Lettre mensuelle – Octobre 2019 - disponible sur omc.ceis.eu

Table des matières

ANALYSES

1. L'INNOVATION EN DEFENSE CYBER : LE MODELE AMERICAIN

Rapidité des progrès technologiques, évolution permanente des menaces, urgence et diversité des besoins opérationnels : s'il y a un domaine dans lequel les cycles d'innovation et de « Time to Market » doivent être accélérés, c'est bien la cybersécurité. Un constat qui a poussé le ministère des Armées à inaugurer à Rennes le 3 octobre 2019 la **Cyberdéfense Factory**, avec l'objectif de stimuler le développement de solutions de cybersécurité innovantes au profit du Commandement de la Cyberdéfense.

Au cœur de cette « fabrique digitale » pilotée par la Direction générale de l'Armement (DGA), un « **Data Lake** » permettant aux différents acteurs impliqués (ministères, ESN, startups, centres de recherche…) de partager des « données d'intérêt cyber ». Un modèle déjà testé depuis quelques mois aux États-Unis avec **DreamPort**[1], le « *Mission Accelerator* » lancé par l'**US Cyber Command** en mai 2018, en partenariat avec le **Maryland Innovation & Security Institute (MISI)**. On peut également citer le modèle britannique qui repose notamment sur deux *Cyber Innovation Centers,* ouverts en 2017 2018 au sein du *Cyber Accelerator Programme du GCFQ* dans le but de faciliter la collaboration entre les gros acteurs industriels, les start-ups et des experts du secteur pour favoriser l'innovation en cybersécurité. En Israël, la création de toutes pièces, au milieu du désert à Beer Sheva, d'une ville entière dédiée à la cybersécurité, procède encore d'un autre modèle.

Cette diversité d'initiatives de natures, d'objectifs et de maturité différentes, reflète la prolifération des projets et structures dédiés à l'innovation de cyberdéfense. L'exemple américain, le plus abouti, a à ce titre servi de modèle à maintes reprises.

DreamPort, bras armé de l'US Cyber Command

Installé à Columbia (Maryland), DreamPort est la version « cyber » de SoftWerx[1], partenariat public-privé du Special Operations Command américain ouvert en 2016. Avec une superficie de 1 500 m², et de 3 000 m² depuis juillet 2019, le centre est d'abord un lieu physique constitué de bureaux, de salles de conférence, de salles de réunions et de plusieurs labs. C'est également **un environnement technique permettant aux entreprises de présenter, de tester et d'intégrer rapidement des prototypes et des solutions**.

L'US Army lance « The Forge »

L'US Army a de son côté lancé en 2018 « The Forge »[1] pour accélérer le développement et l'acquisition de solutions de lutte informatique défensive, ainsi que l'entrainement. Basé à Fort Belvoir (Virginie), l'organisation est dotée à la fois d'une infrastructure physique et virtuelle. Elle a notamment travaillé à la miniaturisation d'un « Defense Cyber Operations Kit », sorte de boîte à outils composée de différentes solutions du secteur privé.

[1] https://dreamport.tech/index.php

Animer la filière et soutenir le développement de l'écosystème

Parmi les missions de DreamPort : **animer la filière et favoriser le développement de l'écosystème**. C'est dans ce contexte que le *« Mission Accelerator »* organise deux fois par an des événements dédiés à la filière cybersécurité. À côté de séminaires de sensibilisation, il accueille surtout la **AvengerCon**, événement de référence de la communauté cyber militaire américaine, dont la 4ème édition a eu lieu les 17 et 18 octobre 2019.

Les Cyber Innovation Centers britanniques

Conçus comme des lieux « inclusifs » et complémentaires, ils comprennent un accélérateur, des espaces dédiés à l'accompagnement (« mentoring », recherche d'investissement…) et doivent permettre aux innovateurs de développer des solutions à des problèmes existants sur la base de besoins formulés, tant par le secteur privé que des différents organes gouvernementaux. Le centre de Cheltenham a pour mission de mettre en contact les expertises du Government Communications Headquarters (GCHQ) et de l'industrie pour répondre aux déficits immédiats du Royaume-Uni en matière de capacités de cybersécurité. La particularité de celui de Londres tient dans la collaboration aussi avec les structures académiques et de recherche, notamment avec les Academic Centers of Excellence et d'autres programmes d'accélération.

L'innovation par l'expérimentation

Le dispositif permet également l'organisation régulière de **Rapid Prototyping Events (RPE),** des hackathons de 3 à 4 jours sur des « défis » lancés par l'US Cyber Command. Les règles[2] sont simples : pas d'information classifiée ; les participants sont libres d'utiliser toutes les technologies qu'ils souhaitent, y compris des technologies propriétaires, dès lors qu'ils ne violent pas les accords de confidentialité (NDA) et dispositions en vigueur dans leurs entreprises. DreamPort se réserve en revanche la possibilité de communiquer les outils et méthodes utilisés en « open source » par la suite. Il s'agit par ailleurs de compétitions de technologies et non pas de « Capture the Flag ».

