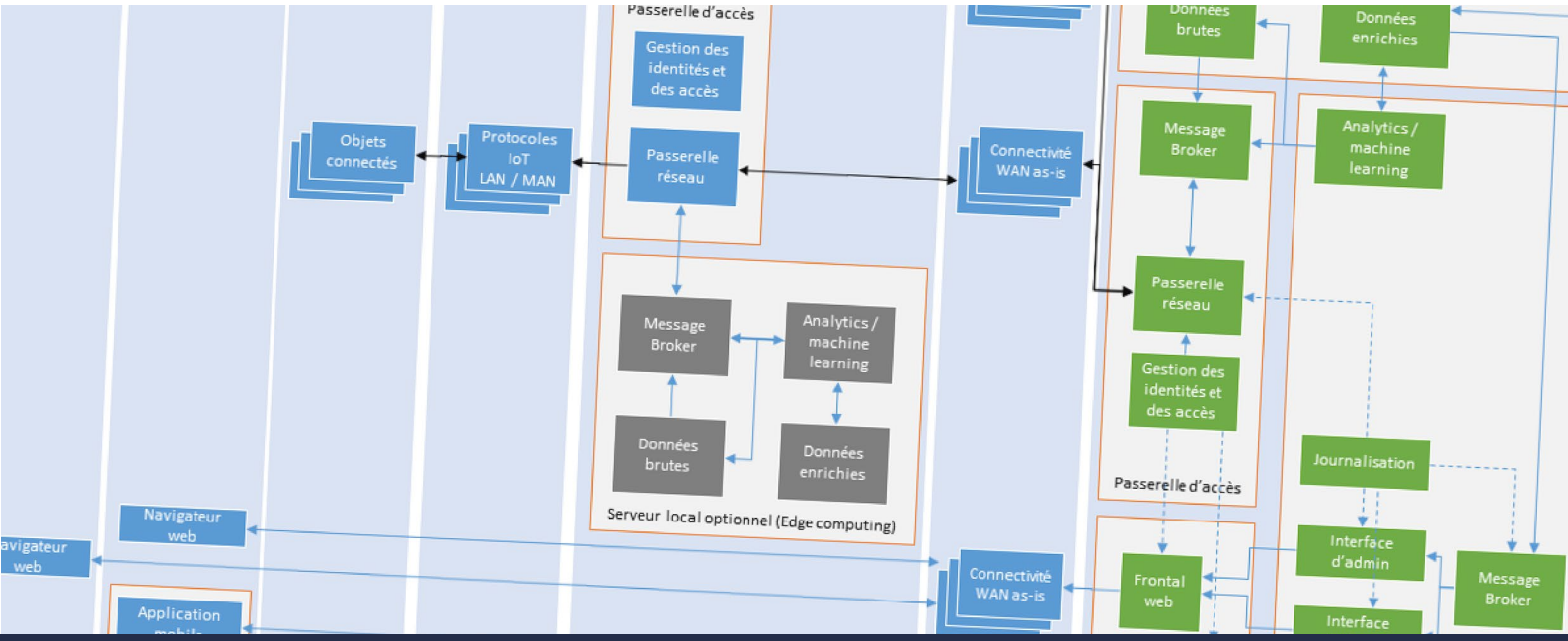
LES FICHES PRATIQUES du CLUSIF - IoT



Quelles questions se poser avec la collecte massive de données ?

1 Peut-on parler de collection massive de données ?

Lorsque l'on aborde l'Internet des Objets sous l'angle de l'acquisition de données il faut essentiellement distinguer deux méthodes de collecte : D'une part des capteurs de mesures qui produisent de faibles volumes de données à intervalle régulier et d'autre part des objets/capteurs générant des flux de données en continu. Peut-on parler d'un risque de collecte massive ?

