MEDICALCITY

WHITEPAPER · SEPTEMBRE 2026 · VERSION 1.0

Le workspace agentique cloud — architecture de référence souveraine pour la santé et l'autonomie

Cinq couches, une infrastructure certifiée HDS, trois positions d'exécution : cloud, edge, embarqué. La face infrastructure du workspace agentique souverain pour la santé et l'autonomie.

Document	Whitepaper MedicalCity · version 1.0 · publication medicalcity.ai
Date	9 septembre 2026
Classification	Public — licence CC BY-NC-ND 4.0
Auteur	Richard YI, président — MEDICALCITY
Société	SASU au capital de 1 299 € · 107 542 565 R.C.S. Paris

LECTURE

Auteur : Richard YI, président — MEDICALCITY. Public : DSI et RSSI d'établissements et de services à domicile, intégrateurs de DPI, hébergeurs, architectes d'entreprise, autorités et investisseurs.

TL;DR (60 secondes)

L'IA médicale ne manque pas de modèles — elle manque d'infrastructure de gouvernance. ***MedicalCity est un workspace agentique souverain pour la santé et l'autonomie*** *: donnée patient unifiée (FHIR), analytique clinique et agents IA gouvernés, avec la conformité HDS / AI Act / EHDS by design. Ce whitepaper décrit la* ***face infrastructure*** *de ce workspace : une pile de référence en* ***cinq couches*** *(Exposition, Agent framework, RAG stack, Data platform, Infra & gouvernance), hébergée sur une* ***infrastructure certifiée HDS en France****, et conçue pour que le cloud ne soit qu'une des* ***trois positions d'exécution*** *— cloud, edge, embarqué — d'une même politique de gouvernance. Le principe directeur tient en une phrase :* ***dans un workspace agentique, la conformité n'est pas une documentation, c'est une condition d'exécution.***

1. Pourquoi un whitepaper « cloud »

Le whitepaper *Le workspace agentique pour la santé et l'autonomie* établit la catégorie : ce qu'est un workspace agentique, pourquoi la santé en a besoin, et ce que le MDM agentique Bayon apporte que les plateformes généralistes n'ont pas. Il répond à la question « quoi ». Celui-ci répond à la question que posent, dès la deuxième réunion, les DSI, les RSSI et les hébergeurs : **« sur quoi cela tourne-t-il, où, et qui en répond ? »**

La question est légitime, et elle est souvent mal posée. Le mot *cloud* fait penser à un lieu — un datacenter, une région, un fournisseur. Pour des données de santé, le cloud est d'abord **un contrat** : un périmètre certifié, une chaîne de responsabilités écrite, une garantie de non-transfert, une réversibilité démontrable, et un journal qui permet de prouver, après coup, que tout cela a été respecté. Un workspace agentique ajoute une exigence propre : les agents **agissent**. Ils appellent des outils, écrivent dans des référentiels, déclenchent des alertes. L'infrastructure ne peut donc pas se contenter d'héberger ; elle doit **contraindre**.

Ce document décrit comment MedicalCity traduit ces exigences en architecture. Il s'appuie sur la pile de référence en cinq couches telle qu'elle est exploitée au 9 septembre 2026, distingue explicitement ce qui est **en place** de ce qui est **cible**, et s'interdit toute promesse de performance commerciale : les valeurs citées sont des **invariants d'architecture**, pas des engagements de service.

---

2. Ce que nous appelons un workspace agentique souverain

> *Un workspace agentique souverain est une plateforme qui unifie, pour un domaine métier exigeant, la donnée, l'analytique et les agents IA, sous une gouvernance audit-grade. Il sert plusieurs personas (du no-code au framework complet) et garantit que toute action agentique est traçable, réversible et conforme aux obligations réglementaires du domaine.*

Trois principes en découlent, et l'infrastructure est jugée à sa capacité à les tenir **par construction** :

1. **Gouvernance audit-grade by design** — chaque action, chaque décision et chaque modification laisse une trace dans un journal chaîné, en insertion seule, conservé sept ans.
2. **Orchestrateur déterministe** — le modèle de langage est un sous-composant borné du workspace, jamais son pilote. Ce qui décide de l'enchaînement des étapes est un moteur de workflow, pas un prompt.
3. **Contrôle humain systématique sur toute action à impact** — matérialisé par une matrice à trois niveaux (AUTO, RECO, OBLIG) et par deux boucles de contrôle, l'une réflexe, l'autre supervisée.

