LE SSPCLOUD : UNE FABRIQUE CRÉATIVE

POUR ACCOMPAGNER LES EXPÉRIMENTATIONS DES STATISTICIENS PUBLICS

Frédéric Comte, Arnaud Degorre**, Romain Lesur****

Environnement d'aide à l'expérimentation sur les nouvelles méthodes de la* data science*, le SSPCloud pour le système statistique public est un ensemble de ressources informatiques permettant de réaliser des prototypes, de tester des traitements statistiques et de s'approprier de nouvelles pratiques de travail. Inscrit dans un courant d'inspiration de type FabLab, il apporte les conditions (im)matérielles pour favoriser la créativité du statisticien et l'aider à valoriser les nouveaux gisements de données. Il s'appuie sur des technologies de l'informatique dans les nuages (le* Cloud computing*) qui renforcent l'autonomie – et la responsabilité – des utilisateurs dans l'orchestration de leurs traitements.

Construit autour d'une communauté ouverte à l'ensemble des statisticiens publics, le SSPCloud se veut un atelier d'apprentissage, où le geste statistique se réinvente à plusieurs. La collaboration s'y trouve facilitée par l'adoption de solutions* open source*, garantissant les possibilités de réutilisation. Le SSPCloud propose un mélange fertile des deux univers professionnels de la statistique et de l'informatique, pour progresser plus particulièrement dans la mise en place de processus répondant aux standards de la reproductibilité appliqués aux traitements de la donnée et aux travaux d'études.

The SSPCloud for the official statistical system is an environment for experimenting with new data science methods. It is a set of computer resources for creating prototypes, testing statistical processing and adopting new work practices. Inscribed in a Fab Lab type of inspiration, it provides the (im)material conditions to encourage the statistician's creativity and help him or her make the most of new data sources. It is based on technologies of Cloud computing which reinforce the autonomy – and the responsibility – of users in the orchestration of their processing.

Built around a community open to all public statisticians, the SSPCloud is intended to be a learning workshop, where the statistical gesture is reinvented by many. Collaboration is facilitated by the adoption of open source solutions, guaranteeing the possibility of reuse. The SSPCloud offers a fertile mix of the two professional worlds of statistics and computer science, to progress more particularly in the implementation of processes that meet the standards of reproducibility applied to data processing and research work.

* Chef de projet, division Innovation instruction technique, DSI, Insee, *frederic.comte@insee.fr*
** Ancien chef de l'unité Innovation et stratégie du système d'information, DSI, Insee, *arnaud.degorre@insee.fr*
*** Chef de la division Innovation instruction technique, DSI, Insee, *romain.lesur@insee.fr*

UNE REFONTE DES APPAREILS DE PRODUCTION DE LA STATISTIQUE PUBLIQUE…

À l'issue de la conférence *The European path towards Trusted Smart Statistics* dédiée à l'émergence d'une société des données, les instances représentatives du système statistique européen adoptaient le 12 octobre 2018 le Mémorandum de Bucarest, posant les principes d'une refonte majeure des appareils de production de la statistique publique dans les différentes nations européennes. Ces principes visent plus particulièrement à doter le système statistique européen des capacités nécessaires pour prendre en compte les nouveaux gisements de données et de méthodes alternatives de traitement du chiffre, et cela dans de multiples dimensions, incluant le cadre juridique, les compétences techniques et les solutions informatiques de mise en œuvre (Eurostat, 2018).

La notion de *Trusted Smart Statistics*[1] embrasse ainsi un vaste ensemble d'évolutions liées à la démultiplication des sources d'information sur la société au sens large, dont les illustrations ne manquent pas :

- appréhender la tension sur le marché du travail *via* l'étude des offres d'emploi en ligne ;
- cartographier finement les mouvements de population au jour le jour, heure par heure, à partir des données de téléphonie mobile ;
- mesurer la vulnérabilité énergétique à partir des données des compteurs connectés d'électricité et de gaz, etc.

Autant d'exemples où la statistique publique est amenée à investir pour valoriser des informations de forme et de nature bien différentes des données d'enquête (Unece, 2013a ; 2013b).

Ces investissements s'accompagnent également d'innovations dans les processus statistiques, pour être en mesure de valoriser la richesse de ces nouvelles sources, mais aussi pour faire face à leur complexité, ou à leurs imperfections, qui impliquent des traitements conséquents de mise en conformité statistique. Au premier rang de ces innovations, figurent les méthodes dites d'apprentissage (*machine learning*) et leurs cas d'usage prometteurs dans les domaines de la codification et de la classification, de l'édition et de l'imputation des données. Le *Machine Learning Project* mis en place par la Commission économique pour l'Europe des Nations Unies (Unece) a ainsi pris la mesure, dans son rapport conclusif, des investissements réalisés à ce jour dans la communauté de la statistique publique : « *De nombreux instituts nationaux de statistique étudient la manière dont le machine learning peut être utilisé pour accroître la pertinence et la qualité des statistiques officielles dans un environnement caractérisé par des demandes croissantes d'informations fiables, des technologies accessibles et en développement rapide et de nombreux concurrents.* » (Unece, 2021).

… RENDUE POSSIBLE PAR DE NOUVELLES CAPACITÉS INFORMATIQUES

Les gisements de données massives, au cœur des *Trusted Smart Statistics*, présentent des caractéristiques qui, du fait de leur volumétrie, de leur vélocité (avec leur vitesse de constitution et de renouvellement) ou de leur variété (données structurées mais aussi non structurées, comme des images), en rendent la manipulation particulièrement complexe.

1. Qu'on pourrait traduire par « statistiques intelligentes et fiables ».

Sont ainsi considérées comme des données massives ou *big data*[2] les informations dont les caractéristiques sont telles qu'elles ne peuvent être facilement collectées, stockées, manipulées ou analysées par des équipements informatiques traditionnels.

Il est alors nécessaire de mettre en place des infrastructures d'un nouveau type, permettant d'anticiper sur les cas d'usage de la statistique publique, voire de les inspirer, dans une approche portée par les opportunités rendues possibles par l'innovation technologique[3].

Derrière la notion d'infrastructure informatique, il faut entendre de multiples strates, qui sont chacune des variables clefs dans la composition des services attendus : les capacités de stockage de la donnée selon différentes méthodes (fichier, objet, base de données, etc.), la puissance de traitement (mémoire vive, CPU[4], GPU[5]) ou encore les services de traitement (logiciels permettant d'exécuter des langages comme *R*, *Python*, etc.), mais aussi des architectures techniques d'**orchestration** dédiées à la mise en relation de ces différentes strates. Par exemple, le calcul distribué est une méthode de traitement de la donnée fondée sur la division d'un problème unique en une multitude de problèmes plus petits, afin de résoudre en parallèle chacun de ces problèmes sur un même centre de calcul (*via* le *multithreading*) ou en les répartissant sur plusieurs centres de calcul liés entre eux (appelés alors un *cluster*). La mise en œuvre effective d'un calcul distribué nécessite de mobiliser plusieurs éléments logiciels, appelés également *framework*, pour gérer l'accès aux données, les répartir entre des nœuds de traitement, d'effectuer toutes les opérations d'analyse nécessaires, puis de livrer les résultats.

