Leçon 3 — Data Analyst, Data Scientist et Data Engineer
==========

Leçon 3 — Data Analyst, Data Scientist et Data Engineer

1. Introduction : trois métiers au cœur de la donnée

Les métiers de la donnée constituent l'une des principales portes d'entrée vers l'intelligence artificielle. Data Analyst, Data Scientist et Data Engineer travaillent ensemble, mais leurs responsabilités ne sont pas identiques. Le data analyst cherche principalement à transformer les données disponibles en informations utiles aux décisions. Le data scientist construit et évalue des modèles statistiques ou d'apprentissage automatique afin de répondre à des problèmes de prédiction, de classification, d'optimisation ou d'exploration. Le data engineer construit les fondations techniques qui permettent de collecter, transporter, transformer, stocker et rendre disponibles les données. Dans une grande organisation, ces fonctions peuvent être très spécialisées. Dans une petite entreprise, elles peuvent être partiellement réunies. Comprendre cette complémentarité est donc indispensable pour choisir une trajectoire professionnelle réaliste.

La présente leçon adopte une approche professionnelle. Elle ne cherche pas seulement à définir trois métiers, mais à expliquer leurs activités quotidiennes, leurs méthodes, leurs outils, leurs livrables, leurs indicateurs, leurs interactions et leurs limites. L'apprenant découvrira également comment analyser un projet réel, choisir un profil adapté, construire un portfolio et organiser une progression de compétences. Les travaux de Foster Provost et Tom Fawcett sur la science des données rappellent notamment que la valeur d'un projet data dépend de la formulation correcte du problème et non de la sophistication isolée d'un algorithme. De même, les pratiques modernes d'ingénierie des données montrent qu'un modèle ne peut être fiable si les données qui l'alimentent sont mal gouvernées.

2. Objectifs pédagogiques

À l'issue de cette leçon, l'apprenant devra pouvoir définir précisément les rôles de data analyst, data scientist et data engineer ; expliquer leurs différences et complémentarités ; décrire le cycle de vie d'une donnée ; identifier les étapes d'un projet analytique ou prédictif ; choisir des indicateurs adaptés ; comprendre les principales familles d'outils ; analyser la qualité des données ; distinguer corrélation et causalité ; expliquer les notions de modèle, entraînement, test et validation ; comprendre les principes des pipelines de données ; analyser une offre d'emploi ; proposer une équipe adaptée à une organisation ; construire un portfolio et élaborer un plan de progression professionnel.

3. La donnée comme matière première de l'IA

Une donnée est une représentation d'un fait, d'une observation, d'un événement ou d'une mesure. Elle peut être numérique, textuelle, temporelle, géographique, audiovisuelle ou relationnelle. Dans un projet d'IA, les données jouent plusieurs rôles : elles permettent de comprendre une situation, d'entraîner un modèle, de vérifier une hypothèse, de mesurer un résultat et de surveiller un système en production. La qualité

d’une donnée dépend de son exactitude, de sa complétude, de sa cohérence, de son actualité, de sa pertinence et de sa traçabilité.

Cette notion de matière première doit cependant être nuancée. Une grande quantité de données ne garantit pas un bon résultat. Des données nombreuses mais biaisées peuvent produire un modèle trompeur. Des données exactes mais mal contextualisées peuvent conduire à une mauvaise décision. Le professionnel de la donnée doit donc toujours poser les questions suivantes : d’où viennent les données ? pourquoi ont-elles été collectées ? quelles populations représentent-elles ? quelles erreurs peuvent-elles contenir ? qui peut les utiliser ? pendant combien de temps ? Ces questions constituent le socle d’une pratique responsable.

4. Le Data Analyst : transformer les données en décisions

Le data analyst travaille principalement sur l’analyse descriptive et diagnostique. Il transforme des données brutes en indicateurs, tableaux de bord, rapports et analyses permettant aux responsables de comprendre une situation. Son travail commence souvent par une question métier : pourquoi les ventes ont-elles diminué ? quels produits progressent ? quels étudiants abandonnent davantage ? quels délais logistiques augmentent ? Une bonne analyse ne répond pas seulement par un chiffre. Elle explique le contexte, les tendances, les limites et les implications possibles.

Le data analyst doit donc posséder une double compétence. La première est technique : tableurs avancés, SQL, visualisation, statistiques descriptives et parfois programmation. La seconde est métier : comprendre les processus, les indicateurs et les décisions auxquelles l’analyse doit contribuer. Un analyste qui produit un graphique élégant mais inutilisable pour la décision n’a pas rempli sa mission. La qualité d’un analyste se mesure à sa capacité à rendre une information fiable, compréhensible et actionnable.

