

Sektorradar: Cloud 3.0 aktører — Analyse av leverandører innen hybrid, multi-cloud og suveren sky for AI-infrastruktur

Mars 2026 — hva som er varmt, stigende, stabilt og avtagende

Pulsrapport for tech-interesserte. Oppdatert basert på ferske signaler (siste 3–6 måneder).

1. Sektorbilde — Cloud 3.0 aktører i dag

Cloud 3.0 representerer en fundamental skifte vekk fra den sentrale, monolittiske public cloud-modellen. Sektoren er nå preget av et heterogent økosystem med hybrid, private og suveren cloud-løsninger som gir organisasjoner kontroll, datatasjeneste og unngåelse av vendor lock-in. Sjøl AI-infrastrukturen er i ferd med å demokratiseres: edge-datasentre nærmer seg brukerne, og «suveren AI» blir en reell forretningskritisk demografi. Leverandørene konkurrerer ikke lenger primært på skala, men på suverenitet, kostnadsoptimalisering og fleksibilitet. Denne sektor modnes raskt drevet av regulering (GDPR, EU AI Act), geopolitiske spenninger og desperat behov for bedre styring av AI-infrastruktur kostnader.

2. Topp 5 prosjekter / selskaper å følge

Navn	Type	Momentum	Hvorfor interessant akkurat nå
CoreWeave	Startup / GPU Cloud	🔥 Varmt	Nylig børsnotering (mars 2025), \$2.6B gjeldfinansiering, \$6.5B avtale med OpenAI utvidet. Har 250k GPUs og er den første til å deploye NVIDIA Rubin i 2026. Ekspanderer med 5 GW AI-fabrikker.
Databricks	Startup / Multi-Cloud Platform	🔥 Varmt	Serie L på \$5B ved \$134B verdsettelse (feb 2026). 65% YoY-vekst, \$5.4B revenue run-rate. Lanserer Lakebase som «system of record» for AI. Fokuserer på multi-cloud agentisk infrastruktur.
CNCF / Kubernetes-økosystemet	OSS / Standard	📈 Stigende	15.6 millioner cloud-native utviklere globalt. KubeCon Europe 2026 (Amsterdam, mars) fokuserer på multi-cloud, AI-integrasjon og kostnadsoptimalisering. Kubernetes er standard for hybrid/multi-cloud.
Ray + vLLM (PyTorch Foundation)	OSS / Distributed AI	📈 Stigende	Ray overføres til PyTorch Foundation, dannet en unified open-source stack med PyTorch + Ray + vLLM. Nye Ray Serve LLM APIs og ray symmetric-run for multi-node LLM serving. Kritisk for distribuert AI across multi-cloud.
Proxmox	OSS / Hypervisor	📈 Stigende	Lanserer Datacenter Manager 1.0 (des 2025) for sentralisert admin av multiple Proxmox-miljøer. Styrker posisjonen som åpen alternativ til VMware/Nutanix. Vinner terreng grunnet VMware-migrering.

3. Nye OSS-prosjekter verdt å følge (< 12 måneder)

Prosjekt	Kjernefunksjon	GitHub-stjerner	Hvorfor spennende
LiteLLM	Unified LLM proxy med multi-provider routing	~13k	Abstraherer 100+ LLM-leverandører bak OpenAI API. Enabler eksperimentering med hybrid local/cloud LLMs uten applikasjons-endring. Kritisk for multi-cloud AI.
Ollama	Local LLM runtime	~80k	Democratizes edge AI ved å kjøre open LLMs lokalt (Llama 3, Mistral, Phi). Fungerer sammen med LiteLLM for managed routing. Høy adoption blant edge-infrastruktur entusiaster.
LocalAI	Comprehensive local AI stack	~25k	Beyond LLMs: multimodal support (tekst, bilde, lyd), autonomous agents (LocalAGI), semantic search (LocalRecall), P2P distribuert inference. Sterkeste kandidat for «offline-first» enterprise AI.

4. Viktige hendelser siste kvartal

- **Des 2025:** Databricks kunngjør Serie L på \$5B ved \$134B verdsettelse; CoreWeave melder \$2.6B gjeldfinansiering.
- **Jan 2026:** NVIDIA investerer \$2B i CoreWeave for å accelerere 5 GW AI-fabrikker. Proxmox lanserer Datacenter Manager 1.0.
- **Feb 2026:** Ray-prosjektet offisielt overtas av PyTorch Foundation; dannes unified stack med vLLM og PyTorch.
- **Mar 2026 (nå):** EU kunngjør EuroStack-initiativ som alternativ til Gaia-X. KubeCon Europe starter 23–26 mars i Amsterdam med fokus på multi-cloud og AI.
- **Aug 2026 (kommende):** EU AI Act full enforcement for high-risk systems begynner; regulatorer forventer dokumentert governance.

