Chapitre 3
Les étapes de création d'une IA

1. Ce que nous allons découvrir et les prérequis

Vous avez désormais bien cerné la notion fondamentale de l'intelligence artificielle consistant à apprendre à partir de la Data. En d'autres termes, cela signifie que si vous n'avez pas de données, vous ne pourrez ni entraîner votre IA ni l'améliorer au fil du temps.

Avant de plonger dans les méandres de la création technique d'une intelligence artificielle, il est essentiel de bien comprendre l'ensemble des étapes nécessaires à la création d'une intelligence artificielle, allant de l'objectif initial à la collecte et la préparation des données, jusqu'à la mise en production de l'IA.

Remarque

La mise en production désigne le processus par lequel un produit, un logiciel, un système informatique ou tout autre élément est déployé dans un environnement opérationnel et utilisé par les utilisateurs finaux.

Ce chapitre vous présentera les fondements de la collecte de données d'un point de vue théorique : les principes de base, les objectifs, le processus de collecte et de préparation des données, le choix et la mise en place des algorithmes, la formation et l'évaluation des différents modèles, puis le déploiement de vos travaux dans des environnements dédiés afin d'en permettre la bonne utilisation.

Remarque

Aucun prérequis n'est nécessaire pour aborder ce chapitre.

2. Comprendre l'objectif

Tout d'abord, vous devez identifier avec précision le domaine ou le problème que vous souhaitez résoudre avec votre future intelligence artificielle. Dans les faits, il n'y a aucune limite puisque cela peut aller de la compréhension des tendances d'achat à la reconnaissance d'image, ou bien encore à l'analyse de marché immobilier, etc.

Cela peut sembler logique au premier abord, mais en comprenant de façon précise votre objectif, vous aurez la capacité de déterminer les types de données nécessaires pour entraîner votre modèle.

Quel problème voulez-vous résoudre ou quelle tâche spécifique l'IA doit-elle accomplir ? Cette compréhension vous guidera dans le choix des algorithmes appropriés.

Par exemple, si vous travaillez sur une tâche de classification d'images, vous aurez besoin d'un algorithme différent de celui requis pour une tâche de prédiction de valeurs numériques.

3. Collecte et la préparation de données

La collecte de données est la toute première étape dans la création d'une intelligence artificielle.

Avant même de penser à l'architecture technique que vous allez déployer ou aux algorithmes les plus contemporains que vous souhaitez utiliser, il est tout bonnement impératif de déterminer l'ensemble des données qui vous seront nécessaires à l'atteinte de vos objectifs. Nous insistons une dernière fois : les données constituent la source d'alimentation de votre intelligence artificielle, leur qualité est essentielle au bon déploiement de votre intelligence artificielle.

© Editions ENI - All rights reserved

Vous noterez également que ces données sont susceptibles de prendre différentes formes toutes aussi variées les unes que les autres telles que des photos, du texte, des documents, des vidéos, des enregistrements audios, des données structurées provenant des systèmes de gestion de bases de données comme Oracle ou SQL Server, ou de bases portant des données non structurées.

Les bases de données dites « relationnelles » sont utilisées depuis plusieurs décennies (1970). Elles disposent d'un langage de requêtage utilisé la majeure partie du temps dénommé SQL (*Structured Query Language*), permettant à lui seul d'interagir avec les données que ce soit à travers un outil dédié (SGBDR - Système de gestion de base de données relationnelle) ou dans une application réalisée par des équipes d'ingénierie logicielle afin de créer, mettre à jour, consommer ou supprimer ces données (CRUD - *Create, Read, Update, Delete*).

Les données y sont stockées dans des tableaux à deux dimensions, aussi appelés « tables ».

Chaque ligne représente un enregistrement (un client, une commande…) et chaque colonne un attribut de cet enregistrement (son nom, sa date, son montant…). Les tables peuvent être reliées entre elles grâce à des clés : la table des commandes contient par exemple le numéro du client concerné, ce qui permet de retrouver en un instant toutes les commandes d'un même client. Ce sont ces « relations » qui ont donné son nom au modèle relationnel.



