Accès aux bases de données d'enquêtes statistiques et de recensements

Approche méthodologique et solutions mises en œuvre à l'Agence Nationale de la Statistique et de la Démographie (ANSD) du Sénégal

Dr Bourama Mane*

Dans un contexte de sortie de crise marquée par la pandémie de la COVID19, les administrations en charge de la production des données statistiques dans les pays africains sont amenées à revoir leurs méthodes de collecte et d'accès aux microdonnées par les utilisateurs (internes et externes). Les orientations dans l'élaboration des politiques publiques et la planification des projets et programmes de développement devraient être basées sur l'utilisation de données statistiques fiables et accessibles pour la prise de décision.

Aujourd'hui, tout le monde est conscient de l'existence de données statistiques au niveau des Instituts Nationaux de Statistique (INS) africains, mais il faut noter que ces données sont généralement dans un état partiellement utilisable, souvent obsolètes ou pas assez désagrégées pour les besoins d'analyses approfondies.

Il s'y ajoute la problématique de l'accès aux microdonnées par les experts des INS en charge du traitement et des analyses approfondies dans un environnement permettant d'assurer une gestion sécurisée des flux de données, le stockage et l'utilisation de solutions innovantes. Ce besoin d'accès aux microdonnées s'inscrit également dans le contexte de l'agenda 2030 des Nations unies qui recommande un Système statistique national agile et réactif pour tous les pays, répondant ainsi aux demandes accrues des utilisateurs de données.

Notre réflexion s'inscrit dans ce contexte et tentera d'apporter une réponse visant à combler les lacunes en termes d'accès aux microdonnées pour des analyses approfondies au niveau des INS africains. Il s'agira de développer des stratégies pour l'utilisation durable de nouvelles méthodes robustes et d'outils innovants qui faciliteront la mise à jour des bases de données, la centralisation des différentes sources de données pour une exploitation continue et en temps réel par les analystes et les spécialistes du traitement des données statistiques, qu'ils soient internes ou externes. Les mesures prises fourniront un cadre permettant l'analyse des données et faciliteront leur diffusion à l'aide d'outils de communication interactifs modernes.

Dans cet article, au-delà du diagnostic de la situation relative à l'accès aux microdonnées, nous présenterons une approche méthodologique pour l'accès sécurisé aux microdonnées de la production statistique du Système Statistique Nationale (SSN) du Sénégal avec l'utilisation de plateformes généralement basées sur l'open-source. Le cas pratique de mise en œuvre concernera les données de la production statistique de l'Agence Nationale de la Statistique et de la Démographie (ANSD) du Sénégal, l'institution en charge de la coordination du SSN.

* Dr Bourama Mane est Chef de la Division Informatique à la Direction des Systèmes d'Information et de la Diffusion, à l'Agence nationale de la Statistique et de la Démographie (ANSD).

INTRODUCTION

De nos jours, la production et la diffusion de données statistiques constituent un enjeu majeur pour beaucoup de pays africains préoccupés par le besoin de préserver leur souveraineté nationale et d'assurer un développement harmonieux.

L'accès à une information statistique fiable est ainsi devenu une nécessité pour la planification des politiques publiques et privées. En effet, les données d'enquêtes et de recensement doivent contribuer à la définition des politiques sur la sécurité alimentaires, les questions de genre, la promotion de la production agricole, l'investissement, la croissance économique, le développement rural, entre autres.

Il se trouve que ces instituts nationaux statistiques rencontrent des difficultés pour la mise en œuvre de ces opérations de collecte de données statistiques car ils dépendent souvent de financements externes.

L'importance et l'utilité de disposer de données statistiques fiables pour ces pays ne sont plus à démontrer mais ils sont souvent confrontés à la problématique d'allocation des ressources budgétaires qui obéissent à de multiples critères et souvent à des priorités conflictuelles. L'avènement de la pandémie de la COVID19 justifie à suffisance l'impérieuse nécessité de disposer de données statistiques en temps réel pour une prise de décision rapide et une adaptation face à des situations inédites.

Par ailleurs, en plus des contraintes liées à la disponibilité de moyens pour la mise en œuvre d'enquêtes nationales ou de recensements, les INS africains font aussi face aux obligations d'ordre juridique et réglementaire pour la diffusion des données statistiques surtout à caractère personnel [3].

A ce titre pour le cas spécifique du Sénégal, il convient de citer :

- Loi n° 2004-21 du 21 juillet 2004 portant organisation des activités statistiques qui définit le cadre pour l'accès et la diffusion des microdonnées à des fins de recherche ;
- Loi n°2012-03 du 03 janvier 2012 modifiant et complétant la Loi n° 2004-21 du 21 juillet 2004 portant organisation des activités statistiques avec nécessité de mettre en place un dispositif permettant de lutter contre les atteintes à la vie privée susceptibles d'être engendrées par la collecte, le traitement, la transmission, le stockage et l'usage des données à caractère personnel ;
- Décret n° 2008-721 du 30 juin 2008 portant application de la loi n° 2008-12 du 25 janvier 2008 sur la protection des données à caractère personnel qui institue une autorité administrative indépendante dénommée **« Commission des Données Personnelles » (CDP)** qui est le garant du respect de la vie privée dans le traitement des données personnelles ;
- L'élaboration des Stratégies nationales des Données, de l'Intelligence Artificielle et du Développement de la Statistique.

Au niveau sous régional, la CEDEAO a aussi adopté l'acte additionnel «A/SA.1/01/10» du 16 février 2010 relatif à la protection des données à caractère personnel. Elle recommande ainsi à ses États membres de mettre en place un cadre légal de protection de la vie privée et professionnelle, consécutive à la collecte, au traitement, à la transmission, au stockage et à l'usage des données à caractère personnel.

Plus généralement pour le continent africain, la Charte Africaine de la Statistique (CAS), qui est entrée en vigueur depuis le 21 mai 2014, ratifiée par plusieurs pays africains, stipule que les administrations en charge de la gestion des données statistiques doivent garantir l'accès aux statistiques africaines. Ce droit d'accès pour tous les utilisateurs, sans aucune restriction, doit être garanti par l'adoption de lois au niveau national. Les fichiers de données peuvent être mis à la disposition des utilisateurs à condition que les lois et les procédures clairement définies soient respectées et que la confidentialité soit garantie.