5. Activités quotidiennes du Data Analyst

Une journée type peut commencer par la vérification d’une demande d’analyse. L’analyste clarifie la question, identifie les sources pertinentes et définit les indicateurs. Il extrait ensuite les données, contrôle leur qualité, réalise des transformations et explore les résultats. Il construit éventuellement un tableau de bord puis présente ses conclusions. Une partie importante du travail consiste à répondre aux questions des utilisateurs et à améliorer progressivement les rapports existants.

Dans un contexte professionnel, l’analyste doit également documenter ses calculs. Une métrique utilisée par la direction doit avoir une définition précise. Par exemple, « chiffre d’affaires mensuel » doit préciser si les remboursements, taxes, commandes annulées ou paiements en attente sont inclus. Sans cette définition, deux services peuvent produire des chiffres différents et chacun croire avoir raison. La documentation est donc une compétence centrale, pas une tâche administrative secondaire.

6. SQL et exploration des données

SQL est une compétence fondamentale pour de nombreux data analysts. Elle permet

d’interroger des bases relationnelles, de filtrer, regrouper, joindre et agréger des données. La maîtrise de SQL ne consiste pas à mémoriser des commandes isolées. Il faut savoir traduire une question métier en logique d’interrogation. Pour analyser les ventes, l’analyste peut devoir relier clients, commandes, produits et dates, puis calculer des indicateurs par période et segment.

L’exploration doit toujours inclure des contrôles. Il faut vérifier les valeurs manquantes, les doublons, les valeurs extrêmes, les incohérences de format et les ruptures temporelles. Un graphique peut aider à repérer une anomalie qu’une requête ne révèle pas immédiatement. L’analyste doit également conserver une trace des transformations afin qu’un collègue puisse reproduire le résultat.

7. Visualisation et communication analytique

La visualisation transforme des résultats numériques en représentations facilitant la compréhension. Un bon graphique répond à une question précise. Une évolution temporelle appelle généralement une représentation différente d’une comparaison entre catégories. Le choix d’un graphique doit tenir compte de la nature de la variable, du nombre de catégories et du message à transmettre.

La communication analytique exige de distinguer résultat et interprétation. Dire que les ventes ont diminué de 12 % est un résultat. Affirmer que la diminution est causée par une campagne marketing est une interprétation qui nécessite des preuves supplémentaires. Le data analyst doit donc utiliser un langage prudent : association, tendance, hypothèse ou relation observée ne signifient pas automatiquement causalité. Cette rigueur protège l’organisation contre des décisions fondées sur des conclusions excessives.

8. Le Data Scientist : expérimenter et modéliser

Le data scientist intervient lorsque le problème nécessite davantage que la description des données. Il peut chercher à prévoir une demande, détecter une anomalie, classer des documents, recommander un produit ou estimer une probabilité. Son travail combine statistiques, programmation, machine learning, expérimentation et connaissance du domaine. Il doit formuler une hypothèse, préparer les données, choisir une méthode, entraîner ou ajuster un modèle, l’évaluer et communiquer ses limites.

Le data scientist doit résister à une tentation fréquente : commencer directement par l’algorithme. La première question est toujours le problème à résoudre. Une organisation peut demander « un modèle d’IA » alors qu’un simple indicateur ou une règle métier répondrait mieux au besoin. La sophistication technique doit être proportionnée à la valeur attendue, au coût et au risque.

9. Statistiques et raisonnement expérimental

Les statistiques fournissent au data scientist un langage pour mesurer l’incertitude et comparer des hypothèses. Moyenne, médiane, variance, distribution, corrélation, échantillonnage et intervalle d’incertitude sont des notions fondamentales. Il faut également comprendre que les données observées peuvent être influencées par la manière dont elles ont été collectées.

Dans une expérimentation, le professionnel doit définir ce qui est comparé, la population concernée et la métrique de succès. Une expérience mal conçue peut produire un résultat statistiquement convaincant mais professionnellement inutile. Le raisonnement expérimental oblige à examiner les biais, les variables confondantes et les effets de sélection. Il constitue donc une compétence durable, même lorsque les outils logiciels changent.

