Data Duplication Removal Technology

Submitted in the partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

IN

GAMING AND GRAPHICS

Submitted by:
SACHIN PAL 21BCG1050
NIKHIL SINGH 21BCG1017
MD. RAJ 21BCG1004

Under the Supervision of:
MAMTA SHARMA

Department of AIT-CSE

DISCOVER . LEARN . EMPOWER

Outline

- Introduction to Project
- Problem Formulation
- Objectives of the work
- Methodology used
- Conclusion
- Future Scope
- References

INTRODUCTION TO PROJECT

Data Duplication Removal Technology refers to the process of identifying and eliminating duplicate data within a dataset or a system. In today's digital age, where vast amounts of data are generated and stored across various platforms, data duplication has become a prevalent issue leading to inefficiencies, increased storage costs, and difficulties in data management.

The introduction of Data Duplication Removal Technology marks a significant advancement in data optimization and management strategies. This technology employs sophisticated algorithms and techniques to detect redundant data instances and remove them, thereby streamlining data storage, enhancing data quality, and improving overall system performance.

Problem Formulation

The problem formulation of Data Duplication Removal Technology involves defining the challenges associated with duplicate data within datasets or systems and developing strategies to effectively address these challenges. Here's a breakdown of the key aspects of the problem formulation:

- Identification of Duplicate Data
- Scalability
- Performance Optimization
- Data Integrity Preservation
- Data Integrity Preservation
- Accuracy and Reliability
- Adaptability to Data Changes
- Compliance and Regulatory Considerations
- Cost-Effectiveness
- User-Friendliness and Accessibility

Objectives of the Work

The primary objective of Data Duplication Removal Technology is to improve the efficiency, accuracy, and reliability of data management processes by identifying and eliminating redundant or duplicate data within datasets or systems. Here are the key objectives of this technology:

- Enhance Data Quality
- Optimize Storage Resources
- Improve System Performance
- Facilitate Data Integration and Migration
- Support Decision-Making and Analytics

Methodology used

Data Duplication Removal Technology utilizes a variety of methodologies, algorithms, and techniques to identify and eliminate duplicate data within datasets or systems. The specific methodologies employed may vary depending on the nature of the data, the size of the dataset, and the requirements of the organization. Here are some commonly used methodologies:

- Duplicate Detection Algorithms
- Data Deduplication Techniques
- Machine Learning and AI
- Probabilistic Matching
- Rule-based Deduplication
- Scalable Processing Architectures

Future Scope

The future scope of Data Duplication Removal Technology is promising, as advancements in data management, storage technologies, and computational capabilities continue to evolve. Here are some key areas where the technology is expected to make significant strides:

- Integration with Big Data and Cloud Computing
- Real-time Duplication Detection
- Enhanced Machine Learning and AI Algorithms
- Blockchain-based Data Integrity
- Privacy-Preserving Techniques
- Automated Data Governance and Compliance
- Cross-platform Compatibility

Conclusion

In conclusion, Data Duplication Removal Technology represents a crucial advancement in data management strategies, offering organizations the means to streamline their data storage, enhance data quality, and improve overall system efficiency. Through sophisticated algorithms, machine learning techniques, and integration with emerging technologies, duplication removal solutions are poised to address the complexities of modern data landscapes and meet the evolving needs of businesses across industries.

By effectively identifying and eliminating duplicate data within datasets or systems, organizations can unlock valuable insights, optimize storage resources, and ensure data integrity and compliance with regulatory requirements. The future scope of Data Duplication Removal Technology holds tremendous potential, with advancements expected in real-time duplication detection, integration with big data and cloud computing environments, privacy-preserving techniques, and industry-specific solutions tailored to diverse organizational needs.

References

- Di Pietro, Roberto and Alessandro Sorniotti, "Proof of ownership for deduplication systems: A secure, scalable, and efficient solution", Computer Communications, 15 May 2016
- M. Bellare,S. Keelveedhi, and T. Ristenpart, "Dupless: Server aided encryption for deduplicated storage", USENIX Security Symposium, 2013
- Harnik, Danny, Alexandra Shulman-Peleg and Benny Pinkas, "Side channels in cloud services, the case of deduplication in cloud storage ", IEEE Security & Privacy 8, 2014.
- Atish Kathpal, Matthew John and Gaurav Makkar, "Distributed Duplicate Detection in Post-Process Data Deduplication", Conference: HiPC , 2011

- X. Zhao, Y. Zhang, Y. Wu, K. Chen, J. Jiang, K. Li, "Liquid: A Scalable Deduplication File System for Virtual Machine Images",IEEE Transactions on Parallel and Distributed Systems, January 2013.

- Stephen J. Bigelow, "Data Deduplication Explained: http://searchgate.org", February, 2018