Enjeux de la protection des données

Pr. Benjamin NGUYEN

INSA Centre Val de Loire

Laboratoire d'Informatique Fondamentale d'Orléans

Co-directeur du GT *Privacy*

https://openclassrooms.com/fr/courses/5280946-protegez-les-donnees-personnelles/

Plan

- **Partie I : La vie privée en Pratique**
  1. Le concept de vie privée
  2. L'exploitation de données personnelles
  3. Contrôler l'accès à vos données
  4. Protection du transfert des données sur les réseaux
- **Partie II : Zoom sur le Règlement Général sur la Protection de Données Personnelles (RGPD)**
  1. Grandes lignes du règlement
  2. Zoom sur le droit à l'oubli
  3. Analyse d'Impact de Vie Privée (AIVP)
- **Partie III : Sécurité Informatique**
  1. Cartographie des attaques
  2. Mécanismes de protection
  3. Techniques avancées de sécurisation du traitement de données
- ***Partie IV : Anonymisation de données***

Partie I : La vie privée en pratique

Le concept de vie privée

Vers une production massive de données personnelles

Les objets connectés (dont les smartphones) produisent des **données sensibles (ou personnelles)**, comme le rythme cardiaque, le poids de l'individu, ou encore sa localisation GPS.

D'après la « loi européenne » nommée ***Règlement général sur la protection des données*** **(ou RGPD) entrée en vigueur en mai 2018**, il faut traiter avec soin ces données, afin de garantir la **protection de la vie privée des individus** (appelée ***privacy*** en anglais).

Qu’est ce que la protection de la vie privée ?

Ce concept est ancien, et correspond à la **possibilité pour chaque individu de ne pas être obligé de communiquer des informations qu’il considère comme étant personnelles**, et n’ayant aucune raison d’atteindre la sphère publique.

Si ce concept est relativement simple et précis, la définition de ce qu’est une information personnelle est un peu plus subjective, et dépend parfois de la culture : par exemple, en France il est assez rare de discuter ouvertement du montant de son salaire, tandis qu’aux États-Unis, c’est un thème qu’il est plus naturel d’aborder.

« Ce n’est pas que j’ai quelque chose à cacher, c’est que je n’ai rien à vous montrer. »

Une définition historique ?

« *Nul ne sera l'objet d'immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, ni d'atteintes à son honneur et à sa réputation. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes.* »

Déclaration universelle des droits de l'homme de 1948

Le contexte de l'époque

Il faut bien se remettre dans le contexte de l'époque, à la sortie de la Seconde Guerre mondiale, et à la **problématique des régimes totalitaristes**, qui cherchaient par tous les moyens possibles de s'assurer de la « loyauté » des individus.

Comme indiqué dans l'article de la Déclaration des droits de l'homme, on imaginait les immixtions comme étant des actions physiques : **intervention dans son domicile, ouverture des lettres, espionnage**, etc.

Le contexte actuel

Avec les technologies numériques, en particulier le web et les objets connectés, **une foultitude de nouvelles sortes d'intrusion** dans la sphère privée sont désormais possibles :

Au travers de nos **smartphones**, de nos **messageries** mail ou instantanées, de nos **GPS**, des **appareils de fitness**, de nos **lunettes connectées**, ou tout simplement des requêtes que nous posons à un **moteur de recherches**, nous sommes amenés à dévoiler des quantités incroyables de données, et ces données sont selon toute vraisemblance des données personnelles.

Prendre conscience du problème

Nous sommes donc en droit de nous interroger sur la légitimité de la collecte et de l'analyse de celles-ci, surtout lorsque ces traitements s'effectuent **à notre insu**, ou sans notre consentement.

Plus que jamais, l'article de 1948 reste pertinent aujourd'hui : nul ne devra être *l'objet d'immixtions arbitraires dans sa vie privée*.

Quelques exemples

- Le site « please rob me » (2010) : https://pleaserobme.com/
- Le scandale Cambridge Analytica (2016)
- La brèche de données de Strava (2018)

2018 : Traces de running de
La base de Bagram (Afghanistan)

Comment se protéger ?

La protection de la vie privée doit donc se faire même si les utilisateurs commettent des imprudences, ou des erreurs. Si on conseille aux automobilistes d'être attentifs et prudents, il n'en reste pas moins que l'utilisation de la ceinture de sécurité est obligatoire.

Une **solution radicale** existe pour protéger sa vie privée : ne pas utiliser les nouvelles technologies, ne pas aller sur Internet, etc.

Cette solution n'est clairement pas satisfaisante.

Nous allons voir qu'en effet **des solutions existent** pour à la fois proposer de nouvelles applications et respecter la vie privée.

L’exploitation de données personnelles

Nous laissons tous des traces …

Nous laissons tous une incroyable trace numérique, au travers de toutes les applications et tous les objets connectés avec lesquels nous interagissons.

Parfois, nous sommes conscients de laisser une trace. Par exemple, lorsque nous nous inscrivons sur un réseau social, nous sommes conscients d'indiquer notre nom, ou peut-être même de fournir d'autres données comme notre âge ou notre adresse. Lorsque nous nous connectons à d'autres amis, nous savons sans doute également que cette information sera stockée par le réseau social.

Une exploitation à notre insu

Ce dont nous nous doutons peut-être moins, c'est que lorsque nous naviguons sur le web, ce même réseau social sur lequel nous venons de nous inscrire va noter très consciencieusement les pages que nous visitons, et ce, même si nous ne nous sommes pas forcément connectés sur le site du réseau.

Pire, le réseau social est **capable de constituer des données sur des personnes qui ne sont même pas membres du réseau**, mais parce qu'elle ont pu avoir accès à des informations personnelles issues par exemple de répertoires mail.

Voir addon : Thunderbeam-Lightbeam (Chrome)

Voir addon : enhanced tracking protection (Firefox) dans **about:protections**

Qu’est ce que Google sait de moi ?

/!\ A faire seul(e) pour plus de confidentialité !

Les traces Google sont disponible sur :
https://myactivity.google.com/myactivity

N’hésitez pas à faire défiler la page, vous vous rendrez compte que des informations très anciennes peuvent être stockées !

Il est intéressant de noter que Google met un petit avertissement en haut de la page « **Vous êtes la seule personne à pouvoir voir ces données** »

Vraiment ?

L’expoitation des données personnelles est un business

Ces données servent à être **exploitées par des techniques d’apprentissage automatique**.

L’apprentissage automatique est une branche de l’intelligence artificielle, qui a besoin pour fonctionner d’un grand nombre de données, afin d’essayer de construire des modèles prédictifs. Ainsi, à l’aide de toutes les données qu’elle possède sur vous, Google est capable de déduire tout un tas d’informations.

Encore une fois, si vous êtes utilisateur de Google, vous pouvez vous rendre sur https://adssettings.google.com/ pour voir ce que Google sait de vous, et de vos centres d’intérêts.

Les données sont-elles vendues ?

Google explique que l'information sur la thématique provient d'un **profilage** basé sur des informations de visionnage de vidéos ou de leur moteur de recherches.

Il est à noter que Google dit explicitement que **vos données personnelles ne sont pas vendues** à d'autres entreprises. Toutefois **elle les possède et les exploite**.

Cas pratique : une petite entreprise qui collecterait des données et indiquerait qu'elle ne les vend pas. Si l'entreprise elle-même est rachetée par une autre entreprise, ce n'est pas une vente d'informations ; néanmoins l'entreprise acquéreuse aura désormais accès à toutes les données personnelles utilisateurs. C'est une des raisons de certains rachats de sociétés dans le domaine du web...

