

Architecture d'entreprise pour 'IA agentique

L'approche de Microsoft pour faire évoluer l'adoption des agents
tout en gardant le contrôle de la sécurité et de la gouvernance



4

Introduction

5

Définition du
centre principal

7

Mise en place d'une
gouvernance des
environnements
d'exécution multiples

7

Plateforme de
données intelligente

8

Capacités fondamentales
pour les produits de
données

9

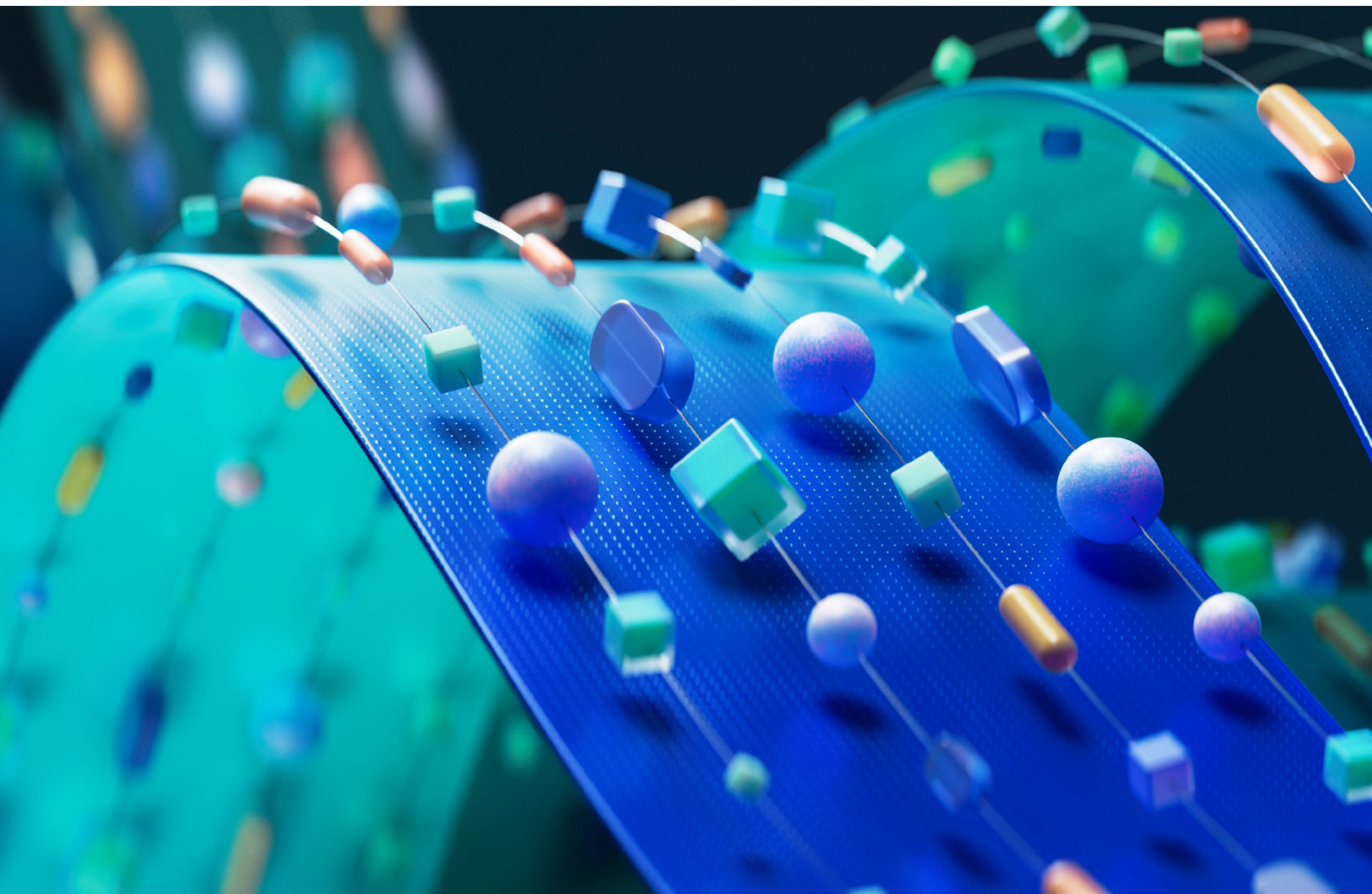
Plan d'application

11

Schémas des zones
de destination et
responsabilités

12

Atelier de Microsoft
et mise en œuvre
progressive du maillage
d'agents



Résumé

Faire évoluer les agents à travers l'organisation en tenant compte de la fiabilité, de la sécurité et de l'efficacité est très difficile. La solution recommandée consiste à mettre en place un maillage d'agents d'entreprise afin de mettre en œuvre un concept de plan d'application pour une plateforme d'agents d'entreprise sensible aux données d'affaires, en utilisant une architecture à trois plans en étoile : (1) un centre principal qui fournit le chemin gouverné unique vers les modèles et outils (identité, application des politiques, inventaire/registre, vérification et télémétrie), (2) une plateforme de données intelligente accessible via des protocoles standards (par exemple, MCP), livrant des produits de données augmentés, c'est-à-dire des actifs de données augmentés par les modèles opérationnels et analytiques qui définissent leur sémantique, et (3) un plan d'application composé de zones de destination standardisées par charge de travail qui hébergent les environnements d'exécution des agents et les intégrations tout en consommant des terminaux médiés par le centre. Dans un premier temps, adoptez cette architecture en déployant les passerelles et l'inventaire du centre, ainsi qu'un plan d'intégration contraignant. Ensuite, intégrez un petit ensemble de serveurs MCP de domaine et quelques charges de travail d'agents représentatives. Enfin, évoluez à l'aide de modèles, de l'automatisation du cycle de vie et des politiques sous forme de code. La solution recommandée crée et gère une place de marché d'agents pour l'entreprise pionnière, en utilisant Microsoft Fabric, Microsoft Foundry et Copilot Studio comme principales plateformes d'agents, tout en prévoyant la possibilité d'étendre à d'autres plateformes.



Introduction

L'IA agentique n'est plus un sujet de recherche. Elle devient une couche opérationnelle de l'entreprise. Dans tous les secteurs, les organisations passent des fonctionnalités d'IA à usage unique à des populations d'agents, des entités logicielles autonomes qui raisonnent, récupèrent, agissent et coordonnent au nom des utilisateurs et des processus d'affaires. Ce changement est comparable en ampleur à l'arrivée de l'Internet public : la valeur ne proviendra pas d'un modèle, d'un cadre ou d'un fournisseur en particulier, mais du tissu conjonctif qui permet aux agents de se découvrir, de partager le contexte, d'invoquer les ressources de l'entreprise en toute sécurité et d'opérer sous une gouvernance vérifiable.

Ce tissu conjonctif se regroupe maintenant en un ensemble émergent de normes ouvertes. Model Context Protocol (MCP) devient le moyen universel pour les agents d'invoquer des outils et d'accéder aux connaissances. Le protocole Agent-à-Agent (A2A) établit comment les agents se découvrent et délèguent le travail au-delà des frontières organisationnelles. Les normes d'identité d'agent basées sur DNS, l'infrastructure à clé publique, OIDC et JWT donnent à chaque agent un principal vérifiable. Les conventions sémantiques d'IA générative d'OpenTelemetry fournissent une infrastructure d'observabilité commune permettant de retracer chaque étape de raisonnement, chaque appel d'outil et chaque décision.

