

Infraestrutura híbrida com banco de dados vetorial: relato de experiência no projeto SB100.

Melissa Ribeiro Romão Dias

Faculdade de Tecnologia (FATEC) Cotia – Centro Paula Souza
melissa.dias@aluno.cps.sp.gov.br

Silvio Fernando Barbieri

Faculdade de Tecnologia (FATEC) Cotia – Centro Paula Souza
silvio.barbieri@cps.sp.gov.br

Resumo

Objetivos: A participação na equipe de banco de dados e infraestrutura computacional do projeto SB100 evidenciou a necessidade de centralizar serviços e dados em um servidor compartilhado. A migração revelou limitações de capacidade computacional frente às demandas das equipes, motivando um estudo de viabilidade técnica e econômica entre infraestrutura local e ambientes em nuvem, com foco na governança de recursos e dados. **Métodos:** Foram levantados os requisitos das equipes e avaliados custo, manutenibilidade, escalabilidade, segurança, backup e suporte, além de aspectos de governança e controle de acesso. **Resultados:** A arquitetura híbrida obteve a maior pontuação na matriz de decisão, combinando servidor local para persistência e bancos de dados vetoriais com recursos em nuvem para demandas intensivas. **Conclusão:** A solução apresentou o melhor equilíbrio para a fase analisada, ampliando a experiência da monitoria para um contexto interdisciplinar de infraestrutura e dados.

Palavras-Chave

Banco de Dados; Banco de Dados Vetorial; Infraestrutura Computacional; Computação em Nuvem; Arquitetura Híbrida.

1. Introdução

Projetos de inteligência artificial e recuperação semântica dependem da integração entre serviços de dados, recursos computacionais e gestão da infraestrutura. Bancos de dados vetoriais permitem armazenar, indexar e recuperar representações de alta dimensionalidade, apoiando aplicações de busca semântica e modelos de linguagem (PAN; WANG; LI, 2024; XIE; LIU; YU, 2023). O Qdrant é uma solução de busca vetorial que permite armazenar e consultar vetores associados a dados e metadados (QDRANT, 2026).

A organização desses serviços envolve controle de versões, usuários, permissões, armazenamento, processamento e comunicação entre sistemas. A comparação entre infraestrutura local, nuvem e soluções híbridas deve considerar elasticidade, compartilhamento de recursos e mensuração do uso (MELL; GRANCE, 2011), além de critérios de qualidade como desempenho, segurança, manutenibilidade e flexibilidade (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2023).

Assim, o objetivo deste relato é apresentar as atividades desenvolvidas na monitoria e no projeto CCD-SB100, incluindo a adoção do Qdrant, a migração para o servidor SB100 e a análise técnica e econômica de alternativas de infraestrutura.

2. Metodologia

O relato foi elaborado com base nas atividades realizadas durante a monitoria e a participação no projeto CCD-SB100, utilizando registros técnicos e resultados produzidos durante o trabalho. Inicialmente, foram levantadas as necessidades relacionadas aos serviços de dados e à infraestrutura compartilhada. O Qdrant foi adotado como banco de dados vetorial principal, e atividades antes realizadas em ambientes locais foram transferidas para o servidor *on-premise* SB100. Nessa etapa, foram organizados o ambiente Linux, o versionamento com Git e GitHub, os repositórios e o gerenciamento de usuários e permissões (GITHUB, 2026).

Em seguida, foram identificadas as necessidades computacionais das equipes, considerando armazenamento, memória, processamento, uso de GPU, escalabilidade e comunicação entre serviços. Foram definidos quatro cenários: hospedagem integral em nuvem, servidor existente com upgrades, solução híbrida e novo servidor dedicado. A comparação utilizou uma matriz de decisão com custo, manutenibilidade, escalabilidade, segurança, backup e garantia e suporte. Os pesos foram definidos segundo as prioridades observadas no projeto: custo recebeu maior peso por causa das restrições orçamentárias; backup recebeu peso intermediário pela necessidade de preservação dos dados; segurança e suporte receberam peso 2; e manutenibilidade e escalabilidade receberam peso 1. Também foram estimados custos para aproximadamente 100 horas mensais, com as configurações de GPU consultadas em cada provedor, sem pressupor equivalência de hardware.

3. Resultados e discussão

A centralização dos serviços do servidor SB100 foi útil para organizar os recursos compartilhados entre as equipes. Isso ajudou a identificar limitações que não eram evidentes antes, quando as atividades estavam distribuídas em ambientes locais. O servidor começou a concentrar os serviços que eram compatíveis com sua capacidade, enquanto as demandas de processamento mais intensivas exigiam alternativas de expansão.