10. Machine learning : de la donnée au modèle

Le machine learning consiste à construire une fonction capable d'inférer une relation à partir de données. En apprentissage supervisé, les observations sont associées à une cible. En classification, la cible représente généralement une catégorie ; en régression, elle représente une valeur numérique. En apprentissage non supervisé, le système recherche des structures sans cible prédéfinie.

Le data scientist doit choisir une métrique adaptée au problème. Pour une classification, l'accuracy peut être insuffisante si les classes sont très déséquilibrées. La précision, le rappel, le score F1, les matrices de confusion ou les métriques spécifiques au domaine peuvent être plus pertinentes. Le choix de la métrique doit refléter le coût des erreurs. Dans un système de détection de fraude, manquer une fraude et déclencher une fausse alerte n'ont pas nécessairement les mêmes conséquences.

11. Entraînement, validation et test

Un modèle doit être évalué sur des données permettant d'estimer sa capacité à généraliser. La séparation entre entraînement et évaluation réduit le risque de mesurer simplement la mémorisation. La validation sert notamment à comparer des configurations et à ajuster les choix. Le jeu de test doit rester protégé autant que possible jusqu'à l'évaluation finale.

Le surapprentissage apparaît lorsqu'un modèle s'adapte excessivement aux données d'entraînement et perd en performance sur de nouvelles observations. Le professionnel peut utiliser des techniques de régularisation, validation croisée ou simplification du modèle selon le contexte. Mais il doit aussi rechercher les causes métier : données trop homogènes, fuite d'information, changement de distribution ou variable artificiellement prédictive.

12. Le Data Engineer : construire les fondations techniques

Le data engineer assure la circulation et la disponibilité des données. Il conçoit des pipelines permettant d'ingérer, transformer, stocker et distribuer des informations. Son travail peut inclure des bases de données, des entrepôts, des lacs de données, des flux temps réel, des systèmes de traitement distribué et des mécanismes de surveillance.

La mission du data engineer est souvent invisible pour les utilisateurs finaux, mais elle conditionne la fiabilité du système. Si les données arrivent avec retard, si les doublons sont fréquents ou si une transformation change sans documentation, l'analyse et les modèles deviennent fragiles. L'ingénieur doit donc penser en termes de fiabilité,

reproductibilité, performance, sécurité et coût.

13. Pipelines et orchestration

Un pipeline décrit une succession d'étapes de traitement. Par exemple, des données de ventes peuvent être extraites d'un système transactionnel, nettoyées, transformées puis chargées dans un entrepôt analytique. L'orchestration permet de planifier et surveiller ces étapes. Un pipeline professionnel doit gérer les erreurs, les reprises, les dépendances et les changements de schéma.

Le data engineer doit aussi distinguer traitement par lots et traitement en flux. Le traitement par lots convient lorsque les données peuvent être regroupées et traitées périodiquement. Le temps réel devient pertinent lorsque la décision dépend d'informations très récentes. Le choix doit être justifié par le besoin et non par l'effet de mode technologique.

14. Qualité, métadonnées et gouvernance des données

La qualité des données doit être contrôlée à plusieurs niveaux. Des tests peuvent vérifier les types, les valeurs autorisées, les volumes attendus et l'absence de certaines anomalies. Les métadonnées décrivent les données : définition, source, fréquence, propriétaire, transformation et conditions d'utilisation. La gouvernance clarifie les responsabilités et les règles d'accès.

Un système mature permet de répondre rapidement à des questions comme : qui est responsable de cette donnée ? quand a-t-elle été mise à jour ? quelles transformations lui ont été appliquées ? quels rapports l'utilisent ? quelles personnes peuvent y accéder ? Cette traçabilité devient particulièrement importante lorsque les données alimentent des décisions sensibles ou des modèles d'IA.

15. Les trois métiers dans un même projet

Considérons une entreprise qui souhaite prévoir les ventes. Le data engineer construit les flux et rend les données disponibles. Le data analyst explore les tendances historiques, définit les indicateurs et identifie les facteurs pertinents. Le data scientist expérimente un modèle de prévision. Après mise en production, les trois métiers peuvent continuer à collaborer : l'analyste suit les résultats, le data engineer surveille les pipelines et le data scientist analyse les performances du modèle.

Cette coopération montre que les métiers ne sont pas trois étapes complètement séparées. Ils forment une boucle. Les problèmes détectés par l'analyste peuvent conduire à une amélioration du pipeline. Les limites du modèle peuvent conduire à rechercher de nouvelles données. Les changements dans les systèmes sources peuvent nécessiter une adaptation de l'ingénierie. Une équipe data efficace organise donc des mécanismes de communication et de responsabilité partagée.

