

Big Data em Astronomia – Desafios e Oportunidades para o Brasil

Relatório da Subcomissão de Big Data para o Plano Nacional de Astronomia Brasileiro (2026–2036)

Comissão Especial de Astronomia

Subcomissão de Big Data

Equipe de Trabalho:

Ana Leonor Chies Santiago Santos (IF-UFRGS)

Angela Cristina Krabbe (IAG-USP)

Claudia Lucia Mendes de Oliveira (IAG-USP)

Clécio Roque De Bom (CBPF)

Felipe Braga Ribas (IF-UTFPR)

Luiz Vitor de Souza Filho (IFSC-USP)

Reinaldo Roberto Rosa (INPE) - Coordenador

Roberto Kalbusch Saito (FSC-UFSC)

Sandro Barboza Rembold (DF-UFSM)

Sumário

Apresentação

Sumário Executivo

1. Cenário Atual de Big Data na Astronomia

- 1.1. Astronomia de Big Data: Conceitos e Tendências
- 1.2. Panorama Nacional: Levantamento da Subcomissão de Big Data

2. Perspectivas para a Próxima Década

- 2.1. Oportunidades e Desafios em Astronomia de Big Data
- 2.2. Identificação de Linhas de Pesquisa Promissoras

3. Questões Orçamentárias e de Fomento à P&D

- 3.1. Orçamentos Empenhados envolvendo Big Data em Astronomia
- 3.2. Estimativas para Investimentos em Big Data em Astronomia

4. Conclusão

- 4.1. Principais Diagnósticos e Recomendações
- 4.2. Big Data como Pilar da Astronomia Brasileira na Próxima Década

Apêndice A

Apêndice B

Referências

Glossário

Apresentação

Este relatório foi elaborado pela Subcomissão de Big Data como contribuição à Comissão Especial de Astronomia (CEA) instituída pelo Governo Federal, com o objetivo de apoiar a formulação do Plano Nacional de Astronomia (PNA) [1] para a década de 2026 a 2036. A proposta central é oferecer uma análise abrangente sobre os desafios e oportunidades relacionados ao uso intensivo de dados na astronomia brasileira, tendo em vista o papel crescente do Big Data como eixo estruturante da pesquisa científica contemporânea. Assim a astronomia brasileira, historicamente comprometida com projetos de relevância internacional, encontra-se diante de um novo ciclo marcado por demandas computacionais complexas, exigências de formação interdisciplinar e necessidade de infraestrutura de ponta. Nesse cenário, é fundamental fortalecer políticas públicas que reconheçam o valor estratégico da astronomia como vetor de inovação, formação científica e inserção global do conhecimento produzido no país.

Trata-se de um contexto onde a integração entre astronomia de ponta, Big Data, ciência de dados e inteligência artificial (IA) é hoje uma realidade irreversível. Projetos astronômicos modernos produzem volumes colossais de dados, cujo armazenamento, acesso e análise requerem o uso intensivo de infraestruturas de armazenamento, algoritmos avançados, aprendizado de máquina e plataformas computacionais de alta performance. A ciência de dados aplicada à astronomia permite a identificação automática de fenômenos cósmicos, classificação de objetos, detecção de padrões e antecipação de eventos, transformando radicalmente os métodos de investigação científica. Nesse novo paradigma, a IA e seus agentes correspondentes deixam de ser apenas uma ferramenta auxiliar e passam a ocupar um lugar central na produção do conhecimento astronômico contemporâneo.

Essa transformação, cujo grande protagonista é o Big Data, abrange desde o planejamento de observações até a análise automatizada de dados em larga escala, especialmente oriundos de levantamentos realizados por telescópios voltados para diferentes janelas do espectro eletromagnético. Nesse contexto, é estratégico que o Brasil preserve e amplie sua participação nas colaborações internacionais em que já possui algum protagonismo, ao mesmo tempo em que consolida novas parcerias relevantes para a comunidade astronômica brasileira. Além disso, é essencial investir no aprimoramento e utilização de equipamentos nacionais de pequeno e médio porte, alinhados às prioridades da comunidade astronômica, para ampliar a capacidade científica e tecnológica instalada no país. Tais investimentos, quando alinhados às prioridades científicas da comunidade astronômica brasileira, podem gerar avanços significativos em ciência

de dados e engenharia aplicada, consolidando uma base nacional robusta para o enfrentamento dos desafios científicos e tecnológicos da astronomia do século XXI.

Diante do cenário exposto acima, o relatório produzido por esta subcomissão descreve o panorama atual do Big Data na astronomia brasileira, apresenta tendências para os próximos anos, identifica linhas promissoras de pesquisa, analisa os investimentos realizados e os recursos projetados, e propõe direções que visam à integração, sinergia e sustentabilidade das iniciativas nacionais. O foco é contribuir com subsídios técnicos e estratégicos para o desenvolvimento de uma comunidade científica coesa, tecnicamente capacitada e geograficamente diversa, preparada para atuar, de forma colaborativa e com protagonismo, na próxima geração de descobertas astronômicas e na implementação bem-sucedida do Plano Nacional de Astronomia.

1. Big Data na Astronomia

1.1. Conceitos e Tendências

Os grandes projetos de observação astronômica atuais, tomando como exemplos dois grandes levantamentos, como o *Vera C. Rubin Observatory-LSST* e o *Square Kilometre Array* (SKA) foram projetados para gerar **petabytes** de dados por ano. Essa produção massiva de dados exige novas abordagens em armazenamento, processamento e análise de dados. Num contexto ainda mais amplo, o termo Big Data em astronomia refere-se às principais propriedades de uma quantidade massiva de dados gerados por telescópios e simulações, exigindo tecnologias avançadas para armazenar, processar e extrair informação científica [2-8]. O *Big Data* [7-9], frequentemente descrito pelos "6 Vs" (volume, velocidade, variedade, valor, veracidade e virtuosismo), impõe desafios únicos para os usuários e sistemas de TI tradicionais, exigindo novas abordagens para gerenciá-los e tratá-los com o objetivo de produzir informação e conhecimento de forma eficaz. Este novo paradigma exige uma sólida capacitação nos três pilares fundamentais da TI: *Hardware*, *Software* e Recursos Humanos ("*Peopleware*"). Devido às suas principais propriedades, ilustradas na Figura 1, a gestão eficiente do Big Data também requer uma abordagem estruturada baseada no *Ciclo de Vida dos Dados* que abrange desde a aquisição até a sua preservação total ou parcial de acordo com protocolos de uso e manutenção do armazenamento.

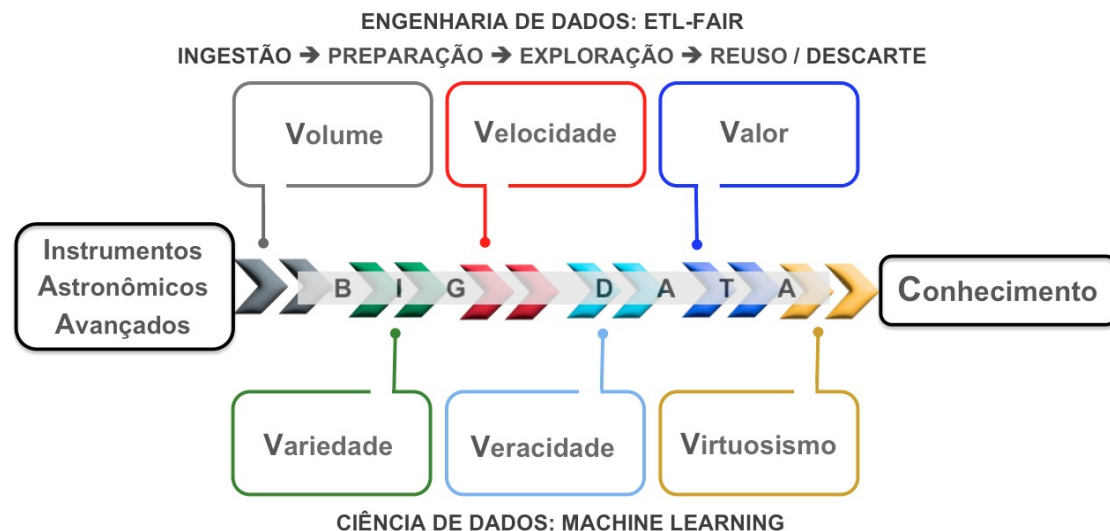


Figura 1 - Diagrama para a Astronomia na era do Big Data representando o fluxo conceitual da geração de conhecimento disruptivo a partir de instrumentos astronômicos avançados.

Principais Propriedades do Big Data em Astronomia:

Volume: Telescópios modernos podem gerar dezenas de **petabytes** de dados por ano, que podem ser armazenados em conjunto com tecnologias como *Data Cubes interativos* (ex. LSST Firefly), e *data lakes* que permitem consulta eficiente e cruzada de grandes volumes compostos por *arrays* multidimensionais.

Velocidade: Os fluxos de dados podem alcançar altas velocidades, exigindo recursos de processamento em tempo real. Em astronomia, altas taxas de produção de dados (ex. 1 Terabyte por noite) devem ser tratadas com agilidade, pois muitas vezes demandam processamento rápido para identificar fenômenos transientes, supernovas, asteroides, ou ainda aqueles relacionados ao MMA (*multimessenger astronomy*).

Variedade: Em geral, Big Data abrange vários tipos de dados, incluindo formatos estruturados, não estruturados e semiestruturados. Em astronomia, um mesmo instrumento pode gerar diferentes tipos e formatos de dados como imagens de alta resolução de campos profundos em altos redshifts, curvas de luz, espectros, e hipercubos dinâmicos derivados de simulações de alto desempenho.

Valor: Quando envolve extração ágil de informação relevante com qualidade a partir da mineração de dados (data mining) em Big Data.

Veracidade: A qualidade e a confiabilidade dos dados e metadados são cruciais para uma análise precisa, exigindo preparação, padronização e limpeza cuidadosas dos dados. Em astronomia, por exemplo, o **IVOA (International Virtual Observatory Alliance)** padroniza serviços e APIs para interoperabilidade com qualidade no padrão FAIR (*Findable, Accessible, Interoperable and Reusable*).

Virtuosismo: Refere-se à excelência em **infraestrutura tecnológica e humana**.

Hardware: data centers robustos, servidores HPC (*High Performance Computing*) e soluções de armazenamento em larga escala; **Software:** ferramentas otimizadas, pipelines paralelos e compatibilidade com infraestruturas modernas; **Peopleware:** equipes técnicas e científicas altamente capacitadas para operar, analisar e inovar com Big Data.

O fluxo conceitual, mostrado na Figura 1, conecta instrumentos astronômicos avançados (incluindo telescópios terrestres e espaciais, redes e *arrays* do século XXI) à geração de conhecimento científico disruptivo por meio da ciência e engenharia de dados. Nesse processo, os **6Vs** do Big Data atuam como eixos fundamentais para caracterizar e enfrentar os desafios tecnológicos e científicos. Assim, o **virtuosismo** acrescentado ao conjunto tradicional de 5Vs reconhece a importância do domínio técnico especializado em *hardware*, *software* e *peopleware*, essenciais para o sucesso da ciência e engenharia de dados em astronomia. Esses elementos estruturais do Big Data sustentam a capacidade de operar grandes sistemas de aquisição, transferência e processamento de dados em ambientes de alta complexidade, característica que evidencia a importância da capacitação de recursos para atuar no domínio do Big Data. Nesse domínio, o processamento dos dados astronômicos é estruturado em etapas alinhadas ao ciclo **ETL** (*Extract, Transform, Load*), abrangendo ingestão, preparação, exploração e, finalmente, reuso ou descarte controlado. Essas etapas estão integradas aos princípios **FAIR**, que asseguram a reutilização eficiente e segura dos dados pela comunidade científica, preservando seu valor científico ao longo do tempo. Em conjunto, ao longo desse fluxo, métodos avançados de ciência de dados e aprendizado de máquina (*machine learning*) desempenham papel central, permitindo identificar padrões complexos, classificar fenômenos e extrair informações de relevância científica a partir de grandes volumes de dados heterogêneos. O resultado é a conversão de dados brutos em conhecimento qualificado, capaz de impulsionar avanços disruptivos na compreensão do Universo.

Com base no conceito de Big Data em Astronomia, surgiram subáreas especializadas como a *astroinformática*, *astroestatística*, engenharia de dados astronômicos e ciência de dados baseada em IA e agentes, que fornecem as

bases computacionais e analíticas para extrair conhecimento de bases heterogêneas e complexas de dados da astronomia.

Um dos destaques recentes é o uso de *brokers* (sistemas que recebem fluxos completos de alertas, como os do LSST: cerca de 60 segundos após cada imagem, filtram, classificam, anotam e enviam aos usuários finais apenas os eventos relevantes), essenciais para filtrar e distribuir alertas de eventos astrofísicos em tempo real, exigindo infraestruturas robustas e agentes inteligentes adaptativos. Conceitos inovadores, como este, não apenas redefinem o papel do cientista na era digital, como também impõem novos desafios às instituições científicas em termos de formação, arquitetura de sistemas, interoperabilidade e acesso a plataformas internacionais de computação distribuída.

