PROVAS DE TI
TUDO PARA VOCÊ PASSAR

Governança de Dados

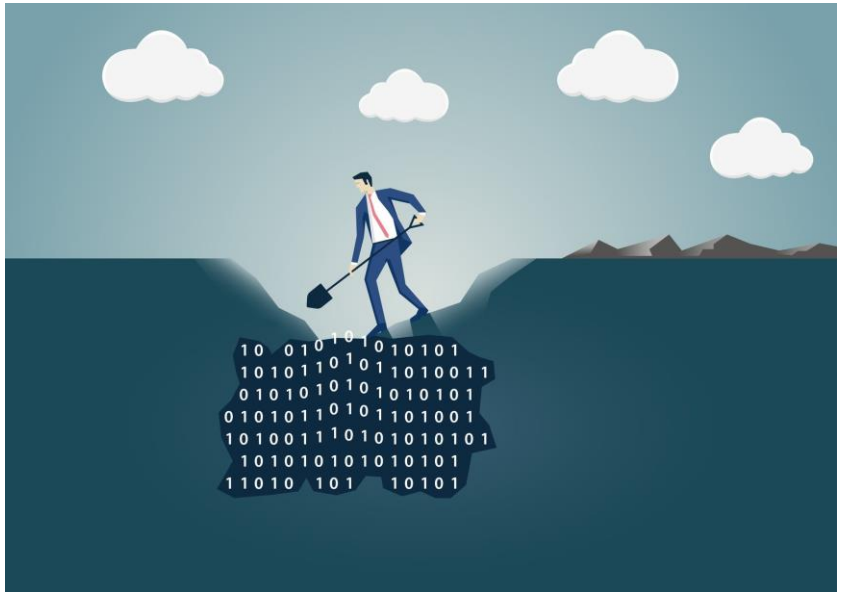
Prof. Rodrigo Macedo

Escopo do Curso

- Governança de Dados (GovData)
- Arquitetura de Dados (ArqData)
- Qualidade de Dados (QuaData)
- Técnicas de Qualidade de Dados
- Crisp DM
- Questões de concursos

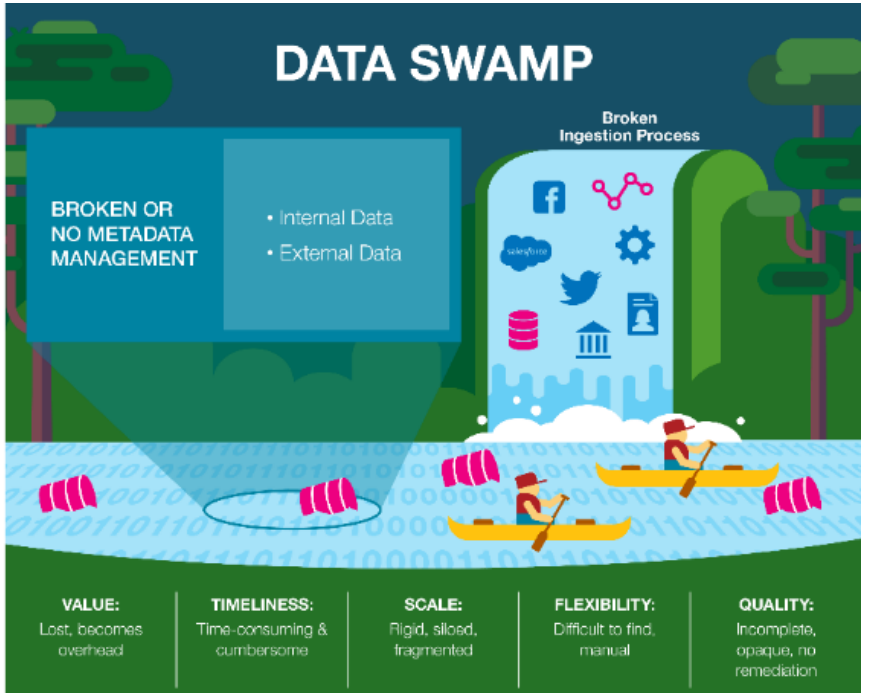
Motivação

- Atualmente, muito se ouve dizer que ‘os dados são o novo petróleo’. A frase, de forma geral, evidencia a importância das informações digitais e digitalizadas no século atual.
- A gestão desses dados é um fator chave para determinar o sucesso ou a dificuldade de implementações das novas funcionalidades, de produtos e serviços ou, em alguns casos, prejuízos para a própria empresa.



Motivação - Data Lake x Data Swamp

- Um data lake ou lago de dados é um sistema ou repositório de dados armazenados em seu formato natural / bruto, geralmente objetos blobs ou arquivos - ênfase em dados não estruturados.
- Um data swamp ou pântano de dados é um lago de dados deteriorado e não gerenciado, inacessível aos usuários pretendidos ou que fornece pouco valor.



Conceitos DataGov

- A governança de dados pode ser definida como uma disciplina que apoia o gerenciamento de dados corporativos.
- Uma de suas funções é alinhar pessoas, processos e tecnologias sob a ótica dos dados. Nesse sentido, o objetivo é determinar papéis, responsabilidades e projetos necessários para a devida gestão das informações estratégicas que transitam por determinada empresa.
- Em muitos casos, a a governança de dados atue como uma autoridade articuladora a fim de estabelecer diretrizes da gestão de dados, liderar iniciativas de melhorias e orquestrar todo esse trabalho.

Conceitos DataGov

- Data Governance geralmente inclui outros conceitos, como Data Quality e outros, para ajudar uma empresa a obter melhor controle sobre seus ativos de dados, incluindo métodos, tecnologias e comportamentos relacionados ao gerenciamento adequado dos dados.
- Em outras palavras, trata-se de uma disciplina-mãe que gere várias outras disciplinas e age como uma espécie de guardiã das informações. Essa estrutura serve de referência para todas as áreas da empresa em caso de necessidades desse tipo. É seu papel, por exemplo, estipular requisitos a respeito dos dados coletados – e que podem ser muitos.
- Também lida com segurança e privacidade, integridade, usabilidade, integração, conformidade, disponibilidade, funções e responsabilidades e gerenciamento geral dos fluxos de dados internos e externos dentro de uma organização.

Conceitos DataGov

DataGov - Importância

- Governança de Dados permite validar, qualificar, distribuir, organizar e armazenar as informações da organização de maneira precisa, ágil e eficiente visando algumas características, como:

1. A utilização de ferramentas para análise e cruzamento de dados para geração de informações para a tomada de decisões.
2. O favorecimento da desburocratização por meio de acesso centralizado a informações de governo para simplificar a oferta de serviços públicos.
3. A ampliação da transparência permitindo a análise de contas públicas para combater fraudes.
4. A adoção de tecnologia de ponta no processamento de grande volume de dados com rápido tempo de resposta.

DataGov - Importância

1. A viabilização da segurança e garantia de sigilo e individualização das bases de dados.
2. A alavancagem da economicidade pelo uso compartilhado de infraestrutura e do consumo de dados para redução de custos.

- Os principais benefícios que uma boa Governança de Dados, pode trazer para as organizações são:

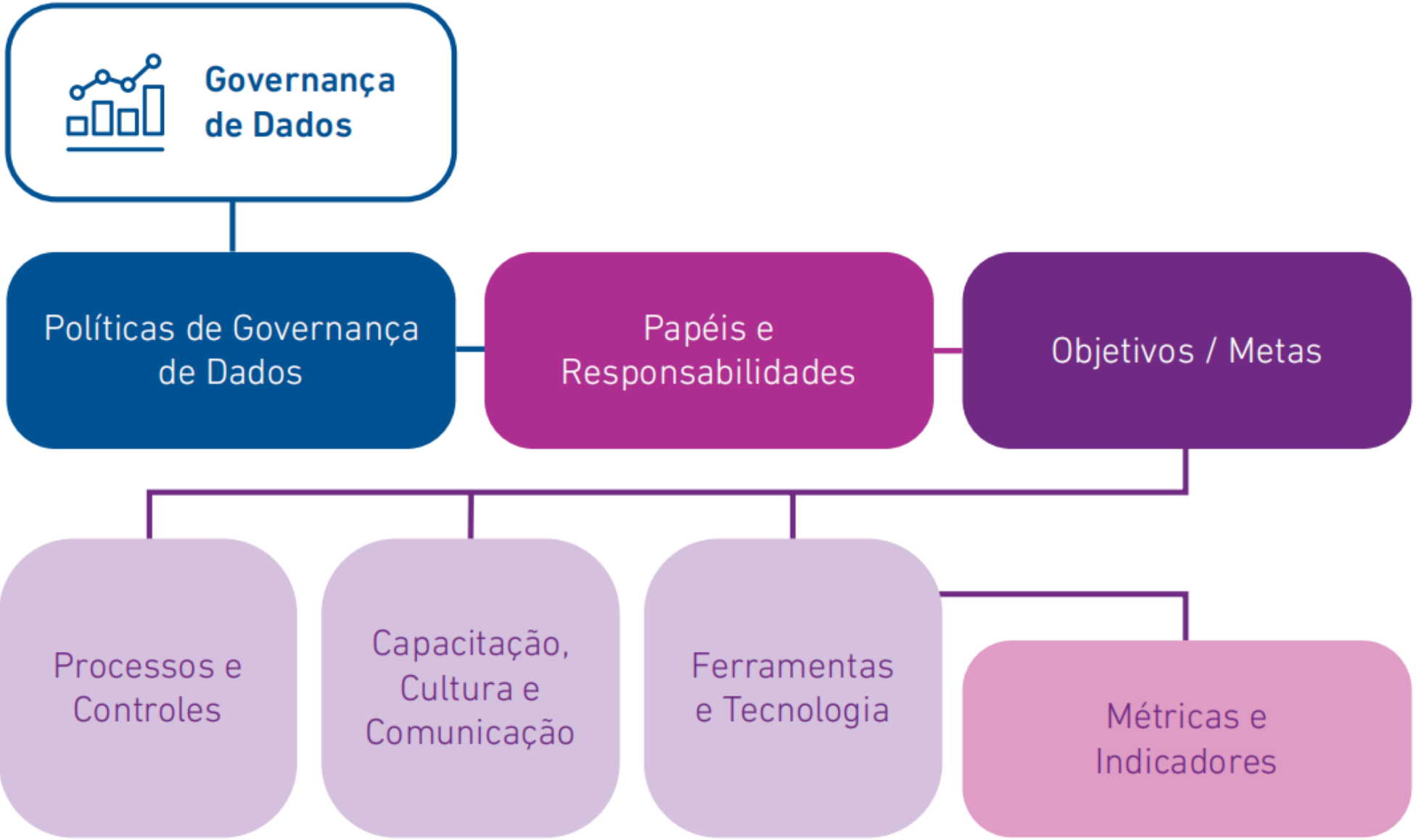
1. Mudança de cultura: dados e informações passam a ser reconhecidos como importantes ativos estratégicos nas organizações.
2. Melhor alinhamento entre as áreas de Tecnologia da Informação e Comunicação (TIC) e de Negócio: esse alinhamento é premissa fundamental para o bom funcionamento da Governança de Dados. Com isso, outras áreas como a de mapeamento de processos e a de desenvolvimento de sistemas podem se beneficiar de alinhamentos já iniciados.