Le pattern n'est pas théorique. Un workspace agentique souverain construit sur ces principes est déployé depuis cinq ans dans la gestion d'actifs, en production multi-entités avec plus de 1 000 utilisateurs. Le passage à la santé n'est pas un saut technologique : c'est une **verticalisation**, avec trois ajouts que le domaine impose — un référentiel de vérité patient (Bayon), une couche réflexe embarquée, et une géolocalisation juridique des actions.

3. La pile de référence en cinq couches

Figure — Workspace agentique — les cinq couches et leur état (pile de référence, 9 septembre 2026)

La pile se lit de haut en bas, de ce que l'utilisateur voit à ce qui le garantit. Chaque couche répond à une question distincte ; aucune ne contourne celle du dessous.

3.1 Exposition — « par où entre-t-on ? »

Brique	Rôle	Garantie
Web UI (trois personas)	Interfaces pour l'utilisateur métier (no-code), l'analyste (low-code) et le développeur (framework complet)	Une seule surface d'accès, un seul modèle de droits
API Gateway	Point d'entrée unique des appels machines et des consoles, authentification déléguée, quotas	Aucun appel direct aux services internes
SDK Python	Construction d'agents et d'outils par les équipes d'intégration	Les mêmes registres, les mêmes contrôles que l'interface
Connecteurs M365 / Teams / Outlook	Insertion du workspace dans l'environnement de travail existant	Identité de l'utilisateur propagée, jamais substituée

La règle de la couche : **les consoles sont des clients comme les autres**. Une console d'administration n'a pas de porte dérobée ; elle passe par le gateway, sous une identité déléguée, et ses actions sont journalisées au même titre que celles d'un agent.

3.2 Agent framework — « qui a le droit de faire quoi ? »

C'est le cœur du produit, et la couche que les plateformes généralistes n'ont pas.

Brique	Rôle	Garantie
Orchestrateur (Argo Workflows)	Exécution déterministe des workflows agentiques, étape par étape	Le LLM ne décide jamais de l'enchaînement
Tools registry, sandboxing, ACL	Catalogue des outils exposés aux agents, isolation d'exécution, contrôle d'accès	Aucun agent n'appelle un système directement
LLM Gateway (multi-modèles, multi-cloud)	Routage des appels vers les modèles, en fonction du niveau de risque et de la juridiction	Aucune donnée de santé identifiante hors périmètre HDS (voir §6)
Cache sémantique, quotas, fallback	Maîtrise des coûts et de la disponibilité des modèles	Dégradation contrôlée, jamais silencieuse
Human-in-the-loop hooks	Points d'arrêt et de validation humaine intégrés au workflow	Matrice AUTO / RECO / OBLIG appliquée à chaque étape

Deux registres soutiennent cette couche. Le **registre de compétences** dit ce qu'un agent a le droit de faire : chaque compétence est enregistrée, versionnée et signée, avec son niveau de risque, son niveau de contrôle humain, ses outils autorisés et son jeu de test. Le **registre des agents** dit qui exécute et s'il est légitime : chaque agent est déclaré, attesté par une empreinte vérifiée à chaque démarrage, doté d'un budget, d'un périmètre et d'un **responsable humain nominatif**.