Deux exemples de RPE organisés par DreamPort depuis sa création :

- Le challenge « **The Chameleon & the Snake** »[3], organisé en septembre 2018 et mai 2019, sur le thème de la diversité de signature des malwares et de leur évaluation dans Microsoft Windows. Le challenge comprenait à la fois un volet « attaquant » qui consistait à créer un outil permettant d'altérer la signature d'un outil sans changer ses fonctionnalités et un volet « défenseur » dont l'objectif était au contraire de développer un outil permettant de détecter automatiquement un exécutable inconnu dans Windows. La compétition a été gagnée en septembre 2018 par les sociétés Northrop Grumman

[2] https://dreamport.tech/RPE005/DreamPort-RPE-005-Introduction-02.pdf

[3] https://dreamport.tech/event-rpe-001-the-chameleon-and-the-snake.php

pour la partie « défense »[4] et Draper pour le volet offensif, Crowdstrike obtenant également une mention « honorable » en défense[5].

- Le challenge « **The Wolf in Sheep's Clothing** »[6] organisé en janvier et février 2019, dont l'objectif était de concevoir des solutions User Activity Monitoring (UAM) pour détecter en temps réel des attaques ou activités non autorisées. Au cœur du problème : le développement de capacités d'analyses prédictives sans configuration préalable et non basées uniquement sur un moteur de règles. Vainqueurs : Jazz Networks[7], IBM, Booz Allen et LogicHub. Les participants bénéficiaient pour cela d'un environnement Windows et Linux mis en place par DreamPort.

BeerSheva, capitale de la cybersécurité

Modèle d'intégration des compétences et capacités des différents acteurs de la cybersécurité, la ville Beer Sheva accueille désormais le CyberSpark, le hub israélien de l'innovation en cybersécurité. Joint-Venture entre le National Cyber Bureau du cabinet du Premier ministre, l'Université Ben Gourion et les principaux industriels du secteur, le lieu rassemble aujourd'hui des start-ups, un accélérateur du fonds d'investissement *Jerusalem Venture Partners* (l'un des plus actifs en matière de cybersécurité), ainsi qu'une vingtaine de centres de R&D de géants comme IBM, Lockheed Martin ou Oracle, mais aussi des bases de l'armée israélienne spécialisées en cybersécurité, notamment l'unité 8200 et le C4I.

Tous ces acteurs sont rassemblés autour de :

- Un centre de recherche ;
- Un R&D hub permettant aux entreprises participantes d'accéder à des appels à projets et des financements ;
- Un centre d'entraînement à la cyberdéfense ;
- Un innovation hub qui facilite l'exposition aux technologies israéliennes les plus avancées ;
- Un incubateur soutenu par l'Agence d'innovation israélienne ;
- Un Intelligence Center par lequel le CERT et les entreprises concernées mettent à disposition des autres acteurs des données de CTI.

Orienter le développement capacitaire

Au-delà des Rapid Prototyping Event, DreamPort a aussi pour objectif d'orienter sur le moyen terme les travaux de R&D du secteur privé autour des grandes priorités, déterminées dans **une liste de challenges techniques définis sous l'égide du J9 de l'US Cyber Command**. Le document publié en mars 2019[8]

[4] https://news.northropgrumman.com/news/releases/northrop-grumman-cybersecurity-team-wins-the-dreamport-rapid-prototyping-competition

[5] https://www.fbcinc.com/e/cyberusa/presentations/DadeScottUnitedStatesCyberCommand.pdf

[6] https://dreamport.tech/event-rpe-003-the-wolf-in-sheeps-clothing.php

[7] https://www.jazznetworks.com/blog/dreamport-winner/

[8] https://www.cybercom.mil/Portals/56/Documents/Technical%20Outreach/Technical%20Challenge%20Problems.pdf?ver=2019-07-02-151118-497

distingue les capacités fonctionnelles (la recherche de vulnérabilités et l'analyse de *malware*) et les « Enabling Technologies », comme le *Machine learning*.

Titre	*Objectifs*
1. Vulnérabilités	
Recherche automatique de vulnérabilités (Magetower)	Développer des capacités de recherche automatique de vulnérabilités sur des fichiers binaires. Le challenge s'inscrit dans le cadre du projet VOLTRON, suite du DARPA Cyber Grand Challenge (2016), qui implique notamment la société ForAllSecure.
Réduction des temps de correction de vulnérabilités	Accélérer et optimiser le processus d'évaluation, de test et de déploiement de patch correctifs grâce à une méthodologie basée sur la nouvelle version des CVEs
Exploitabilité des vulnérabilités	Élaborer de nouvelles méthodes d'analyse permettant d'analyser les logiciels et leurs fonctionnalités pour identifier des vulnérabilités. L'exploitabilité des vulnérabilités doit ensuite être analysée et prouvée.
Vulnérabilités SCADAs	Prouver l'exploitabilité d'une vulnérabilité SCADA.
Recherche API	Utiliser les sources ouvertes pour exploiter les API internet publiquement accessibles.
2. Malwares	
Environnement de Reverse Engineering	Concevoir des environnements de test, y compris off-line, pour analyser des malwares.
Catégorisation de malware	Examiner l'état de l'art en matière de catégorisation rapide de malware, de Reverse Engineering, de corrélation et d'obfuscation.
Lutte contre les malwares polymorphes	Trouver des techniques permettant de détecter les malwares polymorphes et contre le fuzzing et les modifications de signatures.
Analyses des tendances	Conduire des recherches open-source pour identifier les dernières tendances et techniques en matière de malware.
Publication de données défensives	Examiner les possibilités d'utiliser des données classifiées dans des capteurs non classifiés. Publication anonyme de binaires malveillants et de trafic réseau auprès de la communauté open source.
3. Analytics	
Exploration de réseaux	Concevoir des outils permettant d'analyser en permanence des réseaux complexes incluant des équipements, des logiciels mais aussi des systèmes de Command & Control, des flux de données, des protocoles, etc.
Modélisation prédictive	Générer automatiquement des graphes d'attaques permettant de modéliser et de tester les défenses.

