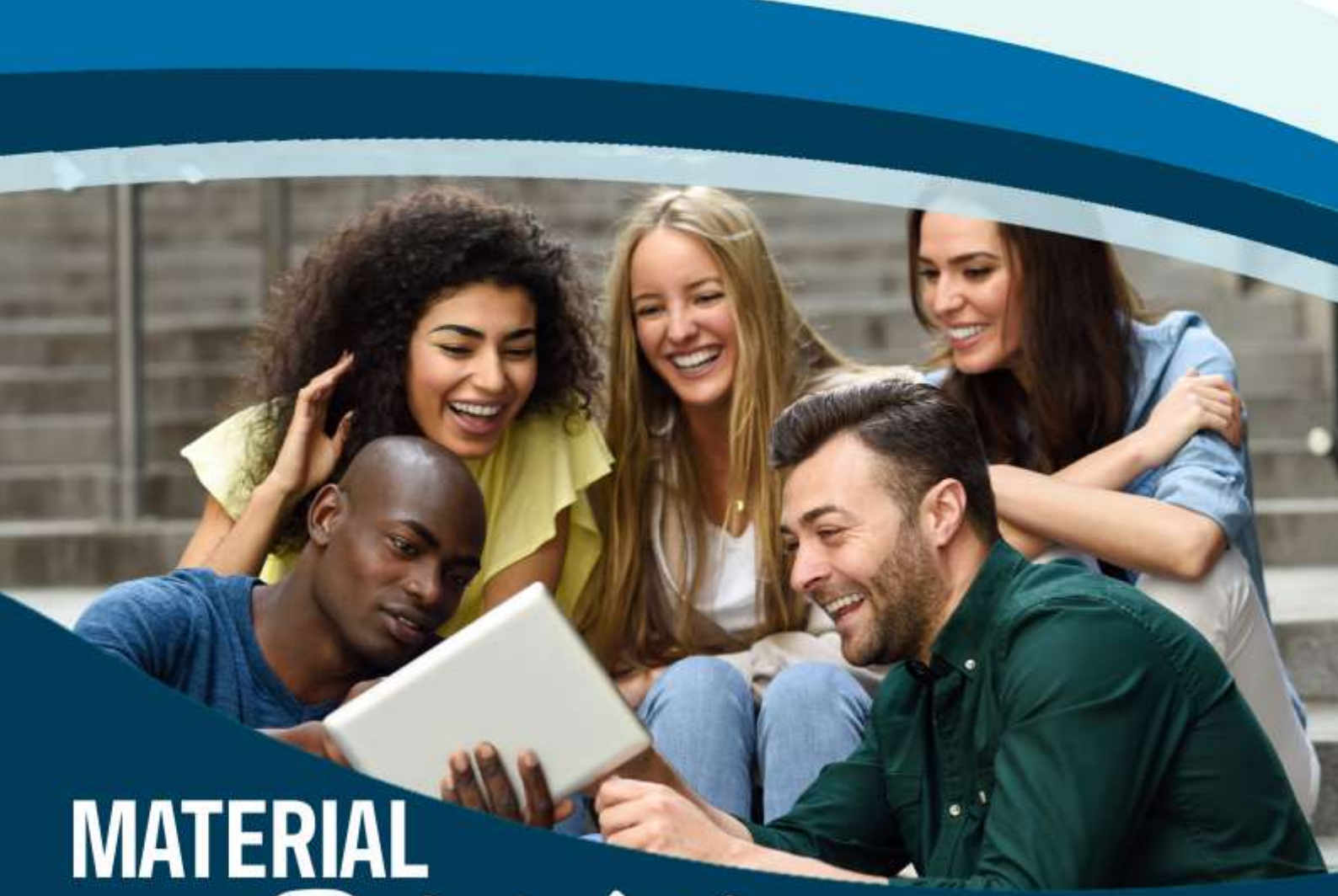


MATERIAL

Didático

ANÁLISE DE DADOS DE ALTA DIMENSÃO

FACUMINAS

CREDENCIADA JUNTO AO MEC PELA PORTARIA
N 3.455 DO DIA 19/11/2003

www.facuminasead.com.br 31 3842-3838

FACUMINAS

SUMÁRIO

FACUMINAS

A história do Instituto Facuminas, inicia com a realização do sonho de um grupo de empresários, em atender a crescente demanda de alunos para cursos de Graduação e Pós-Graduação. Com isso foi criado a Facuminas, como entidade oferecendo serviços educacionais em nível superior.

A Facuminas tem por objetivo formar diplomados nas diferentes áreas de conhecimento, aptos para a inserção em setores profissionais e para a participação no desenvolvimento da sociedade brasileira, e colaborar na sua formação contínua. Além de promover a divulgação de conhecimentos culturais, científicos e técnicos que constituem patrimônio da humanidade e comunicar o saber através do ensino, de publicação ou outras normas de comunicação.

A nossa missão é oferecer qualidade em conhecimento e cultura de forma confiável e eficiente para que o aluno tenha oportunidade de construir uma base profissional e ética. Dessa forma, conquistando o espaço de uma das instituições modelo no país na oferta de cursos, primando sempre pela inovação tecnológica, excelência no atendimento e valor do serviço oferecido.

EVOLUÇÃO DAS INFORMAÇÕES E DOS DADOS

Desde o início do século XXI ocorreram mudanças significativas no âmbito das Tecnologias da Informação e Comunicação (TIC), neste contexto pode-se citar a computação em nuvem, internet das coisas e as redes sociais. O acesso e o uso destas tecnologias fizeram com que a quantidade de dados aumentasse de uma forma contínua e a uma velocidade sem precedentes (CAI; ZHU, 2015).

As tecnologias vêm revolucionando a maneira como as pessoas lidam e interagem com o mundo a seu redor de forma cada vez mais drástica. Desde o surgimento da Internet até sua popularização, o ser humano depende cada vez mais das máquinas para realizar suas tarefas diárias. Com a criação dos smartphones passa-se a estar conectados à Internet 24 horas por dia, 7 dias por semana. Com a Internet das Coisas, muito em breve, vários objetos de nossas casas e trabalho estarão, também, inteiramente conectados.

Todos esses aparatos rastreiam cada passo do que é feito na Web e cada um desses passos é um potencial gerador de dados. Como passa-se cada vez mais tempo online, gera-se progressivamente mais dados, de forma mais rápida. Qualquer tarefa realizada quando se está conectado, desde enviar um e-mail, escrever uma mensagem em uma rede social ou até mesmo deixar o GPS do celular ligado enquanto se movimenta, deixa rastros imperceptíveis que transformam-se em dados.

Gera-se atualmente aproximadamente 1 zettabyte de dados por dia mas, até então, a imensa maioria desse volume era ignorada. Porém, percebeu-se em determinado momento que esses dados podem ser valiosíssimos quando bem aproveitados, principalmente para as empresas, que podem obter informações em tempo real, geradas diretamente por seus clientes.

Para o governo, significam informações sobre como a sociedade se comporta e formas mais fáceis de monitorar a população. Por outro lado, cria-se uma fina barreira entre os benefícios da exploração de dados e a invasão de privacidade.