3.3 RAG stack — « d'où vient ce que l'agent sait ? »

Brique	Rôle	Garantie
Connecteurs sources (Confluence, SharePoint, GED)	Ingestion des corpus documentaires de l'établissement	Périmètre de source déclaré, versionné
Chunking adaptatif (512–1024 tokens, recouvrement 15 %)	Découpage des documents en unités citables	Chaque extrait reste rattachable à sa source
Embeddings multilingues + vector store (Qdrant)	Indexation sémantique	Index reconstruisible depuis les sources
Re-rank (BGE-reranker)	Sélection des passages réellement pertinents	Moins de contexte, mieux cité

La règle de la couche : **tout savoir utilisé par un agent est cité et versionné**. Une recommandation issue d'un référentiel de bonnes pratiques porte la référence du texte, sa version et sa date. Un corpus non cité n'entre pas dans le workspace.

3.4 Data platform — « qu'est-ce qui est vrai ? »

Brique	Rôle	Garantie
Lakehouse Delta sur S3 / MinIO	Stockage analytique en formats ouverts	Réversibilité : les données se relisent sans la plateforme
Médaillon (Bronze / Silver / Gold)	Progression des données brutes vers les produits de données	Chaque niveau est reconstruisible depuis le précédent
Unity Catalog / LakeFS	Catalogue, lignage, versionnement des jeux de données	Une requête répond « d'où vient cette valeur ? »
Spark sur Kubernetes (Databricks)	Calcul distribué, analytique clinique	Isolation par tenant au niveau du calcul
Streaming Spark + Kafka	Événements en flux (DPI, pharmacie, laboratoire)	Le workspace observe, il ne réécrit pas la source

Au-dessus de cette couche, le **MDM agentique Bayon** tient le référentiel de vérité : golden record patient, professionnel et couverture ; résolution d'entités probabiliste, arbitrable, jamais de fusion automatique ; actions réversibles par sagas ; journal chaîné en insertion seule. Bayon est ce qui manque aux workspaces agentiques généralistes : un endroit où l'on peut répondre, de façon opposable, à la question « qu'est-ce qui est vrai ? ».

3.5 Infra & gouvernance — « qui garantit tout cela ? »

Brique	Rôle	Garantie
Kubernetes multi-cluster (GitOps ArgoCD)	Exécution et déploiement déclaratifs	Tout ce qui tourne est décrit dans un dépôt signé
Keycloak (SSO MedicalCity) + Vault + OPA / Kyverno	Identité, secrets, politiques d'admission	Aucune politique n'existe hors du code
OpenTelemetry → Grafana + Prometheus + Loki	Traces, métriques, journaux techniques	Un incident se reconstruit de bout en bout
Terraform + Crossplane + GitLab CI	Infrastructure as code, chaîne d'intégration continue	Reconstruction de l'environnement depuis zéro

La règle de la couche : **l'infrastructure est elle-même un objet d'audit**. Un changement de configuration est un commit, revu, signé et déployé par GitOps ; il n'y a pas de modification manuelle d'un cluster de production. Une suite de tests d'intégration continue est en place.

Statut au 9 septembre 2026 : l'ensemble des composants des cinq couches est opérationnel et confirmé.

4. Ce que « souverain » veut dire ici

Le mot est galvaudé. Nous lui donnons quatre contenus vérifiables.

4.1 Un périmètre certifié, en France

Les données de santé à caractère personnel sont hébergées sur l'infrastructure de **Scaleway**, hébergeur **certifié HDS** (référentiel v2.0) pour les activités 1 à 4 du référentiel — site physique, infrastructure matérielle, plateforme d'hébergement, infrastructure virtuelle — dans la région de Paris. MedicalCity conduit les activités 5 et 6 (administration de la plateforme applicative, sauvegarde) sur le stockage inclus dans ce périmètre certifié, et **n'est pas elle-même titulaire d'un certificat HDS** à la date de ce document. Cette répartition est écrite dans l'annexe d'hébergement remise à chaque client. Nous ne présentons jamais MedicalCity comme « certifiée HDS » ; nous la présentons comme **hébergée sur une infrastructure certifiée HDS**, avec une trajectoire de certification propre.

