Livro Eletrônico

Estratégia
CONCURSOS

Aula 01

Business Intelligence para Concursos - Curso Regular 2018

Professor: Thiago Rodrigues Cavalcanti

AULA 00: Conceitos básicos de Dados, Informação, conhecimento e inteligência.

Sumário

Apresentação do professor

Olá senhoras e senhores,

Sejam bem-vindos a mais um curso de **business intelligence**! Hoje começamos mais uma versão do mais completo curso de business intelligence para concursos públicos. Esse curso, deve cobrir ainda, no conjunto das suas aulas, todo o conteúdo de modelagem multidimensional, ETL, mineração de dados presentes no conteúdo de concurso públicos

Como eu sempre digo é um prazer imenso fazer parte desta equipe de professores do Estratégia Concursos e ter a oportunidade de apresentar um pouco do meu conhecimento e experiência em concursos públicos!

Gostaria, antes de começar de fato o conteúdo teórico desta aula, de me apresentar de forma rápida. Meu nome é Thiago, sou casado, pernambucano, tenho um filho de sete anos. Torço pelo Sport Clube do Recife. Sou cristão. Frequento a IPN – Igreja Presbiteriana Nacional. Me formei em Ciência da Computação pela UFPE. Tenho mestrado em engenharia de software na mesma instituição. Atualmente faço doutorado em economia na UnB.

Frequento academia para manter a forma, mas meu hobby mesmo é pedalar! Decidi vender o carro e viver num desafio intermodal de transporte. Ia para o trabalho de *bike* sempre que possível! Ultimamente tenho usado mais Uber/Cabify do que a magrela, mais isso é um detalhe! A pergunta é: onde eu trabalho? No Banco Central do Brasil!

Fruto de uma trajetória de dois anos de estudos diários. Aposentei as canetas em 2010. Hoje estou de licença do Banco Central para fazer doutorado que começou em março de 2017. Antes de me licenciar eu trabalhava com análise e modelagem de dados.

Minha mais recente experiência com dados, seja na administração ou modelagem, é parte de uma estratégia profissional de alinhar meu trabalho diário como servidor público com minha carreira paralela de professor e consultor de Banco de Dados (BD) e *Business Intelligence* (BI). A ideia é conseguir me especializar cada vez mais no tema, desta nova carreira dentro da TI, que o mercado está denominando de **cientista dos dados (*Data scientist*)**.

Entrei neste universo como professor de concurso há alguns anos. Desde 2012, tenho me dedicado especificamente ao conteúdo de BD e BI. Minhas experiências em cursos presencias aqui em Brasília e em diversas partes do Brasil, bem como na gravação sistemática de aulas on-line me ajudaram a desenvolver um conteúdo exclusivo para os alunos do Estratégia Concursos.

A ideia é desenvolver um material completo, recheado de questões e com diversas dicas para ajudar você no seu objetivo: **ser aprovado e nomeado**!

Para finalizar, não deixe de seguir minha página no Facebook® (profthiagocavalcanti), onde eu publico, sistematicamente, questões comentadas e dicas semanais. Tenho também uma conta no Instagram, lá eu posto motivações e dicas rápidas a respeito do conteúdo de banco de dados e análise de informações. Agora que você já me conhece! Vamos seguir em frente com o nosso curso!

Motivação para o curso

Preparar esse curso é um desafio! Consolidar de forma amigável o conhecimento de business Intelligence para concursos não é uma tarefa fácil! Calibrar o nível do teórico e associar isso a uma didática eficiente tem sido minha meta nos últimos anos. Separamos o conteúdo de forma a segmentar e impulsionar seu aprendizado. Para que você entre na primeira aula com um pouco mais de segurança, vou aproveitar para fazer uma rápida apresentação sobre o assunto.

Você já ouviu falar sobre **Data Science ou ciência dos dados**? É um conceito relativamente recente que agrupa diversas atividades executadas sobre um conjunto de dados. Em especial sobre grandes conjuntos de dados. Para analisar os dados eles precisam estar armazenados e organizados de maneira convenientes para os cientistas dos dados.

Cientistas de dados são uma nova geração de especialistas em análise que têm habilidades técnicas para resolver problemas complexos e a curiosidade de explorar quais são os problemas que precisam ser resolvidos. A solução desses problemas passa por analisar os dados presentes em um banco de dados. Neste curso veremos o passo-a-passo para construção de um banco de dados.

Começaremos pelo básico do básico, analisando uma arquitetura de suporte a decisão e seus componentes. Passaremos por uma conceituação dos elementos presentes na gestão do conhecimento de uma empresa. Neste cenário aproveitaremos para introduzir os diferentes tipos de dados e as características dos mesmos, em especial à forma de armazenamento.

Veremos que existem profissionais dedicados a tarefas específicas. Veremos ainda que a construção de um banco de dados analítico, em especial um data warehouse, passa por algumas etapas bem definidas. Essas etapas criam modelos de dados ou esquemas que permitem um melhor entendimento da estrutura de dados da organização.

