Certification « Expert(e) en science des données »

BLOC 1 : COLLECTER, TRANSFORMER ET SECURISER DES DONNEES

REFERENTIEL D'ACTIVITES *décrit les situations de travail et les activités exercées, les métiers ou emplois visés*	**REFERENTIEL DE COMPETENCES** *identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel d'activités*	**REFERENTIEL D'EVALUATION** *définit les critères et les modalités d'évaluation des acquis*	
		MODALITÉS D'ÉVALUATION	**CRITÈRES D'ÉVALUATION**
		Type d'évaluation : Mise en situation professionnelle réelle ou fictive. **Attendus du candidat :** À partir de l'analyse d'une organisation réelle ou fictive de son choix, le candidat propose une stratégie de collecte, de transformation et de sécurité des données. **Livrable attendu :** Le candidat remet au jury un dossier écrit comprenant :	
A1.1 : Collecte de données structurées et non structurées - Identification des sources de données. - Élaboration d'une stratégie de collecte de données	C1.1.1 : Élaborer une stratégie de collecte de données en définissant les données utiles et nécessaires pour répondre à une problématique, et en identifiant les sources des données afin de cadrer le travail à réaliser pour collecter les données ciblées.	- Une stratégie de collecte de données	La stratégie de collecte présentée permet d'identifier : - Les objectifs de la collecte - les données utiles et nécessaires pour répondre à la problématique - les sources de données - les moyens envisagés

- Mise en place de techniques de collecte de données - Interrogation des bases de données	C1.1.2 : Mettre en œuvre des techniques de collecte de données en exploitant les API[1] externes et les bases de données disponibles, des techniques de web crawling et de web scraping[2] afin de recueillir les données ciblées.	- Un exemple de collecte de données	Un exemple de collecte de données est présenté avec l'utilisation des techniques suivantes : - Web crawling - Web Scraping - requêtes SQL - API externes Pour chaque technique utilisée, la qualité de collecte est démontrée au niveau de : - l'exhaustivité des données collectées - l'exactitude des données collectées - le cadre règlementaire (respect de la propriété intellectuelle et du droit d'auteur notamment)
- Automatisation de la collecte de données	C1.1.3 : Automatiser la collecte de données en mettant en place des tâches planifiées et/ou des flux temps réel, en utilisant des logiciels d'automatisation afin de garantir l'actualisation des données.	- Une méthode d'automatisation de collecte	La méthode d'automatisation utilisée pour la collecte des données est présentée et argumentée. Elle comporte par exemple : - des workflows - des scripts et des librairies - des outils d'automation - des ordonnanceurs L'automatisation mise en place permet également de garantir l'actualisation des données.
A1.2 : Stockage des données structurées et non structurées - Élaboration de la stratégie de stockage - Conception du modèle de données	C1.2.1 : Élaborer la stratégie de stockage des données avec un modèle de données adéquat en intégrant les différents types de données, l'utilisation envisagée (analyse, stockage, disponibilité, accessibilité) et le volume à stocker afin d'organiser le stockage des données.	- Une stratégie de stockage des données - Un modèle de données	La définition d'une stratégie de stockage permet de répondre à l'utilisation envisagée (ex : disponibilité, analyse, stockage, accessibilité). La description formelle du modèle de données permet d'identifier : - l'organisation des données

[1] API = Application Programming Interface
[2] Web crawling et web scraping = exploration et « grattage » du web

			- les règles garantissant leur intégrité - les moyens de les manipuler
- Construction de bases de données et solutions de stockage Big Data	C1.2.2 : Construire une base de données en sélectionnant la technologie (SQL, NoSQL), un système de gestion de base de données (SGBD) et/ou une solution de stockage BIG DATA, en assurant le paramétrage et l'implémentation afin de mettre en œuvre le modèle de données qui garantit la disponibilité et l'intégrité des données.	- Une base de données - Une solution de stockage Big Data	Le choix des solutions de stockage des données (ex : système de bases de données relationnelles, data Warehouse, data lake…) est justifié et il permet de répondre à la problématique du projet. La base de données est organisée suivant un design pattern adapté au besoin.
A1.3 : Structuration, transformation et enrichissement des données - Sélection de la technologie et des outils de transformation des données	C1.3.1 : Sélectionner les technologies et les outils de traitement de données en identifiant les solutions existantes et en comparant leurs avantages et leurs inconvénients afin de traiter efficacement les données collectées.	- Une présentation des outils et des technologies de traitement des données sélectionnées	La présentation des outils et technologies de traitement de données permet d'identifier les avantages et les inconvénients des technologies sélectionnées (ex : Scala, Python, SQL) Les technologies choisies permettent de répondre efficacement à la problématique énoncée.
- Transformation des données (formatage, consolidation, agrégation, profilage, jointure)	C1.3.2 : Transformer les données à l'aide de langage de programmation ou en utilisant des outils dédiés (Talend, Spark…) afin d'obtenir des données nettoyées exploitables.	- Une présentation des données transformées, des méthodes et outils utilisés	La présentation des données transformées permet d'identifier les transformations effectuées depuis les données sources en matière de : - formatage, - consolidation, - agrégation, - profilage, - jointure, - calcul. Les données transformées sont exploitables.

