ELYSIUMPRO

A UNIT OF ELYSIUM GROUPS

FINAL YEAR PROJECT

BIGDATA 2019-2020

TITLES WITH ABSTRACTS

CALL US @

(+91) 9944 7933 98 | (+91) 452 - 424 2842, 424 2843

20 Years of Experience | Automated Services | 24/7 Help Desk Support
Advanced Technologies and Tools | Legitimate Members of all Journals
Quality Project Training | Industry Exposure

ELYSIUMPRO

INSPIRING THE LEADING EDGE TECHNOLOGIES

www.elysiumpro.in

#227, Elysium Campus, Church Road, Anna Nagar,
Madurai - 625020, Tamil Nadu, India

CALL US @

(+91) 9944 7933 98 | (+91) 452 - 424 2842, 424 2843

f /ElysiumPro Project Center / ElysiumPro in / ElysiumPro

ELYSIUM GROUPS

EPRO BD - 001	Hierarchical Density-Based Clustering using MapReduce

Hierarchical density-based clustering is a powerful tool for exploratory data analysis. However, its applicability to large datasets is limited because the computational complexity. In the literature, there have been attempts to parallelize algorithms such as Single-Linkage, which in principle can also be extended to the broader scope of hierarchical density-based clustering, but hierarchical clustering algorithms are inherently difficult to parallelize with MapReduce. In this paper, we discuss why adapting previous approaches to parallelize Single-Linkage clustering using MapReduce leads to very inefficient solutions when one wants to compute density-based clustering hierarchies. Preliminarily, we discuss one such solution, which is based on an exact, yet very computationally demanding, random blocks parallelization scheme. To be able to efficiently apply hierarchical density-based clustering to large datasets using MapReduce, we then propose a different parallelization scheme that computes an approximate clustering hierarchy based on a much faster, recursive sampling approach. This approach is based on HDBSCAN*, the state-of-the-art hierarchical density-based clustering algorithm, combined with a data summarization technique called data bubbles. The proposed method is evaluated in terms of both runtime and quality of the approximation on a number of datasets, showing its effectiveness and scalability.

EPRO BD - 002	A Data Sharing Protocol to Minimize Security and Privacy Risks of Cloud Storage in Big Data Era

A cloud-based big data sharing system utilizes a storage facility from a cloud service provider to share data with legitimate users. In contrast to traditional solutions, cloud provider stores the shared data in the large data centers outside the trust domain of the data owner, which may trigger the problem of data confidentiality. This paper proposes a secret sharing group key management protocol (SSGK) to protect the communication process and shared data from unauthorized access. Different from the prior works, a group key is used to encrypt the shared data and a secret sharing scheme is used to distribute the group key in SSGK. The extensive security and performance analyses indicate that our protocol highly minimizes the security and privacy risks of sharing data in cloud storage and saves about 12% of storage space.

EPRO BD - 003	Haery: a Hadoop based Query System on Accumulative and High-dimensional Data Model for Big Data

Column-oriented stores, known as their scalability and flexibility, are a common NoSQL database implementation and are increasingly used in big data management. In column-oriented stores, a "full-scan" query strategy is inefficient and the search space can be reduced if data is well partitioned or indexed, however there is no pre-defined schema for building and maintaining partitions and indexesat lower cost. We leverage an accumulative and high-dimensional data model, a sophisticated linearization algorithm, and an efficient query algorithm, to solve the challenge of how a pre-defined and well-partitioned data model can be applied to flexible and time-varied key-value data. We adapt a high-dimensional array as the data model to partition the key-value data without additional storage and massive calculation; improve the Z-order linearization algorithm, which map multidimensional data to one dimension while preserving locality of the data points, for flexibility; efficiently build an expansion mechanism for the data model to support time-varied data. The result is Haery, a column-oriented store, based on a distributed file system and computing framework. In experiments, Haery is compared with Hive, HBase, Cassandra, MongoDB, PostgresXL and HyperDex in terms of query performance.

