Ciência de Dados com aplicação em Inteligência de Negócio e em Dados Textuais

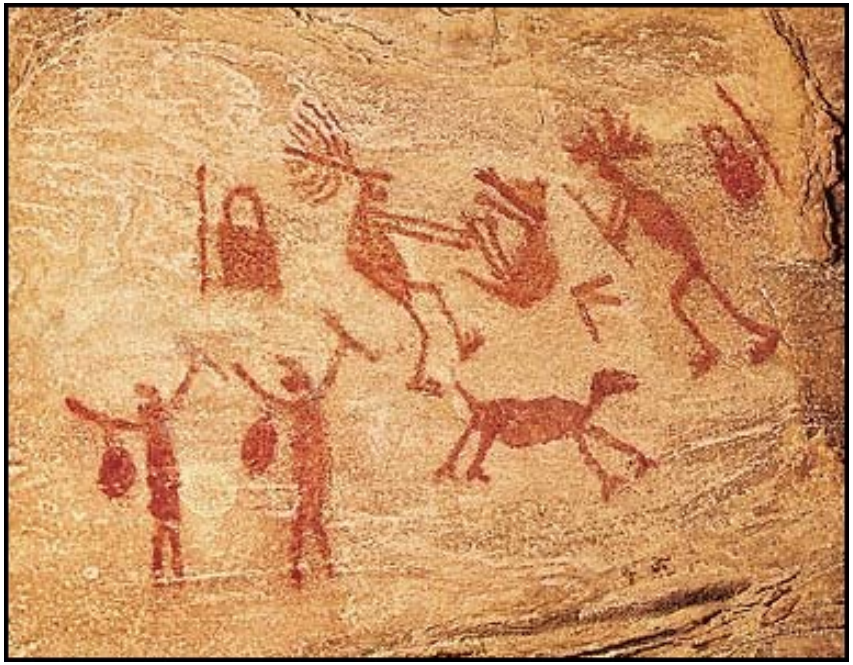
Alex Souza

VI Escola de Verão (MACC)

12 de março de 2019

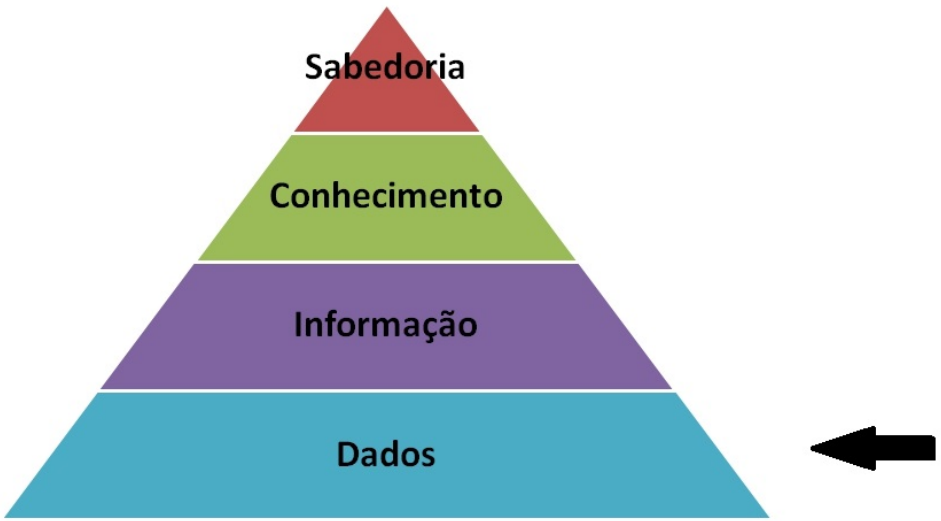
DADOS

ORIGEM



O QUE SÃO DADOS?

Dados são um conjunto de valores ou ocorrências em um estado bruto com o qual são obtidas informações com o objetivo de adquirir benefícios. Existem dois tipos de dados: **estruturados** e **não estruturados**. [1]



[1] https://pt.wikipedia.org/wiki/Dados

O QUE SÃO DADOS?

Os dados **estruturados**, que são dados formatados, organizados em tabelas - linhas e colunas - e são facilmente processados, geralmente é utilizado um sistema gerenciador de banco de dados (SGBD) para armazenar esse tipo de dado, um exemplo são os dados gerados por aplicações empresariais.

bases de dados

O QUE SÃO DADOS?

Os dados **não estruturados** não possuem uma formatação específica e são mais difíceis de serem processados. Por exemplo, mensagens de email, imagens, documentos de texto, mensagens em redes sociais.

FONTE DE DADOS

- Arquivos
  - Físicos
  - Lógicos
- Banco de Dados
- Coleções de Documentos
- Redes sociais
- Sistemas CRM (*Customer Relationship Management*)
- Sistemas ERP (*Enterprise Resource Planning*)

ARQUIVOS FÍSICOS

ARQUIVOS LÓGICOS

- Planilhas Eletrônicas (.xls, .xlsx, .ods)
- Arquivos .CSV (*Comma-separated values*)
- Documentos texto (.txt)
- Arquivos .XML (*Extensible Markup Language*)

BANCO DE DADOS

Um **banco de dados** (em inglês, *database*) é um local onde é possível armazenar dados de maneira estruturada e com a menor redundância possível. Estes dados devem poder ser utilizados por programas e usuários diferentes. [2]

[2] https://br.ccm.net/contents/65-bancos-de-dados

COLEÇÕES DE DOCUMENTOS

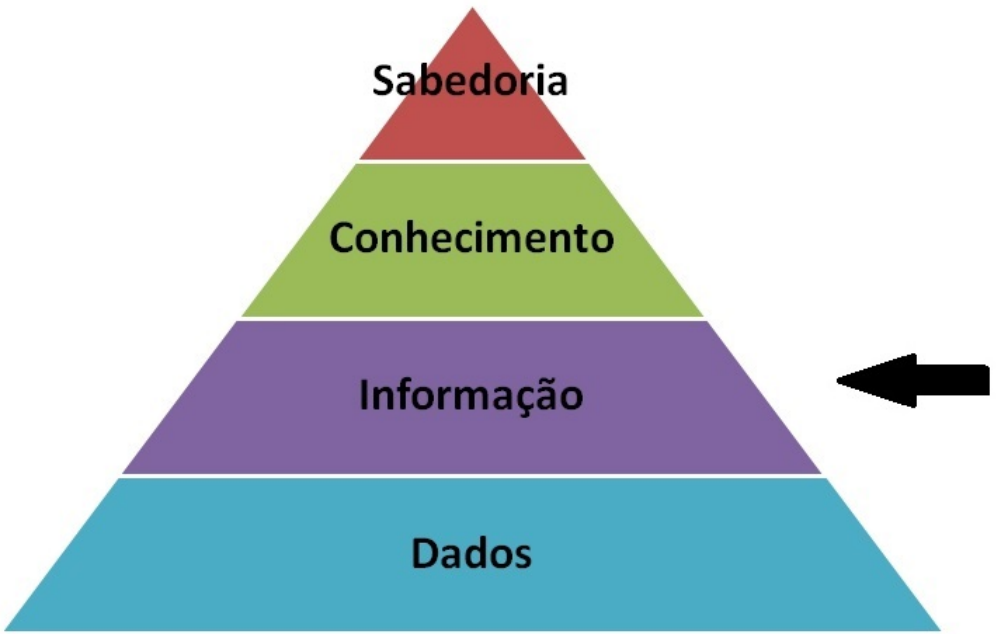
Como o próprio nome diz, é uma coleção de documentos que atende à uma determinada área de atuação (muito utilizada em IA para validação e testes de modelos). Exemplo:

- Recuperação de Informação [3]
  - *20NewsGroups*
  - *Reuters-21578*
  - *NPL*

[3] https://pessoalex.wordpress.com/dados/fontes-de-dados/

O QUE É INFORMAÇÃO?