Les entreprises passent de preuves de concept IA isolées à des portefeuilles d'agents opérant à travers des unités d'affaires, des régions géographiques et des piles technologiques. Le principal défi est de faire évoluer de nombreux agents de façon sécuritaire et prévisible. Un maillage d'agents répond à ce défi en fournissant un modèle à l'échelle de l'entreprise pour la façon dont les agents découvrent et invoquent des modèles, des outils, des données d'affaires et d'autres agents, permettant aux équipes de maintenir une exécution décentralisée (en choisissant l'environnement d'exécution qui convient à chaque charge de travail) tout en utilisant un modèle d'interaction cohérent et gouverné.

En pratique, le maillage centralise les contrôles transversaux qui sont autrement reconstruits (et implémentés de façon incohérente) dans chaque projet : elle établit des passerelles et des registres gouvernés comme voie prise en charge vers les modèles et les capacités exposées au MCP, standardise la propagation de l'identité et la politique d'autorisation, et exige des traces de vérification partagées ainsi que de la télémétrie opérationnelle afin que les actions puissent être tracées de bout en bout à travers les environnements d'exécution. En rendant ces contrôles fondés sur des gabarits et réutilisables, l'entreprise réduit le travail redondant sur la plateforme, évite le contournement de la gouvernance via

une connectivité directe aux dorsaux, améliore l'observabilité et la réponse aux incidents, et peut répondre à grande échelle aux questions de contrôle de base : qui a accédé à quoi, quels outils ont été exécutés, sous quelle identité, et à quel coût.

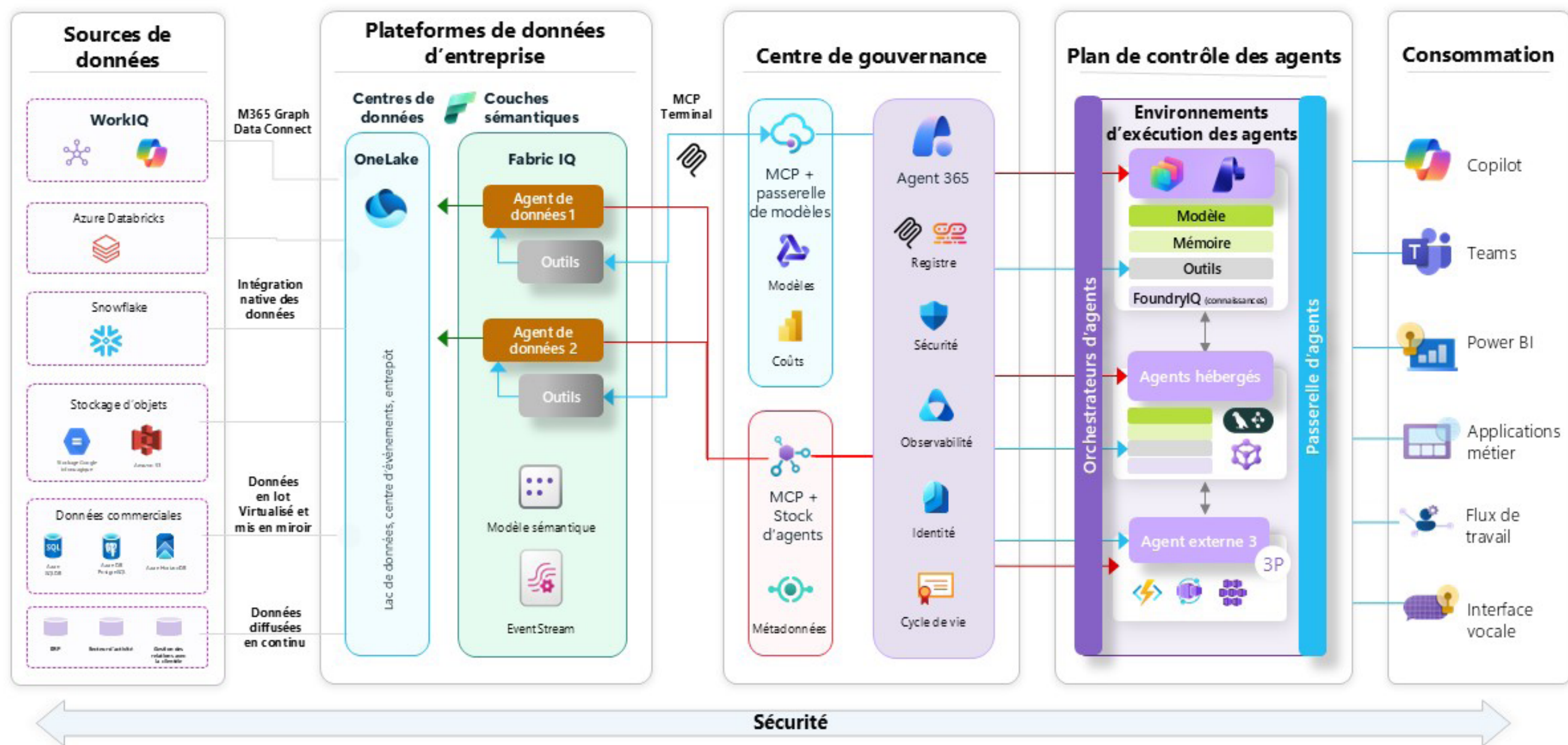
Ce document définit le plan d'application d'un maillage d'agents alimenté par des données d'affaires contextualisées.

Pour que les agents puissent récupérer le contexte de données d'affaires approprié, deux conditions doivent être remplies. Premièrement, les données doivent être dignes de confiance. Tout au long de ce document, nous adoptons donc une architecture de maillage de données, car elle fournit intrinsèquement les capacités de sécurité et de gouvernance nécessaires pour rendre les produits de données fiables et réduire le risque de fuite de données causée par l'IA. Deuxièmement, les données doivent être comprises. À cette fin, les produits de données sont enrichis de représentations de connaissances d'affaires (telles que des ontologies) et soutenus par des moteurs de graphes de données qui capturent et exploitent les relations entre des ensembles de données opérationnelles hétérogènes, une approche de plus en plus réalisée grâce aux graphes de contexte. En conséquence, dans le présent document, les produits de données désignent la combinaison des actifs de données opérationnels et analytiques, des politiques de gouvernance et de sécurité, ainsi que des abstractions et moteurs permettant aux agents de récupérer des données d'affaires contextualisées.

Dans ce modèle, l'objectif est d'atteindre l'autonomie. Les agents développés par les équipes d'application peuvent se connecter de manière sécurisée et autonome aux produits de données, aux outils et à d'autres agents par l'intermédiaire d'une gouvernance centrale. Les équipes de domaine publient des produits de données gouvernés. Les équipes d'application consomment ces produits et les composent en solutions à base d'agents, livrées par le biais d'interfaces utilisateur, de flux de travail et d'intégrations système. Le plan d'application définit les zones de destination en environnement d'exécution, les modèles d'intégration et les contrôles opérationnels nécessaires au déploiement et à l'exploitation de ces solutions en production.

Le document recommande une architecture à trois plans organisée selon une topologie en étoile. Un centre principal fournit des services de gouvernance et de médiation qui standardisent la manière dont les modèles, les outils et les agents sont exposés et consommés. Un plan de données permet aux équipes de domaine de publier des produits de données qui ultimement permettent aux agents de raisonner sur les données de l'entreprise au sein d'une compréhension d'affaires partagée. Un plan d'application fournit des zones de destination spécifiques à chaque charge de travail pour héberger des agents sur différents environnements d'exécution, orchestrer des solutions à agents multiples et s'intégrer aux canaux et aux systèmes d'entreprise.

