TI TOTAL

ÁREA FISCAL E CONTROLE

Professor
Ramon Souza

Tecnologia da Informação

TEORIA

NOÇÕES DE BIG DATA

SUMÁRIO

A nossa aula é bem esquematizada, então para facilitar o seu acesso aos **esquemas**, você pode usar o seguinte índice:

GLOSSÁRIO DE TERMOS

Big Data: conjunto de dados muito grandes ou complexos ou processo de captura, gerenciamento e a análise de dados que vão além dos dados tipicamente estruturados.

Código aberto: código fonte disponibilizado e licenciado com uma licença de código aberto no qual o direito autoral fornece o direito de estudar, modificar e distribuir o software de graça para qualquer um e para qualquer finalidade.

Consulta ad-hoc: consulta sob demanda, sem planejamento prévio.

Data Lake: único repositório dentro de uma empresa com todos os dados brutos.

Predição: sinônimo de previsão.

Processamento paralelo: processamento por várias máquinas ao mesmo tempo.

Variável: representa um elemento arbitrário, não totalmente especificado ou desconhecido

1. NOÇÕES DE BIG DATA

1.1 Conceito de Big Data

Você já parou para pensar na quantidade de dados e informações existentes e que produzidos diariamente seja no âmbito pessoal ou profissional. São produzidos diariament textos, imagens, vídeos, e-mails e muitos outros conteúdos que ampliam a quantidad enorme de dados já existente. Essa imensidão de dados torna inviável o processamento pelas pessoas, isto é, é impossível que o ser humano consiga capturar, analisar, corrigir, pesquisar, compartilhar, gerenciar toda essa grande quantidade de dados e informações.

A análise adequada dos grandes conjuntos de dados permite encontrar novas correlações, como por exemplo: "tendências de negócios no local, prevenção de doenças, combat criminalidade e assim por diante". Cientistas, empresários, profissionais de mídia e publicidade e Governos regularmente enfrentam dificuldades em áreas com grandes conjuntos de dados, incluindo pesquisa na Internet, finanças e informática de negócios. Os cientistas, por exemplo, encontram limitações no trabalho de e-Ciência, incluindo Meteorologia, Genômica, simulações físicas complexas, além de pesquisa biológica e ambiental. Neste contexto, surge o conceito de Big Data.

Big Data é um termo amplamente utilizado na atualidade para nomear **conjuntos de dad muito grandes ou complexos**, que os **aplicativos de processamento de dados tradicionais ainda não conseguem lidar**. Os desafios desta área incluem: análise, captura curadoria de dados, pesquisa, compartilhamento, armazenamento, transferência, visualização e informações sobre privacidade dos dados. Este termo muitas vezes se refere ao **uso de análise preditiva e de alguns outros métodos avançados para extrair valor de dados**. Maior precisão nos dados pode levar à tomada de decisões com mais confiança. Além disso, melhores decisões podem significar maior eficiência operacional, redução de risco e redução de custos.

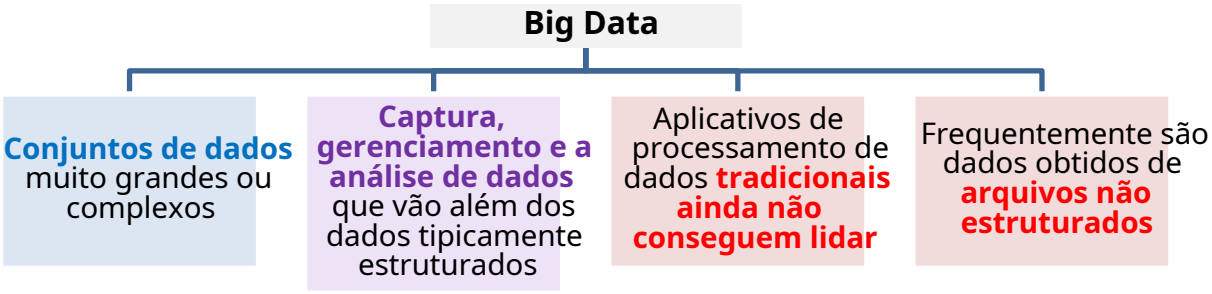
O **Big Data** pode ser definido genericamente como a **captura, gerenciamento e a análise de dados que vão além dos dados tipicamente estruturados**, que podem ser consultado e pesquisados através de bancos de dados relacionais. **Frequentemente são dados obtid de arquivos não estruturados** como vídeo digital, imagens, dados de sensores, arquivos de logs e de qualquer tipo de dados não contidos em registros típicos com campos que podem ser pesquisados.

O **Big Data** possui **variadas fontes de dados** como:

- Dados gerados pelas máquinas (redes de sensores, logs).
- Dispositivos móveis (vídeo, mensagens, fotografias).
- Comunicação máquina a máquina, a “Internet das coisas”.
- Dados em bancos de dados relacionais oriundos das transações da organização.
- Imagens de documentos, etc.

O **objetivo do Big Data** é **propiciar dados e informações que possam ser analisados visando subsidiar tomadas de decisão**.

A primeira informação que você precisa levar para a prova é o conceito de Big Data, então vamos fixá-lo com um **esquema**.



Esquema 1 – Conceito de Big Data.

Com base nos conceitos apresentados, vemos que o **Big Data** tanto pode ser encarado com o **grande volume de dados estruturados e não estruturados que são gerados a cada segundo** ou também como as **tecnologias que são utilizadas para lidar com este grand volume de dados**. Esta segunda acepção é, por vezes, chamada de **Big Data Analytics**

Big Data Analytics é o **trabalho analítico e inteligente de grandes volumes de dados**, estruturados ou não-estruturados, que são **coletados, armazenados e interpretados por softwares de altíssimo desempenho**. Trata-se do cruzamento de uma infinidade de dado do ambiente interno e externo, gerando uma espécie de "bússola gerencial" para tomador de decisão. Tudo isso, é claro, em um **tempo de processamento extremamente reduzido** Pode ser considerada um desdobramento de Big Data.

ESCLARECENDO!!!

Qual a relação de Big Data com Data Mining?

Uma das diferenças entre o data mining e o **Big Data**, é que este segundo engloba a anális de dados não estruturados. Assim, o **Big Data** pode ser encarado como um verdadeiro dat mining em larga escala incluindo dados de um grande volume de fontes, podendo estes dados serem estruturados, semiestruturados ou não estruturados.

E do Big Data com os Data Warehouses?

As ferramentas de **Big Data** podem ser utilizadas para analisar dados presentes também em Data Warehouses, além de outras fontes. Assim, o DW é um repositório de dados que pode ser analisado por ferramentas de data mining ou de **Big Data**.

E do Big Data com Ciência de Dados?

A ciência de dados é o estudo dos dados para extrair insigths significativos para os negócio É uma abordagem multidisciplinar que combina princípios e práticas de matemática, estatística, inteligência artificial e computação para analisar grandes quantidades de dados **Big data** pode ser utilizada pelos cientistas de dados para realizar seu trabalho.

ATENÇÃO!!!

Um conceito recente relacionado à Big Data é o de **Data Lake**. Data Lake é um termo criado pelo CTO (Chief Technical Officer) do Pentaho, James Dixon, para descrever um componente importante no universo da análise de dados e do Big Data. A ideia é ter um **único repositório dentro da empresa**, para que **todos os dados brutos** estejam disponíveis a qualquer pessoa que precise fazer análise sobre eles. O Data Lake não requer que os usuários criem um esquema antes de preparar os dados para armazenamento. Os dados podem ser simplesmente consumidos e o esquema criado e aplicado quando os dados forem usados para análise.

1- (CESPE / CEBRASPE - 2022 – TCE-SC – Auditor de Controle Externo – Ciências da Computação) No que diz respeito a big data e à L ei n . º 12.527/2011 e suas alterações (Lei de Acesso à Informação), julgue o item seguinte.

Big data necessitam de algoritmos de computação mais robustos em comparação aos algoritmos tradicionais de banco de dados, que não são capazes de lidar com os volumes de dados representados em big data.

Resolução:

Big Data é um termo amplamente utilizado na atualidade para nomear **conjuntos de dados muito grandes ou complexos**, que os **aplicativos de processamento de dados tradicionais ainda não conseguem lidar**. Os desafios desta área incluem: análise, captura, curadoria de dados, pesquisa, compartilhamento, armazenamento, transferência, visualização e informações sobre privacidade dos dados.

Gabarito: **Certo**.

2- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Tecnologia da Informação) Com relação a Big Data, julgue o item seguinte.

Os fatores críticos de sucesso da análise de Big Data incluem uma sólida infraestrutura de dados, além de ferramentas analíticas e pessoal habilitado para lidar com elas.

Resolução:

Pessoal, essa questão é bem intuitiva. Logicamente que um Big Data precisa de uma sólida infraestrutura de dados para poder acomodar o grande volume de dados, além de permitir dados de variadas fontes.

Além disso, o big data vai permitir a análise de dados, então é fundamental a existência de ferramentas especializadas e pessoal capaz de usar essas ferramentas.

Gabarito: **Certo**.

3- (COMPERVE - 2020 - TJ-RN - Analista de Suporte Pleno - Banco de Dados)
Big Data surgiu a partir da necessidade de manipular um grande volume de dados e, com isso, novos conceitos foram introduzidos, como o Data Lake, que

a) pode ser considerado um repositório de dados relacionados, sendo, portanto, um armazém de dados orientado por assunto.

b) pode ser considerado um conjunto de bancos de dados relacionais e com relacionament entre tabelas de diferentes esquemas de bancos de dados.

c) é o resultado de sucessivas operações de mineração de dados, sendo um ambiente no q é possível ter relatórios e dashboards de maneira amigável para os analistas de negócio.

d) é projetado para armazenar dados de diversas fontes e formatos, não havendo a necessidade da definição de um esquema de dados para inserir novos itens.

Resolução:

Data Lake é um termo criado pelo CTO (Chief Technical Officer) do Pentaho, James Dixon, para descrever um componente importante no universo da análise de dados e do Bi Data. A ideia é ter um **único repositório dentro da empresa**, para que **todos os dados brutos** estejam disponíveis a qualquer pessoa que precise fazer análise sobre eles. O Data Lake não requer que os usuários criem um esquema antes de preparar os dados pa armazenamento. Os dados podem ser simplesmente consumidos e o esquema criado aplicado quando os dados forem usados para análise.

Dito isto, vamos analisar cada um dos itens:

a) **Incorreto**: o Data Lake não é organizado por assunto.

b) **Incorreto**: o Data Lake não é um banco relacional.

c) **Incorreto**: o Daka Lake não é resultado de mineração.

d) **Correto**: o Data Lake é um conjunto de dados sem definição prévia de esquema e com dados brutos de diferentes fontes.

Gabarito: Letra D.

1.2 Tipos de análise com Big Data

Há quatro tipos de análise de Big Data que se destacam pela usabilidade e potencialidade de seus resultados: análise preditiva, análise prescritiva, análise descritiva e análise diagnóstica.

Vejamos um pouco sobre cada uma delas:

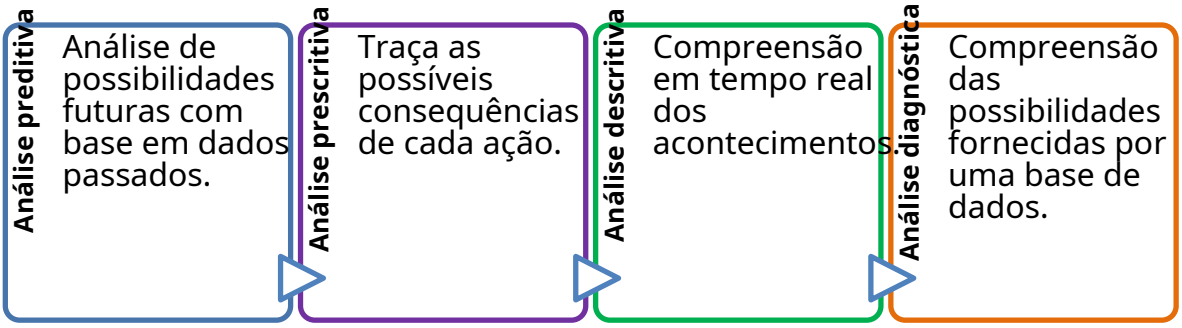
- **Análise preditiva**: **análise de possibilidadesfuturas**. A partir da **identificação de padrões passados** em sua base dados, esse tipo de anális permite aos gestores o mapeamento de possíveis futuros em seus campos atuação. A ideia é deixar de tomar decisões baseadas unicamente na intuição, conseguindo estabelecer um prognóstico mais sólido para cada ação. Responde à pergunta "**O que vai acontecer?**".
  - **Exemplo**: há uma probabilidade de 80% de que percamos X clientes no próximo mês.

- **Análise prescritiva**: **traça as possíveis consequências de cada ação**. É uma forma de definir qual escolha será mais efetiva em determinada situação. O valor dessa análise se dá pela capacidade de numerar determinados padrões e filtrá-lo por especificidades, obtendo um cenário bastante fiel da situação e como cada intervenção responderá. Responde à pergunta "**O que pode acontecer se tomarmos essa medida?**".
  - **Exemplo**: se dermos 3% de desconto aos clientes X no próximo mês, a chance de os perdermos cai para 30%.

- **Análise descritiva**: **compreensão em tempo real dos acontecimentos**. É uma maneira de visualizar os dados, entender como um banco de dados se organiza e o que significa para o presente sem necessariamente relacioná-la com padrões passados ou futuros. Responde à pergunta "**O que está acontecendo?**".
  - **Exemplo**: a empresa pode analisar dados sobre perda de clientes ou mau desempenho em vendas de um determinado produto.

- **Análise diagnóstica**: **compreensão de maneira causal das possibilidades fornecidas por uma base de dados**. Como uma espécie de relatório expandido, quando feita em uma base de dados volumosa, esse tipo de análise permite ainda entender a razão de cada um dos desdobramentos das ações adotadas e, a partir disso, mudar estratégias ineficazes ou reforçar as funcionais. Responde à pergunta "**Por que aconteceu?**".
  - **Exemplo**: chegar à conclusão de que a perda de clientes ocorreu porque o preço dos produtos da companhia estava alto quando comparados com a concorrência.

Vejamos um esquema sobre os tipos de análise em Big Data.



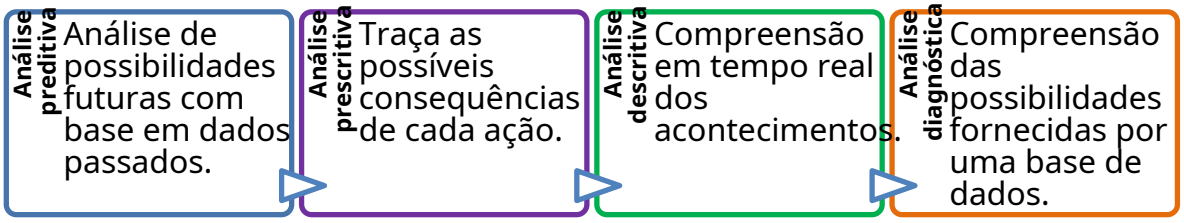
Esquema 2 – Tipos de análises com Big Data.

4- (CESPE / CEBRASPE - 2022 - TCE-RJ – Analista de Controle Externo – Organizacional – Tecnologia da Informação) Acerca dos conceitos de mineração de dados, aprendizado de máquina e bigdata, julgue o próximo item.

A análise prescritiva é empregada na análise de Big Data para relatar acontecimentos e par fazer previsões de comportamentos futuros de indivíduos e processos.

Resolução:

Assertiva se refere a análise preditiva e não a prescritiva. Quanto aos tipos de análise.



Gabarito: **Errado**.

