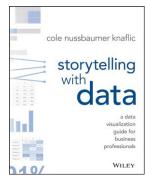
Gestão, Governança e Arquitetura de Dados PUCPR

Marcelo Rosano Dallagassa

Quais são os caminhos (tendências) para a Análise de Dados?

- Foco em resolver questões -> storytelling com os dados
- Insights Destaques importantes sobre a análise de dados.
 São gatilhos para desenvolver ações de melhorias e evolução dos apontamentos realizados.
- Análise Top-down Abordagem geral e depois detalhe os resultados
- Aplicação direta da Ciência de Dados Aprendizagem de Máquina para a descoberta de padrões de forma automática. Auto ML instrumentos que permitem automatizar e reconhecer dinamicamente padrões nos dados do mundo real. Ex. Datarobot, H2O, Datalku, entre outras.
- Aplicação da Mineração de Processos para a descoberta do mapeamento do processo, dentro dos registros (log de eventos).





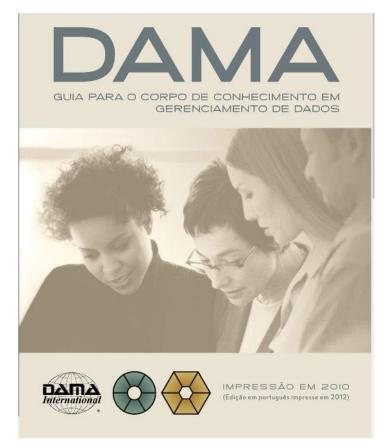


Qual é a estrutura para desenvolver esse ambiente?



Governança e Gestão de Dados

GOVERNANÇA DE DADOS - DAMA-DMBOK ® Data Management Body of Knowledge





John Zachman

Governança e Gestão de Dados

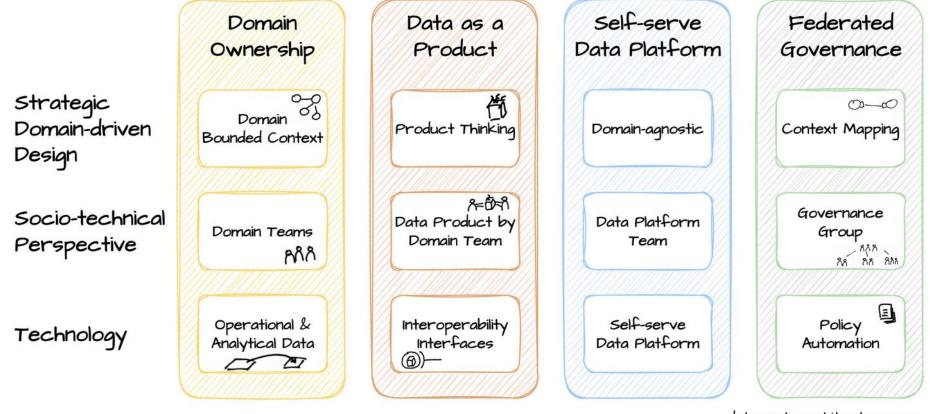




Data Mesh

O conceito *Data Mesh* foi aplicado por **Zhamak Deghani** em **2019**, e apresenta 4 princípios fundamentais que agrupam o conceito

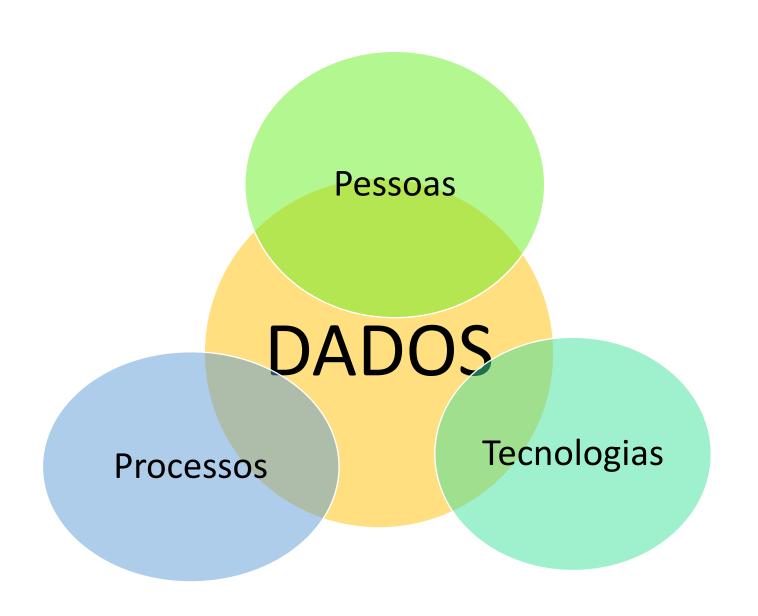
What is Data Mesh?



