Received: 21 July 2020 | Revised: 15 October 2020 | Accepted: 7 November 2020

DOI: 10.1002/cpe.6096

RESEARCH ARTICLE

WILEY

K-ear: Extracting data access periodic characteristics for energy-aware data clustering and storing in cloud storage systems

Xindong You[1] | Tian Sun[1] | Dawei Sun[2] | Xunyun Liu[3] | Xueqiang Lv[1] | Rajkumar Buyya[4]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China

[2]School of Information Engineering, China University of Geosciences, Beijing, China

[3]Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Beijing, China

[4]Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia

Correspondence
Dawei Sun, School of Information Engineering, China University of Geosciences, Beijing, China.
Email: sundaweicn@cugb.edu.cn

Funding information
Australian Research Council (ARC) Discovery Project; National Language Committee of China, Grant/Award Number: ZDI135-53; National Natural Science Foundation of China, Grant/Award Numbers: 61671070, 61972364; Project of Developing University Intension for Improving the Level of Scientific Research, Grant/Award Number: 2019KYNH226; Qin Xin Talents Cultivation Program, Beijing Information Science & Technology University, Grant/Award Number: QXTCP B201908

Abstract

Rapid increase in energy consumption is a serious problem in cloud storage systems. Data accessed in large-scale storage systems usually exhibit temporal and spatial characteristics, which make it possible to reduce energy consumption by clustering data with similar access characteristics for storage in the same zone of cloud storage systems. Existing works usually only focus on the frequency of data access. However, widely existing phenomena show data access with seasonal and tidal characteristics in cloud storage systems. The seasonal and tidal characteristics of data access are extracted thoroughly in this paper. According to the extracted data access characteristics, energy-aware data clustering through a machine learning algorithm (K-ear) is proposed. K-ear classifies data into five seasonal categories according to their seasonal access characteristics and then classifies every seasonal category into three tidal categories according to its tidal access characteristics. The 15 classified categories are stored in different storage zones with different energy and performance modes. Simulation experiments using CloudSimDisk with the constructed mathematic models demonstrate that the proposed K-ear algorithm is more energy-efficient than the default data clustering algorithms in Hadoop and the classical data clustering storage strategy according to the data access frequency (Striping-Based Energy-Aware Strategy).

KEYWORDS

cloud storage system, data access characteristics, data clustering, energy consumption, seasonal characteristics, tidal characteristics

1 | INTRODUCTION

The exponential growth of the volume data is becoming one of the leading causes of high energy consumption in cloud storage systems.[1-3] It has been reported in the literature[4] that the energy consumed by data centers will be more than 1000 TWh during 2013–2025, which will surpass the total energy consumption of Japan and Germany. The energy consumed by data centers, including for cooling equipment, will consume 5% of the total energy consumption worldwide. An even more serious problem is that increasing energy consumption will produce high carbon and Greenhouse Gases emissions,[5,6] which will result in serious environmental pollution. Therefore, reducing energy consumption is one of the hottest issues in the cloud storage domain. Classifying data into different categories according to their access characteristics is an efficient way to enhance energy

efficiency in cloud storage systems. Different data categories are stored in different storage zones, running in different energy and performance modes. This kind of method will lead to less energy consumption in the storage zone, which has a low energy and performance mode while the workload is light. However, only the instant access frequency is considered in the current existing energy-aware data classification strategies. Long periodic access characteristics, such as seasonal characteristics and tidal characteristics, are not considered. Classifying data according to their instant access frequency will lead to frequent data migration, which will result in worse performance. Extracting the long periodic access characteristics is an effective solution, which could allow the low energy and performance running mode for a relatively long time. On the other hand, seasonal and tidal access characteristics obviously appear in cloud storage systems. As shown in Figure 1, the access frequencies (search index) of spring clothing, summer clothing, autumn clothing, and winter clothing exhibit obvious seasonal period characteristics. Other words with seasonal characteristics also have seasonal periodic access frequency.

As Figure 1 is the screenshot from the Baidu website, there are Chinese words in the figure. We list the Chinese words in the Figure 1 from left to right, and top to down, where "春装" means "spring clothings," "夏装"-"summer clothings," "秋装" –"autumn clothings," "冬装" winter clothings. "平均值" –"average value," "搜索指数"-"search indexes."

As Figure 2 is also the screenshot from the Baidu website, there are Chinese words in the figure. We list the Chinese words in the Figure 2 from left to right, and top to down, where "整体趋势" means "whole trend," "PC趋势"-"PC trend," "移动趋势" –"mobile trend," "最近"-"recent," "7天"-"7 days," "30天"-"30 days," "90天"-"90 days," "半 年"-"half a year," "全 部"-"all," "自 定 义"-"custom," "平 均 值" –"average value," "搜 索 指 数"-"search indexes."

As shown in Figure 2, the access frequency of work-related words has the same tidal periodic access characteristics: the peak points usually appear on workdays and the valley points almost always appear on weekends. This kind of weekly tidal period phenomenon always appears in work-related words, documents, and media.

As Figure 3 is also the screenshot from the Baidu website, there are Chinese words in the figure. We list the Chinese words in the figures from left to right, and top to down, where "整体趋势" means "whole trend," "PC趋势"-"PC trend," "移动趋势" –"mobile trend," "最近"-"recent," "7天"-"7 days,"

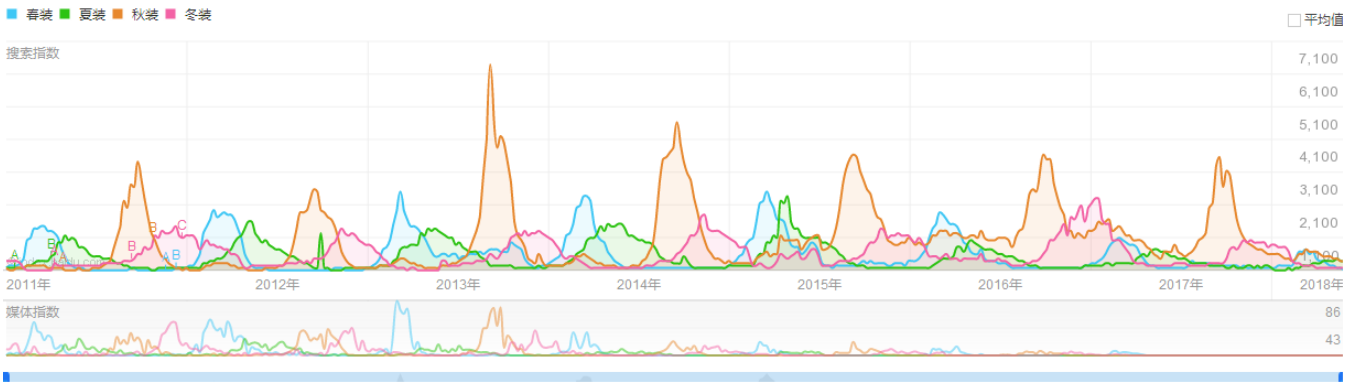

FIGURE 1 Varying curve of the search indexes of the different seasonal clothing in Baidu

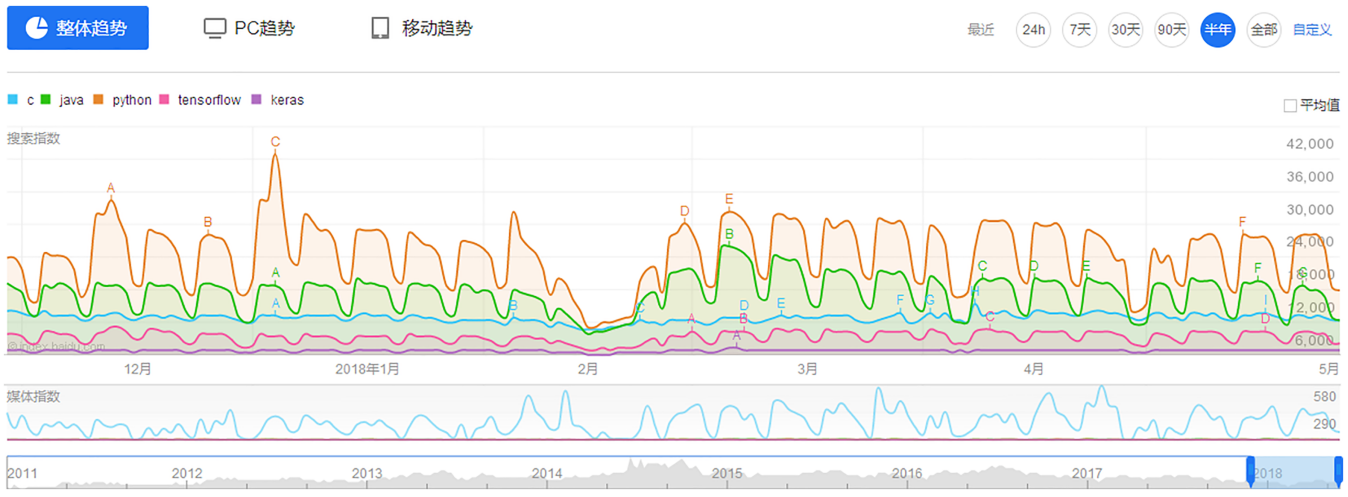

FIGURE 2 Varying curve of the search indexes of work-related words in Baidu

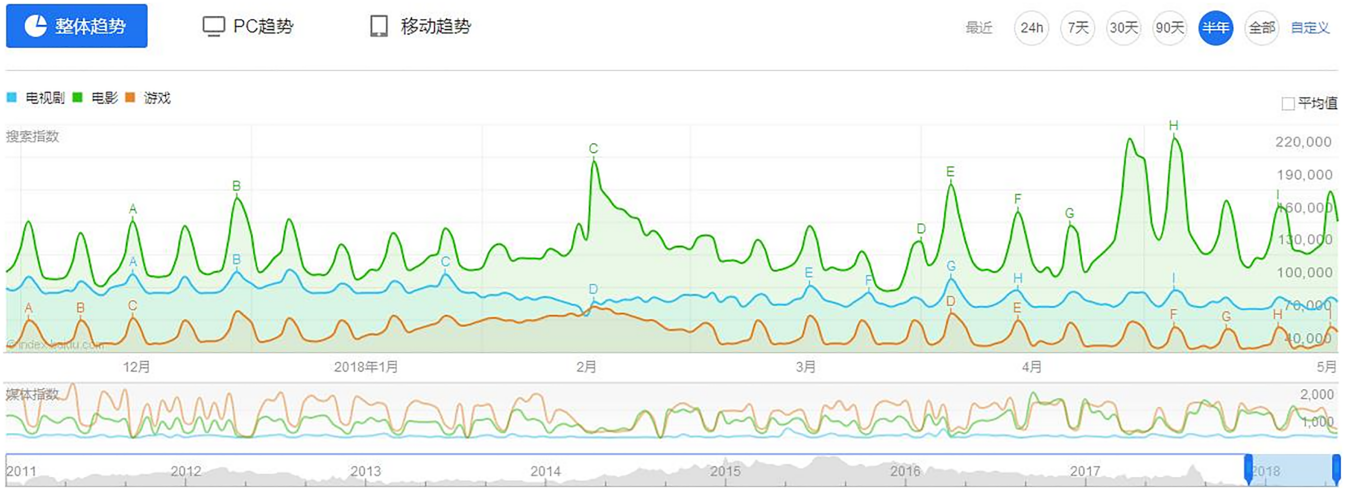

FIGURE 3 Varying curve of the search indexes of entertainment-related words in Baidu

"30天"-"30 days", "90天"-"90 days," "半年"-"half a year," "全部"-"all," "自定义"-"custom," "电视剧"-"TV," "电影"-"movie," "游戏"-"game," "平均值"-"average value," "搜索指数"-"search indexes."

On the other hand, as shown in Figure 3, the access frequency of entertainment-related words also has the same tidal periodic access characteristics. The peak points usually appear on weekends, and the valley points almost always appear on workdays. This kind of weekly tidal period phenomenon always appears in entertainment-related words, documents and media.

Recently, placing data according to the access characteristics is one of the main techniques for energy saving in large-scale storage systems.[7-10] However, only the temporal data access characteristics are considered in the existing literatures.

Based on the above observations and due to a lack of consideration of the periodic data access characteristics of the current existing energy-aware data classification work, we focused on extracting the seasonal and tidal access characteristics for energy-aware data clustering storage in this paper. In this direction, our paper makes the following key contributions:

1. Data seasonal and tidal access characteristics extracting algorithms (SCEA and TCEA) are designed, in which how to store the data access frequency and how to express the seasonal and tidal characteristics are described in detail.
2. The energy-aware data clustering strategy K-ear is designed, in which the framework of K-ear is first constructed and then the data are clustered by an unsupervised machine learning algorithm into different categories. Correspondingly, the cloud storage system is divided into 15 zones to store data with similar access characteristics.
3. The proposed K-ear algorithm is modeled, and the classical data classification SEA and the Hadoop default data placement algorithms are used for comparison through a mathematical method. The constructed mathematics model is the basis for analyzing the energy efficiency of the proposed K-ear strategy and is also used to conduct simulation experiments to verify the advantage of the proposed strategy in energy consumption.
4. Substantial simulation experiments are conducted using extended CloudSimDisk simulator, and the results demonstrate that the proposed K-ear strategy is more energy-efficient than the other two data placement algorithms.

The rest of the paper is organized as follows. In Section 2, we analyze the related energy-aware data classification work. The detailed framework and the algorithms of K-ear are described in Section 3. The mathematical model of the proposed K-ear strategy and the classical energy-aware data classification algorithm SEA are constructed for energy and performance analysis in Section 4. In Section 5, we demonstrate the detailed energy consumption and performance of the proposed K-ear, SEA and Hadoop default data storage algorithms. Conclusions and future work are presented in the final section.