4.2 Aucun transfert hors de l'Espace économique européen

Aucune donnée de santé ne quitte l'Espace économique européen. Tout accès à distance depuis un pays tiers est interdit sauf accord écrit préalable du client et mise en place des garanties requises. L'hébergeur est une société de droit français ; le contrat, le droit applicable et le juge sont français. La qualification SecNumCloud n'est pas revendiquée.

4.3 La chaîne de responsabilités est écrite

L'hébergeur intervient comme sous-traitant ultérieur au sens de l'article 28.4 du RGPD ; la liste des sous-traitants ultérieurs figure au contrat de traitement des données, le client dispose d'un droit d'objection et d'un droit d'audit. Un référent hébergement est nommé. Ce n'est pas une innovation : c'est la base, et elle est trop souvent absente des offres « IA santé ».

4.4 La réversibilité est une propriété technique, pas une clause

Un cloud souverain dont on ne peut pas sortir n'est pas souverain. La pile de référence est construite sur des formats et des interfaces ouverts : tables Delta et Parquet sur stockage objet compatible S3, orchestration Kubernetes, infrastructure décrite en Terraform, politiques exprimées en code, journaux d'audit exportables avec leur chaîne de hachage. Un client peut relire ses données, rejouer son journal et reconstruire son environnement sans MedicalCity.

---

5. Le continuum cloud – edge – embarqué

Le cloud n'est pas le seul endroit où un agent agit. Dans la santé et l'autonomie, une partie des actions s'exécute **au bord du réseau** — poste de travail d'un service à domicile, serveur d'établissement — et une partie **dans l'équipement lui-même** — robot d'assistance, terminal mobile. L'architecture traite ces trois positions comme un continuum régi par une seule politique, avec deux mécanismes.

5.1 La géolocalisation juridique

Chaque compétence, chaque action métier et chaque mapping réglementaire porte un champ de portée juridictionnelle qui le rattache à un ou plusieurs **packs de juridiction** :

Pack	Périmètre	Cadre	Exécution
`juris:FR-HDS`	France	HDS, RGPD, AI Act, EHDS, référentiels de l'ANS	Cloud HDS France, edge, embarqué
`juris:EU`	Espace économique européen	RGPD, AI Act, EHDS	Hébergement EEE
`juris:CN-PIPL`	Chine	PIPL, MLPS	Mode démonstration, sans donnée personnelle ni de santé réelle
`juris:LOCAL-ONLY`	Hors couverture	—	Aucune donnée ne sort de l'équipement ; fail-closed par défaut

Un mapping hors juridiction de la mission est inactif ; l'action qui en dépend est bloquée ou re-routée. Le principe de **double localisation** — site déclaré, signé dans le contexte de mission, contre position observée — fait de toute divergence persistante un motif d'interlock : blocage de l'action et escalade humaine.

5.2 Deux boucles de contrôle

- **Réflexe** — un interlock déterministe local, sous forme d'un bundle préchargé, signé, versionné et daté, évalué avant chaque appel d'outil sensible. Son temps d'évaluation, inférieur à dix millisecondes, est un **invariant d'architecture** : le bundle est en mémoire, l'évaluation est locale, aucun appel réseau n'est sur le chemin. Il est fail-safe : en cas de doute, il refuse.
- **Supervisée** — un humain distant, de la seconde à l'heure, avec un mode *last-known-good* qui maintient la dernière politique valide connue lorsque le lien avec le cloud est dégradé.

Le cloud produit et signe les politiques ; l'edge et l'embarqué les appliquent sans avoir besoin de lui au moment de l'action. C'est ce qui permet d'avoir, sur un robot d'assistance à domicile, la même gouvernance que dans le datacenter — sans dépendre de la qualité de la connexion au moment où elle compte.