Todos esses conceitos serão vistos em detalhes nas próximas páginas. Ao final, teremos nossa tradicional lista de exercícios. Espero conseguir contribuir para a sua aprovação. Vamos em frente?!

Teremos muito trabalho! Por isso, montamos um **curso teórico em PDF**, baseado nas mais diversas bancas apresentando o conteúdo observando as variadas formas de cobrança do mesmo pelas bancas examinadoras.

Teremos ainda videoaulas que apresentam o conteúdo teórico de forma detalhada para algumas partes da matéria. Existe uma força tarefa para gravação de todo o assunto, mas não temos como garantir o término deste trabalho até a data de publicação das aulas. Mas não se preocupe, nosso objetivo é garantir que você tenha capacidade e conhecimento para ser aprovado. Logo, todo conteúdo necessário para a prova estará presente nos PDFs.

Existe ainda a previsão de um curso extensivo com o conteúdo deste curso. Deveremos transmitir várias aulas ao vivo no canal do Estratégia Concursos no YouTube. Fiquem atentos! Eu sempre publico o cronograma no meu perfil do Instagram.

Agora vamos voltar para a nossa aula. Vamos juntos?

Observação importante: este curso é protegido por direitos autorais (copyright), nos termos da Lei 9.610/98, que altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.

Grupos de rateio e pirataria são clandestinos, violam a lei e prejudicam os professores que elaboram os cursos. Valorize o trabalho de nossa equipe adquirindo os cursos honestamente através do site Estratégia Concursos ;-)

Observação importante II: todo o conteúdo deste curso encontra-se completo em nossos textos escritos. As videoaulas, caso existam, visam reforçar o aprendizado, especialmente para aqueles que possuem maior facilidade de aprendizado com vídeos e/ou querem ter mais uma opção para o aprendizado.

Cronograma

Para proporcionar uma visão geral do assunto e fornecer uma linha de ação para o estudo da matéria dividimos o curso em **oito** aulas, sendo esta a aula 01. A aula engloba a parte introdutória da matéria de dados e banco de dados. As demais aulas, seguindo a ementa do curso, são apresentadas abaixo e estão distribuídas como se segue:

EMENTA DO CURSO: Conceitos básicos de Dados, Informação, conhecimento e inteligência. Introdução a Business Intelligence e Data warehouse. ETL - Extração, transformação e Carga. OLAP - On-Line Analitical Processing. Mineração de Dados. Big Data, NoSQL e Hadoop. Gestão de conteúdo (ECM/BPM) e qualidade de dados. ERP – Enterprise Resource Planning

Pois bem, e como serão distribuídas as nossas aulas?

Aula	Conteúdo
Aula 01	Conceitos básicos de Dados, Informação, conhecimento e inteligência.
Aula 02	Introdução a Business Intelligence e Data warehouse
Aula 03	ETL - Extração, transformação e Carga
Aula 04	OLAP - On-Line Analitical Processing
Aula 05	Mineração de Dados
Aula 06	Big Data, NoSQL e Hadoop
Aula 07	Gestão de conteúdo (ECM/BPM) e qualidade de dados
Aula 08	ERP – Enterprise Resource Planning

Definido o cronograma, vamos partir para o conteúdo da nossa aula demonstrativa.

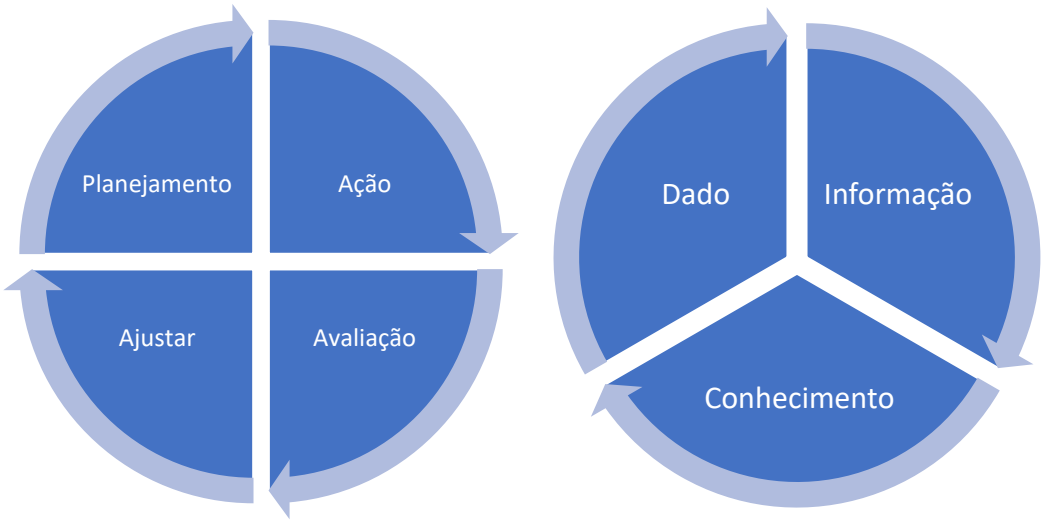
Fundamentos de Gestão do Conhecimento

1.Introdução a gestão do conhecimento

Nas próximas páginas, vamos fazer uma introdução ao assunto de gestão do conhecimento. Esse assunto faz parte da motivação para a utilização dos sistemas de inteligência de negócio que veremos em seguida (nas próximas aulas do nosso curso). A gestão do conhecimento tem por **objetivo fornecer ou aperfeiçoar a capacidade intelectual** das organizações. Não se trata apenas da criação de uma base de dados central que contenha a descrição das experiências e informações da empresa.