Configuration agentique – architecture logique



Hypothèses pour la portée de ce document : la connectivité au réseau privé constitue l'objectif pour les modèles, les passerelles, les agents et les serveurs MCP; l'identité et l'autorisation reposent sur Microsoft Entra ID; l'entreprise exploite un réseau privé sur lequel les services de plateforme partagés sont exécutés dans l'abonnement principal et les environnements d'exécution des agents spécifiques aux charges de travail sont exécutés dans les abonnements en étoile; les produits de données sont définis comme indiqué ci-dessus et sont fournis dans Microsoft Fabric.

Recommandation principale : Recommandation clé: concevez et déployez le plan d'application comme une fonctionnalité de la plateforme, avec des modèles applicables en matière de sécurité, de réseau, d'observabilité, d'automatisation du cycle de vie et de contrats d'intégration. N'autorisez pas les équipes de projet à intégrer directement des terminaux de modèles ou d'outils. Exiger l'utilisation d'interfaces gérées par le centre afin que l'entreprise puisse étendre l'adoption des agents sans compromettre la conformité, la responsabilité ou la fiabilité opérationnelle.

Centre principal

La surface de contrôle de l'entreprise pour les modèles, les outils et les interactions avec les agents.

Le centre principal constitue la surface de contrôle de l'entreprise pour les modèles, les outils et les interactions avec les agents. Il définit le chemin d'accès recommandé permettant aux agents et aux applications d'accéder aux terminaux de modèle, aux serveurs MCP et aux API

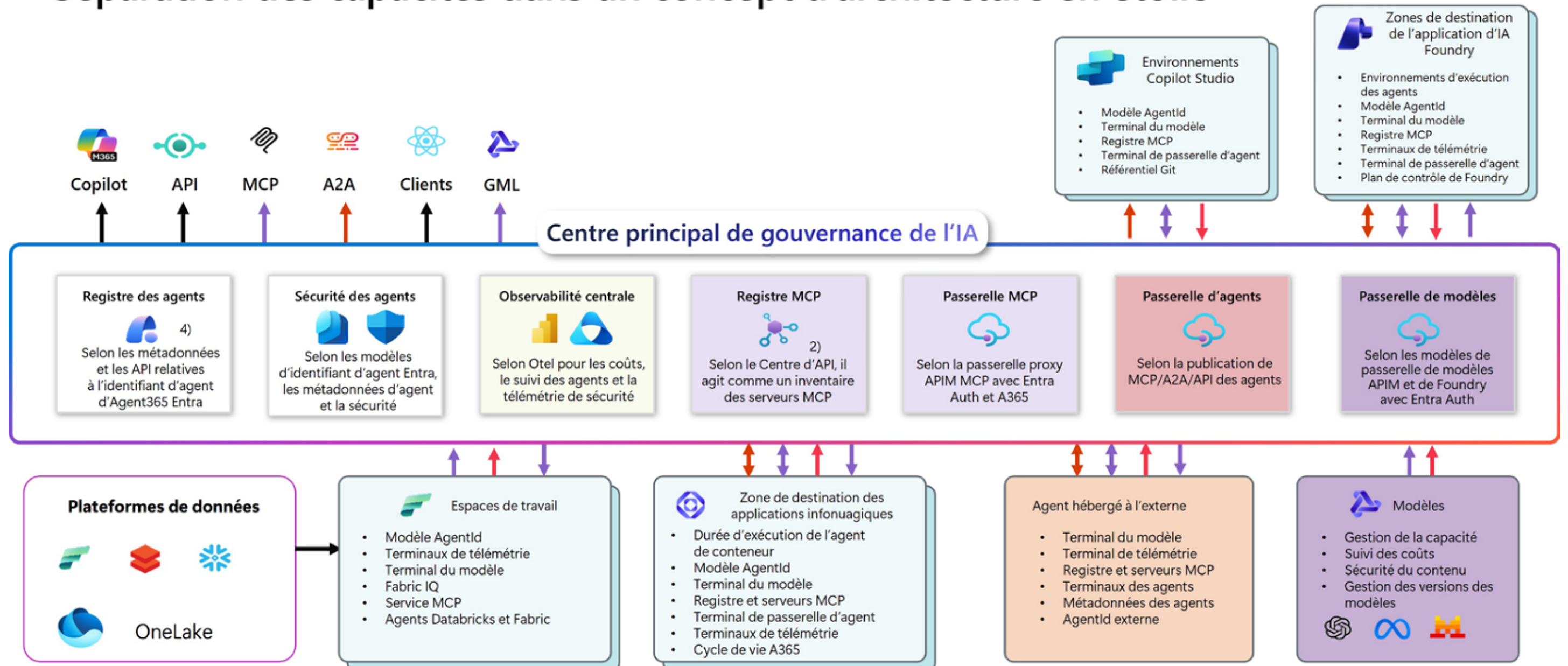
d'entreprise, et applique de manière uniforme les mécanismes d'identité, les politiques, les capacités de vérification et la télémétrie à l'ensemble des composantes de l'architecture en étoile.

Le centre est composé de deux éléments :

- Gestion des API (passerelle de modèles) comme couche de gestion des politiques et de routage
- Inventaire MCP comme système de référence pour les outils et les données

Le premier composant du centre est la passerelle de modèle : le point d'accès d'entreprise pour la consommation de modèles de langage et de services connexes d'IA sans exposer les terminaux bruts ni distribuer les informations d'identification. Mis en œuvre généralement avec Azure API Management en tant que couche de contrôle des politiques et de routage, il centralise la gestion de la capacité (routage, limites de débit, quotas et contrôles régionaux), permet une attribution transparente des coûts par charge de travail consommatrice et fournit un contrat stable qui dissocie les applications des fournisseurs et déploiements sous-jacents. En prenant en charge la gestion des versions des modèles et les responsabilités du cycle de vie, promotion, dépréciation et retrait, au sein de la passerelle, l'équipe de la plateforme peut faire évoluer les déploiements de modèles en toute sécurité, tandis que les équipes d'application intègrent une seule fois et bénéficient d'un débit prévisible et d'un accès gouverné.

Séparation des capacités dans un concept d'architecture en étoile



Le deuxième composant du centre est un inventaire MCP qui sert de système de référence pour les outils et les interfaces d'interaction avec les données exposées avec le protocole MCP, jumelé à une passerelle MCP qui décèle ces serveurs MCP avec Azure API Management. Le protocole MCP standardise la découverte et l'invocation des outils, mais l'entreprise a toujours besoin de rigueur d'inventaire pour documenter la propriété, la finalité, la classification, l'état du cycle de vie, l'authentification requise et les consommateurs autorisés. Microsoft Azure API Center est un exemple de fonctionnalité d'inventaire pour stocker cette API et les métadonnées d'outils, tandis qu'Azure API Management fournit la surface d'interaction médiée permettant aux consommateurs de découvrir et d'appeler les capacités MCP via des terminaux stables et gouvernés, plutôt que de se connecter directement aux adresses des serveurs MCP dorsaux.

