enap RSP

Iniciativas e desafios para prover um ambiente de compartilhamento e análise de dados corporativo: Big Data PE

Eronita Maria Luizines Van Leijden
Cassiane de Fátima dos Santos Bueno
Flávia Danzi d'Amorim
Márcio Alexandre Marques Silva

Agência Estadual de Tecnologia da Informação (ATI-PE), Recife – PE, Brasil

O presente trabalho visa discorrer sobre o processo de desenvolvimento de uma solução para compartilhamento e análise de grandes volumes de dados, chamada Plataforma de Compartilhamento e Análise de Dados (PCAD). Através de um relato de experiência são apresentadas as iniciativas, bem como os desafios e os resultados obtidos até então. A PCAD, implementada através do portal Big Data PE, foi concebida baseada em conceitos e boas práticas de governança de dados e em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD). São ilustrados alguns dos resultados obtidos com a Secretaria Estadual de Saúde e são comentados outros ganhos inerentes à plataforma, dentre os quais: o reuso de papéis, pessoas e de recursos (financeiros, tecnológicos); a simplificação e a disponibilização de processos de compartilhamento e uso dos dados; o cuidado constante com a segurança e a privacidade; a coordenação especialista centralizada; e a inclusão tecnológica de órgãos com menos recursos disponíveis.

Palavras-chave: compartilhamento de dados, Big Data, análise de dados, segurança

DOI: https://doi.org/10.21874/rsp.v73.ia.6885 ISSN: 0034-9240 | E-ISSN: 2357-8017
Submetido: 19 de abril de 2022. Aceito: 2 de maio de 2022.

Iniciativas y desafíos para proporcionar un entorno para compartir y analizar datos corporativos: Big Data PE

El presente trabajo tiene como objetivo discutir el proceso de desarrollo de una solución para compartir y analizar grandes volúmenes de datos, llamada Plataforma de Análisis e Intercambio de Datos (en portugués, PCAD). A través de un relato de experiencia, se presentan las iniciativas, desafíos y resultados obtenidos hasta el momento. La PCAD, implementada a través del portal Big Data PE, fue concebida con base en conceptos y mejores prácticas de gobierno de datos y en conformidad con la Ley General de Protección de Datos Personales (LGPD). Se ilustran los resultados obtenidos con Secretaría de Salud del Estado y se discuten otras ganancias inherentes a la plataforma, entre ellas: la reutilización de papeles, personas y recursos (financieros, tecnológicos); simplificación y disponibilidad de procesos y usos de intercambio de datos; cuidado constante con la seguridad y privacidad; coordinación experta centralizada; y la inclusión tecnológica de agencias con menores recursos disponibles.

Palabras clave: intercambio de datos. Big Data, análisis de datos, seguridad

Initiatives and challenges to provide an environment for sharing and analysis of corporate data: Big Data PE

The present work aims to discuss the process of developing a solution for sharing and analysis of large volumes of data, called Platform for Data Sharing and Analysis (in Portuguese, PCAD). Through a use case report, initiatives, challenges and results obtained so far are presented. The PCAD, implemented through the Big Data PE portal, was conceived based on concepts and good data governance practices and in compliance with the Brazilian General Data Protection Law (in Portuguese, LGPD). Some of the results obtained with SES-PE are illustrated and other gains inherent to the platform are commented, among which: the reuse of roles, people and resources (financial, technological); the simplification and availability of data for sharing and analysis processes; the constant focus on security and privacy; centralized expert coordination; and the technological inclusion of organizations with fewer available resources.

Keywords: data sharing, Big Data, data analysis, safety

1 Introdução

Nos dias atuais, a economia mundial vem comprovando a prevalência dos dados sobre os demais ativos organizacionais. E na administração pública não seria diferente: um setor público orientado a dados os reconhece como um ativo estratégico na concepção e entrega de políticas públicas e serviços (Peña-López, 2020).

Na busca pelo refinamento de processos de disponibilização de dados e pela criação de uma cultura de tomada de decisão com base em evidências, governos precisam desenvolver estruturas sólidas de governança de dados e mecanismos de entrega relacionados a infraestruturas de dados e padrões. Afinal, o compartilhamento de dados pode beneficiar no planejamento estratégico, melhorando a antecipação e resposta às necessidades dos usuários, fornecendo melhores serviços e políticas públicas que promovam integração, acesso, compartilhamento e uso de dados em todo o setor público à luz da legislação vigente (Oecd, 2019) (Garg; Agarwal; Sherry, 2004).

No setor público é comum encontrar uma abundância de dados dispostos em bases distintas que precisam ser compartilhadas entre diversos órgãos. Além disso, é desejável que esses órgãos estejam inseridos nas boas práticas de *compliance* e, assim, consigam apresentar uma boa governança desse processo de compartilhamento; é também natural em iniciativas públicas a crescente dependência de novas tecnologias – como, por exemplo, a que se convencionou chamar *Big Data* – que lidem bem com o rápido crescimento da quantidade e variedade de dados e fontes digitais, que precisam ser coletados, armazenados, recuperados e processados com alta velocidade e segurança.

Até pouco tempo atrás, o governo de Pernambuco atendia às demandas de compartilhamento de forma despadronizada; com papéis, processos, tecnologias e formalizações diferentes a depender dos órgãos envolvidos. Isso levava ao desperdício de recursos digitais, retrabalho das pessoas na operacionalização das demandas e fragilidade na segurança, contribuindo para o surgimento da necessidade de fazer convergir as demandas de compartilhamento para uma solução corporativa e acessível que permitisse o armazenamento e processamento veloz de grandes volumes de dados digitais em formatos variados.

Diante desse cenário, foi desenvolvida uma estratégia de compartilhamento

de dados, que priorizasse o reuso de dados, pessoas, processos e tecnologias, devidamente amparada por arranjos institucionais e regras formalizados em decreto, e implementada na Plataforma de Compartilhamento e Análise de Dados (PCAD), os quais serão apresentados neste trabalho.

2 Objetivos

O principal objetivo deste trabalho é mostrar como foi desenvolvida a estratégia de compartilhamento de dados no Estado de Pernambuco e apresentar como a plataforma foi criada para suportar essa política: (1) racionaliza investimentos do governo estadual em iniciativas de análise de grandes volumes de dados; (2) promove a autonomia aos encarregados ou donos dos dados nas atividades de gestão do compartilhamento; e (3) aborda a questão da segurança e da privacidade dos dados pessoais.