Gestão do conhecimento **pressupõe a adoção de um conjunto de tecnologias e processos** cujo objetivo é apoiar a criação, transferência e aplicação do conhecimento nas organizações. A utilização de bancos de dados internos e externos, de informações dos parceiros e de métricas para quantificar a eficiência da gestão são partes relevantes neste contexto.

Em outras palavras, a gestão do conhecimento é um conjunto de atividades relacionadas a identificação e gestão do aprendizado organizacional. Ela permite as empresas ou órgãos públicos entenderem e organizarem suas estruturas de **informações** e **conhecimento**. Como todo processo de **gestão** devemos ter etapas de planejamento, organização, avaliação e controle (PDCA). A figura abaixo apresenta essa definição de forma organizada.



Para entendermos esse fluxo de forma mais consistente precisamos começar pela formalização de alguns conceitos básicos: **dados, informação,**

conhecimento e inteligência. Esse será nosso objetivo a partir de agora. Faltaremos também de dados estruturados e não estruturados. Vem comigo!

2. Dado, informação, conhecimento e inteligência.

Nosso conteúdo tem início em uma citação: "São múltiplos os desafios de um mundo globalizado, sendo difícil até mesmo obter consenso sobre quais seriam os prioritários. Todavia, um deles – **tornar o conhecimento acessível a todos** – destaca-se como um dos mais importantes devido aos seus efeitos estruturantes".

É com esse texto que se inicia a apresentação do livro Inteligência, Informação e Conhecimento, uma coleção de artigos sobre gestão do conhecimento organizada em 2006 pela UNESCO. Para entender este assunto é preciso conhecer os conceitos básicos que vão subsidiar nosso entendimento, são eles: dados, informação, conhecimento e inteligência.

Dados podem ser definidos como sucessões de fatos brutos, que não foram organizados, processados, relacionados, avaliados ou interpretados, representando apenas partes isoladas de eventos, situações ou ocorrências. Constituem as unidades básicas a partir das quais informações poderão ser elaboradas ou obtidas.

O dado é a forma primitiva que compõe os sistemas de informação. Podemos considerar os dados como a menor partícula estruturada que compõe uma informação. Quando os dados passam por algum tipo relacionamento, avaliação, interpretação ou organização tem-se a geração de **informação**. A partir do momento em que dados são transformados em informações, decisões podem ser tomadas.

A **informação** pode ser definida como um dado acrescido de **contexto, relevância e propósito**. Requer um mínimo de análise para sua obtenção, através de uma avaliação humana. Podemos entender melhor essa ideia usando uma concatenação de dados. "De", "34", "Thiago", "=" e "idade" são dados quando observados individualmente. Porém, o agrupamento dos termos acima de forma lógica nos leva a uma informação: "Idade de Thiago = 34".

A informação é gerada a partir de uma interpretação sobre os dados, estes podem ser contextualizados, categorizados, calculados ou condensados. São fatos sobre uma situação, pessoa ou evento. Podemos ainda transformar os dados em informação com significado filtrando, ordenando, estruturando. Percebam que não existe conclusão, mas a organização dos dados neste momento permite que algum tipo de análise possa ser executado.

RESUMINDO

Para transforma dados em informações precisamos que os mesmos sejam:

(1) precisos e oportunos

(2) específicos e organizados para um propósito

(3) apresentados dentro de um contexto que lhe dê significado e relevância, e

(4) que podem levar a um aumento na compreensão e diminuição da incerteza.

A informação é valiosa porque pode afetar o comportamento, uma decisão ou um resultado. Agora vamos passar as informações por um processo de validação, com foco naquilo que queremos obter. Ao receberem o tratamento adequado a partir de critérios inicialmente definidos, sob ponto de vista estratégico e com um caráter informativo obtemos o **conhecimento**.

Conhecimento é uma informação contextual, relevante e acionável. Segundo **Turban**, conhecimento é a informação em ação. Capital intelectual (ou recursos intelectuais) é outro termo frequentemente usado como sinônimo de conhecimento. Conhecimento é uma informação valiosa da mente, inclui **reflexão, síntese e contexto**. É difícil de estruturar, difícil de capturar em computadores, normalmente é tácito (não explícito) e sua transparência é complexa.

