Arquitetura Digital Inteligência de Negócio do MCTIC

Projeto Temático: Arquitetura Digital Inteligência de Negócio do MCTIC

Relatório anual da modernização da arquitetura digital de inteligência de negócio do MCTI, contemplando os painéis temáticos elaborados

Brasília, DF

Dezembro, 2021

Centro de Gestão e Estudos Estratégicos

Relatório anual da modernização da arquitetura digital de inteligência de negócio do MCTI, contemplando os painéis temáticos elaborados. Arquitetura Digital Inteligência de Negócio do MCTIC. Brasília: Centro de Gestão e Estudos Estratégicos, 2020.

68 p. : il.

1. Ciência, tecnologia e inovação. 2. Plataforma digital de informações. 3. Sistema de apoio à decisão. 4. Inteligência de dados. Título. II. CGEE.

Centro de Gestão e Estudos Estratégicos - CGEE

SCS Quadra 9 – Torre C – 4º andar – salas 401 a 405

Edifício Parque Cidade Corporate

70308-200 - Brasília, DF

Telefone: (61) 3424.9600

http://www.cgee.org.br

Este relatório é parte integrante das atividades desenvolvidas no âmbito do 2º Contrato de Gestão CGEE – 21º Termo Aditivo, Linha de Ação: Apoio Técnico à Gestão Estratégica do SNCTI / Projeto: Arquitetura digital de inteligência de negócios do MCTIC – 8.10.53.05.01.03/MCTIC/2019.

Projeto Temático: Arquitetura Digital Inteligência de Negócio do MCTIC

Relatório anual da modernização da arquitetura digital de inteligência de negócio do MCTI, contemplando os painéis temáticos elaborados

Supervisão

Luiz Arnaldo Pereira da Cunha

Equipe técnica interna

Alberto Akira Okata

Carlos Duarte de Oliveira Junior

Carlson B. de Oliveira (Coordenador)

Marco Antônio Andrade Dias

Marcus Vinícius T. da Cunha

Wagner Alberto Soares Junior

Equipe técnica externa

Adriano Albernaz Golebiowski

Jorge Millis

Raissa Rondon

Sebastião Gonella

Victor Neves Martorelli

SUMÁRIO

1. Introdução

O Ministério da Ciência, Tecnologia e Inovações (MCTI) tem como competências o planejamento, coordenação, supervisão e controle das atividades de ciência, tecnologia e inovação, dentre outras. À Secretaria Executiva – SEXEC, por sua vez, compete supervisionar e coordenar as atividades de formulação e proposição de políticas, diretrizes, objetivos e metas relativas às áreas de atuação do Ministério, atividades naturalmente demandantes de informação de alto valor agregado.

Dentre as atividades do MCTI deve-se ressaltar o papel proeminente no Sistema Nacional de Ciência, Tecnologia e Inovação (SNCTI). Na orquestração das ações buscando maximizar resultados de interesse social e econômico para o Brasil, em meio à complexidade desse Sistema, é fundamental a capacidade de integração da informação distribuída nos atores sistêmicos relevantes para geração de conhecimento e apoio à tomada de decisão.

Este projeto visa, explorar a área de sistemas analíticos modernos (que incluem conceitos tais como *data lake*, *big data*, *business intelligence*) e contribuir para a qualidade da gestão das ações governamentais no SNCTI, aportando conhecimento para a produção de informação estratégica ao MCTI. Como ponto focal do projeto está o desenvolvimento experimental de estruturas tecnológicas e técnicas para a produção de informações e sua apresentação em ambientes virtuais inovadores criados para o apoio à tomada de decisão relacionada a políticas públicas e programas de natureza estratégica.

O objetivo geral do projeto é elaborar e disponibilizar uma Arquitetura digital de inteligência de negócio do MCTI, referenciada como **Arquitetura Digital** no decorrer deste relatório, que consiste em solução metodológica e instrumental (software e hardware) com objetiva dar suporte e prover recursos para armazenamento e tratamento de fontes de dados heterogêneas, visualização de informação e gestão de características de qualidade e segurança sobre o conjunto de dados e informações administradas.

O ambiente deve ser interoperável com sistemas de informação legados do MCTI e fontes de informação externas ao Ministério que sejam consideradas relevantes para o monitoramento e avaliação do desempenho do Sistema Nacional de Ciência, Tecnologia e Inovação.

Esse objetivo geral se decompõe nos seguintes objetivos específicos:

- Evoluir a pesquisa e o desenvolvimento de ambientes digitais de acordo com as especificações feitas pelo MCTI e atores relevantes do SNCTI.
- Disponibilizar um modelo integrado de trabalho sobre ambientes informacionais, interoperáveis, que promova a construção e manutenção de catálogo de fontes de dados e informações do MCTI.
- Disponibilizar meios para construção de análises, produção de dados agregados e indicadores com capacidade para conexão com dispositivos móveis e mobilidade em nuvem, de modo a permitir a expansão da arquitetura de informação.

Neste relatório são apresentados os resultados alcançados no projeto no ano de 2021, contemplando a implementação dos temas estratégicos estabelecidos pelo MCTI para os quais estavam presentes os insumos necessários, materializadas sobre a Arquitetura Digital implantada nas instalações físicas de Tecnologia da Informação do Ministério.

As seções seguintes apresentam os objetivos e metodologias seguidos dos resultados alcançados até o final de 2021, organizados nos temas arquitetura digital, processo de trabalho e painéis experimentais.

2. Objetivos

O objetivo deste relatório é apresentar os resultados do projeto Arquitetura Digital de Inteligência de Negócio do MCTI no ano de 2021, contemplando os painéis temáticos elaborados.

Incorpora, também, o registro dos resultados relativos à arquitetura da informação e tecnológica proposta, processo de trabalho e os demais elementos estruturantes padronizados de gestão de dados, temas centrais da missão conferida ao projeto.

O objetivo geral do projeto é disponibilizar ambiente digital que suporte o armazenamento de fontes de informações heterogêneas e permita a aplicação de metodologias de analise a partir de conjuntos de dados de brutos, dados estruturados, dados parcialmente estruturados e dados não estruturados disponíveis em diferentes formatos. O ambiente deve ser interoperável com sistemas de informação legados do MCTIC e fontes de informação externas ao ministério que sejam consideradas relevantes para o monitoramento e avaliação do desempenho de ambos os sistemas.

Esse objetivo se desdobra nos seguintes objetivos específicos:

- Evoluir a pesquisa e o desenvolvimento de ambientes digitais de acordo com as especificações feitas pelo MCTI e atores relevantes do SNCTI e do sistema de Comunicações.
- Disponibilizar um modelo integrado de trabalho sobre ambientes informacionais, interoperáveis, que promova a construção e manutenção de catálogo de fontes de dados e informações do MCTI.
- Disponibilizar meios para construção de análises, produção de dados agregados e indicadores com capacidade para conexão com dispositivos móveis e mobilidade em nuvem, de modo a permitir a expansão da arquitetura de informação.

3. Metodologia

Para o alcance desses objetivos, a condução do projeto foi estruturada em três linhas de ação principais, conforme apresentado na Figura 1. O detalhamento completo do projeto, contendo descrições das fases e distribuição no tempo, está registrado em (CGEE, 2019 e 2020).

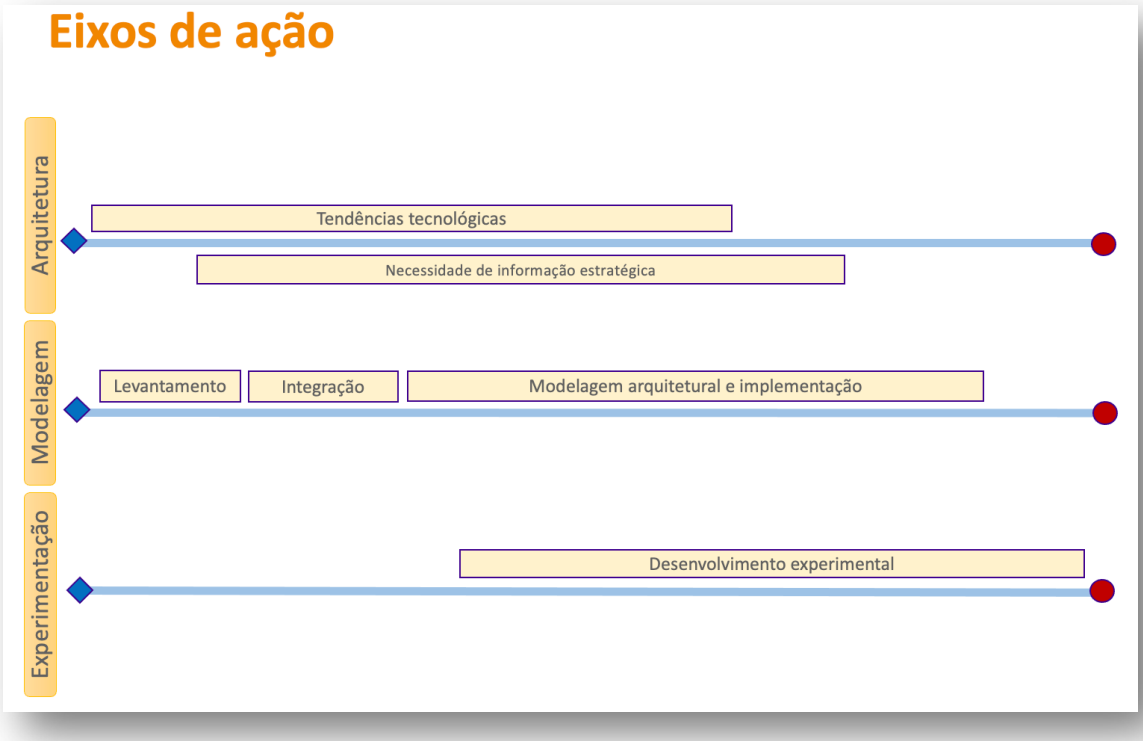


Figura 1 - Plano de trabalho do projeto. Fonte: (CGEE, 2019; 2020).

No eixo de ação Arquitetura são realizadas atividades de pesquisa, conceituação e avaliação de alternativas tecnológicas para a proposição da arquitetura digital de inteligência de negócio. Tem como objetivo a identificação das novas abordagens para tratamento inteligente de dados e geração de informação com alto valor agregado. Esse estudo resulta na avaliação conceitual e tecnológica para arquiteturas e a proposição de um modelo arquitetural que constitua um arcabouço de trabalho para atendimento de necessidades de informação estratégica do MCTI.

No existo Modelagem o foco das atividades é desenvolvimento experimental e implementação voltados para a arquitetura tecnológica e modelo de processos de trabalho. Por meio de um modelo arquitetural é conduzida a organização de processos de trabalho e a seleção, integração e apoio à implementação de instrumentação tecnológica. O resultado principal é um modelo de processo de trabalho contemplando atores, atividades, padrões e infraestrutura de Tecnologia da Informação (TI) para tratamento, análise, curadoria e padrão de qualidade de dados, incluindo catálogo de fontes de dados. Os trabalhos nesse eixo se comportam o detalhamento dos requisitos desenvolvidos no eixo Arquitetura. Para tanto, ações de levantamento das iniciativas de gestão de dados já conduzidas no MCTI e a integração dessas iniciativas são linhas condutoras para construção de evolução e melhorias nos dois modelos – arquitetural e de processo.

Na linha de ação de Experimentação, ambos os modelos construídos nas linhas anteriores e suas implementações práticas são exercitados e implementados. Esses experimentos têm como matéria prima assuntos indicados pelo Departamento Governança Institucional (DGI) da Secretaria Executiva (SEXEC/MCTI), denominados **temas estratégicos**, para os quais foi possível a interlocução com as áreas finais demandantes ou a disponibilização dos dados brutos. Cada tema estratégico exemplifica uma necessidade de informação estratégica do MCTI e dá origem a um produto de informação.

Os temas estratégicos trabalhados até o momento, seus respectivos produtos de informação e situação ao final de 2021 são mostrados na Tabela 1.

Tabela 1 - Temas estratégicos definidos pelo MCTI. Fonte: elaboração própria.

Tema estratégico	**Produto de informação**	**Situação ao final de 2021**
Indicadores COICT	Banco de variáveis	Implementado
Lei do Bem	Painéis de informações (*dashboards*)	Implementada versão 1, implementação de versão 2 em andamento
FNDCT / Fundos Setoriais	Painéis de informações (*dashboards*)	Implementado
Incentivos Fiscais para o Setor de Tecnologias da Informação e Comunicação	Painéis de informações (*dashboards*)	Aguardando disponibilização dos dados brutos

Consistente com as atuais metodologias ágeis, foi explorada a natureza iterativa de elaboração da arquitetura digital e sua implementação. A implantação de fluxos de dados, a identificação e internalização de ferramentas de software, a elaboração e validação de padrões e processos de trabalho, foram ações que se desenvolveram de forma interativa e cíclica com as equipes do Ministério. Isto se traduz em implantações de versões da arquitetura e seu contínuo aprimoramento e evolução à medida em que se desenvolvem as atividades em cada linha de ação da abordagem metodológica.

A abordagem metodológica central para desenvolvimento de artefatos de inteligência de dados foi o Ciclo de Inteligência em CTI do CGEE (CGEE, 2017), representado graficamente na Figura 2.

Figura 2: Ciclo de inteligência em CTI. Fonte: (CGEE, 2017).

Na primeira etapa do ciclo procura-se compreender as necessidades de informação para a tomada de decisão, ou seja, expressar as incertezas e as dificuldades da organização em relação ao seu processo decisório. Estas incertezas são desdobradas em tópicos e questões chave de inteligência conhecidos que, ao serem respondidos, construirão uma estrutura orientadora para a coleta de informações relevantes. Na abordagem do CGEE, essas questões chave são denominadas perguntas norteadoras.

Se inicia nessa primeira etapa ações a elaboração da estrutura orientadora para responder à necessidade de informação, envolvendo a coleta e tratamento de dados e geração de informação, denominada de narrativas de respostas. Essas narrativas expressam alternativas metodológicas, potenciais fontes de dados, necessidades técnicas previstas inicialmente, até as proposições iniciais do "como" atender às perguntas norteadoras. Além disso, faz parte da primeira etapa do ciclo a definição e objetivos, desafios e outras características associadas à tecnologia que poderá ser empregada para atender à necessidade de informação.

Na etapa seguinte, denominada Coleta e Armazenamento de Dados, são executadas as seguintes tarefas:

- obtenção de informações a partir de fontes primárias e secundárias;
- definição dos processos de coleta de informações;
- definição dos modelos analíticos que serão posteriormente utilizados, para planejar a organização do ambiente de armazenamento das informações.

Todos os dados e informações coletados são considerados inteligência bruta e, necessitam ser trabalhados para que o seu valor possa emergir na etapa de análise. Por mais qualidade que tenha uma informação, é muito mais o modo como ela será analisada e utilizada do que apenas a sua captura e disponibilização que determinará sua valia.

Essas ações da segunda etapa podem ajustar, refinar e gerar novas perguntas norteadoras e correspondentes narrativas de respostas.

Na etapa de Análise dos Dados, transformam-se as informações coletadas em um produto de inteligência. O objetivo é definir o melhor ou os melhores métodos de análise das informações para a geração dos produtos de inteligência que se pretende. Neste momento as narrativas de respostas têm suas implementações realizadas com as tecnologias pensadas (ou preparadas) para o projeto em caráter piloto ou experimental. Por meio das informações reunidas, esta etapa visa a identificação de tendências e padrões significativos, ou seja, percepções exclusivas e conexões até então não relacionadas entre

os dados. A condução do CGEE incorpora, também, nesta etapa a experimentação de mecanismos de visualização (por exemplo rascunhos visuais, ou *mock ups*) final sempre que possível.

A etapa da Produção de Resultados e Avaliações envolve a entrega do produto de inteligência, em um formato coerente, claro, objetivo aos clientes finais. Para que o uso ou disseminação dos resultados seja eficiente alguns aspectos precisam ser observados, como por exemplo, o melhor formato do documento a ser entregue pelos profissionais de inteligência para os responsáveis pela tomada de decisão na organização.