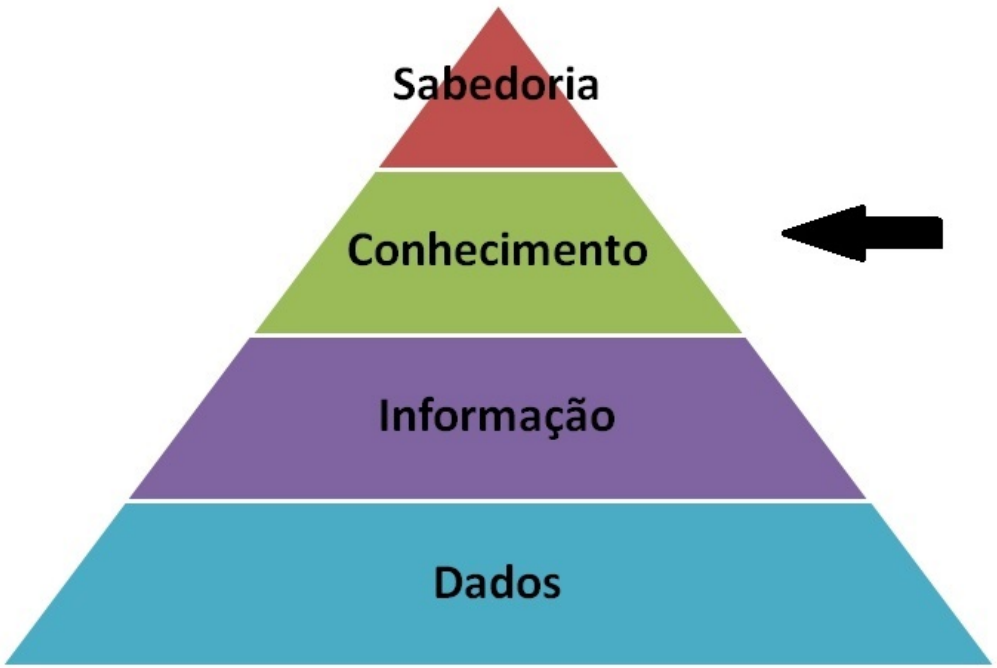
Dados dotados de relevância e propósito. A informação tem um significado e é organizada por algum propósito. [4]



[4] https://pessoalex.wordpress.com/dados/

O QUE É CONHECIMENTO?

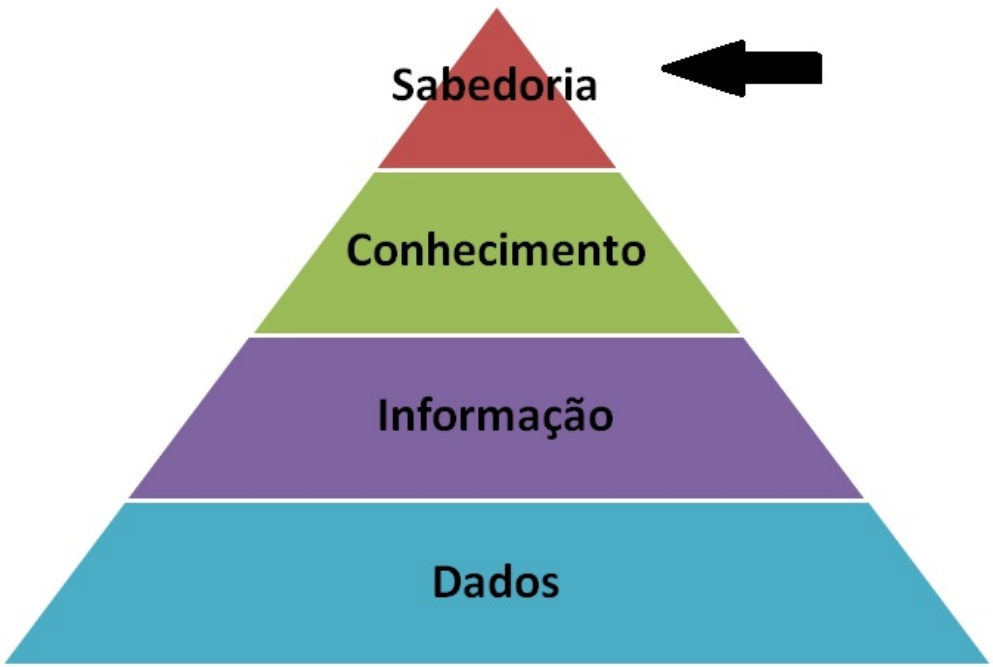
"Refere-se à habilidade de criar um modelo mental que descreva o objeto e indique as ações a implementar, as decisões a tomar."[5] (ex. Business Intelligence)



[5] https://pessoalex.wordpress.com/dados/

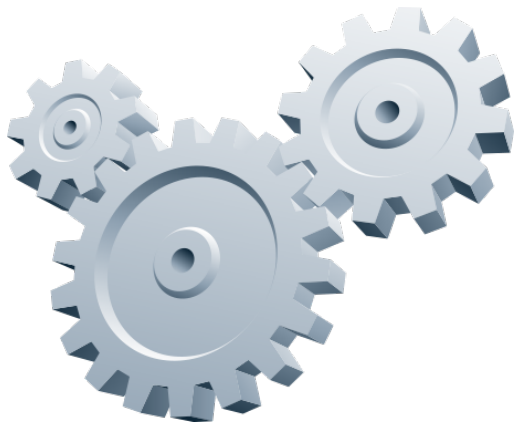
O QUE É SABEDORIA?

Nada mais do que a manipulação do conhecimento. (Nível Gerêncial)



PRATICAR

- Fontes de Dados
  - Arquivos (xls, xml, csv...) [6]
  - Banco de Dados
    - Tabelas



[6] https://github.com/aasouzaconsult/Trabalhando-com-Dados

GESTÃO DE DADOS

A Gestão de Dados visa controlar e alavancar eficazmente o uso dos ativos de dados e sua missão e objetivos são atender e exceder às necessidades de informação de todos os envolvidos (stakeholders) da empresa em termos de disponibilidade, segurança e qualidade. [7]

[7] MOSLEY, M.; BRACKETT, M.; EARLEY, S.; HENDERSON, D. **The DAMA Guide to The Data Management Body of Knowledge: DAMA** - DMBOK Guide. 1. ed. Estados Unidos: Technics Publications, 2009.

DMBOK

DMBOK (Data Management Body of Knowledge) - É um guia de melhores práticas para Governança de Dados[8].

[8] DMBOK - Visão Sintética

Governança de Dados

- Gerência da Arquitetura de dados
- Desenvolvimento de dados
- Gestão de operações de bancos de dados
- Gestão de Segurança de dados
  - GDPR
- Gestão de Dados mestres e de Referência
- Gestão de Data Warehousing e BI
- Gestão de Documentos e conteúdo
- Gestão de Metadados
- Gestão de Qualidade de dados
  - Importância nas empresas
  - Armadilhas no mercado de Business Intelligence

GERÊNCIA DA ARQUITETURA DE DADOS

O objetivo da Gestão da Arquitetura de Dados é:

- Entender as necessidades de informação da empresa
- Desenvolver e manter o modelo corporativo de dados (MCD)
- Analisar e alinhar o MCD com outros modelos de negócios
- Definir e manter uma arquitetura de tecnologia de Dados
- Definir e manter uma arquitetura de integração de dados
- Definir e manter uma arquitetura de Data Warehousing e de Business Intelligence
- **Definir e manter uma taxonomia e padrões de nomes (namespaces) de dados para a empresa**
- Definir e manter uma arquitetura de Metadados

DESENVOLVIMENTO DE DADOS

Tem como objetivo projetar, implementar e manter soluções que satisfaçam as necessidades de dados da empresa. Compreende as atividades focadas em dados dentro do ciclo de desenvolvimento do sistema, incluindo a modelagem de dados, análise de requisitos de dados e projeto, implantação e manutenção de bancos de dados.