Tendências Inerentes ao Big Data em Astronomia:

Essas transformações estão associadas a diversas tendências tecnológicas emergentes motivadas pelo Big Data em astronomia, entre as quais se destacam, além dos *brokers*:

- [Pipelines científicos automatizados](#), que integram aquisição, calibração e análise de dados, reduzindo o tempo entre observação e descoberta;
- [VACs](#) (Value-Added Catalogs), catálogos enriquecidos que complementam os dados brutos de grandes levantamentos com parâmetros físicos e refinados extraídos por equipes especializadas e agentes;
- [Telescópios robóticos](#), operados remotamente e capazes de reagir a eventos em tempo real com base em agentes inteligentes;
- [Computação distribuída e nuvem científica](#), que possibilitam o processamento colaborativo de grandes volumes de dados de forma descentralizada;
- [Simulações numéricas massivas integradas a dados reais](#), que permitem validar modelos astrofísicos e cosmológicos com observações em tempo real;
- [Análise multibanda e multimodal](#), com integração de dados oriundos de diferentes faixas do espectro (óptico, rádio, raios-X etc.);
- [Ciência aberta e reprodutível](#), com uso crescente de ambientes FAIR, notebooks interativos e partilha de dados e códigos.

Essas tendências não apenas redefinem o papel do cientista na era digital, como também impõem novos desafios às instituições científicas em termos de formação, arquitetura de sistemas, interoperabilidade e acesso a plataformas internacionais de computação distribuída.

A próxima década será marcada por um crescimento exponencial do volume de dados em astronomia, impulsionado por uma nova geração de instrumentos científicos de alta resolução, capacidade e desempenho [8]. Telescópios terrestres como o GMT (Giant Magellan Telescope), ELT (Extremely Large Telescope), TMT (Thirty Meter Telescope), além de projetos de levantamento de dados massivos como o LSST (Legacy Survey of Space and Time) e o SKA (Square Kilometre Array), estão projetados para gerar dezenas a centenas de terabytes na escala diária. Não menos importantes na geração de dados, embora em menor quantidade, são também as missões espaciais como o James Webb Space Telescope (JWST) e o Euclid em termos de volume bruto de dados. Por exemplo, o TMT poderá atingir até 90 TB por noite de observação, enquanto o JWST transmite atualmente cerca de 0,24 TB por dia. Ainda que os telescópios espaciais contribuam com dados de altíssima qualidade e resolução, são os observatórios terrestres, com levantamentos contínuos e de larga escala, que estão se consolidando como os principais motores do Big Data na astronomia do século XXI.

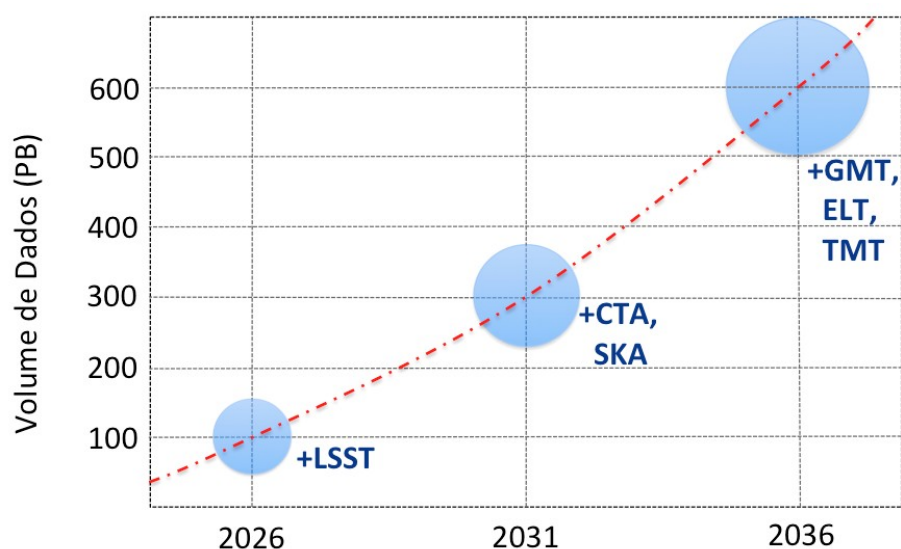


Figura 2 - Projeção do avanço da Astronomia do Big Data para o período de 2026 a 2036.

O crescimento da Astronomia do Big Data, mostrado na Figura 2, será impulsionado pelos grandes projetos astronômicos que estão entrando em operação, parcial ou plenamente, em 2026. Em 2036, o volume acumulado de dados científicos astronômicos ultrapassará a marca de 0.6 exabyte (1 EB = 1.000 petabytes) [3,7]. Essa magnitude consolida a astronomia como uma das áreas científicas mais dependentes de infraestrutura de dados em larga escala, exigindo políticas nacionais robustas de armazenamento, processamento e

acesso distribuído para ambos os casos: onde a participação brasileira já está consolidada ou venha a se consolidar.

1.2. Panorama Nacional: Levantamento da Subcomissão de Big Data

Este relatório incorpora de forma explícita o tema Big Data ao Plano Nacional de Astronomia (PNA). Portanto, trata-se de um conteúdo inédito em relação ao PNA anterior, no qual o Big Data não foi tratado como eixo estruturante. Nesse sentido, essa subcomissão realizou um levantamento inédito junto a uma centena de profissionais que atuam no Brasil em pesquisa, desenvolvimento e inovação em astronomia, com o objetivo de identificar o perfil da participação nacional neste eixo estruturante da Astronomia contemporânea que é o Big Data.

Considerando os critérios técnicos associados ao conceito de Big Data, adota-se aqui uma definição operacional para *Astronomia de Big Data* como sendo aquela cuja geração de dados brutos pode atingir a ordem de 1 PB **por ano**¹[6-9]. Esse limiar permite distinguir de forma objetiva os projetos que impõem demandas específicas de infraestrutura computacional para Big Data, gerenciamento automatizado de dados, e suporte especializado em ciência de dados (Ver Apêndice A). Além do volume, é importante ressaltar que os demais Vs também estão presentes nesse contexto, uma vez que esses dados são altamente variados (imagens, espectros, catálogos derivados), produzidos com grande rapidez (ex. alertas em tempo real), e requerem curadoria, integração, valor científico e excelência em TI para extração de conhecimento. Com esse foco, a Subcomissão de Big Data realizou um mapeamento junto à comunidade astronômica brasileira que revelou um ecossistema ativo, embora fragmentado, de participações, colaborações e iniciativas relacionadas ao acesso e uso de grandes volumes de dados, em parte, já disponibilizados e a serem produzidos nos próximos dez anos pelos instrumentos avançados da astronomia mundial.

¹Equivale à produção de cerca de 100 milhões de fotos 4K de 10 MB, as quais, quando armazenadas em discos de 5 TB, demandam, em dez anos, 2.000 unidades, formando uma torre de 40 metros de altura.

INFRAESTRUTURA OBSERVACIONAL	FAIXA ESPECTRAL	PB / ANO (OPERAÇÃO PLENA - Dados Brutos e Produtos)	PROJETOS & INICIATIVAS (Ano de Início)	INSTITUIÇÕES BRASILEIRAS ENVOLVIDAS (2025)
BDA	RÁDIO	Não foram encontradas estimativas	PBDA-II (2024)	03
BINGO	RÁDIO	Não foram encontradas estimativas	Abdus (2023) Uirapuru (2018)	05
CTAO	GAMA	2.0 - 3.0	LST-BRA (2018) / MST-BRA (2022) / SST-BRA (2023)	13
DESI / M-KPNO	ÓPTICO	0.8 - 1.0	AI-SCoPE (2025) / LIneA Sci. Plat. (2020)	04
GMT / LCO	ÓPTICO / IR	4.0 - 5.0	GMT-Brasil (2020)	21
LSST / Obs. Vera Rubin	ÓPTICO	5.0 - 6.0	AI-SCoPE (2025) BLASST (2025) IDAC-BRASIL (2022) LIneA Sci. Plat. (2020)	26
T250+JPCam / OAJ	ÓPTICO	0.4 - 0.6	J-PAS (2023)	13
T80 / OAJ	ÓPTICO	0.01 - 0.03	J-PLUS (2015)	13
T80S / CTIO	ÓPTICO	0.03 - 0.05	AI-SCoPE (2025) S-PLUS (2017)	24

Tabela 1 - Síntese das infraestruturas observacionais identificadas no levantamento da sub-comissão de Big Data, destacando a participação de instituições brasileiras em projetos e iniciativas que envolvem grande volume e complexidade de dados.

A **Tabela 1** apresenta uma síntese das principais infraestruturas observacionais nas quais a comunidade brasileira mantém participação ativa no contexto do paradigma da Astronomia de **Big Data (ver Apêndice A)**. A tabela sintetiza, por infraestrutura observacional, a **janela espectral**, a **produção estimada (PB/ano, brutos + derivados)**, as **iniciativas com ano de início** e o **número de instituições brasileiras engajadas** em cada projeto. **Cores:** laranja ($\geq 0,5$ PB/ano), amarelo ($< 0,5$ PB/ano), cinza/TBD (ainda sem estimativa). A contagem de instituições corresponde a **engajamentos declarados** e pode incluir **sobreposição** entre projetos.

As infraestruturas observacionais são apresentadas em ordem alfabética, acompanhados da **faixa aproximada de produção anual de dados**, o que facilita uma leitura comparativa das **demandas de infraestrutura** e evidencia **oportunidades de P&D** associadas a cada projeto ou iniciativa. Essa organização permite identificar, de forma objetiva, quais iniciativas requerem **maior capacidade computacional**, **maior integração com redes internacionais** e **maior investimento em recursos humanos especializados**.

Big Data refere-se a conjuntos de dados extensos — caracterizados por grande volume, variedade, velocidade, veracidade e valor — que exigem arquiteturas e tecnologias escaláveis para armazenamento, gestão, processamento e análise eficientes [6]. No contexto astronômico, os levantamentos J-PLUS, J-PAS e S-PLUS podem acumular ~0,7 PB em um ano de operação; embora a taxa diária de aquisição não alcance sistematicamente 10TB/dia (referência comum à infraestruturas observacionais de grande porte como o LSST e SKA), o volume total e a complexidade dos produtos já impõem desafios típicos de ciência de dados em larga escala, justificando sua inclusão no domínio de Big Data em Astronomia. A relevância não se limita ao volume: há desafios substanciais na extração de valor (análise e inferência), na velocidade (plataformas de distribuição que atendem, em média, a centenas de milhares de requisições diárias), na variedade (integração de múltiplas fontes e formatos) e na veracidade (mitigação de ruídos e controle de qualidade). A Tabela 1 lista projetos que exibem várias dessas características.

Conforme detalhado no **Apêndice A**, a estimativa de produção em Astronomia — sobretudo em grandes levantamentos, não se restringe aos dados brutos: a estratégia observacional envolve **exposições repetidas** (profundidade e calibração), **metadados**, **quadros de calibração** e **mosaicos (*coadds*)**; os **produtos derivados** (catálogos, *coadds* finais, fotometria) podem alcançar volumes comparáveis ou superiores; e fatores operacionais — **compressão**, **overheads de pipeline**, **versionamento**, **replicações** e **backups** — ampliam substancialmente o armazenamento requerido. Diante desse ciclo de vida completo e de seus impactos em armazenamento, rede, processamento e governança, é necessário **adotar um limiar mínimo** para enquadramento em Astronomia de Big Data, **definido por critérios objetivos** (multiplicadores de derivados e backups, cadência e repetição de exposições, requisitos de reprocessamento e distribuição) especificados no Apêndice A.

Portanto, para estimar Big Data de forma realista, por exemplo, em J-PLUS e S-PLUS, consideramos todo o ciclo de vida: exposições repetidas (profundidade/calibração), metadados e QA (*Quality Insurance*), frames de calibração, mosaicos/*coadds*, produtos derivados (catálogos, fotometria), artefatos intermediários de pipeline, documentação/portais e réplicas/backups. Na prática, isso introduz um multiplicador global de ~5–10× sobre o volume bruto anual resultando nos valores apresentados na Tabela 1. Mesmo com esses aportes, ambos permanecem abaixo do limiar de 0,5 PB/ano. No entanto, a necessidade de velocidade e processamento em tempo real, diversidade de formatos e um ecossistema tecnológico avançado das respectivas plataformas servem de modelo para como ciência e Big Data se encontram na prática, principalmente no caso do S-PLUS, que é totalmente desenvolvido no Brasil.

É importante destacar que os projetos e iniciativas considerados na Tabela 1 são aqueles que ainda estarão ativos ao longo da próxima década e direcionados à criação de consórcios, centros de dados e plataformas computacionais, com participação brasileira, e destinados ao processamento e análise de grandes volumes de dados astronômicos.

