Research on Monte Carlo application based on Hadoop

Abstract. Monte Carlo method is also known as random simulation method. The more the number of experiments, the more accurate the results obtained. Therefore, a large number of random simulation is required in order to obtain a higher degree of accuracy, but the traditional stand-alone algorithm has been difficult to meet the needs of a large number of simulation. Hadoop platform is a distributed computing platform built on a large data background and an open source software under Apache. It is easier to write and run applications for processing massive amounts of data as an open source software platform. Therefore, this paper takes π value calculation as an example to realize the Monte Carlo algorithm based on Hadoop platform, and get the exact π value with the advantage of Hadoop platform in distributed processing.

1 Introduction

Monte Carlo method is also known as random simulation method and statistical test method, which is a technical methods based on statistical simulation or random simulation for solving the problem [1-3]. "Monte Carlo" is named after Monaco's famous casino - Monte Carlo, which is proposed by Metropolis, Ulam and Von Neumann who studied nuclear weapons during World War II. The basic idea of the Monte Carlo is to establish a probabilistic model or stochastic process, and calculate statistical characteristics of required parameters by observing the process or sampling test, and then use the arithmetic mean as the approximation of the solution [4-5]. In theory, the Monte Carlo method requires a lot of experiments. The more the number of experiments, the more accurate the results obtained. Therefore, a large number of random simulations is required in order to obtain a higher degree of accuracy, but the traditional stand-alone algorithm has been difficult to meet the needs of a large number of simulation.

Hadoop platform is a distributed computing platform built on a large data background and an open source software under Apache [6-9]. It is easier to write and run applications for processing massive amounts of data as an open source software platform. Therefore, Monte Carlo Method is put at the Hadoop platform, which makes it possible to use Monte Carlo

Method to deal with big data. At the same time, it is meaningful to improve the accuracy of the algorithm by using the advantages of Hadoop platform in distributed processing. Organization of the Text

2 Basic principle of Hadoop

The core idea of the Hadoop framework is Map / Reduce. Map / Reduce is a programming model for the calculation of large data volume. It is also an efficient task scheduling model. It divides a task into many fine-grained subtasks, which can be scheduled among idle processing nodes, so that the node with high- speed processing deals with more tasks, so as to avoid extending the completion of the entire task in that slow-speed processing node.

Calculative steps of Map/Reduce are as follows:

(1) The user program calls the MapReduce library, and the input large data set is divided into M data fragments;

(2) There are M Map tasks and R Reduce tasks to be assigned, and master assigns a Map task or Reduce task to an idle work site;

(3) The work site which is assigned a map task reads a data fragment to parse out key/value pairs. The user-defined Map function accepts an input key/value, and then produce a collection of middle key/value;

(4) MapReduce library brings together all intermediate 'value' values of the same intermediate key value I, and then transfer them to the reduce function;

(5) The user-defined Reduce function accepts a set of an intermediate key value I and a related value. The Reduce function combines these 'value' values into a set of smaller 'value' values. It normally outputs only 0 or 1 after calling the Reduce function each time. An intermediate 'value' value is supplied to the Reduce function through an iterator so that the collection of large amounts of 'value' values that can't be fully placed in memory is handled.

The MapReduce workflow is shown in Fig. 1.