Base de données relationnelle : tableau 2 dimensions

Il existe différentes méthodes permettant de concevoir un modèle relationnel. La plus connue en France s'appelle Merise. Née à la fin des années 1970, elle permet aux concepteurs de réaliser un modèle conceptuel de données (MCD) sous forme graphique, sans se soucier à ce stade de la base de données sur laquelle reposera l'application. Son nom viendrait du merisier, un arbre qui ne donne de beaux fruits qu'une fois greffé : une façon de dire qu'une méthode ne porte ses fruits que si elle est adoptée par l'organisation. De nombreux ouvrages y sont consacrés, dont l'excellent guide pratique de Jean-Luc Baptiste aux Éditions ENI. Les bases de données relationnelles offrent encore aujourd'hui bon nombre d'avantages tels que la facilité de déploiement, les très faibles redondances des données, ou bien encore le langage de requêtage SQL uniforme.

Inversement, elles ne sont pas toujours très adaptées à de fortes volumétries de données dès lors que vous commencerez à atteindre des centaines voire des milliers de Go…

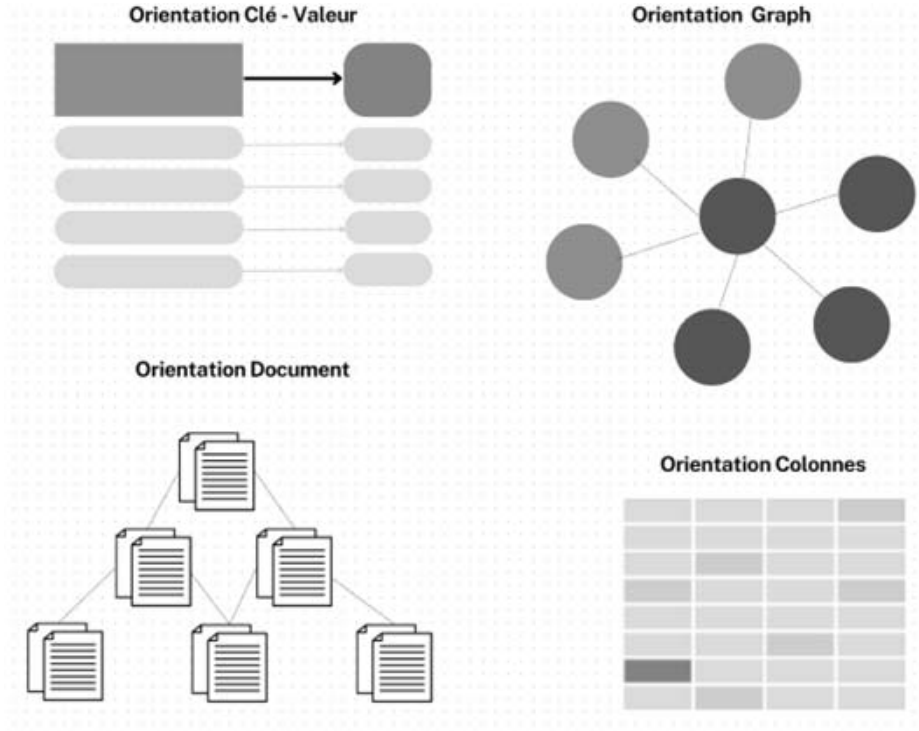
Un concurrent de poids est toutefois venu bousculer la donne dans les années 2000 pour le stockage massif des données : le NoSQL, pour « *Not Only SQL* ». En 2008, Facebook a même rendu open source Cassandra, la base NoSQL que ses ingénieurs avaient conçue pour la recherche dans les messageries de ses utilisateurs. Elle est aujourd'hui maintenue par la fondation Apache et utilisée par Netflix, Apple ou Uber. Pour comprendre cet engouement, sachez que les bases NoSQL ne sont quant à elles pas relationnelles, ce qui sous-entend que les données ne sont plus matérialisées dans des tables comme dans le modèle précédent, offrant une flexibilité accrue dans l'évolution des modèles. Les bases NoSQL conviennent donc parfaitement dans un cadre d'utilisation de données non structurées et lorsque les relations entre ces dernières ne sont pas spécialement importantes, vous en déduirez l'engouement dans le domaine de l'intelligence artificielle.

Cette typologie de bases de données a ainsi pour particularités une rapidité d'accès bien meilleure qu'une base relationnelle classique ainsi qu'une forte flexibilité des modèles.

