MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Avaliando plataformas científicas para análise de dados com computação de alto desempenho

Número do processo institucional: 400077/2022-1

Número do processo individual: 300764/2023-5 (Vigência original: 01/02/2023 a 28/02/2023)

Número do processo individual: 301223/2023-8 (Vigência original: 01/03/2023 a 31/12/2023)

Bolsista: Ebenezer Agyei-Yeboah

Orientador: Rafael Santos

Área: Coordenação de Laboratórios Associados - CTE

Vigência executada da bolsa: 01/02/2023 a 31/10/2023

Modalidade: DA

Natureza do trabalho: Relatório PCI

INPE
São José dos Campos
2023

Resumo

As plataformas de ciência de dados são plataformas de análise do lado do servidor que fornecem ambientes escaláveis, colaborativos, orientados por dados e, às vezes, de aprendizado de máquina para cientistas de dados, pesquisadores e estudantes. Eles oferecem vários recursos, como autenticação, autorização, armazenamento de dados, recursos de computação e ferramentas de programação. No entanto, nem todas as plataformas são igualmente adequadas para diversas necessidades e propósitos. Este relatório analisa plataformas de ciência de dados, como Colab, Kaggle, Posit Cloud, Deepnote, Datalore, Replit, espaços de trabalho DataCamp, RStudio auto-hospedado e web auto-hospedado JupyterHub e SciServer. Este relatório também apresenta um resumo de uma análise comparativa de duas plataformas proeminentes, JupyterHub e SciServer, com foco em sua arquitetura, principais recursos, base de usuários alvo, integração e ecossistema, suporte à comunidade e experiências do usuário. O estudo visa auxiliar pesquisadores, cientistas e analistas de dados na seleção da plataforma mais adequada para seus projetos de análise de dados com computação de alto desempenho. O relatório mostra que o JupyterHub e o SciServer têm características distintas que os diferenciam entre si e de outras plataformas, e que cada plataforma tem seus próprios pontos fortes e fracos, dependendo dos requisitos e preferências específicas dos usuários. O relatório conclui com algumas recomendações e sugestões para futuras pesquisas e desenvolvimento de plataformas de ciência de dados.

LISTA DE FIGURAS

Pag.

LISTA DE TABELAS

Pag.

CONTEÚDO

Pag.

1 SUMÁRIO EXECUTIVO

Este resumo executivo resume a exploração abrangente empreendida no projeto intitulado "Estudo Comparativo e Implementação de uma Plataforma Científica para Análise de Dados Usando Computação de Alto Desempenho". Os objetivos primários deste projeto foram meticulosamente perseguidos e alcançados de forma sistemática.

Os principais objetivos deste projeto são:

- Estudo Comparativo e Testes de Plataformas Científicas:

  O projeto começou com um exame minucioso de várias plataformas de ciência de dados, com foco em seus pontos fortes e fracos. As principais considerações incluíram flexibilidade, facilidade de uso e segurança de acesso a código e dados. Este estudo comparativo estabeleceu a base para a tomada de decisão informada na seleção de uma plataforma ótima.

- Estudo Comparativo das Plataformas JupyterHub e SciServer:

  Com base no estudo comparativo inicial, uma análise detalhada foi conduzida especificamente em duas plataformas proeminentes, JupyterHub e SciServer. A investigação teve como objetivo identificar os requisitos e limitações precisos de cada plataforma. Essa profundidade de análise serviu para destacar os recursos, capacidades e bases de usuários de destino do JupyterHub e do SciServer.

As principais conclusões deste relatório são:

- JupyterHub e SciServer têm características distintas que os diferenciam um do outro e de outras plataformas, e que cada plataforma tem seus próprios pontos fortes e fracos, dependendo dos requisitos e preferências específicas dos usuários.
- O JupyterHub oferece personalização, colaboração multiusuário e compartilhamento de notebooks, enquanto o SciServer se destaca na integração de dados, escalabilidade, ferramentas de colaboração e recursos de visualização.
- O JupyterHub se beneficia de um vasto ecossistema, integrando-se a bibliotecas, estruturas e serviços de ciência de dados, enquanto o SciServer fornece um

ecossistema especializado sob medida para astronomia, integrando-se a bancos de dados astronômicos, ferramentas de simulação e bibliotecas de visualização.

- O JupyterHub tem uma comunidade grande e ativa, enquanto o SciServer oferece suporte personalizado e acesso a especialistas em domínio.

As principais recomendações deste relatório são:

- Para escolher uma plataforma científica para análise de dados com computação de alto desempenho, os usuários devem considerar suas necessidades, preferências e objetivos específicos e avaliar os recursos, benefícios e limitações de cada plataforma de acordo.
- Para melhorar a usabilidade, funcionalidade e acessibilidade das plataformas científicas, os desenvolvedores e provedores devem considerar o feedback e as sugestões dos usuários e da comunidade e implementar as melhorias e aprimoramentos necessários.
- Para promover a pesquisa científica e a colaboração usando plataformas científicas, pesquisadores e analistas de dados devem explorar as oportunidades e desafios do uso de diferentes plataformas e compartilhar suas melhores práticas e experiências com outras pessoas.

2 INTRODUÇÃO

No cenário em constante evolução da exploração orientada por dados e da investigação científica, o papel de plataformas de ciência de dados robustas e versáteis tornou-se primordial. Este relatório, intitulado "Avaliando Plataformas Científicas para Análise de Dados com Computação de Alto Desempenho", busca dissecar e compreender a intrincada dinâmica dessas plataformas. Com foco em escalabilidade, colaboração e ambientes orientados a dados, essa exploração está enraizada no contexto mais amplo do projeto intitulado "Estudo Comparativo e Implementação de uma Plataforma Científica para Análise de Dados Usando Computação de Alto Desempenho".

As plataformas de ciência de dados são plataformas de análise do lado do servidor que fornecem ambientes escaláveis, colaborativos, orientados por dados e, às vezes, de aprendizado de máquina para cientistas de dados, pesquisadores e estudantes. Eles oferecem vários recursos, como autenticação, autorização, armazenamento de dados, recursos de computação e ferramentas de programação. No entanto, nem todas as plataformas são igualmente adequadas para diversas necessidades e propósitos. Escolher a plataforma certa para análise de dados com computação de alto desempenho pode ter um impacto significativo na qualidade, eficiência e eficácia da pesquisa e colaboração.

Em uma era em que pesquisadores, cientistas de dados e estudantes navegam em um vasto mar de plataformas, a necessidade de uma compreensão abrangente de seus pontos fortes, fracos e aplicações é indispensável. Este relatório encapsula a essência desse entendimento, tecendo os resultados de uma exploração aprofundada empreendida para orientar aqueles envolvidos em pesquisa científica e análise de dados.

Os objetivos gerais deste projeto foram meticulosamente alinhados com as demandas em evolução da comunidade científica. Começando com um estudo comparativo meticuloso e testes de várias plataformas científicas, o projeto progrediu para um exame matizado do JupyterHub e do SciServer. Por meio de avaliações rigorosas, esse empreendimento teve como objetivo descobrir as características distintas, a arquitetura e as experiências do usuário que diferenciam essas plataformas.

Ao nos aprofundarmos neste relatório, convidamos os leitores a navegar pelos intrincados caminhos das plataformas de ciência de dados, explorando as facetas da arquitetura, recursos, bases de usuários, integração e suporte da comunidade. Se você é um pesquisador experiente ou um analista de dados iniciante, este relatório busca capacitar

seu processo de tomada de decisão, fornecendo insights valiosos sobre o mundo da computação de alto desempenho para análise de dados. As páginas seguintes desdobram uma narrativa de exploração, análise e contribuição para o discurso contínuo da colaboração científica.

Este relatório apresenta uma análise comparativa de duas plataformas proeminentes, JupyterHub e SciServer, com foco em sua arquitetura, principais recursos, base de usuários alvo, integração e ecossistema, suporte à comunidade e experiências do usuário. O relatório também analisa outras plataformas de ciência de dados, como Colab, Kaggle, Posit Cloud, Deepnote, Datalore, Replit, espaços de trabalho DataCamp, RStudio auto-hospedado e web auto-hospedado, e os compara com JupyterHub e SciServer. O relatório visa auxiliar pesquisadores, cientistas e analistas de dados na seleção da plataforma mais adequada para seus projetos de análise de dados com computação de alto desempenho.

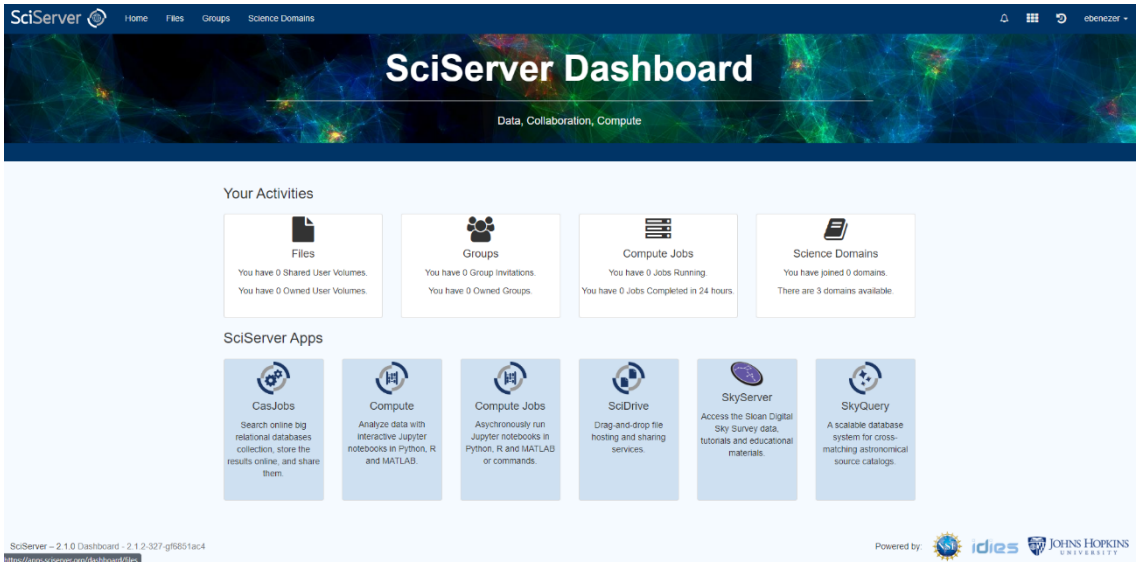
2.1 Histórico

O cenário da ciência de dados foi enriquecido pelo desenvolvimento de plataformas como SciServer e JupyterHub, que permitem aos usuários acessarem e executarem notebooks Jupyter na nuvem, contribuindo com funcionalidades, recursos, capacidades e abordagens distintas para análise de dados, colaboração, escalabilidade e custo. Entender suas semelhanças e diferenças é crucial para pesquisadores, cientistas e analistas de dados na seleção da plataforma que melhor se adapta aos seus requisitos de análise de dados e pesquisa.

O SciServer é uma plataforma científica que fornece ambientes colaborativos, interativos e em modo batch para análise do lado do servidor com conjuntos de dados extremamente grandes usando linguagens como Python, R e SQL em ambientes Jupyter Notebook e RStudio [1]–[3]. Ele utiliza tecnologias de conteinerização como Docker/máquinas virtuais e consiste em vários componentes: Compute, CasJobs, SkyServer, SkyQuery, SciDrive, Login Portal, Dashboard e File Browser [2], [4]. A computação permite que os usuários executem blocos de anotações Jupyter na nuvem usando contêineres do Docker [1], [2], [5]. CasJobs permite que os usuários consultem bancos de dados relacionais usando SQL. O SkyServer fornece aos usuários acesso a dados astronômicos do Sloan Digital Sky Survey (SDSS) [2]. O SkyQuery permite que os usuários cruzem vários catálogos astronômicos usando SQL. SciDrive é o componente que permite aos usuários armazenarem e compartilhar arquivos na nuvem [2], [4]. O Portal de Login permite que

os usuários criem e gerenciem suas contas SciServer. O Painel (consulte Figura 2.1) é onde os usuários têm acesso a todas as ferramentas do SciServer de um só lugar [2], [4]. File Browser é o componente que permite aos usuários gerenciarem seus volumes de armazenamento de arquivos no SciServer [2], [4].

Figura 2.1 - Painel do SciServer



Fonte: [6]

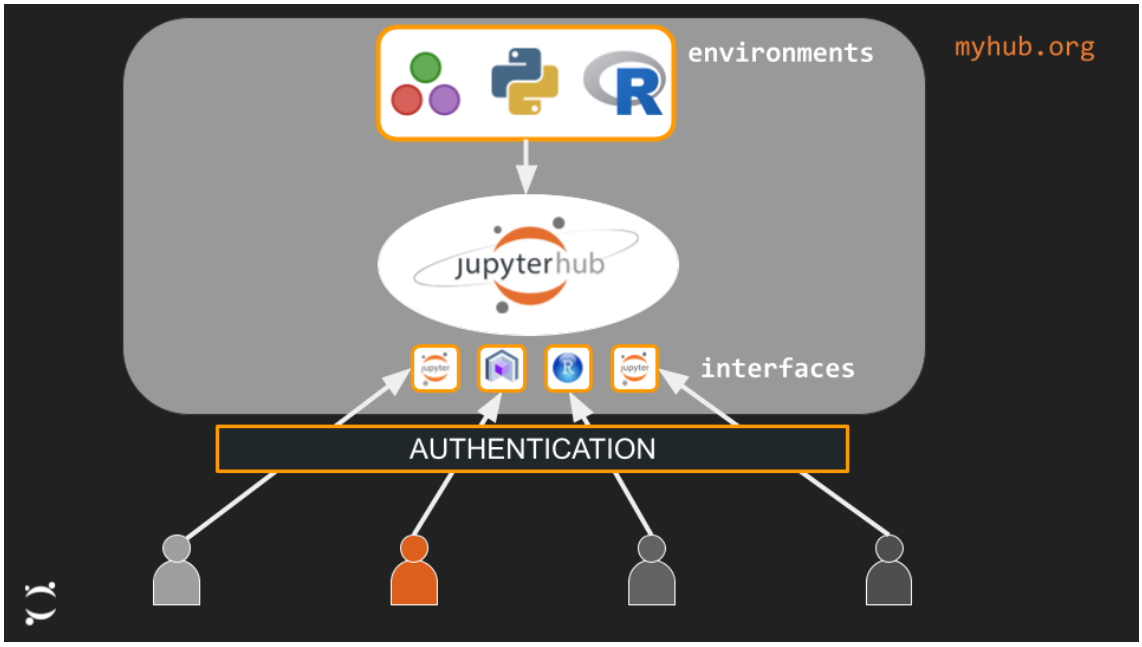
Uma das características de destaque do SciServer é seu fornecimento de armazenamento de arquivos privados e espaço de banco de dados SQL, permitindo que os usuários compartilhem e colaborem em recursos por meio do sistema flexível de controle de acesso a recursos que foi incorporado à plataforma. Suas APIs REST tornam o sistema escalável, colaborativo e orientado por dados, estendendo sua aplicabilidade a vários domínios da ciência, como oceanografia, genômica e ciência da terra [1]. Apesar de seus recursos robustos, o SciServer tem limitações, como suporte restrito para computação GPU afetando seus recursos de aprendizado de máquina [3]. No entanto, seu design geral, construído para ser completo, escalável, modular, portátil e interoperável, o torna um ator significativo no campo da ciência orientada a dados.