- Déploiement d'un processus ETL[3] (Extract, Transform, Load) - Automatisation et orchestration du traitement des données	C1.3.3 : Développer un processus ETL en identifiant les bénéfices des technologies ETL (ex : facilité de développement), en exploitant la technologie ETL préalablement sélectionnée afin d'automatiser l'extraction, la transformation et le chargement de données.	- Un processus ETL - Les solutions utilisés pour l'automatisation et l'orchestration du traitement des données	La solution ETL choisie (ex : Talend, SSIS[4], plateforme AWS Glue) est justifiée avec les bénéfices attendus (ex : facilité de développement, efficacité…) Le processus ETL développé permet d'automatiser et d'orchestrer le traitement des données.
A.1.4 : Sécurisation des données - Définition de la politique de sécurité et de confidentialité des données stockées dans le respect du cadre règlementaire - Évaluation des risques de sécurité des données et audit de sécurité - Gestion des incidents de sécurité	C1.4.1 : Définir la politique de sécurisation des données en évaluant les risques, en qualifiant leur niveau de sensibilité, en identifiant les droits d'accès selon les rôles des différentes parties prenantes et en respectant les exigences légales (ex : RGPD[5]) afin de garantir la bonne utilisation et l'intégrité des données.	- Une politique de sécurité des données	La présentation de la politique de sécurité permet d'identifier : - Les enjeux de sécurité (ex : données sensibles) - Les moyens mis en œuvre pour sécuriser les données (ex : chiffrement des données, surveillance, sauvegarde, différenciation des rôles) La politique de sécurité permet de garantir l'intégrité des données au regard des enjeux identifiés.
- Définition de l'architecture de sécurité des données et mise en œuvre des solution de protection des données	C1.4.2 : Concevoir une architecture sécurisée et robuste, en intégrant des mesures de sécurité multicouches et des contrôles d'accès stricts, en mettant en place des solutions de protection des données (ex : chiffrement) pour sécuriser les données en transit et au repos, en veillant à l'anonymisation des données personnelles afin de répondre aux normes de protection de la vie privée et de sécurité des données.	- Un schéma d'architecture de sécurité	Le schéma d'architecture de sécurité précise : - Les moyens mis en œuvre pour sécuriser les données - Les différentes zones de sécurité avec leurs plans d'adressage - Les flux d'échanges de données L'architecture de sécurité permet d'assurer la protection des données.

[3] ETL = Extract Transform Load

[4] SSIS = SQL Server Integration Services

[5] RGPD = Le règlement général sur la protection des données

BLOC 2 : ANALYSER, ORGANISER ET VALORISER DES DONNEES

REFERENTIEL D'ACTIVITES *décrit les situations de travail et les activités exercées, les métiers ou emplois visés*	**REFERENTIEL DE COMPETENCES** *identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel d'activités*	**REFERENTIEL D'EVALUATION** *définit les critères et les modalités d'évaluation des acquis*	
		MODALITÉS D'ÉVALUATION	**CRITÈRES D'ÉVALUATION**
		Type d'évaluation : Mise en situation professionnelle réelle ou fictive. **Attendus du candidat :** À partir de l'analyse d'une organisation réelle ou fictive de son choix, le candidat propose une stratégie d'analyse, d'organisation et de valorisation des données **Livrable attendu :** Le candidat remet au jury un dossier écrit comprenant :	
A2.1 : Analyse des données - Analyse des besoins, de la problématique et du contexte	C2.1.1 : Analyser les besoins métier et les enjeux exprimés par un commanditaire en réalisant des entretiens exploratoires et en récupérant les informations stratégiques nécessaires afin de cadrer le travail d'analyse des données à produire.	- Une analyse du besoin	L'analyse du besoin permet d'identifier : - Les enjeux et la problématique - Le contexte - L'environnement - Les contraintes (ex : délai, logistique, coût, technique, réglementaire)
- Construction d'un plan d'analyse - Identification des métriques et des indicateurs	C2.1.2 : Définir les axes d'analyse et les métriques en identifiant les données à exploiter, celles disponibles et pertinentes pour traduire la problématique d'entreprise énoncée en problème numérique.	- Une présentation d'un plan d'analyse	Le plan d'analyse décrit les axes et les métriques nécessaires. Il permet de traduire la problématique client en problème numérique.
- Réalisation de requêtes sur une grande quantité de données - Élaboration de calculs pour obtenir les indicateurs	C2.1.3 : Réaliser des requêtes et des calculs en utilisant des outils de dashboarding, des tableurs, des requêtes SQL ou scripts Python afin de produire une analyse des données préalablement collectées.	- Une présentation des requêtes et des résultats sous forme de dashboard	Plusieurs techniques d'analyse sont présentées. (ex : requêtes SQL, Jupyter Notebook, Tableurs, Dashboard) Elles permettent d'obtenir rapidement des résultats justes au regard de la problématique traitée.

