UNIVERSITÉ PARIS 8
VINCENNES-SAINT-DENIS

framatome

Rapport d'Année Alternance

Master : Cyber Sécurité et Sciences des Données

AUTEUR :
ETTADRARTY Mohamed

TITRE :
Conception des KPI pour le suivi de la qualité documentaire d'un projet dans l'industrie nucléaire.

ENTREPRISE :
FRAMATOME
1 Place Jean Millier-Tour Areva
92300 COURBEVOIE

TUTEUR ENTREPRISE :
MAGIONCALDA Roberto

TUTEUR PÉDAGOGIQUE :
BOUBCHIR Larbi

Année Universitaire : 2023/2024

Remerciements

Mes remerciements s'adressent, tout d'abord, à mon tuteur d'alternance, Monsieur MAGIONCALDA Roberto, pour la confiance qu'il m'a accordé lors de l'entretien d'embauche, ainsi que son suivi et son soutien au cours de cette expérience professionnelle.

Je remercie également M. BOUBCHIR Larbi, directeur du parcours Cybersécurité et Science de données (CSSD) , ainsi que tous les autres enseignants qui ont contribué à ma formation académique.

Je tiens à remercier également mes collègues chez Framatome pour leur esprit collaboratif et leur engagement. Le partage de connaissances ont enrichi mon expérience professionnelle.

Résumé

La gestion documentaire dans les projets de Framatome nécessite la création des KPIs pertinents afin d'assurer le suivi et la qualité des processus documentaires. Mon rôle au sein de l'équipe documentation est de concevoir, d'analyser et de fournir des KPIs performants, permettant de surveiller en continu les différents aspects de la documentation. Ces indicateurs sont essentiels pour garantir le respect des standards et pour assurer la gestion maitrisée de l'information et des données tout au long des différentes phases du projet.

En complément de ces missions, je participe également à la mise en place d'un data lake dans le cadre d'un cas d'usage porté par la Business Unit PCM. Cette solution vise à standardiser les données de la BU et à faciliter l'accès à la donnée pour les différents métiers, en s'inscrivant dans une démarche de modernisation et d'optimisation de l'architecture data de l'entreprise.

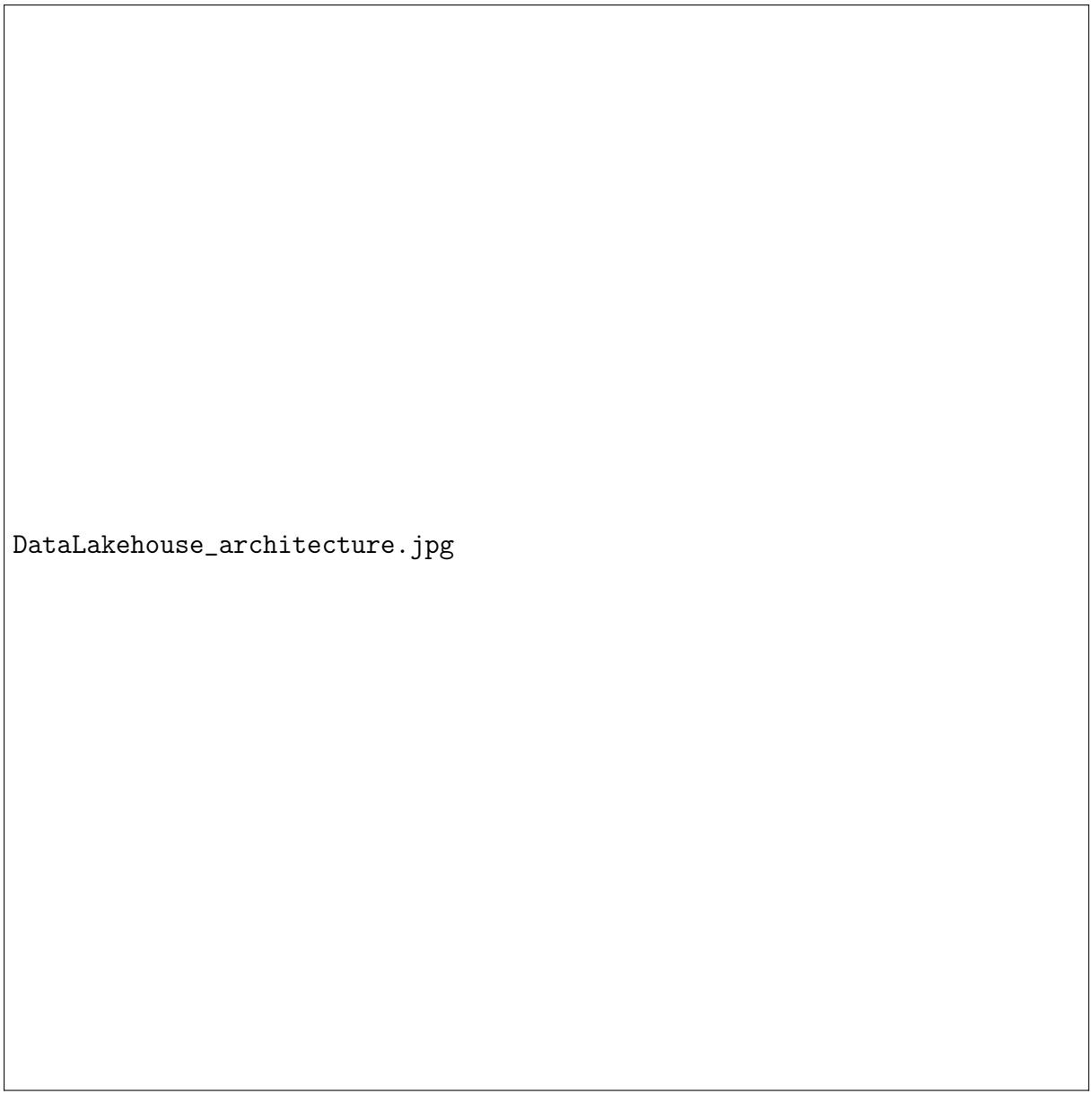


FIGURE 1 – Enter Caption

Sommaire

1 Glossaire

Terme Framatome	Signification (Français/English)
BU	Business Unit
COEDM3	Nom du logiciel documentaire interne à Framatome
DT	Direction Technique
EPR	European Pressurized Reactor
GBR001	Référence pour le projet Hinkley Point C
GBR002	Référence pour le projet Sizewell C
GED	Gestion Electronique Documentaire
HPC	Hinkley Point C
INF	Pour Information
KPI	Key Performance Indicator
LOD	List Of Documents
NHPS	Nuclear Heat Production System
OBS	Pour Observation
PBI	Power BI
PM	Project Manager
PMO	Project Management Office / Project Management Officer
SZC	Sizewell C
UK	United Kingdom
IB	Installed Base
DT	Deriction technique
SOGOUD	Standardisation des Outils et Gouvernance de la Donnée

Table 1 – Signification des termes utilisés chez Framatome

2 Introduction

Pendant ma 1ère année d'alternance chez Framatome, entreprise leader dans le domaine de l'énergie nucléaire, j'ai eu l'opportunité de rejoindre l'équipe documentaire des projets Henkly Point C (HPC) et Sizewell (SCZ) en tant que Data Manager pour contribuer à l'amélioration du suivi et de la gestion des données sur la documentation à travers la création d'indicateurs de performance ou *key performance indicators* en anglais (KPI).

Ce rapport a pour objectif de décrire en détail les missions et les activités réalisées cette année, en mettant l'accent sur le développement et l'implémentation des KPI adaptés aux besoins spécifiques du secteur nucléaire.

Dans un environnement aussi complexe que celui de l'ingénierie nucléaire, les documents techniques sont essentiels pour la bonne gestion du projet, ce qui exige une approche méthodique et rigoureuse pour garantir la qualité et la fiabilité de l'information. Les KPI qu'on a développés contribuent de manière importante au suivi de la performance documentaire, permettant une évaluation précise et continue des processus mis en place.

Ce travail est structuré en différentes sections : après une présentation générale de Framatome et du contexte de mon alternance, on détaille les méthodes et les outils utilisés pour développer les KPI. Pour conclure, on fait une évaluation des résultats obtenus, on présente les compétences acquises également.

2 Description de l'entreprise Framatome

2.1 Histoire

Framatome est créée en 1958 par un groupement de sociétés internationales ayant comme secteur d'activité le nucléaire. En 1975, cette entreprise est choisie comme seul constructeur des centrales nucléaires en France et équipe donc 58 réacteurs à eau pressurisée (REP) français. En 2006, l'entreprise est rebaptisée « Areva NP ». La société est alors spécialisée dans les chaudières nucléaires et les services aux réacteurs. C'est en 2016 qu'un accord est trouvé entre les deux entreprises EDF et Areva pour le rachat de la branche Areva NP qui sera renomée, en 2018, Framatome. Aujourd'hui, Framatome est spécialisée dans :

- La conception, la réalisation et l'amélioration des réacteurs nucléaires ;
- Les équipements et les gros composants pour des nouveaux réacteurs nucléaires ;
- La conception et la fabrication des assemblages de combustibles ;
- Les services aux exploitants de centrales nucléaires, notamment pour la maintenance des réacteurs.