Une solution (juridique) : la collecte limitée des données

RGPD : Article 5 :

Les données à caractère personnel doivent être:

(…)

adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (minimisation des données);

Dans les cas cités précédemment, cette **collecte** est **énorme**, et sans doute **démesurée** par rapport à l'utilité qu'en fait l'individu concerné par les données : on ne recherche pas tous les jours dans ses 10 ans d'historique !

Prenez en compte cette problématique si vous avez à travailler sur des données personnelles !

Contrôler l’accès à vos données

Qui peut accéder à vos données ?

Il existe 4 types d'acteurs :

1. La personne ou des personnes concernées par la donnée elle-même.
2. D'autres personnes ou entités qui sont connues de la personne concernée par la donnée, et avec lesquelles elle souhaite partager cette donnée.
3. Les individus ou entités qui gèrent cette donnée personnelle.
4. Tous les autres individus ou entités (avec qui on ne souhaite a priori pas partager la donnée !)

Les droits de la personnes concernée

Les individus de la catégorie 1 peuvent avoir le droit d'**accéder** à la donnée, de la **modifier** ou même l'**effacer**.

Ces droits sont **garantis dans le RGPD** par plusieurs articles : l'article 12 intitulé « *Transparence des informations et des communications et modalités de l'exercice des droits de la personne concernée* », l'article 16 intitulé « *Le droit à la rectification* », et l'article 17 intitulé « *Le droit à l'effacement* ».

Les personnes avec qui on a partagé la donnée

Souvent, un mécanisme assez simple de **groupes** et de **permissions** est mis en place, ce qui permet de **simplifier le partage**, plutôt que d'être obligé de définir très exactement quels autres utilisateurs peuvent accéder à la donnée.

Ainsi, l'utilisateur pourra définir par des règles simples quels groupes peuvent accéder à telle ou telle ressource. La difficulté ici est de permettre une gestion claire des droits, y compris par des utilisateurs néophytes.

Bien entendu, pour pouvoir bénéficier de ce partage, il faut que ces autres utilisateurs soient enregistrés sur l'application.

Le problème de la portabilité

Nous touchons ici l'une des limites du partage de données dans un contexte d'application propriétaire : l'**obligation qu'ont les individus à utiliser le même système pour pouvoir interagir**.

C'est l'une des raisons de l'introduction du « *droit à la portabilité* », présenté dans l'article 20 du RGPD, qui exprime l'obligation de la part du fournisseur de service d'offrir la possibilité à un utilisateur de récupérer ses données, dans un format lisible, afin de pouvoir changer de fournisseur de services, ou tout simplement récupérer ses données.

Les entités gérant les données

C’est peut-être la **catégorie la plus critique**. En effet, les utilisateurs lui **délèguent la gestion de leurs données personnelles**, parfois au travers d’un contrat, comme l’utilisation d’un espace de stockage payant, ou en général, par un **accord** basé sur des conditions d’utilisation d’un service, en général gratuit.

Les **administrateurs** de ces bases de données, ou de ces serveurs, peuvent donc parfois **avoir accès** à ces données, tout simplement parce qu’ils ont tous les droits sur leur application.

Les questions à se poser

Il faut toujours **s'interroger sur le « business model »** de telles applications, ce qui permet souvent de comprendre pourquoi elles sont « gratuites ».

Les administrateurs sont donc un **facteur de risque**, ou plus précisément des **adversaires** qu'il convient de considérer lorsqu'on développe une application gérant des données personnelles.

Bien sûr, **dans la vaste majorité des cas, les employés et l'entreprise sont honnêtes**, mais il existe néanmoins un risque de tomber sur un employé malhonnête souhaitant récupérer les données pour des objectifs frauduleux.

Réduction du risque ?

Afin de réduire ce risque, il est possible d'utiliser du **chiffrement**, c'est-à-dire utiliser des techniques cryptographiques afin de rendre les données illisibles à toute personne qui n'aurait pas la clé de déchiffrement.

Dans ce cas, le système de stockage conserve **uniquement des données chiffrées et donc inutilisables**, même par l'administrateur du système.

C'est par exemple ce qu'il se passe dans le cas des emails chiffrés : seule la personne destinataire de l'email est capable de le lire, ce qui fait qu'il n'est pas gênant pour la confidentialité que l'email transite ou soit stocké par divers serveurs.

Utilisation du chiffrement ?

Toutefois, utiliser des techniques de chiffrement impose de **gérer des clés de chiffrement**, c'est-à-dire les clés utilisées pour communiquer avec un individu précis. Gérer des clés de chiffrement signifie stocker quelque part ces clés de chiffrement, par exemple sur son ordinateur, ce qui fait qu'**on peut également être victime de piratages divers**...

C'est pourquoi il est suggéré de conserver aussi ses clés sur un **support de stockage externe**, par exemple une **clé USB**, qui serait elle stockée dans une boîte ou même un coffre fort, c'est-à-dire hors connexion.

Ne pas les stocker sur le Cloud !

Et le reste du monde ?

Les entités de cette catégorie peuvent déployer toutes sortes de **stratagèmes** pour essayer de **dérober les données**, allant de l'**attaque** des infrastructures, à la **corruption** des administrateurs, en passant par le **hameçonnage**, ou tout simplement par la récupération de données pour lesquelles le contrôle d'accès aurait été mal protégé !

Ainsi, il est souvent possible, en utilisant des données librement accessibles, de déduire un grand nombre d'informations personnelles sur des individus. Certaines entreprises, appelées des « **data brokers** » en ont fait leur gagne-pain, par exemple le site intelius.com qui propose de fournir les informations personnelles d'un individu, lorsque vous lui fournissez son nom.

Protection des données sur les réseaux

Des données faciles à intercepter

De nombreux logiciels existent, permettant d'observer les données qui transitent sur les réseaux ; l'un des plus connus étant **Wireshark**. En utilisant un tel logiciel, il est **assez facile d'observer** puis d'analyser **les données échangées entre un appareil du réseau et un serveur**. En effet, il est plus rapide d'échanger des données en clair, plutôt que de prendre le temps de les chiffrer.

Le protocole HTTP (Hypertext Transfer Protocol)

Le protocole HTTP, qui permet à une application de dialoguer avec un site web, **n'est pas chiffré par défaut**. Il est donc possible pour n'importe quel observateur du réseau de voir quelles sont les données échangées.

Afin de **sécuriser cette communication**, des protocoles sécurisés ont été développés, comme TLS (*Transport Layer Security*) qui est employé comme une brique utilisée pour implémenter **le protocole HTTPS**, qui permet une communication sécurisée entre une application et un serveur web, sachant que la transmission elle-même peut se faire indépendamment par moyen filaire ou Wi-Fi.

Fonctionnement “simplifié” de TLS

La technique utilisée est basée sur de la **cryptographie à clé publique (ou *asymétrique*) et de cryptographie à clé secrète (ou *symétrique*)**. La partie asymétrique permet d’échanger une clé symétrique, qui est utilisée pour chiffrer les communications, mais le protocole est également utilisé pour garantir, via une signature, l’authenticité du serveur web.

Cryptographie symétrique

La **cryptographie symétrique** est une des techniques de base de la cryptographie. Elle permet à deux entités **qui possèdent toutes les deux une unique clé partagée** de s'échanger des messages qu'elles sont les seules à pouvoir déchiffrer ; mais il faut bien sûr **avoir réussi à se mettre d'accord sur cette clé** au préalable !

Cryptographie asymétrique

La **cryptographie à clé publique** (asymétrique) permet, elle, à un individu de chiffrer un message en utilisant une clé publique (et mise à disposition publiquement). Cette **clé publique**, qui peut être **diffusée sans aucun risque**, est associée à une clé privée connue uniquement par l'entité. C'est cette dernière qui servira à déchiffrer les messages chiffrés en utilisant la clé publique en question.

