Instituto Base dos Dados

2021

Relatório Anual

abr/2022

Equipe 2021

Coordenação (*sudo su*)

Diego Oliveira (Infraestrutura)
Fernanda Scovino (Diretora Executiva / Comunicação)
Frederico Israel (Infraestrutura)
João Carabetta (Infraestrutura)
Ricardo Dahis (Presidente / Dados)

Comunicação

Giovane Caruso
Nayara Moraes

Dados

Crislane Alves
Gustavo Aires
Hevilyn Souza
Lucas Moreira
Matheus Valentim

Infraestrutura

Caio Rogério dos Santos
Vinicius Aguiar
Vitor Mussa

Website

Breno Gomes

Colaboradores

André Masuko (Comunicação)
André Gorenstein (Serviços e Parcerias)
Érica Travain (Comunicação)
Isabella Helter (Infraestrutura)
Matheus Ávila (Serviços e Parcerias)
Pedro Cavalcante (Infraestrutura)
Rodrigo Dornelles (Infraestrutura)
Vanessa Costa (Comunicação)

Carta dos(as) Fundadores(as)

A Base dos Dados surgiu com a missão de universalizar o acesso a dados de qualidade no Brasil. A marca já existia desde 2019, na tentativa de solucionar um problema contínuo nas nossas vidas: gastar muito tempo para achar, baixar, tratar, cruzar e (finalmente) explorar os dados. Desde então, iniciamos o processo de facilitar a busca e análise de importantes bases públicas, quebrando barreiras técnicas para quem já faz e quem quer começar a fazer uso de dados em quaisquer áreas.

O ano de 2021 foi quando demos nosso maior salto. Decidimos que essa ideia era importante demais para depender das mãos de poucas pessoas. Nos tornamos oficialmente o Instituto Base dos Dados (BD) em junho, uma organização sem fins lucrativos, com equipe formal, CNPJ e tudo que tem direito. Essa foi uma evolução natural do que já estávamos construindo desde o início do ano, reunindo uma comunidade engajada na nossa missão e espalhando pelos 5 cantos do Brasil uma forma mais fácil de trabalhar com dados.

Esse extenso trabalho gerou resultados exponenciais. Saímos de 13 mil consultas ao mês em janeiro para o triplo em dezembro (39 mil/mês); mais que triplicamos o número de seguidores no Twitter (de 5 para 16 mil); recebemos um prêmio internacional da Google Cloud e tivemos a oportunidade de apoiar organizações na estruturação e uso inteligente de dados. Esses números são apenas consequência do impacto na vida de milhares de pessoas que encontraram na BD uma aliada diante dos desafios comuns nesse universo de dados.

Por trás de todo esse sucesso existiram dezenas de mãos que abraçaram a ideia conosco e arregaçaram as mangas tratando dados, escrevendo códigos, produzindo conteúdos e análises, dedicando tempo e energia em uma missão construída de forma colaborativa. Só temos a agradecer imensamente a todo o carinho e apoio desses(as) colaboradores(as), e em especial nossos doadores que foram fundamentais para mantermos nossos produtos e expandirmos nossa atuação. Em 2021, a BD se tornou um movimento que aproxima, acolhe e impulsiona pessoas numa rede cada vez mais relevante na produção de conhecimento e expansão do acesso à informação de forma livre, democrática e universal. E esse é só o começo.

Diego Oliveira, Fernanda Scovino, Frederico Israel, João Carabetta e Ricardo Dahis

Gestão e Financeiro

A Base dos Dados se formalizou enquanto organização não governamental em junho de 2021, portanto as informações apresentadas nesta seção se restringem a esse período de tempo.

Equipe

Em 2021 foram realizadas as primeiras contratações de membros da Base dos Dados. Iniciamos o ano apenas com voluntários atuantes e finalizamos com 9 membros formais da organização, além de diversos colaboradores.

- 3 membros(as) contratados(as) de Infraestrutura;
- 2 membros(as) contratados(as) de Comunicação;
- 3 membros(as) contratados(as) de Dados (5 ao todo no ano);
- 1 membro contratado de Website.

Resultados financeiros

Receitas	
Receita de Serviços	500,00C
Receita de Contribuições e Doações *	192.292,51C
Total das Receitas	192.792,51C

Despesas	
Remuneração por Serviços de Terceiros	42.772,36D
Despesas Administrativas Diversas	617,12D
Despesas com Tributos	4.000,00D
Total das Despesas	47.389,48D

Resultado do exercício	
RECEITAS	192.792,51C
DESPESAS + CUSTO	47.389,48D
RESULTADO LÍQUIDO DO EXERCÍCIO	145.403,03

Demonstração do Resultado do Exercício de 28/06/2021 até 31/12/2021

* Durante 2021, tivemos uma parceria com a Fundação Lemann contabilizada como doação pela forma de remuneração realizada. O valor total contabiliza R$50.000,00.