© Editions ENI - All rights reserved

Vous retrouverez ainsi plusieurs typologies de bases de données NoSQL :

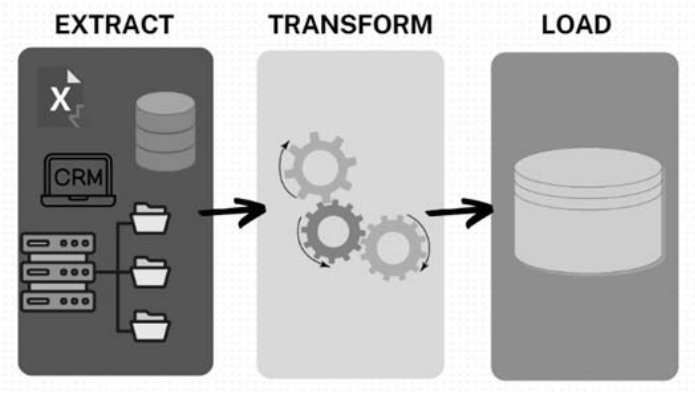
– **Les bases orientées colonnes** : les données y sont stockées sous une forme de paires clé/valeur. Exemple : Cassandra, HBase, Parquet.
– **Les bases orientées documents** : les données y sont stockées dans des documents JSON dans la base de données. Exemple : DynamoDB, MongoDB, Elasticsearch.
– **Les bases orientées clé/valeur** : les données y sont représentées sous une forme de collection clé/valeur. Exemple : Redis, Memcached.
– **Les bases orientées graphes** : les données sont stockées dans une structure de graphe sous forme de propriétés de nœuds, d'arêtes et de données. Exemple : Neo4j, OrientDB, InfluxDB.



Typologies bases NoSQL

La collecte de données

Dès lors que vous aurez identifié les sources de données en adéquation avec votre besoin de création d'intelligence artificielle, vous devez alors les collecter. Cette collecte de données n'est pas à prendre à la légère : elle est susceptible d'être techniquement complexe, nécessitant de fait des compétences techniques d'ingénierie spécifiques. Vous pouvez utiliser des techniques d'extraction automatique à partir d'ETL (processus d'extraction visant à rassembler les données provenant de plusieurs sources) tels qu'**ODI**, **SSIS**, **Talend** ou bien encore **Informatica**.

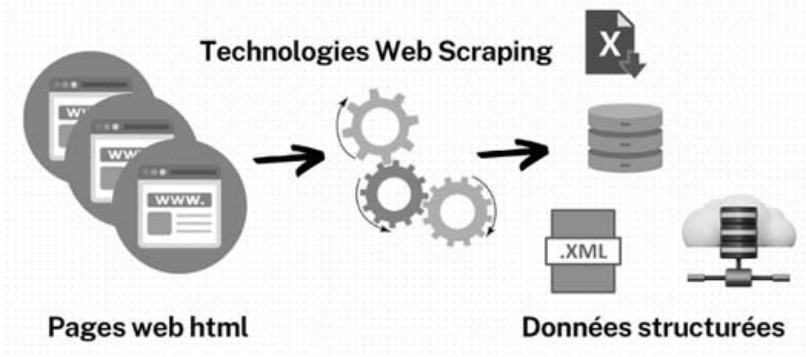


Principe d'extraction de données via ETL

Les techniques de web scraping sont également très utilisées pour vous permettre la récupération de données. Ces techniques consistent à extraire ou plutôt « scraper » du contenu présent sur le Web.

Le scraping vous permettra de récupérer une multitude d'informations qui seront précieuses telles que les adresses e-mail, un numéro de téléphone, des adresses etc., le tout dans l'objectif de les compiler dans une base de données pour en permettre l'exploitation.

© Editions ENI - All rights reserved



Technologie web scraping

Il existe deux manières de faire du web scraping :

– manuellement ;

– automatiquement.

Le scraping manuel, comme son nom l'indique, consiste à copier-coller les informations que vous collectez dans l'objectif de constituer une base de données. C'est un travail rébarbatif et chronophage qui est usuellement utilisé pour de petites quantités de données.

Le scraping automatique permet quant à lui d'explorer et d'extraire des informations depuis le Web. De nombreux outils existent, parmi eux :

– **CaptainData** : permet de récupérer les données depuis un catalogue des principaux sites que vous souhaitez scraper (LinkedIn, X, Leboncoin, TripAdvisor …). Sur chacun des sites mentionnés, l'outil de scraping offre différents scénarios, permettant même d'interagir avec certains d'entre eux, telles que LinkedIn. Ceci a tout de même un coût, comptez 100 €/mois pour vous permettre jusqu'à 20 000 actions.