2 | RELATED WORK

Energy-aware data classification strategies classify data into different categories according to the data access characteristics and then partition the storage system into different zones. Data with similar access characteristic clustering are stored in the same storage zone. Different zones run in different energy and performance modes. Energy consumption savings are achieved by managing the power state of the different zones. T. Xie first proposed a striping-based energy-aware data placement strategy (SEA) in the RAID (Redundant Arrays of Independent Disks) storage system,[11]

in which RAID disks are divided into Hot Disk Zone and Cold Disk Zone. Popular data are stored in the Hot Disk Zone while unpopular data are stored in the Cold Disk Zone. Disks in Hot Disk Zone run in a mode with a high transfer rate and high power rate, while disks in Cold Disk Zone run in a mode with a low transfer rate and low power rate. Analysis and simulation results show that the proposed SEA mechanism saves much energy consumption with little performance loss. Analyzing the access traces in Yahoo showed that the data access patterns in the Hadoop cluster have obvious heterogeneity. R.T. Kaushik et al. designed the GreenHDFS mechanism, in which data are classified into different categories according to temperature. The temperature of data is deduced by their availability and the user's performance requirement. Correspondingly, the Hadoop cluster is divided into Hot Zone and Cold Zone. Simulation experiments conducted on the generated workload based on a real trace of 3 months of data from Yahoo demonstrate that 26% of energy consumption can be reduced by only managing the cold zone at a lower power consumption rate while the system load is light.[12] A similar energy-aware data classification policy, Lightning, was designed by the Kaushik team[13] to reduce the energy consumption of the Yahoo cloud storage system. Inspired by GreenHDFS and the lighting mechanism, we have proposed a green data classification strategy based on anticipation named AGDC, in which the neural network is employed to predict the temperature of data. Based on the predicted temperature of data, they are classified into three categories: cold data, seasonal hot data, and hot data. Correspondingly, the cloud storage system is divided into different zones. Simulation experiments conducted on the GridSim simulator demonstrated that the AGDC mechanism can save approximately 16% of energy consumption at the expense of an increased average response time of 0.005 s. AGDC outperformed the integration general classification algorithm TDCS.[14] In the literature,[15] data are classified into different categories according to their access frequency and regularity. RACK is divided into Active-Zone and Sleep-Zone. Data with different access characteristics are stored in different zones. Simulation experimental results in the MATLAB and GridMix environments show that the energy consumption saved by the proposed algorithm is up to 39.01%. An energy-efficient algorithm that classifies data in the cloud storage system was proposed by Z. Tao et al.[16] that divided the cloud storage system into HotZone, ColdZone, and Reduplication Zone. Data are stored in the zones according to their repetition and activity factor characteristics. The experimental results show that the energy utilization rate improved by 25%. Furthermore, the proposed algorithm performs better when the system load is light. Recently, the SLA (Service Level Agreement) has been considered in more and more literatures for trading off the energy efficiency and the QoS. A dynamic data aggregation algorithm for green cloud computing is proposed in Reference 17. According to the data access pattern, the data are aggregately stored dynamically among nodes. By managing the power state of the storage nodes, energy consumption can be reduced along with the QoS considered. Dr. Long designed static and dynamic file layout and replica and data layout policies to reduce the energy consumption in cloud storage systems.[18] The static file layout strategy (SFLS) first divides data into hot files and big files according to their access frequency and service time. Correspondingly, the disks are divided into different groups. I/O requests are distributed to the different disk groups according to the access frequency and service time. The results obtained from the CloudSim simulator demonstrate that SFLS can save power consumption by over 35% compared to the default HDFS. Moreover, R. Yadav et al. published the related article for minimizing energy consumption and SLA violation in cloud computing.[19-22] Other data placement or data layout strategies for energy efficiency have been published in recent years,[15,19-33] but the period access characteristics also have not been extracted for data clustering storing.

As described before, data classification is an efficient way to reduce energy consumption. However, only the instant data access frequency is considered in all of the above energy-aware data classification strategies, which may cause frequent data migration and result in performance loss. The seasonal and tidal periodic access characteristics of data are thoroughly extracted for data classification storing in our proposed K-ear strategy, which can leave the nodes in a low power state for a relatively long time, leading to energy consumption savings.

A summary of the reviewed related work is presented in Table 1, comparing in terms of whether the access frequency or the periodic access characteristics (seasonal or tidal periodic access characteristics) has been considered or not, the zones have been divided.

TABLE 1 Summary of the reviewed related work compared with our work

			Periodic access characteristics		
Work	**Name**	**Access frequency**	**Seasonal characteristics**	**Tidal characteristics**	**Zones divided**
T. Xie[11]	SEA	✓	–	–	Hot Disk Zone, Cold Disk Zone
R.T. Kaushik et al[12]	GreenHDFS	✓	–	–	Hot Zone, Cold Zone
R.T. Kaushik et al[13]	Lightning	✓	–	–	Hot Zone, Cold Zone
X.D. You et al.[14]	AGDC	✓	✓	–	Cold Zone, Seasonal Hot Zone, Hot Zone
Z. Tao et al.[16]	–	✓	–	–	HotZone, Cold Zone and Reduplication Zone
X. L. Xu[17]	–	✓	–	–	Dynamically aggregating
S. Q. Long[18]	SFLS	✓	–	–	Hot Files Zone, Big Files Zone
Our work	K-ear	✓	✓	✓	Five big zones (according to seasonal characteristics) with 15 small zones (according to tidal characteristics furthermore)

3 | K-EAR: ENERGY-AWARE DATA CLUSTERING STRATEGY

The proposed K-ear strategy consists of two data access characteristics extraction algorithms (Seasonal Characteristics Extracting Algorithm SCEA, and Tidal Characteristics Extracting Algorithm TCEA) and a framework, which are described in the following subsections.

3.1 | Periodic data access characteristics extraction

The data seasonal and tidal access characteristics are extracted by the SCEA and TCEA algorithms, respectively. To explain the SCEA and TCEA algorithms clearly, some definitions are given first.

Representation Dataset: $D = \{d_1, d_2, \cdots, d_m\}$, which is the representation data set of the primary data for clustering, and m is the number of data.

Data seasonal characteristics: They are represented by the vector $\mathrm{SE} = \begin{bmatrix} \mathrm{Se}_1 \\ \mathrm{Se}_2 \\ \vdots \\ \mathrm{Se}_m \end{bmatrix}$. Assume that y is the number of years of data to be collected and that there are four seasonal search index ratios for each year. Therefore, $\mathrm{Se}_i = [se_1, se_2, \cdots, se_{4*y}]$.

Data tidal characteristics: They are represented by the vector $\mathrm{CX} = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_m \end{bmatrix}$ and $cx_i = \begin{bmatrix} \mathrm{p}_{i,1} & \mathrm{v}_{i,1} & \mathrm{p}_{i,2} & \mathrm{v}_{i,2} & \cdots & \mathrm{p}_{i,z} & \mathrm{v}_{i,z} \end{bmatrix}$, in which z is the number of weeks of representation data (there are 52 weeks in 1 year).

The SCEA is described in Algorithm 1, through which the seasonal characteristics of the data stored in the SE vector.

TCEA is described in Algorithm 2, through which the seasonal characteristics of the data are stored in the CX vector.

Algorithm 1. : Seasonal characteristics extracting algorithm

Input: Data representation Set $D = \{d_1, d_2, \cdots, d_m\}$;
The number of the weeks y;

Output: Data Seasonal Characteristic $\mathrm{SE} = \begin{bmatrix} \mathrm{Se}_1 \\ \mathrm{Se}_2 \\ \vdots \\ \mathrm{Se}_{m.} \end{bmatrix}$

Begin

1: **for** each data $d_i \in D$ do
2: Parse the image data from Baidu index page into the Search Index Number of 1 week: $\mathrm{ZS} = \{\mathrm{zs}_1, \mathrm{zs}_2, \cdots, \mathrm{zs}_{y*52}\}$;
3: Calculate the sum of searching index for each season;
4: Initialize the sum_of_season vector to zero;
5: **for** (j=1; $j \le y$*4; j++) // If there are y years, there are y*4 seasons.
6: sum_of_season[j]=0
7: **End for**
8: j=1; from the first year
9: **for**(k=1; $k \le y$*52; k++) //Travel every weeks to calculate the search index number of every seasons in different years
10: sum_of_season[j]+=zs[k] //calculate the index number of every seasons while k is the times of 12.
11: **if** the season ending, Thas is if (k%13==0)
12: go to next season j++;
13: **End if**
14: **End fo**r
15: Calculate the sum of searching index for each year
16: Initialize the sum_of_year vector to zero.
17: **for** (j=1; j$\le y$; j++)
18: sum_of_year[j]=0
19: **End for**
20: i=1; from the first season
21: **for**(k=1; $k \le y$*4; k++) // travel the seasons to calculate the search index of every years.

sum_of_year[i]+=sum_of_season[k]
if the year ending, Thas is if (k%4==0)
go to the next season j++;
End if
End for
Calculate the frequency of searching index for each season
Initialize the frequency_of_season vector to zero
for(k=1; $k \leq y^*4$; k++)
se[k]=0.0;
End for
for(k=1; $k \leq y^*4$; k++)
se[k]=sum_of_season[k]/sum_of_year[k/4+1]// Calculate the frequency of searching index for each season
End for
End for
End

Algorithm 2. : Tidal characteristics extracting algorithm

Input: Data representation Set $D = \{d_1, d_2, \cdots, d_m$
The number of the weeks z;

Output: Data Tidal Characteristics $CX = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_m \end{bmatrix}$

Begin
for each data $d_i \in D$ do
Parse the image data from Baidu index page to the Search Index Number of 1 day: $S = \{s_1, s_2, \cdots, s_{z^*7}\}$;
Find the peak and the valley value of every week;
Initialize the first week k=1 and the index of first week's tidal value x=1;
for j=1; $j \leq z^*7$; j++
Record the index of the weak peak and valley value
Initial value is: p_index=k; v_index=k;
if (s[j]>s[p_index]) p_index=j; // Record the index of the weak peak
End if
if (s[j]<s[v_index]) v_index=j; // Record the index of the valley value
End if
if(j%7==0) //new week will begin, record the week tidal value
if peak value is on the working days (from Monday to Thursday)
That is (p_index-k)%7=1or 2 or 3 or 4 $cx_i[x] = 1$
End if
if peak value is on the weekends (Saturday or Sunday)
That is (p_index-k)%7=0 or 6 $cx_i[x] = 2$
End if
if peak value is on Friday
That is (p_index-k)%7=5 $cx_i[x] = 3$
End if
if valley value is on the working days (from Monday to Thursday)
That is (p_index-k)%7=1or 2 or 3 or 4 $cx_i[x+1] = -1$
End if
if valley value is on the weekends (Saturday or Sunday)
That is (p_index-k)%7=0 or 6 $cx_i[x+1] = -2$
End if

27: **if** valley value is on Friday
28: That is (p_index-k)%7=5 $cx_i[x+1] = -3$
29: **End if**
30: **End if**
31: Reset the initial peak and valley value index of the new week k=j+7;
32: Set the index of the new week's tidal value x=x+2;
33: **End for**
34: **End for**
End

3.2 | Framework and strategy of K-ear

The main procedures of the proposed K-ear are delineated in Figure 4. The seasonal access characteristics of data are extracted by the SCEA algorithm (see Algorithm 1 for details). Then, the machine learning clustering-related algorithm is employed on the extracted seasonal access characteristics. Data are clustered into five categories: spring data, summer data, autumn data, winter data, and other data. Correspondingly, the cloud storage system is divided into five zones: the spring zone, summer zone, autumn zone, winter zone, and other zones. The different data categories are stored in the corresponding storage zones. When entering a certain zone, the TCEA (Algorithm 2 for detail) is utilized to extract the tidal access characteristics of data. Similarly, the machine learning clustering-related algorithm is employed on the extracted tidal access characteristics. The data are further clustered into three categories: work-related data, entertainment data, and other data. Therefore, the cloud storage system is partitioned into 15 zones: the spring-work zone, spring-entertainment zone, spring-other zone, summer-work zone, summer-entertainment zone, summer-other zone, autumn-work zone, autumn-entertainment zone, autumn-other zone, winter-work zone, winter-entertainment zone, winter-other zone, other-work zone, other-entertainment zone, and other-other zone. We assume that a disk has two speed modes: a high-speed mode with a high transfer rate and a high energy-consuming rate and a low-speed mode with a low transfer rate and a low energy consumption rate, as it is a common method for modeling hard-disk power consumption.[11,34-36] According to the different periods of the year, different storage zones run in different speed modes. For example, on a workday in the spring, the spring-work zone, other-work zone and other-other zone run in the high speed mode for performance consideration. The other zones run in low-speed mode to save energy consumption.

The symbols used in the proposed K-ear algorithm are explained in Tables 2 and 3.

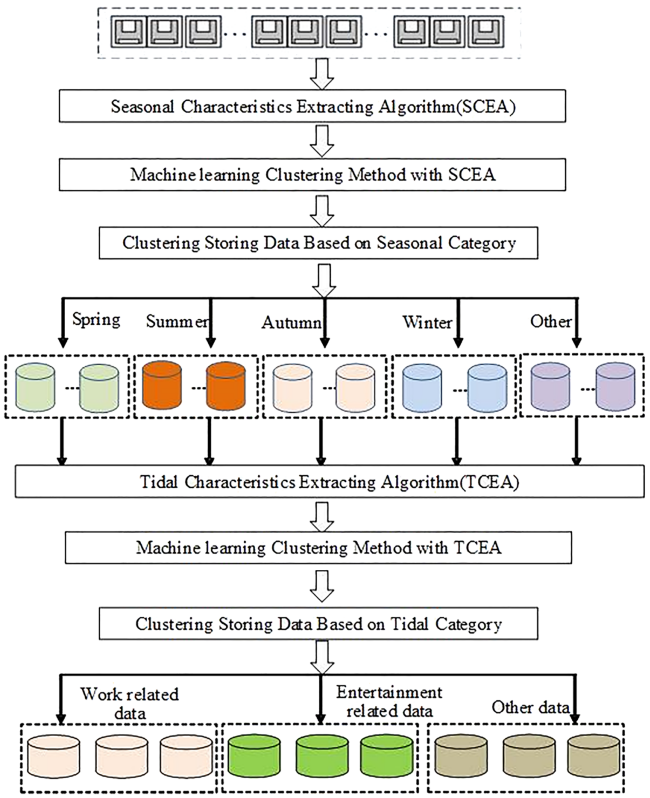

FIGURE 4 Framework of the K-ear strategy

TABLE 2 Symbols in algorithms related to the seasonal data proportion

Symbol	Meaning	Symbol	Meaning
η_s	Ratio of data with spring characteristics	η_{s-w}	Ratio of data with Spring and work characteristics
		η_{s-h}	Ratio of data with Spring and entertainment characteristics
		η_{s-o}	Ratio of data with Spring but without characteristics
η_o	Ratio of data without seasonal characteristics	η_{o-w}	Ratio of data with work but without seasonal characteristics
		η_{o-h}	Ratio of data with entertainment but without seasonal characteristics
		η_{o-o}	Ratio of data without seasonal and tidal characteristics

TABLE 3 The number of disks to store the data with tidal characteristics of every seasonal zone

Parameters	Meanings
$\eta_{s-w} \times n_1$	The number of disks to store the data with spring and work characteristics
$\eta_{s-h} \times n_1$	The number of disks to store the data with spring and entertainment characteristics
$\eta_{s-o} \times n_1$	The number of disks to store the data with spring but without tidal characteristics
$\eta_{o-w} \times n_5$	The number of disks to store the data with work but without seasonal characteristics
$\eta_{o-h} \times n_5$	The number of disks to store the data with entertainment but without seasonal characteristics
$\eta_{o-o} \times n_5$	The number of disks to store the data without work seasonal characteristics

When the *s* in the notation η_s is replaced, respectively, by *m*, *a*, and *w*, notations meaning the related characteristics of the Summer, Autumn, and Winter, respectively.