3 Ambiente corporativo de compartilhamento e análise de dados

A aplicação adequada das Tecnologias da Informação e Comunicação (TIC) pode possivelmente reduzir o número de ineficiências nos processos ao permitir o compartilhamento de arquivos e dados entre os departamentos governamentais, contribuindo, assim, para a eliminação de erros de procedimentos manuais, reduzindo o tempo necessário para as transações (Ndou, 2004).

Garg, Agarwal e Sherry (2004) afirmam que do ponto de vista da eficiência, para bom uso dos dados é melhor capturar os dados apenas uma vez e armazená-los e compartilhá-los para permitir a integração dos usuários entre os diversos interessados. Um único local de compartilhamento pode permitir o acesso a serviços, consultas e transações relacionadas. Esse compartilhamento de dados é fundamental para modernizar o governo, facilitando a entrega eletrônica de serviços governamentais, promovendo um atendimento de qualidade ao cliente, reduzindo a duplicação e erros. Mas esses dados abrangem um volume muito grande de espaço e para processá-los é fundamental o uso de ferramentas de alto desempenho, que possam se beneficiar de computação paralela e distribuída[1] para acelerar o processamento e prover respostas rápidas.

[1] Computação paralela e distribuída: quando um grande volume de dados a serem processados é dividido em partes menores para processamento simultâneo por unidades dedicadas, diminuindo o tempo total de resposta.

No âmbito do governo de Pernambuco, até início de 2021, o compartilhamento de dados entre os órgãos não dispunha de uma estruturação padronizada, o que levava a algumas dificuldades, como: (1) a adoção de processos de formalização e de padrões tecnológicos variados para obtenção de acesso, transmissão e consumo dos dados entre órgãos; (2) o desconhecimento holístico e transparente sobre os dados e informações corporativos gerados no âmbito do governo, eventualmente disponíveis para consumo pelos demais órgãos; (3) a insuficiência de métricas de consumo e mecanismos de *feedback* de usuários que permitissem aferir a qualidade dos dados existentes; e (4) a falta de conhecimento e controle dos compartilhamentos existentes entre os órgãos da administração pública estadual, assim como entre esses órgãos e instituições externas ao governo, sobretudo após a promulgação da Lei Geral de Proteção de Dados Pessoais (LGPD[2]).

Essas dificuldades levaram ao desenvolvimento de um ambiente de compartilhamento de dados no Estado de Pernambuco, cuja estratégia e processo de desenvolvimento estão detalhados nas seções subsequentes deste documento da seguinte forma:

- O primeiro tópico relata os desafios e problemas enfrentados no projeto;
- O tópico seguinte discorre sobre o embasamento legal para a construção do ambiente de compartilhamento em sua origem e sobre como se deu o processo de elaboração da política de compartilhamento de dados;
- O terceiro, descreve as principais funcionalidades da plataforma que corroboram para racionalizar investimentos do governo estadual em iniciativas de análise de dados; e
- O último tópico aborda a documentação gerada e ilustra os principais processos definidos.

3.1 Desafios e problemas enfrentados no projeto

Antes de simplesmente elencar os desafios enfrentados no decorrer do projeto, é importante relatar um breve histórico da evolução do trabalho, o que contextualiza e remete a outros desafios e problemas enfrentados antes e durante o projeto.

Uma vez que a Agência Estadual de Tecnologia da Informação (ATI-

[2] Lei brasileira que busca regulamentar o tratamento de dados pessoais por todos aqueles que, de alguma forma, captam informações sensíveis sobre os indivíduos, seja no meio digital ou não (BRASIL, 2018).

PE) não dispunha de orçamento suficiente para a criação da PCAD, buscou-se engajamento com um órgão que tivesse uma real demanda associada a armazenamento e compartilhamento de grandes volumes de dados e que fosse aderente à política de compartilhamento e análise de dados, de modo que esse órgão entrasse com parte do investimento inicial no projeto através da contratação de uma empresa para desenvolvimento da solução, ao passo que a ATI-PE complementasse o investimento com pessoas dedicadas tanto para a gestão do audacioso projeto quando na gestão da infraestrutura tecnológica a ser implantada. Foi feita, então, uma parceria com a Secretaria Estadual de Saúde (SES-PE), a partir da qual foi obtido êxito não só pelo patrocínio financeiro, mas também pelo comprometimento das pessoas envolvidas.

O trabalho foi coordenado pela ATI-PE através da Gerência de Governança de Dados (GDA) ligada à Diretoria de Tecnologias para Informações Corporativas (DIC) e teve forte parceria dentro da própria agência com a Gerência de Projetos e Gestão da Segurança da Informação (GSI), também ligada à DIC, e com a Gerência de Infraestrutura e Serviços Compartilhados (GIS), ligada à Diretoria de Tecnologia da Informação e Transformação Digital (DTD).

Após os trâmites e aprovações necessárias ao projeto, foi dado seguimento a sua execução com a definição da arquitetura da solução por parte da empresa contratada e da ATI-PE que, através da GIS, forneceu hospedagem do ambiente no Datacenter do Estado, numa infraestrutura de hiperconvergência[3] e topologia de rede[4] especializada que prioriza a segurança e o alto desempenho em aplicações relacionadas a *Big Data*.

No decorrer do projeto houve muita interação e apoio da equipe da GSI nas definições dos requisitos de segurança e privacidade da solução, do termo de uso e da política de privacidade, em conformidade com a LGPD e demais leis e normas correlatas, já nas primeiras versões disponibilizadas da solução.

Ao final do desenvolvimento da plataforma, iniciamos a execução de um plano de divulgação da solução implementada a partir de uma agenda de apresentações internas. Utilizamos também outros canais de comunicação,

[3] Estrutura de TI que combina armazenamento, processamento e rede em um único sistema que pode reduzir a complexidade do Datacenter e aumentar o dimensionamento.

[4] Canal no qual o meio de rede está conectado aos computadores e outros componentes de uma rede de computadores. Essencialmente, é a estrutura topológica da rede, e pode ser descrito física ou logicamente.

como a página da ATI-PE, as redes sociais e informes periódicos da Secretaria de Administração de Pernambuco (SAD-PE). Após essas divulgações, foram feitas algumas operações assistidas com órgãos que apresentavam grande potencial de uso da plataforma, como forma de capacitação *hands-on* e coleta simultânea do *feedback* dos usuários, oferecendo subsídios para correções e melhorias futuras.