EPRO BD - 004	Automated Data Slicing for Model Validation: A Big data - AI Integration Approach

As machine learning systems become democratized, it becomes increasingly important to help users easily debug their models. However, current data tools are still primitive when it comes to helping users trace model performance problems all the way to the data. We focus on the particular problem of slicing data to identify subsets of the validation data where the model performs poorly. This is an important problem in model validation because the overall model performance can fail to reflect that of the smaller subsets, and slicing allows users to analyze the model performance on a more granular-level. Unlike general techniques (e.g., clustering) that can find arbitrary slices, our goal is to find interpretable slices (which are easier to take action compared to arbitrary subsets) that are problematic and large. We propose Slice Finder, which is an interactive framework for identifying such slices using statistical techniques. Applications include diagnosing model fairness and fraud detection, where identifying slices that are interpretable to humans is crucial. This research is part of a larger trend of big data and Artificial Intelligence (AI) integration and opens many opportunities for new research.

EPRO BD - 005	Efficient Data Placement and Replication for QoS-Aware Approximate Query Evaluation of Big Data Analytics

Enterprise users at different geographic locations generate large-volume data that is stored at different geographic datacenters. These users may also perform big data analytics on the stored data to identify valuable information in order to make strategic decisions. However, it is well known that performing big data analytics on data in geographical-located datacenters usually is time-consuming and costly. In some delay-sensitive applications, the query result may become useless if answering a query takes too long time. Instead, sometimes users may only be interested in timely approximate rather than exact query results. When such approximate query evaluation is the case, applications must sacrifice timeliness to get more accurate evaluation results or tolerate evaluation result with a guaranteed error bound obtained from analyzing the samples of the data to meet their stringent timeline. In this paper, we study quality-of-service (QoS)-aware data replication and placement for approximate query evaluation of big data analytics in a distributed cloud, where the original (source) data of a query is distributed at different geo-distributed datacenters. We focus on the problems of placing data samples of the source data at some strategic datacenters to meet stringent query delay requirements of users, by exploring a non-trivial trade-off between the cost of query evaluation and the error bound of the evaluation result. We first propose an approximation algorithm with a provable approximation ratio for a single approximate query. We then develop an efficient heuristic algorithm for evaluating a set of approximate queries with the aim to minimize the evaluation cost while meeting the delay requirements of these queries.

EPRO BD - 006	Fast Communication-efficient Spectral Clustering over Distributed Data

The last decades have seen a surge of interests in distributed computing thanks to advances in clustered computing and big data technology. Existing distributed algorithms typically assume all the data are already in one place, and divide the data and conquer on multiple machines. However, it is increasingly often that the data are located at a number of distributed sites, and one wishes to compute over all the data with low communication overhead. For spectral clustering, we propose a novel framework that enables its computation over such distributed data, with "minimal" communications while a major speedup in computation. The loss in accuracy is negligible compared to the non-distributed setting. Our approach allows local parallel computing at where the data are located, thus turns the distributed nature of the data into a blessing; the speedup is most substantial when the data are evenly distributed across sites. Experiments on synthetic and large UC Irvine datasets show almost no loss in accuracy with our approach while a 2x speedup under all settings we have explored. As the transmitted data need not be in their original form, our framework readily addresses the privacy concern for data sharing in distributed computing.

EPRO BD - 007	Skia: Scalable and Efficient In-Memory Analytics for Big Spatial-Textual Data

In recent years, spatial-keyword queries have attracted much attention with the fast development of location-based services. However, current spatial-keyword techniques are disk-based, which cannot fulfill the requirements of high throughput and low response time. With the surging data size, people tend to process data in distributed in-memory environments to achieve low latency. In this paper, we present the distributed solution, i.e., Skia (Spatial-Keyword In-memory Analytics), to provide a scalable backend for spatial-textual analytics. Skia introduces a two-level index framework for big spatial-textual data including: (1) efficient and scalable global index, which prunes the candidate partitions a lot while achieving small space budget; and (2) four novel local indexes, that further support low latency services for exact and approximate spatial-keyword queries. Skia can support common spatial-keyword queries via traditional SQL programming interfaces. The experiments conducted on large-scale real datasets have demonstrated the promising performance of the proposed indexes and our distributed solution.