And the above symbols have the following relationships: $\eta_s + \eta_m + \eta_a + \eta_w + \eta_o = 1, \eta_{s-w} + \eta_{s-h} + \eta_{s-o} = 1, \eta_{m-w} + \eta_{m-h} + \eta_{m-o} = 1, \eta_{a-w} + \eta_{a-h} + \eta_{a-o} = 1, \eta_{w-w} + \eta_{w-h} + \eta_{w-o} = 1, \eta_{o-w} + \eta_{o-h} + \eta_{o-o} = 1$.

Assume the number of the disks is *n*, and the disks set is defined as below. $K = \{s_1, s_2, \cdots, s_{n1}, m_1, m_2, \cdots m_{n2}, a_1, a_2, \cdots a_{n3}, w_1, w_2, \cdots, w_{n4}, o_1, o_2, \cdots o_{n5}\}$

And the set $S = \{s_1, s_2, \cdots, s_{n1}\}$ $M = \{m_1, m_2, \cdots m_{n2}\}$ $A = \{a_1, a_2, \cdots a_{n3}\}$ $W = \{w_1, w_2, \cdots, w_{n4}\}$ $O = \{o_1, o_2, \cdots o_{n5}\}$ represent the disks store the data with spring, summer, autumn, winter and no seasonal characteristics, respectively, in which $n_1 = \eta_s \times n, n_2 = \eta_m \times n, n_3 = \eta_a \times n, n_4 = \eta_w \times n, n_5 = \eta_o \times n$. And the number of the disks to store the data with tidal characteristics of every seasonal zone are listed in Table 2.

In Table 3, when the *s* in the notation η_s is, respectively, replaced by *m*, *a*, and *w*, && n_1 is, respectively, replcaed by n_2, n_3, n_4 notations meaning the number of the disk to store the data with related characteristics of the Summer, Autumn, and Winter, respectively.

The detailed procedures of the K-ear strategy are illustrated in Algorithm 3.

According to the Algorithm 3, the number of disks running in different modes during different time zones are listed in Table 4.

Algorithm 3. : Energy-aware data classification based on K-means (K-ear)

Input: Data-related characteristics: a collection of *m* data in the set *D*, and the corresponding parameters: $\eta_s, \eta_m, \eta_a, \eta_w, \eta_o$ *and* $\eta_{s-w}, \eta_{s-h}, \eta_{s-o}, \eta_{m-w}, \eta_{m-h}, \eta_{m-o}, \eta_{a-w}, \eta_{a-h}, \eta_{a-o}, \eta_{w-w}, \eta_{w-h}, \eta_{w-o}, \eta_{o-w}, \eta_{o-h}, \eta_{o-o}$.

The exacted data characteristics: the output of the Algorithm 1 and Algorithm 3: Data Tidal Characteristics $CX = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_m \end{bmatrix}$, and the Data Seasonal Characteristics $SE = \begin{bmatrix} Se_1 \\ Se_2 \\ \vdots \\ Se_m \end{bmatrix}$

Disk-related characteristics: A disk array *DISK* with *n* size, every disk with 2-speed mode.

Output: Allocation of the data on the corresponding storage zones.

Begin

1: **for** each data $d_i \in D$ do

2: Based on the Data Seasonal Characteristics $SE = \begin{bmatrix} Se_1 \\ Se_2 \\ \vdots \\ Se_m \end{bmatrix}$, and use the K-means clustering algorithm to classify the data into five classes firstly;

3: **if** d_i has the spring characteristics (**Class 1**)

4: place it evenly into the following storage zone $S = \{s_1, s_2, \cdots, s_{n1}\}$ **(Storage Zone 1)**

5: based on the Data Tidal Characteristics $CX_1 = \begin{bmatrix} cx_{11} \\ cx_{12} \\ \vdots \\ cx_{1\eta_s \times m} \end{bmatrix}$, and use the K-means clustering algorithm to classify the data into three classes;

6: **If** d_i has the working day characteristics (**Sub_Class 1-1**)

7: place the data into the $1 \sim \eta_{s-w} \times n_1$ disks evenly**(Storage Zone 1-1**) // disks numbered by ordering

8: **End if**

9: **If** d_i has the holiday characteristics(**Sub_Class 1-2)**

10: place the data into the $\eta_{s-w} \times n_1 + 1 \sim \eta_{s-w} \times n_1 + \eta_{s-h} \times n_1$ disks evenly; **(Storage Zone 1-2)** // disks numbered by ordering

11: **End if**

12: **If** d_i has no working and holiday characteristics (**Sub_Class 1-3**)

13: place the data into the $\eta_{s-w} \times n_1 + \eta_{s-h} \times n_1 + 1 \sim n_1$ disks evenly **(Storage Zone 1-3)** // disks numbered by ordering

14: **End if**

15: **End if**

16: **if** d_i has the summer characteristics (**Class 2**)

17: place it evenly into the following storage zone $M = \{m_1, m_2, \cdots m_{n2}\}$; **(Storage Zone 2)**

18: based on the Data Tidal Characteristics $CX_2 = \begin{bmatrix} cx_{21} \\ cx_{22} \\ \vdots \\ cx_{2\eta_m \times m} \end{bmatrix}$, and use the K-means clustering algorithm to classify the data into three classes;

19: **If** d_i has the working day characteristics (**Sub_Class 2-1**)

20: place the data into the $1 \sim \eta_{m-w} \times n_2$ disks evenly; // disks numbered by ordering **(Storage Zone 2-1)**

21: **End if**

22: **If** d_i has the holiday characteristics (**Sub_Class 2-2**)

23: place the data into the $\eta_{m-w} \times n_2 + 1 \sim \eta_{m-w} \times n_2 + \eta_{m-h} \times n_2$ disks evenly; **(Storage Zone 2-2)** // disks numbered by ordering

24: **End if**

25: **If** d_i has no working and holiday characteristics (**Sub_Class 2-3**)

26: place the data into the $\eta_{m-w} \times n_2 + \eta_{m-h} \times n_2 + 1 \sim n_2$ disks evenly; // disks numbered by ordering **(Storage Zone 2-3)**

27: **End if**

28: **End if**

29: **if** d_i has the autumn characteristics (**Class 3**)

30: place it evenly into the following storage zone $A = \{a_1, a_2, \cdots a_{n3}\}$; **(Storage Zone 3)**

31: based on the Data Tidal Characteristics $CX_3 = \begin{bmatrix} cx_{31} \\ cx_{32} \\ \vdots \\ cx_{3\eta_a \times m} \end{bmatrix}$, and use the K-means clustering algorithm to classify the data into three classes;

32: **If** d_i has the working day characteristics (**Sub_Class 3-1**)

33: place the data into the $1 \sim \eta_{a-w} \times n_3$ disks evenly; // disks numbered by ordering **(Storage Zone 3-1)**

34: **End if**

If d_i has the holiday characteristics (**Sub_Class 3-2**)
  place the data into the $\eta_{a-w} \times n_3 + 1 \sim \eta_{a-w} \times n_3 + \eta_{s-h} \times n_3$ disks evenly; **(Storage Zone 3-2)**
    // disks numbered by ordering
End if
If d_i has no working and holiday characteristics (**Sub_Class 3-3**)
    place the data into the $\eta_{a-w} \times n_3 + \eta_{s-h} \times n_3 + 1 \sim n_3$ disks evenly // disks numbered by ordering
      (Storage Zone 3-3)
End if
End if
if d_i has winter characteristics; (**Class 4**)
  place it evenly into the following storage zone $W = \{w_1, w_2, \cdots, w_{n4}\}$
    (Storage Zone 4)
  based on the Data Tidal Characteristics $CX_4 = \begin{bmatrix} cx_{41} \\ cx_{42} \\ \vdots \\ cx_{4_{\eta_w} \times m} \end{bmatrix}$, and use the K-means clustering algorithm to classify the data into three classes;
    If d_i has the working day characteristics (**Sub_Class 4-1**)
      place the data into the $1 \sim \eta_{w-w} \times n_4$ disks evenly // disks numbered by ordering;
        (Storage Zone 4-1)
    End if
      If d_i has the holiday characteristics (**Sub_Class 4-2**)
      place the data into the $\eta_{w-w} \times n_4 + 1 \sim \eta_{w-w} \times n_4 + \eta_{w-h} \times n_4$ disks evenly; **(Storage Zone 4-2)**
        // disks numbered by ordering
    End if
    If d_i has no working and holiday characteristics (**Sub_Class 4-3**)
      place the data into the $\eta_{w-w} \times n_4 + \eta_{w-h} \times n_4 + 1 \sim n_4$ disks evenly // disks numbered by ordering
        (Storage Zone 4-3)
    End if
End if
if d_i has no seasonal characteristics (**Class 5**)
  place it evenly into the following storage zone $O = \{o_1, o_2, \cdots o_{n5}\}$;
  (Storage Zone 5)
  based on the Data Tidal Characteristics $CX_5 = \begin{bmatrix} cx_{51} \\ cx_{52} \\ \vdots \\ cx_{5_{\eta_o} \times m} \end{bmatrix}$, and use the K-means clustering algorithm to classify the data into three classes;
    If d_i has the working day characteristics (**Sub_Class 5-1**)
      place the data into the $1 \sim \eta_{o-w} \times n_5$ disks evenly; // disks numbered by ordering
        (Storage Zone 5-1)
    End if
      If d_i has the holiday characteristics (**Sub_Class 5-2**)
      place the data into the $\eta_{o-w} \times n_5 + 1 \sim \eta_{o-w} \times n_5 + \eta_{o-h} \times n_5$ disks evenly; **(Storage Zone 5-2)**
        // disks numbered by ordering
    End if
    If d_i has no working and holiday characteristics (**Sub_Class 5-3**)
      place the data into the $\eta_{o-w} \times n_5 + \eta_{o-h} \times n_5 + 1 \sim n_5$ disks evenly; // disks numbered by ordering
      (Storage Zone 5-3)
    End if
End if
End for
End

TABLE 4 The number of disks running in different modes during different time zones

Time zone	The number of disks running in high mode	The number of disks running in low mode
Workdays in Spring	$H1 = \eta_{s-w} \times n_1 + \eta_{s-o} \times n_1 + \eta_{s-w} \times n_5 + \eta_{s-o} \times n_5$	$L1 = \eta_{s-h} \times n_1 + n_2 + n_3 + n_4 + \eta_{s-h} \times n_5$
Weekends in Spring	$H2 = \eta_{s-h} \times n_1 + \eta_{s-o} \times n_1 + \eta_{s-h} \times n_5 + \eta_{s-o} \times n_5$	$L2 = \eta_{s-w} \times n_1 + n_2 + n_3 + n_4 + \eta_{s-w} \times n_5$
Workdays in Summer	$H3 = \eta_{s-w} \times n_2 + \eta_{s-o} \times n_2 + \eta_{s-w} \times n_5 + \eta_{s-o} \times n_5$	$L3 = \eta_{s-h} \times n_2 + n_1 + n_3 + n_4 + \eta_{s-h} \times n_5$
Weekends in Summer	$H4 = \eta_{s-h} \times n_2 + \eta_{s-o} \times n_2 + \eta_{s-h} \times n_5 + \eta_{s-o} \times n_5$	$L3 = \eta_{s-w} \times n_2 + n_1 + n_3 + n_4 + \eta_{s-w} \times n_5$
Workdays in Autumn	$H5 = \eta_{s-w} \times n_3 + \eta_{s-o} \times n_3 + \eta_{s-w} \times n_5 + \eta_{s-o} \times n_5$	$L5 = \eta_{s-h} \times n_3 + n_1 + n_2 + n_4 + \eta_{s-h} \times n_5$
Weekends in Autumn	$H6 = \eta_{s-h} \times n_3 + \eta_{s-o} \times n_3 + \eta_{s-h} \times n_5 + \eta_{s-o} \times n_5$	$L6 = \eta_{s-w} \times n_3 + n_1 + n_2 + n_4 + \eta_{s-w} \times n_5$
Workdays in Winter	$H7 = \eta_{s-w} \times n_4 + \eta_{s-o} \times n_4 + \eta_{s-w} \times n_5 + \eta_{s-o} \times n_5$	$L7 = \eta_{s-h} \times n_4 + n_1 + n_2 + n_3 + \eta_{s-h} \times n_5$
Weekends in Winter	$H8 = \eta_{s-h} \times n_4 + \eta_{s-o} \times n_4 + \eta_{s-h} \times n_5 + \eta_{s-o} \times n_5$	$L8 = \eta_{s-w} \times n_4 + n_1 + n_2 + n_3 + \eta_{s-w} \times n_5$

4 | MATHEMATIC MODELING

To analyze the energy efficiency advantage of our proposed K-ear strategy, we model the K-ear mathematically in this section. The parameters used during mathematical modeling are listed in following tables. Parameters of the whole system are listed in Table 5. Parameters about the spring season are listed in Table 6. Parameters about the summer season, autumn season, and winter season will be explained correspondingly.

