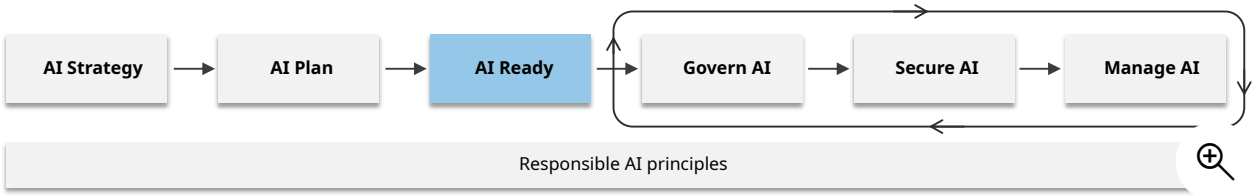
Prêt pour l’IA

Cet article décrit le processus organisationnel pour créer des charges de travail d’IA dans Azure. L’article fournit des recommandations pour prendre des décisions clés en matière de conception et de processus pour adopter les charges de travail d’IA à grande échelle. Il se concentre sur les conseils spécifiques à l’IA pour l’organisation des ressources et la connectivité.



Établir la gouvernance de l’IA

La gouvernance de l'IA nécessite une organisation des ressources et une gestion des politiques appropriées pour garantir des opérations sécurisées, conformes et rentables. Vous devez créer des limites de gouvernance claires pour protéger efficacement les données sensibles et contrôler l’accès aux ressources IA. Voici comment procéder :

1. **Créez des groupes de gestion distincts pour les charges de travail IA exposées à Internet et internes.** La séparation des groupes d’administration établit des limites de gouvernance des données critiques entre les applications IA externes (« en ligne ») et internes uniquement (« d’entreprise »). Cette séparation empêche les utilisateurs externes d’accéder aux données métier internes sensibles tout en conservant les contrôles d’accès appropriés. L’approche s’aligne sur les principes d’architecture de groupe d’administration des zones d’atterrissage Azure et prend en charge l’héritage de stratégie entre les types de charge de travail.

2. **Appliquez des stratégies spécifiques à l’IA à chaque groupe d’administration.** Commencez par des stratégies de base à partir de zones d’atterrissage Azure et ajoutez des définitions Azure Policy pour Foundry, Foundry Tools, Recherche Azure AI et machines virtuelles Azure. L’application des stratégies garantit une gouvernance uniforme de l’IA sur votre plateforme et réduit la surveillance manuelle de la conformité.

3. **Déployez des ressources IA dans des abonnements spécifiques à la charge de travail.** Les ressources IA doivent hériter des stratégies de gouvernance de leur groupe de gestion de charge de travail plutôt que des abonnements de la plateforme. Cette séparation empêche les goulots d’étranglement dans le développement que créent les

contrôles de l'équipe de plateforme et permet aux équipes opérationnelles d'opérer avec une autonomie adéquate. Déployez des charges de travail IA dans des abonnements de zone de lancement d'application au sein des environnements Azure.

Établir un réseau IA

Le réseautage IA englobe la conception de l'infrastructure réseau, les mesures de sécurité et les modèles de transfert de données efficaces pour les charges de travail IA. Vous devez implémenter des contrôles de sécurité et des options de connectivité appropriés pour empêcher les interruptions basées sur le réseau et maintenir des performances cohérentes. Voici comment procéder :

1. **Activez Azure DDoS Protection pour les charges de travail IA accessibles sur Internet.**Azure DDoS Protection protège vos services IA contre les interruptions potentielles et les temps d'arrêt qui provoquent des attaques par déni de service distribuées. La protection DDoS au niveau du réseau virtuel se défend contre les inondations de trafic qui ciblent les applications accessibles sur Internet et conservent la disponibilité du service pendant les attaques.

2. **Sécuriser l'accès opérationnel aux charges de travail IA avec Azure Bastion.** Utilisez un jumpbox et Azure Bastion pour sécuriser l'accès opérationnel aux charges de travail IA et empêcher l'exposition directe d'Internet des interfaces de gestion. Cette approche crée une passerelle sécurisée pour les tâches d'administration tout en conservant l'isolation réseau pour les ressources IA.

3. **Choisissez la connectivité appropriée pour les sources de données locales.** Les organisations qui transfèrent de grandes quantités de données à partir de sources locales vers des environnements cloud ont besoin de connexions à bande passante élevée pour prendre en charge les exigences de performances des charges de travail IA.