Nossos números de 2021

Produtos

417 mil consultas aos dados
Se cada consulta poupou 2 horas de trabalho de alguém, são +3.000 anos poupados.

34 mil acessos mensais* ao site
*Média de outubro a dezembro de 2021, a partir do lançamento da nova plataforma.

249 tabelas tratadas
disponibilizadas no nosso *datalake* público.

2 pacotes de programação
de acesso a dados nas linguagens Python e R.

321 mil linhas de código
adicionadas ao nosso repositório.

25 colaboradores
aprimoraram o código da nossa infraestrutura.

Comunidade e Conteúdo

11 projetos usando a BD
em código aberto no GitHub.

10 matérias usando a BD
em veículos como Estadão, Piauí.

16 artigos e tutoriais
publicados por membros(as) e voluntários(as).

9 palestras em universidades
em todos os cantos do Brasil (remotas).

7 workshops no YouTube
disponibilizados no nosso canal.

4 participações em conferências
de pesquisa, dados e programação.

+1.170 inscritos
10.848 visualizações

+11 mil seguidores no Twitter
+722 membros no Discord

Serviços e Parcerias

3 serviços de análise de dados
junto a ONGs e empresas.

1 apoio a órgão governamental
no tratamento e divulgação de dados.

Prêmios

1 prêmio internacional
Google Cloud Customer Award

1 prêmio brasileiro
XXVI Prêmio Tesouro Nacional 2021

Produtos

Mecanismo de busca

Desde 2019, a Base dos Dados existia como um "Google" de dados públicos brasileiros. O conceito era simples: um site onde são documentadas informações relevantes sobre bases públicas e onde achá-las. Pouco tempo depois, em outubro de 2020, passamos a disponibilizar também os dados já tratados no nosso *datalake* público. Com esse novo produto, nosso mecanismo ganhou ainda mais relevância na busca de dados para análises. A partir de então, trabalhamos em **reformular a plataforma e aprimorar a qualidade dos metadados** disponibilizados para facilitar ainda mais o acesso aos dados.

Nosso grande projeto de 2021 foi reconstruir o site do zero. Contratamos um desenvolvedor *front-end* e organizamos uma equipe multidisciplinar para trabalhar ao longo de 5 meses, desde a ideação até o lançamento do site completo (leia mais **aqui**). Ao longo do processo realizamos 11 entrevistas de usuário, antes e depois do desenvolvimento, buscando a melhor experiência na plataforma. O site foi lançado em 13 de outubro de 2021, quase exatamente 1 ano após o nosso *datalake* e 2 anos após o mecanismo de busca. Do lançamento até o final de 2021, tivemos ao todo **82.725 mil acessos** e **26.599 usuários**.

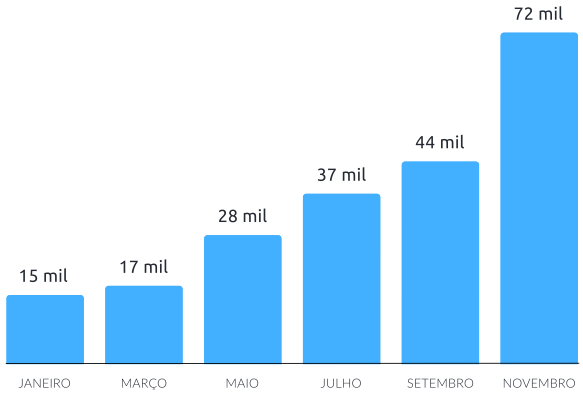
Junto ao desenvolvimento do site, aprimoramos o mecanismo de busca com a ampliação e revisão de metadados. Além de informações gerais sobre bases públicas, passamos a incluir e possibilitar a busca por metadados de conjuntos e tabelas já tratadas no *datalake*. Para tornar essas funcionalidades, realizamos profundas melhorias nos padrões de metadados disponíveis na BD e publicamos nosso primeiro **manual de estilo**.

Datalake público

Um dos grandes focos do nosso trabalho em 2021 foi a **ampliação e sofisticação do nosso *datalake* público**, onde disponibilizamos bases públicas já limpas e integradas de forma gratuita. Através dele é possível acessar e cruzar tabelas de diferentes organizações de maneira simples e rápida.