O SciServer é hospedado pela Universidade Johns Hopkins, e é gratuito para uso acadêmico e não comercial, com um limite de 10 GB de armazenamento por usuário [2], [5] com opções de armazenamento de grandes dados ou computação científica exclusiva

dentro da plataforma. Ele também permite que os usuários escolham entre diferentes ambientes de computação e opções de armazenamento [2], [5]. No entanto, ele só pode gerenciar centenas de usuários simultâneos e tem opções de personalização limitadas [2], [5]. Ele pode ser implantado no local, na nuvem ou em um híbrido de ambos, oferecendo flexibilidade na criação e implantação [1], [8]. O SciServer é usado principalmente para pesquisa em astronomia e mecânica dos fluidos, mas também pode ser usado para outros domínios que exigem análise de dados em larga escala [2], [7].

Por outro lado, o JupyterHub, parte do projeto de código aberto Jupyter, tornou-se sinônimo de experiências de computação interativa e exploratória. Ele suporta mais de 100 linguagens de computador e tem sido amplamente adotado para ciência de dados e aprendizado de máquina [9]. É um sistema JupyterHub multiusuário que gera, gerencia e faz proxy de várias instâncias do servidor de notebook Jupyter de usuário único [8]. Ele gerencia várias instâncias do servidor de notebook Jupyter de usuário único, atendendo a diversos grupos, como estudantes, grupos de trabalho de ciência de dados corporativos, projetos de pesquisa científica e grupos de computação de alto desempenho [8], [9]. Ele consiste em quatro subsistemas: um Hub, um Proxy, um Spawner e um Authenticator [10]. O Hub é o coração do JupyterHub, que gerencia contas de usuário e autenticação, e coordena os Spawners [10]. O Proxy é a parte voltada para o público do JupyterHub, que roteia solicitações para o Hub ou para os servidores de usuário único. O Spawner é o objeto que inicia e para os servidores de usuário único sob demanda. O Authenticator é a classe que gerencia como os usuários podem acessar o sistema [10]. [10]Um modelo conceitual do JupyterHub descrevendo ambientes, hub, interfaces, autenticação e usuários é mostrado em Figura 2.2.

Figura 2.2 – Modelo conceitual do JupyterHub



Fonte: [11]

Sua arquitetura permite que ele seja implantado em qualquer infraestrutura, incluindo infraestrutura de nuvem comercial, VMs e hardware de rede interna, tornando-o personalizável, portátil, flexível e pode escalar até milhares de usuários com o Kubernetes [9], [10], [12], [13]. Ele permite que os usuários personalizem o servidor de notebook, a interface do usuário, a autenticação e o Spawner [10], [12]. O JupyterHub em si é gratuito e de código aberto, fornecendo mais controle sobre a configuração, hardware, dados e ambiente de codificação, no entanto, o custo de implantação do JupyterHub depende do provedor de nuvem ou hardware usado e do número de usuários e recursos necessários.

O JupyterHub geralmente é usado para permitir a colaboração entre equipes pequenas e grandes, como laboratórios de pesquisa, grupos de ciência de dados ou configurações educacionais [14]. Ao contrário do SciServer, o JupyterHub não fornece computação ou armazenamento de dados/infraestrutura de banco de dados, mas se concentra no gerenciamento de sessões de usuários e no controle da infraestrutura de computação [9], [13]. Esta escolha de design dá-lhe uma posição única no panorama das plataformas de ciência de dados [9], [13].

Ambas as plataformas têm seus pontos fortes e limitações únicas. Enquanto o SciServer oferece uma plataforma abrangente com recursos integrados de armazenamento e

computação, o JupyterHub fornece um ambiente mais flexível e personalizável. O projeto do SciServer foi construído para ser completo, escalável, modular, portátil e interoperável, tornando-o significativo na ciência orientada a dados. Em contraste, a natureza de código aberto do JupyterHub e a compatibilidade com outras ferramentas de ciência de dados o tornam uma escolha versátil para vários ambientes de computação e interfaces de usuário.

A comparação entre SciServer e JupyterHub neste projeto fornecerá uma análise aprofundada e iluminará suas características únicas, arquitetura, experiências do usuário, pontos fortes, limitações e aplicações potenciais em vários contextos científicos e educacionais de ciência de dados. Enquanto o SciServer oferece uma plataforma abrangente com recursos integrados de armazenamento e computação, o JupyterHub fornece um ambiente mais flexível e personalizável. Ao explorar as aplicações potenciais dessas plataformas em vários contextos científicos e educacionais, este projeto servirá como um recurso valioso para a comunidade mais ampla de pesquisadores, cientistas e educadores.

3 ESTUDO COMPARATIVO E IMPLEMENTAÇÃO DE PLATAFORMAS CIENTÍFICAS

Plataformas de análise e colaboração de dados ou plataformas de ciência de dados são ferramentas de software que fornecem um ambiente unificado onde os usuários podem acessar, analisar, visualizar e compartilhar dados, promovendo a colaboração e aumentando a produtividade. São ferramentas essenciais que permitem que pesquisadores, cientistas de dados, analistas e profissionais de várias áreas trabalhem juntos em projetos complexos orientados por dados. Essas plataformas geralmente fornecem recursos como controle de versão, segurança, gerenciamento de equipe, armazenamento e processamento, permitindo um fluxo de trabalho mais eficiente. O tipo e a quantidade desses recursos (recursos) a que um usuário/equipe terá acesso depende da plataforma e do plano de assinatura (gratuito/pago).

As plataformas de análise e colaboração de dados podem ajudar pesquisadores, cientistas, empresas, pequenas equipes e indivíduos a aproveitar os dados para vários fins, como relatórios, previsão, otimização, simulação e tomada de decisões. Algumas das principais características dessas plataformas são:

- Integração com ferramentas e bibliotecas: A maioria das plataformas oferece suporte à integração com ferramentas populares de análise de dados, bibliotecas e linguagens de programação, como Python, R, SQL e muito mais. Isso permite que os usuários aproveitem as habilidades e os recursos existentes.
- Ambiente colaborativo: recursos de colaboração, como blocos de anotações compartilhados, edição em tempo real e comentários, permitem que os membros da equipe trabalhem juntos sem problemas, independentemente de sua localização geográfica.
- Escalabilidade e flexibilidade: Essas plataformas geralmente fornecem recursos de computação escalonáveis, permitindo que os usuários lidem com grandes conjuntos de dados e cálculos complexos. Eles podem ser implantados no local, na nuvem ou em ambientes híbridos.
- Segurança e conformidade: garantir a privacidade e a segurança dos dados é uma prioridade, com recursos como controle de acesso, criptografia e conformidade com os padrões do setor.

- Visualização e Relatórios: Ferramentas avançadas de visualização ajudam na representação de dados de maneiras significativas, auxiliando na interpretação e tomada de decisões.
- Comunidade e suporte: Muitas plataformas têm suporte ativo da comunidade, fornecendo fóruns, documentação, tutoriais e assistência especializada.

Existem muitas dessas plataformas disponíveis oferecendo recursos e recursos diferenciados a diferentes preços.

Neste capítulo, uma análise comparativa de várias plataformas de ciência de dados é conduzida, avaliando suas características, capacidades e limitações. As descrições das várias plataformas estudadas são dadas em 3.1.

3.1 Descrição de algumas plataformas

Esta análise abrangente examina um espectro de plataformas de ciência de dados, cada uma oferecendo características, aplicações e adequação distintas para várias atividades científicas. Do amplamente reconhecido Kaggle e Google Colab a ambientes especializados como SciServer, GitHub Codespaces, JupyterHub, Posit Cloud, Replit, DataCamp Workspace, Deepnote, JetBrains Datalore e RStudio Server, exploramos as características diferenciadas que atendem às necessidades multifacetadas de cientistas de dados, pesquisadores e alunos.

3.1.1 Kaggle

O Kaggle, conhecido por sediar competições de ciência de dados, não apenas fornece um vasto campo de aprendizado, mas também opera como uma plataforma colaborativa. Utilizando Jupyter Notebooks no navegador, o Kaggle facilita a hospedagem de dados e recursos de compartilhamento [3], [15]. Além disso, o Kaggle oferece instâncias com GPUs e TPUs gratuitamente, aumentando seu apelo para entusiastas de aprendizado de máquina.

3.1.2 Google Colab

O Google Colab, um ambiente de notebook Jupyter hospedado, se diferencia por oferecer acesso gratuito a recursos de computação, incluindo GPUs e TPUs. Sua natureza baseada em nuvem, juntamente com uma interface de navegador direta, torna-o particularmente

conveniente para análise de dados, aprendizado de máquina e fins educacionais [8], [9], [13], [16], [17].

3.1.3 SciServer

O SciServer, projetado para análise do lado do servidor com extensos conjuntos de dados, destaca-se como um sistema de ciberinfraestrutura totalmente integrado. Suportando análise interativa e em modo batch usando linguagens como Python, o SciServer capacita os pesquisadores a trabalharem perfeitamente com conjuntos de dados científicos massivos dentro da plataforma [1], [18]. Sua abordagem única permite que os cientistas manipulem terabytes ou petabytes de dados sem a necessidade de downloads extensivos.

3.1.4 GitHub Codespaces

O GitHub Codespaces introduz um ambiente de desenvolvimento hospedado na nuvem, oferecendo aos desenvolvedores a flexibilidade de editar e executar código diretamente do navegador ou IDEs locais. Suportando várias linguagens e ambientes, o Codespaces fornece um ambiente de codificação seguro e colaborativo [19], [20]. Sua capacidade de se conectar a redes privadas aumenta ainda mais sua utilidade para vários cenários de desenvolvimento.

3.1.5 JupyterHub

Como parte da colaboração do Projeto Jupyter, o JupyterHub facilita a criação de um Hub multiusuário, gerenciando e fazendo proxy de várias instâncias do servidor de notebook Jupyter de usuário único. Sua flexibilidade na implantação de infraestrutura permite fluxos de trabalho colaborativos de ciência de dados, tornando-se um ativo valioso para projetos de equipe e cursos acadêmicos [21].

3.1.6 Posit Cloud

O Posit Cloud, construído no RStudio IDE, surge como uma plataforma leve e baseada em nuvem para educação e colaboração em ciência de dados. O uso de conteinerização garante a execução segura e independente das saídas, tornando-se um ambiente ideal para ensino e aprendizagem. A interface gráfica do Posit Cloud simplifica a colaboração e o dimensionamento de recursos [22], [23]

3.1.7 Replit

Replit destaca-se como uma plataforma versátil para criação e compartilhamento de software, suportando mais de cinquenta linguagens de programação. Com uma interface baseada em navegador e recursos de colaboração em tempo real, o Replit atende a um público amplo. Seus múltiplos planos de uso, de Free a Business, acomodam necessidades diversas com recursos disponíveis variados [24]–[27].

3.1.8 DataCamp Workspace

O DataCamp Workspace, um bloco de anotações online colaborativo baseado em nuvem, fornece um ambiente para experimentar código, analisar dados e compartilhar insights. Com suporte a R, Python e SQL, o DataCamp Workspace vem com conjuntos de dados prontos para uso, facilitando a transição do aprendizado para a aplicação de habilidades de ciência de dados [28]–[31]. Sua ferramenta no navegador requer zero instalação e download, garantindo acessibilidade em diferentes sistemas operacionais.

3.1.9 Deepnote

O Deepnote, um caderno de ciência de dados on-line gratuito, coloca uma forte ênfase na colaboração em tempo real para Python, SQL e análise sem código. Sua natureza baseada em nuvem permite conexões seguras com várias fontes de dados, e a integração com serviços de nuvem populares aumenta sua versatilidade [32]–[34]. Cada usuário recebe uma página web dedicada, aprimorando a utilidade da plataforma para criação de portfólio e compartilhamento de projetos.

3.1.10 JetBrains Datalore

O JetBrains Datalore, compatível com Jupyter e que oferece assistência de codificação inteligente, suporta Python, Kotlin, Scala e R. Com gerenciamento integrado de ambiente, controle de versão interno e armazenamento, o Datalore atua como uma plataforma colaborativa de ciência de dados. Disponível em nuvem gerenciada e configurações de hospedagem privada, o Datalore atende a uma base diversificada de usuários [35]–[38].

3.1.11 RStudio Server

O RStudio Server facilita o acesso remoto ao R por meio de uma interface baseada em navegador, suportando compartilhamento de código, projetos colaborativos e uso eficiente de recursos de computação. Disponível nas versões open Source e Professional

Edition, o RStudio Server adiciona recursos aprimorados como várias sessões R simultâneas, controle de versão e um painel administrativo para monitorar métricas de desempenho [22], [23], [39], [40]. Sua capacidade de centralizar instalações e configurações do R simplifica o processo de desenvolvimento para usuários em vários locais.