Foi criado um fluxo colaborativo com Git e GitHub (GITHUB, 2026), que inclui repositórios e *branches* para as etapas de desenvolvimento, homologação e produção. A organização dos usuários por equipe e a definição de permissões contribuíram para a rastreabilidade e a segurança do ambiente. O levantamento feito com as equipes mostrou necessidades distintas: enquanto a Equipe 3 precisava de recursos para armazenamento e recuperação dos vetores do Qdrant, outras equipes precisavam de GPU para vetorização, execução de modelos de linguagem e agentes inteligentes.

Diante das necessidades identificadas pelas equipes, foi feita uma análise comparativa de quatro cenários de infraestrutura. Os pesos atribuídos aos critérios da matriz de decisão foram definidos durante a avaliação das alternativas de infraestrutura, considerando a orientação adotada no pro-

jeto. A pontuação final de cada cenário foi obtida pela ponderação dos critérios de custo, manutenibilidade, escalabilidade, segurança, backup e garantia e suporte. Os resultados da avaliação dos cenários estão apresentados na Tabela 1.

Critério	Peso	Cenário 1 Tudo Cloud	Cenário 2 Servidor + Upgrades	Cenário 3 Híbrido	Cenário 4 Novo Servidor
Custo	5	1 (5)	2 (10)	4 (20)	1 (5)
Manutenibilidade	1	5 (5)	2 (2)	3 (3)	3 (3)
Escalabilidade	1	5 (5)	1 (1)	4 (4)	4 (4)
Segurança	2	5 (10)	2 (4)	3 (6)	3 (6)
Backup	3	5 (15)	1 (3)	4 (12)	4 (12)
Garantia e Suporte	2	5 (10)	1 (2)	3 (6)	4 (8)
Total Ponderado	14	50 / 70	22 / 70	51 / 70	38 / 70

Tabela 1 – Pontuação dos cenários de infraestrutura avaliados para o projeto SB100

Fonte: Elaborado pela autora (2026).

O cenário híbrido foi o que obteve mais pontos, com 51 de 70, superando o cenário totalmente em nuvem, que teve 50 pontos, o novo servidor, com 38 pontos, e o servidor existente com upgrades, com 22 pontos. Isso indicou que a combinação entre infraestrutura local e computação em nuvem era a alternativa mais adequada para a fase analisada, principalmente por equilibrar custo, persistência de dados e capacidade de expansão.

Na solução híbrida, o servidor local continua a ser usado para funções como hospedagem do Qdrant, MongoDB, armazenamento de documentos e dados de curadoria e gerenciamento de repositórios internos. Já tarefas que exigem mais capacidade computacional, como vetorização, execução de *LLMs* e agentes inteligentes, podem utilizar recursos de nuvem sob demanda. Para comparar as alternativas de provedores de nuvem diante das demandas de processamento identificadas, foram consideradas as configurações apresentadas na Tabela 2.

Solução	Configuração	Valor estimado (100 h/mês)
Google Cloud	g2-standard-4: 4 vCPU, 16 GiB RAM, 1 NVIDIA L4 (24 GB)	~US\$ 92,44
AWS	g5.xlarge: 4 vCPU, 16 GiB RAM, 1 NVIDIA A10G (22 GiB)	~US\$ 102,03
Microsoft Azure	Standard_NC24ads_A100_v4: 24 vCPU, 220 GiB RAM, 1 NVIDIA A100 (80 GB)	~US\$ 387,01

Tabela 2 – Comparação de configurações de computação em nuvem

Fonte: Elaborada pela autora com base em AWS (2026), Google Cloud (2026) e Microsoft Azure (2026).

As estimativas feitas para cerca de 100 horas mensais de uso mostraram valores aproximados de US\$ 92,44 para Google Cloud, US\$ 102,03 para AWS e US\$ 387,01 para Microsoft Azure. A configuração do Azure utiliza uma VM NC24ads A100 v4, com 24 vCPUs, 220 GiB de memória e uma GPU A100 de 80 GB, portanto possui escala de recursos diferente das configurações Google e AWS. Essa opção foi mantida como referência de uma alternativa de alto desempenho considerada no levantamento; assim, os valores não devem ser interpretados como um benchmark entre GPUs equivalentes, mas como uma comparação de custos de configurações efetivamente consultadas. O estudo, portanto, envolveu não apenas a escolha de tecnologias de banco de dados, mas também a infraestrutura necessária para disponibilizar e sustentar os serviços.