– **Import.IO** : outil de scraping permettant de scraper et d'exporter au format CSV.

– **Scrapy** : framework Python open source permettant d'extraire les données des sites web. Il dispose de nombreux modules complémentaires, en plus d'une très grande communauté.

– **Webscraper** : il s'agit d'une extension du navigateur Google Chrome vous permettant d'extraire les données des sites internet souhaités. La collecte permettra l'extraction sous forme de fichier CSV.
– **Octoparse** : l'outil offre une interface vous permettant de scraper sans avoir à saisir la moindre ligne de code de façon illimitée et le tout gratuitement.

Il est important de noter que, quelle que soit la solution qui vous permettra de collecter des données, cette étape doit être effectuée de manière éthique et légale.

Ce rappel tenait en une ligne dans la première édition de ce livre ; il mériterait aujourd'hui un chapitre entier, tant le sujet a explosé. La collecte massive de données pour entraîner des IA est devenue l'un des terrains juridiques les plus disputés du secteur. Le New York Times poursuit OpenAI et Microsoft depuis décembre 2023 pour avoir utilisé ses articles sans autorisation dans l'entraînement de ChatGPT. Au moment où nous écrivons ces lignes, aucun jugement n'a été rendu sur le fond, mais le tribunal a refusé en 2025 de rejeter la plainte : le procès est considéré comme un test décisif pour toute l'industrie.

Un autre dossier a déjà donné une idée des montants en jeu. En septembre 2025, Anthropic, l'éditeur de Claude, a accepté de verser 1,5 milliard de dollars à des auteurs dont les livres, téléchargés depuis des bibliothèques pirates, avaient servi à entraîner ses modèles, soit environ 3 000 dollars par ouvrage. Le juge avait pourtant estimé que l'entraînement en lui-même relevait de l'usage loyal (fair use) ; c'est la provenance pirate des livres qui a coûté cher. À l'inverse, Getty Images a largement perdu face à Stability AI devant la Haute Cour de Londres en novembre 2025.

© Editions ENI - All rights reserved

En parallèle, une partie du secteur a choisi la négociation plutôt que le contentieux : l'Associated Press, Le Monde, le Financial Times ou le groupe espagnol Prisa (propriétaire d'El País) ont signé des accords de licence rémunérant l'usage de leurs contenus. En Europe, le règlement sur l'IA (AI Act) oblige par ailleurs, depuis août 2025, les fournisseurs de grands modèles à publier un résumé des données utilisées pour les entraîner. Concrètement, si vous êtes en formation ou responsable d'un projet Data & IA, la question « ai-je le droit de collecter et d'utiliser ces données ? » n'est plus une case qu'on coche en bas de page : c'est un sujet de gouvernance à part entière. Vous devrez vous assurer que les données collectées sont représentatives de la réalité à modéliser avec votre intelligence artificielle nouvellement créée. L'importance de la diversité des données permettra à votre modèle de généraliser et de prendre des décisions précises dans des situations réelles.

De plus, il est essentiel de vérifier la qualité des données collectées. Cela peut impliquer la suppression des données incorrectes, dupliquées ou même incohérentes. L'impact de cette non-qualité serait de biaiser le modèle susceptible de conduire à des résultats non fiables et fausser les prédictions de votre IA.

Enfin, vous l'avez constaté à travers les différentes méthodologies d'extraction, les données provenant de différentes sources peuvent être dans des formats variés tels que des documents, des images, du texte... Il est de fait absolument nécessaire de les normaliser et de les formater de manière cohérente afin de faciliter leur traitement ultérieur. Cela peut inclure la conversion de formats, l'élimination des caractères indésirables ou la normalisation des échelles de valeurs.

En résumé, la collecte et la préparation des données sont les fondements mêmes sur lesquels repose la création d'une intelligence artificielle réussie. En comprenant les données dont vous avez besoin, en les collectant de manière éthique et en les préparant de manière adéquate, vous vous assurez d'avoir une base solide pour la suite de votre aventure dans le monde de l'intelligence artificielle.