Tenant compte de ce contexte, l'anonymisation des sources de données s'impose ainsi aux gestionnaires de données pour la préservation du caractère confidentiel des informations qu'elles contiennent. Le traitement consiste à ne pas permettre l'identification d'un individu dans un échantillon. Elle permettra de se conformer aux lois et règlements en vigueur pour le partage des données en assurant le maintien de la confiance des répondants avec la protection des informations à caractère personnel.

Cependant, il convient de noter que l'application des techniques d'anonymisation de microdonnées peut été à l'origine d'une perte d'information qui affecte l'utilité des données. Le principal défi pour les organismes de statistique consistera dans ce cadre à appliquer des techniques optimales qui réduisent les risques de divulgation avec une perte minimale d'information tout en préservant l'utilité des données.

L'application de ces techniques d'anonymisation aboutirait à la mise à disposition des types de données suivants :

- Les microdonnées à l'usage public ;
- Les microdonnées accessibles sur place (data enclave) ;
- Les microdonnées à accès sur licence.

Les microdonnées à usage public contrairement aux deux autres formes d'accès doivent être disponibles en téléchargement libre sans aucune nécessité d'authentification du demandeur.

En conséquence, le besoin de disposer d'un personnel qualifié en mesure d'effectuer ces traitements requis s'impose et constitue un fort enjeu pour certains pays en termes de diffusion de l'information statistique. Aussi, la compréhension du besoin des utilisateurs pour l'accès à l'information se pose avec acuité dans un contexte marqué souvent par l'absence de stratégie en termes de production et de diffusion de données statistiques. Cette situation est aussi à l'origine de l'exploitation marginale des sources de données issues d'enquêtes ou de recensements.

C'est donc là tout l'intérêt pour les instituts nationaux de statistiques d'Afrique, d'adopter des stratégies tendant à mettre en place un dispositif pérenne de production et de diffusion de données statistiques. Ces stratégies sont à mettre dans le cadre d'une vision globale consistant à permettre au pays de fournir des données de meilleure qualité, plus actuelles et désagrégées, en parfaite conformité avec les cadres légaux portant sur le traitement des données individuelles.

Cet article tentera d'apporter une réponse à cette problématique de l'accès aux microdonnées à partir d'un diagnostic effectué sur la base des expériences connues en matière de collecte et de diffusion de microdonnées. Nous présenterons une approche méthodologique d'accès aux microdonnées de la production statistique du Sénégal ainsi que l'implémentation informatique faite pour la centralisation, le stockage, l'analyse avancée et la diffusion des données statistiques.

I. État des lieux de la production et la diffusion de données du système statistique national du Sénégal

Pour faire face au besoin de plus en plus croissant en informations statistiques de qualité, il faut une redynamisation du Système Statistique National (SSN) du Sénégal. La Stratégie Nationale de Développement de la Statistique (SNDS) [4] a été adoptée dans ce cadre et l'objectif principal est de fournir un cadre cohérent et concerté pour concrétiser la vision du SSN par la consolidation des acquis, une meilleure adéquation de l'offre à la demande, l'amélioration de la gouvernance du SSN et de la gestion stratégique du développement de la statistique publique.

La mise en œuvre de la SNDS a permis au SSN de faire des progrès remarquables et d'adhérer à la Norme spéciale de diffusion des données (NSDD) [5] du Fonds Monétaire International (FMI).

Dans ce contexte marqué par la nécessité de mise à disposition de données statistiques fiables et de qualité, la confidentialité portant sur les données à caractère personnel doit être garantie. Cette situation a été à la base de beaucoup d'initiatives appuyées souvent par des partenaires au développement pour la mise en place de plateformes de diffusion de l'information statistique.

1. La plateforme NADA

Le réseau des pays membres de l'IHSN (International Household Survey Network) [6] a initié la mise en place d'une plateforme d'archivage appelée NADA (National Data Archives). Il s'agit d'un portail Web avec un système de catalogage et d'exploration qui permet aux chercheurs de naviguer, rechercher, comparer, demander l'accès et télécharger des informations pertinentes de recensements ou d'enquêtes. À ce jour, plusieurs pays disposent de la plateforme NADA avec un contenu assez fourni de plusieurs dizaines d'enquêtes et de recensements documentés.

Pour le Sénégal, la mise en œuvre du Programme de Statistique Accélérée (PSA) a été le cadre d'appui de la Banque Mondiale pour mettre en place cette plateforme de diffusion des données d'enquêtes et de recensements dénommée ANADS (Archivage Nationale des Données du Sénégal) [7]. La plateforme permet de classer, d'archiver et de diffuser les données de la production statistique. Elle est accessible via le Web et permet l'intégration des données de certaines structures du SSN comme l'Agriculture, l'Elevage, la Pêche, la Santé, l'Education, et les Eaux et Forêts.

Le dispositif répond ainsi aux besoins des utilisateurs pour l'accès aux métadonnées et un archivage centralisé et sécurisé de toute la production de données statistiques. Ces métadonnées ne sont autres que les informations qui décrivent les caractéristiques basiques d'une donnée ou d'un jeu de données, et ce, quel que soit le support de stockage (physique ou numérique).

Les utilisateurs ont ainsi accès aux informations relatives à l'environnement de l'enquête, les rapports et les documents techniques, les questionnaires, les manuels, les conditions et les procédures

d'accès aux micros données, en plus des fréquences des différentes variables.

Cependant, si les métadonnées des enquêtes et recensements sont librement accessibles à partir de la plateforme, l'accès aux microdonnées ou données brutes est exclusivement réservé aux utilisateurs agréés. Ces données brutes sont des résultats immédiats d'observations de variables statistiques qui n'ont pas subi de traitement au niveau statistique. Le traitement de ce type de données nécessite une bonne connaissance de la statistique et des outils d'anonymisation (logiciel SPSS, logiciel SAS, R, etc.). Les utilisateurs souhaitant y accéder doivent ainsi faire une demande en ligne qui sera traitée par application des procédures de diffusion en vigueur.