4. Considerações finais

A experiência de trabalhar como monitora de Banco de Dados, juntamente com a participação no projeto CCD-SB100, permitiu aplicar conhecimentos acadêmicos em uma situação real e interdisciplinar. A centralização dos serviços no servidor SB100 evidenciou limitações de capacidade e levou ao levantamento das necessidades das equipes e à análise de alternativas de infraestrutura.

A comparação demonstrou que a arquitetura híbrida apresentou a maior pontuação na matriz de decisão para a fase analisada, equilibrando custo, capacidade, escalabilidade, segurança, backup e suporte. A experiência também ampliou os conhecimentos sobre Linux, Git, GitHub, gerenciamento de acessos e computação em nuvem, evidenciando que a sustentação dos serviços de dados depende tanto das tecnologias de banco de dados quanto da infraestrutura que as suporta.

5. Limitações e Isenção de Responsabilidade

Este trabalho cita provedores de serviços em nuvem com finalidade exclusivamente técnica e ilustrativa, no contexto da comparação de arquiteturas realizada. Os autores declaram não possuir vínculo profissional, comercial ou financeiro com os provedores mencionados, não havendo conflito de interesses na condução ou apresentação dos resultados.

Os resultados refletem as condições específicas do projeto SB100 — requisitos técnicos, restrições orçamentárias e demandas particulares das squads envolvidas — não devendo ser interpretados como regra genérica aplicável a outros contextos, configurações ou desenhos de projeto. A arquitetura ideal depende de fatores próprios de cada caso, sendo recomendável a condução de estudos de viabilidade específicos para outras aplicações.

Os autores não se responsabilizam por eventual uso, interpretação ou decisão técnica, operacional ou financeira tomada por terceiros com base nas informações aqui apresentadas. O conteúdo possui caráter acadêmico e informativo, cabendo a cada leitor avaliar sua aplicabilidade ao próprio contexto, preferencialmente com apoio de análise técnica especializada.

6. Referências

AMAZON WEB SERVICES. Amazon EC2 accelerated computing instance types: G5. Seattle: AWS, 2026. Disponível em: <https://docs.aws.amazon.com/ec2/latest/instancetypes/ac.html>. Acesso em: 27 ago. 2026.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. Centro Avançado de Pesquisa Smart B100 (CCD-SB100): auxílio à pesquisa 24/00985-1 — Agricultura digital, nutrição vegetal. São Paulo: FAPESP, 2024. Disponível em: <https://bv.fapesp.br/pt/auxilios/116236/centro-avancado-de-pesquisa-smart-b100-ccd-sb100/>. Acesso em: 15 ago. 2026.

GITHUB. Branches. [S. l.]: GitHub Docs, 2026. Disponível em: <https://docs.github.com/en/pull-requests/reference/branches>. Acesso em: 27 ago. 2026.

GOOGLE CLOUD. GPU machine types: G2 machine series. [S. l.]: Google Cloud Documentation, 2026. Disponível em: <https://docs.cloud.google.com/compute/docs/gpus>. Acesso em: 27 ago. 2026.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 25010:2023: systems and software engineering — systems and software quality requirements and evaluation (SQuaRE): product quality model. Geneva: ISO, 2023.

MELL, Peter; GRANCE, Timothy. The NIST definition of cloud computing. Gaithersburg: National Institute of Standards and Technology, 2011. (NIST Special Publication, 800-145). Disponível em: <https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-145.pdf>. Acesso em: 15 ago. 2026.

MICROSOFT. NC_A100_v4 size series: Azure Virtual Machines. Redmond: Microsoft Learn, 2026. Disponível em: <https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/nca100v4-series>. Acesso em: 27 ago. 2026.

PAN, James Jie; WANG, Jianguo; LI, Guoliang. Survey of vector database management systems. The VLDB Journal, [S. l.], v. 33, p. 1597–1621, 2024. DOI: 10.1007/s00778-024-00864-5.

QDRANT. Documentation. [S. l.]: Qdrant, 2026. Disponível em: <https://qdrant.tech/documentation/>. Acesso em: 27 ago. 2026.

XIE, Jiadong; LIU, Yingfan; YU, Jeffrey Xu. A comprehensive survey on vector database: storage and retrieval technique, challenge. arXiv, [S. l.], 2023. DOI: 10.48550/arXiv.2310.11703. Disponível em: <https://arxiv.org/abs/2310.11703>. Acesso em: 15 ago. 2026.