European Journal of Business Startups and Open Society

Vol. 1 No. 1 (2021): EJBSOS ISSN: 2795-9228

http://innovatus.es/index.php/ejbsos

Harnessing Data Science for Enhancing ERP Cloud Security and Data Integrity: A Review

Arjun Reddy Kunduru

Independent Researcher, Pennsylvania, USA

Abstract: Cloud computing provides on-demand services over the internet, offering flexibility and cost savings. However, security and privacy risks hinder wider adoption. This review examines recent data science techniques to boost cloud security and data integrity. Specifically, trust models quantifying risks, task scheduling algorithms like Heterogeneous Earliest Finish Time (HEFT), and metaheuristic optimization methods like genetic algorithms are discussed. The strengths and weaknesses of these approaches are analyzed. Overall, a multi-pronged framework combining trust computation, heuristic scheduling, and metaheuristic optimization emerges as a robust paradigm for balancing security, efficiency, and quality of service in cloud environments. Additional research on computational trust, adaptive scheduling, and scalable optimization can help fully realize the potential of data science for advancing cloud security.

Key words: HEFT, algorithms, Data science, ERP

INTRODUCTION

Cloud computing refers to provisioning computing resources such as storage, servers, applications, and processing power as on-demand services over the internet. Major benefits compared to on-premise infrastructure include flexibility, elastic scalability, and cost savings. However, security and privacy concerns remain foremost barriers to expanded enterprise cloud adoption [1].

Thus, ensuring confidentiality, integrity, and availability of data in the cloud is an active research challenge. This review synthesizes recent data science techniques to enhance security for cloud-based data storage and processing.

First, trust models quantifying risks using multiple factors are covered. Next, algorithms for reliable and efficient cloud task scheduling are examined. Finally, metaheuristic optimization techniques for secure resource allocation are reviewed. For each category, representative methods are analyzed and strengths as well as limitations are highlighted. Overall, the review demonstrates how data science can provide robust security to facilitate cloud adoption across domains like healthcare, finance, and government.

Computational Trust Models

Trust refers to the confidence that a system or service will behave as expected despite vulnerabilities [2]. Computational trust models allow quantifying security risk using various parameters, providing a holistic perspective compared to assessing threats individually. Trust models tailored for cloud environments incorporate relevant factors such as [3]:

By consolidating these factors into a trust score, customers can choose cloud services fulfilling their security requirements. The score can also update dynamically based on runtime monitoring. Next, two sample trust models are reviewed.

Sai Prasad et al. [4] developed a trust model to validate cloud storage integrity. They identified five key attributes: encryption strength, access controls, uptime/availability, audit logs, and geographic redundancy. Each attribute had subfactors weighted based on expert surveys. The final trust score indicates risk of data corruption during storage or transfer. For optimization, the Ant Lion Optimizer (ALO) metaheuristic was used. Experiments found ALO-based trust optimization achieved better accuracy and convergence versus Particle Swarm Optimization (PSO).

Barona et al. [5] surveyed academic trust models for cloud systems. They found confidentiality and compliance were often absent. Thus they proposed a model incorporating: identity management, access control, availability, incident response, and compliance. Fuzzy logic computed trust scores for each factor, aggregated into an overall level. The framework aimed to provide assurance to customers regarding contractual privacy and security guarantees from providers.

In summary, adaptable trust models enable a nuanced view of cloud risks. Both weighted attribute models and fuzzy logic systems have been applied successfully. Challenges remain in selecting appropriate trust factors and balancing model complexity with interpretability. Runtime validation and updating of trust scores warrants further research.

Heuristic Algorithms for Cloud Task Scheduling

Efficient and reliable workload scheduling across distributed cloud resources is vital for quality of service. However, optimal scheduling on heterogeneous, network-based platforms is an NP-hard problem. Heuristic algorithms provide good solutions in reasonable timeframes for large cloud settings [6]. This section reviews two popular heuristics: Heterogeneous Earliest Finish Time (HEFT) and Sufferage.