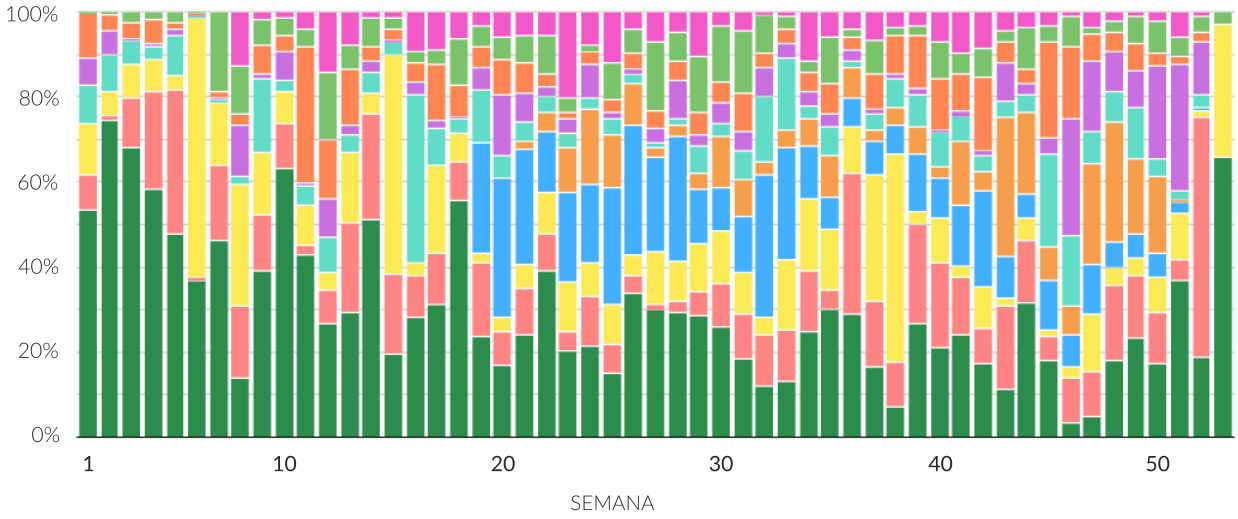
Apenas neste ano, tivemos mais de 417 mil consultas, sendo:

- **329 mil** em SQL, através da interface online do *datalake* no Google BigQuery;
- **54 mil** pelo pacote *basedosdados* em Python;
- **34 mil** pelo pacote *basedosdados* em R.



Número de acesso às bases no *datalake* a cada mês de 2021

Ao longo de 2021, nossa equipe de dados, com ajuda de usuários e colaboradores, limpou, tratou e compatibilizou mais de **60 conjuntos de dados**, adicionando ao todo **249 novas tabelas** no *datalake*. Destacamos abaixo as 10 bases mais acessadas do ano.



Conjunto	Número de consultas
Relação Anual de Informações Sociais (RAIS)	75.323
Diretórios Brasileiros	41.597
Eleições Brasileiras	38.433
Campanha Nacional de Vacinação contra Covid-19	38.279
Censo Escolar	27.372
População Brasileira	25.611
Índice de Desenvolvimento da Educação Básica (Ideb)	23.839
Produto Interno Bruto do Brasil	23.570
Sócios de Empresas - Dados públicos CNPJ	20.123
Censo Demográfico	15.722

Percentual de acesso às bases no *datalake* a cada semana de 2021

Relação Anual de Informações Sociais (RAIS) • Ministério da Economia

A Relação Anual de Informações Sociais (RAIS) é um relatório anual e abrangente de informações socioeconômicas de pessoas jurídicas e outros empregadores, elaborado pelo **Ministério da Economia**. São microdados de vínculos empregatícios e estabelecimentos, agrupados por ano, estado, município, sexo, raça, nacionalidade, tipo de vínculo e mais. No tratamento dessa base, unificamos todos os anos, padronizamos os nomes de variáveis e limpamos valores errados de células.

Diretórios Brasileiros • Base dos Dados

Criamos a base de Diretórios Brasileiros para ser uma referência de centralização de informações de unidades básicas para análises e funciona como um perfil completo de entidades como município, escola, UF, setores censitários e mais. São tabelas que ligam diversos códigos institucionais e informações de diferentes entidades brasileiras.

Eleições Brasileiras • Tribunal Superior Eleitoral

O **Tribunal Superior Eleitoral** (TSE) fornece regularmente dados das eleições brasileiras, que englobam características do eleitorado, candidaturas, resultados eleitorais e processos de prestação de contas. São informações detalhadas dos partidos políticos, desempenho partidário e performance eleitoral dos candidatos em cada município brasileiro. No tratamento dessa base, padronizamos os nomes de variáveis e os formatos para consistência entre anos.

Campanha Nacional de Vacinação contra Covid-19 • Ministério da Saúde

Nós disponibilizamos dados atualizados da Campanha Nacional de Vacinação contra Covid-19, publicados pelo **Ministério da Saúde**, que englobam o número de doses aplicadas por UF e municípios, por um determinado período, por gênero, por faixa etária e por tipo de vacina. No tratamento dessa base nós normalizamos a tabela original, além de atualizar periodicamente os dados.

