Livre blanc

Le guide du Data Catalog

Socle de la gouvernance des données de votre entreprise

DataGalaxy

Sommaire

Le Big Data et le challenge de la gestion des gisements de données

Le **phénomène du Big Data** ne fait que s'intensifier et les entreprises collectent toujours plus de données, sous des formats de plus en plus variés. Le **volume de stockage des données** évolue donc constamment !

Où stocker toute la data accumulée ? Comment construire des gisements de données intelligibles ? Comment éviter les doublons entre les différentes bases de données ? Et comment structurer l'ensemble des informations pour répondre aux besoins de tous les métiers de l'entreprise ?

Le but du **catalogue de données** est de vous aider à répondre à toutes ces questions. Mais pas seulement : il vous permet également de **trouver un objectif business** concret lié au Big Data, et de **définir la gestion des données** de votre entreprise. Nous vous expliquons toutes les subtilités du Data Catalog, l'allié de votre stratégie de Data Governance.

Suivez le guide ! →

Qu'est-ce qu'un Data Catalog ?

Un Data Catalog peut être assimilé à un **inventaire en ligne** des données d'une entreprise. Intelligent et pratique, il facilite la gestion de la data tout en définissant et organisant les métadonnées. Le Data Catalog est l'outil parfait pour **définir une donnée**, ainsi que :

- sa structure,
- sa source,
- sa qualité,
- son utilisation.

Il détermine également la procédure à suivre pour garantir la bonne utilisation de la data. Grâce au Data Catalog, vous obtenez un **ensemble de données uniformisées**, **fiables**, et facilement **actionnables** pour en tirer une valeur business.

Selon Gartner – leader mondial de la recherche technique – les entreprises qui proposent un Data Catalog enrichi à une large variété d'utilisateurs doublent la valeur apportée par leurs données et leurs investissements dans l'analyse de ces dernières.

Le Data Catalog, un incontournable des entreprises Data Driven

Posséder un Data Catalog est indispensable pour les entreprises qui souhaitent tirer un avantage compétitif de leurs données. Mais pour construire une entreprise Data Driven, il faut également **instaurer une stratégie data globale**, et encadrer les équipes pour :

- apprendre à utiliser le catalogue ;
- comprendre la culture qui l'accompagne.

Comprendre les solutions de Data Catalog en 2021

Différents niveaux d'expertise

Il est essentiel de distinguer les Data Catalogs :

- **spécialisés dans la gestion des métadonnées d'un domaine particulier**, comme le marketing par exemple ;
- **inscrits dans une solution plus vaste**, comme les plateformes de fournisseurs d'infrastructures cloud qui proposent aussi quelques fonctionnalités de cartographie des données ;
- **indépendants et complets**.

Pour prendre un virage Data Driven, c'est vers la dernière catégorie qu'il faut vous tourner. Un Data Catalog indépendant et complet doit pouvoir **s'inscrire dans une stratégie data générale**. Pour cela, définissez les différents plans d'action à mettre en place pour répondre à vos besoins (mais sans vous précipiter). Attention, le Data Catalog est un **outil collaboratif**, l'idée est donc d'obtenir une gestion unifiée des données !

L'alliance des métiers et de l'Intelligence Artificielle

Selon Gartner, d'ici 2022, plus de 60 % des projets de Data Catalog traditionnels – c'est-à-dire qui n'utilisent pas le machine learning dans un écosystème Multicloud – ne seront pas livrés à temps. L'Intelligence Artificielle devient donc indispensable pour simplifier et fluidifier la mise en place d'un Data Catalog.

Le Data Catalog augmenté est un outil indépendant, il :

- automatise l'import et l'export des informations,
- enrichit les éléments textuels,
- suggère des liens,
- optimise la classification automatique,
- identifie les doublons...

En revanche, l'attention des différents métiers de l'entreprise est essentielle à son enrichissement. L'IA est un soutien, qui leur fait gagner du temps, pas une solution.

« Au-delà des capacités de collecte et de traitement des métadonnées techniques, ce sont les connaissances et savoirs des métiers que le Data Catalog doit réussir à capturer et restituer. Le Data Catalog, c'est la place du village de la culture data de l'entreprise. »

Lazhar Sellami
Cofondateur de DataGalaxy

Quatre fonctionnalités tournées vers les métiers

Le glossaire métier

C'est une base de connaissance qui permet à tous de définir, maintenir, et partager la **définition des données de l'entreprise**. La classification de ses informations s'appuie sur l'organisation interne de l'entreprise.

Le catalogue de traitements

C'est le mode d'emploi de la transformation des données de l'entreprise – indiquant les règles et traitements à suivre. Il permet d'**identifier** et de **retracer le parcours des données utilisées** (entrantes et sortantes) des chaînes de manipulation de la data.

Le dictionnaire de données

Ce n'est pas qu'un outil pour recenser des définitions : il sert aussi à **inventorier les informations** stockées dans les bases de données. Il partage également la structure, la source de la donnée, etc. À noter qu'il est possible de concevoir des modèles de données à intégrer dans le dictionnaire.

Le catalogue d'usages

Il répertorie les différentes utilisations des données par les équipes de l'entreprise. Il prend en compte, par exemple, les jeux de données mis à disposition des collaborateurs pour faire du reporting en toute indépendance. Ce catalogue permet de **contextualiser la provenance des données**, et d'**analyser l'impact d'une modification** dans les usages.