5- (CESPE - 2018 - TCE-MG - Analista de Controle Externo - Ciência da Computação) Uma empresa, ao implementar técnicas e softwares de big data, deu enfoqu diferenciado à análise que tem como objetivo mostrar as consequências de determinado evento.

Essa análise é do tipo

a) preemptiva. b) perceptiva. c) prescritiva.

d) preditiva. e) evolutiva.

Resolução:

A **análise prescritiva** **traça as possíveis consequências de cada ação**. É uma forma de definir qual escolha será mais efetiva em determinada situação. O valor dessa análise se dá pela capacidade de numerar determinados padrões e filtrá-los por especificidades, obtend um cenário bastante fiel da situação e como cada intervenção responderá.

Gabarito: Letra C.

1.3 Premissas de Big Data

O **Big Data** foi inicialmente conceituado com base em três **premissas básicas**, também conhecidas como 3Vs. Vejamos:

- **Volume**: o Big Data deve possibilitar a **análise de grandes volumes de dados**. Além disso, a tecnologia do Big Data serve exatamente para **lidar com esse volume de dados, guardando-os em diferentes localidades e juntando-os através de software**.
  - O conceito de **volume** no Big Data é melhor evidenciado pelos fatos do cotidiano: diariamente, o volume de troca de e-mails, transações bancárias, interações em redes sociais, registro de chamadas e tráfego de dados em linhas telefônicas. Todos esses servem de ponto de partida para a compreensão do volume de dados presentes no mundo atualmente.
  - Estima-se que em 2014 o volume total de dados que circulavam na internet era de 250 Exabytes (10^{18} bytes) por ano e que todos os dias são criados 2,5 quintilhões de bytes em forma de dados, atualmente 90% de todos os dados que estão presentes no mundo foram criados nos últimos 2 anos. É importante também compreender que o conceito de **volume é relativo a variável tempo, ou seja, o que é grande hoje, pode não ser nada amanhã**. Nos anos 90, um Terabyte (10^{12} bytes) era considerado Big Data. Em 2015, o mundo contava com aproximadamente um volume de informação digital de 8 Zettabytes (10^{21} bytes), um valor infinitamente maior. Já em 2017, estima-se 20 Zettabytes (10^{21} bytes), podendo chegar 40 Zettabytes (10^{21} bytes) em 2020.

- **Velocidade**: o Big Data deve fornecer as **repostas com velocidade e em tempo hábil**. O Big Data serve para **analisar os dados no instante em que são criados**, **sem ter de armazená-los em bancos de dados**.
  - Você cruzaria uma rua vendado se a última informação que tivesse fosse uma fotografia tirada do tráfego circulante de 5 minutos atrás? Provavelmente não, pois a fotografia de 5 minutos atrás é irrelevante, você precisa saber das condições atuais para poder cruzar a rua em segurança. A mesma lógica se aplica a empresas, pois necessitam de dados em atuais sobre seu negócio, ou seja, velocidade.
  - Informação é poder, e assim sendo a velocidade com a qual você obtém essa informação é uma vantagem competitiva das empresas. Velocidade pode limitar a operação de muitos negócios, quando utilizamos o cartão de crédito por exemplo, se não obtivermos uma aprovação da compra em alguns segundos normalmente pensamos em utilizar outro método de pagamento.

- **Variedade**: o Big Data deve ser capaz de lidar com **diferentes formatos de informação**. Os dados podem estar em fontes estruturadas, semi-estruturadas e a grande maioria em fontes não estruturadas.
    - Já pensou na quantidade de informações dispersas em redes sociais? Facebook, Twitter entre outros possuem um vasto e distinto campo de informações sendo ofertadas em público a todo segundo. Podemos observar a variedade de dados em e-mails, redes sociais, fotografias, áudios, telefones e cartões de crédito. Seja qual for a discussão, podemos obter infinitos pontos de vista sobre a mesma. Empresas que conseguem captar a variedade, seja de fontes ou de critérios, agregam mais valor ao negócio.

Ao acrescentar as premissas da Veracidade e Valor, temos o que chamamos de 5 Vs:

- **Veracidade**: um dos pontos mais importantes de qualquer **informação é que ela seja verdadeira**. Com base nas análises e estatísticas de grandes volumes de dados é possível compensar as informações incorretas. Á medida que o volume, a velocidade e a variedade aumentam, a veracidade (confiança ou confiança nos dados diminui. A veracidade refere-se mais à **proveniência ou à confiabilidade da fonte de dados**, seu contexto e a sua utilidade para a análise com base nela.

- **Valor**: os dados do Big Data **devem agregar valor ao negócio**. Sem valor, a informação não tem utilidade. Um valor substancial pode ser encontrado em grandes dados, incluindo a melhor compreensão de seus clientes, direcionando as suas necessidades, otimizando processos e melhorando o desempenho do negócio. importante estar **atento aos custos envolvidos nessa operação**, pois o valor agregado de todo esse trabalho desenvolvido na coleta, armazenamento e análise d todos esses dados tem que compensar os custos financeiros envolvidos.

Além dessas, temos outras premissas trazidas na literatura. Vejamos algumas delas:

- **Variabilidade**: pode ter diferentes sentidos; (1) **número de inconsistências** nos dados; (2) **multiplicidade de dimensões** de dados resultantes de diferentes fontes de dados (também referida como complexidade); (3) **velocidade inconstante** n carga dos dados.

- **Validade**: Semelhante à veracidade, a validade refere-se à **precisão e correção dos dados para o uso pretendido**.

- **Vulnerabilidade**: Big Data traz **novas preocupações de segurança**. Afinal, uma violação de dados em um Big Data é uma grande violação.

- **Volatilidade**: devido à velocidade e ao volume do Big Data, sua volatilidade precisa ser cuidadosamente considerada. É necessário **estabelecer regras para a atualidade e a disponibilidade de dados**, bem como assegurar a recuperação rápida de informações quando necessário.

- **Visualização**: as principais ferramentas de visualização de Big Data enfrentam desafios técnicos devido às limitações da tecnologia na memória e uma escalabilidade, funcionalidade e tempo de resposta fracos. Não se pode confiar em gráficos tradicionais ao tentar traçar um bilhão de pontos de dados, então são necessárias **maneiras diferentes de representar dados** como o agrupamento de dados ou usando mapas de árvores, coordenadas paralelas, diagramas de rede circulares ou árvores de cone.

- **Viscosidade**: mede a **resistência ao fluxo no volume de dados**. Essa resistência pode vir de diferentes fontes de dados, atrito de taxas de fluxo de integração processamento necessário para transformar os dados em insight.

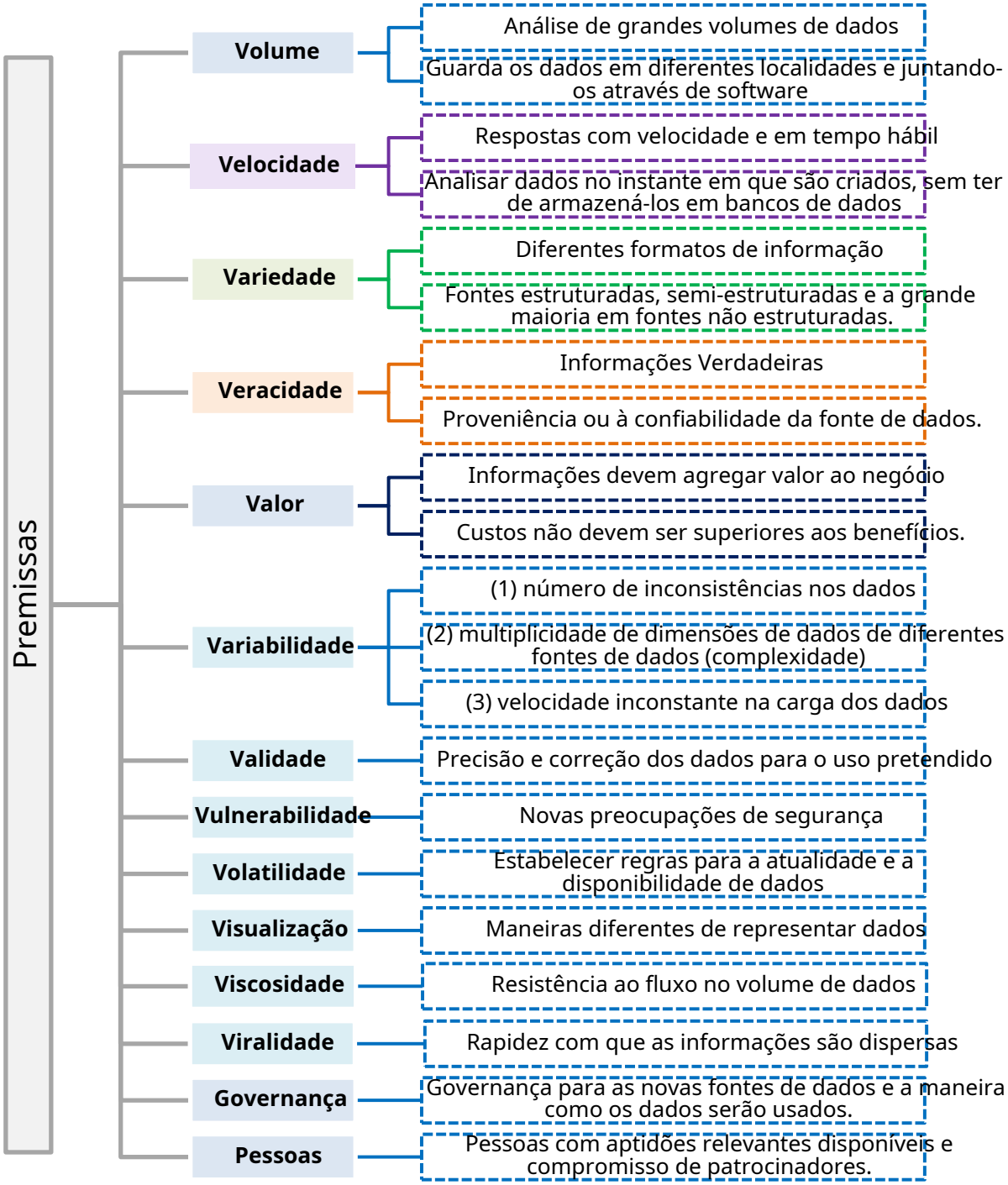
- **Viralidade**: **rapidez com que as informações são dispersas** entre redes de pessoas.

A IBM utiliza, ainda, juntamente com volume, variedade, velocidade, veracidade e valor, a Governança e Pessoas como premissas para avaliar a viabilidade de uma solução de Big Data:

- **Governança**: considerações sobre **governança para as novas fontes de dados e a maneira como os dados serão usados**. Ao decidir implementar ou não uma plataforma de big data, uma organização pode estar olhando novas fontes e novos tipos de elementos de dados nos quais a propriedades não estejam definidas de forma clara. Alguns regulamentos do segmento de mercado regem os dados que são adquiridos e usados por uma organização. Além da questão da governança de TI, também pode ser necessário redefinir ou modificar os processos de negócios de uma organização para que ela possa adquirir, armazenar e acessar dados externos.

- **Pessoas**: **pessoas com aptidões relevantes disponíveis e compromisso de patrocinadores**. É necessário ter aptidões específicas para entender e analisar os requisitos e manter uma solução de big data. Essas aptidões incluem conhecimento do segmento de mercado, conhecimento do domínio e conhecimento técnico sobre ferramentas e tecnologias de big data. Cientistas de dados com conhecimento em modelagem, estatística, analítica e matemática são essenciais para o sucesso de qualquer iniciativa de big data.

Para consolidar as diversas premissas estudadas, vejamos um **esquema**!!!



Esquema 3 – Técnicas de pré-processamento (Navathe).

6- (CESPE / CEBRASPE - 2022 – TCE-RJ – Analista de Controle Externo – Organizacional – Tecnologia da Informação) Julgue o item subsequente, referentes a Big Data e visualização e análise exploratória de dados.

Uma vez que Big Data engloba um grande volume e variedade de dados, o atributo veracidade tem sido inserido nas premissas do conceito para avaliar a confiabilidade e a consistência dos dados da solução.

Resolução:

Em relação **veracidade**, um dos pontos mais importantes de qualquer **informação é que ela seja verdadeira**. Com base nas análises e estatísticas de grandes volumes de dados é possível compensar as informações incorretas. Á medida que o volume, a velocidade e a variedade aumentam, a veracidade (confiança ou confiança nos dados) diminui. A veracidade refere-se mais à **proveniência ou à confiabilidade da fonte de dados**, seu contexto e a sua utilidade para a análise com base nela.

Gabarito: **Certo**.

7- (CESPE / CEBRASPE - 2021 - Polícia Federal - Escrivão de Polícia Federal) Acerca dos conceitos de mineração de dados, aprendizado de máquina e bigdata, julgue o próximo item.

As aplicações de bigdata caracterizam-se exclusivamente pelo grande volume de dados armazenados em tabelas relacionais.

Resolução:

Big data não se caracteriza exclusivamente pelo volume. Possui, ao menos três características fundamentais:

- Volume: o Big Data deve possibilitar a **análise de grandes volumes de dados**. Além disso, a tecnologia do Big Data serve exatamente para **lidar com esse volume de dados, guardando-os em diferentes localidades e juntando-os através de software**.

- Velocidade: o Big Data deve fornecer as **repostas com velocidade e em tempo hábil**. O Big Data serve para **analisar os dados no instante em que são criados, sem ter de armazená-los em bancos de dados**.

- Variedade: o Big Data deve ser capaz de lidar com **diferentes formatos de informação**. Os dados podem estar em fontes estruturadas, semi-estruturadas e a grande maioria em fontes não estruturadas.

Gabarito: **Errado**.

8- (FCC - 2020 - AL-AP - Analista Legislativo - Desenvolvedor de Banco de Dados) Atualmente, diversos dados são coletados pelos sistemas digitais de empresas na internet para constituir Big Data com conteúdo sobre os resultados alcançados por seus produtos e serviços, prestígio da imagem da organização e seus representantes. Porém parte desses dados pode ser falsa ou manipulada por internautas. O tratamento dos dados a fim de qualificá-los antes de disponibilizá-los para a tomada de decisão na empres segundo o conceito das cinco dimensões "V" de avaliação de um Big Data, se refere

a) ao valor.

b) à variedade.

c) à veracidade.

d) à velocidade.

e) ao volume.

Resolução:

Vamos analisar cada um dos itens:

a) **Incorreto**: o valor tem relação com o que o Big Data agrega ao negócio.

b) **Incorreto**: a variedade representa as diferentes fontes de informação.

c) **Correto**: a veracidade é a premissa que visa garantir que a informação seja verdadeira.

d) **Incorreto**: a velocidade tem relação com as respostas hábeis.

e) **Incorreto**: o volume diz respeito à grande quantidade de dados.

Gabarito: Letra C.

9- (COMPERVE - 2020 - TJ-RN - Analista de Suporte Pleno - Banco de Dados) Embora Big Data tenha diferentes definições, há um consenso sobre o modelo dos 3 V's qu correspondem a 3 características. Duas dessas características são:

a) Volume e Velocity.

b) Variety e Value.

c) Viable e Vast.

d) Valid e Verbose.

Resolução:

Os três Vs fundamentais do Big Data são: volume, velocidade e variedade.

Gabarito: Letra A.

1.4 Aplicações de Big Data

As **aplicações do Big Data** e da análise de dados são variadas como:

- Desenvolvimento de mercado.
- Inovação.
- Desenvolvimento de produtos e serviços.
- Eficiência operacional.
- Previsões de demanda de mercado.
- Detecção de fraudes.
- Gerenciamento de riscos.
- Previsão de concorrência.
- Vendas.
- Campanhas de *marketing*.
- Avaliação do desempenho de funcionários.
- Alocação do orçamento anual.
- Estabelecimento de previsões financeiras.
- Gestão de planos médicos.
- Identificação de potenciais compradores.
- Entendimento da base de clientes etc.