Neste mundo globalizado e constantemente conectado, surge então uma nova fonte de poder que não pode mais ser ignorada. Os especialistas a chamam de Big Data

De acordo com Furlan e Laurindo (2017), o crescimento dos dados gerados demandou o desenvolvimento de novas soluções e tecnologias que auxiliassem na sua gestão. E diante desta necessidade surge o big data, propondo novas abordagens para a geração, seleção e manipulação destes grandes volumes dados. O termo big data está relacionado com grandes quantidades de dados, que possuem características distintas, são heterogêneos, providos de diferentes fontes, com controles distribuídos e descentralizados (MCAFEE; BRYNJOLFSSON2012).

Dada a importância das informações disponibilizadas, a qualidade dos dados que geram estas informações tornou-se um dos grandes desafios para as organizações se manterem em um mercado cada vez mais competitivo, sendo que a partir da década de 90 iniciaram-se diversos estudos sobre metodologias e ferramentas para auxiliar no processo de gestão da qualidade dos dados dentro das organizações, e uma daspropposições mais relevantes foi o Total Data Quality Management (TDQM), feita por Madnick e Wang em 1992 (ZHUet al., 2012).

O programa desenvolvido pelo Massachusetts Instituteof Technology - MIT é baseado na estrutura de Gerenciamento de Qualidade Total (TQM) para melhoria da qualidade no domínio da fabricação, proposto por William Edwards Deming em 1982. Suas pesquisas iniciais desenvolveram um modelo que defende a melhoria contínua da qualidade dos dados, seguindo ciclos de definição, medição, análise e melhoria.

A partir do TDQM, várias outras proposições relacionadas as dimensões e atributos da qualidade dos dados foram feitas, porém, a definição de quais critérios a serem adotados depende do contexto em que os mesmos serão aplicados (BATINIet al., 2009).

Outra questão a se considerar remete à diversidade dos dados disponíveis, uma vez que os mesmos são originados a partir de diferentes fontes, causando uma sobrecarga de informação para a sociedade, gerando inúmeras oportunidades de

atuação para os profissionais que atuam na área da gestão da informação (RIBEIRO, 2014).

DADO, INFORMAÇÃO E CONHECIMENTO

Na Ciência da Informação, a conceituação deste termo está quase sempre ligada aos termos Informação e Conhecimento. Por se tratarem de definições com significados muito similares, não há consonância entre os autores da área para defini-los, considerando o aspecto contextual como essencial para distinção, por vezes tênue, dos termos discutidos. (RUSSO, 2010).

Dado é definido por Setzer (2001, não paginado) como “[...] uma sequência de símbolos quantificados ou quantificáveis [...] Com essa definição, um dado é necessariamente uma entidade matemática e, desta forma, é puramente sintático”. Já Angeloni (2003, p.18) afirma que “Os dados são elementos brutos, sem significado, desvinculados da realidade.” Davenport (1998, p.18) define dado como uma “Simples observação sobre o estado do mundo”. Sendo ele facilmente estruturado e obtido por máquinas, além de ser quantificado com frequência. O dado é facilmente transferível. Conclui-se à partir das elucidações de Russo (2010, p.15), que “[...] dados são sinais que não foram processados, correlacionados, integrados, avaliados ou interpretados de qualquer forma, e, por sua vez, representam a matéria prima a ser utilizada na produção de informações”.

A informação, segundo Setzer (2001, não paginado) “[...] é uma abstração informal (isto é, não pode ser formalizada através de uma teoria lógica ou matemática), que está na mente de alguém, representando algo significativo para essa pessoa”. Na definição de Angeloni (2003, p.18) “As informações são dados com significado.” A informação é considerada como “[...] dados processados e contextualizados”. Davenport (1998, p.18) diz que informação são “Dados dotados de relevância e propósito”. A informação requer uma unidade de análise e exige consenso com relação ao seu significado e é imprescindível a mediação humana.

Para Russo (2010, p.15) "[...] informação pode ser entendida como dados processados e contextualizados".

Conhecimento é definido por Davenport (1998, p.19 apud RUSSO, 2010, p.17) como "[...] a informação mais valiosa [...] é valiosa precisamente porque alguém deu à informação um contexto, um significado, uma interpretação". Setzer (2001, não paginado) caracteriza conhecimento como "[...] uma abstração interior, pessoal, de algo que foi experimentado, vivenciado, por alguém". Angeloni (2003, p.18) define:

> O conhecimento pode então ser considerado como a informação processada pelos indivíduos. O valor agregado à informação depende dos conhecimentos anteriores desses indivíduos. Assim sendo, adquirimos conhecimento por meio do uso da informação nas nossas ações. Desta forma, o conhecimento não pode ser desvinculado do indivíduo; ele está estritamente relacionado com a percepção do mesmo, que codifica, decodifica, distorce e usa a informação de acordo com suas características pessoais, ou seja, de acordo com seus modelos mentais. (ANGELONI, 2003, p.18).

O conhecimento de acordo com Davenport (1998, p.18) é uma "Informação valiosa da mente humana. Inclui reflexão, síntese, contexto". É de difícil estruturação, transferência e captura em máquinas. É frequentemente tácito. O conhecimento tácito é capital dos seres humanos, dificilmente transmitido e capturado, dependendo das experiências e vivências da pessoa (RUSSO, 2010, p.19). Russo (2010, p.18) distingue os termos analisados como apresentado na figura a seguir:

Figura 1: Dado x Informação X Conhecimento

Dado	Informação	Conhecimento
Conjunto de letras, números ou dígitos que não contém significado claro.	Dado trabalhado, útil, tratado, com valor significativo atribuído ou agregado a ele com um sentido natural e lógico.	Informação trabalhada por pessoas e pelos recursos computacionais, possibilitando a geração de cenários, simulações e oportunidades.

Fonte: Russo, 2010, pag. 18.

Analisando as definições acima, percebe-se que o dado, por si só, não possui significado claro e precisa portanto, ser capturado e analisado, para que se transforme em informação. Uma característica importante a ser analisada sobre os dados é que antes do Big Data, os dados coletados pelas empresas provinham de fontes internas e eram, majoritariamente, dados estruturados, ou seja, "[...] dados formatados em linhas e colunas numéricas organizadas" (DAVENPORT, 2014, p.113). Esse tipo de dado já vem sendo explorado há um tempo pelas organizações, estando presente em bancos de dados, arquivos sequenciais e com relação de importância. (CIO, 2012 apud CANARY, 2013).

GESTÃO DE DADOS

A gestão de dados tem o objetivo de gerenciar e zelar pelos dados das empresas, tratando-os como um recurso valioso, de modo que as informações possam ser transformadas em valor empresarial e embasar decisões estratégicas. E, para isso, a gestão de dados utiliza processos, profissionais, metodologias e ferramentas.

Atualmente, no contexto da Transformação Digital, um gerenciamento inteligente das informações se torna ainda mais importante. À medida que as organizações são cada vez mais orientadas por processos digitais, a quantidade de dados cresce exponencialmente. Nunca foi tão fácil captar informações de mercado, mas, por outro lado, nunca foi tão crucial protegê-las. Hoje, a segurança dos dados é uma prioridade.

Além disso, empresas que falham em organizar suas grandes quantidades de informações acabam gastando desnecessariamente com armazenagem. Também tendem a arcar com gastos extras com compliance e recursos humanos — o tempo gasto para procurar informações, gerenciar processos e cumprir tarefas aumenta de forma drástica. Entretanto, mais do que tudo isso, a gestão de dados é um alicerce da Transformação Digital. Como sabemos, ela vai muito além da tecnologia: ela está diretamente ligada à cultura organizacional e às operações.