For short, when the *spring* in the labels of the Table 6 is replaced by the *summer*, *autumn*, and *winter*, respectively, the parameters are about the Summer season, Autumn season and Winter season, respectively. Such as *spring_work*_t_h^{active} means the total time of disks running in high mode with active status during workdays in Spring, and the *summer_work*_t_h^{active} means the total time of disks running in high mode with active status during workdays in Summer. Furthermore, when the *S* in the labels of the Table 6 is replaced by the *M*, *A*, *W* respectively, the parameters are about the Summer season, Autumn season and Winter season respectively. Such as *S*_W_h^n means the total access times in high mode during workdays in Spring, and the *M*_W_h^n means the total access times in high mode during workdays in Summer. The replacement rule is the same for the remaining parameters.

According to the proposed K-ear strategy in Algorithm 3, the energy consumption model of K-ear is deduced as below.

$$\begin{aligned} e_{\text{total}} = {} & \textit{spring\_work}\_e_{\text{total}} + \textit{spring\_holiday}\_e_{\text{total}} + \textit{summer\_work}\_e_{\text{total}} + +\textit{summer\_holiday}\_e_{\text{total}} + \textit{autumn\_work}\_e_{\text{total}} + \textit{autumn\_holiday}\_e_{\text{total}} \\ & + \textit{winter\_work}\_e_{\text{total}} + \textit{winter\_holiday}\_e_{\text{total}} \end{aligned} \tag{1}$$

TABLE 5 Meaning of the parameters of the whole system in mathematic models

Paramter	Meaning
τ^h	Transfer rate of disks running in high mode (MB/s)
p^h	Energy consuming rate of disks running in high mode during active status (J/s)
i^h	Energy consuming rate of disks running in high mode during idle status (J/s)
τ^l	Transfer rate of disks running in low mode (MB/s)
p^l	Energy consuming rate of disks running in low mode during active status (J/s)
i^l	Energy consuming rate of disks running in low mode during idle status (J/s)
s'	The average size of the dataset (MB)
e_{total}	The total enery consumption (J)
e_h	Energy consumed by all of the disks running in high mode (J)
e_l	Energy consumed by all of the disks running in low mode (J)
e_h^{active}	Energy consumed by all of the disks running in high mode during active status (J)
e_h^{idle}	Energy consumed by all of the disks running in high mode during idle status (J)
e_l^{active}	Energy consumed by all of the disks running in low mode during active status (J)
e_l^{idle}	Energy consumed by all of the disks running in low mode during idle status (J)
t_h^{active}	Total time of disks running in high mode with active status (Second)
t_h^{idle}	Total time of disks running in high mode with idle status (Second)
t_l^{active}	Total time of disks running in low mode with active status (Second)
t_l^{idle}	Total time of disks running in low mode with idle status (Second)
T	Total service time of a disk (Second/disk)

TABLE 6 Meaning of the parameters about the spring season in mathematic models

$spring\_work\_t_h^{active}$	**Total time of disks running in high mode with active status during workdays in Spring (s)**
$spring\_holiday\_t_h^{active}$	Total time of disks running in high mode with active status during weekends in Spring (s)
$spring\_work\_t_h^{idle}$	Total time of disks running in high mode with idle status during workdays in Spring (s)
$spring\_holiday\_t_h^{idle}$	Total time of disks running in high mode with idle status during weekends in Spring (s)
$spring\_work\_t_l^{active}$	Total time of disks running in low mode with active status during workdays in Spring (s)
$spring\_holiday\_t_l^{active}$	Total time of disks running in low mode with active status during weekends in Spring (s)
$spring\_work\_t_l^{idle}$	Total time of disks running in low mode with idle status during workdays in Spring (s)
$spring\_holiday\_t_l^{idle}$	Total time of disks running in low mode with idle status during weekends in Spring (s)
$spring\_work\_e_{total}$	Energy consumption during workdays in Spring (J)
$spring\_holiday\_e_{total}$	Energy consumption during weekends in Spring (J)
$spring\_work\_e_h^{active}$	Energy Consumption of disks running in high mode with active status during workdays in Spring (J)
$spring\_holiday\_e_h^{active}$	Energy Consumption of disks running in high mode with active status during weekends in Spring (J)
$spring\_work\_e_l^{active}$	Energy Consumption of disks running in low mode with active status during workdays in Spring (J)
$spring\_holiday\_e_l^{active}$	Energy Consumption of disks running in low mode with active status during weekends in Spring (J)
$spring\_work\_e_h^{idle}$	Energy Consumption of disks running in high mode with idle status during workdays in Spring (J)
$spring\_holiday\_e_h^{idle}$	Energy Consumption of disks running in high mode with idle status during weekends in Spring (J)
$spring\_work\_e_l^{idle}$	Energy Consumption of disks running in low mode with idle status during workdays in Summer (J)
$spring\_holiday\_e_l^{idle}$	Energy Consumption of disks running in low mode with idle status during weekends in Summer (J)
spring_work_T	Total service time of a disk during workdays in Spring (Second/disk) = 5 T/28
spring_holiday_T	Total service time of a disk during weekends in Spring (Second/disk) = 2 T/28
$S\_W_h^n$	Total access times in high mode during workdays in Spring
$S\_W_l^n$	Total access times in low mode during workdays in Spring
$S\_H_h^n$	Total access times in high mode during weekends in Spring
$S\_H_l^n$	Total access times in low mode during weekends in Spring

The intuition behind the Equations (1) is that the total energy consumption of the systems is the summary of the energy consumption of the workdays in spring ($spring\_work\_e_{total}$), holidays in spring ($spring\_holiday\_e_{total}$), workdays in summer ($summer\_work\_e_{total}$), holidays in summer ($summer\_holiday\_e_{total}$), workdays in autumn ($autumn\_work\_e_{total}$), holidays in autumn ($autumn\_holiday\_e_{total}$), workdays in winter ($winter\_work\_e_{total}$), holidays in winter ($winter\_holiday\_e_{total}$). And there are calculated by the following formulas

$$spring_{work_e\,total} = spring\_work\_e_h^{active} + spring\_work\_e_h^{idle} + spring\_work\_e_l^{active} + spring\_work\_e_l^{idle} \tag{2}$$

$$summer\_work\_e_{total} = summer\_work\_e_h^{active} + summer\_work\_e_h^{idle} + summer\_work\_e_l^{active} + summer\_work\_e_l^{idle} \tag{3}$$

$$autumn\_work\_e_{total} = autumn\_work\_e_h^{active} + autumn\_work\_e_h^{idle} + autumn\_work\_e_l^{active} + autumn\_work\_e_l^{idle} \tag{4}$$

$$winter\_work\_e_{total} = winter\_work\_e_h^{active} + winter\_work\_e_h^{idle} + winter\_work\_e_l^{active} + winter\_work\_e_l^{idle} \tag{5}$$

$$spring\_holiday\_e_{total} = spring\_holiday\_e_h^{active} + spring\_holiday\_e_h^{idle} + spring\_holiday\_e_l^{active} + spring\_holiday\_e_l^{idle} \tag{6}$$

$$summer\_holiday\_e_{total} = summer\_holiday\_e_h^{active} + summer\_holiday\_e_h^{idle} + summer\_holiday\_e_l^{active} + summer\_holiday\_e_l^{idle} \tag{7}$$

$$autumn\_holiday\_e_{total} = autumn\_holiday\_e_h^{active} + autumn\_holiday\_e_h^{idle} + autumn\_holiday\_e_l^{active} + autumn\_holiday\_e_l^{idle} \tag{8}$$

$$winter\_holiday\_e_{total} = winter\_holiday\_e_h^{active} + winter\_holiday\_e_h^{idle} + winter\_holiday\_e_l^{active} + winter\_holiday\_e_l^{idle} \tag{9}$$

When

$$\boldsymbol{spring\_work\_e_h^{active}} = \mathrm{p}^h \times spring\_work\_t_h^{\mathrm{active}} = \mathrm{p}^h \times S\_W_h^n \times \mathrm{s}'/(\tau^h \times H1) \tag{10}$$

$$\boldsymbol{spring\_work\_e_h^{idle}} = i^h \times (spring\_work\_T \times H1 - spring\_work\_t_h^{\mathrm{active}}) = i^h \times (\frac{5T}{28} \times H1 - S\_W_h^n \times \mathrm{s}'/(\tau^h \times H1) \tag{11}$$

$$\boldsymbol{spring\_holiday\_e_h^{active}} = \mathrm{p}^h \times spring\_holiday\_t_h^{\mathrm{active}} = \mathrm{p}^h \times S\_H_h^n \times \mathrm{s}'/(\tau^h \times H2) \tag{12}$$

$$\boldsymbol{spring\_holiday\_e_h^{idle}} = i^h \times (spring\_holiday\_T \times H2 - spring\_holiday\_t_h^{\mathrm{active}}) = i^h \times \left(\frac{2T}{28} \times H2 - S_{Hh}^{\ n} \times \mathrm{s}'/(\tau^h \times H2)\right) \tag{13}$$

And the other formulas to calculate the valuses related to the Summer season, Autumn season and Winter season are simluar to the formulas from (10) to (13), as they accorrding to the above replacement rules plus that H1 and H2 are replcaed by H3 and H4, H5 and H6, H7, and H8.

$$\boldsymbol{spring\_work\_e_l^{active}} = \mathrm{p}^l \times spring\_work\_t_l^{\mathrm{active}} = \mathrm{p}^l \times S\_W_l^n \times \mathrm{s}'/(\tau^l \times L1) \tag{14}$$

$$\boldsymbol{spring\_work\_e_l^{idle}} = i^l \times (spring\_work\_T \times L1 - spring\_work\_t_l^{\mathrm{active}}) = i^l \times \left(\frac{5T}{28} \times L1 - S\_W_l^n \times \mathrm{s}'/(\tau^l \times L1)\right) \tag{15}$$

$$\boldsymbol{spring\_holiday\_e_l^{active}} = \mathrm{p}^l \times spring\_holiday\_t_l^{\mathrm{active}} = \mathrm{p}^l \times S\_H_l^n \times \mathrm{s}'/(\tau^l \times L2) \tag{16}$$

$$\boldsymbol{spring\_holiday\_e_l^{idle}} = i^l \times (spring\_holiday\_T \times L2 - spring\_holiday\_t_l^{\mathrm{active}}) = i^l \times \left(\frac{2T}{28} \times L2 - S\_H_l^n \times \mathrm{s}'/(\tau^l \times L2)\right) \tag{17}$$

And the other formulas to calculate the valuses related to the Summer season, Autumn season and Winter season are simluar to the formulas from (14) to (17), as they accorrding to the above replacement rules plus that L1 and L2 are replcaed by L3 and L4, L5 and L6, L7, and L8.

Therefore,

$$\begin{aligned}
\mathrm{e}_{\mathrm{total}} = {} & \mathrm{p}^h \times S_W_h^n \times \mathrm{s}'/(\tau^h \times H1) + i^h \times (\frac{5T}{28} \times H1 - S_W_h^n \times \mathrm{s}'/(\tau^h \times H1) + \mathrm{p}^h \times S_H_h^n \times \mathrm{s}'/(\tau^h \times H2) + i^h \times (\frac{2T}{28} \times H2 - S_H_h^n \times \mathrm{s}'/(\tau^h \times H2) \\
& + \mathrm{p}^h \times M_W_h^n \times \mathrm{s}'/(\tau^h \times H3) + i^h \times (\frac{5T}{28} \times H3 - M_W_h^n \times \mathrm{s}'/(\tau^h \times H3) + \mathrm{p}^h \times M_H_h^n \times \mathrm{s}'/(\tau^h \times H4) + i^h \times (\frac{2T}{28} \times H4 - M_H_h^n \times \mathrm{s}'/(\tau^h \times H4) \\
& + \mathrm{p}^h \times A_W_h^n \times \mathrm{s}'/(\tau^h \times H5) + i^h \times (\frac{5T}{28} \times H5 - A_W_h^n \times \mathrm{s}'/(\tau^h \times H5) + \mathrm{p}^h \times A_H_h^n \times \mathrm{s}'/(\tau^h \times H6) + i^h \times (\frac{2T}{28} \times H6 - A_H_h^n \times \mathrm{s}'/(\tau^h \times H6) \\
& + \mathrm{p}^h \times W_W_h^n \times \mathrm{s}'/(\tau^h \times H7) + i^h \times (\frac{5T}{28} \times H7 - W_W_h^n \times \mathrm{s}'/(\tau^h \times H7) + \mathrm{p}^h \times W_H_h^n \times \mathrm{s}'/(\tau^h \times H8) + i^h \times (\frac{2T}{28} \times H8 - W_H_h^n \times \mathrm{s}'/(\tau^h \times H8) \\
& + \mathrm{p}^l \times S_W_l^n \times \mathrm{s}'/(\tau^l \times L1) + i^l \times \left(\frac{5T}{28} \times L1 - S_W_l^n \times \mathrm{s}'/(\tau^l \times L1)\right) + \mathrm{p}^l \times S_H_l^n \times \mathrm{s}'/(\tau^l \times L2) + i^l \times \left(\frac{2T}{28} \times L2 - S_H_l^n \times \mathrm{s}'/(\tau^l \times L2)\right) \\
& + \mathrm{p}^l \times M_W_l^n \times \mathrm{s}'/(\tau^l \times L3) + i^l \times \left(\frac{5T}{28} \times L3 - M_W_l^n \times \mathrm{s}'/(\tau^l \times L3)\right) + \mathrm{p}^l \times M_H_l^n \times \mathrm{s}'/(\tau^l \times L4) + i^l \times \left(\frac{2T}{28} \times L4 - M_H_l^n \times \mathrm{s}'/(\tau^l \times L4)\right) \\
& + \mathrm{p}^l \times A_W_l^n \times \mathrm{s}'/(\tau^l \times L5) + i^l \times \left(\frac{5T}{28} \times L5 - A_W_l^n \times \mathrm{s}'/(\tau^l \times L5)\right) + \mathrm{p}^l \times A_H_l^n \times \mathrm{s}'/(\tau^l \times L6) + i^l \times \left(\frac{2T}{28} \times L6 - A_H_l^n \times \mathrm{s}'/(\tau^l \times L6)\right) + \mathrm{p}^l \\
& \times W_W_l^n \times \mathrm{s}'/(\tau^l \times L7) + i^l \times \left(\frac{5T}{28} \times L7 - W_W_l^n \times \mathrm{s}'/(\tau^l \times L7)\right) + \mathrm{p}^l \times W_H_l^n \times \mathrm{s}'/(\tau^l \times L8) + i^l \times \left(\frac{2T}{28} \times L8 - W_H_l^n \times \mathrm{s}'/(\tau^l \times L8)\right)
\end{aligned} \tag{18}$$