Big Data afeta organizações em praticamente todas as indústrias. Vejamos como alguns ramos podem se beneficiar:

- **Banco**: com grandes quantidades de informações fluindo partir inúmeras fontes, bancos são desafiados a encontrar maneiras novas e inovadoras de gerenciar big data. Ao mesmo tempo em que big data é importante para compreender os clientes e aumentar sua satisfação, é igualmente importante para minimizar os riscos e fraudes enquanto mantém uma conformidade regulatória. Big Data traz ótimos insights, mas exige que as instituições financeiras estejam um passo à frente neste jogo, com análises avançadas.

- **Governo**: quando as organizações governamentais são capazes de aproveitar e aplicar análises em big data, elas progridem significativamente quando se trata de gerenciar serviços públicos, lidar com o congestionamento ou a prevenir a criminalidade. Mas, enquanto existem muitas vantagens com o uso de big data, os governos também devem abordar as questões de transparência e privacidade das informações.

- **Manufatura**: armados com uma visão que big data pode fornecer, os fabricantes poder aumentar a qualidade e a produção, minimizando o desperdício - processos que fundamentais no mercado altamente competitivo de hoje. Mais e mais fabricantes estão trabalhando em uma cultura baseada em análise de dados, o que significa que eles pod resolver problemas mais rapidamente e tomar decisões de negócios mais ágeis.

- **Ensino**: educadores armados com uma visão orientada a dados podem ter um impacto significativo sobre os sistemas escolares, estudantes e currículos. Analisando big data, eles podem identificar alunos em risco, assegurar que os estudantes estão progredindo de forma adequada, e podem implementar um sistema melhor de avaliação e apoio aos professores e diretores.

- **Saúde**: Registros de pacientes. Planos de tratamento. Informações de prescrição. Quando se trata de cuidados com a saúde, tudo precisa ser feito rapidamente, c precisão e, em alguns casos, com suficiente transparência para satisfazer as regulamentações rigorosas desta indústria. Quando grandes quantidades de dados são geridas de forma eficaz, os prestadores de cuidados de saúde podem descobrir insights escondidos que melhoram o atendimento ao paciente.

- **Varejo**: A construção de relacionamento com o cliente é fundamental para o setor de varejo - e a melhor maneira de gerenciar este relacionamento é gerenciando big data. Os varejistas precisam saber a melhor maneira de vender aos clientes, a maneira mais eficaz de lidar com transações, e a maneira mais estratégica de aumentar o número de negócios repetidos. Big data permanece no coração de todas essas coisas.

10- (CESPE - 2015 - TRE-GO - Técnico Judiciário - Área Administrativa) Julgue o item subsecutivo, acerca de procedimentos de segurança e educação a distância (EAD). A Big Data pode ser utilizada na EAD para se entender as preferências e necessidades de aprendizagem dos alunos e, assim, contribuir para soluções mais eficientes de educa mediada por tecnologia.

Resolução:

O **objetivo do Big Data** é **propiciar dados e informações que possam ser analisados visando subsidiar tomadas de decisão**. Assim, em um contexto de Educação à Distância, é perfeitamente possível analisar as preferências e necessidades dos alunos para, com bas na tomada de decisão, melhorar as soluções para a aprendizagem destes alunos.

Educadores armados com uma visão orientada a dados podem ter um impacto significativo sobre os sistemas escolares, estudantes e currículos. Analisando big data, eles poder identificar alunos em risco, assegurar que os estudantes estão progredindo de forma adequada, e podem implementar um sistema melhor de avaliação e apoio aos professores diretores.

Gabarito: Certo.

1.5 Padrões utilizados em soluções de Big Data

Neste tópico trataremos de diversos conceitos interessantes e que ajudar a entender universo de Big Data, como as possíveis formas de consumir, processar, acessar e armazer os dados. Além disso, diversas questões sobre big data versam sobre assuntos tratados ne tópico, ainda que não referenciem diretamente o termo padrões. Sendo assim, verm brevemente os padrões existentes.

A IBM definiu um conjunto de **padrões** que servem para **visualizar como os componente devem ser projetados e onde eles devem ser posicionados funcionalmente**. Os padrõe também ajudam a **definir a arquitetura da solução de big data**. A IBM classificou os padrões em atômicos, compostos e de soluções.

- Os **padrões atômicos** ajudam a identificar a **maneira em que os dados são consumidos, processados, armazenados e acessados** por problemas recorrentes em um contexto de big data.

- Os **padrões compostos** são classificados com base na **solução de ponta a ponta**. Os **padrões compostos** são **mapeados para um ou mais padrões atômicos** para resolver um determinado problema de negócios.

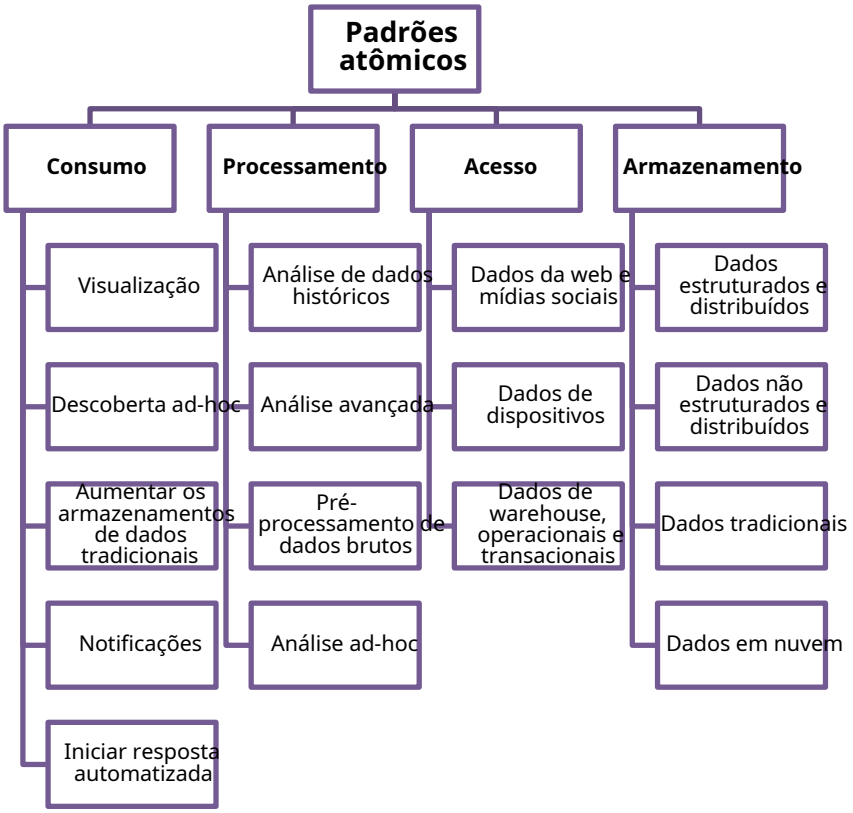
- Os **padrõesde solução**tambémajudama **definir o melhor conjuntode componentes com base em se o problema de negócios** precisa da descoberta e exploração de dados, de análise previsível e propositada ou de análise acionável.

Os **padrões atômicos** são os que fornecem as bases para a solução de big data, por isso, c discutiremos em maiores detalhes. Os padrões compostos e de solução são mais abranger e variados, muitas vezes utilizando uma composição de padrões atômicos para defini solução de big data.

ATENÇÃO!!!

Além de poderem ser cobrados de maneira direta, os padrões podem ser cobrados questões de maneira indireta. Assim, é importante ficar atento as possibilidades trazidas por estes padrões.

Vejamos inicialmente um **esquema** dos padrões atômicos, que serão detalhados adiante.



Esquema 4 – Padrões atômicos para Big Data.

Padrões de consumo

Vamos abordar inicialmente os **padrões de consumo** que lidam com as várias formas em que o resultado da análise de dados é consumido.

- **Padrão de visualização**: a forma tradicional de visualizar dados se baseia em gráficos, painéis e relatórios de resumo. Essas abordagens tradicionais não são sempre a melhor maneira de visualizar os dados.

  Os requisitostípicos para visualizaçãode big data, incluindoos requisitos emergentes, são listados abaixo:

  - Realizar análise em tempo real e exibição de dados de fluxo
  - Extrair dados de forma interativa, com base no contexto
  - Executar procuras avançadas e obter recomendações
  - Visualizar informações paralelamente
  - Ter acesso a hardware avançado para necessidades futuristas.

  O objetivo é fazer com que seja **mais fácil consumir os dados de forma intuitiva**, portanto os relatórios e painéis devem oferecer visualização full-HD e vídeos interativos 3D e também devem fornecer aos usuários a capacidade de controlar os resultados e atividades de negócios a partir de um aplicativo.

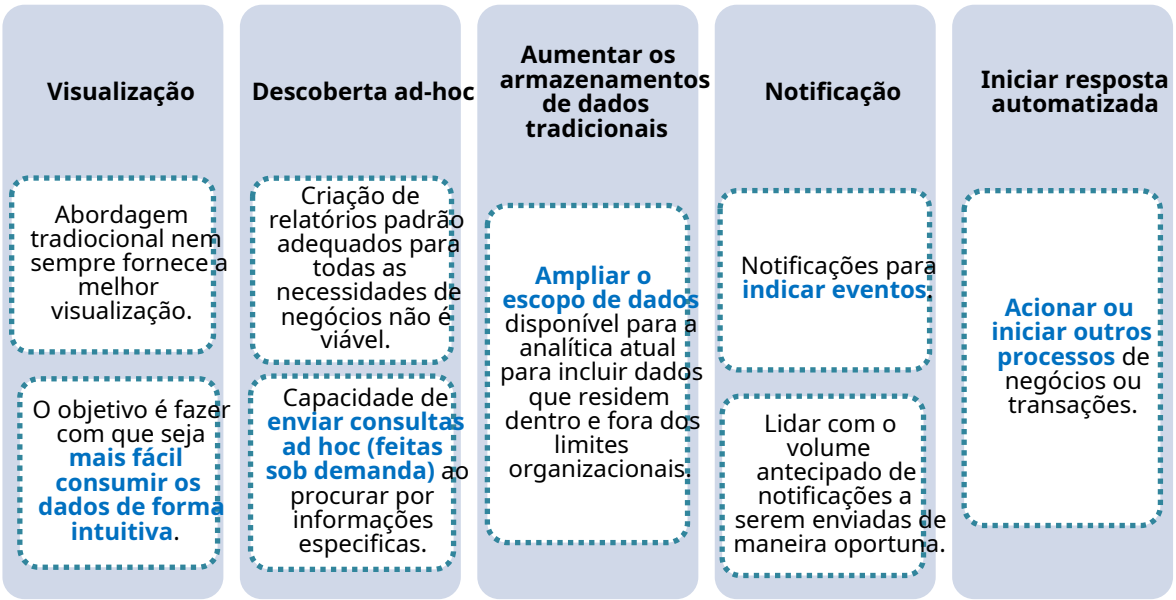
- **Padrão de descoberta ad-hoc**: em muitos casos, a criação de relatórios padrão que sejam adequados para todas as necessidades de negócios não é viável, pois a empresas têm requisitos de consultas de dados de negócios diversas. Os usuários podem precisar da **capacidade de enviar consultas ad hoc (feitas sob demanda) ao procurar por informações especificas**, dependendo do contexto do problema. A análise ad hoc pode ajudar os cientistas de dados e os principais usuários corporativos a entender o comportamento dos dados de negócios.

- **Aumentar os armazenamentos de dados tradicionais**: ajuda a **ampliar o escopo de dados disponível para a analítica atual para incluir dados que residem dentro e fora dos limites organizacionais**, como dados de mídia social, que podem melhorar os dados principais. Ao ampliar o escopo para incluir novas tabelas de fatos, dimensões e dados principais nos armazenamentos existentes e adquirir dado de clientes a partir de mídia social, uma organização pode obter um insight mais profundo do cliente.

- **Padrão de notificação**: os insights de big data permitem que as pessoas, negócios e máquinas ajam instantaneamente usando **notificações para indicar eventos**. A plataforma de notificação deve ser capaz de lidar com o volume antecipado de notificações a serem enviadas de maneira oportuna.

- **Padrão para iniciar uma respostaautomatizada**: os insights de negócios derivados do big data podem ser usados para **acionar ou iniciar outros processos de negócios ou transações**.

O **esquema** a seguir resume os padrões de consumo.



Esquema 5 – Padrões de consumo.

Padrões de processamento

O big data pode ser processado quando os dados estão em repouso ou em movim Dependendo da complexidade da análise, os dados podem não ser processados em tempo real. Esse padrão lida com como o big data é processado em tempo real, quase em tempo real ou em lote.

Os **padrões de processamento** são:

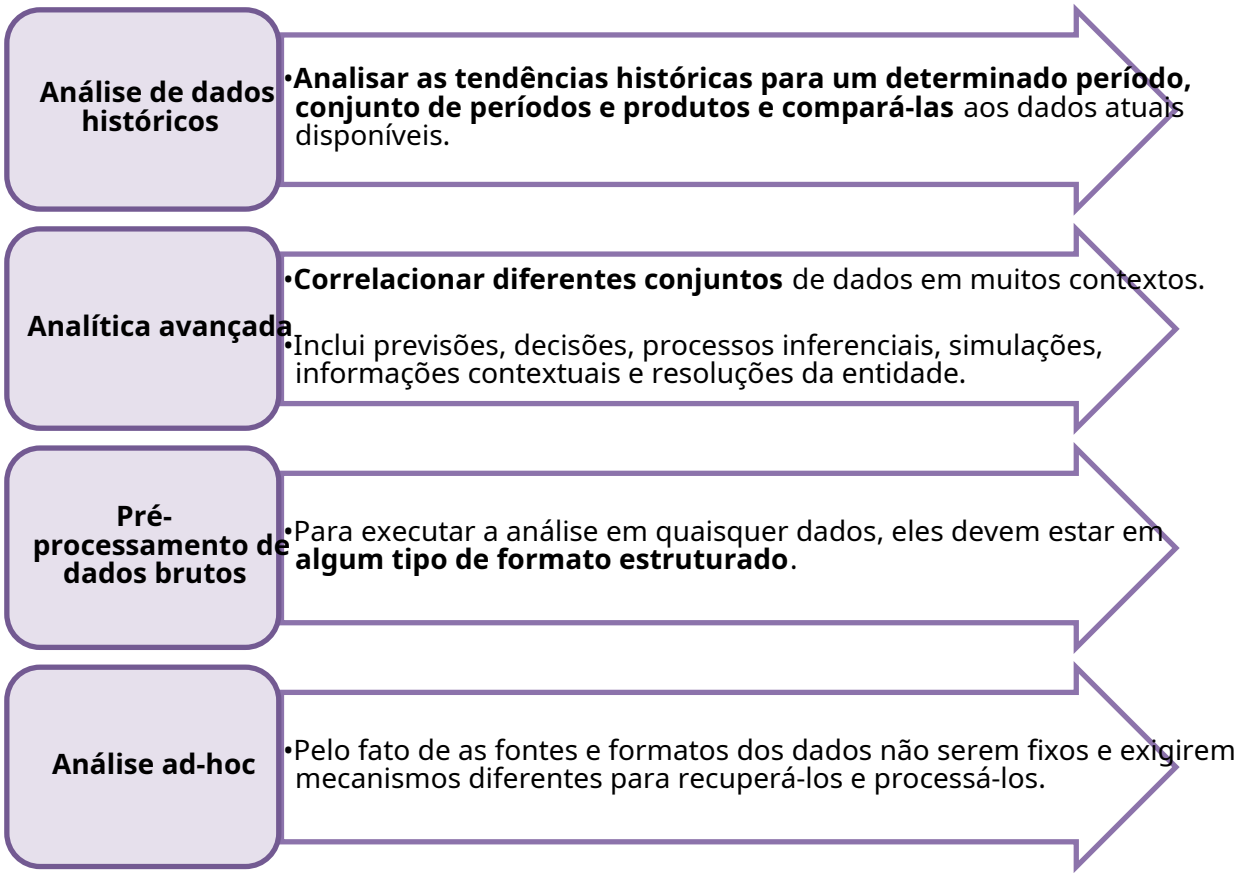
- **Padrão de análise de dados históricos**: análise de dados históricos tradicional é **limitada a um período predefinido de dados**, que normalmente depende das políticas de retenção de dados. Após esse período, geralmente os dados são arquivados ou limpos em virtude de limitações de armazenamento e processamento A análise histórica envolve **analisar as tendênciashistóricas para um determinado período, conjunto de períodos e produtos e compará-las** aos dados atuais disponíveis.