5. Hype vs. substans

Teknologi / begrep	Hype-nivå (1-5)	Substans-nivå (1-5)	Vurdering
Suveren cloud / Digital sovereignty	4	4	Reell substans: regulering (GDPR, AI Act), geopolitikk driver etterspørsel. Gaia-X slo an, men EuroStack viser markedet vil differentiated løsninger.
Multi-cloud og «cloud arbitrage»	5	4	Høyt hypet, solid grunnlag. 77% av backend-devs bruker cloud-native tech. Automatisert «shopping» av compute across clouds er reelt nå med verktøy som Databricks.
Edge AI og distributed LLMs	4	3	Stor hype rundt inferens på edge. Substans er der for inference (Ollama, vLLM), men training distribuert across edge er fremdeles hard. Vekst på 30–35% årlig.
AI Agentic infrastructure (agent bricks, autonomous systems)	5	2	Ekstrem hype. Databricks snakker «agent bricks», men faktisk kapabilitet for multi-agent coordination across clouds er immature. Venter på 2026-2027 modning.
GPU cloud og «AI factories»	4	5	Høy hype, men fundamental realitet. NVIDIA, Microsoft, Meta investerer hundre mrd i GPU-

			infrastruktur. CoreWeave/Anyscale gjør det reelt. Flaskehals er strøm og kjøling, ikke teknologi.
Zero-trust og hybrid cloud governance	3	4	Moderat hype, solid substans. DORA og EU Data Act tvinger implementering. Likevel kompleks og ofte sub-optimal i praksis.

6. Hvor beveger de kloke hodene seg?

De beste ingeniørene migrerer mot edge-compute, distributed AI, og suveren cloud-løsninger. CoreWeave, Databricks og Anyscale rekrutterer aggressivt fra Google Cloud og AWS. Ray-prosjektets move til PyTorch Foundation signaliserer at open-source distributed AI er der talenten og ressursene går. CNCF-økosystemet (Kubernetes, service mesh, observability) absorberer ingeniørene som flykter VMware-monopolet.

Akademia fokuserer på federated learning, differential privacy, og edge inference — fundament for suveren AI. Berkeley, CMU og Mila (Montreal) fører forskningen innen distributed systems for AI. Gleichzeitig trekker store AI-labs (Meta, OpenAI, Anthropic) talenter vekk fra tradisjonelle cloud-leverandører.

Tegn til talent-drain fra AWS er rar; AWS AI-divisjon styrkes. Microsoft ser migrering på hybrid-fronten da Azure Stack får prioritet. GCP sliter litt på hiring for multi-cloud, men sterkt på open-source (Kubernetes, TensorFlow).

7. Risikofaktorer for sektoren

- **Regulatorisk:** EU AI Act enforcement fra august 2026 kan blokkere eller bremse cloud-tjenester dersom de ikke oppfyller documentation og governance krav. GDPR-tolking av «AI data subjects» og model rights fortsatt uklart.
- **Teknisk:** Multi-cloud-operasjoner er operasjonelt komplekst; vendor-lock-in byttes med operational fragmentation. Dataoverføring på tvers av clouds er en flaskehals for latency-sensitive AI workloads.
- **Markedsmessig:** GPU-mangel og strømkappinger kan hindre skalering av AI-fabrikker. Fornylsesenergi og infrastruktur kan ikke følge etterspørsel før 2027-2028.
- **Avhengighetsrisiko:** Suveren cloud-løsninger avhenger av NVIDIA-GPUs som er primært fabrikert i Taiwan. Geopolitisk usikkerhet rundt chip-eksport kan splitte infrastruktur-markedet.
- **Økonomisk:** «AI-fabrikk»-modellen (5+ GW datasenter) krever titalls miliarer i kapital. Gjeldfinansiering (CoreWeave-modellen) er volatil; rentetøkninger kan påvirke lønnsomhet.

8. Radar-sammendrag

Inspirert av ThoughtWorks Technology Radar.

ADOPT — bruk nå

- Kubernetes for multi-cloud orchestration
- Ray + vLLM for distributed AI inference
- Hybrid cloud (mix av on-prem og public cloud)
- Zero-trust security for cloud governance
- Observability tools (logging, tracing) for multi-cloud

TRY — eksperimenter

- Ollama / LiteLLM for hybrid LLM routing
- Databricks for multi-cloud AI data platform
- Edge AI inference at the network edge
- Open-source sovereign cloud (OVHcloud, Hetzner)
- Proxmox Datacenter Manager for private cloud

ASSESS — følg med

- LocalAI og andre offline-first AI stacks
- Federated learning for privacy-preserving AI
- GPU-as-a-service commodity market (CoreWeave model)
- EU AI Act enforcement og compliance tooling
- Decentralized / P2P cloud infrastructure

HOLD — avvent

- AI-agent orchestration frameworks (too immature)
- Gaia-X as primary cloud strategy (use EuroStack)
- Blockchain-basert cloud governance
- Proprietary «AI-specific» databases (immature)
- Multi-region active-active setups (operationally fragile)

9. Anbefaling — 3 ting å gjøre de neste 30 dagene

1. **Kartlegg dine cloud-avhengigheter:** Lag en oversikt over hvilke tjenester som kjører hvor (AWS, Azure, GCP, on-prem), data-flows mellom dem, og potensielle vendor lock-in-risikoer. Prioriter high-impact workloads. Bruk denne som baseline for multi-cloud-strategi.
2. **Eksperimenter med Ray + vLLM:** Sett opp en test med open-source LLM (f.eks. Ollama) og Ray Serve for distribuert inference. Se hvordan latency og kostnader endres. Dette er fremtiden for AI-infrastruktur — bedre du forstår det nå.
3. **Audit mot EU AI Act:** Dokumenter hvilke AI-systemer dine som klassifiseres som high-risk under EU AI Act (enforcement august 2026). Møt et team og begynn governance-planen. Bedrifter som starter nå har 5 måneders forsprang.