16. Différences essentielles entre les trois profils

Le data analyst répond principalement à la question « que se passe-t-il et que pouvons-nous comprendre ? ». Le data scientist ajoute « que peut-on prévoir, classer ou modéliser ? ». Le data engineer se demande « comment rendre les données fiables, disponibles et exploitables à l'échelle nécessaire ? ». Ces formulations sont simplifiées mais utiles pour distinguer les orientations dominantes.

Dans la réalité, les frontières évoluent. Un analyste avancé peut programmer des modèles simples. Un data scientist peut construire certaines parties d'un pipeline. Un data engineer peut réaliser des analyses exploratoires. L'objectif d'une carrière n'est donc pas de défendre des frontières rigides, mais de développer une spécialité principale tout en comprenant les interfaces.

17. Outils et environnements professionnels

Les outils varient selon les organisations. Pour l'analyse, les tableurs, SQL et outils de visualisation restent courants. Pour la science des données, Python ou R, bibliothèques statistiques et machine learning sont largement utilisés. Pour l'ingénierie, les bases de données, systèmes de traitement, outils d'orchestration, contrôle de version, conteneurs et plateformes cloud peuvent intervenir.

L'apprenant ne doit pas chercher à maîtriser simultanément tous les outils. Il doit comprendre les catégories et choisir un environnement cohérent. Un portfolio bien construit avec quelques technologies maîtrisées vaut généralement mieux qu'une liste de dizaines d'outils connus superficiellement. La capacité à apprendre rapidement une nouvelle technologie est elle-même une compétence professionnelle.

18. Étude de cas : prévision de la demande

Une chaîne de distribution souhaite prévoir la demande hebdomadaire. Le data engineer collecte ventes, stocks, calendriers et données externes. Le data analyst examine les tendances saisonnières et construit des indicateurs. Le data scientist compare plusieurs méthodes de prévision et mesure l'erreur. Le responsable métier vérifie que les résultats sont compatibles avec les réalités commerciales.

Le projet doit également définir le coût des erreurs. Une sous-estimation peut provoquer une rupture de stock, tandis qu'une surestimation peut immobiliser des ressources. La meilleure méthode n'est donc pas nécessairement celle qui minimise une métrique abstraite, mais celle qui répond au coût réel des décisions. Cette étude montre l'importance du dialogue entre données et métier.

19. Étude de cas : ONG et suivi des bénéficiaires

Une ONG souhaite améliorer le suivi de ses programmes. Le data engineer structure les données provenant de plusieurs projets. Le data analyst construit des indicateurs de couverture et de progression. Le data scientist peut étudier des facteurs associés à certains résultats si la qualité et la légitimité des données le permettent. Les responsables protection et programme contrôlent les risques liés aux personnes concernées.

Dans ce cas, la compétence technique ne suffit pas. Les données peuvent concerner des personnes vulnérables. Il faut donc appliquer des principes de minimisation, sécurité, accès contrôlé et utilisation proportionnée. L'équipe doit également éviter de présenter une corrélation comme une cause individuelle. Le projet devient un exemple concret de coopération entre expertise data et responsabilité professionnelle.

20. Étude de cas : administration publique

Une administration veut analyser les délais de traitement des dossiers. Le data engineer consolide les sources. Le data analyst identifie les étapes où les délais augmentent. Le data scientist peut éventuellement construire un modèle de prévision de charge. Les responsables métier examinent les résultats avant toute décision. La gouvernance vérifie les conditions d'accès et la traçabilité.

Le projet doit éviter qu'un indicateur soit utilisé pour pénaliser automatiquement un agent sans analyse du contexte. Une hausse du délai peut résulter d'une surcharge exceptionnelle, d'un changement réglementaire ou d'une défaillance informatique. Les professionnels de la donnée doivent donc fournir des outils d'aide à la décision plutôt que masquer la complexité derrière un score unique.

21. Méthode professionnelle de conduite d'un projet data

Une méthode robuste peut suivre huit étapes : cadrage du problème, inventaire des sources, contrôle de qualité, exploration, définition de la méthode, expérimentation, déploiement éventuel et suivi. À chaque étape, un livrable doit être produit. Le cadrage fournit une question précise. L'inventaire fournit une cartographie des sources. Le contrôle fournit un diagnostic de qualité. L'exploration fournit des constats. L'expérimentation fournit des résultats comparables.