As plataformas de análise e colaboração de dados desempenham um papel vital na pesquisa científica e na educação:

- Facilitar a colaboração interdisciplinar.
- Acelerando o processo de pesquisa através de automação e código reutilizável.
- Permitir o acesso e o compartilhamento de grandes conjuntos de dados e recursos de computação de alto desempenho.
- Melhorar a qualidade, a reprodutibilidade e a transparência na pesquisa.
- Realização de análise e visualização de dados complexos utilizando diversas ferramentas e métodos.

As plataformas de análise e colaboração de dados estão na vanguarda da pesquisa científica moderna e das aplicações da indústria. Eles [41]–[43]capacitam as equipes a trabalharem juntas de forma eficiente, alavancar ferramentas e recursos avançados e contribuir para descobertas e inovações inovadoras. Plataformas como JupyterHub e SciServer exemplificam esses recursos, cada um atendendo a necessidades e domínios específicos, e seu estudo comparativo pode fornecer insights valiosos para usuários e organizações.

3.2 COMPARAÇÃO DE ALGUMAS PLATAFORMAS DE CIÊNCIA DE DADOS

Existem muitas plataformas e tecnologias de ciência de dados no mercado hoje. Cada fornecedor oferece recursos, serviços e tipos de recursos ou recursos diferentes e exclusivos. Alguns oferecem serviços e recursos gratuitamente, enquanto outros cobram taxas mensais/anuais ou oferecem opções de pagamento conforme o uso (PAYG) para usar seus recursos e serviços. Os recursos e funcionalidades diferem de plataforma para plataforma. Tabela 3.2 e Tabela 3.3 Listar várias plataformas e tabular suas características e serviços, e comparar preços. Abaixo estão as definições ou descrições de recursos/serviços que estão sendo comparados nas tabelas.

Para auxiliar na análise, diversas plataformas de ciência de dados baseadas em critérios-chave foram avaliadas. As plataformas consideradas incluem Kaggle, Google Colab, SciServer, GitHub Codespaces, JupyterHub, Posit Cloud, Replit, DataCamp Workspace, Deepnote, JetBrains Datalore e RStudio Server.

Os recursos e serviços analisados, incluindo acesso baseado na Web, implantação de infraestrutura, preços, recursos de computação e muito mais, são definidos em Tabela 3.1.

Tabela 3.1 – Definições/descrições de recursos e serviços

Recurso/Serviço	**Descrição**
Baseado na Web	Permite edição, execução e compartilhamento de notebooks remotamente através de uma interface de navegador. Os usuários acessam recursos com qualquer dispositivo e conexão com a internet.
Implantação de infraestrutura	Local: hardware, software e funções alojados nas instalações da organização. Nuvem: Serviços hospedados por um provedor de serviços de nuvem externo. Híbrido: uma combinação de serviços locais e de nuvem pública.
Precificação	Planos/camadas gratuitos ou pagos determinam o acesso a recursos como GPU, CPU, tempo de computação, armazenamento, colaboração etc.
Modelos de Pagamento	Assinatura: Taxas mensais para acesso. PAYG: Pagar pelos serviços/recursos utilizados.
Recurso de computação/computação	Engloba poder de processamento, memória, rede, armazenamento, necessário para computação.
Limites da CPU	Especifica o número de CPUs ou vCPUs disponíveis para executar programas. Pode depender do plano ou processo.
Limites da GPU	Especifica o acesso a unidades de processamento gráfico (GPUs). Camadas mais altas geralmente fornecem acesso a GPUs mais poderosas.
CARNEIRO	A Memória de Acesso Aleatório armazena temporariamente dados para o tempo de execução do programa. Limitado pelo plano subscrito.
Limites de armazenamento	Alocação de espaço de armazenamento na nuvem para fácil acesso e compartilhamento de dados digitais. Varia de acordo

	com os planos, de uma fração de gigabyte a armazenamento ilimitado.
Acesso aos dados	Diversas maneiras pelas quais os cientistas de dados se conectam aos dados, incluindo arquivos simples, RDBs, GitHub, provedores de nuvem, bancos de dados SQL, etc. A acessibilidade pode depender do plano subscrito.
Linguagens/Kernel	Número de linguagens de programação ou kernels suportados pela plataforma. Varia de suporte a um único idioma a vários idiomas.
Compartilhamento	As plataformas permitem o compartilhamento de código e dados criando links ou URLs para que outras pessoas acessem. Pode definir níveis de acesso, como edição ou visualização.
Colaboração	Recursos que permitem que mais de uma pessoa trabalhe em um bloco de anotações ou projeto. O nível de colaboração varia, desde membros da equipe até qualquer pessoa com um link. A colaboração em tempo real pode estar disponível.
Controle de versão/código-fonte	Prática de rastreamento e gerenciamento de alterações no código. Pode usar Git ou outros sistemas. Algumas plataformas oferecem controle ilimitado do código-fonte, enquanto outras o limitam com base no nível de assinatura.
Tipo de plataforma	Aberto: Fornece flexibilidade para os usuários escolherem linguagens e ferramentas de programação. Fechado: os usuários devem aderir à linguagem específica da plataforma, às ferramentas de GUI e aos pacotes de modelagem do fornecedor.
Código aberto	Indica se o código da plataforma é acessível publicamente (código aberto) ou proprietário.
Configuração	Refere-se à configuração do servidor para plataformas como JupyterHub, SciServer, RStudio auto-hospedado e Web auto-hospedado.
Uso	Descreve a facilidade de uso, normalmente envolvendo interfaces gráficas e recursos de computação pré-alocados.

FONTE: Autor

Tabela 3.2 – Comparação de algumas funcionalidades e serviços de plataformas de ciência de dados selecionadas atualmente.

Plataforma	Baseado na Web	Implantação de infraestrutura	Precificação	Modelos de Pagamento	Recurso de computação/computação	Limites da CPU	Limites da GPU	CARNEIRO	Limites de armazenamento
Kaggle	Sim	Nuvem	Mix de Grátis e Pago	Assinatura, PAYG	Computação, GPU, TPU	CPUs virtuais	Acesso Limitado	Sim	Sim
Google Colab	Sim	Nuvem	Livre	PAGAMENTO	Computação, GPU, TPU	CPUs virtuais	Acesso Limitado	Sim	Sim
SciServer	Sim	Combinação de local e nuvem	Grátis, pago	Assinatura, PAYG	Computação, GPU	CPUs virtuais	Solicitação formal	Sim	Sim
Espaços de código do GitHub	Sim	Nuvem	Mix de Grátis e Pago	PAGAMENTO	Calcular	CPUs virtuais	Sem acesso	Sim	Sim
JupyterHub	Sim	Combinação de local e nuvem	Grátis, pago	Subscrição	Calcular	CPUs virtuais	Acesso Limitado	Sim	Sim
Posit Cloud	Sim	Nuvem	Mix de Grátis e Pago	Assinatura, PAYG	Computação, GPU	CPUs virtuais	Acesso Limitado	Sim	Sim
Replit	Sim	Nuvem	Mix de Grátis e Pago	PAGAMENTO	Calcular	CPUs virtuais	Acesso Limitado	Sim	Sim
Espaço de trabalho do DataCamp	Sim	Nuvem	Grátis, pago	Subscrição	Calcular	CPUs virtuais	Sem acesso	Sim	Sim

Nota profunda	Sim	Nuvem	Livre	PAGAMENTO	Calcular	CPUs virtuais	Acesso Limitado	Sim	Sim
JetBrains Datalore	Sim	Combinação de local e nuvem	Grátis, Pago	Assinatura, PAYG	Computação, GPU	CPUs virtuais	Solicitação formal	Sim	Sim
Servidor RStudio	Sim	Combinação de local e nuvem	Grátis, Pago	Subscrição	Calcular	CPUs virtuais	Sem acesso	Sim	Sim

FONTE: Autor

Tabela 3.3 – Comparação de alguns recursos e serviços de plataformas de ciência de dados selecionadas hoje – continuou (mais plataformas).

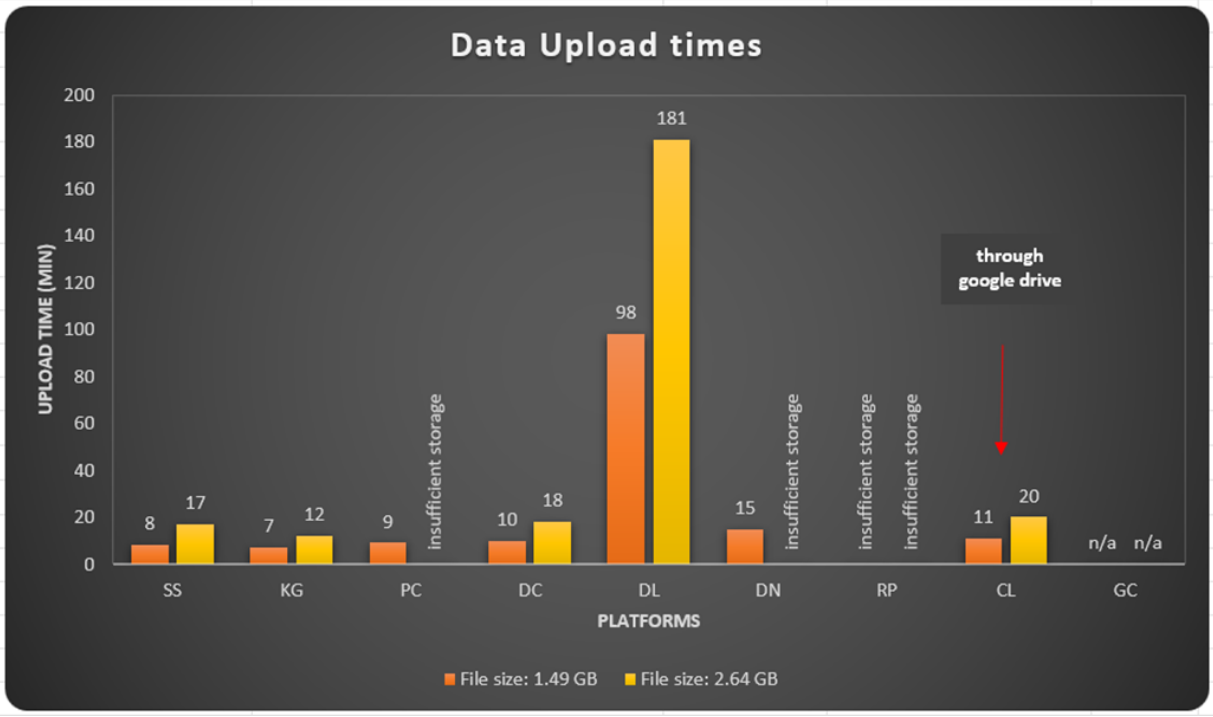
Plataforma	**Acesso aos dados**	**Linguagens/Kernel**	**Compartilhamento**	**Colaboração**	**Controle de versão/código-fonte**	**Tipo de plataforma**	**Código aberto**	**Configuração**	**Uso**
Kaggle	Fontes Diversas	Python, R	Sim	Sim (limitado)	Sim (Git)	Fechado	Não	N/A	GUI, Notebooks
Google Colab	Fontes Diversas	Pitão	Sim	Sim (em tempo real)	Sim (Git)	Fechado	Não	N/A	GUI, Notebooks
SciServer	Fontes Diversas	Python, R, SQL	Sim	Sim (em tempo real)	Sim (Git)	Abrir	Sim	Requer a configuração do servidor	GUI, Notebooks, CLI
Espaços de código do GitHub	GitHub, Git	Múltiplo	Sim	Sim (em tempo real)	Sim (Git)	Fechado	Não	N/A	VS Code, Notebooks

JupyterHub	Fontes Externas	Múltiplo	Sim	Sim (em tempo real)	Sim (Git)	Abrir	Sim	Requer a configuração do servidor	Cadernos, CLI
Posit Cloud	Fontes Externas	R	Sim	Sim (em tempo real)	Sim (Git)	Fechado	Não	N/A	GUI, Notebooks
Replit	Fontes Externas	Múltiplo	Sim	Sim (em tempo real)	Sim (Git)	Fechado	Não	N/A	GUI, Notebooks
Espaço de trabalho do DataCamp	Fontes Externas	R, Python, SQL	Sim	Sim (limitado)	Sim (Git)	Fechado	Não	N/A	GUI, Notebooks
Nota profunda	Serviços em Nuvem	Python, SQL	Sim	Sim (em tempo real)	Sim (Git)	Fechado	Não	N/A	GUI, Notebooks
JetBrains Datalore	Fontes Externas	Python, Kotlin, Scala, R	Sim	Sim (em tempo real)	Sim (Git)	Fechado	Não	N/A	GUI, Notebooks
Servidor RStudio	Fontes Externas	R	Sim	Sim (limitado)	Sim (Git)	Fechado	Não	Requer a configuração do servidor	GUI, Notebooks

FONTE: Autor

Na comparação dessas plataformas, outros testes foram realizados para avaliar seus desempenhos e capacidades entre si. Entre esses testes incluem carregar arquivos e comparar seus tempos de upload e se eles suportam esse tipo de arquivo para começar. Outro teste foi rodar códigos python na plataforma para comparar as velocidades de execução. Figure 3.1 mostra um gráfico de barras comparando os tempos de upload de arquivos em minutos de dois arquivos de tamanhos 1,5 e 2,6 GB .nc e arquivos de imagem .tiff, respectivamente. O n/a significa que os arquivos não puderam ser carregados na plataforma. Outras plataformas forneceram armazenamento suficiente para o upload, outras ainda não suportaram os tipos de arquivo.

Figure 3.1 – Comparando os tempos de upload de dados em minutos



FONTE: Autor

4 ESTUDO COMPARATIVO ENTRE JUPYTERHUB E SCISERVER

No cenário em constante evolução das plataformas de ciência de dados, escolher o ambiente certo é crucial para a colaboração eficaz, a utilização de recursos e o sucesso do projeto. Esta seção se aprofunda em um estudo comparativo abrangente com foco em duas plataformas proeminentes: JupyterHub e SciServer. O JupyterHub, conhecido por sua natureza aberta e flexível, contrasta com o SciServer, conhecido por suas capacidades robustas em servir comunidades científicas. Essa exploração visa desvendar as características, funcionalidades e atributos de desempenho distintivos dessas plataformas, fornecendo insights valiosos para pesquisadores, cientistas de dados e tomadores de decisão que buscam o ambiente de ciência de dados ideal para seus projetos.