DataGov - Importância

1. A gestão das operações de captura, armazenamento, proteção, planejamento, controle e garantia da qualidade dos ativos de dados é centralizada em uma única estrutura, permitindo a redução de custos e a otimização do uso dos recursos.
2. Redução da quantidade de informações redundantes.
3. Reutilização de dados corporativos e/ou compartilhados, por meio do gerenciamento de dados mestre e dados de referência.
4. Conhecimento de dados e informações utilizados por meio da adoção de um vocabulário único sobre as definições dos dados: ampliação e melhoria da disseminação do conhecimento entre as pessoas – passagem do capital intelectual para o capital estrutural.

DataGov - Importância

DataGov - Princípios



DataGov - Princípios

1. **Políticas de Governança de Dados:** As políticas oferecem a **visão** e as **diretrizes** sobre conceitos, regras, responsabilidades, restrições e premissas relacionadas aos processos de dados. Elas orientam os funcionários da empresa sobre as iniciativas na condução de atividades de Gestão de Dados. É nas políticas que estão definidas **as responsabilidades, as circunstâncias, as aprovações e/ou validações.**

- Em empresas menores, é recomendado que a política cubra todo o tratamento de dados, observando sempre os pontos mais críticos da Lei, como tratamento de dados de menores de idade, dados pessoais sensíveis, entre outros.
- Já nas empresas com mais colaboradores e/ou com número maior de processos estruturados, é importante avaliar a elaboração de políticas específicas para tópicos que podem ser mais críticos, como a captação ou aquisição de dados externos, o armazenamento e o descarte dos dados,

DataGov - Princípios

1. **Objetivos e Metas:** A Governança de Dados deve garantir que a empresa atinja os objetivos estabelecidos para os seus dados, o que trará, como consequência, maiores chances de a empresa alcançar suas metas estratégicas e de negócios. Para isso, os responsáveis devem impulsionar o **desenvolvimento** de **Processos** e **Controles** que garantam a execução dos pontos observados nas políticas. Os processos dependerão de **Capacitação**, **Cultura** e **Comunicação** para serem claramente entendidos, assimilados, difundidos e executados, e devem fazer uso de **Ferramentas** e **Tecnologias** capazes de auxiliar tais tarefas.

DataGov - Princípios

1. **Capacitação, Cultura e Comunicação:** A comunicação é fundamental em um programa de governança sólido e funcional, e deve ser uma missão permanente por parte dos times envolvidos. Ela é importante para disseminar os conceitos definidos no programa de Governança de Dados aos funcionários da organização. A capacitação das pessoas envolvidas no uso de dados, o entendimento do papel e das responsabilidades de cada função no ciclo de vida do dado na organização e a disseminação da cultura de governança de dados são atividades inter-relacionadas.

DataGov - Princípios

1. **Ferramentas e Tecnologia:** As ferramentas tecnológicas ajudam a apoiar a execução das atividades diárias da Governança, o armazenamento e o controle das informações relacionadas e, em alguns casos, ajudam até a executar os fluxos de trabalho que relacionam as pessoas com o programa de governança de dados.

- Uma ferramenta indispensável é um Inventário ou **Catálogo de Dados**, que ajuda a aumentar o conhecimento da empresa a respeito dos dados sob seu poder.
- **Inventário/Catálogo de Dados:** ferramenta em que ficam documentados todos os dados da organização. Eles podem ser de cunho técnico (nome da tabela, tipo de dado por coluna da tabela, quantidade de colunas da tabela, origem do dado, tipo de ingestão, entre outros).

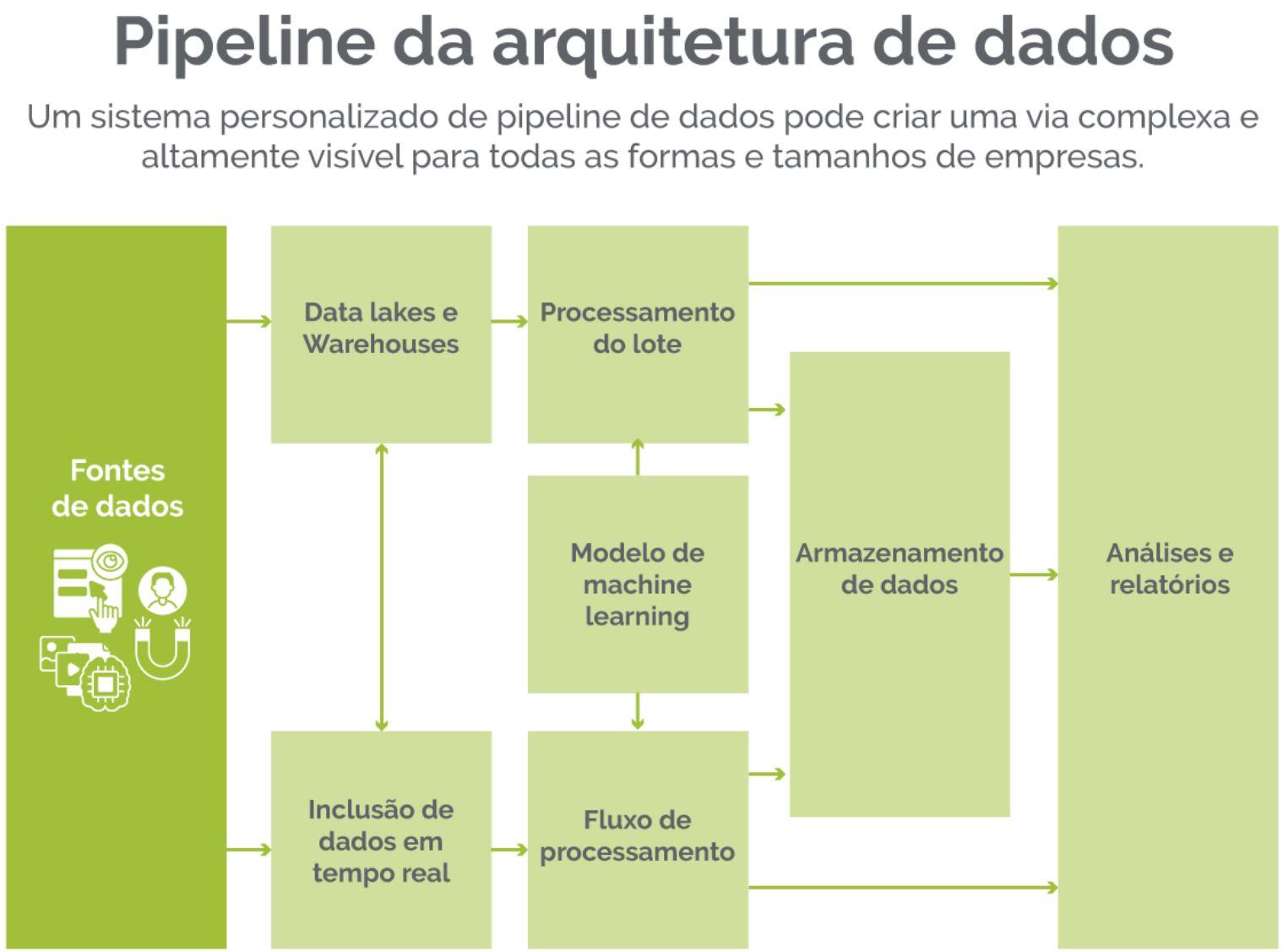
ArqData- Conceitos

- Arquitetura de dados é a estrutura dos componentes de dados de uma organização - considerados sob diferentes níveis de abstração, suas inter-relações, bem como os princípios, diretrizes, normas e padrões que regem seu projeto e evolução ao longo do tempo.
- Envolve o processo de gerenciamento dos ativos informacionais e o projeto de dados usado para definir uma determinada situação futura, incluindo o subsequente planejamento necessário para alcançar tal estado.
- Arquitetura dos dados é a maneira que as empresas decidem fazer a organização dos seus ativos digitais.