GESTÃO DE OPERAÇÕES DE BANCOS DE DADOS

Tem como objetivo planejar, controlar a apoiar os ativos de dados ao longo do seu ciclo de vida, indo desde a criação e aquisição (obtenção) até o arquivamento final (archiving) e eliminação (purge). A estrutura é:

- Suporte a Bancos de dados
  - Implementar e controlar ambientes de Bancos de Dados
  - Planejar para Recuperação de dados (Recovery)
  - **Realizar Backup e Recovery de Bancos de Dados**
  - Monitorar e ajustar aspectos de **performance de Bancos de Dados**
  - Planejar a retenção de dados

GESTÃO DE OPERAÇÕES DE BANCOS DE DADOS

- Gerência de tecnologia de dados
  - Avaliar tecnologias de dados
  - Instalar e administrar tecnologias de dados
  - Controlar e acompanhar aspectos de licenças de tecnologia de dados

GESTÃO DE SEGURANÇA DE DADOS

O objetivo é planejar, desenvolver e executar as políticas de segurança e procedimentos a fim de prover a adequada autenticação, acesso e auditoria de dados e informações. A estrutura é:

- Definir Política de **segurança** de dados
- Definir Procedimentos e controles de segurança de dados
- Monitorar autenticação de usuários e comportamento de acesso
- Auditar a segurança dos dados
- Classificar o grau de confidencialidade das informações

SEGURANÇA - LGPDP - LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS

A lei Nº 13.709 [9], estabelece que organizações públicas e privadas só poderão coletar dados pessoais, como nome, endereço, e-mail, idade, estado civil e situação patrimonial, se tiverem consentimento do titular. A solicitação deverá ser feita de maneira clara para que o cidadão saiba exatamente o que vai ser coletado, para quais fins e se haverá compartilhamento.

[9] Lei Nº 13.709

SEGURANÇA - LGPDP - LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS

Dados considerados “sensíveis”, que dizem respeito a crenças religiosas, posicionamentos políticos, características físicas, condições de saúde ou vida sexual, terão **utilização mais restrita**.
Nenhuma organização poderá fazer uso deles para fins discriminatórios. Também será necessário garantir que eles serão **devidamente protegidos** (multa por vazamento e descumprimento).

GESTÃO DE DADOS MESTRES E DE REFERÊNCIA

O objetivo é planejar, implementar e controlar atividades para garantir consistência de dados Mestres e de Referência.
Os dados Mestres são os dados fundamentais de uma empresa e envolvem clientes, fornecedores, colaboradores, contas, locais, entre outros.
Os dados de Referência são dados relacionados com códigos, como estado, país, status de um pedido, entre outros, e servem como elementos para categorizar/classificar outros dados.

GESTÃO DE DATA WAREHOUSING E BI

Tem como objetivo planejar, implementar e controlar processos para prover dados de suporte à decisão e apoio a colaboradores envolvidos em produção de relatórios, consultas e análises. A estrutura é:

- Entender as necessidades de informações analíticas (BI)
- Implementar os DW e DataMarts
- Implementar as ferramentas de BI e de Interface para usuários
- Processar os dados para o ambiente de BI
- Monitorar e ajustar os processos de DW

GESTÃO DE DOCUMENTOS E CONTEÚDO

O objetivo é planejar, implementar e controlar atividades para armazenar, proteger e acessar dados encontrados em arquivos eletrônicos e registros físicos (texto, gráficos, imagens, áudio e vídeo), ou seja, o foco em dados não estruturados, não armazenados em sistemas relacionais. Assim como documentação do ambiente.

GESTÃO DE METADADOS

O objetivo é planejar, implementar e controlar atividades que viabilizem um fácil acesso aos metadados integrados e de qualidade.
Um conceito simples e metafórico de metadado é aquela plaquinha que fica ao lado dos "rechauds" (bandejas), nos restaurantes de comida à quilo, indicando o nome do prato, detalhes da sua composição complementar, a sua localização.
Também quando se pensa num catálogo de biblioteca, entende-se com sentido mais computacional o conceito de metadados, ou seja, aqueles elementos que ajudam a entender os objetos, a sua composição, o seu relacionamento, a sua localização, entre outros.

GESTÃO DE QUALIDADE DE DADOS

O objetivo é planejar, implementar e controlar atividades que apliquem técnicas de gerência de qualidade de dados para medir, avaliar, melhorar e garantir a adequação dos dados ao seu uso pretendido. A estrutura de atividade desta função é:

- Desenvolver e promover aspectos de conscientização sobre Qualidade de Dados
- Definir requisitos de Qualidade de Dados
- Análise e avaliação de Qualidade de Dados
- Definir regras de negócios para Qualidade de Dados
- Medir e monitorar continuamente a Qualidade de Dados
- Corrigir os defeitos de Qualidade de Dados
- Projetar e implementar procedimentos operacionais de Gerência de Qualidade de Dados.

QUALIDADE - IMPORTÂNCIA NAS EMPRESAS

Essas informações são com base em uma pesquisa sobre Qualidade de Dados em empresas Brasileiras[10]:

- Falta de conhecimento
- Não é dado a devida Importância (outras prioridades)
- Falta de Apoio da Alta Gestão
- Falta de Conscientização por parte da empresa em geral

[10] A importância da Qualidade dos Dados nas Empresas

QUALIDADE - ARMADILHAS NO MERCADO DE BUSINESS INTELLIGENCE

Abaixo, algumas das principais armadilhas [11]:

- Falhas na Aquisição e Gerenciamento dos Dados (TI não alinhada ao negócio, Falhas de gerenciamento)
- Data Cleansing (Falta de Tratamento dos Dados - Imprecisão das Informações)
- Integração de Dados (Multiplas fontes - Complexidade de Integração)
- Enriquecimento de Dados (Poucos dados e pouca variedade, falta de dados de redes sociais)
- Treinamento para uso da Ferramenta (Pessoas despreparadas)

[11] https://blog.toccato.com.br/qualidade-da-informacao-e-dados-armadilhas-do-mercado-bi/

PRATICAR

- SGBD (SQL Server)
  - Administração (Visão Geral)
  - Bancos de Dados (Sistema — Usuário)
  - Segurança
  - Backup / Restore
  - Montagem de um ambiente (Banco de dados)
  - Monitoramento (SQL Profiler)
  - Visualização de Dados - BI (Introdução)

BUSINESS INTELLIGENCE

Refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios. [12]

[12] Business Intelligence

BUSINESS INTELLIGENCE

O que é?

- ***Data Driven, Data Lake*** **(Novos conceitos)**
- Mascaramento de dados
- ETL (Extração, Transformação e Carga)
- Visualização de Dados

DATA DRIVEN (CULTURA EM DADOS)

Cultura *data driven* consiste basicamente em tomar decisões embasadas em dados, ou seja, existe quando uma empresa organiza seus processos e métricas com base em dados reais, fugindo assim de decisões embasadas em intuição, instinto, exemplos passados, achismos ou heurísticas.

DATA DRIVEN (CULTURA EM DADOS)

Algumas empresas que tem essa cultura[13]:

- **NETFLIX** (*Stranger Things* - essa série foi escrita totalmente baseada em dados - baseada em filmes dos anos 90)
- **MARVEL** (O Filme Vingadores: Guerra Infinita - Foi baseado em dados dos filmes do universo Marvel(Homem de Ferro, Os Vingadores, Pantera Negra e etc...))