datamesh-architecture.com

Fonte: https://www.datamesh-architecture.com/

Governança de Dados



Governança de Dados



Definição

 Controle e melhoria eficaz do uso dos ativos de dados atendendo e excedendo as necessidades de informações de todos os envolvidos (stakeholders) da empresa em termos de disponibilidade, segurança e qualidade



Objetivos

- Definir, aprovar e comunicar a estratégia de dados, políticas, arquitetura, procedimentos e métricas.
- Acompanhar e exigir o cumprimento das normas e políticas de conformidade, padrões, arquitetura e procedimento
- Patrocinar e acompanhar as entregas de projetos de gestão de dados
- Gerenciar e resolver problemas relacionados aos dados
- Entender e promover o valor dos ativos de dados

Analisar, definir e manter o modelo corporativo de dados (MCD) **ARQUITETURA DE DADOS** Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)

Projetar e manter a implementação dos dados **DESENVOLVIMENTO DE DADOS** Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)

Planejar, Controlar e apoiar os ativos dos dados durante o ciclo de vida **GESTÃO DE OPERAÇÃO DE DADOS** Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)

Planejar, desenvolver e executar as políticas de segurança e procedimentos **GESTÃO DE SEGURANÇA DE DADOS** Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)

Planejar, implementar e controlar as atividades para garantir a consistência de dados Mestres e de Referência. **GESTÃO DE DADOS MESTRES E DE REFERÊNCIA** Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)





Planejar, implementar e controlar atividades que viabilizem um fácil acesso aos metadados **GESTÃO DE METADADOS** Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)

Planejar, implementar e controlar atividades que apliquem técnicas de gerência de qualidade de dados **GESTÃO DE QUALIDADE DE DADOS** Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)

Diagnóstico da Governança de Dados

Entendimento da situação atual (AS-IS) Definição da situação futura (TO-BE)

Definição do Roadmap

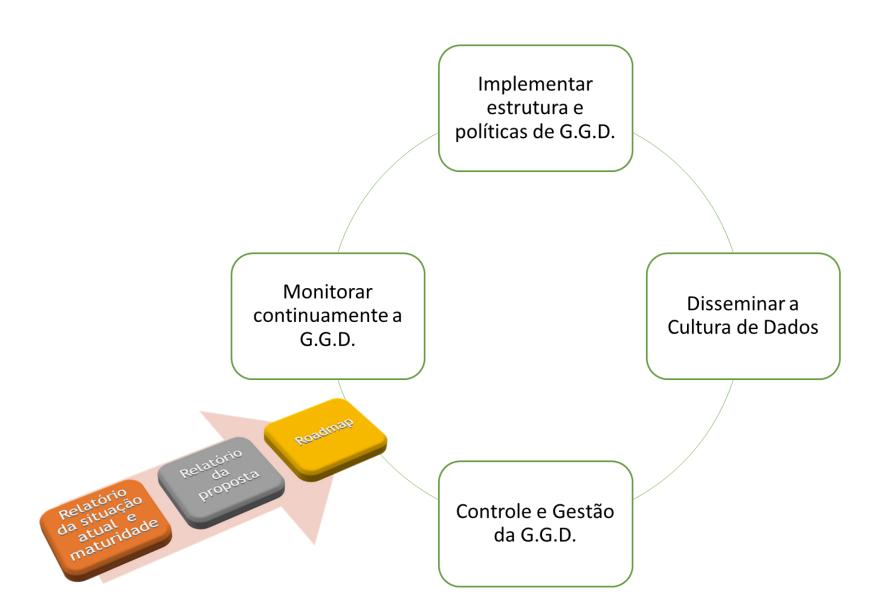
Entregas do Diagnóstico da Governança de Dados

Relatório da situação atual e maturidade

Relatório da proposta de gestão de dados

Roadmap para gestão e governança de dados

Roadmap.



Estudo de Caso - Governança de Dados

Desenvolva um diagnóstico (assessment) sobre governança de dados da empresa **Ortomagic**, indústria de calçados ortopédicos, e algumas recomendações para seguir no desenvolvimento do seu projeto de gestão de dados, ou seja, um roadmap para elevar o nível de maturidade da situação atual (AS-IT) para a desejável (TO BE).



A indústria de calçados Ortomagic possui 250 funcionários, e não tem filiais. Por meio de vendedores externos (representantes comerciais), recebe pedidos de vendas, encaminhados eletronicamente e adicionados automaticamente no sistema de gestão, que realiza a provisão da matéria-prima para sua produção por meio de uma ordem de fabricação.



