Ternopil, TNTU press._7p._____

SUMMARY

In the thesis work was illustrated MapReduce is attractive because it abstracts parallel and distributed concepts in such a way that it allows novice programmers to take advantage of cluster computing without needing to be familiar with associated complexities such as data dependency, mutual exclusion, replication, and reliability. However, the challenge is that problems must be expressed in such a way that they can be solved using MapReduce. This often involves carefully designing inputs and outputs of MapReduce problems as often outputs of one MapReduce are used as inputs to another.

Keywords: TECHNOLOGY, MAPREDUCE, DATA CLUSTERING, K-MEANS, HADOOP, PARALLELIZATION