    - **Utilisez Azure ExpressRoute pour le transfert de données en volume élevé.**Azure ExpressRoute fournit une connectivité dédiée pour les volumes de données élevés, le traitement en temps réel ou les charges de travail nécessitant des performances cohérentes. ExpressRoute inclut une fonctionnalité FastPath qui améliore les performances du chemin de données en contournant la passerelle ExpressRoute pour des flux de trafic spécifiques.

    - **Utilisez la passerelle VPN Azure pour le transfert de données modéré.**La passerelle VPN Azure fonctionne bien pour les volumes de données modérés, le transfert de données peu fréquent ou lorsque l'accès à Internet public est

nécessaire. La passerelle VPN offre une configuration plus simple et une opération économique pour les jeux de données plus petits par rapport à ExpressRoute. Utilisez la topologie et la conception appropriées pour vos charges de travail IA, notamment le VPN de site à site pour la connectivité intersite et le VPN point à site pour l'accès sécurisé aux appareils.

Établir la fiabilité de l'IA

La fiabilité de l'IA nécessite un placement stratégique de région et une planification de redondance pour garantir des performances et une haute disponibilité cohérentes. Les organisations doivent traiter l'hébergement de modèle, la localité des données et la récupération d'urgence pour maintenir des services IA fiables. Vous devez planifier votre stratégie de déploiement régional pour éviter les interruptions de service et optimiser les performances. Voici comment procéder :

1. **Déployez des endpoints IA dans plusieurs régions pour des charges de travail de production.** Les charges de travail IA de production nécessitent l'hébergement dans au moins deux régions pour fournir une redondance et garantir une haute disponibilité. Les déploiements multirégions permettent un basculement et une récupération plus rapides pendant les défaillances régionales. Pour Azure OpenAI dans Foundry, utilisez des déploiements globaux qui acheminent automatiquement les demandes vers des régions avec une capacité disponible. Pour les déploiements régionaux, implémentez gestion des API Azure pour équilibrer la charge des demandes d'API sur les points de terminaison IA.

2. **Vérifiez la disponibilité du service IA dans les régions cibles avant le déploiement.** Les différentes régions fournissent différents niveaux de disponibilité du service IA et de prise en charge des fonctionnalités. Vérifiez la disponibilité du service Azure par région pour vérifier que vos services IA requis sont disponibles. Les modèles de déploiement Azure OpenAI incluent des options mondiales standard, mondiales approvisionnées, régionales standard et régionales approvisionnées, avec différents modèles de disponibilité régionale.

3. **Évaluez les limites de quota régionaux et les besoins en capacité.** Les outils foundry ont des limites d'abonnement régionales qui affectent les déploiements de modèles à grande échelle et les charges de travail d'inférence. Contactez le support Azure de manière proactive lorsque vous prévoyez des besoins de capacité qui dépassent les quotas standard pour empêcher les interruptions de service pendant la mise à l'échelle.

4. **Optimisez le placement des données pour les applications de génération**

augmentée par récupération. L'emplacement de stockage des données affecte considérablement les performances des applications dans les scénarios RAG. La colocalisation des données avec des modèles IA dans la même région réduit la latence et améliore l'efficacité de la récupération des données, bien que les configurations interrégions restent viables pour des besoins métier spécifiques.

5. **Répliquez les ressources IA critiques dans les régions secondaires pour assurer la continuité de l'activité.** La continuité des activités nécessite la réplication de modèles affinés, de jeux de données RAG, de modèles formés et de données d'apprentissage dans des régions secondaires. La réplication des ressources permet une récupération plus rapide pendant les pannes et maintient la disponibilité du service dans différents scénarios d'échec.

Établir une fondation pour l'IA

Une fondation pour l'IA fournit l'infrastructure de base et la hiérarchie des ressources qui soutiennent les charges de travail d'IA dans Azure. Elle comprend la mise en place d'environnements évolutifs et sécurisés qui s'alignent sur les besoins en gouvernance et en exploitation. Une fondation solide pour l'IA permet un déploiement et une gestion efficaces des charges de travail d'IA. Elle assure également la sécurité et la flexibilité pour la croissance future.