A gestão de dados eficiente gera vantagem competitiva para as empresas. Apesar de a maioria dos benefícios serem diferentes em cada empresa, alguns ainda são comuns à maioria. Entre eles podemos destacar:

– Melhor alinhamento entre as áreas de tecnologia e de negócio.

-Conhecimento dos dados utilizados na empresa através da adoção de um vocabulário único sobre as definições dos dados que circulam na empresa.

-Entendimento das principais necessidades de dados e informações da empresa, fornecendo um importante subsídio para estabelecer o planejamento para absorção, criação e/ou transformação de novos dados e informações.

-Melhoria na qualidade e confiabilidade dos dados e informações através do uso de dados cada vez mais claros, precisos, íntegros, integrados, pertinentes e oportunos.

-Criação da cultura do uso de indicadores de processo e qualidade dos dados.

-Reutilização de dados considerados corporativos, contribuindo dessa forma para a melhoria da qualidade dos dados e também reduzindo os esforços, tempos e custos do desenvolvimento de novas aplicações.

-Redução dos riscos e falhas no desenvolvimento dos sistemas e aplicações.

-Eliminação ou redução drástica na quantidade de informações redundantes, contribuindo para reduzir os esforços em manter íntegras as informações que antes eram redundantes.

-Estabelecimento de mecanismos formais de segurança, acesso e disponibilização de dados e informações a quem realmente necessita.

-Aumento da produtividade das pessoas que utilizam os dados e as informações.

Sem uma gestão efetiva dos dados, a evolução desta cadeia não é atingida, portanto, para atingir os objetivos é fundamental a disciplina atuar nos estágios

iniciais da cadeia. Por esta razão o nome da disciplina é Gestão de Dados e não Gestão das Informações ou Gestão do Conhecimento. Porém, valem ressaltar que, dependendo do nível da maturidade da empresa, as ações de gestão para a evolução da cadeia podem se estabelecer em outros níveis.

A análise de dados tem sido cada vez mais importante para o sucesso, sendo você um pequeno empresário ou uma grande indústria. Para que a sua empresa se destaque e deixe a concorrência para trás, é necessário reunir e analisar informações sobre clientes e o mercado como um todo. Sendo assim, em diversos aspectos, fazer o gerenciamento de dados de forma inteligente é absolutamente essencial para o sucesso dos negócios e para atingir um nível sustentável de crescimento.

Se puder resumir em uma palavra o que é gerenciamento de dados, podemos definir como Organização. Esse processo é baseado em coletar, validar, armazenar e garantir a segurança dos dados para poderem ser transformados, de fato, em informações úteis. Decisões importantes fundamentadas em dados crus podem ser muito perigoso, por isso deve-se ir além e analisar cada detalhe da base de dados para qualificar o processo de tomada de decisões, que não deve ser realizado de maneira precipitada.

Quando todas as informações das quais você precisa estão bem organizadas e devidamente armazenadas, de forma integrada e ordenada, arquivos como projetos, balanços financeiros, prospecções de vendas e outros documentos importantes são encontrados com maior facilidade no sistema toda vez que você precisar analisá-los.

A segurança de dados é um ponto de extrema importância para qualquer organização. Quando a proteção das informações é abalada, seja no caso de um vazamento, seja diante de um roubo de dados, o próprio futuro do negócio fica comprometido.

É inevitável que, para fazer bons negócios e sustentar o crescimento, a empresa tenha a internet como uma ferramenta de trabalho no dia a dia. Entretanto,

quem está sempre conectado também está em constante risco de perder seus dados, por diversos motivos.

A seguir, elencamos algumas das principais ameaças à segurança da informação, passíveis de ocorrerem quando não há um adequado gerenciamento de dados. Confira:

1. **Ciber ataques :**Infelizmente, toda e qualquer empresa está sujeita aos ataques de pessoas de má-fé e cibercriminosos em geral. Entre outras possibilidades, eles podem invadir a rede da sua organização, prejudicando a integridade dos dados. Outra ação maléfica causada por esses criminosos é o envio de vírus, que podem danificar por completo os seus softwares.
2. **Configurações de segurança desatualizadas:** Existem diversas soluções em sistemas de dados que possibilitam a personalização de medidas de segurança e restrição de acessos por nível de usuário, a fim de proteger sua estrutura de dados e permitir que só usuários autorizados acessem determinadas pastas e arquivos importantes. Porém, infelizmente muitas empresas simplesmente não se aproveitam dessa funcionalidade, ficando mais vulneráveis a problemas e falhas de segurança.
3. **Desastres:** Em 2017, em um dos mais graves atentados cibernéticos mundiais, o ransonware WannaCry fez milhares de vítimas em mais de 150 países, entre pequenos negócios a grandes corporações. O ransonware é um tipo de vírus que sequestra os dados de um computador ou notebook e não libera o acesso a eles até que um resgate seja pago.

Nesse sentido, se uma empresa não conta com um programa de recuperação de desastres, por meio do qual ela se prepara para catástrofes como desastres naturais, falha humana (acidental ou não), ataques cibernéticos, quebra de equipamentos, entre outros, os dados gerados no decorrer das atividades são perdidos.

Para auxiliar você a garantir a segurança das informações de sua empresa, é necessário ações que para garantir a eficiência do gerenciamento de dados. Como backups, diversas empresas adotam uma periodicidade mensal (ou maior) para seus backups. Entretanto, considere a imensa quantidade de dados que são gerados e processados durante esse tempo: o mais adequado é realizar backups diários, ao final de cada jornada de trabalho. Isso pode ser feito de forma programada, sem que você precise executar o processo manualmente todos os dias.

O uso da cloud computing tem se tornado uma tendência cada vez mais presente no mercado. A praticidade e o dinamismo que essa tecnologia oferece vem beneficiando empresas de todos os portes e segmentos. De fato, é compreensível que muitas empresas julguem necessário ter alguma forma de backup físico (como alternativa e uma segunda opção de garantia), porém, além de dispensar o uso de aparelhos físicos (como HDs), o backup na nuvem realmente garante a preservação e integridade dos dados.

Outro elemento essencial que é proporcionado pelo gerenciamento de dados na nuvem é a disponibilização de modo mais fácil, dinâmico e rápido das informações, inclusive, para um número mais amplo de pessoas se necessário. Compartilhar informações entre todas as pessoas na empresa e agentes externos, fica mais simples com os dados disponibilizados na nuvem. O tempo que se ganha com o uso da tecnologia em nuvem é gigantesco. Com as informações sincronizadas e acessíveis praticamente de modo instantâneo, agiliza-se o fluxo de atividades, otimizando o tempo.

Implemente ferramentas de monitoramento e controle, as ferramentas para monitoramento e controle serão, de fato, muito úteis para averiguar as suspeitas de ataques ou vazamentos e, além disso, reforçar o sigilo interno, ou seja, garantir que as pessoas só tenham disponibilidade às informações de acordo com seu nível de autorização de acesso.