Intégration de données et automatisation	Développer des méthodologies permettant de capter et d'analyser des jeux de données en vue de développer l'automatisation de certaines tâches.
Analyse comportementale	Concevoir des techniques permettant de spécifier l'état normal d'un réseau ou d'un système pour ensuite identifier d'éventuels comportements déviants.
Machine learning défensif	Concevoir des modèles de Machine learning défensifs pour caractériser et détecter des infections par des malwares inconnus.
Comparaison des plateformes d'analyse de données	Développer les coopérations avec d'autres agences gouvernementales, l'industrie et le monde académique pour construire des jeux de données et des méthodes d'analyse permettant d'analyser les plateformes d'analyse, les modèles, les techniques.
4. Implant	
Intelligence artificielle pour l'analyse de vulnérabilités et la détection de menaces	Développer l'utilisation de l'intelligence artificielle pour automatiser l'analyse de vulnérabilités et la détection de menaces
Redirection de trafic et obfuscation	Développement de techniques permettant de rediriger/obfusquer un trafic réseau précis sans qu'un adversaire ne détecte la modification
Objets connectés	Conduire des recherches open source et du reverse engineering sur des objets connectés grand public.
Persistance de code/fonctionnalité	Développer des solutions permettant de maintenir un code ou une fonctionnalité au sein d'un réseau quand différentes actions ont été faites pour désactiver le code ou la fonctionnalités (reboot du système par exemple).
Diversité / survivabilité	Développement de nouvelles méthodes pour renforcer la résilience. L'utilisation de logiciels et de librairies communs, voire d'infrastructures partagées, augmente les risques de vulnérabilité. Parmi les réponses, les stratégies de « Moving Target Defense ».
Environnements de tests	Conception d'environnements de simulation permettant de tester des solutions en dehors d'environnement sécurisés.
5. Situational awareness	
Malware situational awareness	Développer des solutions permettant de suivre les malwares, d'établir les corrélations et de comprendre les écosystèmes pour faciliter l'attribution des opérations
Visualisation	Concevoir des représentations holistiques combinant les données de trafic réseau et les métadonnées.
6. Développement de capacités	
Prototypage rapide	Concevoir des environnements collaboratifs permettant de tester et d'expérimenter des solutions et de développer une communauté unifiée pour les niveaux de classification les plus bas.
7. Personnes	

Identification d'avatars	Détecter l'utilisation d'avatars et d'informations fausses lors des enregistrements.
Détection de fausses représentations	Détecter l'utilisation de techniques permettant d'éviter l'identification et la détection.
Collecte et anonymisation de données réelles	Collecter et exploiter des données de trafic réseau réelles pour les rejouer dans des environnements simulés
8. Chasse	
Exercice de « chasse »	Identifier les meilleures pratiques, TTPs et données nécessaires pour conduire des opérations de « chasse » défensives.
9. Mission management	
Automatisation des analyses de risque	Automatiser les analyses de risques lors de la conception d'une opération et leur suivi.
Partage de données	Collecter, stocker et transférer des données à l'échelle peta ou exa.
Installation automatique	Changer, mettre à jour ou installer un logiciel de façon automatique.
10. Attaque	
Puissance de calcul	Identifier des façons innovantes de mettre à disposition du DoD des capacités de calcul permettant des attaques en brute force ou des calculs massivement parallèles sans entraver les besoins opérationnels.
11. Sécurité	
Nouvelles approches de défense	Faire évoluer les méthodes traditionnelles de défense, notamment avec des approches de type « Zero Trust ».
Meilleures pratiques pour des processeurs, OS, protocoles, API plus sûrs et résilients	Identifier les outils, techniques, méthodes permettant d'intégrer la sécurité de façon native à chaque étape de conception et développement.
Protection des éléments les plus critiques	Protéger les éléments clés d'un réseau ou d'un système, en prenant acte du fait qu'il est impossible de protéger l'ensemble de son terrain.
User activity monitoring	Détecter les attaques internes et activités non autorisées grâce à l'analyse comportementale.
12. Blockchain	
Identification des adversaires	Identifier des adversaires et attribuer des opérations en exploitant et analysant la blockchain ou les crypto-monnaies qu'elle supporte.
Crypto-monnaie défensive	Développer et expérimenter des prototypes pour contrer l'utilisation par l'adversaire de crypto-monnaies et de minage.

À noter que la Defense Information Systems Agency (DISA) lance également régulièrement des appels à contribution[9].

Les « Nodes » australiens

AustCyber, le dispositif australien dédié au développement du secteur de la cybersécurité, a déployé sur l'ensemble du territoire des Nodes « nœuds » un réseau d'innovation en cybersécurité. Dans le cadre d'un dispositif et d'une feuille de route nationale, les Nodes mettent en contact les écosystèmes locaux dans des espaces qui rassemblent les gouvernements, le milieu universitaire et les centres de recherche.

