Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH Open Access

An efficient hybrid optimization of ETL process in data warehouse of cloud architecture

Lina Dinesh[1*] and K. Gayathri Devi[2]

Abstract

In big data, analysis data is collected from different sources in various formats, transforming into the aspect of cleansing the data, customization, and loading it into a Data Warehouse. Extracting data in other formats and transforming it to the required format requires transformation algorithms. This transformation stage has redundancy issues and is stored across any location in the data warehouse, which increases computation costs. The main issues in big data ETL are handling high-dimensional data and maintaining similar data for effective data warehouse usage. Therefore, Extract, Transform, Load (ETL) plays a vital role in extracting meaningful information from the data warehouse and trying to retain the users. This paper proposes hybrid optimization of Swarm Intelligence with a tabu search algorithm for handling big data in a cloud-based architecture-based ETL process. This proposed work overcomes many issues related to complex data storage and retrieval in the data warehouse. Swarm Intelligence algorithms can overcome problems like high dimensional data, dynamical change of huge data and cost optimization in the transformation stage. In this work for the swarm intelligence algorithm, a Grey-Wolf Optimizer (GWO) is implemented to reduce the high dimensionality of data. Tabu Search (TS) is used for clustering the relevant data as a group. Clustering means the segregation of relevant data accurately from the data warehouse. The cluster size in the ETL process can be optimized by the proposed work of (GWO-TS). Therefore, the huge data in the warehouse can be processed within an expected latency.

Keywords Swam intelligence, Grey wolf optimizer, Tabu search, ETL, Latency, Data warehouse

Introduction

ETL (Extraction Transformation Load) plays a vital role in the data warehouse. It gathers the data from various sources and applies the transformation process to implement effective performance. The most essential and complicated ETL process is the transformation stage. Before loading the data into the data warehouse, some processing techniques have employed in the transformation stage. The load stage transfers the transformed data into the data warehouse [1]. In general, the process of the transformation stage is replacing the missing attributes, deletion of irrelevant data columns, typecasting of data, aggregation of data values, removal of null values in the queue, and so on.

The issues in handling data stored in the storage devices are managing the data, processing data, storage cost of the data/cluster, and data security. The ETL process is needed to cleanse data, eliminate null values, replace the missing attributes, etc. In the ETL process, before loading the data into the data warehouse, at the transform phase, data must be appropriately handled by eliminating irrelevant data columns, reduction in repeated data available in the database and collecting data are in various formats; therefore, the normalization process is needed [2].

The literature studies mention many issues in the ETL process. The big data has coincided with enormous

*Correspondence:
Lina Dinesh
linadinesh@sece.ac.in
[1] Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore 641202, India
[2] Department of Electronics and Communication Engineering, Dr.N.G.P.Institute of Technology, Coimbatore 641048, India

different features. In work [2], the author has addressed ETL for big data cost-effectively using a data aggregation model. The cloud-ETL model is used to handle big data. The main problem was identified in the literature as high dimensionality, and the model does not address complex data structures. By taking this issue as main consideration, the proposed research article is developed.

This research defines all three Extract-Transform-Load situations concerning handling complex data maintenance in the load operation. We proposed efficient hybrid optimization of the transformation process in the cloud-based architecture of the data warehouse for data maintenance. The first two processes of extract and transform have been widely concentrated [2], but loading the complex data without redundancy and handling dimensionality reduction in different data formats is a challenging research problem. In the transform phase, this paper proposed two things: reducing original data size by high dimensionality reduction of data using the swarm intelligence algorithm of the grey-wolf optimizer. After that, data clustering is needed by applying tabu search to quickly access data. Tabu search effectively segregates the similarity of data to form a cluster.

The ETL is evaluated using three different algorithms: GWO, tabu search, and hybrid GWO-TS with a clustering approach. The traditional tools of Spark have been used for dimensional data processing. The proposed architectures are validated using Amazon AWS cluster with large 50 GB to 2 TB datasets. The architecture is demonstrated on prediction problems with thousands of features generated on the ETL process. The proposed methodology is performed with sales teams to eliminate service interruptions. Whole ETL data processing is evaluated using sales production settings, and the outcome is improved compared to traditional works.

The loading of the data process in the ETL is gathering more real-time data and optimizing it for processing the data to reduce time analysis [3]. The main contribution of these works is:

- In the transformation phase, a High dimensionality reduction process of data using the swarm intelligence algorithm of the grey-wolf optimizer in the ETL process is done.
- Efficient Data clusters are generated with various nodes by applying the tabu search algorithm.
- The proposed ETL process analyzes the storage cost and efficient data process.

The paper has been organized as follows: Review of literature section discusses the Review of Literature, Proposed GWO-TS methodology section describes the transformation of data in the ETL process, Result & discussion section results and analysis, and Conclusion section describes the conclusion and future work.

Review of literature

In the digital world, a massive volume of data is available in various complex formats, collected from multiple sources like private sectors, corporate companies, government sectors, etc. Collecting, storing, analyzing, and accessing complex data helps many users make better business and company decisions. To achieve it better, the ETL process accomplishes these data processes. Data is shared in distributed cloud architecture, and cost optimization is essential in data loading in cloud-based architecture [4]. Zdravevski et al. [5] describe ETL data scenarios based on aggregation in pre-defined time intervals. The data used in the ETL collects user logs for further processing. Mayo et al. [6] propose analyzing ETL clinical data, storing electronic healthcare data, and planning treatment details. Belo [7] describes the ETL process of real-time analysis of relational algebra for secure computing. Parul et al. [8] propose optimized ETL processing data performance in preparing the data report for deciding their business activities.

In the data report analysis, construct a data warehouse for storing data of heterogeneous type with cleansing and reformatting value. This is based on the concept of ETL, and it is one of the essential components of the data warehouse. The major consumes of the ETL process in terms of time complexity and storage cost of the data warehouse [9]. The data sources in the data warehouse are ERP systems, online transaction processing systems, and Customer Relationship Management systems (CRM). To process the data in the application mentioned above, the required data will be in unstructured format and structured formats like web pages, spreadsheets, images, textual data, etc. [10, 11].