2. La plateforme open data

La Banque Africaine de Développement a aussi contribué à l'amélioration de la politique de diffusion des indicateurs statistiques en apportant son appui à la mise en place d'une plateforme Open Data [8]. Il s'agit d'un portail Web qui présente les données sous forme de graphiques avec une possibilité de téléchargement des jeux de données en format standard (CSV, Excel, etc.). Cette plateforme a été mise en place suite à l'adhésion à la Norme Spéciale de Diffusion des Données (NSDD) [5] lancée en 1996 par le FMI.

Le respect de cette norme, qui impose un certain nombre de contraintes pour la production et la diffusion des données statistiques nationales, montre un certain degré de maîtrise et une exigence de qualité dans la gestion des statistiques.

3. Autres solutions de diffusion de données statistiques

Au-delà des initiatives précédemment citées, d'autres plateformes de diffusion ou d'exploitation des données statistiques ont aussi été mises en place. Pour le cas spécifique du Sénégal, le portail dénommé AGRIDATA [9] pour la diffusion des indicateurs et microdonnées relatifs à l'Agriculture a été développé et mis en ligne pour l'accès à l'information statistique agricole.

La plateforme offre un accès à plusieurs sources de données exhaustives et fiables, sur l'agriculture au Sénégal. L'objectif vise à apporter au secteur agricole sénégalais des statistiques agricoles pertinentes et de qualité pour constituer la base d'une politique de développement économique reposant sur des bases factuelles, tout en responsabilisant les agriculteurs et les autres acteurs de la chaîne de valeur.

L'initiative OPAL (Open Algorithms) [10] vient aussi s'ajouter à cet ensemble pour l'exploitation des données téléphoniques des opérateurs. Le projet vise à fournir à des acteurs tiers des indicateurs de mobilité de façon fiable et éthique. Sa mise en œuvre devrait aboutir au développement :

1) d'une plateforme web pour donner accès aux indicateurs de mobilité ;
2) des indicateurs de mobilité ajustés pour certains biais et mis à l'échelle de la population ;
3) d'un modèle d'affaire pérenne permettant de rendre la solution durable et non uniquement dépendante des fonds des bailleurs.

II. Limites des méthodes de diffusion

De ce qui précède, il convient de noter que les solutions pour l'accès à l'information statistique ne manquent pas pour les producteurs de statistique officielle. Cependant, les solutions existantes ont toutes le trait commun de ne pas permettre l'accès direct aux microdonnées de la production statistiques pour les analyses approfondies avec un respect strict des mesures de protection des données à caractère personnel.

Hors, la raison d'être des INS est de rendre disponible, pour tous les usagers, de la manière la plus large possible, les données statistiques qu'ils produisent. Tenant compte de cette exigence de qualité de service souhaitée, on note un réel besoin de valoriser les produits de la production statistique pour atteindre les objectifs et satisfaire les différents utilisateurs.

À ce titre, au-delà de l'absence de laboratoire de microdonnées répondant aux normes, il convient de noter d'une part, le temps très long de traitement des demandes d'accès aux microdonnées avec notamment le renseignement du formulaire d'engagement nécessaire avant toute mise à disposition. D'autre part, l'application des techniques d'anonymisation [11], conformément aux règles et lois en vigueur [12] contribue pour la plupart à impacter la qualité des données pouvant ainsi compromettre les analyses statistiques issues de ces données.

Pour accomplir leur mission de service publique, les INS doivent ainsi s'inscrire dans une nouvelle dynamique d'exploitation des technologies de l'information pour permettre l'accès aux données de la production statistique, notamment pour ses experts, les structures de l'administration, les chercheurs, les universitaires, les bailleurs de fonds, les associations et ONG. Les initiatives prises dans ce cadre pour le Sénégal font l'objet de cet article qui présentera les nouvelles orientations prises en matière d'accès aux microdonnées et les solutions préconisées.

III. Approches méthodologiques

Les difficultés rencontrées pour l'accès continu aux microdonnées a permis à l'ANSD du Sénégal d'identifier deux groupes d'utilisateurs de la production statistique : les experts statisticiens de l'Agence et les utilisateurs externes. Le personnel interne constitué d'experts statisticiens soumis à une déontologie stricte est souvent confronté aux mêmes difficultés d'accès aux bases de données individuelles consolidées et apurées, lui permettant d'effectuer des analyses approfondies dans les délais requis. Le respect des règles et lois en vigueur voudrait qu'un expert, malgré son appartenance à la structure productrice ne soit en mesure de copier les données sources à l'état brut en local sur son ordinateur de travail, ce qui n'est pas le cas aujourd'hui.

L'approche proposée pour cette catégorie d'utilisateurs des données de la production statistique consiste à mettre en place un environnement distant permettant de travailler sur les bases de données volumineuses de manière efficace, sécurisée et sans contrainte en termes d'accès et de récupération des fichiers de sortie (Outputs). Le dispositif devrait permettre à l'utilisateur de se servir de ses programmes d'analyses et de scripts pour exécuter les requêtes sur toutes les bases de données auxquelles il a l'autorisation d'accéder.