La cryptographie à clé publique est bien plus **lente** que la cryptographie symétrique. Aussi, la plupart des protocoles mettant en jeu des clés publiques ne cherchent en fait qu'à échanger une clé secrète, qui sera ensuite utilisée pour transférer de gros volumes de données.

Cryptographie : cacher le contenu de l'enveloppe

Avec ces techniques de chiffrement, on voit qu'on est capable de **protéger le contenu** d'un message échangé entre un objet connecté A et une application web B.

Par contre, un observateur serait tout de même capable de savoir que **A et B sont en train de communiquer entre eux** !

TOR : cacher avec qui on communique

Imaginez que vous envoyiez un **message à l'intérieur d'une enveloppe**. Sur cette enveloppe, vous inscrivez l'adresse du **destinataire X**.

Si vous donnez cette enveloppe à votre facteur, même s'il ne l'ouvre pas, il saura que vous êtes en communication avec X.

Supposons que vous mettiez cette enveloppe à l'intérieur d'une autre enveloppe, et que vous envoyiez cette plus grosse enveloppe à **une autre personne Y**, qui sera chargée d'ouvrir l'enveloppe et de poster la lettre qu'elle y trouvera. Dans ce cas, le facteur saura que vous communiquez avec Y, et que Y communique avec X, mais ne verra pas que vous communiquez directement avec X.

C'est ainsi que fonctionne, dans ses grandes lignes, le réseau TOR, mais avec un nombre plus important d'enveloppes, d'où le nom de système de routage en oignon, puisque les multiples enveloppes peuvent être assimilées à des pelures d'oignon.

Les limites de TOR

Le système TOR est **sécurisé si la majorité des intermédiaires sont honnêtes**, c'est-à-dire s’ils ne communiquent pas entre eux.

Pour obtenir une telle condition, le réseau TOR fait donc appel à de multiples entités différentes tout autour du globe.

Bien entendu, la **vitesse de transmission** d’une information via TOR est **beaucoup plus lente** qu’une transmission en direct, mais elle permet de totalement dissimuler nos communications. Le débit est également souvent réduit.

Conclusion partie I

Dans cette partie, nous avons vu que des **données personnelles sont présentes dans la plupart des applications informatiques** de notre époque, en particulier dès qu'on commence à utiliser des objets connectés.

D'un point de vue légal (cf. Règlement général sur la protection des données, ou **RGPD**) et d'un point de vue éthique, il est important de traiter ces données de manière respectueuse.

Nous avons vu dans cette partie quels étaient **les risques** liés à une gestion qui ne serait pas respectueuse des utilisateurs, et pointé **certaines pratiques courantes** qui peuvent poser question.

Nous avons également donné **quelques pistes de réponse** concernant les questions de contrôle d'accès et de transfert des données.

Partie II : Le RGPD

Grandes lignes du règlement

Retour historique

Le concept de la vie privée est très ancien, puisqu'on peut le faire remonter à **Aristote**, qui distinguait la sphère publique des activités politiques (de la cité) et la sphère privée, avec les activités de la famille et du foyer.

La question de la protection de la vie privée est plus récente, et date de **la fin du XIXe siècle**, où elle avait été définie à l'époque par deux juristes, **Samuel D. Warren** et **Louis Brandeis**, dans un article scientifique intitulé « The Right to Privacy » (le droit à la vie privée), qui avait été rédigé suite au développement de la presse écrite et de la photographie.

La loi “informatique et liberté” de 1978

Sur le plan informatique, la loi générale pionnière est la **loi française 78-17 du 6 janvier 1978** relative à l’informatique, aux fichiers et aux libertés, appelée plus couramment « loi Informatique et Libertés ».

Il est très intéressant de voir que cette loi a subi très peu de modifications, malgré le fait qu’elle ait été conçue à l’aube de l’informatique personnelle, et bien loin du traitement massif de données effectué de nos jours.

Déjà en 1978, le législateur français avait l’expérience de ce que pouvait être le traitement des « fichiers » informatiques, en particulier à l’aune du **fichage des personnes sous Vichy**. La plus grande précaution était donc de mise, et même aujourd’hui, les garanties de la loi de 1978 sont excellentes.

La directive européenne 95/46/CE

En 1995, L'Europe s'est dotée d'une directive, la **directive 95/46/CE** sur la protection des données personnelles. La loi française issue de la directive européenne gérait le problème général des données personnelles, tandis qu'aux États-Unis, il existe de multiples lois pour gérer toutes sortes de données : les données médicales, les données de crédit, etc

La **directive** est un instrument de nature législative utilisé par l'Union européenne pour prendre des mesures. La directive passe par deux étapes avant de produire ses effets : une fois votée par les institutions européennes, elle doit ensuite être transposée par les Etats membres dans leur droit national, à la différence du **règlement**, qui s'applique directement.

Le Règlement Général sur la Protection des Données (RGPD /GDPR)

Depuis le 25 mai 2018, le **Règlement général sur la protection des données** personnelles est entré en vigueur. Il s'agit d'une loi européenne, s'appliquant dans tous les pays de l'Union.

Au final, il reprend dans les grandes lignes les concepts tracés par la loi française de 1978, à un petit détail près : **les peines pour non-respect de la loi sont désormais dissuasives**, y compris pour les acteurs majeurs d'Internet, pour qui les peines précédentes de quelques centaines de milliers d'euros n'étaient qu'une paille.

Protections principales du RGPD

- la **loyauté** (qui s'assure que le traitement effectué est raisonnable dans sa mise en place ou « n'abuse pas ») ;
- le **consentement** (qui permet d'informer la personne qu'on va collecter ses données, dans quel but, et de recueillir son accord) ;
- le **droit d'accès** (rectification et oubli, qui permettent aux personnes de modifier ou faire disparaître des données les concernant) ;
- le **droit à la notification** (qui oblige un responsable de traitement d'informer l'utilisateur s'il est victime d'un piratage) ;
- l'obligation de construire des systèmes informatiques « **Privacy by design** », c'est-à-dire prendre en compte la problématique de la protection des données dès la conception du système en réalisant une EIVP ;
- des **amendes dissuasives** (jusqu'à 10 millions d'euros, ou 4 % du chiffre d'affaires mondial, si ce montant est supérieur à 10 millions d'euros).

La finalité du traitement (loyauté)

RGPD : Les données doivent être :

1. traitées de manière licite, loyale et transparente au regard de la personne concernée (licéité, loyauté, transparence);
2. collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités; le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré, conformément à l'article 89, paragraphe 1, comme incompatible avec les finalités initiales (limitation des finalités);
3. adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (minimisation des données);

Le consentement

- Libre et éclairé

- Que penser du fait de cliquer sur des conditions d'utilisation ?
- Que penser des bandeaux sur les cookies ?
- Comment retirer son consentement une fois donné ?
- …

Zoom sur le droit à l'oubli

Un concept juridique

Dans le cadre d'une condamnation, une fois qu'un individu a purgé sa peine, il est considéré comme « quitte » avec la société, et on ne doit plus lui tenir rigueur des événements passés. Le droit à l'oubli est donc **le droit de « passer à autre chose »** ou le droit d'oublier les erreurs faites dans le passé.

Dans la société prénumérique, ce droit pouvait s'exercer *de facto*. En effet, il n'était pas forcément simple d'exhumer des photos ou des documents présentant les erreurs passées d'un individu. Même la plupart des archives ne doivent être gardées qu'un certain temps.

Toutefois, le web possède la caractéristique d'**hypermnésie**, c'est-à-dire qu'il a la capacité de se rappeler de tout. On le voit en particulier avec le fait d'éplucher ce qu'a pu dire telle ou telle personne il y a des années sur un réseau social.