4.1 JupyterHub: arquitetura, recursos e experiências do usuário

O JupyterHub é um ambiente multiusuário que permite aos usuários interagirem com notebooks Jupyter e outras ferramentas baseadas no Jupyter. Ele permite que os usuários executem vários notebooks Jupyter em um único servidor, com cada usuário tendo seu próprio ambiente isolado. Ele é frequentemente usado em ambientes educacionais e de pesquisa para fornecer uma plataforma centralizada para vários usuários trabalharem em projetos de análise, programação e pesquisa de dados. Aqui está uma visão geral de seus recursos:

- Ele suporta vários métodos de autenticação, como OAuth, GitHub, LDAP, etc [10].
- Ele pode ser implantado em diferentes infraestruturas, como nuvem, Kubernetes, Docker, etc [10].
- Ele pode ser personalizado com diferentes kernels, interfaces de usuário e extensões [10]
- Ele pode ser ampliado para cima ou para baixo para atender a demanda dos usuários [14].
- Ele pode ser integrado a outras ferramentas e serviços, como bancos de dados, sistemas de arquivos e armazenamento em nuvem [12].

- permite que vários usuários tenham seus próprios ambientes isolados para executar notebooks Jupyter e outras ferramentas de computação interativas.
- suporta vários métodos de autenticação, permitindo que as instituições se integrem com seus sistemas de gerenciamento de usuários existentes.
- O spawner permite que os administradores definam como os ambientes de usuário são criados, como o uso de contêineres ou máquinas virtuais.
- Contas de usuário, permissões e recursos podem ser gerenciados centralmente.
- pode ser dimensionado para acomodar um grande número de usuários, tornando-o adequado para ambientes educacionais e de pesquisa.
- O ambiente de cada usuário é separado, proporcionando segurança e evitando interferências entre os processos dos usuários.
- pode gerenciar recursos para garantir a alocação justa de poder de computação entre os usuários.

A arquitetura JupyterHub foi projetada para fornecer a cada usuário de um grupo um servidor Jupyter Notebook. Para conseguir isso, a arquitetura usa os três subsistemas principais a seguir:

- Hub: é o componente principal que gerencia a autenticação do usuário, a geração e o gerenciamento de sessões de usuários individuais e o roteamento de solicitações para as instâncias corretas do servidor Notebook. Ele interage com o autenticador do usuário para gerenciar o login e o gerenciamento de usuários.
- Proxy: projetado para rotear solicitações de usuários para seus próprios servidores de notebook e gerenciar a autenticação.
- Servidor de notebook de usuário único: projetado para executar o código do usuário e fornecer a interface do usuário.
- Autenticador: O autenticador é responsável por autenticar usuários e determinar suas permissões. Ele pode ser configurado para integrar com vários métodos de autenticação, como LDAP, OAuth, GitHub, Google etc.

- Spawner: O spawner cria e gerencia as instâncias individuais do servidor Notebook do usuário. Cada usuário tem seu próprio ambiente isolado para trabalhar. O spawner pode ser configurado para criar contêineres (Docker, Kubernetes), máquinas virtuais ou outros ambientes de computação.
- Instâncias do servidor Notebook: cada usuário tem sua própria instância isolada do servidor Jupyter Notebook, onde pode criar e gerenciar blocos de anotações Jupyter, executar código e executar análise de dados.

O JupyterHub tem sido usado por muitas organizações e comunidades para vários fins. Alguns exemplos são:

- O curso Data 8 na UC Berkeley usa o JupyterHub para ensinar ciência de dados a mais de 1.000 alunos por semestre [44]
- O projeto Pangeo usa o JupyterHub para fornecer uma plataforma escalável para pesquisa e educação em geociências.
- O projeto Binder usa JupyterHub para permitir que os usuários criem e compartilhem ambientes computacionais reproduzíveis a partir de repositórios do GitHub.

4.2 SciServer: Arquitetura, Recursos e Experiências do Usuário

O SciServer é uma plataforma científica que consiste em vários componentes e ferramentas que trabalham em conjunto para fornecer um ambiente colaborativo para análise do lado do servidor com conjuntos de dados extremamente grandes. A arquitetura do sistema do SciServer pode ser resumida da seguinte forma:[1]

- Portal SciServer: O principal ponto de entrada para os usuários acessarem os recursos e ferramentas do SciServer. Ele também fornece serviços de autenticação e autorização de usuário.
- SciServer Compute: O componente que permite aos usuários criarem e executar blocos de anotações Jupyter, sessões RStudio e scripts de linha de comando usando contêineres do Docker. Ele também fornece acesso a volumes de dados públicos e volumes de dados do usuário.[1][45]

- SciServer CasJobs: O componente que permite aos usuários consultarem bancos de dados usando SQL e armazenar os resultados em tabelas MyDB. Ele também fornece acesso ao SkyQuery, um serviço de correspondência cruzada para catálogos astronômicos.[1][45]
- Arquivos SciServer: O componente que permite aos usuários carregarem, baixar e gerenciar arquivos no sistema de arquivos do SciServer. Ele também fornece acesso a pastas de arquivos públicos e pastas de arquivos de usuário.[45]
- SciServer Groups: O componente que permite aos usuários criarem ou ingressar em grupos e compartilhar seus recursos com seus colaboradores de forma privada. Ele também fornece um modo de exibição de grupo que lista todos os recursos compartilhados entre os membros do grupo.[45]

4.3 ANÁLISE COMPARATIVA E ACHADOS

Esta seção apresenta uma análise comparativa detalhada das arquiteturas do JupyterHub e do SciServer, destacando seus respectivos pontos fortes e limitações. O impacto das diferenças de arquitetura na escalabilidade, no gerenciamento de recursos e no desempenho é examinado. O JupyterHub e o SciServer são ambientes de pesquisa colaborativos para ciência orientada a dados em larga escala. Eles têm algumas diferenças e semelhanças em suas características e funcionalidades. Aqui estão alguns deles:[14]
O JupyterHub é um serviço de código aberto que cria servidores de notebook Jupyter baseados em nuvem sob demanda, enquanto o SciServer é um sistema de ciberinfraestrutura totalmente integrado que engloba ferramentas e serviços relacionados para permitir que os pesquisadores lidem com big data científico.[46][18]

JupyterHub suporta vários kernels que permitem aos usuários trabalharem com diferentes linguagens de programação e frameworks. O SciServer fornece imagens de computação, que são máquinas virtuais gratuitas que os usuários podem personalizar com seus próprios pacotes de software.[14]

O JupyterHub permite que os usuários acessem o JupyterLab ou o Jupyter Notebooks remotamente sem instalar software em sua máquina local. O SciServer permite que os usuários trabalhem com terabytes ou petabytes de dados científicos sem a necessidade de baixar grandes conjuntos de dados.

O JupyterHub é focado principalmente em fornecer um ambiente de computação comum para os usuários através de qualquer navegador da web. O SciServer também está focado em fornecer acesso a conjuntos de dados hospedados de várias disciplinas, como astronomia, biologia, oceanografia etc.

O JupyterHub tem distribuições diferentes projetadas para diferentes cenários de implantação, como o The Littlest JupyterHub (TLJH) para implantações menores em uma única máquina virtual ou Zero to JupyterHub com Kubernetes (Z2JH) para implantações maiores em um cluster de máquinas. O SciServer não tem distribuições diferentes, mas fornece um serviço de logon único que permite que os usuários façam login com sua instituição de ensino ou conta do Google.[47][12], [46]

O JupyterHub fornece conectividade que permite que os usuários se conectem com a infraestrutura necessária para suas sessões, como GPUs, CPUs, memória etc. O SciServer também fornece conectividade, mas também oferece serviços de armazenamento que permitem aos usuários carregarem seus próprios dados ou acessar conjuntos de dados públicos hospedados pelo SciServer.

O JupyterHub permite o acesso remoto ao JupyterLab, bem como aos Jupyter Notebooks, que são ambientes interativos baseados na Web para escrever e executar código. O SciServer também permite o acesso remoto ao Jupyter Notebooks, mas também suporta outras ferramentas, como o CasJobs, que é uma interface baseada na web para consultar grandes bancos de dados.

4.3.1 Cenários de uso

Os cenários de uso do JupyterHub e do SciServer refletem sua aplicabilidade mais ampla na colaboração em educação e pesquisa (JupyterHub) e o foco especializado em astronomia e pesquisa em ciência de dados (SciServer). Os recursos de cada plataforma se alinham com as necessidades exclusivas de seus usuários e cenários pretendidos. Tabela 4.1 fornece alguns cenários de uso do JupyterHub e SciServer.

Tabela 4.1 – Alguns casos de uso do JupyterHub e SciServer.

Cenários de uso	**JupyterHub**	**SciServer**
Educação	Comumente utilizado em instituições de ensino para ensino de programação e codificação colaborativa.	Permite que educadores de astronomia e ciência de dados forneçam aos alunos acesso a ferramentas e conjuntos de dados especializados.
Colaboração em Pesquisa	Facilita a codificação colaborativa e o trabalho de projeto entre pesquisadores em várias áreas.	Oferece uma plataforma colaborativa para astrônomos e cientistas de dados conduzirem pesquisas, analisarem dados e colaborarem em projetos.
Ambiente Multiusuário	Fornece um ambiente compartilhado para vários usuários trabalharem em projetos separados simultaneamente.	Suporta acesso multiusuário a conjuntos de dados astronômicos, permitindo que os pesquisadores trabalhem juntos dentro de uma plataforma unificada.
Ambientes Customizados	Permite que os administradores configurem ambientes personalizados para projetos, cursos ou equipes de pesquisa específicos.	Fornece um ambiente sob medida otimizado para a pesquisa em astronomia, oferecendo ferramentas e recursos especializados.
Análise e Exploração	Suporta análise, visualização e exploração de dados por meio de blocos de anotações Jupyter.	Permite que os pesquisadores analisem dados astronômicos, realizem simulações
Grandes conjuntos de dados	Pode gerenciar grandes conjuntos de dados, mas o armazenamento e o gerenciamento de dados podem exigir configuração adicional.	Oferece armazenamento de dados em escala de petabytes, ideal
Ferramentas Especializadas	Oferece uma ampla gama de ferramentas de programação e análise de dados de uso geral	Fornece ferramentas e recursos especializados relevantes para a pesquisa em astronomia
Projetos Colaborativos	Codificação colaborativa e compartilhamento de cadernos para projetos/pesquisas em grupo.	Pesquisa colaborativa usando conjuntos de dados compartilhados, ferramentas de análise
Fluxos de trabalho científicos	Criação e compartilhamento de fluxos de trabalho científicos reproduzíveis usando cadernos.	Suporta a criação de fluxos de trabalho científicos personalizados

Pesquisa em Astronomia	Embora não seja específico da astronomia, pode ser adaptado para vários domínios de pesquisa.	Projetado explicitamente para pesquisa em astronomia, oferecendo acesso direto a conjuntos de dados astronômicos e ferramentas específicas de domínio.

FONTE: Autor

4.3.2 Custo e manutenção

O SciServer é gratuito em seu nível básico para qualquer pessoa que se registre. O SciServer oferece um número limitado de recursos gratuitos para cada usuário, como 10 GB de espaço em disco, 2 GB de RAM e 4 núcleos de CPU. Provedores de dados e outros usuários com maiores requisitos de recursos, como placas de vídeo ou nós para computação científica exclusiva com armazenamento extra de arquivos ou banco de dados, podem hospedá-los no SciServer ou implantar uma instância do SciServer em seu próprio ambiente. O JupyterHub é de código aberto e está disponível gratuitamente. No entanto, o custo de implantação do JupyterHub pode variar significativamente dependendo de fatores como o ambiente de hospedagem escolhido, os recursos computacionais necessários, as necessidades de armazenamento de dados e o tamanho da base de usuários.

Tanto o JupyterHub quanto o SciServer têm estruturas de custos diferentes. Os custos associados ao JupyterHub estão principalmente relacionados à configuração e manutenção da infraestrutura, enquanto os custos do SciServer abrangem recursos aprimorados, armazenamento de dados, personalização, suporte e treinamento. Ao implementar essas plataformas, as organizações devem considerar cuidadosamente as despesas associadas à infraestrutura de nuvem, espaço de armazenamento e quaisquer licenças de software adicionais ou serviços de suporte necessários.

Para garantir a eficiência de custos, é essencial realizar uma análise completa dos padrões de uso e requisitos de escalabilidade, permitindo a alocação eficaz de recursos e estratégias de otimização de custos. A escolha entre os dois depende das necessidades dos usuários e considerações orçamentárias. Tabela 4.2 fornece alguns dos custos associados à instalação, implementação, manutenção e uso do JupyterHub e do SciServer.

Tabela 4.2 - Custo de implementação, manutenção e uso do JupyterHub.

Custar	JupyterHub	SciServer
Custo do software	open-source e disponível gratuitamente.	O acesso básico ao SciServer é gratuito, oferecendo ferramentas e recursos limitados.
Custo de Infraestrutura	Envolve custos relacionados à configuração, hospedagem e manutenção do servidor	fornecendo recursos e recursos aprimorados com base nas necessidades do usuário.
Custo de armazenamento de dados	Dependendo da solução de armazenamento escolhida e dos requisitos de capacidade.	
Custo da nuvem	depende do provedor de nuvem, do tipo de serviço, da região, do uso, etc.	depende do provedor de nuvem, do tipo de serviço, da região, do uso, etc.
Custo de personalização	custos de personalização e manutenção de ambientes de usuário, dependendo da complexidade das configurações.	Customização dentro do ambiente SciServer sem custos adicionais de customização.
Custo de manutenção	depende da frequência e complexidade das atualizações, patches, backups, etc.	O custo de manutenção também depende da frequência e complexidade das atualizações, patches, backups, etc.
Custo de treinamento	Custos de treinamento podem ser incorridos para que administradores e usuários entendam a configuração e as funcionalidades da plataforma.	O SciServer oferece recursos e materiais de treinamento como parte de seu serviço, reduzindo custos adicionais de treinamento.