LE RAPPORT DU STATISTICIEN À L'INFORMATIQUE ÉVOLUE, AVEC PLUS DE POUVOIRS… ET DE RESPONSABILITÉS

Ces infrastructures se destinent à des professionnels du traitement de la donnée qui sont appelés à développer, pour leurs besoins propres, des briques techniques et à les assembler en un processus intégré (ou *pipeline*). L'appellation de *data scientist*[6] traduit, parmi ses multiples acceptions, cette implication accrue du statisticien dans l'élaboration puis l'orchestration informatique de son traitement, au-delà des seules phases de conception ou de recette. Les nouvelles infrastructures de *data science* prennent en compte ce rôle étendu de ses utilisateurs, en leur accordant des possibilités d'action plus large qu'une infrastructure conventionnelle.

« L'appellation de data scientist *traduit, parmi ses multiples acceptions, cette implication accrue du statisticien dans l'élaboration puis l'orchestration informatique de son traitement. »*

2. Pour une caractérisation des *big data*, le lecteur pourra par exemple consulter les publications du *National Institute of Standards and Technology* (NIST) *Big Data Public Working Group* (NIST, 2017).
3. Voir les deux modèles d'alignement stratégique d'une organisation (Anderson et Ventrakaman, 1990) : un alignement « descendant » des solutions informatiques sur la base des besoins portés par les processus métiers et un alignement « ascendant » des processus métiers qui saisissent des opportunités de transformation rendues possibles par les évolutions informatiques.
4. *Central Processing Unit,* désigne la plupart du temps le processeur d'un ordinateur. On peut le traduire en français par unité centrale de traitement (UCT) ou unité centrale de calculs.
5. *Graphics Processing Unit*, ou processeur graphique en français.
6. « *Le data scientist* […] *effectue des tâches complexes dans le traitement des données. Il est capable de traiter des données variées et de mettre en place des algorithmes optimisés de classification, de prédiction sur des données numériques, textuelles ou d'images.* […] *Le data scientist utilise des outils de programmation et doit savoir optimiser ses calculs pour les faire tourner rapidement en exploitant au mieux les capacités informatiques (serveur local, cloud, CPU, GPU, etc.)* » (Dinum et Insee, 2021).

À cet élargissement du domaine d'intervention, correspond également un renforcement de la place des travaux de prototypage dans l'activité des statisticiens publics, à travers des démarches de va-et-vient entre la conception d'un traitement statistique et sa mise en œuvre. Certes, les traitements statistiques ont toujours été des processus évolutifs, appelés à connaître des adaptations à l'épreuve des données. Cette dimension est toutefois accentuée s'agissant de données qui, contrairement à des fichiers d'enquête, ne présentent pas nativement les qualités attendues en termes de stabilité des concepts et de précision des mesures. Par exemple, les données de téléphonie mobile peuvent comporter des évolutions dans les informations de localisation, liées aux méthodes de triangulation à partir de la position des antennes-relais, dont l'implantation évolue au fil des ans. Ou encore, l'interprétation des images satellitaires ou aériennes pour établir des indicateurs statistiques comme la tâche urbaine se fonde sur des clichés dont la qualité dépend des saisons ou des conditions météorologiques. Enfin, l'analyse des offres d'emploi en ligne peut être conduite pour en déduire des éléments sur les compétences requises, sur des champs textuels dont le contenu est hétérogène d'un recruteur à un autre, pour décrire un même métier.

Ce type de données appelle en outre à conduire des traitements statistiques de nature évolutive, fondés sur des algorithmes apprenants, plutôt que des opérations déterministes. Le comportement de l'algorithme est ainsi conçu pour évoluer au fur et à mesure qu'il « apprend » des données qui lui sont transmises, et son orchestration nécessite de disposer, au sein de l'infrastructure informatique, de services et de briques techniques pensées en conséquence. Le statisticien doit non seulement concevoir l'algorithme de traitement, mais aussi s'intéresser à sa mise en œuvre, prenant ainsi à son compte des éléments qui relèvent habituellement de l'exploitation informatique.

SORTIR DU CADRE POUR MIEUX ACCUEILLIR LES DÉMARCHES EXPLORATOIRES

La forte évolutivité des nouveaux gisements de données appelle à engager, en complément des projets informatiques structurants, des approches plus agiles, conduites dans des calendriers courts : fondées sur du tâtonnement, des essais et erreurs, des intuitions, elles privilégient une opportunité pour laquelle il s'agit de faire le point sur les possibilités offertes, plutôt que d'expertiser la solution optimale pour un objectif défini à l'avance. Ces démarches de prospection privilégient l'expérimentation avec une mise en place accélérée d'un premier prototype, afin de faire la preuve du concept[7] de façon empirique, sans forcément traiter toutes les facettes d'un sujet.

Les expérimentations peuvent toutefois buter sur une difficulté pour accéder aux ressources nécessaires à leur réalisation… naturellement plus faciles à accorder à un projet qui présente, par ses études préalables, toutes les garanties souhaitées par des instances d'arbitrage. L'essor des « laboratoires » dans les grandes organisations publiques comme privées est une réponse pour contrebalancer la place prédominante des investissements *via* les projets structurants, en proposant des espaces de créativité laissant la part belle au prototypage et à l'expérimentation. Un laboratoire, ou « lab », a pour principale caractéristique de proposer une « autre façon d'inventer »[8] par rapport au processus de R&D dominant. Pour cela, il est, au moins en partie, en dehors des règles habituelles de fonctionnement de l'organisation, donc en dehors des procédures classiques de décision.

7. *Proof of concept* (POC) en anglais.
8. « *Les Open labs constituent un lieu et une démarche portés par des acteurs divers, en vue de renouveler les modalités d'innovation et de création par la mise en œuvre de processus collaboratifs et itératifs, ouverts et donnant lieu à une matérialisation physique ou virtuelle* » (Mérindol, Bouquin, Versailles *et alii*, 2016).