- **Padrão de analítica avançada**: o big data fornece enormes oportunidades de obter insights criativos. É possível **correlacionar diferentes conjuntos de dados em muitos contextos**.A descobertadessesrelacionamentosequer técnicas e algoritmos complexos inovadores. A análise avançada inclui previsões, decisões, processos inferenciais, simulações, identificações de informações contextuais e resoluções da entidade.

- **Padrão para pré-processar dados brutos**: a extração de dados a partir de dados não estruturados, como imagens, áudio, vídeo, feeds binários ou até mesmo texto, é uma tarefa complexa e precisa de técnicas como aprendizado de máquina e processamento de idioma natural, etc. O outro grande desafio é como verificar a precisão e a exatidão do resultado de tais técnicas e algoritmos. **Para executar a análise em quaisquer dados, eles devem estar em algum tipo de formato estruturado**. Os dados não estruturados acessados de várias fontes podem se armazenados como estão e, em seguida, transformados em dados estruturados (por exemplo, JSON) e novamente armazenados nos sistemas de armazenamento de big data. O texto não estruturado pode ser convertido em dados estruturados ou semiestruturados. Da mesma forma, os dados de imagem, áudio e vídeo precisam ser convertidos nos formatos que podem ser usados para análise.

- **Padrão de análise ad hoc**: o processamento de consultas ad hoc no big data traz desafios diferentes daqueles incorridos ao realizar consultas ad hoc em dados estruturados pelo fato de as **fontes e formatos dos dados não serem fixos exigirem mecanismos diferentes para recuperá-los e processá-los**. Embora as consultas ad hoc simples possam ser resolvidas pelos provedores de big data, na maioria dos casos, elas são complexas porque os dados, algoritmos, formatos resoluções da entidade devem ser descobertos dinamicamente.

Vamos resumir os padrões de processamento em um **esquema**.



Esquema 6 – Padrões de processamento.

Padrões de acesso

Embora existam **muitas fontes de dados e formas em que os dados podem ser acessados** em uma solução de big data, algumas formas comuns podem ser discutidas pel seguintes padrões:

- **Padrão de acesso à web e mídia social**: a Internet é a fonte de dados que fornece muitos dos insights produzidos atualmente. **A web e a mídia social são úteis em praticamente todas as análises**, mas são necessários mecanismos de acesso diferentes para obter esses dados. A web e a mídia social são a fonte de dados mais complexa de todas em virtude de sua enorme variedade, velocidade e volume. Há aproximadamente de 40 a 50 categorias de websites e cada uma requer um tratamento diferente para acessar esses dados.

- **Padrão de dados gerados por dispositivos**: o conteúdo gerado por dispositivos **inclui dados de sensores**. Os **dados são detectados a partir das origens de dados**, como informações sobre o clima, medições elétricas e dados sobre poluição, e capturados pelos sensores. Os dados podem ser fotos, vídeos, texto e outro formatos binários.

- **Padrão de dados de warehouse, operacionais e transacionais**: É possível **armazenar os dados de warehouse, operacionais e transacionais existentes para evitar a limpeza ou o arquivamento deles** (em virtude de limitações de armazenamento e processamento) ou para reduzir a carga no armazenamento tradicional quando os dados são acessados por outros consumidores. Para a maioria das empresas, os dados principais, operacionais, transacionais e as informações de warehouse estão no centro de qualquer analítica. Esses dados, se aumentados com os dados não estruturados e externos disponíveis em toda a Internet ou por meio de sensores e dispositivos inteligentes, podem ajudar as organizações a obterem insights precisos e executarem analíticas avançadas.

Vejamos um **esquema** para sintetizar os padrões de acesso.

Padrão de acesso à web e mídia social

- A web e a mídia social são **úteis em praticamente todas as análises**, mas são necessários mecanismos de acesso diferentes para obter esses dados.

Padrão de dados gerados por dispositivos

- Inclui **dados de sensores**.
- Os dados são **detectados a partir das origens** de dados.

Padrão de dados de warehouse, operacionais e transacionais

- É possível **armazenar os dados de warehouse, operacionais e transacionais** existentes para evitar a limpeza ou o arquivamento deles ou para reduzir a carga no armazenamento tradicional quando os dados são acessados por outros consumidores.

Esquema 7 – Padrões de acesso.

Padrões de armazenamento

Os **padrões de armazenamento** ajudam a **determinar o armazenamento adequado para diversos formatos e tipos de dados**. Os dados podem ser armazenados como estão, com relação a pares de valores de chave ou em formatos predefinidos. Os padrões de armazenamento são:

- **Padrão de armazenamento para dados não estruturados e distribuídos**: a maior parte do big data não é estruturada e pode conter informações que podem ser extraídas de diferentes formas para diferentes contextos. Na maioria das vezes, **os dados não estruturados devem ser armazenados como estão**, em seu formato original.

- **Padrão de armazenamento para dados estruturados e distribuídos**: os dados estruturados incluem aqueles que chegam da fonte de dados e já estão em u formato estruturado e os dados não estruturados que foram pré-processados em ur formato como JSON. Esses **dados convertidos devem ser armazenados para evitar a frequente conversão de dados brutos para dados estruturados**.

- **Padrão de armazenamento para armazenamentos de dados tradicionais**: o armazenamento de dados tradicional **não é a melhor opção para armazenar big data**, mas nos casos em que as empresas estão realizando a exploração de dados inicial, elas podem optar por **usar o data warehouse, o sistema relacional e outros armazenamentos de conteúdo existentes**. Esses sistemas de armazenamento existentes podem ser usados para armazenar os dados que sã compilados e filtrados usando a plataforma de big data.

- **Padrão de armazenamento para armazenamento em nuvem**: muitos provedores de infraestrutura da nuvem possuem recursos de armazenamento estruturado e não estruturado distribuídos. As tecnologias de big data são um pouco diferentes das perspectivas de configurações, manutenção, gerenciamento de sistemas e programação e modelagem tradicionais. Além disso, as qualificações necessária para implementar as soluções de big data são raras e caras. As **empresas explorando as tecnologias de big data podem usar soluções de nuvem que fornecem o gerenciamento de sistemas, manutenção e armazenamento de big data**. A grande vantagem de associar big data à cloud computing é **reduzir o custos de uma infraestrutura de TI para armazenar e processar os dados**.

Vamos fechar o entendimento dos padrões de armazenamento com um **esquema**.

Dados não estruturados e distribuídos	Dados estruturados e distribuídos	Armazenamento de dados tradicionais	Armazenamento em nuvem
•Dados não estruturados devem ser armazenados como estão.	•Dados convertidos devem ser armazenados para evitar a frequente conversão de dados brutos para dados estruturados.	•Usar o data warehouse, o sistema relacional e outros armazenamentos de conteúdo existentes.	•Uso de soluções de nuvem que fornecem o gerenciamento de sistemas, manutenção e armazenamento de big data

Esquema 8 – Padrões de armazenamento.

11- (CESPE - 2015 - TCU - Auditor Federal de Controle Externo) No que concerne a data mining (mineração de dados) e big data, julgue o seguinte item.

Devido à quantidade de informações manipuladas, a (cloud computing) computação em nuvem torna-se inviável para soluções de big data.

Resolução:

Muito pelo contrário. A grande vantagem de **associar big data à cloud computing** é reduzir os custos de uma infraestrutura de TI para armazenar e processar os dados.

Big Data em relação à nuvem representa um desafio tecnológico pois demanda atenção à infraestrutura e tecnologias analíticas. **Processamento de massivos volumes de dados pode ser facilitado pelo modelo de computação em nuvem**, desde que, é claro, que este imenso volume não seja transmitido repetidamente via Internet.

Gabarito: **Errado**.

12- (CESPE - 2014 - TJ-SE - Analista Judiciário - Banco de Dados) Julgue os itens que se seguem, no que se refere a Big Data.

Em soluções Big Data, a análise dos dados comumente precisa ser precedida de um transformação de dados não estruturados em dados estruturados.

Resolução:

Para executar a análise em quaisquer dados, **eles devem estar em algum tipo de formato estruturado**. Os dados não estruturados acessados de várias fontes podem ser armazenados como estão e, em seguida, transformados em dados estruturados (por exemplo, JSON) e novamente armazenados nos sistemas de armazenamento de big data. O texto não estruturado pode ser convertido em dados estruturados ou semiestruturados. Da mesma forma, os dados de imagem, áudio e vídeo precisam ser convertidos nos formatos que podem ser usados para análise. Além disso, a precisão e exatidão da analítica avançada que usa algoritmos preditivos e estatísticos dependem da quantidade de dados e algoritmos usados para treinar os modelos. A lista a seguir mostra os algoritmos e atividades necessários para converter dados não estruturados em estruturados:

- Classificação de texto e documento;
- Extração de recurso;
- Segmentação de texto e imagem;
- Correlacionamento de recursos, variáveis e tempos e, em seguida, extração de valores com o tempo;
- Verificação de precisão do resultado usando técnicas como a matriz de confusão e outras atividades manuais;

Gabarito: **Certo**.

13- (QUADRIX - 2022 – CRECI 11 – Profissional de Atividades Estratégicas – Analista de Tecnologia da Informação) Acerca das noções de Big Data, julgue o item.

Com a tecnologia do Big Data, é possível que uma empresa virtualize seus dados para que possam ser armazenados em nuvem, obtendo, assim, um melhor custo-benefício.

Resolução:

Uma das formas de armazenamentos de dados de Big Data é na nuvem. Muitos provedores de infraestrutura da nuvem possuem recursos de armazenamento estruturado e não estruturado distribuídos. As tecnologias de big data são um pouco diferentes das perspectivas de configurações, manutenção, gerenciamento de sistemas e programação e modelagem tradicionais. Além disso, as qualificações necessárias para implementar a soluções de big data são raras e caras. As **empresas explorando as tecnologias de big data podem usar soluções de nuvem que fornecem o gerenciamento de sistema manutenção e armazenamento de big data**. A grande vantagem de associar big data à cloud computing é **reduzir os custos de uma infraestrutura de TI para armazenar e processar os dados**.

Gabarito: **Certo**.

2. HADOOP

O **Hadoop** é um framework de código aberto, implementado em Java e **utilizado para o processamento e armazenamento em larga escala, para alta demanda de dados utilizando máquinas comuns**. Os serviços do Hadoop fornecem armazenamento, processamento, acesso, governança, segurança e operações de dados.

Entre os benefícios de utilizá-lo pode-se destacar:

- **Código aberto**: todo projeto de software livre de sucesso tem por trás uma comunidade ativa. No projeto Apache Hadoop, essa **comunidade é composta por diversas empresas e programadores independentes partilhando seus conhecimentos no desenvolvimento de melhorias, funcionalidades e documentação**. Sendo um software de código aberto, tem por princípio a garantia das quatro liberdades aos seus usuários: liberdade para executar o programa para qualquer propósito; liberdade de estudar como o programa funciona e adaptá-lo para as suas necessidades; liberdade de redistribuir cópias do programa; e liberdade para modificar o programa e distribuir essas modificações, de modo que toda a comunidade se beneficie.

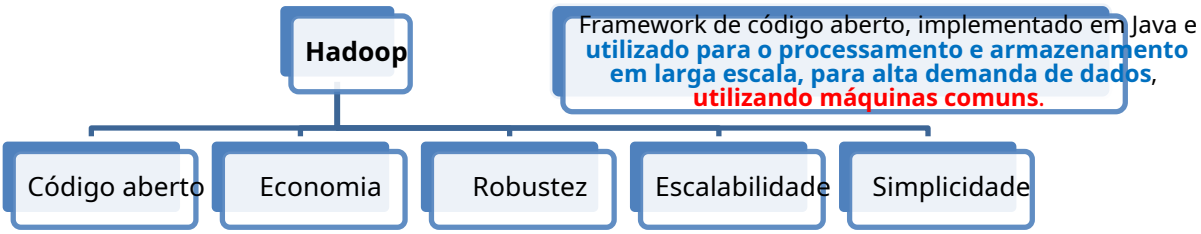
- **Economia**: podemos apontar neste quesito 3 formas de economia.
  - Primeiro, por ser um software livre, com um esforço relativamente pequeno é possível **implantar, desenvolver aplicações e executar Hadoop sem gastar com aquisição de licenças** e contratação de pessoal especializado.
  - Segundo, a possibilidade realizar o processamento da sua massa de dados **utilizando máquinas e rede convencionais**.
  - Por último, existe uma outra alternativa econômica dada pela existência de serviços em nuvem, como a Amazon Elastic MapReduce (EMR), que permite a **execução de aplicações Hadoop sem a necessidade de implantar seu próprio aglomerado de máquinas**, alugando um parque virtual ou simplesmente pagando pelo tempo de processamento utilizado;

- **Robustez**: como ele foi projetado para ser executado em hardware comum, ele já **considera a possibilidade de falhas frequentes nesses equipamentos e oferece estratégias de recuperação automática** para essas situações. Assim, sua implementação disponibiliza mecanismos como replicação de dados, armazenamento de metadados e informações de processamento, que dão uma maior garantia para que a aplicação continue em execução mesmo na ocorrência de falhas em algum recurso;

- **Escalabilidade**: enquanto as demais aplicações similares apresentam dificuldade em aumentar a quantidade de máquinas utilizadas no processamento e/ou aumentar o conjunto de dados, precisando em alguns casos até reescrever todo o código-fonte da aplicação, Hadoop permite obter escalabilidade de forma relativamente simples. **Mudanças no ambiente implicam em pequenas modificações em um arquivo de configuração**. Dessa forma, o trabalho para preparar um ambiente contendo mil máquinas não será muito maior do que se este fosse de dez máquinas. O aumento no volume de dados só fica limitado aos recursos, espaço em disco e capacidade de processamento, disponíveis nos equipamentos do aglomerado, sem a necessidade de alteração da codificação;

- **Simplicidade**: Hadoop **retira do desenvolvedor a responsabilidade de gerenciar questões relativas à computação paralela, tais como tolerância a falhas, escalonamento e balanceamento de carga**, ficando estas a cargo do próprio arcabouço. O Hadoop descreve suas operações apenas por meio da funções de mapeamento (Map) e de junção (Reduce). Dessa forma, o foco pode ser mantido somente na abstração do problema, para que este possa ser processado no modelo de programação MapReduce. O Hadoop fornece uma solução excelente e simples para as empresas que necessitam processar seu dados de diversas maneiras para crescer, ou, em alguns casos, continuar dominando o mercado. Não à toa grandes corporações como o Facebook e o Twitter utilizam o Hadoop.