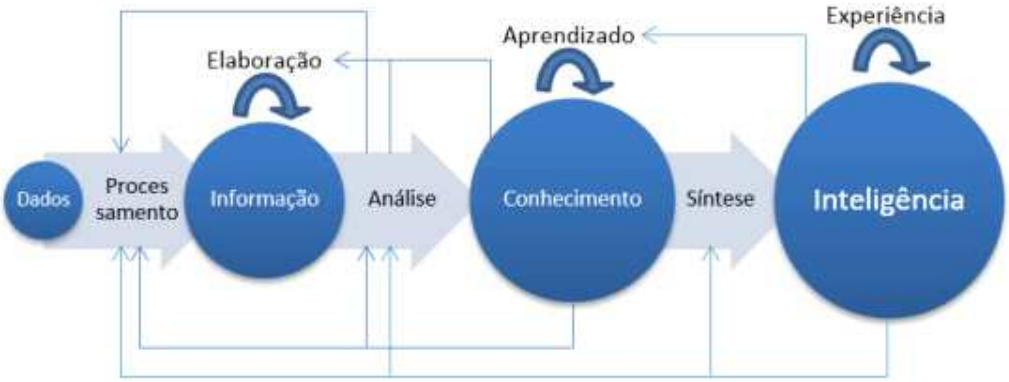
Veja abaixo um quadro comparativo entre conhecimento, informação e dados:

ESCLARECENDO

Dados	Informação	Conhecimento
Simples observações sobre o estado do mundo	Dados dotados de relevância e propósito	Informação valiosa da mente humana. Inclui reflexão, síntese, contexto
▪ Facilmente estruturado ▪ Facilmente obtido por máquinas ▪ Freqüentemente quantificado ▪ Facilmente transferível	▪ Requer unidade de análise ▪ Exige consenso em relação ao significado ▪ Exige necessariamente a mediação humana	▪ De difícil estruturação ▪ De difícil captura em máquinas ▪ Freqüentemente tácito ▪ De difícil transferência

Faltou apresentar a inteligência! Para exibi-la de forma contextualizada e aproveitar para expor algumas características do processo quem tem inicio na aglomeração de dados brutos vamos apresentar uma figura abaixo e em seguida tecer alguns comentários sobre a mesma.

TOME NOTA!



Dados compreendem a classe mais baixa de informação e incluem os itens que representam fatos, textos, gráficos, imagens estáticas, sons, segmentos de vídeo analógicos ou digitais etc. Os dados são coletados, por meio de processos organizacionais, nos ambientes interno e externo. Em suma, dados são sinais que não foram processados, correlacionados, integrados, avaliados ou interpretados de qualquer forma. Esta classe representa a matéria-prima a ser utilizada na produção de informações.

A próxima classe é a da informação propriamente dita. Nesta, os dados passam por algum tipo de processamento para serem exibidos em uma forma inteligível às pessoas que irão utilizá-los. Processar dados inclui a revelação de fotografias de um filme, as transmissões de rádio transformadas em um formato de relatório padronizado, a exibição de arquivos de computador como texto ou gráfico em uma tela, a grade de coordenadas em um mapa etc. O processo de transformação envolve a aplicação de procedimentos que incluem formatação, tradução, fusão, impressão e assim por diante. A maior parte deste processo **pode ser executada automaticamente**.

Uma vez que dados tenham sido transformados em informações, pelo menos em uma interpretação inicial, é possível refinar as informações mediante um processo de **elaboração**. As informações resultantes deste processo incluem características adicionais do problema, geram hipóteses, consequências das hipóteses, sugerem soluções para problemas, explanação e justificativas de sugestões, crítica de argumentos, entre outras saídas. Portanto, a transformação de dados em informações deve ser vista simplificadamente, como um tipo de pré-processamento de um processo de elaboração.

O próximo nível é o do **conhecimento**, que pode ser definido como sendo informações que foram analisadas e avaliadas sobre a sua **confiabilidade**, sua **relevância** e sua **importância**. Neste caso, o conhecimento é obtido pela

interpretação e integração de vários dados e informações para iniciar a construção de um quadro de situação.

O processo de transformação é realizado por meio de avaliação de dados e informações. Os insumos provenientes das diversas fontes são analisados e combinados na síntese de um produto final, o conhecimento. É por meio do conhecimento que **aqueles que assessoram as decisões buscam uma compreensão mais efetiva da situação problema**.

O conhecimento **não é estático**, modificando-se mediante **a interação com o ambiente**, sendo este processo denominado **aprendizado**. Uma visão mais ampla é que o aprendizado é a integração de novas informações em estruturas de conhecimento, de modo a torná-las potencialmente utilizáveis em processos futuros de processamento e de elaboração. Além disto, conhecimentos novos podem resultar de um processo de inferência na própria estrutura do conhecimento.

O nível mais alto desta hierarquia é **a inteligência**, que pode ser entendida como sendo a informação como oportunidade, ou seja, **o conhecimento contextualmente relevante que permite atuar com vantagem no ambiente considerado**. Também pode ser vista como o conhecimento que foi sintetizado e aplicado a uma determinada situação, para ganhar maior profundidade de consciência da mesma.