The technical challenge of the data warehouse is handling real-time ETL processing data, and some techniques are required to perform the extraction phase in the real-time data. The extract process in the ETL includes the Enterprise Application Integration system, triggers, log sniffing, and timestamping [12]. The research-oriented ETL process mainly focuses on conceptual-based data [13], physical level [14], and logical design data process [15]. For handling high-level automation processing of data, ETL implements conceptual-based data modeling [16]. In the automatic loading of data, ETL's SysML abstract data process is implemented [17] for transforming data into the data warehouse. The

Table 1 Survey on the ETL process

Author Name	Description
Gant et al. [19] (2020)	ETL based on academic-related data analysis
Kartick et al. [20] (2020)	ETL automation using machine learning process.
Souibgui et al. [4] (2019)	Extract, Transform, and Load (ETL) process has been done to analyze data for the improvable quality of data.
Ghasemaghaei M et al. [21] (2019)	To make decision support system of model based on Structured model for Data Quality of big data analysis and in ETL
Sang-Su Kim et al. [22] (2019)	ETL process of spatial Information in Heterogeneous type of data.
Timmerman, Y et al. [23] (2019)	ETL-based data quality system for a decision support system.
Taleb et al. [24] (2019)	Unstructured Data processing in ETL to provide data quality.
Cichy, C et al. [25] (2019)	Analysis of data quality in the ETL process.
Günther, L. et al. [26] (2019)	To make a decision-processing system of users in the analysis of ETL.
Tian et al. [27] (2019)	Automatic verification for ETL data processing

ETL process transforms textual data based on the BPMN language, automatically updating the commercial tool process [18]. Table 1 describes the survey on the ETL process.

Recently, the satellite data process requires ETL for data maintenance. The article [28] addresses the increasing significance of satellite data in environmental applications. Still, it lacks an exploration of the challenges of handling the massive volume of remote sensing data in real-time. The proposed software solution for data preprocessing is promising, focusing primarily on data ingestion efficiency. However, the article misses an opportunity to delve deeper into the analytics and utilization of preprocessed data, leaving a research gap in understanding how these improved datasets contribute to more accurate environmental monitoring and decision-making.

The research article [29] addresses the significant challenge of handling large datasets and highlights the need for more sophisticated computations in data-intensive environments. It introduces a novel machine-learning approach for reducing dimensional space in large datasets involving data merging, ETL processing, PCA algorithm application, and dashboard visualization. The significant contribution lies in the five-phase hybrid architecture, demonstrated through a promising case study with an epileptic seizure recognition database, offering potential applicability across various domains. The review of various ETL works is discussed in the article [30]. The ETL has wide applications in database processing and maintenance.

The article [31] discusses the critical role of data quality in research information systems (RIS) that integrate data from various sources. It emphasizes the importance of data cleansing and harmonization during the extract, transform, and load (ETL) processes to ensure accurate and reliable research information. The paper focuses on presenting the data transformation process within RIS. It addresses the challenge of controlling and improving data quality during integration, highlighting its relevance in maintaining the integrity of research data. The article [32] highlights the importance of accurate and timely Liquidity Coverage Ratio (LCR) reporting for banks in Indonesia, emphasizing the challenges faced by those still relying on semi-automated processes. It discusses an ETL-based automation method developed with a waterfall software development model to streamline daily reporting. Based on Basel III frameworks, the proposed methodology offers a viable solution for banks to efficiently complete LCR reporting, representing a valuable contribution to the banking industry's regulatory compliance efforts.

The article [33] acknowledges the significance of data analytics in modern organizations. It highlights the crucial role of data integration, particularly through Extract Transform and Load (ETL) processes, in deriving valuable insights. It aims to contribute by conducting a systematic literature review, focusing on approaches, quality attributes, research depth, and challenges in ETL solutions. The study offers valuable insights and trends analysis, which can benefit ETL researchers and practitioners in various domains, demonstrating the importance of keeping pace with evolving ETL methodologies and practices.

The article [34] addresses the challenge of handling large-scale video streams efficiently and cost-effectively, proposing a novel approach called Video

Extract-Transform-Load (V-ETL) akin to data warehousing. The authors introduce Skyscraper, a tailored system for V-ETL, designed to reduce costs while maintaining data throughput. Skyscraper adapts video ingestion pipelines by optimizing sampling rates and resolutions, utilizing on-premises compute, and leveraging cloud resources for peak workloads. Experimental results demonstrate that Skyscraper offers significant cost reductions compared to state-of-the-art systems, providing both cost-efficiency and robustness in large-scale video analytics.

The article [35] highlights the increasing significance of Big Data applications for organizations seeking competitiveness and insights. It emphasizes the critical role of data ingestion and preparation as the initial steps in Big Data projects. The paper reviews various widely used Big Data ingestion and preparation tools, providing insights into their features, advantages, and use cases. Its primary goal is to assist users in making informed choices by helping them select the most suitable tools based on their specific needs and application requirements, thus serving as a valuable resource for those navigating the complex landscape of Big Data technologies. A recent swarm intelligence optimization model, particle swarm optimization, is discussed in [36, 37]. The optimized data handling is introduced in detail. Big data requires ETL [38, 39] to provide efficient scalability in data processing. The complex data handling in big data using ETL has piqued broad interest in recent research. The complex data processing and transformation with redundancy has a key challenging aspect.

A dynamic spectrum allocation scheme suggests that this allocation is not fixed but can change in real-time based on network conditions and demands [40, 41]. Random optimization methods are a class of algorithms used to find optimal solutions to problems where the objective function is not known precisely but can be evaluated at different points in the search space [42–44]. The Wright-Fisher process is a mathematical concept used in population genetics. It models how the genetic composition of a population changes over generations due to random events, such as mutations and genetic drift [45, 46].

The research constraints from the literature are cost-efficient ETL process, scalability issues, high dimensional data processing, and complex data handling. The proposed research focused on the above problems and offered novel techniques in ETL data handling.

Proposed GWO-TS methodology

In many organizations, decision-making is based on the data report analysis of data stored in the data warehouse. For that data report analysis, the ETL (Extract, Transform, and Load) tool plays a vital role in efficiently consolidating data in the data warehouse. This proposed work GWO-TS contains two phases. In Phase 1, Dimensionality reduction of data using Grey Wolf Optimizer. In Phase 2, Clustering the data using Tabu search is done.

