Programme de l'atelier "Sécuriser les données"

Lionel Le Folgoc : Disponibilité et durabilité des données pour les Serveurs Personnels de Données

Une quantité croissante de données personnelles est rassemblée sur les serveurs des administrations, hôpitaux, compagnies d’assurance, etc. Bien que cette centralisation assure la disponibilité, la durabilité et la facilité d'interrogation des données, ces avantages doivent être pondérés avec les risques encourus sur la vie privée. Les Serveurs Personnels de Données, combinant la sécurité d'une carte à puce avec la capacité de stockage d'une clé USB, constituent une alternative intéressante pour restaurer le contrôle de l'utilisateur sur ses données. Malgré leur faible puissance et leur connectivité réduite, ils doivent être en mesure d'interagir avec les serveurs externes. Il reste un certain nombre d'aspects à adresser dans cette optique, en particulier le rétablissement de services traditionnels des bases de données : augmenter la disponibilité des données en respectant les contrats Hippocratiques (contrôle d'accès, audit, etc.), et assurer leur durabilité sans divulgation des données sensibles.

Mohamed JAWAD : Data Privacy in Structured P2P Systems

P2P systems are increasingly used in distributed networks for efficient, scalable data sharing. Popular applications focus on massive file sharing, however, advanced applications will appear where professional communities (e.g., medical or research communities) will share private or sensitive data. Currently, in P2P systems, peers can easily violate data privacy by using data for malicious purposes (e.g., fraudulence, profiling). To prevent such behavior, a well accepted principle has been proposed which states that data owners should specify the purpose for which their data will be collected. Due to the decentralized control and the potentially large number of participants, in general, peers do not trust each other and it is hard to enforce data privacy. In our work, we apply the Hippocratic database principles and reputation techniques to support purpose and trust notions in structured P2P systems. Hippocratic databases enforce purpose-based privacy while reputation techniques guarantee trust. We propose a P2P data privacy model which defines the basic concepts of data privacy in P2P systems. We also present the algorithms of PriServ, a DHT-based P2P privacy service which supports this model and prevents data privacy violation. A prototype of PriServ is under construction and will be tested on the largely distributed platform Grid5000.

Shaoyi Yin : Database principles for Flash-based devices

NAND Flash memory stands out as the best adapted storage medium for a wide spectrum of mobile and embedded devices (PDAs, cell phones, sensors, smart cards, USB keys, etc). The reasons for this success include shock resistance, size, energy consumption, performance and the ease of on-chip integration. However, NAND Flash memory exhibits specific hardware characteristics. Reads and writes can be done at a page granularity, but writes are more time and energy consuming than reads. In addition, a page cannot be rewritten before erasing the complete block containing it, a costly operation. Finally, a block wears out after 10^6 to 10^7 repeated write/erase cycles. Due to this, updates are usually performed “out-of-place”, meaning that a page is copied in another location before being modified, entailing an address translation and garbage collection mechanisms. All these constraints make database management very challenging.

Database storage and indexation models dedicated to NAND Flash have recently been proposed in the literature. The commonalities of all state of the art methods are: (1) to delay the data and index updates thanks to a log with the objective to group updates related to a same page; (2) to build a RAM index to speed up the lookup of a key in the log and (3) to commit the log updates with a given frequency in order to bound the log size. Changing the commit frequency entails complex trade-offs between the number of reads and writes, the Flash occupancy and the RAM consumption and none of the proposed solutions exhibits a good behavior on all metrics.

The objective of my work is to investigate a different approach where the complete database is organized in pure sequential append-only data structures. The benefit is avoiding “out-of-place” updates provided a simple buffering strategy is followed. The question is now how to look up database records with acceptable performance? A first idea is building sequential summaries of database attributes and doing the lookups in these summaries. This introduces a new trade-off between the compression ratio of a summary and its accuracy (nb: Bloom filters is an example of summarization technique exhibiting such trade-off). A second idea consists in partitioning a sequential structure vertically in order to decrease its scanning cost. This introduces a new trade-off between the number of partitions, the RAM usage and the Flash occupancy, with a direct impact on the read and write costs. The problem is making these two simple ideas effective in practice and studying their impact on the data storage, the data indexation, the buffering strategy and the transaction protocols.

Yanli GUO : Confidentiality and tamper-resistance of embedded databases

Ubiquitous and pervasive computing introduces the need for embedding and managing data in ever lighter and specialized computing devices. In this context, SMIS is designing a full-fledged database engine embedded in a new form of hardware secure device called SPT(secure Portable Token).Embedding database systems in such devices is very challenging given the high volume of data to process and the particular characteristics of NAND Flash. In addition, data confidentiality and data integrity must be guaranteed. These two properties have to be enforced in an unusual context where the SPT microcontroller provides tamper-resistance guarantees for the processing and the storage of a small amount of data while the SPT mass storage area (i.e., the NAND Flash) is external to the microcontroller and can then be tampered. Hence, there is a strong need for preventing information disclosures and protecting the data integrity in FLASH thanks to cryptographic techniques.