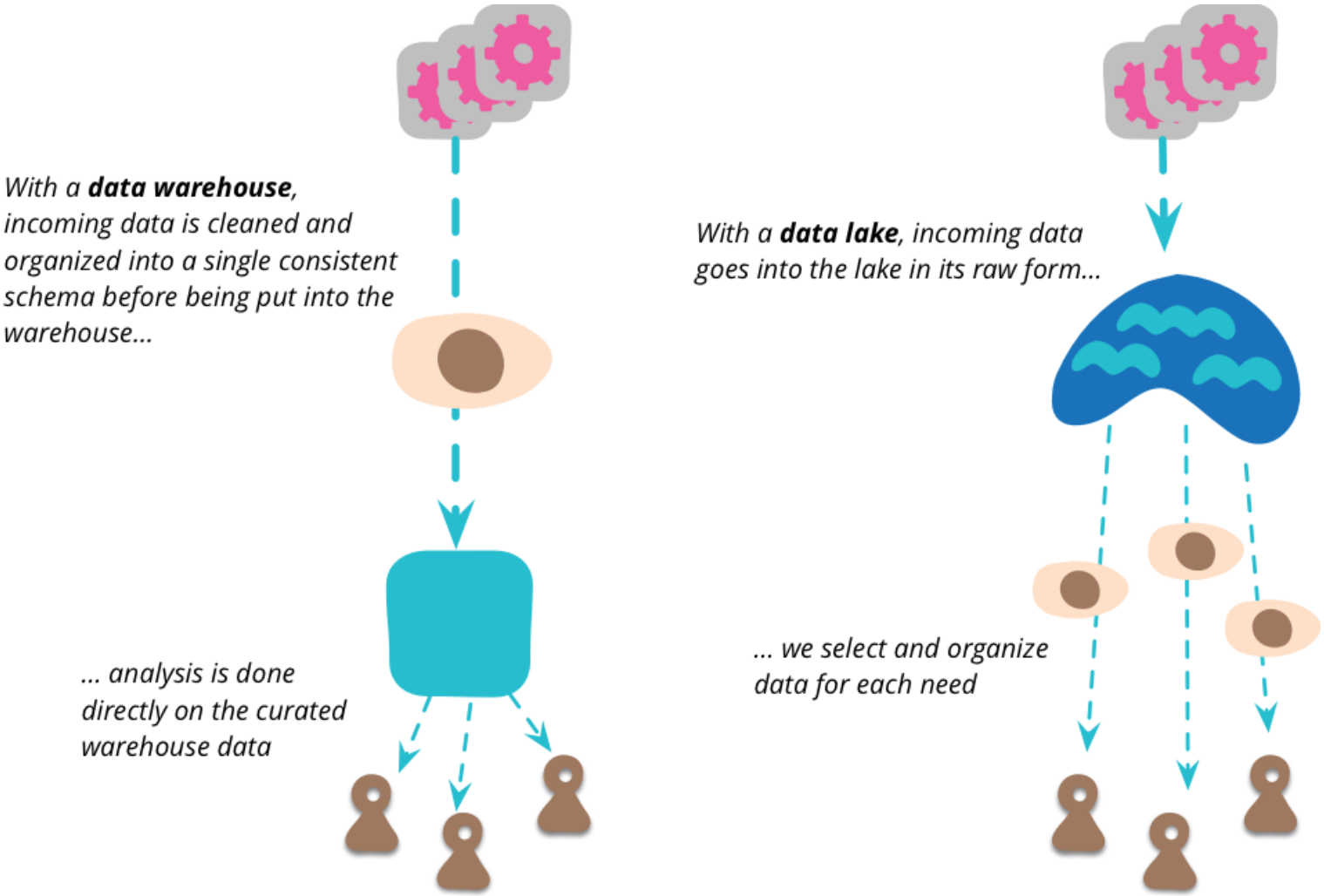
[13] https://inteligencia.rockcontent.com/cultura-data-driven/

DATA LAKE (LAGO DE DADOS)

"*Data Lake* pode ser definido como armazenamento centralizado, consolidado e persistente de dados brutos, não modelados e não transformados de múltiplas fontes, sem um esquema pré-definido explícito." [14]

[14] http://www.cienciaedados.com/data-lake-a-evolucao-do-armazenamento-e-processamento-de-dados/

DATA LAKE (LAGO DE DADOS)



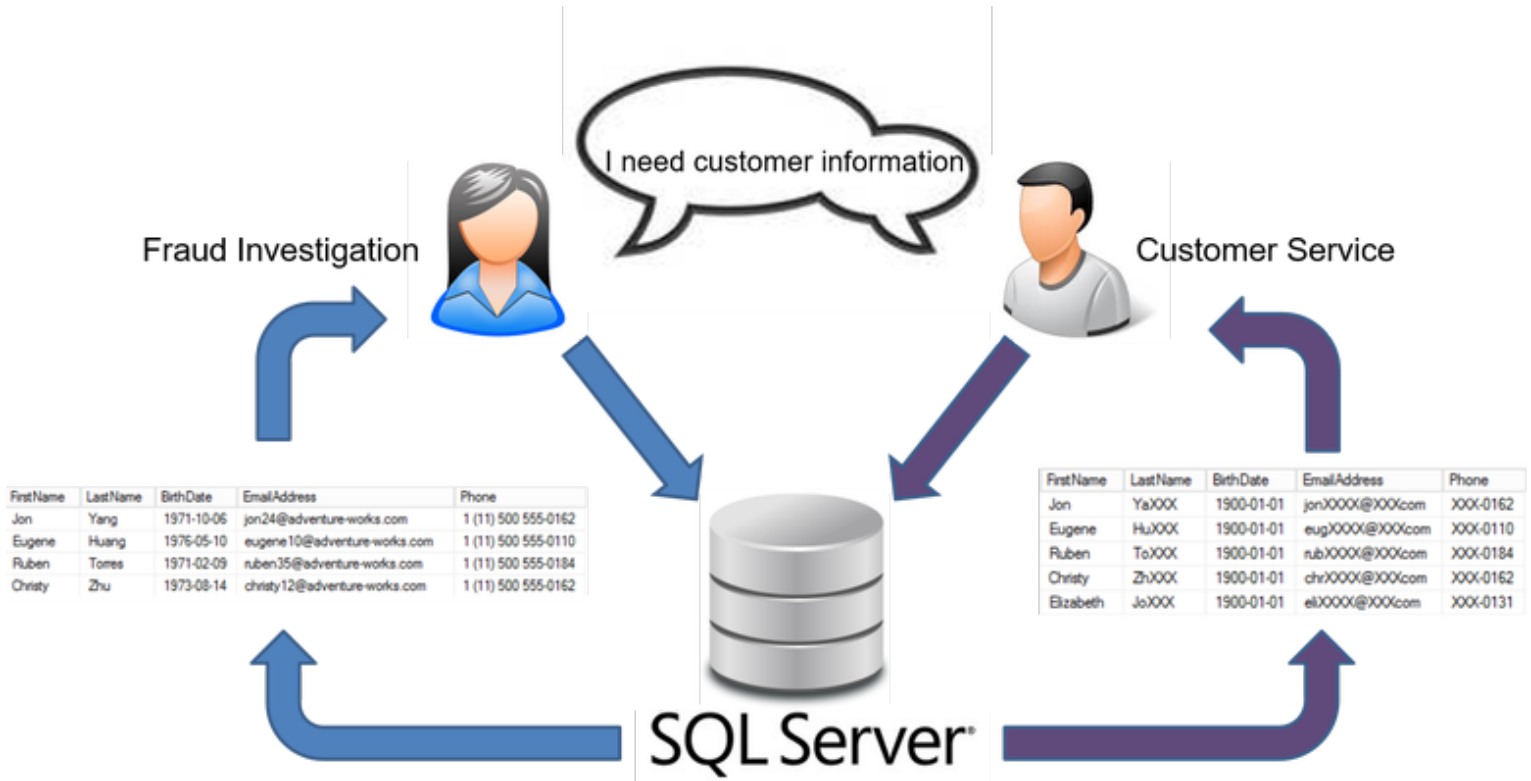
BUSINESS INTELLIGENCE

O que é?