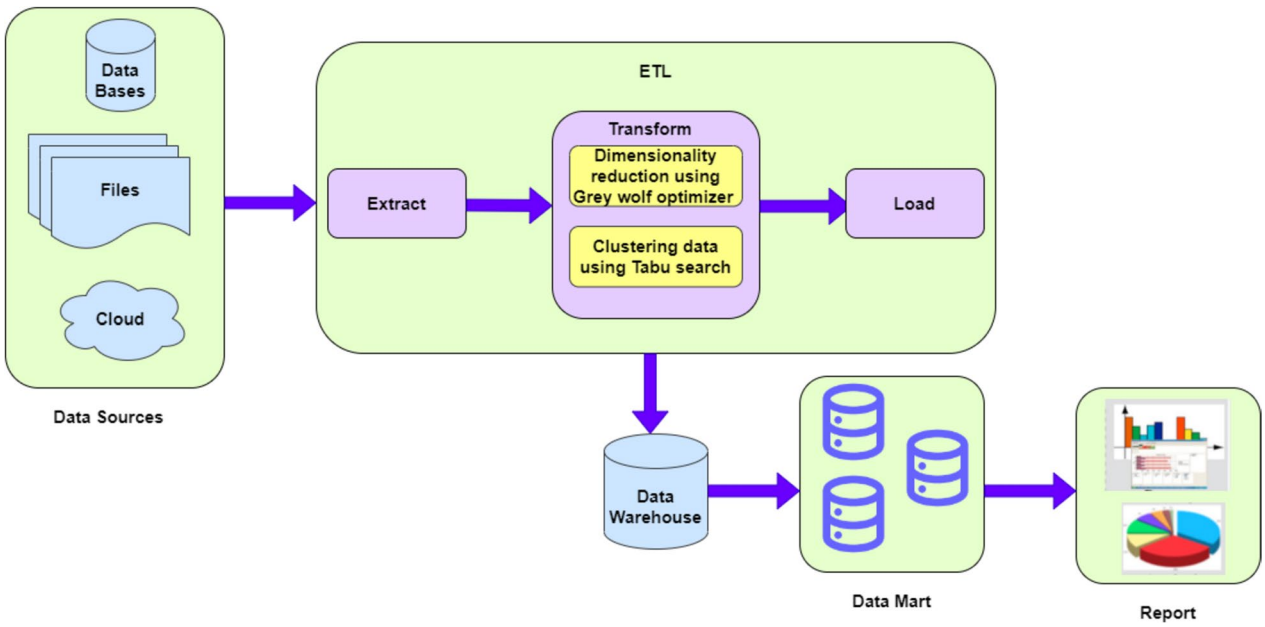

Fig. 1 Architecture of proposed work GWO-TS

The architecture of the proposed work is given in Fig. 1.

In Fig. 1, data is collected from various sources in various formats which is stored in the data warehouse. For data analysis, the efficient accessing of the information ETL process is required.

Data sources

Data is collected from various sources in different formats like structured data, and unstructured data through a cloud-based architecture. These databases are stored in the data warehouse and used for the decision-making system of the particular organization. To improve the decision-making system, apply the ETL process.

ETL process

The process of ETL can be described below. The ETL process is crucial for data integration, ensuring that data from diverse sources is standardized, cleaned, and structured consistently for reporting, analytics, and decision-making purposes. ETL tools and platforms automate many aspects of this process, making it more efficient and reliable in handling large volumes of data.

Extract

Data is extracted from various sources like files, databases, E-Commerce sites, social networking sites, healthcare organizations, private sectors, etc. It is also called a data provider to the databases. These data are used only for accessibility.

Transform (Proposed)

The extracted data are transformed into a standardized format and it can be used by many organizations. In this work, this transformation of data involved two different types of phases.

Phase 1: Dimensionality reduction of data using Grey Wolf Optimizer
Phase 2: Clustering the data using Tabu search.

Phase 1: Dimensionality reduction of data using Grey Wolf Optimizer In the transform stage, reducing the high dimensionality of data by implementing the following steps:

Step 1: Eliminating the duplicate entries of data.
Step 2: To improve the data quality by including the missing attributes, and replacing the null values.
Step 3: Conversion of data format into Unicode.
Step 4: Apply to normalize and de-normalize of data into the desired dimensionality for entering into the data warehouse.

To implement the reducing high dimensionality of data by using swarm intelligence algorithm (SIA) of grey wolf optimizer. SIA is an artificial intelligence based on nature-inspired techniques. It deploys the collective, emerging behavior of interacting with multiple agents by following certain rules. The characteristics of SIA are:

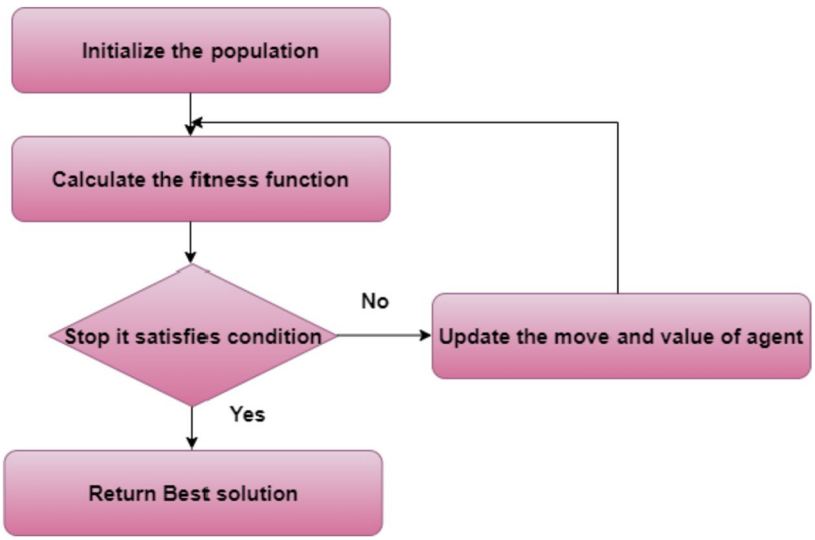

Fig. 2 General architecture of SIA

➢ It can share the information with each other via agents.
➢ Self- organization behavior of collective intelligence of every agent.
➢ In the dynamic change environment, it can be adaptable and immediate responsibility.
➢ It can be supported in real-time problems

In SIA algorithms, an individual agent from population executes in search space and communicate with one another. Every agent produces the possible solution and based on the fitness function value solution is evaluated. Figure 2 shows that general architecture of SIA.

Grey Wolf Optimizer (GWO)

In this work, we are using swarm intelligence of grey wolf optimizer (GWO) technique for the reduction of dimensionality of data. GWO starts with randomly initialize the agents in the wolf pack. The position of each agent is assigned by vector $w_i(k)$.

$$w_j(t) = [w_j^1(t) \ldots..w_j^d(t) \ldots.w_j^q(t)]^T, j = 1, 2, \ldots N. \quad (1)$$

Here $w_j^d(t)$ is the position of j^{th} agent in the d^{th} dimension, $d = 1, \ldots q, k$ is the current iteration index value, and it extends up to maximum number of iterations.The inspiration of grey wolf in the social intelligence prefers to live as a group of minimum 5 to 12 individual agents. GWO technique is used for solving the problems related with continuous and real-time optimization problems. To simulate the leadership hierarchy of wolf pack and its social behaviour of hunting the prey is done by four levels namely alpha (α), beta (β), delta (δ) and omega (ω).