Cette démarche évite de transformer le projet en suite d'essais techniques sans finalité. Elle facilite également la communication avec les décideurs. Un responsable peut comprendre où se situe le projet, quelles hypothèses restent ouvertes et quelles décisions doivent être prises. La documentation doit être maintenue tout au long du cycle de vie.

22. Sécurité et protection des données

Les professionnels data doivent intégrer la sécurité dès la conception. Les accès doivent être limités selon les responsabilités. Les données sensibles doivent être protégées et les échanges sécurisés. Les environnements de développement, test et production doivent être distingués lorsque le risque le justifie. Les secrets techniques ne doivent pas être intégrés directement dans les scripts ou documents publics.

La protection des données implique également de réfléchir à la finalité. Une donnée disponible n'est pas automatiquement une donnée que l'on peut utiliser pour n'importe quel projet. Le professionnel doit comprendre les politiques internes et les obligations applicables dans son contexte. Lorsque le projet concerne des données personnelles ou sensibles, l'expertise juridique et conformité doit être associée suffisamment tôt.

23. Éthique, biais et équité

Les biais peuvent provenir des données, de la sélection des variables, du processus de collecte, des objectifs ou de l'interprétation. Un modèle peut produire des performances globales satisfaisantes tout en étant moins fiable pour certains groupes. L'évaluation doit donc être adaptée au contexte et aux conséquences possibles.

Le professionnel doit documenter les limites et mettre en place des contrôles appropriés. Il faut éviter deux extrêmes : considérer tout modèle comme dangereux par principe ou considérer tout résultat numérique comme neutre. La responsabilité consiste à examiner les preuves, les usages et les conséquences. Cette approche rejoint les travaux sur la gouvernance responsable de l'IA et les cadres professionnels de gestion des risques.

24. Indicateurs de performance des métiers data

Pour un data analyst, les indicateurs peuvent inclure fiabilité des rapports, adoption des tableaux de bord, délai de production et satisfaction des utilisateurs. Pour un data scientist, on peut suivre performance du modèle, robustesse, stabilité, valeur métier et coût d'exploitation. Pour un data engineer, on peut mesurer disponibilité des pipelines, taux d'échec, fraîcheur des données, temps de traitement et coût d'infrastructure.

Il faut cependant éviter de réduire chaque métier à un seul indicateur. Un pipeline rapide mais souvent erroné n'est pas performant. Un modèle très précis mais trop coûteux ou impossible à expliquer peut être inadapté. Les indicateurs doivent donc couvrir qualité, efficacité, fiabilité, valeur et risque.

25. Travail en équipe et communication

Les équipes data réussissent lorsqu'elles établissent des responsabilités claires. Le data analyst doit pouvoir expliquer ses besoins en données. Le data engineer doit signaler les contraintes techniques. Le data scientist doit communiquer les limites du modèle. Le responsable métier doit préciser les décisions et conséquences. Les réunions doivent être orientées vers des questions concrètes et des livrables.

La documentation joue un rôle central. Un dictionnaire de données, des notes d'expérimentation, des règles de transformation et des décisions d'architecture permettent de préserver la connaissance. Sans documentation, l'organisation devient dépendante de quelques individus et les erreurs se reproduisent.

26. Lire une offre d'emploi Data Analyst

Une offre de data analyst doit être examinée selon les missions plutôt que les mots-clés. Recherchez les livrables : tableaux de bord, analyses ad hoc, reporting, indicateurs, études de performance. Identifiez les outils obligatoires et les compétences métier. Vérifiez également le niveau d'autonomie. Un poste qui demande seulement de produire des rapports n'a pas le même profil qu'un poste qui doit définir les indicateurs avec la direction.

Pour se préparer, l’apprenant peut construire trois projets : une analyse SQL documentée, un tableau de bord et une étude complète répondant à une question métier. Chaque projet doit présenter données, méthode, résultats, limites et recommandations. Ces éléments constituent des preuves concrètes de compétence.

27. Lire une offre de Data Scientist

Une offre de data scientist peut demander statistiques, programmation, machine learning, expérimentation et communication. Il faut vérifier si le poste est réellement orienté recherche, développement de modèles, analyse ou mise en production. Certaines annonces regroupent des responsabilités qui relèvent plutôt de plusieurs métiers.