ArqData- Conceitos

- As informações coletadas precisam ser organizadas do mesmo modo que as pessoas costumam organizar os seus utensílios pessoais, para que sejam úteis para os profissionais da gestão. A informação em formato eletrônico precisa ser acessada com segurança.
- Essas informações devem ser usadas de forma inteligível e no tempo apropriado, para que gerem resultados satisfatórios.
- A arquitetura de dados está presente na formação de uma visão estratégica para os negócios, tendo em vista que a sua organização e planejamento mantêm os processos alinhados com os objetivos da empresa.

ArqData- Conceitos



ArqData- Benefícios

1. **Disponibilidade:** Disponibilidade é o termo que resume melhor a importância da arquitetura dos dados no contexto empresarial. Por meio dela, as funções dos sistemas e softwares estarão disponíveis quando se fizer necessário. Ou, na pior das hipóteses, poderão ser acessadas em tempo hábil, garantindo processos decisórios menos sujeitos ao acaso.
2. **Controle no compartilhamento dos dados:** Em certas empresas, o regime de trabalho colaborativo e em rede exige que dados, formulários, planilhas e arquivos sejam compartilhados de forma controlada. Nesse aspecto, é preciso que a estrutura de dados da companhia seja, ao mesmo tempo, fácil de ser acessada e segura, a ponto de impedir que pessoas não autorizadas tenham privilégios indevidos.

ArqData- Benefícios

1. **Desempenho:** Todo software depende do acesso ao seu respectivo banco de dados para funcionar bem. Quando há falhas nessa função, todo o sistema colapsa e se torna inútil. Dessa forma, a arquitetura dos dados é a melhor garantia de que seus sistemas estarão sempre prontos para dar respostas quando forem exigidas.
2. **Facilidade de consulta:** terá amplitude de dados e acesso das informações por todos os setores. Terá o acesso aos indicadores relevantes mais facilmente, se optar, assim como os demais dados.
3. **Auxilia a tomada de decisões:** esse é um dos principais ganhos quando se realiza análise de dados com as informações arquitetadas. A arquitetura de dados traz tudo que uma organização necessita para ter esse grande volume de informações estruturado.

ArqData- Princípios

- Como qualquer segmento ligado à tecnologia digital, a arquitetura dos dados está em constante evolução.
- Diante disso, uma arquitetura de dados só traz resultados quando se orienta por seis princípios:

1. **Segurança:** Pelos bons princípios de governança de dados, todo sistema deve ser seguro o bastante para que as informações em uma companhia sejam acessíveis apenas às pessoas autorizadas.Dessa forma, a arquitetura deve observar mecanismos de proteção contra acessos indevidos ou invasores, franqueando os dados críticos somente àqueles que forem credenciados para isso.

ArqData- Princípios

1. **Flexibilidade:** Uma boa arquiteura, deve ser moldado conforme as necessidades. Afinal, na transformação digital, é preciso que a arquitetura dos dados tenha certa elasticidade, permitindo que os sistemas evoluam e sejam escaláveis.Há casos, ainda, em que é necessário conceder novos acessos ou autorizações de uso não previstas.Então, quanto mais ela antecipar essas demandas, melhor.
2. **Colaboração:** O modelo de gestão horizontal é cada vez mais uma tendência. Por isso, as empresas precisam de soluções que permitam gerir, acessar e tratar seus dados por múltiplos times e colaboradores.
3. **Orientada para resultados:** Para ter maior efetividade, uma arquitetura deve ser orientada pelas metas do negócio.

ArqData- Princípios

1. **Inteligência:** Em uma arquitetura, espera-se que os dados possam ser usados para tomada de decisões. Esse é o princípio por trás do conceito de Business Intelligence (BI), pelo qual as atividades são pautadas por decisões tomadas a partir de dados estruturados. Logo, é tarefa dos profissionais de arquitetura de banco de dados garantir que a empresa terá à sua disposição não apenas dados em estado bruto, mas informação útil sempre que precisar.
2. **Automação:** Não dá para imaginar uma ferramenta digital que abra mão de processos automatizados. Por esse princípio, a arquitetura dos dados assume o compromisso de gerar soluções que sejam efetivas e em níveis máximos de automação.

ArqData- Atividades

As atividades de um arquiteto de dados envolvem uma série de responsabilidades associadas ao escopo da estratégia de dados de uma empresa que abrange uma combinação de plataformas **locais** e serviços de aplicativos e dados em **nuvem**.

- Descrever os padrões e princípios de dados que governam o gerenciamento de dados em ambientes de dados, incluindo locais híbridos e várias nuvens;
- Definir os tipos de estruturas de gerenciamento de dados a serem usados, incluindo RDBMSes para processamento transacional e operacional; data warehouses, data marts e data lakes para processamento analítico; e ferramentas de consulta e visualização de dados do usuário final;
- Considerar as demandas operacionais e as expectativas de desempenho, bem como os custos, e planejar uma estratégia para gerenciar dados e aplicativos, cada vez mais na nuvem;

ArqData- Atividades

- Documentar como os dados fluem de pontos de origem e aquisição em sistemas e aplicativos e supervisionar o desenvolvimento, gerenciamento e monitoramento de pipelines de dados;
- Descrever técnicas e processos de integração de dados e selecionar ferramentas para implementação e supervisão dos esforços de integração;
- Definir políticas de proteção de dados e selecionar as tecnologias certas para implementar as políticas;
- Monitorar, auditar e relatar a conformidade com os padrões de dados internos, regulamentos e políticas definidas externamente e expectativas de desempenho.

ArqData- Atividades

- Implementar um catálogo de dados para listar ativos de dados corporativos junto com suas características, onde esses ativos estão localizados, controles de acesso e classificação da sensibilidade dos dados;
- Supervisionar o uso de ferramentas e tecnologias de modelagem de dados, orientar os modeladores de dados no desenvolvimento de seus modelos, supervisionar os processos de modelagem de dados e manter um repositório de metadados para capturar “inteligência de dados” sobre o cenário de dados corporativos;
- Supervisionar a seleção e implementação de ferramentas de gerenciamento de dados que se alinham aos processos e metodologias de desenvolvimento;

QuaData- Conceitos

- O gerenciamento de qualidade de dados (DQM) refere-se ao conjunto de práticas de negócios que envolvem o emprego das pessoas, processos e tecnologias certos para obter insights acionáveis a partir das informações disponíveis.
- Uma estrutura de integração e qualidade de dados bem estabelecida garante que o fluxo do processo de qualidade de dados seja mantido durante todo o ciclo de vida dos dados.
- Por exemplo, como parte de um plano de gerenciamento de qualidade de dados corporativos, os usuários especificam certas verificações de qualidade de dados ao longo da jornada de dados para elimine quaisquer inconsistências ou erros e garantir dados confiáveis para processos analíticos e de inteligência de negócios.