Prey hunting process of wolf is conducted as follows: search and chasing the prey, encircle it and hunting the prey. At first, grey wolves search and track the prey, then alpha(α), leads other wolves to encircle the prey and covered in all directions. The responsible of Beta (β) wolf passing the prey message of alpha (α) wolf to remining other wolves in the pack and helps alpha (α) wolf in hunting the prey. The final level of omega (ω) wolves are allowed to eat food at the end. Remaining wolves are considered as delta (δ) and its duties are caretaker of wounded wolves, act as a defender from external attacks of enemies and contributes for hunting the prey. After completion of hunting prey under the guidance of leaders α, β, and δ wolves. Now,each wolf has to update its position by getting help from leading wolves of α, β, and δ. Now, the leading wolves are responsible for searching prey in all directions. The leading wolf provides the best solution and optimized guidance to get prey in each iteration. The pseudocode of GWO is given below:

Input: Data from various sources

Output: Dimensionally of data is reduced (Select only best position of fitness value)

Step 1. Generate dimension search space by randomly initializing the population of grey wolf, represented by M agents with its position.

Step 2: Initially, the index value of iteration $t = 0$and maximum iteration is t_{max}.

Step 3:while it does not reach t_{max}iteration do

Step 4: Evaluate the first best fitness values of α , second best fitness valueβ, third best fitness valueδ.

Step 5: Calculate the search coefficient by using

$$p_l^d(t) = p^d(t)\left(2r_{1l}^d - 1\right),$$

$$q_l^d(t) = \left(2r_{2l}^d\right), l \in \{\alpha, \beta, \delta\} \quad (2)$$

Here r_{1l}^d and r_{2l}^d are distributed uniformly within the range of random numbers $0 \le r_{1l}^d \le 1$, $0 \le r_{2l}^d \le 1$, $d = 1,2, \ldots., n$. The coefficient $p^d(t)$ are linearly decrease from 2 to 0 and increase the iteration number.

$$p^d(t) = 2[1 - (t-1)/t_{max} - 1], d = 1,2, \ldots., n. \quad (3)$$

Step 6: For each wolf do

Step 7: The wolf (agent) is moved to new position and update its position by using:

$$w^{ld}(t+1) = w^{ld}(t) - p^d(t)r_l^{ld}(t), d = 1,2, \ldots, m, and\ l = 1,2, \ldots., N, l \in \{\alpha, \beta, \delta\}. \quad (4)$$

The updated position of wolf is obtained by arithmetic mean value of α, β, δ wolf is :

$$w_i^t\left((t+1) = w^{\alpha d}(t+1) + w^{\beta d}(t+1) + w^{\delta d}(t+1)\right)/3 \quad (5)$$

The vector solution of Eqn (5)can be defined as:

$$w_i^t\left((t+1) = w^{\alpha}(t+1) + w^{\beta}(t+1) + w^{\delta}(t+1)\right)/3 \quad (6)$$

Step 8: the iteration index value t is incremented and it continues until it reaches t_{max}.

Step 9: return w^{α}.

Algorithm 1. Grey Wolf Optimizer

In the algorithm 1, it α wolf, β wolf, δ wolf have better knowledge about the data collected from various sources and it eliminates the irrelevant, missing attributes, cleansing and produce the best information. By this way its dimensionality gets reduced.

Phase 2: Clustering the data using Tabu search The process of clustering is grouping of same patterns of data as a group. Tabu search algorithm is based on partition clustering. In partition clustering, same patterns of data are partition into clusters and split the same pattern into groups by increasing it clusters. This type of partitional cluster is prototype-based clustering. Here each cluster is defined as prototype, and sum of distance between prototype and pattern is considered as objective function. In general, prototype represents centre of the cluster. In this work, K-means algorithm with tabu search. K-means algorithm is one of the types of prototype-based clustering. In each iteration, it reduces the average distortion. Therefore, K-means algorithm trapped the local optima and K-means algorithm is described below:

Step 1: Randomly choose M clustering pattern as initial centroid of clusters.
Step 2: Assume $iter = 0$.
Step 3: In each clustering pattern, identify the nearest centroid. Put Y_j is the cluster (or partition set) $parti_i$. Take $centr_i$ is the nearest centroid to Y_j.
Step 4: After partition the sets $parti_i$ and $1 \geq i \geq n$; increment by n and evaluate the overall average distortion is given by:

$$Disto_n = \frac{1}{ct} \sum_{i=1}^{N} \sum_{j=1}^{ct} Disto\left(Y_j^i, centr_i\right) \quad (7)$$

Here $parti_i = \{Y_1^i, Y_2^i, \ldots . Y_{ct_i}^i\}$, ct_i is the number of clustering patterns.
Step 5: In all disjoint partitioned set, evaluate the centroid by using:

$$centr_i = \frac{1}{ct} \sum_{j=1}^{ct_i} Y_j^i \quad (8)$$

Step 6: If $(Disto_{n-1} - Disto_n)/Disto_n > \varepsilon$ go to Step 3; else terminate the program.
Where, ε is small distortion threshold value.

Tabu search technique is a metaheuristic algorithm and designed to optimize the problem of ETL process