L’apprenant doit examiner les données disponibles, les types de problèmes, les métriques et l’environnement de déploiement. Un portfolio pertinent peut présenter une classification, une régression, une prévision ou une détection d’anomalies, mais le projet doit surtout démontrer le raisonnement : choix de la méthode, comparaison, validation, interprétation et limites.

28. Lire une offre de Data Engineer

Une offre de data engineer peut mentionner bases de données, pipelines, orchestration, cloud, traitement distribué, qualité et sécurité. Le candidat doit comprendre les architectures et être capable d’expliquer comment une donnée passe de la source à l’utilisateur ou au modèle.

Un portfolio peut contenir un pipeline complet sur des données publiques : extraction, transformation, chargement, tests, documentation et surveillance simple. L’objectif est de montrer la fiabilité et non seulement la capacité à écrire du code. Un projet reproductible avec une documentation claire constitue une preuve particulièrement utile.

29. Construire une matrice de compétences

Une matrice peut comporter les colonnes compétence, niveau actuel, niveau cible, preuve et prochaine action. Pour le data analyst, les compétences peuvent inclure SQL, statistiques descriptives, visualisation, communication et compréhension métier. Pour le data scientist, on ajoute modélisation, validation, programmation et expérimentation. Pour le data engineer, on ajoute architecture, pipelines, bases de données, orchestration et fiabilité.

La matrice doit être régulièrement actualisée. Une compétence n’est pas acquise parce qu’un cours a été suivi. Elle est acquise lorsqu’une personne peut produire un résultat fiable dans une situation donnée. Les projets pratiques sont donc indispensables.

30. Exercice : choisir le bon métier

Imaginez une entreprise qui possède des données de ventes mais aucun système analytique fiable. Elle souhaite d’abord comprendre ses performances puis prévoir la demande.

Identifiez les responsabilités nécessaires. Expliquez pourquoi un data engineer est nécessaire avant ou en parallèle de certaines analyses. Décrivez ensuite le travail du data analyst puis celui du data scientist. Pour chaque profil, indiquez trois livrables et deux risques.

31. Exercice : diagnostic d’un mauvais modèle

Un modèle prédit les ventes avec une excellente performance sur les données d’entraînement mais une performance faible sur les nouvelles données. Proposez au moins cinq hypothèses : surapprentissage, fuite d’information, changement de distribution, données insuffisantes, mauvais découpage temporel ou variables instables. Pour chaque hypothèse, proposez une vérification. Cet exercice doit conduire l’apprenant à raisonner par diagnostic plutôt qu’à changer immédiatement d’algorithme.

32. Exercice : diagnostic d’un mauvais pipeline

Un tableau de bord affiche parfois des chiffres différents selon l’heure de consultation. L’équipe soupçonne un problème de pipeline. Proposez une méthode d’investigation : vérifier la fraîcheur des données, les doublons, les horaires d’exécution, les dépendances, les transformations et le cache. Définissez ensuite trois indicateurs de surveillance. L’objectif est de montrer que la qualité d’une analyse dépend aussi de l’infrastructure qui l’alimente.

33. Étude de cas professionnelle intégrée

Une entreprise de commerce électronique souhaite réduire les ruptures de stock. Le data engineer doit intégrer ventes, stocks, fournisseurs et délais. Le data analyst étudie les tendances, les produits saisonniers et les ruptures historiques. Le data scientist teste des méthodes de prévision. Le responsable logistique valide les hypothèses et décide des actions. Le product manager coordonne le projet.

Le projet est évalué selon plusieurs critères : précision des prévisions, réduction des ruptures, coût d’immobilisation, stabilité du pipeline et adoption par les équipes. Une approche strictement technique pourrait optimiser une métrique sans améliorer le résultat commercial. L’étude montre pourquoi les trois métiers doivent rester reliés à la décision métier.

34. Étude de cas : agriculture et prévision des rendements

Une organisation agricole veut utiliser des données météorologiques, historiques de production et informations agronomiques pour estimer les rendements. Le data engineer structure les sources. Le data analyst étudie les relations historiques. Le data scientist construit plusieurs modèles. Les agronomes interprètent les variables et contrôlent la plausibilité des résultats.

Une difficulté importante est la représentativité. Un modèle entraîné sur certaines régions peut fonctionner moins bien ailleurs. Le projet doit donc mesurer les performances selon les contextes et documenter les limites. Cette étude illustre

l’importance de l’expertise sectorielle dans les métiers de la donnée.