EPRO BD - 008	K-nearest Neighbors Search by Random Projection Forests

K-nearest neighbors (kNN) search is an important problem in data mining and knowledge discovery. Inspired by the huge success of tree-based methodology and ensemble methods over the last decades, we propose a new method for kNN search, random projection forests (rpForests). rpForests finds nearest neighbors by combining multiple kNN-sensitive trees with each constructed recursively through a series of carefully chosen random projections. As demonstrated by experiments on a wide collection of real datasets, our method achieves a remarkable accuracy in terms of fast decaying missing rate of kNNs and that of discrepancy in the k-th nearest neighbor distances. rpForests has a very low computational complexity as a tree-based methodology. The ensemble nature of rpForests makes it easily parallelized to run on clustered or multicore computers; the running time is expected to be nearly inversely proportional to the number of cores or machines. We give theoretical insights on rpForests by showing the exponential decay of neighboring points being separated by ensemble random projection trees when the ensemble size increases. Our theory can also be used to refine the choice of random projections in the growth of rpForests; experiments show that the effect is remarkable.

EPRO BD - 009	Transfer to Rank for Top-N Recommendation

In this paper, we study top-N recommendation by exploiting users' explicit feedback such as 5-star numerical ratings, which has been overlooked to some extent in the past decade. As a response, we design a novel and generic transfer learning based recommendation framework coarse-to-fine transfer to rank (CoFiToR), which is a significant extension of a very recent work called transfer to rank (ToR). The key idea of our solution is modeling users' behaviors by simulating users' shopping processes. Therefore, we convert the studied ranking problem to three subtasks corresponding to three specific questions, including (i) whether an item will be examined by a user, (ii) how an item will be scored by a user, and (iii) whether an item will finally be purchased by a user. Based on this new conversion, we then develop a three-staged solution that progressively models users' preferences from a coarse granularity to a fine granularity. In each stage, we adopt an appropriate recommendation algorithm with pointwise or pairwise preference assumption to answer each question in order to seek an effective and efficient overall solution. Empirical studies on two large and public datasets showcase the merits of our solution in comparison with the state-of-the-art methods.

EPRO BD - 010	New Scheduling Algorithms for Improving Performance and Resource Utilization in Hadoop YARN Clusters

The MapReduce framework has become the defacto scheme for scalable semi-structured and un-structured data processing in recent years. The Hadoop ecosystem has evolved into its second generation, Hadoop YARN, which adopts fine-grained resource management schemes for job scheduling. Nowadays, fairness and efficiency are two main concerns in YARN resource management because resources in YARN are shared and contended by multiple applications. However, the current scheduling in YARN does not yield the optimal resource arrangement, unnecessarily causing idle resources and inefficient scheduling. It omits the dependency between tasks which is extremely crucial for the efficiency of resource utilization as well as heterogeneous job features in real application environments. We thus propose a new YARN scheduler which can effectively reduce the makespan (i.e., the total execution time) of a batch of MapReduce jobs in Hadoop YARN clusters by leveraging the information of requested resources, resource capacities and dependency between tasks. For accommodating heterogeneity in MapReduce jobs, we also extend our scheduler by further considering the job iteration information in the scheduling decisions. We implemented the new scheduling algorithm as a pluggable scheduler in YARN and evaluated it with a set of classic MapReduce benchmarks. The experimental results demonstrate that our YARN scheduler effectively reduces the makespans and improves resource utilizations.