Assume B=$\mathrm{p}^h \times \mathrm{s}'/\tau^h$ C=$\mathrm{p}^l \times \times \mathrm{s}'/\tau^l$

The Energy Consumption Model can be simplified as:

$$\begin{aligned}
\mathrm{e}_{\mathrm{total}} = {} & (S_{W_h}^{\ n}/H1 + S_{H_h}^{\ n}/H2 + M_{W_h}^{\ n}/H3 + M_{H_h}^{\ n}/H4 + A_{W_h}^{\ n}/H5 + A_{H_h}^{\ n}/H6 + W_{W_h}^{\ n}/H7 + W_{H_h}^{\ n}/H8) \times B + i^h \times (\frac{5T}{28} \times H1 - S_W_h^n \times \mathrm{s}'/(\tau^h \times H1) \\
& + i^h \times (\frac{2T}{28} \times H2 - S_H_h^n \times \mathrm{s}'/(\tau^h \times H2) + i^h \times (\frac{5T}{28} \times H3 - M_W_h^n \times \mathrm{s}'/(\tau^h \times H3) + i^h \times (\frac{2T}{28} \times H4 - M_H_h^n \times \mathrm{s}'/(\tau^h \times H4) + i^h \\
& \times (\frac{5T}{28} \times H5 - A_W_h^n \times \mathrm{s}'/(\tau^h \times H5) + i^h \times (\frac{2T}{28} \times H6 - A_H_h^n \times \mathrm{s}'/(\tau^h \times H6) + i^h \times (\frac{5T}{28} \times H7 - W_W_h^n \times \mathrm{s}'/(\tau^h \times H7) + i^h \\
& \times (\frac{2T}{28} \times H8 - W_H_h^n \times \mathrm{s}'/(\tau^h \times H8) + (S_{W_l}^{\ n}/L1 + S_{H_l}^{\ n}/L2 + M_{W_l}^{\ n}/L3 + M_{H_l}^{\ n}/L4 + A_{W_l}^{\ n}/L5 + A_{H_l}^{\ n}/L6 + W_{W_l}^{\ n}/L7 + W_{H_l}^{\ n}/L8) \times C + i^l \\
& \times \left(\frac{5T}{28} \times L1 - S_W_l^n \times \mathrm{s}'/(\tau^l \times L1)\right) + i^l \times \left(\frac{2T}{28} \times L2 - S_H_l^n \times \mathrm{s}'/(\tau^l \times L2)\right) + i^l \times \left(\frac{5T}{28} \times L3 - M_W_l^n \times \mathrm{s}'/(\tau^l \times L3)\right) + i^l \\
& \times \left(\frac{2T}{28} \times L4 - M_H_l^n \times \mathrm{s}'/(\tau^l \times L4)\right) + i^l \times \left(\frac{5T}{28} \times L5 - A_W_l^n \times \mathrm{s}'/(\tau^l \times L5)\right) + i^l \times \left(\frac{2T}{28} \times L6 - A_H_l^n \times \mathrm{s}'/(\tau^l \times L6)\right) + i^l \\
& \times \left(\frac{5T}{28} \times L7 - W_W_l^n \times \mathrm{s}'/(\tau^l \times L7)\right) + i^l \times \left(\frac{2T}{28} \times L8 - W_H_l^n \times \mathrm{s}'/(\tau^l \times L8)\right)
\end{aligned} \tag{19}$$

Correspondingly, the energy consumption model of the Hadoop default data placement strategy is as follows, which is without classification:

$$e'_{\text{total}} = e'^{\text{active}}_{h} + e'^{\text{idle}}_{h}$$

$$e'^{\text{active}}_{h} = p^h \times (S\_W_h^n + S\_H_h^n + M\_W_h^n + M\_H_h^n + A\_W_h^n + A\_H_h^n + W\_W_h^n + W\_H_h^n + S\_W_l^n + S\_H_l^n + M\_W_l^n + M\_H_l^n + A\_W_l^n + A\_H_l^n + W\_W_l^n + W\_H_l^n) \times s'/(\tau^h \times n)$$

$$= (S\_W_h^n + S\_H_h^n + M\_W_h^n + M\_H_h^n + A\_W_h^n + A\_H_h^n + W\_W_h^n + W\_H_h^n + S\_W_l^n + S\_H_l^n + M\_W_l^n + M\_H_l^n + A\_W_l^n + A\_H_l^n + W\_W_l^n + W\_H_l^n) \times B/n$$

$$e'^{\text{idle}}_{h} = i^h \times (T \times n - (S\_W_h^n + S\_H_h^n + M\_W_h^n + M\_H_h^n + A\_W_h^n + A\_H_h^n + W\_W_h^n + W\_H_h^n + S\_W_l^n + S\_H_l^n + M\_W_l^n + M\_H_l^n + A\_W_l^n + A\_H_l^n + W\_W_l^n + W\_H_l^n) \times s'/(\tau^h \times n)$$

Therefore,

$$e'_{\text{total}} = (S\_W_h^n + S\_H_h^n + M\_W_h^n + M\_H_h^n + A\_W_h^n + A\_H_h^n + W\_W_h^n + W\_H_h^n + S\_W_l^n + S\_H_l^n + M\_W_l^n + M\_H_l^n + A\_W_l^n + A\_H_l^n + W\_W_l^n + W\_H_l^n) \times B/n + i^h \times (T \times n - (S\_W_h^n + S\_H_h^n + M\_W_h^n + M\_H_h^n + A\_W_h^n + A\_H_h^n + W\_W_h^n + W\_H_h^n + S\_W_l^n + S\_H_l^n + M\_W_l^n + M\_H_l^n + A\_W_l^n + A\_H_l^n + W\_W_l^n + W\_H_l^n) \times s'/(\tau^h \times n) \quad (20)$$

And the Energy Consumption model of SEA classification[11] can be deduced by the following formula:

$$e_{\text{total}_{\text{sea}}} = e_{\text{hot}} + e_{\text{cold}} = e^{\text{active}}_{\text{hot}} + e^{\text{idle}}_{\text{hot}} + e^{\text{active}}_{\text{cold}} + e^{\text{idle}}_{\text{cold}} = p^h \times n_h \times s'/(\tau^h \times \eta_h \times n) + i^h \times (T \times \eta_h \times n - n_h \times s'/(\tau^h \times \eta_h \times n)) + p^l \times n_c \times s'/(\tau^l \times \eta_c \times n) + i^l \times (T \times \eta_c \times n - n_c \times s'/(\tau^l \times \eta_c \times n)) \quad (21)$$

where η_h is the ratio of hot (popular) data and η_c is the ratio of cold (unpopular) data. n_h is the total access time of hot data, and n_c is the total access time of cold data.

As the mathematical models of the K-ear, Hadoop-default and SEA are stated in formula (19), (20), (21), respectively. We can deduce that the time complexity of the three algorithms depends on the time to calculate the access frequency of the data. That is to say, the time complexity of the three algorithms is $O(n)$. On the other hand, space complexity of the three algorithms depends on the space to store the access frequency of the data. That is to say, the space complexity of the three algorithms is also $O(n)$. As the analyzed above, the advantage of the three algorithms are usually depends on the energy consumption savings.

Based on the energy consumption model, simulation experiments with the K-ear algorithm and the SEA and Hadoop default algorithms will be conducted in the following section.

5 | PERFORMANCE EVALUATION

To evaluate the energy efficiency of the proposed K-ear strategy, we generate the workload according to the real access trace from the wiki (wiki workload) and assume a disk with two speed modes. The different transfer rates and energy consumption rates are extracted from article.[11] All of the simulation experiments are conducted in CloudSimDisk,[37] which builds on the CloudSim[38] simulator and adds storage simulation capabilities. And the detailed experimental parameters about the software and hardware configuration are listed in Table 7.

The general parameters utilized in the experiments are listed in Table 8.

TABLE 7 Software and hardware configuration in the experiments

Equipment/software	Type/version
CPU	Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz 3.3 GHz
Memory Size	4.0GB
Hard Disk	1 TB(TOSHIBA DT01ACA100 ATA Device)
Network Card	Realtek PCIe GBE Family Controller
Operating System	Windows 10
Energy-aware Disk Simulator	CloudSimDisk1.0
Cloud Environment Simulator	CloudSim 4.0
Programming Platform	Eclipse-Java-Luna-SR2

TABLE 8 General parameters value in the simulation experiments

Paramter	Value	Paramter	Value
p^h	30.26 J/s	i^l	2.17 J/s
i^h	5.26 J/s	τ^l	9.3 Mb/s
τ^h	31 Mb/s	n	1000
p^l	21.33 J/s	T	31,536,000 s

Disks with two speed modes, the workload and the disk zone, are modeled in the CloudSimDisk environment. The energy efficiency of the three algorithms is evaluated by setting the following ratios: (1) The ratio of the high speed disk utilization to the cloud storage system utilization, (2) the ratio of the data with seasonal characteristics to the data without seasonal characteristics, (3) the ratio of the data with tidal characteristics to the data without tidal characteristics, and (4) the ratio of the hot data to the cold data.

5.1 | The impact of the different ratios of high-speed disk utilization on cloud storage system utilization

The parameters corresponding to the values set in the experiment are listed in Table 9.

In Table 9, The ratios of the other seasons with the different characteristics are the same as the Spring season, with the *s* in the notation η_s is, respectively, replaced by *m*, *a*, and *w*.

In order to calculate the energy consumption of the SEA algorithm, the ratio of the hot data to cold data is set to 4:6 in this experiment. Obtained experiment results are demonstrated in the following tables (Tables 10–12), when the ratios of the high disk utilization to the whole system utilization are 1.6, 1.8, 2.0, respectively.

As shown in the Table 10, the ratio of the high disk utilization to the whole system utilization is 1.6, energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is more than 40%, while compared to the SEA, the saved energy consumption is about 11%. Moreover, we have found that the system utilization has little impact on the energy consumption for the three algorithms.

As shown in the Table 11, when the ratio of the high disk utilization to the whole system utilization is 1.8, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 42%, while compared to the SEA, the saved energy consumption is also about 11%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms in this experiment.

As shown in the Table 12, when the ratio of the high disk utilization to the whole system utilization is 2.0, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 42%, while compared to the SEA, the saved energy consumption is also about 11%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms in this experiment.

The above experimental results show that under the different ratios of high-speed disk utilization to system utilization, the K-ear strategy consumes the least energy and the Hadoop default strategy consumes the most energy. The amount of energy consumed by SEA is between the amounts

TABLE 9 Parameters and values set in this experiment

Parameter	Value	Parameter	Value
Ratio of the data with Spring characteristics η_s	0.2	Ratio of data with Spring and workday characteristics η_{s-w}	0.3
		Ratio of data with Spring and weekend characteristics η_{s-h}	0.3
		Ratio of data with Spring characteristics but without tidal characteristics η_{s-o}	0.4
Ratio of the data without seasonal characteristics η_o	0.2	Ratio of data with workday characteristics but without seasonal characteristics η_{o-w}	0.3
		Ratio of data with weekend characteristics but without seasonal characteristics η_{o-h}	0.3
		Ratio of data without seasonal and tidal characteristics η_{o-o}	0.4

TABLE 10 Energy consumption of the three data placement strategies when the ratio of the high disk utilization to the whole system utilization is 1.6

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	95,998.62547	165,958.2	107658.606
0.11	96,026.6813	165,966.084	107,683.3049
0.12	96,054.73713	165,973.968	107,708.0039
0.13	96,082.79295	165,981.852	107,732.7029
0.14	96,110.84878	165,989.736	107,757.4019
0.15	96,138.90461	165,997.62	107,782.1009
0.16	96,166.96043	166,005.504	107,806.7999
0.17	96,195.01626	166,013.388	107,831.4989
0.18	96,223.07209	166,021.272	107,856.1979
0.19	96,251.12792	166,029.156	107,880.8969
0.2	96,279.18374	166,037.04	107,905.5959
0.21	96,307.23957	166,044.924	107,930.2949
0.22	96,335.2954	166,052.808	107,954.9939
0.23	96,363.35123	166,060.692	107,979.6929
0.24	96,391.40705	166,068.576	108,004.3919
0.25	96,419.46288	166,076.46	108,029.0909
0.26	96,447.51871	166,084.344	108,053.7899
0.27	96,475.57453	166,092.228	108,078.4889
0.28	96,503.63036	166,100.112	108,103.1879
0.29	96,531.68619	166,107.996	108,127.8869
0.3	96,559.74202	166,115.88	108,152.5859
0.31	96,587.79784	166,123.764	108,177.2849
0.32	96,615.85367	166,131.648	108,201.9838
0.33	96,643.9095	166,139.532	108,226.6828
0.34	96,671.96532	166,147.416	108,251.3818
0.35	96,700.02115	166,155.3	108,276.0808
0.36	96,728.07698	166,163.184	108,300.7798
0.37	96,756.13281	166,171.068	108,325.4788
0.38	96,784.18863	166,178.952	108,350.1778
0.39	96,812.24446	166,186.836	108,374.8768

consumed by the K-ear strategy and the Hadoop default algorithm. Moreover, the energy consumption of the cloud storage system only slightly increased as the ratio increased.

5.2 | The impact of the different ratios of data with seasonal characteristics to data without seasonal characteristics

The ratio of the high-speed disk utilization to the system utilization is set to 2.0, and the ratio of hot data to cold data is set to 4:6 in all of the experiments (experiment 1, experiment 2, and experiment 3) in this subsection. The other parameter values are listed in Tables 13 and 14.