O cadastro de clientes utilizados pelos vendedores engloba tanto clientes ativos, prospecção de negócios como alimentados pelo setor de vendas, via sistema de vendas. Quando finalizada a fabricação dos calçados, são geradas as notas fiscais de vendas, com base nos itens produzidos. A inspeção de qualidade final dos produtos e a expedição geram o certificado do produto e o documento de transporte para a entrega ao cliente final.



A empresa conta com uma equipe de TI, que desenvolveu seu sistema administrativo, vendas e de produção, e há um profissional dedicado para a administração e a arquitetura de dados. Não possui um sistema de BI/DW e pretende implementá-lo em breve, com a contratação de uma consultoria especializada (VOCÊ!).



No levantamento de informações, VOCÊ percebeu, que não há cuidadores/curadores de dados, nem cultura sobre qualidade e segurança, ocasionando muitas vezes falha nos processos, ou seja, emissão de notas fiscais com dados incorretos, envio de correspondência aos clientes com endereços errados, falha de contato com fornecedores e compra de matéria-prima inadequada por falta de qualidade no cadastro de materiais/insumos.

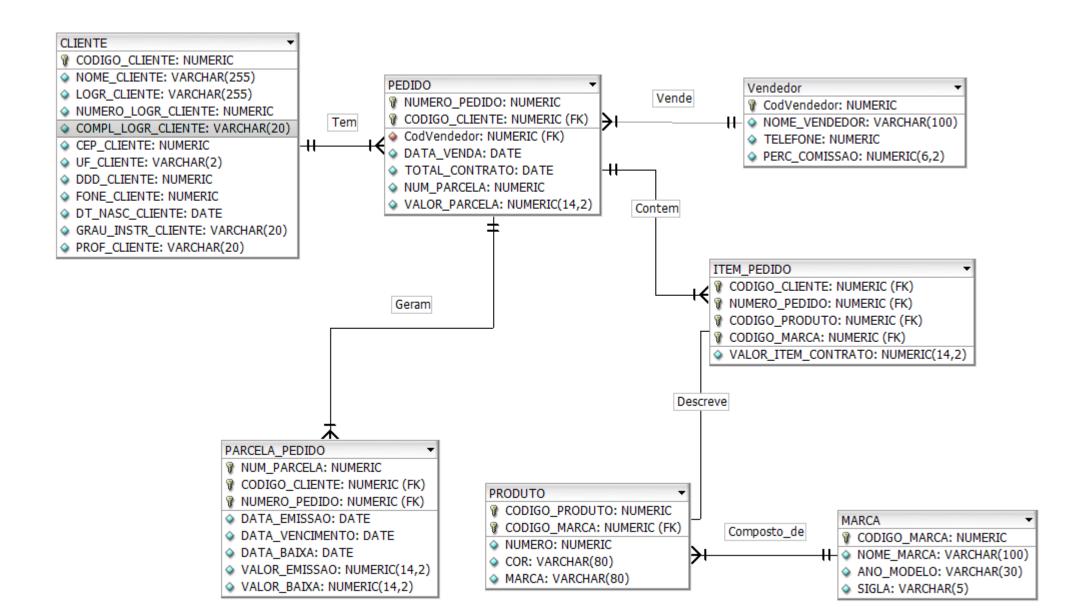


FUNÇÕES DO DMBOK - Levantamento de informações



Áreas de conhecimento na Gestão de Dados – Segundo DMBOK – fonte: Barbieri (2013)

Modelo de Dados Transacional - Ortomagic



SUGESTÃO DE PONTUAÇÃO SOBRE O NÍVEL DE MATURIDADE (adaptação DMBOK).

Score - Visão situação atual (AS-IT)

- 0 Não possui nenhuma iniciativa
- 1- As atividades geralmente são reativas e não preventivas, e inexiste processos definidos
- 2- Os processos existem e são padronizados e gerenciados
- 3- Além de atender o nível 2, os processos são constantemente avaliados e identifica-se oportunidade de melhorias

1 Governança de Dados

Existe alguma área específica para a Governança de Dados?

Há algum gestor responsável pela Governança de Dados?

Existe algum projeto formal para tratar ações de Gestão de Dados?

2 Gerência da Arquitetura de Dados

Existe profissional responsável pela arquitetura de dados para a definição do modelo corporativo de dados?

É utilizado alguma ferramenta CASE para modelagem de dados?

Foi desenvolvido e mantido o modelo de dados corporativo (MCD)?

3 Desenvolvimento de Dados

Existe a modelagem lógica e física dos projeto de sistemas da organização?

Existe na organização os profissionais (DBA's) responsáveis pelo gerenciamento e manutenção dos dados para os diversos ambientes da organização?

