

Jornadas de Dados

Data Lake Institucional



Brasília, março de 2022.

I. CONCEITO

O que é uma jornada de dados?

É um itinerário técnico a ser percorrido por uma área de negócio dentro de uma iniciativa de data & analytics. Esta jornada é definida a partir do objetivo principal do projeto.

As jornadas contempladas no âmbito do datalake institucional são: Jornada de ingestão, jornada de transformação, jornada de laboratório e jornada de dataviz.

Caráter orientativo

- Objetivo de negócio: Ingestão de dados – Jornada de Ingestão
- Objetivo de negócio: Imprimir novas regras de negócio a um ativo pré-existente no datalake – Jornada de transformação
- Objetivo de negócio: Explorar e preparar conjuntos de dados a título de experimentação – Jornada de laboratório
- Objetivo de negócio: Desenvolvimento de painéis e análises visuais – Jornada de Dataviz

II. VISÃO GERAL DAS JORNADAS DE DADOS NO DATA LAKE

IV. Ingestão

Engenharia de dados

- Quando o dado não está no data lake (raw)
- Necessidade de iniciativa com investimento da área de negócio e participação do analista de relacionamento
- Executado, geralmente, por meio de Fornecedores
- Necessária a inclusão/atualização de ativos no catálogo de dados
- Origem dos dados via sistemas (OLTP) ou Crawlers (bots)
- Necessário o uso do cofre de senhas para acesso ao DL e para desenvolvimento de Crawlers
- Necessária a utilização dos padrões de desenvolvimento de projetos de dados

*etapa de escrita no DL

III. Transformação

Engenharia de dados

- Quando o dado já está no data lake (raw)
- Necessidade de iniciativa com investimento da área de negócio e participação do analista de relacionamento
- Executado, geralmente, por meio de Fornecedores
- Necessária a inclusão/atualização de ativos no catálogo de dados
- Necessária a definição de regras de negócio
- Necessária a utilização dos padrões de desenvolvimento de projetos de dados

*etapa de escrita no DL

II. Laboratório

Ciência de dados

- Quando o dado já está no data lake
- Necessário solicitar ao data steward o acesso ao ativo de interesse
- Realizado por meio da ferramenta institucional Databricks
- Recomendado o uso do Pyspark
- A área de negócio é protagonista
- Possibilidade de carga de dados pelo próprio usuário
- Possibilidade de realizar junção de dados, consultas, criar relatórios e machine learning
- Possibilidade de testar e converter (por meio de iniciativa) em pipeline institucional.

*etapa de consulta ao DL

I. Visualização de Dados

Business Intelligence

- Quando o dado já está no data lake (Biz)
- Possibilidade de uso do databricks (CSV, PDF, HTML)
- Tableau (modelo self service BI)
- Pode ser realizado por fornecedor ou analista da área de negócio
- Boas práticas não recomendam lançar mão de dados granulares para estruturação de projetos de dataviz
- Os projetos de dados devem estar alinhados com a Política de governança de self service BI.

*etapa de consulta ao DL

Obs: As etapas de escrita geram informações que devem ser contempladas no catálogo de dados

III – COMPORTAMENTO DOS DADOS NO DATALAKE INSTITUCIONAL

DADOS BRUTOS



ORDENADOS



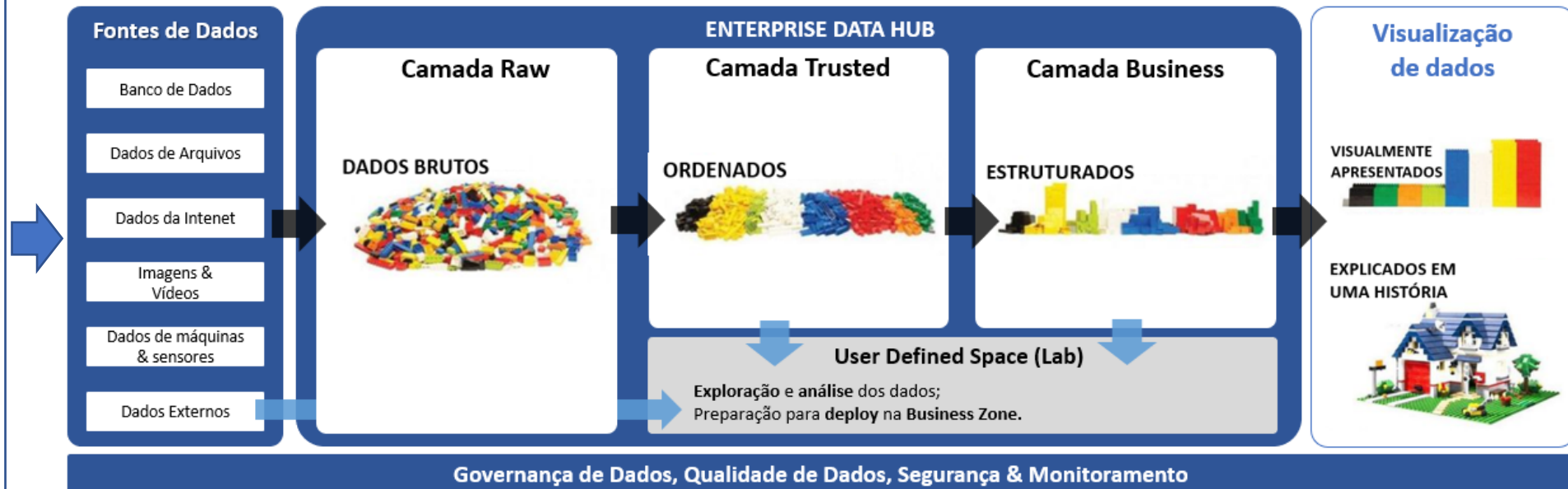
ESTRUTURADOS



VISUALMENTE APRESENTADOS



EXPLICADOS EM UMA HISTÓRIA



COORDENAÇÃO DE DATA & ANALYTICS - STI