- Élaboration des modèles statistiques d'analyse de données - Conception et réalisation des tests d'hypothèses	C2.1.4 : Élaborer des modèles statistiques et des tests d'hypothèses en modélisant des relations entre les variables, en évaluant la pertinence des résultats des simulations afin de valider ou réfuter des hypothèses.	- Une méthodologie de tests statistiques	La méthodologie présentée comporte : - La formulation d'une ou plusieurs hypothèses - Le test statistique associé - L'interprétation des résultats Les résultats obtenus par la méthodologie présentée permettent de valider ou réfuter l'hypothèse initiale.
A2.2 : Visualisation des données, interprétation et communication des résultats - Visualisation des données (ex : graphiques, tableaux de bord, rapports)	C2.2.1 : Représenter les données en choisissant les modèles de représentation les plus adaptés (ex : histogramme, Heat map, nuage de points) et en utilisant des outils de représentation adaptés (ex : Office, power BI) afin de permettre la compréhension et l'exploitation des données par le public visé.	- La visualisation des résultats de l'analyse	Le choix des outils de mise en forme et des représentations est justifié avec les bénéfices attendus (ex : lisibilité, facilité d'utilisation) La mise en forme (ex : couleurs, légendes, échelles, graphiques) permet de communiquer les données au public ciblé avec clarté et justesse. Les choix de mise en forme prennent en compte les spécificités des personnes en situation de handicap le cas échéant (ex : formes, contrastes, choix des couleurs)
- Interprétation et communication des résultats - Présentation de recommandations	C2.2.2 : Présenter des recommandations, en préparant son discours et des arguments, en structurant son analyse sur les données représentées afin d'aider les décideurs à établir leurs stratégies.	- Une présentation de recommandations	La présentation des recommandations est structurée, synthétique et argumentée. Les recommandations permettent d'éclairer le commanditaire pour l'aider dans sa prise de décision.
A2.3 : Support et accompagnement des utilisateurs - Formation des utilisateurs à l'utilisation des données et des outils de visualisation	C2.3.1 : Former les utilisateurs à l'utilisation des données et des outils de visualisation en analysant le besoin de montée en compétences et en élaborant des supports de formation et de sensibilisation adaptés afin de permettre aux utilisateur de maitriser l'exploitation des données.	- Un support de formation	L'enjeu et le sujet de la formation ou de la sensibilisation sont présentés. Le support est adapté au sujet et permet de monter en compétences le public visé (ex : présentation power point, newsletter, mail de bonnes pratiques…)

- Rédaction de la documentation	C2.3.2 : Rédiger la documentation technique d'utilisation du système d'analyse de données en identifiant le public concerné, en détaillant le fonctionnement du système d'analyse de données afin d'assurer la traçabilité et la transmission aux utilisateurs.	- Une documentation technique	La documentation technique comporte : - La description des sources de données (ex : origine, périmètre) - La description des méthodes de calculs - La description technique et fonctionnelle des indicateurs La documentation permet la compréhension, la transmission et la reproductibilité de l'analyse de données.