La deuxième catégorie de demandeurs concerne les utilisateurs externes à l'ANSD. Il s'agit pour l'essentiel des consultants internationaux, des chercheurs, des universitaires, des experts des structures de l'administration, des bailleurs, ONG, entre autres. La méthode d'accès aux données de la production statistique préconisée pour cette catégorie consistera à mettre en place un dispositif informatique structuré en trois phases :

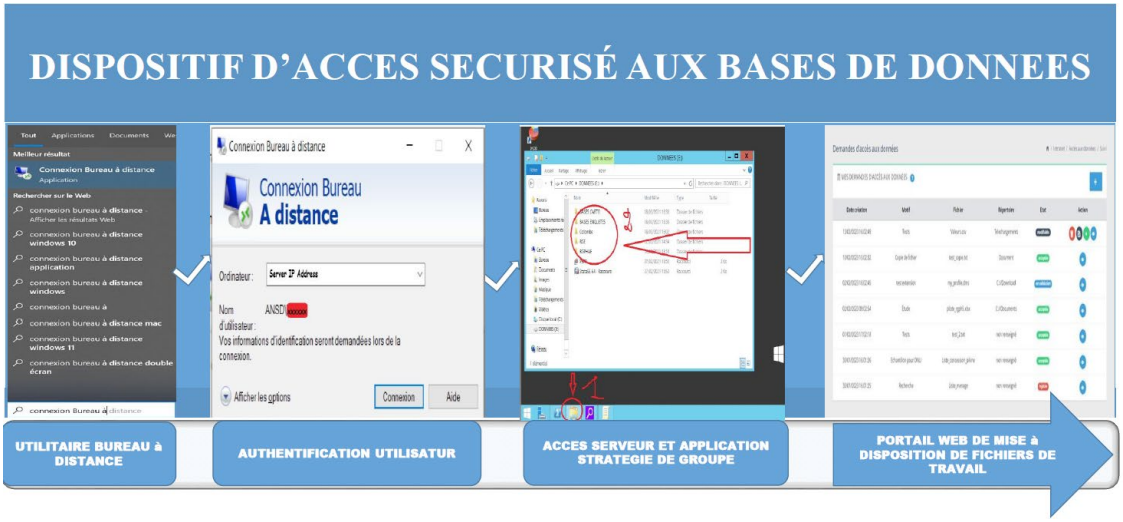
1) mise à disposition de différentes sources de données de l'ANSD, dans des formats variés (Spss, Stata, Csv, BD structurées, etc.) ;
2) la définition du cadre de stockage des données et de l'accès centralisé via une infrastructure sécurisée ;
3) le traitement des données avec possibilité d'accès multiples et la diffusion des résultats des analyses.

La mise en œuvre informatique de ces méthodes est décrite dans les prochaines séquences de cet article.

IV. Résultats de la mise en œuvre informatique

1. Mise en place d'un environnement de travail sécurisé

Pour satisfaire la demande des utilisateurs en termes d'accès sécurisé aux données de la production statistique et renforcer les performances d'exécution des requêtes dans des données de masse, le dispositif décrit ci-dessous a été mis en place afin d'offrir de nouvelles possibilités d'exploitation des données statistiques. Le dispositif se présente comme suit :



Description de la procédure d'accès aux données par les utilisateurs :

a. Le service informatique a mis en place un serveur performant en appliquant des stratégies de groupe qui bloquent la copie de fichiers ou de répertoires de l'ordinateur local vers le serveur et vice versa ;
b. L'organisation de l'entreprise autorise seulement l'administrateur des bases de données et de l'infrastructure à disposer des paramètres d'accès administrateur de ce serveur ;
c. L'utilisateur se connecte préalablement par VPN pour s'assurer du cryptage des informations qui

transiteront lors de la transaction client-serveur. Un compte utilisateur et un mot de passe VPN lui seront communiqués avec une durée de vie limitée. Il suivra ainsi une procédure de connexion VPN pour l'accès à l'infrastructure informatique ;

d. Après avoir réussi l'étape de la connexion VPN, l'utilisateur peut accéder au serveur en se servant de l'utilitaire « Bureau à Distance » de Windows ;

e. L'utilisateur doit encore utiliser un compte utilisateur et un mot de passe d'ouverture d'une session sur le serveur. Ces paramètres d'accès sont aussi préalablement communiqués à l'utilisateur par l'administrateur de bases de données et de l'infrastructure informatique de l'ANSD ;

f. Après cette ***double authentification***, l'utilisateur accède ainsi au contenu du serveur avec les logiciels de traitement déjà installés, lui permettant d'effectuer ses travaux de traitement de données et de stocker ses fichiers Outputs dans un répertoire dédié ;

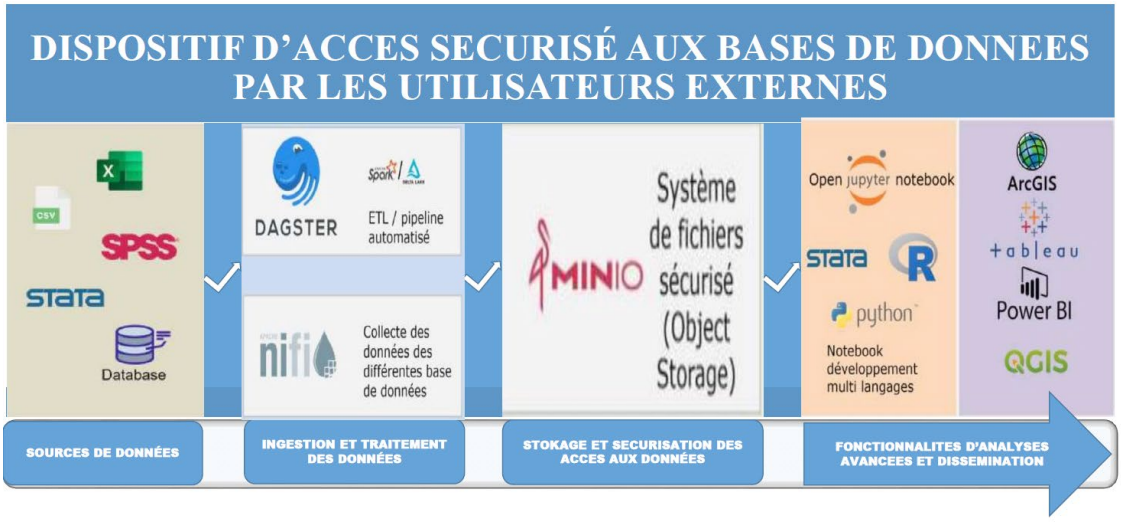
g. L'utilisateur se connecte enfin sur un portail Web dédié pour formuler une demande de récupération de ses fichiers de travail ;

h. Cette demande sera validée par un responsable désigné de l'institution (Directeur des Systèmes d'Information et de la Diffusion – DSID) en offrant ainsi à l'utilisateur, la possibilité de télécharger ses outputs de données. En effet, le DSID en rapport avec ses services effectue une vérification préalable du contenu des outputs avant de les rendre disponible via un lien de téléchargement. Les scripts, programmes et autres prérequis de fichiers dont dispose l'utilisateur peuvent être préalablement copiés sur le serveur par l'administrateur, après vérification pour des raisons sécuritaires.