A etapa da Avaliação da Informação tem dois objetivos:

- avaliar se o processo desenhado foi eficiente do ponto de vista da elaboração do produto de inteligência. Diz respeito ao desempenho de cada uma das etapas que compõem o ciclo de inteligência, isto é, se o melhor método de análise foi escolhido, se a escolha das fontes de informação poderia ter sido mais bem direcionada, se o formato do produto foi o mais adequado e assim por diante;
- avaliar a eficiência deste produto para o cliente final, ou seja, verificar os resultados práticos obtidos com o uso dos produtos gerados para o cliente de inteligência.

Estas duas avaliações são imprescindíveis tanto para o aprimoramento do processo quanto para a sua sobrevivência. A consolidação e o reconhecimento da utilidade deste processo só são possíveis a partir dos resultados de seus produtos na tomada de decisão. Caso as atividades do processo terminem na produção de resultados, a organização terá somente adquirido informação, uma vez que a inteligência somente ocorre quando os resultados do processo são utilizados na definição das ações organizacionais.

4. A Arquitetura Digital de Inteligência de Negócios do MCTI

A Arquitetura Digital de Inteligência de Negócios do MCTI tem como conceito central a constituição de um ambiente digital para gestão de dados que:

- suporte o armazenamento de fontes de informações heterogêneas,
- permita a aplicação de metodologias de tratamento e análise de dados a partir de conjuntos de dados de brutos, e
- viabilize a produção e visualização de informação com alto valor agregado para subsídio à tomada de decisão dos gestores e analistas do MCTI.

A arquitetura digital tem uma missão inerentemente conectada que consiste em promover e contribuir para o alcance dos pilares de gestão de dados - integridade, confiabilidade, disponibilidade, autenticidade.

4.1. Modelo Arquitetural

O cerne desse ambiente digital de informação é o modelo arquitetural em camadas que estrutura o processo geral de **tratamento de dados** e **produção de informação**, alinhado com os requisitos de gestão de dados, em especial sua governança. O **produto de informação**, ou seja, a informação propriamente dita que atende à demanda estabelecida pelo usuário final, subsidia a **geração de conhecimento** para esse usuário. Adota-se, como significado de "geração de conhecimento", a tomada de decisão, a geração de ideias ou a realização de ações tornadas possíveis pelo subsídio da informação.

Cada camada do modelo organiza elementos estruturantes que compartilham de um determinado nível semântico associado a um objetivo de negócio. Ou seja, as camadas mais baixas agrupam conjuntos de dados e lógicas de tratamento de dados mais gerais com objetivo de receber dados brutos e prepará-los para o reuso. As camadas mais altas incorporam as necessidades específicas de informação, ou seja, objetivos de negócio, que se refletem em tratamentos e conjuntos de dados resultantes semanticamente alinhados com necessidades específicas de tomada de decisão. Por fim, a cada camada do modelo é acrescentado valor agregado ao dado até alcançar o produto de informação.

O modelo arquitetural funciona como um arcabouço lógico para a construção de fluxos de coleta e tratamento de dados e geração de informação com alto valor agregado. Esses fluxos de dados se expressam no encadeamento de transformações de dados realizadas com auxílio informatizado (ferramentas de *big data*). Cada transformação aplicada envolve a aplicação de regras de negócio para alcance de objetivos específicos e contemplam, também, aspectos de qualidade de dados (integridade, confiabilidade, disponibilidade, autenticidade).

O arcabouço de trabalho do modelo arquitetural, funciona também como um mapa. Permite a localização de conjuntos de dados, rotinas automatizadas de tratamento de dados, ferramentas de *big data* e padrões e rotinas de governança de dados. Associa esses elementos com a infraestrutura de Tecnologia da Informação que provê o poder computacional e de armazenamento e orienta a elaboração dos processos de trabalho, explicitando os atores envolvidos.

A Figura 3 apresenta as quatro camadas do modelo arquitetural e seus objetivos principais.

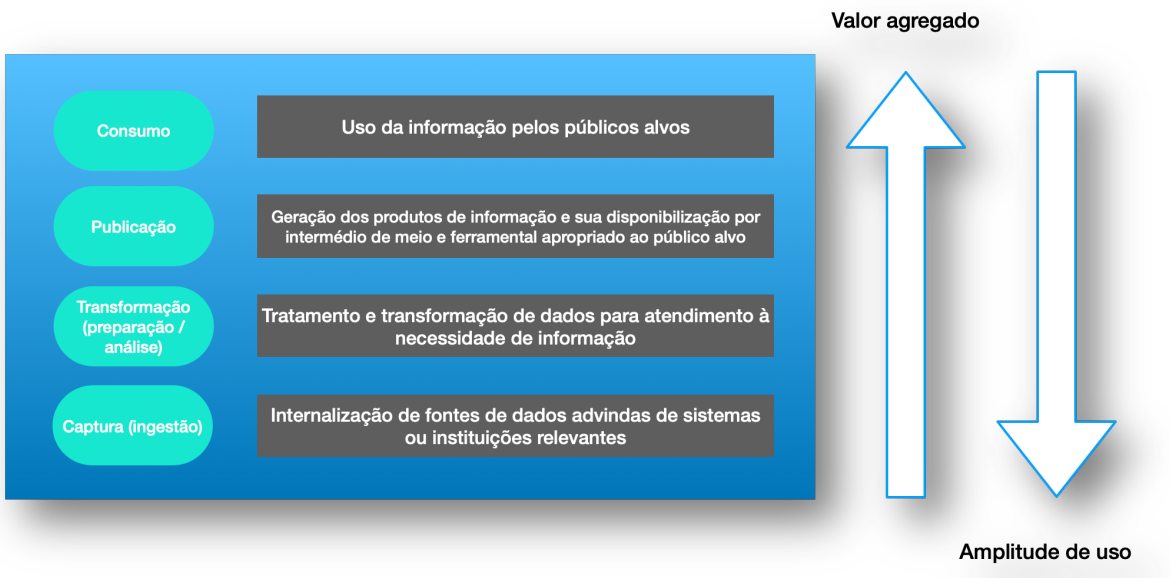


Figura 3 - Camadas do modelo arquitetural e características principais. Fonte: elaboração própria.

Na Figura 4 são apresentados os conceitos arquiteturais associados às camadas da arquitetura digital. Como se observa, fronteiras que delimitam o escopo do sistema de informação analítica, foco deste projeto, dos demais sistemas transacionais internos do MCTI ou de outras instituições. Essa distinção posiciona e define o conceito de "fonte de dados" usado neste projeto, que consiste em um provedor de dados externo ao sistema analítico representado pela Arquitetura Digital. Assim, uma fonte de dados poder ser, por exemplo, um provedor pode ser um sistema transacional interno ou externo ao MCTI, um serviço ou repositório digital de dados pagos ou aberto.

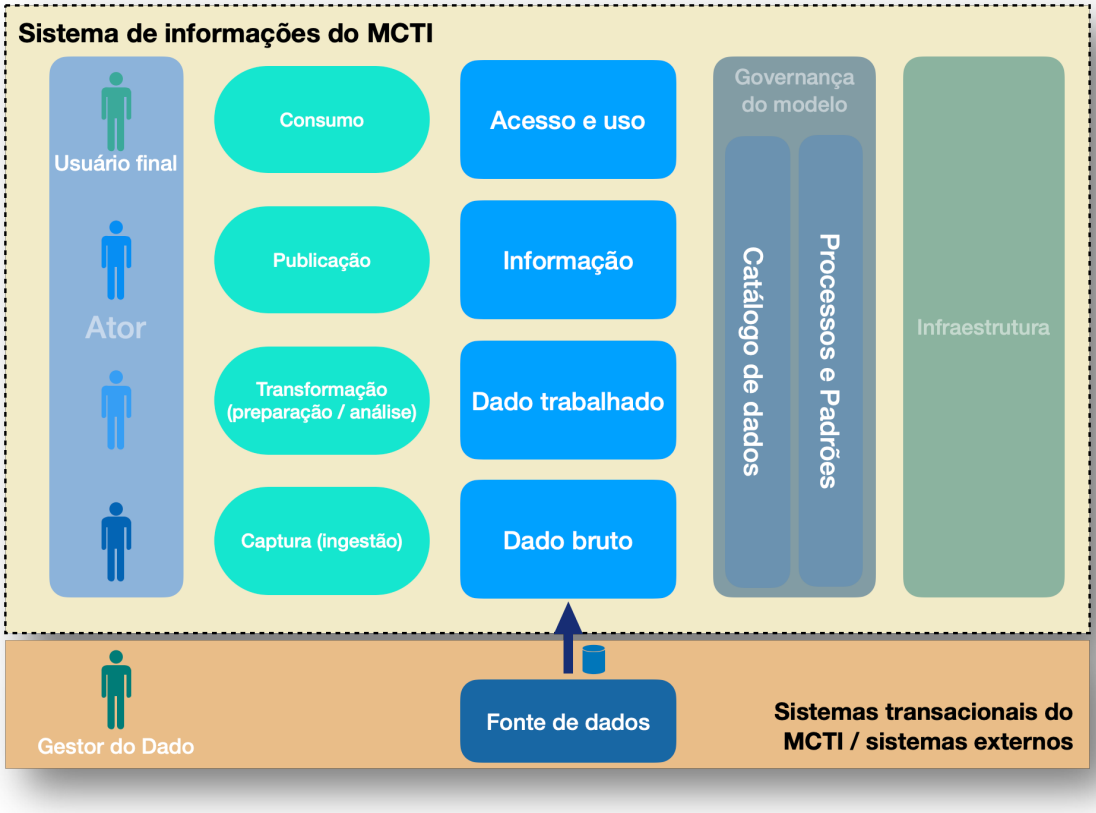


Figura 4 - Visão geral de componentes e fronteiras do modelo arquitetural. Fonte: elaboração própria.

O objetivo da modelagem contempla, também, o estabelecimento das bases do "Sistema de informações do MCTI" iniciativa interna do Ministério descrita em (MCTI, 2020), que tem como missão a organização dos dados e informações dispersas no Órgão e suas entidades vinculadas e integração com objetivo de prover subsídio à tomada de decisão nos níveis estratégicos. O aparato digital e o processo de trabalho que deriva do modelo arquitetural proposto tem como objetivo subsidiar esse sistema analítico de informações,

o qual está assentado sobre o conjunto de sistemas transacionais internos do MCTI ou externos ao MCTI (incluindo aí aqueles sistemas transacionais das suas unidades vinculadas). Ainda que os sistemas transacionais / externos não façam parte do modelo arquitetural, eles se comunicam com a arquitetura digital por meio de fontes de dados. Nesse ponto de contato entre o modelo e o mundo "externo" ocorrem atividades articulação com o responsável pela fonte de dados (Gestor do Dado).

4.2. Elementos do modelo arquitetural

4.2.1. Atores

Na ponta superior da arquitetura está o **Usuário final**. Representado como um **Ator** que se relaciona com a camada de consumo da informação do modelo. Representa os variados tipos de interesses e necessidades de informação de gestores do MCTI, analistas, técnicos do Ministério, bem como pessoas externas interessadas nos dados e informações produzidas pelo Órgão.

Em termos gerais, um **Ator** representa a pessoa humana que interage com a arquitetura digital em conformidade com sua necessidade, objetivos e alinhado com direitos e deveres estabelecidos na **Governança do modelo**.

O **Gestor do Dado** é um ator, que mesmo estando fora da fronteira do sistema analítico, é de suma importância pois detém o poder de decisão e os meios de disponibilizar um conjunto de dados de interesse do usuário final. Esse ator pode ser uma pessoa ou unidade organizacional interna ao MCTI, outras instituições ou provedores de dados, via serviços pagos ou gratuitos, como por exemplo, os portais de dados abertos dos governos. As articulações entre o usuário interessado no dado e o responsável pelo dado demandado é uma ação externa ao modelo arquitetural, por isso o posicionamento externo desse ator em relação à fronteira do modelo. Entretanto, esse ator precisa ser conhecido e documentado, como parte dos metadados do conjunto de dados que ele disponibiliza.

Atores interagem com a arquitetura digital por meio do consumo ou produção dos elementos típicos de cada camada. As ações de consumo ou produção podem ser

intermediadas por ferramentas de software, que compõem o leque de opções disponibilizados na Infraestrutura, e são regulados pelas definições e padrões estabelecidos pela Governança de Dados.

Os atores são caracterizados sob dois aspectos principais: (a) sua relação formal com o MCTI e (b) sua relação com a arquitetura digital.

No contexto da Política de Dados Abertos (MCTI, 2020a e BRASIL, 2020), assunto conduzido pelo DGI, foram definidas responsabilidades centrais para sua execução. Essas definições originam direitos e deveres a serem observados nas suas relações com a arquitetura digital. Assim, se estabelece, no contexto da relação formal com o MCTI, os perfis relacionados na Tabela 2, onde são apontados os principais perfis e suas características relevantes no relacionamento com a arquitetura digital. A relação não é exaustiva e recomenda-se as referências citadas para um detalhamento completo.

Tabela 2 - Extrato de atores definidos na minuta de portaria de dados abertos do MCTI. Fonte: elaboração própria a partir de (MCTI, 2020a).

Ator (MCTI, 2020a) – Capítulo II – Art. 6º	Características
Área de Inteligência de Negócio e Informação	No regimento interno vigente corresponde ao DGI, e tem foco de atuação nos assuntos qualidade de dados, metadados, *data* warehousing e *business intelligence* (em conjunto com a área de Tecnologia da Informação), e gestão de conteúdos e documentos (em apoio à Área de Comunicação Social), do complexo de governança de dados segundo (DAMA, 2017).
Área de Tecnologia da Informação	No regimento interno vigente corresponde do DTI, e é responsável pelos assuntos *data* warehousing e *business intelligence* (em conjunto com a Área de Inteligência de Negócio), operação e armazenamento de dados, segurança de dados, integração e interoperabilidade de dados, modelagem e projeto de dados,
Área de Comunicação Social	Gestão de conteúdos e documentos (DAMA, 2017), no que tange a comunicação digital por meio da internet (Web) e redes sociais.

Unidades organizacionais	Por meio de seus gestores, analistas, técnicos, as unidades organizacionais do MCTI trazem consigo a demanda de informação, bem como a disponibilização de dados (fontes de dados). Além disso, atua no fornecimento de requisitos para os produtos de informação desejados, ou seja, estabelece a necessidade de informação. Conforme as habilidades e conhecimentos de seus representantes, pode atuar na realização, propriamente dita, de modelagem e projeto de dados (capítulo da governança de dados).

Sob o enfoque do relacionamento do ator com a arquitetura digital, a Tabela 3 apresenta os principais perfis operacionais e suas características com o objetivo de evidenciar as competências relevantes e as ações típicas em cada caso.

Tabela 3 - Perfis operacionais de atores previstos no Modelo Arquitetural. Fonte: elaboração própria.

Camada	**Ator**	**Características**
Externo	Gestor de dados	Ator externo à fronteira do modelo arquitetural que é responsável por fonte de dados de interesse para a arquitetura digital.
Captura	Engenheiro de dados	Planejamento, execução e gestão de atividades de extração, tratamento e ingestão de dados com uso de tecnologias de *big data*, observando políticas e padrões pertinentes. Faz parte da Equipe de Gestão de Dados.
Captura	Administrador de infraestrutura de rede	Planejamento, execução e gestão de infraestrutura de Tecnologia da Informação, incluindo hardware, software e redes de computadores. Faz parte da Equipe de Infraestrutura de TI.
Transformação	Analista de dados, Cientista de dados	Planejamento, execução e gestão de atividades de modelagem, tratamento, enriquecimento, armazenamento e disponibilização de dados e informações, com uso de tecnologias de *big data* ou ciência de dados, observando políticas e padrões pertinentes. Faz parte da Equipe Analítica.