35. Étude de cas : éducation

Une plateforme de formation veut comprendre les facteurs associés à l’abandon. Le data analyst commence par décrire les parcours. Le data engineer vérifie que les événements pédagogiques sont correctement enregistrés. Le data scientist peut construire un modèle de risque. Mais le résultat ne doit pas devenir une étiquette définitive sur un apprenant. Il doit servir à identifier des situations où un accompagnement pourrait être utile.

Les responsables pédagogiques doivent pouvoir interpréter les indicateurs et contester une conclusion lorsque le contexte l’exige. Le projet devient ainsi un exemple d’IA augmentant l’action humaine plutôt que la remplaçant.

36. Méthode de documentation d’un projet

Une documentation professionnelle devrait contenir le contexte, la question, les sources, les transformations, les hypothèses, les méthodes, les résultats, les limites et les décisions. Pour un modèle, ajoutez les données d’entraînement et d’évaluation, les métriques, les versions et les conditions de déploiement. Pour un pipeline, ajoutez les dépendances, la fréquence, les règles de reprise et les contrôles de qualité.

La documentation doit être compréhensible par plusieurs publics. Une fiche technique détaillée peut être accompagnée d’un résumé exécutif. Cette capacité à adapter le niveau de détail est importante pour évoluer vers des fonctions de responsabilité.

37. Portfolio du Data Analyst

Un portfolio de data analyst peut commencer par une analyse d’un jeu de données public. Le candidat doit présenter une question métier, une préparation des données, des analyses, des visualisations et des recommandations. Un second projet peut montrer un tableau de bord. Un troisième peut présenter une analyse plus approfondie avec une comparaison de segments ou de périodes.

Chaque projet doit préciser les limites. Le candidat doit être capable d’expliquer pourquoi certains indicateurs ont été choisis et ce qu’ils ne permettent pas de conclure. Cette transparence renforce la crédibilité professionnelle.

38. Portfolio du Data Scientist

Le portfolio du data scientist doit montrer une démarche expérimentale. Un projet peut comparer plusieurs modèles sur un problème de classification. Le candidat doit expliquer les données, le prétraitement, la séparation des jeux, les métriques, les résultats et les limites. Un autre projet peut traiter une prévision temporelle ou une détection d’anomalies.

La priorité doit être donnée au raisonnement. Un notebook rempli de code mais sans

explication ne démontre pas suffisamment la compétence. Le lecteur doit comprendre la question, les choix effectués et les conséquences professionnelles du résultat.

39. Portfolio du Data Engineer

Le portfolio du data engineer peut présenter un pipeline reproductible. Il doit montrer la source, l'extraction, la transformation, le stockage, les tests et la documentation. Le candidat peut également présenter un schéma d'architecture et expliquer les choix de technologie. La gestion des erreurs et la surveillance sont des éléments particulièrement valorisants.

Un projet simple mais fiable est préférable à une architecture extrêmement complexe dont le fonctionnement n'est pas maîtrisé. Le portfolio doit montrer la capacité à construire une solution adaptée au besoin et aux ressources.

40. Travail individuel : projet complet

Choisissez un domaine : agriculture, commerce, éducation, santé, tourisme, administration, finance ou ONG. Définissez une question pouvant être traitée par les données. Identifiez au moins trois sources possibles. Proposez le rôle du data engineer, du data analyst et du data scientist. Définissez cinq indicateurs, trois risques et cinq étapes de projet. Produisez ensuite une fiche de deux à trois pages présentant votre solution.

41. Préparation à l'évaluation continue

Pour préparer l'évaluation, l'apprenant doit être capable de distinguer les responsabilités des trois métiers, expliquer le rôle de SQL, définir un pipeline, expliquer entraînement et test, identifier un problème de surapprentissage, choisir une métrique appropriée et reconnaître l'importance de la qualité des données. Il doit également savoir appliquer ces notions à un cas professionnel.

La mémorisation de définitions ne suffit pas. Chaque notion doit être reliée à un exemple. Si une question demande quel profil construire un pipeline, il faut pouvoir expliquer pourquoi et dans quelles circonstances le data scientist ou l'analyste peut néanmoins intervenir.

42. Compétences humaines et professionnelles

Les trois métiers nécessitent communication, curiosité, rigueur, capacité d'apprentissage et travail en équipe. Le data analyst doit souvent expliquer simplement des résultats complexes. Le data scientist doit défendre une méthode tout en reconnaissant ses limites. Le data engineer doit communiquer les contraintes d'architecture et les incidents. La collaboration exige écoute et capacité à négocier les priorités.