Censo Escolar • Inep

O Censo Escolar é a mais importante pesquisa estatística educacional brasileira, em que cada escola do país responde a diversas perguntas sobre a própria escola, sobre as turmas, os professores e os

alunos que ali frequentam. São 4 tabelas para diferentes níveis de agregação: uma a nível de escola, uma a nível de turma, uma para docentes e outra para os alunos.

Disponibilizamos dados da população a nível de município, estado e de todo o país, desde o ano de 1991 até 2020. As informações são estimadas pelo **IBGE**, a partir da pesquisa de Projeções da População para o Brasil e dados do Censo Demográfico, e publicados pelo **Inep**.

O processo de limpeza da base incluiu a junção e a compatibilização entre vários anos distintos, harmonizando as variáveis e seus valores. Outra parte crucial foi a criação de variáveis para melhor compatibilização da tabela com outras tabelas deste dataset e do próprio catálogo da Base dos Dados.

Índice de Desenvolvimento da Educação (Ideb) • Inep

O IDEB reúne o resultado do fluxo escolar e as médias de desempenho nas avaliações. São tabelas por região, estado, município e escolas, disponibilizadas pelo **Inep**. Organizamos e separamos os dados para garantir sua consistência.

Produto Interno Bruto do Brasil • IBGE

Disponibilizamos a série histórica do Produto Interno Bruto (PIB) dos municípios brasileiros, um sistema de indicadores municipais com importantes informações socioeconômicas, publicada pelo **IBGE**. Na limpeza desses dados nós mudamos a unidade de medida para R$ 1.

Sócios de Empresas - Dados públicos CNPJ • Ministério da Economia

A base de sócios de empresas reúne dados do **Ministério da Economia** com informações de administradores de CNPJ, como a inscrição de sócios, o cadastro de empresas e holdings, até a relação de CNPJs e CNAEs de empresas. Nessa base realizamos o tratamento de strings, a inclusão do ano de entrada na sociedade e a renomeação das variáveis de acordo com nossos padrões de dados.

Censo Demográfico • IBGE

O Censo Demográfico é a mais complexa operação estatística realizada por um país, com um

levantamento que consiste na visita a todos os domicílios, reunindo informações de condições de vida da população. Disponibilizamos em nosso *datalake* público a série histórica com dados de 1970 até do último censo, em 2010, originalmente publicada pelo **IBGE**. Na limpeza desses dados nós seguimos o pacote do DataZoom e o próprio dicionário do IBGE.

Pacotes de programação

Além de disponibilizar os dados através do nosso *datalake* público, desenvolvemos e mantemos pacotes de programação para ampliar o acesso a esses dados. Os pacotes foram construídos nas linguagens de código aberto, fortemente utilizadas para análise de dados: Python e R.

O pacote ***basedosdados* em R** foi lançado em 2021, permitindo o acesso, manipulação e cruzamento das bases disponíveis no *datalake* pelo ambiente da linguagem. Dentre as funcionalidades, são destaque:

- Requisição e download de todas as tabelas do *datalake*;
- Possibilidade de listar todas as tabelas e conjuntos disponíveis no *datalake*.

Ao longo de 2021, lançamos também 4 grandes releases de novas funcionalidades do pacote ***basedosdados* em Python** (versão **1.2.1 a 1.6.0**), incluindo:

- Possibilidade de listar todas as tabelas e conjuntos disponíveis no *datalake*;
- Possibilidade de listar todas as colunas de uma tabela sem precisar baixar os dados;
- Melhorias do fluxo e validação de metadados;
- Melhorias nos fluxos de upload de dados;
- Criação do ambiente de desenvolvimento e produção do *datalake* (leia mais **aqui**);
- Automatizações na checagem da qualidade de dados e metadados.

Serviços e Parcerias

Nossa área de Serviços e Parcerias foi criada para garantir a sustentabilidade de receita da organização, de forma a manter e expandir nossos produtos de forma gratuita. A prestação de serviços de dados e estruturação de parcerias vem sendo chave não só na organização financeira da Base dos Dados, mas também no fortalecimento da nossa posição de referência em disponibilização

e tratamento de dados de qualidade.

Atuamos em **3 frentes de serviços** para ajudar pessoas e organizações a extraírem o máximo de valor de dados:

- **Captura de dados**;
- **Análise de dados**;
- **Consultoria de dados**.

Além de serviços pagos, realizamos também parcerias com instituições governamentais, prestando serviço pro bono de grande valor público. Em 2021, prestamos ao todo 3 serviços de estruturação e análise de dados e apoiamos 1 órgão governamental.