1.1 Les flux de données discrets

Les objets de type « capteur de mesure » sont conçus pour collecter des données précises : température, vitesse, pression, luminosité, présence, etc. Ces différents capteurs collectent généralement peu de données et sur une fréquence variable qui va dépendre de l'usage. Par exemple, un constructeur d'ascenseurs utilise des capteurs connectés qui transmettent les mesures suivantes : la position de l'ascenseur ; son état : en montée, en descente, ouverture des portes ; la température et le degré d'humidité dans la cabine ou encore la masse de la cabine. Dans ce cas, le volume de données collectées quotidiennement au niveau d'une batterie d'ascenseur est relativement faible. En revanche, dès lors que ces informations sont agrégées et traitées au niveau d'un immeuble, d'une ville ou d'un pays le volume de données peut vite être qualifié de « massif ».

1.2 Les flux de données continus

D'autres types d'objets connectés capturent massivement des données en continu : des flux audio ou vidéo par exemple. L'essor des caméras IP dans l'espace public depuis le milieu des années 2000 est impressionnant. A titre d'exemple la SNCF recensait en 2015 : plus de 12.000 caméras installées en gare et plus de 23.000 caméras embarquées, dont la grosse majorité est déployée en Ile de France. Et ces chiffres sont en augmentation constante[1]. Ces objets connectés émettent quotidiennement sur leur réseau IP (sécurisé) des téraoctets de données, on peut ici aisément parler de collecte massive.

[1] SNCF Open Data - Recensement des caméras de vidéosurveillance

2 Quelles sont les caractéristiques de ces collectes massives ?

2.1 Caractéristiques des capteurs de données discrètes

La taille de chaque message échangé entre ces objets de type capteur et un concentrateur ou un serveur est généralement faible, quelques octets (voir ci-dessous un exemple de message JSON issu de l'AWS IoT Developer Guide).

```
{
    "deviceid" : "iot123",
    "temp" : 54.98,
    "humidity" : 32.43,
    "coords" : {
        "latitude" : 47.615694,
        "longitude" : -122.3359976
    }
}
```

L'effet de production massive de données est induit par soit une très grande fréquence d'émission de ces messages courts, soit par un déploiement important de ce type d'objets ou naturellement une combinaison des deux. La perte d'un de ces messages (par perte de connectivité réseau, altération des données en cours de transfert, arrêt du capteur, etc.) peut paraître insignifiante dans tout ce flux d'informations échangées. Cette absence d'information ou ce manque de données est d'autant plus complexe à détecter que la volumétrie globale est importante. En revanche, il faut être capable d'identifier ces manques de données. Car même si la plupart du temps cela correspond à une défaillance de capteur, il peut aussi s'agir d'une attaque avancée discrète qui n'attire pas l'attention.

2.2 Caractéristiques des flux vidéo

Les caractéristiques des flux vidéo vont directement impacter l'activité et la charge, donc la bande passante du réseau IP sur lequel sont connectées les différentes caméras. D'autre part les obligations légales de stockage des séquences vidéo impliquent de définir à la fois le bon dimensionnement, les autorisations d'accès appropriées ainsi que les politiques d'archivage et de rétention des données. Par exemple, le volume de données collecté sur un réseau de vidéosurveillance, est dépendant des principales caractéristiques suivantes :

- Nombre de capteurs (caméras) connectés au réseau ;
- Notion d'images par seconde : l'œil humain perçoit un mouvement à partir de 10 à 12 images affichées successivement. Plus ce nombre est élevé plus la séquence filmée est fluide mais en contrepartie : plus la masse de données transférée est élevée. Les caméras IP utilisent habituellement des taux de 15 à 60 FPS (Frame Per Second) ;
- Résolution de l'image : s'exprime en nombre de pixels qui constituent une image (largeur x hauteur) varie généralement entre 352 x 240 (CIF) à 1920 x 1080 (Full HD) et voire beaucoup plus avec les formats ultra HD 4K ou même 8K ;
- Type de compression vidéo : Algorithme de compression qui permet de réduire le volume des données vidéos échangées. On parle de formats d'encodage de type Motion JPEG, MPEG-4, H.264 et récemment H.265[2]
- Complexité de la scène : les conditions d'éclairage et la quantité de mouvement (ciel bleu ou paysage urbain, foule, etc.) influent sur le volume de données transféré ;
- Durée exigée pour la conservation des données.