Le nœud du New South Wales organise dans ce cadre les *Industry Discovery Days*, imaginés pour mettre en contact le milieu du R&D avec des industriels ayant exprimés des besoins spécifiques, et pour leur permettre de réfléchir ensemble à une réponse collaborative. Chaque événement est conçu sur mesure pour répondre aux besoins des industriels participants et fait appel à des chercheurs sélectionnés par le Node pour leur expertise. Leur objectif est non seulement d'apporter une réponse concrète à des déficits ou besoins en matière de cybersécurité mais aussi, de façon plus générale, de favoriser le développement d'un écosystème collaboratif.

La gouvernance des données, nouvel enjeu de l'innovation de défense cyber.

L'innovation en matière de cyberdéfense, notamment en matière d'intelligence artificielle, se heurte cependant à la profusion et à l'éclatement des données disponibles. Le général de corps armée Jack Shanahan, chef du Joint Artificial Intelligence Center (JAIC), constate ainsi que le Department of Defense (DoD) a 24 fournisseurs de données cyber qui utilisent chacun des formats différents[10]. 80% du travail consiste à labelliser et préparer les données, souligne le général Shanahan. C'est la raison pour laquelle le JAIC, en partenariat avec la National Security Agency (NSA) et l'US Cyber Command, travaille à l'élaboration d'un « **Data Governance Framework** » ou schéma de gouvernance des données cyber sur le modèle de ce qui a déjà été fait sur le projet Maven (exploitation de l'intelligence artificielle en matière d'analyse vidéo).

De la profusion à la confusion ?

Ce schéma est d'autant plus nécessaire que **le nombre de hubs dédiés à l'innovation au sein du DoD a explosé ces dernières années** : 25 consortiums de type PIA (Partnership Intermediary Agreement - support contractuel utilisé notamment par DreamPort pour recourir aux services d'un acteur privé pour l'animation d'un hub), 16 accords de partenariat public-privé, plus de 100 agences dédiées à l'innovation, etc. Sur ce total**, 20 seraient dédiées à des thématiques « cyber »**[11].

[9] https://dreamport.tech/call-for-white-papers.php

[10] https://www.defenseone.com/technology/2019/09/pentagon-nsa-laying-groundwork-ai-powered-cyber-defenses/159650/

[11] https://www.govexec.com/technology/2019/09/so-many-innovation-hubs-so-hard-find-them/159798/

2. L'OPEN DATA AU SERVICE DE L'INNOVATION : MODE D'EMPLOI

L'**Open Data**, concept qui a émergé et s'est développé au début des années 2000, doit permettre de **rendre accessible à tous, gratuitement, des volumes de données considérables collectées ou générées par les services publics ou les entreprises privées d'utilité publique**. Un modèle innovant tant par le dispositif technique et organisationnel qu'il suppose, que (ou surtout) par l'ambition sur laquelle il repose : ouvrir, partager et échanger des données collectées mais rarement ou insuffisamment valorisées. On comprend aisément l'apport que représente l'Open Data pour l'innovation : non seulement il donne accès à des données de nature et de contenus d'une grande variété et d'une extrême richesse, mais il permet aussi à des acteurs d'écosystèmes n'ayant a priori pas grand-chose en commun d'échanger et de travailler conjointement à des solutions innovantes, en croisant les expertises, les compétences, les perspectives, les habitudes de travail, les ressources…

La Défense n'est pas en reste avec le lancement début 2019 par la Direction générale du numérique et des systèmes d'information et de communication (DGNUM) de la Fabrique Numérique et plus récemment, par le Commandement de la cyberdéfense et de la DGA, de la **Cyberdéfense Factory**. Au centre de ce dispositif, un **Data Lake** (« lac de données »), espace de stockage et gestion de Big Data, qui permettra de **centraliser et fluidifier l'accès et le partage de données brutes**. Côté privé, les initiatives fleurissent également. On peut par exemple citer le **Data Shaker de la SNCF (2014)**, mené au « catalyseur d'innovations » NUMA à Paris sur le thème du Big Data, avec des startups travaillant sur des sujets comme l'information aux voyageurs ou l'optimisation des trajets, ou encore **DataPoste (2013)**, événement d'une journée destiné à inciter des startups, des développeurs et quelques entreprises partenaires à imaginer des services innovants à partir de données postales.

L'appel à innovation du Defence and Security Accelerator (DASA) britannique

Partant du constat que les équipements de défense génèrent un nombre considérable de jeux de données dont la collecte et l'analyse représentent un atout considérable (facilitation de la prise de décision, optimisation du MCO, élaboration de scénarios, simulation d'environnements…), le DASA a ouvert un nouvel Innovation Focus Area (IFA) consacré à la Défense, d'une capacité de 100 00 livres. Il a par la suite lancé en juin 2019 un appel à innovation pour des projets d'Open Data permettant d'analyser et partager rapidement des données structurées et non structurées issues de sources multiples tout en conservant les niveaux de classifications de sécurité.

Mais l'ouverture et le partage de données n'est pas chose facile, ni anodine. **Ouvrir et partager des données pose un certain nombre de questions, d'ordre à la fois organisationnel** (quelle gouvernance ? quelle infrastructure ? quelles données méritent d'être ouvertes ?), **juridique** (comment s'assurer par exemple que les modalités d'ouverture et de partage respectent la législation liée à la protection des données personnelles ?) et **pratique** (quel vecteur et quel format de diffusion ?). Une série de contraintes que l'émergence de « schémas de gouvernance » permet de recenser et visualiser pour accompagner les initiatives d'ouverture des données.