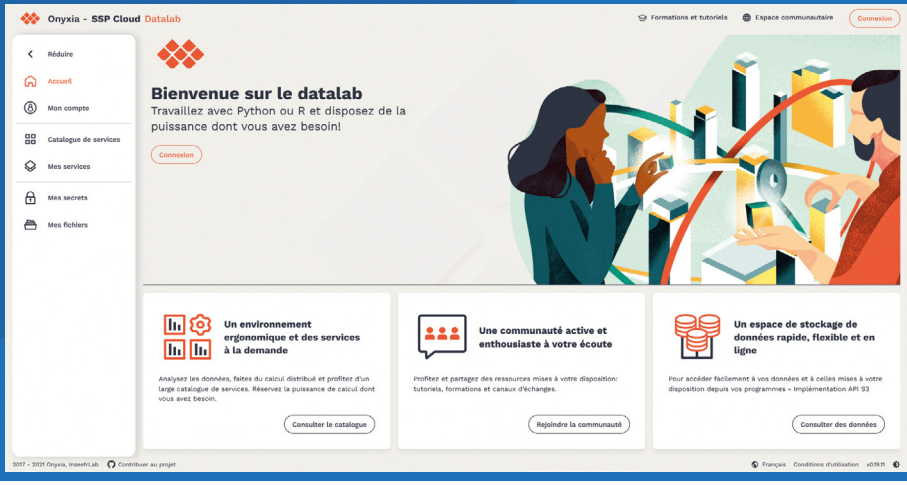
DE L'OMBRE À LA LUMIÈRE : L'ÉMERGENCE D'UNE NOUVELLE INFRASTRUCTURE POUR LE SSP

Conçu pour répondre aux besoins d'outillage en *data science*, le SSPCloud ***(figure 1)*** est une nouvelle infrastructure mise en place par l'Insee pour offrir, sur le versant informatique, un environnement « lab » propice à l'engagement d'expérimentations : sur les nouveaux gisements de données et les nouvelles méthodes de traitement, plus particulièrement sur des calculs distribués sur données massives, comme les données de téléphonie mobile pour calculer des populations présentes à des échelles géographiques fines (Suarez Castillo *et alii*, 2020) ; sur des méthodes d'apprentissage, comme des systèmes apprenants de codification automatique de libellés de professions dans des nomenclatures.

D'abord créé « en dehors de l'organisation » ***(encadré 1)***, le SSPCloud trouve ses origines en 2017, avec la participation d'une équipe de l'Insee au *hackathon New Techniques and Technologies for Statistics* (NTTS). Celle-ci joue le rôle de déclencheur : l'équipe parvient à proposer un traitement de données massives intégrant à la volée des restitutions sous forme de *data visualisation*, mais rapporte dans son retour d'expérience d'importantes limitations techniques rencontrées dans les infrastructures alors à leur disposition. Répondre à ce besoin sera le fil conducteur des travaux engagés par un collectif d'agents de la direction du système d'information. L'opportunité offerte par la disponibilité d'équipements informatiques décommissionnés de leur usage initial permettra, quelques mois plus tard, de créer le premier prototype fonctionnel d'une plateforme d'expérimentation en *data science*, qui évoluera ensuite pour devenir le SSPCloud.

Le dispositif a continué d'évoluer et de s'enrichir, jusqu'à recevoir plusieurs soutiens institutionnels internes[9] et externes[10], et s'ouvrir à l'ensemble du système statistique public en octobre 2020. Pour mettre en exergue cette volonté d'ouverture et faire référence au paradigme technologique utilisé, l'appellation de SSPCloud a été retenue pour ce nouveau service mutualisé entre l'Insee et les services statistiques ministériels.

Figure 1. Pour faire un tour sur le SSPCloud (https://datalab.sspcloud.fr)



9. La démarche a reçu le soutien progressif de l'institut, avec la création des labs statistique (SSP Lab) et informatique (division Innovation et instruction technique) de l'Insee, comptant en leur sein les agents ayant œuvré à la constitution de cet écosystème expérimental.
10. Le SSPCloud a également bénéficié d'un financement *via* le Fonds de transformation ministériel de Bercy en 2019, et de l'apport de trois entrepreneurs d'intérêt général, suite à un appel à projet remporté auprès de la direction du Numérique en 2020.

UNE FABRIQUE CRÉATIVE, OUVERTE EN LIBRE SERVICE

Le SSPCloud se présente comme une « fabrique créative », ou FabLab[11], de la statistique publique. Ouvert en libre service, il propose un accès direct à l'ensemble des matériels physiques (comme des cartes graphiques), logiques (comme des *framework* et briques logicielles), utiles à une activité de *data science* ***(figure 2)***. Le statisticien public y dispose des composants techniques nécessaires pour prototyper et tester un processus de traitement, de bout en bout.

L'accès à ces ressources est immédiat : il n'est pas nécessaire, par exemple, de demander le provisionnement préalable d'un espace de stockage, ou d'un environnement de recette, les technologies utilisées rendent autonome l'utilisateur pour en disposer.

Les réalisations conduites sur le SSPCloud ne font pas l'objet d'une régulation ou d'un contrôle[12] : l'utilisateur a le libre choix des éléments qu'il mobilise, sans être contraint par un cadre technique de cohérence qui viendrait délimiter le champ des possibles.

Pour rester une fabrique créative, le SSPCloud évolue par ailleurs en permanence, avec l'installation de nouveaux logiciels et leur mise à jour en continu[13]. L'utilisateur peut contribuer à élargir le catalogue de services et introduire de nouvelles briques techniques dont ses pairs pourront bénéficier. Le SSPCloud place ainsi le statisticien au cœur de la conception et du développement de ses futurs processus statistiques.

Encadré 1. « C'est plus marrant d'être un pirate que de s'engager dans la marine » (S. Jobs)

L'informatique et l'innovation de l'ombre

Ces mots de Steve Jobs permettent d'appréhender l'une des motivations amenant à l'apparition de démarches d'innovation conduites « en secret » au sein d'une organisation. S'interrogeant sur les ressorts de l'innovation dans les entreprises et les administrations, Donald A. Schon a introduit la notion de contrebande d'innovation (*bootlegging*) (Schon, 1963). Le *bootlegging* est défini comme une recherche dans laquelle des individus motivés organisent secrètement un processus d'innovation, sans la permission officielle des instances de direction, mais pour le bénéfice de l'entreprise. Dans le domaine des systèmes d'information, le même phénomène a pu être observé avec la *Shadow IT* – terme désignant les systèmes d'information réalisés et mis en œuvre au sein d'organisations sans approbation de la direction du système d'information.

Prenant conscience du potentiel de créativité des collaborateurs et de la nécessité de les accueillir, les grandes organisations visent désormais à accompagner les démarches « masquées » d'innovation, en apportant des ressources techniques d'usage libre sans avoir à rendre de compte (Robinson et Stern, 1997). Cette mise en lumière de « l'innovation de l'ombre » permet alors de capter l'esprit créatif des agents, de façon plus horizontale.

11. Dans l'imaginaire collectif, les laboratoires sont peuplés d'éprouvettes, de réactifs en tout genre, de machines sophistiquées, dans des assemblages parfois déconcertants, aux réactions inattendues… L'appellation FabLab permet de rappeler la place prise par ce matériel d'expérimentation dans le SSPCloud, et les créations qui peuvent découler de leur libre usage.
12. Tant qu'elles s'inscrivent dans les conditions générales d'usage. Ces dernières portent sur le type de données pouvant être traitées sur le SSPCloud. Pour en savoir plus, voir les Conditions générales d'utilisation du SSPCloud (*https://www.sspcloud.fr/tos_fr.md*).
13. Le dispositif doit aider à imaginer, demain, ce que pourraient être les futures chaînes de production de la statistique. Pour reprendre l'image du laboratoire de chimie, il s'agit de travailler non seulement sur la découverte d'une nouvelle molécule, mais aussi sur le procédé permettant de l'obtenir de façon industrielle.