---

6. Les données de santé et les modèles

La question la plus fréquente des RSSI est simple : **« où vont les prompts ? »** La réponse tient en trois règles, appliquées par le LLM Gateway et vérifiables dans le journal.

1. **Aucune donnée de santé identifiante ne quitte le périmètre HDS.** Les modèles qui traitent de la donnée patient sont exécutés dans le périmètre certifié, ou ne reçoivent que des données pseudonymisées selon une politique versionnée. Le choix est fait par le gateway en fonction de la juridiction et du niveau de risque de la compétence, pas par l'agent.
2. **Multi-modèles, par conception.** Le gateway route vers plusieurs modèles, avec cache sémantique, quotas et repli. Aucun agent ne dépend d'un fournisseur unique ; un modèle peut être remplacé sans réécrire une compétence.

3. **Le modèle n'est jamais la source de vérité.** Ce qu'un agent affirme sur un patient vient de Bayon ; ce qu'il cite vient d'un corpus versionné ; ce qu'il a le droit de faire vient des registres. Le modèle formule, il n'établit pas.

Ces règles ont une conséquence architecturale assumée : la plateforme accepte un coût et une latence supérieurs à ceux d'un appel direct à un modèle public, en échange d'une réponse opposable à « où est passée cette donnée ? ».

---

7. La gouvernance audit-grade, vue de l'infrastructure

7.1 Le journal

Le journal d'audit est **chaîné, en insertion seule**, conservé sept ans. Chaque étape d'une action agentique y écrit une ligne horodatée — et géo-estampillée pour les exécutions embarquées — avec l'identité de l'agent, la compétence invoquée, l'outil appelé, le niveau de contrôle humain appliqué et, le cas échéant, le validateur. Un refus de l'interlock est journalisé comme une action. Il n'y a pas d'étape silencieuse.

Le journal est plus qu'une preuve : **il est le corpus annoté**. Chaque arbitrage humain devient un exemple daté ; les régularités deviennent des règles candidates, soumises à validation avant d'entrer dans un registre.

7.2 La matrice de contrôle humain

Niveau	Signification	Exemple
AUTO	L'agent exécute ; l'action est journalisée	Lecture d'un référentiel, simulation d'une interface
RECO	L'agent propose ; un humain valide avant exécution	Mise à jour d'un dossier, rapprochement d'identités
OBLIG	Validation humaine obligatoire, tracée nominativement	Assouplissement d'une politique, modification de la matrice elle-même

En identification patient, sous un seuil de confiance de 0,7, l'étape est bloquée et escaladée : la plateforme préfère une attente humaine à une fusion automatique.

7.3 L'infrastructure comme objet d'audit

La gouvernance ne s'arrête pas aux agents. Le même principe — *rien n'existe hors du code, tout changement est signé* — s'applique aux clusters, aux politiques d'admission, aux secrets et aux pipelines. Un auditeur peut reconstruire l'état de la plateforme à une date donnée à partir du dépôt GitOps et du journal, et comparer ce qui était déclaré à ce qui tournait.

Ces mécanismes répondent directement à l'article 12 (tenue de registres) et à l'article 14 (contrôle humain) du règlement sur l'intelligence artificielle, et outillent la notification d'incident grave de l'article 73.

---

8. Multi-tenant et isolation

Le cas d'usage d'intégration DPI multi-tenant est celui qui sollicite le plus l'infrastructure : plusieurs dossiers patients informatisés du marché, plusieurs établissements, un même workspace.

- **Isolation par tenant** à chaque couche : espaces de noms et politiques d'admission distincts au niveau Kubernetes, catalogues et jeux de données séparés au niveau de la data platform, index vectoriels et corpus par tenant, budgets et périmètres d'agents par tenant dans les registres.
- **Contrat de données versionné** : chaque branchement de DPI est décrit par un contrat — schéma, cardinalités, terminologies (SNOMED CT, LOINC, CIM-10), identité INS — versionné et testé. Un changement de contrat est une nouvelle version, pas une modification en place.
- **Connectique de santé** : FHIR R4, HL7 v2, DICOMweb ; chaque appel est audité et rattaché à un tenant.
- **Le workspace observe, il ne réécrit pas la source** : le DPI reste le système d'enregistrement ; le workspace tient le référentiel de vérité consolidé et propose, sous contrôle humain, les corrections à reporter.