BANCO DE DADOS

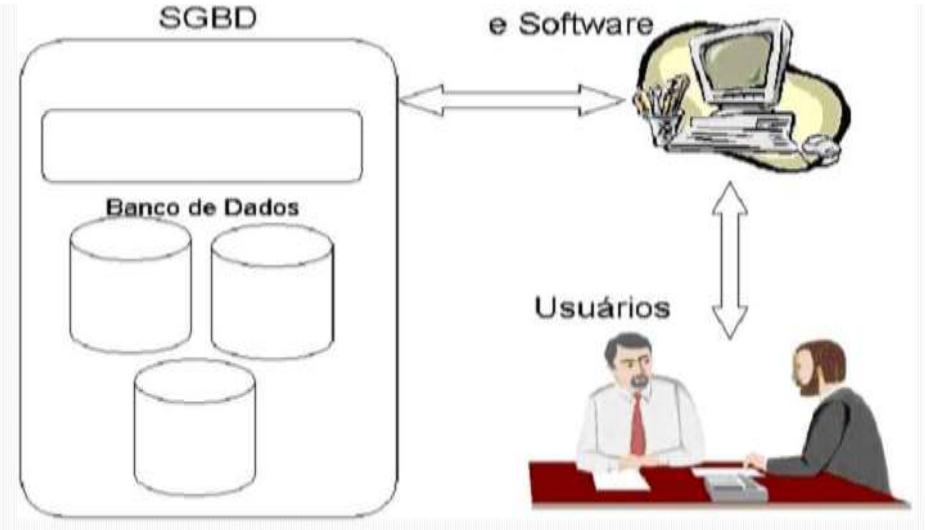
Segundo Korth, um banco de dados “é uma coleção de dados inter-relacionados, representando informações sobre um domínio específico”, ou seja, sempre que for possível agrupar informações que se relacionam e tratam de um mesmo assunto, posso dizer que tenho um banco de dados.

Podemos exemplificar situações clássicas como uma lista telefônica, um catálogo de CDs ou um sistema de controle de RH de uma empresa.

Já um sistema de gerenciamento de banco de dados (SGBD) é um software que possui recursos capazes de manipular as informações do banco de dados e interagir com o usuário. Exemplos de SGBDs são: Oracle, SQL Server, DB2, PostgreSQL, MySQL, o próprio Access ou Paradox, entre outros.

Por último, temos que conceituar um sistema de banco de dados como o conjunto de quatro componentes básicos: dados, hardware, software e usuários. Date conceituou que “sistema de bancos de dados pode ser considerado como uma sala de arquivos eletrônica”. A Figura 2 ilustra os componentes de um sistema de banco de dados.

Figura 2: Componentes de um sistema de banco de dados



Os objetivos de um sistema de banco de dados são o de isolar o usuário dos detalhes internos do banco de dados (promover a abstração de dados) e promover a independência dos dados em relação às aplicações, ou seja, tornar independente da aplicação, a estratégia de acesso e a forma de armazenamento.

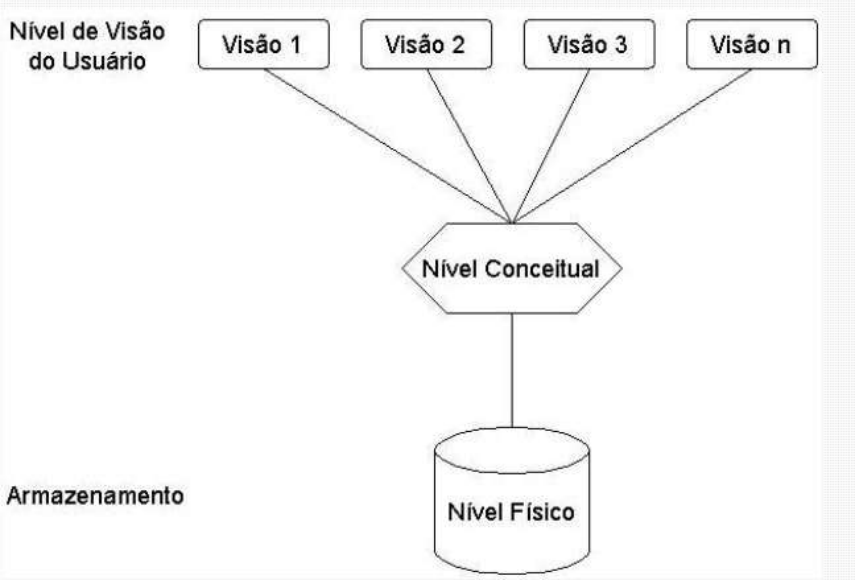
ABSTRAÇÃO DE DADOS

O sistema de banco de dados deve garantir uma visão totalmente abstrata do banco de dados para o usuário, ou seja, para o usuário do banco de dados pouco importa qual unidade de armazenamento está sendo usada para guardar seus dados, contanto que os mesmos estejam disponíveis no momento necessário.

Esta abstração se dá em três níveis (Figura 3):

- Nível de visão do usuário: as partes do banco de dados que o usuário tem acesso de acordo com a necessidade individual de cada usuário ou grupo de usuários;
- Nível conceitual: define quais os dados que estão armazenados e qual o relacionamento entre eles;
- Nível físico: é o nível mais baixo de abstração, em que define efetivamente de que maneira os dados estão armazenados.

Figura 3 Níveis de abstração



PROJETO DE BANCO DE DADOS

Todo bom sistema de banco de dados deve apresentar um projeto, que visa a organização das informações e utilização de técnicas para que o futuro sistema obtenha boa performance e também facilite infinitamente as manutenções que venham a acontecer. O projeto de banco de dados se dá em duas fases:

- Modelagem conceitual;
- Projeto lógico.

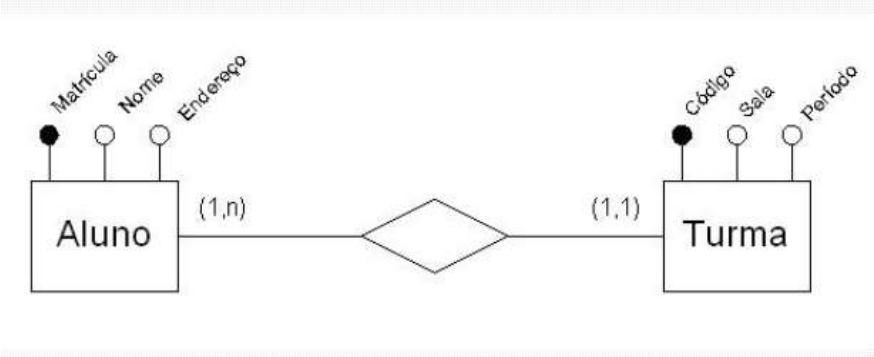
Estas duas etapas se referem a um sistema de banco de dados ainda não implementado, ou seja, que ainda não exista, um novo projeto. Para os casos em que o banco de dados já exista, mas é um sistema legado, por exemplo, ou um sistema muito antigo sem documentação, o processo de projeto de banco de dados se dará através da utilização de uma técnica chamada de Engenharia Reversa, que será visto em outra oportunidade.

MODELO CONCEITUAL

É a descrição do BD de maneira independente ao SGBD, ou seja, define quais os dados que aparecerão no BD, mas sem se importar com a implementação que se dará ao BD. Desta forma, há uma abstração em nível de SGBD.

Uma das técnicas mais utilizadas dentre os profissionais da área é a abordagem entidade-relacionamento (ER), onde o modelo é representado graficamente através do diagrama entidade-relacionamento (DER) (Figura 4).

Figura 4. Exemplo de diagrama entidade-relacionamento



O modelo acima, entre outras coisas, nos traz informações sobre Alunos e Turmas. Para cada Aluno, será armazenado seu número de matrícula, seu nome e endereço, enquanto para cada turma, teremos a informação de seu código, a sala utilizada e o período.