Pour les serveurs MCP, deux points non négociables s'appliquent : ils doivent être enregistrés dans l'inventaire et consommés avec la passerelle. L'enregistrement nécessite des métadonnées complètes (nom de la personne propriétaire et de la personne contact opérationnelle, balises de classification/conformité, état du cycle de vie/intention de gestion des versions et périmètre d'utilisation déclaré). L'intégration de passerelle nécessite une authentification

basée sur Microsoft Entra ID sans secrets partagés, une validation automatisée des politiques (contrôles de jetons et de revendications, règles de contrôle d'accès et toute contrainte requise en matière de sécurité ou de gestion des données), ainsi que des champs obligatoires d'observabilité et de vérification afin que chaque invocation d'outil puisse être tracée et attribuée de façon cohérente. Les serveurs MCP qui ne satisfont pas à ces exigences sont considérés comme hors du périmètre jusqu'à ce qu'ils puissent être corrigés.

Une propriété essentielle du centre principal est qu'il sépare ce qui doit être centralisé de ce qui peut rester fédéré. La politique d'identité, l'application des règles de base, l'inventaire et la télémétrie centrale doivent être gérés de façon centralisée, car toute incohérence dans ces domaines devient un risque systémique. Les implémentations d'outils, les produits de données et la logique des agents devraient rester fédérés, car c'est là que la connaissance du domaine et la rapidité de livraison comptent le plus. Le centre fonctionne donc comme une plateforme d'habilitation et d'application, plutôt que comme un environnement central de développement. Il ne concurrence pas les équipes de domaine; il crée un substrat cohérent sur lequel les équipes de domaine et d'application peuvent bâtir.

L'automatisation du cycle de vie, la traçabilité des vérifications et la manière dont le centre permet la gouvernance dans les environnements d'exécution multiples

Les composantes techniques du centre n'apportent de valeur à l'entreprise que lorsque la gouvernance est mise en œuvre comme une discipline du cycle de vie. Dans un maillage d'agents, la gouvernance n'est pas une révision unique; c'est une propriété continue qui doit s'appliquer à mesure que les agents évoluent, que les outils changent et que les produits de données sont versionnés. Le centre a donc besoin d'un modèle d'automatisation du cycle de vie qui définit comment les artefacts sont publiés, validés, promus entre les environnements et retirés.

Un modèle de cycle de vie pratique commence par un contrat de publication et de consommation : un artefact déclaratif conservé dans le contrôle de source qui enregistre une connexion de modèle, un terminal du serveur MCP ou un terminal d'agent dans l'inventaire de l'entreprise et la configuration de la passerelle, tout en définissant également l'interface stable à laquelle les consommateurs s'intègrent. Le contrat associe une URL de passerelle et un schéma d'authentification à des métadonnées requises telles que la propriété, la classification des données, les consommateurs visés et les contacts opérationnels, afin que les équipes d'application puissent dépendre d'un terminal et d'une définition d'interface cohérents sans avoir à connaître les adresses dorsales.

La traçabilité des vérifications et le principe de politique sous forme de code constituent les contrôles incontournables de la plateforme : l'entreprise doit être en mesure de reconstituer ce qu'un agent a fait, quels outils et produits de données il a utilisés, sous quelle identité et sous quelles décisions de politique de sécurité, avec une télémétrie pouvant être corrélée de bout en bout à travers les environnements d'exécution. Pour rendre cette traçabilité contraignante plutôt qu'aspirante, les politiques sont centralisées, versionnées et appliquées de façon cohérente au niveau de la passerelle (par exemple, via des politiques de gestion des API Azure pour la validation des jetons, les vérifications de revendications, les contraintes de routage et les limites de débit) et sur le reste de la plateforme grâce à Azure Policy sur les ressources Azure sous-jacentes. Ensemble, l'application de la passerelle et les lignes de base de politiques au niveau des ressources assurent des chemins d'accès réglementés, une sémantique cohérente d'authentification et d'autorisation, ainsi que des preuves fiables pour la vérification, la réponse aux incidents, et l'attribution des coûts et des risques.

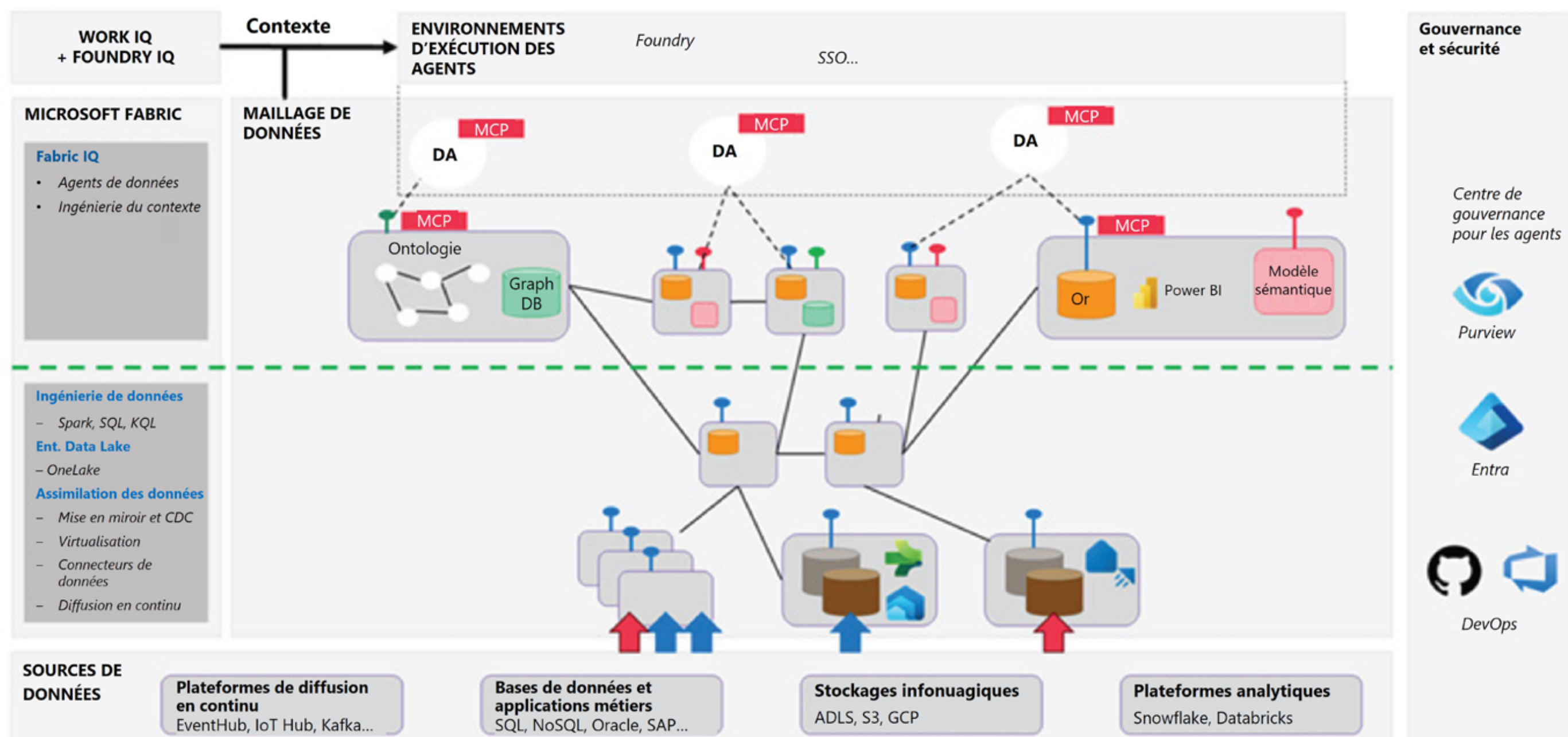
Le centre devient encore plus important lorsque l'entreprise cherche à connecter la gouvernance des agents aux structures de gouvernance organisationnelle plus larges. Un modèle axé sur le registre d'agents, où les agents sont inventoriés, gouvernés et administrés avec des contrôles du cycle de vie, s'aligne sur l'approche de Microsoft Agent 365 en tant que concept de plan de contrôle pour les agents. Même si une entreprise adopte Microsoft Agent 365 graduellement, concevoir l'inventaire du centre et la sémantique du cycle de vie de façon à pouvoir s'aligner avec les schémas de registre d'agents et de contrôle d'accès réduit les frictions futures d'intégration. Cela renforce également l'attente de l'entreprise selon laquelle un « agent non enregistré » est un agent non autorisé, et que chaque agent doit avoir un propriétaire, des autorisations et un statut de cycle de vie clairs.

