

REFERENTIEL D'ACTIVITES décrit les situations de travail et les	REFERENTIEL DE COMPETENCES identifie les compétences et les connaissances, y compris	définit les crite	REFERENTIEL D'EVALUATION ères et les modalités d'évaluation des acquis
activités exercées, les métiers ou emplois visés	transversales, qui découlent du référentiel d'activités	MODALITÉS D'ÉVALUATION	CRITÈRES D'ÉVALUATION
	BC1. Concevoir et dé	velopper une architecture de stock	age de données
	has a large		
A1 : Conception d'architecture de de stockage de données	BC1.1. Concevoir et développer une base de données relationnelle en réponse aux besoins d'un client en	• •	Une base de données relationnelle est mise en place : C1 : Le schéma de données relationnelle répond au besoin exprimé par le client
A2 : Développement d'une architecture de stockage de données.	vue de la mise à disposition de ses données structurées pour un tout en utilisant les technologies et les langages de requêtes adaptés	jury d'évaluation.	C2 : Un model relationnel traduit un schéma de données utilisant le modèle « Entités- Association » C3 : Les systèmes de gestion de bases de données mobilisent les technologies appropriées
A3 : Construction de lac de données A4 : Traitement et Analyse de données	aux développements envisagés	Création d'une base de données Evaluation : Présentation orale et démonstration par binôme devant ur jury d'évaluation	C4 : Le langage SQL permet de manipuler les données pour un traitement analytique : création, insertion, consultation, mise à jour et suppression C5 : Les contraintes d'intégrité, les déclencheurs et les transactions sont implémentés
	BC1.2. Concevoir et développer une base de données non-relationnelle en vue de la mise à disposition des données semis-structurées et nonstructurées pour un traitement analytique ou d'intelligence artificielle tout en utilisant les technologies et les langages de requêtes adaptés	• •	Une base de données non-relationnelle est mise en place : C1 : Le langage NoSQL utilisé permet de manipuler les données semi-structurées pour leur traitement analytique C2 : Les systèmes de gestion de bases de données non relationnelles qui sont mobilisés permettent un traitement analytique et des requêtes adaptés C3 : Une base de données clé-valeurs qui répond à un besoin spécifique est mise en place C4 : Une base de données orientée documents est développée C5 : Une base de données orientée colonnes est développée



	Evaluation : Présentation orale et démonstration par binôme devant un jury d'évaluation	C6 : Une base de données orientée graphes est développée
BC1.3. Concevoir et construire un lac de données « Datalake » en choisissant les architectures, les indicateurs de performance et les solutions de stockage appropriées afin d'intégrer les données provenant des systèmes d'information multivariés : Données non structurées, semistructurées, structurées et des solutions de stockage de données massives ; redondantes.		Un « datalake » est mis en place C1 : L'architecture technique permet d'intégrer des données non structurées (par exemple des fichiers de tout type) - Des données semi-structurées (par exemple des fichiers JSON) - Des données structurées C2 : La solution proposée est distribuée et permet le stockage de données massives Elle fournit un stockage de données redondant C3 : Les indicateurs de performance permettent de monitorer l'état de la solution
BC1.4. Créer une API en utilisant les technologies qui permettent de rendre accessible les données en augmentant l'efficacité et la praticité des applications et des services	Challenge (hackathon) : construction d'un data lake avec un API RESTful permettant l'accès aux données E1.8. Création d'une base de données Evaluation : Présentation orale et	Un API WEB RESTful est fonctionnel C1: Les technologies utilisées pour créer l'API sont appropriées C2: L'architecture déployée rends accessible les données C3: Le niveau de complexité de l'API est adapté aux données à traiter C4: Les mécanismes d'authentification sont déployés C5: Les autorisations et les quotas imposés aux utilisateurs de l'API sont faciles à comprendre et à intégrer C6: L'accès à l'API est rapide C7: Le processus d'intégration via l'API est automatisé C8: La sémantique de l'API (la syntaxe des fonctions fournies) est claire, compréhensible et documentée C9: Un contrôle de version de l'API est mise en place