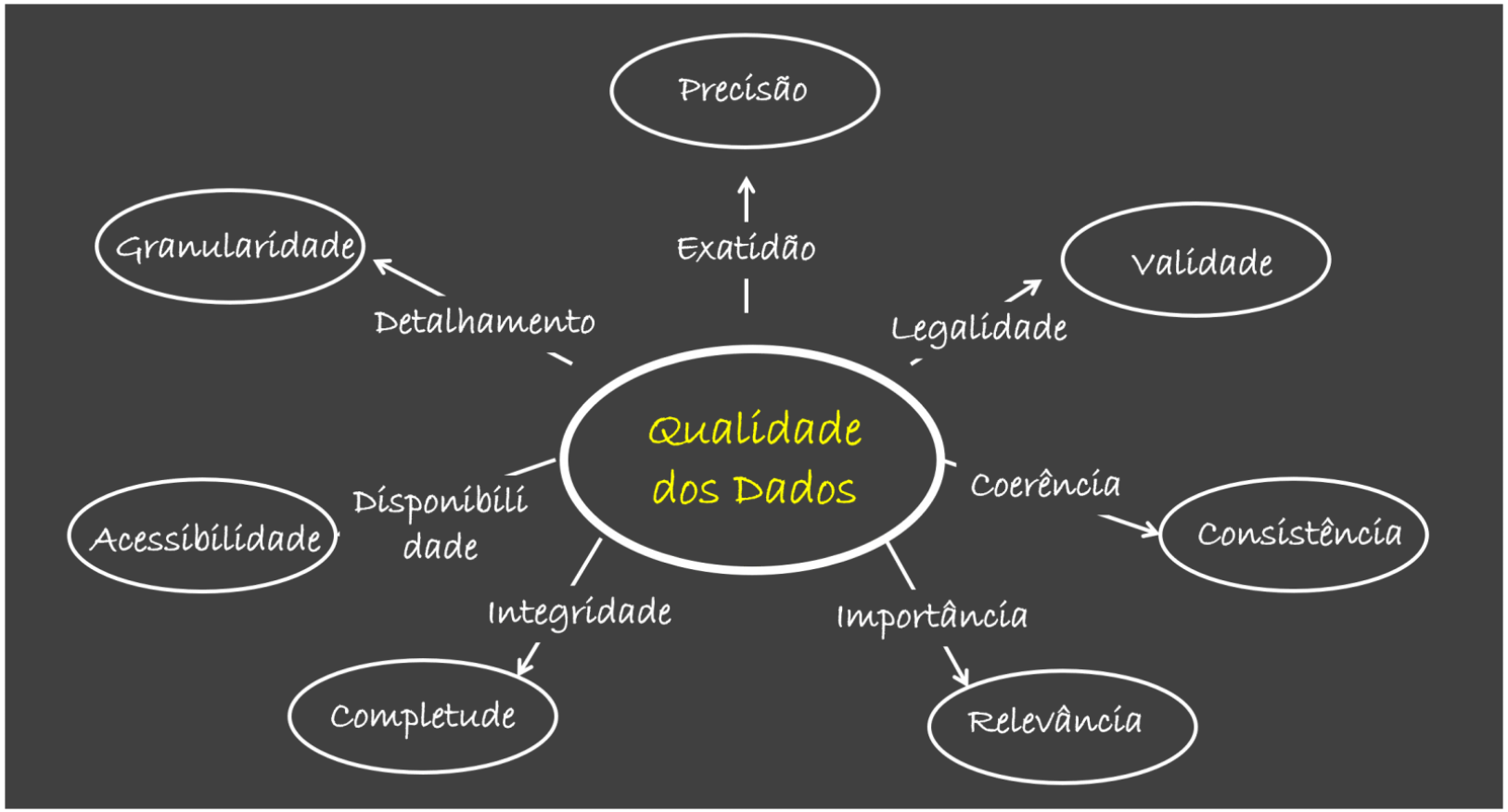
QuaData- Conceitos

- Em um sistema de informações, a qualidade dos dados é uma avaliação da precisão, atualização e consistência das informações disponíveis em uma base de dados.
- Grande parte dos problemas de qualidade dos dados em sistemas de informações empresariais está ligada a nomes digitados incorretamente, números trocados ou códigos faltantes e ocorre durante a entrada de dados.
- Esses erros ficam mais comuns quando as empresas transferem parte dos seus dados para internet e permite que clientes e fornecedores insiram seus dados no site, e isso efetue alterações no sistema interno.
- Nesses casos, pode ser aplicado **limpeza** e **padronização** que representa o processo de deletar e corrigir, dentro de um banco de dados as infomações incompletas, incorretas, com formatação inadequadas. O data cleansing não corrige apenas os dados, mas também reforça a consistência entre diferentes conjuntos de dados oriundos de sistema de informação independente

QuaData- Conceitos

QuaData- Características

- Ter um conjunto bem definido de métricas de avaliação de gerenciamento de qualidade de dados em vigor é vital para avaliar o desempenho das iniciativas de gerenciamento de qualidade de dados de uma empresa.



QuaData- Características

1. **plenitude:** indica se os dados coletados são suficientes para tirar conclusões. Isso pode ser avaliado assegurando que não haja informações ausentes em nenhum conjunto de dados.
2. **consistência:** garante que os dados em todos os sistemas de uma organização sejam sincronizados e reflitam as mesmas informações. Um exemplo de dados consistentes inclui o registro da data de remessa no mesmo formato de data da planilha de informações de um cliente.
3. **precisão:** implica se os dados que foram coletados representam com precisão o que deveriam. Isso pode ser medido contra dados de origem e validados contra regras de negócios definidas pelo usuário.

QuaData- Características

1. **singularidade:** envolve garantir que não haja duplicatas presentes nos dados. Por exemplo, a falta de dados exclusivos pode resultar no envio de vários emails para um único cliente devido a registros duplicados.
2. **validade:** mede se os dados atendem aos padrões ou critérios definidos pelo usuário de negócios. Por exemplo, uma empresa pode colocar uma verificação de qualidade de dados corporativos no campo de quantidade do pedido, ou seja, 'Quantidade do Pedido> = 0', pois a quantidade do pedido negativa implica em informações inválidas.
3. **disponibilidade:** é preciso poder acessar um dado de forma ágil sempre que necessário.

QuaData- Ferramentas

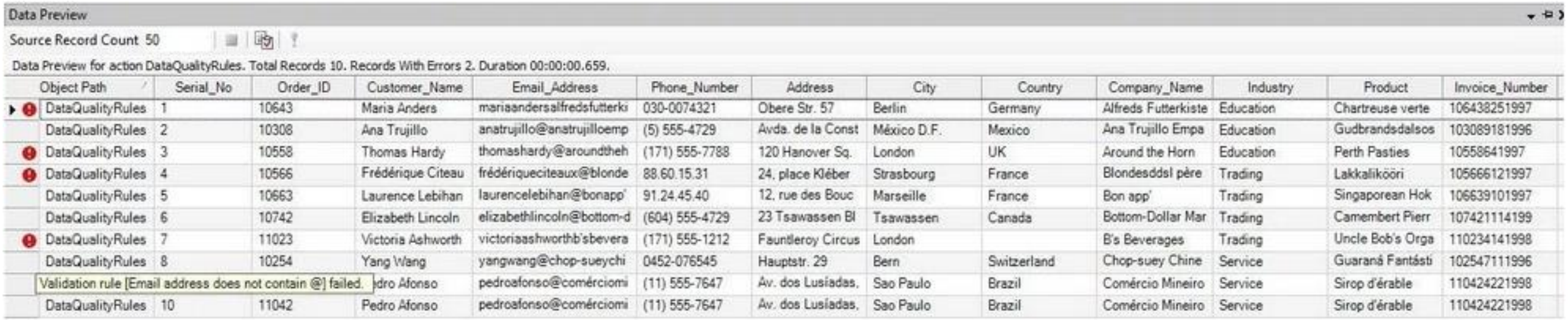
- As ferramentas de gerenciamento de qualidade de dados são tecnologias usadas para identificar, compreender e corrigir quaisquer falhas nos dados.
- Tais ferramentas oferecem suporte à tomada de decisões e aos processos de negócios para uma governança de dados eficiente.
- Alguns fatores importante para selecionar a ferramenta adequada:

1. **Perfil de dados e funcionalidade de limpeza:** Uma ferramenta eficaz de qualidade de dados deve incluir perfil de dados recursos. Além de ajudar a automatizar a identificação de metadados e fornece visibilidade clara dos dados de origem para identificar quaisquer discrepâncias e utilizar recursos de limpeza e resolvê-los antes que os dados sejam carregados em um destino

QuaData- Ferramentas

1. **Verificações de qualidade de dados:** É esperado que o software contenha objetos e regras integrados ao fluxo de informações para monitorar e relatar quaisquer erros que possam ocorrer durante o processamento de dados. Eles garantem que os dados sendo processados sejam validados com base em regras de negócios definidas para garantir a integridade dos dados.
2. **Gerenciamento de linhagem de dados:** auxilia no gerenciamento da linhagem de dados, que ajuda a controlar e analisar o fluxo de informações, descrevendo a origem dos dados e sua jornada, como as etapas em que os dados foram transformados ou gravados no destino.
3. **Conectividade com várias fontes de dados:** oferecer suporte para dados em qualquer formato e complexidade, sejam estruturados ou não estruturados, planos ou hierárquicos, legados ou modernos.

QuaData- Ferramentas


Data Preview

Source Record Count 50

Data Preview for action DataQualityRules. Total Records 10. Records With Errors 2. Duration 00:00:00.659.

Object Path	Serial_No	Order_ID	Customer_Name	Email_Address	Phone_Number	Address	City	Country	Company_Name	Industry	Product	Invoice_Number
DataQualityRules	1	10643	Maria Anders	mariaandersalfredsfutterki	030-0074321	Obere Str. 57	Berlin	Germany	Alfreds Futterkiste	Education	Chartreuse verte	106438251997
DataQualityRules	2	10308	Ana Trujillo	anatrujillo@anatrujilloemp	(5) 555-4729	Avda. de la Const	México D.F.	Mexico	Ana Trujillo Empa	Education	Gudbrandsdalsos	103089181996
DataQualityRules	3	10558	Thomas Hardy	thomashardy@aroundtheh	(171) 555-7788	120 Hanover Sq.	London	UK	Around the Horn	Education	Perth Pasties	10558641997
DataQualityRules	4	10566	Frédérique Citeau	frédériqueciteaux@blonde	88.60.15.31	24, place Kléber	Strasbourg	France	Blondesddsl père	Trading	Lakkalikööri	105666121997
DataQualityRules	5	10663	Laurence Lebihan	laurencelebihan@bonapp'	91.24.45.40	12, rue des Bouc	Marseille	France	Bon app'	Trading	Singaporean Hok	106639101997
DataQualityRules	6	10742	Elizabeth Lincoln	elizabethlincoln@bottom-d	(604) 555-4729	23 Tsawassen Bl	Tsawassen	Canada	Bottom-Dollar Mar	Trading	Camembert Pierr	107421114199
DataQualityRules	7	11023	Victoria Ashworth	victoriaashworthb'sbevera	(171) 555-1212	Fauntleroy Circus	London		B's Beverages	Trading	Uncle Bob's Orga	110234141998
DataQualityRules	8	10254	Yang Wang	yangwang@chop-sueychi	0452-076545	Hauptstr. 29	Bern	Switzerland	Chop-suey Chine	Service	Guaraná Fantásti	102547111996
Validation rule [Email address does not contain @] failed.			dro Afonso	pedroafonso@comérciomi	(11) 555-7647	Av. dos Lusíadas,	Sao Paulo	Brazil	Comércio Mineiro	Service	Sirop d'érable	110424221998
DataQualityRules	10	11042	Pedro Afonso	pedroafonso@comérciomi	(11) 555-7647	Av. dos Lusíadas,	Sao Paulo	Brazil	Comércio Mineiro	Service	Sirop d'érable	110424221998

Nessa imagem, vemos que um dos registros estava errado por causa do endereço de e-mail incorreto

QuaData- Técnicas

- A fim de conferir uma maior qualidade dos dados, podemos aplicar diversas técnicas, dentre elas:

1. Profiling.
2. Data Matching.
3. Deduplication.
4. Data Cleasing.
5. Data enrichment

QuaData- Técnicas

- **Data Profiling:** é uma atividade de Qualidade de Dados que é caracterizada pelo uso de técnicas analíticas sobre os dados para conhecer o seu conteúdo, estrutura e qualidade. Em outras palavras, é uma espécie de exame (diagnóstico) multidisciplinar a respeito dos dados existentes.
- Geralmente se divide em quatro métodos:

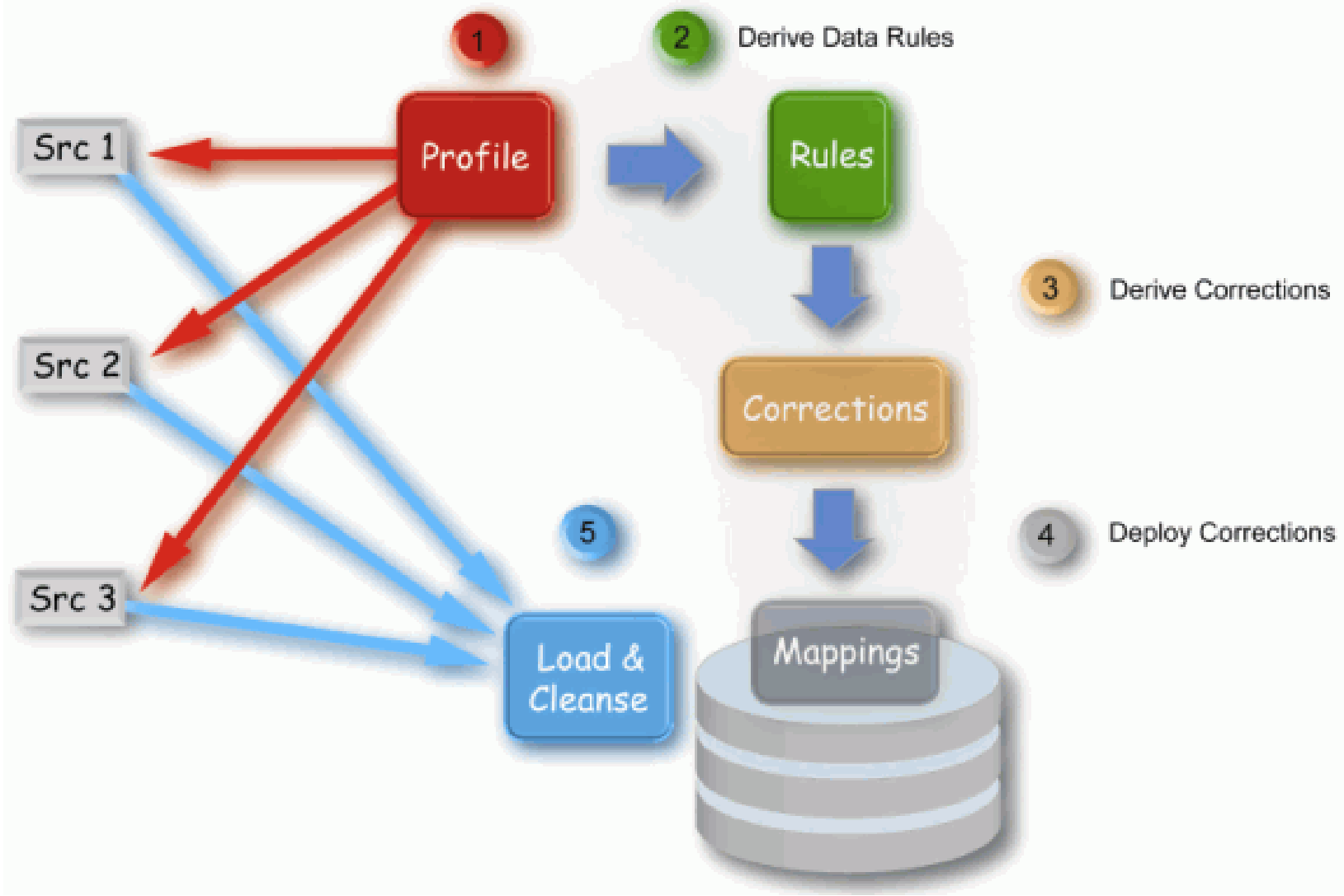
1. **Column profiling:** analisa uma tabela e conta o número de vezes que cada valor aparece em cada coluna. Uma utilização desse método é encontrar padrões e distribuição de frequência em uma coluna de dados.
2. **Validação de regra de dados:** utiliza a criação de perfil de dados de maneira proativa para verificar se as instâncias e conjuntos de dados estão em conformidade com as regras predefinidas.

QuaData- Técnicas

1. **Cross-column profiling:** esse método possui dois processos. A análise de chave examina coleções de valores de atributos, procurando uma possível chave primária. A análise de dependência é um processo mais complexo que determina se há relacionamentos ou estruturas incorporadas em um conjunto de dados. Esses métodos, combinados ou não, auxiliam a expor dependências entre atributos de dados na mesma tabela, facilitando o mapeamento de dados em um Data Lake.
2. **Cross-table profiling:** utiliza a análise de chave estrangeira, que é a identificação de registros órfãos e a determinação de diferenças semânticas e sintáticas, para examinar os relacionamentos dos conjuntos de colunas em diferentes tabelas. Isso pode ajudar a reduzir a redundância, mas também identificar conjuntos de valores de dados que podem ser mapeados juntos.

QuaData- Técnicas

- **Data Profiling:**



QuaData- Técnicas

- **Data Matching:** é a tarefa de encontrar, em bancos de dados, registros que se referem à mesma entidade. Normalmente, tais registros são oriundos de diferentes conjuntos de dados e não possuem nenhum identificador comum, mas também há casos em que estas técnicas são aplicadas para detectar registros duplicados em um único banco de dados.
- Identificar registros correspondentes em múltiplos bancos de dados é uma tarefa desafiadora por várias razões. Em primeiro lugar, tais registros normalmente não possuem nenhum atributo que deixe óbvio quais se referem às mesmas entidades, sendo necessário, portanto, analisar outros atributos que dão identificações parciais, tais como nomes e datas de nascimento (para pessoas) ou títulos e marcas (para produtos).

QuaData- Técnicas

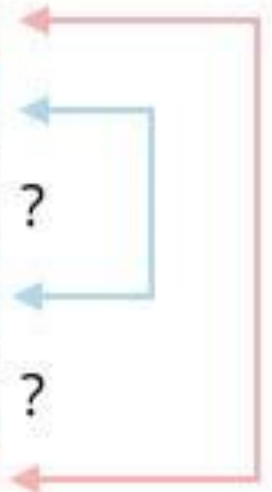
- **Data Matching:**

1. **Listas de email marketing**: muitas empresas utilizam listas de emails de clientes para anunciar seus produtos ou serviços, as quais comumente contém dados inconsistentes ou mesmo mais de um registro para alguns determinados clientes. Neste caso, técnicas de linkagem de registros podem ser usadas para identificar clientes duplicados e evitar que uma oferta seja enviada mais de uma vez para um mesmo cliente.
2. **Censo:** o governo coleta uma grande quantidade de informações sobre a população através de diferentes órgãos governamentais, os quais frequentemente utilizam padrões de dados distintos e armazenam as informações de maneira totalmente descentralizada. Relacionar estas bases de dados permite ao governo produzir relatórios estatísticos diversos e entender melhor muitos aspectos do país.

QuaData- Técnicas

- **Data Matching**

Name	Date of birth (DD/MM/YYYY)	Address	
Lisa Perry	11/05/1984	67, Brown St.	
Jack Davis	29/03/1958	789, Saxon St.	
Michael Lewis	*NULL*	2065, Salisburry Av.	?
Jack Davis Miller	29/03/1958	Saxon Street 789	
Sarah Boyle	26/04/1990	*NULL*	?
Lisa Terry	01/05/1984	67, Brown St.	
Sara Rachel Boyle	*NULL*	90 Gourdon Court	?



QuaData- Técnicas

- **Deduplicação de dados:** é uma espécie de conceito de compactação de dados que permite diminuir o volume de dados armazenados. O trabalho consiste em eliminar cópias de dados armazenados, ao invés de utilizar técnicas como a compactação em arquivos RAR ou ZIP.
- Duas abordagens sobre essa técnica:

1. **Deduplicação em nível de arquivo:** funciona checando se os mesmos objetos (arquivos) já estão armazenados.
2. **Deduplicação em nível de bloco:** usa a mesma abordagem que a deduplicação em nível de arquivo, mas aqui os objetos são blocos de dados.