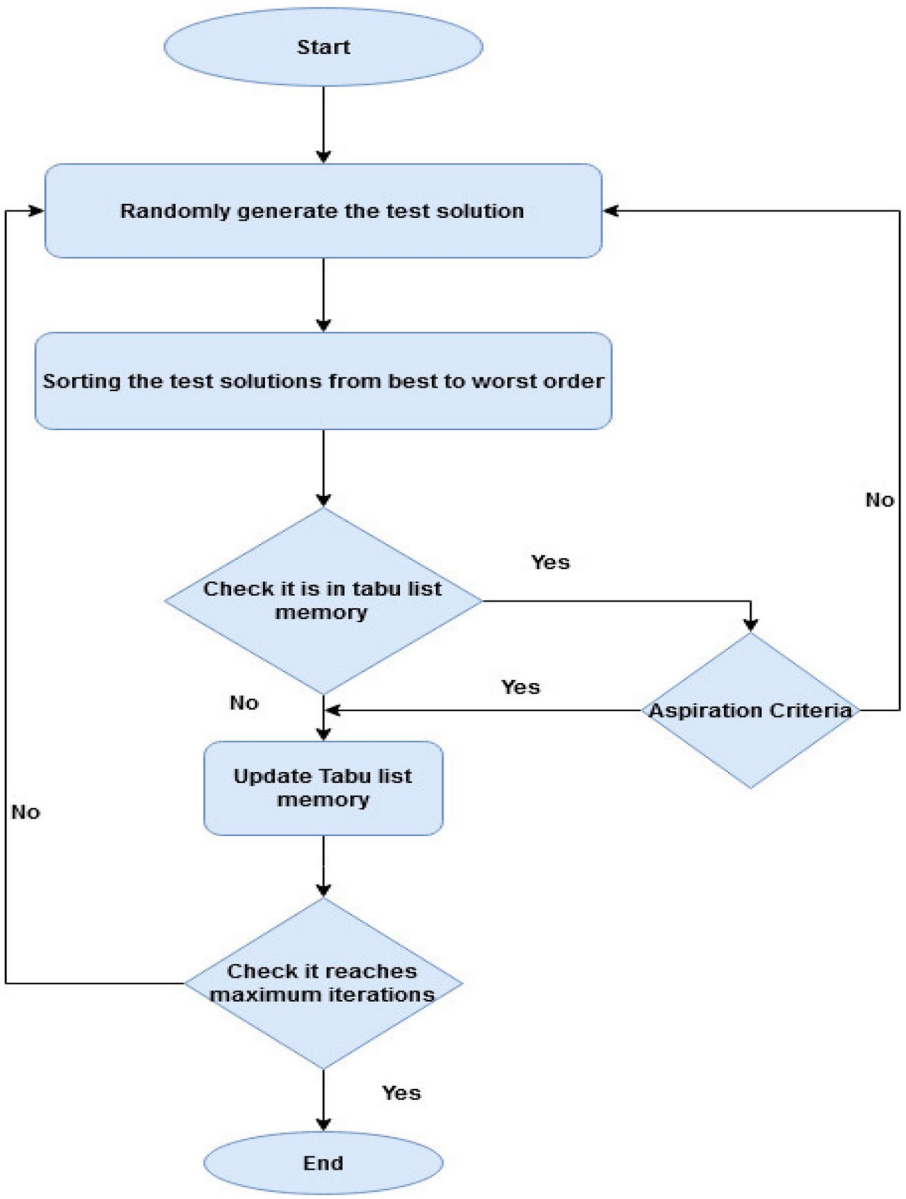

Fig. 3 Procedure of Tabu search

in transform the data at a less time, and occupies minimum storage space by the reduction of dimensionality. In the tabu search, it optimizes the sequence of moves and choose the test solution. This algorithm uses neighbourhood pattern for exploring the search space. It has a short-term memory called as tabu. It contains list

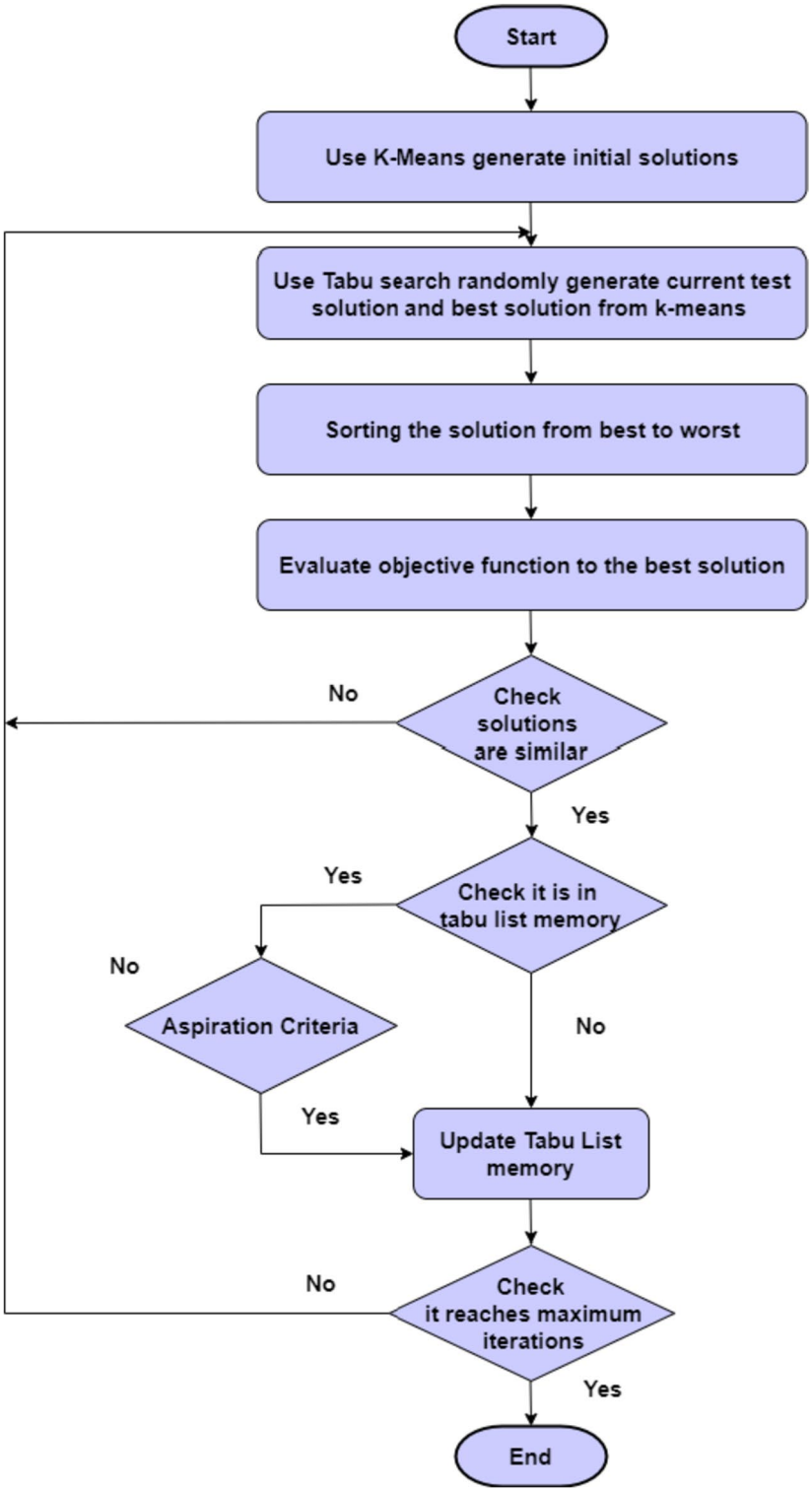

Fig. 4 Outline of K-means with Tabu search

of forbid moves and prevents to and fro movements between already available solutions in the tabu lit and it is called as cycling. The drawback of using k-means algorithm is local minima. In order to overcome this, tabu search allows the poor move solution to yield better performance by improving the objective function. In the Tabu list memory it holds the best solution in the search.