Figure 2. Le SSPCloud vu de l'intérieur : dans la mécanique du nuage

Sur sa partie physique, la «fabrique» SSPCloud s'appuie sur une ferme d'une quinzaine de serveurs…

… pour une capacité totale de 6 To de RAM, 120 To de stockage, 300 cœurs (600 threads), de 18 GPU Nvidia T4 et d'un débit réseau intra cluster de 20 Gb/s.

Sur sa partie logique, elle propose un libre assemblage de services :

- environnements de développement (*Visual Studio Code*), dont des environnements spécialisés dans les langages de traitement de données (*RStudio, Jupyter*) ;
- systèmes de gestion de base de données (*PostgreSQL*, *MongoDB*) ;
- *frameworks* et composants dédiés à des traitements avancés de *data science* (*Apache Spark* pour les données massives, *TensorFlow* pour le *machine learning*, *fast.ai* pour le *deep learning*, *RAPIDS AI* pour l'utilisation de la puissance de GPU) ;
- moteurs avancés de recherche (*Elasticsearch*) ;
- gestion d'espaces sécurisés *via* la solution de gestion de secrets *Vault*.

DES CHOIX TECHNOLOGIQUES EN FAVEUR DE LA SCALABILITÉ

Dans ses partis pris technologiques, le SSPCloud s'appuie sur une architecture répondant à plusieurs sources d'influence de l'informatique contemporaine.

La capacité de dimensionner des traitements – ce que désigne la notion de scalabilité[14] – est une attente clef du *data scientist*, pour stocker et manipuler des données massives, mais aussi pour disposer de ressources computationnelles adaptées aux méthodes de *machine learning*.

Ce besoin trouve sa réponse dans la distribution du traitement sur plusieurs centres de calcul, par exemple avec une ferme de serveurs – une approche au cœur des technologies du *Cloud computing* (l'informatique en nuage). Le fournisseur de service, également appelé *Cloud provider*, met à disposition à la demande des ressources physiques (CPU, mémoire, disques, GPU) fournies par des serveurs distants, partagés entre plusieurs clients, de façon à ce que le service puisse être redimensionné au fur et à mesure du besoin exprimé, en mutualisant de vastes ensemble de serveurs.

Pour être mise en œuvre, cette approche nécessite également d'adapter l'orchestration des traitements : ce fut en particulier l'objectif poursuivi dans le développement du *framework Hadoop*[15]. L'idée principale est la colocalisation du traitement et de la donnée : si le fichier source se trouve réparti sur plusieurs serveurs (approche horizontale), chaque section du fichier source est directement traitée par un processus de la machine hébergeant cette section pour éviter les transits réseaux entre les serveurs.

Le SSPCloud a adopté une solution qui s'inscrit dans le sillage de ces travaux, en retenant le *framework Apache Spark* conçu comme une méthode pour accélérer le traitement des systèmes *Hadoop* ***(encadré 2)***. Il a d'ailleurs été utilisé à cette fin, au cours du premier semestre 2021, pour prototyper une architecture alternative à celle pré-existante, s'agissant du processus de traitement des données de caisse dans le cadre de l'élaboration de l'indice des prix à la consommation[16]. Une accélération des traitements pouvant atteindre jusqu'à un facteur 10 a ainsi été obtenue, pour des opérations qui prenaient jusqu'alors plusieurs heures d'exécution.

UN ENVIRONNEMENT ET DES RESSOURCES ACCESSIBLES EN TOUT LIEU

Dans une infrastructure *Cloud*, l'ordinateur de l'utilisateur devient un simple point d'accès pour exécuter des applications ou des traitements sur une infrastructure centrale. Le SSPCloud a été conçu sur ce modèle, en permettant aux statisticiens publics de s'y connecter depuis n'importe quel poste, dès lors que ce dernier accède à Internet. Indépendant des infrastructures propres de l'Insee ou des services statistiques ministériels, il ne nécessite donc pas d'être relié à un réseau local de l'administration d'appartenance, ce qui lui permet en outre d'accueillir des utilisateurs venus d'autres horizons (par exemple, des membres d'instituts statistiques européens, des universitaires, etc.).

14. En informatique matérielle et logicielle et en télécommunications, l'extensibilité ou scalabilité désigne la capacité d'un produit à s'adapter à un changement d'ordre de grandeur de la demande, en particulier sa capacité à maintenir ses fonctionnalités et ses performances en cas de forte demande.
15. Le *framework Hadoop* a été publié par Doug Cutting et l'entreprise *Yahoo* sous la forme d'un projet *open source* en 2008, en s'inspirant des travaux de *Google* (Dean et Ghemawat, 2004).
16. Voir (Leclair, 2019).

L'infrastructure du SSPCloud apporte ainsi un service dit ubiquitaire, accessible en tout lieu, depuis n'importe quel terminal doté d'un navigateur internet (ordinateur, tablette, téléphone, etc.).

Encadré 2. Comment l'écosystème de référence peut être détrôné en quelques années

Il faut imaginer, plutôt qu'un logiciel monolithique, tout un écosystème d'applications et de composants, en évolution permanente… jusqu'à ce qu'un nouvel écosystème l'emporte.

Début des années 2000

Les solutions de traitement de données massives voient leurs origines dans les travaux conduits pour les moteurs de recherche sur le *web*, avec les premières investigations de la fondation *Apache* pour des systèmes d'indexation de contenus massifs (projets *Lucene* et *Nutch*).

2004

Jeffrey Dean et Sanjay Ghemawat, employés chez Google, créent l'algorithme *MapReduce* pour paralléliser les traitements de grands volumes de données sur plusieurs serveurs.

2006

Doug Cutting, qui a mené les projets *Lucene* et *Nutch* avant de rejoindre *Yahoo*, crée un nouveau système de fichier distribué qu'il combine avec *MapReduce*, et nomme ce *framework Hadoop*…

Le *framework Hadoop* est destiné à créer des applications distribuées, au niveau du stockage des données et de leur traitement.

2009

…*Hadoop* est légué en 2009 à la fondation *Apache*…

…émergence d'*Apache Spark*, conçu par Matei Zaharia au sein de l'université de Californie à Berkeley…

Là où le *MapReduce* travaille par étape, *Spark* peut travailler sur la totalité des données et exécute toutes les opérations d'analyse en temps réel.

D'abord pensé en complémentarité à *Hadoop*, *Spark* peut aussi être utilisé avec d'autres méthodes de stockage, comme *Amazon S3*… et constituer ainsi le cœur d'un nouvel écosystème.

Voir (Dean et Ghemawat, 2004).