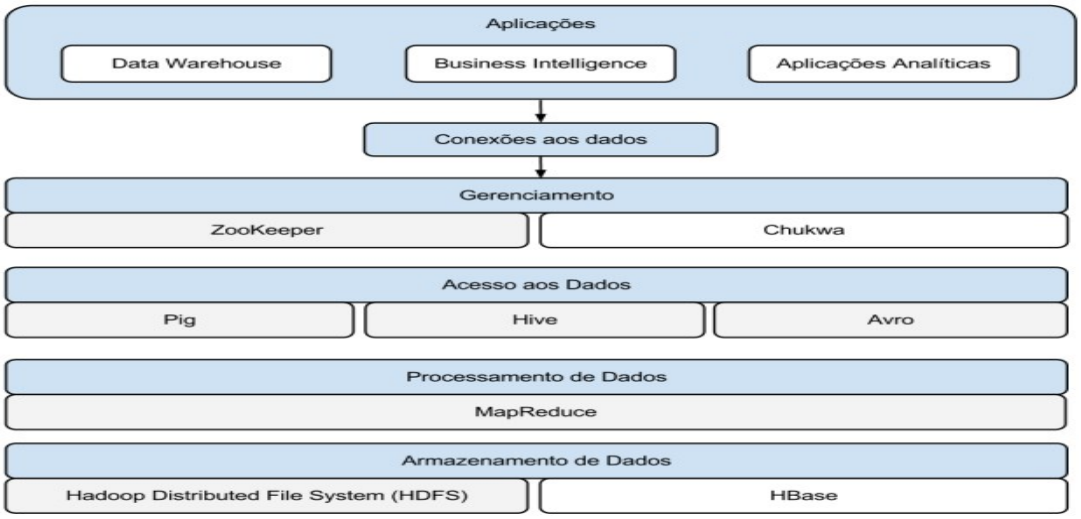
O esquema a seguir resume o que é o hadoop e quais as suas vantagens!



Esquema 9 – Hadoop.

Os elementos chave do Hadoop são o **modelo de programação MapReduce** e o **sistema de arquivos distribuído HDFS**. Entretanto, em meio a sua evolução, novos subprojetos, cada um para uma proposta específica da aplicação, foram incorporados ao seu ecossisten tornando a infraestrutura do arcabouço cada vez mais completa.

Os **subprojetos do Hadoop** e sua posição na anatomia do framework podem ser visualizados no esquema a seguir.



Esquema 10 – Subprojetos do Hadoop.

A seguir, vamos falar um pouco sobre estes subprojetos, dando maior ênfase ao HDFD e ao MapReduce que são os dois principais.

Subprojetos para gerenciamento

O **Chukwa** é o **sistema especialista em coleta e análise de logs em sistemas de larga escala**. Utiliza HDFS para armazenar os arquivos e o MapReduce para geração de relatórios. Possui como vantagem um kit auxiliar de ferramentas, muito poderoso e flexíve que promete melhorar a visualização, monitoramento e análise dos dados coletados.

O **ZooKeeper** é o arcabouço criado pelo Yahoo! em 2007 com o objetivo de fornecer um **serviço de coordenação para aplicações distribuídas de alto desempenho**, que provê meios para facilitar as seguintes tarefas: configuração de nós, sincronização de processos distribuídos e grupos de serviço.

Subprojetos para acesso aos dados

O **Pig** é uma **linguagem de alto nível orientada a fluxo de dados e um arcabouço de execução para computação paralela**. Sua utilização não altera a configuração do aglomerado Hadoop, pois é utilizado no modo client-side, fornecendo uma linguagem chamada Pig Latin e um compilador capaz de transformar os programas do tipo Pig em sequências do modelo de programação MapReduce.

O **Hive** é o arcabouço desenvolvido pela equipe de funcionários do Facebook, tendo tornado um projeto de código aberto em agosto de 2008. Sua principal funcionalidade é fornecer uma infraestrutura que permita utilizar Hive QL, uma linguagem de consult similar a SQL bem como demais conceitos de dados relacionais tais como tabelas, colunas linhas, para **facilitar as análises complexas feitas nos dados não relacionais de uma aplicação Hadoop**.

Existe também uma definição que trata o **Hive** como um **datawarehouse distribuído** que facilita o uso de grandes conjuntos de dados. Nesse caso, ele seria enquadrado como um subprojeto para armazenamento dos dados.

O **Avro** é o **sistema de seriação de dados baseado em schemas**. Sua composição é dada por um repositório de dados persistentes, um formato compacto de dados binários e supo a chamadas remotas de procedimentos (RPC).

Subprojetos para processamento de dados

O **Hadoop MapReduce** é um **modelo de programação e um arcabouço especializado no processamento de conjuntos de dados distribuídos em um aglomerado computacional (cluster)**.

O **MapReduce** é uma excelente solução para o processamento paralelo de dados devido ao fato de serem inerentemente paralelos. O **programador não precisa realizar nenhum tip de programação extra para garantir que os processos serão processados paralelamente**.

O **MapReduce** possui duas fases de processamento: o **Map e o Reduce**. A primeira fase, a fase de **mapeamento**, é responsável pelo **processamento primário dos dados de entrad** Então, os resultados dessa fase são enviados para a **função de redução** como entradas. Então, o **resultado final** é realizado pela fase de redução e **enviado para arquivos que conterão esses resultados**. O escalonamento dos processos é feito internamente pe Hadoop, e o desenvolvedor nem sequer fica sabendo como isso é realizado.

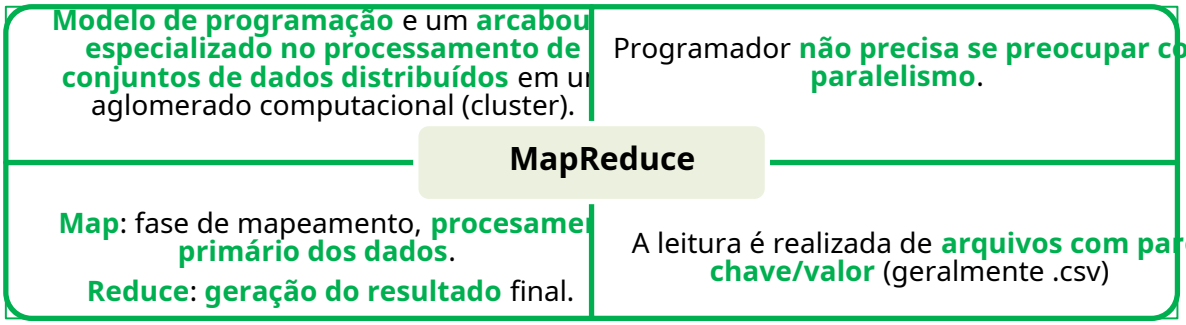
Normalmente, os programas MapReduce leem dados de arquivos em formato .csv, que são capazes de separar os dados mais ou menos com o mesmo formato que se vê em bases de dados de qualquer tipo, onde cada valor representa uma coluna de dados. É poss ler dados de arquivos de texto ou qualquer outro tipo, desde que haja uma lista de dados d entrada que possa ser transformada em pares chave/valor, que é o que a aplicação entender posteriormente.

O **MapReduce** funciona da seguinte forma:

1. A entrada da aplicação é uma lista de pares chave/valor.
2. Então, esses pares são selecionados um a um e processados, cada um gerando um par (chave/lista) de valores. Os detalhes dessa transformação é que normalmente definem o que o programa MapReduce faz.
3. Essa nova lista de pares é selecionada como entrada pela função Reducer e é agregada de alguma forma, gerando uma saída final.

Um exemplo muito simples e comum de programação paralela é a contagem de palavras em vários documentos diferentes. Sem utilizar o MapReduce, o desenvolvedor estaria envolvido em uma série de problemas que são inerentes ao processamento paralelo dados. Porém, utilizando o MapReduce, o Hadoop cuida automaticamente disso para programador, evitando que o mesmo tenha que se preocupar com problemas de escalonamento e o local que os dados se encontram no sistema de arquivos.

Sendo o MapReduce um dos principais subprojetos do Hadoop, vejamos um **esquema** para fixar as informações sobre ele.



Esquema 11 – MapReduce.

Subprojetos para armazenamento de dados

O **Hadoop Distributed File System (HDFS)** é um **sistema de arquivos distribuído nativo do Hadoop**. Permite o armazenamento e transmissão de grandes conjuntos d dados em máquinas de baixo custo. Possui mecanismos que o caracteriza como um sistem altamente tolerante a falhas.

O **HDFS** é um projeto da Apache Software Foundation e um subprojeto do projeto Apache Hadoop. O Hadoop é ideal para armazenar grandes quantidades de dados, do porte terabytes e pentabytes, e usa o **HDFS** como **sistema de armazenamento**. O HDFS permite a conexão de nós (computadores pessoais padrão) contidos nos clusters por meio dos qua os arquivos de dados são distribuídos. É possível acessar e armazenar os arquivos de dado como um sistema de arquivos contínuo. O acesso aos arquivos de dados é gerenciado de u modo em fluxo, o que significa que aplicativos ou comandos são executados diretamente por meio do modelo de processamento MapReduce. O **HDFS** é tolerante a falhas e disponibiliza acesso de alto rendimento a grandes conjuntos de dados.

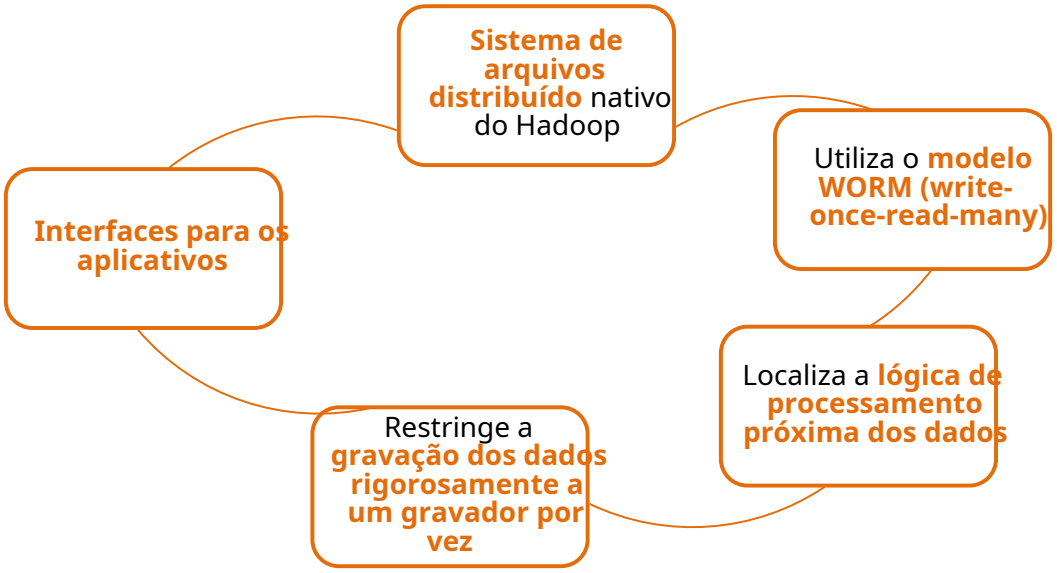
O **HDFS** tem muitas similaridades com outros sistemas de arquivos distribuídos, mas é diferente em vários aspectos. Uma diferença notável é o **modelo WORM (write-once-read-many)** do HDFS que afrouxa as exigências do controle de simultaneidade, simplifica a persistência de dados e habilita acesso de alto rendimento.

Outro atributo exclusivo do **HDFS** é o argumento que, normalmente, é **melhor localizar a lógica de processamento próxima dos dados**, ao invés de mover os dados para o espaç do aplicativo.

O **HDFS** **restringe a gravação dos dados rigorosamente a um gravador por vez**. Os bytes são sempre anexados ao final do fluxo e há a garantia de que os fluxos de bytes serã armazenados na ordem gravada.

O **HDFS** fornece **interfaces para os aplicativos** a fim de movê-los para perto de onde se localizam os dados. O HDFS disponibiliza uma interface de programação de aplicativos (API) Java™ e um wrapper nativo em linguagem C para a API Java. Além disso, é possível usar um navegador da web para buscar arquivos no HDFS.

Vejamos um esquema sobre o HDFS.



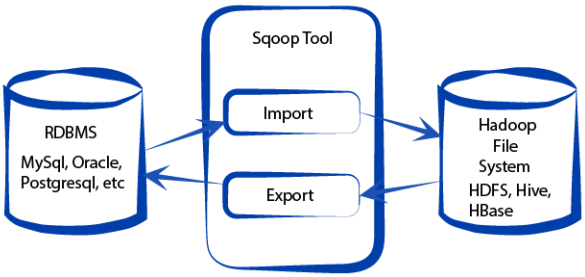
Esquema 12 – HDFS.

Outro subprojeto relacionado ao armazenamento de dados é o **Hbase**, que é um banco de dados criado pela empresa Power Set em 2007, tendo posteriormente se tornando u projeto da Apache Software Foundation. Considerado uma versão de código aberto d banco de dados BigTable, criada pela Google, é um **banco de dados distribuído e escalável que dá suporte ao armazenamento estruturado e otimizado para grandes tabelas**.

Sqoop (SQL to Hadoop)

Um outro projeto interessante é o Sqoop. Esse é um projeto de alto nível do Apache.

O **Sqoop (SQL to Hadoop)** é um aplicativo de interface de linha de comando para **transferência de dados entre bancos de dados relacionais e Hadoop**. Ele pode ser usad tanto para importação de dados para o Hadoop quanto para a exportação de dados para o banco relacional.



Flume

Um outro projeto de alto nível do Apache é o Flume. O **Flume** é um software distribuído, confiável e disponível para **coletar, agregar e mover com eficiência grandes quantidades de dados de log**.

Componentes do Hadoop

Uma **execução típica de uma aplicação Hadoop** em um aglomerado utiliza cinco processos diferentes: **NameNode, DataNode, SecondaryNameNode, JobTracker e TaskTracker**. Os três primeiros são integrantes do sistema de arquivo HDFS, e os dois últimos do modelo de programação MapReduce. Os componentes **NameNode, JobTracker e SecondaryNameNode** são **únicos para toda a aplicação**, enquanto que o **DataNode e TaskTracker** são **instanciados para cada máquina**.

Vejamos um pouco sobre cada um destes componentes:

- **NameNode**: tem como responsabilidade **gerenciar os arquivos armazenados no HDFS**. Suas funções incluem mapear a localização, realizar a divisão dos arquivos em blocos, encaminhar os blocos aos nós escravos, obter os metadados dos arquivos controlar a localização de suas réplicas. Como o NameNode é constantemente acessado por questões de desempenho, ele mantém todas as suas informações em memória. Ele integra o sistema HDFS e fica localizado no nó mestre da aplicação, juntamente com o JobTracker.

- **SecondaryNameNode**: utilizado para **auxiliar o NameNode a manter seu serviço, e ser uma alternativa de recuperação no caso de uma falha do NameNode**. Sua única função é realizar pontos de checagem (checkpointing) do NameNode em intervalos pré-definidos, de modo a garantir a sua recuperação e atenuar o seu tempo de reinicializaçã

- **DataNode**: enquanto o NameNode gerencia os blocos de arquivos, são os DataNodes que **efetivamente realizam o armazenamento dos dados**. Como o HDFS é um sistema de arquivos distribuído, é comum a existência de diversas instâncias do DataNode em uma aplicação Hadoop, para que eles possam distribuir os blocos d arquivos em diversas máquinas. Um DataNode poderá armazenar múltiplos blocos inclusive de diferentes arquivos. Além de armazenar, eles precisam se reportar constantemente ao NameNode, informando quais blocos estão guardando bem como todas as alterações realizadas localmente nesses blocos.