Portanto, a inteligência resulta da síntese de corpos de conhecimentos, são usados julgamento e intuição daquele que toma decisões e uma visualização completa da situação é obtida. Nesta etapa os tomadores de decisão aplicam ao conhecimento gerado, suas habilidades, suas competências de negócio e vivência na organização, para identificar direções estratégicas, tais como: novos projetos de pesquisa, acordos de cooperação, transferência de tecnologia e ações e reações da concorrência.

Idealmente, o entendimento da situação apoia a tomada de decisão a partir da visualização do cenário e cria as condições para que o planejamento possa ser realizado e as ações efetivadas. Além disso, podem ser revelados fatores críticos, possibilitando a antecipação a eventos, mediante o reconhecimento das consequências de novos ou iminentes efeitos de uma decisão. Por tudo isso, **a inteligência** deve ser a base do processo decisório, mesmo considerando que raramente é possível alcançar a compreensão total.

A transformação de conhecimento em inteligência é realizada por meio de síntese, sendo uma habilidade puramente humana baseada na experiência e intuição, que vai muito além da capacidade de qualquer sistema especialista ou de inteligência artificial. Síntese simplesmente não pode ser reduzida a procedimentos ou regras, por não considerarem o complexo.

Por fim, a **experiência** pode ser definida como **a efetividade da inteligência de uma organização**, que é aperfeiçoada pelas decisões tomadas e considerada geradora de algum tipo de vantagem. Sendo assim, a experiência agrega valor ao processo decisório de uma organização, por refletir toda a capacidade em atuar no ambiente competitivo.

2.1. Conhecimento tácito e explícito

Observando os conceitos acima apresentados é possível prever que uma empresa pode gerir de forma eficiente e informatizada o conhecimento. Devemos, portanto, viabilizar a conversão de conhecimento para em seguida geri-lo de maneira adequada.

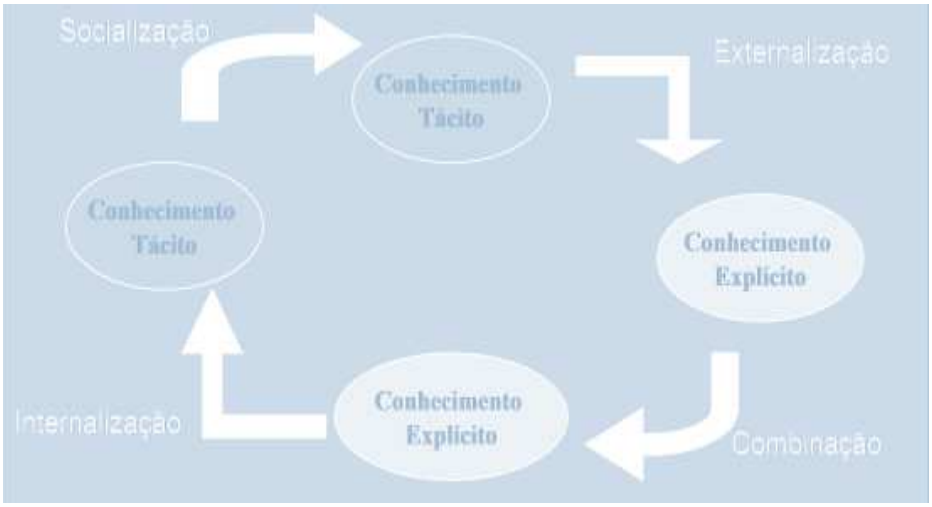
O conhecimento pode ser dividido em **tácito**, armazenado nas pessoas e adquirido através de experiências de vida, e **explícito** ou codificado, encontrado na forma de texto, formalizado e pode ser facilmente transmitido através do um meio físico. Desta forma, há quatro formas básicas pelas quais o conhecimento pode ser transmitido:

1º **Socialização** que é a interação do conhecimento através da troca de ideias e do compartilhamento de experiências, por meio de observação ou/e da prática.

2º **Externalização** neste caso o conhecimento pessoal passa para o domínio público por meio de uma documentação. Este pode ser reutilizado por outro, portanto deve ser classificado.

3º **Internalização** adquirido a partir de leitura. Vem de alguma publicação escrita que pode estar disponível em livro, revistas, na Internet ...

4º **Combinação** neste caso o conhecimento explícito de diferentes fontes é agrupado ou misturado visando agregar valor para gerar um novo conhecimento.



Vejam a figura acima as formas de conversão do conhecimento. Agora que já temos o conhecimento a respeito do assunto, vamos fazer uma rápida questão do CESPE para fixarmos o conteúdo.

1. Ano: 2016 Banca: CESPE Órgão: TCE-SC Cargo: Auditor de TI

Julgue os itens a seguir, acerca de dado, informação, conhecimento e inteligência.