BLOC 3 : ELABORER ET PILOTER UN PROJET DATA

REFERENTIEL D'ACTIVITES *décrit les situations de travail et les activités exercées, les métiers ou emplois visés*	**REFERENTIEL DE COMPETENCES** *identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel d'activités*	**REFERENTIEL D'EVALUATION** *définit les critères et les modalités d'évaluation des acquis*	
		MODALITÉS D'ÉVALUATION	**CRITÈRES D'ÉVALUATION**
		Type d'évaluation : Mise en situation professionnelle réelle ou fictive. **Attendus du candidat :** À partir de l'analyse d'une organisation réelle ou fictive de son choix, le candidat élabore un projet Data. **Livrable attendu :** Le candidat remet au jury un dossier écrit comprenant :	
A3.1 : Élaboration et cadrage du projet DATA - Analyse des contraintes, des menaces et des opportunités - Définition des objectifs et du périmètre du projet	C3.1.1 : Définir les objectifs à atteindre et le périmètre du projet, en analysant les contraintes techniques et réglementaires, en étudiant le contexte et les enjeux afin de dimensionner le projet en termes de délai et budget.	- Le cadrage du projet	Le cadrage du projet permet d'identifier : - La problématique - Les objectifs et les livrables du projet - Le cadre règlementaire le cas échéant - Les contraintes et les points de vigilance - Les enjeux RSE le cas échéant.
- Dimensionnement du projet - Budgétisation du projet - Étude de faisabilité	C3.1.2 : Dimensionner le projet en évaluant la charge de travail et les ressources nécessaires (humaines, matérielles) au regard des exigences attendues et des contraintes préalablement définies afin d'estimer le temps et le budget nécessaires à la faisabilité du projet.	- Le dimensionnement du projet	Le dimensionnement du projet comporte : - Les ressources humaines nécessaires - Les ressources matérielles et logistiques - Un chiffrage du projet (coût et délai) - Une analyse de la faisabilité Le dimensionnement du projet permet d'atteindre les objectifs de qualité, coût, délai fixés par le commanditaire.

- Rédaction de la documentation projet (ex : Cahier des charges fonctionnel et technique, charte éthique)	C3.1.3 : Rédiger la documentation projet, en identifiant les parties prenantes concernées, en prenant en compte l'ensemble des caractéristiques du projet, afin de clarifier et formaliser les attendus.	- La documentation projet	La documentation projet est en adéquation avec le cadrage du projet et permet de présenter l'ensemble des caractéristiques du projet. (ex : cahier des charges, spécifications techniques, fonctionnelles) La documentation est rédigée dans un vocabulaire compréhensible par les parties prenantes.
A3.2 : Pilotage du projet DATA - Choix de la méthodologie projet - Planification - Allocation des ressources	C3.2.1 : Planifier l'exécution du projet en organisant la répartition et l'ordonnancement des activités, le planning prévisionnel de réalisation et les ressources nécessaires à son exécution, en prenant en considération les personnes en situation de handicap afin de suivre les différentes phases du projet.	- Le planning projet	Le choix de la méthodologie de gestion de projet est justifié avec les bénéfices attendus (ex : Kanban, Scrum, Lean). L'outil utilisé pour la planification (ex : diagramme de Gantt, rétroplanning) est compatible avec la méthodologie projet choisie. Le planning du projet est découpé en phases, en tâches ou lots. Il permet de visualiser les différentes phases du projet (ex : collecte données, analyse, restitution). Les tâches sont assignées aux différents membres de l'équipe selon leurs compétences (matrice RACI[6], RASCI[7]...) et tiennent compte des personnes en situation d'handicap. Les points de vigilance sont soulignés (ex : chemin critique, compétences rares)
- Construction d'un outil de pilotage - Définition des indicateurs de suivi de la performance	C3.2.2 : Suivre l'avancement du projet en mettant en place un outil de suivi (logiciel de suivi, tableau de bord), en définissant les indicateurs (qualitatifs et/ou quantitatifs) pour chaque jalon défini dans le planning, en réalisant des reportings et des comptes	- Un outil de suivi de projet - Un tableau de bord	L'outil de suivi (ex : logiciel de suivi, tableau de bord) permet de piloter le projet en adéquation avec la méthodologie projet choisie. Le choix des indicateurs qualitatifs et quantitatifs est argumenté.