Descrição de Projetos e Iniciativas

Em suma, a Astronomia de Big Data no Brasil apresenta atualmente um conjunto de projetos e iniciativas de destaque, em sua maioria concentrados na região Sudeste - mas que engloba participação de outras regiões, com destaque para o Sul - que refletem o seu avanço e o seu potencial de crescimento. A seguir, com base no levantamento realizado pela subcomissão, apresenta-se um detalhamento entre os projetos e iniciativas que se destacam no cenário da Astronomia de Big Data:

A) Projetos e Iniciativas envolvendo Centros de Dados e Plataformas para Big Data

(A1) AI-SCoPE (Artificial Intelligence for Smart Cosmic Exploration) [41] – Desenvolvido no **Centro Brasileiro de Pesquisas Físicas (CBPF)**, o AI-SCoPE é uma plataforma inovadora que funciona como interface para um *data lake*, desenvolvido no Laboratório de Inteligência Artificial do CBPF (LAB-IA/CBPF)[44] para a comunidade. O sistema foi projetado para lidar com o futuro recebimento de múltiplos petabytes de dados provenientes de múltiplos surveys tais como DESI (Dark Energy Spectroscopic Instrument) [18], LSST [12], L4 (La Silla Schmidt Southern Survey) [35], ZTF (Zwicky Transient Facility) [34] e 4MOST [36], com ênfase em catálogos públicos incorporando inteligência artificial para busca e cruzamento de catálogos, recuperação de imagens e requisições de processamento em tempo real. A meta é maximizar a eficiência dos usuários no acesso e uso dos dados. A plataforma também concentra desenvolvimento de Agentes de Inteligência Artificial que vão recomendar e alertar sobre possíveis observações para redes de telescópios, executar modelagem automatizada de eventos astrofísicos, em particular transientes rápidos e MMA.

(A2) BLASST (Brazilian Laboratory for Astronomical STudies) [38], uma iniciativa que envolve uma equipe multidisciplinar de profissionais com grande experiência em diversas áreas da astronomia, não se restringindo somente ao software e hardware. O BLASST funcionará de forma similar mas complementar ao AI-SCoPE no sentido de distribuição de dados, mas o diferencial será a criação de catálogos de valor agregado para uso geral da comunidade, assim como desenvolvimento de novos métodos, que usam o estado da arte atual em

software, LLMs (large language models) [9] e agentes, para multiplicar o impacto dos novos telescópios e grandes projetos. O BLASST conterá catálogos de diferentes levantamentos e em especial do LSST e possibilitará também a subtração de imagens do LSST em tempo real e fotometria forçada, além de complementar e expandir as capacidades que o AI-SCoPE já tem, com possibilidade de integração de inteligência artificial à análise dos dados do LSST e de outros levantamentos.

(A3) IDAC-BRA-LSST (*Independent Data Access Center – Brasil - LSST*), também chamado *Brazil Lite IDAC* [10], é um centro independente de acesso a dados do **LSST**, coordenado pelo **LineA – Laboratório Interinstitucional de e-Astronomia** [11], em parceria com o Observatório Vera C. Rubin [12]. Integrado a uma rede internacional de centros IDAC distribuídos em todos os continentes, o IDAC-BRA-LSST constitui uma contribuição *in-kind* do Brasil ao projeto, oferecendo uma infraestrutura dedicada ao armazenamento, gerenciamento e disponibilização de dados científicos do LSST. No modelo Lite IDAC, o centro brasileiro é responsável por hospedar e distribuir os **VACs derivados das imagens combinadas (co-added)**, além de produtos auxiliares como mapas de profundidade, footprints e catálogos de redshifts fotométricos, assegurando sua integração às plataformas globais de acesso do Rubin Observatory. Sua infraestrutura conta com servidores **PostgreSQL** de alto desempenho, sistemas de armazenamento de grande porte distribuídos e também conectados à rede com capacidade escalável até a ordem de 10 petabytes, clusters de processamento de mais de **500 núcleos de CPU** (HPE Apollo 2000) e suporte a ambientes modernos de análise, como **Kubernetes** e **Jupyter Notebooks**, complementados pelo acesso a recursos de computação de alto desempenho do **supercomputador Santos Dumont instalado no LNCC (Laboratório Nacional de Computação Científica)** [29], interligados por rede de **100 Gbps** para garantir a transferência eficiente de dados. O acesso a esses recursos é mediado pelo **LineA Science Platform** [42], uma infraestrutura que disponibiliza ambientes interativos e ferramentas especializadas, como **JupyterHub**, serviços de consulta via **TAP (Table Access Protocol)**, visualizadores de céu (*Sky Viewer*) e de objetos (*Target Viewer*), além de portais específicos para exploração de dados (ex.: SDSS Sky Server, LSST Photo-z Server, Portal de Ocultação Solar).

(A4) S-PLUS (Southern Photometric Local Universe Survey) - é um projeto internacional de Big data em Astronomia iniciado em 2015 [19], liderado por brasileiros, que atualmente atende o maior número de usuários (229 - de 52 instituições inscritos no wiki - e centenas de não inscritos) e é aberto a toda a comunidade. Seu objetivo é mapear 9 300 graus quadrados do céu do hemisfério

sul em 12 filtros fotométricos, produzindo imagens e catálogos de milhões de objetos. O projeto conta com um telescópio robótico dedicado de 86 cm, contando com 3 servidores e 5 estações de processamento de alta capacidade que compõem um data lake e que fornece diferentes tipos de dados como catálogos, imagens, documentações, serviços de processamento em tempo real e APIs. O splus.cloud [19] é usado por mais de 800 usuários em todo o mundo. Com 40 artigos publicados, 30 teses e dissertações defendidas, 5 data releases internacionais de imagens e catálogos que cobrem cerca de metade do céu do sul, o S-PLUS é uma preparação para a próxima etapa de dados que virão do LSST bem como um pathfinder para sistemas que serão associados como telescópios maiores, de tamanho 1.5-2m, como o T150. O S-PLUS é um projeto que exemplifica de forma clara o conceito de Big Data aplicado à Astronomia. Além dos volumes de dados brutos totais mostrados na Tabela 1, o caráter de Big Data do S-PLUS se revela em diferentes dimensões. Na velocidade, a plataforma de distribuição splus.cloud lida com uma média de centenas de milhares de requisições por dia e já registrou picos de até 4,5 milhões de acessos em apenas 24 horas, chegando a mais de 50 TB em um único dia. Além disso, parte das requisições envolve dados gerados em tempo real, como a criação de imagens coloridas a partir das observações em múltiplos filtros, exigindo processamento dinâmico e entrega imediata. Na variedade, o projeto precisa gerenciar não apenas imagens brutas e calibradas, mas também catálogos fotométricos, metadados e produtos derivados, cada qual com requisitos diferentes de armazenamento e acesso.

B) Projetos e Iniciativas envolvendo outras Infraestruturas

(B1) BINGO [13] e BDA (Brazilian Decimetric Array) [14] são projetos na faixa de rádio, cujos volumes e cadências de dados ainda não foram dimensionados, mas que poderão representar uma contribuição significativa para a radioastronomia de Big Data no Brasil, especialmente no mapeamento de grandes áreas do céu, na caracterização de estruturas em escalas cósmicas e nos estudos de transientes.

(B2) CTA (Cherenkov Telescope Array) [17] e demais observatórios dedicados à astrofísica de altas energias deverão gerar fluxos substanciais de dados de raios gama e radiação Cherenkov, com exigências específicas de análise em tempo real. Projetos do CTA, como LST, MST e SST contam com a participação brasileira in-kind [17].

(B3) J-PAS [20] e J-PLUS [21] são projetos de levantamentos ópticos conduzidos a partir de dois telescópios, o **Telescópio Javalambre Survey Telescope (T250)**, de 2,5 metros de diâmetro, e outro de 80 cm, localizados no Observatório Astrofísico de Javalambre (OAJ), Espanha. Esses levantamentos produzem catálogos fotométricos amplos e de alta qualidade, essenciais para complementar observações de grandes levantamentos internacionais. O volume de dados gerado é estimado em **10–15 terabytes por ano**, valor que, embora abaixo da escala de dezenas de petabytes de outros grandes projetos, representa um recurso estratégico para calibração cruzada e enriquecimento de catálogos no contexto do Big Data em astronomia.

(B4) Laboratório Nacional de Astrofísica (LNA) – O LNA, órgão do MCTI, mantém uma série de programas e instrumentos voltados à observação astronômica e ao monitoramento sistemático de fenômenos celestes, abrangendo desde astrometria e fotometria de precisão até espectroscopia e radioastronomia. Entre os ativos de maior destaque estão o **Southern Astrophysical Research Telescope (SOAR)**, localizado no Chile, e o **Observatório Gemini**, com telescópios de 8,1 m situados no Havaí (Gemini North) e no Chile (Gemini South). O Brasil, por meio do LNA, participa ativamente da operação e do uso científico desses telescópios, garantindo acesso a tempo de observação e integração com grandes levantamentos internacionais. Embora nem todos os programas associados a esses instrumentos operem na escala típica de **Big Data**, eles desempenham papel essencial na manutenção de séries temporais históricas, no desenvolvimento de instrumentação científica, na formação de recursos humanos e na criação de sinergias com projetos de maior escala de dados. Parte dessas atividades também possui potencial de integração com iniciativas multibanda e multimessageiras, com o BRICS Astronomy, reforçando a contribuição brasileira no ecossistema global de pesquisa astronômica.

(B5) GMT-Brasil - A iniciativa GMT-Brasil [15] representa a participação científica do Brasil no Telescópio Gigante Magalhães (Giant Magellan Telescope-GMT) [16], coordenada pelo IAG/USP e financiada pela FAPESP, com foco no desenvolvimento de instrumentação e na colaboração internacional em astronomia de ponta.

C) Projetos e Iniciativas envolvendo Grandes Grupos e Consórcios

(C1) BPG-LSST – Grupo de Participação Brasileiro (BPG) com acesso a dados restritos do LSST (data right holders) [10, 11]. Atualmente conta com cerca de 120 membros, sendo 24 Pesquisadores Principais (PI) tendo em seu grupo até 4 pesquisadores juniores (JAs), pós-docs, estudantes de pós-graduação ou graduação. Os pesquisadores associados têm sua atuação coordenada pelo LIneA em uma estrutura organizacional definida pelo LSST. Composto por pesquisadores de 22 instituições diferentes em 10 estados brasileiros, distribuídos em 8 colaborações científicas independentes: Galáxias; Estrelas, Via Láctea e Volume Local; Sistema Solar; Energia Escura; Núcleos Galácticos Ativos; Transientes e Estrelas Variáveis; Lentes Fortes; e Informática & Estatística. O BPG-LSST atuará através do IDAC-Brasil-LSST (Centro de Comissionamento Brasileiro do LSST), que é a infraestrutura brasileira de apoio à análise de dados do projeto, contando com o apoio de parceiros do LIneA como RNP, ON e LNCC.

(C2) BRICS Astronomy Working Group (BAWG) - No contexto multilateral, o Brasil é membro do BAWG [37] que coordena ações conjuntas em pesquisa, desenvolvimento tecnológico e formação de recursos humanos entre os países do bloco. Apoia iniciativas brasileiras no Square Kilometre Array (SKA) por meio de colaborações técnicas e industriais, como o programa SKA Connect, que integra indústria, universidades e centros de pesquisa para desenvolver soluções compatíveis com as demandas do observatório. No contexto multilateral, participa do BRICS Intelligent Telescope and Data Network (BITDN), que busca conectar telescópios e centros de dados por meio de uma rede inteligente para levantamentos coordenados e detecção de transientes, e a participação na iniciativa Open Universe [46] chancelada pela UNOOSA (ONU Office of Outer Space Affairs) [47], voltada para capacitação e ampliação ao acesso aberto e a usabilidade de dados científicos. Essas ações incluem ainda programas de capacitação, *hackathons* e treinamentos em análise de dados astronômicos, facilitando a inserção brasileira em redes internacionais de ciência e tecnologia envolvendo futuramente Big Data.

(C3) CLAST (Consortium for Large Astronomical Survey Tasks) [38] - Esse é um consórcio de mais de 40 membros, de 14 instituições brasileiras e instituições estrangeiras, reunindo astrônomos, físicos, estatísticos, engenheiros e cientistas da computação. Esse consórcio é responsável por planejar, financiar e coordenar o BLASST (Brazilian Laboratory for Astronomical Survey STudies). O consórcio tem por princípios organizacionais a definição de estratégias comuns, otimização da participação em grandes levantamentos e instrumentos e a maximização do impacto das pesquisas individuais, principalmente por meio da disseminação

entre a comunidade astronômica brasileira de técnicas avançadas de Inteligência Artificial (IA), incluindo a construção de plataformas de IA, ferramentas de Aprendizado de Máquina, incluindo agentes de IA e LLMs, para acelerar as pesquisas científicas análises de alto nível.

(C4) SPANET (São Paulo Astronomy Network) [39] - é uma aliança entre universidades, instituições e profissionais para impulsionar a Astronomia e o desenvolvimento da Instrumentação Astronômica e Ciência de Dados no estado de SP. A rede busca alcançar isso através da coordenação de ações que aumentem a visibilidade da produção brasileira no cenário nacional e internacional, promovam a educação científica na sociedade e impulsionem o desenvolvimento tecnológico e econômico. A SPAnet contribui ativamente para grandes projetos internacionais com foco na investigação do universo em todo o espectro eletromagnético e considerando iniciativas com tecnologias emergentes, abrangendo assim projetos de Big Data em Astronomia. Alguns exemplos de projetos relacionados à SPANET: BINGO, GMT, CTAO, J-PAS, J-PLUS e S-PLUS.