EPRO BD - 011	P-MOD: Secure Privilege-Based Multilevel Organizational Data-Sharing in Cloud Computing

Cloud computing has changed the way enterprises store, access, and share data. Big data sets are constantly being uploaded to the cloud and shared within a hierarchy of many different individuals with different access privileges. With more data storage needs turning over to the cloud, finding a secure and efficient data access structure has become a major research issue. In this paper, a Privilege-based Multilevel Organizational Data-sharing scheme (P-MOD) is proposed that incorporates a privilege-based access structure into an attribute-based encryption mechanism to handle the management and sharing of big data sets. Our proposed privilege-based access structure helps reduce the complexity of defining hierarchies as the number of users grows, which makes managing healthcare records using mobile healthcare devices feasible. It can also facilitate organizations in applying big data analytics to understand populations in a holistic way. Security analysis shows that P-MOD is secure against adaptively chosen plaintext attack assuming the DBDH assumption holds. The comprehensive performance and simulation analyses using the real U.S. Census Income dataset demonstrate that P-MOD is more efficient in computational complexity and storage space than the existing schemes.

EPRO BD – 012	Deadline-Aware Map Reduce Job Scheduling with Dynamic Resource Availability

As MapReduce is becoming ubiquitous in large-scale data analysis, many recent studies have shown that the performance of MapReduce could be improved by different job scheduling approaches, e.g., Fair Scheduler and Capacity Scheduler. However, most exiting MapReduce job schedulers focus on the scenario that MapReduce cluster is stable and pay little attention to the MapReduce cluster with dynamic resource availability. In fact, MapReduce cluster resources may fluctuate as there is a growing number of Hadoop clusters deployed on hybrid systems, e.g., infrastructure powered by mix of traditional and renewable energy, and cloud platforms hosting heterogeneous workloads. Thus, there is a growing need for providing predictable services to users who have strict requirements on job completion times in such dynamic environments. In this paper, we propose, RDS, a Resource and Deadline-aware Hadoop job Scheduler that takes future resource availability into consideration when minimizing job deadline misses. We formulate the job scheduling problem as an online optimization problem and solve it using an efficient receding horizon control algorithm. To aid the control, we design a self-learning model to estimate job completion times. We further extend the design of RDS scheduler to support flexible performance goals in various dynamic clusters. In particular, we use flexible deadline time bounds instead of the single fixed job completion deadline. We have implemented RDS in the open-source Hadoop implementation and performed evaluations with various benchmark workloads.

EPRO BD - 013	Enabling Ternary Hash Tree based Integrity Verification for Secure Cloud Data Storage

Cloud Computing enables the remote users to access data, services and applications in on-demand from the shared pool of configurable computing resources. On the other hand, it is not easy for the cloud users to identify whether Cloud Service Provider's (CSP) tag along with the data security legal expectations. So, cloud users could not rely on CSP's in terms of trust. So, it is significant to build a secure and efficient data auditing framework for increasing and maintaining cloud users trust with CSP. In this work, we proposed a novel public auditing framework for securing cloud storage based on Ternary Hash Tree (THT) and Replica based Ternary Hash Tree (R-THT), which will be used by TPA to perform data auditing. Differing from existing work, the proposed framework performs Block-level, File-level and Replica-level auditing with tree block ordering, storage block ordering for verifying the data integrity and ensuring data availability in the cloud. In addition, the framework also supports error localization with data correctness, dynamic updates with block update, insert and delete operations in cloud. The results shows that the proposed secure cloud auditing framework is highly secure and efficient in storage, communication and computation costs.

EPRO BD - 014	Cost-Effective Cloud Server Provisioning for Predictable Performance of Big Data Analytics

Cloud datacenters are underutilized due to server over-provisioning. To increase datacenter utilization, cloud providers offer users an option to run workloads such as big data analytics on the underutilized resources, in the form of cheap yet revocable transient servers (e.g., EC2 spot instances, GCE preemptible instances). Though at highly reduced prices, deploying big data analytics on the unstable cloud transient servers can severely degrade the job performance due to instance revocations. To tackle this issue, this paper proposes iSpot, a cost-effective transient server provisioning framework for achieving predictable performance in the cloud, by focusing on Spark as a representative Directed Acyclic Graph (DAG)-style big data analytics workload. It first identifies the stable cloud transient servers during the job execution by devising an accurate Long Short-Term Memory (LSTM)-based price prediction method. Leveraging automatic job profiling and the acquired DAG information of stages, we further build an analytical performance model and present a lightweight critical data checkpointing mechanism for Spark, to enable our design of iSpot provisioning strategy for guaranteeing the job performance on stable transient servers. Extensive prototype experiments on both EC2 spot instances and GCE preemptible instances demonstrate that, iSpot is able to guarantee the performance of big data analytics running on cloud transient servers while reducing the job budget by up to 83.8 percent in comparison to the state-of-the-art server provisioning strategies, yet with acceptable runtime overhead.