- **JobTracker**: assim como o NameNode, o JobTracker também possui uma **função de gerenciamento**, porém, nesse caso, o controle é realizado **sobre o plano de execução das tarefas a serem processadas pelo MapReduce**. Sua função então é designa diferentes nós para processar as tarefas de uma aplicação e monitorá-las enquan estiverem em execução. Um dos objetivos do monitoramento é, em caso de falha identificar e reiniciar uma tarefa no mesmo nó ou, em caso de necessidade, em um nó diferente.

- **TaskTracker**: processo **responsável pela execução de tarefas MapReduce**. Assim como os DataNodes, uma aplicação Hadoop é composta por diversas instâncias d TaskTrackers, cada uma em um nó escravo. Um TaskTracker executa uma tarefa Map ou uma tarefa Reduce designada a ele. Como os TaskTrackers rodam sobre máquinas virtuais, é possível criar várias máquinas virtuais em uma mesma máquina física, de forma a explorar melhor os recursos computacionais.

Vamos esquematizar as funções desses processos.

NameNode

•**Gerenciar os arquivos armazenados** no HDFS.

SecondaryNameNode

•**Auxiliar o NameNode** a manter seu serviço.
•Ser uma **alternativa de recuperação** no caso de uma falha do NameNode.

DataNode

•Efetivamente realizam o **armazenamento dos dados**.

JobTracker

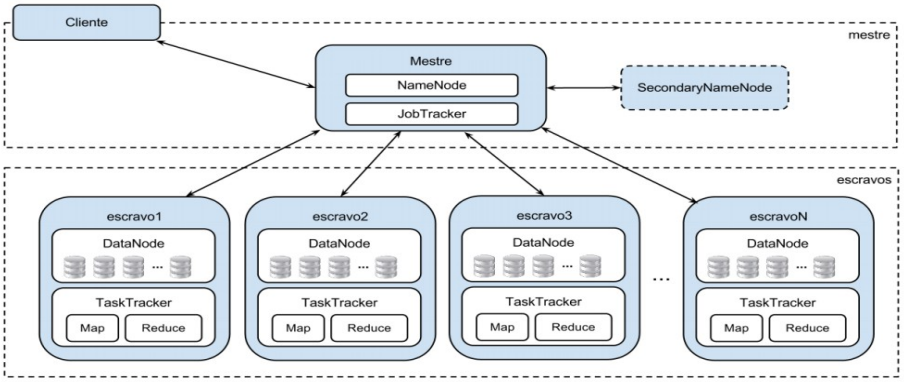
•Função de **gerenciamento sobre o plano de execução das tarefas** a serem processadas pelo MapReduce.

TaskTracker

•Responsável pela **execução de tarefas MapReduce**.

Esquema 13 – Processos Hadoop.

No esquema a seguir, pode-se observar de forma mais clara como os processos da arquitetura do Hadoop estão interligados.



Esquema 14 – Arquitetura dos Processos Hadoop.

Inicialmente nota-se uma separação dos processos entre os nós mestre e escravos. primeiro contém o NameNode, o JobTracker e possivelmente o SecondaryNameNode. Já o segundo, comporta em cada uma de suas instâncias um TaskTracker e um DataNod vinculados respectivamente ao JobTracker e ao NameNode do nó mestre.

Um cliente de uma aplicação se conecta ao nó mestre e solicita a sua execução. Ne momento, o JobTracker cria um plano de execução e determina quais, quando e quantas vezes os nós escravos processarão os dados da aplicação. Enquanto isso, o NameNo baseado em parâmetros já definidos, fica encarregado de armazenar e gerenciar as informações dos arquivos que estão sendo processados. Do lado escravo, o TaskTracker executa as tarefas a ele atribuídas, que ora podem ser Map ora Reduce, e o DataN armazena um ou mais blocos de arquivos. Durante a execução, o nó escravo também preci se comunicar com o nó mestre, enviando informações de sua situação local.

Paralelamente a toda essa execução, o SecondaryNameNode registra pontos de checagem dos arquivos de log do NameNode, para a necessidade de uma possível substituição no cas de o NameNode falhar.

14- (CESPE / CEBRASPE - 2021 - SERPRO - Analista - Especialização: Ciência de Dados) Sobre processamento de dados, julgue o item a seguir.

MapReduce divide o conjunto de dados de entrada em blocos independentes que sã processados pelas tarefas de mapa de uma maneira completamente paralela. Essa estrutu classifica as saídas dos mapas, as quais são, então, inseridas nas tarefas de redução.

Resolução:

O **MapReduce** possui duas fases de processamento: o **Map e o Reduce**. A primeira fase, a fase de **mapeamento**, é responsável pelo **processamento primário dos dados de entrad** Então, os resultados dessa fase são enviados para a **função de redução** como entradas. Então, o **resultado final** é realizado pela fase de redução e **enviado para arquivos que conterão esses resultados**. O escalonamento dos processos é feito internamente pe Hadoop, e o desenvolvedor nem sequer fica sabendo como isso é realizado.

Normalmente, os programas MapReduce leem dados de arquivos em formato .csv, que são capazes de separar os dados mais ou menos com o mesmo formato que se vê em bases de dados de qualquer tipo, onde cada valor representa uma coluna de dados. É poss ler dados de arquivos de texto ou qualquer outro tipo, desde que haja uma lista de dados d entrada que possa ser transformada em pares chave/valor, que é o que a aplicação entender posteriormente.

O **MapReduce** funciona da seguinte forma:

4. A entrada da aplicação é uma lista de pares chave/valor.
5. Então, esses pares são selecionados um a um e processados, cada um gerando um par (chave/lista) de valores. Os detalhes dessa transformação é que normalmente definem o que o programa MapReduce faz.
6. Essa nova lista de pares é selecionada como entrada pela função Reducer e é agregada de alguma forma, gerando uma saída final.

Gabarito: **Certo**.

15- (CESPE / CEBRASPE - 2021 - SERPRO - Analista - Especialização: Ciência de Dados) Julgue o próximo item, relativo à tecnologia de big data e ao Hadoop.

Apesar de ser uma tecnologia de código aberto disponibilizada pela ASF (Apache Software Foundation), o Hadoop também é oferecido por distribuidores comerciais, de maneira que fornecedores oferecem distribuições específicas que incluem não só ferramentas administrativas adicionais, mas também suporte técnico.

Resolução:

Pessoal, o Hadoop é de código aberto, tendo por princípio a garantia das quatro liberdades aos seus usuários: liberdade para executar o programa para qualquer propósito; liberdade de estudar como o programa funciona e adaptá-lo para as suas necessidades; liberdade de redistribuir cópias do programa; e liberdade para modificar o programa e distribuir essas modificações, de modo que toda a comunidade se beneficie.

Porém, isso não impede que empresas explorem comercialmente um serviço que inclua o Hadoop juntamente com outras ferramentas e/ou suporte técnico. Isso pode ocorrer com várias ferramentas de código aberto e não é diferente com o Hadoop.

Gabarito: **Certo**.

16- (CESPE / CEBRASPE - 2021 - SEFAZ-CE - Auditor Fiscal de Tecnologia da Informação da Receita Estadual) Em relação a big data e analytics, julgue o próximo item.

Hive e Sqoop são subprojetos do Hadoop destinados a queries e data warehousing, respectivamente.

Resolução:

O **Hive** fornece uma infraestrutura para utilizar a Hive QL, que é uma linguagem de consulta. Existe uma definição que trata o Hive como DW distribuído que facilita o uso de grandes conjuntos de dados.

O **Sqoop** é usado para transferência de dados entre bancos relacionais e o Hadoop.

Logo, houve uma inversão na questão.

Gabarito: **Errado**.

17- (CESPE / CEBRASPE - 2021 - SEFAZ-CE - Auditor Fiscal de Tecnologia da Informação da Receita Estadual) Julgue o seguinte item, relativo a ferramentas de BI e banco de dados NoSQL.

O Hadoop pode ser configurado em clusters de servidores para implementação de projeto de big data, podendo o ZooKeeper ser utilizado nesse caso como provedor de servi centralizado para fornecer informações de configuração, sincronização e serviços de grupo nesses clusters.

Resolução:

O **ZooKeeper** é o arcabouço criado pelo Yahoo! em 2007 com o objetivo de fornecer um **serviço de coordenação para aplicações distribuídas de alto desempenho**, que provê meios para facilitar as seguintes tarefas: configuração de nós, sincronização de processos distribuídos e grupos de serviço.

Gabarito: **Certo**.

18- (INSTITUTO AOCP - 2020 - MJSP - Engenheiro de Dados - Big Data) Assim como o Hadoop foi desenvolvido para possibilitar o processamento em lote de gran volume de dados, também surgiram tecnologias com suporte ao processamento em tempo real de Big Data, como o

O HDFS é o sistema de arquivos do Hadoop. Ele possui uma arquitetura mestre-escravo na qual um servidor é responsável por fazer todo o gerenciamento de metadados do siste Dentro da arquitetura do Hadoop, como se denomina esse servidor?

a) NameNode.

b) DataNode.

c) HDFSnode.

d) LinkNode.

e) TraceNode.

Resolução:

Vamos analisar cada um dos itens:

a) **Correto**: NameNode é o componente com responsabilidade **gerenciar os arquivos armazenados no HDFS**.

b) **Incorreto**: DataNode é o componente que **efetivamente realizam o armazenamento dos dados**.

c) **Incorreto**: HDFSnode não é um componente do Hadoop.

d) **Incorreto**: LinkNode não é um componente do Hadoop.

e) **Incorreto**: TraceNode não é um componente do Hadoop.

Gabarito: Letra A.

3. SPARK

O **Spark** é um **framework para processamento de Big Data construído com foco em velocidade, facilidade de uso e análises sofisticadas**. O Spark tem muitas vantagens se comparado as outras tecnologias de Big Data e do paradigma MapReduce, como o Hadoop e o Storm.

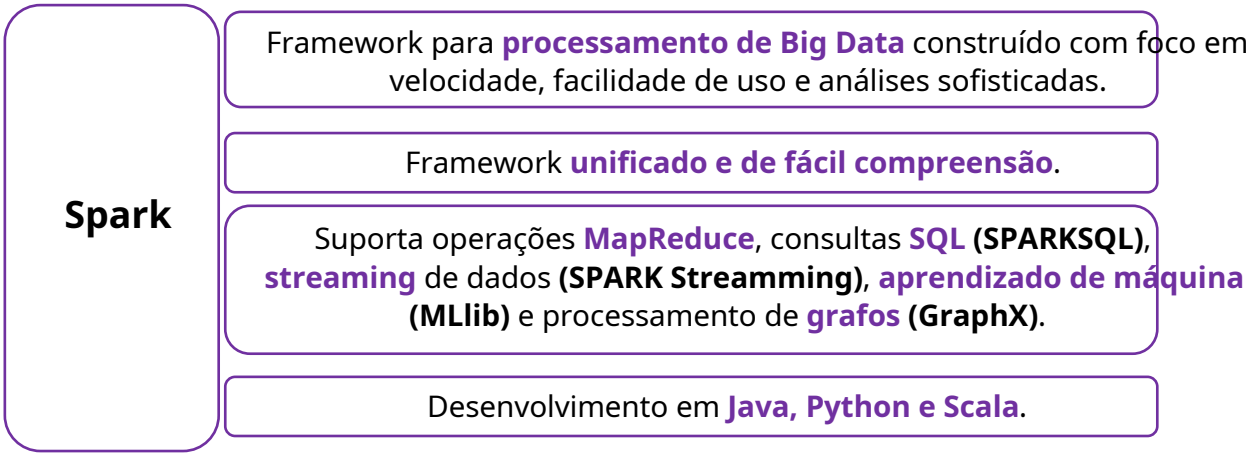
Inicialmente, o **Spark** oferece um **framework unificado e de fácil compreensão pa gerenciar e processar Big Data** com uma variedade de conjuntos de dados de diversas naturezas (por exemplo: texto, grafos, etc), bem como de diferentes origens (batch c streaming de dados em tempo real).

O Spark permite que aplicações em clusters Hadoop executem até 100 vezes mais rápido em memória e até 10 vezes mais rápido em disco, desenvolver rapidamente aplicações em Java, Scala ou Python. Além disso, vem com um conjunto integrado de mais de 80 operadores de alto nível e pode ser usado de forma interativa para consultar dados diretamente do console.

Além das **operações de Map/Reduce, suporta consultas SQL, streaming de dados, aprendizado de máquina e processamento de grafos**. Desenvolvedores podem usar ess recursos no modo stand-alone ou combiná-los em um único pipeline.

O Spark tem diversos componentes para diferentes tipos de processamentos, todos construídos sobre o Spark Core, que é o componente que disponibiliza as funções básicas para o processamento como as funções map, reduce, filter e collect. Entre eles destacam-s

- **Spark Streamming**: possibilita o processamento de fluxos em tempo real;
- **GraphX**: realiza o processamento sobre grafos;
- **SparkSQL:** para a utilização de SQL na realização de consultas e processamento sobre os dados no Spark;
- **MLlib**: biblioteca de aprendizado de máquina, com deferentes algoritmos para as mais diversas atividades, como clustering.



Esquema 15 – Spark.

19- (INSTITUTO AOCP - 2020 - MJSP - Engenheiro de Dados - Big Data) Assim como o Hadoop foi desenvolvido para possibilitar o processamento em lote de gran volume de dados, também surgiram tecnologias com suporte ao processamento em tempo real de Big Data, como o

a) Hadoop RTime.

b) Kubernetes.

c) Elasticsearch.

d) Spark.

e) RealStorm.

Resolução:

O **Spark** é um **framework para processamento de Big Data construído com foco em velocidade, facilidade de uso e análises sofisticadas**. O Spark tem muitas vantagens se comparado as outras tecnologias de Big Data e do paradigma MapReduce, como o Hadoop e o Storm.

Inicialmente, o **Spark** oferece um **framework unificado e de fácil compreensão pa gerenciar e processar Big Data** com uma variedade de conjuntos de dados de diversas naturezas (por exemplo: texto, grafos, etc), bem como de diferentes origens (batch o streaming de dados em tempo real).

Gabarito: Letra D.

4. VISUALIZAÇÃO E ANÁLISE EXPLORATÓRIA

Visualização de dados

A **visualização de dados** é o estudo da **representação visual dos dados**, definidos como informações que podem ser abstraídas de forma esquemática, incluindo atributos ou variáveis das unidades de informação.

O **principal objetivo da visualização de dados** é **comunicar a informação de maneira clara e efetiva utilizando meios gráficos**. Isto não significa que a visualização de dados necessita ter um visual muito sofisticado ou bonito. Para transmitir ideias efetivamente, tanto a forma estética quanto as necessidades funcionais precisam estar equilibradas promovendo a compreensão de um complexo conjunto de dados, comunicando seus principais aspectos de uma forma mais intuitiva. No entanto, projetistas muitas vezes não conseguem alcançar um equilíbrio entre beleza e funcionalidades, criando lindas visualizações, que, no entanto, deixam de servir ao seu principal objetivo - a comunicação de informações. As **visualizações de dados corretamente delineadas providenciam perspectivas-chave sobre conjuntos de dados complexos, através de formas que são significativas e intuitivas**.

A visualização de dados está intimamente relacionada com os gráficos de informação, visualização científica e com gráficos estatísticos. Atualmente, a visualização de dados é muito prática e uma área vital de pesquisas, ensino e desenvolvimento. O termo une os campos da visualização científica e da visualização da informação.

No contexto de business intelligence e big data, as **técnicas modernas de visualização fornecem condições de identificar padrões ou correlações de dados antes invisíveis**. Fazendo as perguntas certas, é possível identificar coisas que estão acontecendo, ou que irão acontecer, se identificarmos corretamente as tendências.