2.2 Composition de l'entreprise

Aujourd'hui, Framatome regroupe environ 18000 salariés à travers 20 pays et sur plus de 60 sites. Elle est à l'origine de la fabrication des équipements de 92 centrales nucléaires dans le monde. Cette entreprise est regroupée en différentes Business Units (BU) et chacune représente des activités clés de la société :

- Base Installée (IB) : Produits et services pour la maintenance, la modernisation et la prolongation de la durée d'exploitation des centrales ;
- Combustible (FL) : Développement, conception, licensing et fabrication d'assemblage de combustible et de composants pour les réacteurs ;
- Projets & Composants (PCM) : Conception et fabrication des composants lourds et mobiles de l'îlot nucléaire. Gestion et exécution des projets de nouvelles constructions de réacteurs nucléaires ;
- Direction Technique et Ingénierie (DTI) : Développement, conception et licensing des chaudières nucléaires ;

— Contrôle-commande (I&C) : Conception et fabrication de technologies d'automatisation et d'instrumentation pour une exploitation sûre, durable et économique des centrales nucléaires.

Sites de Framatome en France :

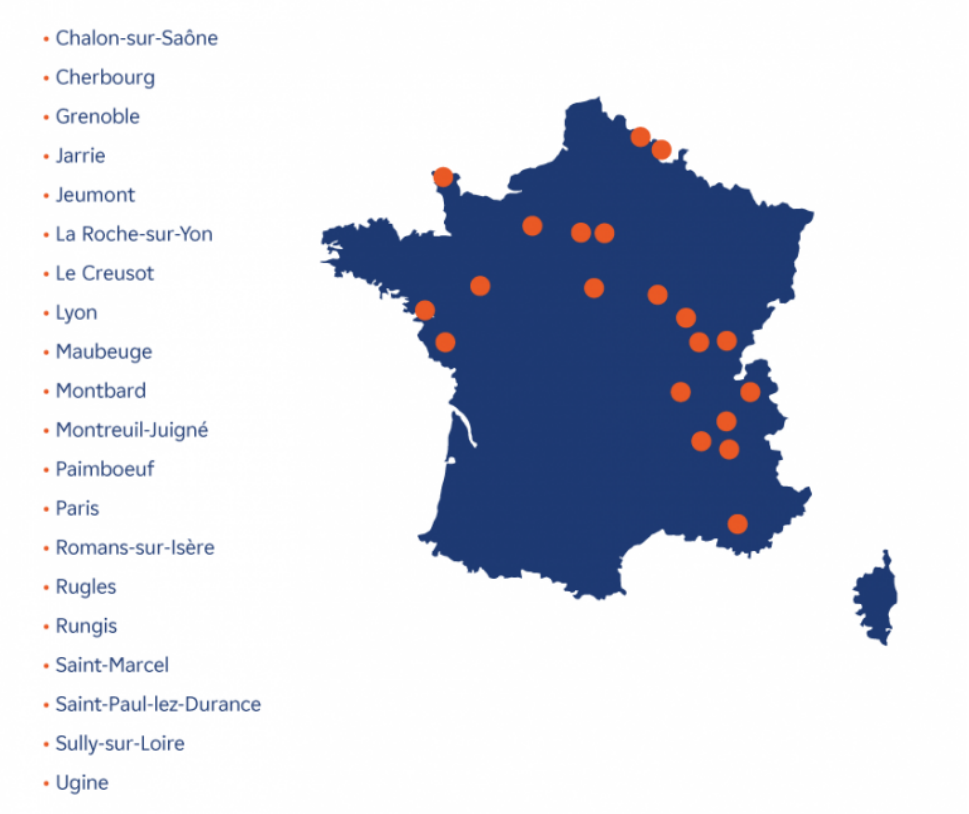


FIGURE 2 – Principaux sites de Framatome en France

3 Contexte et enjeux

3.1 Programme UK

Le programme UK englobe deux projets de construction de centrales nucléaires de type EPR (European Pressurized Reactor) en Angleterre, dirigées par EDF Energy. Le premier projet, la centrale nucléaire de Hinkley Point C (HPC), est en cours de construction. Le deuxième projet, celui de Sizewell C (SZC), est en phase d'étude et de conception, et c'est sur ces deux projets que je travaille cette année. Ces deux sites ont pour objectif de fournir de l'électricité à plus de 12 millions de foyers pendant 60 ans et de contribuer à la neutralité carbonne du pays. La livraison d'Hinkley Point C est prévue entre 2029 et 2031, et celle de Sizewell C est prévue autour de 2035.

3.2 Importance de la gestion documentaire

Dans le cadre de projets industriels majeurs, tels que la construction de centrales nucléaires, le volume des documents techniques à gérer est très important, et les question réglementaires, de conformité et de qualité sont complexes. En conséquence, les ingénieurs ne sont plus en mesure de maitriser leurs documents. Il faut avoir une équipe documentaire dédiée à la gestion de cette documentation qui est un facteur clé pour l'avancement du projet.

Ainsi, la documentation technique doit être organisée rigoureusement, pilotée et mise à jour tout au long du projet selon les différentes phases de conception, de construction, et d'exploitation.

Une gestion documentaire efficace permet de garantir que l'ensemble des acteurs du projet disposent des informations à jour, fiables et intègres, réduisant ainsi les risques d'erreurs, de retards et de non-conformités. Cela est particulièrement important dans l'industrie nucléaire, où la conformité aux normes de sécurité et la sureté sont essentielles pour la réussite du projet.

3.3 Besoin des KPIs pour le suivi documentaire

La mise en place d'indicateurs clés de performance (KPI) est indispensable pour une gestion et un suivi efficaces des documents, c'est-à-dire, pour le pilotage de la documentation : ils permettent de connaitre l'avancement du projet, d'identifier les points de blocage, d'alerter en cas de dysfonctionnement, et ils contribuent à une meilleure planification. Ils permettent par la suite de mettre en place des actions correctives.

Enfin, les KPI sont des outils essentiels pour contrôler, évaluer et améliorer les performances des processus liés aux documents, en veillant à ce que tous les documents soient traités efficacement et répondent aux normes de sécurité et de qualité exigées dans ce domaine. L'importance des KPI dans le pilotage documentaire réside dans leur capacité à fournir des informations claires et mesurables sur l'efficacité des processus et des systèmes en place. En suivant des KPI spécifiques, les organisations peuvent identifier rapidement les problèmes potentiels, respecter les normes élevées de qualité et de conformité, et s'assurer que toute la documentation contractuelle est à jour et accessible.

4 Missions

4.1 Détermination des besoins

Ma première mission consiste à comprendre et à déterminer les besoins des équipes métiers (équipe Project Manager Office, PMO, et équipe documentation). Avec ces équipes, on organise des réunions de cadrage, auxquelles j'assiste régulièrement, afin de discuter des processus actuels, identifier les défis rencontrés, et définir les objectifs à atteindre en ciblant les points le plus importants et les besoins d'amélioration en matière de suivi de projets et de gestion documentaire. Ces réunions entre équipes sont indispensables pour comprendre les besoins initiaux des PMO et de l'équipe documentaire, qu'il faut savoir ensuite traduire en exigences fonctionnelles précises pour la création ou l'amélioration des KPI en priorisant les fonctionnalités critiques selon les contraintes techniques.

Une fois que les KPI sont développés, on présente aux parties prenantes pour validation les conclusions et le travail effectue aux parties prenantes pour sa validation. Il faut que les KPI atteignent leurs exigences préalablement définies. Ensuite, si les résultats ou les indicateurs fournis ne sont pas satisfaisants à 100%, il faut implémenter les ajustements nécessaires en fonction des retours pour assurer l'alignement avec les attentes des équipes.

4.1 Collecte de données

En tant que Data Manager, la collecte de données est une partie fondamentale de mon travail. Les données que j'ai traitées provenaient de diverses sources hétérogènes, chacune ayant ses propres caractéristiques en termes de structure, de format et de mises à jour.

Sources de données :

— *SQL Server* : J'ai extrait des données de plusieurs bases de données relationnelles hébergées sur SQL Server. Les requêtes qu'on a élaborées étaient complexes et visaient à agréger et à filtrer les données pertinentes pour l'analyse.

— *COEDM* : COEDM3 est le logiciel de gestion électronique des documents (GED), et il est utilisé dans toute l'entreprise. La documentation d'un projet comprend au moins :

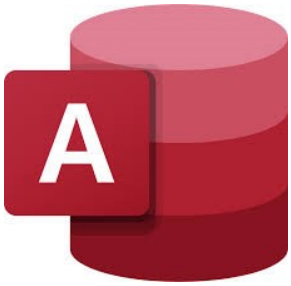
  — Les documents techniques reçus et envoyés aux clients, partenaires et fournisseurs

  — Les documents techniques internes créés dans le cadre d'un projet

  — Les communications officielles reçues et envoyées (courriel)

— *SharePoint* : Les données de SharePoint consistaient principalement en des fichiers Excel. J'ai utilisé des connecteurs Power BI pour extraire automatiquement ces données, en veillant au respect des autorisations d'accès et aux mises à jour en temps réel.

— *BDD Access* : Certaines données étaient stockées dans des bases de données Microsoft Access. L'importation de ces données a souvent nécessité une pré-transformation pour les rendre compatibles avec d'autres sources (Harmonisation des données).



— *SAP* : SAP est un système intégré de planification des ressources d'entreprise (ERP) largement utilisé par les organisations pour gérer divers processus d'entreprise, notamment les finances, la logistique et les ressources humaines.

— 

La phase de collecte des données a permis de rassembler des informations cruciales à partir de multiples sources, telles que SQL Server, COEDM3, SharePoint, Microsoft Access et SAP. Chaque source a nécessité une approche spécifique d'extraction et de transformation pour assurer la cohérence et la qualité des données recueillies.