MODELO LÓGICO

Descreve o BD no nível do SGBD, ou seja, depende do tipo particular de SGBD que será usado. Não podemos confundir com o Software que será usado. O tipo de SGBD que o modelo lógico trata é se o mesmo é relacional, orientado a objetos, hierárquico, etc. Abordaremos o SGBD relacional, por serem os mais difundidos. Nele, os dados são organizados em tabelas (Quadro 1).

Quadro 1. Exemplo de tabelas em um SGBD relacional

Aluno		
mat_aluno	nome	endereco
1	Cecília Ortiz Rezende	Rua dos Ipês, 37
2	Abílio José Dias	Avenida Presidente Jânio Quadros, 357
3	Renata Oliveira Franco	Rua Nove de Julho, 45

Turma		
cod_turma	sala	periodo
1	8	Manhãnte
2	5	Noite

O modelo lógico do BD relacional deve definir quais as tabelas e o nome das colunas que compõem estas tabelas. Para o nosso exemplo, poderíamos definir nosso modelo lógico conforme o seguinte:

```
Aluno(mat_aluno, nome, endereco)
Turma (cod_turma, sala, periodo)
```

É importante salientar que os detalhes internos de armazenamento, por exemplo, não são descritos no modelo lógico, pois estas informações fazem parte do modelo físico, que nada mais é que a tradução do modelo lógico para a linguagem do software escolhido para implementar o sistema.

CONCEITOS BÁSICOS DE DADOS DE ALTA DIMENSÃO

Um tipo específico de dados, onde existe um número muito grande de dimensões, que chega a ser comparável com o tamanho de amostra, ou então, algumas vezes até maior do que ele, é conhecido como dados com altas dimensões. Um exemplo prático disto é a presença de muitos genes, mas poucos pacientes com uma determinada doença. Este tipo de dado não pode ser tratado pelas técnicas tradicionais existentes, que não suportam essa alta dimensionalidade dos dados. Veremos a seguir, os problemas que este tipo de dado pode causar.

Em termos gerais, problemas com alta dimensionalidade resultam do fato de que um número fixo de pontos torna-se cada vez mais “esparso” quando o número de dimensões vai aumentando. Para tornar a visualização deste problema mais compreensível, vamos considerar 100 pontos distribuídos aleatoriamente segundo uma distribuição uniforme no intervalo [0,1].

Se este intervalo é dividido em 10 células, é altamente provável que cada célula contenha algum ponto. Entretanto, se mantivermos fixo este número de pontos, porém, distribuídos em um quadrado unitário (onde, logicamente cada ponto passa a ser bidimensional), mantendo a divisão proposta anteriormente (discretização de 0,1 para cada dimensão), desta vez teremos 100 células bidimensionais e é razoável propor que alguma das células ficará vazia.

Partindo para um exemplo tridimensional, teremos 1000 células, resultando numa quantidade muito maior de células vazias do que células com pontos, visto que o número de células é muito maior do que de pontos. Os dados começam a se “perder no espaço” à medida que aumentamos as dimensões.

No caso do agrupamento de dados (análise de cluster) com altas dimensões, o problema da dimensionalidade afeta principalmente a medida de distância ou de similaridade. A maioria das técnicas de análise de cluster depende desta medida e, em geral, agrupam objetos mais próximos em grupos separados. O mesmo

problema será encontrado quando utilizamos medidas de similaridade para detectar padrões semelhantes em imagens multidimensionais.

O comportamento das distâncias em dados com altas dimensões vem sendo estudado há alguns anos. É mostrado por Beyer et al (1998) que para alguns tipos de distribuições de dados, a distância relativa entre o ponto mais próximo e o ponto mais distante de um determinado ponto escolhido ao acaso vai a zero quando o número de dimensões aumenta. É comum se dizer então que as distâncias entre pontos se tornam relativamente uniformes em dados com altas dimensões. O mesmo problema é encontrado ao se utilizar a distância absoluta ao invés da relativa, onde, segundo Hinneburg et al (2000), para dados com mais de duas dimensões, o uso da distância entre pontos é insignificante para análise de cluster.

Em resumo, isto tudo mostra que dados com altas dimensões definitivamente não podem receber o mesmo tratamento que dados com poucas dimensões e, portanto, necessitam de diferentes abordagens. O conhecimento deste tipo de dificuldade apresentada é de grande interesse para uma grande gama de áreas da ciência. Para tal entendimento, é importante que alguns conceitos destas áreas sejam definidos também.

Na área genética, em 1990, se inicia o Projeto Genoma Humano, com um financiamento inicial de 50 bilhões de dólares e duração prevista de 15 anos. Entre os seus principais objetivos, encontram-se sequenciar e decodificar todo o DNA do genoma humano. Genoma é uma sequência completa de DNA. Esta abordagem foi pioneira na obtenção de dados com altas dimensões. A partir deste projeto, inúmeras outras pesquisas no ramo genético começaram a se desenvolver.

Um exemplo do que foi citado acima é a criação de um termo muito utilizado na área veterinária, que é o proteoma. Ele vem da união das palavras PROTEína e genOMA e é o conjunto de proteínas expressas por algum genoma. Segundo definição de Wilkins et at (1995), em termos gerais, proteoma é o equivalente protéico ao genoma. Porém, o genoma de um indivíduo é praticamente constante, independente de qual célula está sendo analisada, enquanto que o seu proteoma varia bastante de célula para célula (neurônio e linfócitos, por exemplo)

Para a análise de proteoma, muito útil nas áreas de genética, medicina, biociências e veterinária, é utilizada uma técnica conhecida como eletroforese, que se baseia na migração das moléculas carregadas, numa solução, em função da aplicação de um campo elétrico. Esta técnica foi primeiramente utilizada no ano de 1937. Porém, uma nova técnica, em 1975 começa a se desenvolver, gerando imagens bidimensionais.

Esta técnica se chama eletroforese bidimensional e tem como principal objetivo a detecção de proteínas, assim como a sua quantidade, em determinadas células estudadas. Com o surgimento desta técnica, começou a surgir a necessidade de algum tipo de análise que comparasse diferentes imagens (chamadas de géis) para se verificar a eficiência de algum novo tratamento celular, já que este tipo de 15 dado apresenta altas dimensões. No entanto, estes géis são uma representação bidimensional de uma imagem originalmente composta por três dimensões.

Ainda na área genética, a análise da sequência de DNA vem se desenvolvendo de forma crescente. Cada sequência de DNA apresenta nucleotídeos, que são compostos ricos em energia e que auxiliam os processos metabólicos. O DNA apresenta nucleotídeos compostos por duas bases púricas (adenina e guanina, representadas pelas letras A e G, respectivamente) e duas bases pirimídicas (citosina e timina, representadas pelas letras C e T, respectivamente). Para as análises realizadas em enormes micro arranjos do DNA, são realizados cortes em sequências alvo específicas, chamados de sítios. Estes sítios então serão estudados, deixando o trabalho mais objetivo e prático.

Já na área química, para a detecção de determinados tipos de elementos presentes em algum objeto, são utilizados gráficos com espectros, chamados de espectrogramas. Determinados picos no gráfico podem determinar a presença de algum tipo de elemento. A mesma necessidade de comparação de imagens surge aqui, onde temos diversos espectrogramas e queremos verificar se seguem o mesmo padrão de comportamento com relação aos seus elementos químico.