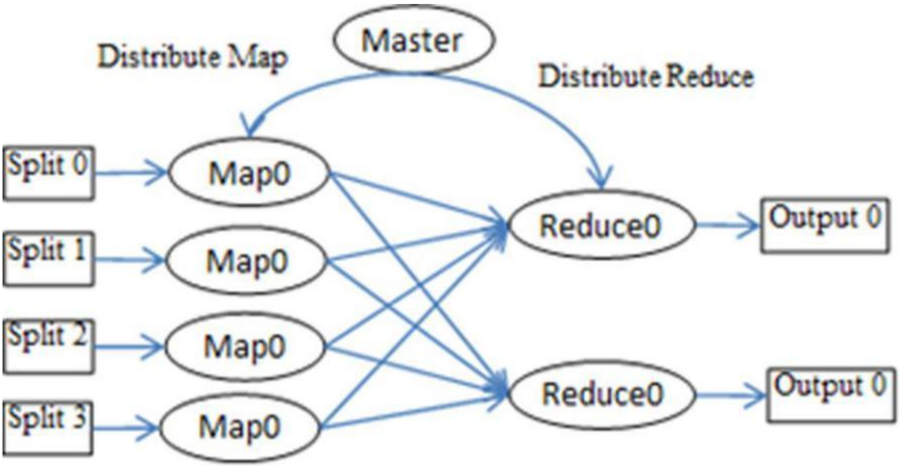

Fig. 1. The MapReduce Workflow

3 The proposed Algorithm

Monte Carlo has a high application value. Of course, it is also very significant in the field of mathematics research. The prototype of Monte Carlo method can be traced back to Buffon's needle problem in the late 19th century, which is an important problem in the classical probability that the value of π is determined by calculating the frequency of the needle onto the ground and combining the known exact relational expression. The Monte Carlo method is used to reproduce the value of π through imitating the idea of the ancients.

3.1 Basic idea

A circle B is got based on a square A (As is shown in Fig. 2). The area of A is given, so the area of B could be got by calculating the ratio of area (k=SB/SB). Then the value of π could be got by the formula (SB=πR2). Therefore, the value of k is the most critical.

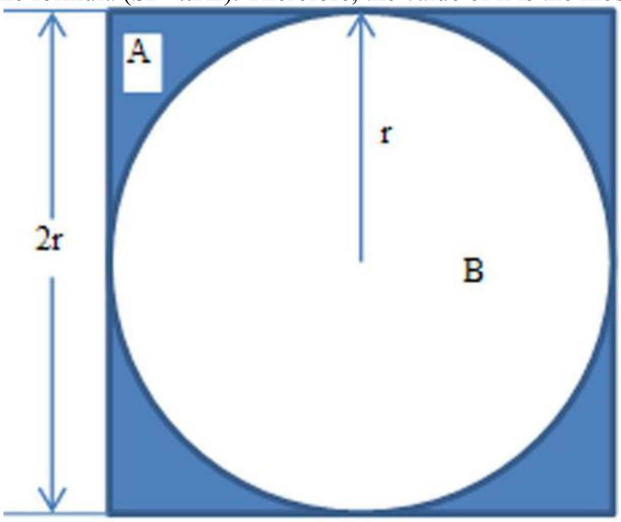

Fig. 3. Monte Carlo Simulation

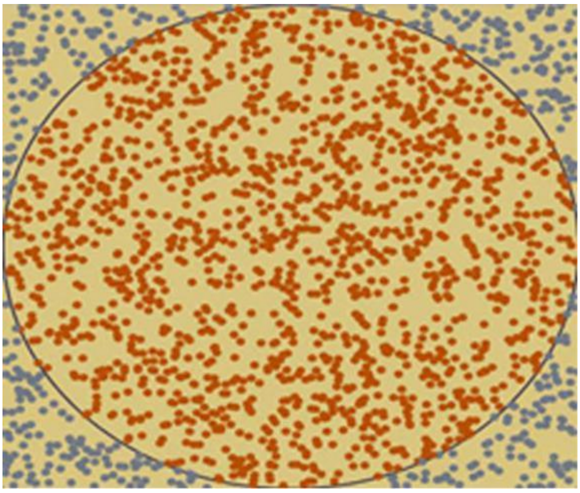

Fig. 2. The Square and The Inscribed Circle

A and B can be seen as an infinite number of points, and all points within B are within A. n points are randomly generated. If there are m points in B (The side length of A is assumed to be 10, and the center of A acts as the origin of the coordinate system. It falls in the B as long as the vertical and horizontal coordinates of the random point meet $x^2+y^2\leq100$), you can approximate the value of k (As is shown in Fig. 3), that is, the value of π can be obtained by k=m/n.