EPRO BD - 015	CHARON: A Secure Cloud-of-Clouds System for Storing and Sharing Big Data

We present CHARON, a cloud-backed storage system capable of storing and sharing big data in a secure, reliable, and efficient way using multiple cloud providers and storage repositories to comply with the legal requirements of sensitive personal data. CHARON implements three distinguishing features: (1) it does not require trust on any single entity, (2) it does not require any client-managed server, and (3) it efficiently deals with large files over a set of geo-dispersed storage services. Besides that, we developed a novel Byzantine-resilient data-centric leasing protocol to avoid write-write conflicts between clients accessing shared repositories. We evaluate CHARON using micro and application-based benchmarks simulating representative workflows from bioinformatics, a prominent big data domain. The results show that our unique design is not only feasible but also presents an end-to-end performance of up to 2.5x better than other cloud-backed solutions.

EPRO BD - 016	Low Latency Big Data Processing without Prior Information

Job scheduling plays an important role in improving the overall system performance in big data processing frameworks. Simple job scheduling policies do not consider job sizes and may degrade the performance when jobs of varying sizes arrive. More elaborate job scheduling policies assume that complete information about job sizes is available from the prior runs. In this paper, we design and implement an efficient and practical job scheduler to achieve better performance even without prior information about job sizes. The superior performance of our job scheduler originates from the design of multiple level priority queues, where jobs are demoted to lower priority queues if the amount of service consumed so far reaches a certain threshold. In this case, our new job scheduler can effectively mimic the shortest job first scheduling policy without knowing the job sizes in advance. To demonstrate its performance, we have implemented our job scheduler in YARN and validated its performance with experiments in real testbeds and large-scale trace-driven simulations. Our experimental and simulation results show that our new job scheduler can reduce the average job response time of the Fair scheduler by up to 45% and achieve better fairness at the same time.

EPRO BD - 017	PISCES: Optimizing Multi-Job Application Execution in MapReduce

Nowadays, many MapReduce applications consist of groups of jobs with dependencies among each other, such as iterative machine learning applications and large database queries. Unfortunately, the MapReduce framework is not optimized for these multi-job applications. It does not explore the execution overlapping opportunities among jobs and can only schedule jobs independently. These issues significantly inflate the application execution time. This paper presents PISCES (Pipeline Improvement Support with Critical chain Estimation Scheduling), a critical chain optimization (a critical chain refers to a series of jobs which will make the application run longer if any one of them is delayed), to provide better support for multi-job applications. PISCES extends the existing MapReduce framework to allow scheduling for multiple jobs with dependencies by dynamically building up a job dependency DAG for current running jobs according to their input and output directories. Then using the dependency DAG, it provides an innovative mechanism to facilitate the data pipelining between the output phase (map phase in the Map-Only job or reduce phase in the Map-Reduce job) of an upstream job and the map phase of a downstream job. This offers a new execution overlapping between dependent jobs in MapReduce which effectively reduces the application runtime. Moreover, PISCES proposes a novel critical chain job scheduling model based on the accurate critical chain estimation.

EPRO BD - 018	T-PCCE: Twitter Personality based Communicative Communities Extraction System for Big Data

The identification of social media communities has recently been of major concern, since users participating in such communities can contribute to viral marketing campaigns. In this work we focus on users' communication considering personality as a key characteristic for identifying communicative networks i.e. networks with high information flows. We describe the Twitter Personality based Communicative Communities Extraction (T-PCCE) system that identifies the most communicative communities in a Twitter network graph considering users' personality. We then expand existing approaches in users' personality extraction by aggregating data that represent several aspects of user behaviour using machine learning techniques. We use an existing modularity based community detection algorithm and we extend it by inserting a post-processing step that eliminates graph edges based on users' personality. The effectiveness of our approach is demonstrated by sampling the Twitter graph and comparing the communication strength of the extracted communities with and without considering the personality factor. We define several metrics to count the strength of communication within each community. Our algorithmic framework and the subsequent implementation employ the cloud infrastructure and use the MapReduce Programming Environment. Our results show that the T-PCCE system creates the most communicative communities.