Un droit contesté aux US

Aux États-Unis, l'accent est mis sur la **liberté d'expression et la liberté d'informer**, qui s'oppose parfois au droit à l'oubli : si quelque chose s'est passé, on serait en droit de pouvoir en parler.

Comment le droit à l'oubli peut-il donc s'exprimer dans un monde qui n'oublie rien, et où les lois sont différentes ?

Le cas des moteurs de recherche

En réalité, c'est moins la présence sur le web de l'information qui gène, que la capacité simple à la retrouver via des moteurs de recherche.

Si vous voulez savoir des informations sur un individu, la première chose qui vient à l'esprit est de taper son nom dans la barre de recherche Google. Jusqu'à récemment, Google ne filtrait aucun des résultats que pouvait retourner une recherche.

Le cas *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González*

Contexte :

En 2014, un espagnol, **Mario Costeja, a attaqué la filiale espagnole de Google** parce que le moteur de recherches continuait de retourner des résultats datant de 1998 de petites annonces du journal "La Vanguardia", sur le fait qu'il avait été obligé de vendre sa maison pour des raisons d'impayés. Il faisait valoir que plus de 15 ans après, il n'y avait plus aucune raison d'avoir accès à cette information qui lui était préjudiciable.

Le cas *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González*

Complexité du problème :

Google n'est finalement que le **vecteur d'accès à l'information** que Costeja voulait voir disparaître. Serait-il légitime de demander au journal d'effacer la page ? Sans doute pas, le document original du journal pouvant être pertinent pour des raisons historiques. Par contre, qu'une recherche sur le nom de l'individu retourne un lien vers cette annonce est plus gênant.

Verdict

La cour de justice a donné raison au plaignant et a demandé à Google **d'effacer les liens** vers ces données et de **mettre en place une procédure** permettant à n'importe qui de demander l'effacement de ses données personnelles sur la base du droit à l'oubli.

Vous pouvez faire le test vous-même : tapez le nom d'une personne dans le moteur de recherches, vous verrez apparaître en bas de l'écran une mention indiquant que *certains résultats peuvent être omis en raison de la loi européenne*.

Du droit à l'oubli à la censure

Toutefois, il convient de **s'interroger sur le pouvoir octroyé aux moteurs de recherches** : en vertu d'une loi européenne, toutes les réponses à une requête donnée ne sont par retournées.

Finalement, quelle différence avec le fait qu'en Chine une requête sur le nom d'un dissident ne retourne aucune page non plus ? La différence entre le droit à l'oubli et la censure est donc ténue.

Par ailleurs, ce droit à l'oubli est en vérité un droit au « **déréférencement** », c'est-à-dire que la donnée elle-même n'est pas effacée, mais c'est le lien vers cette donnée – une requête Google – qui est retiré, même si dans les faits, ce qui compte c'est effectivement d'être capable de retrouver et d'accéder à cette donnée.

De l’implémentation du droit à l’oubli

Notons enfin que c’est actuellement Google qui décide quelles données elle déréférence. Elle sait donc qu’un individu a demandé expressément qu’une donnée soit déréférencée, et doit conserver cette information.

N’est-ce pas un peu paradoxal ?

Comment se fait-il que les état délèguent cette responsabilité capitale de décision à une entité privée ?

Le paradoxe du droit à l'oubli (Paradoxe d'Erostrate)

Il suffit de taper le nom de Costeja dans un moteur de recherches pour se rendre compte que le risque d'essayer de faire appliquer le droit à l'oubli est que l'exercice de ce droit peut vous mettre sous les feux des projecteurs !

« — Je le connais votre type, me dit-il. Il s'appelle Érostrate. Il voulait devenir illustre et il n'a rien trouvé de mieux que de brûler le temple d'Éphèse, une des Sept Merveilles du monde.

— Et comment s'appelait l'architecte de ce temple ?
— Je ne me rappelle plus, confessa-t-il, je crois même qu'on ne sait pas son nom.
— Vraiment ? Et vous vous rappelez le nom d'Érostrate ? Vous voyez qu'il n'avait pas fait un si mauvais calcul. »

Jean-Paul Sartre, *Le Mur*

L’Analyse d’Impact de Vie Privée

Pourquoi faire une AIVP ? (ou EIVP)

La réalisation d'une AIVP est **indispensable** lors de la mise en place d'un système de traitement de données personnelles, comme par exemple le logiciel de gestion de données de type « running » exploitant des données issues d'une montre connectée.

Elle permet de vérifier la conformité des traitements mis en œuvre.

Elle permet de vérifier qu'on a bien pris en compte l'ensemble des risques, et procédures de réduction de risque.

Elle pourra être opposée au tribunal en cas de litige.

La méthode CNIL

Afin de réaliser cette AIVP, nous allons nous baser sur l'utilisation de la **méthode et du logiciel proposés par la CNIL**, et disponibles en ligne.

Vous pourrez télécharger le logiciel (disponible sur les plateformes usuelles), ainsi que les trois guides qui décrivent la démarche, les modèles pour formaliser l'étude, et le catalogue de mesures pour traiter les risques issus de l'analyse.

La démarche de l'AIVP est que d'une part on protège des droits « fondamentaux » en quelque sorte non négociables, et que d'autre part on cherche à gérer les risques sur la vie privée, en proposant les meilleures mesures possibles pour y arriver.

Etapes de l'AIVP

Une AIVP se découpe en 4 grandes parties :

1. **Délimiter et décrire le contexte** des traitements considérés dans l'analyse.
2. **Analyser des mesures** garantissant le respect des principes fondamentaux : proportionnalité et nécessité du traitement, protection des droits.
3. **Évaluer les risques** sur la vie privée liés à la sécurité des données et vérifier qu'ils sont correctement traités.
4. **Valider**.

L’exemple : application de “running”

La montre connectée (e.g. Android) appartenant à l'utilisateur, client de la société PrivateRun. L'utilisateur a installé un logiciel propriétaire de PrivateRun qui permet de récupérer des informations sur la géolocalisation de la personne la portant (position et vitesse) et de les stocker en local sur la montre (ou le portable connecté à la montre), mais également de pouvoir les uploader sur un serveur de PrivateRun.

Des serveurs (ou cloud) loués par PrivateRun à la société SuperCloudProvider sur lesquels la société PrivateRun a installé un logiciel permettant d'effectuer des traitements sur les données uploadées par les utilisateurs sur ce cloud, et de les présenter sur un site web hébergé sur le même cloud. Le site web contient une partie privée, contenant les données de toutes les trajectoires d'un utilisateur, accessible via un login et mot de passe, et une partie publique avec des données qui sont disponibles à tout le monde, sans login.

PrivateRun s'autorise à mettre à jour le logiciel sur la montre connectée à distance, après accord de l'utilisateur.

Etape I : Contexte du traitement

La première étape concerne le **contexte des traitements**. Il faut tout d'abord identifier le **responsable du traitement**, puis décrire la **nature, la portée, le contexte, la finalité et les enjeux** du traitement.

Ici, le responsable de traitement sera l'entreprise « running », qui va collecter les données issues de la montre, avec comme objectif de mesurer le parcours exact, le temps mis, et calculer les calories brûlées.

L'objectif de l'application est de fournir ensuite ces données à l'utilisateur, qu'il puisse les comparer à celles de ses amis, et de publier les trajectoires de manière anonyme sur le site.

Etape I : Données concernées et destinataires

On décrit ensuite de manière détaillée les **données personnelles concernées, les destinataires, et la durée de conservation**.