Cette logique visant à concilier scalabilité et ubiquité a également influencé le choix des technologies de stockage de la donnée. Le SSPCloud s'appuie sur S3 (*Simple Storage Service*)[17], un stockage dit « objet » – un objet se composant d'un fichier, d'un identifiant et de métadonnées, le tout de taille arbitraire, non bornée. Chaque dépôt de données (ici appelé un *bucket*) se trouve consultable directement avec un adressage unique (une URL par dépôt[18]) et des services d'accès par API[19]. Pensé pour offrir un dimensionnement adaptable, optimisé pour lancer des calculs intensifs[20], le stockage S3 apporte également des propriétés intéressantes pour faciliter l'accès aux données : ainsi, il comporte une API de sélection pour n'appeler dans une requête qu'un sous-ensemble d'un fichier, même si celui-ci est compressé ou chiffré. Surtout, il est le complément naturel à des architectures fondées sur des environnements dits conteneurisés, pour lesquels il apporte une couche de persistance, des modalités de connexion facilitées sans compromettre la sécurité, voire en la renforçant par rapport à un système de stockage traditionnel.

LA CONTENEURISATION POUR MAÎTRISER LES CONDITIONS D'EXÉCUTION

Dans le monde de l'informatique, un conteneur[21] est un regroupement logique de ressources, qui permet d'encapsuler et d'exécuter des ensembles logiciels, par exemple une application, des bibliothèques et d'autres dépendances réunies en un seul package. La conteneurisation ***(encadré 3)*** propose un système logique d'isolation, apportant une réponse à deux problématiques majeures des environnements de traitement de données :

- elle permet de gérer la concurrence d'accès aux ressources physiques (CPU, mémoire) entre les différents utilisateurs, en organisant la répartition des capacités entre les conteneurs ;
- elle assure une complète indépendance du contenu de chaque conteneur, ce qui permet de construire une large palettes de services et logiciels sans exposer les utilisateurs à des problèmes de compatibilité entre librairies.

Dans le cas d'un environnement dédié à l'expérimentation, le mécanisme des conteneurs permet de gérer une forte évolutivité de l'offre de services : en effet, contrairement à une plateforme centrale monolithique qui implique pour chaque utilisateur d'adapter son code à la montée de version logicielle du socle, un dispositif de conteneur permet à chacun de maîtriser les composants qu'il utilise. Ce dispositif permet de rejouer, dans des conditions techniques maîtrisées, l'intégralité d'un traitement.

17. En 2006, *Amazon web service* ouvre un service de stockage en ligne fondé sur une nouvelle technologie, désigné par l'appellation *Amazon S3*. Le logiciel *open source Minio* permet de construire et déployer un service S3 ne dépendant pas d'*Amazon*.
18. Une URL (*Uniform Resource Locator*, littéralement « localisateur uniforme de ressource ») est une chaîne de caractères qui permet d'identifier une ressource du *World Wide Web* par son emplacement.
19. Une API (*Application Programming Interface*) est un ensemble standardisé de méthodes par lequel un logiciel offre des services à d'autres logiciels. On parle d'API à partir du moment où une entité informatique cherche à agir avec un système tiers, et que cette interaction se fait en respectant les contraintes d'accès définies par le système tiers.
20. Contrairement au stockage HDFS qui s'appuie sur la colocalisation de la donnée et des unités de traitement, S3 permet de relier librement des conteneurs avec des données qui n'y sont pas physiquement adossées. Cette logique est pertinente pour des traitements qui nécessitent d'allouer une grande puissance de calcul de façon dynamique, comme pour le *machine learning*.
21. Voir *https://lwn.net/Articles/256389/*.

Encadré 3. Les techniques de conteneurisation renforcent l'autonomie du *data scientist*

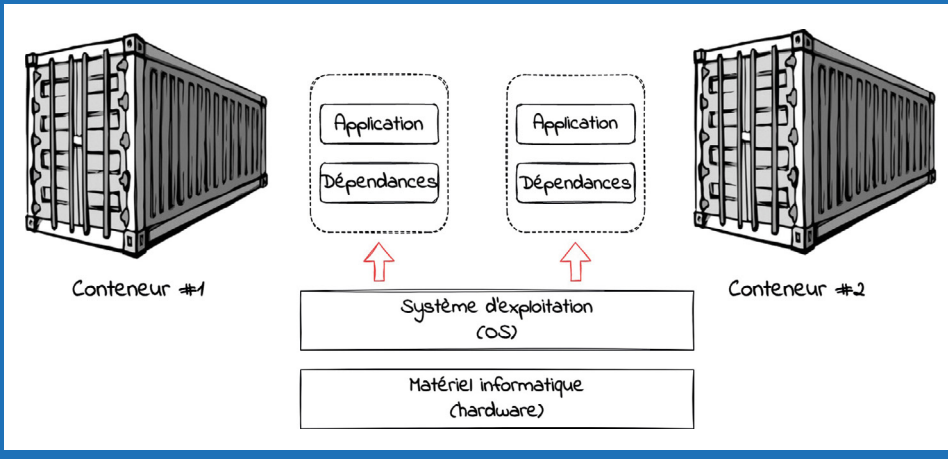
Un système d'information se compose de **services** (serveur *web*, base de données, etc.) qui s'exécutent en continu et de **tâches** dont l'exécution est planifiée de manière régulière ou ponctuelle. Le code nécessaire à l'exécution des services et des tâches s'appuie sur un **système d'exploitation**, qui organise l'accès aux ressources physiques (CPU, RAM, disques, etc.). En pratique, l'intervention d'acteurs spécialisés dans l'exploitation des infrastructures informatiques est nécessaire pour préparer et mettre à jour les systèmes sur lesquels s'exécute le code. Ainsi, dans l'organisation conventionnelle d'un SI, la mise en production et la maintenance d'un code (application complète ou script de traitement) nécessitent des interactions fortes entre les développeurs du code et les exploitants des infrastructures informatiques.

La **conteneurisation apporte une alternative** en créant des « bulles » propres à chaque service, tout en ayant recours à un même système d'exploitation support. Cette isolation assure la portabilité du code d'un environnement de développement à l'environnement d'exploitation, en maîtrisant l'ensemble des dépendances. La conteneurisation peut être harmonisée sur un ensemble de serveurs.

Pour qu'un utilisateur puisse demander, en toute autonomie, l'exécution d'un code, des ensembles logiciels assurent alors une fonction d'**orchestrateur** : *Kubernetes* optimise ainsi l'allocation des ressources physiques pour un ensemble de conteneurs et facilite leur mise en relation. Ce dispositif gère la **scalabilité** d'un traitement, par exemple en dupliquant un conteneur si nécessaire pour tenir la charge. Il gère aussi la **portabilité**, en déplaçant au besoin un conteneur dans un autre groupe de ressources (d'autres outils permettent même de les déplacer d'un *cluster* de calcul à un autre). Ces deux caractéristiques permettent de gérer des calculs intensifs avec les technologies *big data*.

Le SSPCloud correspond à un paradigme dit d'infrastructure codée (*infrastructure as a code*).