Instituto Alziras

O **Instituto Alziras** é uma organização sem fins lucrativos com a missão de ampliar e fortalecer a presença de mulheres, em toda sua diversidade, na política e na gestão pública. Nossa parceria se deu na produção de análises com dados de eleições brasileiras do Tribunal Superior Eleitoral para o Instituto. Todas as agregações e análises de dados foram feitas via SQL na plataforma BigQuery, com consultas que passaram por um processo de filtragem dos anos e informações de interesse.

EB Capital Educação

A **EB Capital** é uma empresa brasileira de investimentos focada na criação de valor investindo em empresas da economia que atendem às lacunas estruturais brasileiras. Nós fornecemos serviços de tratamento e análise de dados e ajudamos a empresa a responder perguntas decisivas sobre tendências em educação no Brasil. Para isso lançamos mão de nossas tabelas já tratadas do Censo Escolar, PNAD Contínua, dentre outras, disponíveis no *datalake*.

Fundação Lemann

A **Fundação Lemann** é uma organização filantrópica que trabalha para garantir educação de qualidade para crianças brasileiras e apoiar líderes focados no desenvolvimento social do país. Colaboramos com a Fundação na criação de metas e painéis para acompanhamento de indicadores-chave da educação brasileira. Trabalhamos também em conjunto com a equipe da Fundação Lemann

para organizar conjuntos de dados como os do SAEB.

SEDUC (pro bono)

Em parceria com o Escritório de Evidências, uma iniciativa da **Secretaria da Educação do Estado de São Paulo** (SEDUC) que procura ajudar na geração e disseminação de conhecimento científico para políticas educacionais, nós tratamos, compatibilizarmos e disponibilizamos os dados do Índice de Desenvolvimento da Educação do Estado de São Paulo (IDESP), do Indicador de Nível Socioeconômico (INSE) e do Fluxo Escolar no Estado de São Paulo.

Comunidade

Em menos de um ano, reunimos uma comunidade enorme não só de usuários, mas também um grande público que nos acompanha nas redes sociais. Isso foi fruto de um extenso trabalho de conteúdo, participação e engajamento em diferentes núcleos. Nossas principais frentes e conteúdos se dividiram em:

- **Redes sociais**;
- **Conferências e palestras**;
- **Produção de conteúdo**;
- **Clipping**.

Redes sociais

Trabalhamos na criação de uma rotina de conteúdo e comunicação nas redes, publicando gráficos, análises, *quizzes*, memes e novidades dos nossos produtos. Além disso, lançamos nosso canal do YouTube, página no LinkedIn e publicação de artigos com colaboradores da BD.

Rede social		Total de seguidores
Twitter	↑ 11.060	16.183
LinkedIn	↑ 1.742	1.742
Discord	↑ 722	722
YouTube	↑ 1.171	1.711

Twitter

Reunindo uma comunidade com mais de 16 mil seguidores no final de 2021, nosso perfil no Twitter se tornou um dos principais canais para divulgações e interações com nossos usuários. Ao longo do ano, ajudamos a espalhar a palavra da Base dos Dados através de *tweets* informativos, compartilhamento de análises de usuários, análises, gráficos, brincadeiras lúdicas, como nosso *quiz* "Que dado é esse?", e — por que não? — vários memes. Só em 2021, foram 11.060 novos seguidores em nosso perfil.

LinkedIn

Com perfil mais profissional e institucional, o LinkedIn também foi uma plataforma que exploramos para criar maior conexão com nossa comunidade e divulgar nossas atividades.

Discord

Além de ser a principal ferramenta de comunicação para organizar nossa equipe remotamente, nosso servidor no Discord se tornou um amplo espaço de discussão, debate e interação com usuários. Por lá tiramos dúvidas, compartilhamos nossas atividades e organizamos diferentes canais de discussão sobre a Base dos Dados e o universo dos dados abertos. Chegamos no final de 2021 com mais de 722 membros no Discord.

YouTube

O YouTube se tornou uma ferramenta estratégica para divulgação de nossos workshops e tutoriais em vídeo. Além de organizar eventos remotos ao vivo e com a participação de nossa comunidade, disponibilizamos vídeos para quem busca mais informações sobre como acessar e explorar nosso *datalake* público. Só em 2021, acumulamos 1.711 novos inscritos em nosso canal.