---

9. Exploitation

- **Observabilité** — traces, métriques et journaux techniques unifiés par OpenTelemetry vers Grafana, Prometheus et Loki ; un incident se reconstruit de bout en bout, de l'appel utilisateur à l'appel d'outil.
- **Sauvegarde et reprise** — conduites dans le périmètre certifié ; la politique de sauvegarde et le plan de reprise sont annexés au contrat client.
- **Chaîne d'intégration continue** — suite de tests d'intégration en place ; un déploiement qui ne passe pas les tests de gouvernance (registres, politiques, journal) ne passe pas en production.
- **Incidents** — un incident qui relève d'un incident grave au sens de l'article 73 déclenche le circuit de notification outillé par le journal.
- **Évaluation continue des modèles** — les compétences portent leur jeu de test ; un changement de modèle est évalué avant d'être routé.

---

10. Ce que cette architecture n'est pas

La confiance se gagne aussi par ce que l'on refuse de dire.

- MedicalCity **n'est pas un hébergeur certifié HDS** ; elle est hébergée sur une infrastructure certifiée HDS et conduit les activités applicatives dans ce périmètre, avec une certification propre en trajectoire.
- La pile **ne revendique pas** la qualification SecNumCloud.
- Un agent de surveillance dont l'alerte constituerait une décision clinique entrerait dans le champ du règlement sur les dispositifs médicaux ; MedicalCity le prend en compte dans le niveau de contrôle humain et ne présente aucun de ses agents comme un dispositif médical.
- Les scénarios de robotique d'assistance sont des **scénarios illustratifs prospectifs**, pas des références clients.

- Les valeurs de temps citées sont des invariants d'architecture, pas des engagements de niveau de service.

11. Feuille de route

Jalon	Cible	Contenu
v0.1	T1 2027	Workspace santé sur la pile de référence, premier cas d'usage de cadrage de dossiers
v0.5	T2 2027	Multi-tenant, pilotes en France
v1.0	T4 2027	Quatre cas d'usage marché, certification HDS propre en cours
v1.5	2028	Extension au continuum embarqué (robotique d'assistance)

Ces jalons sont des cibles, présentées comme telles.

12. Conclusion

Un workspace agentique pour la santé ne se juge pas à la qualité de ses modèles mais à la solidité de ce qui les entoure : un périmètre certifié, une chaîne de responsabilités écrite, des registres qui disent qui a le droit de faire quoi, un journal qui prouve ce qui a été fait, et une couche réflexe qui refuse quand il faut refuser — y compris hors ligne. Les cinq couches décrites ici existent pour cela. Le cloud en est le lieu de production ; il n'en est pas la limite.

Demander une démonstration — medicalcity.ai · richard.yi@medicalcity.ai

Références réglementaires

1. Règlement (UE) 2024/1689 établissant des règles harmonisées concernant l'intelligence artificielle (AI Act), notamment articles 12, 14 et 73.
2. Règlement (UE) 2025/327 relatif à l'espace européen des données de santé (EHDS).
3. Règlement (UE) 2016/679 relatif à la protection des données (RGPD), notamment article 28.
4. Référentiel de certification « Hébergeur de données de santé » (HDS) v2.0 — Agence du Numérique en Santé ; article L. 1111-8 du Code de la santé publique.
5. Règlement (UE) 2017/745 relatif aux dispositifs médicaux (MDR).

MEDICALCITY — SASU au capital de 1 299 € · 107 542 565 R.C.S. Paris · 66 avenue des Champs-Élysées, 75008 Paris · medicalcity.ai · richard.yi@medicalcity.ai