Plateforme de données intelligente

Le plan de données fournit les capacités fondamentales permettant aux produits de données appartenant au domaine d'être publiés et consommés avec une sémantique prévisible. Pour que les agents puissent raisonner de façon cohérente à propos des données d'affaires, celles-ci doivent être à la fois dignes de confiance et clairement comprises. Un maillage de données fournit la base de la confiance en appliquant la gouvernance, la propriété et la sécurité au niveau des produits de données individuels. En plus de cette base fiable, les modèles de connaissances sont essentiels pour aider les agents à interpréter à la fois les données analytiques et opérationnelles, leur permettant d'assembler le bon contexte d'affaires plutôt que des faits isolés. Cela permet aux agents de raisonner de manière holistique à travers des données spécifiques au domaine provenant de plusieurs produits de données gérés indépendamment. Pour fonctionner à grande échelle, ces agents de données et modèles de connaissance devraient être exposés via des protocoles d'accès standard (p. ex., MCP) et gouvernés de façon centralisée par un centre de gouvernance, garantissant la cohérence, le contrôle et la traçabilité à travers l'organisation.

Dans ce contexte, Microsoft Fabric constitue la base d'une plateforme de données intelligente pour la prise de décision agentique. Il réunit plusieurs composantes clés : un lac de données d'entreprise, gouvernance et sécurité des données, modèles de connaissances, agents de données et interfaces d'accès.

Plateforme de données intelligente pour les agents



Fonctionnalités fondamentales pour des produits de données détenus par le domaine pouvant être consommés par des agents et des applications

OneLake de Microsoft Fabric offre un lac de données d'entreprise qui unifie les données opérationnelles à travers plusieurs sources, permettant des analyses et fournissant une base cohérente et évolutive pour les applications agentiques. Fabric offre plusieurs schémas d'ingestion natifs : la réplication miroir des bases de données permet de répliquer en continu les données provenant des bases de données sur place et dans le nuage, des raccourcis offrent un accès virtuel à des plateformes externes telles que Databricks, Snowflake et aux principaux stockages d'objets nuage sans déplacement de données, et l'ingestion par lots ou en continu est prise en charge par Fabric Pipelines et Real Time Intelligence. Toutes les données dans OneLake sont stockées dans un format commun (Apache Delta) et peuvent être consultées avec SQL ou Spark, avec un accès sécurisé appliqué par Microsoft Entra ID et des autorisations granulaires jusqu'au niveau des lignes et des colonnes.

À mesure que les agents se déploient au sein d'une organisation et agissent de façon autonome sur l'information d'affaires, la gouvernance des données devient le fondement de la confiance, garantissant que les décisions des agents

reposent sur des données exactes et autorisées. Cela s'aligne naturellement sur les principes du maillage de données, où les données sont organisées en produits appartenant au domaine, avec une propriété claire, des garanties de qualité, une gouvernance explicite et observable ainsi que des politiques d'accès. Microsoft Fabric permet ce modèle en prenant en charge des produits de données gouvernés au sein d'une plateforme unifiée, tandis que Microsoft Purview offre des capacités complètes de gouvernance tout au long du cycle de vie des données. Ensemble, Fabric et Purview offrent la découverte et le catalogage des données, la classification automatisée et l'étiquetage de sensibilité, le suivi de la lignée de bout en bout, des contrôles d'accès fondés sur des politiques, ainsi que la vérification de l'accès et de l'utilisation des données. Ces capacités garantissent que les agents consomment uniquement des données autorisées, que les entrées dans les décisions des agents sont traçables et inspectables, et que les organisations peuvent démontrer leur conformité, expliquer les résultats et appliquer les contrôles de sécurité de façon cohérente à travers les flux de travail des agents.

Les modèles de connaissances permettent de transformer les données d'affaires en contextes spécifiques à une tâche, l'un des défis les plus difficiles pour les agents.

Sans une compréhension explicite de la signification métier, les agents interprètent souvent mal les relations entre ensembles de données, ce qui mène à des inférences erronées et des hallucinations incontrôlées qui limitent fondamentalement les déploiements autonomes et évolutifs. Parce que les tables de données n'encodent pas la signification métier, elles sont souvent directement intégrées dans les requêtes : bien que cette approche fonctionne pour des cas isolés, elle demeure fragile, difficile à tester et coûteuse à maintenir à travers les agents et les scénarios. Fabric IQ comble cette lacune en traitant les modèles de connaissance comme des actifs partagés et gouvernés : en représentant explicitement les entités commerciales, les relations et les mesures sous une forme lisible par machine, les modèles de connaissances ancrent le raisonnement des agents dans des structures vérifiables plutôt que sur des suppositions inférées. Cet ancrage réduit les hallucinations et permet l'explicabilité : les décisions des agents peuvent être retracées à des concepts d'affaires bien définis, rendant le comportement inspectable, vérifiable et fiable à grande échelle.

Fabric IQ fournit deux types complémentaires de modèles de connaissances :

- Les ontologies représentent des modèles opérationnels en définissant explicitement les entités d'affaires et les relations qui les relient. Fondés sur des technologies de graphes, ils permettent aux agents de raisonner sur des données connectées en traversant des relations complexes. Ces structures basées sur les graphes forment des Context Graphs, permettant une intelligence opérationnelle telle que l'identification de clients dans une région spécifique qui sont associés à des achats récents et à des problèmes de support produit non résolus.
- Les modèles sémantiques représentent la connaissance analytique en définissant explicitement les métriques et les dimensions ainsi que les règles qui régissent leur agrégation, leur calcul et leur interprétation. Fondés sur des concepts établis d'intelligence d'affaires, ils fournissent aux agents une représentation explicite et lisible par machine de l'analytique d'affaires. En tirant parti du stockage en colonnes optimisé, les modèles sémantiques supportent efficacement l'agrégation et les schémas d'analyse basés sur les dimensions de l'entreprise.