São projetados os modelos de dados para representar a preservação de fundamentos dos ambientes transacionais (Atomicidade, Consistência, Isolamento e Durabilidade)?

4

Gestão de Operações de Banco de Dados Existem processos para administração, otimização e performance do banco de dados?

Existe uma política para tratamento e controle dos dados durante todo o seu ciclo de vida ?

Existem cuidados de mascaramento ou embaralhamento para cuidados em relação a segurança da informação?

5 Gestão de Segurança de Dados

Existem políticas e padrões de segurança, com a definição de perfis de usuário para atender requisitos de autenticação, autorização, acesso e auditoria?

Há mapeamentos dos processos/dados com a identificação de dados sensíveis e pessoais?

Estão bem definidos os processos de segurança de dados?

6

Gestão de Dados Mestres e de Referência Há alguma estratégia de implementação e controle de dados mestre e dados de referência?

As regras de negócios estão bem definidas para automatizar a qualidade dos dados mestres (consistência)?

Existem glossário de terminologia de negócios (tesauro) para identificar e validar a abrangência de utilização dos dados de referência?

Gestão de Big Data, DW e BI

Existem iniciativas de modelagem de dados específicas (data marts) para a criação de ambientes de sistemas de apoio a decisão?

Há padronização e normalização desses ambientes para refletir o cuidado nos processos de ETL (extração, transformação e carga de dados)?

Há algum ambiente específico de análises preditivas utilizando-se de aprendizagem de máquina?

8

Gestão de Documentos e Conteúdo A gestão de documentos utiliza-se de ferramentas que garantem a governança de dados futuros?

A coleta de informações e dados pessoais e sensíveis utiliza-se de controles por meio de ferramentas?

As impressões de documentos, quando necessária, são observados os controles de gestão desses documentos?

9

Gestão de Metadados Há definição dos termos de negócios nas aplicações que utilizam-se dos dados?

Existe iniciativas de centralizar e unificar diversas fontes de metadados?

Há algum processo para desenvolver e manter os padrões de metadados?

10

Gestão de Qualidade de Dados Há definição de requisitos, métricas e regras de negócios sobre a qualidade dos dados?

É realizado a medição e controle de forma contínua da qualidade de dados?

Existem processo de garantia e qualidade dos dados?

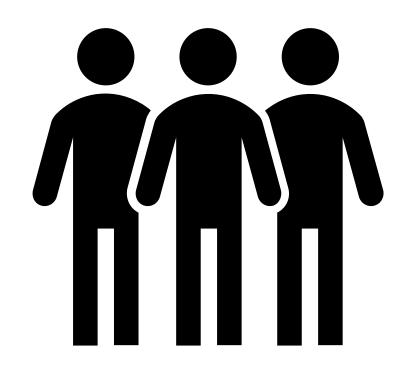
11

Integrações e Interoperabilid ade Como são realizadas as integrações dos dados entre os clientes e fornecedores?

Existe algum processo padronizado de integrações de sistemas?

Existe coesão entre os diversos sistemas, padrões de interoperabilidade entre eles?

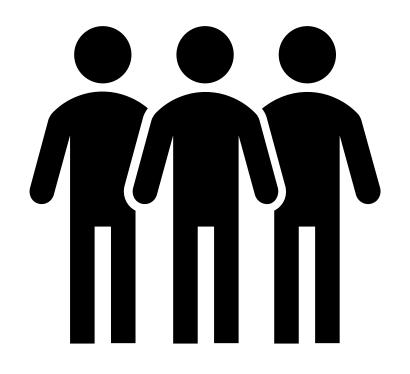
Estrutura Governança e Gestão de Dados



CDO – Chief Data Officers

Comitê de Governança de Dados

Estrutura Governança e Gestão de Dados



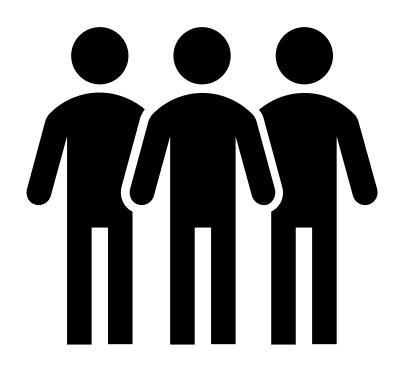
Data Owners – Curadores Proprietários

Curador de dados de projetos

Curador de dados técnicos

Curadores de dados

Estrutura Governança e Gestão de Dados



Arquitetura de dados

"O conjunto de representações descritivas que são necessárias para criar um objeto "John Zachman