EPRO BD - 019	Cyber security in Big Data Era: From Securing Big Data to Data-Driven Security

"Knowledge is power" is an old adage that has been found to be true in today's information age. Knowledge is derived from having access to information. The ability to gather information from large volumes of data has become an issue of relative importance. Big Data Analytics (BDA) is the term coined by researchers to describe the art of processing, storing and gathering large amounts of data for future examination. Data is being produced at an alarming rate. The rapid growth of the Internet, Internet of Things (IoT) and other technological advances are the main culprits behind this sustained growth. The data generated is a reflection of the environment it is produced out of, thus we can use the data we get out of systems to figure out the inner workings of that system. This has become an important feature in cybersecurity where the goal is to protect assets. Furthermore, the growing value of data has made big data a high value target. In this paper, we explore recent research works in cybersecurity in relation to big data. We highlight how big data is protected and how big data can also be used as a tool for cybersecurity. We summarize recent works in the form of tables and have presented trends, open research challenges and problems. With this paper, readers can have a more thorough understanding of cybersecurity in the big data era, as well as research trends and open challenges in this active research area.

EPRO BD – 020	Practical Privacy-Preserving MapReduce Based K-Means Clustering Over Large-Scale Dataset

Clustering techniques have been widely adopted in many real world data analysis applications, such as customer behavior analysis, targeted marketing, digital forensics, etc. With the explosion of data in today's big data era, a major trend to handle a clustering over large-scale datasets is outsourcing it to public cloud platforms. This is because cloud computing offers not only reliable services with performance guarantees, but also savings on in-house IT infrastructures. However, as datasets used for clustering may contain sensitive information, e.g., patient health information, commercial data, and behavioral data, etc, directly outsourcing them to public cloud servers inevitably raise privacy concerns. In this paper, we propose a practical privacy-preserving K-means clustering scheme that can be efficiently outsourced to cloud servers. Our scheme allows cloud servers to perform clustering directly over encrypted datasets, while achieving comparable computational complexity and accuracy compared with clusterings over unencrypted ones. We also investigate secure integration of MapReduce into our scheme, which makes our scheme extremely suitable for cloud computing environment. Thorough security analysis and numerical analysis carry out the performance of our scheme in terms of security and efficiency. Experimental evaluation over a 5 million objects dataset further validates the practical performance of our scheme.

EPRO BD - 021	RoD: Evaluating the Risk of Data Disclosure Using Noise Estimation for Differential Privacy

Differential privacy (DP) is a notion of big data privacy protection that offers protection even when an attacker has arbitrary background knowledge in advance. DP introduces noise, such as Laplace noise, to obfuscate the true value in a dataset while preserving its statistic properties. However, a large amount of Laplace noise added into a dataset is typically defined by the discursive scale parameter of Laplace distribution. The privacy budget ϵ in DP has been theoretically interpreted, but the implication on the risk of data disclosure (RoD) in practice has not yet been well studied. In this paper, we define and evaluate RoD in a dataset with either numerical or binary attributes for numerical or counting queries with multiple attributes based on noise estimation. Through confidence probability of noise estimation, we provide a method to select the ϵ for DP and associate differential privacy with k-anonymization. Finally, we show the relationship between the RoD and ϵ as well as between ϵ and k in our experimental results. To the best of our knowledge, this is the first study using the quantity of noise as a bridge to evaluate RoD for multiple attributes and determine the relationship between DP and k-anonymization.