L'environnement conteneurisé est créé uniquement *via* les spécifications des scripts : le *data scientist* peut lui-même définir son environnement de travail, les ressources à y allouer, les logiciels à inclure (par exemple, *R*) et toutes les bibliothèques utiles à son traitement (par exemple, des *packages R*). La logique du SSPCloud est ainsi de proposer au statisticien d'assurer à la fois les fonctions de conception (écriture des traitements), de déploiement (écriture du conteneur qui encapsule le traitement) et d'exploitation (écriture de l'orchestration du conteneur).



Le SSPCloud est fondé sur l'utilisation de conteneurs, en s'appuyant sur l'environnement *open source Kubernetes*[22] pour déployer et gérer les services conteneurisés. Ce choix technologique, parfaitement en phase avec les environnements *Cloud*, répond en outre naturellement aux besoins de scalabilité : la ressource affectée à un conteneur est en effet paramétrable, et des traitements complexes peuvent être répartis au sein d'un *cluster* et pris en charge par plusieurs conteneurs en parallèle. En conséquence, les environnements dédiés aux données massives sont désormais de plus en plus souvent organisés *via* le recours à un mariage du *Cloud computing* et de l'orchestration de conteneurs. Les *frameworks big data* comme *Apache Spark* ou *Dask* fonctionnent en parfaite synergie avec les conteneurs : le traitement massif s'y trouve découpé dans une multitude de petites tâches et le conteneur est sans aucun doute la meilleure façon de déployer ces petites opérations unitaires, en minimisant la ressource requise[23].

ACCOMPAGNER LE GESTE INFORMATIQUE DU STATISTICIEN

L'ensemble des technologies réunies au sein du SSPCloud amène à revoir, en profondeur, le geste professionnel du statisticien dans son usage de l'environnement informatique. Au tournant des années deux-mille, la micro-informatique connaissant son apogée, une grande partie des ressources techniques d'un agent de l'Insee étaient locales, ou du moins dans une logique d'accès local : le statisticien avait, sur sa machine, son code et son logiciel de traitement, ainsi qu'un accès aux données *via* un système de partage de fichiers. Cet environnement a conduit, dans une certaine mesure, à l'essor d'une gestion manuelle des traitements réalisés « en self » par le statisticien.

Lorsque, dans une démarche de rationalisation des infrastructures informatiques, des systèmes de traitement mutualisés ont été de nouveau privilégiés, ces derniers ont cherché à préserver l'expérience utilisateur en proposant d'accéder à un « bureau distant ». Par exemple, l'Architecture Unifiée Statistique (AUS), un centre de calcul interne à l'Insee, propose une forme de transition entre une informatique locale et une informatique centralisée : elle concentre toutes les ressources sur des serveurs centraux, mais recrée un poste virtuel par agent, ce qui l'amène à garder la même pratique, largement manuelle, dans l'élaboration de ses scripts et dans l'exécution de ses traitements.

> *« Ce cadre permet d'adopter des pratiques vertueuses, pour mieux séparer le code, les données et le processus. »*

L'expérience utilisateur au sein du SSPCloud est radicalement différente, et rend impérative l'appropriation d'autres gestes. L'utilisateur ne dispose pas d'un bureau distant, et doit apprendre à composer avec des ressources qui, dans leur conception, sont évanescentes et n'existent qu'au moment de leur mobilisation effective. Ce cadre permet d'adopter des pratiques vertueuses, pour mieux séparer le code, les données et le processus, dont la persistance sera gérée selon des technologies différentes. Il lui faut également apprendre à organiser leur orchestration, dans un cadre qui concilie autonomie et automatisation. Le statisticien voit ainsi sa pratique se rapprocher de celle d'un développeur.

22. L'environnement *Kubernetes* est né dans *Google Cloud* : il y a été développé et publié en *open source* en 2014.
23. Alors qu'un dispositif de machines virtuelles (VM) nécessite d'installer un système d'exploitation complet pour chaque VM, les conteneurs sont des unités bien plus légères, partageant un même noyau de système d'exploitation.

RENDRE PÉRENNE CE QUI N'EST PAS PERSISTANT

La première transformation du geste du statisticien porte sur sa capacité à organiser la persistance des éléments qu'il mobilise dans son traitement. Dans le SSPCloud, l'ensemble des services sont dits « non-persistants » : ils sont conçus pour être désactivés lorsque l'utilisateur n'en a plus l'usage. Tout ce qui a été développé au sein de ce service disparaît à cette occasion – plus précisément, toutes les ressources conçues au sein d'un conteneur s'effacent avec l'extinction de ce dernier – contrairement à la pratique sur un poste local où l'utilisateur garde une trace de ses fichiers sur son espace de stockage, de même qu'avec un poste distant adossé à un répertoire de fichiers.

L'utilisateur du SSPCloud doit veiller à organiser la pérennité des ressources qu'il crée – à commencer par son programme informatique. Les fonctionnalités d'un outil de contrôle de version comme *Git*[24], utilisé avec une forge collaborative[25], permet au statisticien de gérer son code en dehors de l'environnement dans lequel il travaille, et d'y recourir depuis n'importe quel terminal. L'utilisation de *Git* permet en outre d'accroître la traçabilité des traitements et de les archiver au fur et à mesure qu'ils sont conçus, participant ainsi à l'amélioration continue des processus statistiques. D'autres services peuvent alors être adossés au dépôt du code source, pour « construire » le *pipeline* statistique, tester son intégrité, produire ses modules – c'est la fonction première d'une forge logicielle, un incontournable du génie logiciel qui devient également un élément pivot dans l'environnement de travail du statisticien. Le SSPCloud est prévu pour fonctionner de pair avec des forges logicielles – il comporte en son sein une instance privée du logiciel *GitLab*, et permet également d'appeler les services des instances publiques de *GitHub* et *GitLab*.

ÊTRE AUTONOME DANS L'ORCHESTRATION DE SON TRAITEMENT

La deuxième transformation apportée par le SSPCloud a trait à la méthode d'orchestration des traitements. L'utilisateur de cette infrastructure est appelé à construire de lui-même l'environnement technique adapté à son besoin, sans avoir à solliciter des équipes informatiques. À cette fin, il lui faudra apprendre à spécifier, de façon programmatique, les différents paramètres qui vont composer son environnement d'exécution.

> *« L'utilisateur de cette infrastructure est appelé à construire de lui-même l'environnement technique adapté à son besoin, sans avoir à solliciter des équipes informatiques. »*

Cette approche est possible grâce à l'utilisation des technologies *Cloud* et de la conteneurisation. En effet, les technologies *Cloud* sont agnostiques quant aux services qui y sont déployés et les conteneurs quant à eux sont façonnables à loisir par l'utilisateur, qui en définit lui-même le contenu et les ressources allouées. La possibilité offerte au statisticien de déployer en quelques secondes des bases de données, un *cluster* de calcul et un environnement de traitement statistique, qu'il a lui-même choisis voire conçus lui permet la plus grande créativité[26].