It Randomly select each move and generate the current best solution. For that move, generate random number $0 \leq rnd \leq 1$. If $rnd \geq pthresh_t$ then it is assigned to cluster j. This cluster j also generated randomly. But it is not assigned to the same cluster of current solution. $pthresh_t$ represents that predefined threshold value. For selecting the current best solution, the current new test solution satisfies the aspiration criteria and which is better than current best solution in the tabu list memory. Then current new test solution is considered as current best solution and included in the tabu list memory. If all test solutions are available in the tabu list memory then for generating test solution again from the current best solution. Figure 3 shows that procedure for Tabu search algorithm.

In this work, tabu search algorithm is combined with K-means for generate the cluster. Figure 4 shows that outline procedure of K-means with Tabu search algorithm.

To improve the clustering tabu search algorithm is implemented in k-means algorithm. The improved version of tabu search algorithm avoid cycling and local minima. In all iterations it provides same best solution for each move then replace the current best solution with best solution of all iterations and update the tabu list memory.

Load

Transforming this best solution of data from GWO-TS implementation, into targeted data base like data ware house and it is used for various purposes. Loading is the final step in the ETL process. During this phase, the transformed data is loaded into a target system, such as a data warehouse or database.

Result & discussion

In this section our proposed work is implemented by using the data set of AWS ([47], https://aws.amazon.com/ec2/instance-types/). It is implemented with four ETL scenarios with GWO-TS. Table 2 shows that details about data set for experimental evaluation of four scenarios. The cluster size for all scenarios are 5,10,15,20,25,50, 100 nodes. The experimental evaluation is implemented by using the algorithms of Grey wolf optimizer (GWO) [48], Tabu Search (TS) [49], GWO-TS.

Table 2 Details of data set

Data Source Type	ETL data set in CSV Format
Data Source Columns	31
Data aggregated Columns	62
Data unaggregated Columns	86
Distributed data storage S3 objects	550
Data Source Size (GB)	53
Data Source Rows	137 M
Destination Unaggregated rows	137 M
Destination Unaggregated size (GB)	94

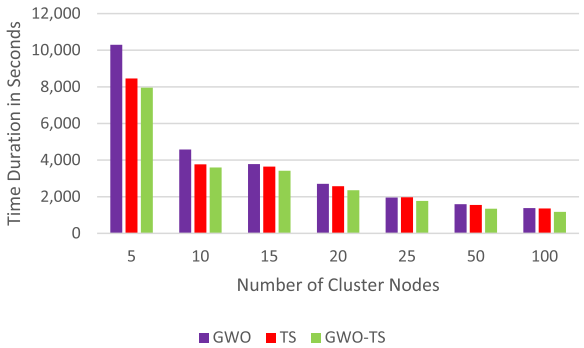

Fig. 5 Time duration of ETL

Table 3 After applying reduction in high dimensionality of data

Algorithm	Original Data Source Size in GB	After Applying reduction in high dimensionality (GB)
GWO	53	51.89
TS	53	52.11
GWO-TS	53	50.45

Table 2 shows that details of data set in CSV format. Figure 5 shows that duration time of ETL data set used in various algorithm.

Experimental analysis process implemented with swarm intelligence of grey wolf optimizer, Tabu search, and it is compared with our proposed work GWO-TS. Our proposed work produces a minimum time duration, and also, when the nodes in the cluster increase, it gives a minimum time duration [50, 51]. The cluster size cost optimization is implemented in source data size and the time required to finish the ETL process. Figure 5 shows that time duration in seconds has been experimented with starting from a small cluster size with five nodes, ten nodes, 15 nodes, 20 nodes, 25 nodes, 50 nodes, and 100 nodes. In Table 2, the

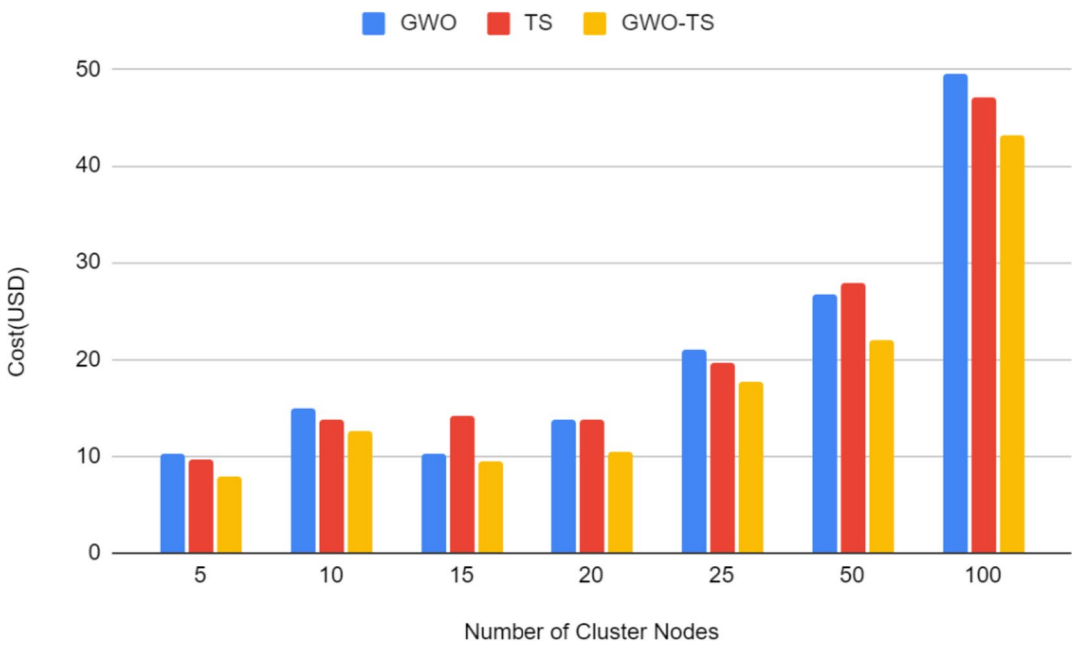

Fig. 6 Cost of ETL process

source data size is 53 GB and the time required to implement the ETL process depends on selecting several nodes in the cluster size. Table 3 shows the comparison of various algorithms in reducing the high dimensionality of data.

From the Table 3, it observes that reduction of data source size after applying the reduction of high dimensionality of data using proposed work gives better performance. Figure 6 shows that cost of ETL process in various algorithms.

Figure 6 shows that the cost of the ETL process depends on cluster size within the stipulated period (1 h) and using various algorithms. Our proposed work gives better performance compared with other algorithms. It requires the minimum cost of the ETL process. When the number of nodes increases, it creates the most cost-effectiveness. Figure 7 shows that average iteration needed to produce the best result.