90 Define-se informação como significado, ou seja, como registros icônicos e simbólicos — fonéticos ou numéricos — e signos — linguísticos, lógicos ou matemáticos —, por meio dos quais se representam atos, conceitos ou instruções.

91 O atributo de inteligência depende mais da qualidade da informação disponível do que da sua quantidade, tendo, portanto, natureza qualitativa.

Comentário: Vamos analisar cada uma das alternativas acima.

[90] Vejamos uma definição formal de informação: Informação é todo o conjunto de dados devidamente ordenados e organizados de forma a terem significado. Dados por sua vez pode ser definido como uma representação dos fatos, conceitos ou instruções de uma maneira normalizada que se adapte à comunicação, interpretação e processamento pelo ser humano ou através de máquinas. Os dados são representados por símbolos como por exemplo as letras do alfabeto: a, b, c , etc, mas não são em si a informação desejada. Vejam que a alternativa troca o significado de dados e informação, estando, portanto, incorreta.

[91] Existe uma estrutura baseada nas noções de qualidade versus quantidade, classificando a informação em quatro níveis, na medida em que é agregado valor a cada nível, e nos apresenta essa estrutura através da pirâmide informacional, dado, informação, conhecimento e inteligência. Assim, explica que o atributo da inteligência é mais de caráter qualitativo que quantitativo, uma vez que depende mais da qualidade da informação adquirida do que de sua quantidade, ao contrário dos dados, onde a quantidade de informação é mais importante que sua qualidade aparente. Desta forma a alternativa encontra-se correta.

Gabarito: E C.

3. Dados estruturados e não estruturados.

Quando resolvemos classificar os dados, uma das taxonomias possíveis é aquela que divide os dados de acordo com a sua estruturação. Neste contexto segregamos os dados em **estruturados**, **semiestruturados** e **não estruturados**. Vamos tentar definir cada um deles:

Dados estruturados são aqueles quem possuem a mesma estrutura de representação rígida e previamente projetada, ou sejam, existe um esquema que estabelece algumas características dos dados que serão armazenados. Esses dados são organizados e gravado em um banco de dados.

Perceba que os dados estruturados são definidos por um esquema ou um formato padrão para cada um dos atributos. Podemos, portanto, observar que existe uma mesma estrutura de representação para cada grupo de item de dados (relação) que é rígida e definida previamente por meio de um esquema.

Vamos tentar abstrair um pouco essa definição. Imagine uma tabela com apenas duas colunas nome e telefone. O nome deve possuir no máximo 20 caracteres e o telefone deve ser um valor numérico de 9 algarismos. Veja que para incluir um par, nome/telefone, nesta tabela você deve respeitar o tipo e tamanho dos campos.

Outro ponto interessante é que a tabela tem apenas dois atributos ou colunas, essa característica foi definida pelo esquema. O esquema é semelhante a uma forma de bolo. Não temos como inserir outra informação, como endereço se não reformularmos o esquema para aceitar esse novo atributo. Enfim, perceba que temos uma rigidez quando aos valores que podem ser manipulados quando trabalhamos com dados estruturados.

Para flexibilizar essa situação foram criados os dados semiestruturados.

Dados semiestruturados. Essa segunda categoria de dados geralmente não é mantida em um banco de dados (embora a maioria dos bancos de dados relacionais admita o tipo XML, seguindo a recomendação do padrão SQL). Os dados semiestruturados apresentam uma organização bastante heterogênea, o que pode dificultar as consultas a esses dados.

Os dados não são estritamente tipados[1], mas também não são completamente desestruturados. O que então faz com que ele seja classificado desta forma?

Os dados semiestruturados são dados onde o esquema de representação está presente de forma explícita ou implícita, podendo ser, inclusive, auto descritivo.

[1] Um tipo de dados em uma linguagem de programação ou em um banco de dados são os domínios básicos que podem ser atribuídos a cada uma das colunas de uma tabela. Os tipos de dados mais utilizados são os numéricos e os de caracteres.

Uma análise do dado dever ser feita para que a sua estrutura seja identificada e extraída. Como exemplos de dados semiestruturados podemos citar: XML – *eXtensible Markup Language*, RDF – *Resource Description Framework*, OWL – *Web Ontology Language*. A figura abaixo apresenta a diferença entre um texto plano e um conjunto de dados semiestruturado.

Exemplo de texto e seu correspondente XML

Catálogo de endereços João Silva Rua Carijós, 135 Belo Horizonte, MG 30.000 Brasil 31 3335-5556 (preferido) 31 3549-4446 joaosilva@net.com.br José Almeida jalmeida@net.com.br	<?xml version="1.0"?> <catálogo de endereços> <entrada> <nome> João Silva </nome> <endereço> <rua> Carijós, 135</rua> <estado> MG </estado> < cep> 30.000 </cep> <país> Brasil </pais> <endereço> <telefone preferido="true">31 3335-4456</telefone> <telefone> 31 3594-4446 </telefone> <email> joaosilva@net.com.br </email> </entrada> <entrada> <nome><prim>José</prim> <sobren>Almeida</sobren> <email> jalmeida@net.com.br </email> </entrada> </catalogo de endereço>

Como características destes dados podemos listar:

1. Definição à posteriori – os esquemas são definidos após a existência dos dados por meio da investigação de suas estruturas particulares.