FONTE: Autor

4.3.3 Fontes de dados e formatos suportados.

JupyterHub e SciServer são plataformas versáteis que suportam uma ampla gama de fontes de dados e formatos, tornando-os ferramentas inestimáveis para ciência de dados e pesquisa. Os usuários podem trabalhar perfeitamente com diversas fontes de dados, incluindo bancos de dados estruturados, arquivos CSV, JSON, XML e muito mais. Além disso, facilitam a integração com repositórios de dados científicos, permitindo o acesso a bancos de dados astronômicos, conjuntos de dados climáticos, informações genômicas e outras fontes de dados especializadas. Seu suporte para vários formatos de dados, como matrizes NumPy, Pandas DataFrames e bibliotecas de visualização interativa, garante a compatibilidade com ferramentas comuns de manipulação e análise de dados. Essa flexibilidade capacita os pesquisadores a explorarem e analisar eficientemente dados de múltiplos domínios, promovendo a colaboração interdisciplinar e acelerando as descobertas científicas.

O JupyterHub oferece suporte a uma variedade de back-ends de banco de dados via SQLAlchemy, como PostgreSQL, MySQL e SQLite. Ele também permite que os usuários acessem volumes de dados públicos hospedados na nuvem ou em seu próprio hardware, como Introdução, SDSS DR16 e LSST DC22. O JupyterHub pode executar blocos de anotações Jupyter contendo código em Python, R e MATLAB, bem como outras linguagens que possuem kernels Jupyter. Os Jupyter Notebooks usam o formato de arquivo .ipynb, que é um documento JSON que contém texto, código, saída e metadados. O JupyterHub também pode funcionar com outros formatos de arquivo, como .txt, .csv, .fits, .excel, .json, .xml e .hdf5. O JupyterHub permite que os notebooks Jupyter leiam e gravem dados de várias fontes, como arquivos locais, armazenamento em nuvem, bancos de dados, serviços web etc. JupyterHub também suporta dados de imagem na saída da célula e como arquivos em formatos como PNG, JPEG, GIF, SVG etc..

O SciServer fornece acesso a mais de dois petabytes de vários conjuntos de dados hospedados em domínios e disciplinas como astronomia, biologia, oceanografia, mecânica dos fluidos etc. Esses conjuntos de dados podem ser consultados ou acessados usando comandos SQL, APIs ou CasJobs, como SDSS DR16, Gaia DR2 e DESI DR92. O SciServer também oferece volumes de dados públicos que podem ser montados em contêineres criados pelos usuários, como Introdução, SDSS DR16 e LSST DC22. O SciServer usa o SciServer Compute para executar blocos de anotações Jupyter contendo código em Python ou R. O SciServer Compute também suporta o formato de arquivo .ipynb para notebooks Jupyter, bem como outros formatos de arquivo como .txt, .csv, .fits e .hdf5. O SciServer também permite que os usuários carreguem seus conjuntos de dados ou os importem de fontes externas, como Dropbox, Google Drive etc..

4.3.4 Ferramentas e ambientes de análise

O SciServer e o JupyterHub oferecem um amplo conjunto de ferramentas de análise e ambientes de computação versáteis, atendendo às diversas necessidades de cientistas de dados, pesquisadores e educadores. O SciServer fornece um ecossistema poderoso para análise de dados astronômicos, incluindo acesso integrado a ferramentas populares como Python, R e SQL, enquanto o JupyterHub oferece uma ampla variedade de kernels e bibliotecas para suportar a exploração e visualização interativa de dados. Ambas as plataformas possibilitam a criação de cadernos Jupyter, fomentando pesquisas reprodutíveis e colaborativas. Os usuários podem aproveitar ambientes pré-configurados

ou personalizar seus próprios, incorporando bibliotecas especializadas e pacotes de software. Além disso, com suporte para GPU e computação paralela, essas plataformas facilitam simulações e cálculos complexos, tornando-os recursos indispensáveis para investigações científicas que abrangem uma infinidade de domínios.

O JupyterHub oferece suporte a uma variedade de back-ends de banco de dados via SQLAlchemy, como PostgreSQL, MySQL e SQLite. Ele também permite que os usuários acessem volumes de dados públicos hospedados na nuvem ou em seu próprio hardware. O JupyterHub pode executar blocos de anotações Jupyter contendo código em Python, R e MATLAB, bem como outras linguagens que possuem kernels Jupyter. O JupyterHub pode ser usado para atender a uma variedade de interfaces de usuário, incluindo o Jupyter Notebook, Jupyter Lab., RStudio, nteract e muito mais. O JupyterHub pode ser implantado com tecnologia de contêiner moderna, como o Kubernetes, permitindo que ele seja dimensionado e mantido de forma eficiente para um grande número de usuários.

O SciServer usa uma arquitetura baseada em Docker/VM para fornecer análise interativa e em lote do lado do servidor com linguagens de script como Python e R em vários ambientes, incluindo Jupyter (notebooks), RStudio e linha de comando, além da análise de dados tradicional baseada em SQL. O SciServer hospeda mais de dois petabytes de dados científicos em uma variedade de disciplinas, como astronomia e mecânica dos fluidos. Ele fornece acesso a bancos de dados que podem ser consultados com comandos SQL ou CasJobs, bem como volumes de dados públicos que podem ser montados em contêineres criados pelos usuários.

4.3.5 Recursos de colaboração e comunicação

O JupyterHub e o SciServer promovem a colaboração e a comunicação entre pesquisadores e cientistas de dados, fornecendo recursos robustos para compartilhar e interagir com dados e análises. Com os blocos de anotações Jupyter em sua essência, ambas as plataformas permitem que os usuários criem e compartilhem documentos interativos que combinam código, texto e visualizações. Esses documentos podem ser facilmente compartilhados com colegas ou colaboradores, fomentando pesquisas transparentes e reprodutíveis. O SciServer oferece um espaço de trabalho colaborativo, permitindo que os usuários compartilhem dados, códigos e análises com colegas astrônomos e cientistas dentro de sua comunidade. Além disso, ambas as plataformas

suportam a colaboração em tempo real por meio da integração com sistemas de controle de versão, como o Git, e facilitam a comunicação entre as equipes de pesquisa por meio de fóruns de discussão e funcionalidade de bate-papo. Esses recursos colaborativos e comunicativos são fundamentais para acelerar o ritmo da descoberta científica e promover esforços interdisciplinares de pesquisa.

O JupyterHub suporta colaboração em tempo real (RTC), onde vários usuários podem trabalhar com o mesmo servidor Jupyter e ver as edições uns dos outros. Os usuários também podem compartilhar cópias estáticas de blocos de anotações por um link. O JupyterHub pode ser usado para atender a uma variedade de interfaces de usuário, como Jupyter Notebook, Jupyter Lab., RStudio, nteract e muito mais. O JupyterHub pode ser implantado com tecnologia de contêiner moderna, como o Kubernetes, permitindo que ele seja dimensionado e mantido de forma eficiente para muitos usuários.

O SciServer usa uma arquitetura baseada em Docker/VM para fornecer análise interativa e em lote do lado do servidor com linguagens de script como Python e R em vários ambientes, incluindo Jupyter (notebooks), RStudio e linha de comando. O SciServer hospeda mais de dois petabytes de dados científicos em uma variedade de disciplinas, como astronomia e mecânica dos fluidos. O SciServer permite equipes interdisciplinares globais e facilita o compartilhamento de conhecimento dentro da comunidade. O SciServer permite que os usuários criem grupos e compartilhem arquivos, conjuntos de dados, imagens de computação etc. com outros membros do grupo. Os usuários também podem comentar em arquivos e conjuntos de dados para fornecer comentários ou fazer perguntas.

4.3.6 Desempenho e escalabilidade

O JupyterHub e o SciServer exibem desempenho e escalabilidade louváveis, tornando-os adaptáveis a uma ampla gama de cenários de pesquisa e análise de dados. O JupyterHub gerencia com eficiência várias sessões de usuário em um único servidor, otimizando a alocação de recursos e garantindo experiências de usuário suaves, mesmo com grandes bases de usuários. Além disso, a capacidade de dimensionar horizontalmente implantando várias instâncias do JupyterHub pode gerenciar demandas crescentes sem problemas. O SciServer, com sua infraestrutura robusta, pode acomodar conjuntos de dados astronômicos massivos e simulações de computação intensiva, fornecendo recursos de

computação de alto desempenho. Ambas as plataformas são projetadas para aproveitar os recursos de computação em nuvem, permitindo que os usuários aumentem ou diminuam a escala de acordo com suas necessidades computacionais específicas. Essa escalabilidade, combinada com seus recursos de melhoria de desempenho, torna o JupyterHub e o SciServer opções confiáveis para pesquisadores e cientistas de dados que trabalham em projetos de tamanhos e complexidades variados.

O JupyterHub usa uma arquitetura de três camadas que permite que ele seja executado na nuvem ou em hardware local ou máquinas virtuais. Isso permite que ele atenda a um ambiente de ciência de dados pré-configurado para qualquer usuário no mundo com opções de personalização e escalabilidade. O JupyterHub oferece suporte a uma variedade de back-ends de banco de dados via SQLAlchemy, como PostgreSQL, MySQL e SQLite. Ele também permite que os usuários acessem volumes de dados públicos hospedados na nuvem ou em seu próprio hardware. O JupyterHub pode executar blocos de anotações Jupyter contendo código em Python, R e MATLAB, bem como outras linguagens que possuem kernels Jupyter.

O JupyterHub pode ser usado para atender a uma variedade de interfaces de usuário, incluindo o Jupyter Notebook, Jupyter Lab, RStudio, nteract e muito mais. O JupyterHub pode ser implantado com tecnologia de contêiner moderna, como o Kubernetes, permitindo que ele seja dimensionado e mantido de forma eficiente para um grande número de usuários. No entanto, o JupyterHub também faz algumas escolhas incomuns na forma como se conecta ao banco de dados, favorecendo a simplicidade e o desempenho de processo único em detrimento da escalabilidade horizontal (várias instâncias do Hub). O JupyterHub pode ser usado para equipes de até dois e bases de usuários de até 10.000. A escalabilidade do JupyterHub depende em grande parte da infraestrutura na qual ele é implantado.

O JupyterHub pode configurar a quantidade de CPU, memória, espaço em disco e largura de banda de rede que cada usuário pode usar. Os limites podem ser definidos no número de usuários simultâneos ou sessões por usuário. O JupyterHub inicia um proxy que encaminha todas as solicitações para o Hub por padrão. O proxy também gerencia o balanceamento de carga e os recursos de segurança, como criptografia SSL.

Balanceadores de carga externos ou proxies também podem ser usados para recursos mais avançados.

O SciServer usa uma arquitetura baseada em Docker/VM para fornecer análise interativa e em lote do lado do servidor com linguagens de script como Python e R em vários ambientes, incluindo Jupyter (notebooks), RStudio e linha de comando. A arquitetura baseada em Docker/VM permite que ele seja executado em um cluster de servidores ou clusters Kubernetes. Isso permite que ele gerencie grandes conjuntos de dados e usuários simultâneos com alta disponibilidade e confiabilidade. O SciServer hospeda mais de dois petabytes de dados científicos em uma variedade de disciplinas, como astronomia e mecânica dos fluidos. Ele fornece acesso a bancos de dados que podem ser consultados com comandos SQL ou CasJobs, bem como volumes de dados públicos que podem ser montados em contêineres criados pelos usuários.

O SciServer também tem um aplicativo Compute Images que fornece máquinas virtuais gratuitas para executar notebooks Jupyter, pré-instalados com pacotes de software. O SciServer permite equipes interdisciplinares globais e facilita o compartilhamento de conhecimento dentro da comunidade. O SciServer permite uma nova abordagem que permitirá aos pesquisadores trabalharem com terabytes ou petabytes de dados científicos, sem a necessidade de baixar grandes conjuntos de dados. No entanto, o SciServer pode não oferecer o mesmo nível de personalização e flexibilidade que o JupyterHub em termos de interfaces de usuário e provedores de nuvem.

4.3.7 Segurança e privacidade do SciServer e JupyterHub

O JupyterHub e o SciServer priorizam a segurança e a privacidade para proteger dados confidenciais de pesquisa e garantir a integridade das investigações científicas. O JupyterHub fornece mecanismos de autenticação e autorização para controlar o acesso, permitindo a integração com provedores de identidade e soluções de logon único. Ele também oferece isolamento do usuário, garantindo que o ambiente de cada usuário permaneça separado e seguro. O SciServer emprega protocolos de segurança robustos e controles de acesso para proteger dados astronômicos e conteúdo gerado pelo usuário. Ambas as plataformas facilitam a comunicação criptografada, ajudando a proteger os dados em trânsito. Além disso, os administradores podem monitorar e auditar as atividades do usuário, aprimorando a supervisão de segurança. Ao se concentrar nessas

medidas de segurança e práticas conscientes da privacidade, o JupyterHub e o SciServer capacitam os pesquisadores a colaborarem e analisar dados com confiança, ao mesmo tempo em que aderem a rígidos padrões de proteção e conformidade de dados.