In Table 13, the ratios of the other seasons with the different characteristics are the same as the Spring season, with the s in the notation η_s is, respectively, replaced by m, a, and w.

TABLE 11 Energy consumption of the three data placement strategies when the ratio of the high disk utilization to the whole system utilization is 1.8

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	95,998.72826	165,958.2	107,647.5193
0.11	96,026.79436	165,966.084	107,671.1096
0.12	96,054.86047	165,973.968	107,694.7
0.13	96,082.92657	165,981.852	107,718.2903
0.14	96,110.99268	165,989.736	107,741.8806
0.15	96,139.05878	165,997.62	107,765.4709
0.16	96,167.12489	166,005.504	107,789.0613
0.17	96,195.19099	166,013.388	107,812.6516
0.18	96,223.2571	166,021.272	107,836.2419
0.19	96,251.32321	166,029.156	107,859.8323
0.2	96,279.38931	166,037.04	107,883.4226
0.21	96,307.45542	166,044.924	107,907.0129
0.22	96,335.52152	166,052.808	107,930.6033
0.23	96,363.58763	166,060.692	107,954.1936
0.24	96,391.65373	166,068.576	107,977.7839
0.25	96,419.71984	166,076.46	108,001.3742
0.26	96,447.78595	166,084.344	108,024.9646
0.27	96,475.85205	166,092.228	108,048.5549
0.28	96,503.91816	166,100.112	108,072.1452
0.29	96,531.98426	166,107.996	108,095.7356
0.3	96,560.05037	166,115.88	108,119.3259
0.31	96,588.11647	166,123.764	108,142.9162
0.32	96,616.18258	166,131.648	108,166.5065
0.33	96,644.24868	166,139.532	108,190.0969
0.34	96,672.31479	166,147.416	108,213.6872
0.35	96,700.3809	166,155.3	108,237.2775
0.36	96,728.447	166,163.184	108,260.8679
0.37	96,756.51311	166,171.068	108,284.4582
0.38	96,784.57921	166,178.952	108,308.0485
0.39	96,812.64532	166,186.836	108,331.6389

As the value set as in Tables 12 and 13, the obtained experimental results are demonstrated in Table 15.

As shown in the Table 16, when the experimental parameters are set as in Tables 12 and 13, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is also fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 36%, while compared to the SEA, the saved energy consumption is also about 1.3%, which is almost the same. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms in this experiment.

The obtained experimental results are demonstrated in Table 17.

As shown in the Table 18, when the experimental parameters are set as in Tables 13 and 16, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is also fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 40%, while compared to the SEA, the saved energy consumption is also about 7%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms in this experiment.

Seasonal characteristics related parameters' value set in experiment 3 are listed in Table 18

TABLE 12 Energy consumption of the three data placement strategies when the ratio of the high disk utilization to the whole system utilization is 2.0

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	95,998.83104	165,958.2	107,636.4326
0.11	96,026.90742	165,966.084	107,658.9143
0.12	96,054.98381	165,973.968	107,681.396
0.13	96,083.06019	165,981.852	107,703.8776
0.14	96,111.13658	165,989.736	107,726.3593
0.15	96,139.21296	165,997.62	107,748.841
0.16	96,167.28934	166,005.504	107,771.3226
0.17	96,195.36573	166,013.388	107,793.8043
0.18	96,223.44211	166,021.272	107,816.286
0.19	96,251.5185	166,029.156	107,838.7676
0.2	96,279.59488	166,037.04	107,861.2493
0.21	96,307.67126	166,044.924	107,883.7309
0.22	96,335.74765	166,052.808	107,906.2126
0.23	96,363.82403	166,060.692	107,928.6943
0.24	96,391.90042	166,068.576	107,951.1759
0.25	96,419.9768	166,076.46	107,973.6576
0.26	96,448.05318	166,084.344	107,996.1393
0.27	96,476.12957	166,092.228	108,018.6209
0.28	96,504.20595	166,100.112	108,041.1026
0.29	96,532.28234	166,107.996	108,063.5843
0.3	96,560.35872	166,115.88	108,086.0659
0.31	96,588.4351	166,123.764	108,108.5476
0.32	96,616.51149	166,131.648	108,131.0292
0.33	96,644.58787	166,139.532	108,153.5109
0.34	96,672.66426	166,147.416	108,175.9926
0.35	96,700.74064	166,155.3	108,198.4742
0.36	96,728.81702	166,163.184	108,220.9559
0.37	96,756.89341	166,171.068	108,243.4376
0.38	96,784.96979	166,178.952	108,265.9192
0.39	96,813.04618	166,186.836	108,288.4009

TABLE 13 General parameters' values for the different ratios of the data with seasonal characteristics and the data without seasonal characteristics

Parameter	Value
Ratio of the data with Spring and workday characteristics $\eta_{s-\mathrm{w}}$	0.3
Ratio of the data with Spring and weekend characteristics $\eta_{s-\mathrm{h}}$	0.3
Ratio of the data with Spring but without tidal characteristics $\eta_{s-\mathrm{o}}$	0.4
Ratio of the data with workday characteristics but without seasonal characteristics $\eta_{o-\mathrm{w}}$	0.3
Ratio of the data with weekend characteristics but without seasonal characteristics $\eta_{o-\mathrm{h}}$	0.3
Ratio of the data without seasonal and tidal characteristics $\eta_{o-\mathrm{o}}$	0.4

TABLE 14 Seasonal characteristics related parameters' value set in experiment 1

Parameter	Value
Ratio of the data with Spring characteristics η_s	0.15
Ratio of the data with Summer characteristics η_m	0.15
Ratio of the data with Autumn characteristics η_a	0.15
Ratio of the data with Winter characteristics η_w	0.15
Ratio of the data without seasonal characteristics η_o	0.4

TABLE 15 Energy consumption of the three data placement strategies in experiment 1

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	106,182.9264	165,958.2	107,636.4326
0.11	106,206.2268	165,966.084	107,658.9143
0.12	106,229.5272	165,973.968	107,681.396
0.13	106,252.8276	165,981.852	107,703.8776
0.14	106,276.128	165,989.736	107,726.3593
0.15	106,299.4285	165,997.62	107,748.841
0.16	106,322.7289	166,005.504	107,771.3226
0.17	106,346.0293	166,013.388	107,793.8043
0.18	106,369.3297	166,021.272	107,816.286
0.19	106,392.6301	166,029.156	107,838.7676
0.2	106,415.9305	166,037.04	107,861.2493
0.21	106,439.2309	166,044.924	107,883.7309
0.22	106,462.5313	166,052.808	107,906.2126
0.23	106,485.8317	166,060.692	107,928.6943
0.24	106,509.1321	166,068.576	107,951.1759
0.25	106,532.4325	166,076.46	107,973.6576
0.26	106,555.7329	166,084.344	107,996.1393
0.27	106,579.0333	166,092.228	108,018.6209
0.28	106,602.3337	166,100.112	108,041.1026
0.29	106,625.6341	166,107.996	108,063.5843
0.3	106,648.9345	166,115.88	108,086.0659
0.31	106,672.2349	166,123.764	108,108.5476
0.32	106,695.5353	166,131.648	108,131.0292
0.33	106,718.8357	166,139.532	108,153.5109
0.34	106,742.1361	166,147.416	108,175.9926
0.35	106,765.4365	166,155.3	108,198.4742
0.36	106,788.7369	166,163.184	108,220.9559
0.37	106,812.0373	166,171.068	108,243.4376
0.38	106,835.3377	166,178.952	108,265.9192
0.39	106,858.6381	166,186.836	108,288.4009

TABLE 16 Seasonal characteristics-related parameters' value set in experiment 2

Parameter	Value
Ratio of the data with Spring characteristics η_s	0.18
Ratio of the data with Summer characteristics η_m	0.18
Ratio of the data with Autumn characteristics η_a	0.18
Ratio of the data with Winter characteristics η_w	0.18
Ratio of the data without seasonal characteristics η_o	0.28

TABLE 17 Energy consumption of the three data placement strategies in experiment 2

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	100,074.2443	165,958.2	107,636.4326
0.11	100,100.5878	165,966.084	107,658.9143
0.12	100,126.9314	165,973.968	107,681.396
0.13	100,153.2749	165,981.852	107,703.8776
0.14	100,179.6184	165,989.736	107,726.3593
0.15	100,205.9619	165,997.62	107,748.841
0.16	100,232.3054	166,005.504	107,771.3226
0.17	100,258.6489	166,013.388	107,793.8043
0.18	100,284.9924	166,021.272	107,816.286
0.19	100,311.3359	166,029.156	107,838.7676
0.2	100,337.6794	166,037.04	107,861.2493
0.21	100,364.0229	166,044.924	107,883.7309
0.22	100,390.3664	166,052.808	107,906.2126
0.23	100,416.7099	166,060.692	107,928.6943
0.24	100,443.0534	166,068.576	107,951.1759
0.25	100,469.3969	166,076.46	107,973.6576
0.26	100,495.7404	166,084.344	107,996.1393
0.27	100,522.0839	166,092.228	108,018.6209
0.28	100,548.4275	166,100.112	108,041.1026
0.29	100,574.771	166,107.996	108,063.5843
0.3	100,601.1145	166,115.88	108,086.0659
0.31	100,627.458	166,123.764	108,108.5476
0.32	100,653.8015	166,131.648	108,131.0292
0.33	100,680.145	166,139.532	108,153.5109
0.34	100,706.4885	166,147.416	108,175.9926
0.35	100,732.832	166,155.3	108,198.4742
0.36	100,759.1755	166,163.184	108,220.9559
0.37	100,785.519	166,171.068	108,243.4376
0.38	100,811.8625	166,178.952	108,265.9192
0.39	100,838.206	166,186.836	108,288.4009

TABLE 18 Seasonal characteristics related parameters' value set in experiment 3

Parameter	Value
Ratio of the data with Spring characteristics η_s	0.12
Ratio of the data with Summer characteristics η_m	0.12
Ratio of the data with Autumn characteristics η_a	0.12
Ratio of the data with Winter characteristics η_w	0.12
Ratio of the data without seasonal characteristics η_o	0.52

The obtained experimental results are demonstrated in Table 19.

As shown in the Table 19, when the experimental parameters are set as in Tables 12 and 17, the energy consumed by the SEA is least, and the most is the Hadoop-default. While the energy consumed by the proposed K-ear is fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 32%, while compared to the SEA, the energy consumed by K-ear is increased is about 4%, which has little difference. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms in this experiment.

As shown in the abovementioned experimental results, the K-ear and SEA strategies consume less energy than the Hadoop-default strategy under different ratios of data with seasonal characteristics to data without seasonal characteristics. When the data has a higher ratio of seasonal characteristics of 6:4, K-ear slightly outperforms SEA in energy consumption. It can also be found that the lower the ratio of seasonal characteristics (48:52), SEA is more energy-efficient than K-ear but with little difference.

TABLE 19 Energy consumption of the three data placement strategies in experiment 3

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	112,284.6623	165,958.2	107,636.4326
0.11	112,304.225	165,966.084	107,658.9143
0.12	112,323.7877	165,973.968	107,681.396
0.13	112,343.3504	165,981.852	107,703.8776
0.14	112,362.913	165,989.736	107,726.3593
0.15	112,382.4757	165,997.62	107,748.841
0.16	112,402.0384	166,005.504	107,771.3226
0.17	112,421.6011	166,013.388	107,793.8043
0.18	112,441.1638	166,021.272	107,816.286
0.19	112,460.7264	166,029.156	107,838.7676
0.2	112,480.2891	166,037.04	107,861.2493
0.21	112,499.8518	166,044.924	107,883.7309
0.22	112,519.4145	166,052.808	107,906.2126
0.23	112,538.9772	166,060.692	107,928.6943
0.24	112,558.5398	166,068.576	107,951.1759
0.25	112,578.1025	166,076.46	107,973.6576
0.26	112,597.6652	166,084.344	107,996.1393
0.27	112,617.2279	166,092.228	108,018.6209
0.28	112,636.7906	166,100.112	108,041.1026
0.29	112,656.3532	166,107.996	108,063.5843
0.3	112,675.9159	166,115.88	108,086.0659
0.31	112,695.4786	166,123.764	108,108.5476
0.32	112,715.0413	166,131.648	108,131.0292
0.33	112,734.6039	166,139.532	108,153.5109
0.34	112,754.1666	166,147.416	108,175.9926
0.35	112,773.7293	166,155.3	108,198.4742
0.36	112,793.292	166,163.184	108,220.9559
0.37	112,812.8547	166,171.068	108,243.4376
0.38	112,832.4173	166,178.952	108,265.9192
0.39	112,851.98	166,186.836	108,288.4009

5.3 | The impact of different ratios of data with tidal characteristics to data without tidal characteristics

The energy consumption of the three strategies with different ratios of data with tidal characteristics is tested in this subsection. The common parameters used in the different experiments are listed in Table 20.

And the tidal characteristics-related parameters used in the first experiment are listed in Table 21.

The ratios of the other seasons with the different characteristics are the same as the Spring season, with the *s* in the notation η_s is, respectively, replaced by *m*, *a*, and *w*.

As the parameters of the first experiment set as the above tables, the obtained experimental results are shown in Table 22.

As shown in the Table 22, when the ratio of the data with tidal characteristics to the data without tidal characteristics is 4:6, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is also fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 40%, while compared to the SEA, the saved energy consumption is also about 7%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms.

Tidal characteristics related parameters used in the second experiment are listed in Table 23.

In Table 23, the ratios of the other seasons with the different characteristics are the same as the Spring season, with the *s* in the notation η_s is, respectively, replaced by *m*, *a*, and *w*.

As the parameters of the second experiment set as the above tables, the obtained experimental results are shown in Table 24.

As shown in the Table 24, when the ratio of the data with tidal characteristics to the data without tidal characteristics is 5:5, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is also fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 41%, while compared to the SEA, the saved energy consumption is also about 9%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms in this experiment.

Tidal characteristics related parameters used in the third experiment are listed in Table 25.

In Table 24, the ratios of the other seasons with the different characteristics are the same as the Spring season, with the *s* in the notation η_s is, respectively, replaced by *m*, *a*, and *w*.

As the parameters of the third experiment set as the above tables, the obtained experimental results are shown in Table 26.