2. Mise en place d'un portail d'analyse des données pour les utilisateurs

Pour ce qui concerne le besoin d'accès des utilisateurs, les externes en particulier, c'est dans le cadre du projet Data4Now [13] que l'ANSD a expérimenté la mise en œuvre de la méthode ci-dessous décrite. Cette initiative codirigée par la Division de statistique de l'ONU (UNSD), la Banque mondiale, le Partenariat Mondial pour les Données sur le Développement Durable (GPSDD) et le Réseau des solutions de développement durable (SDSN), vise à développer les capacités des responsables des INS à fournir les informations utiles et fiables aux décideurs locaux et nationaux pour réaliser l'Agenda 2030.

Il a ainsi été retenu dans le cadre de ce projet, l'implémentation de cette plateforme informatique personnalisée qui permettra de disposer d'une méthode unifiée de stockage et de traitement de données massives provenant de diverses sources pour centraliser et disséminer l'ensemble des données de la production statistique. L'architecture du dispositif mis en place repose sur l'utilisation d'utilitaires open sources et se présente comme suit :



Description du dispositif mis en place :

1. Sources de données : Identification et centralisation de toutes les données à mettre dans la plateforme ;
2. Ingestion et traitement des données : Utilisation du logiciel libre de gestion des flux de données « Apache NiFi » [14] : il permet de construire, de déployer et de surveiller les flux de données afin de gérer les pipelines de données en temps réel et d'automatiser le mouvement des données entre des systèmes disparates ;
3. Stockage et accès sécurisé aux données avec l'utilitaire Open source MINIO [15] : À partir des utilisateurs du domaine Active Directory de l'entreprise, l'utilitaire MINIO permet de gérer les comptes

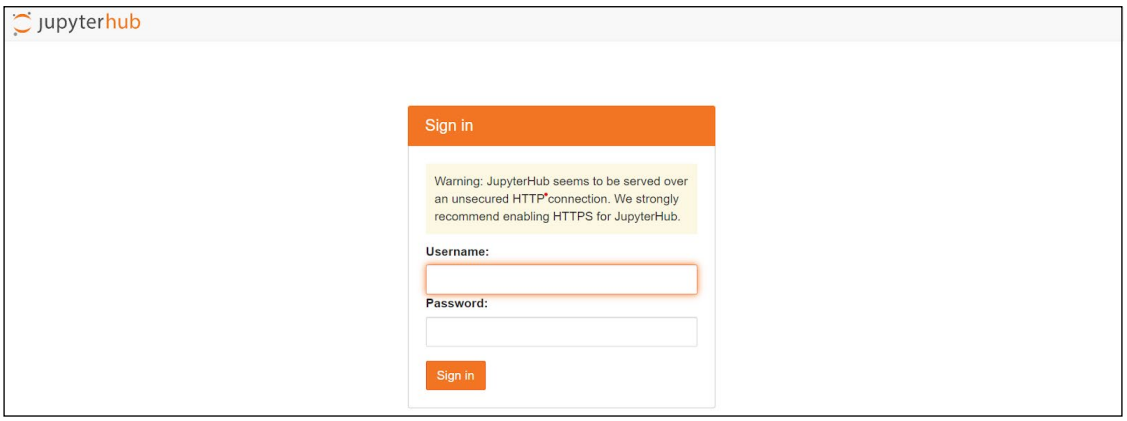
utilisateurs, d'optimiser les conteneurs dans lequel sont stockés des objets au sein de l'environnement (buckets) et de migrer d'énormes quantités de données vers l'emplacement de stockage centralisé appelé « datalake ». On peut ainsi contrôler les accès (créer, supprimer et consigner les objets) ;

4. Transformation des données avec le logiciel open source « Dagster » [16] : il permet la mise en place et déploiement du script de transformation des données et la création de pipelines de données dans le but de les rendre plus exploitables ;
5. Analyses avancées avec les utilitaires Open Source Minio SDK et Jupyterlab [17] : ils permettent l'utilisation d'interface de programmation d'applications communément appelé API, de générer une clé d'API, d'installer le kit de développement logiciel SDK R, de stocker et d'extraire des fichiers. L'utilitaire jupyterlab permet par ailleurs d'utiliser un navigateur pour lire et écrire un fichier CSV, R.data et stata dans le datalake ;
6. Dissémination des données avec le logiciel open source QGIS [20] : ce logiciel permet d'élaborer des résultats sous forme de graphiques pour la diffusion de l'information statistique.

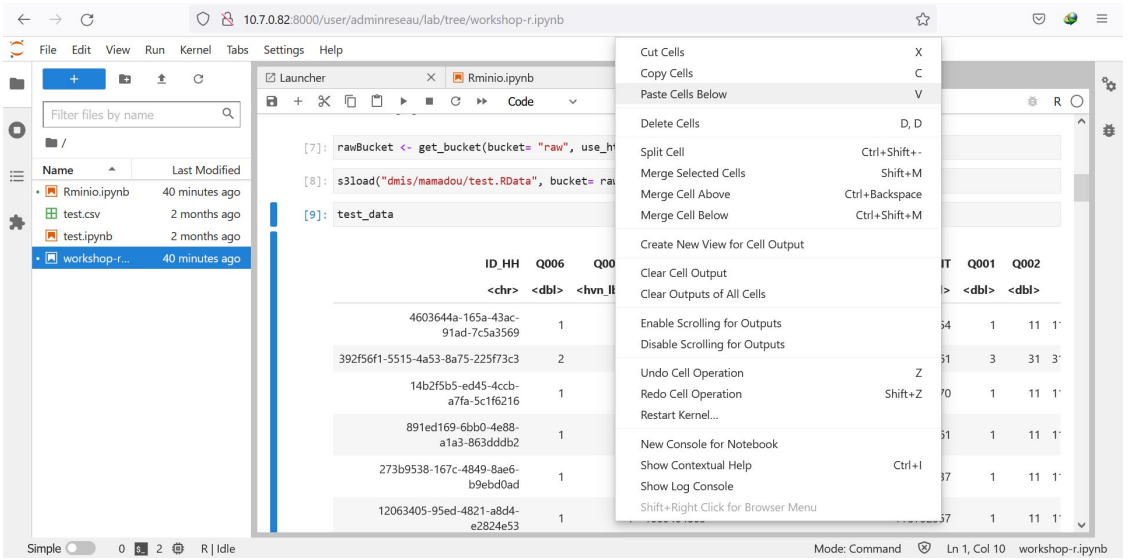
Description de la procédure d'accès aux données par les utilisateurs :