QuaData- Técnicas

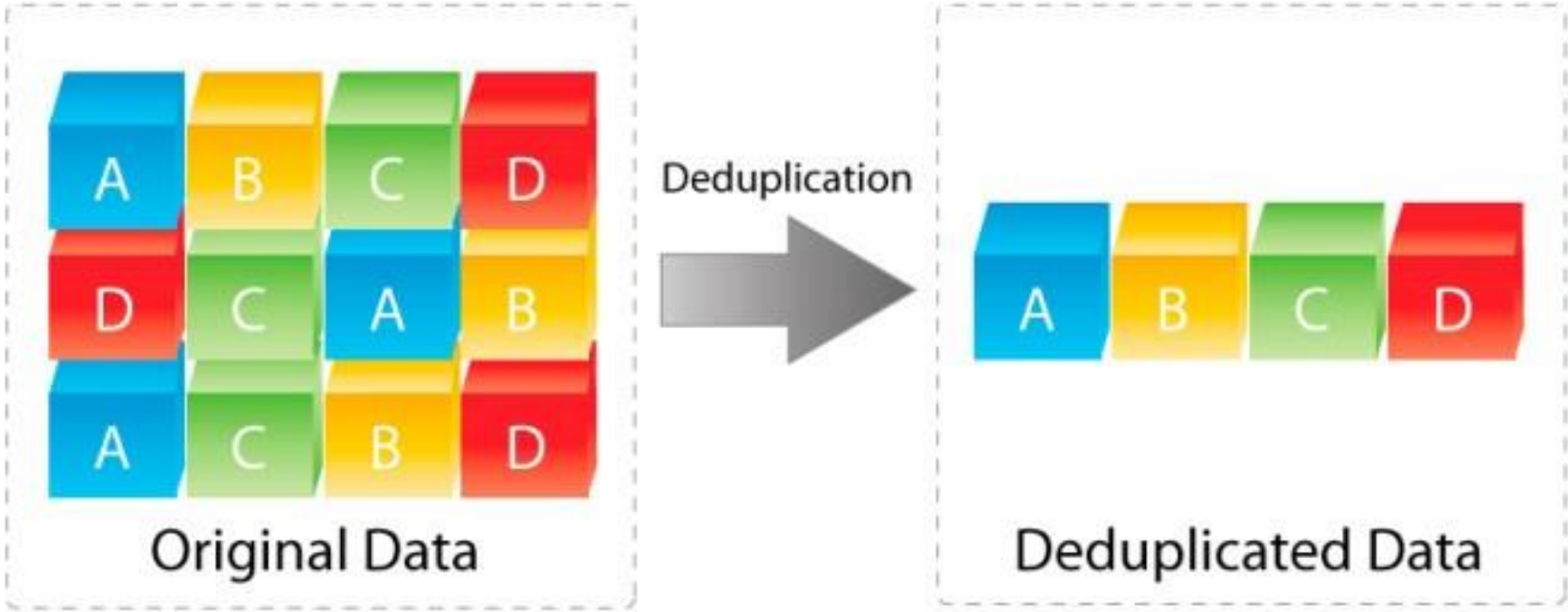
1. **Deduplicação em nível de arquivo:** permite ignorar o armazenamento de cópias de vários arquivos – elas são apenas substituídas pelo “link” do arquivo original. Através das “impressões digitais”, sequência de caracteres única em cada arquivo, dos objetos é checado se o arquivo já está colocado no armazenamento. Essa técnica de impressão digital geralmente é baseada em métodos de hashing ou atributos de arquivo (depende da solução de deduplicação utilizada). Em média, a deduplicação em nível de arquivo permite economizar até 80% em espaço de armazenamento. Tipos específicos de arquivos também influenciam a eficiência da dedução de dados: imagens ou arquivos de áudio tendem a ser únicos e não podem se beneficiar da deduplicação; já documentos, modelos e arquivos internos do sistema possuem uma boa taxa de deduplicação.

QuaData- Técnicas

1. **Deduplicação em nível de bloco:** A deduplicação em nível de bloco é mais profunda e verifica a exclusividade dos blocos de todos os arquivos. Quando um arquivo é modificado, o sistema armazena somente partes (blocos) alterados do arquivo original, como cada bloco tem sua própria identificação (normalmente gerada via algoritmo de hash), o sistema comparar com os metadados “já armazenados”. Essa abordagem permite economizar ainda mais espaço (a taxa de redução utilizando deduplicação em nível de bloco pode chegar a 95%), mas requer mais computação pois o número de objetos (blocos) a serem processados é muito maior.

QuaData- Técnicas

- **Deduplicação de dados**



QuaData- Técnicas

- **Limpeza de Dados:** Limpeza de dados é o processo de analisar a qualidade de dados em uma fonte de dados, aprovando/rejeitando as sugestões manualmente pelo sistema e fazer alterações assim aos dados. A limpeza de dados na Qualidade de dados inclui um processo auxiliado por computador que analisa a conformidade dos dados em relação ao conhecimento de uma base de dados de conhecimento, e um processo interativo que permite que o administrador de dados examine e modifique resultados de processo auxiliado por computador para garantir que a limpeza de dados seja executada exatamente como desejado.

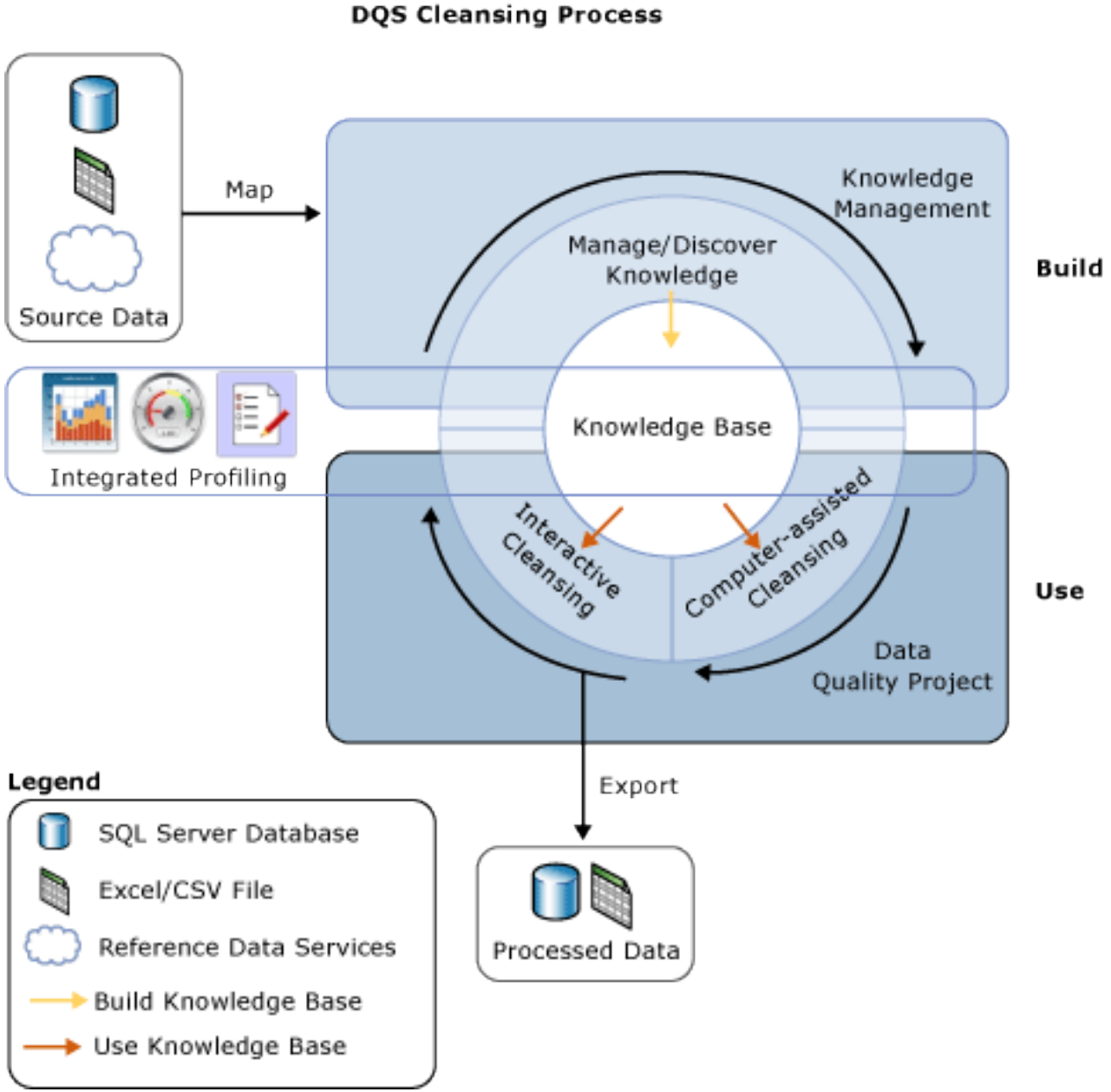
QuaData- Técnicas

- **Limpeza de Dados - Benefícios:**

1. Identifica dados incompletos ou incorretos em sua fonte de dados (arquivo do Excel ou banco de dados do SQL Server) e então corrige ou o alerta sobre os dados inválidos.
2. Oferece uma interface de assistente simples, intuitiva e consistente para que o usuário navegue pelos dados e inspecione erros em um conjunto muito grande de dados.
3. Oferece processo de duas etapas para limpar os dados: auxiliado por computador e interativo. O processo por computador usa o conhecimento em uma base de conhecimento de DQS para processar os dados automaticamente e sugere substituições/correções. A próxima etapa, interativa, permite que o administrador de dados aprove, rejeite ou modifique as alterações propostas pelo DQS durante a limpeza auxiliada por computador.

QuaData- Técnicas

- **Limpeza de Dados**



QuaData- Técnicas

- **Enriquecimento de Dados:** O enriquecimento de dados é o processo de combinar dados próprios de fontes internas com dados externos de outros sistemas internos ou dados de terceiros de fontes externas. Os dados do cliente começam na forma bruta, independentemente da fonte, seja o tráfego do site, mídia social ou listas de e-mail. Quando os dados do cliente são coletados, eles são armazenados em um armazenamento de dados central e são amplamente inúteis. Os dados brutos são limpos e estruturados, antes de serem enriquecidos com dados externos para adicionar informações úteis adicionais. O enriquecimento de dados torna os dados mais úteis, agregando valor a eles. Ajuda as marcas a compreender melhor seus clientes e obter percepções mais profundas de suas vidas. Existem muitas maneiras de enriquecer os dados. Um excelente exemplo de enriquecimento de dados seria enriquecer os dados de vendas internas com dados de anúncios de terceiros para obter uma melhor compreensão da eficácia dos anúncios.