Dado o exposto, salientamos que todo o projeto havia sido planejado nos seis meses que antecederam os eventos iniciais da pandemia da Covid-19 em nosso estado, atingindo em cheio o planejamento e a priorização de recursos em nossa instituição parceira, a SES-PE, e elencamos a seguir uma série de desafios enfrentados no decorrer do projeto:

- Orçamento insuficiente para alavancar o projeto corporativo completo;
- A definição do escopo e do embasamento legal, obtidos em grande parte após a elaboração e publicação da Política Estadual de Compartilhamento de Dados;
- O pouco domínio sobre as tecnologias a serem integradas para prover uma solução que fosse *open-source*[5], expansível e segura;
- A dificuldade na obtenção de infraestrutura que comportasse tamanha solução;
- A manutenção do alinhamento entre as diversas áreas envolvidas;
- A conformidade com a LGPD e demais leis e normas de acesso a dados vigentes; e
- A comunicação com os órgãos estaduais para divulgação da PCAD.

Além disso, um desafio constante ainda enfrentado e relacionado à divulgação é a contínua necessidade de demonstração da vantajosidade e dos benefícios trazidos com o uso da plataforma, em detrimento das soluções já existentes ou de outras específicas a serem desenvolvidas internamente nos órgãos. Em resumo, tentar mobilizar pessoas responsáveis pelos dados para terem em mente que de um lado existe uma demanda muito clara por uma grande quantidade de dados armazenados em silos e que podem e devem ser compartilhados com um ou mais órgãos e que, de outro lado, já existe uma solução corporativa disponível que provê ferramentas e processos para o compartilhamento de maneira estruturada, coordenada, padronizada, segura e

[5] Em português, código aberto: software cujo criador permite a qualquer pessoa utilizá-lo de graça, incluindo modificá-lo e distribuí-lo, para qualquer finalidade; no caso concreto, reduziu custos mensais por usuário.

em conformidade com as normas vigentes.

3.2 Embasamento legal

Em Pernambuco, apesar da abundância de dados disponíveis nas diversas secretarias e entidades vinculadas, foi somente no início de 2021 que se dispôs sobre uma política estadual de compartilhamento de dados suportada por uma solução que permitisse o armazenamento e processamento veloz de grandes volumes de dados digitais em formatos variados. O Decreto Estadual nº 50.474, de 29 de março de 2021, foi construído justamente para este fim, de dispor sobre a política e, ao mesmo tempo, criar uma plataforma de compartilhamento e análise de dados que a suportasse.

Inspirado no Decreto Federal nº 10.046, de 09 de outubro de 2019 (Brasil 2019), o decreto estadual também cria níveis de compartilhamento de dados, sendo feitas as devidas adequações à realidade estadual, inclusive de nomenclaturas. No decreto estadual as regras gerais de compartilhamento foram adaptadas às necessidades mais essenciais dos órgãos de Pernambuco naquele momento, dispensando, em sua primeira versão, a criação de um cadastro base do cidadão e aproveitando estruturas estaduais deliberativas de TI já existentes.

De maneira resumida, o decreto estadual dispõe sobre a política de compartilhamento de dados através da definição dos seguintes itens fundamentais:

- Três tipos de instituição que tratam e/ou fazem parte de qualquer compartilhamento de dados, que são: a Instituição Coordenadora de Compartilhamentos - ICC, a Instituição Compartilhadora dos Dados - ICD e a Instituição Usuária dos Dados - IUD, e seus respectivos papéis e competências;
- Três níveis de compartilhamento de dados: amplo, corporativo e especial; e
- Regras de compartilhamento gerais e específicas para cada um dos níveis de compartilhamento definidos.

O mesmo decreto designou a ATI-PE como Instituição Coordenadora de Compartilhamentos – ICC, então "responsável por coordenar as disponibilizações de compartilhamento, pela guarda do catálogo dos dados compartilhados, pela gestão e pelo funcionamento da Plataforma de Compartilhamento e Análise de Dados" (Pernambuco 2021, artigo 2º, inciso XVII).

A redação do decreto estadual foi feita primariamente pela ATI-PE, com revisões e contribuições do Comitê Técnico de Governança de Dados (CTGD) e do Comitê Executivo de Governança de Dados (CEGD), ambos integrantes do Sistema Estadual de Informática do Governo - SEIG e formados a partir da Lei nº 16.379, de 06 de junho de 2018 (Pernambuco 2018). A publicação do decreto só se deu após a aprovação do texto por esses referidos comitês, ouvidos os representantes técnicos de cada membro integrante.

Foi com base nesse decreto estadual que o ambiente de compartilhamento e análise de dados foi construído, conforme descrito na próxima seção, onde é detalhado em que consiste e como foi implementada a plataforma.

A PCAD tem como propósito mediar o consumo próprio das instituições participantes, bem como o compartilhamento de dados em si, introduzindo a privacidade e a proteção de dados desde a sua idealização (*Privacy by Design*), em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD), com a Política Estadual de Proteção a Dados Pessoais (Pernambuco PEPDP, 2020) e com a Política Estadual de Segurança da Informação (Pernambuco PESI, 2020), além de considerar conceitos e boas práticas de governança de dados.

3.3 Construção da solução

A Plataforma de Compartilhamento e Análise de Dados é o conjunto centralizado de recursos e ambientes de TIC que viabiliza compartilhamentos, análises e tratamentos de dados e intercâmbios de dados. É composta por recursos da Plataforma de Informações Corporativas e recursos da Plataforma de Interoperabilidade, Figura 1.