2. Estrutura irregular – não existe um esquema padrão para os dados, as coleções de dados são definidas de maneiras diferentes, podendo conter informações incompletas.

3. Estrutura implícita – muitas vezes uma estrutura implícita.

4. Estrutura parcial – apenas parte dos dados disponíveis podem ter uma estrutura.

Agora podemos passar para uma análise dos dados não estruturados.

Dados não estruturados. De forma simples são inclusos aqui os dados que não possuem uma estrutura definida. Podem ser listados como exemplo documentos, textos, imagens e vídeos. Esses, geralmente, não possuem descrição para suas estruturas, nem implicitamente. A grande maioria dos dados disponíveis na Web são classificados dentro destes formatos.

Tendo em vista a grande quantidade de dados (**volume**), uma imensa **variedade** de dados e a necessidade de processamento cada vez mais rápida (**velocidade**) surge o conceito de Big Data, que será assunto de uma aula posterior (caso seja um assunto cobrando no seu concurso).

Esse assunto foi cobrado pelo CESPE no concurso do TCE-PB, vejamos:

2. Ano: 2018 Banca: CESPE Órgão: TCE-PB Cargo: Auditor de Constas Públicas Questão: 96

Com relação a dados estruturados e não estruturados, assinale a opção correta.

A Dados não estruturados podem ser caracterizados por arquivos de diversos tipos — textos, imagens, vídeos, entre outros —, cujas estruturas não são descritas implicitamente.

B Por padrão, documentos do tipo XML (eXtensible Markup Language) são estruturados.

C Dados não estruturados de um mesmo grupo possuem as mesmas descrições e, consequentemente, os mesmos atributos.

D Por padrão, dados não estruturados são organizados em blocos semânticos.

E A alta heterogeneidade facilita as consultas aos dados não estruturados, desde que estes estejam ligados por ponteiros.

Comentário: Vamos analisar cada uma das alternativas acima.

A primeira alternativa trata de dados não estruturados, cujas estruturas não são descritas nem **explicitamente, nem implicitamente.** Veja que o examinador tentou confundir o candidato usando apenas o termo implícito. Isso é o que acontece com os dados **semiestruturados**, onde não existe uma definição a priori dos arquivos, contudo você consegue deduzir características do arquivo. Agora vamos analisar o que está errado nas demais alternativa.

Os documentos XML são considerados semiestruturados. Eles não têm um tamanho definido para seus elementos, mas os termos possuem "tags" que descrevem os atributos presentes nos arquivos. Sendo assim, a alternativa b está incorreta. Já a letra c, inclui atributos nos dados não estruturados. Os dados não estruturados tendem a possuir uma forma mais livre, não tabular, são dispersos e não facilmente recuperáveis. Esses dados requerem uma intervenção deliberada para dar sentido aos mesmos. E-mails, documentos, páginas da web e arquivos (sejam eles texto, áudio e / ou vídeo) em locais dispersos são exemplos de dados não estruturados.

Tente abrir um documento de música ou vídeo em um notepad. Perceba que não existe nenhum bloco semântico facilmente identificado dentro do arquivo. Existe um termo denominado Web Semântica que é uma tentativa de prover soluções e tecnologias adequadas para resolver o problema da integração de dados, em especial quando usamos dados estruturados e não estruturados.

Por fim, a última afirmação é totalmente equivocada. A heterogeneidade é um dos fatores que dificulta as consultas em dados não estruturados. Podemos resumir alguns pontos de dados estruturados e não estruturados na tabela abaixo:

Característica	Dados estruturados	Dados não estruturados
Associação	Organizado	Difuso e disperso
Aparência	Formalmente definido	Definição livre
Acessibilidade	De fácil acesso e consulta	Dificil de cosultar e acessar

Gabarito: A

4. Dados Abertos

Vejamos, a seguir, o que o governo federal tem a dizer sobre dados abertos. O conteúdo foi extraído de http://dados.gov.br/dados-abertos/.

Segundo a definição da **Open Knowledge Foundation**, "dados são abertos quando **qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los**, estando sujeito a, no máximo, a exigência de creditar a sua autoria e compartilhar pela mesma licença"

Isso geralmente é satisfeito pela publicação dos dados em formato aberto e sob uma licença aberta. Os dados abertos também são pautados por **três leis** e **oito princípios**.

AS TRÊS LEIS – O especialista em políticas públicas e ativista dos dados abertos David Eaves propôs as seguintes "leis":

1. "Se o dado não pode ser encontrado e indexado na Web, ele não existe"

2. "Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado"

3. "Se algum dispositivo legal não permitir sua replicação, ele não é útil"

As leis foram propostas para os Dados Abertos Governamentais, mas pode-se dizer que elas se aplicam aos Dados Abertos de forma geral.