24. *Git* est un logiciel de gestion de versions, permettant de garder trace de toutes les modifications apportées à un code, pour l'ensemble des contributeurs. Il fonctionne de façon décentralisée : le code informatique développé est stocké sur l'ordinateur de chaque contributeur du projet, et le cas échéant sur des serveurs dédiés.
25. Une forge est un ensemble d'outils de travail, initialement conçus pour le développement de logiciels, utiles plus généralement pour d'autres types de projets comme l'écriture de codes statistiques. Pensée pour le travail à plusieurs mains, une forge collaborative permet d'organiser des processus de contribution impliquant plusieurs acteurs.
26. Le SSPCloud offre, au travers de son catalogue de services, des outils standards qui répondent à la majorité des besoins. Pour autant, l'utilisateur peut concevoir ses propres outils et les partager. Il est ainsi libre de tester un nouveau logiciel ou une nouvelle librairie *open source.*

Cette approche permet également de concevoir l'orchestration d'une suite d'opérations élémentaires. Ainsi, le statisticien peut concevoir un ordonnancement de tâches telles que par exemple, récupérer toutes les nuits des données issues du *web*, en extraire les informations pertinentes, les stocker dans des tables de données et alimenter une application interactive qui sera mise à jour et déployée automatiquement.

Cette logique, poussée jusqu'à son terme, a par exemple permis à une équipe de statisticiens de l'Insee de déployer un service de distancier (calcul de distances routières) utilisable par la communauté SSPCloud pour outiller des analyses spatiales. Ils ont ainsi à la fois transformé des données, déployé en toute autonomie un serveur offrant une API de calcul des distances routières et enfin, développé et déployé une application interactive permettant à un utilisateur de calculer des distances entre différents lieux. Fait remarquable, aucune intervention d'un exploitant informatique n'a été requise pour réaliser l'ensemble de ces tâches[27].

PROGRESSER DANS LA MISE EN PLACE D'UNE STATISTIQUE REPRODUCTIBLE

L'enjeu d'une pleine maîtrise du contexte d'exécution vaut tout autant pour les activités d'études que pour les activités de production statistique. Dans le cas de travaux académiques[28], il est par exemple exigé, pour des publications se fondant sur l'exploitation de données, de répondre à un nouveau critère de scientificité[29] : la possibilité pour un tiers de reproduire à l'identique l'ensemble des résultats publiés.

L'atteinte de cette reproductibilité demande donc de concevoir des solutions de traitement du chiffre qui puissent être rejouées d'une part, partagées avec des tiers d'autre part. Les technologies de conteneurs répondent parfaitement à cette attente, en décrivant de façon normée et datée toutes les ressources d'un ensemble de traitements, description qui peut être publiée sur des espaces ouverts (par exemple, des registres de conteneurs) et reliées à la publication du code source d'un traitement (par exemple, sur une forge logicielle publique comme *GitHub* ou *GitLab*).

> *« Le data scientist procède désormais d'une documentation active, où la description même des tâches à conduire va être menée de façon à automatiser leur mise en œuvre. »*

Cette exigence se diffuse progressivement au monde de la statistique publique, déjà attachée au fait de pouvoir rendre compte, par la documentation méthodologique et les rapports qualité, du processus d'élaboration d'un indicateur. Là où le statisticien public était amené à constituer une documentation passive des étapes de son processus, dans l'optique de rendre compte d'une conformité, le *data scientist* procède désormais d'une documentation active, où la description même des tâches à conduire va être menée de façon à automatiser leur mise en œuvre. La capacité à décrire les environnements de traitement de façon programmatique, avec des « contrats » de conteneurisation, est au cœur de cette démarche.

27. Voir également ***encadré 3***.
28. La détection d'erreurs majeures dans les calculs d'un article des économistes de renom Carmen Reinhart et Kenneth Rogoff, publié en 2010 et conduisant à un *erratum* en 2013, a appelé à rehausser les exigences académiques en matière de scientificité des publications économiques et statistiques.
29. En écho aux critères de scientificité de Karl Popper, ce dernier insistant sur la nécessité de pouvoir confronter des théories à des éléments de vérifications, sous forme d'expérience. Par extension, il s'agit d'avoir des éléments vérifiables pour s'assurer de l'exactitude des calculs effectués par des auteurs en démonstration de leur thèse.

S'OUVRIR POUR FACILITER LE PARTAGE DES SAVOIRS ET SAVOIR-FAIRE

En plus d'être une infrastructure informatique, le SSPCloud a été pensé comme un « lieu »[30] facilitant les rencontres entre des acteurs forts de compétences et d'expériences différentes. L'accès au SSPCloud depuis Internet rend possible son accès à l'ensemble du système statistique public – et plus largement, à l'ensemble des acteurs souhaitant s'inscrire dans un courant de partage de leur connaissance.

Date symbolique de cette volonté collaborative, la pré-ouverture du SSPCloud a été conduite le 20 mars 2020, au tout début du premier confinement lié à la crise pandémique, une période où beaucoup d'agents publics n'avaient plus d'accès à leur environnement administratif de travail. La version de test du SSPCloud a alors été dédiée à l'accueil de l'ensemble des agents publics, quelle que soit leur administration d'appartenance, pour leur apporter un espace de formation en distanciel aux logiciels de *data science*.

Couvrant tout autant des étapes de découverte (initiation à des langages) que des cas d'usages plus avancés (*machine learning*, calcul distribué), l'espace pédagogique du SSPCloud a immédiatement rencontré son public, en outillant par la même occasion une communauté d'agents dédiés à l'apprentissage de *R* et *Python*, la communauté *Spyrales*[31]. Le dispositif des conteneurs permet en particulier de proposer, aisément, des systèmes de tutoriels interactifs, en associant dans des environnements sur mesure les logiciels, les exercices et jeux de données, les exemples de programmes, sous la forme par exemple d'un *notebook*. Le SSPCloud est devenu, par la même occasion, la terre virtuelle d'accueil d'évènements à visée pédagogique, organisés en distanciel en 2020 et en 2021, à l'instar d'un jeu sérieux de formation en *R*[32].

BÉNÉFICIER DE LA MISE EN COMMUN EN PRIVILÉGIANT L'*OPEN SOURCE*

L'élargissement des publics bénéficie plus généralement à l'innovation, en facilitant la coopération entre des équipes qui, autrement, évoluent dans des systèmes d'information étanches.

L'essor de l'*open source* est venu apporter un cadre juridique et une méthode de travail permettant d'accompagner cette volonté de partage. La statistique a été marquée par l'essor des logiciels libres durant ces 20 dernières années : c'est ainsi que *R* et maintenant *Python* sont devenus les langages de référence pour la statistique en lieu et place, par exemple, de *SAS*®.