A visualização pode ocorrer através de diversas ferramentas, desde simples planilhas de dados, quadros, gráficos, figuras, dashboards e softwares. As ferramentas OLAP podem ser utilizadas para a visualização de dados em um modelo multidimensional.

De forma **esquemática** temos:

Visualização de dados		
Conceito: representação visual dos dados	**Objetivo**: comunicar a informação de maneira clara e efetiva utilizando meios gráficos.	**Ferramentas**: ex.: planilhas de dados, quadros, gráficos, figuras, dashboards, softwares, OLAP e outras.

Esquema 16 – Visualização de dados.

Análise Exploratória de Dados

A **análise exploratória de dados (AED)** emprega **grande variedade de técnicas gráficas e quantitativas** (inclusive a análise de regressão), visando maximizar a obtenção de informações ocultas nas estruturas, descoberta de variáveis importantes nas tendências e nas variações, detecção de comportamentos anômalos de fenômenos, teste validade hipóteses assumidas, escolha de modelos e determinação de número otimizado de variáve

Os softwares atualmente disponíveis possibilitam que esta técnica se constitua em u ferramenta para **descobrir quais tendências, relações e padrões poderiam estar oculto em uma coleção de dados analisados**. Os investigadores deveriam iniciar sua análise pelo exame dos dados disponíveis, e só depois decidir sobre qual técnica aplicar para equaciona o problema e depois procurar a equação que melhor os represente e interprete.

A **finalidade da Análise Exploratória de Dados** é **examinar os dados previamente à aplicaçãode qualquertécnica estatística**.Desta forma o analista consegueum entendimento básico de seus dados e das relações existentes entre as variáveis analisadas

A Análise Exploratória de Dados é bastante utilizada na estatística através da análise de variáveis principalmente por meio tabelas (tabelas de frequência) e gráficos (histogramas, árvores, diagramas de dispersão, diagramas de caixa, entre outros).

No contexto de business intelligence, a mineração de dados é uma poderosa ferramenta p realizar a análise exploratória de dados, pois constitui-se de um processo que utiliza técnic de estatística, matemática e inteligência artificial para extrair e identificar informações útei e subsequentes conhecimentos (ou padrões) em grandes conjuntos de dados.

Análise exploratória de dados	**Conceito:** abordagem que emprega **grande variedade de técnicas gráficas e quantitativas** para analisar dados.
	Finalidade: examinar os dados previamente à aplicação de qualquer técnica estatística a fim de **descobrir quais tendências, relações e padrões** poderiam estar ocultos em uma coleção de dados analisados.
	Ferramentas: tabelas (tabelas de frequência) e **gráficos** (histogramas, árvores, diagramas de dispersão, diagramas de caixa, entre outros). **Mineração de dados** é usada no contexto de BI.

Esquema 17 – Análise Exploratória de Dados (AED).

As Variáveis na Análise Exploratória de Dados

As características estudadas pela análise exploratória de dados são chamadas **variáveis**. D modo mais completo, chamamos de **variável** toda **característica que se pretende avalia estatisticamente em um determinado conjunto de elementos** (amostra ou população).

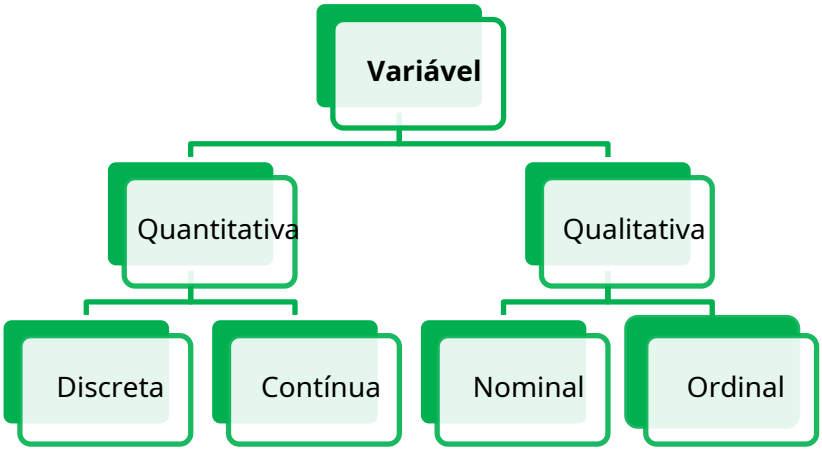
As variáveis podem, quanto ao **nível de mensuração**, ser quantitativas ou qualitativas.

A **variável quantitativa** é aquela que **pode ser expressa em termos numéricos** (e altura, peso, número de ocorrências, dias, etc.). Elas podem ser de duas espécies:

- **Discretas**: são aquelas **expressas em números contáveis**. O número de valores possíveis é **finito ou "enumerável"**. Ex.: número de alunos em uma sala de aula, número de veículos por pessoa.
- **Contínuas**: são aquelas **mensuráveis em um intervalo**, ou seja, podem assumir qualquer número (inteiro ou não) dentro de um intervalo de valores. Ex.: distância percorrida, tempo para executar uma tarefa.

A **variável qualitativa** é aquela **expressa** não por valores, mas **por um atributo**. P exemplo: nacionalidade, religião, profissão, sexo. Elas também podem ser divididas em:

- **Nominais**: são aquelas em que **não se pode estabelecer uma ordem para elas**. Ex.: cor dos olhos, profissão.
- **Ordinais**: são aquelas em que **é possível estabelecer uma ordem**. Ex.: escala de frequência (pouco, médio, muito).



Esquema 18 – Tipos de variáveis (níveis de mensuração).

Quanto ao nível de manipulação, temos as variáveis dependentes ou independentes:

Uma **variável independente** é uma variável que representa uma **grandeza que está send manipulada em um experimento**. São as variáveis manipuladas.

Uma **variável dependente** representa uma **grandeza cujo valor depende de como variável independente é manipulada**. São apenas medidas ou registradas.

De maneira esquemática temos:

Variável independente	Variavél dependente
• Grandeza **manipulada em um experimento**. • São **manipuladas**.	• Grandeza cujo valor **depende de como a variável independente é manipulada**. • São **medidas ou registradas**.

Esquema 19 – Tipos de variáveis (nível de manipulação).

20- (CESPE / CEBRASPE - 2021 - TCE-RJ - Analista de Controle Externo - Especialidade: Tecnologia da Informação) Com relação aos conceitos de análise de dado e informações, julgue o item a seguir.

No nível de mensuração da análise exploratória de dados, as variáveis são classificadas cor dependentes e independentes.

Resolução:

Galera, no **nível de mensuração** as variáveis são classificadas como quantitativas e qualitativas.

A **variável quantitativa** é aquela que **pode ser expressa em termos numéricos** (ex altura, peso, número de ocorrências, dias, etc.).

A **variável qualitativa** é aquela **expressa** não por valores, mas **por um atributo**. P exemplo: nacionalidade, religião, profissão, sexo.

A classificação como dependente e independente é no nível de manipulação.

Uma **variável independente** é uma variável que representa uma **grandeza que está send manipulada em um experimento**. São as variáveis manipuladas.

Uma **variável dependente** representa uma **grandeza cujo valor depende de como variável independente é manipulada**. São apenas medidas ou registradas.

Gabarito: **Errado**.

5. ESQUEMAS DE AULA

Conceito de Big Data

Big Data

- **Conjuntos de dados** muito grandes ou complexos
- **Captura, gerenciamento e a análise de dados** que vão além dos dados tipicamente estruturados
- Aplicativos de processamento de dados **tradicionais ainda não conseguem lidar**
- Frequentemente são dados obtidos de **arquivos não estruturados**

Tipos de análises com Big Data

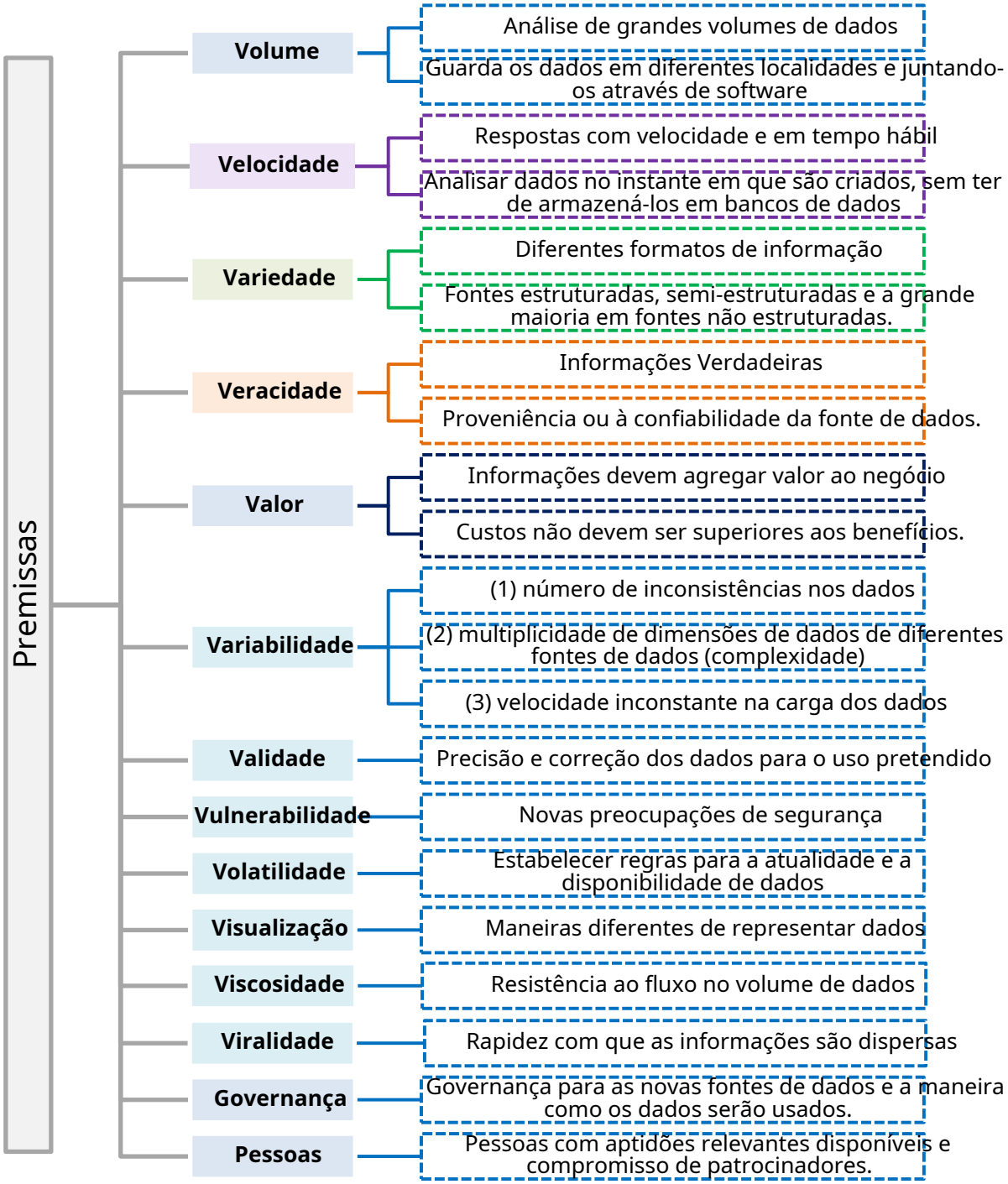
Análise preditiva — Análise de possibilidades futuras com base em dados passados.

Análise prescritiva — Traça as possíveis consequências de cada ação.

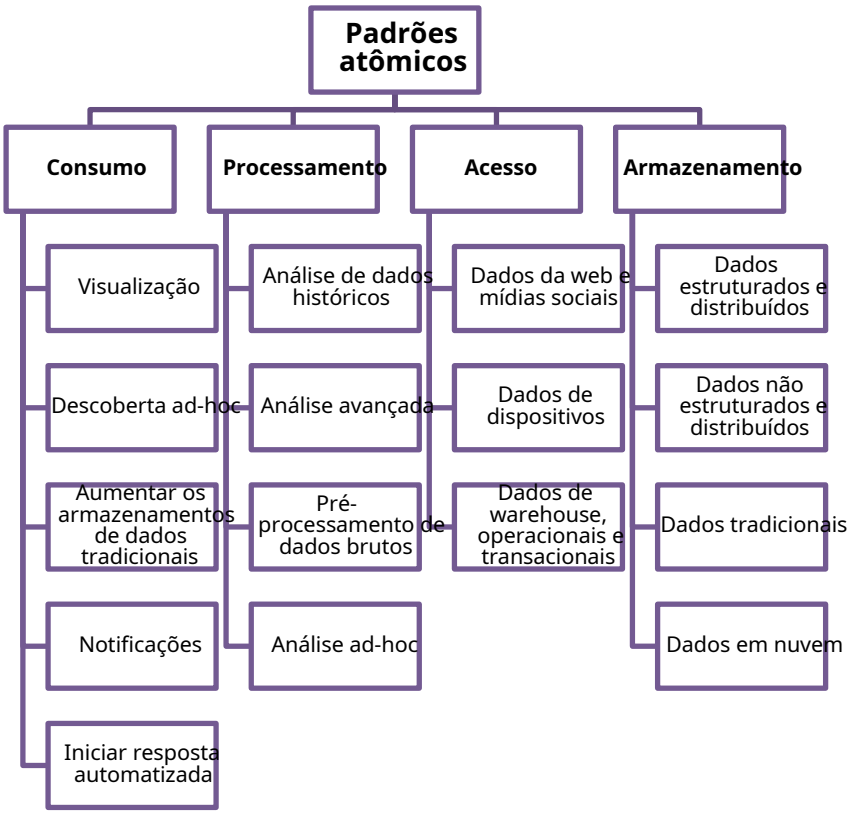
Análise descritiva — Compreensão em tempo real dos acontecimentos.

Análise diagnóstica — Compreensão das possibilidades fornecidas por uma base de dados.

Técnicas de pré-processamento (Navathe)



Padrões atômicos para Big Data



Padrões de consumo

Visualização	Descoberta ad-hoc	Aumentar os armazenamentos de dados tradicionais	Notificação	Iniciar resposta automatizada
Abordagem tradiocional nem sempre fornece a melhor visualização.	Criação de relatórios padrão adequados para todas as necessidades de negócios não é viável.	**Ampliar o escopo de dados** disponível para a analítica atual para incluir dados que residem dentro e fora dos limites organizacionais.	Notificações para **indicar eventos**.	**Acionar ou iniciar outros processos** de negócios ou transações.
O objetivo é fazer com que seja **mais fácil consumir os dados de forma intuitiva**.	Capacidade de **enviar consultas ad hoc (feitas sob demanda)** ao procurar por informações específicas.		Lidar com o volume antecipado de notificações a serem enviadas de maneira oportuna.	

Padrões de processamento

Análise de dados históricos
- **Analisar as tendências históricas para um determinado período, conjunto de períodos e produtos e compará-las** aos dados atuais disponíveis.

Analítica avançada
- **Correlacionar diferentes conjuntos** de dados em muitos contextos.
- Inclui previsões, decisões, processos inferenciais, simulações, informações contextuais e resoluções da entidade.

Pré-processamento de dados brutos
- Para executar a análise em quaisquer dados, eles devem estar em **algum tipo de formato estruturado**.

Análise ad-hoc
- Pelo fato de as fontes e formatos dos dados não serem fixos e exigirem mecanismos diferentes para recuperá-los e processá-los.

Padrões de acesso

Padrão de acesso à web e mídia social

- A web e a mídia social são **úteis em praticamente todas as análises**, mas são necessários mecanismos de acesso diferentes para obter esses dados.