Figura 1 | Composição da Plataforma

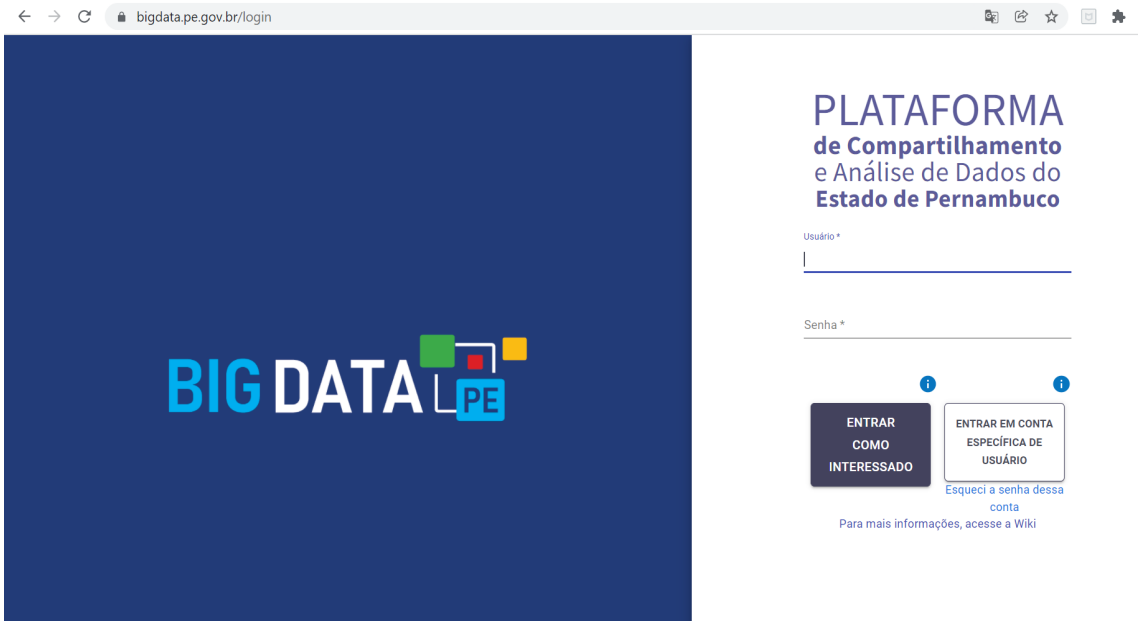
Fonte: elaborado pelos autores.

A Plataforma de Interoperabilidade[6] é um conjunto de tecnologias e ferramentas digitais que viabiliza a gestão, o monitoramento, a configuração e a implementação de interoperabilidade entre sistemas digitais para o compartilhamento de dados entre instituições (Pernambuco, 2021). Embora esta plataforma faça parte da PCAD, ela não está inserida no escopo deste trabalho.

A Plataforma de Informações Corporativas é um conjunto de tecnologias e ferramentas digitais que possibilita a coleta, o armazenamento, a classificação, o tratamento, o processamento, a análise e o compartilhamento de grandes volumes de dados (Pernambuco, 2021). Para tanto, foi desenvolvida a solução Big Data PE, cujo acesso está disponível através do portal *bigdata.pe.gov.br*. Somando-se ao portal, a plataforma se utiliza de ferramentas que favorecem a construção de soluções específicas de análise de dados para sistemas de *Business Intelligence* (*BI*) e *Business Analytics* (*BA*) corporativos.

O portal *bigdata.pe.gov.br* é uma aplicação web restrita à rede interna *pe.gov.br*, com controle de usuário próprio e autenticação de dois fatores, onde o usuário recebe um código de autenticação por e-mail para digitá-lo no portal e ter acesso à plataforma. Sua tela principal, de login, está ilustrada na Figura 2.

Figura 2 | Tela de login Portal Big Data PEE

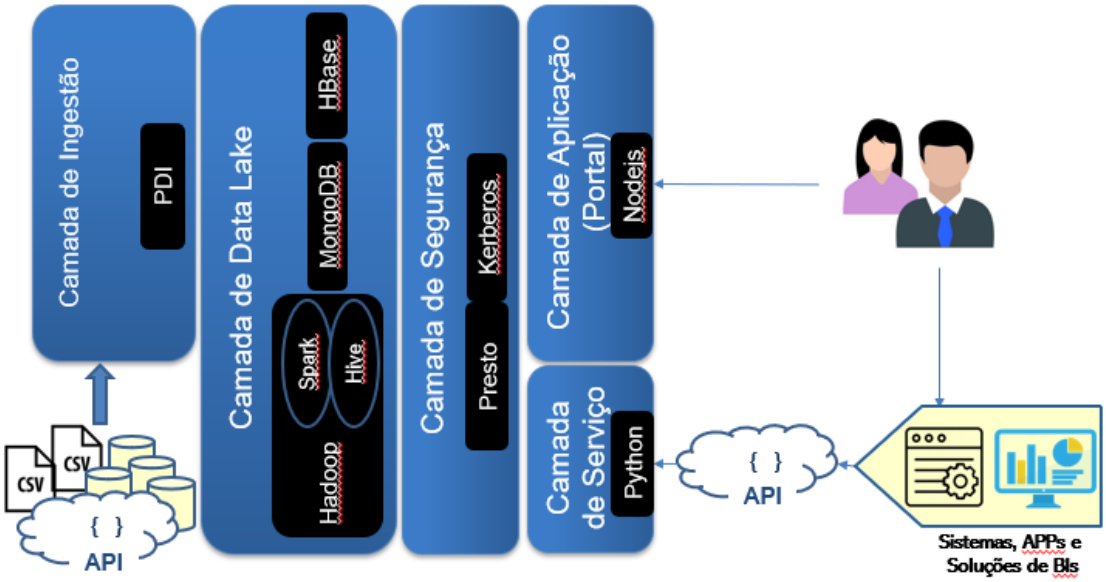


Fonte: Captura de tela da aplicação Big Data PE, extraída pelos autores.

[6] Capacidade de diversos sistemas e organizações trabalharem em conjunto, de modo a garantir que pessoas, organizações e sistemas computacionais troquem dados (Pernambuco, 2021).

A Figura 3 apresenta a arquitetura macro da solução composta de 5 camadas: ingestão, *Data Lake*, segurança, aplicação e serviço.

Figura 3 | Arquitetura Macro do Portal Big Data PE



Fonte: elaborado pelos autores.

A camada de ingestão é por onde os dados são inseridos como conjuntos de dados na plataforma, provenientes de três possíveis tipos de conexão: bancos de dados padrão SQL, *web services* ou arquivos delimitados por vírgulas (CSV). Utiliza o módulo PDI (*Pentaho Data Integration*) do Pentaho como uma de suas principais tecnologias.