4.3 Outils utilisés

Power Query

Power Query est un moteur de transformation et de préparation des données. Power Query est fourni avec une interface graphique permettant d'obtenir des données à partir de sources, et avec l'éditeur Power Query qui permet d'appliquer des transformations. Étant donné que le moteur est disponible dans de nombreux produits et services, la destination où les données seront stockées dépend de l'endroit où Power Query a été utilisé. Avec cet outil, on peut effectuer un traitement des données de type extraction, transformation et chargement (ETL).

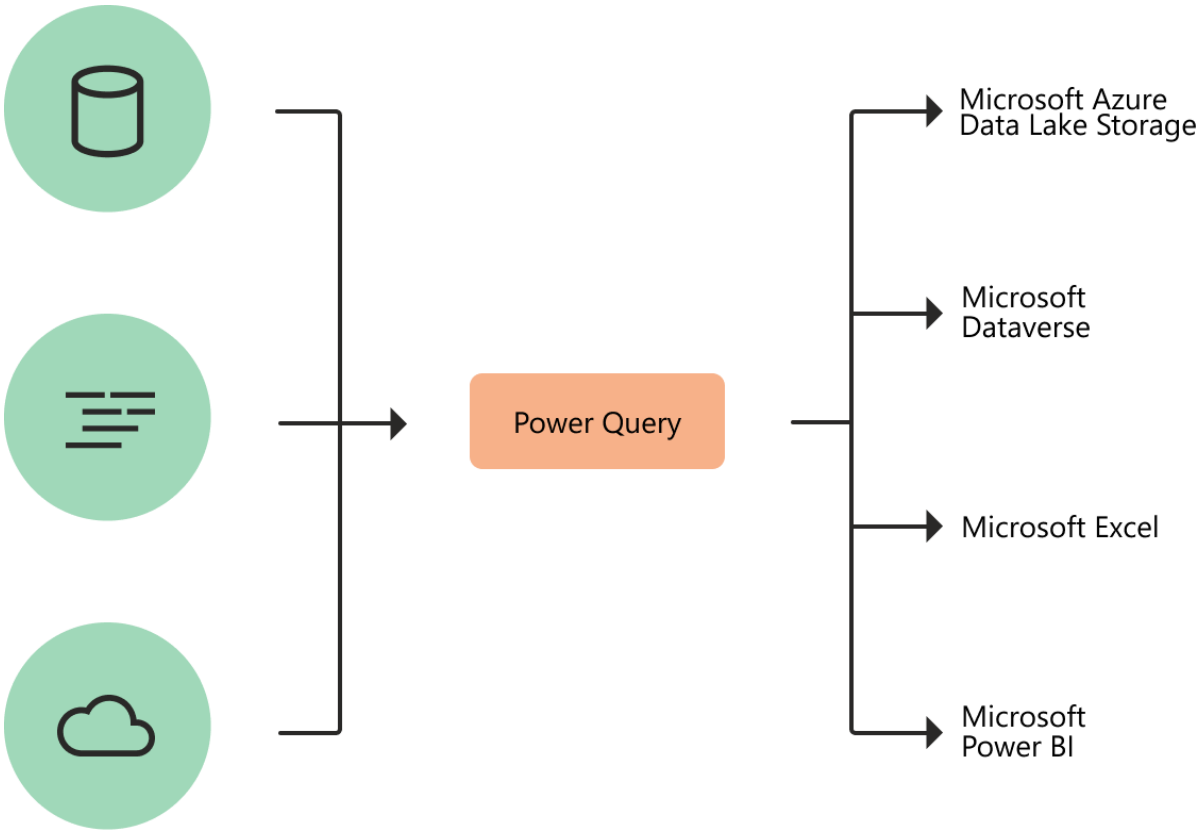


FIGURE 3 – Fonctionnement de Power Query

Power BI

La visualisation des données est l'un des aspects essentiels qui permet de représenter d'énormes volumes de données sous une forme simple et facile à comprendre. Selon Dzemyda, Kurasova et Žilinskas (2012), les entreprises et les organisations du 21e siècle produisent de grandes quantités de données et, par conséquent, la visualisation des données est nécessaire pour rendre ces données utiles. En d'autres termes, il est essentiel de simplifier les données collectées pour leur donner un sens. L'un des outils qui joue un rôle déterminant dans la visualisation des données est Microsoft Power BI. En référence à Aspin (2014), Power BI, qui techniquement parlant fait partie de Sharepoint en ligne, fournit une plate-forme permettant de télécharger des classeurs Excel dans le cloud et de partager des rapports avec les destinataires souhaités. De plus, ces classeurs Excel peuvent être mis à jour automatiquement, ce qui assure un flux fluide en ce qui concerne le traitement des données. À cet égard, Power BI peut permettre à différentes parties intéressées de discuter des rapports.

De plus, Power Bi dispose d'outils qui permettent aux personnes qui tra-

vaillent sur un projet d'utiliser Power Query(langage m) pour partager des requêtes ainsi que des routines de traitement de données complexes. Dans cette optique, les entreprises et les organisations peuvent utiliser efficacement des ressources telles que le temps en éliminant le besoin de duplication des efforts qui pourrait émerger lorsque les employés travaillent dans des « silos de données » (Aspin, 2014).

En plus d'économiser des ressources commerciales, la minimisation des cas de duplication des données permet d'éliminer les problèmes d'intégrité des données qui pourraient survenir en cours de route. Il est important de noter qu'en tant qu'outil de tableau de bord, Power BI peut être déployé sur un certain nombre de plateformes, notamment les systèmes d'exploitation Android et Windows. Cela signifie que les utilisateurs de cet outil peuvent l'utiliser comme une application mobile ou en déplacement. Notez que toutes ces fonctions de Power BI sont accessibles via un tableau de bord interactif, faisant ainsi de cette application de visualisation un meilleur outil de tableau de bord.

Composants de Power BI :

Power BI est constitué de plusieurs éléments qui fonctionnent ensemble, dont ces trois éléments de base :

- Une application de bureau Windows appelée Power BI Desktop.
- Un service SaaS (Software as a Service) en ligne appelé service Power BI.
- Des applications Power BI Mobile pour des appareils Windows, iOS et Android.

Langage DAX

DAX (Data Analysis expressions) est un langage d'expression de formule utilisé dans les applications Analysis Services, Power BI et Power Pivot dans excel. Les formules DAX incluent des fonctions, des opérateurs et des valeurs qui permettent d'effectuer des requêtes et des calculs complexes sur des données de colonnes et tables associées dans des modèles de données tabulaires.

J'ai largement utilisé le langage pour le développement des rapports, en particulier pour la création et le calcul de mesures complexes. Grâce à DAX,

j'ai pu élaborer des formules puissantes permettant de réaliser des calculs dynamiques et contextuels sur les données.

Modélisation des données :

Pour s'assurer que les tableaux de bord nous donnent des informations précises et utiles, il est essentiel de modéliser les tables contenant les données et d'établir les relations appropriées entre elles. Les relations entre les tables (comme les relations un à un (1,1), un à plusieurs (1,*), plusieurs à un (*,1) et plusieurs à plusieurs (*,*)) sont importantes pour le bon fonctionnement du tableau de bord (DASHBOARD).

Comment définir les relations entre les tables ?

La première chose à faire est d'analyser les tables, on commence par examiner la structure de chacune pour mieux comprendre le type de données que chaque table contient, comme les informations sur les clients, les documents ou les détails des produits. Ensuite, on va identifier les champs qui sont clés primaires de chaque table et qui seront utilisés pour créer les relations.

Définition des relations :

• Un à un (1,1) : Ce type de relation se produit lorsqu'un enregistrement d'une table est lié à exactement un enregistrement d'une autre table. Par exemple, chaque employé d'une table « Employés » peut avoir un enregistrement unique dans une table « Détails de l'employé ». Dans ce cas, le champ clé (comme EmployeeID) doit être unique dans les deux tables.

• Une à plusieurs (1,*) : Il s'agit d'une relation courante dans laquelle un enregistrement unique dans une table peut être lié à plusieurs enregistrements dans une autre table. Par exemple, un client unique dans une table « Customers » peut avoir plusieurs commandes dans une table « Orders ». Le « CustomerID » serait une clé primaire dans la table « Customers » et une clé étrangère dans la table « Orders ».

• Plusieurs à un (*,1) : Il s'agit essentiellement de l'inverse de la relation « un à plusieurs », dans laquelle plusieurs enregistrements d'une table renvoient à un seul enregistrement d'une autre table. Par exemple, plusieurs employés (dans une table « Employés ») peuvent être liés à un seul département (dans une table « Départements »). Le « DepartmentID » serait une clé étrangère dans la table « Employés ».

• Plusieurs à plusieurs (*,*) : Dans certains cas, plusieurs enregistrements d'une table peuvent être liés à plusieurs enregistrements d'une autre table. Ce type de relation nécessite généralement une table intermédiaire ou « jonction » pour gérer efficacement les relations. Par exemple, dans une table « Cours » et une table « Étudiants », un étudiant peut s'inscrire à plusieurs cours et un cours peut avoir plusieurs étudiants. Une table de jonction telle que « Enrollments » contiendrait à la fois « StudentID » et « CourseID » pour gérer cette relation.