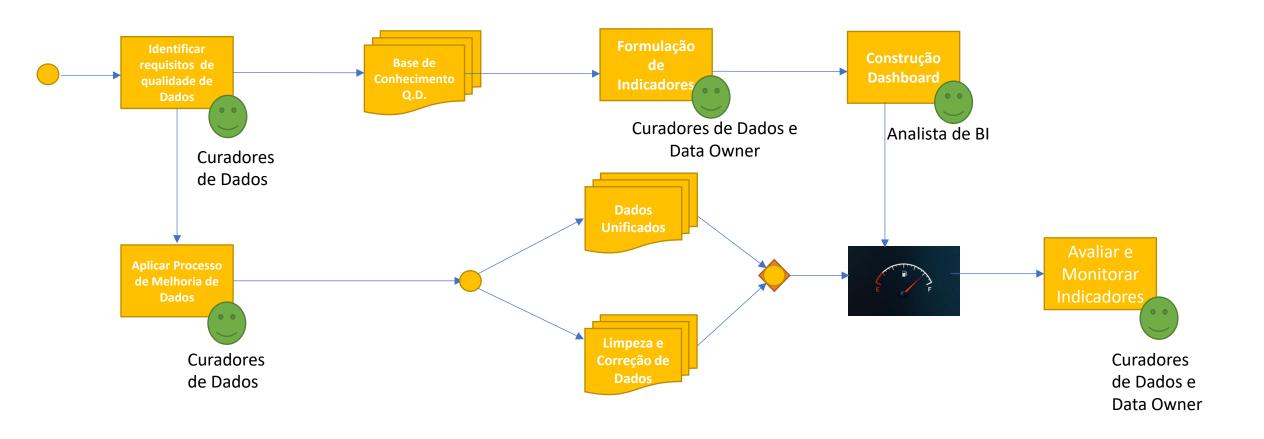
Papel do arquiteto de dados:

Alinhar os recursos e investimentos com a estratégia de negócios

Integrar recursos para atingir de forma mais eficaz os objetivos de negócios

Ajudar na gestão da mudança corporativa.

Proposta de Processo de Garantia da Qualidade de Dados



Fonte: o próprio autor (2022)



Qualidade de Dados - Exemplos

Unicidade	Dois clientes diferentes com o mesmo CNPJ?
Completude	Falta a UF no endereço?
Validade	Valor preenchido com caractere de 20 posições?
Razoabilidade	A UF do Cliente corresponde a UF de entrega do Pedido de Venda?
Precisão Acurácia	O valor total da venda corresponde ao total dos itens do pedido?
Integridade	O valor do item corresponde ao valor de tabela?
Temporalidade	Data da entrega do pedido de venda é superior a data da solicitação do pedido?

Próximas Etapas da Atividade:

- 1. Tabule os resultados de cada função do DMBOK, utilize o score de 0 a 3 para cada item.
- 2. Crie um gráfico radar para representar quais são os pontos mais distantes entre os desejáveis
- 3. Escreva um resumo conclusivo, apresentando a situação atual (AS-IT), a situação desejável (TO-BE) e quais os próximos passos como sugestão para a implantação da governança de dados na **Ortomagic**
- 4. Proponha ainda, uma estrutura funcional/pessoas para compor a área de Governança de Dados
- 5. Defina uma política de qualidade de dados para a **Ortomagic**

Nível de Maturidade - OrtoMagic?

Função DMBOK	Score Max	Score Ating
1 - Governança de Dados	3	0
2 - Gerência da Arquitetura de Dados	3	0
3 - Desenvolvimento de Dados	3	0
4 -Gestão de Operações de Banco de Dados	3	0
5 - Gestão de Segurança de Dados	3	0
6 - Gestão de Dados Mestres e de Referência	3	0
7 - Gestão de Big Data, DW e BI	3	0
8 - Gestão de Documentos e Conteúdo	3	0
9 - Gestão de Metadados	3	0
10 - Gestão da Qualidade de Dados	3	0
11 - Integrações e Interoperabilidade	3	0
Score Total	33	0

Data Lakehouse

Ambiente que reúne as estruturas de armazenamento em larga escala de dados estruturados (Data Warehouse) com os dados não estruturados (Data Lake) e oportuniza uma estrutura de alta capacidade de gerenciamento e processamento de forma otimizada