De tels schémas, comme celui, ci-dessous établi par la société DLT, doivent prendre en compte l'ensemble du processus, depuis l'identification des données disponibles, leur intégration dans un « Data Lake » et leur exploitation grâce à différentes briques : les technologies de stockage, le « **Master Data Management** » et

les fonctionnalités associées (catalogue, gestion de la qualité, accès…), les outils de reporting, les « Data Sciences » et la gouvernance des données (protection de la vie privée, sécurité, conformité).

Source : https://www.dlt.com/file/big-data-frameworkpng

Ces schémas permettent de passer en revue, en amont, toutes les questions qui peuvent se poser en vue de la mise en place d'un projet d'Open Data.

L'exemple du datahub HyperThought[12]

*En partenariat avec l'entreprise privée MarkLogic, l'Air Force Research Laboratory (AFRL) a notamment développé un Data Hub opérationnel nommé « **HyperThought ».** Cette plateforme collaborative opérée au profit de l'US Air Force, notamment dans le cadre de la conception d'avions et d'autre outils, est déployée à la fois dans les milieux académique et industriels dans le but de **désenclaver l'information[13].***

On peut classer les enjeux liés à la mise en place d'une démarche Open Data en quatre catégories : la cartographie des données, l'organisation du dispositif, la sécurité, et la gouvernance.

1. Sélection et cartographie des données disponibles

La première étape de la mise en place d'un Data Lake consiste à dresser un état des lieux des données disponibles, à la fois en interne et sur les plateformes d'Open Data existantes. Il s'agit donc d'abord de déterminer quelles sont les données à partager, et ce en fonction de l'objectif recherché. L'organisation initiatrice doit répondre à la question de la pertinence des données qui doivent être partagées. Par exemple dans le cas de la Cyberdéfense Factory, le ministère des Armées devra répondre à la question : **que sont les données de cyberdéfense ?** Dans leur sens le plus large, celles-ci peuvent être définies comme les

[12] « U.S. Air Force Research Lab Creates MarkLogic-Based Data Platform to Make Safer and More Powerful Planes, Tools, and More », MarkLogic [en ligne].

[13] Holly Jordan, « HyperThought poised to break down barriers in information sharing », Wright Patterson AFB [en ligne], 20 mars 2017.

données collectées et exploitées par un État pour le développement de mesures techniques et non techniques visant à la défense dans le cyberespace de ses systèmes d'information jugés d'importance vitale. Il s'agit donc de données de natures très variées, issues de capteurs et de sources diverses, produites par différents services.

La nature et le format des données partagées doit également faire l'objet d'une attention particulière. On considère généralement que pour une exploitation et une réutilisation efficace des données, le dispositif doit contenir :

- Les données sources/brutes,
- Les sources/brutes éventuellement retraitées par des applications techniques,
- L'historisation des données,
- Les métadonnées.

C'est ce dispositif qui permet à la fois de donner confiance aux utilisateurs dans la donnée qu'ils utilisent, mais aussi de contextualiser, donner du sens, et établir des liens et des relations entre toutes ces données. Une fois les données à partager identifiées, il convient ensuite de déterminer à la fois l'endroit (ou les endroits) où elles sont stockées, sous quel format elles sont actuellement stockées, et comment elles peuvent être récupérées, afin de pouvoir être facilement intégrées au dispositif.

A cet effet, le Data Lineage, ou référentiel des transformations, a pour objectif de cartographier le système d'information pour visualiser le cycle de vie d'une donnée et de comprendre ainsi de quelle source elle provient et quelles transformations elle a éventuellement subie.

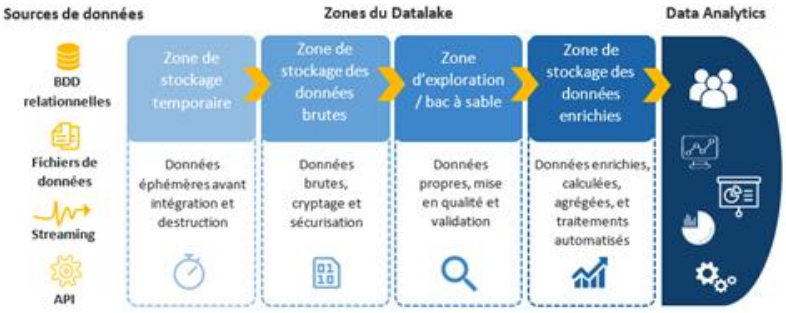
Le Data Science Campus

L'Office National de Statistiques britannique a lancé en 2018 à Newport un nouveau Campus dédié à la science des données. Dans le cadre d'une démarche de modernisation de l'utilisation des statistiques, le Campus est destiné, dans un premier temps, à la collecte de données de natures diverses issues à la fois de détecteurs et radars routiers, des équipements de téléphonie mobile et des satellites, dans le but de mesurer l'activité de l'économie britannique. Les données, une fois collectées, seront mises à profit de projets d'utilité publique ou sociale, conçus en collaboration avec les universités, l'industrie, le gouvernement et la société civile.