Conferências e palestras

MARÇO

- **1º Datathon BD 2021**: Inspirados(as) no tema do *Open Data Day 2021*, resolvemos abrir espaço para programadores, jornalistas, pesquisadores e entusiastas de dados pensarem conosco como podemos identificar ou combater desigualdades no Brasil a partir de dados públicos. Esse foi nosso **primeiro Datathon**, em que mais de 30 grupos participaram produzindo análises dentro do tema Desenvolvimento Igualitário, usando dados da BD. Os vencedores foram o grupo UFRJ Analytica (Erica Ferreira, Pedro Boechat, Pedro Borges e Rafael Ribeiro), que analisaram de que forma diferenças no acesso a uma educação de qualidade se manifestam em diferentes regiões do país, e o Felipe Macedo Dias, que analisou se a inflação foi maior em municípios mais beneficiados pelo Auxílio Emergencial.

MAIO

- **Apresentações em universidades:** FGV-RIO, Universidade Federal do Amazonas (UFA), Universidade Federal de Viçosa (UFV).
- **Curso de programação em R do Cebrap.lab**: Marcamos presença no curso de Programação em R do cebrap.lab, programa de cursos aplicados de métodos, técnicas e ferramentas de pesquisa em ciências sociais do Centro Brasileiro de Análise e Planejamento (CEBRAP). Além de apresentar a BD, mostramos aos participantes do curso como acessar nosso banco de dados público pelo ambiente da linguagem.
- **Workshop com a Escola de Dados:** Fernanda Scovino, co-fundadora da BD, apresentou um **workshop** com a Escola de Dados, ensinando a acessar nosso *datalake* público em Python e a explorar os dados históricos de desmatamento da Amazônia Legal. Ela mostrou como explorar a base de desmatamento do PRODES/INPE, além de apresentar exemplos de gráficos que podem ser gerados e como cruzar essa base com a Pesquisa Pecuária Municipal.

JUNHO

- **Palestra seminário internacional de estatística com R:** Participamos na **5ª edição do Seminário Internacional de Estatística com R**, nos dias 9 e 10 de junho de 2021. Durante a nossa palestra, apresentaremos mais a fundo o nosso pacote em R. Além de mostrar o funcionamento básico e desenvolver aplicações conjuntas com outras bibliotecas, exploramos possibilidades da utilização do pacote com programação funcional e com técnicas de regressão.

JULHO

- **Live no Curso-R:** Nossa equipe apresentou a BD em um papo descontraído sobre a origem do projeto, nossas dificuldades e conquistas. Tivemos também um ***live coding*** para demonstrar o

potencial do nosso pacote em R.

AGOSTO

- **EAESP - Aula Curso de Políticas Públicas Rudi Rocha**
- **Apresentação no Tesouro Nacional**
- **EESP - III Conferência de Ciência de Dados**
- **3º Domingo de Dados da Abraji:** Fernanda Scovino, co-fundadora da BD, participou da sessão *Tira Dúvidas: jornalismo com R e Python para jornalismo*, falando sobre nosso pacote em Python e análises de dados com nosso *datalake* público. O evento aconteceu no 3º Domingo de Dados, parte da programação do Congresso Internacional de Jornalismo Investigativo da Abraji.
- **3º Encontro Brasileiro de *Data Science* da FGV:** Nosso co-fundador, Ricardo Dahis, participou do 3º Encontro Brasileiro de *Data Science* pela FGV EESP para falar sobre a experiência da BD e o papel dos dados abertos na pesquisa e desenvolvimento de soluções para o país.

SETEMBRO

- **Apresentações em universidades:** Universidade Federal Fluminense (UFF), Universidade Federal de Pernambuco (UFPE), Universidade Federal do Mato Grosso do Sul (UFMS), **Comunicação Social da PUC-Rio**.

OUTUBRO

- **Python Brasil 2021**: João Carabetta, co-fundador da BD, apresentou um tutorial que te mostra como analisar 250GB em segundos usando Python e a Base dos Dados.
- **Semana de Sociologia do Programa de Pós da UnB:** João Carabetta, co-fundador da BD, foi um dos palestrantes na Mesa Redonda sobre **Desafios Metodológicos: como fazer pesquisa na pandemia de Covid-19**.

NOVEMBRO

- **LatinR:** Marcamos presença também no LatinR de 2021, Conferência Latinoamericana sobre uso de R em pesquisa e desenvolvimento, com participação na sessão **Periodismo de datos, datos abiertos y visualización** (Jornalismo de dados, dados abertos e visualização). Além de apresentar o pacote em R da BD, discutimos sobre a coleta e uso de dados abertos do Banco Mundial e sobre a democratização de dados públicos de Juiz de Fora (JF em Dados).
- **BD na Trilha Google do CODA BR 2021:** Fernanda Scovino, co-fundadora da BD, apresentou na sexta edição da Conferência Brasileira de Jornalismo de Dados e Métodos Digitais (CODA BR) um workshop sobre como explorar nosso *datalake* pelo Google BigQuery. Ela explicou como

acessar informações de mais de 70 conjuntos de dados completos como RAIS, CAGED, Censo, TSE, eleições etc.