Publicação	Cientista de dados, Analista de visualização de dados	Planejamento, execução e gestão de atividades de modelagem, tratamento, enriquecimento, disponibilização e visualização de dados, com uso de tecnologias de visualização e de *big data*, observando políticas e padrões pertinentes. Faz parte da Equipe Analítica.
Consumo	Usuário final	Interação com os produtos de dados e de informação, observando políticas e padrões pertinentes.

É importante ressaltar que a relação do ator com a arquitetura digital pode se dar com qualquer camada do modelo, conforme sua necessidade e conhecimento.

4.2.2. Camadas

A estruturação do modelo em camadas reconhece distinções semânticas nos elementos constituintes do repositório digital resultante da implementação física da Arquitetura Digital. Essas distinções semânticas implicam em diferenças no tratamento dos objetos contidos em cada camada em relação a competências, ferramentas, padrões e processos. Assim, o modelo busca categorizar e agrupar os fatores comuns por camadas e otimizar a gestão do ambiente de dados e informações resultante.

No que tange a dados, elemento central da arquitetura, a Tabela 4 apresenta a categorização dos objetos. A demais dimensões da arquitetura (competências, ferramentas, padrões e processos) são alinhadas para o melhor tratamento e gestão do tipo de dados característico de cada camada.

Tabela 4 - Tipos de dados e informações característicos das camadas do Modelo Arquitetural. Fonte: elaboração própria.

Camada	**Elementos típicos (Dados)**
Captura	Fontes de dados em formato bruto. Exemplo: Arquivos CSV, tabelas no *SQL Server*, API de conexão com a fonte de dados externa.
Transformação	Dados modelados e trabalhados, em qualquer nível de prontidão para seu uso final, e armazenados em gerenciadores de dados apropriados para os objetivos finais de uso. Exemplos: tabelas no *SQL Server*, índice no

	ElasticSearch, arquivos de intermediários ou finais de ferramentas tais como RStudio e Python.
Publicação	Informação trabalhada, que pode estar armazenada em repositório ou disponível por meio de ferramentas de software para visualização. Produto de dados e Produto de informação. Exemplos: variáveis, métricas e indicadores preparados para apresentação, persistidos ou disponíveis por meio de painéis de informação, ou outros meios de visualização, serviços Web disponíveis para acesso pelo usuário interessado.
Consumo	Produto de informação nas mãos do usuário final, por exemplo: gráficos, mapas, tabelas, relatórios disponíveis em sítios web interativos com painéis gráficos ou georreferenciados, ou conjunto de dados disponível na página de Dados Abertos do MCTI; interfaces de acesso por programas (API) para interação entre computadores.

Em cada camada um leque de métodos e ferramentas são mais adequadas em vista dos tipos de dados e seus propósitos. Na camada de "Ingestão", as tarefas são dominadas por métodos de extração e movimento de dados, com uso de métodos de web scraping, interoperabilidade entre sistemas, desenvolvimento de rotinas de ETL (do inglês *Extract, Transform, Load* - extração, tratamento e carga de dados).

Já na camada de "Transformação", as tarefas são mais sofisticadas. Envolvem o pensar em como responder à necessidade de informação do usuário final. Portanto levam a necessidade de levantamento de requisitos, projetos de sistemas (analíticos), modelagem de dados, projeto de interfaces e implementação dos artefatos de softwares produção de informação com valor agregado. Entram em cena metodologias também sofisticadas de desenvolvimento de sistemas analíticos, *business intelligence*, aprendizado de máquina, dentre outras. No exercício conduzido neste projeto, foi adotado a metodologia o Ciclo de Inteligência Estratégica (CGEE, 2017) para na camada de "Transformação".

Na camada de "Publicação" o foco é a visualização de informação, com seu repertório de alternativas gráficas de visualização de informação complexa, e ferramentas atuais que facilitam a criação de visualizações ricas com integração de diferentes formas de prover informação (gráficos, mapas, tabelas etc.). Foi utilizada a ferramenta Microsoft PowerBI em consonância com escolhas já realizada pelo MCTI para essa função.

4.2.3. Governança do modelo

A definição de **Governança de dados** utilizada pelo Governo Federal é o "exercício de autoridade e controle que permite o gerenciamento de dados sob as perspectivas do compartilhamento, da arquitetura, da segurança, da qualidade, da operação e de outros aspectos tecnológicos" (BRASIL, 2019, Art. 2o - Inciso XV). Essa definição também é encontrada em (DAMA, 2017) onde são detalhadas as disciplinas associadas, mostradas na Tabela 5.

Tabela 5 - Disciplinas de governança de dados. Fonte: (DAMA, 2017), tradução própria.

Disciplina	
Data Architecture	Arquitetura de dados
Data modelling and design	Modelagem e projeto de dados
Data storage and operations	Gestão de ferramentas e armazenamento de dados
Data security	Segurança de dados
Data integration and interoperability	Integração e interoperabilidade de dados
Document content and management	Gestão de conteúdo e documentos (digitais)
Reference and master data	Gestão de dados corporativos
Data warehousing and business intelligence	Gestão de armazéns de dados e inteligência de negócio
Metadata Management	Gestão de metadados
Data quality	Qualidade de dados

O universo de atividades e entregáveis contidos nessas disciplinas é muito amplo e uma abordagem exaustiva dessas disciplinas não é objetivo ou escopo do projeto atual. O projeto busca agregar conhecimento e atividades exploratórias às iniciativas em curso no MCTI, assim como, integração incremental com os investimentos já realizados, contribuindo para o aumento do nível de maturidade de gestão de dados do Ministério.

Na camada denominada Governança do Modelo a atenção é dirigida aos itens mostrados na Figura 5. Assim, algumas disciplinas de governança de dados são abordadas diretamente, outras parcialmente ou indiretamente.

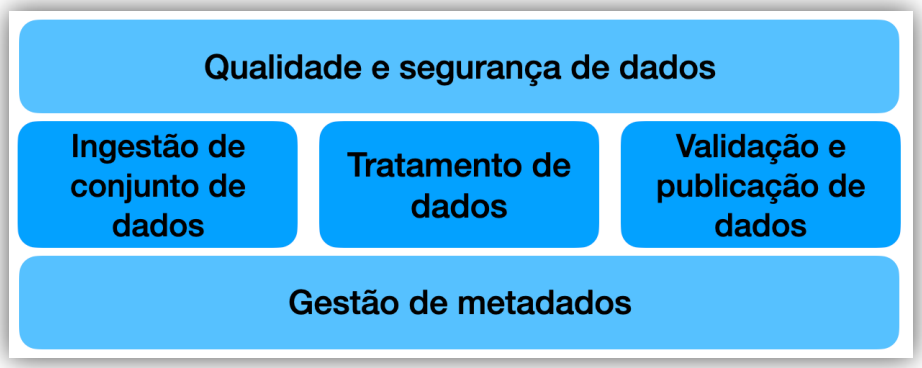


Figura 5 - Atividades em foco na Governança do Modelo. Fonte: elaboração própria.

Dentro da governança do modelo se encontra um dos elementos fundamentais da arquitetura digital, o **Catálogo de dados**. Esse componente, que no DAMA (2017) está associado ao conceito de repositório de metadados (*Metadata Repository*), tem a importante e necessária função de prover informação descritiva sobre o acervo de dados e informações digitais da organização em detalhes suficientes para que o usuário (final ou técnico) saiba o que está disponível e como acessá-los. Um catálogo de dados é implementado como um repositório digital de dados com funcionalidade de pesquisa e exploração dos conjuntos de dados disponíveis na arquitetura digital.

Outro elemento constituinte da governança do modelo é o conjunto de padrões e processos elaborados para conferir reuso, critérios de qualidade e organizar o trabalho sobre a arquitetura digital. A adoção de padrões tem como foco definir os critérios para documentação de conjuntos de dados, seu tratamento e experiência do usuário, assim como requisitos de qualidade de dados e segurança da informação.

A organização do trabalho, com foco em repetitividade e aumento de produtividade, é alcançada por meio do esclarecimento dos processos de trabalho. Esses processos estruturam o fluxo de tarefas e decisões recomendadas desde o estabelecimento da demanda de um usuário final até o provimento do produto de informação, passando pelas tarefas intermediárias de desenho e implementação da solução da demanda, enriquecimento da arquitetura com novos recursos (conjuntos de dados reutilizáveis por

outros usuários) e alinhamento dos resultados (intermediários ou finais) aos requisitos de qualidade e segurança da informação.

4.2.4. Infraestrutura

O componente **Infraestrutura** do modelo arquitetural representa *hardware, software* e redes de comunicação de dados (local, externa ou internet) onde se materializam todo o arranjo de coleta, tratamento e visualização de dados informações.

No projeto atual, a infraestrutura consiste nos seguintes elementos principais:

- um repositório digital: dados e informações armazenados e as respectivas ferramentas de gestão desses dados,
- um conjunto de processos de tratamento de dados implementados em software, bem como o conjunto de ferramentas de software que permitem a execução desses tratamentos e visualização de dados, e
- *hardware* para disponibilizar capacidade de armazenamento, de processamento e de comunicação digital de dados.

Esse conjunto integrado de elementos é referenciado como **plataforma digital**. Neste projeto a implementação da plataforma digital é realizada pelo Departamento de Tecnologia da Informação (DTI) do MCTI com o apoio técnico do CGEE.

Atualmente, com a evolução tecnológica na área da computação em nuvem (seja ela nuvem privada, híbrida ou pública) o detalhamento de configurações de computadores e capacidades de redes de comunicação se mostra extremamente dinâmica. As capacidades podem variar conforme a demanda colocada pelo conjunto de usuários sobre a plataforma digital. Além disso, podem ser providas de forma física (com acréscimos de componentes de *hardware*) ou virtual, ou seja, alocações dinâmicas, permanentes ou temporárias, de capacidade de *hardware* interna ou externa à infraestrutura física do Ministério.

O componente infraestrutura da arquitetura digital representa o conjunto integrado de hardware, software e serviços de nuvem que provê o suporte concreto para o uso das

modernas técnicas e métodos de tratamento inteligente do volume de dados necessários para a implementação da inteligência de negócio para o MCTI

4.3. Arquitetura tecnológica

4.3.1. *Data lake* do MCTI

A implementação resultante do modelo arquitetural em 2021 é apresentada na Figura 6, em termos de seus componentes tecnológicos. As ferramentas de software dessa arquitetura tecnológica são descritas resumidamente na Tabela 6. O conjunto das ferramentas selecionadas atendem a um requisito fundamental estabelecido pelo Ministério, a saber, alinhamento com a infraestrutura de Tecnologia da Informação já disponível no MCTI. Com isso é maximizado os investimentos já realizados e promovida a evolução da plataforma digital para atendimento dos novos requisitos advindos do ambiente de inteligência de dados.

O cerne da arquitetura apresentada é um *cluster* de contêineres Linux orquestrados pelo Kubernetes que hospeda o Microsoft Big Data SQL Server. Esse *cluster* disponibiliza o ambiente de processamento de transações e capacidade de armazenamento de dados que implementa o ***data lake* do MCTI**.

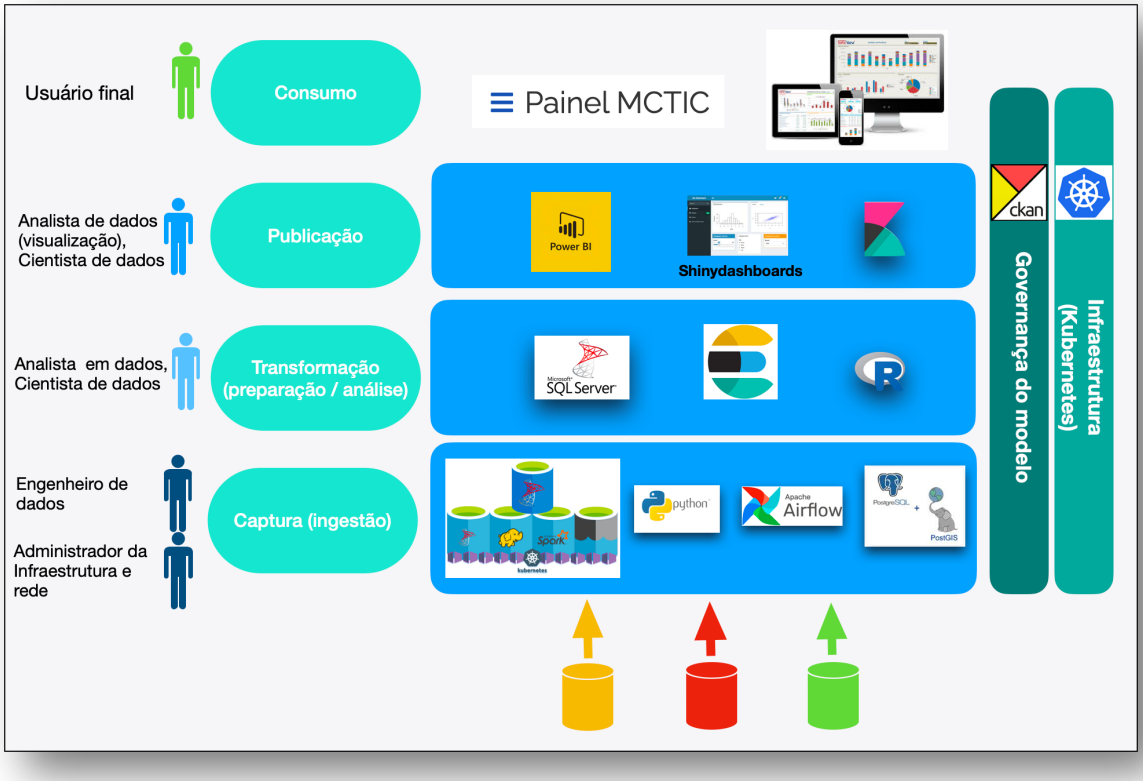


Figura 6 - Visão geral de componentes arquiteturais com respectivas implementações técnicas. Fonte: elaboração própria.

Tabela 6 - Detalhamento dos elementos tecnológicos da implementação inicial do modelo arquitetural. Fonte: elaboração própria.

Componente tecnológico	**Função principal**
Apache AirFlow	Automação de processos de ETL (*Extract, Transform and Load*).
PostgreSQL / PosGIS	Gerenciador de banco de dados relacional. Armazenamento e acesso de dados relacionais e georreferenciados.
MS SQL Server Big Data Clusters	Disponibilização do Hadoop de forma integrada com o, e abstraída pelo, MS SQL Server, o qual é um outro gerenciador de banco de dados relacional. Provê o ferramental de software do cerne de armazenamento e tratamento de grandes volumes de dados, com alta capacidade de processamento com uso intenso de paralelismo.

ElasticSearch e Kibana (ELK)	Sistema integrado de coleta de dados e provimento de mecanismos de busca e visualização, incluindo visualização georreferenciada.
R Studio	Ambiente de desenvolvimento integrado para a ferramenta estatística R. R é uma linguagem de computação estatística e gráficos.
ShinyDashboards	Framework web para criação de aplicativos interativos com linguagem R.
Power BI	Ferramenta e serviço na internet para elaboração e disponibilização de visualizações interativas com recursos de business intelligence.
Site https://paineis.mctic.gov.br/	Página web do MCTI cuja intenção é concentrar a disponibilização de informação sobre atividades e resultados de Ciência, Tecnologia e Inovação para público externo.

Os temas estratégicos trabalhados estão hospedados nessa plataforma digital. Como mencionado no capítulo da metodologia, foi utilizado o processo incremental e iterativo da abordagem da metodológica adotada (Ciclo de Inteligência em CTI) para os desenvolvimentos experimentais dos temas estratégicos. A iteratividade e o incremento de cada rodada do ciclo de inteligência impulsionaram a identificação e integração de componentes da arquitetura, produzindo evolução técnica e operacional da arquitetura digital. Essa abordagem permitiu a internalização de novas tecnologias diretamente na infraestrutura de TI do Ministério, suavizando a curva de aprendizado e garantido que os resultados se materializem internamente ao Ministério desde sua concepção. Evita-se com isso a necessidade de atividades posteriores de transferência de conhecimento e tecnologias.