	BC2 : Concevoir, développer et	t déployer une solution de traiteme	nt des données massives
	BC2.1. Concevoir en s'appuyant sur		Une architecture de traitement distribué de données est proposée
•	une veille technologique et mettre	Conception d'une application	C1 : Une veille technologique des Framework big data est réalisée
	en œuvre une architecture	de traitement distribué.	C2 : L'architecture proposée répond au besoin exprimé par le
	distribuée répondant au besoin du	E2.2. Évaluation : Présentation de	client en termes de traitement C3 : Les ressources mobilisées sur le cluster de calcul en termes
A2 : Développement de solution données	client pour traiter les données massives en entreprise en utilisant	l'architecture par binôme devant un	de puissance de calcul et de mémoire sont suffisantes
dominees	les technologies de traitement	jury et les autres apprenants.	C4 : Les environnements logiciels de traitement de données sont
A3 : Déploiement d'une solution de	les technologies de traitement	July et les autres apprenants.	adaptés
traitement des données massives			C5 : La solution proposée permet de traiter (en batch) et
			d'analyser l'ensemble de données disponibles
A4 : Transformation de données			C6 : L'architecture proposée est résiliente à la panne du système
issues de sources différentes			C7 : La solution assure une scalabilité horizontale
	BC2.2. Implémenter un système	E2.3. Étude de cas réalisée en amont	Une solution de streaming distribué est mise place :
	distribué en utilisant des		C1: Une veille technologique permet d'identifier et mobiliser les
	technologies de streaming	Conception d'une application de traitement distribué.	solutions de streaming adaptées
	identifiées à partir d'une veille <i>pour</i> traiter des données sur une période	traitement distribue.	C2: La solution proposée permet de traiter (en micro batch) et d'analyser l'ensemble de données collectées sur une période
	précise ou en temps quasi réel	E2.4. Evaluation : Soutenance Orale	données (en secondes ou millisecondes)
A7 : Création et automatisation de	precise ou en temps quasi reer	et démonstration de la solution par	C3 : La solution proposée permet de traiter (en temps réel) et
tests		binôme devant un jury et les autres	d'analyser l'ensemble de données au fur et à mesure de leur
		apprenants. Rapport écrit en binôme	·
		examiné par un jury.	



provenant de différentes sources en prenant en compte la variété de données pour faire de l'analytique à échelle (intégration, formatage, manipulation, stockage données multidimensionnelles)	Conception d'une solution d'intégration et de transformation de données. Évaluation : E2.6. Evaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme	Les données sont extraites, transformées et mise à disposition : C1 : L'utilisation des outils d'informatique décisionnelle permet de transformer et formater les données selon un format précis C2 : Les outils utilisés permettent l'intégration de données C3 : Les CUBES permettent la représentation et la manipulation des données multidimensionnelles
BC2.4. Optimiser la performance des pipelines en utilisant les techniques d'intégration et de mise en scène adéquates pour le traitement des données massives	E2.8. Evaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme examiné par un jury.	Les pipelines sont développés C1: Les pipelines développés permettent d'intégrer des sources de données provenant de différentes sources C2: - un accès rapide aux données - L'ajout de nouvelles sources de données C3: Les indicateurs mis en place permettent de mesurer la performance de transfert de données C4: L'ordonnanceur mobilisé est adapté aux technologies utilisées C5: La performance du pipeline de données est monitorée C6: Les pipelines permettent l'ajout de nouvelles sources de données