A camada de *Data Lake* utiliza o Hadoop, o MongoDB e o HBase para armazenar os dados, e é através da camada de serviço que os dados podem ser consumidos via APIs (*Application Programming Interface*), sistemas e soluções de *Business Intelligence (BI).*

A camada de segurança utiliza o Presto integrado com o Kerberos. O Presto oferece ao *Data Lake* a possibilidade de consumo de dados nos diferentes repositórios de dados (HDFS, MongoDB, HBase) com uma única interface de consulta em SQL. O Kerberos tem a função de incrementar a segurança do *Data Lake*, oferecendo a funcionalidade de autenticação a serviços que não a possuem por padrão, como é o caso do Presto.

A camada de aplicação usa a API Node.js, que possibilita a construção de aplicações *web* em geral. É nessa aplicação *web* que os usuários interagem com o portal Big Data PE para acessar os serviços providos por ele.

A plataforma possui cinco possíveis perfis de usuário: Interessado, Usuário, Usuário Externo, Controlador e Superusuário, detalhados no Quadro 1.

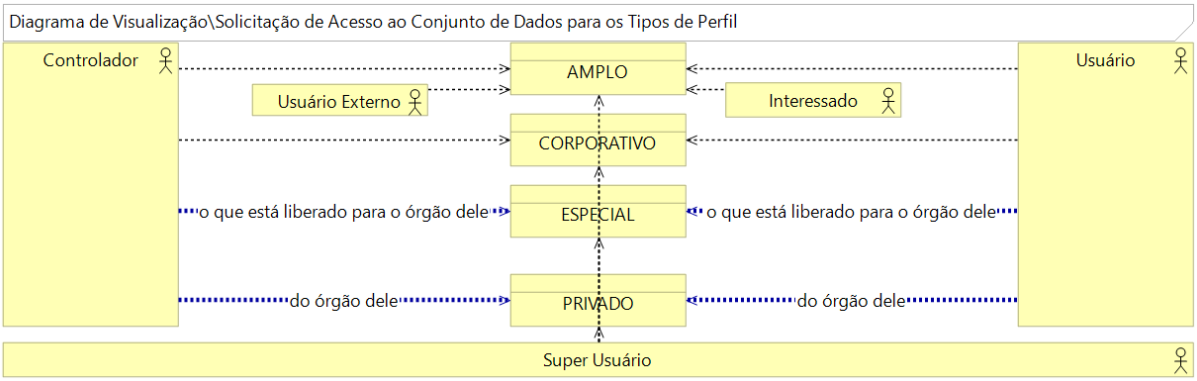
Quadro 1 | Perfis de Usuários

Perfil na plataforma	Instituição que representa	Explicação/Exemplo de usuário
Interessado	Candidato a IUD	Agente público de órgão ou entidade da APE que possua credencial de acesso aos sistemas da rede corporativa do Estado e que é cadastrado automaticamente quando da primeira tentativa de acesso à plataforma com essa credencial.
Usuário / Usuário Padrão	IUD	Agente público de órgão ou entidade da APE, podendo ser também um sistema ou aplicação. Tem permissão de usufruir dos principais serviços da plataforma.
Usuário Externo	IUD	Perfil adicional criado para representar os agentes externos.
Controlador[7]	ICD	Servidor de órgão ou entidade da APE responsável pela catalogação e compartilhamento dos conjuntos de dados do seu órgão criados por ele.
Superusuário	ICC	Servidor da ATI que tem permissão de administrar o ambiente e o uso da plataforma.

Fonte: Wiki da plataforma na ATI (2022).

As permissões de visualização dos conjuntos de dados para solicitação de acesso seguem os critérios de nível de compartilhamento definidos pelo controlador (ICD), dono dos dados, conforme ilustrado pelo diagrama da Figura 4.

Figura 4 | Visualização de Conjuntos de Dados por Perfil



Fonte: elaborado pelos autores.

Os conjuntos que o usuário visualiza são os conjuntos que ele pode solicitar acesso de acordo com o seu perfil, seu órgão e algumas características

[7] O perfil de Controlador na plataforma não tem nenhuma relação com o papel de controlador definido na LGPD.

dos conjuntos como nível de compartilhamento, instituição compartilhadora e órgãos permitidos, por exemplo.

Os conjuntos de dados caracterizados como amplo, tratam de dados que não possuem nenhuma restrição de acesso e são possíveis candidatos a dados abertos. Os conjuntos de dados com nível de compartilhamento corporativo tratam de dados protegidos por sigilo, mas com concessão de acesso a todos os órgãos. Já os conjuntos de dados caracterizados como especial tratam de dados protegidos por sigilo, mas com concessão de acesso individual a cada instituição interessada, segundo critérios e regras de segurança adicionais. O nível privado foi criado na plataforma apenas para abranger os conjuntos de dados a serem utilizados interna e exclusivamente em iniciativa da própria instituição de origem dos dados.

Na plataforma, um conjunto que possui dados pessoais só pode ser caracterizado como privado ou especial, evitando assim a possibilidade de ocorrência de compartilhamentos inadvertidos para os quais indicação de finalidade de consumo não tenha sido providenciada.

Após ter permissão de acesso ao portal, o usuário pode realizar principalmente duas ações: compartilhar e consumir conjuntos de dados. O compartilhamento de dados se dá através da inserção ou ingestão de conjuntos de dados na plataforma por um controlador (ICD) para posterior consumo dos dados por um usuário (IUD).

Os dados são inseridos na plataforma de acordo com sua estrutura (metadados) definida na criação do conjunto de dados associado, o qual é devidamente catalogado no portal pelo controlador do órgão dono dos dados (ICD).

Um órgão pode selecionar dados para consumo próprio ou para compartilhamento com outros órgãos. O consumo é feito mediante o acesso que cada usuário pode ter aos conjuntos de dados disponibilizados na plataforma, observados os níveis de compartilhamento e as regras de restrição de acesso aos mesmos. Para obtenção de acesso a um conjunto de dados, existe um processo formal a ser seguido dentro e fora da plataforma, de forma que todo o controle de acesso aos dados seja realizado por um controlador do órgão dono do conjunto de dados. O acesso, quando liberado, fornece ao usuário solicitante uma chave única (*token* de acesso) para que ele possa consumir os dados do conjunto. Este

token é uma codificação que representa uma associação entre usuário e conjunto de dados, de forma que apenas o usuário dono do *token* tem acesso ao mesmo.