Ici par exemple, il s'agira de s'interroger sur la possibilité de partager les données avec d'autres utilisateurs. Mais dans tous les cas, le site « running » aura accès à toutes les données.

Etape II : Respect des principes fondamentaux

Dans la deuxième étape, on vérifie la conformité aux principes fondamentaux : **finalité** (explicite et légitime), **fondement** (légal et pas de détournement), **consentement, minimisation des données, qualité des données et durées de conservation**.

Dans notre exemple, la finalité est assez claire. En revanche, si par exemple on voulait se servir des données pour proposer des publicités de chaussures de sport, ou recommander des salles de sport proches des trajectoires observées, bref **proposer des services suite à un profilage**, ce serait un **détournement de finalité**.

Il faudrait dans ce cas, pour respecter le critère de **consentement**, indiquer tous ces traitements à l'utilisateur. De plus, si au final on veut profiler un utilisateur pour lui proposer une recommandation géolocalisée, le principe de **minimisation** indique qu'on n'a peut-être pas besoin d'informations aussi précises que la position et le temps...

Etape II : Respect des principes fondamentaux

D'autres droits, qu'on a évoqués précédemment, doivent être vérifiés ici, comme le **droit à la portabilité, ou le droit de rectification ou encore le droit à l'oubli**.

On est ici déjà en train de réaliser **l'évaluation des mesures de protection** des droits des personnes : on doit s'assurer de l'information des personnes (loyauté et transparence du traitement), et du recueil du consentement.
D'autres droits, qu'on a évoqués précédemment, doivent être vérifiés ici, comme le **droit à la portabilité, ou le droit de rectification ou encore le droit à l'oubli**.

Qu'avons-nous prévu pour un utilisateur souhaitant se retirer du système ? Allons-nous pouvoir effacer ses données ? Pouvons-nous lui donner un export de toutes ses trajectoires dans un format lisible ?

Etape III : Analyse de risque

Dans la troisième étape, nous menons une analyse des risques liés à la sécurité des données. Un risque dépend de deux facteurs : la **gravité** et la **vraisemblance**.

L'évaluation d'un risque est souvent le **produit de ces deux facteurs**.

Considérons le risque d'une **attaque informatique de la part de l'administrateur** du système. La gravité serait très sévère puisqu'on perdrait l'ensemble des données. La vraisemblance va dépendre des précautions prises lors du recrutement, voire de la rémunération de cet individu. Si on souhaite réduire ce risque, on pourra par exemple chercher à éviter qu'une seule personne puisse avoir accès ET au système ET à la base de données.

On pourra aussi **auditer** le système, c'est-à-dire avoir un système qui tracerait et permettrait à l'individu de remonter jusqu'à l'entreprise qui a volé ses informations. On pourra citer comme mesure de prévention le chiffrement des données, l'anonymisation, le cloisonnement, ou le contrôle d'accès.

Etape III : Acceptation du risque

Une fois l'étude des risques et des contre-mesures effectuée, il s'agit pour le responsable de traitement **d'accepter ou de refuser le risque**.

Si on refuse le risque, alors il faut reboucler et rajouter des mesures de protection supplémentaires.

Etape IV : Validation

La dernière étape de l'AIVP est de la faire **valider** par le responsable de traitement, en particulier voir s'il accepte le risque. En effet, si jamais l'événement redouté advient, il faudra se défendre au tribunal.

Bien entendu, le risque minimal n'existe qu'en l'absence de traitement. Un juge évaluera donc l'ensemble des critères, la licité du traitement, sa proportionnalité, et ne sera bienveillant avec le responsable du traitement que s'il juge que le **risque était acceptable**.

Conclusion Partie II

Dans cette partie, nous avons présenté très rapidement certains concepts de **protection** de la vie privée qui doivent être légalement pris en compte lors de la **réalisation** d'une application informatique. Nous avons également détaillé dans ses grandes lignes la **technique de réalisation** d'une analyse d'impact de vie privée (AIVP).

Nous réaliserons une AIVP en TD.

Partie III : Sécurité Informatique

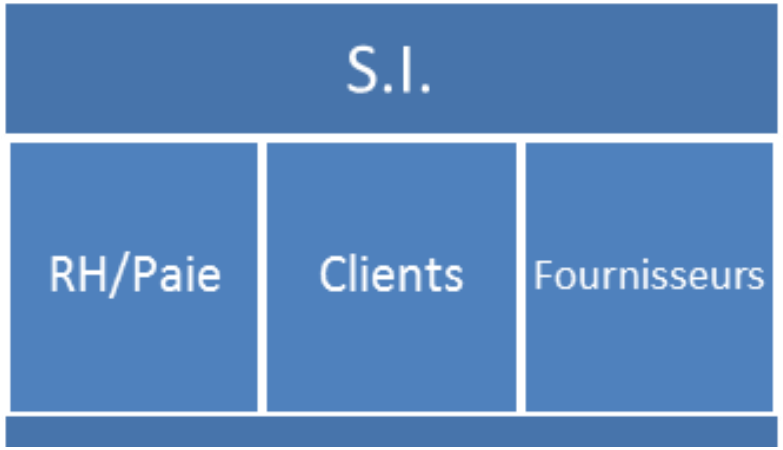
Cartographie des attaques

Les risques des systèmes d'information

Commençons par identifier les risques liés aux systèmes d'information.

Qu'est-ce qu'un système d'information ?

On parle de système d'information, ou *SI*, pour désigner un ou plusieurs logiciels utilisés pour gérer des données. En général, ce sont des données d'une entreprise, d'une administration, etc.



Le système d'information concerne une entreprise de vente. Il est composé de trois éléments :

1. des informations autour des ressources humaines et de la paie ;
2. des informations autour des clients ;
3. des informations pour des fournisseurs.

Le système de gestion des base de données (SGBD) : au coeur du SI

Plus précisément, nous nous intéressons au SGBD : **système de gestion de base de données**. C'est une brique centrale de ce système d'information.

Il y a de nombreux **SI** qui contiennent des données sensibles :

1. les SI **nationaux** comme les impôts, les fichiers de police, les dossiers médicaux partagés, etc. ;
2. les SI **d'entreprises**, de banques ou encore d'assurances ;
3. les bases de données **personnelles** constituées par les individus eux-mêmes (Drive, Dropbox) sur lesquelles ils vont stocker des factures, des fichiers personnels ;
4. les bases de données **"ambiantes"** générées à partir de capteurs. Cela peut être une base de données qui stocke les passages au niveau tourniquet comme le Pass Navigo, ou bien une base de données de téléphone mobile qui stocke les appels passés à partir de certaines antennes, ou encore une base de données de télésurveillance utilisant des capteurs à l'intérieur d'une maison, pour retransmettre ou stocker des informations à propos des personnes qui sont entrées dans la maison.

Quelques attaques contre les SI

Un SI contient des données structurées ayant une forte valeur ajoutée. Très souvent, de gros volumes de données sont stockés dans un SI.

En général, lorsque les "hackers" s'introduisent dans les SI, ils vont chercher à y dérober des informations personnelles, financières, etc. Parfois, il peut simplement y avoir destruction des données ou prise en otage des données avec un "rand software", par exemple. Les données vont être chiffrées et l'utilisateur devra payer une certaine somme d'argent pour se les voir rendre.

Objectif : Le vol de données

En regardant plus précisément les vols de données, nous voyons que, depuis ces dernières années, nous avons affaire à des chiffres astronomiques.

- En 2015, il y avait à peu près 3 000 brèches de données, c'est-à-dire des pertes de données ou des intrusions dans des SI qui ont exposé près de 736 millions d'enregistrements constituant des données personnelles.
- En 2021, il n'y avait plus que 1500 brèches, pour plus de 5 milliards d'enregistrements !