Por sua importância central no *data lake* do MCTI, vale o detalhamento da arquitetura tecnológica do MS SQL Server Big Data Clusters, mostrado na Figura 7 e comentado em seguida.

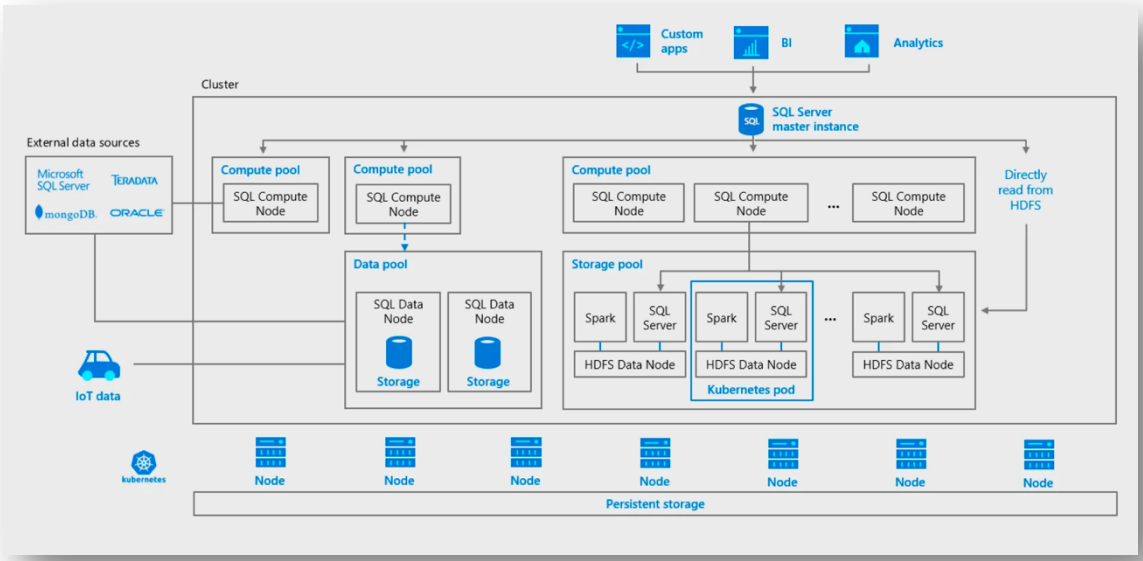


Figura 7 - Módulos da implementação física do Microsoft Big Data SQL Server. Fonte: (WRIGHT, 2018).

A infraestrutura tem como base o Clusters de Big Data SQL Server da Microsoft, isso em estrutura de cluster de contêineres Linux orquestrados pelo *Kubernetes*.

Dentre outros componentes, o cluster Big Data SQL Server possui componentes do Hadoop e seu ecossistema, tal como HDFS para armazenamento distribuído, o Spark e seus componentes para o processamento distribuído e o Apache Kafka para o barramento de mensagens, isso além do próprio SQL Server e os componentes de gerenciamento do cluster.

Foi uma solução com certa novidade da Microsoft quando de sua implantação no MCTI (2020) que apresentou resultados satisfatórios até o momento. Apresenta facilidade de implementação, gerenciamento, integração dos componentes internos e externos e segurança, em especial em instalações que já utilizam do sistema gerenciador de banco de dados SQL Server da Microsoft. O MCTI já possuía um relacionamento comercial com a fabricante da solução, facilitando assim a aquisição e a evolução da arquitetura com os produtos do fabricante.

A solução permite a implementação de um importante conceito no mundo *big data*, denominado virtualização de dados (*data virtualization*). Virtualização de dados (ROUSE, 2019) é uma abordagem de gestão de dados que permite que uma aplicação acesse e manipule dados sem necessidade de conhecer ou implementar os detalhes técnicos de acesso ao dado, como por exemplo, qual o formato do dado e onde o dado está localizado. Uma camada de abstração é provida, na forma de um servidor de virtualização de dados (DMBOK, 2020, p.294), de tal forma que a aplicação acessa diretamente o dado sem a necessidade de remanejamentos de dados ou cópias. Essa abordagem facilita e, em alguns casos, dispensa a realização de ETL.

No caso do Big Data SQL Server, essa facilidade é disponibilizada por meio do SQL Server (vide Figura 8) executa a função de abstração do ambiente Hadoop e outros gerenciadores de bases de dados, realizando as operações de ETL e integração de dados internamente e provendo a figura de virtualização de dados para o usuário. Assim, o desenvolvedor de uma aplicação de BI pode consultar tabelas de dados de diferentes formatos de origem ou tipos e fabricantes de gerenciadores de dados, sem se preocupar se a implementação física está em uma base de dados SQL Server, Postgres, NoSQL, Hadoop HDFS, arquivos CSV etc.

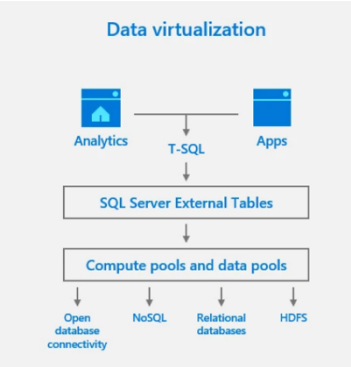


Figura 8 - Estrutura lógica da virtualização de dados com uso do MS Big Data SQL Server. Fonte: (WRIGHT, 2018).

4.3.2. Catálogo de dados

Outro componente fundamental de arquitetura digital de um *data lake* é a ferramenta de gestão de catálogo de dados. Como já mencionado anteriormente, o catálogo de dados

tem a importante e necessária função de prover informação descritiva sobre o acervo de dados e informações digitais da organização em detalhes suficientes para que o usuário (final ou técnico) saiba o que está disponível e como acessá-los.

No ano de 2021 foram conduzidos estudos de alternativas de ferramentas de catálogos de dados. Os requisitos estabelecidos junto a Ministério são apresentados na Tabela 7.

Tabela 7 - Requisitos de seleção de ferramenta de catálogo de dados. Fonte: elaboração própria.

Funcionalidade	Descrição da funcionalidade	Atendimento obrigatório ou não
Ferramenta livre ou proprietária	Ferramentas livre podem ser baixadas, modificadas e utilizadas na instalação física do Ministério sem custos ou dependência de um fornecedor. Ferramentas proprietárias requerem procedimento de aquisição da licença ou subscrição de uso junto a um fornecedor e, normalmente, não permite modificações ou acesso ao código fonte da ferramenta. Para o MCTI esse requisito é satisfeito se a ferramenta for livre, ou disponibilizar uma versão livre.	Sim
Repositórios de metadados / Metadata repositories	Usado para documentar e manusear metadados e realizar análises utilizando metadados. Organizações também podem utilizar repositórios para publicar informações sobre fontes de dados reutilizáveis, as quais fornecem ao usuário a capacidade de fazer buscas em metadados.	Sim
Glossário de negócio / Business glossary	Um repositório que é usado para comunicar e manusear os termos de negócio na empresa. Além de suas definições e os relacionamentos entre eles.	Não

Caminho do dado / Data lineage	Especifica a origem do dado e por onde ela caminha ao decorrer do tempo. Também descreve o que acontece ao dado ao percorrer este caminho. O caminho do dado ajuda a mapear os pontos chaves da origem de um dado que sirva algum propósito em particular.	Sim
Análise de impacto / Impact analysis	Cobre detalhes da dependência de informações ou o impacto que a alteração de uma fonte de dados pode causar.	Não
Manutenção de regras / Rule management	Automatiza a validação de regras de negócio que são amarradas a elementos de dados e metadados associados. Permite por exemplo criar regras que os metadados tem de seguir para garantir confiança na fonte de dados. Por exemplo: Avisar que existem dados de idade com número negativo.	Não
Arcabouços semânticos / Semantic framework	Inclui suporte para taxonomias, modelos de entidade relacionamento e modelos como UML por exemplo.	Não
Ingestão de metadados e tradução / Metadata ingestion and translation	Permitir a ingestão de metadados, ou seja, o mapeamento dos bancos, de várias fontes de dados como: Aplicações de etl, Ferramentas de BI, ferramentas de modelagem, dbms, nosql, hadoop. As ferramentas devem possuir a habilidade de identificar, documentar e manter os relacionamentos entre os metadados inseridos.	Sim
Controle de acesso a dados e metadados / Data and metadata access control	Permitir o controle por usuário no acesso aos conjuntos de dados ou aos seus metadados.	Sim
Perfil dos dados / Data profilling	Perfil dos dados é o processo de revisar as estruturas dos dados, conteúdo e relacionamentos. Perfil dos dados envolve coletar estatísticas descritivas, analisar tipos de dados, tamanho de campos e padrões.	Não

Interface amigável para o usuário	Interface com usuário fácil e intuitiva, com curva suave de aprendizado.	Não
Contagem de downloads de bases	Funcionalidade específica que permite identificar quantas vezes um conjunto de dados foi acessado (baixado) por um usuário.	Sim

Foram identificadas as ferramentas indicadas na Tabela 8 e o resultado do estudo de atendimento aos requisitos estabelecidos foi a escolha da ferramenta CKAN[1]. Além de atender aos requisitos importantes obrigatórios estabelecidos, é uma ferramenta usada extensivamente no Governo Federal (sítio de Portal Brasileiro de Dados Abertos, assim como em vários dos Ministérios e outros órgãos públicos). Outro benefício da ferramenta é sua capacidade de customização e integração com outros sistemas via sua API.

Tabela 8 - Ferramentas de catálogo de dados estudadas no projeto. Fonte: elaboração própria.

Ferramenta	**Detalhes**
Apache atlas	https://atlas.apache.org/
Amundsen	https://www.amundsen.io/
CKAN	https://ckan.org/
Dataverse	https://dataverse.org/
Socrata	https://dev.socrata.com/data/
Google Cloud Data Catalog	https://cloud.google.com/data-catalog
Talend Data Catalog	https://www.talend.com/products/data-catalog/
Ovaledge Data Catalog	https://www.ovaledge.com/data-catalog
Informatica Data Catalog	https://www.informatica.com/products/data-catalog/enterprise-data-catalog.html
Aginity	https://www.aginity.com/active-analytics-catalog/
Waterline Data	https://www.trifacta.com/partners/waterline-data/
Linkedin DataHub	https://github.com/linkedin/datahub
Kylo	https://buildmedia.readthedocs.org/media/pdf/kylo/master/kylo.pdf

[1] https://ckan.org/ - Sistema de gerenciamento de dados de código aberto utilizado para implementação de repositório integrado e portais de dados.

Magda	https://magda.io/
Cdap	https://github.com/cdap-guides/cdap-bi-guide

5. Padrões e Processos

Um resultado importante do desenvolvimento experimental dos temas estratégicos foi a validação e refinamento da estrutura de governança do modelo. A implantação da ferramenta de catálogo de dados traz consigo a necessidade de definições de padrões para dar nome aos objetos armazenados no *data lake* e quais atributos devem ser utilizados para descrever esses objetos – metadados. Por outro lado, a condução do Ciclo de Inteligência em CTI para o desenvolvimento dos temas estratégicos evidenciou quais atores, atividades e pontos de interação com o ambiente organizacional do Ministério são necessários. Trazer à tona essas características práticas do trabalho de implementação de produtos de informação com o uso do *data lake* instruiu a elaboração do modelo de processo de trabalho.

Neste capítulo, são mostrados os resultados alcançados para esses dois assuntos centrais dentre o conjunto das preocupações preconizadas na literatura especializada e das boas práticas de mercado atuais.

5.1. Nomenclatura, metadados, linhagem de dados

No *data lake* do MCTI dois aspectos relativos a conjunto de dados receberam atenção especial no contexto da governança do modelo. Como descrever conjuntos de dados e como descrever as transformações aplicadas sobre os dados.

Para a descrição dos conjuntos de dados é levado em consideração que é necessário dar nomes a esses arquivos, aos seus campos e descrever sua semântica, ou seja, seu significado para a lógica de negócio. Assim, foi produzido um padrão de nomenclatura para arquivos e campos, aplicável aos seus diferentes tipos, que também leva em consideração em qual camada do modelo arquitetural o conjunto de dados reside.

Para descrever o significado para o negócio, foi elaborado um padrão de metadados, que também leva em consideração as camadas do modelo arquitetural.

Essas definições de metadados fazem parte do documento **Padrão para nomenclatura de objetos em *data lake***, que se encontra no Anexo I deste relatório. Esse padrão ainda está em pleno exercício em vista da continuidade dos trabalhos de desenvolvimento dos temas estratégicos e, portanto, o padrão constitui um documento vivo que continua a ser refinado até a conclusão do projeto.

A descrição das transformações sobre os dados, conhecido como **linhagem de dados** (*data lineage*), será detalhada nas seções subsequentes. Descrever transformações sobre os dados está intimamente relacionado com a metodologia de inteligência de dados utilizada para atender uma necessidade de informação. Por isso, no decorrer dos trabalhos de desenvolvimento dos temas estratégicos, essa descrição foi feita no bojo da documentação de especificação dos painéis. Um padrão mínimo com objetivo de ser utilizado independente de metodologia aplicada está em processo de construção.

5.2. Modelo de processo

O atendimento à demanda de informação dos usuários finais com o uso do *data lake* do MCTI envolve atividades administrativas e técnicas. O primeiro tipo se refere à interação entre unidades e atores organizacionais na formulação da demanda, autorizações e providências de cunho administrativo. As atividades técnicas, por sua vez, são aquelas realizadas para o atendimento propriamente dito da necessidade de informação apresentada pelo usuário e são conduzidas com o apoio da arquitetura digital.

Para ambos os casos foi modelado um **processo de inteligência de negócio alinhado com a arquitetura de inteligência estratégica do MCTI**.

5.2.1. Modelo do processo de trabalho

O modelo geral do processo de trabalho é apresentado na Figura 9, onde se percebe integração de ações dos usuários finais e dos departamentos DGI e DTI e outros atores relevantes no contexto de gestão de dados. O processo de trabalho organiza e descreve um fluxo de tarefas que:

– Contempla as atividades necessárias para a elaboração de produtos de informação para subsídio à tomada de decisão dos gestores do MCTI;
– Promove o atendimento aos pilares de gestão de dados (integridade, confiabilidade, disponibilidade, autenticidade); e
– Está integrado com as características regimentais do Ministério.

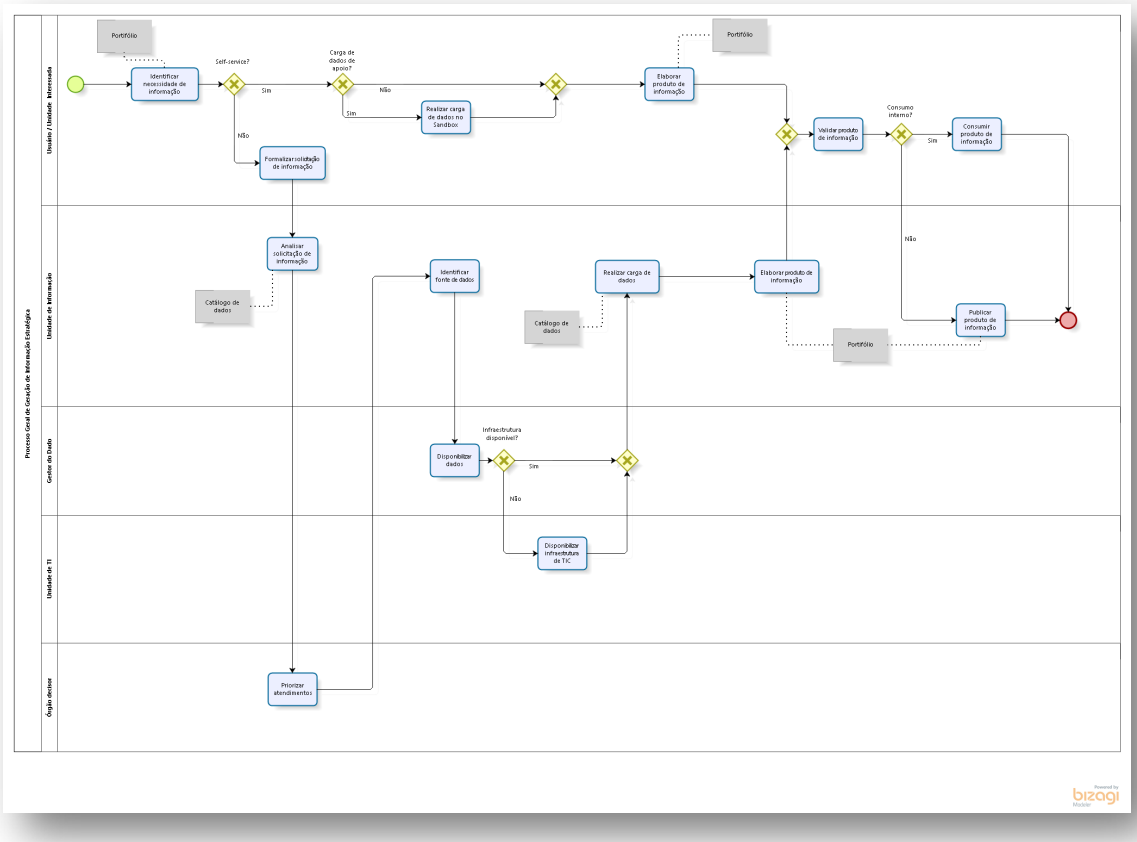


Figura 9 - Modelo de processo de trabalho sobre a Arquitetura Digital. Fonte: elaboração própria.