Les autres qualités essentielles du Data Catalog

Environnement Multicloud et connectivité

Le Data Catalog doit pouvoir **fonctionner dans un Multicloud**, voire avoir été conçu directement pour un tel environnement. La **connectivité vers d'autres outils de l'écosystème informatique** est essentielle. Elle permet au Data Catalog de récupérer ou d'envoyer des informations dans les systèmes existants de l'entreprise.

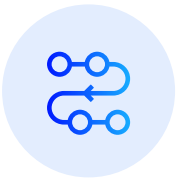
Dimension collaborative

Le Data Catalog encourage la collaboration entre les différents métiers de l'entreprise ! Il sert notamment à :

- **déterminer des accès uniques** pour les administrateurs, les responsables des modèles et bases de données et les utilisateurs ;
- **assigner des tâches** à réaliser ;
- **commenter les données** ;
- **partager des informations**...

Moteur de recherche

Le **moteur de recherche** est la fonctionnalité idéale pour gagner du temps : les informations sont trouvées grâce à l'indication de requêtes précises. Et pour affiner les recherches, il est aussi possible de mettre en place des **filtres spécifiques**.



Data Lineage

Le **Data Lineage** est la représentation visuelle de l'ensemble du cycle de vie d'une donnée. Il permet de comprendre le contexte de la data : sa source, ses différentes transformations, et ses usages finaux. Il est aussi possible de **consulter tous les éléments d'une donnée** en même temps à l'aide de la vue à 360° !

L'utilisation du Data Catalog en 4 étapes

1

Découvrir

- Faire une recherche sémantique
- Parcourir / filtrer
- Visualiser les relations sémantiques

2

Comprendre et enrichir

- Définir et contextualiser
- Étiqueter et annoter

3

Contribuer et gouverner

- Réaliser le Data Lineage
- Collaborer
- Certifier
- Documenter et publier

4

Consommer

- Remettre en question ou réutiliser
- Transférer les informations sur l'outil de son choix (outil BI par exemple)

Le Data Catalog : pour quels métiers ?

Le Chief Data Officer : le chef d'orchestre de la Data Culture

Le **Chief Data Officer** (CDO) s'occupe de la gestion et de la stratégie data de l'entreprise. Il assure notamment la mise en place d'une **gouvernance des données**.

Dans la réalisations de ses tâches quotidiennes, le Data Catalog lui permet de :

- **concevoir un langage commun** à tous les collaborateurs de l'entreprise pour échanger sur les données ;
- définir les **conditions d'utilisation** des données ;
- créer un **glossaire** ;
- **comprendre** et **appréhender les besoins** de tous les métiers.

Business Intelligence Manager : le coordinateur des projets d'analyse data

Le **Business Intelligence Manager** a pour mission de collecter et d'analyser les données. Il est responsable de la mise en place de l'écosystème analytique de l'entreprise.

Le Data Catalog est pour lui un outil essentiel puisqu'il lui offre une **donnée fiable**, qu'il peut alors analyser et restituer aux décideurs de l'entreprise. Il lui sert aussi à :

- retracer la **provenance de la donnée** ;
- consulter les **résultats des analyses** faites par les métiers pour en vérifier la source ;
- suivre les différentes **transformations** de la donnée.

Data Scientist : le Data Analyst du Big Data

Les **Data Scientists** sont responsables de la **gestion de la donnée** de l'entreprise. Ils doivent créer (et tester) des algorithmes pour disséquer et analyser la data afin d'en comprendre toutes ses composantes.

Le Data Scientist a besoin d'un Data Catalog enrichi pour :

- **analyser des données fiables**, à la source vérifiée ;
- **échanger avec les Data Bakers** ;
- **concevoir des indicateurs clés**, compréhensibles par l'ensemble de l'entreprise.

Les référents métiers de l'entreprise

Les référents métiers – ou **Data Prosumers** – doivent pouvoir **analyser les données** de manière autonome et créer un reporting personnalisé pour prendre des décisions. En ce sens, ils ont besoin d'un accès (restreint ou complet) au Data Catalog de l'entreprise.

De nombreux métiers sont concernés par l'utilisation de la donnée : le Data Catalog indique ses sources, ses différentes utilisations au sein de l'entreprise, les référents à consulter, etc. Et comme ils ont accès au glossaire, ils peuvent utiliser le **langage expert des Data Bakers** et échanger avec eux.

Dépasser le cloisonnement IT et métiers

Le Data Catalog permet d'**améliorer la collaboration** entre l'Intelligence Artificielle et les différents référents métiers, mettant un terme à la séparation de ces deux départements. Une entreprise Data Centric permet à l'ensemble des collaborateurs de **gérer les données** qui les concernent.

Le Data Catalog est l'outil idéal pour maîtriser ses données d'entreprise, mais aussi pour en tirer une forte valeur business.

Dans les marchés ultra-concurrentiels, il est indispensable !

Chez DataGalaxy, nous sommes convaincus que la réussite d'un projet Data Catalog repose sur une approche métier pragmatique. Nous avons donc conçu notre DataCatalog 360° en conséquence. Le plus ? Nous pouvons accompagner les dirigeants et les équipes pour les aider à instaurer une stratégie et une culture data globale dans l'entreprise.

Je découvre le DataCatalog 360°

DataGalaxy