[6] RACI = Responsible, Accountable, Consulted, Informed
[7] RASCI = Responsible, Accountable, Support, Consulted, Informed

	rendus de réunion afin d'anticiper les aléas éventuels.		Les indicateurs permettent de suivre l'avancement du projet, le respect des délais et la maitrise des coûts.
A3.3 : Management d'équipe - Constitution de l'équipe projet - Évaluation des compétences et du besoin de montée en compétences	C3.3.1 : Évaluer les besoins en compétences de l'équipe projet, en collaborant avec le service Ressources Humaines, en établissant un plan de développement des compétences et en orientant les membres de l'équipe vers des formations adaptées, afin de renforcer l'équipe responsable de mener à bien le projet DATA.	- Un plan de développement des compétences	Les compétences à mobiliser dans le cadre du projet sont identifiées. Une grille d'évaluation des compétences actuelles et des compétences à acquérir est commentée. Un plan de développement des compétences adapté au projet est établi et détaillé. Il permet de monter en compétences le public visé. Des formations sont préconisées en fonction des besoins du projet et du profil des membres de l'équipe. Les modalités de formation sont adaptées pour prendre en considération les spécificités liées au handicap des personnes formées. (ex : aménagement matériel, temps supplémentaire…)
- Animation et management de l'équipe projet	C3.3.2 : Piloter l'équipe projet en affectant les missions à réaliser, en prenant en compte les spécificités des membres de l'équipe, en intégrant les spécificités d'un contexte multiculturel, international, en utilisant les différentes techniques de communication et d'animation managériale pour favoriser le bon fonctionnement de l'équipe.	- Les outils de communication et managériaux utilisés	La charge de travail est répartie sur l'ensemble de l'équipe de manière équilibrée. Les outils collaboratifs et les routines managériales utilisés sont détaillés et justifiés. Ils permettent de garantir le bon fonctionnement et la collaboration des membres de l'équipe projet. Les personnes en situation de handicap sont prises en compte (ex : bonne intégration, poste de travail adapté).
- Gestion du risque - Réajustement et arbitrage	C3.3.3 : Procéder aux arbitrages et aux réajustements nécessaires à partir de l'analyse des écarts entre le prévisionnel et l'état du projet à date, en utilisant des outils d'aide à la décision (ex : logigramme) afin de garantir la performance du	- La présentation d'un cas d'arbitrage rencontré au cours du projet	La problématique qui nécessite un arbitrage est exposée avec ses conséquences potentielles. Les options possibles pour y remédier sont détaillées.

	projet dans le respect des objectifs de qualité, coûts et délai.		La décision d'arbitrage est argumentée et permet de résoudre la problématique.
A3.4 : Veille, éthique et gouvernance des données - Veille technologique - Veille réglementaire (RGPD[8], Patriot Act, Data Act, IA Act)	C3.4.1 : Mettre en place un système de veille technologique et réglementaire en matière de science des données et d'Intelligence Artificielle à l'aide de recherches documentaires, de plateformes de partage, de webinars afin d'être alerté des évolutions qui impacteraient les pratiques métier.	- Une méthodologie de veille	Le choix de la méthodologie de recueil de l'information est argumenté avec les bénéfices attendus. (Ex : Utilisation d'un outil d'automatisation de la veille, inscription à des salons, réseaux de professionnels) Le résultat d'une action de veille est présenté et permet d'identifier : - L'impact engendré sur les pratiques métier - Les avantages et les inconvénients de cette évolution métier
- Prise en compte de la réglementation et des normes - Définition et mise en œuvre d'une gouvernance des données responsable	C3.4.2 : Intégrer dans ses pratiques métier les enjeux en termes de données responsables, de responsabilité sociétale et environnementale (RSE), de sécurité, d'éthique et de confidentialité des données en se tenant informé des évolutions du cadre juridique, à travers une recherche documentaire ou en étant accompagné par des juristes afin d'agir dans le respect de la législation.	- Un plan d'actions relatif aux enjeux RSE, de sécurité, d'éthique et de confidentialité	Les enjeux de la science des données en termes de RSE sont détaillés. Les arbitrages de priorisation pour établir un plan d'actions sont précisés et justifiés. Ce plan d'actions précise : - Le sujet traité (ex : confidentialité, réduction des émissions CO2) - L'action mise en œuvre - Les délais envisagés - Une estimation des coûts - Les résultats attendus

[8] RGPD = Le règlement général sur la protection des données (RGPD)

BLOC 4 : CONCEVOIR ET OPERER UNE INFRASTRUCTURE DATA (Spécialité DATA ENGINEER)