Padrão de dados gerados por dispositivos

- Inclui **dados de sensores**.
- Os dados são **detectados a partir das origens** de dados.

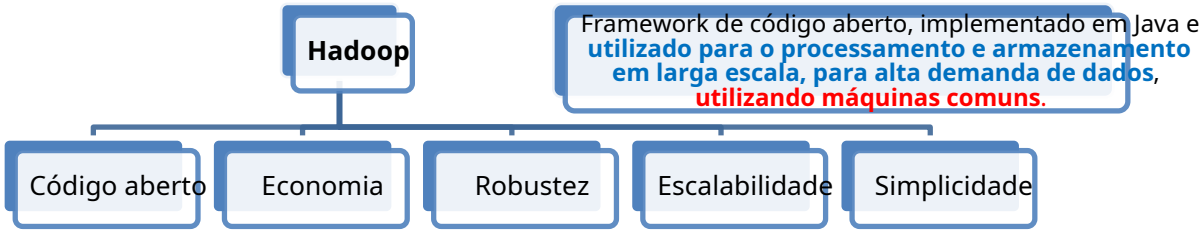
Padrão de dados de warehouse, operacionais e transacionais

- É possível **armazenar os dados de warehouse, operacionais e transacionais** existentes para evitar a limpeza ou o arquivamento deles ou para reduzir a carga no armazenamento tradicional quando os dados são acessados por outros consumidores.

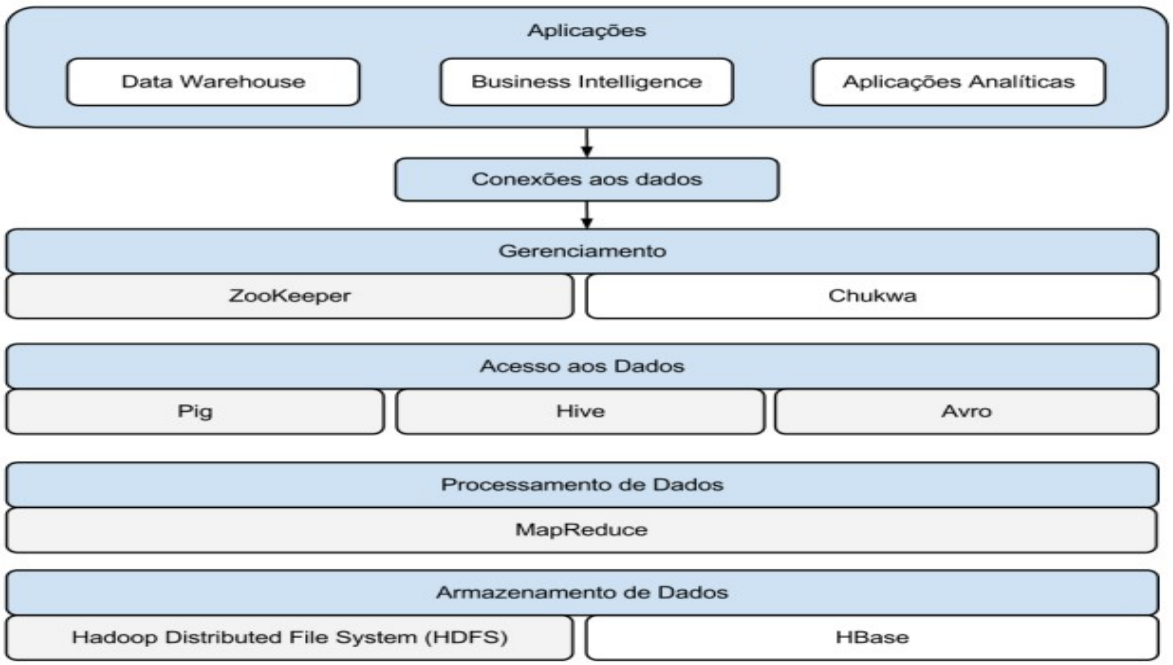
Padrões de armazenamento

Dados não estruturados e distribuídos	Dados estruturados e distribuídos	Armazenamento de dados tradicionais	Armazenamento em nuvem
•Dados não estruturados devem ser armazenados como estão.	•Dados convertidos devem ser armazenados para evitar a frequente conversão de dados brutos para dados estruturados.	•Usar o data warehouse, o sistema relacional e outros armazenamentos de conteúdo existentes.	•Uso de soluções de nuvem que fornecem o gerenciamento de sistemas, manutenção e armazenamento de big data

Hadoop



Subprojetos do Hadoop



MapReduce

Modelo de programação e um **arcabou especializado no processamento de conjuntos de dados distribuídos** em u aglomerado computacional (cluster).

Programador **não precisa se preocupar com o paralelismo**.

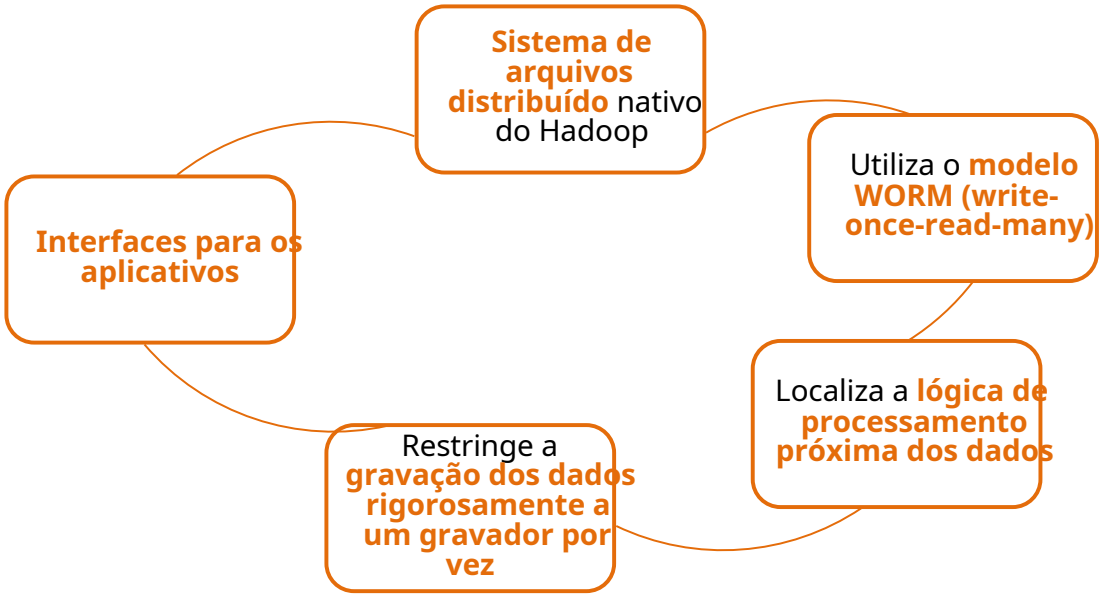
MapReduce

Map: fase de mapeamento, **procesame primário dos dados**.

Reduce: **geração do resultado** final.

A leitura é realizada de **arquivos com pares chave/valor** (geralmente .csv)

HDFS



Processos Hadoop

NameNode

- **Gerenciar os arquivos armazenados** no HDFS.

SecondaryNameNode

- **Auxiliar o NameNode** a manter seu serviço.
- Ser uma **alternativa de recuperação** no caso de uma falha do NameNode.

DataNode

- Efetivamente realizam o **armazenamento dos dados**.

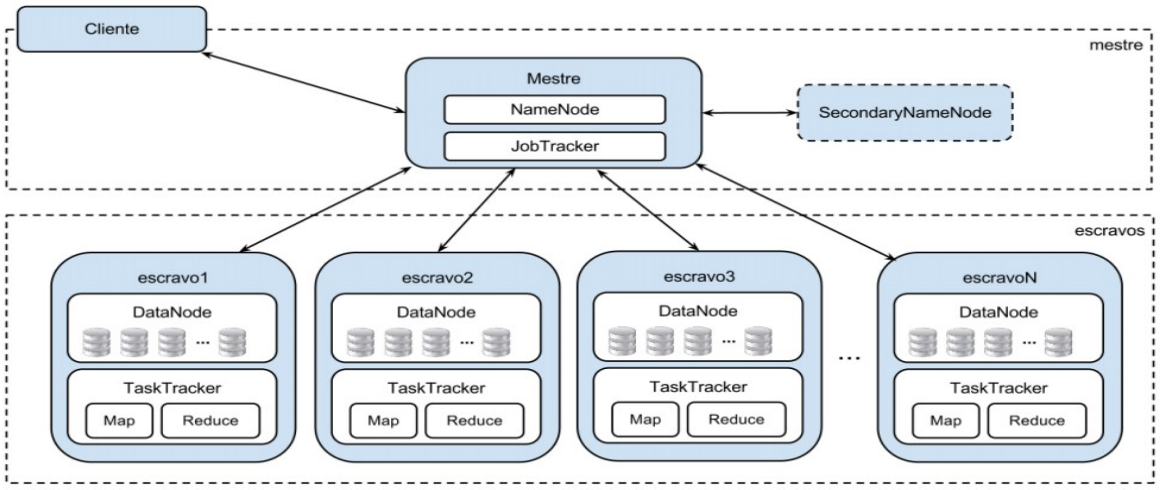
JobTracker

- Função de **gerenciamento sobre o plano de execução das tarefas** a serem processadas pelo MapReduce.

TaskTracker

- Responsável pela **execução de tarefas MapReduce**.

Arquitetura dos Processos Hadoop



Spark

Spark

- Framework para **processamento de Big Data** construído com foco em velocidade, facilidade de uso e análises sofisticadas.
- Framework **unificado e de fácil compreensão**.
- Suporta operações **MapReduce**, consultas **SQL (SPARKSQL)**, **streaming** de dados **(SPARK Streamming)**, **aprendizado de máquina (MLlib)** e processamento de **grafos (GraphX)**.
- Desenvolvimento em **Java, Python e Scala**.

Visualização de dados

Visualização de dados

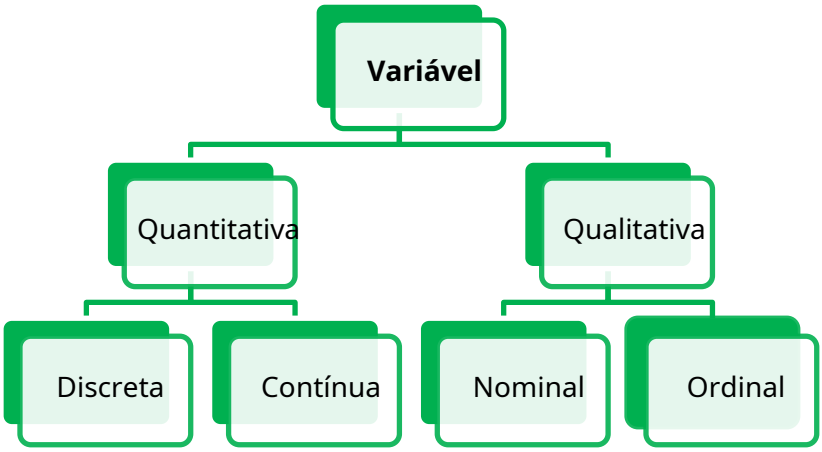
Conceito: representação visual dos dados	**Objetivo**: comunicar a informação de maneira clara e efetiva utilizando meios gráficos.	**Ferramentas**: ex.: planilhas de dados, quadros, gráficos, figuras, dashboards, softwares, OLAP e outras.

Análise Exploratória de Dados (AED)

Análise exploratória de dados

- **Conceito:** abordagem que emprega **grande variedade de técnicas gráficas e quantitativas** para analisar dados.
- **Finalidade:** **examinar os dados previamente** à aplicação de qualquer técnica estatística a fim de **descobrir quais tendências, relações e padrões** poderiam estar ocultos em uma coleção de dados analisados.
- **Ferramentas:** **tabelas** (tabelas de frequência) e **gráficos** (histogramas, árvores, diagramas de dispersão, diagramas de caixa, entre outros). **Mineração de dados** é usada no contexto de BI.

Tipos de variáveis (níveis de mensuração)



Tipos de variáveis (níveis de manipulação)

Variável independente	Variavél dependente
• Grandeza **manipulada em um experimento**. • São **manipuladas**.	• Grandeza cujo valor **depende de como a variável independente é manipulada**. • São **medidas ou registradas**.

6. REFERÊNCIAS

AVILA, Thiago. **O que faremos com os 40 trilhões de gigabytes de dados disponíveis em 2020?** Disponível em: <http://thiagoavila.com.br/sitev2/dados-abertos/o-que-faremos-com-os-40-trilhoes-de-gigabytes-de-dados-disponiveis-em-2020/> Acesso em: 19 dez. 2017.

BIG DATA BUSINESS. **Big Data Analytics**: você sabe o que é? Disponível em: <http://www.bigdatabusiness.com.br/voce-sabe-o-que-e-big-data-analytics/> Acesso em: 19 dez. 2017.

BIG DATA BUSINESS. **Tipos de análise de Big Data:** você conhece todos os 4? Disponível em: <http://www.bigdatabusiness.com.br/conheca-os-4-tipos-de-analises-de-big-data-analytics/> Acesso em: 22 mar. 2019.

BROCKE, Jan Vom; ROSEMANN, Michael. **Metodologia de Pesquisa**. 5ª ed. São Paulo: AMGH Editora, 2013.

CIENCIA E DADOS. **Data Lake, a fonte do Big Data**. Disponível em: < http://www.cienciaedados.com/data-lake-a-fonte-do-big-data/> Acesso em: 15 jan. 2021.

FERNANDES, Aguinaldo Aragon; DE ABREU, Vladimir Ferraz. **Implantando a Governança de TI**: Da estratégia à Gestão de Processos e Serviços. Brasport, 2014.

FAROLBI. **Conheça os 4 tipos de análise do Big Data Analytics!** Disponível em: < https://farolbi.com.br/conheca-os-4-tipos-de-analise-do-big-data-analytics/> Acesso em: 15 jan. 2021.

GOLDMAN, Alfredo et al. **Apache Hadoop**: conceitos teóricos e práticos, evolução e novas possibilidades. XXXI Jornadas de atualizaçoes em informatica, p. 88-136, 2012.

HANSON, J. **Uma Introdução ao Hadoop Distributed File System**. Disponível em: <https://www.ibm.com/developerworks/br/library/wa-introhdfs/index.html> Acesso em: 19 fev. 2018.

IBM. **Como saber se uma solução de big data é ideal para sua organização**. Disponível em: <https://www.ibm.com/developerworks/br/library/bd-archpatterns2/index.html> Acesso em: 19 dez. 2017.

IBM. **Entendendo padrões atômicos e compostos de soluções de big data**. Disponível em: <https://www.ibm.com/developerworks/br/library/bd-archpatterns4/index.html: Acesso em: 20 dez. 2017.

MACHADO, Henrique. **Hadoop MapReduce: Introdução a Big Data**. Disponível em: <https://www.devmedia.com.br/hadoop-mapreduce-introducao-a-big-data/30034>. Acesso em: 19 fev. 2018.

PENCHINKALA, Srini. **Big Data com Apache Spark - Parte 1: Introdução**. Disponível em: <https://www.infoq.com/br/articles/apache-spark-introduction> Acesso em: 19 fev. 2018.

UFRJ. **Os 5 V's do Big Data**. Disponível em: <https://www.gta.ufrj.br/grad/15_1/bigdata/vs.html> Acesso em: 19 dez. 2017.

SIEWERT, Sam B. Big data in the cloud: data velocity, volume, variety veracity. **IBM developersWorks**. July 2013.

TDWI. **The 10 Vs of Big Data**. Disponível em: <https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx> Acesso em: 19 dez. 2017.