No Anexo III deste relatório são apresentados os sub processos associados ao processo geral.

5.2.2. Processo de inteligência de dados e linhagem de dados

O processo de inteligência de dados envolve as tarefas específicas de transformação de dados na informação desejada pelo usuário. Tem como componentes resultantes um método de transformação, artefatos de software que implementam esse método e produtos

de dados (conjuntos de dados) intermediários e produtos de informação, os resultados que atendem a necessidade de informação do usuário.

Os métodos podem transitar desde as tradicionais metodologias de sistemas de inteligência de negócio (BI) até as novas metodologias de inteligência artificial, passando pelos métodos estatísticos e aprendizagem de máquinas.

No modelo arquitetural esse processo pode perpassar camadas e é organizado em sub processos de desenvolvimento e implantação de fluxos de captura, transformação de dados e produção de visualizações, conforme mostrado na Figura 10. O principal sub processo relacionado com inteligência de dados é "Elaborar produto de informação".

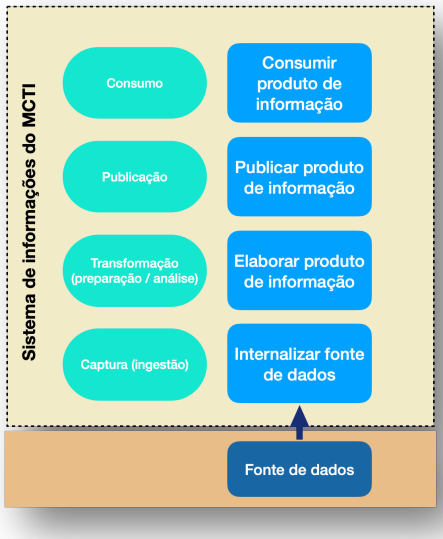


Figura 10 - Tarefas do processo de inteligência de dados e relação com as camadas do modelo arquitetural. Fonte: elaboração própria.

De forma geral esse é o cerne do processo de ingestão e tratamento de dados, e visualização de informação. Cada instância de processo é aplicada à demanda específica de informação, no caso atual, a cada um dos temas estratégicos definidos. Em cada instância de execução desse processo consiste na integração de componentes metodológicos utilizados, a saber, fontes e conjuntos de dados, métodos de tratamento de dados, variáveis, métricas e indicadores, opções de visualização ou disponibilização da informação, incluindo ferramentas eletrônicas. Todos esses componentes são

arregimentados para a elaboração de um todo coerente, um fluxo de transformação de dados que objetiva o atendimento à necessidade de informação.

Com objetivo de descrever e documentar esses fluxos de dados, foram elaboradas **linhagens de dados** (*data lineage*) que consistem em descrições de fluxo de transformação de dados desde o ponto de sua origem até o ponto de seu uso (STEENBEEK, 2019; DAMA, 2017). Este produto é associado a cada instância de execução do processo de inteligência de dados, de forma a manter uma trilha de informação sobre como um dado é utilizado, assim como por quais transformações passou um dado que faz parte de um painel (produto de informação).

6. Temas estratégicos

O exercício de validação da Arquitetura Digital produziu insumos importantes sobre a implementação e internalização dos modelos no MCTI. A partir dos exercícios de implementação dos temas estratégicos foi possível construir uma primeira aproximação para a arquitetura da informação subjacente à ação do Ministério. Outro resultado foi a disponibilização de painéis temáticos, que sob uma óptica mais prática, representa a aplicação e validação dos modelos, padrões processos de trabalho e ferramentas tecnológicas em produtos de informação implantados e disponibilizados para usuários internos.

Os desenvolvimentos experimentais realizados têm como matéria prima assuntos indicados pelo Departamento Governança Institucional (DGI) da Secretaria Executiva (SEXEC/MCTI), denominados temas estratégicos. Cada tema estratégico exemplifica uma necessidade de informação estratégica do MCTI e dá origem a um produto de informação. Portanto, requerem interlocução com as áreas finais demandantes bem como a disponibilização dos dados brutos. Esses pré-requisitos implicaram em redefinições do conjunto de temas estratégicos selecionados e, em alguns casos, a necessidade de suspensão temporária das atividades.

6.1. Arquitetura da informação

A arquitetura da informação da Arquitetura Digital de Inteligência de Negócios do MCTI é a estruturação temática dos ambientes de dados de caráter corporativo de modo a organizar a conjuntos de dados do *data lake* do MCTI e subsidiar a elaboração de política de governança de dados. Requisitos tais como, definição de gestores dos dados, controle de acesso e auxílio na busca e reuso dos dados disponíveis no *data lake* são otimizados por meio da clara delimitação das categorias de informação inerentes à missão do Ministério.

A arquitetura em questão é materializada por meio de Áreas Temáticas, categorias amplas de assuntos abrigados na competência regimental do Ministério e sua forma de

organização. As áreas temáticas são pontos de convergência dos conjuntos de dados, métricas e indicadores correlatos aos respectivos temas que residem na plataforma digital.

As categorias definidas pelas Áreas Temáticas dão origem às políticas próprias, adequada ao tema e à área gestora principal, aplicadas sobre seus respectivos conjuntos de dados. Essas políticas contemplam atividades tais como: controle de acesso, critérios de curadoria dos dados, padrões de nomenclatura e de metadados. Além disso, as categorias funcionam como qualificadores dos dados (domínios em uma base de dados) produzindo uniformidade necessária para integração de dados advindos de diferentes fontes de dados.

A proposta tem como origem atividades e ideias debatidas no contexto do projeto junto às equipes do DGI e do DTI do MCTI, insumos advindos do Modelo Arquitetural proposto neste projeto e do "Guia Operacional: Regras e Padrões para a classificação de conjunto de dados no Datalake", produzido pela COGCD/DGI-MCTI (COGCD, 2020).

A estrutura geral do conteúdo da arquitetura da informação, ou seja, o conteúdo de informação é organizado em três nível principais, apresentados a seguir.

6.1.1. Assuntos e Temas

O primeiro e segundo níveis de organização estão associados aos assuntos organizacionais do MCTI (Tabela 9).

Tabela 9 - Estrutura do conteúdo da arquitetura da informação inicial. Fonte: (COGCD, 2020).

ASSUNTO	TEMAS
POLÍTICAS SETORIAIS DE CT&I	– Política Nuclear – Política Espacial – Política de Energia, Petróleo e Mineração – Política de Inclusão Digital e Tecnologias Assistivas – Política de C&T para o Desenvolvimento Sustentável (Meio Ambiente/ Clima) – Política de Projetos na Fronteira Tecnológica – Política de Bens Sensíveis – Política de Transformação Digital – Bioma – Bioecnomia – Biotecnologia
	– Sistema Nacional de Ciência, Tecnologia e Inovação (SNCTI) – Propriedade Intelectual

ASPECTOS TRANSVERSAIS DA POLÍTICA DE CT&I	– Fomento em Ciência e Tecnologia – Extensão e Serviços Tecnológicos (ACTC[2]) – Instrumentos de incentivos Fiscais e apoio ao P&D Empresarial – Fundos CTI – Ações de Empreendedorismo
ACOMPANHAMENTO DE ATIVIDADES DO MCTI	– Acompanhamento das Unidades de Pesquisa – Acompanhamento das Entidades Vinculadas

6.1.2. Ações Institucionais

Na medida da necessidade de maior detalhamento categórico, utiliza-se o terceiro nível de granularidade para uma classificação mais detalhada que aborda Ações institucionais. Na Tabela 10 são mostradas as ações identificadas, organizadas por Assuntos e Temas.

Tabela 10 - Detalhamento das áreas temáticas. Fonte: (COCDG, 2020).

Assuntos	Temas	Ações institucionais
Políticas Setoriais de CT&I	Política Nuclear	Radiofármacos e Radioisótopos P&D Nuclear Proteção radiológica social e ambiental Segurança Nuclear Controle de Materiais Nucleares
	Política Espacial	Defesa Cibernética Setor Aeroespacial: – Veículos Lançadores – Base Lançamento / Monitoramento
	Política de Energia, Petróleo e Mineração	
	Política de Inclusão Digital e Tecnologias Assistivas	Emissão de CO2 Mudanças Climáticas
	Política de C&T para o Desenvolvimento Sustentável (Meio Ambiente/ Clima)	
	Política de Projetos na Fronteira Tecnológica	Novos Materiais Fotônica Biologia Sintética Nanotecnologia Materiais Avançados

[2] Atividade Científica e Técnica Correlata

	Política de Bens Sensíveis	
	Política de Transformação Digital	Manufatura Avançada Indústria 4.0 Inteligência Artificial IoT Cidades Inteligentes Agricultura 4.0 (de precisão)
	Bioma	
	Bioeconomia	
	Biotecnologia	
Aspectos Transversais da Política de CT&I	Sistema Nacional de Ciência, Tecnologia e Inovação (SNCTI)	Indicadores Nacionais
	Propriedade Intelectual	Núcleo de Informação Tecnológica (NIT)
	Fomento em Ciência e Tecnologia	Formação de Recursos Humanos em C&T Infraestrutura de Pesquisa Infraestrutura de Apoio à Inovação Redes de Pesquisa
	Extensão e Serviços Tecnológicos (Atividade Científica e Técnica Correlata - ACTC)	
	Instrumentos de incentivos Fiscais e apoio ao P&D Empresarial	Incentivos Fiscais: – Lei do Bem – Lei de Informática – Rota 2030 – PADIS Subvenção Econômica Fundo Perdido
	Fundos CTI	FNDCT
	Ações de Empreendedorismo	
Acompanhamento de atividades do MCTI	Acompanhamento das Unidades de Pesquisa	
	Acompanhamento das Entidades Vinculadas	

As áreas temáticas apresentadas acima se encaixam em um espaço de informação mais amplo, mostrado na Figura 11, que explicita um conjunto de macro categorias aplicável na classificação de conjuntos de dados permitindo a construção de uma gestão unificada da informação que é produzida ou circula no MCTI.

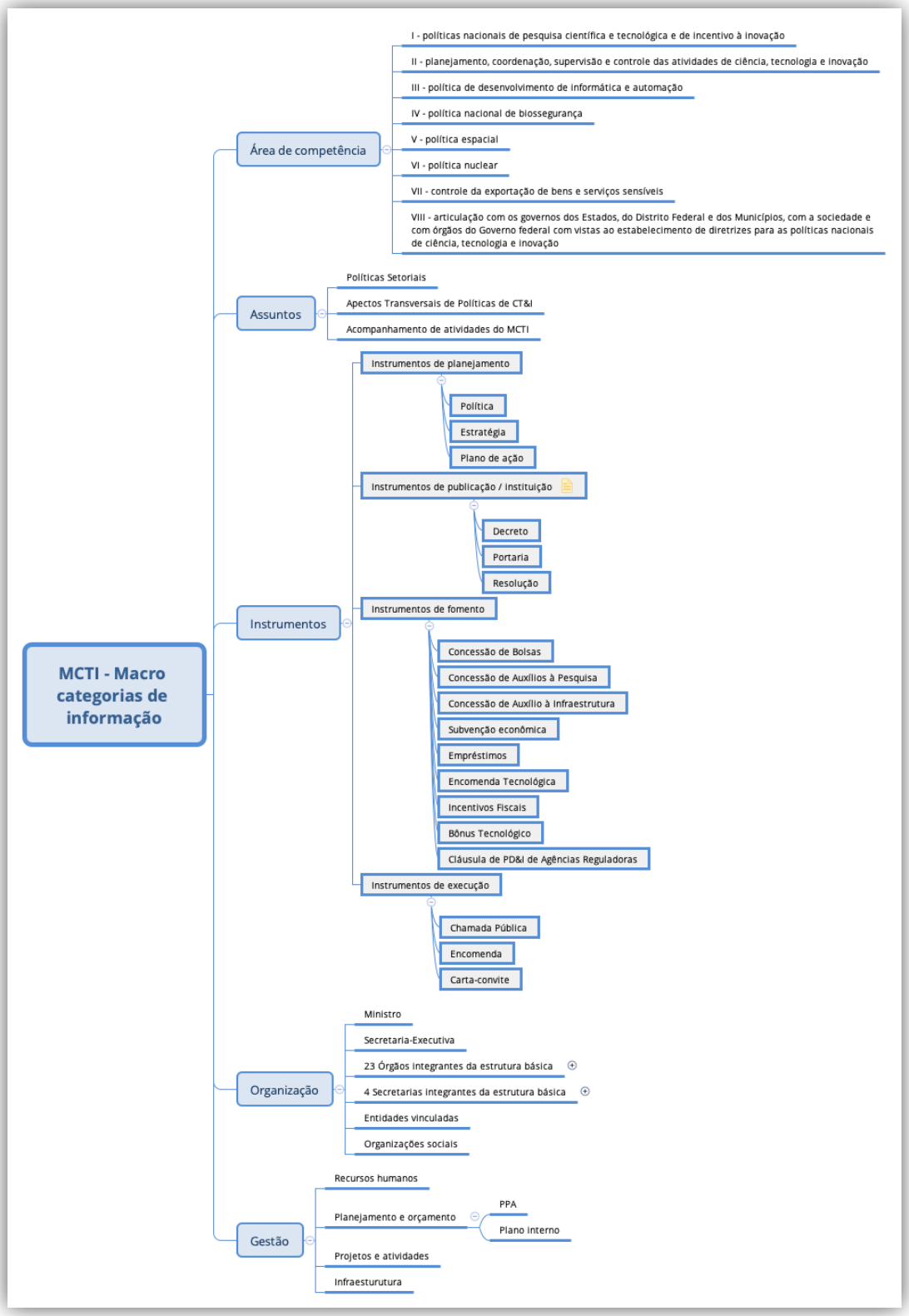


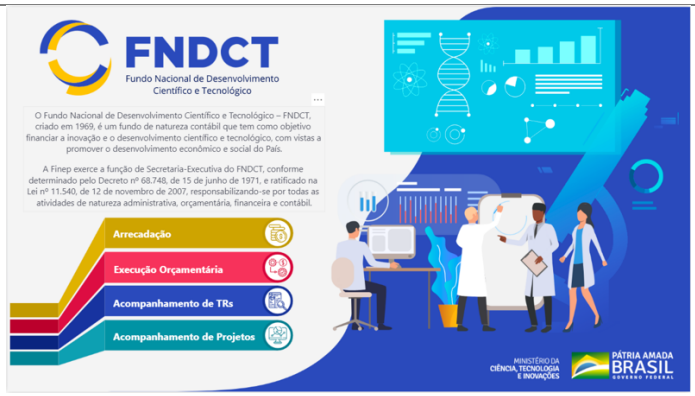
Figura 11 - Macro categorias de arquitetura da informação no MCTI. Fone: elaboração própria.