REFERENTIEL D'ACTIVITES *décrit les situations de travail et les activités exercées, les métiers ou emplois visés*	**REFERENTIEL DE COMPETENCES** *identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel d'activités*	**REFERENTIEL D'EVALUATION** *définit les critères et les modalités d'évaluation des acquis*	
		MODALITÉS D'ÉVALUATION	**CRITÈRES D'ÉVALUATION**
		Type d'évaluation : Mise en situation professionnelle réelle ou fictive. **Attendus du candidat :** À partir de l'analyse d'un projet réel ou fictif, le candidat présente la stratégie de conception d'une architecture DATA et de maintien d'une infrastructure DATA. **Livrable attendu :** Le candidat présente une soutenance comprenant :	
A4.1 : Analyse des besoins et définition de l'architecture - Analyse des besoins métiers - Analyse des contraintes et de l'environnement du projet	C4.1.1 : Analyser l'environnement du projet en recueillant les besoins métiers, les volumes de données à traiter, en réalisant un état des lieux des composants existants afin d'orienter le choix de conception de l'architecture DATA à mettre en œuvre.	- Un rapport d'analyse	Les rapport d'analyse permet d'identifier : - Les besoins - Les enjeux du projet - L'environnement - Les contraintes (ex : coût, délais, complexité d'implémentation) - L'état de l'existant Ce rapport d'analyse permet de cadrer le travail de conception et de déploiement de l'architecture DATA.
- Choix de l'infrastructure, des différents composants et des langages de programmation	C4.1.2 : Sélectionner l'ensemble des composants et technologies de l'infrastructure en étudiant les solutions existantes, en vérifiant leur compatibilité et les normes en vigueur afin de concevoir une architecture DATA correctement dimensionnée pour le projet.	- Une présentation des composants de l'architecture DATA	La présentation du choix des composants comporte : - La liste des composants et des technologies sélectionnés (ex : ressources dans le cloud, serveurs dédiés, systèmes requis, applicatifs souhaités, …) - Les avantages attendus (ex : scalabilité, performance, facilité d'implémentation…)

			- Les points de vigilance (ex : vendor lock-in) - Une estimation des coûts Ces composants et ces technologies permettent de concevoir une architecture DATA qui répondent efficacement au besoin exprimé par le commanditaire
A4.2 : Conception et déploiement de l'infrastructure DATA - Conception de l'architecture d'entrepôt de données (ex : modélisation, diagramme)	C4.2.1 : Concevoir une architecture d'entrepôt de données en s'appuyant sur le cahier des charges, en sélectionnant les composants appropriés afin d'optimiser le stockage des données en termes de rapidité, de sécurité et d'accessibilité.	- Un schéma de données	Le schéma de données permet d'identifier : - Le type de données - Les modalités d'accès aux données - L'organisation des données
- Conception des pipelines de données, des workflows automatisés et des outils permettant de traiter des données de masse	C4.2.2 : Mettre en place des pipelines DATA temps réel ou asynchrones à l'aide d'outils BIG DATA (ex : DATA brick, plateforme DBT[9], Snowflake), Cloud et/ou On premise afin d'automatiser la transformation et la transmission les données.	- Des pipelines de traitement de la donnée	3 méthodes de traitement de la donnée sont présentées. Elles sont réalisées avec : - un pipeline temps réel (ex : SQL, Python) - un orchestrateur (ex : Apache Airflow) - des calculs distribués (ex : Spark) Les traitements présentés permettent de produire les DATA demandées.
- Mise en place d'outils d'intégration et de déploiement continu	C4.2.3 : Automatiser l'intégration et le déploiement des composants en utilisant des outils d'intégration et de développement continu, afin d'industrialiser la mise en production de l'architecture DATA.	- Un pipeline CI/CD	Le pipeline CI/CD présenté permet d'automatiser les tâches d'intégration et de déploiement continu.
A4.3 : Supervision et exploitation de l'infrastructure DATA - Mise en place d'outils de supervision et de tests	C4.3.1 : Mettre en place un système de supervision et d'alertes en déployant des outils de supervision, et en déterminant les indicateurs de suivi pertinents	- Une présentation du système de supervision	La présentation du système de supervision permet d'identifier : - Les éléments et les indicateurs à surveiller - Les choix et la configuration des outils de supervision