2. Organisation de l'espace de partage

Les données collectées et stockées dans un même espace peuvent être structurées, semi-structurées ou non structurées. Il faut donc trouver un moyen de les faire cohabiter pour les rendre exploitables, et dans le cas d'un Data Lake d'éviter que le gisement ne se transforme en marécage… Cette étape qui consiste à séparer à la fois logiquement et physiquement les données est indispensable pour préserver l'intégrité, la sécurité et l'organisation d'un Data Lake, et par conséquent pour faciliter la gestion des données et sécuriser les informations. On distingue généralement 4 zones :

- une zone de stockage temporaire,
- une zone de stockage des données brutes,
- une zone « bac à sable » dédiée à l'exploration
- une zone de stockage des données enrichies utilisables par les applications métiers.



Les différentes zones de stockage d'un Datalake

Source : https://www.solutions-numeriques.com/expertise-mc2i-gouvernance-dun-datalake-ne-laissez-rien-au-hasard/

Certains opposent à ce système l'inconvénient de stocker des informations redondantes, mais il a toutefois pour avantage de laisser la possibilité d'appliquer à ces différentes zones des mesures de sécurité et une gouvernance différenciée.

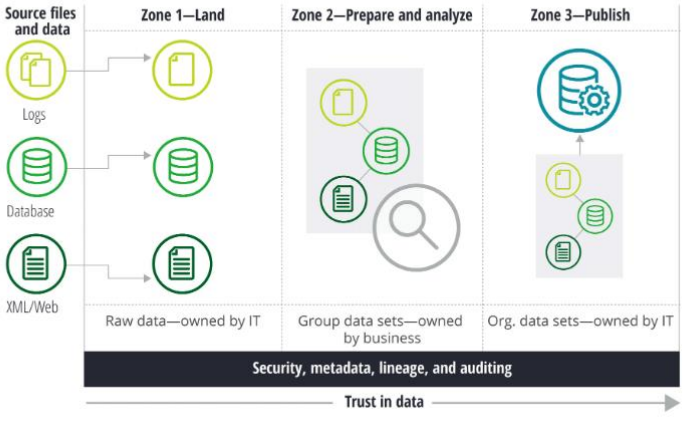
3. Sécurisation des jeux de données

- **Gestion des accès**

La question des accès aux données partagées est en effet l'un des enjeux clés des modèles d'Open Data. Au-delà d'une mesure de sécurité, la gestion des accès permet aussi à chaque participant de n'avoir accès qu'à l'information dont il a besoin, et donc d'améliorer son « expérience utilisateur ». Dans le cadre d'un Data Lake par exemple, il est possible de contrôler les accès de façon très précise, du fichier à la feuille de calcul et même à la ligne ou la colonne. La mise en place de restrictions sur les différentes zones du Data Lake est également indispensable pour minimiser les risques dans la manipulation ou l'extraction de données. Ainsi, l'accès aux zones de stockage temporaire et de stockage des données brutes doit être strictement limitée. Plus les données sont traitées, transformées et sécurisées, plus l'accès peut être élargi. Deloitte propose le modèle suivant :

FIGURE 1

Creating data "zones" with different verification requirements can enhance analytical reliability and accuracy



Source: Deloitte analysis.

- **Protection des données**

  - **Open Data et RGPD**

Intuitivement, le principe même du Data Lake et de l'Open Data semblent contradictoires avec les obligations en termes de protection des données prévues par RGPD. D'abord avec la notion de « minimisation », qui prévoit que le responsable d'un traitement ne doit traiter que des informations nécessaires à la finalité qu'il poursuit. Ensuite avec les dispositions relatives à la conservation des données, puisque celles-ci doivent être supprimées dès que l'objectif pour lequel ils sont collectés a été atteint. L'Open Data, au contraire, suppose la collecte et le stockage permanant et continu de volumes de données considérable sans objectif pré-défini, et sans même savoir si elles seront réutilisées. Et ce d'autant qu'il mêle dans un même espace des données sensibles et données obsolètes. Et pourtant, rassembler dans un lieu unique une série de données d'intérêt permet aussi de les protéger, de les gouverner et les auditer de façon sécurisée. Sans compter qu'il est aussi plus facile pour un individu de demander l'application de droits tels que le droit d'opposition au profilage, à l'oubli ou à la portabilité des données, si celles-ci sont concentrées plutôt qu'éparpillées.

  - **Open Data et confidentialité**

Le caractère sensible des données personnelles fait toutefois l'objet d'une attention particulière. Pour les protéger, deux solutions peuvent être envisagées[14]. L'**anonymisation** retire aux données leur caractère personnel grâce à un traitement technique visant à rendre impossible et de manière irréversible l'identification d'un individu. Dans le cadre d'un Data Lake, cette procédure n'est pas obligatoire pour le stockage mais le devient en cas d'exploitation des données[15]. Les données anonymisées ne sont plus soumises aux dispositions du RGPD[16]. Il existe deux techniques d'anonymisation[17]:

- la **randomisation**, qui vise à protéger le jeu de données du risque d'inférence en modifiant ses attributs de sorte à les rendre moins précises, tout en conservant la répartition globale ;
- et la **généralisation**, qui permet notamment d'éviter l'individualisation d'un jeu de données, en modifiant les échelles afin de s'assurer qu'ils soient communs à un ensemble de personnes.