As shown in the Table 26, when the ratio of the data with tidal characteristics to the data without tidal characteristics is 2:8, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is also fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 37%, while compared to the SEA, the saved energy consumption is also about 3.5%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms.

TABLE 20 Common parameters used in the following three experiments

Parameter	Value
Ratio of the data with Spring characteristics η_{s}	0.2
Ratio of the data with Summer characteristics η_{m}	0.2
Ratio of the data with Autumn characteristics η_{a}	0.2
Ratio of the data with Winter characteristics η_{w}	0.2
Ratio of the data without seasonal characteristics η_{o}	0.2
Ratio of the high speed disk utilization to the system utilization	2.0
Ratio of the hot data to the cold data	4 : 6

TABLE 21 Tidal characteristics-related parameters used in the first experiment

Parameter	Value
Ratio of the data with Spring and workday characteristics $\eta_{s-\mathrm{w}}$	0.2
Ratio of the data with Spring and weekend characteristics $\eta_{s-\mathrm{h}}$	0.2
Ratio of the data with Spring but without tidal characteristics $\eta_{s-\mathrm{o}}$	0.6
Ratio of the data with workday but without seasonal characteristics $\eta_{o-\mathrm{w}}$	0.2
Ratio of the data with weekend but without seasonal characteristics $\eta_{o-\mathrm{h}}$	0.2
Ratio of the data without seasonal and tidal characteristics $\eta_{o-\mathrm{o}}$	0.6

TABLE 22 Energy consumption of the three data placement strategies when the ratio of the data with tidal characteristics to the data without tidal characteristics is 4:6

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	99,880.22558	165, 958.2	107,636.4326
0.11	99,906.65646	165, 966.084	107,658.9143
0.12	99,933.08734	165, 973.968	107,681.396
0.13	99,959.51822	165, 981.852	107,703.8776
0.14	99,985.94909	165, 989.736	107,726.3593
0.15	10,0012.38	165, 997.62	107,748.841
0.16	100,038.8108	166, 005.504	107,771.3226
0.17	100,065.2417	166, 013.388	107,793.8043
0.18	100,091.6726	166, 021.272	107,816.286
0.19	100,118.1035	166, 029.156	107,838.7676
0.2	100,144.5344	166, 037.04	107,861.2493
0.21	100,170.9652	166, 044.924	107,883.7309
0.22	100,197.3961	166, 052.808	107,906.2126
0.23	100,223.827	166, 060.692	107,928.6943
0.24	100,250.2579	166068.576	107,951.1759
0.25	100,276.6888	166, 076.46	107,973.6576
0.26	100,303.1196	166, 084.344	107,996.1393
0.27	100,329.5505	166, 092.228	108,018.6209
0.28	100,355.9814	166, 100.112	108,041.1026
0.29	100,382.4123	166, 107.996	108,063.5843
0.3	100,408.8431	166, 115.88	108,086.0659
0.31	100,435.274	166, 123.764	108,108.5476
0.32	100,461.7049	166, 131.648	108,131.0292
0.33	100,488.1358	166, 139.532	108,153.5109
0.34	100,514.5667	166, 147.416	108,175.9926
0.35	100,540.9975	166, 155.3	108,198.4742
0.36	100,567.4284	166, 163.184	108,220.9559
0.37	100,593.8593	166, 171.068	108,243.4376
0.38	100,620.2902	166, 178.952	108,265.9192
0.39	100,646.721	166, 186.836	108,288.4009

TABLE 23 Tidal characteristics-related parameters used in the second experiment

Parameter	Value
Ratio of the data with Spring and workday characteristics η_{s-w}	0.25
Ratio of the data with Spring and weekend characteristics η_{s-h}	0.25
Ratio of the data with Spring but without tidal characteristics η_{s-o}	0.5
Ratio of the data with workday but without seasonal characteristics η_{o-w}	0.25
Ratio of the data with weekend but without seasonal characteristics η_{o-h}	0.25
Ratio of the data without seasonal and tidal characteristics η_{o-o}	0.5

TABLE 24 Energy consumption of the three data placement strategies when the ratio of the data with tidal characteristics to the data without tidal characteristics is 5:5

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	97,939.76338	165,958.2	107,636.4326
0.11	97,967.04052	165,966.084	107,658.9143
0.12	97,994.31766	165,973.968	107,681.396
0.13	98,021.5948	165,981.852	107,703.8776
0.14	98,048.87194	165,989.736	107,726.3593
0.15	98,076.14907	165,997.62	107,748.841
0.16	98,103.42621	166,005.504	107,771.3226
0.17	98,130.70335	166,013.388	107,793.8043
0.18	98,157.98049	166,021.272	107,816.286
0.19	98,185.25763	166,029.156	107,838.7676
0.2	98,212.53477	166,037.04	107,861.2493
0.21	98,239.8119	166,044.924	107,883.7309
0.22	98,267.08904	166,052.808	107,906.2126
0.23	98,294.36618	166,060.692	107,928.6943
0.24	98,321.64332	166,068.576	107,951.1759
0.25	98,348.92046	166,076.46	107,973.6576
0.26	98,376.19759	166,084.344	107,996.1393
0.27	98,403.47473	166,092.228	108,018.6209
0.28	98,430.75187	166,100.112	108,041.1026
0.29	98,458.02901	166,107.996	108,063.5843
0.3	98,485.30615	166,115.88	108,086.0659
0.31	98,512.58329	166,123.764	108,108.5476
0.32	98,539.86042	166,131.648	108,131.0292
0.33	98,567.13756	166,139.532	108,153.5109
0.34	98,594.4147	166,147.416	108,175.9926
0.35	98,621.69184	166,155.3	108,198.4742
0.36	98,648.96898	166,163.184	108,220.9559
0.37	98,676.24612	166,171.068	108,243.4376
0.38	98,703.52325	166,178.952	108,265.9192
0.39	98,730.80039	166,186.836	108,288.4009

TABLE 25 Tidal characteristics-related parameters used in the third experiment

Parameter	Value
Ratio of the data with Spring and workday characteristics η_{s-w}	0.1
Ratio of the data with Spring and weekend characteristics η_{s-h}	0.1
Ratio of the data with Spring but without tidal characteristics η_{s-o}	0.8
Ratio of the data with workday but without seasonal characteristics η_{o-w}	0.1
Ratio of the data with weekend but without seasonal characteristics η_{o-h}	0.1
Ratio of the data with Spring and workday characteristics η_{s-w}	0.8

TABLE 26 Energy consumption of the three data placement strategies when the ratio of the data with tidal characteristics to the data without tidal characteristics is 2:8

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	103,759.5632	165,958.2	107,636.4326
0.11	103,784.1429	165,966.084	107,658.9143
0.12	103,808.7226	165,973.968	107,681.396
0.13	103,833.3023	165,981.852	107,703.8776
0.14	103,857.882	165,989.736	107,726.3593
0.15	103,882.4617	165,997.62	107,748.841
0.16	103,907.0413	166,005.504	107,771.3226
0.17	103,931.621	166,013.388	107,793.8043
0.18	103,956.2007	166,021.272	107,816.286
0.19	103,980.7804	166,029.156	107,838.7676
0.2	104,005.3601	166,037.04	107,861.2493
0.21	104,029.9398	166,044.924	107,883.7309
0.22	104,054.5194	166,052.808	107,906.2126
0.23	104,079.0991	166,060.692	107,928.6943
0.24	104,103.6788	166,068.576	107,951.1759
0.25	104,128.2585	166,076.46	107,973.6576
0.26	104,152.8382	166,084.344	107,996.1393
0.27	104,177.4179	166,092.228	108,018.6209
0.28	104,201.9976	166,100.112	108,041.1026
0.29	104,226.5772	166,107.996	108,063.5843
0.3	104,251.1569	166,115.88	108,086.0659
0.31	104,275.7366	166,123.764	108,108.5476
0.32	104,300.3163	166,131.648	108,131.0292
0.33	104,324.896	166,139.532	108,153.5109
0.34	104,349.4757	166,147.416	108,175.9926
0.35	104,374.0553	166,155.3	108,198.4742
0.36	104,398.635	166,163.184	108,220.9559
0.37	104,423.2147	166,171.068	108,243.4376
0.38	104,447.7944	166,178.952	108,265.9192
0.39	104,472.3741	166,186.836	108,288.4009

The results of the above three experiments demonstrate that our proposed K-ear strategy is more energy-efficient than the SEA algorithm when the ratio of data with tidal characteristics is high. Furthermore, K-ear has an energy efficiency advantage over the Hadoop default strategy whenever the ratio is high or low.

5.4 | The impact of the different ratios of hot data to cold data

We set the different ratios of hot data to cold data while leaving the other parameters fixed. The energy consumption of the three strategies is evaluated according to the different ratios. The common parameters used in the following three experiments are listed in Table 27.

In Table 27, the ratios of the other seasons with the different characteristics are the same as the Spring season, with the s in the notation η_s is, respectively, replaced by m, a, and w.

Results obtained from simulation experiments when the ratio of hot data to cold data is set as 4:6 are shown in Table 28.

TABLE 27 Common parameters used in the following three experiments

Parameter	Value	Parameters	Value
Ratio of the data with Spring characteristics η_s	0.2	Ratio of data with Spring and workday characteristics η_{s-w}	0.3
		Ratio of data with Spring and weekend characteristics η_{s-h}	0.3
		Ratio of data with Spring characteristics but without tidal characteristics η_{s-o}	0.4
Ratio of the data without seasonal characteristics η_o	0.2	Ratio of data without seasonal characteristics but with workday characteristics η_{o-w}	0.3
		Ratio of data without seasonal characteristics but with weekend characteristics η_{o-h}	0.3
		Ratio of data without seasonal characteristics and tidal characteristics η_{o-o}	0.4
Parameter		Value	
Ratio of the high speed disk utilization to the system utilization		2.0	

TABLE 28 Energy consumption of the three data placement strategies when the ratio of the hot data to the cold data is 4:6

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	95,998.83104	16,5958.2	107, 636.4326
0.11	96,026.90742	165,966.084	107658.9143
0.12	96,054.98381	165,973.968	107,681.396
0.13	96,083.06019	165,981.852	107,703.8776
0.14	96,111.13658	165,989.736	107726.3593
0.15	96,139.21296	165,997.62	107748.841
0.16	96,167.28934	166,005.504	107,771.3226
0.17	96,195.36573	166,013.388	107,793.8043
0.18	96,223.44211	166,021.272	107,816.286
0.19	96,251.5185	166,029.156	107,838.7676
0.2	96,279.59488	166,037.04	107,861.2493
0.21	96,307.67126	166,044.924	107,883.7309
0.22	96,335.74765	166,052.808	107,906.2126
0.23	96,363.82403	166,060.692	107928.6943
0.24	96,391.90042	166,068.576	107,951.1759
0.25	96,419.9768	166,076.46	107,973.6576
0.26	96,448.05318	166,084.344	107996.1393
0.27	96,476.12957	166,092.228	108,018.6209
0.28	96,504.20595	166,100.112	108,041.1026
0.29	96,532.28234	166,107.996	108,063.5843
0.3	96,560.35872	166,115.88	108,086.0659
0.31	96,588.4351	166,123.764	108,108.5476
0.32	96,616.51149	166,131.648	108,131.0292
0.33	96,644.58787	166,139.532	108,153.5109
0.34	96,672.66426	166,147.416	108,175.9926
0.35	96,700.74064	166,155.3	108,198.4742
0.36	96,728.81702	166,163.184	108,220.9559
0.37	96,756.89341	166,171.068	108,243.4376
0.38	96,784.96979	166,178.952	108,265.9192
0.39	96,813.04618	166,186.836	108,288.4009

As shown in the Table 28, when the ratio of the hot data to the cold data is 4:6, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is also fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 42%, while compared to the SEA, the saved energy consumption is also about 11%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms.

Results obtained from simulation experiments when the ratio of hot data to cold data is set as 3:7 are shown in Table 29.

As shown in the Table 29, when the ratio of the hot data to the cold data is 3:7, the energy consumed by the K-ear is least, and the most is the Hadoop-default. While the energy consumed by the SEA is also fall in between. Compared to the Hadoop-default, the average energy consumption saved by the K-ear is about 42%, while compared to the SEA, the saved energy consumption is also about 2%, which is almost the same. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms.

Results obtained from simulation experiments when the ratio of hot data to cold data is set as 2:8 are shown in Table 30.