Modèles de publication avec MCP, régissant la lecture par opposition à la modification et reliant la sémantique à la recherche

Les agents de données Microsoft Fabric sont conçus et testés de façon systématique pour récupérer, corrélérer et raisonner sur les données d'entreprise à travers des conversations de données en plusieurs étapes et à sources multiples. Ils traduisent les requêtes en langage naturel en plans de raisonnement structurés, orchestrant la récupération contrôlée de données à travers des sources telles que les modèles de connaissances Fabric IQ, les tables Delta dans OneLake et les index Azure AI Search. En combinant les faits récupérés avec une sémantique d'affaires explicite, Data Agents assemblent un contexte ancré qui soutient un raisonnement fiable et des résultats cohérents et pertinents pour la tâche. La couche d'exposition est ce qui rend les produits de données fiablement consommables tant par les agents que par les applications. En fournissant des surfaces d'interaction stables, telles que des terminaux de requête, des agents de données et des protocoles compatibles avec les agents, il assure une sémantique prévisible indépendante de la plateforme sous-jacente ou de l'environnement d'exécution, permettant aux agents de raisonner de façon cohérente à travers des données analytiques et opérationnelles provenant de multiples domaines. À grande échelle, cette couche d'exposition doit être construite sur des protocoles d'accès standards comme MCP et gouvernée par un centre de gouvernance centralisé. MCP permet un accès découvrable et réutilisable aux produits de données et aux modèles de connaissance, donnant aux agents une visibilité sur les données disponibles ainsi que sur leur contexte sémantique et de gouvernance. La gouvernance centrale impose un contrôle d'accès cohérent, des politiques, de la lignée et de la gestion du cycle de vie à travers des environnements hétérogènes, tandis que le traitement des serveurs MCP comme des produits opérationnels de première classe introduit une discipline autour de la gestion des versions, de l'observabilité et de la dépréciation. Ensemble, ces capacités étendent les principes de Data Mesh au plan application, permettant la réutilisation à l'échelle de l'entreprise des capacités d'agents avec cohérence, contrôle et traçabilité, sans dégénérer en intégrations ponctuelles.

Plan d'application

Le plan application est la capacité d'entreprise qui transforme les actifs gouvernés en solutions déployées. Il fournit les environnements d'exécution dans lesquels les agents s'exécutent, les mécanismes d'intégration qui connectent les agents aux outils et aux produits de données, ainsi que les contrôles opérationnels nécessaires à la fiabilité en production.

Dans une architecture en étoile d'entreprise, chaque zone de destination d'application constitue une frontière d'abonnement spoke qui héberge des ressources par workload sous la propriété d'une équipe de projet IA. L'équipe de projet est responsable de la livraison de la solution et des opérations à l'intérieur de sa zone, tandis que l'équipe centrale de la plateforme administre les services de gouvernance partagée tels que la passerelle de modèle, la passerelle MCP et les inventaires d'entreprise. Cette séparation reflète la manière dont les entreprises s'organisent et fonctionnent, car l'équipe centrale impose les contrôles de base pendant que les équipes de projet construisent et exécutent les charges de travail.

Une zone de destination d'application robuste doit fournir une base standardisée pour l'identité, le réseautage, la gouvernance, la surveillance et la gestion des coûts. Elle doit inclure une utilisation cohérente des identités gérées, des limites claires en matière de contrôle d'accès en fonction du rôle, l'intégration du réseau privé au centre et au plan de données, ainsi que des pipelines de télémétrie standardisés. L'objectif est que les équipes d'application puissent déployer des solutions d'agents prêtes à la production sans violer les politiques de l'entreprise et sans réinventer l'infrastructure fondamentale pour chaque projet.

Le plan d'application doit supporter plusieurs formats d'hébergement d'agents, car les scénarios d'entreprise nécessitent différents compromis entre la commodité, le contrôle, l'extensibilité et la responsabilité opérationnelle. Ces formats sont :

- Agents sans code et à faible code
- Agents hébergés liés à Microsoft Foundry
- Agents hébergés dans des conteneurs personnalisés (hors de Microsoft Foundry)

La première catégorie regroupe les agents sans code et à faible code implémentés sous forme d'agents de requête et de flux de travail au sein de Microsoft Foundry Agent Service et associés à un projet Microsoft Foundry. Dans cette catégorie, les équipes utilisent des capacités gérées pour l'orchestration, l'évaluation et la télémétrie dans le périmètre du projet. Ils exploitent des outils exposés via MCP, un ancrage à des sources de connaissances connectées, et une observabilité à l'échelle du projet pour le traçage et l'évaluation.

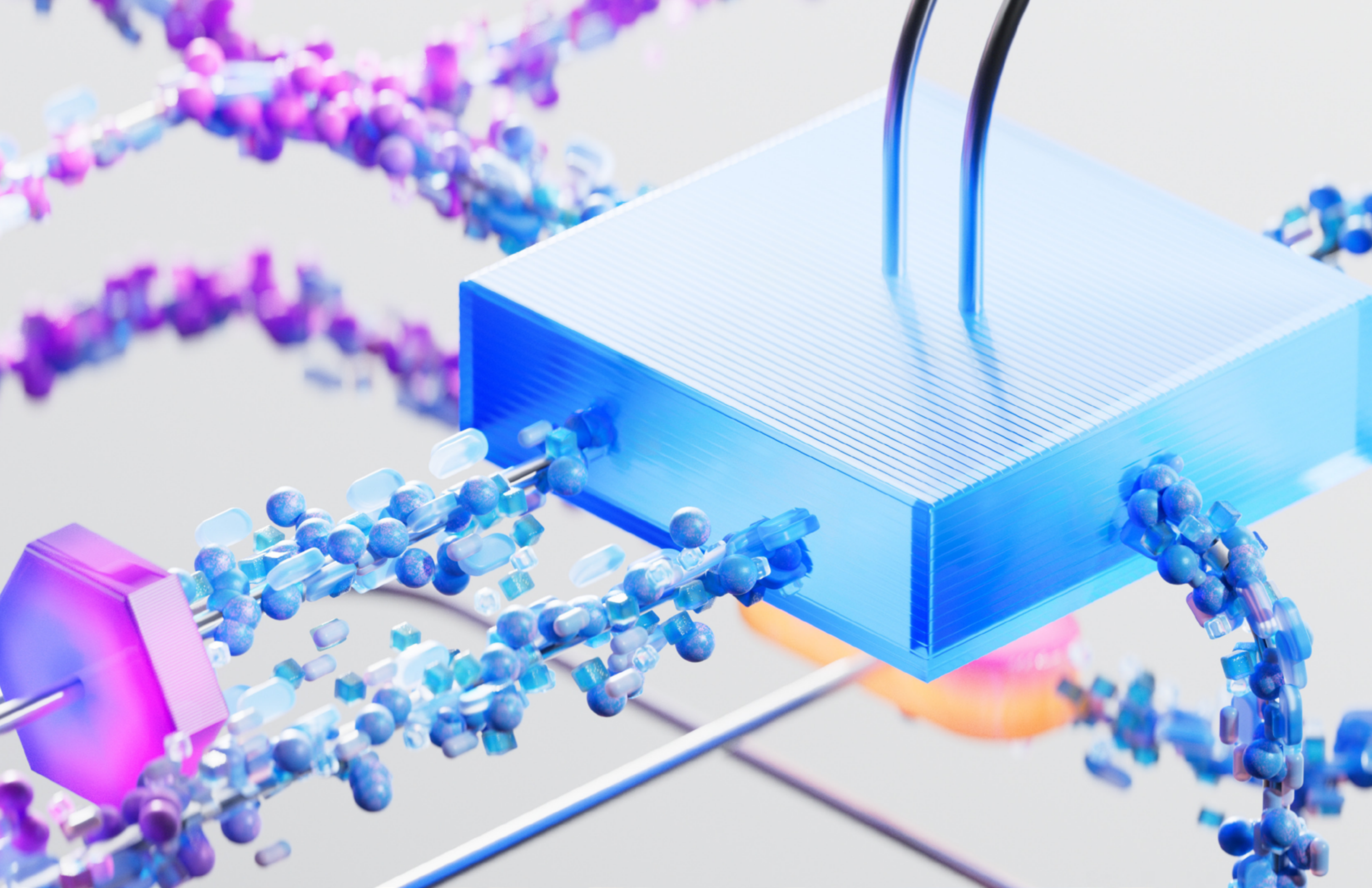
La deuxième catégorie concerne des agents hébergés, associés à un projet Microsoft Foundry, mais exécutés sous forme de workloads conteneurisés construits à partir de code. Dans cette catégorie, les équipes développent des agents avec des pratiques d'ingénierie logicielle avancées, les conditionnent dans des conteneurs et les déploient dans des environnements d'exécution contrôlés, tout en s'intégrant à l'orchestration et à la gouvernance au niveau du projet. Ce modèle est utile lorsque les équipes ont besoin de dépendances personnalisées, de logiques d'intégration d'outils sur mesure, de comportements d'orchestration spécialisés ou d'un contrôle d'exécution spécifique qui