Utiliser une zone d'atterrissage Azure

Une zone d'atterrissage Azure est le point de départ recommandé qui prépare votre environnement Azure. Elle fournit une configuration prédéfinie pour les ressources de la plateforme et des applications. Une fois la plateforme en place, vous pouvez déployer des charges de travail d'IA dans des zones d'atterrissage d'applications dédiées.

Si votre organisation utilise des zones d'atterrissage Azure pour les charges de travail, continuez à les utiliser pour les charges de travail qui utilisent l'IA. Vous déployez vos charges de travail IA dans des zones d'atterrissage d'applications standard, comme vous le feriez pour toute autre charge de travail. Consultez l'IA dans les zones d'atterrissage Azure. La figure 2 ci-dessous illustre comment les charges de travail d'IA s'intègrent dans une zone d'atterrissage Azure.

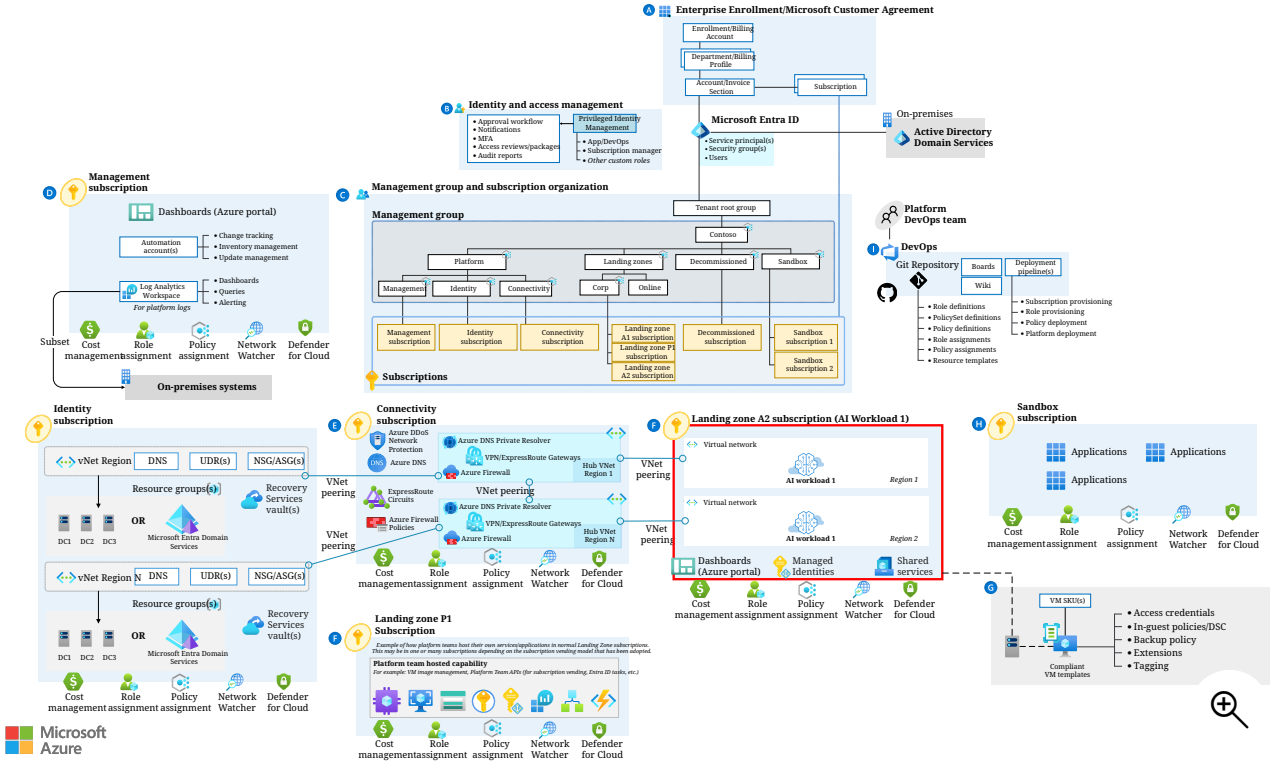


Figure 2. Charge de travail d'IA dans une zone d'atterrissage Azure.