- **Apresentação no E-Vigilância 2021:** Apresentamos como é possível utilizar nosso *datalake* para pesquisar sobre a pandemia no E-vigilância 2021, conferência nacional interdisciplinar sobre inovação na vigilância de doenças transmissíveis organizada pela Fiocruz.

DEZEMBRO

- **Apresentações em universidades:** Universidade Federal de Pernambuco (UFP).
- **Brazil Stata Conference**
- **Webinar no Ipea:** Nosso co-fundador, Ricardo Dahis, participou como debatedor no Webinar **Um Guia para o Uso dos Painéis da Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua**. O evento foi organizado e apresentado pelo Instituto de Pesquisa Econômica Aplicada (Ipea).

Produção de conteúdo

YouTube

- **Workshop: SQL + BigQuery**: Logo no início de 2021, lançamos nosso primeiro workshop ao vivo. João Carabetta, co-fundador da BD, apresentou o *datalake* público da BD e como usar SQL para acessar mais de 250GB da RAIS pelo BigQuery.

- **Workshop: Aplicações no R**: Ricardo Dahis, co-fundador da BD, apresentou o *datalake* público da BD e possíveis aplicações da ferramenta R.

- **Workshop: Aplicações no Python**: Realizamos também um workshop ao vivo para demonstrar como acessar e utilizar os dados da BD através do nosso pacote Python. O apresentador dessa terceira live foi o Fred Israel, co-fundador da BD.

- **Workshop: Dados do Diário Oficial da União**: Henrique Xavier, colaborador da BD, apresentou um workshop demonstrando como explorar os dados do Diário Oficial da União (DOU), disponíveis e atualizados diariamente em nosso *datalake* público.

- **Workshop: Aprenda a acessar dados públicos em R**: Pedro Cavalcante, da nossa equipe de infra, apresentou um workshop ao vivo mostrando como explorar os dados tratados e disponíveis no *datalake* público da BD através do pacote em R.

- **Workshop: Esquenta Brasileirão: Construindo estatísticas do futebol brasileiro com a BD**: Moacyr Alvim, professor da FGV, apresentou como analisar com os dados do campeonato disponíveis no nosso *datalake* público em Python, em um workshop inédito em nosso canal do YouTube.

- **SiconfiBD: Acesso prático a dados de finanças públicas**: Fernando Barbalho, cientista de dados no Tesouro Nacional, apresentou seu pacote siconfiBD, que permite você acessar e cruzar dados do Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro com diversas outras bases da BD.

Tutoriais

- **Introdução ao pacote Python**: Aprenda a acessar nosso *datalake* público;

- **BD em Python 102** (continuação da Introdução ao pacote Python);

- **Google BigQuery (SQL) 101**: Acesse diversas bases de dados públicas com uma simples consulta SQL;

- **Como acessar dados públicos em R**: Guia prático para utilizar nosso *datalake* na linguagem R;

- **Diretórios Brasileiros**: Como essa base facilita sua análise;

- **Como acessar dados da BD no Power BI**: Veja como acessar o *datalake* público da Base dos Dados no Power BI para criar gráficos, visualizações e *dashboards*;

- **Escreva na Base dos Dados**: Publique seu artigo, tutorial ou análise em nossa página;

- **Explorando o Censo Escolar com a BD+**: Uma maneira prática de analisar a mais importante pesquisa estatística educacional do Brasil.

Artigos e análises

- **Analisando preços de combustíveis com a BD+**: Veja como analisar a variação média dos preços de combustíveis no Brasil com valores corrigidos pelo IPCA;

- **O Soberano mítico**: Entenda o processo de limpeza e tratamento dos dados do Siconfi;

- **Analisando a frota brasileira com a BD**: Saiba quais são as cidades com mais carros por habitantes no Brasil;

- **Analisando dados de vacinação contra COVID-19 com a BD**: Veja como utilizar o *datalake* público da Base dos Dados para criar um gráfico de vacinação da sua cidade.

- **Em busca de dados LGBTQIA+**: Relembre nossa campanha e confira os dados LGBTQIA+ que você já pode acessar pela BD;

- **Como funciona o sistema de inserção de dados na BD?**: Conheça nossa infraestrutura de inserção de dados e veja como você pode melhorar seu portfólio contribuindo;

- **O Brasil nas Olimpíadas**: Um panorama da performance brasileira nos jogos olímpicos ao longo dos anos;

- **Um site feito a várias mãos**: Conheça o projeto colaborativo para desenvolver uma plataforma que facilita ainda mais seu trabalho com dados.