n'est pas disponible dans des environnements totalement sans code. Il permet également une composition mixte, dans laquelle les flux de travail à faible code délèguent des tâches précises à des agents hébergés en code avancé, ce qui permet d'allier de manière pragmatique un montage rapide et une personnalisation poussée.

La troisième catégorie regroupe des agents personnalisés hébergés sur conteneurs hors de Microsoft Azure AI Foundry, incluant des agents déployés sur Azure Container Apps, Azure Kubernetes Service, ou potentiellement dans des environnements non Azure. Dans cette catégorie, l'entreprise a toujours besoin d'une intégration avec les fonctionnalités du centre avec des protocoles standard comme REST, MCP et des modèles de communication d'agent à agent qui exploitent les API et les fonctionnalités de Foundry. L'objectif architectural est que ces agents continuent d'être gérés par des modèles et des outils basés sur un centre de consommation, en utilisant des schémas d'identité qui évitent la diffusion d'informations confidentielles, et en transmettant des données de télémétrie pouvant être corrélées avec le modèle global d'observabilité du réseau maillé.

Un point crucial pour le plan d'application est que l'accès aux modèles et aux outils doit être médiatisé par le centre principal, en utilisant des schémas pilotés par des gabarits plutôt que des terminaux et des informations d'identification intégrés dans chaque agent. Cela garantit la cohérence pour l'application des politiques et la vérifiabilité, permet aux équipes de la plateforme de gérer de manière centralisée la capacité, la configuration de sécurité et le contrôle des coûts, et permet aux équipes d'application de se mettre rapidement au travail en instanciant des modèles approuvés et des mises en œuvre de référence (pour l'hébergement, l'intégration aux terminaux gérés par le centre et aux produits de données exposés au MCP, et la télémétrie) au lieu de devoir repenser l'infrastructure pour chaque projet.





Modèles d'intégration interprojets d'agents et responsabilités opérationnelles de la zone de destination

La caractéristique essentielle d'un plan d'application dans un maillage d'agents est qu'il permet l'interaction entre projets, temps d'exécution et frontières organisationnelles sans tomber dans le couplage point à point. Un plan d'application mature propose des patrons d'intégration qui prennent en charge quatre types de scénarios courants tout en préservant les principes de gouvernance.

Les scénarios sont :

- Flux de travail d'enrichissement et d'indexation centré sur les données
- Assistants interactifs au service de l'utilisateur
- Flux de travail séquentiel et de groupe
- Collaboration d'agent à agent et planification dynamique

La première famille de scénarios est l'enrichissement centré sur les données et les flux de travail d'indexation. Dans ces scénarios, les agents procèdent à l'ingestion, l'analyse, la fragmentation, la classification et l'enrichissement de contenus structurés et non structurés. Le résultat est généralement une surface de récupération gouvernée, telle

qu'un index de recherche, un jeu de données enrichi ou un référentiel de connaissances organisé. Ces flux de travail nécessitent souvent un traitement par lots, une intégration avec des produits de la couche de données, ainsi qu'une gestion rigoureuse des identités et de la vérification, car ils peuvent porter sur des contenus sensibles. Le plan d'application devrait fournir des modèles de conception pour exécuter ces flux de travail comme des processus durables, avec planification explicite, tentatives et télémétrie, plutôt que comme des clavardages interactifs ponctuels.

La deuxième famille de scénarios regroupe des assistants interactifs destinés aux utilisateurs, intégrés dans des canaux tels que Microsoft Teams ou des applications Web. Dans ces scénarios, la latence, la propagation de l'identité et l'expérience utilisateur sont les préoccupations principales. Le plan d'application devrait fournir des modèles standards pour connecter l'identité du canal à celle de l'agent, préserver le contexte au nom de l'utilisateur lorsque nécessaire, et s'assurer que les appels d'outils et l'accès aux produits de données respectent les droits d'accès des utilisateurs. Ce modèle bénéficie de la combinaison de Microsoft Copilot Studio comme couche de canal et d'orchestration avec des agents spécialisés implémentés dans Microsoft Foundry Agent Service ou des environnements de conteneurs lorsqu'un contrôle plus poussé est nécessaire.

La troisième famille de scénarios concerne les flux de travail séquentiels et de groupe qui mettent en œuvre des processus d'affaires avec un ou plusieurs agents. Ces flux de travail peuvent être de longue durée, nécessiter des approbations humaines et s'intégrer à des systèmes d'entreprise tels que la gestion de la relation client, la planification des ressources d'entreprise ou les plateformes de gestion des services informatiques. Le plan d'application devrait fournir des motifs d'orchestration qui externalisent l'état et la logique du flux de travail plutôt que de les cacher dans des requêtes. Il devrait prendre en charge des étapes avec intervention humaine comme états explicites du flux de travail, et il devrait prendre en charge des chemins d'escalade pour la conformité et la responsabilité.

La quatrième famille de scénarios est la collaboration d'agent à agent et la planification dynamique. Dans ces scénarios, un agent décompose un objectif en tâches, identifie des agents spécialisés adaptés, délègue les tâches et agrège les résultats. Ce modèle devient sûr à l'échelle de l'entreprise uniquement lorsque les agents identifient et invoquent d'autres agents avec des registres et des passerelles régulés, plutôt que d'utiliser un partage informel de terminaux. Le plan d'application doit donc s'intégrer à un concept central de passerelle ou de registre d'agents et préserver la sémantique de l'identité et des politiques lors des appels entre agents. Dans un maillage mature, cela inclut le transfert dynamique, le routage basé sur les capacités et des contraintes explicites qui empêchent un agent de déléguer des tâches à des agents non autorisés ou non fiables. La valeur d'entreprise réside dans le fait que les agents spécialisés deviennent des blocs de construction réutilisables et que les solutions complexes deviennent des compositions de capacités de domaine plutôt que des systèmes monolithiques.

La zone de destination doit fournir un environnement standardisé pour héberger et exploiter des applications natives de l'IA construites sur Microsoft Foundry, Azure Container Apps et Azure Kubernetes Service, en assurant que les préoccupations liées à l'identité, au réseau, à la gouvernance, à la surveillance et à la gestion des coûts sont mises en œuvre de manière cohérente. Il convient de traiter Microsoft Azure AI Foundry comme un plan de contrôle pour l'hébergement de modèles, l'orchestration des requêtes et l'intégration de l'exécution des agents, tout en considérant les environnements d'exécution de conteneurs comme des frontières d'isolation pour le déploiement de la logique d'orchestration, des services d'agents compatibles MCP et des intergiciels d'application. Le modèle accélérateur de zone de destination fournit une posture de réseau privé délibérée, utilisant des réseaux virtuels, des terminaux privés, des groupes de sécurité réseau et des schémas privés de résolution DNS afin d'éliminer les chemins d'entrée publics lorsque cela est possible entre

les agents, les passerelles, les services de télémétrie et les magasins de récupération gérés. Les identités gérées devraient constituer le mécanisme par défaut d'accès entre services afin d'éviter la distribution d'identifiants. La zone de destination fournit par défaut des signaux de type OpenTelemetry et des pipelines de télémétrie standardisés, afin que les traces et les métriques puissent être corrélées à travers le maillage.