- *Data Driven, Data Lake* (Novos conceitos)
- **Mascaramento de dados**
- ETL (Extração, Transformação e Carga)
- Visualização de Dados

MASCARAMENTO DE DADOS

Nos permite mascarar e ocultar informações sensíveis a determinados usuários de forma simples e prática. [15]



FirstName	LastName	BirthDate	EmailAddress	Phone
Jon	Yang	1971-10-06	jon24@adventure-works.com	1 (11) 500 555-0162
Eugene	Huang	1976-05-10	eugene10@adventure-works.com	1 (11) 500 555-0110
Ruben	Torres	1971-02-09	ruben35@adventure-works.com	1 (11) 500 555-0184
Christy	Zhu	1973-08-14	christy12@adventure-works.com	1 (11) 500 555-0162

FirstName	LastName	BirthDate	EmailAddress	Phone
Jon	YaXXX	1900-01-01	jonXXXX@XXXcom	XXX-0162
Eugene	HuXXX	1900-01-01	eugXXXX@XXXcom	XXX-0110
Ruben	ToXXX	1900-01-01	rubXXXX@XXXcom	XXX-0184
Christy	ZhXXX	1900-01-01	chrXXXX@XXXcom	XXX-0162
Elizabeth	JoXXX	1900-01-01	eliXXXX@XXXcom	XXX-0131

[15] Mascaramento de dados ($\geq$ SQL Server 2016)

BUSINESS INTELLIGENCE

O que é?

- *Data Driven, Data Lake* (Novos conceitos)
- Mascaramento de dados
- **ETL (Extração, Transformação e Carga)**
- Visualização de Dados

ETL (EXTRAÇÃO, TRANSFORMAÇÃO E CARGA)

ETL [16] - do inglês *Extract, Transform and Load* - tem como objetivo:

- Extrair informações de diversas fontes de dados distintas
- Transformação desses dados conforme regras de negócios (Limpeza, Qualidade, Consolidação de dados e etc)
- Carregar esses dados extraídos e Transformados para um repositório de dados periodicamente (Data Warehouse e etc)

[16] ETL

BUSINESS INTELLIGENCE

O que é?

- *Data Driven, Data Lake* (Novos conceitos)
- Mascaramento de dados
- ETL (Extração, Transformação e Carga)
- **Visualização de Dados**

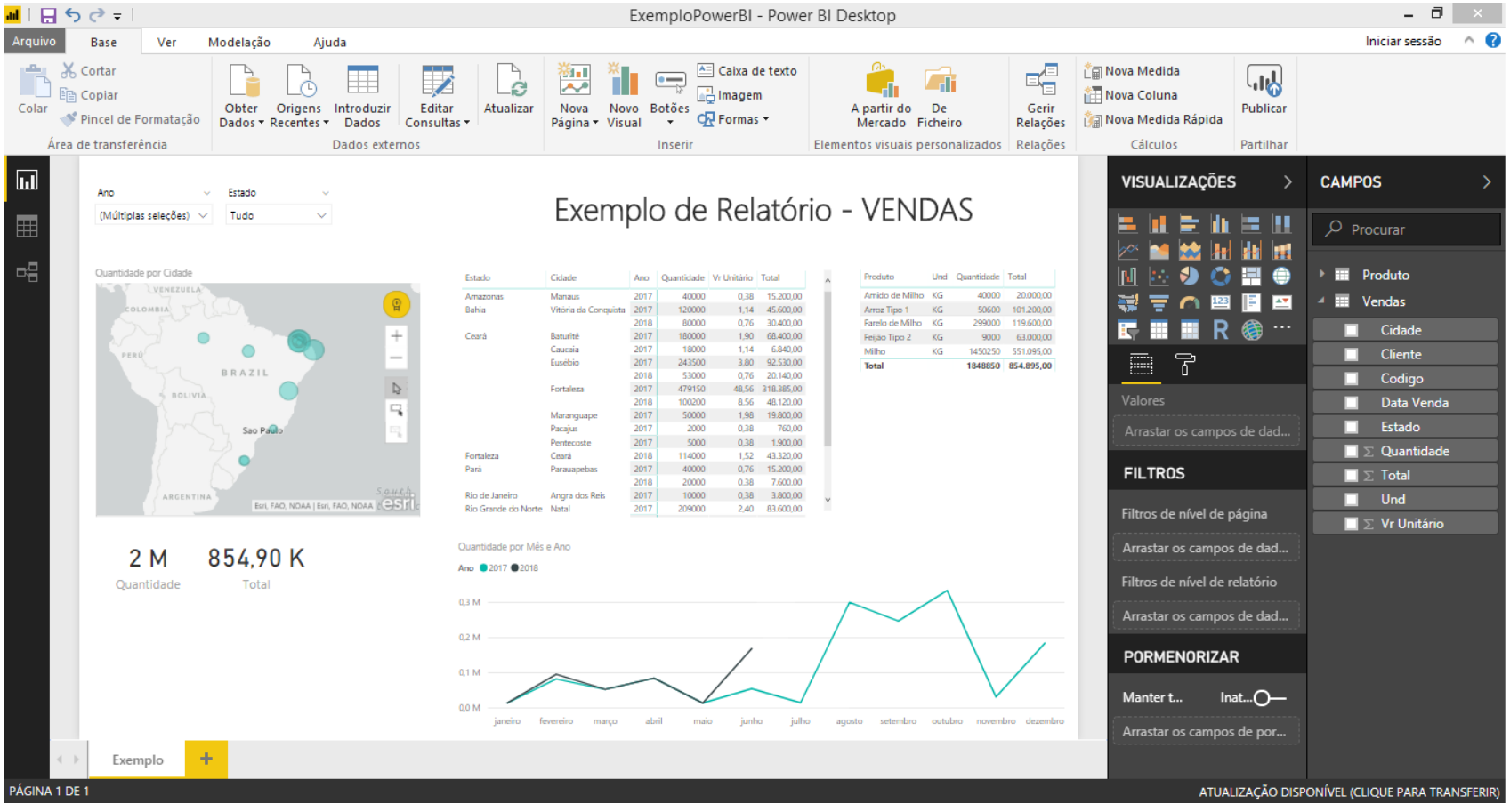
VISUALIZAÇÃO DE DADOS

"A visualização de dados consiste na representação gráfica de informações e dados. Usando elementos visuais, como diagramas, gráficos e mapas, a visualização de dados é uma forma acessível de ver e entender exceções, tendências e padrões nos dados."[17]

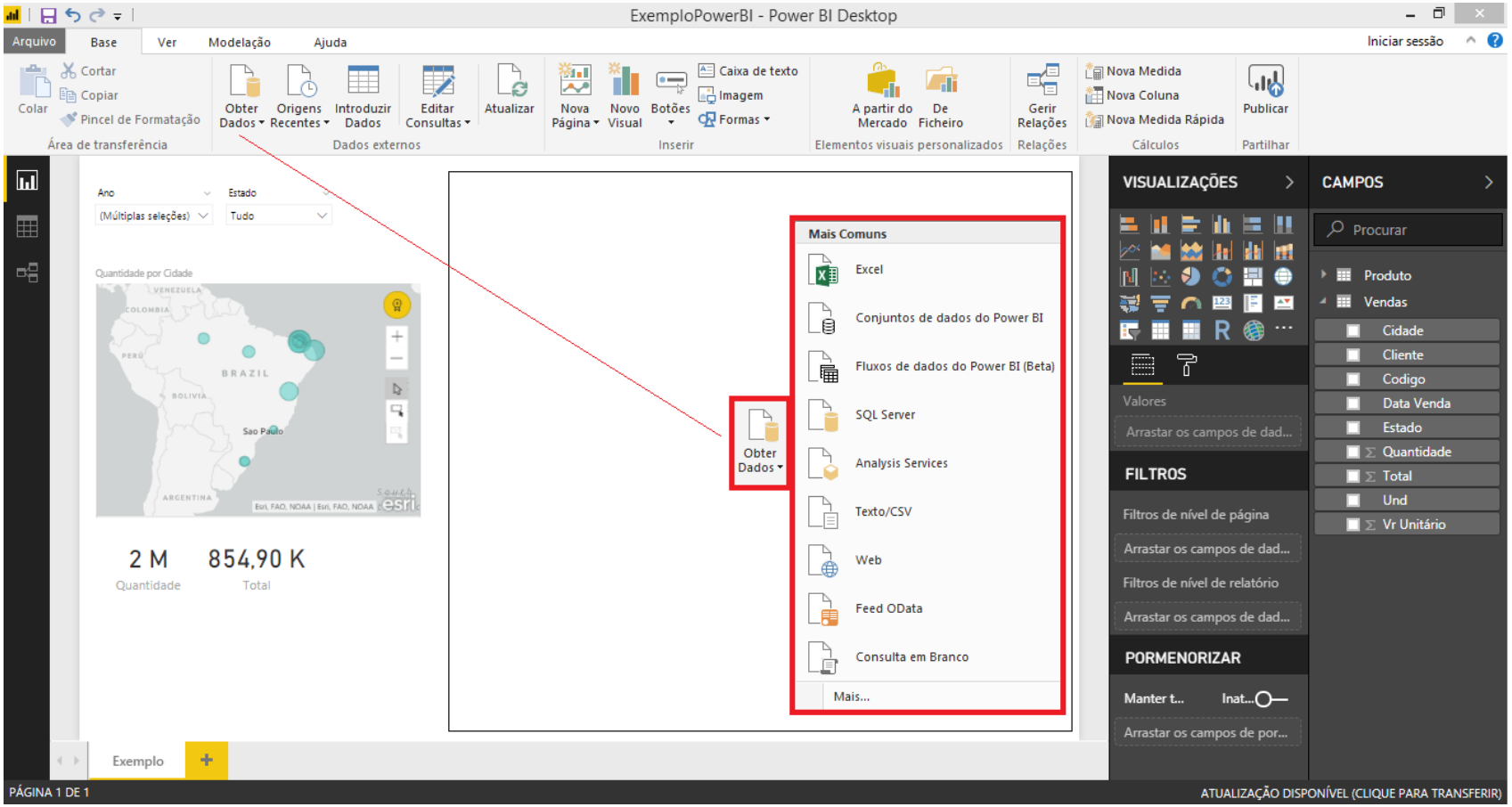
Relatórios, Painéis não precisam de explicação, eles mesmo devem se explicar.