O JupyterHub oferece suporte à autenticação por meio de uma variedade de métodos, como OAuth, LDAP, Kerberos e PAM. Ele também permite que os usuários configurem a criptografia SSL/TLS para comunicação segura entre o navegador e o servidor. O JupyterHub pode ser implantado com a tecnologia de contêiner moderna, como o Kubernetes, permitindo isolar processos e dados do usuário em pods separados. O JupyterHub também permite que os usuários executem servidores de usuário único em seus próprios subdomínios, o que fornece proteção entre origens entre servidores[48][49]. No entanto, o JupyterHub também expõe alguns possíveis riscos de segurança, como cross-site scripting (XSS), cross-site request forgery (CSRF) e execução remota de código (RCE), que podem ser explorados por usuários mal-intencionados ou hackers. Portanto, os usuários do JupyterHub devem seguir as práticas recomendadas, como usar senhas fortes, atualizar o software regularmente e evitar a execução de código não confiável.[49]

O SciServer usa uma arquitetura baseada em Docker/VM para fornecer análise interativa e em lote do lado do servidor com linguagens de script como Python e R em vários ambientes, incluindo Jupyter (notebooks), RStudio e linha de comando. O SciServer hospeda mais de dois petabytes de dados científicos em uma variedade de disciplinas, como astronomia e mecânica dos fluidos. Ele fornece acesso a bancos de dados que podem ser consultados com comandos SQL ou CasJobs, bem como volumes de dados públicos que podem ser montados em contêineres criados pelos usuários. O SciServer também tem um aplicativo Compute Images que fornece máquinas virtuais gratuitas para executar notebooks Jupyter, pré-instalados com pacotes de software.

O SciServer permite equipes interdisciplinares globais e facilita o compartilhamento de conhecimento dentro da comunidade. O SciServer implementa um sistema de controle de acesso baseado em função (RBAC) que permite aos usuários criarem grupos e atribuir permissões para acessar dados e recursos. O SciServer também criptografa dados em trânsito e em repouso usando algoritmos SSL/TLS e AES-256. O SciServer usa criptografia HTTPS para todo o tráfego da Web e exige que os usuários façam login com um nome de usuário e senha ou um serviço de autenticação de terceiros, como Google ou

GitHub. A plataforma também permite que os usuários controlem as permissões de acesso de seus arquivos e conjuntos de dados e fornece logs de auditoria para rastrear as atividades do usuário. No entanto, o SciServer pode não oferecer o mesmo nível de privacidade que o JupyterHub em termos de propriedade e controle de dados do usuário, pois o SciServer pode coletar, armazenar e usar dados do usuário para fins de pesquisa ou compartilhá-los com terceiros. Portanto, os usuários do SciServer devem ler atentamente os termos de serviço e a política de privacidade antes de utilizar a plataforma.

Algumas das considerações de segurança estão listadas abaixo e resumidas em Tabela 4.3:

- Criptografia SSL: Este é um recurso que habilita HTTPS para comunicação segura entre o navegador e o servidor. Ele requer um certificado SSL válido e uma chave privada.
- Segredo do cookie: Esta é uma chave para criptografar cookies do navegador que armazenam informações de autenticação do usuário. Deve ser uma cadeia de caracteres aleatória de pelo menos 32 bytes.
- Token de autenticação de proxy: esse é um token usado para o Hub e outros serviços para autenticar o proxy. Deve ser uma cadeia de caracteres aleatória de pelo menos 32 bytes.
- Contas de usuário do sistema: Cada usuário do JupyterHub recebe sua própria conta de usuário Unix criada quando inicia seu servidor pela primeira vez. Essas contas podem ser isoladas umas das outras usando namespaces Linux ou contêineres do Docker.
- Usuários off-boarding com segurança: Quando você exclui usuários do console de administração do JupyterHub, suas contas de usuário Unix não são removidas. Esses usuários devem ser removidos manualmente para liberar espaço em disco ou impedir o acesso não autorizado.
- Por usuário /tmp: /tmp é compartilhado por todos os usuários na maioria dos sistemas de computação, e isso tem sido uma fonte consistente de problemas de segurança. Os diretórios /tmp por usuário devem ser usados para evitar o acesso a arquivos entre usuários ou a execução de códigos mal-intencionados.
- O SciServer requer que os usuários façam login com suas credenciais e acessem seus dados e ferramentas.
- O SciServer fornece um recurso chamado 'Grupos' para que os usuários compartilhem seus recursos com seus colaboradores de forma privada.
- O SciServer fornece dois pools de armazenamento em rede para armazenar dados do usuário que podem ser acessados e usados por todos os aplicativos SciServer: Storage e Temporary.

- A SciServer aconselha os usuários a não armazenarem dados sensíveis ou confidenciais no SciServer, pois não podem garantir sua segurança ou privacidade.

Tabela 4.3 – Algumas considerações de segurança JupyterHub e SciServer

Considerações de segurança	**JupyterHub**	**SciServer**
Autenticação	métodos: PAM, OAuth, OAuth2, LDAP, GitHub, Google, etc.	**métodos**: protocolos OAuth2, Google, Facebook, OpenID Connect. tokens para acessar ferramentas e serviços.
Autorização	mecanismos: funções, escopos, grupos, casos de uso, Keycloak, Auth0, etc.	Mecanismos: funções, permissões
Segurança do Servidor	Requer que os administradores gerenciem a segurança do servidor	Garante a segurança do servidor como parte da plataforma gerenciada
Encriptação	Certificados SSL/TLS para criptografia de dados e comunicação.	Certificados SSL/TLS para criptografia de dados e comunicação
Isolamento	contêineres ou VMs como Docker, Kubernetes, etc.	contêineres/VMs como Docker, Kubernetes, etc.
Segurança de colaboração	depende da segurança do servidor subjacente e dos métodos de autenticação do usuário implementados.	oferece colaboração segura por meio de acesso controlado a conjuntos de dados compartilhados, ferramentas de análise e recursos de projeto colaborativo.
Proteção de Dados do Usuário	Os administradores devem garantir a privacidade dos dados e a conformidade com os regulamentos de proteção de dados.	Concentra-se na proteção dos dados do usuário, na garantia da conformidade com os padrões de proteção de dados e na proteção de dados confidenciais de pesquisa.
Segurança de Acesso Remoto	Requer métodos de acesso remoto seguros (por exemplo, VPNs) para impedir o acesso não autorizado ao servidor JupyterHub.	Oferece acesso remoto seguro através da interface baseada na web
Auditoria e registro em log	Suporte a auditoria de atividades e eventos do usuário usando logs ou métricas.	Implementa recursos de auditoria e registro.

FONTE: Autor

4.3.8 As melhores práticas e recomendações para usar o SciServer e o JupyterHub de forma eficaz.

Para usar o SciServer e o JupyterHub de forma eficaz, é crucial aderir a algumas práticas recomendadas e recomendações que otimizam os fluxos de trabalho de pesquisa e análise de dados. Comece definindo claramente os objetivos e os requisitos de dados do seu projeto para garantir que as ferramentas e configurações apropriadas sejam selecionadas. Mantenha backups regulares do trabalho e dos dados para proteger contra possíveis perdas de dados. Colabore de forma eficiente aproveitando sistemas de controle de versão como o Git para gerenciamento de código e documentação. Priorize a segurança implementando métodos de autenticação fortes e controles de acesso. Monitore continuamente o uso de recursos para dimensionar os recursos conforme necessário e otimizar os custos. Abrace a reprodutibilidade documentando análises em cadernos Jupyter e compartilhando-as com colaboradores. Atualize e mantenha regularmente seu software e bibliotecas para se beneficiar dos recursos e patches de segurança mais recentes. Por fim, mantenha-se envolvido com a comunidade de usuários e busque suporte quando necessário, pois tanto o SciServer quanto o JupyterHub têm comunidades ativas que podem fornecer insights e assistência valiosos. Seguir essas práticas recomendadas ajudará você a aproveitar ao máximo essas plataformas poderosas para seus esforços de pesquisa e análise de dados.

SciServer e JupyterHub são plataformas que permitem a ciência colaborativa orientada por dados usando várias ferramentas e ambientes. Para usá-los de forma eficaz, os usuários devem seguir algumas práticas recomendadas e recomendações, como:

- Planejar e definir os objetivos, escopo e resultados esperados do projeto. Isso ajudará a escolher a plataforma, as ferramentas, as fontes de dados e os formatos mais adequados para a análise.
- Explore os dados e recursos disponíveis em ambas as plataformas e familiarize-se com os recursos e funcionalidades que elas oferecem. Isso ajudará a otimizar os fluxos de trabalho de acesso, processamento e visualização de dados.
- Documente e anote o código e os resultados em blocos de anotações Jupyter ou outros formatos. Isso ajudará a garantir a reprodutibilidade, a transparência e a qualidade da análise.

- Colabore e comunique-se com outros usuários e grupos em ambas as plataformas. Isso ajudará a compartilhar ideias, feedback e melhores práticas, bem como a alavancar o conhecimento e a experiência coletiva da comunidade.
- Siga as diretrizes de segurança e privacidade em ambas as plataformas e proteja seus dados e credenciais contra acesso não autorizado ou uso indevido. Isso ajudará a evitar potenciais riscos ou violações que possam comprometer sua análise ou reputação.

4.3.9 As vantagens e desvantagens do SciServer e do JupyterHub para diferentes cenários de ciência de dados

JupyterHub e SciServer oferecem vantagens e desvantagens distintas que atendem a diferentes necessidades de pesquisa. O JupyterHub se destaca por fornece um ambiente flexível, interativo e colaborativo para análise e pesquisa de dados, tornando-o ideal para equipes que trabalham em diversos projetos. Seu suporte para várias linguagens de programação, bibliotecas e ferramentas de colaboração em tempo real fomenta a criatividade e a inovação. No entanto, gerenciar a infraestrutura e garantir a escalabilidade pode ser um desafio para organizações menores sem recursos de TI dedicados.

Por outro lado, o SciServer é adaptado para astrônomos e pesquisadores que lidam com grandes conjuntos de dados, oferecendo ferramentas especializadas e recursos otimizados para pesquisas astrofísicas. Seus ambientes pré-configurados simplificam as tarefas de análise de dados, mas seu foco específico no domínio pode limitar sua aplicabilidade mais ampla. Ambas as plataformas priorizam a segurança e a privacidade, mas os usuários ainda devem aderir às melhores práticas para garantir a integridade dos dados e a conformidade com os padrões de pesquisa. Em última análise, a escolha entre JupyterHub e SciServer depende dos requisitos específicos de pesquisa e restrições de recursos, com cada plataforma oferecendo pontos fortes e compensações únicas.

Aqui estão algumas vantagens e desvantagens do JupyterHub.

Vantagens:

- Ele suporta uma variedade de backends de banco de dados via SQLAlchemy, como PostgreSQL, MySQL e SQLite. Isso dá aos usuários mais flexibilidade e escolha em termos de armazenamento e gerenciamento de dados.

- Ele permite que os usuários acessem volumes de dados públicos hospedados na nuvem ou em seu próprio hardware. Isso dá aos usuários mais controle e propriedade sobre seus dados e recursos.
- Ele pode executar notebooks Jupyter contendo código em Python, R e MATLAB, bem como outras linguagens que possuem kernels Jupyter. Isso dá aos usuários mais versatilidade e diversidade em termos de linguagens de programação e frameworks.
- Ele pode ser usado para servir uma variedade de interfaces de usuário, incluindo o Jupyter Notebook, Jupyter Lab., RStudio, nteract e muito mais. Isso dá aos usuários mais opções e preferências em termos de experiência do usuário e funcionalidade.
- Ele pode ser implantado com tecnologia de contêiner moderna, como o Kubernetes, permitindo que seja personalizado, dimensionado e mantido de forma eficiente para pequenos e grandes números de usuários, cursos acadêmicos e infraestrutura de grande escala. Isso dá aos usuários mais confiabilidade e desempenho em termos de disponibilidade e qualidade do serviço.
- Ele aproveita o poder do Jupyter Notebook, que é uma ferramenta popular para ciência de dados que oferece recursos como codificação interativa, visualização de dados, documentação e colaboração.
- Permite o gerenciamento de sessões de computação interativa de múltiplos usuários.
- Ele fornece conectividade com a infraestrutura necessária para as sessões dos usuários, como computação em nuvem ou clusters de supercomputação.
- Ele simplifica a transição da pesquisa para a produção usando uma linguagem de uso geral como o Python, que é amplamente usada em muitos domínios

Desvantagens:

- Ele faz algumas escolhas incomuns na forma como se conecta ao banco de dados, favorecendo a simplicidade e o desempenho de processo único em detrimento da

escalabilidade horizontal (várias instâncias do Hub). Isso pode limitar sua capacidade de gerenciar alta simultaneidade e balanceamento de carga.

- Ele pode exigir mais tarefas de instalação e manutenção do que um servidor Jupyter Notebook autônomo, especialmente para personalização ou dimensionamento.
- Ele pode não fornecer o mesmo nível de suporte e recursos que um poderoso ambiente de desenvolvimento integrado (IDE), como conclusão de código, depuração, teste ou refatoração.
- Ele pode incentivar a escrita de código em células em vez de funções/classes/objetos, o que pode levar a código duplicado, pouca modularidade e dificuldade em reutilizar ou compartilhar código.
- Ele pode não ser compatível com algumas bibliotecas ou estruturas que exigem ambientes ou dependências específicos.
- Ele pode não ser seguro o suficiente para dados ou aplicativos confidenciais, pois depende de um navegador da Web e um servidor proxy para comunicação.
- Ele pode não ser adequado para fluxos de trabalho complexos que envolvem várias etapas ou estágios, como pré-processamento de dados, treinamento de modelo, avaliação, implantação e monitoramento.
- Pode não ser ideal para tarefas de alto desempenho que exigem alto uso de memória ou processamento paralelo.
- O JupyterHub pode exigir algumas habilidades técnicas e conhecimento para configurar e manter, especialmente para o uso de recursos avançados, como autenticação, segurança e Spawners.

Aqui estão algumas vantagens e desvantagens do JupyterHub.

Vantagens:

- Ele usa uma arquitetura baseada em Docker/VM para fornecer análise interativa e em lote do lado do servidor com linguagens de script como Python e R em vários ambientes, incluindo Jupyter (notebooks), RStudio e linha de comando. Isso

oferece aos usuários mais conveniência e eficiência em termos de fluxos de trabalho de análise de dados.