O portal *bigdata.pe.gov.br* foi estruturado em oito itens principais, conforme ilustrado na Figura 5.

Figura 5 | Tela Principal do Portal Big Data PE

Fonte: Captura de tela da aplicação Big Data PE, extraída pelos autores.

A depender do perfil de usuário, diferentes itens e funcionalidades estarão disponíveis nos menus do portal, mas apenas o superusuário pode ter acesso a todos os itens e funções.

Os itens "Catálogo de Conjuntos de Dados Compartilhados", "Gestão de Dados", "Gerir Pedido de Acesso" concentram as principais funções do portal para prover o compartilhamento de dados que são as funções de solicitação de acesso a conjuntos de dados disponibilizados no catálogo para o órgão, de gestão de conjuntos de dados e de gestão do acesso dos usuários aos conjuntos de dados.

Os demais itens são de manutenção de cadastro (usuários e instituições), consultas estatísticas e de gestão de pedidos de mudança em conjuntos de dados feitos por controladores. Apesar dos controladores poderem configurar seus conjuntos de dados, as configurações ainda precisam ser revistas por um superusuário antes de serem efetivadas, devido à necessidade de verificação de disponibilidade na infraestrutura compatível com o pedido de mudança.

A plataforma possui também uma série de configurações que podem ser feitas pelos superusuários, como, por exemplo, o gerenciamento dos *drivers* de

bancos de dados, alguns controles de ciclo de vida do usuário e a atualização das versões dos documentos de termo de uso e de política de privacidade do portal Big Data PE.

Uma outra funcionalidade embutida no portal diz respeito à comunicação com os usuários. A cada operação relevante realizada no portal, um e-mail é automaticamente enviado para as pessoas interessadas, por exemplo, quando um usuário solicita acesso a um conjunto de dados, o controlador responsável pelo conjunto automaticamente recebe uma mensagem de e-mail com informações sobre a solicitação, de forma que ele não precisa ficar acessando a plataforma frequentemente para saber se houve alguma requisição que deva tratar.

Todo evento gerado por um usuário na plataforma é registrado, para fins de consulta e auditoria, atendendo aos requisitos de segurança e de rastreamento das ações dos usuários no portal.

3.4 Documentação e processos

Em virtude de a equipe ser bastante reduzida, houve uma preocupação maior com a definição dos processos e elaboração da documentação para que os órgãos que desejam compartilhar e/ou consumir dados possam ter uma maior autonomia no entendimento e uso da plataforma. As informações estão publicadas na Wiki interna da ATI (Ati, 2022), onde é possível ter acesso a toda a documentação disponível para o usuário. Além de reunir informações descritivas sobre a plataforma, também contém links para: manuais, documentos de apoio e *templates* de ofícios, termos de compromisso e formulários.

A Wiki e os termos de uso e de privacidade do portal Big Data PE podem ser acessados a partir do próprio link da Wiki ou a partir do portal *bigdata.pe.gov.br*.

Existem algumas situações no processo de compartilhamento de dados que requerem o uso de uma ferramenta de apoio para registros e formalizações corporativos. No caso da PCAD, essa ferramenta é o Sistema Eletrônico de Informações (SEI), conforme mostra a Figura 6.

Figura 6 | Macroprocessos de adesão à plataforma

Fonte: elaborado pelos autores.

Dos quatro processos de adesão ilustrados acima, três exigem formalização via SEI. Isso se deve à necessidade de registrar na ferramenta corporativa de gestão de documentos e processos eletrônicos do Estado as solicitações de adesão, utilizando-se dos instrumentos já existentes na mesma e *templates* definidos especificamente para esses casos.

Essa formalização normalmente é necessária apenas para a entrada do usuário de um determinado perfil na plataforma. Contudo, para pedido de acesso a conjunto de dados de nível especial, por envolver compartilhamento de dados sensíveis, pessoais ou não, também é uma situação que requer prévia abertura de processo no SEI.

Além dos processos de solicitação de adesão, existem os processos de solicitação de acesso a dados (consumo), que são fundamentais e estão ilustrados no diagrama de macroprocessos na Figura 7.

Figura 7 | Macroprocessos de Consumo de Dados

Fonte: elaborado pelos autores.

Dentre esses processos de consumo, apenas um exige a formalização via SEI, que é quando uma solicitação de acesso ao dado é negada e o usuário decide abrir recurso.

4 Conquistas e resultados

Completando seu primeiro ano de funcionamento em março deste ano, já é possível compartilhar diversos resultados e conquistas obtidos com a construção desse ambiente, cujos benefícios de alto desempenho, baixo custo de adoção, segurança, conformidade legal e o próprio potencial de habilitar órgãos em novas iniciativas em compartilhamento e análise de dados entre si parece ter aguçado a procura por novas adesões e começa a justificar todo o esforço empreendido até aqui.

Em termos de capacidade de armazenamento e processamento compartilhados, temos mais de 20 máquinas integradas e expansíveis, monitoradas 24 horas por 7 dias da semana, utilizando cinco tecnologias de banco de dados *open-source* prontas para serem exploradas pelos órgãos, permitindo a ingestão dos principais formatos e bancos de dados em uso, o que potencializa a redução de custos mediante o reuso de dados, pessoas e ferramentas.

A relevância em se usar um ambiente especializado de *Big Data* está na capacidade em se processar uma quantidade enorme de informações numa

tecnologia destinada para esse fim, o que ajuda a otimizar os processos de recuperação de dados e suas análises. Como exemplo prático, na Secretaria da Saúde do Estado de Pernambuco (SES-PE) existia a necessidade de se consumir dados de alguns sistemas do Ministério da Saúde (MS), compartilhando-os tanto internamente na própria SES-PE, como com outros órgãos do Estado e inclusive com instituições externas ao Poder Executivo Estadual (Tribunal de Contas do Estado, prefeituras etc.).

Antes do uso da plataforma, a SES-PE obtinha os dados da base do MS, armazenando-os temporariamente em um banco de dados convencional e consumindo-os diretamente desse banco de forma recorrente, necessitando de atuação humana para gerar e compartilhar os dados em arquivos delimitados por vírgulas (CSV) e/ou através do uso de planilhas.