Adeel Anjum : Providing K-anonymity using spatial index methods

Many research organizations e.g. medical organizations use to publish their data for either research activities or research funding. Releasing this kind of data about individuals without risking their privacy has been an important problem. To obviate personal identification, many organizations use to remove the uniquely identifying information like name, SSN etc from the published data. However, this purification of data might not be helpful in keeping the secrecy of given individuals. This gave rise to the need of a mechanism to publish sensitive individual data keeping their privacy intact. Thus, the term K-anonymization came into existence.

K-anonymization, though exists abundantly in today's literature, is a generalization model proposed in early work on data anonymization. A table satisfies k-anonymity if every record in the table is indistinguishable from at least k-1 other records in its public release. This simple principle determines an equivalence relation on the data and is sufficient to prevent the disclosure of identity with a probability of 1 / k.

Though, k-anonymity is the most basic approach in dealing with the anonymization of data, other more sophisticated approaches have emerged in recent years to address the problem of revealing identities of individuals, not taken into account by the k-anonymity. Among these approaches are: 1) l-Diversity that aims to ensure that the sensitive values are well represented in each equivalence class and 2) m-Unicity which forces each sensitive value to be unique in a single equivalence class.

Since the emergence of k-anonymization, several algorithms have been proposed in order to provide the basis to cater the problem of revealing individual identities. It has also been shown that spatial indexing techniques e.g. R-Tree can provide basis to efficient and high quality data anonymization. Besides R-tree spatial indexing technique, BANG (Balanced-And-Nested-Grid) is special kind of Grid File spatial index structure which differs from R-tree in that it partitions the data space into block regions by successive binary divisions while R-tree splits space with hierarchically nested, and possibly overlapping, minimum bounding rectangles (MBRs, otherwise known as bounding boxes, i.e. rectangle, what the R in R-tree stands for).

The first objective of our research is to make use of a spatial index structure i.e. BANG file to anonymize the data and compare this anonymization technique with the one that uses R-tree spatial index as its basis. We further plan to ameliorate our research using a chosen spatial indexing technique to contend with the problems during INSERTS and UPDATES of an anonymizing dataset.

Tristan ALLARD : Safe Anonymization of Data Hosted in Smart Tokens

Un nombre croissant d’enquêtes et d’articles dans l’actualité mettent en évidence l’échec des serveurs de bases de données à sauvegarder réellement la privacité des données confidentielles. Sans même considérer les attaques, internes ou externes, de simples négligences mènent souvent à des dévoilement massifs de données. Un nouveau type de dispositifs, appelés Secure Portable Token (SPT), combinant la sécurité d’une carte à puce avec les capacités de stockage FLASH, autorise et rend crédibles des alternatives à la centralisation systématique des données personnelles. Chacun peut stocker ses données personnelles (e.g., son dossier médical) dans son propre SPT sous son contrôle, et ne jamais les dévoiler en clair au monde extérieur. Cependant, cette nouvelle gestion des données personnelles entre en conflit avec les outils d'aide à la décision qui supposent habituellement une certaine centralisation des données. Cet article adresse précisément ce problème en proposant d'adapter le modèle traditionnel de publication de données confidentielles (Privacy-Preserving Data Publishing - PPDP) à un environnement constitué d'un grand nombre de SPTs sécurisés, se connectant rarement à une infrastructure disponible mais non digne de confiance. Cette combinaison unique d'hypothèses rend le problème fondamentalement différent de tout problème PPDP déjà traité.

Aurélien Faravelon : Préservation de la vie privée et fouille de données médicales

La fouille de données, ou *data mining*, permet aujourd'hui d'établir des patrons de données, c'est à dire de repérer des régularités dans les données pouvant être utilisées dans un but prédictif. En médecine, l'établissement de tels patrons permet, par exemple, l'aide au diagnostic ou la prédiction du coût du séjour d'un patient dans un hôpital.

Devant la valeur prédictive et la multiplicité des acteurs ayant intérêt à extraire les motifs de données médicales, nous proposons une double approche, philosophique et informatique visant à établir les enjeux épistémologiques et éthiques de la fouille de données médicales et à concevoir un système permettant de préserver la vie privée des patients auxquels les motifs de données peuvent être appliqués.

Au cours de cette présentation, j'esquisserai les composants de ce système en montrant comment, à partir de la formalisation des motifs de données médicales et des profils utilisateurs, on peut à évaluer le niveau de confiance d'un utilisateur et à modifier les motifs de données si nécessaire. Mon approche consiste en effet à dire qu'un motif de données n'est utile que s'il décrit une situation suffisamment précise et arrivant dans un temps raisonnable et qu'en jouant sur l'utilité des motifs on peut parvenir à protéger la vie privée des individus.