Au-delà de l'usage de langages et logiciels *open source*, les statisticiens ont plus généralement pris l'habitude de partager leurs programmes : ainsi, la publication d'un article proposant une nouvelle méthode de traitement des données est désormais quasi systématiquement

30. La notion d'infrastructure renvoie à l'idée d'un axe de passage (l'autoroute) où les utilisateurs se suivent ; la notion de lieu renvoie à un espace de jonction (la place publique) où les utilisateurs se rencontrent.
31. La communauté *Spyrales* a été constituée en mars 2020 pour aider les agents publics à se former aux langages *R* et *Python*, en facilitant la mise en relation de personnes désireuses de se former et de tuteurs/formateurs et en cataloguant des ressources pédagogiques.
32. Conçu sur les principes des « jeux sérieux », le *FuncampR* propose une approche ludique de l'apprentissage de *R*. Au cours d'un jeu vidéo, le stagiaire est amené à résoudre des énigmes dont la réponse est apportée dans des tutoriels *R*. Le *FuncampR* est une formation déployée en ligne sur le SSPCloud en utilisant la technologie des conteneurs.

accompagnée par la publication, par les mêmes auteurs, d'une librairie *R* ou *Python* implémentant leurs méthodes. Le partage du savoir statistique s'effectue donc de façon duale, à la fois scientifique mais aussi technique.

Le SSPCloud a fait le choix de fonder ses services, exclusivement, sur des briques *open source* – logiciels statistiques, systèmes de gestion de base de données, outils de développement, etc. Il s'agit en effet de s'assurer de la portabilité des travaux menés sur le SSPCloud vers n'importe quel autre environnement de *data science*, sans barrière de propriété, là où l'introduction de logiciels commerciaux conduit inéluctablement à limiter les possibilités de réutilisation.

De façon emblématique, l'interface graphique du SSPCloud est elle-même le fruit d'un projet collaboratif *open source*[33] développé à cette occasion par l'Insee. Ce projet logiciel, intitulé *Onyxia*, permet à l'utilisateur de lancer ses services, de gérer ses dépôts de données, de paramétrer ses ressources ***(figure 1)***. Elle est conçue pour pouvoir être déployée sur d'autres infrastructures utilisant des technologies de conteneurs, et peut ainsi être réutilisée par d'autres directions informatiques désireuses de constituer des services de même nature.

NOUER DES PARTENARIATS TECHNIQUES, POUR INSPIRER ET S'INSPIRER

Fidèle à sa vocation de fabrique créative, le SSPCloud se veut une pierre de touche pour aider le statisticien à découvrir de nouvelles technologies, mais aussi pour permettre aux ingénieurs informatiques d'imaginer de nouvelles architectures de traitement du chiffre.

Bac-à-sable pour ses utilisateurs finaux, le SSPCloud l'est également pour ses concepteurs, dans une posture de veille sur les nouvelles opportunités techniques à prendre en compte. La démarche du SSPCloud vise ainsi à nouer des partenariats technologiques avec des tiers, tant au niveau du système statistique européen qu'auprès d'autres acteurs spécialisés dans la manipulation et l'exploration de données (centres de recherche, observatoires). L'ensemble du projet, fondé sur l'*open source*, permet d'en imaginer une réutilisation par de futurs partenaires, créant des instances propres tout en contribuant à améliorer, avec leurs retours d'expérience, la conception d'ensemble du SSPCloud. C'est ainsi qu'Eurostat a mis en place en 2021, sur la base du code source partagé, un environnement expérimental de traitement de la donnée, engageant la dynamique d'échanges au cœur de l'intention portée par les équipes du SSPCloud.

33. Voir *https://github.com/InseeFrLab/onyxia-ui*.

BIBLIOGRAPHIE

ANDERSON, John C. et VENKATRAMAN, N., 1990. *Strategic alignment: a model for organizational transformation through information technology.* [en ligne]. Novembre 1990. Center for Information Systems Research, Massachusetts Institute of Technology. CISR WP N°217. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://dspace.mit.edu/bitstream/handle/1721.1/49184/strategicalignme90hend.pdf*.

DEAN, Jeffrey et GHEMAWAT, Sanjay, 2004. *MapReduce: Simplified Data Processing on Large Clusters.* [en ligne]. OSDI'04, Sixth Symposium on Operating System Design and Implementation, pp. 137-150. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://research.google/pubs/pub62.pdf*.

DINUM et INSEE, 2021. *Évaluation des besoins de l'État en compétences et expertises en matière de donnée.* [en ligne]. Juin 2021. Insee, Direction interministérielle du numérique. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.numerique.gouv.fr/uploads/RAPPORT-besoins-competences-donnee.pdf*.

EUROSTAT, 2018. Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics). In : *site d'Eurostat*. [en ligne]. 12 octobre 2018. 104e conférence des directeurs généraux des instituts nationaux statistiques (DGINS). [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/european-statistical-system/-/dgins2018-bucharest-memorandum-adopted*.

MÉRINDOL, Valérie, BOUQUIN, Nadège, VERSAILLES, David W. et alii, 2016. *Le Livre blanc des Open Labs. Quelles pratiques ? Quels changements en France ?* [en ligne]. Mars 2016. Futuris, PSB. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.researchgate.net/profile/Ignasi-Capdevila/publication/301356114_Les_open_labs_de_la_recherche_et_de_l'enseignement_superieur/links/571501de08aedc7cbcc99e6d/Les-open-labs-de-la-recherche-et-de-lenseignement-superieur.pdf*.

NIST, 2017. *NIST Big Data Program*. [en ligne]. Mis à jour le 11 janvier 2017. The National Institute of Standards and Technology (NIST). U.S. Department of Commerce. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://bigdatawg.nist.gov*.

ROBINSON, Alan G. et STERN, Sam, 1997. *Corporate creativity: How innovation and improvement actually happen*. Éditions Berrett-Koehler.

SCHON, Donald A., 1963. *Champions for Radical New Inventions*. Mars-avril 1963. Harvard Business Review.

SUAREZ CASTILLO, Milena, SEMECURBE, François, LINO, Galiana, COUDIN, Élise et POULHES, Mathilde, 2020. *Que peut faire l'Insee à partir des données de téléphonie mobile ? Mesure de population présente en temps de confinement et statistiques expérimentales.* [en ligne]. 15 avril 2020. Blog Insee. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://blog.insee.fr/que-peut-faire-linsee-a-partir-des-donnees-de-telephonie-mobile-mesure-de-population-presente-en-temps-de-confinement-et-statistiques-experimentales/*.

UNECE, 2013a. *Utilisation des « données massives » dans les statistiques officielles. Note du secrétariat.* [en ligne]. 18 mars 2013. Groupe de haut niveau sur la modernisation de la production et des services statistiques. Conférence des statisticiens européens des 10-12 juin 2013, Genève. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://statswiki.unece.org/download/attachments/77170614/Big%20Data%20Published%20version%20FR.pdf?version=1&modificationDate=1370507699015&api=v2*.

UNECE, 2013b. *Big data and modernization of statistical systems, Report of the Secretary-General.* [en ligne]. 20 décembre 2013. 45th Statistical Commission. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://unstats.un.org/unsd/statcom/doc14/2014-11-bigdata-e.pdf*.

UNECE, 2021. *HLG-MOS Machine Learning Project.* [en ligne]. Mis à jour le 13 octobre 2021. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project*.