HEFT is a list scheduling technique which prioritizes tasks based on upward rank and assigns them to resources minimizing finish time. Upward rank represents the longest path from a task to an exit node, incorporating computation costs on candidate machines and data transfer times between resources [7]. HEFT iterates over the ranked list, allotting each task to the machine giving earliest finish.

Sufferage ranks tasks by the difference in completion time between best and second-best machine assignments. Intuitively, tasks with higher sufferage value benefit more from optimal allocation. Sufferage seeks to maximize scheduler flexibility by prioritizing high impact decisions [8].

Both offer polynomial time complexity and efficiently generate schedules for large DAG workflows on heterogeneous cloud platforms. Bahmani et al. [9] evaluated HEFT and Sufferage as part of a cloud workflow engine. HEFT performed better overall, with Sufferage comparable on some graph structures. The study used simulated annealing to further refine the heuristic schedules. Extensions like combining Sufferage ranking with HEFT assignment have been proposed.

Limitations include lack of optimality guarantees. Performance depends heavily on careful weighting of computation costs and data transfer times. Heuristics may also not adapt well to dynamic cloud conditions with fluctuating resource availability and network performance.

However, heuristics remain essential tools for real-world cloud scheduling due to their speed, scalability, and near-optimal outcomes.

Metaheuristic Algorithms for Resource Optimization

Metaheuristics provide high-level problem-solving frameworks integrating heuristics with randomization and iterative improvement to circumvent local optima [10]. Popular metaheuristic algorithms like evolutionary computation and swarm intelligence have been applied to optimize secure and efficient cloud resource allocation. This section examines genetic algorithms and particle swarm optimization as examples.

Genetic algorithms emulate natural selection. Candidate solutions are encoded as chromosomes within a population. Chromosomes reproduce via crossover and randomly mutate between generations. Solutions with higher fitness have higher likelihood of selection for reproduction. Over successive generations, the population converges to near-optimal solutions.

Sindhu and Muneeswaran [11] developed a genetic algorithm for workflow scheduling across virtual machines (VMs). The chromosome encoded task-to-VM mappings. Fitness incorporated metrics like makespan, cost, energy, and resource utilization. Crossover exchanged mappings between solutions. Mutation altered the assigned VM for a random task. Results showed the genetic algorithm minimized makespan and energy versus baseline heuristics.

Particle swarm optimization (PSO) models the social dynamics of organisms like bird flocking. Particles representing solutions traverse the search space based on personal best positions and the global best. This enables rapid convergence to optima.

Yassa et al. [12] used PSO to optimize VM placement and workload distribution, encoding public/private VM mixes and task assignments in particle positions. Fitness aimed to minimize cost under workload deadlines and security levels. On simulation data, PSO provided significantly cheaper solutions than round-robin or random allocation with comparable quality of service. However, computational overhead grew exponentially with problem size.

In summary, metaheuristics effectively navigate large, complex cloud optimization search spaces. However efficiency and scalability challenges remain, especially for real-time adaptations. Hybrid models applying metaheuristics to refine heuristic solutions appear promising. More research is essential to deploy academic algorithms in production cloud environments.

Challenges and Future Outlook

Despite extensive research, leveraging data science to advance cloud security and privacy still faces many open challenges:

- Trust computation: Methods to quantify security risk based on multiple dynamic factors remain in development. How to best select, combine, and update relevant trust metrics based on the cloud service and risk appetite needs further study.
- Scheduler design: More work is needed on heuristic schedulers accommodating heterogeneous, distributed cloud architectures including edge resources. Adaptivity to changing conditions and online multi-objective optimization are crucial but lacking in current schedulers.
- Scalable optimization: Metaheuristics like GA and PSO have high overhead for large-scale systems with thousands of tasks and resources. Lightweight approximations suitable for real-time

control warrant investigation. Hybrid models applying metaheuristics judiciously may improve scalability.