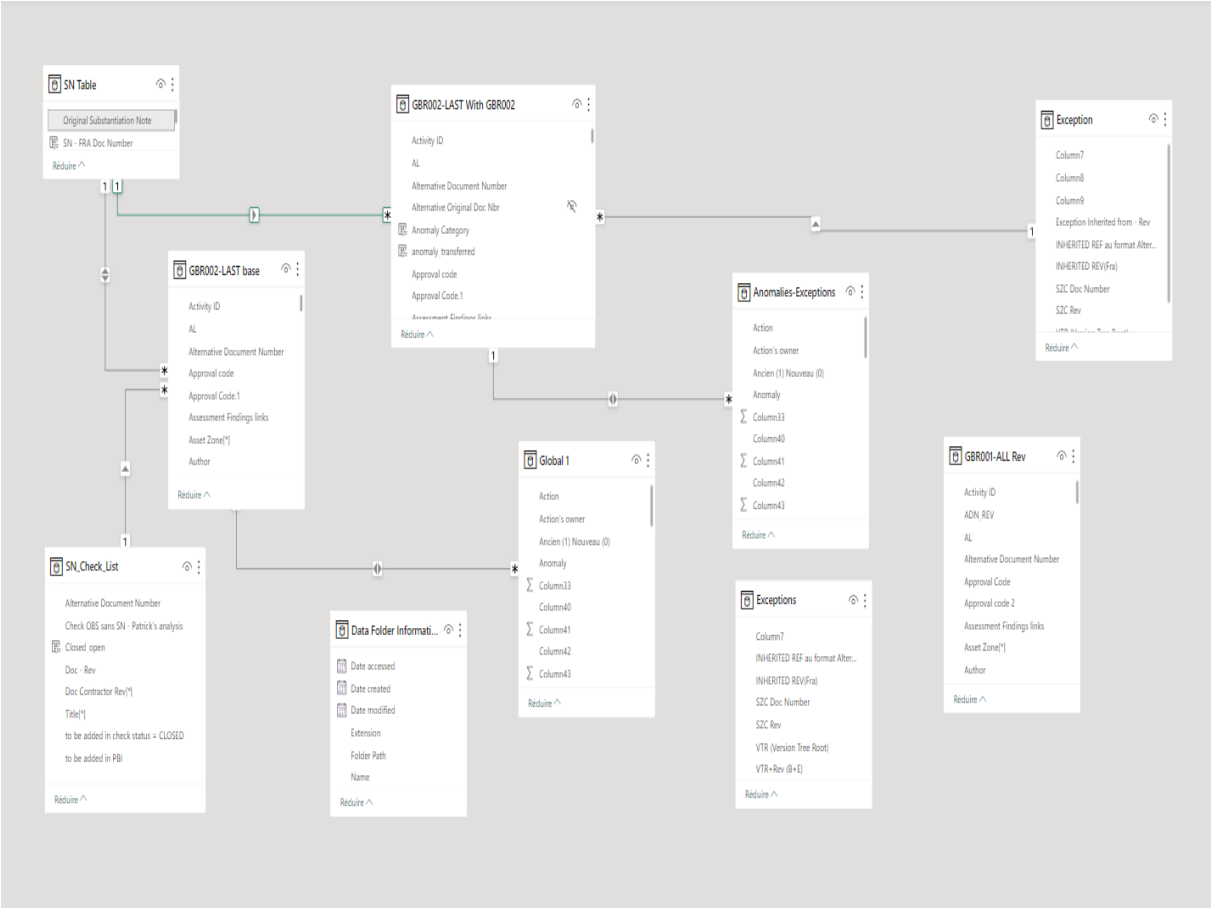


FIGURE 4 – Le modèle de données associé au KPI Réplication SZC

4.4 Création des KPI

Pour créer des KPI, il est essentiel de maîtriser toutes les étapes indiquées précédemment. Il faut commencer par la collecte des données, puis effectuer le prétraitement des données dans Power Query, et enfin créer les tableaux de bord dans Power BI.

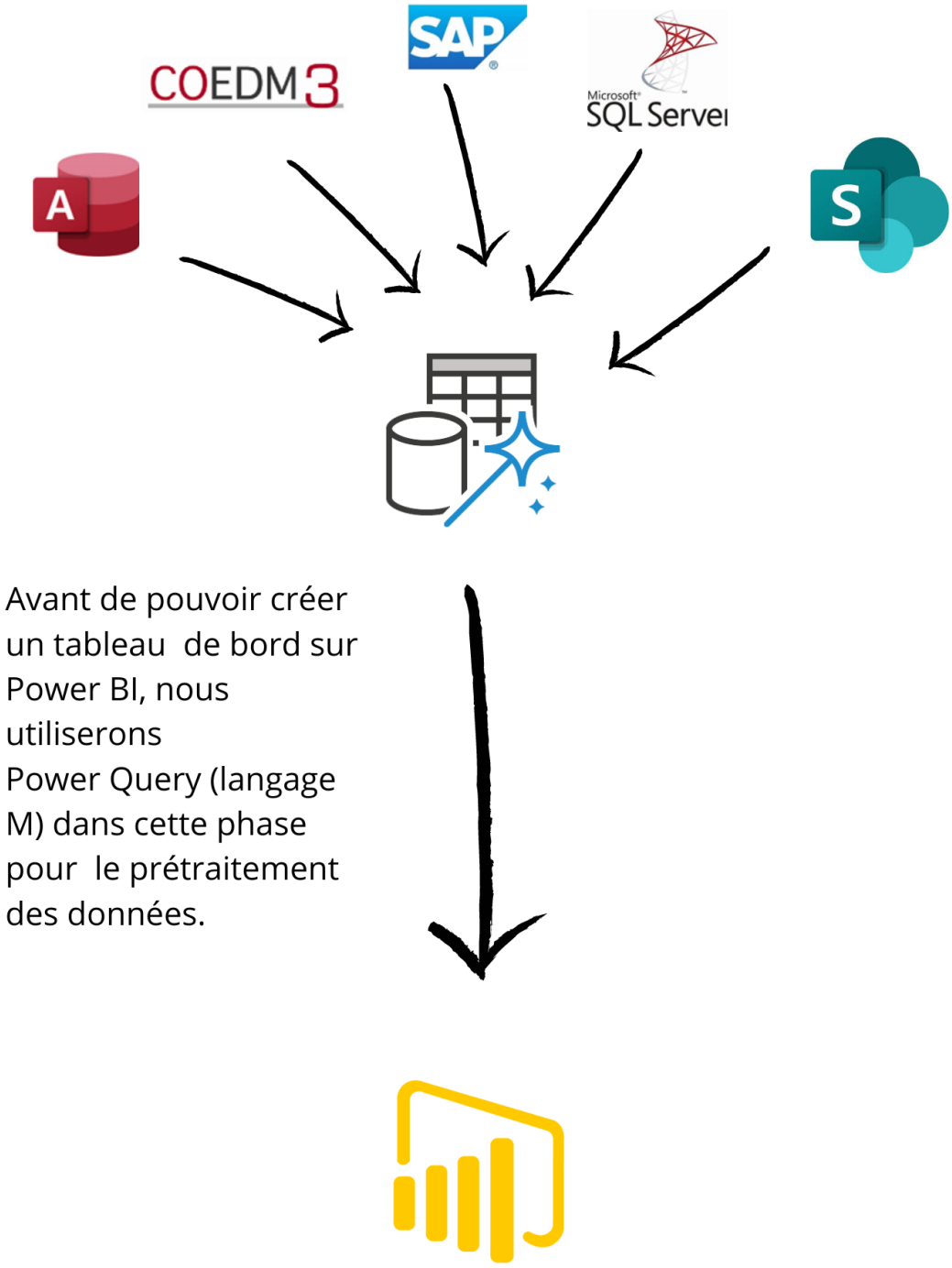


FIGURE 5 – Le processus de création de KPI

4.4 Participation à la création d'un Data Lake

4.4.1 Présentation des architectures Data

4.4.1.1 Data Lake

Un data lake est un emplacement de stockage centralisé qui contient des big data sous un format brut et granulaire provenant d'un grand nombre de sources. Il peut stocker des données structurées, semi-structurées ou non structurées, ce qui signifie que les données peuvent être conservées sous des formats plus souples pour une utilisation ultérieure. Lorsqu'il importe les données, le data lake les associe à des identificateurs et des balises de métadonnées pour une récupération plus rapide.

Imaginé par James Dixon, responsable des technologies (CTO) chez Pentaho, le terme « data lake » sous-entend que les données sont stockées en vrac et sous forme brute par contraste avec les données propres et traitées qui sont stockées dans les data warehouses traditionnels.

En général, les data lakes sont configurés sur un cluster de serveurs standard peu coûteux et évolutifs. Ce type de configuration permet de stocker des données dans le data lake (au cas où elles seraient nécessaires plus tard) sans avoir à se préoccuper de la capacité de stockage disponible. Les clusters peuvent être déployés sur site ou dans le cloud.

Les data lakes ne doivent pas être confondus avec les data warehouses. En effet, ils présentent des différences notoires qui peuvent constituer des avantages non négligeables pour certaines entreprises, en particulier à un moment où les big data et les processus des big data sont en train de migrer des solutions sur site vers le cloud.