Si les « données » produites par ce type de capteur sont habituellement destinées à être visualisées par des opérateurs (humains), on trouve de plus en plus de solutions logicielles qui visent à exploiter ces données pour des applications de sécurité notamment. On parle alors de solutions de vidéo « intelligence » et les données (métadonnées) exploitées permettent de détecter des visages, des objets ainsi que leur mouvement (trajectoire). L'usage de ces données doit répondre à un encadrement légal très strict.

3 Quels sont les principaux risques associés ?

Les objets connectés n'ont pas généralement vocation à stocker localement l'information, la donnée va être traitée, agrégée et véhiculée sur un réseau. Plus la masse de données traitée est grande plus les manipulations, interprétations auront un impact important sur le fonctionnement global du système d'objets.

3.1 La nécessité de nettoyer ou transformer les données

Un réseau de capteurs peut être constitué d'appareils de différents modèles provenant de différents constructeurs, n'ayant pas tous les mêmes caractéristiques techniques. Il peut être nécessaire de préparer la donnée, la convertir ou éventuellement la nettoyer avant de la communiquer. Dans un bâtiment « connecté » des capteurs de température peuvent fournir une donnée en degrés Celsius ou en degré Fahrenheit par exemple, la donnée doit être normalisée avant d'être utilisée. De même la précision mathématique peut varier selon les équipements. Il faut également prendre en compte l'absence de mesure ou la production de mesures aberrantes. Par exemple, comment faut-il considérer les valeurs « 0 », « -1 » ou l'absence de valeur dans ce contexte ? Il est donc très important de s'assurer que chaque donnée produite est utilisée de manière

[2] International Telecommunication Union - Standard de compression vidéo H.265 - REC-H.265

consistante sur tout le réseau d'objets, depuis le capteur jusqu'au consommateur de l'information.

3.2 Risques de mauvaise interprétation des résultats

Ce risque n'est pas directement lié à une utilisation malveillante des objets connectés, mais peut conduire à provoquer un comportement défaillant du système. Une application qui utilise des données issues de multiples capteurs aura probablement besoin, dans le cadre de ses traitements, d'effectuer des calculs de somme, moyenne, écart type, variance par exemple. Il est donc important de vérifier que les valeurs retournées par chaque capteur répondent à des caractéristiques précises comme l'unité de mesure et la précision mathématique. Il est également important de s'assurer que la valeur obtenue est comprise dans une plage de valeurs acceptables et de définir comment traiter les écarts ou les valeurs aberrantes. Par exemple : que signifie une valeur de 80°C captée dans une cage d'ascenseur : début d'incendie ou capteur défectueux ? Autre exemple : Un bâtiment connecté reçoit les valeurs de température d'une cinquantaine de capteurs, tous indiquent une température ambiante autour de +20°C mais un seul retourne une valeur de -800°C, la température moyenne calculée se situe alors autour de +4°C. Faut-il envoyer un ordre à la chaudière pour remonter globalement la température ? Si des mesures de contrôle ne sont pas mise en place, on peut imaginer que la prise de contrôle d'un ou plusieurs objets du réseau pourrait conduire à modifier le comportement global du système. Dans le cas d'accumulation massive d'information, les algorithmes d'apprentissage automatique (Machine Learning - ML) peuvent extraire du système un schéma de comportement « normal » et faire ressortir toutes les déviations atypiques. Les développements de l'Intelligence Artificielle (IA) dans ce domaine peuvent participer à détecter les anomalies et même anticiper des événements.