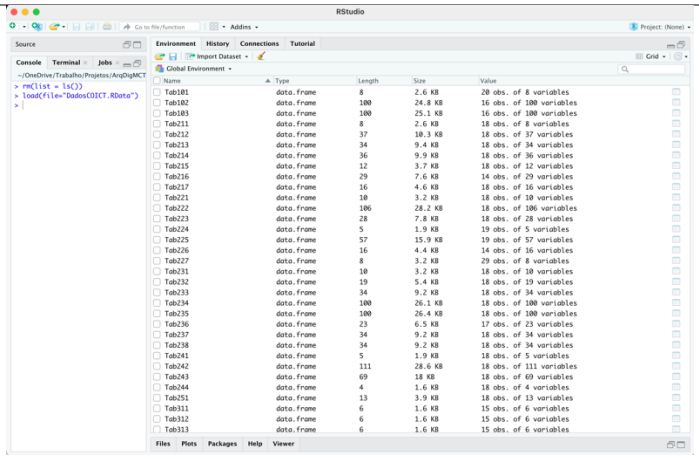
6.2. Painéis temáticos

Outro resultado alcançado no desenvolvimento experimental foram os painéis de informação, propriamente ditos, para os temas estratégicos selecionados. Os produtos de informação são mostrados na Tabela 11 e estão disponibilizados no endereço da internet abaixo indicado:

https://paineis.mcti.gov.br

Tabela 11 – Produtos de informação resultantes no período 2020/2021. Fonte: elaboração própria.

Tema estratégico	Tela de entrada	Tipo
Lei do Bem		Painéis de informação estratégica
FNDCT		Painéis de informação estratégica

Indicadores da COICT		Base de variáveis

6.3. Indicadores da COICT

O caso específico do tema estratégico Indicadores da COICT – Coordenação de Indicadores de Ciência, Tecnologia e Inovação do Departamento de Governança Institucional, pelas características específicas da demanda, teve condução, resultados e forma de documentação diferenciada.

A demanda, nesse caso, consistiu em captar e internalizar o conjunto de indicadores e variáveis já existentes e disponíveis na página Web do Ministério "Indicadores Nacionais de Ciência, Tecnologia Inovação"[3]. A expectativa do demandante era “apoio do CGEE para incorporar a base de dados da COIND[4] e demais bases à Infraestrutura da DTI” (MCTI, 2020).

Essa expectativa resultou nas seguintes especificações e critérios de aceitação:

- Conversão de formato, de arquivos no formato CSV, para uma base de dados de variáveis facilmente manipulável pela ferramenta estatística R / RStudio.
- Integração das variáveis em um único repositório, conservando ao máximo os metadados existentes na fonte de dados origem, de modo a documentar as variáveis.

[3] https://antigo.mctic.gov.br/mctic/opencms/indicadores/indicadores_cti.html .

[4] Antiga denominação da COICT/DGI.

Esses objetivos orientaram uma condução de trabalhos distinta dos demais temas estratégicos, uma vez que as variáveis já constituem a informação desejada, não havendo uma necessidade de informação a ser atendida. Assim, a condução se voltou para o trabalho de captura de dados a partir de uma página web (*web scraping*[5]), a ser realizado uma única vez, tratamento (conversão de forma, extração de metadados e outras características, limpeza de caracteres especiais) e ingestão na Arquitetura Digital.

O resultado foi a construção de um repositório com 5.083 variáveis, dispostos nos formatos JSON e RData e documentados no formato JSON[6] com os seguintes atributos: nome da tabela original, título, subtítulo, descrição, fontes de dados, notas de rodapé, variáveis (metadados: nome para a variável gerado automaticamente, título da coluna da tabela original referente à variável, referências de rodapé específicas da variável).

6.4. Situação atual dos temas estratégicos

No decorrer do segundo semestre de 2021 foi redefinido o conjunto de temas estratégicos em decorrência de fatos relevantes fora da governabilidade do projeto. Em deliberação conjunta com o Departamento de Governança Institucional (DGI), adotou-se como prioridade uma revisão do Painel da Lei do Bem (segunda versão para o tema estratégico em questão), e a elaboração de painéis para a Lei de Informática - Lei 8.248/91 e para o Programa de Apoio ao Desenvolvimento Tecnológico da Indústria de Semicondutores – Padis - instituído pela Lei no 11.484, de 2007.

Esses painéis se encontram em processo de desenvolvimento. Na referência do Ciclo de Inteligência em CTI, já foi executada a fase 1 e no momento as atividades em execução se relacionam com o provimento dos dados brutos (fase 2).

[5] Web scraping: raspagem web, é uma forma de mineração que permite a extração de dados de sites da web convertendo-os em informação estruturada para posterior análise. (Wikipedia).

[6] JSON (JavaScript Object Notation - Notação de Objetos JavaScript) é uma formatação leve de troca de dados. Para seres humanos, é fácil de ler e escrever. Para máquinas, é fácil de interpretar e gerar. (http://json.org/json-pt.html)

7. Conclusões e próximos passos

Até o final de 2021 foram alcançados marcos importantes do projeto, relacionados abaixo:

- Elaboração e revisão de modelo arquitetural para sistemas analíticos com uso de *data lake*, frente à experiência prática nas dependências do MCTI.
- Plataforma digital implementada, pelo DTI/SEXEC/MCT com apoio do CGEE, contemplando a infraestrutura de hardware e as ferramentas de software que implementam as funções de ingestão de dados, fluxos de tratamento de dados, visualização de informação, catálogo de dados e controle de acesso.
- Arquitetura de dados, padrões e processo de trabalho integrado com a plataforma digital implementada.
- Implantação de três (dos oito) temas estratégicos previstos, a saber: Lei do Bem, FNDCT e Indicadores da COICT.
- Estruturação inicial de arquitetura da informação que contextualiza os temas estratégicos implementados e apoia a curadoria de dados para a Arquitetura Digital do MCTI.
- Implementação parcial da segunda versão dos painéis da Lei do Bem, e projeto dos painéis da Lei de Informática e Programa Padis.

Em paralelo foi concluída nova versão do modelo de processos alinhado com a Arquitetura Digital, debatida com o Departamento de Tecnologia da Informação, incluindo o desenho da arquitetura de informação que o *data lake* implementa considerando os quatro temas estratégicos trabalhados. Outro resultado alcançado foi a definição do CKAN como ferramenta de catálogo de dados e sua implantação na infraestrutura de TIC do Ministério com apoio do CGEE. Essa decisão é relevante tendo em vista a importância que um catálogo de dados tem para uma arquitetura digital para *data lake*.

Os resultados alcançam objetivos maiores do projeto na medida que experimenta o suporte e o armazenamento a fontes de informações heterogêneas. Além disso, proporciona arcabouço instrumental organizado e controlado para aplicação de

metodologias de tratamento e análise de dados de brutos, de variados formatos. Os exercícios também permitiram a experimentação da interoperabilidade da Arquitetura Digital com sistemas de informação legados do MCTIC e fontes de informação externas ao ministério.

Os temas estratégicos definidos para o trabalho enriqueceram a Arquitetura Digital com:

- Definição de padrão de documentação para linhagem de dados;
- Ajustes no alinhamento de conteúdo nas camadas arquiteturais;
- Estabelecimento dos primeiros conjuntos de dados, nas camadas de Ingestão e Transformação;
- Validação da integração entre as ferramentas de uso cotidiano do MCTI com a Arquitetura Digital;
- Execução das tarefas previstas no modelo de trabalho, com esclarecimentos sobre objetivos das tarefas, sua natureza e atores envolvidos e responsáveis;
- Homogeneização do conhecimento sobre as dimensões do ciclo de vida de dados e a relação com a Arquitetura Digital, assim como os conceitos, componentes da arquitetura, modelo de trabalho e responsabilidades.

A continuidade do projeto em 2022 requer a revisão de cronograma de trabalho e nova rodada de definições de temas estratégicos que seja compatível com o tempo disponível para o projeto e a disponibilidade de engajamento das unidades organizacionais envolvidas com os temas selecionados. Esses últimos desenvolvimentos experimentais produzirão os insumos para a elaboração das versões finais dos produtos previstos para o projeto.

8. Referências Bibliográficas

BRASIL, 2019 BRASIL. DECRETO Nº 10.046, DE 9 DE OUTUBRO DE 2019. Diário Oficial da União. Brasília, DF. 10 de outubro de 2019.

BRASIL, 2020 BRASIL. Ministério da Ciência, Tecnologia, Inovações e Comunicações. Plano de dados abertos: 2020/2021. Departamento de Governança Institucional. Julho/2020.

CGEE, 2017 CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS. Desenho e detalhamento do primeiro nível do metaprocesso Inteligência Estratégica em CTI. Brasília, DF: Centro de Gestão e Estudos Estratégicos, 2017.

CGEE, 2018 CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS. Uma análise dos resultados da Lei do Bem: com base nos dados do FormP&D. Resumo Executivo. Brasília: Centro de Gestão e Estudos Estratégicos, 2018.

CGEE, 2019 CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS. Plano de trabalho para o desenvolvimento da arquitetura digital de inteligência de negócio do MCTIC. In: Arquitetura digital de inteligência de negócios do MCTIC. Brasília, DF: Centro de Gestão e Estudos Estratégicos, dezembro / 2019.

CGEE, 2020 CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS. Projeto Arquitetura Digital de Inteligência de Negócio do MCTIC - Execução do plano de trabalho. In: Arquitetura digital de inteligência de negócios do MCTIC. Brasília, DF: Centro de Gestão e Estudos Estratégicos, maio / 2020.

COGGD, 2020 COORDENAÇÃO DE GESTÃO E GOVERNANÇA DE DADOS. Guia Operacional: Regras e Padrões para a classificação de conjunto de dados no Datalake. Brasília, DF: COGGD/DGI-MCTI. Outubro 2020.

DAMA, 2017 DAMA International. DAMA-DMBOK Data Management Body of Knowledge. 2º Edition. Basking Ridge, NJ/USA: Technics Publications, 2017.

MCTI, 2020 MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES - MCTI. Sistema de informações do MCTI. Brasília, DF: Departamento de Governança Institucional (DGI). Apresentação PPT. Sistema_de_Informações_do_MCTI(2).pptx. Agosto de 2020.

MCTI, 2020a MCTI. Documento SEI/MCTI 5569274 – Minuta de portaria. MCTI. Arquivo "MINUTA_PORTARIA_DADOS_ABERTOS.pdf".

MPOG, 2018 MINISTÉRIO DO PLANEJAMENTO, DESENVOLVIMENTO E GESTÃO. Plano de Dados Abertos do Ministério do Planejamento. Disponível em https://plano.dados.planejamento.gov.br/#-glossário. Acessado em 10/12/2021. Versão 3.0.1 de julho/18

ROUSE, 2019 Rouse, Margaret. "Data Virtualization". TechTarget/WhatIs.com. Página disponível na Internet em https://searchdatamanagement.techtarget.com/definition/data-virtualization. Acessado em 10/11/2020. Última atualização em maio/2019.

STEENBEEK, 2019 STEENBEEK, Irina. Data Lineage 101. Data Crossroads. Disponível em https://datacrossroads.nl/category/series/data-lineage-101/. Último acesso em 17/10/2020. Março de 2019.

WRIGHT, 2018 Wright, T. Introducing Microsoft SQL Server 2019 Big Data Clusters. SQL Server Blog. Microsoft. Setembro 2018. Link acessível em https://cloudblogs.microsoft.com/sqlserver/2018/09/25/introducing-microsoft-sql-server-2019-big-data-clusters/. Acessado em 01/10/2020.

--- glossário ---
HOUAISS, A. Dicionário eletrônico Houaiss da língua portuguesa. Versão 1.0. [s. l.]: Objetiva, 2001.

SETZER, V. W. Dado, informação, conhecimento e competência. Datagrama, São Paulo v. 10, 2001. Disponível em: <http://www.ime.usp.br/~vwsetzer>. Acesso em:12 jun. 2004. Coleção Ensaios Transversais.

Anexo I – Padrão para nomenclatura de objetos em *data lake*

Padrão para nomenclatura de objetos em *datalake*

Referência: PS-MCTIC - Guia Operacional – Regras e Padrões de Modelagem de Dados - Versão 1.2 (DTI/SEXEC/MCTI)

NOME DE OBJETO

A proposta incorpora a definição já utilizada para objetos de banco de dados SGBD, estendendo-a para componente das camadas da arquitetura digital.

Camada	Componente da Camada	Tipo de Objeto	Nomenclatura recomendada (*)	Exemplo
Captura e Transformação	Arquivos e Pastas	PASTA (DIRETÓRIO)	Sigla da unidade organizacional + "_" + ano relativo ao assunto tratado	COIND_2020
		ARQUIVO	Nome curto relacionado o conteúdo do arquivo	indicadores2019.csv
	HDFS (**)	PASTA (DIRETÓRIO)	Sigla da unidade organizacional + "_" + ano relativo ao assunto tratado.	
		ARQUIVO	Nome curto relacionado o conteúdo do arquivo	
Transformação	SGBDR	TABELA	TB + Nome da Tabela	TBPessoa TBOrgaoConveniado
		PRIMARY KEY	PK+ Nome da Tabela (sem prefixo TB)	PKPessoa PKPerfil
		FOREIGN KEY	FK + Nome da Tabela Origem + Nome da Tabela (FK) (sem prefixos TB)	FKUFOrgao FKUFRegiao
		CHECK CONSTRAINT	CC + Nome da tabela sem TB + Nome da Coluna	CCFuncionarioSexo
	ElasticSearch (***)	ÍNDICE	ID-TEMA + IN + Nome do índice	
		RELATÓRIO	ID-TEMA + RL + Nome do relatório	
		MAPA	ID-TEMA + MP + Nome do mapa	
		PAINEL	ID-TEMA + PN + nome do painel	
		CANVAS	ID-TEMA + CN + nome do canvas	
	R Studio	ARQUIVOS	Nome curto relacionado o conteúdo do arquivo	
Publicação	PowerBI			
	Kibana (**)	RELATÓRIO	ID-TEMA + RL + Nome do relatório	

		MAPA	ID-TEMA + MP + Nome do mapa	
		PAINEL	ID-TEMA + PN + nome do painel	
		CANVAS	ID-TEMA + CN + nome do canvas	
	Shiny Dashboards			
Consumo	API			

(*) O nome do objeto tem limitação de 20 caracteres ASCII puro, isto é, sem uso de acentuação, cedilha, espaço em branco ou caracteres especiais (TAB, *new line* etc.).

(**) Recomenda-se adotar a seguinte estrutura mínima dentro da pasta principal:

Subdiretório	Objetivo	Observações
/data	Dado bruto	Somente leitura para usuários
/user/<nome usuário>	Diretórios "home" para usuários que trabalham com os dados	
/etl	Programas e scripts de tratamento dos dados	Pode ser organizado internamente por grupos / projeto que atuam sobre os dados
/tmp	Arquivos temporários	

(***) Como o ElasticSearch tem apenas um *namespace*, utiliza-se o identificador de SUBCLASSE (da proposta DGI contida em "Guia Operacional:
Regras e Padrões para a classificação de conjunto de dados no Datalake" para identificar objetos de um tema.

NOME DE ATRIBUTO DE OBJETO

A proposta incorpora a definição já utilizada para atributos, flexibilizando a definições de nomenclatura para possibilitar o uso do nome em minúsculo, por exemplo: “cdfuncionario”, “qtacesso”.