Petit bémol toutefois, le risque de ré-identification par recoupement demeure, et augmente même avec le volume de données. D'autre part, en les transformant, en en changeant le contenu et le structure, l'anonymisation est aussi susceptible d'entraîner une perte de qualité des données.

Quant à la **pseudonymisation,** elle permet de rendre impossible l'identification d'une personne sans avoir recours à des informations supplémentaires. Elle permet ainsi de traiter les données d'individus sans pouvoir les identifier directement. Concrètement, elle consiste à remplacer un attribut par un autre au sein d'un enregistrement, c'est-à-dire par exemple de remplacer les données directement identifiantes (nom, prénom...)

---

[14] Xavier Biseul, « Comment mettre son data lake au service de la conformité au RGPD », *ZDNet* [en ligne], 11 mai 2018, [consulté le 04 novembre 2019].

[15] Sauf si un texte de loi autorise la publication des données, qu'elles figurent dans la liste du Code des relations entre le public et l'administration (CRPA), ou que les personnes concernées donnent leur accord.

[16] « L'anonymisation des données, un traitement clé pour l'open data », *CNIL* [en ligne], 17 octobre 2019.

[17] *Op. cit. CNIL*, 2019.

par des données indirectement identifiantes (alias, numéro dans un classement…). Contrairement à l'anonymisation, il s'agit d'un processus réversible.

Le Data 61 australien : sécuriser les jeux de données

Dans le cadre de l'Agenda national pour la science et l'innovation, le CSIRO, l'agence australienne pour la science et l'innovation, travaille avec diverses agences gouvernementales australiennes via son bras armé pour les questions de data science et de numérique, Data 61, sur des projets de R&D destinés à accroitre le nombre de jeux de données accessibles aux autres agences et au grand public. C'est à l'occasion d'une collaboration avec le Bureau Australien des Statistiques que Data61 a développé le prototype Protary « Protecting Against Re-Identification », une API qui permet aux utilisateurs de générer des statistiques « confidentialisés ».

- **Gouvernance**

Pour gérer ces données et superviser les opérations et les modalités de partage, plusieurs nouvelles fonctions ont vu le jour :

- Le **Data Engineer** et le **Data Architect** sont chargés de l'infrastructure du Data Lake, à la fois sur le volet « connectique », tant avec les sources de données que vers les utilisateurs, que sur le volet « conception » de l'infrastructure informatique ;
- Le **Data Owner**, propriétaire et responsable des données dont il a la charge, a pour mission de gérer leur collecte, leur stockage et leur protection. C'est lui notamment qui cartographie les données, qui en contrôle l'accès, en coordonne la protection et qui met en place un référentiel pour les contextualiser.
- Le **Data Steward**, responsable référent de la gouvernance de la donnée, est responsable de l'organisation et de la gestion des données. C'est lui qui documente les données et l'ensemble des processus, traitements et contrôles qui leur sont appliqués. En d'autres termes, il est garant de la qualité des données, en partie grâce au Data Lineage.

L'ouverture et le partage de données dans le cadre de la mise en place d'un projet d'Open Data doit donc faire l'objet d'une planification précise et d'un encadrement rigoureux. Elle nécessite, en amont, une préparation minutieuse permettant d'identifier et de cartographier les données pouvant ou devant être ouvertes et partagées. Elle suppose ensuite la mise en place d'une équipe dédiée responsable de la gouvernance du dispositif, et chargée notamment de veiller à la sécurité des données ainsi partagées. Dans le cas du gisement de données qui constitue le cœur de la Cyberdéfense Factory et qui doit favoriser l'innovation de cyberdéfense et l'élaboration à la fois d'outils techniques (cyber-armes, diversification des capacités cybernétiques des armées…) et de mesures non techniques (amélioration du fonctionnement du ministère, planification et conduite des opérations…), il est essentiel que la politique de données profite tant aux armées qu'à l'ensemble de l'écosystème. Potentiel espace de collaboration, le Data Lake est une opportunité de prolonger le partage de données en partage de compétences entre civils militaires, si ce n'est une première étape vers l'échange de personnels entre sphères publique et privée.

FOCUS INNOVATION

Gorille de Cyber-Detect© : l'analyse morphologique pour la détection des attaques

La société

Cyber-Detect est une « spin-off» d'un projet recherche initié dès 2007 au sein du laboratoire LORIA (Laboratoire lorrain de recherche en informatique et ses applications) établi par le CNRS, l'INRIA et l'Université de Lorraine.

Constatant qu'il devenait **nécessaire d'endiguer les attaques capables de passer outre les filtres des antivirus classiques**, l'équipe de recherche a entrepris de développer les algorithmes qui fonderont **la suite logicielle Gorille de détection des attaques inédites ou obfusquées (dissimulées)**.

Cyber-Detect est fondé en 2017 en s'appuyant sur une équipe propre de **douze personnes** (ingénieurs, marketing etc.) ainsi que sur les conseillers scientifiques du laboratoire LORIA.

L'innovation

La solution Gorille repose **sur l'analyse morphologique© qui désassemble les codes binaires contenus dans les exécutables pour identifier leur comportement et leurs fonctions et, in fine, déceler les éléments malveillants**. En d'autres termes, Gorille agit comme une sonde qui décompose les exécutables entrant sur le réseau et les décortique : le logiciel pourrait, par exemple, détecter la fonctionnalité de chiffrement d'un rançongiciel inconnu des antivirus.