As shown in the Table 30, when the ratio of the hot data to the cold data is 2:8, the energy consumed by the SEAis least, and the most is the Hadoop-default. While the energy consumed by the proposed K-ear is fall in between. Compared to the Hadoop-default, the average energy

TABLE 29 Energy consumption of the three data placement strategies when the ratio of the hot data to the cold data is 3:7

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	95,998.83104	165,958.2	97,939.76338
0.11	96,026.90742	165,966.084	97,967.04052
0.12	96,054.98381	165,973.968	97,994.31766
0.13	96,083.06019	165,981.852	98,021.5948
0.14	96,111.13658	165,989.736	98,048.87194
0.15	96,139.21296	165,997.62	98,076.14907
0.16	96,167.28934	166,005.504	98,103.42621
0.17	96,195.36573	166,013.388	98,130.70335
0.18	96,223.44211	166,021.272	98,157.98049
0.19	96,251.5185	166,029.156	98,185.25763
0.2	96,279.59488	166,037.04	98212.53477
0.21	96,307.67126	166,044.924	98,239.8119
0.22	96,335.74765	166,052.808	98,267.08904
0.23	96,363.82403	166,060.692	98,294.36618
0.24	96,391.90042	166,068.576	98,321.64332
0.25	96,419.9768	166,076.46	98,348.92046
0.26	96,448.05318	166,084.344	98,376.1976
0.27	96,476.12957	166,092.228	98,403.47473
0.28	96,504.20595	166,100.112	98,430.75187
0.29	96,532.28234	166,107.996	98,458.02901
0.3	96,560.35872	166,115.88	98,485.30615
0.31	96,588.4351	166,123.764	98,512.58329
0.32	96,616.51149	166,131.648	98,539.86043
0.33	96,644.58787	166,139.532	98,567.13756
0.34	96,672.66426	166,147.416	98,594.4147
0.35	96,700.74064	166,155.3	98,621.69184
0.36	96,728.81702	166,163.184	98,648.96898
0.37	96,756.89341	166,171.068	98,676.24612
0.38	96,784.96979	166,178.952	98,703.52326
0.39	96,813.04618	166,186.836	98,730.80039

TABLE 30 Energy consumption of the three data placement strategies when the ratio of the hot data to the cold data is 2:8

System utilization	Energy consumption of K-ear (KJ)	Energy consumption of Hadoop-default (KJ)[12]	Energy consumption of SEA(KJ)[11]
0.1	95,998.83104	165,958.2	88,231.10544
0.11	96,026.90742	165,966.084	88,261.97918
0.12	96,054.98381	165,973.968	88,292.85293
0.13	96,083.06019	165,981.852	88,323.72667
0.14	96,111.13658	165,989.736	88,354.60042
0.15	96,139.21296	165,997.62	88,385.47416
0.16	96,167.28934	166,005.504	88,416.3479
0.17	96,195.36573	166,013.388	88,447.22165
0.18	96,223.44211	166,021.272	88,478.09539
0.19	96,251.5185	166,029.156	88,508.96914
0.2	96,279.59488	166,037.04	88,539.84288
0.21	96,307.67126	166,044.924	88,570.71662
0.22	96,335.74765	166,052.808	88,601.59037
0.23	96,363.82403	166,060.692	88,632.46411
0.24	96,391.90042	166,068.576	88,663.33786
0.25	96,419.9768	166,076.46	88,694.2116
0.26	96,448.05318	166084.344	88,725.08534
0.27	96,476.12957	166,092.228	88,755.95909
0.28	96,504.20595	166,100.112	88,786.83283
0.29	96,532.28234	166,107.996	88,817.70658
0.3	96,560.35872	166,115.88	88,848.58032
0.31	96,588.4351	166,123.764	88879.45406
0.32	96,616.51149	166,131.648	88,910.32781
0.33	96,644.58787	166,139.532	88,941.20155
0.34	96,672.66426	166147.416	88,972.0753
0.35	96,700.74064	166,155.3	89,002.94904
0.36	96,728.81702	166,163.184	89,033.82278
0.37	96756.89341	166,171.068	89,064.69653
0.38	96,784.96979	166,178.952	89,095.57027
0.39	96,813.04618	166,186.836	89,126.44402

consumption saved by the K-ear is about 42%, while compared to the SEA, the energy consumed by K-ear is increased is about 7%. Moreover, we also have found that the system utilization has little impact on the energy consumption for the three algorithms in this experiment.

As shown in the above comparative experimental results, it can be seen that when the ratio of hot data to cold data is higher, the K-ear strategy performs better than the SEA algorithm. When the ratio is 4:6, K-ear has an obvious advantage over SEA. When the ratio is 3:7, the performance of K-ear and SEA are almost the same. When the ratio is 2:8, SEA slightly outperforms K-ear.

6 | CONCLUSIONS AND FUTURE WORK

An energy-aware data clustering strategy, K-ear, is proposed in this paper, in which the seasonal and tidal characteristics of data access are extracted thoroughly. The machine learning algorithm is applied to cluster data into different categories, and based on the categories, the data are stored in different storage zones. During the different time zones, some storage zones run in high performance mode to satisfy the performance requirement

and the remaining storage zones run in low energy consumption mode to save energy consumption. To analyze and evaluate the energy efficiency of the proposed K-ear strategy, the famous classical SEA strategy and default data placement strategy in Hadoop are used for comparison, and mathematical models are constructed for the three strategies. Moreover, substantial simulation experiments are conducted in the CloudSimDisk simulator from different perspectives with different ratios. Compared with the mainstream data placement strategy (Hadoop default), K-ear and SEA are more energy efficient. The proposed K-ear strategy outperforms the classical SEA algorithm in most cases. Only when the ratio of hot data is very low is SEA more energy-efficient than K-ear. As a whole, compared with the other evaluated algorithms, the proposed K-ear data clustering storing strategies extracted the data access characteristics more thoroughly and classified the data into fine granularity categories, which will achieve more energy consumption savings with different disk running modes. Furthermore, as algorithm of SEA outperform the algorithms of Greedy,[39] SP (Sort Partitions),[40] HP (Hybrid Partition),[40] and PVFS (Parallel Virtual File System).[41,42] And the Hadoop default is more energy efficient than the algorithm of Datacenter without energy management. One part of our future work, we will explore how to combine the advantages of the different data placement strategies to achieve higher energy consumption reduction in cloud storage systems.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China under Grants (No. 61671070, 61972364.), National Language Committee of China under Grants ZDI135-53, Australian Research Council (ARC) Discovery Project and Project of Developing University Intension for Improving the Level of Scientific Research–No.2019KYNH226, Qin Xin Talents Cultivation Program, Beijing Information Science & Technology University No.QXTCP B201908.

AUTHOR CONTRIBUTION

Xindong You: Conceptualization, Methodology, Validation, Writing original draft, Funding acquisition. **Tian Sun:** Formal analysis, Investigation, Data curation, Writing - review and editing. **Dawei Sun:** Validation, Investigation, Writing - review and editing. **Xueqiang Lv:** Conceptualization Supervision, Writing - review and editing, Funding acquisition. **Xunyun Liu**: Data curation, Investigation, Writing - review and editing. **Rajkumar Buyya:** Conceptualization, Writing - review and editing, Supervision, Funding acquisition.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Xindong You https://orcid.org/0000-0002-3351-4599
Dawei Sun https://orcid.org/0000-0003-3137-6257
Rajkumar Buyya https://orcid.org/0000-0001-9754-6496

REFERENCES

1. Aujla GS, Kumar N. MEnSuS: an efficient scheme for energy management with sustainability of cloud data centers in edge–cloud environment. *Future Gener Comput Syst*. 2018;86(9):1279-1300.
2. Kaushik R. *Energy Management Costs for a Data Center*. USA Patent, Current Assignee: International Business Machines Corp. Granted in 2018.
3. Varghese B, Buyya R. Next generation cloud computing: new trends and research directions. *Future Gener Comput Syst*. 2018;79(3):849-861.
4. Mills M. *The Cloud Begins with Coal: Big Data, Big Networks, Big Infrastructure, and Big Power*. West Virginia: National Mining Association & American Coalition for Clean Coal Electricity; 2013.
5. Popa D, Pop F, Serbanescu C, Castiglione A. Deep learning model for home automation and energy reduction in a smart home environment platform. *Neural Comput Appl*. 2019;31(5):1317-1337.
6. Envantage. *New Report Reveals Warehouses Overspend on Energy by £190m*. New York: Data by. World Resources Institute;August 14,; 2013.
7. Uzaman SK, Rehman Khan A, Shuja J, Maqsood T. A systems overview of commercial data centers: initial energy and cost analysis. *Int J Inf Technol Web Eng*. 2019;14(1):42-65.
8. Marahatta A, Pirbhulal S, Zhang F, Parizi RM, Raymond Choo KK, Liu Z. Classification-based and energy-efficient dynamic task scheduling scheme for virtualized cloud data center. *IEEE Trans Cloud Comput*. 2019:1–1.
9. Lin H-Y, Yang S-Y. A smart cloud-based energy data mining agent using big data analysis technology. *Smart Sci*. 2019;7(3):175-183.
10. Asgari S, Moazamigoodarzi H, Tsai PJ, et al. Hybrid surrogate model for online temperature and pressure predictions in data centers. *Future Gener Comput Syst*. 2020;114:531-547.
11. Xie T. SEA: a striping-based Energy-aware strategy for data placement in RAID-structured storage systems. *IEEE Trans Comput*. 2008;57(6):748–761.
12. Kaushik RT, Bhandarkar M. GreenHDFS: towards an energy-conserving, storage-efficient, hybrid Hadoop compute cluster. Paper presented at: Proceedings of the 2010 International Conference on Power Aware Computing and Systems. HotPower '10, USENIX Association; 2010; Berkeley, CA: 1–9.
13. Kaushik RT, Cherkasova L, Campbell R, Nahrstedt R. Lightning: self-adaptive, energy-conserving, multi-zoned, commodity green cloud storage system. Paper presented at: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing HPDC '10 . ACM; 2010; New York, NY: 332-335.
14. You XD, Dong C, Zhou L, Huang J, Jiang CF. Anticipation-based green data classification strategy in cloud storage system. *Appl Math Inf Sci*. 2015;9(4):2151-2160.

15. Liao B, Yu J, Sun H, et al. Energy-efficient algorithms for distributed storage system based on data storage structure reconfiguration. *J Comput Res Dev*. 2013;50(1):3-18.
16. Zhang T, Liao B, Sun H, Li FG, Ji JH. Energy-efficient algorithm based on data classification for cloud storage system. *J Comput Appl*. 2014;34(8):2267-2273.
17. Xu XL, Yang G, Li LJ, Wang RC. Dynamic data aggregation algorithm for data centers of green cloud computing. *Syst Eng Electron*. 2012;34(9):1923–1929.
18. Long SQ. *Research on Data Layout Strategies for Cloud Storage System* [PhD thesis]. South China University of Technology; 2014.
19. Yadav R, Weizhe Z. MeReg: managing energy-SLA tradeoff for green mobile cloud computing. *Wirel Commun Mob Comput*. 2017;2017:1–11.
20. Yadav R, Zhang W, Chen H, Guo T. Mums: energy-aware vm selection scheme for cloud data center. Paper presented at: 2017 28th International Workshop on Database and Expert Systems Applications (DEXA). IEEE; 2017: 132-136.
21. Yadav R, Zhang W, Kaiwartya O, Singh PR, Elgendy IA, Tian YC. Adaptive energy-aware algorithms for minimizing energy consumption and SLA violation in cloud computing. *IEEE Access*. 2018;6:55923-55936.
22. Yadav R, Zhang W, Li K, Liu C, Shafiq M, Karn NK. An adaptive heuristic for managing energy consumption and overloaded hosts in a cloud data center. *Wirel Netw*. 2020;26(3):1905-1919.
23. Reddy R, Kathpal A, Basak J, Katz R. Data layout for power efficient archival storage systems. Paper presented at: HotPower'15 Proceedings of the Workshop on Power-Aware Computing and Systems; 2015.
24. Song Z, Wang T, Li T, Energy Consumption GY. Optimization data placement algorithm for MapReduce system. *J Softw*. 2015;26(8):2091-2110.
25. Ahuja SP, Muthiah K. Survey of state-of-art in green cloud computing. *Int J Green Comput*. 2016;7(1):25–36.
26. Wu Q, Deng QY, Ganesh L. Dynamo: Facebook's data center- wide power management system. Paper presented at: 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA); 2016.
27. Wang LZ, Ma Y, Yan JN, Chang V, Zomaya AY. pipsCloud: High-performance cloud computing for remote sensing big data management and processing. *Future Gener Comput Syst*. 2018;78, Part 1:353-368.
28. Song J, He HY, Yu G, Pierson JM. Modulo based data placement algorithm for Energy Consumption optimization of MapReduce system. *J Grid Comput*. 2018;16(3):409-424.
29. Tran XT, Do TV, Rotter C, Hwang D. A new data layout scheme for Energy-efficient MapReduce processing tasks. *J Grid Comput*. 2018;16(2):285-298.
30. Ebadi Y, Navimipour NJ. An energy-aware method for data replication in the cloud environments using a Tabu search and particle swarm optimization algorithm. *Concurr Comput Pract*. 2018;31:e4757.
31. Vales R, Moura J, Marinheiro R. Energy-aware and adaptive fog storage mechanism with data replication ruled by spatio-temporal content popularity. *J Netw Comput Appl*. 2019;135(6):84-96.
32. Li CL, Wang YP, Chen Y, Luo YL. Energy-efficient fault-tolerant replica management policy with deadline and budget constraints in edge-cloud environment. *J Netw Comput Appl*. 2019;143(1):152-166.
33. Khan AA, Zakarya M, Khan R, Rahman I, Khan M. An energy, performance efficient resource consolidation scheme for heterogeneous cloud datacenters. *J Netw Comput Appl*. 2020;150(15):102497.
34. Chatradhi S. *Hard Drive for Low Power Energy Efficiency in Disk Storage*. San Jose: Hitachi Global Storage Technologies; 2009.
35. Hylick A, Sohan R, Rice A, Jones B. An analysis of hard drive energy consumption. Paper presented at: 2008 IEEE International Symposium on Modeling, Analysis and Simulation of Computers and Telecommunication Systems; 2008; Baltimore, MD: 1–10. https://doi.org/10.1109/MASCOT.2008.4770567.
36. Zedlewski J, Sobti S, Garg N, Fengzhou Z,et al. Modeling hard-disk power consumption. Paper presented at: Proceedings of FAST '03:2nd USENIX Conference onFile and Storage Technologies; March 31–April 2, 2003; San Francisco, CA.
37. Louis B, Mitra K, Saguna S, Åhlund C. CloudSimDisk: energy-aware storage simulation in CloudSim. Paper presented at: Proceedings of the 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC); 2015; Limassol, Cyprus.
38. Calheiros R, Ranjan R, Beloglazov A, Rose C, Buyya R. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Softw Pract Exp*. 2011;41(1):23-50.
39. Graham RL. Bounds on multiprocessing timing anomalies. *SIAM J Appl Math*. 1969;7(2):416-429.
40. Lee LW, Scheuermann P, Vingralek R. File assignment in parallel I/O systems with minimal variance of service time. *IEEE Trans Comput*. 2000;49(2):127-140.
41. Latham R, miller N, Ross R, Carns P. A next-generation parallel file system for Linux clusters: an introduction to the second parallel virtual file system. *Linux World Mag*. 2004;2(1):56-59.
42. Dehghani-Sanij AR, Tharumalinga E, Dusseault MB, Fraser R. Study of Energy storage systems and environmental challenges of batteries. *Renew Sustain Energy Rev*. 2019;104(4):192-208.

How to cite this article: You X, Sun T, Sun D, Liu X, Lv X, Buyya R. K-ear: Extracting data access periodic characteristics for energy-aware data clustering and storing in cloud storage systems. *Concurrency Computat Pract Exper*. 2021;33:e6096. https://doi.org/10.1002/cpe.6096