(C5) T80-Sul - Consórcio composto de 229 membros de 24 instituições brasileiras e outras 28 internacionais que utilizam o telescópio robótico brasileiro no Cerro Tololo Interamerican Observatory [40]. Iniciado em 2016, conta com 6 data releases abertos a toda a comunidade brasileira (sendo que 5 deles abertos ao mundo, ver splus.cloud). Organiza anualmente inúmeros eventos científicos e em especial escolas de IA.

D) Projetos e Iniciativas envolvendo Uso integrado de Base de Dados

D1) Uso de levantamentos e bases de dados públicos internacionais consolidadas – Muitos membros da comunidade astronômica brasileira utilizam extensivamente dados provenientes de grandes levantamentos públicos, como **Zwicky Transient Facility (ZTF)**, **Dark Energy Survey (DES)**, **Sloan Digital Sky Survey (SDSS)**, **Euclid** e **Prime Focus Spectrograph (PFS)**. Além disso, há intenso uso de dados de telescópios espaciais e de levantamentos de rádio voltados para cosmologia, incluindo projetos baseados em medições da **Radiação Cósmica de Fundo (CMB)**. Esses recursos, amplamente acessíveis, possibilitam que pesquisadores brasileiros desenvolvam investigações de ponta, muitas vezes em sinergia com dados obtidos em colaborações internacionais nas quais o Brasil já participa ativamente.

D2) Integrações Multibanda e Multimessageira – Essas linhas de pesquisa, e outras que aqui não foram mencionadas, fortalecem-se com a abordagem combinada da **astrofísica multibanda** e **multimessageira**, que integra

observações em múltiplos comprimentos de onda (óptico, infravermelho, rádio, raios gama) com mensageiros não eletromagnéticos, como **ondas gravitacionais** e **partículas de alta energia**. No Brasil, destacam-se grupos de pesquisa envolvidos em colaborações internacionais como o **LIGO Scientific Collaboration** e iniciativas correlatas voltadas para a detecção e análise de sinais de ondas gravitacionais, essenciais para a compreensão de eventos astrofísicos extremos. Essa integração potencializa respostas rápidas a alertas transientes, enriquecimento de VACs, priorização de segmentos robóticos e correlações estatísticas em larga escala, conectando iniciativas ópticas (J-PLUS/J-PAS/S-PLUS), de altas energias (CTA) e de rádio (BINGO, BDA, SKA), além do uso estratégico de bases de dados públicas internacionais.

D3) International Virtual Observatory Alliance (IVOA) [43] - Essa iniciativa continua sendo uma referência global para a padronização e interoperabilidade de dados astronômicos, viabilizando a integração de grandes volumes de informações coletadas por telescópios e missões distintas. Por meio de protocolos como TAP (Table Access Protocol), VOTable, SIA (Simple Image Access) e ConeSearch, o IVOA estabeleceu a infraestrutura que permite a cientistas consultar e cruzar catálogos, imagens e espectros em escala massiva, sem barreiras institucionais ou geográficas. Embora o termo *Virtual Observatory* tenha perdido força, sendo progressivamente substituído por abordagens mais amplas de astroinformática e pelos princípios de dados FAIR, sua base tecnológica permanece ativa e essencial. Projetos de grande porte como o LSST, GMT, SKA e Gaia continuam a adotar esses padrões, reforçando sua relevância. No caso da astronomia brasileira, a participação em colaborações internacionais torna fundamental a adesão às normas do IVOA, garantindo não apenas acesso pleno a dados globais, mas também a possibilidade de desenvolver ferramentas e serviços interoperáveis capazes de colocar o país em posição de destaque na astronomia de Big Data.

O conjunto de projetos e iniciativas descritos acima demonstra não apenas a capacidade técnica instalada no país, mas também a necessidade de fortalecer redes de colaboração e ampliar o investimento em infraestrutura e recursos humanos, de forma a assegurar que a comunidade astronômica brasileira possa atuar com protagonismo na era do Big Data.

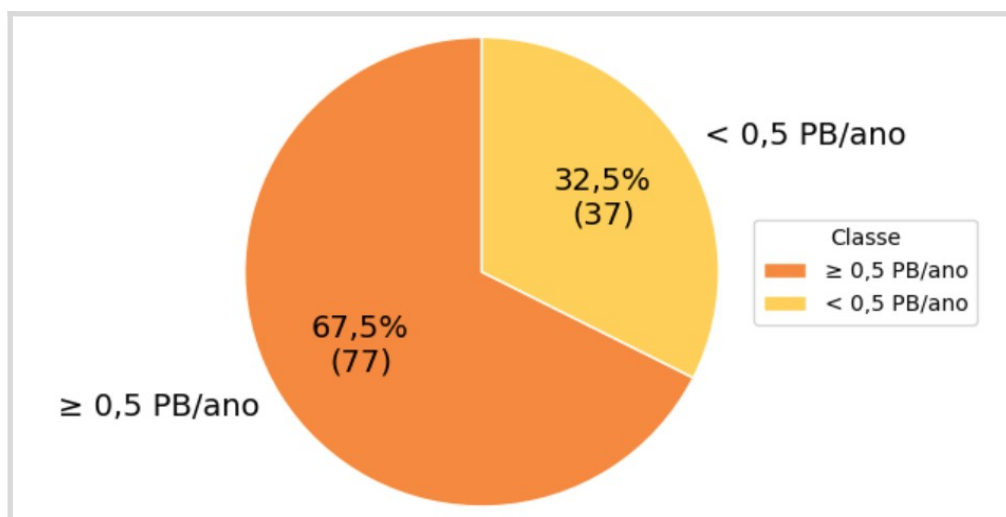


Figura 3 - Engajamento institucional em projetos e iniciativas envolvendo Astronomia de Big Data de acordo com os valores da Tabela 1.

O gráfico da Figura 3 apresenta a distribuição aproximada de instituições brasileiras, públicas e privadas, que atuam em astronomia por meio de projetos e iniciativas listados na Tabela 1, nos quais foram identificadas 114 instituições envolvidas com projetos e iniciativas cujos volumes de dados gerados são conhecidos. Os valores refletem a proporção de envolvimento institucional em cada caso, considerando tanto atividades de pesquisa e desenvolvimento quanto suporte técnico e operacional. Este resultado indica que, em 2025, cerca de 67,5% das instituições brasileiras possuem pesquisadores e grupos atuando em Astronomia de Big Data onde o volume de dados ultrapassa 0.5 PB/ano, evidenciando uma base significativa de engajamento e demanda por competência nacional nessa área estratégica.

2. Perspectivas para a Próxima Década

A astronomia brasileira expandiu-se significativamente nas últimas décadas e apresenta hoje uma comunidade diversa, com crescente inserção internacional e capacidade técnica reconhecida em diferentes áreas. A expectativa para os próximos dez anos é de consolidação e avanço em áreas estratégicas, alavancadas por novos levantamentos de dados e pela participação ativa em grandes colaborações globais. Este cenário será fortemente impulsionado pela adoção de abordagens integradas de astroinformática, ciência de dados e inteligência artificial, associadas a novas plataformas de observação em solo e no

espaço, geradoras de volumes sem precedentes de informação científica atingindo a ordem de 1 Exabyte de dados gerados até 2040.

2.1. Oportunidades e Desafios em Astronomia de Big Data

Entre as oportunidades estratégicas para a próxima década de Big Data em Astronomia, destacam-se as chamadas e acordos de colaboração institucional que possibilitem a participação efetiva de astrônomos brasileiros em P&D associadas aos grandes telescópios do século XXI, garantindo acesso antecipado aos dados ainda na fase de comissionamento e fomentando a elaboração de iniciativas e projetos que explorem plenamente o potencial do Big Data. Essas colaborações devem também abrir espaço para que o Brasil desenvolva e lidere a concepção de instrumentos complementares e soluções avançadas de tecnologia da informação voltadas à ciência e engenharia de dados astronômicos.

Com base nas tendências internacionais e no diagnóstico nacional apresentados nas seções anteriores, identificam-se ainda como oportunidades:

- (i) desenvolvimento e operação de *pipelines* científicos automatizados, integrando aquisição, calibração e análise de dados;
- (ii) criação e manutenção de *brokers* e VACs voltados a grandes levantamentos, com interoperabilidade internacional;
- (iii) implantação de *data centers* de alto desempenho e plataformas de computação distribuída alinhadas aos princípios FAIR (*Findable, Accessible, Interoperable, Reusable*);
- (iv) aplicação avançada de inteligência artificial e aprendizado de máquina para classificação de objetos, detecção de padrões e previsão de eventos astronômicos;
- (v) integração de dados multibanda e multimodais, abrangendo rádio, óptico, infravermelho e raios-X;

(vi) execução de simulações numéricas massivas integradas a dados observacionais;

(vii) fortalecimento de práticas de ciência aberta, com reprodutibilidade e compartilhamento de código e dados.

Entre os desafios, destacam-se a necessidade de expandir e modernizar a infraestrutura nacional de computação de alto desempenho, formar e reter quadros técnicos especializados em engenharia e ciência de dados astronômicos, e estabelecer estratégias coordenadas para integração entre centros de pesquisa e acesso a dados internacionais.

2.3. Identificação das Linhas de Pesquisa Promissoras

Dentre as principais linhas de pesquisa em astrofísica com relevância para o paradigma de Big Data na próxima década, destacam-se:

2.3.1. Cosmologia e Estrutura em Larga Escala: Investigação de tensões cosmológicas fundamentais, como a Constante de Hubble (H_0) e a Amplitude de Flutuações de Matéria (S_8); estudos sobre matéria escura e energia escura, utilizando levantamentos como LSST, DES, Euclid e CMB-S4; análise de alinhamentos intrínsecos e lentes gravitacionais fracas, combinando observações e modelagem estatística avançada.

2.3.2. Formação e Evolução de Galáxias: Classificação morfológica de galáxias com redes neurais profundas e agentes autônomos para detecção de anomalias; estudos da evolução química e do conteúdo de poeira com dados multibanda (S-PLUS, J-PLUS, J-PAS); investigação de galáxias em aglomerados e estruturas cósmicas de grande escala; análise de objetos raros e núcleos ativos variáveis (AGNs).

2.3.3. Astrofísica Estelar e de Exoplanetas: Estudos de estrelas Be, variáveis cataclísmicas, sistemas binários e objetos compactos; caracterização de atmosferas de exoplanetas gigantes com dados de missões como Roman e JWST; arqueologia estelar e dinâmica galáctica com dados do Gaia, THEIA, S-PLUS e levantamentos correlatos.

2.3.4. Transientes e Fenômenos Anômalos: Aplicação de *machine learning* e uso de *brokers* e *plataforma de recomendação de follow-up* (ex.: FINK, AI-SCope) para detecção e classificação automática em tempo quase real; Identificação de supernovas raras, kilonovas, transientes rápidos, eventos cataclísmicos e microlentes usando surveys como LSST e ZTF.

2.3.5. Sistema Solar e Pequenos Corpos: Predição e análise de ocultações estelares, com potencial de bilhões de eventos anuais a partir do LSST; estudos sobre propriedades de corpos menores (objetos binários, famílias, cores, curvas de luz) e objetos interplanetários.

2.3.6. Integrações Multibanda e Multimessageira: Combinação de observações em diferentes faixas do espectro (óptico, infravermelho, rádio, raios gama) com detecção de ondas gravitacionais e partículas de alta energia; participação brasileira em colaborações como LIGO e detectores correlatos.

2.3.7. Infraestrutura e Formação em Astroinformática (*linha transversal a todas as anteriores*): Desenvolvimento de infraestrutura para armazenamento e processamento de grandes volumes de dados (Big Data); criação de *pipelines* automatizados, plataformas interoperáveis e bancos de dados abertos; formação de profissionais especializados em ciência de dados aplicada à astronomia, com competências em HPC, IA e engenharia de software científico.

3. Questões Orçamentárias e de Fomento à P&D

Os investimentos em Big Data em Astronomia devem ser estruturados com critérios que assegurem infraestrutura robusta, capacitação contínua em engenharia de dados e astroinformática e retenção de recursos humanos especializados. A curto e médio prazo, é essencial investir no desenvolvimento de portfólios de serviços disponibilizados em plataformas abertas para a comunidade astronômica, incluindo curadoria e atualizações de catálogos legados e novos. A manutenção e criação de centros de e-astronomia requer reforço contínuo em tecnologia de bancos de dados, armazenamento e processamento, com integração de soluções em nuvem. É estratégico fortalecer a infraestrutura de rede por meio da RNP (Rede Nacional de Ensino e Pesquisa) [28] e demais redes, prevendo conexões escaláveis a partir de 100 Gbps, implantação de roteadores de última

geração com DMZ (Demilitarized Zone), switches e transceivers modernos. Com base em benchmarks internacionais, o investimento mínimo anual estimado para infraestrutura de data centers (“AI-ready”) gira em torno de US\$ 5–10 milhões, enquanto aprimoramento de rede demanda de US\$ 1–5 milhões por ano, totalizando uma estimativa mínima média de US\$ 10 milhões por ano.