OS OITO PRINCÍPIOS - Em 2007, um grupo de trabalho de 30 pessoas reuniu-se na Califórnia, Estados Unidos da América, para definir os princípios dos Dados Abertos Governamentais. Chegaram a um consenso sobre os seguintes 8 princípios:

Completos. Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo, mas não se limitando a, documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, reguladas por estatutos.

Primários. Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível (maior nível de detalhe possível), e não de forma agregada ou transformada.

Atuais. Os dados são disponibilizados o quão rapidamente seja necessário para preservar o seu valor.

Acessíveis. Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis.

Processáveis por máquina. Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado.

Acesso não discriminatório. Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro.

Formatos não proprietários. Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo.

Livres de licenças. Os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

Além disso, o grupo afirmou que a conformidade com esses princípios precisa ser verificável e uma pessoa deve ser designada como contato responsável pelos dados.

Apesar dos princípios terem sido pensados para os Dados Abertos Governamentais, pode-se aplicá-los, também, a Dados Abertos de modo geral (com a possível exceção do primeiro, já que este trata de dados do poder público).

CINCO MOTIVOS PARA A ABERTURA DOS DADOS - A publicação "**5 motivos para a abertura de dados na Administração Pública**" elaborada pelo Tribunal de Contas da União, apresenta razões para que as organizações públicas invistam em iniciativas de abertura de dados governamentais. Os cinco motivos para a abertura dos dados são:

Transparência na gestão pública;

Contribuição da sociedade com serviços inovadores ao cidadão;

Aprimoramento na qualidade dos dados governamentais;

Viabilização de novos negócios;

Obrigatoriedade por lei.

Visite e conheça o portal de dados abertos do Governo Federal, em http://dados.gov.br/, e, como fonte complementar, temos o manual de dados abertos, traduzido de opendatamanual.org. Agora, antes de passarmos para os conceitos básicos de banco de dados, vamos fazer mais uma questão do CESPE.

3. Ano: 2016 Banca: CESPE Órgão: TCE-SC Cargo: Auditor de TI

A respeito de dados estruturados, não estruturados e abertos, julgue os itens subsequentes.

92 Dados abertos são os dados de livre utilização, reutilização e redistribuição, exigindo-se, no máximo, créditos à autoria e compartilhamento pela mesma licença.

93 Em se tratando de dados estruturados, a informação de esquema está mesclada aos valores dos dados, e cada objeto de dados pode ter atributos diferentes, que não são conhecidos com antecedência. Essa característica os diferencia de dados não estruturados.

Comentário: Vamos, mais uma vez, comentar as assertivas acima.

[92] Dados abertos (open data) trouxe à ideia de que certos dados devem estar disponíveis para que todos usem e publiquem, sem restrições de direitos autorais e patentes ou outros mecanismos de controle. Segundo a *Open Definition*, dados abertos são **dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa** – sujeitos, no máximo, à exigência de atribuição à fonte original e ao compartilhamento pelas mesmas licenças em que as informações foram apresentadas. Analisando a definição, podemos marcar a alternativa como **correta**.

[93] Dados estruturados são organizados em linhas e colunas, geralmente são encontrados em banco de dados relacionais (falaremos sobre eles na aula 02 do nosso curso), são eficientes quanto à recuperação e processamento. Cada coluna deve ter o mesmo tipo de dados para todos os valores diferentes de nulo[2], assim como as linhas devem possuir registros que satisfação as restrições de integridade descritas na definição da tabela. A definição da questão é confusa e se assemelha com o contexto de dados semiestruturados. Assim, podemos concluir que a afirmação está incorreta.

Gabarito: C E.

[2] Valores nulos correspondem ao conjunto vazio. Suponha que um campo em uma tabela Excel não esteja preenchido, se formos traduzir esse fato para uma tabela em um banco de dados relacional em deve ser preenchido com NULL ou nulo.

Considerações finais

Chegamos, pois, ao final da aula demonstrativa do curso de Business Intelligence! As videosaulas referentes ao conteúdo visto até aqui já estão gravadas e devidamente disponibilizadas na área do aluno. Espero que gostem!

A continuação deste assunto encontra-se na próxima aula. Falaremos um pouco sobre Modelagem Multidemesional e Data Warehouse. Espero reencontrar você como um aluno efetivo.

Forte abraço, bons estudos e até breve!

Que Deus abençoe!

0

Thiago Cavalcanti

Referências

Fiz uma lista com alguns links de referências caso você queria se aprofundar um pouco.

i. Fundamentals of Database Systems - Ramez Elmasri, Sham Navathe Addison-Wesley, 2011 - Computers - 1172 pages
ii. Introdução a sistemas de bancos de dados - By C. J. Date - Elsevier Brasil, 2004 - 865 pages
iii. Sistema de Banco de Dados - Abraham Silberschatz, Henry F. Korth, S. Sudarshan - Editora: ELSEVIER BRASIL

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.