In the analysis of Fig. 7 which shows that, the size of the original data is 53 GB. After applying Algorithm 1, it reduces into 50.45 GB. To produce the best solution in the original data size which more time consuming and in the reduced data size requires minimum time consumption along with reduce number of iterations. Figure 8 shows that quality of improvement compare it with swarm intelligence algorithm.

As shown in Fig. 8, the ETL data set contains various data from different sources in varied structures and formats. After applying the swarm intelligence algorithm of the grey wolf, the optimizer reduces the original data size. The cost of the operation takes the quality of

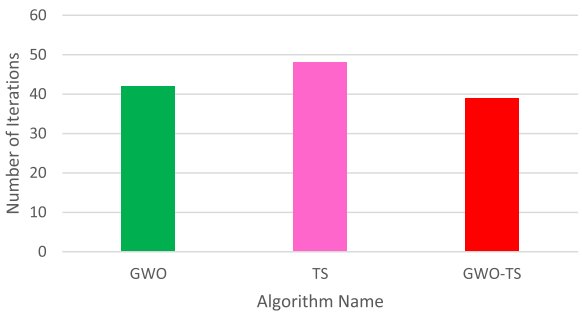

Fig. 7 Average iteration

improvement done in the nodes of clusters by using the proposed work GWO-TS. Therefore, it needs less storage space, less storage cost, and requires minimum computation time. Compared with our proposed work, GWO-TS is faster than other algorithms and produces an optimal solution that effectively recognizes the data stored in the cluster's nodes.

Conclusion

In this paper, we have efficiently proposed a hybrid optimization of the cloud-based architecture of the ETL process in the data warehouse. They extract the data and transform it in the data warehouse using distributed cloud storage that uses the nodes in the cluster. Before loading data into the data warehouse, the transform stage reduces the original data size by using

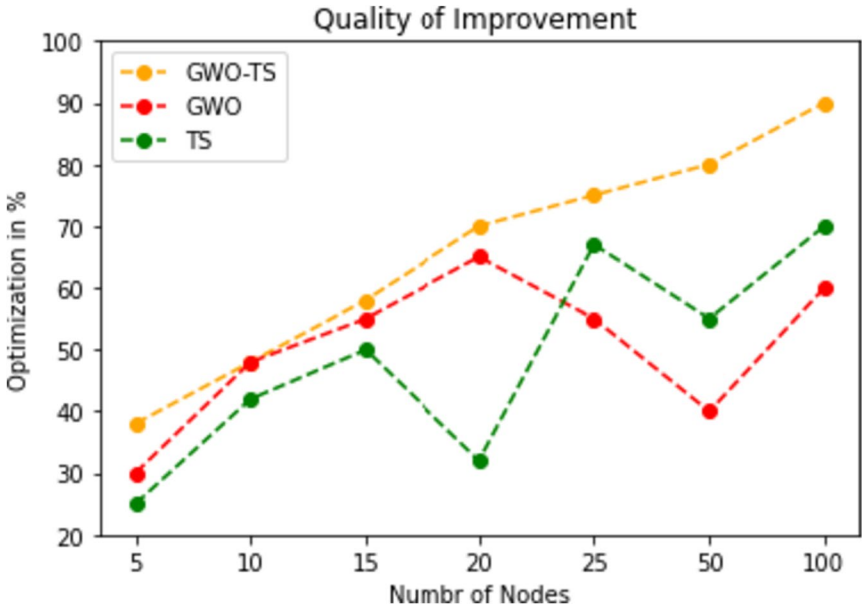

Fig. 8 Quality of improvement

the swarm intelligence algorithm of the grey wolf optimizer. After reducing the data size, it contains various data types in different formats. Our proposed work resulted in less storage space, less cost, minimum computation time, improviser in quality, and cluster size optimization. In the future, this work may be extended up to applying other swarm intelligence algorithms for the dynamic real-time application of the ETL process.

Authors' contributions
Lina Dinesh- Execution and Implementation, K. Gayathri Devi – Drafting.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 July 2023 Accepted: 7 December 2023
Published online: 08 January 2024

References

1. Zdravevski E, Lameski P, Dimitrievski A, Grzegorowski M, Apanowicz C (2019) Cluster-size optimization within a cloud-based ETL framework for Big Data. In: 2019 IEEE International Conference on Big Data (IEEE BigData 2019), at Los Angles, USA, pp 3754–3763
2. Aziz O, Anees T, Mehmood E (2021) An efficient data access approach with queue and stack in optimized hybrid join. IEEE Access 9:41261–41274.
3. Mehra KK et al (2017) Extract, transform and load (ETL) system and method. U.S. patent no. 9
4. Souigbui M, Augui F, Zammali S, Cherfi S, Yahia SB (2019) Data quality in ETL process: a preliminary study. Procedia Comput Sci 159:676–687. Elsevier
5. Zdravevski E, Apanowicz C, Stencel K, Slezak D (2019) Scalable cloud-based ETL for self-serving analytics. In: Perner P (ed) Advances in data mining: applications and theoretical aspects. 19th Industrial Conference, ICDM 2019. Springer International Publishing, Cham, pp 387–394
6. Mayo C et al (2016) Taming big data: implementation of a clinical use-case driven architecture. Int J Radiat Oncol Biol Phys 96(2):E417-8
7. Belo VS (2015) Using relational algebra on the specification of real world ETL processes. Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, (CIT/IUCC/DASC/PICO), IEEE International Conference on. IEEE, Liverpool, pp 861–866
8. Parul SN, Teggihalli S (2015) Performance optimization forextraction, transformation, loading and reporting of data. In: Communication Technologies (GCCT), 2015 Global Conference on. IEEE, Thuckalay, pp 516–519
9. Vassiliadis P (2009) A survey of extract - transform - load technology. Int J Data Warehous Min 5(3):1–27
10. Vassiliadis P, Simitsis A (2009) Extraction, transformation, and loading. In Encyclopedia of database systems. Springer, pp 1095–1101
11. Liu C, Wu T, Li Z, Ma T, Huang J (2022) Robust online tensor completion for IoT streaming data recovery. In: IEEE transactions on neural networks and learning systems
12. Zhou X, Zhang L (2022) SA-FPN: an effective feature pyramid network for crowded human detection. Appl Intell 52(11):12556–12568
13. Li S, Chen H, Chen Y, Xiong Y, Song Z (2023) Hybrid method with parallel-factor theory, a support vector machine, and particle filter optimization for intelligent machinery failure identification. Machines 11(8):837
14. Liang X, Huang Z, Yang S, Qiu L (2018) Device-free motion & trajectory detection via RFID. ACM Trans Embed Comput Syst 17(4):78
15. Cao B, Zhao J, Gu Y, Fan S, Yang P (2020) Security-aware industrial wireless sensor network deployment optimization. IEEE Trans Industr Inform 16(8):5309–5316
16. Skoutas D, Simitsis A (2006) Designing ETL processes using semantic web technologies. In: Proceedings of the 9th international ACM workshop on data warehousing and OLAP, USA. pp 67–74