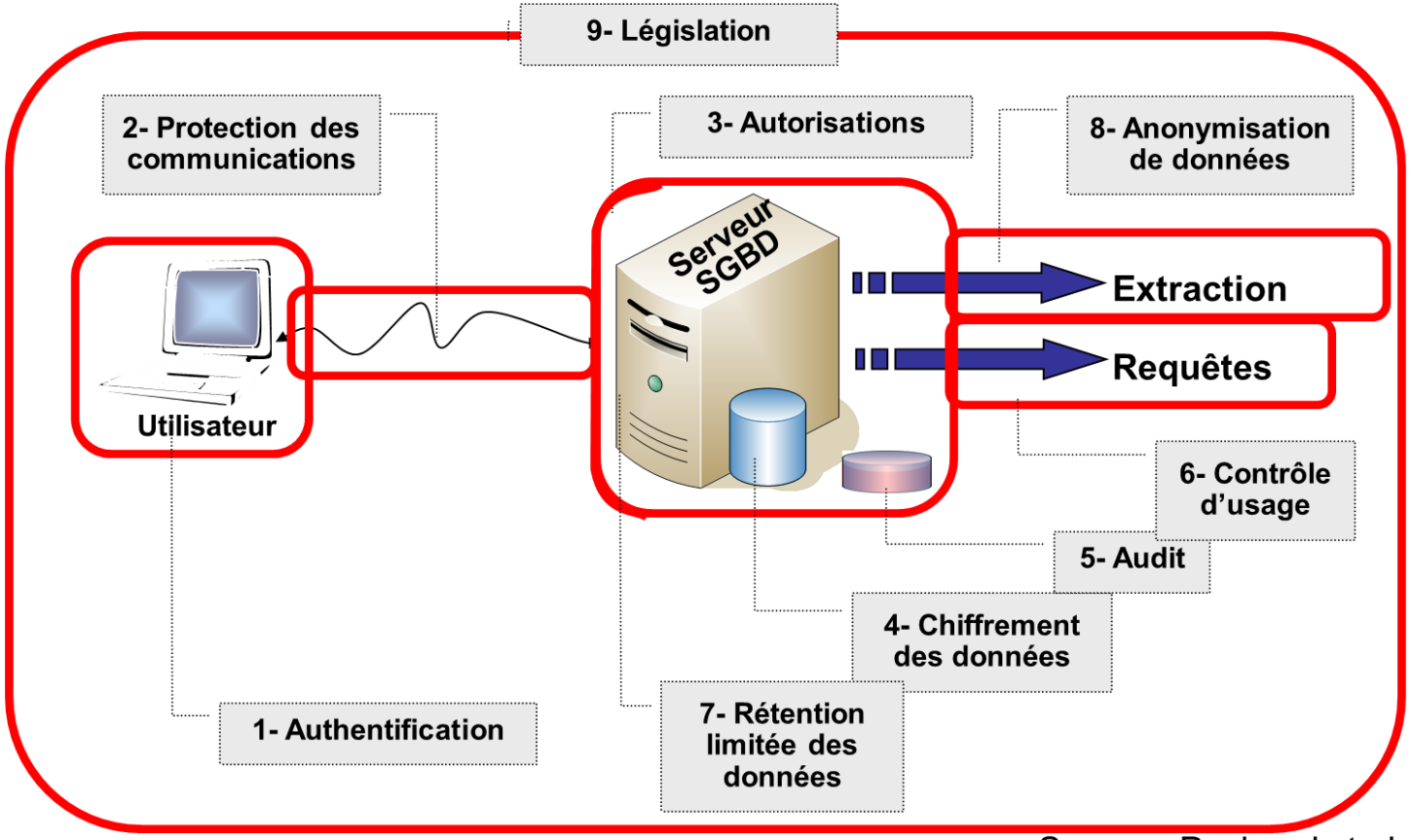
Le coût « ex post » d'une brèche est entre $100 et $500 par enregistrement.

Divers types d’attaques

1. La première attaque vient d'une **vulnérabilité logicielle**, donc une **vulnérabilité du SGBD** lui-même, permettant une intrusion car le logiciel a été mal programmé, mal conçu en termes de sécurité.
2. Le même type d’attaque peut exister au niveau du **système d’exploitation**, par exemple Windows. Ainsi, le système d'exploitation peut avoir une faille de sécurité permettant à des utilisateurs non autorisés de lancer certains programmes ou d’observer le contenu de la mémoire, et donc de dérober des informations.
3. La méthode du **vol de mot de passe** est très prisée actuellement. Il s’agit d’essayer de récupérer le mot de passe d’un utilisateur, ou même de l’administrateur de la base de données, par diverses techniques. La principale utilisée est le "phishing" : on vous envoie un faux e-mail, en espérant que vous allez cliquer sur un site web et rentrer votre mot de passe.
4. Il y a de nombreuses autres attaques comme celles des **États eux-mêmes**. Avec le scandale "Prism" par exemple, nous avons su que la NSA cherchait à s’introduire dans la plupart des bases de données du monde entier pour observer ce qu’il s’y produisait.

Mécanismes de protection

Mécanismes de protection



Source : Pucheral et al.

L’authentification

L’authentification sert à s’assurer que la personne qui se connecte sur le système d’information est bien une personne donnée. Par la suite, cette personne aura des droits en **lecture** ou en **écriture** sur les données. Ces droits vont dépendre des **règles de contrôle d’accès**. La problématique de l’authentification est de savoir si la personne qui se connecte est bien la personne qu’elle prétend être.

Fonctionnement de l’authentification

Il existe **trois** méthodes permettant d’identifier une personne.

1. La première méthode consiste à regarder **ce que la personne possède**. C’est ce qu’il se passe lorsque vous arrivez devant un coffre-fort à la banque. Dans le monde informatique, il existe le "token matériel" ou le "token physique" : cela peut être, par exemple, une sorte de carte à puces ou bien un dispositif que vous pouvez brancher sur votre ordinateur et qui va vous authentifier.
2. Le deuxième point est l'authentification à l’aide d’une **information que vous connaissez**. C’est la technique classique du mot de passe ou de la question secrète.
3. Enfin, le troisième point est l'authentification avec **ce que vous êtes** : nous parlons ici de biométrie. Vous pouvez vous authentifier à l’aide d’une empreinte digitale, d’une empreinte rétinienne, etc.

Criticité de l'authentification

On voit que **sécuriser l'authentification** est crucial. Il faut, en particulier, que les mots de passe utilisés ou choisis qui sont associés à un login soient sûrs et changés souvent. En effet, il y a énormément d'attaques et beaucoup de bases de données contiennent des informations sur des mots de passe. Lorsqu'un système se fait attaquer, les pirates récupèrent une grosse base de données de mots de passe qu'ils peuvent réutiliser.

Dès lors que vous attaquez un système, c'est beaucoup plus **simple** et beaucoup plus rapide de tester des mots de passe qui ont déjà été utilisés une fois, plutôt que d'essayer d'en inventer. C'est la raison pour laquelle il est demandé à ce que les mots de passe soient changés souvent : il existe un risque que le mot de passe, qui est secret, se retrouve en clair dans la nature.

Quelle technique adopter ?

L’authentification biométrique n’est pas forcément plus sûre. Sur les téléphones, par exemple l’iPhone, l’authentification biométrique est possible en posant un doigt sur le capteur ; mais si jamais ce système ne fonctionne pas, le téléphone va basculer sur le système classique qui est celui du mot de passe.