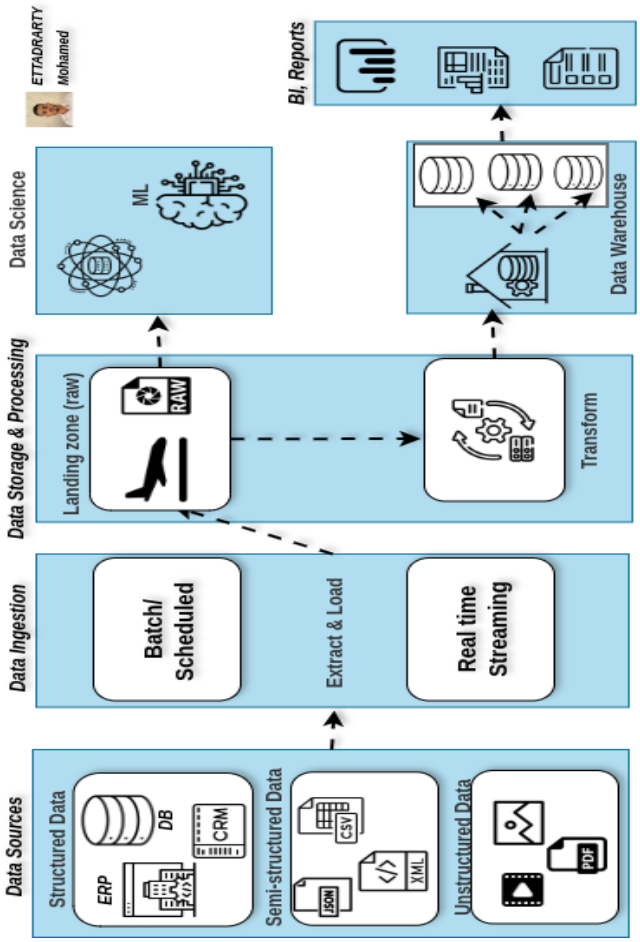


FIGURE 6 – Architecture de Data Lake

La figure 6 présente l'architecture typique d'un Data Lake traditionnel, organisée en plusieurs couches successives.

La Figure 6 présente l'architecture d'un Data Lake traditionnel, c'est-à-dire un système de stockage massif capable de centraliser toutes les données d'une organisation, quel que soit leur format ou leur origine. Cette architecture repose sur les étapes suivantes :

- **Sources de données** : Le Data Lake collecte des données depuis des systèmes variés. On distingue :
  - Les **données structurées**, issues des bases de données relationnelles, des systèmes ERP (Enterprise Resource Planning), ou des outils CRM (Customer Relationship Management). Ces données sont bien organisées en tables et colonnes.
  - Les **données semi-structurées**, comme les fichiers

JSON, XML ou CSV. Elles possèdent une certaine organisation mais ne respectent pas forcément un schéma strict.
    - Les **données non structurées**, telles que les documents PDF, images, vidéos, ou fichiers audio. Ce type de données représente souvent une grande partie du volume total mais nécessite des traitements spécifiques pour en extraire de la valeur.
- **Ingestion des données** : L'ingestion désigne le processus d'introduction des données dans le Data Lake. Elle se fait via deux mécanismes :
    - Le **batch** ou l'ingestion programmée (scheduled), qui consiste à importer les données à intervalles réguliers (quotidien, hebdomadaire, etc.).
    - Le **streaming en temps réel**, qui permet de traiter les données dès leur génération, ce qui est utile dans des contextes comme la supervision d'équipements ou l'analyse de logs en direct.
- **Stockage et traitement des données** :
    - Les données sont d'abord stockées dans une **landing zone** (zone brute), aussi appelée *RAW zone*, sans aucune transformation ni nettoyage. Cela permet de conserver une trace fidèle des données d'origine.
    - Ensuite, les données peuvent être transférées vers une zone de **transformation**, où elles subissent diverses opérations : nettoyage, enrichissement, jointures, conversions de formats, etc. Ces étapes préparent les données pour leur usage futur, en les rendant plus fiables et exploitables.
- **Exploration et valorisation des données** : Deux cas d'usage principaux sont identifiés :
    - La **Data Science**, où les données (brutes ou transformées) sont utilisées pour entraîner des modèles de Machine Learning (ML). Ces modèles permettent de produire des prédictions, classifications, ou détections automatiques à partir des données historiques.
    - Le **Data Warehousing**, qui consiste à déplacer les données nettoyées vers un entrepôt de données struc-

turé. Cet espace est optimisé pour des requêtes analytiques complexes, souvent exécutées via SQL.

- **Science des données (Data Science)** : Une fois les données disponibles dans le Data Lake, elles peuvent être exploitées par des data scientists pour mener des analyses avancées. Ces experts extraient des échantillons depuis la zone brute (RAW) ou transformée pour les explorer, nettoyer et enrichir selon leurs besoins. Ensuite, ils entraînent des modèles de **Machine Learning** (ML) ou d'**intelligence artificielle** sur ces données. Les cas d'usage sont multiples : détection de fraude, maintenance prédictive, segmentation de clients, etc. Cette étape nécessite souvent un environnement flexible (comme Jupyter, Databricks ou PySpark) et des bibliothèques spécialisées (Scikit-learn, TensorFlow, etc.).
- **Restitution des données (BI)** : Enfin, les données transformées ou agrégées sont exploitées dans des outils de Business Intelligence, comme Power BI, Tableau ou Qlik. Ces outils permettent de construire des rapports dynamiques et des tableaux de bord interactifs, à destination des métiers (marketing, finance, direction, etc.).

4.4.1.2 Data WareHouse

Un data warehouse est un type de système de gestion de données conçu pour permettre et faciliter les activités de business intelligence (BI), en particulier l'analytique. Les data warehouses sont uniquement destinés à effectuer des requêtes et des analyses. Ils contiennent souvent de grandes quantités de données historiques. Les données contenues dans un data warehouse proviennent généralement d'un large éventail de sources telles que les fichiers journaux d'application et les applications transactionnelles.

Un data warehouse centralise et consolide de grandes quantités de données provenant de plusieurs sources. Ses capacités analytiques permettent aux entreprises de tirer de leurs données de précieuses informations commerciales leur permettant d'améliorer leur processus de prise de décision. Au fil du temps, il crée un enregistrement historique qui peut s'avérer inestimable pour les data scientists et les analystes métiers. En raison de ces capa-

cités, un data warehouse peut être considéré comme la source unique d'informations fiables d'une entreprise.

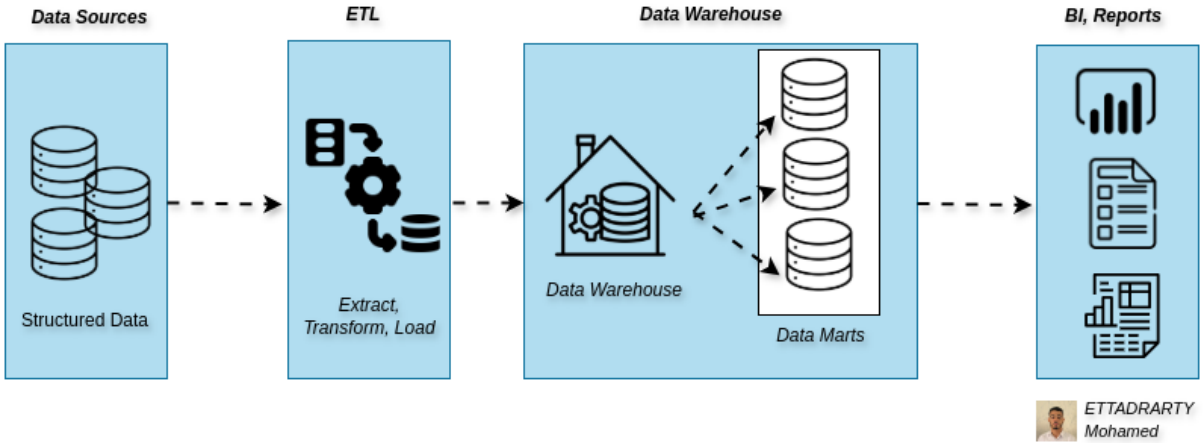


FIGURE 7 – Architecture de Data Warehouse

La Figure 7 illustre l'architecture classique d'un **Data Warehouse**, un système centralisé conçu pour le stockage et l'analyse de données principalement structurées. Contrairement aux Data Lakes, le Data Warehouse repose sur des schémas rigides et une forte gouvernance des données. Cette architecture est généralement structurée comme suit :

- **Sources de données** : Le Data Warehouse tire ses données de différentes sources transactionnelles, principalement structurées. On y trouve :
  - Des **bases de données relationnelles** (comme MySQL, SQL Server, Oracle) utilisées dans les applications métier.
  - Des systèmes **ERP** (Enterprise Resource Planning) centralisant la gestion des processus internes : finance, production, logistique, etc.
  - Des outils **CRM** (Customer Relationship Management) qui gèrent les interactions clients : ventes, marketing, service client.

  Ces systèmes produisent en continu des données fiables mais réparties et hétérogènes.
- **ETL (Extract, Transform, Load)** : Cette étape est essentielle pour assurer la qualité et la cohérence des données :
  - **Extract** : extraction des données depuis les différentes sources mentionnées.

- **Transform** : nettoyage, filtrage, normalisation des formats, enrichissement avec des règles métier (ex. : calcul d'indicateurs clés, conversion de devises, etc.).
- **Load** : chargement des données transformées dans le Data Warehouse, souvent par lots (batch).

L'ETL permet d'industrialiser la chaîne de traitement, tout en garantissant une intégrité forte des données.

- **Data Warehouse (Entrepôt de données)** : Il s'agit du cœur de l'architecture. Les données y sont stockées selon des modèles dimensionnels (modèle en étoile ou en flocon), facilitant les requêtes analytiques. Le Data Warehouse est :
  - **centralisé**, pour offrir une vision unique des données de l'entreprise,
  - **historisé**, afin de conserver les données dans le temps (notamment pour les analyses temporelles),
  - **optimisé** pour la lecture, via des index, agrégats et partitions.