[9] DBT = Data Build Tool

	afin de s'assurer du bon fonctionnement des composants de l'architecture DATA.		- La visualisation des indicateurs - Le système d'alertes Le système de supervision permet de surveiller le bon fonctionnement de l'infrastructure.
- Administration des systèmes (DataOPS)	C4.3.2 : Exploiter les systèmes et les équipements de l'infrastructure DATA en respectant les procédures d'administration et de maintien en condition opérationnelle afin de garantir l'intégrité et la disponibilité des données de l'organisation.	- Une feuille de route d'exploitation	La feuille de route d'exploitation présentée comporte : - les tâches à réaliser (ex : expiration et renouvellement des certificats) - les échéances (ex : tests, mises à jour) - la planification de la maintenance - les points de vigilance La feuille de route d'exploitation permet de maintenir en condition opérationnelle l'infrastructure DATA.
- Rédaction de la documentation technique	C4.3.3 : Rédiger la documentation fonctionnelle et technique en identifiant le public concerné et les objectifs attendus, en prenant en compte l'ensemble des caractéristiques de l'infrastructure DATA afin de transférer les procédures et les modalités de fonctionnement aux différents utilisateurs.	- Une documentation technique	Le choix de la documentation technique présentée (ex : procédure de configuration) permet une bonne utilisation de l'infrastructure DATA.
A4.4 : Maintenance et sécurisation de l'infrastructure DATA - Création d'un cahier de recettes - Réalisation des tests de fonctionnalités	C4.4.1 : Élaborer le cahier de recette en rédigeant les scénarios de tests et les résultats attendus afin de détecter les anomalies de fonctionnement et les régressions éventuelles.	- Un cahier de recettes et de tests	Le cahier de recettes et de tests reprend l'ensemble des fonctionnalités attendues. Les tests fonctionnels, structurels et de sécurité exécutés sont conformes au plan défini.
- Investigation et résolution d'un incident technique	C4.4.2 : Résoudre un incident technique en investiguant la source du problème et en déployant une méthode de résolution afin de rétablir la disponibilité du service.	- Une méthodologie d'investigation et de traitement d'un incident.	Un exemple d'incident est présenté avec la méthodologie d'investigation appliquée. La méthodologie permet d'identifier : - la nature du problème - les actions à mettre en œuvre selon les

			scénarios - la communication auprès des différentes parties prenantes - les résultats attendus L'application de la méthodologie permet de résoudre l'incident technique.

BLOC 5 : CONCEVOIR ET DEPLOYER DES MODELES D'APPRENTISSAGE AUTOMATIQUE (Spécialité DATA SCIENTIST)

REFERENTIEL D'ACTIVITES *décrit les situations de travail et les activités exercées, les métiers ou emplois visés*	**REFERENTIEL DE COMPETENCES** *identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel d'activités*	**REFERENTIEL D'EVALUATION** *définit les critères et les modalités d'évaluation des acquis*	
		MODALITÉS D'ÉVALUATION	**CRITÈRES D'ÉVALUATION**
		Type d'évaluation : Mise en situation professionnelle réelle ou fictive. **Attendus du candidat :** À partir de l'analyse d'un projet réel ou fictif, le candidat présente la stratégie de conception et de déploiement de modèles d'apprentissage automatique ~~de machine learning.~~ **Livrable attendu :** Le candidat présente une soutenance comprenant :	
A5.1 : Analyse du besoin et résolution de problèmes complexes - Analyse du besoin et du contexte	C5.1.1 : Analyser la problématique et le contexte d'un commanditaire en réalisant des entretiens exploratoires, des questionnaires et une analyse de l'existant afin de lui apporter une réponse appropriée.	- Une analyse du besoin	L'analyse du besoin exprimé par un commanditaire comporte une description : - De la problématique, - De l'environnement et du contexte - Des contraintes (ex : coût, délai, complexité) - Des points de vigilance identifiés le cas échéant L'analyse du besoin permet de conclure à la faisabilité du projet.
- Résolution d'un problème complexe à l'aide de modèles d'apprentissage automatique	C5.1.2 : Cadrer la stratégie de résolution du problème, en utilisant des algorithmes, en traduisant le problème en un problème d'optimisation afin de le résoudre avec les outils des modèles d'apprentissage automatique.	- Une présentation de la stratégie de résolution du problème	L'angle de résolution du problème est justifié et le type d'algorithme utilisé est argumenté. (ex : segmentation, recommandation, classification régression). La stratégie présentée permet : - d'identifier le type de données à acquérir - de résoudre efficacement le problème complexe