Por fim, na área de astronomia, com o avanço tecnológico, milhões de dados surgem de telescópios. São, portanto, tiradas diversas fotos e gerados milhares de gráficos que precisam de uma análise mais sofisticada devido a sua alta dimensionalidade. Nos próximos capítulos serão abordadas técnicas estatísticas que vem sendo desenvolvidas com o intuito de comparar grupos de imagens ou então agrupar informações de dados com altas dimensões.

ANÁLISE DE SIMILARIDADE

A técnica de eletroforese bidimensional teve as suas primeiras citações feitas por O'Farrell (1975). Com esta técnica, é possível separarmos as proteínas primeiramente utilizando os seus pontos isoelétricos e depois no seu peso molecular. Isto gera uma imagem bidimensional (chamada de gel), onde, como pode ser observado, manchas escuras indicam a presença de uma proteína. Um dos maiores interesses deste tipo de técnica é detectar diferentes expressões protéicas.

Para isto ser feito, é necessário algum tipo de medida de similaridade robusta e precisa entre as manchas. Devido à complexidade biológica, física e química do processo, a localização da mesma mancha protéica pode diferir de maneira tanto global quanto local, o que torna quase impossível o registro perfeito das imagens. Isso torna ainda mais fácil a detecção de diferenças entre duas imagens. Entretanto, este tipo de análise é de extremo interesse, já que traz informações muito pertinentes para biólogos.

A partir destas ideias, Xin e Zhu (2009) propuseram um método simples e preciso para medir a similaridade entre manchas baseado em múltiplas informações. Este método proposto tem inspiração no princípio de atração e explora simultaneamente a distância entre as manchas, a intensidade das manchas e a informação de padrão da mancha para, de forma precisa e automática, relacionar manchas em duas imagens bidimensionais.

O princípio de atração da lei universal de gravidade da física é utilizado, este princípio diz que cada ponto de massa atrai cada outro ponto de massa por um ponto de força ao longo da linha que intersecta os dois pontos; e que a força é proporcional ao produto das duas massas e inversamente proporcional ao quadrado da distância entre os pontos de massa. O método de similaridade de manchas proposto é baseado nestas ideias encontradas no princípio de atração.

Suponha que Ir e If são duas imagens bidimensionais de entrada, representando, respectivamente, as imagens de referência e de flutuação; suponha que são dados dois conjuntos de manchas bidimensionais φr = {sri|i=1,2,..., N} e φf = {sfj|j=1,2,..., M}, onde N e M representam o número de manchas em Ir e If , respectivamente. Seja c um operador que dá as coordenadas (x, y) para o centróide da mancha s, ou seja, c(sri) = (xri, yri), são as coordenadas do centróide para a i-ésima mancha sri da imagem de referência, 17 e c(sfj) = (xfj, yfj) as coordenadas do centróide para a j-ésima mancha sfj na imagem de flutuação. De mesmo modo, g(c(sri)) e g(c(sfj)) denotam as intensidades dos centróides (chamados de níveis cinza) em Ir e c, respectivamente.

REDUÇÃO DE DIMENSÕES

Uma alternativa bastante razoável ao nos depararmos com dados com altas dimensões consiste em encontrarmos uma maneira de reduzirmos a dimensionalidade destes dados, para então realizarmos alguma análise estatística. Para tal, existem dois tipos de técnicas utilizadas: transformação das características (feature transformation) e seleção das características (feature selection).

Técnicas de transformação das características têm como objetivo reduzir um conjunto de dados em menos dimensões através da combinação dos atributos originais. Este tipo de técnica é muito eficiente em encobrir estruturas latentes em conjuntos de dados. Entretanto, são menos eficientes quando há uma quantidade elevada de atributos irrelevantes “escondidos”, visto que preservam a distância

relativa entre objetos. Além disso, as novas características serão formadas pelas combinações das originais e isso pode causar certo transtorno na hora de interpretá-las.

Técnicas de seleção das características selecionam apenas as dimensões mais relevantes de um conjunto de dados. No caso particular de uma possível análise de cluster posteriormente, uma limitação deste tipo de técnica ocorre quando os grupos são formados em diferentes subespaços. Este tipo de problema motivou a criação de algoritmos que usam ideias das técnicas de transformação das características e posteriormente selecionam apenas os subespaços relevantes para cada cluster separadamente.

Transformação das Características

Transformações das características são muito utilizadas em dados com altas dimensões. Estes métodos incluem técnicas como análise de componentes principais e decomposição em valores singulares. As transformações geralmente preservam as distâncias relativas originais dos objetos. Desta maneira, o conjunto de dados é resumido através da combinação linear dos atributos, mostrando estruturas latentes.

Geralmente é utilizado antes de alguma outra análise dos dados, permitindo o uso das novas características criadas. Apesar de na maioria das vezes ser útil, este tipo de técnica não deixa de levar nenhum dos atributos originais em consideração. Pelo contrário, a informação das dimensões irrelevantes é preservada, fazendo com que estas técnicas não sejam eficientes para revelar clusters, por exemplo, quando há um grande número de atributos irrelevantes.

Outra desvantagem de se usar combinações dos atributos é a dificuldade de interpretação. Por causa disto, transformações das características são mais adequadas para conjuntos de dados que apresentem a maioria das dimensões sendo relevantes para uma análise posterior, mas muitas delas sendo redundantes ou altamente correlacionadas.

Seleção das Características

Técnicas de seleção das características têm como objetivo descobrir os atributos de um conjunto de dados que são mais relevantes para uma análise posterior. São técnicas poderosas e altamente utilizadas para a redução da dimensionalidade para níveis mais manejáveis.

A seleção das características consiste em procurar através de vários subconjuntos de características e avaliar cada um destes subconjuntos utilizando algum critério, como os de Pena et al (2001) ou de Yu e Liu (2003). As estratégias mais comuns de procura são as buscas sequenciais através do espaço de características podendo ser feita tanto para frente (forward) quanto para trás (backward).

ANÁLISE DE CLUSTER EM IMAGENS

Uma boa alternativa quando trabalhamos com dados com altas dimensões e necessitamos de alguma forma agrupar este tipo de dados é a Análise de Cluster em imagens, que pode ser realizada tanto com os dados brutos ou padronizados quanto com os dados após a realização de uma redução de dimensões. Neste trabalho serão apresentadas apenas algumas técnicas de cluster (agrupamento) baseadas em grids.

Em sua forma mais básica, agrupamento baseado em grids é relativamente simples:

a) Divida o espaço de amplitude dos dados em (hiper) células retangulares, por exemplo, particionando todo o intervalo de dados de cada dimensão em células de tamanhos iguais.

b) Descarte as células menos densas. Isto resulta em uma definição de densidade baseada em clusters, isto é, regiões altamente densas representam clusters, enquanto que regiões menos densas representam ruído. Geralmente esta é uma boa suposição,

entretanto, quando os clusters apresentam densidades muito diferentes, podemos ter problemas com este tipo de abordagem.

c) Combine células adjacentes com alta densidade para formar clusters. Se as regiões mais densas são adjacentes, então elas podem ser juntadas para formar um único cluster.