[17] www.tableau.com/pt-br/learn/articles/data-visualization

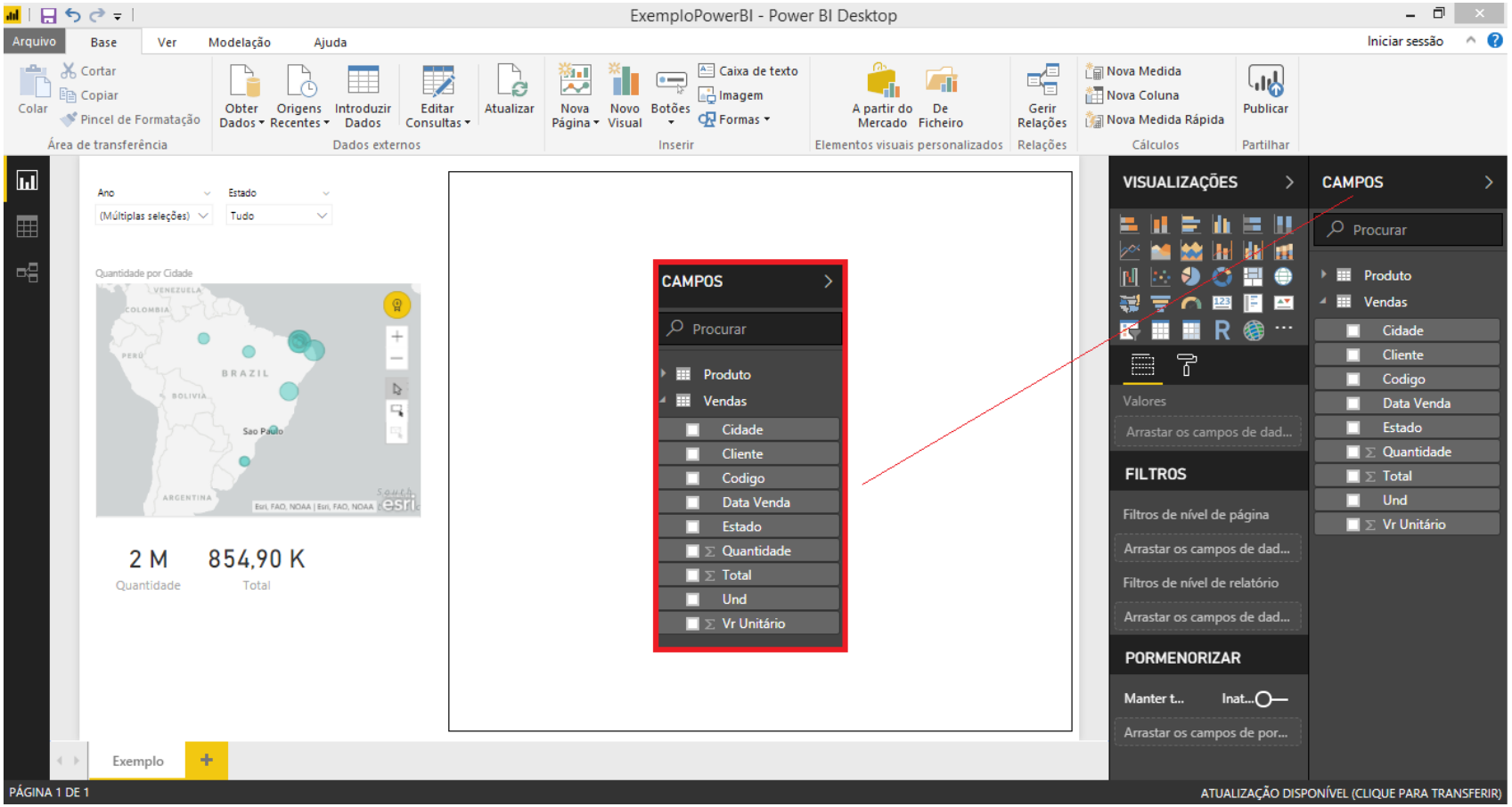
VISUALIZAÇÃO DE DADOS - POWER BI DESKTOP