Rede / Provedor	Capacidade Típica (Gbps)	Estimativa de Custo Anual (USD)	Aplicabilidade na Astronomia de Big Data
RNP	10–100 (> 400)	0 (custos cobertos por verbas públicas)	Principal backbone nacional para tráfego de dados científicos
Rednosp	100–400	0 (custos cobertos por verbas públicas)	Principal backbone para tráfego de dados científicos entre o Estado de São paulo e Europa e EUA.
AmLight	100–400	0 (custos cobertos por convênios e ag. internacionais)	Transferência de dados de telescópios no Chile e África do Sul
RedCLARA	10–100	Até ~50 mil (taxa de consórcio e manutenção mínima)	Cooperação latino-americana e acesso à Europa
GÉANT / Internet2	100–400	0 (custos cobertos por consórcios internacionais)	Interligação com supercomputadores e centros de dados globais
Backbones Privados (Claro, Vivo, TIM, Algar)	10–400	500 mil – 3 milhões	Rotas alternativas ou projetos com grande volume interno
Cloud Providers (AWS, Google Cloud, Azure)	10–100	200 mil – 2 milhões	Armazenamento, processamento e distribuição de dados

Tabela 2 - RNP e Redes Alternativas para Comunicação de Grande Volume de Dados.

É fundamental assegurar, no Brasil, uma infraestrutura de rede de alta capacidade e alta disponibilidade, interligada às redes acadêmicas internacionais — condição estratégica para a participação nacional em grandes observatórios. No caso específico dos telescópios instalados no Chile, a continuidade de uma malha nacional de dados com conectividade internacional robusta pode gerar vantagens decisivas para a comunidade científica brasileira. Para atingir esse objetivo, é necessário investir de forma sustentada na RNP e nas redes estaduais e institucionais, bem como manter e ampliar os contratos de interconexão com as redes globais. A Tabela 2 apresenta um comparativo entre a RNP e redes equivalentes que podem funcionar como canais adicionais para aumentar a resiliência e reduzir gargalos no fluxo do Big Data em astronomia.

3.1. Orçamentos Empenhados envolvendo Big Data em

Astronomia

Nesta subseção são apresentados, sempre que disponíveis, os valores já empenhados em projetos e iniciativas estratégicas mencionados neste relatório. A ênfase recai sobre plataformas, centros de dados e instrumentos que compõem a espinha dorsal da atuação brasileira no paradigma de Big Data em Astronomia.

As três principais iniciativas nacionais voltadas ao **LSST (AI-SCoPE, BLASST e IDAC-Brasil)** representam investimentos expressivos e complementares no ecossistema de dados astronômicos:

i) AI-SCoPE (Artificial Intelligence for Smart Cosmic Exploration): A referida plataforma para análise massiva de dados em múltiplos surveys, recebeu investimentos através de projetos FINEP para o Laboratório de Inteligência Artificial na Física do CBPF de aproximadamente R\$ 24 milhões. Esse investimento envolve infraestrutura e pessoal especializado para dar suporte às demandas de hardware e software do Laboratório.

ii) BLASST (Brazilian Lab of Astronomical Survey Studies): Iniciativa multidisciplinar voltada à criação de **catálogos de valor agregado (VACs)**, subtração de imagens em tempo real e fotometria forçada, além do desenvolvimento de novos métodos baseados em *Large Language Models (LLMs)* e agentes inteligentes. Complementa as funções do AI-SCoPE e busca multiplicar o impacto dos grandes levantamentos. A criação deste laboratório só foi possível graças ao financiamento de mais de **US\$ 25 milhões** já aplicados pela FAPESP nos últimos 10 anos em fibras ópticas para comunicação da comunidade científica e especialmente para o LSST. Em adição, em 2025 a FAPESP investiu **R\$ 576 mil**, concedidos diretamente aos pesquisadores do Estado de São Paulo com data-rights no LSST. Atualmente o BLASST conta com **40 pesquisadores** de **11 instituições** (CBPF, IAG-USP, IF-USP, IFT-UNESP, IME-USP, ON, UESC, UFRJ, UFRGS, UFS e UFSC).

iii) IDAC-BRA (Independent Data Access Center – Brasil): Centro de dados e análise voltado ao armazenamento e processamento dos dados do LSST, integrado a uma rede internacional de *IDACs*. O acordo **LIneA/SLAC-DOE**, avaliado em **US\$4,4 milhões (~R\$ 24 milhões)**, foi assinado em 2024. Já foram investidos **~R\$10 milhões** em equipamentos, e há **R\$ 6 milhões** aprovados pela FINEP para a próxima fase. Participam como *data right holders* pesquisadores de mais de 25 instituições brasileiras, incluindo UEL, UFU, UnB, UNESP, UFRGS, UTFPR, USP, UFSM, INPE, UFF, UFRJ, CBPF, LNA, ON e outras. Após o período de comissionamento, de dois anos, tanto os dados como help-desk especializado estarão disponíveis para a comunidade astronômica brasileira.

Outros projetos relevantes, que não são diretamente relacionados com o LSST, incluem:

S-PLUS: Projeto internacional liderado por brasileiros, baseado no telescópio robótico T80 e em um *data lake* distribuído (IAG, IMT e CBPF) com **500 TB** de dados. Possui 229 usuários ativos em 52 instituições brasileiras e internacionais e mais de **800 usuários globais que utilizam o banco de dados (splus.cloud)**. O investimento inicial no telescópio foi de **US\$ 4 milhões**, e cerca de **R\$ 500 mil** foram aplicados na curadoria e distribuição de dados.

T150: Projeto futuro, no infravermelho próximo, aberto a toda a comunidade brasileira. Proposta de **US\$1,5 milhão** em negociação, condicionada a contrapartidas de outras fontes.

J-PAS e J-PLUS: Levantamentos ópticos no hemisfério norte, com telescópios de 2,5 m e 0,8 m, respectivamente, sob liderança brasileira parcial. Banco de dados gerido pelo Centro de Estudios de Física del Cosmos de Aragón (Espanha). Já receberam **mais de US\$ 5 milhões** de agências brasileiras.

GMT (Giant Magellan Telescope): Participação brasileira atualmente concentrada em serviços de software para equipes instrumentais, dado que o telescópio se encontra na fase de *preliminary design*. O investimento total do Brasil no GMT é de **US\$ 50 milhões**, mas a fração destinada a dados ainda não foi definida.

CTAO (Cherenkov Telescope Array Observatory): Mais de **50 pesquisadores brasileiros** já atuam na análise dos primeiros dados. Investimento nacional da ordem de **R\$ 20 milhões**, incluindo máquinas para processamento local e ações para conectividade internacional (Chile–Europa). Há expectativa de implantação de uma central de análise no Brasil.

Outros projetos internacionais de interesse (Einstein Telescope, CSST, Euclid, Nancy Grace Roman, SWGO, FINK, ZTF, BRICS Astronomy): Participação brasileira incipiente ou sem valores consolidados informados até o momento.

3.2. Estimativas para Investimentos em Big Data em Astronomia

O enfrentamento dos desafios do Big Data na Astronomia brasileira nos próximos dez anos exigirá um plano de investimentos consistente, coordenado e capaz de fortalecer uma comunidade científica integrada e distribuída geograficamente. A Astronomia, além de seu impacto cultural e educacional, desempenha papel estratégico no desenvolvimento de tecnologias de ponta e na formação de recursos humanos altamente qualificados. Sua natureza interdisciplinar e o uso intensivo de ciência de dados fazem com que a área esteja entre as ciências exatas que mais impulsionam inovações no século XXI.

Nesse contexto, a consolidação da atuação brasileira em grandes colaborações internacionais, como LSST, SKA, Gaia, GMT, além de outras, depende de investimentos contínuos tanto em Infraestrutura Técnica quanto em Recursos Humanos (RH) especializados.

Recursos Humanos: É essencial investir na formação, capacitação, contratação e manutenção de profissionais especializados em engenharia e ciência de dados, astronomia computacional com uso de HPC e engenharia de software baseada em MLOps (Machine Learning Operations). Também se destaca a necessidade de fomentar linhas interdisciplinares emergentes, como *astroinformática* e *astroestatística* baseadas em IA, para sustentar o ciclo completo de tratamento e análise de grandes volumes de dados astronômicos.

Infraestrutura Técnica: O fortalecimento e a implementação de centros de dados universitários e institucionais são prioridades, incluindo soluções para armazenamento, processamento e conectividade de alta velocidade. Isso deve ser acompanhado por ações efetivas para integração com redes e iniciativas internacionais, garantindo interoperabilidade e acesso competitivo aos dados globais.

Com base em projeções atuais, recomenda-se que o orçamento para a próxima década contemple:

i) aumento progressivo de bolsas e contratações (taxa anual estimada de 2,5%),

ii) aquisição e atualização de hardware compatível com demandas de Big Data (crescimento médio de 15% ao ano), e

iii) manutenção e evolução contínua de software científico especializado (crescimento médio de 8% ao ano).

Esses investimentos devem também abordar três eixos estratégicos transversais:

Inclusão e Descentralização: desenvolver políticas para envolver grupos de todas as regiões do país, estimulando redes colaborativas regionais e promovendo o acesso equitativo a recursos e dados, considerando governança horizontal regional representativa.

Conectividade e Sustentabilidade: garantir conexões de alta velocidade entre centros de pesquisa e assegurar a sustentabilidade técnica e financeira das bases de dados e infraestruturas de longo prazo.

Compartilhamento de Softwares e Ferramentas: Desenvolvimento e manutenção de *pipelines* científicas, sistemas interoperáveis e ambientes de dados abertos seguindo os princípios FAIR.

Nesse sentido, as Redes Colaborativas Nacionais desempenham papel central, envolvendo instituições estratégicas como o LNCC e o supercomputador Santos Dumont, que oferecem suporte crítico em HPC, e sociedades científicas como SAB, SBF, SBC e SBMAC, capazes de articular políticas públicas integradas para ciência, tecnologia e inovação aplicadas ao Big Data em Astronomia.

Considerando esses fatores, estima-se que o cenário nacional demande investimentos da ordem de R\$ 500 milhões a R\$ 1 bilhão nos próximos 10 anos, distribuídos entre infraestrutura de HPC, contratação e formação de pessoal, desenvolvimento e manutenção de software.

O fortalecimento da atuação brasileira em Big Data na Astronomia para a próxima década exigirá portanto investimentos estruturados em recursos humanos, infraestrutura tecnológica e conectividade, de modo a viabilizar uma comunidade científica integrada, distribuída geograficamente e preparada para enfrentar os desafios impostos por grandes levantamentos e colaborações internacionais.

A estimativa preliminar para um cenário nacional distribuído entre 3 a 5 centros de dados regionais ou temáticos é de R\$ 220 a 500 milhões para um período de 5 a 10 anos, distribuídos em quatro eixos:

- I. **Hardware** (HPC, Armazenamento, Infraestrutura): R\$ 100 a 150 milhões;
- II. **Software** (desenvolvimento MLOps, licenças, manutenção): R\$ 20 a 50 milhões;
- III. **Recursos Humanos** (equipes técnicas, programas especiais, bolsas): R\$50 a 100 milhões;
- IV. **Conectividade** (enlaces, equipamentos de rede): R\$ 50 a 100 milhões

Embora este plano de investimentos seja concebido para um horizonte de 10 anos, considera-se estratégico apresentar projeções estendidas para 20 anos, de forma a nortear discussões e subsidiar o planejamento de longo prazo para o próximo Plano Nacional de Astronomia (PNA). Essas projeções contemplam três fases principais, distribuídas ao longo de duas décadas, envolvendo hardware, software, recursos humanos e conectividade, considerando também a manutenção e a inserção brasileira em novas colaborações ao longo da próxima década.

Nesse cenário, no curto prazo (1 a 5 anos), o foco é a instalação inicial de sistemas de computação de alto desempenho (HPC) e armazenamento de grande capacidade, com investimentos de R\$60 a 80 milhões por centro. Desse montante, R\$15 a 20 milhões para o desenvolvimento de ferramentas; R\$25 a R\$30 milhões para sistemas de software considerando a formação inicial de equipes e a concessão de bolsa; e R\$15 a 20 milhões para contratação de enlaces de alta velocidade e infraestrutura básica de rede.

No médio prazo (6 a 10 anos), prevê-se a expansão e atualização tecnológica do hardware (R\$20 a R\$30 milhões por centro), a consolidação de sistemas de software e adoção de novas tecnologias (R\$20 a R\$30 milhões), o fortalecimento das equipes técnicas e científicas (R\$30 a R\$40 milhões) e o aumento da capacidade e resiliência da conectividade (R\$20 a R\$30 milhões).

No longo prazo (11 a 20 anos), a meta é garantir a renovação e modernização da infraestrutura com nova geração de equipamentos (R\$30 a R\$40 milhões), reengenharia e integração de software (R\$25 a R\$30 milhões), consolidação de carreiras técnicas e científicas (R\$40 a R\$50 milhões) e manutenção e integração internacional da conectividade (R\$20 a R\$30 milhões).