Tipo Atributo	Nomenclatura	Exemplo	Uso	Equivalência com tipo SQL
Código	CD+Descrição	CDFuncionario CDCPF	Identificadores numéricos ou alfanuméricos estruturados	Numeric, int, char, smallint, tinyint
Data/Hora	DT+Descrição	DTNascimento	Determinação de dia, mês, ano e horário	datetime
Número	NR+Descrição	NRImovel	Informação de grandeza numérica	numeric, int, smallint, tinyint
Nome	NO+Descrição	NOTrabalhador	Denominação própria. Nomes de pessoas, cidades	char, varchar (mínimo 70)
Valor	VL+Descrição	VLSalario	Valores monetários	numeric (nn,2)
Tipo	TP+Descrição	TPAssunto	Caracterização de atributos (domínios) Domínios até 5 valores usar Check Constraint	numeric, smallint, tinyint
Sigla	SG+Descrição	SGOrgao	Atributos codificadores do tipo caracter	char, varchar
Identificador (auto-incremento)	ID+Descrição	IDUsuario	Atributos identificadores, inclusive gerados automaticamente pelo BD ou sistema	numeric, int, smallint, bigint
Situação/status	ST+Descrição	STAndamento	Status (até 5 valores)	smallint, char(1), bit

Indicador	IC+Descrição	ICPago	Duas condições contrapostas. Booleanos (set ou reset – apenas 1 bit)	bit NOT NULL
Endereço	ED+Descrição	EDFiscal	Atributos alfanuméricos contendo informações de endereço do tipo rua, bairro, número, etc.	varchar, char
Matrícula	MT+Descrição	MTFuncionario	Atributos codificadores alfanuméricos	char, varchar , numeric
Memo	MM+Descrição	MMParecer	Texto sem limite de tamanho	text (> 8k) varchar (< 8k)
Descrição	DS+Descrição	DSClasse	Descrição com limite de tamanho	char, varchar (Mínimo 100)
Texto	TE+Descrição	TEParecer	Texto com limite de tamanho	varchar
Taxa	TX+Descrição	TXComissao	Taxas	numeric
Quantidade	QT+Descrição	QTAcessos	Número de unidades	Numeric, int,smallint, tinyint
Link	LK+Descrição	LKPagina	Atributos alfanuméricos contendo endereços de páginas WEB	varchar
Binário	BI+Descrição	BIFoto	Atributos BLOB (imagens, vídeo, som)	image
Endereço Eletrônico	EE+Descrição	EEUsuario	Atributos alfanuméricos contendo endereços de e-mail	varchar
Sequencial	SQ+Descrição	SQCarteira	Sequencial	numeric
Observação	OB+Descrição	OBCargo	Observações genéricas textuais sem formatação definida	varchar
Percentual	PC+Descrição	PCSalario	Percentual de valores	numeric
Total	TT+Descrição	TTSalarios	Totalização de valores	numeric
Média	MD+Descrição	MDSalarios	Média de valores	numeric

Anexo II – Proposta para metadados

Metadados para CONJUNTO DE DADOS	Descrição do metadado	Captura	Transformação	Publicação	Consumo
Nome	Nome do conjunto de dados, por exemplo, o nome do arquivo (ou conjunto de dados).	**X**	**X**	**X**	**X**
Título	Titulo do conjunto de dados. Expressão que indica o assunto contido no conjunto de dados.	**X**	**X**	**X**	**X**
Subtítulo	Detalhe complementar sobre o assunto contido no conjunto de dados				
Descrição	Descrição do conjunto de dados.	**X**	**X**	**X**	**X**
Gestor do dado	Pessoa ou organização responsável por aprovar e providenciar a disponibilização do dado	**X**		**X**	
Fonte de dados (Origem do dado/Autoria/Mantenedor)	Indicação da origem do conjunto de dados. Sistema transacional do MCTI ou sistema externo ao MCTI que originou o conjunto de dados.	**X**	**X**	**X**	**X**
Periodicidade de atualização	Indicação da frequência de disponibilização do conjunto de dados na sua origem. Exemplo: diária, semanal, mensal etc.			**X**	**X**

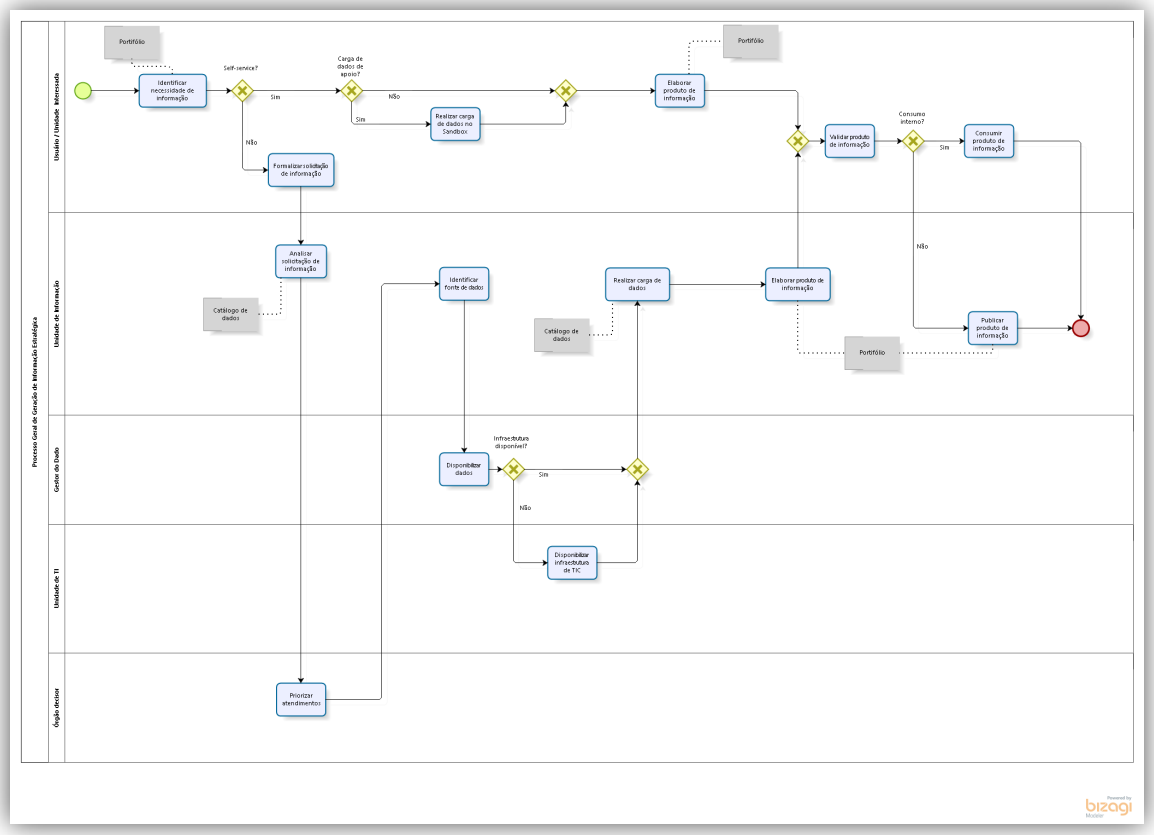
Ano base / cobertura temporal	Indicação do período temporal coberto no conjunto de dados.			**X**	**X**
URL do conjunto de dados	Endereço Web (na forma de URL) onde o conjunto de dados poderá ser acessado.	**X**		**X**	**X**
Descritivo metodológico	Método(s) ou processo(s) de tratamento aplicado ao dado contido no conjunto de dados.		**X**	**X**	
Palavra-chave (tag)	Palavra (simples ou composta) que identifica ou clarifica o conjunto de dados. Utilizado para indexação e facilitação de buscas.			**X**	
Licença	Informação sobre direitos autorais e disponibilidade de acesso e uso do conjunto de dados, expresso na forma do tipo de licença afeita ao conjunto de dados. Por exemplo: Creative Commons, MIT license, Open Software License, W3C License etc.			**X**	**X**
Formato de dado	Descrição do formato do dado em relação a suas características digitais. Por exemplo: texto (TXT, CSV), documento (DOC, PDF, HTML), planilha (ODS, XLSX, CSV), conjuntos de documentos (JSON, XML), dados georreferenciados, tabela de banco de dados relacional, arquivo produzido por aplicativo (R Studio,	**X**			

	PowerBI) etc. Detalhes podem ser acrescentados para descrever seu conteúdo.				
Local de armazenamento	Indicação do local onde o conjunto de dados está armazenado. Por exemplo: Schema de SGBDR, Repositório em nuvem, pasta no HDFS, pasta compartilhada de rede etc.	**X**	**X**	**X**	
Ponto de conexão (datasource)	Endereço digital de conexão com o conjunto de dados. Normalmente utilizado para bases de dados (relacionais ou documentais), repositório em nuvem e serviços de informação disponíveis na Web.		**X**		
Controle de acesso	Características de segurança no armazenamento e acesso ao conjunto de dados.	**X**	**X**	**X**	
Data última atualização	Data e hora em que a fonte de dados foi criada, gerada ou atualizada.	**X**		**X**	**X**
Data de disponibilização na fonte de dados	Data, e se possível, hora, da disponibilização do dado na Fonte de dados.	**X**			

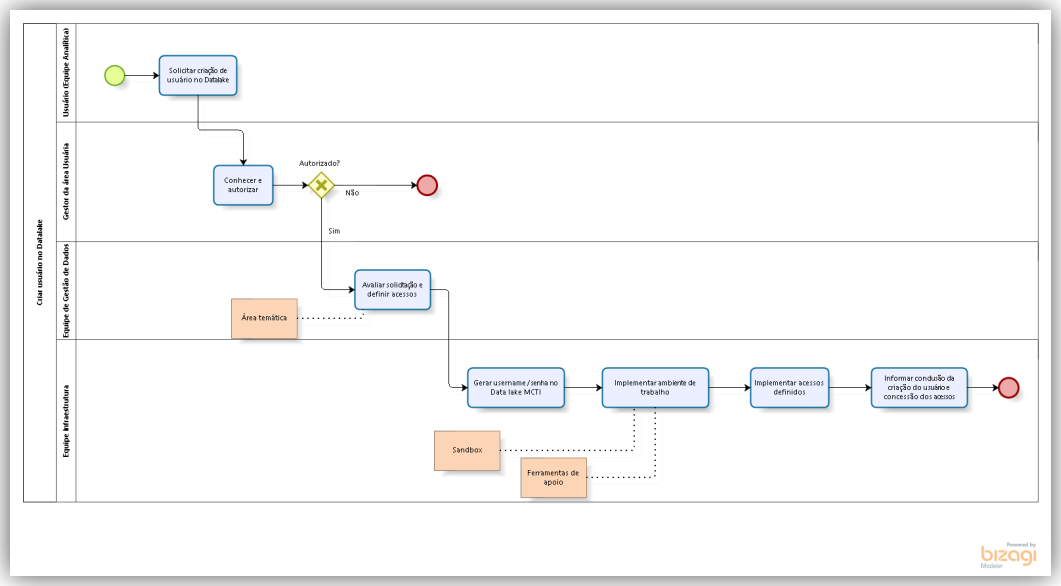
Metadados para ATRIBUTOS	Descrição do metadado	Captura	Transformação	Publicação	Consumo
Nome (Campo / atributo / variável)	Nome do atributo. Sempre que possível adotar o padrão de nomenclatura para campos de dados.	**X**	**X**		**X**
Descrição	Descrição do atributo dentro do respectivo conjunto de dados.		**X**		**X**
Formato	Formato do dado contido no atributo. Exemplo: caracteres, texto, inteiro, decimal, valor monetário, data, hora etc.	**X**	**X**		**X**
Tamanho	Tamanho do atributo em caracteres. Normalmente aplicado a campos caracteres ou texto.	**X**	**X**		**X**
Controle de acesso	Características de segurança no armazenamento e acesso ao atributo no respectivo conjunto de dados.	**X**	**X**		**X**

Anexo III – Processo de trabalho

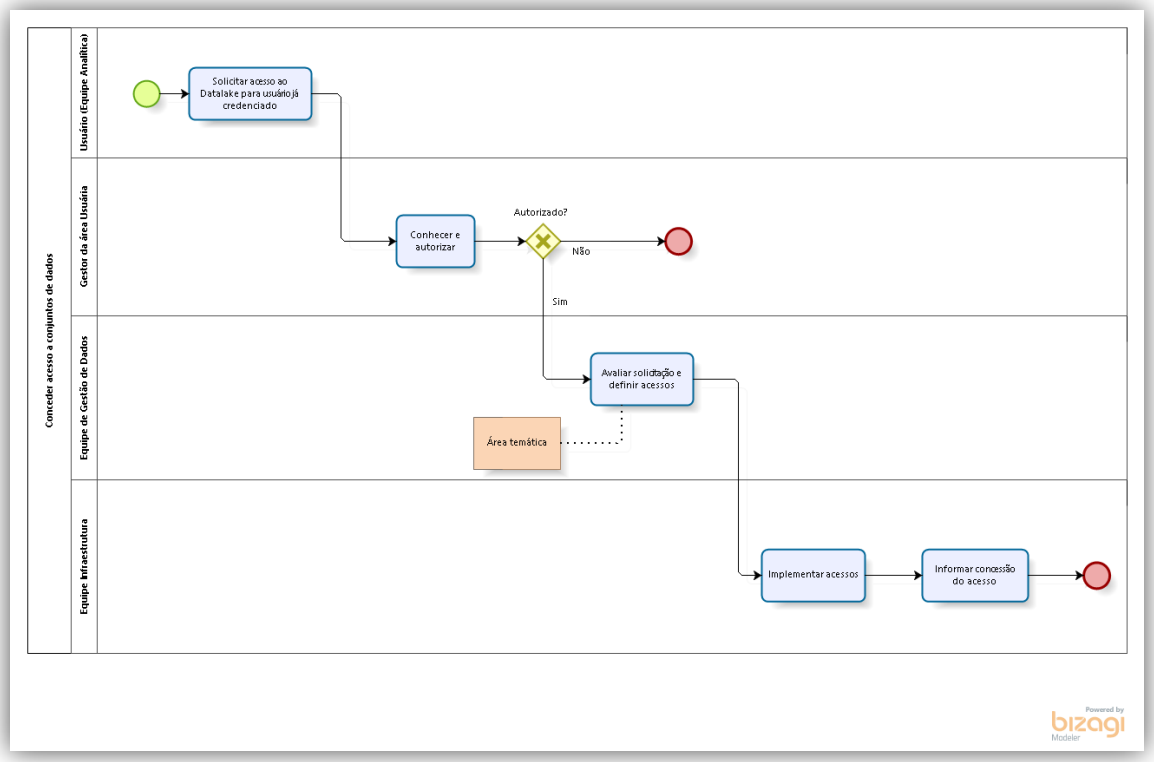
Processo geral – inclui o solicitante e instâncias de decisão e priorização inter-departamental.