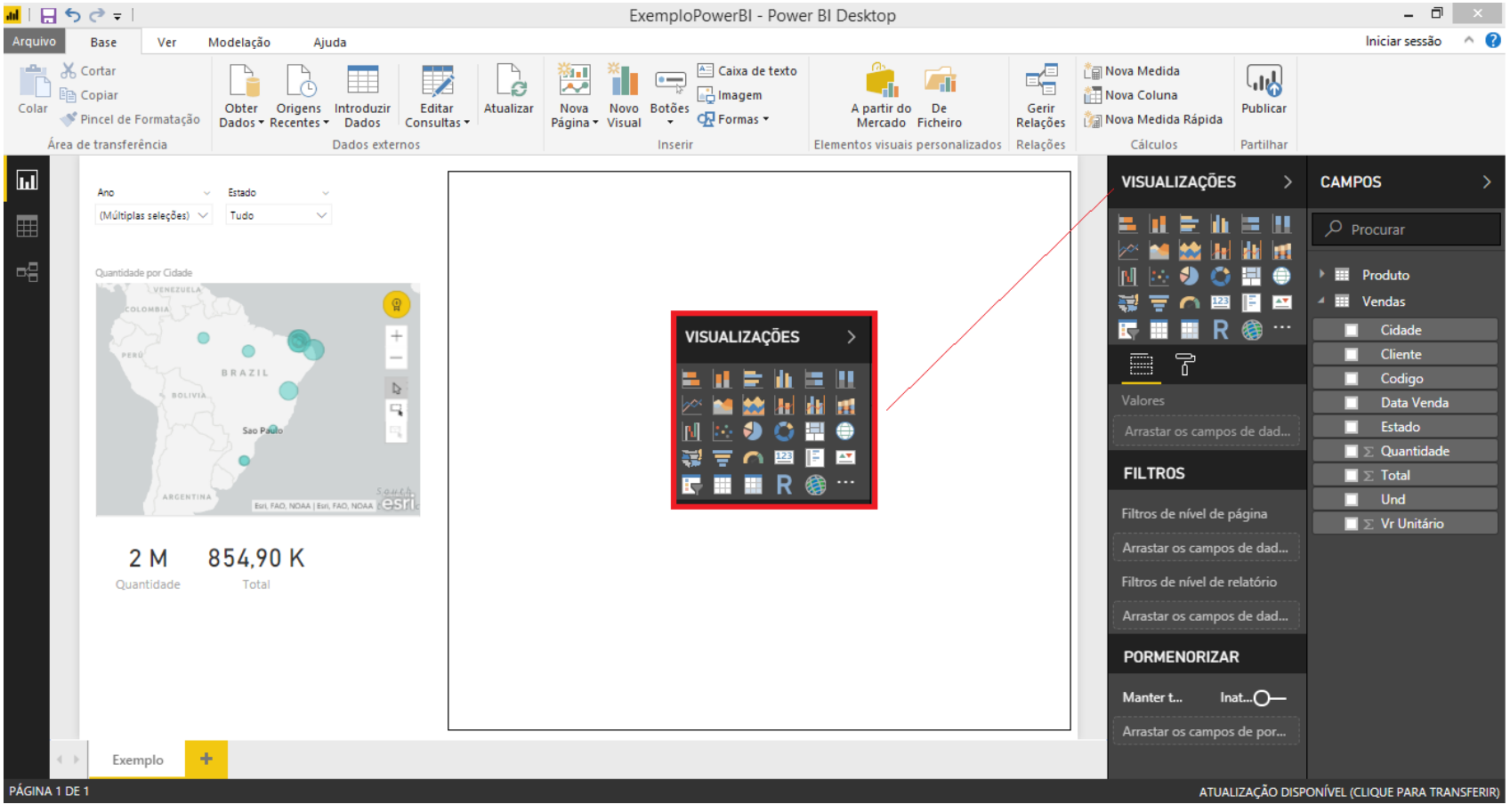
VISUALIZAÇÃO DE DADOS - OBTER DADOS



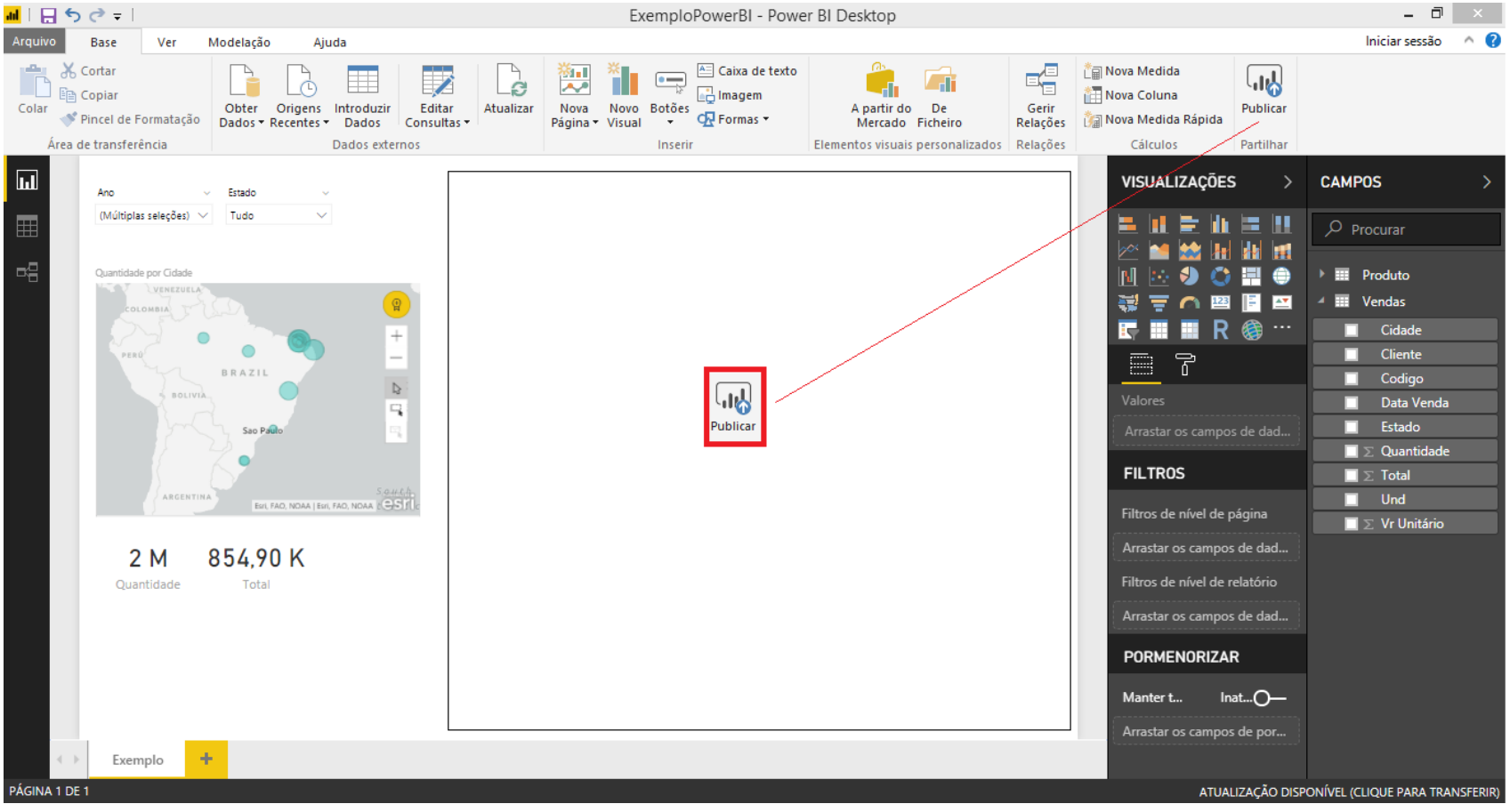
VISUALIZAÇÃO DE DADOS - CAMPOS



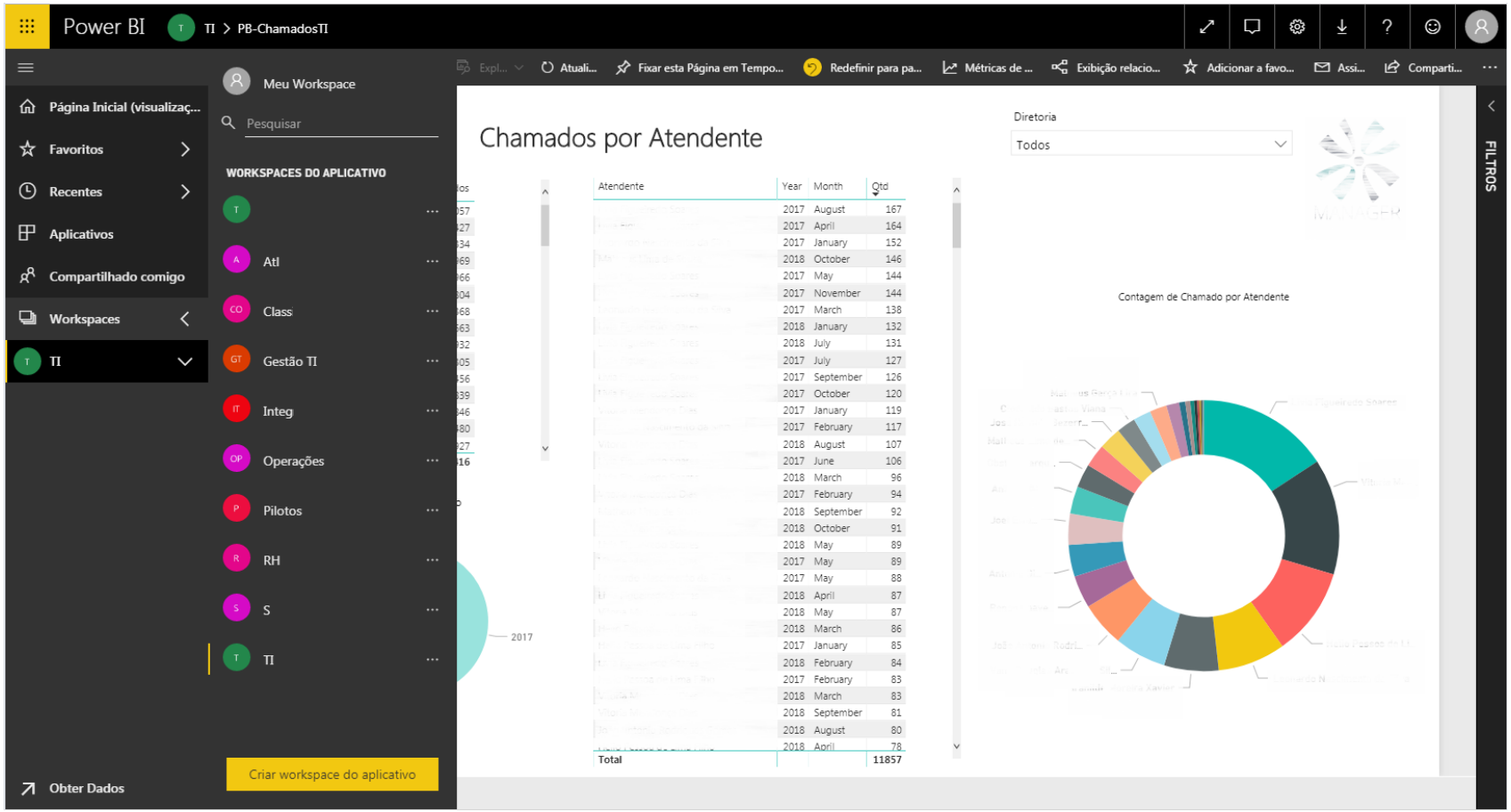
VISUALIZAÇÃO DE DADOS - VISUALIZAÇÃO



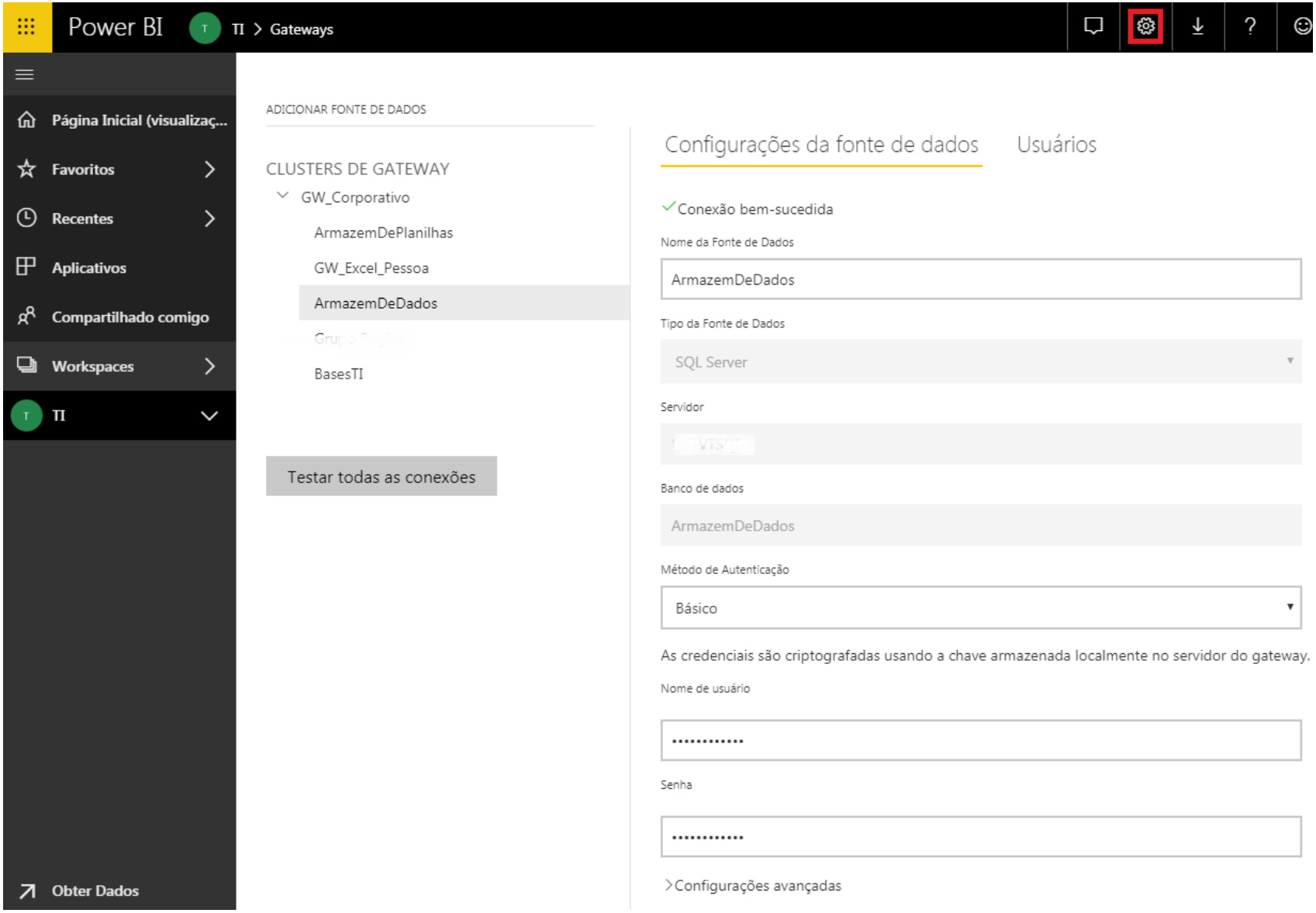
VISUALIZAÇÃO DE DADOS - PUBLICAR



VISUALIZAÇÃO DE DADOS - POWER BI WEB

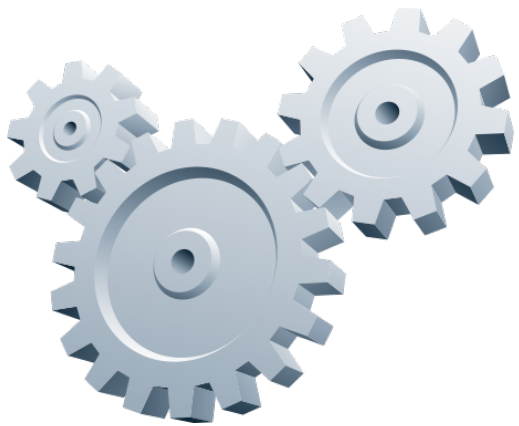


VISUALIZAÇÃO DE DADOS - GATEWAY



PRATICAR

- Montagem ambiente - SQL Server (Planilhas + Tabelas)
- Qualidade de Dados
- Mascaramento de Dados
- ETL
  - Extração
  - Transformação
  - Carga
- Visualização de Dados com PowerBI



MINERAÇÃO DE DADOS (MD)

É o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

MINERAÇÃO DE DADOS (MD)

Pode-se diferenciar o BI da MD como dois patamares distintos de atuação.

- O primeiro busca subsidiar a empresa com conhecimento novo e útil acerca do seu meio ambiente e funciona no plano estratégico;
- O segundo visa obter a partir dos dados operativos brutos, informação útil para subsidiar a tomada de decisão nos escalões médios e altos da empresa e funciona no plano tático(mais específico).

MINERAÇÃO DE DADOS (MD)

MD X IA [18]

- MD - Tem o foco na descoberta de propriedades desconhecidas nos dados;
- IA - Tem foco na predição, baseado em características conhecidas e aprendidas pelos dados de treinamento;

[18] Diferença entre Data Mining e Machine Learning

DICAS

- Site com informações (Dados, BD, BI, IA e RI).
- Dados.
- Fontes de Dados.
- Banco de Dados.
- SQL Server.
- BI.
- PowerBI.
- IA.
- Recuperação de Informação (RI).

OBRIGADO

UNIVERSIDADE ESTADUAL DO CEARÁ