Exemplo de Ambientes Data Lakehouse





- 1. Modo de Serviço: PaaS -Plataforma de análise de Big Data baseada no Apache Spark.
- **2. Plataforma Cloud:** AWZ, Azure, GCP, aquisição deste serviço à parte.
- **3. Escalabilidade:** Escalabilidade horizontal e gerenciamento de recursos automático.
- **4. Linguagem de consulta:** Suporte para várias linguagens, incluindo Python, Scala, Java e R.
- **5. Recursos:** Processamento de dados em lote e streaming, aprendizado de máquina, análise de texto, gráficos e consultas SQL, ambiente colaborativo de notebooks.
- **6. Casos de uso:** Análise avançada, processamento de dados em tempo real, aprendizado de máquina, ETL e pipelines de dados.
- 7. Facilidade de uso: curva de aprendizagem maior, com exigência da experiência dos profissionais em Apache Spark.
- **8. Escabilidade:** com base em Apache Spark
- Manutenção: Média, em função da dependência dos Cloud Provider
- **10. Funcionalidade:** Necessidade de instalação no Cloud Provider do deployer
- 11. Arquitetura: Baseado no Apache Spark (Arquivos Delta)

- 1. Modo de Serviço: SaaS Data Warehouse baseado em nuvem.
- **2. Plataforma Cloud:** AWS, Azure e GCP, contratação embutida com o Serviço snowflake.
- 3. Escalabilidade: Automática de armazenamento e computação, separação entre computação e armazenamento.
- 4. Linguagem de consulta: SQL (principal), Java, Python, Scala
- 5. Recursos: Armazenamento e gerenciamento de dados em larga escala, suporte para processamento de dados estruturados e semiestruturados, fácil integração com ferramentas de visualização e análise.
- **6. Casos de uso:** Análise de dados, business intelligence, relatórios e dashboards.
- 7. Facilidade de uso: baseado em sql padrão, fácil de utilizar e com gestão automática.
- 8. Escabilidade: independente de armazenamento e processamento.
- 9. Manutenção: Baixa
- 10. Funcionalidade: sem a necessidade de instalação do deployer
- 11. Arquitetura: Separa armazenamento e processamento

Exemplo de Ambientes Data Lakehouse



O Databricks, focado em Apache Spark, é ideal para processamento de dados e machine learning, exigindo mais conhecimentos técnicos e configurações. Possui limitações em relação a formato de dados e flexibilidade. As cobranças são individualizadas em relação ao Cloud Provider. Para aplicação de políticas de segurança e governança de dados atua com o Unit Catalog.



Snowflake, focado no conceito de Data Warehouse (SQL), destaca-se pela simplicidade e eficiência de entregas, suportando dados estruturados e semi-estruturados e disponível em uma única plataforma e sem dependência de instalação de serviços e cobranças em cloud provider. A cobrança é no modelo "pay-as-you-go", ou seja, pagamento conforme a utilização. Tem embutida na sua arquitetura a criação de políticas de segurança e governança de dados

Algumas Ferramentas para Ambientes Data Lakehouse

Ingestão















Armazenamento













Processamento











Orquestração











Monitoramento









Entrega







Ciência de Dados













Business Intelligence x Analytics

Objeto	Business Intelligence	Business Analytics
Finalidade	Relatórios, KPIs, Dashboards	Identificação de Padrões, correlações, modelos preditivos
Processo	Estático, comparativo	Exploratório, experimental, visual
Qualidade dos dados	Alta	Baixa ou Média (exige processo de limpeza e transformação)
Tipo de Análise	Descritiva, retrospectiva	Preditiva, Prescritiva
Tipo de respostas	O que aconteceu?	O que pode acontecer?

Ferramentas Analytics







alteryx















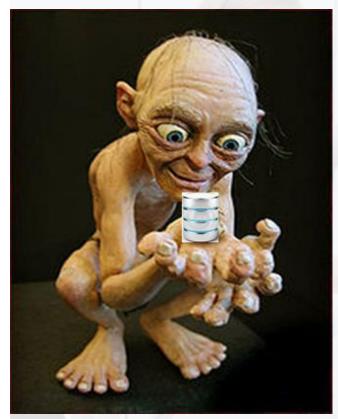




 Extração não trivial da informação implícita, previamente desconhecida e potencialmente útil a partir dos dados.

 Exploração e análise, por meio semiautomáticos ou automáticos, de grande quantidade de dados a fim de descobrir padrões significativos.

Business Analytics



"É um processo não trivial para identificar padrões que sejam válidos e previamente desconhecidos, potencialmente úteis e compreensíveis, visando melhorar o entendimento do problema, possibilitando o auxílio no processo de tomada de decisões" (FAYYAD et al., 1996).

A descoberta de uma informação preciosa que pode-se transformar em um conhecimento novo

Ciência de Dados

Justificativas

Grande volume de dados armazenados:

- ✓ Web data, e-commerce
- ✓ Compras (departamentos, Supermercados)
- ✓ Bancos, transações com cartão de crédito,...

Outras:

- ✓ Computadores mais acessíveis e mais potentes
- ✓ Mercado mais competitivo
- ✓ Fornecimento de produtos e serviços com mais qualidade, personalizados e com menores custos.