	BC2.5. Automatiser la création, les tests, l'intégration et le déploiement des pipelines de données en s'appuyant sur une veille technologique qui permet d'identifier et de mobiliser les solutions pour maximiser l'efficience et réduire le 'time to market' tout en utilisant les technologies de containerisation et d'ordonnancement.	Automatiser la création, les tests, l'intégration et le déploiement des pipelines de données E2.10. Evaluation : Soutenance Orale et démonstration de la solution par	La création, les tests, l'intégration et le déploiement des pipelines de données sont automatisées : C1: La veille technologique permet d'identifier et mobiliser les technologies de contrôle de version, containerisation, testing, intégration et déploiement continue adaptées C2: Les pipelines développés sont containerisés C3: Les outils d'orchestration permettent l'intégration et le déploiement continu et rapide du produit C4: Les différents types de tests mobilisés permettent de vérifier et valider la solution déployée
	 BC3. Implémenter et optimiser de	 s solutions de stockage et de traiten	nent de données sur le cloud
	2 cop.cc. c. cpc. ac		
A1 : Conception de solutions de traitement de données	BC3.1. Mettre en œuvre des solutions de stockage de données dans le cloud pour permettre aux entreprises d'explorer leurs	E3.1. Certification externe par un fournisseur du cloud public (e.g. Azure, Aws, Google)	Une solution de stockage des données est mise en place. C1: Une solution data Lake est implémentée: Plusieurs types de fichiers sont choisis pour répondre aux besoins de stockage et d'archivage
A2 : Mise en œuvre de solutions de stockage de données dans le cloud	·	E3.2. Étude de cas réalisé en amont: Conception et déploiement d'une solution de stockage et de	C2 : Une stratégie de réplication de données est mise en place C3 : Les métadonnées sont maintenues
A3 : développement de solution de traitement de données	types de stockage et d'archivages appropriés	traitement sur le cloud. E3.3. Evaluation : Soutenance	C4 : Les données non relationnelles sont mises à disposition et accessibles C5 : Un schéma de données relationnel est mis en place
A4 : Exploitation de données		Orale et démonstration de la solution par binôme devant un jury	23 . 2 23
A5 : Mise à disposition de données à des fins d'utilisation		et les autres apprenants. Rapport écrit en binôme examiné par un jury.	



A6 : Sécurisation de données			
Ao : Securisation de données			
A7 : Stratégie et Politique de sécurisation de données A8 : conformité règlementaire A9 : protection des données et des accès utilisateurs A10 : optimisation des coûts		fournisseur du cloud public (e.g.	Une solution de traitement des données dans le cloud est conçue et développée C1: Les services de gestion et de transformation des données dans le cloud sont bien mobilisés C2: Des pipelines de traitement des données sont développées et permettent l'extraction et la transformation de la donnée C3: Des outils de contrôle de version des pipelines sont implémentés C4: Un service de traitement de données en temps réel est bien identifié et implémentée
	li ·	E3.7. Étude de cas : Mise en place d'une politique de sécurité de données dans le cloud. E3.8. Evaluation : Rapport écrit par binôme évalué par un jury.	Une solution de sécurité des données dans le cloud est mise en place. C1: Un chiffrement de données stockées est mis en place Les données en cours de traitement sont chiffrés afin de garder leur sécurité C2: La politique de conservation de données est conforme aux règlements en vigueur



	BC3.4. Optimiser les solutions de stockage et de traitement des données dans le cloud en définissant des indicateurs de performance pour assurer la disponibilité des services et optimiser les coûts	E3.9. Certification externe par un fournisseur du cloud public (e.g. Azure, Aws, Google) E3.10. Étude de cas : Conception et déploiement d'une solution de stockage et de traitement sur le cloud. Evaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme examiné par un jury.	C3: Le temps de réponse est testé et monitoré en permanence C4: La performance des requêtes est mesurée en permanence C5: Un service de logs est implémenté C6: Des indicateurs de performance des clusters sont définies et implémentés C7: Les services d'optimisation de coûts sont bien mis en place
--	---	---	--

BC4. Implémenter des méthodes d'intelligence artificielle pour modéliser et prédire de nouveaux comportements et usages.