No Quadro 2 é possível comparar os ganhos obtidos pela SES-PE ao aderir à PCAD:

Quadro 2 | Experiência na SES

Sistema	Extração		Compartilhamento	
	SEM uso plataforma	**COM uso plataforma**	**SEM uso plataforma**	**COM uso plataforma**
e-SUS VE[8]	Limitada a blocos de 10 mil linhas, durava em torno de 1 hora e meia para carregar toda a base.	Base completa, com aproximadamente 2,7 milhões de linhas, carregada em apenas três minutos.	Atuação humana para gerar e compartilhar os dados em arquivos csv e/ou planilhas.	Automatizado, utilizando chave de segurança (*token*) para compartilhar os dados.
GAL[9]	Limitada a blocos de 20 mil linhas, durava um pouco mais de 1 hora para carregar toda a base.	Base completa, com aproximadamente 1,4 milhão de linhas, carregada em apenas nove minutos.		

Fonte: elaborado pelos autores (2021).

A existência de uma demanda constante dos órgãos por informações oriundas de sistemas corporativos do Estado propiciou uma das grandes conquistas do uso da plataforma, que é a construção de BIs corporativos a partir dos conjuntos de dados compartilhados. Tais conjuntos, uma vez ingeridos no ambiente, podem ser reutilizados por iniciativas de análise de outros órgãos e

[8] e-SUS VE: ferramenta de registro de notificações de casos suspeitos de Covid-19.
[9] GAL: sistema aplicado aos exames e ensaios de amostras de origem humana, animal e ambiental.

permitir a criação de conhecimento novo, antes restrito aos silos de geração de informações de cada secretaria. A seguir, relacionamos os principais sistemas corporativos com projetos de ingestão de seus conjuntos de dados em curso e potencial de atendimento a todos os órgãos e entidades do Poder Executivo em 2022:

- Sistema Eletrônico de Informações (SEI): parceria com a Secretaria da Fazenda (Sefaz-PE) para a ingestão de dados e a criação de um BI corporativo, incluindo a integração com dados selecionados do planejamento e da execução financeira provenientes do Sistema Eletrônico Integrado de Informações Fazendárias (e-Fisco);
- Sistema PE-Integrado[10]: parceria com a Secretaria de Administração do Estado de Pernambuco (SAD-PE) para a ingestão e o compartilhamento seguro de conjuntos de dados selecionados dos módulos de compras e contratos;
- Sistema de Folha de Pagamento: demanda da Secretaria da Controladoria-Geral do Estado (SCGE-PE) para disponibilização de bases selecionadas para atividades de auditoria interna e exportação simplificada de informações de interesse do cidadão no Portal da Transparência.

A plataforma favorece também a inclusão tecnológica de órgãos menos providos de pessoas, orçamento e/ou recursos de infraestrutura, pois os habilita a realizarem seus projetos de dados com mais robustez, desempenho e segurança, custeando apenas o que já não esteja disponível ou adequadamente dimensionado na plataforma, como, por exemplo, em soluções específicas de análise de dados ou pontos de compartilhamento de dados[11].

Nesse sentido, há um projeto piloto em andamento, em parceria com a Secretaria de Planejamento e Gestão (Seplag-PE), que envolve a disponibilização da ferramenta *Power BI Report Server* para uso corporativo de forma integrada, usando dados da plataforma como fonte de dados na construção de painéis internos e para consumo de informações estratégicas pelos diversos órgãos contemplados. Neste caso, a Seplag-PE dispõe de várias pessoas habilitadas no desenvolvimento de painéis, mas enxergou a vantagem de migrar a solução

[10] Sistema integrado de gestão de compras, contratos, licitações, patrimônio e almoxarifado do Estado de Pernambuco.

[11] Artigo 2º inciso XI - ponto de compartilhamento de dados: recurso digital e disponível em rede, em cada extremidade do processo de compartilhamento, que permite a interoperabilidade entre sistemas/aplicações ou o intercâmbio/compartilhamento de dados (Pernambuco, 2021).

anterior para um ambiente seguro e corporativo, dispensando custos com licenciamento por usuário e contando com acesso potencial à mais bases, de forma facilitada. Essa iniciativa ainda auxilia na publicação de painéis existentes de análise de dados, elaborados pela Seplag-PE para outras secretarias, provendo a infraestrutura necessária, sem ônus adicionais para os órgãos envolvidos.

Com as recentes funcionalidades implementadas na plataforma para gestão de usuários externos, vislumbramos ainda a possibilidade de resultados relevantes em termos de segurança, transparência, colaboração e pronto atendimento a demandas de instituições externas ao Governo de Pernambuco. Exemplos importantes dessa capacidade disponível são: (1) o uso da plataforma para compartilhar informações ou prestar contas a outros entes e órgãos de controle e de defesa da sociedade, como no caso de prefeituras, do Tribunal de Contas do Estado e do Ministério Público; e (2) a disponibilização dos dados em ambientes de inovação aberta.

Especificamente sobre a inovação aberta, vale ressaltar que uma das principais dificuldades encontradas em iniciativas do tipo é o caráter experimental e menos formal na disponibilização de dados aos participantes, o que, mais à frente, pode gerar um gargalo conhecido nas áreas de tecnologia do governo para conversão de protótipos outrora aprovados em ferramentas prontas e integradas com os sistemas originais para uso no ambiente real. Tais etapas tendem a ser otimizadas com o uso da plataforma combinada à experiência de nossa equipe de especialistas.

Em termos de uso seguro dos dados, a plataforma aplicou o conceito de *Privacy by Design* para garantir sua aderência à LGPD e demais leis e normas de acesso a dados vigentes. Fez uso também de ferramentas de segurança para o acesso à plataforma e aos dados que, juntamente aos processos bem definidos, resguardam não só os agentes públicos responsáveis pelos dados a serem compartilhados, como também os agentes públicos e outros colaboradores externos que fazem o seu consumo.

A partir do momento em que é disponibilizado um ambiente de compartilhamento centralizado como a plataforma Big Data PE, observa-se a possibilidade de melhoria na qualidade e fidedignidade dos dados, proveniente de algumas das seguintes características:

- Automatização e padronização da geração dos conjuntos de dados;

- Centralização do catálogo dos dados e especificação das descrições dos dados (metadados) pelo próprio responsável dono dos dados e entendedor do serviço prestado;
- Responsabilidade pela geração dos dados facilitada para a instituição compartilhadora, com procedimento mais transparente;
- Reuso dos conjuntos pelas próprias secretarias e vinculadas, garantindo maior precisão das informações compartilhadas; e
- Custo acessível que permite inclusão e adoção da cultura de decisão orientada a dados.