- Validation: Most techniques are evaluated only on simulated cloud environments. Reproducible testbeds and studies on real-world platforms are necessary to validate security and performance gains.

- Automated security: Incorporating data science techniques into auto-scaling, self-healing cloud frameworks could enable "security by design". Advances are needed in trust feedback loops, online learning, and software integration.

By addressing these challenges through cross-disciplinary collaboration, the full potential of data science for advancing cloud security can be realized over the next decade.

Conclusion

This review synthesized recent data science approaches to enhance cloud security and data integrity:

Trust models aggregating relevant factors like access policies and data integrity provide a nuanced risk perspective compared to evaluating threats in isolation. Adaptable frameworks allow incorporating security metrics tailored to an organization's cloud risk profile.Heuristic algorithms such as HEFT and Sufferage enable efficient task scheduling across distributed cloud resources within reasonable timeframes. HEFT prioritizes critical path tasks while Sufferage schedules highest impact tasks first. Extensions refining heuristic solutions using metaheuristics can further boost performance. Metaheuristic optimization leverages AI to navigate large, multi-objective cloud optimization search spaces and skirt local optima. Genetic algorithms mimic natural selection while particle swarm optimization reflects swarm behavior. Both capably optimize secure, efficient resource allocation, albeit with current efficiency and scalability shortcomings.

Overall, data science has already improved cloud security practices through trust management, intelligent workload scheduling, and automated optimization. However, progress on computational trust models, adaptive scheduling, scalable optimization, validation, and self-protecting systems is important to fully harness the potential of data science for advancing cloud security. As cloud adoption continues accelerating across industries, translating these technologies from research into practice will become increasingly critical.

References:

1. K. Hashizume et al., "An analysis of security issues for cloud computing", J. Internet Services and Applications 4, 5 (2013).
2. S. Nepal et al., "Trustworthy processing of healthcare big data in hybrid clouds", IEEE Cloud Computing 2, 2 (2015), 78-84.
3. R.K.L. Ko et al., "TrustCloud: A framework for accountability and trust in cloud computing", IEEE World Congress on Services 2 (2011), 584-588.
4. A.V.H. Sai Prasad et al., "A trust model of cloud scheduling based on data integrity using ant lion optimizer", Turkish Journal of Computer and Mathematics Education 12, 13 (2021), 4876-4886.

5. R. Barona and N. Anita, "A survey on data breach challenges in cloud computing security: Issues and threats", 2017 IEEE International Conference on Circuit, Power and Computing Technologies (ICCPCT) (2017), 1-7.
6. H. Arabnejad and J.G. Barbosa, "List scheduling algorithm for heterogeneous systems by an optimistic cost table", IEEE Transactions on Parallel and Distributed Systems 25, 3 (2013), 682-694.
7. H. Topcuoglu et al., "Performance-effective and low-complexity task scheduling for heterogeneous computing", IEEE Transactions on Parallel and Distributed Systems 13, 3 (2002), 260–274.
8. M.A. Iverson et al., "Statistical prediction of task execution times through analytic benchmarking for scheduling in a heterogeneous environment", IEEE Transactions on Computers 48, 12 (1999), 1374–1379.
9. B. Bahmani et al., "Task clustering for scheduling DAGs on heterogeneous cloud systems", IEEE 6th International Conference on Cloud Computing (CLOUD) (2013), 612-619.
10. A.Y.S. Lam and V.O.K. Li, "Chemical-reaction-inspired metaheuristic for optimization", IEEE Transactions on Evolutionary Computation 14, 3 (2010), 381-399.
11. S. Sindhu and K. Muneeswaran, "Optimization of cloud workflow scheduling using hybrid ga-aha algorithm", Journal of Communications Software and Systems 14, 4 (2018), 356–363.
12. S. Yassa et al., "Multi-objective approach for energy-aware workflow scheduling in cloud computing environments", The Scientific World Journal 2013 (2013), 350934.