  Il sert de socle pour tous les besoins analytiques de l'entreprise.
- **Data Marts (Magasins de données)** : Ce sont des sous-ensembles du Data Warehouse, extraits pour répondre à des besoins métiers spécifiques. Par exemple :
  - un data mart pour le marketing, avec les données de campagne et de comportement client,
  - un autre pour la finance, avec les indicateurs de performance financière.

  Cela permet aux équipes de travailler plus efficacement avec des données ciblées.
- **Restitution et reporting (BI)** : Les données sont exploitées à travers des outils de Business Intelligence comme **Power BI**, **Tableau** ou **Excel**, permettant :
  - la génération de rapports automatisés et partagés,
  - la création de tableaux de bord dynamiques et interactifs,
  - la visualisation des KPI (Key Performance Indicators) en temps réel.

Ces outils permettent aux décideurs de prendre des décisions fondées sur des données fiables, accessibles et mises à jour.

4.4.1.3 Data Lakehous

Un data lakehouse est une plateforme de données qui combine le stockage flexible des data lakes avec les capacités analytiques haute performance des data warehouses (entrepôts de données).

Les data lakes et les data warehouses sont généralement utilisés ensemble. Les data lakes servent de réceptacle global pour les nouvelles données, tandis que les data warehouses appliquent une structure en aval à ces données.

Cependant, coordonner ces systèmes pour fournir des données fiables peut être coûteux en termes de temps et de ressources. Les longs temps de traitement contribuent à l'obsolescence des données, et les couches supplémentaires d'ETL (extraction, transformation, chargement) introduisent des risques pour la qualité des données.

Les data lakehouses compensent les défauts des data warehouses et des data lakes grâce à des fonctionnalités qui forment un meilleur système de gestion des données. Ils associent la structure des données des entrepôts avec le coût réduit et la flexibilité des data lakes.

Les data lakehouses permettent aux équipes de données d'unifier leurs systèmes de données disparates, d'accélérer le traitement des données pour des analyses avancées (comme l'apprentissage automatique (ML)), d'accéder efficacement aux big data, et d'améliorer la qualité des données.

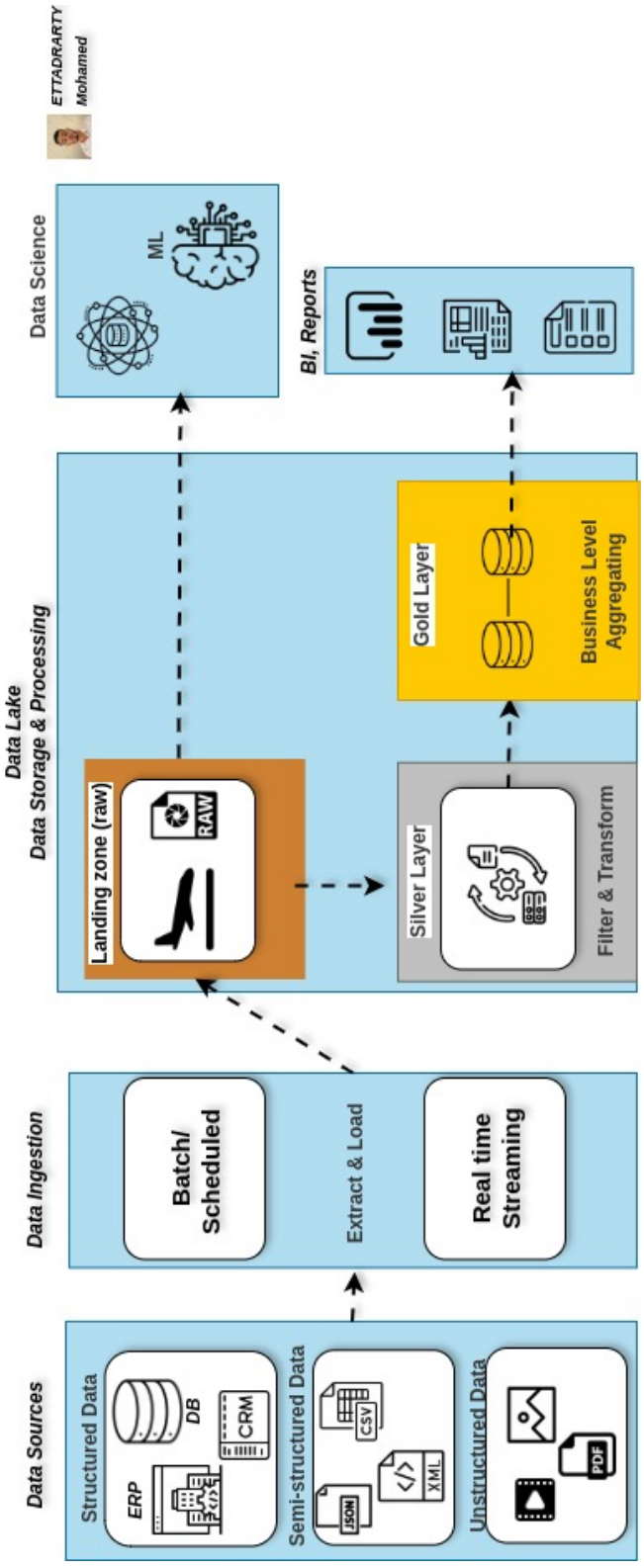


FIGURE 8 – Architecture de Data Lakehouse

La Figure 8 représente l'architecture d'un **Data Lakehouse**, un modèle hybride combinant les avantages des **Data Lakes** (flexibilité, scalabilité, stockage multi-format) avec ceux des **Data**

Warehouses (structure, gouvernance, performance analytique). Cette architecture moderne est pensée pour gérer un large éventail de types de données, tout en répondant aux exigences de la Business Intelligence et de la Data Science.

- **Stockage dans le Data Lake** :
  - **Landing Zone (Raw)** : première zone de dépôt, elle contient les données brutes telles qu'elles ont été reçues. Cette couche assure la traçabilité et la conservation des données originales.
  - **Silver Layer** : les données y sont filtrées, nettoyées et transformées. C'est ici qu'ont lieu les opérations de normalisation et d'enrichissement.
  - **Gold Layer** : couche finale contenant les données agrégées et prêtes à l'analyse. Elle sert de base fiable pour la création d'indicateurs métier.

  Ce découpage en couches favorise la séparation des préoccupations (brut / transformé / analytique) et garantit la qualité des données.

4.4.1.4 Objectif de Data Lake

Le projet de mise en place d'un Data Lake chez Framatome s'inscrit dans une démarche de transformation numérique visant à exploiter le plein potentiel des données d'entreprise. Ce Data Lake a pour vocation de répondre aux enjeux suivants :

- **Améliorer la productivité et l'efficacité opérationnelle** : La centralisation des données permet d'éliminer les silos d'information et d'automatiser de nombreux traitements aujourd'hui réalisés manuellement. L'objectif est de réduire le temps nécessaire pour accéder, préparer et exploiter les données, tout en minimisant les risques d'erreurs liées à la manipulation humaine.
- **Renforcer la prise de décision par la donnée** : Le Data Lake fournit un socle commun de données consolidées (données de référence ou « gold data »), issues de multiples sources hétérogènes. Les utilisateurs peuvent ainsi accéder à des tableaux de bord interactifs, réaliser des analyses multidimensionnelles, et prendre des décisions sur la base de données fiables, à jour, et contextualisées.

- **Favoriser l'innovation et le développement de nouveaux services** : Grâce à une infrastructure big data évolutive, le Data Lake permet de mettre en œuvre des cas d'usage avancés comme la maintenance prédictive, le contrôle qualité automatisé, ou l'optimisation des procédés industriels. Il constitue également un environnement propice à la création de POC (proof of concept) rapides et à la démocratisation des usages IA/ML.
- **Renforcer la gouvernance des données et la conformité réglementaire** : La plateforme intègre des mécanismes de gestion des accès, de traçabilité des données (data lineage), et de classification selon la sensibilité (Export Control, RGPD, etc.). Cela permet de garantir un usage sécurisé et conforme des données, avec des capacités de monitoring et d'audit.
- **Favoriser le partage et l'ouverture des données** : Le Data Lake facilite le partage transversal des données entre départements (ingénierie, production, qualité, etc.) et permet également l'intégration de données externes (clients, fournisseurs, partenaires). Il ouvre la possibilité de fournir des services à valeur ajoutée à des tiers via des APIs sécurisées.
- **Assurer la scalabilité et la résilience de l'infrastructure** : Conçu pour gérer des volumes croissants et des données de nature variée (structurées, non structurées, images, vidéos, documents), le Data Lake repose sur une architecture cloud ou hybride offrant haute disponibilité, montée en charge, et traitements parallèles à grande échelle.

Cette plateforme représente ainsi un levier stratégique pour Framatome, à la fois pour moderniser la gestion de la donnée, renforcer l'efficacité des processus internes, et explorer de nouveaux leviers de valeur ajoutée.