5 Considerações finais

A plataforma Big Data PE permanece em constante processo de evolução, seja pela implementação de pontos de melhoria identificados pela equipe da ATI-PE, seja pela necessidade de expansão de funcionalidade ou capacidade provenientes de projetos concretos patrocinados por alguma secretaria parceira interessada, como nos casos mencionados.

Na esteira de desenvolvimento da plataforma há projetos para tratar a integração de dados classificados como amplos com o portal de dados abertos, a anonimização prévia dos dados em iniciativas de inovação aberta, o aproveitamento de dados da plataforma em integrações de sistemas por interoperabilidade e a melhoria das condições tecnológicas para uso dos dados abrigados na plataforma em soluções de *Business Intelligence* (*BI*) e *Business Analytics* (*BA*) corporativas.

Com a implantação da plataforma e a adesão gradativa das secretarias e demais entidades do Governo de Pernambuco, esperamos oferecer ao Poder Executivo Estadual e às instituições externas interessadas: menor custo total, maior produtividade sem perda de padronização nas iniciativas de compartilhamento e consumo dos dados, em conformidade com a LGPD e dispondo de maior segurança no armazenamento, tratamento e transmissão de suas informações, com a devida autonomia para decidir sobre a gestão dos compartilhamentos ofertados, de forma acessível e, ainda assim, simplificada.

Esperamos que este trabalho seja útil como ponto de partida e/ou um convite à discussão de experiências com equipes de outros estados da federação; esses gestores de negócio e de tecnologia da informação que vivem, como nós,

a dificuldade de superar a rotina do dia a dia, com todas as restrições e desafios já conhecidos, e fazer espaço para uma agenda de projetos estruturadores e tão necessários, como o que compartilhamos até aqui.

Referências

Ati. *Wiki da Plataforma de Compartilhamento e Análise de Dados*. Disponível em: https://www.wiki.pe.gov.br/mediawiki/index.php/Plataforma_de_Compartilhamento_e_An%C3%A1lise_de_Dados. Acesso em: 02/02/2022.

Brasil. *Lei nº 13.709, de 14 de agosto de 2018,* Lei Geral de Proteção de Dados Pessoais (LGPD).

Brasil. *Decreto nº 10.046, de 09 de outubro de 2019*. Dispõe sobre a governança no compartilhamento de dados no âmbito da administração pública federal e institui o Cadastro Base do Cidadão e o Comitê Central de Governança de Dados.

Garg, Kapil Mohan; Agarwal, Nikhil; Sherry, Arun Mohan. Maneuvering Best Business Practices to Improve the Quality of E-Governance Services. *Journal of Internet Banking and Commerce* 9, 2004.

Ndou, Valentina. E–Government for developing countries: opportunities and challenges. *The electronic journal of information systems in developing countries*, v. 18, n. 1, p. 1-24, 2004.

Oecd (2019), *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies, OECD Publishing, Paris.

Peña-López, Ismael *et al. Digital Government Index*: 2019 results. 2020.

Pernambuco. *Lei nº 16.379, de 06 de junho de 2018*. Altera a Lei nº 12.985, de 2 de janeiro de 2006, que dispõe sobre o Sistema Estadual de Informática de Governo - SEIG.

Pernambuco PEPDP. *Decreto nº 49.265, de 06 de agosto de 2020*. Institui a Política Estadual de Proteção de Dados Pessoais do Poder Executivo Estadual em consonância com a Lei Federal nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais).

Pernambuco PESI. *Decreto nº 49.914, de 10 de dezembro de 2020*. Institui a Política Estadual de Segurança da Informação - PESI, no âmbito da administração pública estadual.

Pernambuco. *Decreto nº 50.474, de 29 de março de 2021*. Dispõe sobre a Política Estadual de Compartilhamento de Dados e cria a Plataforma de Compartilhamento e Análise de Dados dos órgãos e entidades da administração direta e indireta do Poder Executivo Estadual.

Pernambuco. *Sistema Eletrônico de Informações (SEI).* Disponível em: https://www.sei.pe.gov.br. Acesso em: 02/02/2022.

Eronita Maria Luizines Van Leijden
https://orcid.org/0000-0002-8434-7954
Mestre em Engenharia da Computação pela Universidade de Pernambuco (POLI/UPE). Especialista em Ciência de Dados e Analytic pela Escola Politécnica de Pernambuco (POLI/UPE) e em Banco de Dados pela AESO Barros Melo. Analista em Gestão de TIC na Agência Estadual de Tecnologia da Informação (ATI-PE).
eronita.leijden@ati.pe.gov.br.

Cassiane de Fátima dos Santos Bueno
https://orcid.org/0000-0002-7129-8591
Bacharel em Ciência da Computação pela Universidade Federal de Pernambuco (UFPE). Especialista em Ciência da Computação pela Universidade Federal de Pernambuco (CIN/UFPE). Analista em Gestão de TIC, Supervisora de Governança de Dados na Agência Estadual de Tecnologia da Informação (ATI-PE)
cassiane.bueno@ati.pe.gov.br.

Flávia Danzi d'Amorim
https://orcid.org/0000-0002-8806-5137
Especialista em Tecnologia da Informação em Gestao de Projetos pela Universidade Federal de Pernambuco (UFPE). Bacharel em Ciência da Computação pela Universidade Católica de Pernambuco (UNICAP). Analista em Gestão de TIC, Gerente de Governança de Dados na Agência Estadual de Tecnologia da Informação (ATI-PE).
flavia.danzi@ati.pe.gov.br.

Márcio Alexandre Marques Silva
https://orcid.org/0000-0002-6884-6956
Especialista em Tecnologia para Negócios: AI, Data Science e Big Data (PUCRS) e em Gestão Pública (UFRPE). Gestor Governamental da Secretaria de Administração de Pernambuco (SAD-PE), Diretor de Tecnologias para Informações Corporativas na Agência Estadual de Tecnologia da Informação (ATI-PE).
marcio.marques@sad.pe.gov.br.