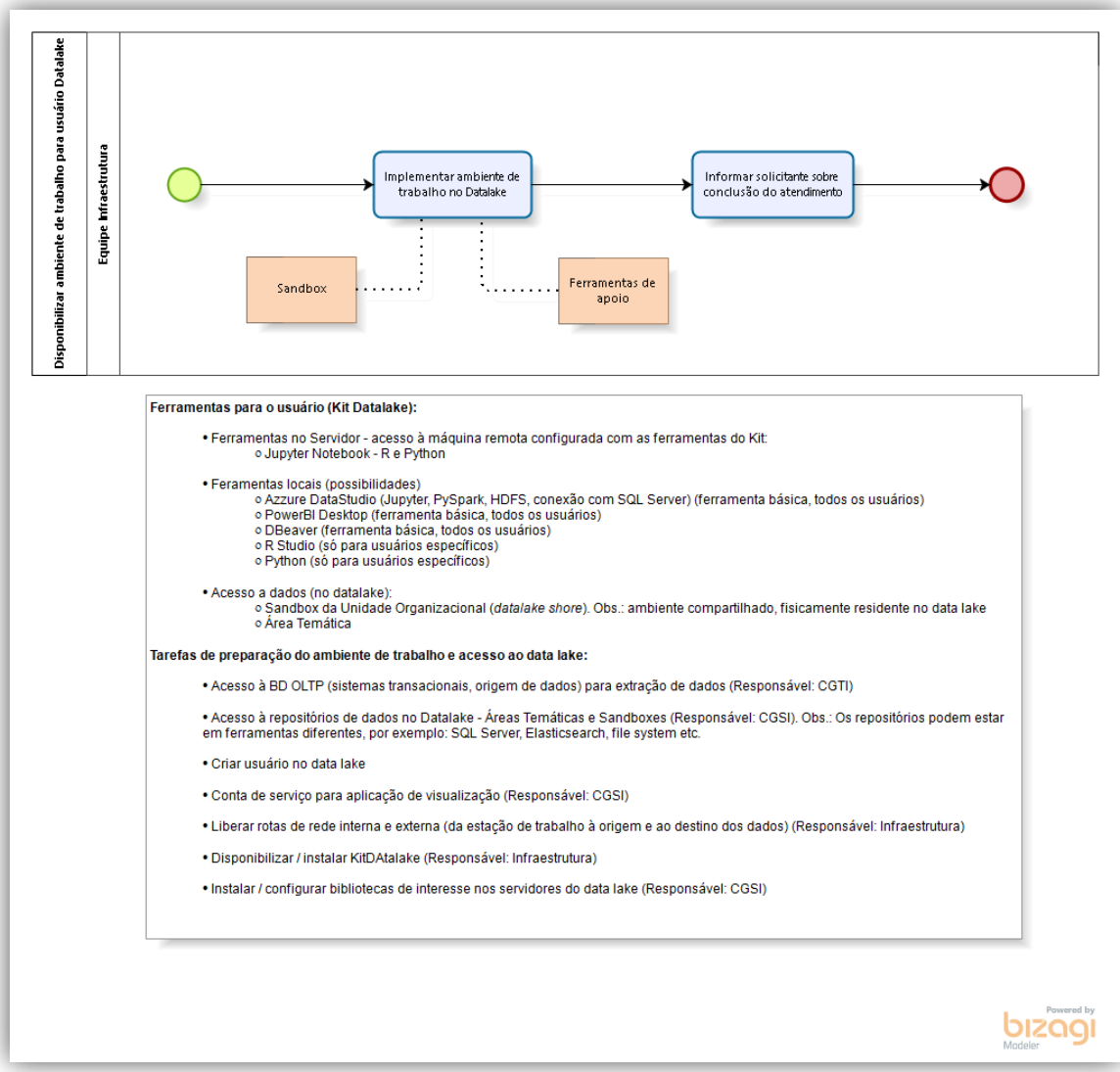
Inserção de um novo usuário no *data lake*



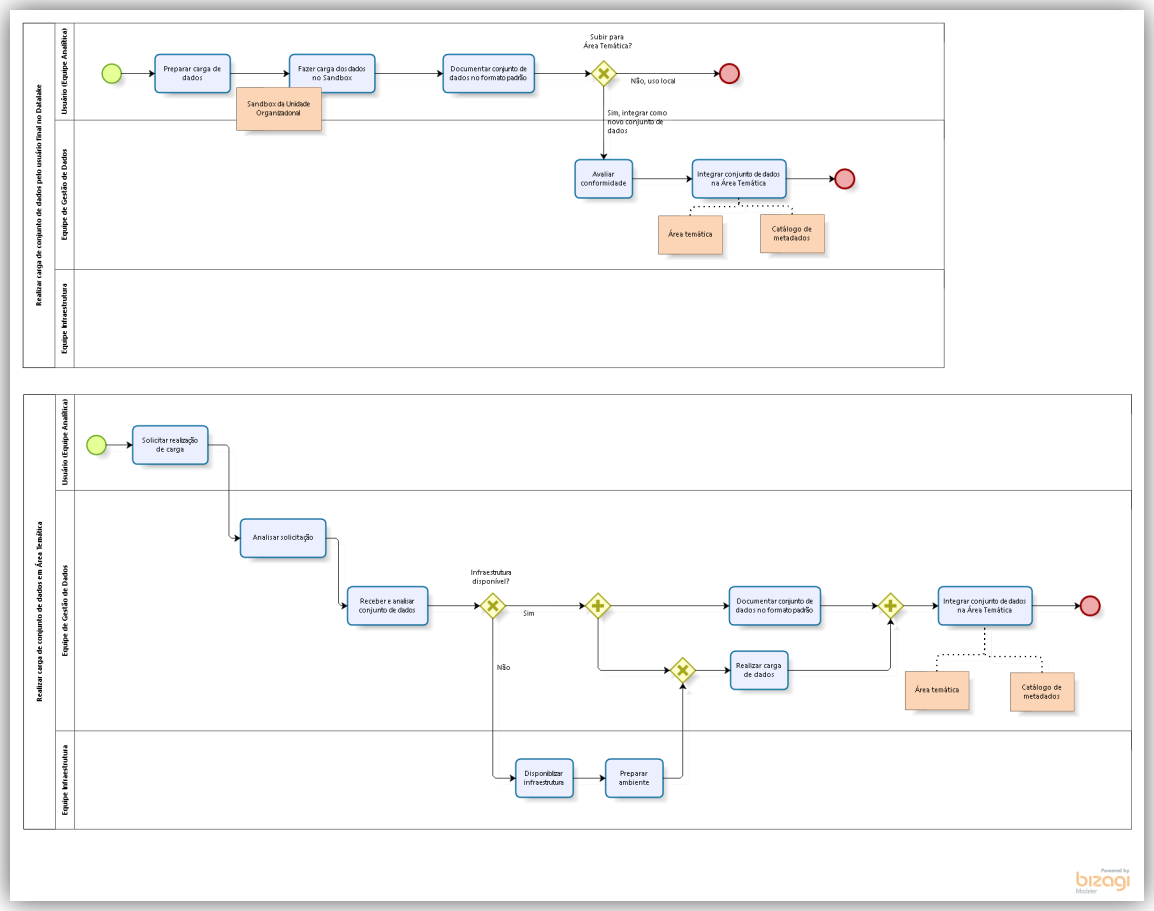
Concessão de acesso para usuários



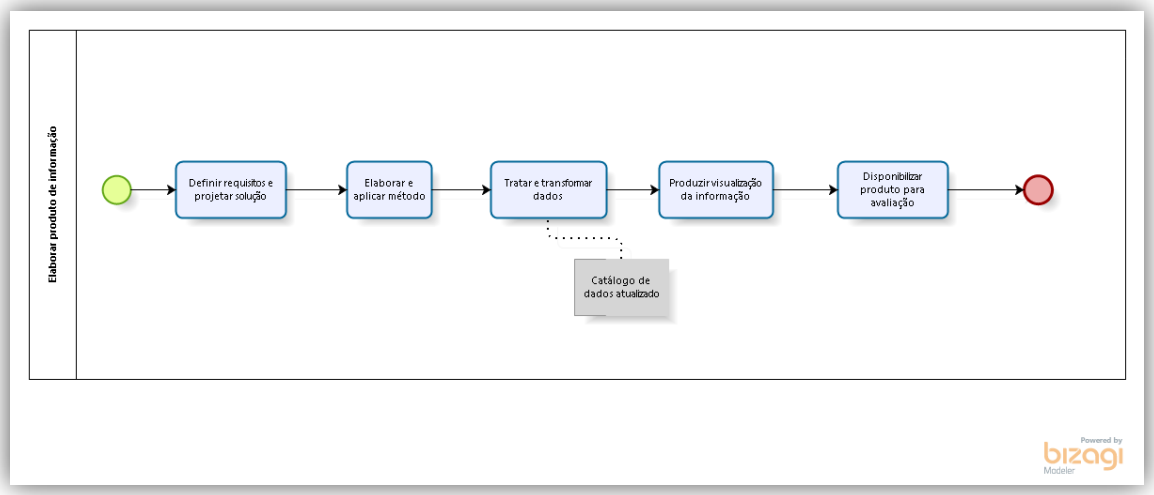
Disponibilização de instrumentos de trabalho para usuário do *data lake*



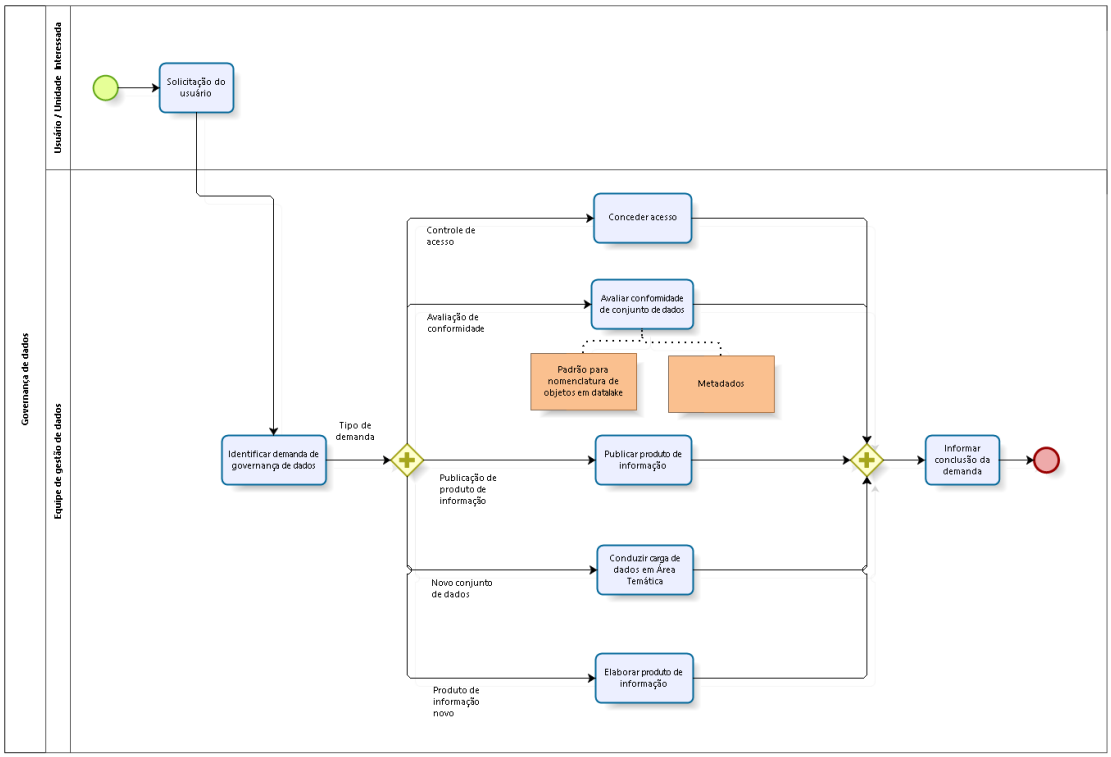
Processos de realização de ingestão de dados no *data lake*



Elaboração de produtos de informação – exemplo com metodologia utilizada no projeto.



Governança do modelo:


Governança de dados
Usuário / Unidade Interessada
Equipe de gestão de dados
Solicitação do usuário
Identificar demanda de governança de dados
Tipo de demanda
Controle de acesso
Conceder acesso
Avaliação de conformidade
Avaliar conformidade de conjunto de dados
Padrão para nomenclatura de objetos em datalake
Metadados
Publicação de produto de informação
Publicar produto de informação
Novo conjunto de dados
Conduzir carga de dados em Área Temática
Produto de informação novo
Elaborar produto de informação
Informar conclusão da demanda

Anexo IV – Glossário

Termo	Uso no projeto	Referências complementares
Arquitetura digital	*Modelo lógico (desenho) organização dos atores e elementos que compõem a solução tecnológica proposta, bem como os relacionamentos entre esses elementos e atores. Inclui os elementos tecnológicos, atores e processo de trabalho*	(DAMA, 2017)
Big data	*Conceito amplo que aborda as alterações tecnológicas ocorridas na última década que objetivam habilitar e viabilizar a geração, armazenamento e análise de crescentes volumes de dados. Os critérios mais utilizados para caracterizar a aplicação de big data é endereçar grandes volumes de dados, ampla variedade de dados, velocidade de produção de resultados. Outros critérios utilizados lidam com a veracidade da informação e com o seu valor (impactos resultantes do uso das informações).*	Elaboração própria a partir de (DAMA, 2017).
Business intelligence	*Categoria de análise de dados que tem como alvo a compreensão das atividades e oportunidades organizacionais por meio de informação sobre produtos, serviços e clientes subsidiando a tomada de decisões para atingir objetivos estratégicos. As ferramentas utilizadas para a realização dessas análises, consideradas ferramentas de* business intelligence *(BI), implementam suporte à consulta, mineração e visualização de dados. Dentre os instrumentos mais utilizados para BI está os* data warehouses*, repositórios de grandes quantidades de dados estruturados com foco na análise de negócio, criação de relatórios e painéis de informação. No histórico recente o instrumental de BI foi complementado com a introdução de* data lakes*, que permitem a realização de análises aprofundadas de dados (análise preditiva, analítica sobre dados não estruturados e geração de* insight*) a partir de aprendizado de máquina sobre dados estruturados ou não estruturados.*	Elaboração própria a partir de (DAMA, 2017).

Conjunto de dados	*Um arquivo digital de dados de qualquer formato (estruturado ou não estruturado), armazenado em um meio digital e acessível por intermédio de uma ferramenta de software.*	
Dado	*Sequência de símbolos ou valores, representados em algum meio, produzidos como resultado de um processo natural ou artificial. Entende-se que dados são observações ou o resultado de uma medida (por investigação, cálculo ou pesquisa) de aspectos característicos da natureza, estado ou condição de algo de interesse, que são descritos através de representações formais e, ao serem apresentados de forma direta ou indireta à consciência, servem de base ou pressuposto no processo cognitivo.*	(HOUAISS, 2001; SETZER, 2001)
Data lake	*Um ambiente digital onde um vasto volume de dados de diferentes tipos e estruturas pode ser ingerido (internalizado), armazenado, acessado e analisado. Faz parte de um* Data lake *um conjunto variado e integrado de ferramentas de hardware e software com foco em tratamento de dados. Uma atividade crítica associada a* data lakes *é a administração de metadados para garantir a organização, correção e consistência dos dados ingeridos e disponibilizados aos usuários.*	Elaboração própria a partir de (DAMA, 2017).
Ferramenta	*Artefato de software adquirido, internalizado (software livre) ou desenvolvido internamente, que implementa uma funcionalidade.*	
Fonte de dados	*Um provedor de dados externo ao sistema analítico representado pela Arquitetura Digital. Um provedor pode ser um sistema transacional interno ou externo ao MCTI, serviços ou repositórios digitais de dados pagos ou aberto.*	Definição do projeto.
Governança de dados	*Exercício de autoridade e controle que permite o gerenciamento de dados sob as perspectivas do compartilhamento, da arquitetura, da segurança, da qualidade, da operação e de outros aspectos tecnológicos*	(BRASIL, 2019)
Informação	*Dados, processados ou não, que podem ser utilizados para produção e transmissão de conhecimento, contidos em qualquer meio, suporte ou formato;*	(BRASIL, 2019)

Metadado	*Conjunto de informações descritivas sobre os dados, incluindo as características do seu levantamento, produção, qualidade e estrutura de armazenamento, essenciais para promover a sua documentação, integração e disponibilização, bem como possibilitar a sua busca e exploração.*	(MPOG, 2018)
Objeto do repositório digital	*Um item residente no repositório digital. Pode ser um arquivo, uma pasta digital, uma tabela de banco de dados, um esquema de banco de dados, índices, painéis, gráficos etc. Qualquer um destes itens sobre o qual há interesse em administrar e tornar disponível para uso.*	
Plataforma digital	*Implementação do arcabouço conceitual da arquitetura digital no que tange aos elementos de Tecnologia da Informação. Conjunto integrado e funcional de hardware e software que implementa fisicamente a Arquitetura Digital do MCTI e hospeda o data lake do Ministério.*	
Produto de dados	*Conjunto de dados tratados com uma finalidade preestabelecida, disponibilizado via qualquer meio digital, que pode ser incorporado ou usado por outra aplicação digital.*	
Produto de informação	*Conjunto de dados tratados, disponibilizado no seu estado final de valor agregado para consumo humano, e que atende uma ou mais necessidades de informação.*	
Repositório digital; ou repositório digital de referência	*Parte da Plataforma digital que implementa a função de armazenamento de dados brutos (fontes de dados originais) e trabalhados (produtos intermediários ou finais) que requeiram armazenamento. Está sujeito a controle de acesso.*	