3.2 Algorithm flow

(1) 30 million between the 10 random points are generated with using the random function, and stored in three txt file;

(2) The point of the first step is read with using the map function, and the right point is found to output to the intermediate file;

(3) The intermediate file of the second step is read with using the reduce function, and the number of points within the circle is counted to calculate the value of π. (The algorithm flow is shown in Fig. 4)

3.3 Test environment and result

This experiment is mainly carried out on the Hadoop platform, and the configuration is as follows:

Hadoop cluster is composed of four computers, one of which is as a master node, and three other are as slave nodes. The operating system of node is centos7. The version of Hadoop is 2.7.2, 64 bit. The computer environment configuration is Intel (R) Core (TM) i7-3770 CPU 3.4GHZ, the memory is 8GB, and the operating system is 64-bit. Development tools: eclipse.

In this paper, 30 million between the 10 random points are generated with using the Monte Carlo random simulation. 3 nodes 30 million points (10 million points per node) are calculated respectively on 3 nodes with using Hadoop model, and the result is π = 3.14130944. The running time is 65000ms, and running time will have a certain difference in each run due to Hadoop's current situation. The more accurate of the value of π is seen from the data, but the time has no advantage. The reason of this result is relatively small experimental data, and the master node takes some time to distribute the task. Therefore, the advantage of cluster can’t be reflected when the data set is too small. The time of distribution can be ignored when the data set is relatively large, so the advantage of the cluster will be very obvious.

4 Summary

In this paper, the classical Monte Carlo stochastic algorithm is used to calculate the value of πunder the parallel environment of Hadoop. This parallelization and cluster environment can effectively deal with massive amounts of data, and it can be widely used to analyze in a variety of large data environment.

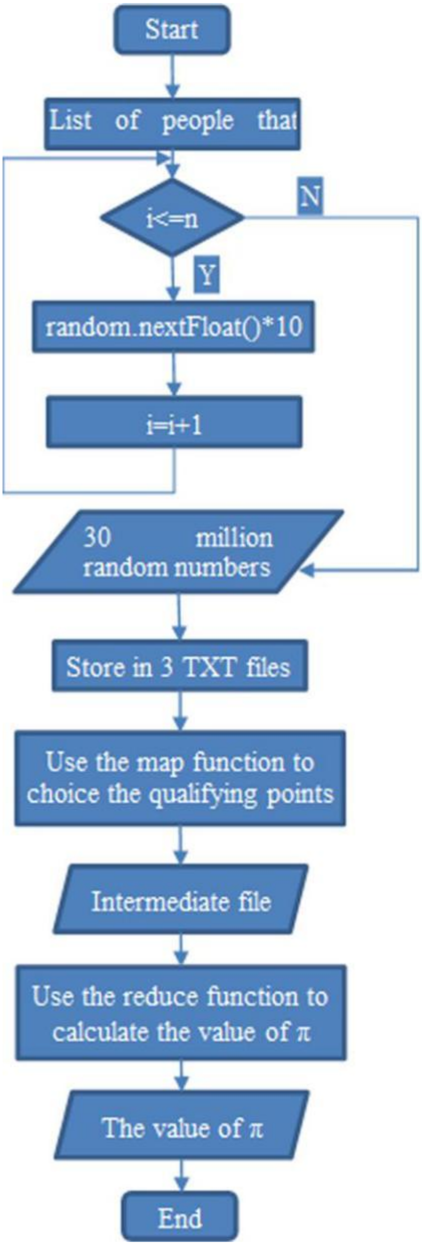
Start
List of people that
i<=n
N
Y
random.nextFloat()*10
i=i+1
30 million random numbers
Store in 3 TXT files
Use the map function to choice the qualifying points
Intermediate file
Use the reduce function to calculate the value of π
The value of π
End

Fig. 4. Algorithm Flow

This research was financially supported by the the National Natural Science Foundation of China (U1431102).