 A1: Extraction de données depuis différentes sources A2: analyse exploitation de données avec outils appropriés A3 transformation de données selon les besoins définis A4: construction de tableaux de bord, reporting A5: communication de données et d'indicateurs en temps réel 	pour faire l'analyse et le reporting selon les besoins des différents métiers	aux apprenants de manipuler les types d'outils suivants : - Outils d'extraction, de transformation et de stockage de données - Outils logiciels de traitement de données massives E4.2. Évaluation : Soutenance Orale par binôme devant un jury et les autres apprenants. Rapport de TP écrit en binôme examiné par un jury.	C1: Les environnements logiciels de traitement de données sont adaptés aux besoins prédéfinis tC2: Les logiciels d'extraction et transformation de données sont mobilisés dans leurs fonctions de base C3: Les plateformes de data management sont mises en œuvre à partir du cloud C4: Les fonctions avancées de nettoyage de données sont exploitées et expliquées C5: Les données transformées sont prêtes à l'analyse et permettent un apprentissage automatique
	infographique visuelle en construisant des tableaux de bord	Mise en place d'une application interactive avec de tableaux de bord <u>Évaluation</u> : Présentation du tableau de bord	Une communication infographique visuelle est bien élaborée : C1 : Les environnements logiciels d'analyse de données massives sont déployés C2 : Les visualisations sont informatives et font parler les données C3 : Les graphiques sont chargés affichés en temps réel C4 : Les fonctions des logiciels de visualisation de données sont exploitées pleinement C5 : Le tableau de bord est responsif et interactif Un modèle prédictif est développé et déployé :
	prédictif pour identifier de	LT.T. WINGE CIT SICUATION .	C1: Les données sont bien chargées est prétraitées.



nouveaux comportements et usages Proposé via une plateforme de C2 : Les algorithmes sont testés avec des outils et méthodes en implémentant des algorithmes compétition data science de connus d'apprentissage automatiques type kaggle, consiste à : - Identifier C3 : Les codes implémentés en utilisant un langage supervisés le type de problème et définir une ladapté fonctionnent sans erreur C4: Le processus d'apprentissage automatique est mis en œuvre stratégie de modélisation de avec les techniques appropriées données. -Implémenter les modèles de C5: Les algorithmes et leur fonctionnement sont clairement machine Learning et faire de la expliqués sélection de variables (feature sélection). -Comparer et analyser les performances des modèles et conclure les règles de décision pour les futures observations. F4.5. Évaluation : Soumission des résultats de prédiction sur un ensemble de données test et obtenir un score du modèle proposé. Soumission d'un rapport expliquant la stratégie suivie et le choix du modèle. Le rapport sera examiné par un jury d'évaluation comprenant un représentant opérationnel professionnel et un enseignant spécialisé



BC4.5. Évaluer la performance d'un modèle de Machine Learning en analysant ses résultats et en comparant avec d'autres modèles afin d'implémenter la solution la plus convenable à un cas d'usage

E4.6. Mise en situation:

Proposé via une plateforme de compétition data science de type kaggle, consiste à : - Identifier le type de problème et définir une stratégie de modélisation de données.

-Implémenter les modèles de machine Learning et faire de la sélection de variables (feature sélection).
-Comparer et analyser les performances des modèles et conclure les règles de décision pour les futures observations.

F4.7. Évaluation :

Soumission des résultats de

prédiction sur un ensemble de données test et obtenir un score du modèle proposé.
Soumission d'un rapport expliquant la stratégie suivie et le choix du modèle. Le rapport sera examiné par un jury d'évaluation comprenant un représentant opérationnel professionnel et un enseignant spécialisé

Une comparaison de plusieurs modèles a été effectuée :

- C1 : Plusieurs modèles de Machine Learning sont développés, évalués, préparés et mise en œuvre
- C2 : Des tentatives d'amélioration des prédictions sur le test sont soumises sur la plateforme dédiée à la mise en situation et expliquées.
- C3 : Des méthodes pour éviter le sur-apprentissage (overfitting) sont déployées et expliquées

BC5 : Concevoir une stratégie de management et de gouvernance de données pour transformer les données en informations créatrices de valeur