Existem algumas preocupações óbvias com relação aos métodos de agrupamento baseados em grids. Como os grids criados são quadrados ou retangulares, em muitos casos eles não irão se encaixar perfeitamente no formato do cluster. Para resolver este tipo de problema, pode-se aumentar o número de grids, de forma a deixá-los menores e assim aproximar-se da forma real do cluster. Entretanto, há um preço a se pagar por isto, e neste caso, será o aumento do tempo de trabalho computacional. Outro problema que pode surgir a partir deste aumento de grids é a aparição de “buracos” dentro dos clusters, devido ao tamanho muito reduzido que estes irão assumir, ainda mais se estivermos trabalhando com um tamanho de amostra não tão grande (resultando em uma menor quantidade de pontos).

Além do alerta descrito anteriormente a respeito deste tipo de técnica, existem sérios problemas quando a dimensionalidade dos dados aumenta muito. Para se perceber isto basta imaginar o caso em que cada dimensão é dividida em apenas 2 grids. Sendo d o número de dimensões presentes, teremos 2d células. Dados com 30 dimensões iriam, então, utilizar no mínimo 1 bilhão de células. Inclusive para grandes conjuntos de dados, a maioria das células ficaria vazia.

Outro problema é encontrar clusters entre as dimensões. Para compreender isto, imagine que cada ponto em um dos clusters é aumentado com muitas variáveis adicionais e que os valores atribuídos aos pontos nestas dimensões são uniformemente e aleatoriamente distribuídos. Então quase todos os pontos irão “cair” em células separadas no novo espaço alto dimensional. Assim, percebemos que grupos de pontos podem estar presentes em apenas alguns subespaços do modelo com altas dimensões.

TÉCNICAS DE AGRUPAMENTO DE DADOS

A seguir, serão mostradas quatro técnicas mais comuns de agrupamento de dados baseado em grids que estão disponíveis: CLIQUE, MAFIA, DENCLUE e OptiGrid.

CLIQUE

CLIQUE, como pode ser visto em Agrawal et al (1998), é um algoritmo para agrupamento que tenta lidar com os problemas citados anteriormente e que tem a sua abordagem baseada na seguinte observação: uma região que é densa em um determinado subespaço deve criar regiões densas quando projetado em dimensões menores.

Ao iniciar com intervalos unidimensionais densos, é possível encontrar potenciais intervalos bidimensionais densos e, ao inspecionar estes, encontrar os verdadeiros intervalos. Este procedimento pode ser estendido para se encontrar regiões densas em qualquer subespaço de forma muito mais eficiente do que formar células correspondentes a todos os possíveis subespaços de dimensões e então procurar pelas unidades densas nestas células. Porém, o CLIQUE ainda necessita de habilidade de descobrir regiões densas para reduzir os subespaços investigados. Além disso, sua complexidade computacional, mesmo sendo linear no número de pontos de dados, é não-linear no número de dimensões.

MAFIA

MAFIA (Merging Adaptative Finite Intervals And is more than a clique), como pode ser visto em Harasha et al (1999), é um refinamento da abordagem CLIQUE. Esta técnica encontra melhores clusters e alcança maior eficiência ao utilizar grids não uniformes.

Mais especificamente, ao invés de arbitrariamente dividir os dados em intervalos pré-determinados e pré-espaçados, MAFIA particiona cada dimensão usando um número variável de intervalos que se adaptam e que melhor refletem a distribuição dos dados naquela dimensão.

Conceitualmente, MAFIA começa com um grande número de pequenos intervalos para cada dimensão e então combina intervalos adjacentes de densidades similares para terminar com um menor número de intervalos maiores. Assim, um grid utilizando a abordagem MAFIA é bem representado pelo grid.

DENCLUE

Uma diferente abordagem do mesmo problema é fornecida pelo método DENCLUE (DENsity CLUstEring), em estudo feito por Hinnenburg e Keim (1998). O DENCLUE é um agrupamento por densidade que leva em conta uma abordagem mais formal ao método baseado pelas densidades para modelar a densidade total de um conjunto de pontos como a soma de funções de “influência” associadas a cada ponto. A função de densidade total terá picos locais, como por exemplo, local de densidade máxima e, a partir destes picos locais, chegaremos aos clusters de forma simples.

Especificamente, para cada ponto de dado, encontraremos o pico mais próximo associado a ele. O conjunto de todos os pontos associados a um particular pico (chamado de densidade atratora local) se torna um cluster. Entretanto, se a densidade em um pico local é muito baixa, então os pontos no cluster associado a este pico são considerados ruídos e então descartados. Além disso, se um pico local puder ser conectado a um segundo pico local por um trajeto de pontos e a densidade em cada ponto deste trajeto é maior do que um limiar mínimo, então os clusters associados a estes pontos se fundem. Com isso, clusters com qualquer formato podem ser encontrados.

OptiGrid

Apesar das características atraentes do DENCLUE em espaços com poucas dimensões, esta abordagem não lida tão bem com os dados a medida que a dimensão aumenta ou quando há presença de ruído. Por isso, os mesmos criadores do DENCLUE, Hinnenburg e Keim (1999) desenvolveram o OptiGrid.

O algoritmo descrito pelos autores segue seis passos, que serão resumidamente citados a seguir:

1) Para cada dimensão:

a) Faça um histograma dos dados. Note que isto é equivalente a contar os pontos em um grid uniforme unidimensional impostos nos valores;

b) Determine o nível de ruído. Isto pode ser feito através de uma inspeção manual do histograma, se a dimensionalidade não é muito alta. Em caso contrário, este processo necessita ser automatizado;

c) Encontre o ponto máximo mais à direita e mais à esquerda e o q-1 máximo entre eles (onde q é o número de partições dos dados que nós procuramos e todas estas partições podem estar em uma dimensão);

d) Escolha q, o mínimo entre os máximos encontrado no passo anterior. Estes pontos representam localizações para possíveis cortes, isto é, localizações onde o hiperplano pode ser posicionado para particionar os dados. Escolher células pouco densas minimiza a chance de cortes através de um cluster;

e) Dê um escore para cada corte potencial, por exemplo, pela sua densidade.

2) De todas as dimensões, selecione os melhores q cortes, isto é, aqueles cortes com menores densidades.

3) Utilizando estes cortes, crie um grid que particiona os dados.

4) Encontre as células mais densas e adicione elas para a lista de clusters.

5) Refine a lista de clusters.

6) Repita as etapas 1-5 utilizando cada cluster.

Em resumo, a abordagem OptiGrid é semelhante à MAFIA no sentido de que cria um grid utilizando uma partição dependente dos dados. Entretanto, o OptiGrid não se preocupa em localizar o melhor subespaço para usar esta partição. Ele localiza potenciais clusters entre o conjunto de células formadas através do seu plano de cortes. Com relação à eficiência, este tipo de abordagem é muito melhor.

Porém, também existem alguns problemas neste tipo de técnica como, por exemplo, o fato do número de cortes necessários ser bastante vago. Com o objetivo de solucionar este tipo de problema, novos tipos de abordagens estão sendo desenvolvidas, como o PDDP (Power-Delay-Direction Profile).

REFERENCIAS

REDAÇÃO. **O que é Big Data?** Disponível em<:https://canaltech.com.br/big-data/o-que-e-big-data/>. Acesso em 03 de maio de 2020.

BATINI, Carlo; SCANNAPIECA, Monica. **Data Quality:** Concepts, Methodologies and Techniques. New York. Springer, 2006

BATINI, Carlo et al. **Methodologies for Data Quality Assessment and Improvement**. ACM ComputingSurveys, n.3, v.41, 2009, p. 1-52.