- Ele fornece acesso a bancos de dados que podem ser consultados com comandos SQL ou CasJobs, bem como volumes de dados públicos que podem ser montados em contêineres criados pelos usuários. Isso dá aos usuários mais funcionalidade e flexibilidade em termos de manipulação e exploração de dados.
- Ele implementa um sistema de controle de acesso baseado em função (RBAC) que permite aos usuários criarem grupos e atribuir permissões para acessar dados e recursos. Isso dá aos usuários mais segurança e privacidade em termos de compartilhamento de dados e colaboração.
- Permite a comparação de conjuntos de dados e descobrir novas conexões entre eles.
- Ele fornece acesso mundial a grandes conjuntos de dados de simulação e técnicas de processamento inovadoras.
- Ele oferece um sistema de armazenamento de dados científicos baseado em nuvem que se integra com outras ferramentas do SciServer
- Ele adapta ferramentas existentes como CasJobs, SkyServer e SkyQuery para facilitar o uso no SciServer.
- Ele adiciona recursos colaborativos, como compartilhamento de dados, blocos de anotações e comentários com outros usuários.
- Ele oferece suporte à análise do lado do servidor com consultas SQL e blocos de anotações Python próximos aos dados.
- O SciServer tem a vantagem de fornecer aos usuários acesso a grandes conjuntos de dados com curadoria de vários domínios da ciência, como astronomia, biologia, oceanografia, etc. Os usuários também podem carregar seus próprios conjuntos de dados ou usar conjuntos de dados públicos de outras fontes.
- O SciServer também tem a vantagem de permitir que os usuários executem uma análise do lado do servidor com Python, R, MATLAB, etc. sem a necessidade de baixar grandes conjuntos de dados

Desvantagens:

- Ele pode não oferecer o mesmo nível de personalização e flexibilidade que o JupyterHub em termos de interfaces de usuário e provedores de nuvem. Os usuários podem ter que se adaptar aos ambientes e ferramentas predefinidos que o SciServer oferece.
- Ele pode não oferecer o mesmo nível de privacidade que o JupyterHub em termos de propriedade e controle de dados do usuário.
- Pode exigir algumas habilidades técnicas e familiaridade com SQL e Python para usar o SciServer de forma eficaz.
- O SciServer pode ter algumas limitações nos tipos e formatos de conjuntos de dados que podem ser carregados ou acessados[2]. Os usuários também podem precisar aprender a usar ferramentas e interfaces específicas para trabalhar com os conjuntos de dados no SciServer.
- O SciServer pode não oferecer suporte a algumas linguagens ou bibliotecas que os usuários podem querer usar para suas análises. Os usuários também podem ter menos controle sobre seu ambiente de computação e recursos no SciServer.

Embora o JupyterHub e o SciServer ofereçam inúmeras vantagens, eles também vêm com certas limitações. O JupyterHub, apesar de sua flexibilidade, pode consumir muitos recursos, especialmente ao acomodar muitos usuários, o que pode exigir gerenciamento cuidadoso de recursos e planejamento de escalabilidade. Também pode faltar ferramentas especializadas e repositórios de dados para necessidades de pesquisa específicas do domínio, o que pode ser uma limitação para certas disciplinas científicas. O SciServer, embora seja adequado para pesquisas astrofísicas, pode não ser tão versátil para pesquisadores de outras áreas, já que seu foco é principalmente em dados e ferramentas relacionados à astronomia. Além disso, ambas as plataformas exigem um grau de conhecimento técnico para configuração e manutenção, o que pode representar desafios para usuários com recursos de TI limitados. Por fim, garantir a segurança e a privacidade dos dados continua sendo uma responsabilidade compartilhada entre os provedores e usuários da plataforma, exigindo a adesão cuidadosa às melhores práticas para mitigar potenciais riscos. Apesar dessas limitações, tanto o JupyterHub quanto o SciServer

permanecem ferramentas poderosas quando utilizadas dentro de seus respectivos domínios e com uma consideração cuidadosa de seus pontos fortes e fracos.

4.3.10 Resumo dos Resultados Comparativos

Esta seção resume a análise comparativa, destacando as distintas características, pontos fortes e limitações do JupyterHub e do SciServer. As descobertas fornecem insights valiosos para pesquisadores, cientistas e analistas de dados na seleção da plataforma mais adequada para suas necessidades específicas de análise de dados e colaboração.

A análise comparativa entre JupyterHub e SciServer revela características distintas que atendem a diversas necessidades dos usuários e domínios de pesquisa especializados. O JupyterHub, uma plataforma multiusuário versátil, se destaca na colaboração em educação e pesquisa ao oferecer um ambiente personalizável para codificação e ensino colaborativos. Ele possui um ecossistema robusto com extensas contribuições da comunidade, tornando-se uma escolha ideal para instituições de ensino e equipes de pesquisa em várias disciplinas.

Por outro lado, o SciServer surge como uma solução focada, principalmente destinada a apoiar astrônomos, cientistas de dados e pesquisadores envolvidos em pesquisas astronômicas intensivas em dados. Sua plataforma baseada na web elimina a necessidade de instalações individuais, permitindo acesso contínuo a ferramentas de análise especializadas e grandes conjuntos de dados astronômicos. A arquitetura do SciServer simplifica o acesso ao armazenamento de dados em escala de petabytes, aprimorando os recursos de colaboração e análise dentro da comunidade astronômica.

Em termos de integração e ecossistema, o JupyterHub se integra bem a um amplo espectro de linguagens de programação, ferramentas e serviços de terceiros, graças à sua popularidade e desenvolvimento orientado pela comunidade. Em contraste, o SciServer, embora tenha um conjunto mais focado de integrações, fornece uma conexão perfeita com conjuntos de dados astronômicos massivos, especialmente do Sloan Digital Sky Survey (SDSS), contribuindo para as necessidades especializadas dos astrônomos.

Ambas as plataformas exibem forte apoio da comunidade, mas com focos distintos. O JupyterHub se beneficia de uma comunidade diversificada e oferece extensa documentação, tornando-se uma escolha amigável para uma ampla gama de usuários. O SciServer conta com a experiência colaborativa de seu fórum de usuários e o suporte da

equipe do SciServer, garantindo que astrônomos e cientistas de dados tenham recursos e orientação personalizados.

Em termos de custo, o JupyterHub é de código aberto e gratuito, mas os usuários devem gerenciar os custos de infraestrutura do servidor e serviços complementares. O SciServer, por sua vez, oferece uma camada de acesso básico gratuita e níveis premium baseados em assinatura, tornando-se uma escolha viável para aqueles que buscam recursos e ferramentas aprimoradas sem o ônus do gerenciamento de infraestrutura.

Ambas as plataformas têm limitações: a complexidade de configuração e a natureza de propósito geral do JupyterHub podem limitar sua especialização, enquanto o foco de nicho do SciServer em astronomia pode restringir sua aplicabilidade mais ampla a pesquisadores em outros campos. Além disso, os usuários de ambas as plataformas podem enfrentar curvas de aprendizado associadas a ferramentas especializadas, mas os usuários do JupyterHub podem acessar recursos de aprendizado abundantes em vários domínios.

Em essência, a escolha entre JupyterHub e SciServer depende das necessidades e contextos específicos dos usuários. A versatilidade do JupyterHub atende àqueles que buscam codificação colaborativa e ensino entre disciplinas, enquanto os recursos centrados na astronomia do SciServer o tornam uma plataforma ideal para astrônomos e cientistas de dados que exigem acesso contínuo a ferramentas de análise especializadas e grandes conjuntos de dados astronômicos. Tabela 4.4 dá uma visão geral de alguns dos principais pontos discutidos.

Tabela 4.4 - Visão geral de alguns dos principais recursos e características do JupyterHub e SciServer comparados.

Item	**JupyterHub**	**SciServer**
Arquitetura	Servidor multiusuário para notebooks Jupyter.	Uma plataforma de análise e colaboração científica.
Cenários de uso	Educação, pesquisa, codificação colaborativa.	Plataforma de pesquisa em astronomia e ciência de dados.
Instalação	Requer configuração em um servidor ou ambiente de nuvem.	Plataforma baseada na Web, sem necessidade de instalação direta.
Custar	Livre e de código aberto.	Acesso gratuito a determinadas ferramentas e níveis de assinatura.
Principais características	Suporte multiusuário, ambientes personalizáveis.	Armazenamento de dados em escala de petabytes e ferramentas de análise.
Base de usuários de destino	Instituições de ensino, equipes de pesquisa.	Astrônomos, cientistas de dados, pesquisadores.

Integração	Integra-se com vários sistemas de autenticação.	Integra-se com dados SDSS e ferramentas de análise.
Apoio à comunidade	Comunidade ativa de código aberto.	Suporte da equipe SciServer, fórum de usuários.
Limitações	Requer manutenção do servidor e configuração complexa.	Especializado em astronomia e curva de aprendizado.

FONTE: Autor

5 CONCLUSÕES

Este relatório discute duas plataformas de ciência de dados, JupyterHub e SciServer, que são usadas para pesquisa científica e análise de dados. Essas plataformas têm características únicas que as diferenciam das demais do mercado. O relatório fornece uma visão geral das ferramentas de análise do lado do servidor, sua arquitetura, recursos, experiências do usuário, funcionalidades e aplicabilidade em diversos cenários de ciência de dados no âmbito da computação de alto desempenho.

O relatório conclui que ambas as plataformas oferecem soluções valiosas para pesquisa orientada a dados, permitindo que os usuários acessem conjuntos de dados grandes e complexos, realizem análises do lado do servidor com várias ferramentas e ambientes e compartilhem seus resultados e insights por meio de visualizações. Ambas as plataformas oferecem flexibilidade e opções de personalização para atender a diferentes necessidades e preferências do usuário.

O JupyterHub é uma plataforma amplamente adotada e versátil que pode ser adaptada a diferentes provedores de nuvem e hardware, tornando-a adequada para um amplo espectro de usuários e disciplinas científicas. Sua força está na flexibilidade e personalização, capacitando os usuários a personalizarem seus ambientes de acordo com suas necessidades específicas, tornando-o ideal para esforços de pesquisa colaborativa abrangendo diferentes domínios.

O SciServer é um ecossistema especializado que atende às necessidades únicas de astrônomos e astrofísicos. Ele é finamente ajustado para hospedar conjuntos de dados específicos da astronomia e domínios relacionados, fornecendo um conjunto abrangente de ferramentas e recursos. Os recursos notáveis incluem a capacidade de criar contêineres

de computação personalizados, promovendo um alto grau de personalização, e recursos colaborativos, como espaços de trabalho compartilhados e espaços de armazenamento.

A análise comparativa destaca os caminhos divergentes percorridos por essas plataformas. A especificidade de domínio e os recursos personalizados do SciServer contrastam com o apelo mais amplo e as extensas opções de personalização do JupyterHub. A escolha entre as duas plataformas depende de fatores como base de usuários alvo, recursos de integração, necessidades de escalabilidade e suporte da comunidade, exigindo consideração cuidadosa de pesquisadores, cientistas e analistas de dados.

O relatório conclui que tanto o JupyterHub quanto o SciServer continuam a evoluir, introduzindo novos recursos e aprimoramentos à medida que as outras plataformas e novas ainda estão surgindo. Essa natureza dinâmica garante que pesquisadores e cientistas sempre tenham acesso a ferramentas em evolução que se alinham com suas necessidades e objetivos em evolução. O relatório serve como uma bússola, orientando os usuários a navegarem pelo intrincado terreno da seleção de plataformas, incentivando a exploração e promovendo uma abordagem informada para uma era em que a computação de alto desempenho é parte integrante do progresso científico.

6 TRABALHOS FUTUROS

A implementação das plataformas (JupyterHub e/ou SciServer) em um ambiente de testes e documentação são alguns dos futuros trabalhos fundamentais para moldar o caminho e a evolução da implementação prática dessas plataformas para implantação futura. A atividade 3 dos objetivos do projeto envolve a execução prática dessas plataformas em um ambiente de teste controlado, permitindo experiência prática e avaliação de suas funcionalidades. Esta fase de implementação ilumina potenciais desafios, pontos fortes e áreas de melhoria na implantação das plataformas. É uma exploração que desvenda os meandros de suas funcionalidades.

Com base nessa implementação prática, a Atividade 4 traça o curso para documentação abrangente e processos de teste. A documentação abrange informações detalhadas sobre procedimentos de instalação, configuração de bancos de dados de usuários temporários e protocolos de acesso a bancos de dados de instituições científicas. Além disso, aprofunda as intrincadas características relacionadas ao compartilhamento de dados e códigos, bem como os mecanismos de publicação. Esses esforços robustos de documentação estabelecem as bases para as etapas subsequentes, incluindo a preparação e apresentação de seminários e tutoriais. A disseminação do conhecimento por meio de seminários e tutoriais torna-se uma via fundamental para compartilhar insights, melhores práticas e a experiência adquirida com um público mais amplo, promovendo uma comunidade colaborativa e informada. À medida que essas atividades se desenrolam, o contínuo refinamento e expansão dessas plataformas são antecipados, impulsionando a evolução dos ambientes colaborativos de pesquisa e análise de dados.

Além disso, a Atividade 4 estende seu escopo à exploração de recursos avançados, particularmente aqueles relacionados ao compartilhamento e publicação de dados e código. Ao documentar de forma abrangente esses aspectos, a comunidade de pesquisa obtém insights sobre o potencial colaborativo das plataformas e as metodologias para disseminar os resultados da pesquisa. Os preparativos para seminários e tutoriais funcionam como uma ponte para compartilhar esse conhecimento com um público mais amplo, fomentando uma cultura de colaboração e troca de conhecimento.

7 ATIVIDADES EXTRAS

Em meio ao foco principal do projeto, me envolvi ativamente em atividades complementares que não apenas ampliaram o escopo de minhas contribuições, mas também melhoraram minhas habilidades em diversos domínios. Uma parte considerável dos meus esforços foi dedicada à criação de código Python destinado a gerar patches de imagem sob medida para aplicações de aprendizado de máquina. Isso envolveu um processo meticuloso de fatiar e processar grandes imagens de satélite em patches menores e mais gerenciáveis, preparando as bases para esforços subsequentes de aprendizado de máquina.

No âmbito do aprendizado de máquina, meu foco principal era criar código Python para gerar patches de imagem. Esse processo tem importância no treinamento de modelos de aprendizado de máquina, especialmente em cenários onde grandes imagens de satélite ou conjuntos de dados precisam ser segmentados em seções menores e gerenciáveis. A capacidade de criar patches de imagem de forma eficiente não só contribui para a otimização do treinamento do modelo, mas também aumenta a capacidade do modelo de discernir padrões e recursos dentro de regiões específicas de interesse.