A1 : Définition de stratégie de	BC5.1. Définir une gouvernance de	E.5.1. Étude de cas :	Une gouvernance de données est définie:
gestion de données	données en mettant en place les		C1 : Les rôles sont bien définis et présentés
	politiques et les standards afin		C2 : Les responsabilités sont bien déterminées
A2 application de politique et de	d'établir les rôles, les responsabilités	·	C3 : Les catégories des données sont identifiées
standards appropriés pour la	et la propriété des données	, ,	C4 : Les données « maîtres » sont identifiées
gestion des données		examiné par un jury.	C5 : Un dictionnaire de données est disponible et accessible
A3 : définition de responsabilité et		examine par arrigary.	
de rôles dans la gestion des			
données			
A4 : veille et amélioration des			
pratiques en gestion de données	BC5.2. Mettre en place les bonnes	F 5 2 Étuda da cas : mise en nlace de	Les règles et les bonnes pratiques pour respecter les régulations en
par rapport à la règlementation en	· · · · · · · · · · · · · · · · · · ·		
vigueur	appliquant les régulations en		C1 : L'objectif et la finalité du traitement des données est précisé
	vigueur pour protéger les données	d'un cas d'usage	avec clarté
A5 : vérification de la qualité et la	et respecter la vie privée dans une		C2 : Un recensement de données en entreprise est effectué, tracé
fiabilité des données	logique de transparence.	E.5.3. Évaluation : Soutenance Orale et démonstration de la solution par	et enregistré dans un registre
A6 : valorisation de l'utilisation de la		·	La stratégie d'information et de transparence à l'égard des
donnée		_ ·	personnes dont les données sont traitées est correctement
		examiné par un jury.	mobilisée
A6 élaboration d'une stratégie et			On sait :
une sécurisation des données, politique d'accès			 Qui a accès aux données Combien de temps les données sont conservées
politique d'acces			- Pourquoi les données sont collectées
			- Si les données sont transférées hors de l'UE
			C1 : Les personnes dont leurs données sont traitées disposent les
			droits d'accès, de rectification, d'opposition, d'effacement.



donnée permet d'atteindre les objectifs business en faisant un audit de qualité et en proposant des plans de prévention et de	E5.5. Évaluation : Soutenance Orale	Des bonnes pratiques pour assurer la qualité des données sont mises en place C1: Les outils de gestion de qualité de données sont mobilisés C2: Les indicateurs de qualité de données sont identifiés et implémentés C3: Les indicateurs de performance liés à l'ensemble du dispositif de sécurité des données sont élaborés dans la démarche d'audit ainsi que les critères de services liés à l'utilisation de ces données. C4: La data linéage permet de renseigner sur les différents usages appliqués à une même donnée C5: Des processus de prévention de la non-qualité en amont sont mises en place C5: Des processus de remédiation implémentés permettent de corriger les anomalies C6: Les processus de remédiation implémentés permettent de corriger les anomalies
BC5.4. Identifier l'approche stratégique pour faire de la donnée	E5.6. Étude de cas :	Une approche stratégique de la gestion de données est définie



l'approche dans un plan de	et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme	C1 : Le plan de communication permet une prise de conscience de l'enjeu et d'importance les collaborateurs à la data C2 : Le modèle de maturité permet de mesurer le niveau d'utilisation des données
sécurité afin de protéger les données en utilisant les technologies adéquates permettant de parer aux menaces et en mettant en place une politique d'accès aux	du passage devant le jury E.5.9. Évaluation : Soutenance Orale et démonstration de la solution par binôme devant un jury et les autres apprenants. Rapport écrit en binôme examiné par un jury.	Une politique de sécurité des données est mise en place C1: Les technologies de chiffrement de données mobilisés permettent de protéger les données C2: Les menaces internes et externes sont identifiées et documentées C3: Les règles d'identification et d'authentification sont bien définies C4: La gestion des identités et des accès n'autorise que les personnes éligibles à accéder à la donnée selon leur rôle et responsabilité C5: Les plans de réponse aux incidents permettent de réagir en cas d'attaque C6: La veille technique permet d'implémenter les recommandations de l'agence national de sécurité d'information