BATINI, Carlo. et al. **From Data Quality to Big Data Quality**. Journal of Database Management, v. 26, n. 1, 2015, p. 60–82.

BECKER, David; MCMULLEN, Bill; KING, Trish Dunn. Big Data, Big Data Quality Problem. **In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA,** 2015, Santa Clara. Anais eletrônicos... Santa Clara: 2015. p.2644-2653

CAI, Li; ZHU, Yangyong. **The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.** Data Science Journal, v. 14, n. 0, 2015, p. 2. Dis

CIANCARINI, Paolo; POGGI, Francesco; RUSSO, Daniel. Big Data Quality**: a Roadmap for Open Data. 2ND IEEE INTERNATIONAL CONFERENCE ON BIG DATA COMPUTING SERVICE AND APPLICATIONS**, 2., 2016, Oxford. Anais eletrônicos… Praga: 2016.

DATE, C.J.; Int. a **Sistemas de Bancos de Dados**, tradução da 4a.edição norte-americana, Editora Campus, 1991.

ENDLER, Gregor; BAUMGAERTEL, Philipp; LENZ, Richard. Pay-as-you-go data quality improvement for medical centers. **In: CONFERENCE ON EHEALTH - HEALTH INFORMATICS MEETS EHEALTH**, 2013, Vienna. Anais eletrônicos… Vienna: 2013.

RDBCI: **Revista Digital Biblioteconomia e Ciência da Informação RDBCI** : Digital Journal of Library and Information Science DOI 10.20396/rdbci.v16i1.8650412 © RDBCI: Rev. Digit. Bibliotecon. Cienc. Inf. Campinas, SP v.16 n.1 p. 194-210 jan./abr. 2018 [207] p.13-18. Disponível em: . Acessoem: 7 jul. 2017.

ERL, Thomas; KHATTAK, Wajid; BUHLER, Paul. **Big Data Fundamentals:** Concepts, Drivers & Techniques. Boston: Prentice Hall, 2016.

FREITAS, Patrícia Alves de et al. **Information Governance, Big Data and Data Quality.** In: IEEE 16TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND ENGINEERING (CSE), 16., 2013, Sydney. Anais eletrônicos… Sydney: 2013. p.1142- 1143.

FURLAN, Patricia Kuzmenko; LAURINDO, Fernando José Barbin. **Agrupamentos epistemológicos de artigos publicados sobre big data analytics.** Transinformação, v. 29, n. 1, 2017, p. 91-100.

GANAPATHI, Archana; CHEN, Yanpei. **Data Quality: Experiences and Lessons from Operationalizing Big Data**. 4TH IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 4., 2016, Washington. Anais eletrônicos… Washington: 2016.

GANDOMI, Amir; HAIDER, Murtaza. **Beyond the hype: Big data concepts, methods, and analytics.** International Journal of Information Management, v. 35, n. 2, 2015, p. 137– 144.

HARYADI, AdiskaFardaniet al. **Antecedents of Big Data Quality An Empirical Examination in Financial Service Organizations.** 4TH IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 4., 2016, Washington. Anais eletrônicos… Washington: 2016.

HAZEN, Benjamin T. et al. **Data quality for data science, predictive analytics, and big data in supply chain management:** An introduction to the problem and suggestions for research and applications. International Journal of Production Economics, v. 154, 2014, p. 72–80.

JUDDOO, Suraj. **Overview of data quality challenges in the context of Big Data.** In: INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND SECURITY (ICCCS), 2015, Pamplemousses. Anais eletrônicos… Pamplemousses : 2015..

KAISLER, Stephen et al. **Big Data: Issues and Challenges Moving Forward**. In: XLVI HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 46., Maui, 2013.

KELLING, Steve et al. **Taking a `Big Data' approach to data quality in a citizen science project**. AMBIO, v. 44, n. 4, 2015, p. S601–S611.

KWON, Ohbyung; LEE, Namyeon; SHIN, Bongsik. **Data quality management, data usage experience and acquisition intention of big data analytics.** International Journal of Information Management, v. 34, n. 3, 2014, p. 387–394.

LANEY, Doug. **Application Delivery Strategies.** META Group, 2001. Disponível em: . Acessoem: 7 jul. 2017.

KORTH, H.F. e SILBERSCHATZ, A.; **Sistemas de Bancos de Dados,** Makron Books, 2a. edição revisada, 1994.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big Data. **The management revolution.** Harvard BuinessReview, v. 90, n. 10, 2012 p. 61–68.

MERINO, Jorge et al.**A Data Quality in Use model for Big Data.** Future Generation Computer Systems, v. 63, 2016, p.123-130.

PAIM, Isis; NEHMY, Rosa Maria Quadros, GUIMARÃES, César Geraldo. **Problematização do conceito "Qualidade" da Informação**. Perspectivas em Ciência da Informação, v. 1, n. 1, 1996, p. 111–119.

PORTAL DE PERIÓDICOS DA CAPES/MEC. 07 jun. 2017.

RAO, Dhana; GUDIVADA, Venkat N.; RAGHAVAN, Vijay V. **Data Quality Issues in Big Data.** In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2015, Santa Clara. Anais eletrônicos... Santa Clara: 2015.

RIBEIRO, Claudio José Silva. **Big Data:** os novos desafios para o profissional da informação. Informação & Tecnologia, v. 1, n. 1, 2014, p. 96–105.

RDBCI: Revista Digital Biblioteconomia e Ciência da Informação RDBCI : **Digital Journal of Library and Information Science** DOI 10.20396/rdbci.v16i1.8650412 © RDBCI: Rev. Digit. Bibliotecon. Cienc. Inf. Campinas, SP v.16 n.1 p. 194-210 jan./abr. 2018 [209]

SADIQ, Shazia; PAPOTTI, Paolo. Big Data Quality - Whose problem is it? 32ND **IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE**), 32., 2016, Helsinki. Anais eletrônicos… Helsinki: 2016.

SAHA, Barna; SRIVASTAVA, Divesh. Data Quality: The other Face of Big Data. **In: IEEE 30TH INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE)**, 30., 2014, Chicago. Anais eletrônicos…Chicago: 2014.

SOMASUNDARAM, G.; SHRIVASTAVA, Alok**. Armazenamento e Gerenciamento de Informações:** Como armazenar, gerenciar e proteger informações digitais. Porto Alegre: Bookman. 2011. 472p.

TALEB, Ikbalet al. **Big Data Quality: A Quality Dimensions Evaluation**. 13TH IEEE INT CONF ON UBIQUITOUS INTELLIGENCE AND COMP, 13., 2016, Toulouse. Anais eletrônicos… Toulouse: 2016. Disponível em: . Acesso em: 7 jul. 2017.

VALENTE, Nelma T. Zubek; FUJINO, Asa. **Atributos e dimensões de qualidade da informação nas Ciências Contábeis e na Ciência da Informação**: um estudo comparativo. Perspectivas em Ciência da Informação, v. 21, n. 2, 2016, p. 141–167. Disponível em: . Acesso em: 16 mar. 2017.

VIANNA, William Barbosa; DUTRA, Moisés Lima; FRAZZON, Enzo Morosini. **Big data e gestão da informação: modelagem do contexto decisional apoiado pela sistemografia**. Informação & Informação, v. 21, n. 1, 2016, p. 185.