4.4.1.5 Architecture de Data Lake

L'architecture illustrée à la Figure 8 représente la solution que nous sommes en train de mettre en place au sein de Framatome. Il ne s'agit donc pas d'une architecture générique ou théorique, mais d'une implémentation concrète et adaptée aux besoins

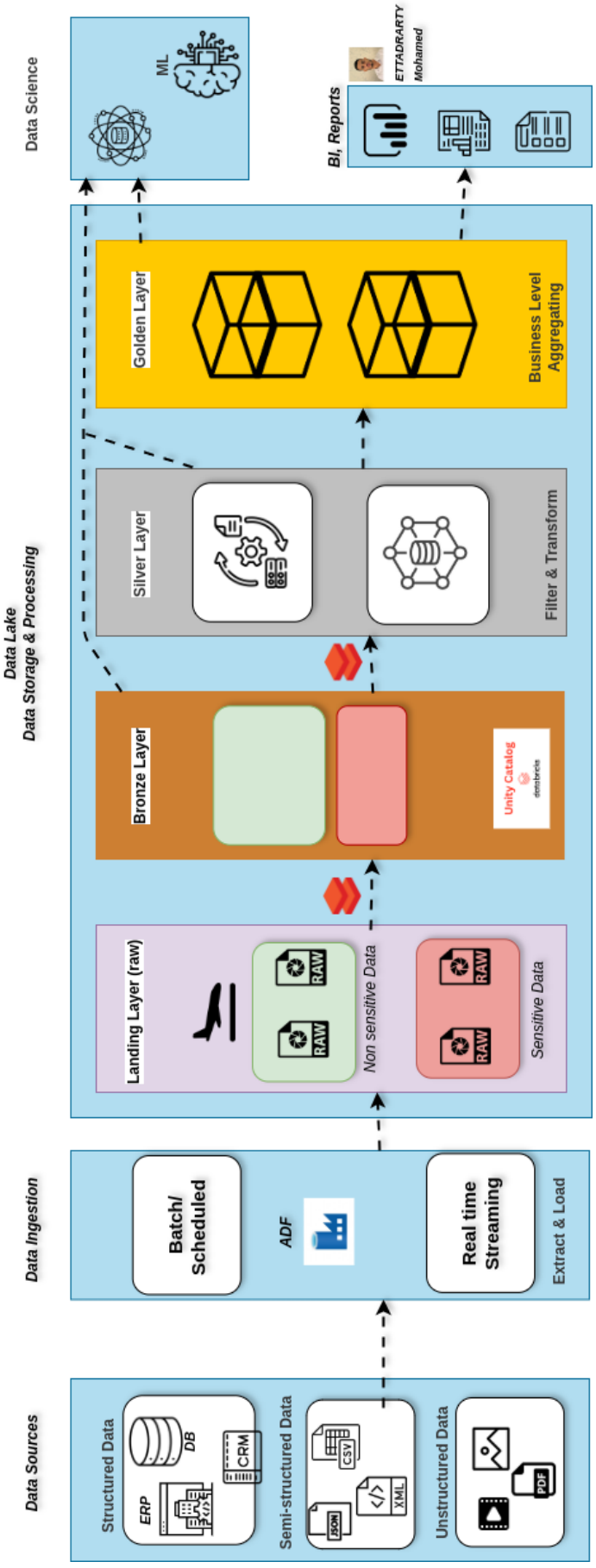


FIGURE 9 – SOGOUD Architecture

opérationnels d'un environnement orienté données (*data-driven*). Cette solution s'inscrit dans le paradigme moderne du **Data Lakehouse**, qui conjugue les avantages du Data Lake (souplesse de stockage) et de l'entrepôt de données (performance analytique).

Analyse de l'architecture mise en place

1. Ingestion de données multi-formats et multi-fréquences

L'architecture est conçue pour intégrer des données issues de sources hétérogènes :

- **Structurées** : ERP, bases de données relationnelles, systèmes CRM.
- **Semi-structurées** : fichiers JSON, XML, CSV.
- **Non structurées** : documents PDF, images, etc.

Deux modes d'ingestion sont mis en œuvre :

- **Batch planifié** pour les flux à fréquence régulière.
- **Streaming temps réel** pour les cas d'usage nécessitant de faibles latences.

Cette approche hybride garantit la flexibilité nécessaire pour traiter à la fois les traitements différés et les événements temps réel.

2. Cloisonnement des données dès l'ingestion

Dès la couche d'atterrissage (*Landing Layer*), un cloisonnement est opéré entre les **données sensibles** et les **données non sensibles**. Cette séparation permet :

- De renforcer la sécurité et la confidentialité,
- D'appliquer des politiques de conformité adaptées (notamment vis-à-vis du RGPD),
- De limiter les risques de mauvaise manipulation ou de fuite de données.

3. Traitement en trois couches : Bronze, Silver, Gold

L'organisation des données suit un modèle en trois couches, garantissant à la fois la traçabilité, la qualité et la performance :

- **Bronze Layer** : données brutes historisées, non transformées, servant d'archive.
- **Silver Layer** : données nettoyées et enrichies, prêtes pour les analyses intermédiaires.
- **Gold Layer** : données consolidées, agrégées et prêtes pour les analyses métier ou scientifiques.

Ce modèle favorise un traitement progressif, conforme aux principes du *data lineage* et du *data quality management*.

4. Gouvernance et sécurité avec Unity Catalog

La mise en place de **Unity Catalog** dans l'environnement Databricks permet de centraliser :

- La gestion des métadonnées,
- Les politiques d'accès aux données,
- Les audits et le suivi des consultations.

C'est un levier essentiel pour garantir une gouvernance de données efficace dans un contexte multi-équipes.

5. Valorisation des données : BI et Data Science

L'architecture alimente deux grands types de cas d'usage :

- **Business Intelligence** : tableaux de bord, indicateurs, visualisation des performances via des outils comme Power BI.
- **Data Science et Machine Learning** : préparation et exploration des données via des notebooks, modélisation prédictive, etc.

Cette double orientation garantit à la fois une réponse aux besoins immédiats des métiers, et un socle robuste pour les projets d'analyse avancée.

4.4.1.6 Mes contributions

test Au cours de cette mission, j'ai été mobilisé sur plusieurs volets du projet data, en lien avec les équipes techniques et

les référents métier. Mes contributions s'organisent autour de deux axes principaux : un travail en cours sur la spécification des sources de données, et des travaux à venir autour de la modélisation et de la restitution des données.

0.0.1 Spécification technique des sources de données (en cours)

Actuellement, je suis chargé de la **rédaction des spécifications techniques pour chaque source de données** à intégrer dans le système décisionnel.

Ce travail comprend plusieurs dimensions :

- **Analyse du besoin métier** : compréhension des objectifs de chaque source, du contexte fonctionnel et des usages prévus des données.
- **Identification des données pertinentes** : sélection des attributs à exploiter, description de leur signification, leur typage, leur format (CSV, JSON, Excel, etc.), et leur fréquence d'actualisation (quotidienne, hebdomadaire, événementielle, etc.).
- **Documentation technique** : formalisation des informations sous forme de fiches techniques, pour chaque source, à destination des équipes de développement.
- **Gestion des contraintes** : prise en compte des aspects de sécurité, de confidentialité (ex : données sensibles ou personnelles), ainsi que des spécificités d'accès (via API, fichiers partagés, bases de données, etc.).

Ce travail est essentiel pour garantir une ingestion et une exploitation cohérente et conforme des données dans les différentes couches du data lake.

0.0.2 Travaux prévus

Dans les prochaines phases du projet, je participerai activement aux étapes suivantes :

- **Modélisation des données dans la couche Silver** :

- Nettoyage et traitement des données brutes en vue d'une standardisation,
- Structuration des données selon les besoins métier,
- Création de tables intermédiaires prêtes à l'analyse.

— **Construction de la couche Gold** :

- Agrégation des données Silver pour produire des indicateurs de performance,
- Préparation des jeux de données finaux pour les utilisateurs métier.

— **Mise en place de rapports Power BI** :

- Conception de tableaux de bord dynamiques,
- Création de visualisations interactives répondant aux besoins exprimés par les utilisateurs métier,
- Paramétrage des sources de données et des filtres pour faciliter la prise de décision.

À noter : l'ingestion initiale des données en provenance des différentes sources (fichiers, bases, API, etc.) est gérée par l'équipe ***Data Platform (DPIT)****, via la* ***Landing Zone*** *et la* ***couche Bronze****, pour des raisons de sécurité et de conformité avec les standards internes (environnement* ***SOGOUD****).*

4.4.1.7 Résultats et bénéfices

Grâce à cette architecture, l'entreprise a pu :

- Gagner du temps dans la préparation et l'accès aux données.
- Améliorer la fiabilité et la traçabilité des indicateurs produits.
- Faciliter la collaboration entre les équipes techniques et métiers autour de données partagées.

4.4.1.8 Compétences acquises

Cette mission m'a permis de renforcer mes compétences techniques (Azure, Python, Spark, SQL), mais aussi de développer des aptitudes en gestion de projet, en communication avec les métiers, et en documentation technique.

5 Résultats

5.1 KPI réalisés

Au cours de ma 1ér année d'alternance, on a créé plusieurs indicateurs de performance clés (KPI) qui étaient essentiels pour l'équipe de documentation et PMO afin de surveiller les processus documentaires et de garantir le respect des délais et des procédures. Pour ce faire, on a identifié des indicateurs clés en collaboration avec les membres des équipes métiers pour comprendre leurs besoins. Ensuite, on a conçu et mis en œuvre ces indicateurs dans Power BI.

Les KPIs résultants ont fourni des informations en temps réel, permettant aux équipes de suivre les progrès de manière efficace et de s'assurer que la documentation des projets était constamment alignée aux normes établies par les équipes.