Sur l’iPhone, l’authentification biométrique est plutôt utilisée comme une **authentification plus rapide** par rapport à l'action de taper un mot de passe. Éventuellement, elle est aussi plus **discrète**, puisque lorsque vous tapez votre mot de passe, il est possible qu'une personne se trouvant à côté de vous puisse essayer d’entrapercevoir le code que vous êtes en train de taper.

Comme discuté précédemment, vous pouvez vous authentifier avec quelque chose que vous possédez, comme le **"token"** ou encore le téléphone. Prenons pour exemple les modèles de paiement en ligne des banques comme le "**3D Secure**" qui est très souvent utilisé pour s’assurer que vous êtes bien la personne propriétaire du compte en banque. Un code à rentrer sur un site web vous est ainsi envoyé sur votre téléphone.

Authentification “forte”

On parle d’**authentification "forte"** lorsqu’un système est protégé par au moins **deux moyens d’authentification différents** comme, par exemple, un mot de passe et de la biométrie.

Pour l'iPhone, ce n’est pas une authentification "forte", puisque c’est mot de passe **ou** biométrie. Très souvent, mot de passe et dispositif physique sont combinés.

C’est le cas de la carte bancaire où, pour effectuer un paiement, vous devez être en possession de la carte et vous devez connaître en plus le code PIN à taper pour autoriser un paiement.

Le contrôle d'accès

Une fois qu'un individu est authentifié par le système, donc identifié, le contrôle d'accès va indiquer **ce que cette personne peut faire** : lire ou accéder à certaines données, modifier des données existantes, effacer des données ou encore créer de nouvelles données.

Il existe de nombreux modèles de contrôle d'accès :

- le contrôle d'accès obligatoire (***MAC***, en anglais) ;
- le contrôle d'accès discrétionnaire (**DAC**) ;
- le contrôle d'accès basé sur les rôles (**RBAC**) ;
- le contrôle d'accès basé sur les attributs (**ABAC**).

Concepts généraux du contrôle d’accès

Le contrôle d’accès définit pour chaque utilisateur et pour chaque objet de la base de données **ce que l'utilisateur a le droit de faire**. L'objet de la base de données peut être au niveau de l’enregistrement, au niveau de la table, au niveau de la vue ou encore au niveau d’une requête.

Ainsi, pour chaque couple utilisateur et objet, vous allez avoir la possibilité de lire l’objet ou bien d’écrire, c’est-à-dire modifier l'objet. Ici, la granularité varie selon l'utilisateur : par exemple, sur une table, vous pourrez peut-être lire certains attributs et pas d’autres, ou bien écrire certains attributs et pas d’autres.

Modèle DAC

Le **modèle DAC** est peut-être le plus simple : il permet aux utilisateurs de transférer leurs droits à d'autres, en estimant que chaque utilisateur va créer ses propres informations.

Ainsi, par exemple, vous stockez un fichier sur un SGBD et vous autorisez d'autres utilisateurs à lire ou modifier ce fichier. Ce type de contrôle d'accès est par exemple utilisé sur Dropbox et d'autres systèmes de stockage de fichiers en ligne.

Modèle MAC

À l'inverse, le **modèle MAC** définit une structure extrêmement rigide d'accès, ou non, à certaines données, selon des **critères d'accréditation**.

Chaque donnée va être qualifiée par un critère d'accréditation, par exemple "confidentiel", "top secret", "secret défense", et les utilisateurs eux-mêmes vont avoir des droits.

Selon les droits des utilisateurs, donc selon leur niveau d'accréditation, ils pourront accéder ou non à certaines données.

Modèle RBAC

Le **modèle RBAC** (le modèle de contrôle d'accès basé sur les rôles) est celui qui est le plus souvent utilisé dans les SGBD. C'est un système qui simplifie l'utilisation du contrôle d'accès par la création de **rôles** qui sont ensuite associés à des utilisateurs.

Ainsi, un rôle va avoir le droit d'effectuer un certain nombre d'actions sur les objets d'un SGBD, et chaque utilisateur va être associé à un ou plusieurs rôles qu'il pourra ensuite activer.

Modèle ABAC

Il existe d'autres modèles de contrôle d'accès plus récents, plus exotiques comme le **modèle de contrôle d'accès basé sur les attributs**.

Ce modèle va utiliser les valeurs des attributs des utilisateurs eux-mêmes pour leur donner des droits.

Prenons l'exemple classique de l'accès à une vidéo uniquement si l'utilisateur a plus d'un certain âge. Ici, nous allons donner à chaque objet une règle associée aux attributs, et nous allons ensuite comparer les attributs des utilisateurs lorsqu'ils cherchent à accéder ou à modifier tel objet de la base de données.

L’audit

L’audit des opérations sur une base de données permet de **connaître toutes les opérations qui ont été effectuées** par les utilisateurs et les administrateurs.

Il permet de mettre en évidence des comportements "louches" d’utilisateurs, en regardant s'ils ont accédé à des données auxquelles normalement ils n’ont pas accès.

Prenons l'exemple d'un accès non autorisé à des fichiers de police par des personnes ne travaillant pas sur une enquête en cours, et qui souhaitent obtenir des informations sur des individus.

Fonctionnement de l'audit

Il y a une **séparation** entre les **droits de l'administrateur de la base de données**, qui peut effectuer n'importe quel type d'opération sur les tables métier de la base, et les **droits de l'administrateur système**.

En effet, l'audit stocke automatiquement toutes les informations de la base dans un fichier auquel l'administrateur du SGBD n'a pas accès.

En revanche, l'administrateur du système lui-même, donc l'administrateur du système d'exploitation, peut avoir accès à ces informations, à ce fichier de logs qu'il peut éventuellement modifier.

Que stocker dans les logs ?

Dans les fichiers de logs, il est donc possible de conserver **toutes les traces** des accès ou des modifications des données. Cela peut être utile pour essayer de comprendre pourquoi une erreur s'est produite.

Dans tous ces accès, des informations sont également stockées.

Une question se pose alors : **qui a le droit d'avoir accès à ces informations de logs ?** En général, cet accès va être extrêmement restreint.

L'accès aux logs est uniquement exploité en cas d'accidents, par exemple dans le cas d'une panne ou d'une attaque pour essayer de reconstituer l'ensemble des événements qui se sont produits.

Conclusion : la sécurité par conception (Security by design)

Le concept de la sécurité et de la confidentialité "by design", c'est-à-dire par conception, est mobilisé pour sécuriser les SI.

Il s'agit de **prendre en compte la sécurité et la confidentialité** au moment où vous concevez votre système. En effet, une bonne sécurité et une bonne confidentialité des données doivent se prévoir à l'avance.

Il est en réalité beaucoup plus difficile et compliqué de rajouter après coup tous ces éléments, qui ont un impact important sur l'architecture du logiciel.

Techniques avancées de sécurisation du traitement de données

Enjeux et objectifs du DBaaS (Database as a service)

Poursuivons avec l’exemple de l'utilisateur qui a installé l'application "Running App" de la société "RAC" sur son téléphone. Les **données de l'utilisateur** (sa position GPS, ses données cardiaques, etc.) vont être **exportées** chez l’éditeur "RAC" qui va les **traiter**, vraisemblablement sur un serveur ou sur plusieurs serveurs de type cloud.

Des **fournisseurs de services** peuvent proposer à "RAC" de traiter eux-mêmes les données. C'est ce qu’on appelle le "**DBaaS**" ou "Data Base as a Service" (en français, *base de données comme un service*).

Le gestionnaire du DBaaS : un sous traitant de confiance ?

Il existe donc des entreprises qui fournissent ce service de **stockage** et de **traitement de données**. Dans ce cas, les données de l'utilisateur se servant de "Running App" vont être envoyées à "Running App Corp" qui va ensuite **déléguer** et **stocker le traitement de ces informations sur des serveurs tiers**.