A abordagem, para estimativa de recursos, por fases garante: (i) início coordenado com foco em infraestrutura e capacitação, (ii) adaptação tecnológica e expansão escalável no médio prazo, e (iii) sustentação científica e técnica, de longo prazo, com autonomia e protagonismo internacional.

Considerando a instalação de três a cinco centros de dados distribuídos regionalmente, o investimento nacional estimado totaliza de R\$0,7 a R\$1,4 bilhões ao longo de 20 anos, assegurando autonomia científica e tecnológica para consolidar a participação brasileira nas colaborações internacionais estratégicas em Astronomia e Astrofísica.

Estimativa Nacional (3 a 5 centros):

Curto prazo (1–5 anos): R\$ 150–250 milhões;

Médio prazo (6–10 anos): R\$ 200–400 milhões;

Longo prazo (11–20 anos): R\$ 350–750 milhões;

Total nacional (1-20 anos): R\$ 0,7 a 1,4 bilhões²

Portanto, o fortalecimento da atuação brasileira em Big Data na Astronomia para a próxima década exigirá investimentos estruturados da ordem de 1 bilhão de reais, em recursos humanos, infraestrutura tecnológica e conectividade, de modo a viabilizar uma comunidade científica integrada, distribuída geograficamente e preparada para enfrentar os desafios impostos por grandes levantamentos e colaborações internacionais.

As projeções apresentadas neste relatório foram obtidas exclusivamente a partir (i) do formulário respondido pela comunidade (≈ 100 contribuições

2Os valores estimados são aproximações e dependem de refinamento e detalhamento referenciado dentro do paradigma TCO (ver Apêndice B).

qualificadas) e (ii) de valores correntes de mercado para infraestrutura de Big Data aplicável a volumes $\geq 0,5$ PB/ano (armazenamento, rede e computação). Embora este estudo não tenha adotado formalmente o modelo de Custo Total de Propriedade (TCO, CAPEX+OPEX) [46], recomendamos sua aplicação nos próximos ciclos orçamentários. Para projetos com $\geq 0,5$ PB/ano, o TCO captura custos de armazenamento em camadas, rede $\geq 10\text{--}40$ Gb/s, HPC/reprocessos, plataforma de dados (pipeline, catálogo, portal), pessoal e energia/manutenção, além de réplicas/backups e contingências. Propõe-se estimar volumes com multiplicadores do ciclo de vida ($\approx 5\text{--}10\times$) e de replicação ($\approx 2\text{--}3\times$), prever renovações em 4–5 anos e apresentar cenários Base/Alto/Conservador para 2026–2036.

Valores de referência sobre recursos orçamentários, obtidos em documentos públicos de investimento em C&T e Astronomia [24–30], usados para comparação com as estimativas aqui apresentadas, são discutidos no Apêndice B.

4. Conclusão

4.1. Diagnósticos e Recomendações

O trabalho realizado pela Subcomissão de Big Data evidencia que o Brasil dispõe de uma comunidade astronômica tecnicamente qualificada, com inserção internacional crescente e participação relevante em projetos estratégicos de grande porte. Contudo, o avanço efetivo da Astronomia de Big Data no país depende de ações coordenadas que articulem **infraestrutura, recursos humanos e integração institucional**.

A partir do levantamento feito com mais de uma centena de profissionais que atuam no Brasil em P&D em astronomia constatou-se que:

- i) Uma parcela significativa já está envolvida em projetos relevantes de Big Data, incluindo **brokers**, **VACs**, observação automatizada e análise multibanda.
- ii) Existem iniciativas promissoras de **data centers e plataformas computacionais**, como o AI-Scope [41], o BLASST [38], o IDAC-Brasil [10] e o *LineA Science Platform* [42], que já operam ou estão em fase de implantação, mas em diferentes estágios de maturação.

iii) Persistem gargalos críticos, como a carência de **infraestrutura de HPC**, a escassez de profissionais especializados em **engenharia e ciência de dados astronômicos** e a falta de estratégias nacionais coordenadas para formação e manutenção de profissionais a longo prazo (> 5 anos), integração e acesso a dados internacionais.

Diagnósticos Específicos:

No contexto da Astronomia de Big Data, dois elementos estruturantes são essenciais para a descoberta científica: **os Data Centers (DCs) e as Plataformas de Análise Avançada com Inteligência Artificial (IA)**, incluindo soluções baseadas em *Machine Learning* e *Agentes Inteligentes*.

a) Data Centers (DCs)

- *Hardware*: Muitos DCs nacionais ainda operam com infraestrutura limitada, enfrentando desafios de escalabilidade, capacidade de armazenamento (petabytes), redundância de dados e gargalos de I/O. A ingestão e consulta simultânea de dados em volumes massivos requer arquitetura paralela e balanceamento de carga.
- *Software*: Sistemas de gestão e análise de dados frequentemente não estão adaptados a workflows astronômicos complexos ou à integração com pipelines automatizados, dificultando o uso em tempo real.
- *Peopleware*: Há carência de grupos formados por profissionais com formação interdisciplinar que combina conhecimentos em astrofísica, engenharia de dados, computação de alto desempenho (HPC) e gestão de infraestrutura. Essa lacuna compromete a eficiência, manutenção e evolução tecnológica dos DCs.
- DACs (Data Access Control): Adequar DCs ao paradigma de DACs, que são sistemas de última geração para gerir e autorizar o acesso aos dados gerados pela infraestrutura observacional, garantindo que apenas utilizadores autorizados tenham acesso e controle ao Big Data gerado. Garantindo assim uma interação segura, organizada e controlada com os dados e a informação correspondente.

b) Plataformas de Análise com IA (por exemplo: VACs e Brokers)

- *Hardware*: A análise automatizada de eventos astronômicos em tempo real requer infraestrutura robusta com GPUs/TPUs de alto desempenho, baixa latência e alto throughput, principalmente para inferência em milhões de alertas diários.
- *Software*: Os modelos de IA precisam ser constantemente atualizados, interpretáveis e auditáveis, adaptando-se a novos tipos de dados e mantendo a confiabilidade sob diferentes condições. É essencial garantir interoperabilidade entre múltiplas fontes e plataformas.
- *Peopleware*: Há carência de profissionais com formação interdisciplinar que combina conhecimentos em astrofísica, engenharia de dados, computação de alto desempenho (HPC) e gestão de infraestrutura. Essa lacuna pode comprometer a eficiência, manutenção e evolução tecnológica dos DACs.

c) Papel das Agências de Fomento:

No eixo **Projetos, Iniciativas e Recursos Humanos**, ressalta-se o papel central das **agências de fomento brasileiras**, CAPES, CNPq, FINEP e Fundações Estaduais de Amparo à Pesquisa (FAPs), na sustentação de programas e ações voltados ao Big Data em Astronomia. A **coordenação e integração** entre essas agências é estratégica para otimizar investimentos, evitar redundâncias e maximizar impactos.

As modalidades de apoio atualmente disponíveis (auxílios individuais, bolsas, projetos temáticos, redes de pesquisa e programas institucionais) podem e devem ser ajustadas para contemplar explicitamente as demandas em **infraestrutura computacional, engenharia de dados e capacitação técnica** necessárias à ciência de dados astronômicos em larga escala.

Constata-se também demandas por **chamadas e editais** voltados a projetos bilaterais e multilaterais envolvendo o Brasil e parceiros internacionais, favorecendo o intercâmbio de tecnologias, dados e competências. Essa abordagem é essencial para consolidar a presença brasileira em **colaborações globais de grande porte**, permitindo não apenas o acesso a dados estratégicos, mas também a liderança em componentes de hardware, software e soluções inovadoras de TI aplicadas à astronomia.

Recomendações:

Frente a esse diagnóstico, destacam-se as seguintes **recomendações estratégicas**:

Eixo 1 — Fortalecimento de Infraestrutura

- Implantar **centros de armazenamento e bancos de dados e infraestrutura para processamento massivo de dados nacionais** interoperáveis com redes globais;
- Investir em plataformas abertas com suporte a análise massiva, *brokers* e VACs integrados;
- Investir em infraestrutura modular e escalável (com suporte a storage distribuído e sistemas paralelos), uso de containers e microserviços para maior flexibilidade e manutenção, e formação contínua de equipes técnicas e científicas interdisciplinares.

Eixo 2 — Prospecção e Capacitação de Recursos Humanos

- Criar programas permanentes de formação interdisciplinar em IA aplicada à astrofísica, HPC e astroinformática;
- **Organização de encontros temáticos e workshops nacionais e internacionais** sobre Astronomia de Big Data, incentivando a troca de experiências, o mapeamento de tendências e a articulação de parcerias;
- Estimular a mobilidade de pesquisadores entre centros, criando e fortalecendo redes colaborativas. Em especial, com políticas de incentivo que estimulem sinergia, interoperabilidade e compartilhamento de recursos entre grupos e centros de P&D em Astronomia e Computação;
- **Prospecção e formação de talentos** em múltiplos níveis, promovendo capacitação não apenas de jovens pesquisadores, mas também de cientistas seniores e quadros técnicos, assegurando atualização contínua frente às mudanças tecnológicas.

Eixo 3 — Políticas de Fomento

- Lançar **editais específicos para astronomia de Big Data**, contemplando infraestrutura, software e formação;
- Valorizar propostas colaborativas e abrangentes, respeitando as diferentes vocações e expertises dos grupos e casos científicos;
- Incentivar colaborações bilaterais e multilaterais, garantindo participação ativa e acesso aos dados em fases iniciais de operação;

- Promover políticas que permitam trazer e manter no Brasil cientistas de dados e engenheiros de software do sul global (e.g. Índia);
- A estrutura de financiamento deve incluir fundos plurianuais, consórcios interinstitucionais, cooperação internacional e mecanismos de avaliação e transparência ao longo do ciclo;
- **Promoção de INCTs dedicados à Astronomia de Big Data**, garantindo financiamento de longo prazo e foco em inovação tecnológica, ciência de dados e integração internacional;
- Para garantir a sustentabilidade das iniciativas em Big Data astronômico no Brasil ao longo da próxima década, é fundamental adotar uma abordagem de planejamento financeiro estruturado e escalonado. As agências de fomento devem articular fundos plurianuais integrados, com editais específicos para infraestrutura computacional, desenvolvimento de software científico, plataformas de IA e formação técnica continuada;
- A criação de consórcios nacionais multi-institucionais pode facilitar o compartilhamento de recursos e ampliar o impacto coletivo. Além disso, a busca ativa por financiamento internacional colaborativo (via parcerias com observatórios como LSST, SKA, ESO, CTA e agências como a NSF ou ESOF) pode diversificar as fontes de apoio. Finalmente, a implementação de mecanismos de monitoramento e avaliação periódica, com indicadores técnicos e científicos bem definidos, permitirá ajustes estratégicos ao longo do ciclo decenal.

Eixo 4 — Integração Nacional e Governança

- **Criação de uma Rede Nacional de Astronomia e Astrofísica**, sob coordenação do MCTI, com a missão de: (i) Organizar **white papers**, relatórios técnicos e workshops para definir prioridades de P&D; (ii) Integrar e alinhar políticas de fomento entre CNPq, FINEP, FAPs e programas internacionais; (iii) Garantir a **representação e o protagonismo brasileiro** nas principais colaborações científicas internacionais da próxima década;
- Promover governança distribuída, garantindo representatividade regional e autonomia orçamentária.

Diagnóstico e Recomendações sobre Integração:

Embora a comunidade astronômica brasileira seja **vibrante e diversificada**, com múltiplas iniciativas relacionadas a **Big Data**, ainda se observam oportunidades para **aprofundar a colaboração e otimizar esforços em projetos de P&D**.

Existem no país grupos e iniciativas com grande potencial de **sinergia**, mas nem sempre plenamente articulados entre si, o que reforça a necessidade de fortalecer a **articulação entre os diferentes atores** e ampliar a **cooperação regional**. Nesse contexto, ações que promovam **parcerias mais amplas e colaborativas** podem potencializar recursos já existentes, ampliar o impacto dos resultados e consolidar uma comunidade científica mais **inclusiva, coesa e cooperativa**.

Para enfrentar esse desafio, é necessário promover políticas e ações que incentivem a criação de redes de colaboração bem articuladas, e o fortalecimento e consolidação de centros de excelência em big data e *astro informática*. Além disso, recomenda-se que futuros editais de fomento valorizem explicitamente propostas cooperativas, com participação multi-institucional e estratégias claras de uma boa articulação entre os diferentes núcleos de pesquisa e as redes de colaboração.

Essa mudança de paradigma poderá ampliar a eficiência no uso dos recursos públicos, favorecer a formação de capital humano altamente qualificado em escala nacional e maximizar o impacto da ciência astronômica brasileira no cenário global.

A efetiva implementação dessas recomendações permitirá ao Brasil não apenas manter, mas ampliar sua relevância científica global, garantindo **capacidade técnica, autonomia estratégica e liderança em áreas-chave da Astronomia de Big Data**.