L'outil ne puise donc pas dans les répertoires de signatures de fichiers des antivirus, mais est en mesure de détecter en temps réel un exécutable malveillant quand bien même la menace serait une APT[1] ayant étudié et anticipé les solutions de sécurité mises en place par sa cible.

La suite logicielle Gorille complète ainsi les systèmes de défense logiciels classiques (antivirus, firewall) et doit permettre de déceler :

- Les attaques inconnues, qu'elles utilisent les variantes d'un code malveillant existant ou non ;
- Les attaques ciblées visant une organisation ou des individus ;
- Les attaques persistantes, le piratage furtif et continu.

Les applications

La suite logicielle **Gorille Expert et Gorille Réseau** permet, de :

- de sécuriser des datacenters et des sites physiques en agissant comme un sonde d'analyse morphologique en temps réel ;

- de réduire les risques d'exploitation des vulnérabilités via des Common Vulnerabilities Exposures (CVE), notamment des logiciels ; et d'assurer le maintien en condition opérationnelle (McO) et en condition de sécurité (McS).
- de renforcer les équipes de réponse à incident et de reverse en leur fournissant des éléments de compréhension des menaces sophistiquées et des pistes de remédiation, voire d'attribution, grâce à la décomposition des fonctionnalités des logiciels malveillants (Gorille Expert) ;
- La suite logicielle Gorille est conçue pour s'adapter et s'intégrer aux solutions de cybersécurité préexistantes et notamment les applications d'orchestration.

L'actualité

La start-up Cyber-Detect a bénéficié du **programme d'accélération de Thalès dans le cadre de la Station F** qui lui a notamment permis d'accélérer l'automatisation de sa solution d'analyse morphologique. La version automatisée, Gorille Réseau, s'adresse à un public plus large que Gorille Expert puisqu'elle automatise le traitement de la menace.

Enfin, la startup présentera sa solution lors de salons tels que la **European Cyberweek** de Rennes, le **Forum International de la Cybersécurité (FIC)** à Lille, ou encore le prestigieux **CES** de Las Vegas.

CALENDRIER

12-13/12 : Forum de Paris pour la Paix

Le Forum de Paris sera l'occasion d'une série de conférences et interventions sur des thématiques de cybersécurité et cyberdéfense parmi lesquelles :

- Premier anniversaire de l'appel de paris pour la confiance et la sécurité dans le cyberespace : pour une approche multi-acteurs des enjeux de cybersécurité
- Faire progresser la cyberstabilité : rapport final de la commission mondiale sur la stabilité du cyberespace
- Mise en œuvre des principes et valeurs de l'appel de paris : des initiatives concrètes
- Comment partager les données utiles ? Défricher la terre promise des data commons
- Novdigital democracy charter : pour le bien-être des sociétés démocratiques
- Solutions fusionnées : vers des principes de gouvernance communs pour l'intelligence artificielle?
- Liberté d'expression sans malveillance : contrer les fakes news et les contenus haineux

Pour plus d'informations : https://parispeaceforum.org/fr/

ACTUALITÉ

Inauguration de la Cyberdéfense Factory

La ministre des Armées, Florence Parly, a inauguré le 3 octobre la **Cyberdéfense Factory** à Rennes.

Antenne de l'Innovation Défense Lab et pilotée par la Direction générale de l'Armement (DGA), la création de la Cyberdéfense Factory témoigne de l'engagement du ministère des Armées pour **accélérer le développement de solutions innovantes au profit du Commandement de la cyberdéfense**. Il s'agit donc à la fois de capter les innovations du secteur civil, mais aussi de mettre à disposition des acteurs de l'innovation l'expertise du ministère des Armées.

Pour ce faire, la Cyberdéfense Factory a été conçue comme **un plateau collaboratif réunissant tous les acteurs de l'innovation : universitaires, start-ups et PME, et opérationnels de la défense**, qui pourront dans ce lieu unique échanger et partager des données et des compétences. Au cœur de ce dispositif, un **Data Lake**, ou gisement de données, qui permettra au ministère des Armées de **mettre à disposition de tous les participants des données d'intérêt cyber qui pourront être utilisées pour tester des solutions innovantes et développer de nouveaux algorithmes**.

La **Direction Générale des Relations Internationales et de la Stratégie (DGRIS)** propose les analyses politiques et stratégiques contribuant à renforcer l'appréciation des situations et l'anticipation. Elle soutient la réflexion stratégique indépendante, en particulier celle menée par les instituts de recherche et organismes académiques français et étrangers. Elle contribue au maintien d'une expertise extérieure de qualité sur les questions internationales et de défense.

A ce titre, la **DGRIS** a confié à **CEIS** la réalisation de cet **Observatoire du Monde Cybernétique**, sous le numéro de marché 1502492543. Les opinions développées dans cette étude n'engagent que leur auteur et ne reflètent pas nécessairement la position du Ministère de la Défense.

Ministère des Armées
Direction Générale des Relations Internationales et de la Stratégie
60 Boulevard du Général Martial Valin – CS21623 – 75 509 Paris Cedex 15

CEIS
Tour Montparnasse – 33, avenue du Maine – BP 36 – 75 755 - Paris Cedex 15
Téléphone : 01 45 55 00 20
E-mail : omc@ceis-strat.com