Dans les deux cas, la **confiance** que peut avoir l'utilisateur envers le traitement de ses données est **relative**.

Il s'agit d'une **confiance contractuelle**, c'est-à-dire qu'elle ne prévient pas les problèmes mais qu'elle permet, éventuellement, de les résoudre au tribunal si jamais l'utilisateur constate un souci dans l'exploitation de ses données.

Protocoles PIR

Définition : Les protocoles de récupération d'informations confidentielles permettent à un client de récupérer une information d'une base de données stockée sur un serveur, **sans que ce serveur sache quelle est l'information qui a été transférée**.

PIR “trivial”

Il existe une solution triviale permettant au client de récupérer l'information en question. Elle consiste à **chiffrer toute la base de données** et, quelle que soit la requête, **retourner l’intégralité de la base de données chiffrée** à l’utilisateur.

Naturellement, ce concept de chiffrement de la base de données est **intrinsèque à l’approche**. Le fournisseur de service (celui qui stocke la base de données) ne doit évidemment pas pouvoir lire l’information qui est stockée sur le serveur.

Solution “simple” (?)

Considérons que la base de données est définie comme une liste de bits $K = 0, 1^n$. L'utilisateur souhaite récupérer le $i^{ème}$ élément de cette liste, soit k_i qui vaudra soit 0 soit 1.

Pour ce faire, l'utilisateur va construire **2 ensembles** E_1 et E_2 qui vont être des **éléments aléatoires** compris entre 0 et n. E_1 ne contiendra pas l'élément *i*. E_2 sera exactement le même ensemble que E_1, mais avec l'élément *i* en plus.

Dans cette solution, il y a 2 serveurs qui, chacun, connaissent l'**intégralité de la base de données** ; et c'est de là que va jaillir l'inconnu pour le serveur.

L'utilisateur envoie de manière aléatoire E_1 et E_2 à chacun de ces serveurs.

Les serveurs vont calculer le **ou exclusif**, soit **XOR**, de tous les bits d'indices contenus dans E_1 et dans E_2. Ensuite, les serveurs vont retourner cette information à l'utilisateur qui n'aura plus qu'à **calculer le ou exclusif des 2 valeurs** qui lui ont été retournées. Le serveur 1 retourne à l'utilisateur 0 ou 1 et le serveur 2 lui retourne également 0 ou 1. De cette manière, l'utilisateur va très exactement récupérer la **valeur** de ce bit d'information, grâce au calcul $k_i = E_1 \ XOR \ E_2$.

Primitives sécurisées distribuées

Considérons un **ensemble d'individus** qui possèdent des **informations** et qui souhaitent effectuer un **calcul global** à l'aide de leurs données, sans qu'aucun des individus ne connaisse la valeur des autres individus.

Cela pourrait être pour calculer la **distance totale de course** sur l'application "Running App", sans qu'aucun des utilisateurs ne sache exactement quelle distance a été courue par les autres utilisateurs.

Primitives sécurisées distribuées

Plusieurs primitives ont été définies : par exemple

- la somme sécurisée ;
- l'union ensembliste sécurisée ;
- la taille de l'ensemble intersection sécurisé ;
- le produit scalaire sécurisé.

Exemple : somme sécurisée modulo 50

Fonctionnement :

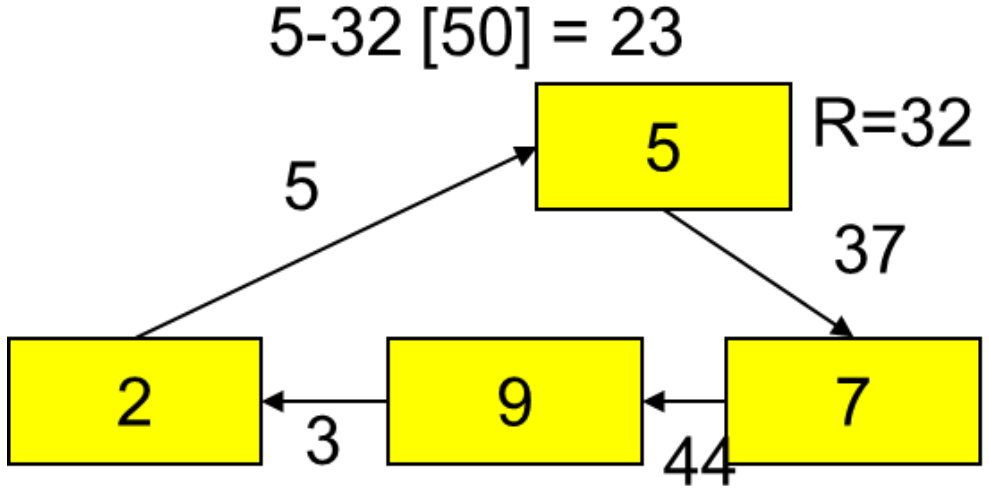
Calculons la **somme modulo 50** de la distance totale parcourue par 3 utilisateurs de l'application "Running App".

Un utilisateur a couru **5** km. Cet individu va tirer au sort aléatoirement un nombre compris entre 0 et 49 ; ici il tire la valeur **32**.

L'utilisateur ajoute sa valeur, à savoir 5, ce qui fait **37** et la transmet au prochain individu. Ce dernier rajoute sa propre valeur qui est de **7**, ce qui fait **44** et la transmet au troisième individu.

Celui-ci rajoute sa valeur qui est de **9**.

Ce résultat est transmis à l'individu suivant qui ajoute sa propre valeur de distance, soit **2**. Puis, il retourne cette valeur à l'**utilisateur de départ** qui va soustraire le nombre initial qu'il connaît pour obtenir la valeur totale de cette somme, à savoir **23**.



Conclusion

Pour conclure cette partie, retenez que plusieurs **techniques** existent pour permettre le **traitement sécurisé de données**, sans faire aucune hypothèse de confiance dans les serveurs qui traitent les données, si ce n'est qu'ils effectuent des calculs corrects.

Une des solutions (non détaillées) sur laquelle travaillent les chercheurs est la cryptographie homomorphe, qui permet de faire toutes sortes d'opérations sur des données chiffrées.

Une autre approche pour effectuer des traitement sécurisés est d'utiliser l'**anonymisation**.

Les principes généraux du RGPD :

Les données à caractère personnel doivent être:

1. traitées de manière licite, loyale et transparente au regard de la personne concernée (licéité, loyauté, transparence);
2. collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités; le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré, conformément à l'article 89, paragraphe 1, comme incompatible avec les finalités initiales (limitation des finalités);
3. adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (minimisation des données);
4. exactes et, si nécessaire, tenues à jour; toutes les mesures raisonnables doivent être prises pour que les données à caractère personnel qui sont inexactes, eu égard aux finalités pour lesquelles elles sont traitées, soient effacées ou rectifiées sans tarder (exactitude);
5. conservées sous une forme permettant l'identification des personnes concernées pendant une durée n'excédant pas celle nécessaire au regard des finalités pour lesquelles elles sont traitées; les données à caractère personnel peuvent être conservées pour des durées plus longues dans la mesure où elles seront traitées exclusivement à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques conformément à l'article 89, paragraphe 1, pour autant que soient mises en œuvre les mesures techniques et organisationnelles appropriées requises par le présent règlement afin de garantir les droits et libertés de la personne concernée (limitation de la conservation);
6. traitées de façon à garantir une sécurité appropriée des données à caractère personnel, y compris la protection contre le traitement non autorisé ou illicite et contre la perte, la destruction ou les dégâts d'origine accidentelle, à l'aide de mesures techniques ou organisationnelles appropriées (intégrité et confidentialité);