QuaData- Técnicas

- **Enriquecimento de Dados - Benefícios:**

1. O enriquecimento de dados economiza dinheiro porque você não armazena informações que não são úteis para o seu negócio. Em vez disso, você aprimora os dados internos com fontes externas de dados para o benefício da sua organização.
2. Os dados enriquecidos promovem comunicações personalizadas e aumentam a probabilidade de relacionamentos significativos com o cliente e oportunidades de negócios. Com dados relevantes do cliente, sua empresa pode desenvolver estratégias de comunicação que atendam às preferências e necessidades do cliente.
3. Dados redundantes têm um custo significativo para a empresa. Isso resulta em perda de receita, perda de clientes e danos à reputação. Dados redundantes são comuns nas organizações porque elas não têm certeza dos dados a serem liberados e dos dados a serem mantidos.

QuaData- Técnicas

- **Enriquecimento de Dados**

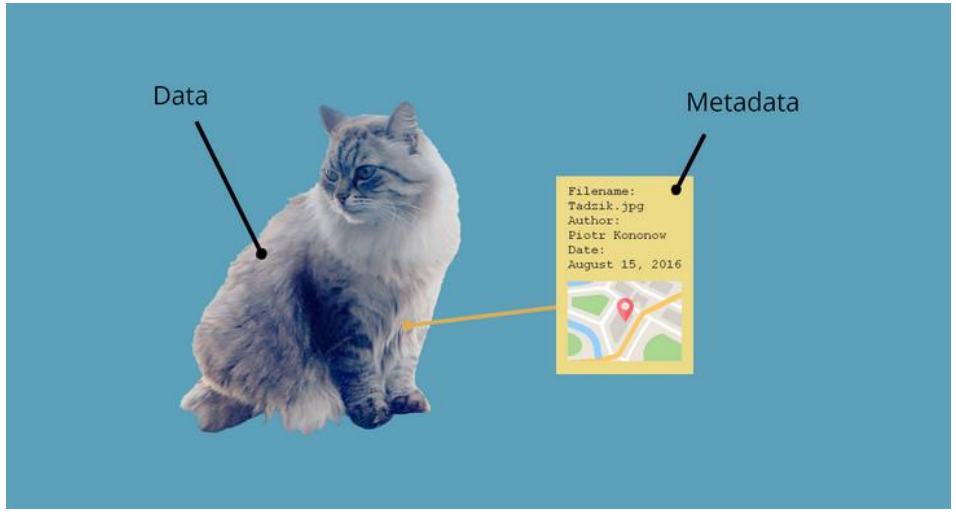
MetaDados

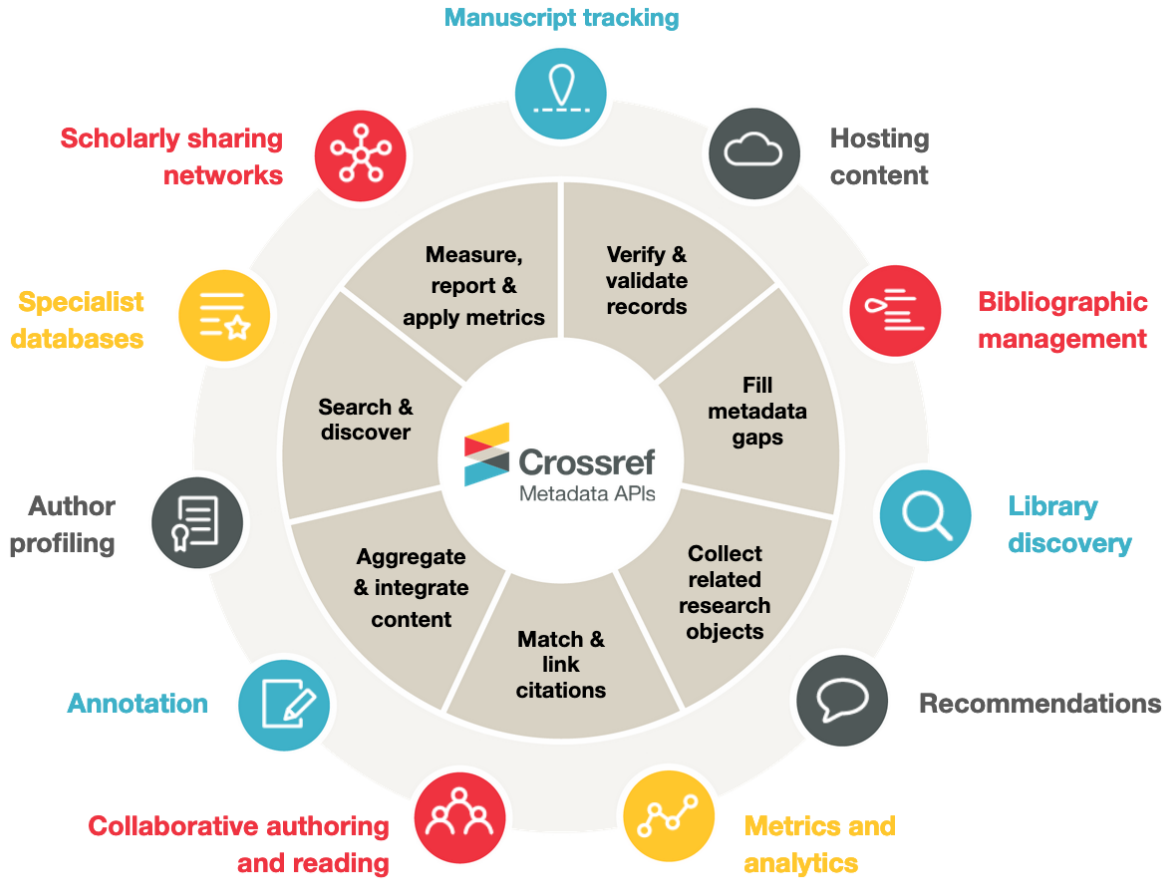
- Metadados, ou Metainformação, são dados sobre outros dados. Um item de um metadado pode dizer do que se trata aquele dado, geralmente uma informação inteligível por um computador.
- Os metadados facilitam o entendimento dos relacionamentos e a utilidade das informações dos dados.
- Metadados são indispensáveis para a comunicação entre computadores, mas podem ser inteligíveis também por humanos. Todos os dados descritivos de um documento, físico ou digital, sobre autor, data de criação, local de criação, conteúdo, forma, dimensões e outras informações são metadados.

MetaDados - Aplicações

- Web semântica, é uma web "inteligente", capaz de conceder um significado a um arquivo que será disponibilizado para outros utilizadores, podendo ser usado como fonte de pesquisa.
- A importância dos metadados para a websemântica está basicamente ligada à facilidade de recuperação dos dados, uma vez que estes terão um significado e um valor bem definidos. Nesse sentido, todos os documentos publicados na web devem ser catalogados.
- Aumento da segurança das informações, visando a Lei Geral de Proteção de Dados. Pois os metadados podem ser utilizados sem ferir políticas de privacidade;
- Melhora na análise de dados, pois a empresa será mais eficiente ao encontrar, capturar, cruzar e analisar os dados.

MetaDados


Data
Metadata
Filename:
Tadzik.jpg
Author:
Piotr Kononow
Date:
August 15, 2016


Manuscript tracking
Hosting content
Bibliographic management
Library discovery
Recommendations
Metrics and analytics
Collaborative authoring and reading
Annotation
Author profiling
Specialist databases
Scholarly sharing networks
Measure, report & apply metrics
Verify & validate records
Fill metadata gaps
Collect related research objects
Match & link citations
Aggregate & integrate content
Search & discover
Crossref
Metadata APIs

CRISP-DM

- É um modelo de processo de mineração de dados que descreve abordagens comumente usadas por especialistas em mineração de dados para atacar problemas.
- Possui 6 fases:

1. Entender o Negócio.
2. Entender os Dados.
3. Preparação dos Dados.
4. Modelagem.
5. Avaliação.
6. Implantação.

Entendimento do Negócio

- Foca em entender o objetivo do projeto a partir de uma perspectiva de negócios, definindo um plano preliminar para atingir os objetivos.

Essa fase fornece três artefatos de saída:

1. Background: explique a situação da empresa e como o projeto vai ser direcionado para solucionar o problema;
2. Objetivo do projeto: informe qual o objetivo maior que seu projeto tem;
3. Critério de sucesso: deixe bem claro qual será a métrica que ditará se seu projeto atingiu o sucesso ou não.

Entendimento dos Dados

- Como você deve saber bem, coletar e tratar o dado é uma tarefa responsável por mais de 70% do tempo gasto em um projeto por Data Scientists, e é exatamente sobre isso que essa fase e a próxima dizem respeito. Aqui, você deverá coletar, descrever — usando estatísticas — , explorar e verificar a qualidade do seu dado.