Construisez un environnement d'IA

Si vous n'utilisez pas de zone d'atterrissage Azure, suivez les recommandations de cet article pour créer votre environnement d'IA. Le schéma suivant montre une hiérarchie de ressources de base. Il segmente les charges de travail IA internes et les charges de travail IA accessibles sur Internet. Les charges de travail internes utilisent des politiques pour refuser l'accès en ligne des clients. Cette séparation protège les données internes contre l'exposition aux utilisateurs externes. Le développement d'IA devrait utiliser un jumpbox pour gérer les ressources et les données d'IA.

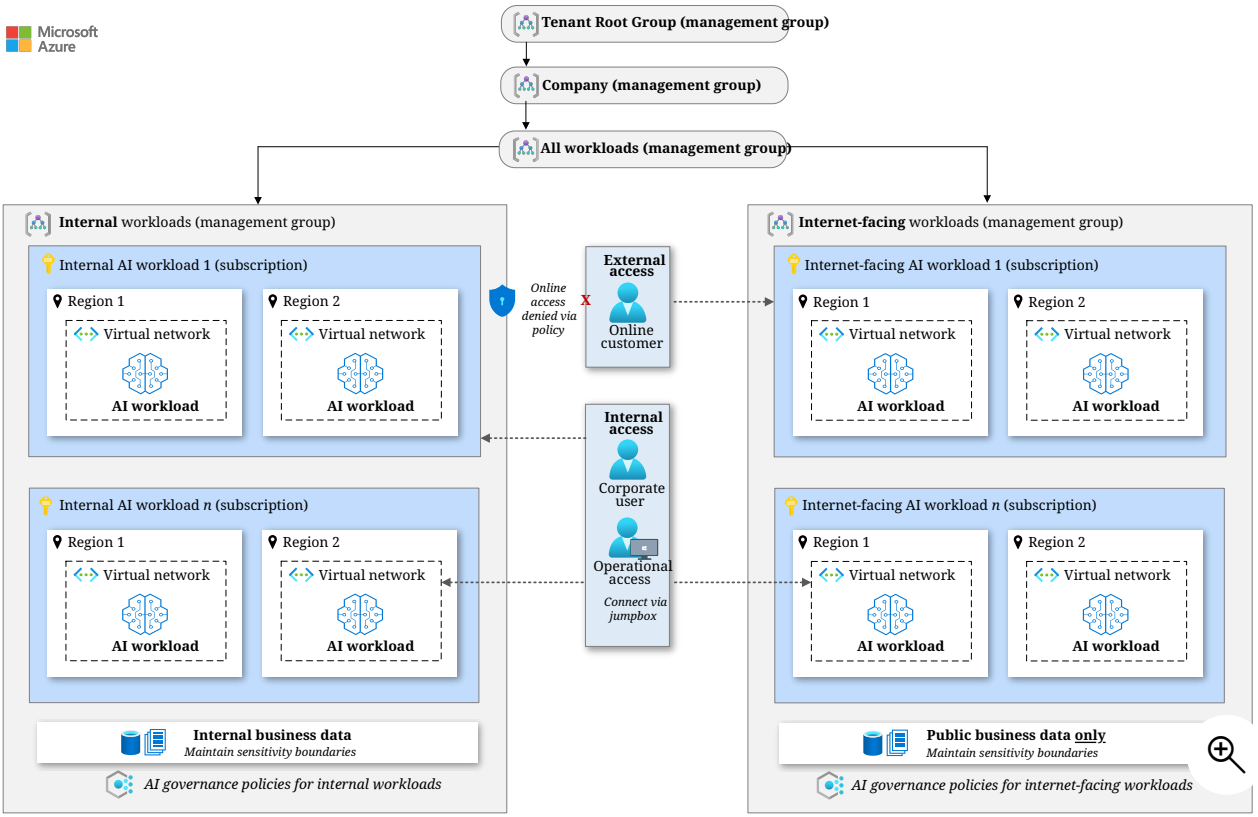


Figure 3. Hiérarchie des ressources de base pour les charges de travail d'IA.

Étapes suivantes

L'étape suivante consiste à construire et déployer les charges de travail d'IA dans votre environnement d'IA. Utilisez les liens suivants pour trouver les conseils architecturaux qui répondent à vos besoins. Commencez par les architectures de plateforme en tant que service (PaaS). PaaS est l'approche recommandée par Microsoft pour adopter l'IA.

Conseils architecturaux pour l'IA PaaS

Conseils architecturaux pour l'IA IaaS

Last updated on 03/07/2025