EPRO BD - 022	A Review of Judgment Analysis Algorithms for Crowdsourced Opinions

The crowd-powered systems have been shown to be highly successful in the current decade to manage collective contribution of online workers for solving different complex tasks. It can also be used for soliciting opinions from a large set of people working in a distributed manner. Unfortunately, the online community of crowd workers might involve non-experts as opinion providers. As a result, such approaches may give rise to noise making it hard to predict the appropriate (gold) judgment. Judgment analysis is in general a way of learning about human decision from multiple opinions. A spectrum of algorithms has been proposed in the last few decades to address this problem. They are broadly of supervised or unsupervised types. However, they have been readdressed in recent years having focus on different strategies for obtaining the gold judgment from crowdsourced opinions, viz., estimating the accuracy of opinions, difficulties of the problem, spammer identification, handling noise, etc. Besides this, investigation of various types of crowdsourced opinions to solve complex real-life problems provide new insights in this domain. In this survey, we provide a comprehensive overview of the judgment analysis problem and some of its novel variants, addressed with different approaches, where the opinions are crowdsourced.

EPRO BD - 023	Towards Thwarting Template Side-channel Attacks in Secure Cloud Deduplications

As one of a few critical technologies to cloud storage service, deduplication allows cloud servers to save storage space by deleting redundant file copies. However, it often leaks side channel information regarding whether an uploading file gets deduplicated or not. Exploiting this information, adversaries can easily launch a template side-channel attack and severely harm cloud users' privacy. To thwart this kind of attack, we resort to the k-anonymity privacy concept to design secure threshold deduplication protocols. Specifically, we have devised a novel cryptographic primitive called "dispersed convergent encryption" (DCE) scheme, and proposed two different constructions of it. With these DCE schemes, we successfully construct secure threshold deduplication protocols that do not rely on any trusted third party. Our protocols not only support confidentiality protections and ownership verifications, but also enjoy formal security guarantee against template side-channel attacks even when the cloud server could be a "covert adversary" who may violate the predefined threshold and perform deduplication covertly. Experimental evaluations show our protocols enjoy very good performance in practice.

EPRO BD - 024	Secure Encrypted Data with Authorized Deduplication in Cloud

In this paper, we propose a novel secure role re-encryption system (SRRS), which is based on convergent encryption and the role re-encryption algorithm to prevent the privacy data leakage in cloud and it also achieves the authorized deduplication and satisfies the dynamic privilege updating and revoking. Meanwhile, our system supports ownership checking and achieves the proof of ownership for the authorized users efficiently. Specifically, we introduce a management center to handle with the authorized request and establish a role authorized tree (RAT) mapping the relationship of the roles and keys. With the convergent encryption algorithm and the role re-encryption technique, it can be guaranteed that only the authorized user who has the corresponding role re-encryption key can access the specific file without any data leakage. Through role re-encryption key updating and revoking, our system achieves the dynamic updating of the authorized user’s privilege. Furthermore, we exploit the dynamic count filters (DCF) to implement the data updating and improve the retrieval of ownership verifying effectively. We conduct the security analysis and the simulation experiment to demonstrate the security and efficiency of our proposed system.

EPRO BD - 025	Updatable Block-Level Message-Locked Encryption

Deduplication is widely used for reducing the storage requirement for storage service providers. Nevertheless, it is unclear how to support deduplication of encrypted data securely until the study of Bellare et. al. on message-locked encryption (MLE, Eurocrypt 2013). While updating (shared) files is natural, existing MLE solutions do not allow efficient update of encrypted files stored remotely. Even modifying a single bit requires the expensive way of downloading and decrypting a large ciphertext (then re-uploading). This paper initiates the study of updatable block-level MLE, a new primitive in incremental cryptography and cloud cryptography. Our proposed provably-secure construction is updatable with computation cost logarithmic in the file size. It naturally supports block-level deduplication. It also supports proof-of-ownership which protects storage providers from being abused as a free content distribution network. Our experiments show its practical performance relative to the original MLE and existing non-updatable block-level MLE.

ELYSIUMPRO

A UNIT OF ELYSIUM GROUPS

THANK YOU!