Preparação dos dados

- Essa etapa dá subsídios para a etapa posterior da modelagem. Possui quatro tarefas:

1. **Data Selection:** aqui você vai selecionar os dados que serão usados no modelo. Por exemplo, talvez você não queira usar outliers, ou todas as colunas da tabela. Escolha tudo que serão relevantes para seu modelo, e não esqueça de documentar o motivo de escolhe-los;
2. **Data Cleaning:** é bem provável que seu dado não virá da melhor forma possível. Datas em formato incorreto e números inteiros sendo interpretados como string, são só alguns dos exemplos de sujeira que vão ser encontrados no seu dado. É nessa hora que você irá tratá-los;
3. **Construct Data:** talvez nem todos os dados que você precise estará a sua disposição. É possível que você tenha que criar novos dados para seu modelo. Por exemplo, talvez você precise de um campo ou coluna no seu dado que diga se uma determinada data é feriado, ou qual dia da semana ela representa;
4. **Integrating Data:** essa tarefa é necessária quando você precisa juntar duas fontes de dados diferentes.

Modelagem

- É nesse momento que você irá realizar a construção do seu modelo. Essa fase consiste em escolher seu algoritmo — que pode ser desde Árvores de Decisão, até redes neurais — , criar o modelo em si e utilizar seus parâmetros. Você pode criar diferentes modelos e compará-los na próxima fase.

Avaliação

- Hora de avaliar os resultados de seu modelo. Se lembra dos critérios de sucesso que você definiu lá na primeira fase? Tá na hora de verificar se ela foi atingida. Se não foi, é necessário voltar a primeira fase e entender o que deu de errado, determinar um novo escopo e tentar novamente. Caso tudo dê certo, avance para a sexta e última fase.

Implantação

- Nesta fase final é hora de colocar seu modelo em produção, para que possa ser usado. O deployment coloca fim ao seu projeto, mas lembre-se de sempre monitorar os resultados e adaptar o modelo sempre que necessário.
- Pode envolver o trabalho de um profissional da área de Engenharia de Software nessa etapa.

Q1) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

A acurácia na qualidade de dados está diretamente associada à de duplicação, uma vez que indica que há exclusividade da fonte de dados e de suas entidades, de forma a garantir a precisão dos dados na vida real.

Q1) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

A acurácia na qualidade de dados está diretamente associada à de duplicação, uma vez que indica que há exclusividade da fonte de dados e de suas entidades, de forma a garantir a precisão dos dados na vida real.

Q2) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

O gerenciamento de qualidade de dados inclui a definição de padrões e métricas sobre os dados, porém dispensa o gerenciamento do ciclo de vida desses dados.

Q3) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

Definir processos de qualidade para modelos de dados implica analisar as regras de negócios fundamentais dos processos, bem como observar a qualidade dos dados, com a finalidade de garantir a conformidade da regra com o processo.

Q4) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

A fim de gerenciar a qualidade de dados, o MDM se serve de ferramentas de validação dos dados que ajudam na visualização de todo o fluxo de gestão dos dados mestres, o que torna possível, de maneira rápida, a identificação de quaisquer desvios em relação à política de dados da empresa.

Q2) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

O gerenciamento de qualidade de dados inclui a definição de padrões e métricas sobre os dados, porém dispensa o gerenciamento do ciclo de vida desses dados. ERRADO.

Q3) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

Definir processos de qualidade para modelos de dados implica analisar as regras de negócios fundamentais dos processos, bem como observar a qualidade dos dados, com a finalidade de garantir a conformidade da regra com o processo. CERTO.

Q4) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

A fim de gerenciar a qualidade de dados, o MDM se serve de ferramentas de validação dos dados que ajudam na visualização de todo o fluxo de gestão dos dados mestres, o que torna possível, de maneira rápida, a identificação de quaisquer desvios em relação à política de dados da empresa. ERRADO.

Q5) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

Uma boa prática em relação à qualidade de dados é verificar a precisão dos dados, validando-se se estão corretamente representados, e a sua consistência, avaliando-se se há integridade cruzada entre duas ou mais fontes que armazenem o mesmo dado.

Q6) [CESPE TCE RJ 2021] Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

Na primeira fase do CRISP-DM (cross industry standard process for data mining), há o entendimento dos dados para que se analise a qualidade destes.

Q5) [CESPE SERPRO 2021] Julgue o próximo item, relativo à qualidade de dados.

Uma boa prática em relação à qualidade de dados é verificar a precisão dos dados, validando-se se estão corretamente representados, e a sua consistência, avaliando-se se há integridade cruzada entre duas ou mais fontes que armazenem o mesmo dado. CERTO.

Q6) [CESPE TCE RJ 2021] Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

Na primeira fase do CRISP-DM (cross industry standard process for data mining), há o entendimento dos dados para que se analise a qualidade destes. ERRADO.

Q7) [CESPE TCE RJ 2021] Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

Na primeira fase do CRISP-DM (cross industry standard process for data mining), há o entendimento dos dados para que se analise a qualidade destes.

Q8) [CESPE TCE PE 2017] Julgue o seguinte item, que se refere a CRISP-DM (Cross-Industry Standard Process of Data Mining).

Durante a fase de entendimento do negócio, busca-se descrever claramente o problema, fazer a identificação dos dados e verificar se as variáveis relevantes para o projeto não são interdependentes.

Q9) [CESPE TCE RJ 2021] Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

A fase de implantação do CRISP-DM (cross industry standard process for data mining) só deve ocorrer após a avaliação do modelo construído para atingir os objetivos do negócio.

Q7) [CESPE TCE RJ 2021] Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

Na primeira fase do CRISP-DM (cross industry standard process for data mining), há o entendimento dos dados para que se analise a qualidade destes. ERRADO.

Q8) [CESPE TCE PE 2017] Julgue o seguinte item, que se refere a CRISP-DM (Cross-Industry Standard Process of Data Mining).

Durante a fase de entendimento do negócio, busca-se descrever claramente o problema, fazer a identificação dos dados e verificar se as variáveis relevantes para o projeto não são interdependentes. ERRADO.

Q9) [CESPE TCE RJ 2021] Com relação a noções de mineração de dados e Big Data, julgue o item que se segue.

A fase de implantação do CRISP-DM (cross industry standard process for data mining) só deve ocorrer após a avaliação do modelo construído para atingir os objetivos do negócio. CERTO.

Q10) [CESPE Ministerio Economia 2020] Julgue o seguinte item, a respeito de big data.

A etapa de modelagem do modelo CRISP-DM permite a aplicação de diversas técnicas de mineração sobre os dados selecionados, conforme os formatos dos próprios dados.

Q11) [CESPE TCE PA 2016] Julgue o item subsequente, acerca de segurança da informação de um SGBD e de um BI (Business Intelligence).

CRISP-DM é uma metodologia proprietária que identifica as fases Business Understanding e Data Understanding na implantação de um projeto de data mining.

Q12) [CESPE SERPRO 2021] Acerca do CRISP-DM, julgue a assertiva a seguir, como certa ou errada.

Para o atendimento à necessidade I, deve-se implantar a CRISP-DM, cujas etapas são Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Aplicação.

Q10) [CESPE Ministerio Economia 2020] Julgue o seguinte item, a respeito de big data.

A etapa de modelagem do modelo CRISP-DM permite a aplicação de diversas técnicas de mineração sobre os dados selecionados, conforme os formatos dos próprios dados.
CERTO.

Q11) [CESPE TCE PA 2016] Julgue o item subsequente, acerca de segurança da informação de um SGBD e de um BI (Business Intelligence).

CRISP-DM é uma metodologia proprietária que identifica as fases Business Understanding e Data Understanding na implantação de um projeto de data mining.
ERRADO.

Q12) [CESPE SERPRO 2021] Acerca do CRISP-DM, julgue a assertiva a seguir, como certa ou errada.

Para o atendimento à necessidade I, deve-se implantar a CRISP-DM, cujas etapas são Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados,

Q13) [CESPE Ministerio da Economia 2020] No que se refere à mineração de dados, julgue o item a seguir.

No modelo CRISP-DM, a fase na qual se planejam todas as atividades para carga dos dados é denominada entendimento dos dados.

Q14) [CESPE TCM-BA 2018] Assinale a opção correta a respeito do CRISP-DM.

a) CRISP-DM é uma suíte de ferramentas proprietárias que vem se tornando um padrão da indústria para mineração de dados, uma vez que fornece um plano completo e tecnologias para a realização de um projeto de mineração de dados.

b) A verificação da qualidade dos dados é uma atividade da fase de entendimento dos dados.

c) Durante a fase de preparação dos dados, é realizado um inventário de requisitos, suposições e restrições de recursos.

d) Na fase de avaliação dos dados, são realizadas as atividades de identificar valores especiais dos dados e catalogar seu significado.

e) Na fase de preparação dos dados, são realizadas as atividades de analisar o potencial de implantação de cada resultado e estimar o potencial de melhoria do processo atual.

Q13) [CESPE Ministerio da Economia 2020] No que se refere à mineração de dados, julgue o item a seguir.

No modelo CRISP-DM, a fase na qual se planejam todas as atividades para carga dos dados é denominada entendimento dos dados. ERRADO.

Q14) [CESPE TCM-BA 2018] Assinale a opção correta a respeito do CRISP-DM.

a) CRISP-DM é uma suíte de ferramentas proprietárias que vem se tornando um padrão da indústria para mineração de dados, uma vez que fornece um plano completo e tecnologias para a realização de um projeto de mineração de dados.

b) A verificação da qualidade dos dados é uma atividade da fase de entendimento dos dados.

c) Durante a fase de preparação dos dados, é realizado um inventário de requisitos, suposições e restrições de recursos.

d) Na fase de avaliação dos dados, são realizadas as atividades de identificar valores especiais dos dados e catalogar seu significado.

e) Na fase de preparação dos dados, são realizadas as atividades de analisar o potencial de implantação de cada resultado e estimar o potencial de melhoria do processo atual.