Enfin, l'accélérateur de zone de destination permet l'automatisation du déploiement à l'aide de modèles d'infrastructure-as-code et de pipelines de livraison continue. Il automatise le marquage des ressources, l'assignation de politiques, la configuration de la surveillance et les contrôles de sécurité de base afin que la conformité demeure cohérente et vérifiable. En mettant en œuvre ces responsabilités sous forme de modèles, les équipes d'application peuvent innover rapidement au sein de leurs architectures en étoile tout en respectant le cadre de gouvernance de l'entreprise.

Concept d'intégration : comment construire le maillage d'agents par le biais d'un atelier et d'une livraison incrémentale avec Microsoft

Un maillage d'agents d'entreprise ne s'implémente pas efficacement par la simple rédaction d'un document d'architecture ou la création d'une plateforme centrale prise isolément. Il est mis en œuvre grâce à un atelier axé sur la prise de décision et à une série contrôlée d'itérations qui permettent de valider les hypothèses à l'aide de charges de travail réelles. Le concept d'intégration décrit ici connecte le centre principal, le plan de données et le plan d'application dans un modèle de livraison pratique qu'une entreprise peut exécuter conjointement avec les équipes de livraison de Microsoft.

La première étape consiste à planifier le centre principal comme fondation, car le centre définit les décisions de gouvernance non négociables qui orientent toutes les architectures en étoile. L'entreprise devrait établir le schéma de passerelle, le modèle d'inventaire MCP et les interfaces de médiation de passerelle dès le début, même si cela commence avec un périmètre minimal viable. L'objectif n'est pas de construire un centre parfait dès le départ; il s'agit plutôt d'offrir un chemin facilité qui rend la consommation gouvernée plus facile que le contournement de la gouvernance. Le centre doit également définir le gabarit par défaut qui encode les politiques de sécurité de base et les exigences de traçabilité de vérification. Lorsque le modèle existe sous forme de modèles déployables, les équipes d'application et de données peuvent être intégrées avec des résultats prévisibles et sans débats de conception répétés sur le cycle de vie et la gouvernance, et non sur des ensembles de données ponctuels.



La troisième étape consiste à intégrer un petit groupe d'équipes d'applications et à mettre en œuvre des scénarios de solution représentatifs qui sollicitent le maillage. Ces scénarios doivent inclure au moins un assistant interactif auservice de l'utilisateur intégré à un canal comme Microsoft Teams, au moins un pipeline d'enrichissement centré sur les données qui exploite des produits du plan de données et au moins un scénario de processus métier axé sur les flux de travail, comprenant l'appel d'outils et des étapes facultatives impliquant uneintervention humaine. Ces implémentations devraient être choisies délibérément afin de forcer des décisions concernant les patrons d'hébergement, les flux d'identités, la gouvernance des outils et la corrélation de l'observabilité.

Les concepts fondamentaux qui permettent la mise à l'échelle sont l'identité des agents, les références de politique, la sémantique de l'observabilité et la discipline en matière de modèles. L'identité de l'agent assure que chaque agent possède une propriété claire, des autorisations, et un état de cycle de vie, et que les actions peuvent être attribuées même lors d'une exécution autonome ou semi-autonome. Les bases de politiques garantissent que l'authentification, l'autorisation, les contrôles de sécurité et les contraintes d'accès aux données sont cohérents dans tous les environnements d'exécution. La sémantique de l'observabilité permet à l'entreprise de tracer les interactions à travers les

agents, les outils et les produits de données, de corréler les défaillances et d'attribuer les coûts et les risques. L'utilisation de modèles permet de transformer ces concepts en une réalité opérationnelle plutôt qu'en simples directives ambitieuses, permettant aux équipes d'étendre l'adoption grâce à une instanciation répétable.

Le résultat final de ce document est un modèle cohérent et réalisable pour construire une plateforme d'entreprise dans laquelle les agents collaborent et communiquent via des contrats régis, les produits de données sont exposés comme des capacités MCP réutilisables, et les équipes d'application livrent des solutions intelligentes à travers les canaux et les systèmes sans compromettre les contrôles d'entreprise. La mise en œuvre devrait se faire progressivement. L'entreprise devrait commencer avec un centre minimal et un petit nombre de produits et d'agents publiés, puis élargir l'inventaire, renforcer les politiques, améliorer la corrélation de la télémétrie et déployer les gabarits à mesure que l'adoption progresse. L'architecture doit être traitée comme un système vivant qui évolue avec les capacités du produit, les apprentissages de l'entreprise et les contraintes réglementaires. Une équipe du centre d'excellence devrait prendre en charge l'architecture vivante, maintenir les gabarits et le modèle de gouvernance, et améliorer continuellement l'expérience d'intégration, tandis que les équipes de domaine et d'application conservent leur autonomie pour innover au sein de leurs architectures en étoile.

Utiliser Microsoft Agent Factory pour instaurer votre propre architecture d'entreprise pour les agents

Microsoft Agent Factory est un programme unifié qui permet aux organisations de créer, déployer et faire évoluer des agents d'IA tout au long du cycle de vie en utilisant des outils comme Copilot Studio, Microsoft Foundry, GitHub Copilot et Microsoft Fabric. Il simplifie l'adoption grâce à un modèle commercial unique basé sur la consommation (P3), ce qui réduit la complexité liée aux licences et permet des dépenses prévisibles et flexibles.

La plateforme intègre plusieurs couches d'intelligence, Work IQ, Fabric IQ et Foundry IQ, afin de fournir aux agents des informations sur le contexte organisationnel, les données d'entreprise et un raisonnement avancé, ce qui améliore la précision et la pertinence. Elle prend en charge un modèle de développement basé sur les rôles, permettant ainsi aux utilisateurs métier et aux développeurs de collaborer à l'aide d'outils à faible code et à code avancé.

Avec l'accès à des Forward Deployed Engineers et à des programmes complets de perfectionnement, les organisations peuvent accélérer le passage de l'expérimentation à la production. Les mécanismes intégrés de gouvernance et de sécurité garantissent un fonctionnement fiable des agents à grande échelle, ce qui aide les organisations à obtenir des résultats commerciaux mesurables.

Voir Microsoft Agent Factory →

Source :

1 <https://learn.microsoft.com/azure/cloud-adoption-framework/ai/>

2 <https://learn.microsoft.com/azure/cloud-adoption-framework/ai-agents/>

© Microsoft Corporation, 2026. Tous droits réservés. Le présent document est fourni « tel quel ». Les informations et les points de vue exprimés dans le document, y compris les adresses URL et autres références à des sites Web, sont susceptibles d'être modifiés sans préavis. Vous assumez tous les risques liés à son utilisation. Le présent document ne vous donne pas les droits juridiques propres à la propriété intellectuelle de tout produit Microsoft. Vous pouvez photocopier et utiliser ce document à titre de référence interne.