4.2. Big Data como Pilar da Astronomia Brasileira na Próxima

Década

O advento do Big Data consolidou-se como um dos marcos transformadores da ciência contemporânea, e na Astronomia esse impacto é ainda mais evidente. A crescente produção de dados em petabytes por ano, associada aos grandes telescópios e levantamentos globais, coloca a comunidade científica diante de desafios e oportunidades sem precedentes.

O Brasil, com sua tradição de excelência em astronomia, astrofísica e cosmologia, encontra-se diante de uma encruzilhada histórica: **ou consolida suas capacidades em ciência e engenharia de dados, integrando-se plenamente às redes globais de produção e análise de informação científica, ou corre o**

risco de ver seu protagonismo reduzir-se diante da rápida evolução tecnológica mundial.

Os diagnósticos e recomendações aqui apresentados deixam claro que **o Big Data não é apenas um recurso de apoio, mas um pilar central para o avanço da Astronomia Brasileira na próxima década.** A integração de infraestrutura de alto desempenho, plataformas interoperáveis, recursos humanos qualificados e políticas de fomento coordenadas será determinante para que o país amplie sua relevância e se posicione como um parceiro estratégico nas principais colaborações científicas internacionais.

Nesse cenário, a consolidação de uma Rede Nacional de Astronomia e Astrofísica de Big Data, articulada com centros de supercomputação, universidades, institutos de pesquisa e redes de alta velocidade, não apenas fortalecerá a produção científica e ampliará o impacto da astronomia na educação, na cultura e no desenvolvimento tecnológico, como também poderá impulsionar a geração de spin-offs (empresas derivadas), pois investimentos em TI de Big Data fomentam a produção de plataformas reutilizáveis, padrões abertos e competências de engenharia de dados que se traduzem em startups, transferência de tecnologia, novos serviços e empregos qualificados, com efeitos multiplicadores em setores como saúde, agricultura, clima e cidades inteligentes.

Ao final, investir de forma contínua e estratégica no Big Data em Astronomia significa **investir no conhecimento e no futuro do país**, garantindo que o Brasil participe não apenas como observador, mas como protagonista na exploração e compreensão do Universo.

Apêndice A - Limiares para a Astronomia de Big Data

Com base nas propriedades enumeradas abaixo, constata-se que geração de dados brutos acima de 0,1 PB/ano é o ponto em que requisitos de armazenamento, rede, HPC e governança mudam de escala e passam a demandar soluções de classe *petabyte* (Petabyte-Class Storage (ver Glossário)). Portanto, ao dimensionar Big Data em astronomia, é crucial explicitar o Fator de Amplificação de Dados (FAD), que multiplica o volume bruto para estimar o footprint efetivo anual de cada levantamento, incorporando *derivados* (coadds, catálogos, fotometria), metadados/QA, artefatos intermediários, versionamento, replicações/backups e políticas ILM/HSM.

Além FAD, o dimensionamento deve contemplar: I/O e reproprocessamento (paralelismo e *throughput* sustentado em escala PB), rede e distribuição (links $\geq 10\text{--}40$ Gb/s, *caches/mirrors* e políticas de *egress*), consulta e análise (particionamento espacial/temporal, índices e motores distribuídos), operação e custos (armazenamento em camadas quente/morna/fria, ILM/HSM, CAPEX/OPEX e equipe) e poder discriminativo (critérios objetivos que separam projetos com impacto infraestrutural elevado). Esses vetores, em conjunto com o FAD, determinam a classe Big Data do projeto e orientam requisitos de armazenamento, HPC, rede e orçamento.

Propriedades considerando 5 PB/ano:

(i) FAD: 0,5 PB/ano de brutos pode atingir valores de 1–3 PB/ano com *coadds*, catálogos, metadados, versões e 2–3× em réplicas/backups, montantes que demandam storage em escala *petabyte* contínuo e escalável e uso de tecnologias de armazenamento em ambientes de Big Data/HPC (objeto + fita / Lustre / Ceph).

(ii) I/O e reproprocessamento: ler 0,5 PB uma única vez a 1 GB/s consome ~6 dias; a 500 MB/s, ~12 dias. Reprocessos anuais exigem paralelismo massivo e filas HPC.

(iii) Rede e distribuição: transferir 0,5 PB leva ~46 dias a 1 Gb/s sustentado; na prática, demanda enlaces ≥ 40 Gb/s, *caches* e *mirrors*.

(iv) Consulta e análise: catálogos derivados atingem $10^9\text{--}10^{10}$ entradas, pedindo particionamento (HEALPix), índices espaciais/temporais e SQL distribuído/engines de *data lake* além de outras tecnologias de curadoria por agentes.

(v) Operação e custos: esse patamar impõe SLA de curadoria, pipelines versionados, AAI, DOIs e orçamentos de capex/opex típicos de centros de dados dedicados.

(vi) Poder discriminativo: separa levantamentos médios (J-PLUS, S-PLUS, SDSS histórico) daqueles que operam próximos ao regime petabyte/ano (p.ex., J-PAS em anos intensos), alinhando o termo “Big Data” ao impacto infraestrutural e não só ao bruto.

Comparação, exceto FAD, com 0,1 PB/ano (≈ 100 TB/ano):

- Reprocessamento: ler 0,1 PB a 1 GB/s ≈ 28 h; 0,5 PB ≈ 6 dias. Múltiplos *passes* tornam inviável sem forte paralelismo.
- Rede: mover 0,1 PB a 1 Gb/s ≈ 9 dias; 0,5 PB ≈ 46 dias — na prática exige $\geq 10\text{--}40$ Gb/s, *caches* e *mirrors*.
- Backups: LTO-9 (~ 18 TB nativo): 0,1 PB ≈ 6 fitas; 0,5 PB ≈ 28 fitas (duas cópias $\rightarrow 12$ vs 56).
- Armazenamento: 0,1 PB cabe num NAS único bem equipado; 0,5 PB normalmente pede objeto + fita e operações distribuídas.
- Análise: em 0,5 PB catálogos e *logs* já pressionam particionamento/índices distribuídos e orquestração de pipelines.

Conclusão: 0,1 PB/ano é de “alto volume institucional” e pode, sem considerar os produtos, ser tratado ainda fora da classe PSC; 0,5 PB/ano marca o salto operacional típico de Big Data (armazenamento, rede, HPC e governança de classe *petabyte*). Por isso recomendamos o uso de 0,5 PB como limiar mais discriminativo. Acima desse patamar, o ciclo de vida completo (exposições repetidas, metadados/QA, calibrações, mosaicos/coadds, catálogos/fotometria, artefatos intermediários, documentação/portais), somado a réplicas e backups, eleva o *footprint* anual típico para $\sim 1\text{--}3$ PB. Esse volume exige Armazenamento em Classe Petabyte (PCS (camadas quente/morna/fria), rede $\geq 10\text{--}40$ Gb/s para ingestão/distribuição, HPC para processamento cíclico (passadas múltiplas sobre centenas de TB–PB) e governança FAIR. Abaixo desse nível (p.ex., 0,1 PB/ano), vários cuidados ainda se aplicam, mas sem o mesmo salto infraestrutural e de custos.

É importante salientar que para obter os valores apresentados na Tabela 1, compilamos ordens de grandeza publicadas/estimadas por projeto e

ajustamos para refletir o ecossistema completo de dados (bruto + derivados). Assim, os valores apresentados na Tabela 1 são faixas indicativas por ano e podem variar com cadência observacional, eficiência de compressão, políticas de versionamento e janela operacional. Para planejamento de infraestrutura, recomenda-se considerar multiplicadores (3–10×) sobre o volume bruto devido a derivados, reprocessamentos e backups, além de políticas de retenção e espelhamento.

Apêndice B - Valores de Referência

Panorama Nacional: Atualmente, o Brasil investe cerca de 1% do seu Produto Interno Bruto (PIB) em ciência, tecnologia e inovação (CT&I), o que corresponde a aproximadamente R\$ 89,5 bilhões por ano, segundo dados da Apufsc-Sindical e do jornal *Valor Econômico*. Essa proporção é considerada modesta quando comparada a países desenvolvidos, como Alemanha e Estados Unidos, cujos investimentos superam 3% do PIB. O Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT) é a principal fonte de financiamento para pesquisa e desenvolvimento no país, e seus aportes vêm crescendo nos últimos anos — em 2023, foram investidos R\$ 9,96 bilhões, com previsão de R\$ 12,72 bilhões para 2024, conforme informações da *Agência Brasil*. O Ministério da Ciência, Tecnologia e Inovação (MCTI) mantém programas estratégicos para impulsionar a inovação e o desenvolvimento tecnológico, como o *Programa Mais Inovação* e o *Programa Centelha*. No campo emergente da inteligência artificial, o Plano Brasileiro de Inteligência Artificial (PBIA) prevê investimentos da ordem de R\$23 bilhões até 2028, segundo a *Agência Brasil*. Paralelamente, o governo federal direciona recursos a projetos de telecomunicações e outras áreas estratégicas, visando acelerar a transformação digital e aumentar a competitividade da indústria nacional, conforme divulgado pelas agências federais de fomento e notícias.

Pesquisa e Desenvolvimento: O Brasil investe cerca de 1,2% do PIB em pesquisa e desenvolvimento, um valor que pode ser aumentado para 2,5% conforme projetos como a PEC (Proposta de Emenda à Constituição) que prevê essa alocação. **Projetos de Apoio:** O Ministério da Ciência, Tecnologia e Inovação (MCTI) tem destinado mais de R\$26,3 bilhões do Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT) para fortalecer a ciência brasileira, incluindo projetos de Astronomia. **Grande Telescópios:** O Brasil é um membro fundador do Giant Magellan Telescope (GMT), que terá uma capacidade de observação 200 vezes maior que os telescópios atuais, e o investimento do estado de São Paulo de 50 milhões de dólares garante cerca de 4% do tempo anual de observação para cientistas. Nos últimos cinco anos, o investimento médio anual em tecnologia no Brasil é de cerca de US\$50 bilhões. Este valor inclui investimentos em software, hardware e serviços, e coloca o Brasil em 10º lugar no ranking mundial de investimentos em TI e HPC. A tendência é que os investimentos em transformação digital continuem crescendo na taxa média de 15% ao ano, com foco em áreas como Nuvem, Big Data & Analytics, Inteligência Artificial, Segurança da Informação, Robótica e Internet das Coisas.

Astronomia: As informações utilizadas para as estimativas apresentadas na seção 3.2 estão baseadas em fontes disponíveis na internet sobre investimentos em astronomia no Brasil e projeção do custo de HPC para os próximos dez anos. **Fonte Google:** O investimento anual em Astronomia no Brasil não é precisamente divulgado em um valor único, mas sim em diversos

projetos e iniciativas, como o investimento em telescópios, projetos de pesquisa e desenvolvimento, e apoio a estudantes e pesquisadores. O orçamento anual para os telescópios SOAR, Gemini e CFHT, por exemplo, é de aproximadamente US\$2.58 milhões desde 2009. Além disso, o Brasil investe em pesquisa e desenvolvimento cerca de 1,2% do PIB, e há projetos de aumentar essa porcentagem para 2,5%. Detalhes do investimento para **Telescópios**: o investimento anual em telescópios como SOAR, Gemini e CFHT é de US\$ 2.58 milhões, com o Gemini a representar cerca de 1,652 milhões, o SOAR 0,509 milhões e o CFHT 0,424 milhões.