3.3 Le détournement d'usage par agrégation

Les multiples capteurs d'un réseau peuvent émettre leurs données sur des fréquences variables et dans des volumes extrêmement variés. Selon l'architecture du réseau les protocoles de communication utilisés et finalement les cas d'usage, il peut être intéressant d'agréger toutes ces données sur des passerelles intermédiaires et localisées avant de les transmettre vers les applications consommatrices. Les données provenant de multiples capteurs et envoyées à une application peuvent être consolidées pour apporter une vue globale, soit au niveau d'un système connecté : un véhicule, un logement, un bâtiment, ou plus largement d'une zone géographique par exemple. Le fait d'agréger les données peut apporter une information essentielle comme un état, un statut comme par exemple : bâtiment fermé (température globale inférieure à une moyenne et détecteurs ne signalant aucune présence dans le bâtiment) ou encore zone de circulation encombrée (multiples capteurs dans la même zone, mais position des véhicules au ralenti). Si l'intérêt de l'information consolidée ne fait pas de doute, la mise à disposition de cette donnée peut conduire à des détournements d'usage. Début 2018, un chercheur a révélé sur Twitter que la carte créée par le système de géolocalisation d'une l'application de Fitness développée par Strava-Labs mettait en évidence des itinéraires empruntés par des soldats à proximité de leur base au cours de leur entrainement. Cette information éminemment sensible aurait pu être utilisée pour organiser des attaques[3]. Lors du développement d'applications impliquant des réseaux d'objets connectés, il est important d'essayer d'avoir une vue la plus globale possible. Cela permet d'anticiper des usages non désirés en simulant le traitement d'échantillons significatifs de données.

Un autre exemple avec l'application de Fitness Polar, montre qu'à partir du tracé des entrainements de course on pouvait identifier plus de 6 400 utilisateurs s'entrainant près de lieux sensibles tels que la NSA, la Maison Blanche, le MI6 à Londres, la DGSE à Paris ou encore le GRU à Moscou[4].

Si un objet publie les données qu'il collecte, cela représente un risque pour le respect de la vie privée. Il faut s'assurer d'un consentement explicite et compris. Dans le cas d'une collecte massive, le risque se trouve accru.

3.4 Vol de données

Le traitement massif des données diffusées par tous ces capteurs connectés est de plus en plus contrôlé et encadré. Depuis Mai 2018, la réglementation Européenne RGPD fixe un cadre strict à l'utilisation des données à caractère personnel. Il n'en reste pas moins vrai que ces données sont toujours collectées et stockées et que d'autres types de données non visées par ces règlements peuvent aussi se révéler « sensibles » pour une entreprise. L'accès frauduleux à ces informations peut conduire à des actes malveillants. L'ENISA (European Union Agency for Network and Information Security) publie un recueil de bonnes pratiques de l'IoT dans le contexte de l'industrie 4.0 « Smart Manufacturing »[5], mais qui s'avèrent utiles globalement dans le monde des objets connectés. En particulier, le chapitre 4 décrit un certain nombre de mesures de sécurité et de bonnes pratiques qu'il est élémentaire de suivre pour se protéger d'un accès incontrôlé à un grand nombre de données à considérer comme « sensibles ». Comme par

[3] France Info - Strava-Labs, une application de fitness dévoile l'emplacement de sites militaires

[4] ZDNet L'application Polar expose la localisation de personnel militaire

[5] ENISA - Good Practices for Security of Internet of Things in the context of Smart Manufacturing

exemple la recommandation TM-26 qui préconise d'anonymiser et de sécuriser toutes les données personnelles directes ou indirectes traitées au sein de l'entreprise, par exemple par le contrôle d'accès et le chiffrement basés sur les rôles, après avoir pris en compte toutes les exigences légales pertinentes.

ⓘ Point de vue du RSSI

Un grand nombre de mesures déjà utilisées par les DSI pour contenir les risques associés aux traitements des données peuvent être appliqués au monde de l'Internet des Objets. De même les mesures de sécurité préconisées dans le monde de l'Industrie 4.0 tendent à rejoindre toutes ces recommandations. Il est important de s'inspirer des bonnes pratiques de base et de se concentrer sur les risques spécifiques liés à l'activité qui s'appuie sur des objets connectés. Pour finalement s'assurer qu'une vue globale et consolidée des données n'expose pas à un usage détourné.