Voici quelques exemples des KPIs qu'on a développé :

KPI performances SZC :

Contexte : Framatome travaille principalement avec des clients pour la construction de centrales nucléaires. Pour chaque composant livré, les clients fournissent un retour d'information en utilisant des codes spécifiques : A (Approuvé), B (Approuvé avec commentaires) ou C (Refusé). Dans ce contexte, il était essentiel de mettre au point un KPI permettant de suivre et d'évaluer les performances de l'entreprise avec chaque client. Cet indicateur a été conçu pour analyser les taux d'approbation, identifier les domaines à améliorer et garantir que nos produits répondent toujours aux attentes des clients.

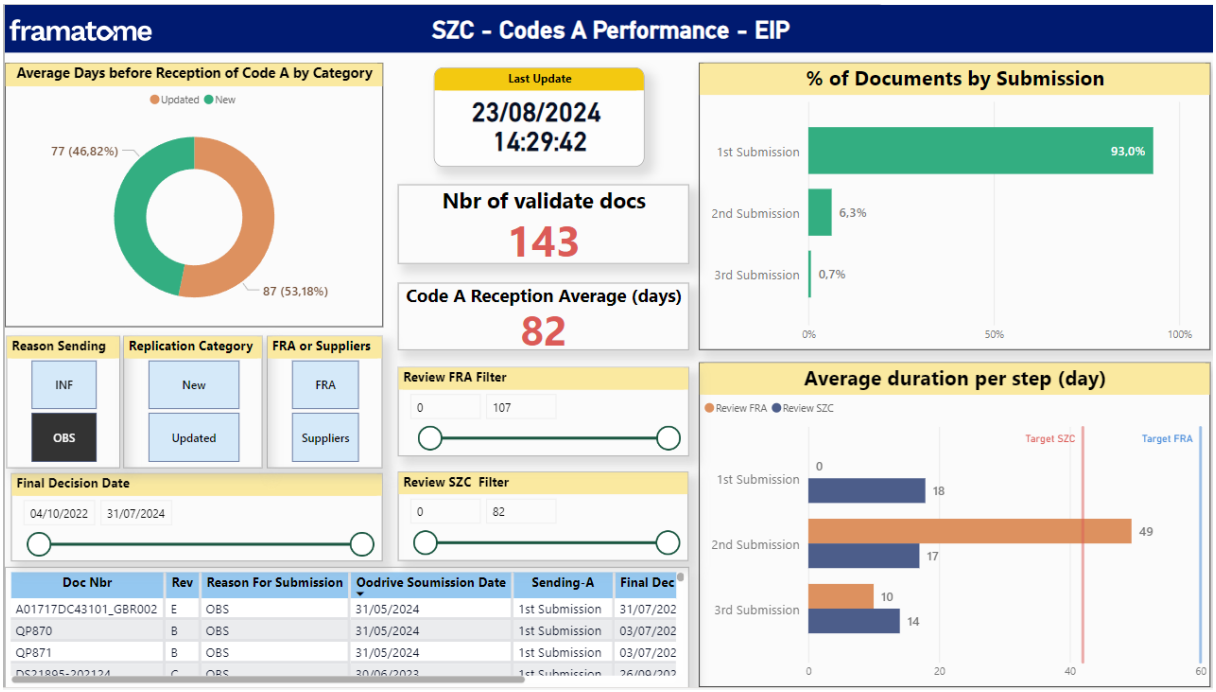


FIGURE 10 – capture d'écran d'un onglet du KPI SZC - Performances

KPI Réplication SZC :

Contexte : SZC est une réplication du projet HPC, ce qui signifie que tous les documents (chaque document représentant un composant) sont répliqués. Cela m'a amené, en collaboration avec les équipes du PMO et de la DTI, à développer un indicateur de contrôle qualité et de perferfance clé (KPI) pour suivre la réplication de chaque composant du projet HPC dans le projet SZC. Le défi était l'indisponibilité des données nécessaires pour tracer chaque document reproduit, ce qui nous a amené à créer un processus permettant une comparaison précise entre les deux projets et faire la liason entre les documents. Ce fut un processus compliqué à mettre en oeuvre parce qu'il n'y avait pas de liens direct entre les champs de deux bases des deux projets. On partait des métadonnées de la base de SZC pour créer le pointeur qui nous permet de réaliser la liaison entre les deux base et enfin le KPI en question.

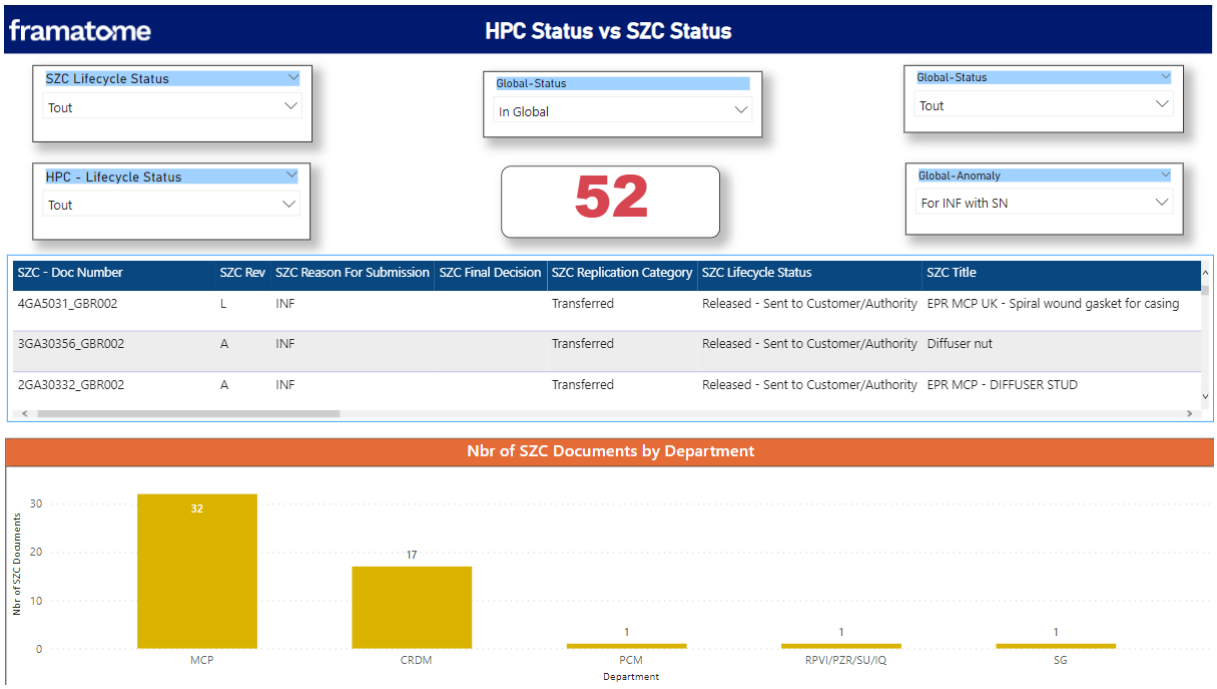


FIGURE 11 – capture d'écran d'un onglet du KPI Replication

6 Conclusion

6.1 Conclusion

Les indicateurs de performance (KPI) créés ont eu un impact significatif sur la gestion documentaire de Framatome. En apportant une meilleure visibilité sur les processus documentaires, ces KPI ont permis une gestion plus efficace, facilitant une prise de décision plus rapide à tous les niveaux de l'entreprise. Les indicateurs développés ont permis de suivre en temps réel des éléments clés tels que les délais de traitement des documents, les volumes de documents traités, et les écarts par rapport aux objectifs fixés. Cela a non seulement amélioré le contrôle des processus, mais aussi permis d'identifier rapidement les goulots d'étranglement et les zones nécessitant des actions correctives.

Grâce à ces KPI, Framatome a pu s'approcher de ses objectifs de performance et de d'atteindre la conformité documentaire, notamment dans le cadre du programme UK, où une gestion rigoureuse et transparente est essentielle. Les informations claires et précises fournies par les KPI ont contribué à renforcer la collaboration entre les équipes, à améliorer la communication et à assurer une meilleure traçabilité des documents. En outre, ces indicateurs ont aidé à aligner les activités documentaires sur les exigences réglementaires et contractuelles, ce qui est crucial pour garantir la qualité et la sécurité des

opérations.

Bibliographie

[1] Microsoft, *Power BI - Microsoft*, https://www.microsoft.com/fr-fr/power-platform/products/power-bi,

[2] Microsoft, *Qu'est-ce que Power Query ? - Microsoft Learn*, https://learn.microsoft.com/fr-fr/power-query/power-query-what-is-power-query,

[3] Framatome, *Framatome - Site officiel*, https://www.framatome.com/fr/,

[4] EDF, *EDF - Électricité de France*, https://www.edf.fr/,

[5] DAX Guide, *DAX Guide*, https://dax.guide/,

[6] PHData, *Data Modeling Fundamentals in Power BI - PHData*, https://www.phdata.io/blog/data-modeling-fundamentals-in-power-bi/,

[7] Framatome, *Global Search - Framatome*, https://globalsearch.framatome.corp/app/com_fram_app_globalsearch/#/home,

[8] Wiki Framatome, *Wiki - Framatome*, https://wikimonde.com/article/Framatome,

[9] ladrome, *ladrome*, https://www.ladrome.fr/wp-content/uploads/2020/03/framatome-nov-19-pres-cerca-v131119.pdf,

[10] learn.microsoft, *learn.microsoft*, https://learn.microsoft.com/fr-fr/power-query/power-query-what-is-power-query

[11] talend, *talend.learn*, https://www.talend.com/fr/resources/guide-data-lake/

[12] Oracle, *Oracle.Learn*, https://www.oracle.com/fr/database/what-is-a-data-warehouse/