L'accès utilisateur se fait via la plateforme de serveur multi-utilisateurs dénommée Jupyter Hub. Il permet à plusieurs utilisateurs d'accéder au seul serveur. Cela facilite la collaboration entre les utilisateurs et permet de partager facilement des projets et des analyses de données.

Les utilisateurs qui ont accès à cette interface sont ceux de l'entreprise (Domaine Active Directory) et ceux qui ont été créés directement sur JupyterHub (utilisateurs externes).



Après authentification, l'utilisateur accède à l'environnement sécurisé d'analyse des données où il ne peut exploiter que les bases de données dont il a l'autorisation. Les API R, Python et Stata lui permettent d'exploiter tout le potentiel des données sans avoir la possibilité de copier les données sources (brutes ou anonymisées) sur l'ordinateur local. L'option « Copy Output to Clipboard » a été désactivée pour interdire la copie des données depuis l'interface Jupyterlab, comme l'illustre la prise d'écran suivante :



V. Discussion

Les méthodes d'accès aux données adoptées et mises en œuvre par l'ANSD permettent une centralisation et un accès sécurisé à toutes les données de la production statistique. Les experts statisticiens de l'ANSD ainsi que les utilisateurs externes ont désormais un environnement sécurisé de travail leur permettant d'effectuer des analyses approfondies sur les données d'enquêtes et de recensements.

Cependant, des difficultés d'ordre conceptuel demeurent toujours pour certains, considérant qu'après avoir participé à toutes les phases de mise en œuvre d'une enquête ou d'un recensement, l'expert statisticien se retrouve dans une situation où il est tenu de demander à accéder aux données qu'il a lui-même produites.

De ce point de vue, cette façon de voir pourrait constituer un frein pour l'utilisation du dispositif. Pour cerner la propension de telles idées, il convient de rappeler qu'une entreprise pour fonctionner et atteindre ses objectifs a besoin d'une organisation avec des missions et rôles assignés à chaque agent.

À ce titre, la fonction d'administrateur de bases de données ou d'administrateur de l'infrastructure informatique attribue de facto à un agent, la charge de ne permettre l'accès aux données de l'entreprise qu'à une liste d'utilisateurs fournis par l'autorité et de veiller à la sécurisation des données stockées sur les serveurs conformément aux règles et lois en vigueur.

Les utilisateurs externes ont désormais la possibilité d'accéder aux microdonnées et d'effectuer les analyses souhaitées. La question qui se pose est de savoir la plus-value que l'ANSD pourrait tirer de cette mise en œuvre. De prime abord, cela pourrait être l'accès rapide à l'information statistique pour la prise de décision en temps opportun et pour l'élaboration des politiques publiques et privées.

Plus généralement, des initiatives d'accès sécurisés et de protection des données ont été misses en œuvre en France à l'INSEE, notamment pour les chercheurs externes. L'institut de Statistiques du Canada (StatCan) a aussi entrepris un vaste chantier dans ce domaine avec la mise en place d'un guichet unique pour l'accès à certains types de données statistiques. La spécificité de la solution mise en place à l'ANSD réside dans son caractère innovant en se basant sur la performance de l'infrastructure informatique pour offrir un service sécurisé d'accès fluide aux microdonnées.

VI. Conclusion

Le principe d'élaboration de ces méthodes d'accès aux données de l'ANSD est basé sur les pratiques quotidiennes de diffusion et les difficultés rencontrées après la mise en œuvre des projets d'enquêtes et de recensement.

S'appuyant sur les procédures de diffusion existantes, nous sommes partis d'un état des lieux sur la production statistique, la gestion du stockage et la définition des règles de sécurité pour la diffusion de l'information. Nous avons par la suite élaboré les stratégies se conformant au mieux aux dispositions réglementaires pour cerner les contours des méthodes d'accès aux données présentées dans cet article.

Elles se fondent sur l'organisation de la structure avec une acceptation du rôle assigné à chaque acteur mais aussi sur une combinaison intelligente entre la mission de service public de l'Agence et le développement d'une nouvelle approche utilisateur. La démarche adoptée place ainsi l'usager au cœur du système de production et de diffusion des données statistiques.

Références

[1] B. Mane, I. Fall, M. S. Camara, et A. Bah, « Ontological modeling approach for statistical databases publication in linked open data », présenté à Third International Congress on Information and Communication Technology: ICICT 2018, London, 2019, p. 277-292.

[2] « Agenda 2030 des Nations Unies ». [En ligne]. Disponible sur: https://sdgs.un.org/goals

[3] B. Mane, I. Fall, M. S. Camara, et A. Bah, « Definition of the database anonymization method for open data », présenté en 2017 Intelligent Systems and Computer Vision (ISCV), 2017, p. 1-7.

[4] « Une Stratégie Nationale de Développement de la Statistique - SNDS ». https://www.statsenegal.sn/

[5] « Normes du FMI pour la diffusion des données ». https://www.imf.org/fr/About/Factsheets/Sheets/2016/07/27/15/45/Standards-for-Data-Dissemination.

[6] « International Household Survey Network (IHSN) », *Western Michigan University*, 28 février 2019. https://wmich.edu/globalstudies/ihsn.

[7] « Home ». https://anads.ansd.sn/index.php/home (consulté le 26 février 2023).

[8] « openAFRICA ». https://africaopendata.org/.