17. Peng Y, Zhao Y, Hu J (2023) On the role of community structure in evolution of opinion formation: a new bounded confidence opinion dynamics. Inf Sci 621:672–690
18. Zhao K, Jia Z, Jia F, Shao H (2023) Multi-scale integrated deep self-attention network for predicting remaining useful life of aero-engine. Eng Appl Artif Intell 120:105860
19. Mhon GGW, Kham NSM (2020) ETL pre-processing with multiple data sources for academic data analysis. In: IEEE Conference on Computer Applications (ICCA). pp 1–5
20. Mondal KC, Biswas N, Saha S (2020) Role of machine learning in ETL automation
21. Ghasemaghaei M, Calic G (2019) Can big data improve firm decision quality? The role of data quality and data diagnosticity. Decis Support Syst 120:38–49
22. Kim S-S, Lee W-R, Go J-H (2019) A study on utilization of spatial information in heterogeneous system based on Apache NiFi. pp. 1117–1119
23. Timmerman Y, Bronselaer A (2019) Measuring data quality in information systems research. Decis Support Syst 126(February):113138
24. Taleb I, Serhani MA, Dssouli R (2019) Big data quality assessment model for unstructured data. In: 13th International Conference on Innovations in Information Technology, IIT 2018. pp 69–74
25. Cichy C, Rass S (2019) An overview of data quality framework. IEEE Access 7:24634–24648
26. Günther LC, Colangelo E, Wiendahl HH, Bauer C (2019) Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. Procedia Manuf 29:583–591
27. Tian Q, Liu M, Min L, An J, Lu X, Duan H (2019) An automated data verification approach for improving data quality in a clinical registry. Comput Methods Programs Biomed 181:104840
28. Semlali BEB, El Amrani C, Ortiz G (2020) SAT-ETL-Integrator: an extract-transform-load software for satellite big data ingestion. J Appl Remote Sens 14(1):018501
29. Terol RM, Reina AR, Ziaei S, Gil D (2020) A machine learning approach to reduce dimensional space in large datasets. IEEE Access 8:148181–148192
30. Galici R, Ordile L, Marchesi M, Pinna A, Tonelli R (2020) Applying the ETL process to blockchain data. Prospect and findings. Information 11(4):204
31. Azeroual O, Saake G, Abuosba M (2019) ETL best practices for data quality checks in RIS databases. Informatics 6(1):10
32. Hendayun M, Yulianto E, Rusdi JF, Setiawan A, Ilman B (2021) Extract transform load process in banking reporting system. MethodsX 8:101260
33. Nwokeji JC, Matovu R (2021) A systematic literature review on big data extraction, transformation and loading (etl). In: Intelligent computing: proceedings of the 2021 computing conference, volume 2. Springer International Publishing, pp 308–324
34. Kossmann F, Wu Z, Lai E, Tatbul N, Cao L, Kraska T, Madden S (2023) Extract-transform-load for video streams. Proc VLDB Endow 16(9):2302–2315
35. Alwidian J, Rahman SA, Gnaim M, Al-Taharwah F (2020) Big data ingestion and preparation tools. Mod Appl Sci 14(9):12–27
36. Ul Hassan N, Bangyal WH, Ali Khan MS, Nisar K, Ag. Ibrahim AA, Rawat DB (2021) Improved opposition-based particle swarm optimization algorithm for global optimization. Symmetry 13(12):2280
37. Fan W, Yang L, Bouguila N (2022) Unsupervised grouped axial data modeling via hierarchical Bayesian nonparametric models with Watson distributions. IEEE Trans Pattern Anal Mach Intell 44:9654-68
38. Zhang X, Wen S, Yan L, Feng J, Xia Y (2022) A hybrid-convolution spatial–temporal recurrent network for traffic flow prediction. Comput J c171
39. Li B, Zhou X, Ning Z, Guan X, Yiu KC (2022) Dynamic event-triggered security control for networked control systems with cyber-attacks: a model predictive control approach. Inf Sci 612:384–398
40. Wu H, Jin S, Yue W (2022) Pricing policy for a dynamic spectrum allocation scheme with batch requests and impatient packets in cognitive radio networks. J Syst Sci Syst Eng 31(2):133–149
41. Wang Y, Han X, Jin S (2022) MAP based modeling method and performance study of a task offloading scheme with time-correlated traffic and VM repair in MEC systems. Wireless Networks 29:47-68
42. Zhang J, Tang Y, Wang H, Xu K (2022) ASRO-DIO: Active subspace random optimization based depth inertial odometry. IEEE Trans Robot 1–13
43. Ni Q, Guo J, Wu W, Wang H, Wu J (2022) Continuous influence-based community partition for social networks. IEEE Trans Netw Sci Eng 9(3):1187–1197
44. Xu Y, Chen H, Wang Z, Yin J, Shen Q, Wang D et al (2023) Multi-factor sequential re-ranking with perception-aware diversification. Paper presented at the KDD '23
45. Tan J, Jin H, Hu H, Hu R, Zhang H et al (2022) WF-MTD: Evolutionary decision method for moving target defense based on Wright-Fisher process. In: IEEE transactions on dependable and secure computing
46. Cheng B, Wang M, Zhao S, Zhai Z, Zhu D et al (2017) Situation-aware dynamic service coordination in an IoT environment. IEEE/ACM Trans Netw 25(4):2082–2095
47. Mathew S (2017) Overview of Amazon Web Services. Accessed 6 Apr 2019
48. Zhang J, Liu Y, Li Z, Lu Y (2023) Forecast-assisted service function chain dynamic deployment for SDN/NFV-enabled cloud management systems. IEEE Syst J 17:4371–4382
49. Yuan H, Yang B (2022) System dynamics approach for evaluating the interconnection performance of cross-border transport infrastructure. J Manag Eng 38(3):04022008
50. Guo F, Zhou W, Lu Q, Zhang C (2022) Path extension similarity link prediction method based on matrix algebra in directed networks. Comput Commun 187:83–92
51. Li Q, Lin H, Tan X, Du S (2020) Consensus for multiagent-based supply chain systems under switching topology and uncertain demands. IEEE Trans Syst Man Cybern 50(12):4905–18

Publisher's Note