Em conjunto com a geração de patches de imagens, mergulhei no desenvolvimento de código Python dedicado a rastrear e classificar navios em torno de caminhos de derramamento de petróleo no oceano perto das costas brasileiras. Esse esforço englobou o desenvolvimento de algoritmos para identificar e avaliar a culpabilidade dos navios em incidentes de derramamento de óleo. Essa empreitada envolveu manipulação de dados geográficos e temporais, análise de dados e raciocínio espacial. O código Python não só facilitou o rastreamento de navios como também introduziu um mecanismo de classificação, destacando as embarcações consideradas mais culpadas com base em sua proximidade com a crise ambiental.

Essas atividades extras, sob a orientação e supervisão do Prof. Roberto Garcia, foram componentes integrais que fortaleceram minha proficiência em programação Python, manipulação de dados e uma sólida introdução ao aprendizado de máquina e suas aplicações. Além disso, a mentoria recebida do Prof. Roberto Garcia desempenhou um papel fundamental na navegação pelas complexidades dessas tarefas, transformando essas

atividades extras em experiências de aprendizado inestimáveis que se estenderam além dos objetivos imediatos do projeto.

8 GRUPO DE TRABALHO DE ASTRONOMIA DO BRICS (BAWG) E HACKATHON

Durante meu tempo neste Projeto, tive a oportunidade, através dos esforços do Prof. Rafael Santos, Prof. Roberto García e Prof. Ulisses Barres, de participar e participar do The BAWG 2023 Hackathon que ocorreu como parte integrante do 9th Annual BRICS Astronomy Working Group Workshop, realizado de 18 a 19 de outubro de 2023, na Cidade do Cabo, África do Sul. A iniciativa colaborativa de Ciência de Dados e Machine Learning Hackathon foi realizada no terceiro e quartos dias do Workshop. O Hackathon em si reuniu uma coorte de entusiastas de dados, especialistas em aprendizado de máquina e novatos, e astrônomos no auditório do Observatório Astronômico da África do Sul, na Cidade do Cabo. Focado em avançar o aspecto de dados da proposta emblemática do BRICS "BRICS Intelligent Telescope and Data Network", o Hackathon centrou-se em uma tarefa desafiadora: desenvolver um pipeline de aprendizado de máquina para classificação binária de galáxias formadoras de estrelas e núcleos de galáxias ativas a partir do conjunto de dados do catálogo de vários comprimentos de onda MIGHTEE-COSMOS.

O evento durou 1,5 dias, começando com introduções e sessões de fundo sobre astronomia, ciência de dados e aprendizado de máquina. Os participantes, formando seis equipes com 3-5 membros cada, mergulharam no desafio comum de classificar galáxias com base em imagens de vários comprimentos de onda. Aproveitando um grande conjunto de dados da pesquisa MIGHTEE, as equipes aplicaram diversas técnicas de aprendizado de máquina, incluindo redução e normalização de dimensionalidade. O Hackathon utilizou a instalação de nuvem de pesquisa intensiva em dados de ponta sul-africana, fornecendo um ambiente padronizado para codificação colaborativa.

Notavelmente, o Hackathon foi tão competitivo quanto colaborativo em uma atmosfera amigável. As equipes tiveram 24 horas para elaborar e apresentar suas soluções, que foram posteriormente avaliadas com base em critérios como precisão, criatividade, escalabilidade e apresentação. A natureza colaborativa do Hackathon foi reforçada pela oferta de tutoriais on-line antes do evento, mentoria durante o Hackathon e um local projetado para facilitar o networking entre os participantes. O sucesso geral do BAWG 2023 Hackathon reflete seu papel fundamental em impulsionar a inovação colaborativa dentro da comunidade de astronomia intensiva em dados do BRICS, marcando um passo

significativo em direção à realização dos objetivos do Telescópio Inteligente e da Rede de Dados do BRICS.

REFERÊNCIAS

[1] TAGHIZADEH-POPP, M.; KIM, J.; LEMSON, G.; MEDVEDEV, D.; RADDICK, M.; SZALAY, A.; THAKAR, A.; BOOKE,R J.; CHHETRI, C.; DOBOS, L.; RIPPIN, M. SciServer: A science platform for astronomy and beyond. **Astronomy and Computing**, v.33, Oct. 2020. doi: 10.1016/j.ascom.2020.100412.

[2] SCISERVER. **SciServer – Collaborative data-driven science**. https://sciserver.org/. Acesso em: 9 ago. 2023.

[3] THE TECH SPREE. **What Is Kaggle? The New platform Simply Explained.** 9 nov. 2021. Disponível em: https://www.thetechspree.com/what-is-kaggle-the-new-platform-simply-explained/. Acesso em: 11 nov. 2022.

[4] SCISERVER. **How to use SciServer – SciServer**. Disponível em: https://www.sciserver.org/support/how-to-use-sciserver/. Acesso em: 11 nov. 2022.

[5] SCISERVER. **SciServer Compute**. Disponível em: https://sciserver.org/about/compute/. Acesso em: 31 ago. 2023.

[6] SCISERVER. **SciServer Dashboard: Data, Collaboration, Compute**. SciServer. Disponível em: https://apps.sciserver.org/dashboard/. Acesso em: 5 set. 2023.

[7] SCISERVER. **Astronomy – SciServer**. Disponível em: https://sciserver.org/integration/astronomy/. Acesso em: 9 ago. 2023.

[8] DIVYA, S. **All About Using Jupyter Notebooks and Google Colab**. Data Science Central. 26 mar. 2019. Disponível em: https://www.datasciencecentral.com/all-about-using-jupyter-notebooks-and-google-colab/. Acesso em: 11 nov. 2022.

[9] PRABANJAN, R. **Google Colab - Everything you Need to Know**. - Scaler Topics," Scaler Academy. 27 ago. 2021. Disponível em: https://www.scaler.com/topics/what-is-google-colab/. Acesso em: 11 nov. 2022.

[10] JUPYTER DOCUMENTATION 4.1.1 ALPHA DOCUMENTATION. **Architecture**. Disponível em: https://docs.jupyter.org/en/latest/projects/architecture/content-architecture.html. Acesso em: 5 set. 2023.

[11] THE JUPYTER BOOK COMMUNITY. **JupyterHub.** 2022. Disponível em: https://jupytearth.org/jupyter-resources/ecosystem/jupyterhub.html Acesso em: 31 ago. 2023.

[12] JUPYTERHUB FOR KUBERNETES. **The JupyterHub Architecture - Zero to JupyterHub with Kubernetes**. Disponível em: https://z2jh.jupyter.org/en/stable/administrator/architecture.html. Acesso em: 9 ago. 2023.

[13] TUTORIALSPOINT. **Google Colab - Quick Guide**. Disponível em: https://www.tutorialspoint.com/google_colab/google_colab_quick_guide.htm. Acesso em: 11 nov. 2022.

[14] RUN:AI. **JupyterHub: A Practical Guide**. Disponível em: https://www.run.ai/guides/machine-learning-operations/jupyterhub. Acesso em: 19 fev. 2023.

[15] PRZYBYLA, M. **Why Does Everyone Use Kaggle? …here's why a data scientist should**. Medium. 9 jul. 2020. Disponível em: https://towardsdatascience.com/why-does-everyone-use-kaggle-db1bdf1f1b1a. Accessed: 11 nov. 2022.

[16] GOOGLE RESEARCH. **Welcome To Colaboratory - Colaboratory.** Disponível em: https://colab.research.google.com/#scrollTo=Nma_JWh-W-IF. Acesso em: 11 nov. 2022.

[17] BONNER, A. **Getting Started With Google Colab.** Medium. Jan. 1, 2019. Disponível em: https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c. Acesso em: 11 nov. 2022.

[18] SCISERVER. **About – SciServer.** Disponível em: https://www.sciserver.org/about/. Acesso em: 11 nov. 2022.

[19] GITHUB DOCS. **GitHub Codespaces overview.** Disponível em: https://docs.github.com/en/codespaces/overview. Acesso em: 19 fev. 2023.

[20] MICROSOFT. **Notebooks at Microsoft - Visual Studio.** Disponível em: https://visualstudio.microsoft.com/vs/features/notebooks-at-microsoft/. Acesso em: 19 fev. 2023.

[21] PROJECT JUPYTER TEAM. **JupyterHub Documentation: User Guide**. Release 0.7.1. 2018. 64p.

[22] POSIT. **Posit Cloud**. Disponível em: https://posit.cloud/learn/guide. Acesso em: 19 fev. 2023.

[23] POSIT. **Posit Cloud**. Disponível em: https://posit.cloud/plans/free. Acesso em: 19 fev. 2023.

[24] REPLIT Docs. **Introduction to Replit**. Disponível em: https://docs.replit.com/getting-started/intro-replit. Acesso em: 19 fev. 2023.

[25] REPLIT. **Pricing**. Disponível em: https://replit.com/pricing. Acesso em: 19 fev. 2023.

[26] REPLIT. **Templates**. Disponível em: https://replit.com/templates. Acesso em: 19 fev. 2023.

[27] JAYANTI. **Top 10 Google Colab Alternatives for Machine Learning Engineers in 2023.** Analytics Insight. Nov. 14, 2022. Disponível em: https://www.analyticsinsight.net/top-10-google-colab-alternatives-for-machine-learning-engineers-in-2023/. Acesso em: 19 fev. 2023.

[28] DATACAMP. **What is DataCamp Workspace?** Support. Disponível em: https://support.datacamp.com/hc/en-us/articles/4680748580631-What-is-DataCamp-Workspace-. Acesso em: 19 fev. 2023.

[29] DATACAMP. **What is DataCamp Workspace?** Workspace Docs. Disponível em: https://workspace-docs.datacamp.com/. Acesso em: 19 fev. 2023.

[30] DATACAMP. **R Packages: A Beginner's Tutorial**. DataCamp. março de 2019. Disponível em: https://www.datacamp.com/tutorial/r-packages-guide. Acesso em: 31 ago. 2023.

[31] DATACAMP. **R Packages: A Beginner's Tutorial**. DataCamp. março de 2019. Disponível em: https://www.datacamp.com/tutorial/r-packages-guide. Acesso em: 31 ago. 2023.

[32] DEEPNOTE DOCS. **Welcome to Deepnote.** Disponível em: https://deepnote.com/docs. Accesso em: 19 fev. 2023.

[33] BATAPATI, T. **3 Reasons Why Deepnote Is Powerful Than Famous Jupyter Notebooks.** Medium. Nov. 09, 2020. Disponível em: https://medium.datadriveninvestor.com/3-reasons-why-

deepnote-is-powerful-than-famous-jupyter-notebooks-43aee38419f5. Accesso em: 19 fev. 2023.

[34] PRAKASH, Y. **Here's How To Use Jupyter Notebooks on Steroids** — with Deepnote. Medium. Dec. 13, 2021. Disponível em: https://towardsdatascience.com/heres-how-to-use-jupyter-notebooks-on-steroids-with-deepnote-c35251222358. Acesso em: 19 fev. 2023.

[35] JETBRAINS. **Notebooks**. Disponível em: https://www.jetbrains.com/datalore/features/notebooks/. Acesso em: 19 fev. 2023.

[36] JETBRAINS . **Datalore Features**. Disponível em: https://www.jetbrains.com/datalore/features/. Acesso em: 19 fev. 2023.

[37] JETBRAINS. **Quick start tutorial.** Datalore Documentation. 28 nov. 2022. Disponível em: https://www.jetbrains.com/help/datalore/datalore-quickstart.html. Accesso em: 19 fev. 2023.

[38] JETBRAINS. **Datalore Pricing: Choose Your Datalore Plan.** Disponível em: https://www.jetbrains.com/datalore/buy/. Acesso em: 19 fev. 2023.

[39] POSIT. **RStudio Server**. Disponível em: https://posit.co/download/rstudio-server/. Acesso em: 19 fev. 2023.

[40] POSIT. **Posit Cloud**. Disponível em: https://posit.cloud/plans/free. Acesso em: 19 fev. 2023.

[41] GARTNER. **What Is Data and Analytics?** Disponível em: https://www.gartner.com/en/topics/data-and-analytics. Acesso 30 ago. 2023.

[42] JENN DAUGHERTY. **Why Today's Companies Need Collaborative Data Analysis.** Mode. 15 de dezembro de 2022. Disponível em: https://mode.com/blog/collaborative-data-analytics/. Acesso em: 30 ago. 2023.

[43] TRŪATA. **The power of data collaboration and the rise of collaborative analytics**. Trūata. 6 de maio de 2022. Disponível em: https://www.truata.com/articles/data-collaboration-and-the-rise-of-collaborative-analytics/. Acesso em: 30 ago. 2023.

[44] PROJECT JUPYTER. **JupyterHub.** Disponível em: https://jupyter.org/hub. Acesso: 9 nov. 2022.

[45] SCISERVER. **Architecture**. Disponível em: https://sciserver.org/about/architecture/. Acesso em: 30 ago. 2023.

[46] COLLEGE OF DATA SCIENCE AND SOCIETY. **Choosing the Right JupyterHub Infrastructure**. College of Data Science and Society - Berkeley. Disponível em: https://data.berkeley.edu/choosing-right-jupyterhub-infrastructure Acesso em: 31 ago. 2023.

[47] THE LITTLEST JUPYTERHUB. **Installing on your own server**. https://tljh.jupyter.org/en/latest/install/custom-server.html. Acesso em: 30 ago. 2023.

[48] JUPYTERHUB. **JupyterHub and OAuth**. Project Jupyter Contributors. 2023. Disponível em: https://jupyterhub.readthedocs.io/en/stable/explanation/oauth.html# Acesso em: 3 set. 2023.

[49] JUPYTERHUB. **Security Overview**. Project Jupyter Contributors. 2023. Disponível em: https://jupyterhub.readthedocs.io/en/stable/explanation/websecurity.html Acesso em: 3 set. 2023).