[9] « Plateforme AgriData du Sénégal ». http://agridata.ansd.sn:8080/index.html#/fr/home.

[10] « OPAL Project », *OPAL Project*. https://www.opalproject.org.

[11] M. Bergeat, « Anonymisation de donnees individuelles: bien calees, bien protegees? ». http://jms.insee.fr/files/documents/2015/s09_2_acte_v2_bergeat_jms2015.pdf

[12] D. T. GROUPE, « Article 29» sur la protection des données », *Doc. Trav. Sur Quest. Prot. Données Liées Aux Droits Propr. Intellect.*, vol. 10092, n° 05.

[13] « Data4Now | Department of Economic and Social Affairs ». https://sdgs.un.org/partnerships/data4now.

[14] « Apache NiFi ». https://nifi.apache.org/.

[15] « MinIO | High Performance, Kubernetes Native Object Storage ». https://min.io/.

[16] « Dagster | Cloud-native orchestration of data pipelines ». https://dagster.io/.

[17] « Software Development Kits (SDK) — MinIO Object Storage for Linux ». https://min.io/docs/minio/linux/developers/minio-drivers.html.

[18] « Tableau: Business Intelligence and Analytics Software », *Tableau*. https://www.tableau.com/node/62770.

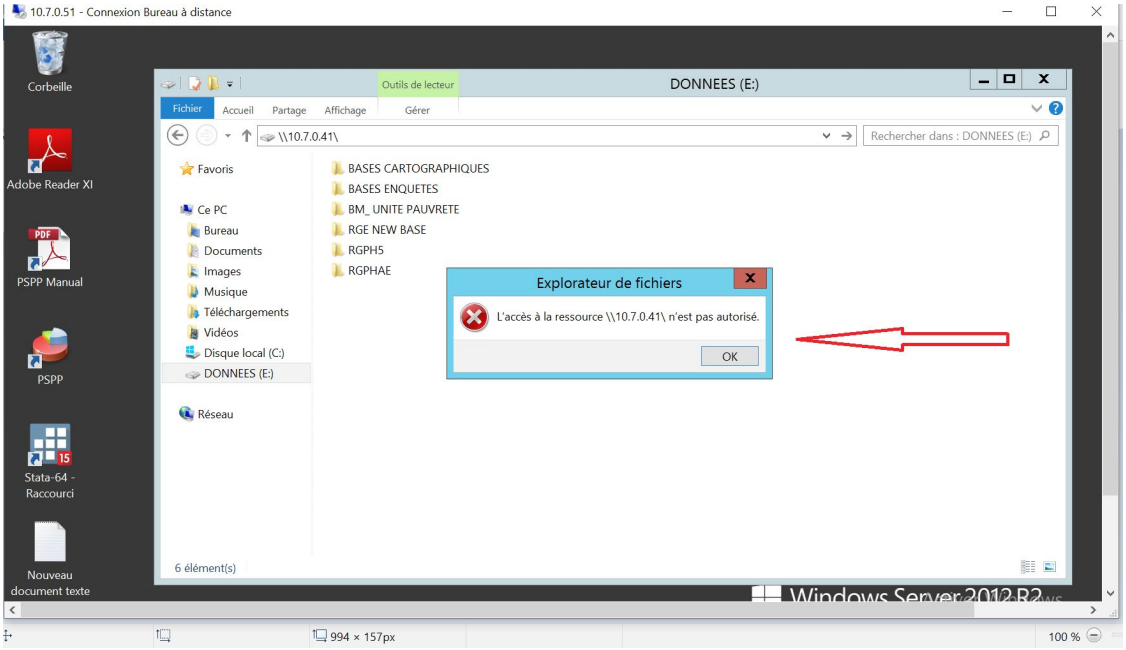
[19] « Data Visualization | Microsoft Power BI ». https://powerbi.microsoft.com/en-us/

[20] « Welcome to the QGIS project! » https://www.qgis.org/en/site/.

Annexe : Illustration de mise en œuvre

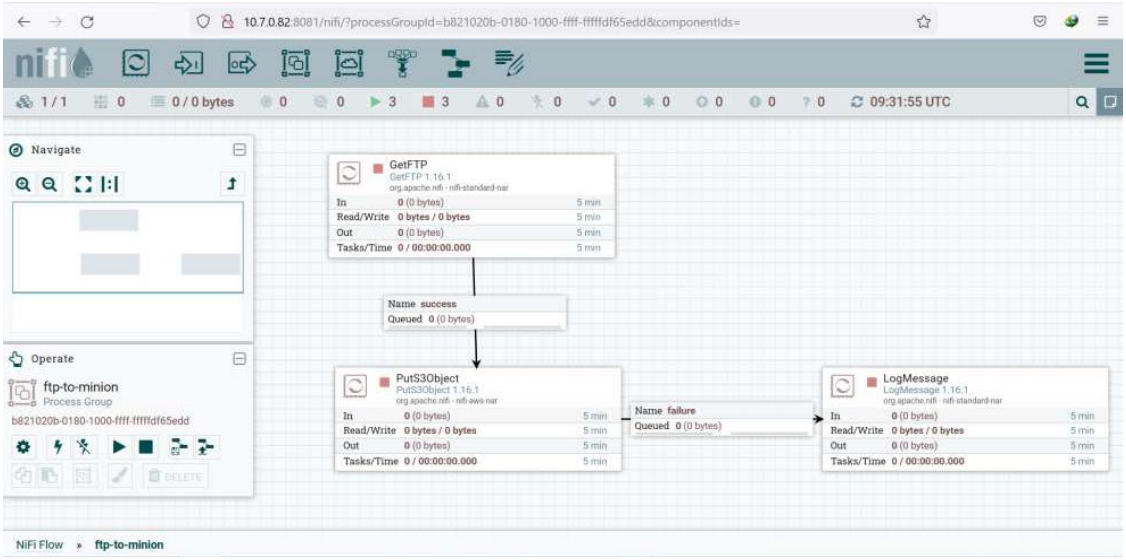
1. Serveur d'accès distant interne

Copie par réseau non autorisé à partir du serveur d'accès distant :