La rigueur documentaire est également une compétence humaine : elle montre que le professionnel pense au collectif et à la continuité de l'activité. Dans une

organisation, une solution qui dépend entièrement de la mémoire d’une personne représente un risque opérationnel.

43. Évolution des métiers avec l’IA générative

L’IA générative transforme aussi les métiers data. Les assistants peuvent accélérer l’écriture de requêtes, la documentation, l’exploration, la génération de code ou l’explication de résultats. Cependant, le professionnel reste responsable de la vérification. Une requête générée automatiquement peut être syntaxiquement correcte mais logiquement erronée. Un code produit par une IA peut contenir une faille. Une interprétation peut être trop confiante.

La compétence émergente consiste donc à utiliser l’IA comme assistant tout en conservant une capacité indépendante de contrôle. Les meilleurs professionnels développeront une combinaison de maîtrise technique, jugement critique et capacité à orchestrer des outils.

44. Carrière et trajectoires possibles

Une personne peut entrer par l’analyse de données puis évoluer vers la data science, le produit ou la gouvernance. Une personne disposant d’une formation informatique peut commencer par l’ingénierie des données puis évoluer vers le machine learning engineering ou MLOps. Un professionnel métier peut développer une spécialisation analytique sans devenir ingénieur à temps plein.

La trajectoire dépend des aptitudes et du contexte. Il n’est pas nécessaire que tout le monde devienne data scientist. Une organisation a besoin de plusieurs profils complémentaires. Une spécialisation claire et bien démontrée peut être plus utile qu’une ambition vague de devenir « expert IA ».

45. Plan de progression sur douze mois

Un débutant peut consacrer les trois premiers mois aux fondamentaux : tableur, statistiques descriptives, SQL et culture data. Les mois quatre à six peuvent être consacrés à un portfolio analytique. Les mois sept à neuf peuvent introduire Python et le machine learning selon l’objectif. Les mois dix à douze peuvent approfondir une spécialité et construire un projet complet. Chaque période doit comporter une production vérifiable.

Pour un profil data engineer, la progression peut mettre davantage l’accent sur bases de données, programmation, pipelines, contrôle de version, cloud et orchestration. Pour un analyste, la visualisation, la communication et la compréhension métier peuvent être prioritaires. Le plan doit donc être personnalisé.

46. Ressources vidéo et apprentissage complémentaire

Pour compléter la leçon, l’apprenant peut consulter des conférences universitaires et professionnelles sur la science des données, les statistiques, le machine learning et

l’ingénierie des données. Les ressources de cours universitaires ouverts, conférences de laboratoires de recherche et présentations professionnelles sont particulièrement utiles. Pour chaque vidéo, il est recommandé de relever l’auteur, la date, trois idées principales, une notion technique nouvelle et une application possible dans son contexte professionnel.

47. Bibliographie indicative

Pour approfondir, consulter notamment Foster Provost et Tom Fawcett, Data Science for Business, pour le raisonnement orienté décision ; Wes McKinney, Python for Data Analysis, pour l’analyse avec Python ; Jake VanderPlas, Python Data Science Handbook, pour les outils de science des données ; Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, pour l’apprentissage automatique ; Martin Kleppmann, Designing Data-Intensive Applications, pour les systèmes de données ; et les publications professionnelles du NIST, de l’OCDE et de l’UNESCO sur la gestion des risques et la gouvernance de l’IA. L’apprenant doit vérifier les éditions disponibles et compléter ces références avec des ressources actualisées.

48. Synthèse finale et transition

Data Analyst, Data Scientist et Data Engineer sont trois métiers complémentaires qui participent à une même chaîne de valeur. L’analyste transforme les données en compréhension et en indicateurs. Le scientifique des données expérimente et construit des modèles lorsque la situation le justifie. L’ingénieur des données construit les fondations permettant aux informations de circuler de manière fiable. Dans les organisations modernes, ces rôles travaillent avec des responsables métier, product managers, ingénieurs logiciels, spécialistes sécurité et fonctions de gouvernance.

La compétence durable n’est pas la maîtrise d’un outil particulier. C’est la capacité à comprendre un problème, sélectionner une méthode adaptée, vérifier la qualité des données, produire un résultat reproductible, expliquer les limites et mesurer la valeur créée. Cette logique prépare directement le module suivant, consacré à l’ingénierie IA, au machine learning engineering et au MLOps.