Newsletters

Em 2021 começamos a produzir nossa newsletter informativa, a *BDletter*, com divulgações, novidades e atualizações da Base dos Dados. Da primeira edição até a décima, reunimos 721 inscritos que recebem mensalmente atualizações sobre as bases já tratadas e disponíveis em nosso *datalake*, eventos e conteúdos sobre como explorar e analisar conjuntos de dados grandes e complexos de maneira simplificada.

- **BDletter #1** - Bem-vindo(a) à primeira newsletter da Base dos Dados!

- **BDletter #2** - 1º Datathon BD, como explorar dados do Diário Oficial da União e mais!

- **BDletter #3** - Lançamento do pacote em R e muitos dados novos em nosso *datalake*.

- **BDletter #4** - Lançamentos dados do Censo Escolar, workshops, tutoriais e mais!

- **BDletter #5** - Campanha #OndeEstãoDadosLGBTQIA, BD de cara nova e mais!

- **BDletter #6** - Dados históricos dos Jogos Olímpicos, como acessar a BD com Power BI e mais!

- **BDletter #7** - Nova versão do pacote em R, como usar nossa base de Diretórios Brasileiros para cruzar diferentes bases e mais!

- **BDletter #8** - Novo site no ar, dados do novo CAGED tratados e prontos para análise, BD na Python Brasil 2021 e mais!

- **BDletter #9** - Uma maneira mais prática de acessar e explorar dados de finanças públicas, BD no CODABR 2021 e mais!

- **BDletter #10** - Prêmio Tesouro Nacional 2021, dados de segurança pública do Rio de Janeiro em 3 linhas de código e mais!

Clipping

Nossos dados foram protagonistas de diversas matérias jornalísticas para importantes veículos de comunicação do Brasil. Além de uma ótima fonte para pautas, diversos jornalistas utilizam a Base dos Dados para acessar e apurar facilmente informações públicas. Ganhamos destaque também em uma **reportagem do Ijnet** sobre as iniciativas brasileiras que oferecem dados gratuitamente para jornalistas.

- Estadão: **Falta segunda dose da AstraZeneca em metade dos postos da cidade de SP**;

- Revista Piauí: **Minas Gerais é o estado com maior desigualdade salarial entre docentes homens e mulheres**;

- Revista Piauí: **Professores ganham 12% a mais que professoras no Brasil**;

- Revista Piauí: **Elas na sala de aula**;

- Nexo Jornal: **A covid-19 acabou com o 'efeito mandante' no Brasileirão?**;

- Tribuna de Minas: **Quase metade dos juiz-foranos mortos por Covid-19 em junho tem menos de 60 anos**;

- AcidadeOn: **Campinas recolhe lotes suspensos da Coronavac para devolver ao Estado**;

- O Expresso: **Um perfil dos venezuelanos e haitianos em Curitiba**;

- O Expresso: **Os novos núcleos imigrantes de Curitiba**;

- Midiamax (UOL): **PT e MDB são maiores partidos em número de filiados ativos em Campo Grande**.

Prêmios

Google Customer Award

No dia 12 de outubro, recebemos o **Google Cloud Customer Awards** na categoria de Impacto Social. O prêmio inédito foi anunciado no evento internacional, **Google Cloud Next'21**, e tem como objetivo reconhecer as implementações mais inovadoras e transformadoras do Google Cloud ao redor do mundo. Fomos a única organização brasileira a receber a premiação na categoria de Impacto Social, que também reconheceu outras iniciativas que usam tecnologia para promover mais abertura e transparência.

XXVI Prêmio Tesouro Nacional 2021

Conquistamos o 1º lugar na categoria Soluções do **XXVI Prêmio Tesouro Nacional 2021**. O prêmio tem como objetivo expandir as fronteiras do conhecimento em finanças públicas, promovendo a normalização de temas específicos quando tratados consistentemente pela pesquisa científica. Além disso, a premiação busca reconhecer o impacto e repercussão de artigos e soluções na Administração Pública.

A Base dos Dados foi selecionada por conta de nosso trabalho compatibilizando informações de despesas e receitas orçamentárias do Setor Público Brasileiro. Disponibilizamos os **dados do Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro** (Siconfi) municipal de 1989 até 2020, tratados e compatibilizados em nosso *datalake* público. Foi um enorme trabalho de compatibilização entre anos e tabelas para facilitar ao máximo a vida de nossos usuários. Padronizamos nomes, valores, contas, e criamos um identificador único por conta. Ou seja, agora é possível acessar dados de receitas correntes de São Paulo para 32 anos com apenas 3 linhas de código. Preparamos também um **artigo** para demonstrar em detalhes como foi o trabalho para disponibilizar essa base.