Referências

1. PNA, 2010-2020. <https://old.lna.br/PNA-FINAL.pdf>
2. **Pathways to Discovery in Astronomy and Astrophysics for the 2020s**
<https://nap.nationalacademies.org/catalog/26141/pathways-to-discovery-in-astronomy-and-astrophysics-for-the-2020s>
3. Astronomy decadal plan 2026–2035
<https://www.science.org.au/supporting-science/science-policy-and-analysis/decadal-plans-for-science/astro2035>
4. Mahabal et al., AstroInformatics: Recommendations for Global Cooperation
<https://arxiv.org/pdf/2401.04623>
5. Barres Almeida et al., Information technology & astronomical data in Brazil: Perspectives and proposals, SAB, <https://arxiv.org/abs/2006.08703>
6. Big Data by IEEE; NIST; ISO-IEC . <https://bigdataieee.org/> ;
<https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r1.pdf> ;
<https://cdn.standards.iteh.ai/samples/68305/cd3c6bd81a1341bdba5bef2a53174968/ISO-IEC-20546-2019.pdf>
7. Rosa, R.R. Data Science Strategies for Multimessenger Astronomy, An. Acad. Bras. Ciênc. 93 (suppl 1), 2021 <https://doi.org/10.1590/0001-3765202020200861>
8. Rosa, R.R. Astronomia do futuro: ciência de dados e tecnologia da informação, Revista Brasileira de Astronomia, No 4, 2019. SAB.
<https://sab-astro.org.br/wp-content/uploads/2022/01/RBA-4online.pdf>
9. Alvo, M., Statistical Inference and Machine Learning for Big Data, Springer, 2022.
<https://link.springer.com/book/10.1007/978-3-031-06784-6>
10. IDAC-BRasil-LSST. <https://www.linea.org.br/idac>

11. LIneA. <https://www.linea.org.br/sobre-linea> / <https://www.linea.org.br/inct>
12. LSST Corporation / Vera C. Rubin Observatory. *Legacy Survey of Space and Time*. Disponível em: <https://www.lsst.org> .
13. BINGO <https://link.springer.com/article/10.1007/s10686-020-09666-9>
<https://www.gov.br/inpe/pt-br/assuntos/ultimas-noticias/radiotelescopio-bingo>
14. BDA. https://en.wikipedia.org/wiki/Brazilian_Decimetric_Array ;
<https://www.sciencedirect.com/science/article/abs/pii/S0273117799005918>
15. GMT-Brasil. <https://www.gmt.iag.usp.br/en/gmt-braz>
<https://www.gmt.iag.usp.br/gmtbro/equipe>
16. Giant Magellan Telescope (GMT) Organization. *Project Overview*. Disponível em: <https://www.gmto.org> .
17. Cherenkov Telescope Array Observatory (CTAO). *Science with CTA*. Disponível em: <https://www.cta-observatory.org> .
<https://www.ctao.org/partners/in-kind-contributors/>
18. DESI. <https://www.desi.lbl.gov/data-systems/>
19. S-PLUS <https://www.splus.iag.usp.br/>
<https://revistapesquisa.fapesp.br/en/star-fossil/>
splus.cloud: <http://splus.cloud>
20. J-PAS. <https://www.j-pas.org/>
<https://www.j-pas.org/collaboration>
21. J-PLUS. <https://www.j-plus.es/>
<https://www.j-plus.es/collaboration>
22. Gemini Observatory. *Brazilian Participation in the Gemini Partnership*. Disponível em: <https://www.gemini.edu> .
23. Square Kilometre Array (SKA) Observatory. *Project Overview*. Disponível em: <https://www.skao.int> .
24. Apufsc-Sindical. *Investimento em Ciência e Tecnologia no Brasil*. Disponível em: <https://www.apufsc.org.br> .
25. Agência Brasil. *FNDCT investe R\$ 9,96 bilhões em 2023*. Disponível em: <https://agenciabrasil.ebc.com.br> .

26. Agência Brasil. *Plano Brasileiro de Inteligência Artificial (PBIA)*. Disponível em: <https://agenciabrasil.ebc.com.br> .
27. Agência Gov. *Investimentos do Governo Federal em Transformação Digital e Inovação*. Disponível em: <https://www.gov.br> .
28. National Research and Education Network (RNP). *Serviços e Projetos para a Ciência Brasileira*. Disponível em: <https://www.rnp.br> .
29. LNCC – Laboratório Nacional de Computação Científica. *Supercomputador Santos Dumont*. Disponível em: <https://sdumont.lncc.br> .
30. Ministério da Ciência, Tecnologia e Inovação (MCTI). *Programas e Projetos de Fomento à Pesquisa*. Disponível em: <https://www.gov.br/mcti> .
31. ESO. <https://www.eso.org> .
32. DES. <https://arxiv.org/pdf/2205.10939>
33. HEPIX. <https://w3.hepix.org/benchmarking/HS06.html>
34. ZTF. <https://www.ztf.caltech.edu/#>
35. L4. <https://sites.northwestern.edu/l4/>
36. 4MOST <https://www.4most.eu/cms/home/>
37. BAWG <https://www.gov.br/inpe/en/events/bawg-workshop>
38. BLASST / CLAST. <https://people.if.usp.br/abramo/pt-br/node/49>
39. SPANET. <https://www.spanet.iag.usp.br/home>
40. T80-SOUTH. <https://noirlab.edu/public/programs/ctio/t80-south-telescope/>
41. AI-SCoPE. <aiscope.cbpf.br>
42. LIneA Science Platform. <https://scienceplatform-dev.linea.org.br/lsp>
43. IVOA. <https://www.ivoa.net/>
44. LAB-IA. <http://labia.cbpf.br>
45. Rednesp. <https://rednesp.br/>

46. Wynstra, F., Hurkens, K. (2005). Total Cost and Total Value of Ownership. In: Perspektiven des Supply Management. Springer, Berlin, Heidelberg.
https://link.springer.com/chapter/10.1007/3-540-27018-3_21
47. Open Universe. <https://openuniverse.cbpf.br/>
48. UNOOSA. <https://www.unoosa.org/>
https://link.springer.com/chapter/10.1007/3-540-27018-3_21

Glossário

AI-SCoPE (Artificial Intelligence for Smart Cosmic Exploration) — Plataforma que integra um *data lake* astronômico com inteligência artificial para busca, filtragem e análise de dados.

Agentes Autônomos — Programas ou sistemas que executam tarefas de forma independente, aprendendo e adaptando-se a novas situações, como classificação automática de eventos astronômicos ou coordenação de observações.

Astroestatística — Aplicação de métodos estatísticos avançados para análise e interpretação de dados astronômicos complexos incluindo as abordagens de aprendizagem de máquina e mineração de dados para Big Data.

Astroinformática — Campo interdisciplinar que une astronomia, ciência de dados e computação de alto desempenho para análise de grandes volumes de dados astronômicos.

Big Data — Conjunto de técnicas e tecnologias para lidar com grandes volumes, velocidades e variedades de dados, incluindo sua veracidade, valor e variabilidade, além de outras propriedades complementares.

BLASST (Brazilian Lab of Astronomical Survey Studies) — Plataforma brasileira de análise em astrofísica avançada com foco em *Value-Added Catalogs* (VACs) e inteligência artificial.

BPG (Brazilian Participation Group) – Grupo de Participação Brasileira (BPG) com acesso à dados restritos do LSST via LIneA. Atualmente conta com cerca de 120 membros, sendo 24 Pesquisadores Principais (PI) de 22 instituições em 10 estados, distribuídos em 8 colaborações científicas. Cada PI podendo ter até 4 pesquisadores juniores (JAs) em seu grupo (pós-docs, estudantes de pós-graduação ou graduação).

Broker — Sistema que recebe fluxos de alertas astronômicos (ex.: LSST), filtra eventos, classifica-os, adiciona metadados e redistribui resultados relevantes para os usuários.

Ceph - plataforma de armazenamento distribuído que oferece **objeto (S3/Swift via RGW), bloco (RBD) e ficheiro (CephFS)**. Escala em *commodity hardware*, com **replicação/erasure coding** e *self-healing*; ideal para “dados mornos” e **repositórios duráveis**.

Coadds (de *co-added images* ou *coaddition*) - são imagens resultantes da **combinação estatística de múltiplas exposições da mesma região do céu**. As imagens são registradas astrométrica e fotometricamente e depois combinadas (média/mediana/soma ponderada por ruído ou seeing), com **rejeição de outliers** (ex.: sigma-clipping), **homogeneização da PSF** e **subtração de fundo**, para ganhar **profundidade e SNR** e mitigar artefactos. Geralmente produzem-se também **mapas de peso/variância** e **catálogos “coadd”**

CTA (Cherenkov Telescope Array) — Rede internacional de telescópios para detecção de raios gama de alta energia.

DAC (Data Access Center) — Centro de dados que armazena, processa e disponibiliza grandes volumes de dados científicos.

DEO - Escritório de Ciência do Departamento de Energia dos Estados Unidos da América.

FAIR — Princípios para dados científicos: *Findable* (Encontráveis), *Accessible* (Acessíveis), *Interoperable* (Interoperáveis), *Reusable* (Reutilizáveis).

FAD (Fator de Amplificação de Dados): coeficiente que multiplica o **volume bruto** para estimar o **footprint efetivo anual** de um levantamento, incorporando *derivados* (coadds, catálogos, fotometria), **metadados/QA**, **artefatos intermediários**, **versionamento**, **replicação** e **backups**.

FNDCT — Fundo Nacional de Desenvolvimento Científico e Tecnológico, principal fonte de financiamento à pesquisa no Brasil.

HPC (High-Performance Computing) — Computação de alto desempenho para processar grandes volumes de dados ou simulações complexas.

HSM (Hierarchical Storage Management) – Mecanismo **técnico** que **migra automaticamente** dados entre camadas (quente→morna→fria e vice-versa), mantendo *stubs* e **recall transparente** quando um ficheiro é acedido. Responde **como** mover, integrando disco/SSD, objeto e fita.

IDAC-BRA (Independent Data Access Center - Brazil) — Centro de acesso independente para armazenamento e processamento de dados de grandes levantamentos como o LSST. O IDAC-BRA está atualmente em instalação pelo LIneA, como contribuição *in-kind*, o que permitiu, entre outros, a oferta de 24 vagas de PI para acesso aos dados restritos desse levantamento.

ILM (Information Lifecycle Management) – Conjunto de **políticas e processos** que governam todo o ciclo de vida dos dados — criação, uso, retenção, arquivamento e descarte. Define **o quê** mover, **quando** e **porquê**, com base em critérios como idade, frequência de acesso, custo, compliance/retention, sensibilidade, versão etc.

J-PAS / J-PLUS — Projetos de mapeamento do céu com múltiplos filtros fotométricos no hemisfério norte.

Lab-IA - Laboratório de Inteligência Artificial aplicada à Física do Centro Brasileiro de Pesquisas Físicas. É uma infraestrutura aberta e multiusuário voltada para projetos de pesquisa de alta complexidade em Inteligência Artificial em áreas como Física, Astrofísica, Geofísicas e suas aplicações.

LineA (Laboratório Interinstitucional de e-Astronomia) - é um laboratório multiusuário que opera um centro de e-ciência, prestando apoio técnico e científico a pesquisas que demandam o processamento e a análise de grandes volumes de dados. É responsável por coordenar as atividades do BPG-LSST.

LSST (Legacy Survey of Space and Time) — Levantamento de 10 anos no Observatório Vera C. Rubin, projetado para gerar dezenas de petabytes de dados.

Lustre - sistema de ficheiros paralelo POSIX, típico de supercomputadores. Entrega **altíssima largura de banda** para reprocessamento e *scratch* (múltiplos nós a ler/escrever em paralelo). Arquitetura com **MDS/OSS/OST**; ótimo para “dados quentes” e I/O intensivo.

MLOps (Machine Learning Operations) — Conjunto de práticas para desenvolvimento, implantação e manutenção de modelos de aprendizado de máquina de forma automatizada e escalável, aplicadas aqui à análise astronômica.

MMA (Multimessenger Astronomy) — Astronomia multimessageira, que combina diferentes mensageiros cósmicos (luz, partículas, ondas gravitacionais).

Peopleware — Peopleware, em tecnologia da informação, refere-se à parte humana envolvida no desenvolvimento, implementação e uso de sistemas de hardware e software. É o reconhecimento de que o sucesso de qualquer projeto tecnológico depende tanto das ferramentas (hardware e software) quanto das pessoas que as criam, mantêm e utilizam.

Quality Assurance (QA):. No contexto de Big Data em Astronomia, refere-se ao **conjunto de processos** que asseguram que os dados e os produtos derivados cumprem padrões definidos. Inclui: validação de **metadados**, checks de **integridade e consistência**, métricas automáticas (seeing/FWHM, PSF, SNR, fundo do céu), verificação **astrométrica/fotométrica** (zeropoints, desvios), **flags de artefatos** (rastros de satélite, *cosmic rays*), auditoria do **pipeline**, monitorização temporal e critérios de **aceitação/rejeição**.

QC (Quality Control) é a inspeção de **amostras/produtos**; **QA** foca no **processo**. Ambos geram documentos e metadados que contam no volume total.

RNP (Rede Nacional de Ensino e Pesquisa) — Rede acadêmica de alta velocidade que conecta universidades e centros de pesquisa brasileiros.

SLAC - Stanford Linear Accelerator Center (expandir)

SKA (Square Kilometre Array) — Projeto internacional para o maior radiotelescópio do mundo, com centenas de petabytes de dados anuais.

Petabyte-class storage (PCS) - Em Big Data, PCS refere-se à temperatura da camada de armazenamento para a qual “**quente / morna / fria**” designa **camadas de armazenamento** segundo frequência de acesso, latência e custo: a) **Quente (hot)**: acesso muito frequente e baixa

latência. Tipicamente NVMe/SSD e/ou FS paralelo (ex.: Lustre/GPFS). Uso: *scratch*, processamento diário, QA, reprocessos intensivos. **Mais caro** e rápido. b) **Morna (warm)**: acesso recorrente, mas menos crítico. Normalmente **objeto** (Ceph/S3, MinIO) e HDD. Uso: produtos estáveis (coadds, catálogos), partilha e distribuição. **Custo intermédio**. c) **Fria (cold)**: acesso raro, com recuperação mais lenta. Arquivo em **fita (LTO)** ou “deep archive” em nuvem. Uso: *raw* históricos, *snapshots* de releases, cópias de segurança. **Mais barato** por TB. Os dados circulam entre camadas por políticas de **ILM/HSM** (ciclo de vida), equilibrando **desempenho x custo x durabilidade**.

VAC (Value-Added Catalog) — Catálogo de dados enriquecidos, derivados de levantamentos astronômicos brutos, contendo parâmetros físicos e informações complementares calculadas.

ZTF (Zwicky Transient Facility) — Projeto de levantamento rápido do céu para detecção de eventos astronômicos transitórios.