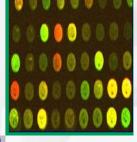


Ciência de Dados



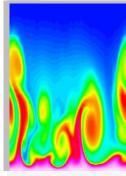
Do Ponto de Vista Científico

- ✓ Dados coletados a enormes velocidades (GB/hora)
- ✓ Sensores remotos em satélites
- √ Telescópios investigando o universo
- ✓ Micro arranjos gerando dados de expressão
- ✓ gênica
- ✓ Simulações científicas gerando terabytes de informações



Técnicas tradicionais inviáveis para dados brutos A Ciência de dados auxilia os cientistas em:

- ✓ Classificação e segmentação de dados
- ✓ Na formulação de hipóteses.





Ciência de Dados

Como serão os custos de produção do próximo ano?

Quais as Doenças que mais acometem os nossos Pacientes?

Será que trata-se de um caso de fraude?

Como será o comportamento desse cliente nos próximos 6 meses?

Business Intelligence x Analytics

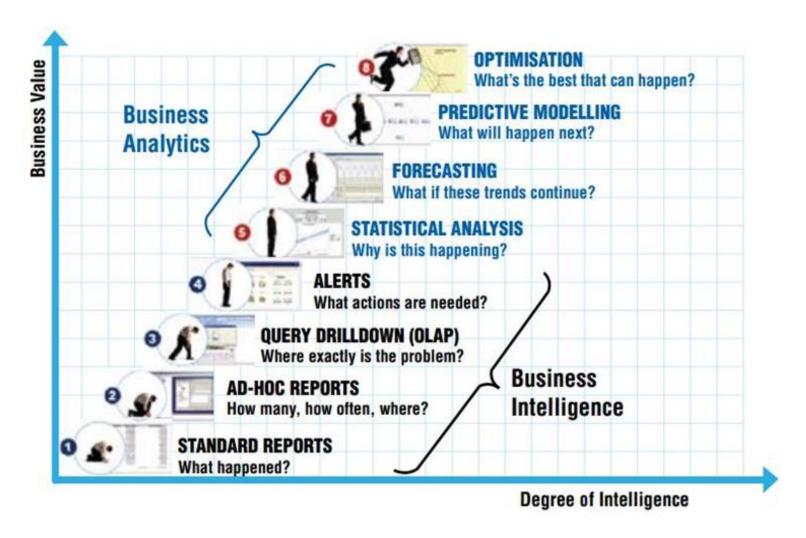


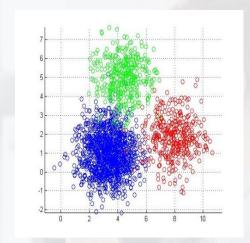
Figura 2- BI x BA - Michael Walken - http://www.rosebt.com/1/post/2012/09/eight-levels-of-analytics-for-competitive-advantage.html



Aprendizagem de Máquina - Métodos

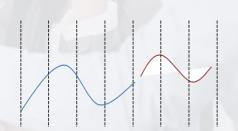
Métodos Descritivos

Encontrar padrões compreensíveis por humanos para descrever os dados



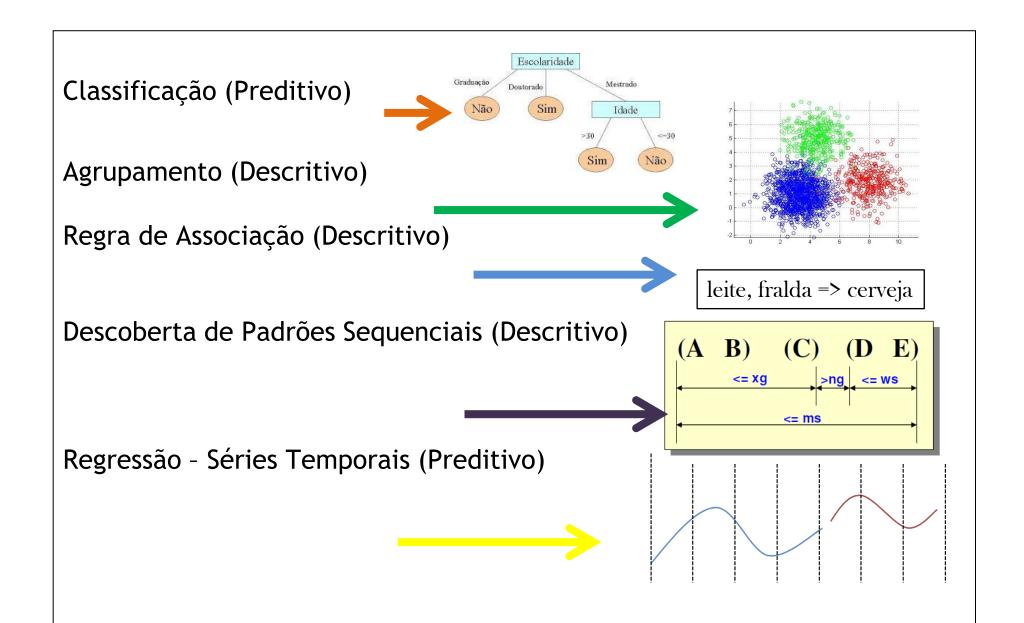
Métodos Preditivos

Usar algumas variáveis para prever valores desconhecidos ou futuros de outras variáveis