- Sélection des outils et des algorithmes (ex : les librairies, le langage de programmation, les technologies, l'infrastructure)	C5.1.3 : Sélectionner les technologies, les outils et les algorithmes en identifiant les différentes solutions disponibles et en comparant leurs avantages et leurs inconvénients afin de répondre à la problématique du commanditaire au regard des contraintes du projet.	- Une présentation des technologies et des outils sélectionnés	Le choix des technologies et outils est justifié et permet d'identifier les avantages en termes de : - compatibilité, - coût, - simplicité d'utilisation, - performance, - maintenabilité. La présentation des technologies et des outils sélectionnés permet d'implémenter la solution.
A5.2 : Développement de modèles d'apprentissage automatique - Construction des variables	C5.2.1 : Construire des variables en utilisant des langages de programmation (ex : Python, Scala, R, Julia…) en exploitant des bibliothèques d'analyse de données afin de fournir les meilleures variables au modèle d'apprentissage automatique.	- Un jeu de données exploitables pour le modèle d'apprentissage automatique	Le jeu de données comporte : - La ou les variables dépendantes - Les variables prédictives envisagées pour la modélisation Le jeu de données permet d'entrainer le modèle d'apprentissage automatique.
- Sélection des variables	C5.2.2 : Sélectionner les variables en identifiant les différentes méthodes de sélection de variables possibles, en utilisant des méthodologies d'apprentissage automatique afin d'optimiser la performance du modèle.	- Des méthodes de sélection de variables	Le choix des méthodes de sélection de variables utilisées est argumenté (ex : test statistique, réduction de dimension, élimination récursive, méthode incorporée) Les méthodes de sélection de variables démontrent la pertinence de la liste de variables choisies.
- Construction et entrainement des modèles d'apprentissage automatique (ex : Machine learning, Deep Learning)	C5.2.3 : Entrainer un modèle d'apprentissage automatique à l'aide de librairies (ex : Scikit-learn XGBoost, TensorFlow, PyTorch) afin d'obtenir des modèles capables de prédictions sur de nouvelles données inconnues.	- Un entrainement d'un modèle d'apprentissage automatique	La présentation de l'entrainement d'un modèle d'apprentissage automatique précise les librairies utilisées. L'entrainement présenté permet d'identifier : - le bon déroulement de l'entrainement - la performance du modèle entrainé L'entrainement fonctionne et produit un modèle d'apprentissage automatique capable d'inférences.

- Optimisation de la performance et des hyperparamètres des modèles d'apprentissage automatique - Comparaison et évaluation de la performance des modèles d'apprentissage automatique	C5.2.4 : Optimiser la performance des modèles d'apprentissage automatique en modifiant les hyperparamètres et en analysant les prédictions afin de répondre au mieux à la problématique du commanditaire.	- Une méthode d'optimisation des modèles d'apprentissage automatique.	La méthode d'optimisation présentée comporte : - les leviers d'optimisation (ex : DATA, hyperparamètres, infrastructure) - les moyens utilisés (ex : librairies, hardware) - la comparaison des performances des modèles d'apprentissage automatique La méthode d'optimisation présentée permet d'obtenir un gain de performance.
A5.3 : Déploiement et automatisation des modèles d'apprentissage automatique - Sauvegarde des modèles d'apprentissage automatique (ex : versioning, package)	C5.3.1 : Sauvegarder le modèle d'apprentissage automatique entrainé à l'aide d'outils de sérialisation, virtualisation, containerisation, versioning afin de pouvoir le déployer dans des environnements de production.	- Une méthode de sauvegarde	La méthode de sauvegarde permet d'identifier : - les librairies utilisées - les modalités de sauvegarde - les modalités de réutilisation du modèle La méthode permet de sauvegarder le modèle dans un format qui permet sa réutilisation et son déploiement.
- Déploiement des modèles d'apprentissage automatique - Développement d'un processus d'intégration et de déploiement continu.	C5.3.2 : Déployer des modèles d'apprentissage automatique en utilisant des API et des outils CI/CD[10] afin de le mettre en production.	- Un processus CI/CD	Le processus CI/CD permet d'automatiser : - l'enregistrement du modèle dans un registre - les tests du modèle - l'évaluation des performances - le déploiement du modèle
- Mise en place d'outils de supervision et de détection de DATA Drift.	C5.3.3 : Superviser le système Machine Learning en sélectionnant des outils de monitoring et en les exploitant afin de détecter les dérives et les bugs du modèle d'apprentissage automatique.	- Un système de monitoring de la performance	Le système de monitoring permet de suivre la performance du modèle au cours du temps. Il permet de remonter les alertes.
- Automatisation des tâches et des processus d'apprentissages automatiques	C5.3.4 : Automatiser les tâches inhérentes au cycle de vie d'un système d'apprentissage automatique à l'aide de pipelines et des outils adaptés afin de maintenir la performance du modèle d'apprentissage automatique.	- Un système de collecte de données	Le système de collecte des données présenté permet : - d'historiser les prévisions - de collecter de nouvelles données d'apprentissage - de mettre à jour des modèles Machine

[10] CI /CD = Continuous Integration / Continuous Deployment

			Learning.