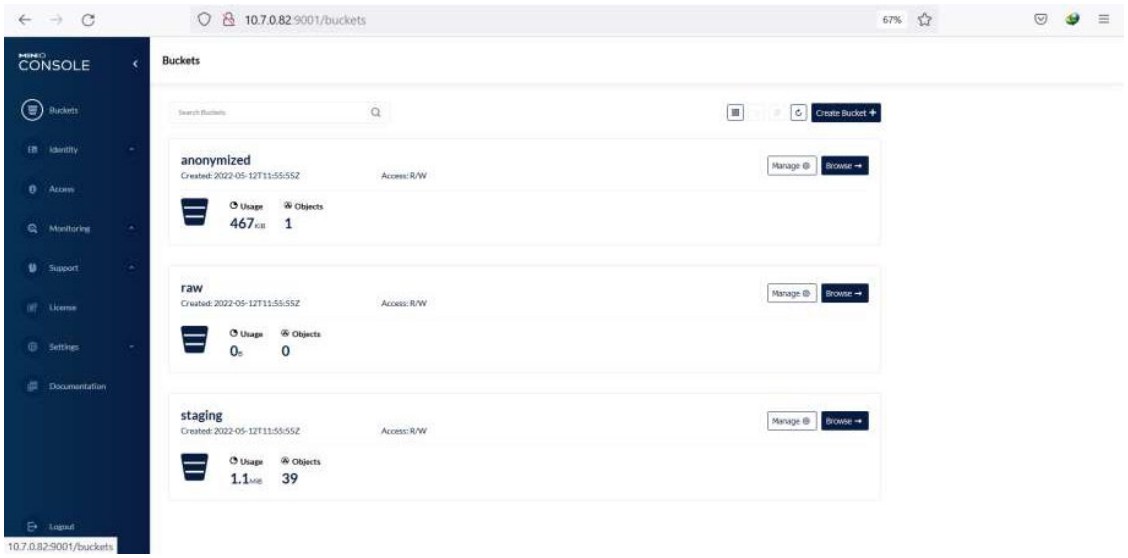


2. Plateforme d'accès externe

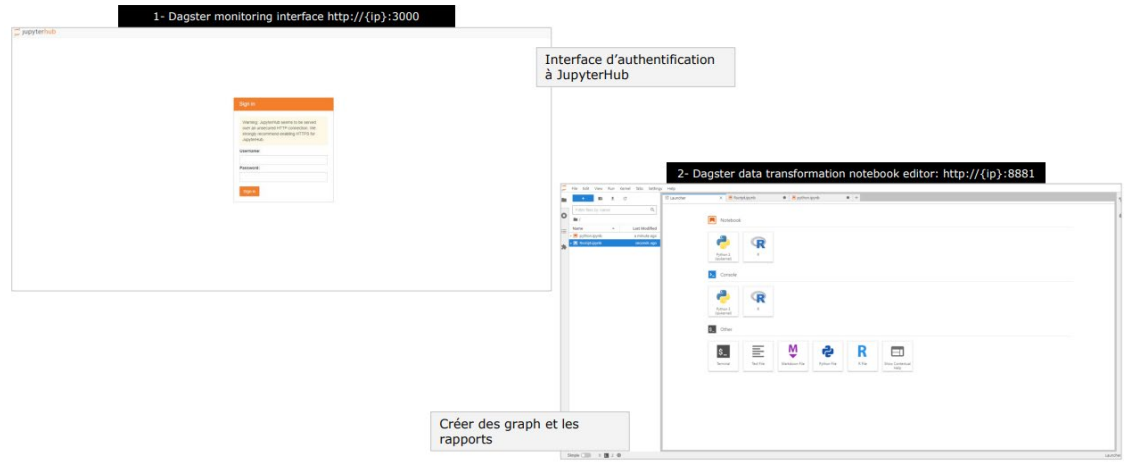
2.1. Data ingestion (Apache NiFi)

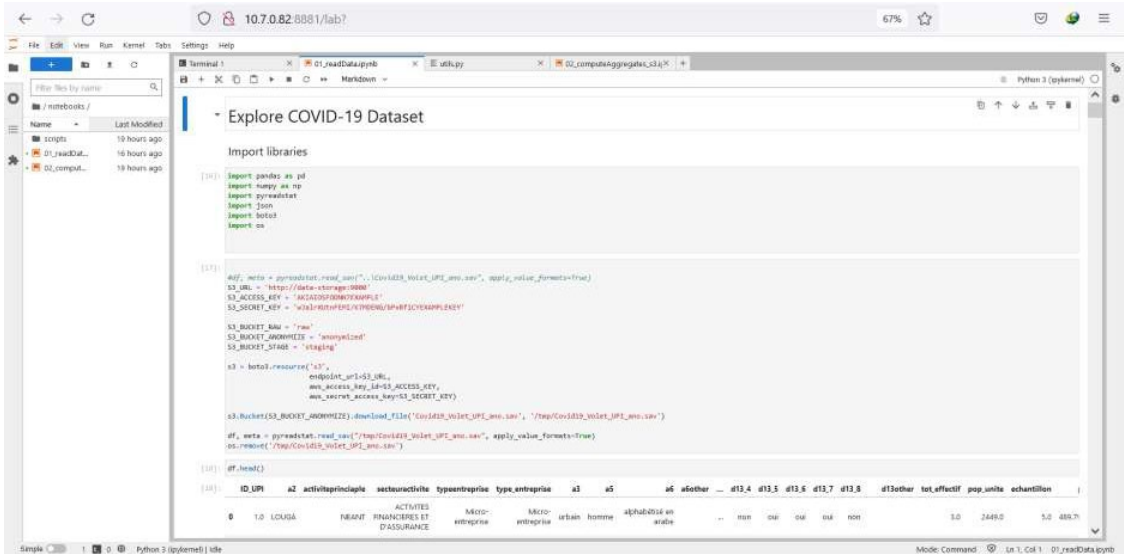


2.2. Data storage (minIO)



2.3. Data modeling and analysis (Jupiter notebook, Python)





2.4. Data dissemination (Dashboard)