Tarefas de Aprendizagem de Máquina



Dúvidas



Referências

- Date, Christopher J., Introdução a Sistemas de Bancos de Dados, Campus, 2005
- Silberschatz, Abraham; Korth, Henry F., Database System Concepts, Elsevier Brasil, 2016
- Barbieri, C. Governança de Dados, Práticas, Conceitos e Novos Caminhos. Alta Books. 2019
- Barbieri, C. Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK)
- Elmasri, R.; Navathe, S. **Sistema de Banco de Dados**, 4 ed. São Paulo, Pearson Addison Wesley, 2005.
- HARJINDER, Gill S.; PRAKASH Rao C. The Official Guide to Data Warehousing, Indianapolis, USA: QUE Corporation, 1996. 382p

Referências

- Kimbal, Ralph. The Data Warehouse Toolkit. New York, USA: John Wiley & Sons, Inc., 1996. 388 p.
- Kaplan R, Norton D. Organização Orientada para a Estratégia. 8 ed. Rio de Janeiro, Elsevier, 2000.
- Tomsen, E. OLAP: Construindo sistemas de informações multidimensional. 2 ed. Rio de Janeiro, Campus, 2002.
- Albright, et al: Data Analysis e Decision Making. 2 ed. CA, USA, Brooks/Cole, 2003.
- Turban, E. Decision Support and Expert Systems. 4 ed. New Jersey, USA, 1988.

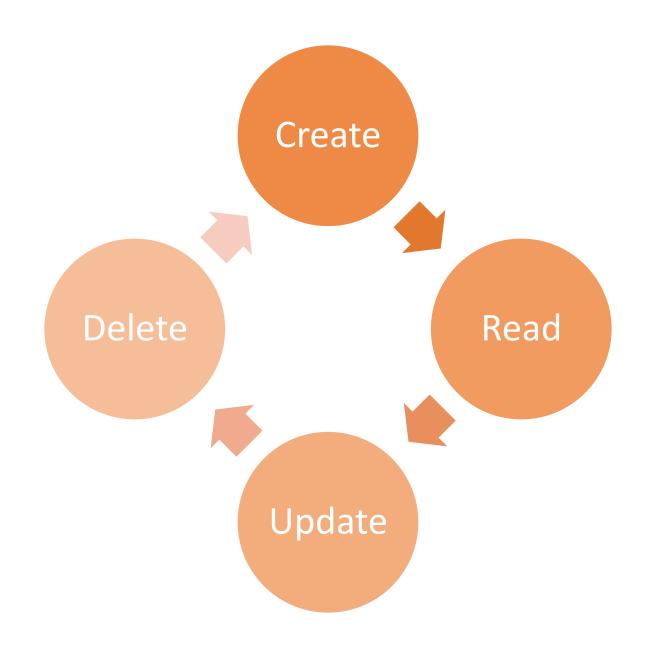
Relatório - Ortomagic - Situação Atual

- Inexiste na Ortomagic uma área forma de Governança de Dados, apesar de iniciativas da TI sobre arquitetura de dados.
- Não há uma estratégia e processo de gestão de dados, sem a presença de um gestor responsável para a governança e curadoria de dados
- O Modelo de dados corporativo (sistema de gestão) é administrado pela área de TI, utilizando-se de ferramenta Case para a criação dos modelos conceituais e administração do ambiente.
- A estrutura de DBA e suporte desenvolveu processos para o gerenciamento de banco de dados, por meio de documentação especifica o que traz uma segurança em relação a administração, modelagem e desenvolvimento de dados.

Relatório - Ortomagic - Situação Atual

- Na ortomagic, não há um processo de gestão de documento, pois as faturas são impressas e não há um processo de gestão automatizada desses documentos.
- As terminologias (tesauros) existem em visões departamentais, sem a centralização de uma documentação corporativa
- O processo de qualidade de dados inexiste na Ortomagic, as figuras e papeis de curadoria precisam ser estabelecidos em conjunto com o estabelecimento dos processos e políticas.

MATRIZ CRUD



MATRIZ RACI

