Nome **: Angemydelson Saint-Bert**

Questões/Respostas: Big Data (BRISA)

 **Hadoop é uma das tecnologias mais proeminentes no ecossistema de Big Data. Desenvolvida pelo Apache Software Foundation, é uma estrutura de código aberto projetada para armazenar, processar e analisar grandes volumes de dados distribuídos em clusters de servidores. Uma das principais características do Hadoop é a sua abordagem de processamento _____. Ele divide tarefas de processamento em várias máquinas em um cluster, o que permite escalabilidade horizontal para lidar com grandes quantidades de dados. Esse modelo distribuído também garante tolerância a falhas, pois, caso uma máquina falhe, as tarefas são realocadas para outros nós.**

Questão 1:

a.em tempo real

b.individual

c.em lote

d.em série

e.sequencial

 **Marque se as afirmações a seguir são verdadeiras (V) ou falsas (F) com base no tópico sobre Modelagem e Gerenciamento de Dados. 1. ( ) A modelagem de dados no Big Data nunca requer esquemas flexíveis, pois os dados sempre seguem uma estrutura rígida. 2. ( ) A denormalização é aplicada para otimizar o desempenho de consultas em larga escala, evitando a necessidade de várias junções de tabelas. 3. ( ) A modelagem de dados colunar não é adequada para sistemas de Big Data, pois não oferece**

vantagens em termos de compressão e consulta. 4. ( ) Sistemas de arquivos distribuídos, como o Hadoop Distributed File System (HDFS), não são utilizados para armazenamento escalável de dados em clusters de computadores. 5. ( ) Indexação é uma técnica irrelevante para melhorar o desempenho das consultas em grandes conjuntos de dados.
Questão 2:

a.V V F V F

b.F V F V F

c.V F V F F

d.F V F F V

e.V F F V V

 **Qual dos seguintes setores utiliza o Big Data para analisar riscos de crédito, prever tendências do mercado e melhorar a eficiência na gestão de ativos?**
Questão 3:

a.Finanças e Seguros

b.Saúde e Medicina

c.Internet das Coisas (IoT)

d.Educação

e.Setor Público

 **Qual é a principal finalidade do Elasticsearch no ecossistema de Big Data?**
Questão 4:

a.Armazenamento e recuperação de dados não estruturados.

b.Análise de logs de servidores.

c.Processamento de dados em tempo real.

d.Geração de relatórios de business intelligence.

e.Integração com o Apache Hadoop.

 **Qual das seguintes afirmações sobre as metodologias utilizadas no desenvolvimento de projetos Big Data está correta, com base no texto? 1. ( ) A metodologia Agile é mais adequada para projetos de Big Data devido à sua ênfase na colaboração entre equipes de TI, análise de dados e negócios. 2. ( ) A metodologia CRISP-DM é uma abordagem centrada no ser humano, focada em compreender as necessidades dos usuários finais e criar soluções orientadas para o cliente. 3. ( ) A metodologia LEAN é aplicada em projetos de Big Data para lidar especificamente com desafios de escalabilidade, gerenciamento de dados em larga escala e processamento distribuído. 4. ( ) A metodologia CRISP-BigData é uma variação da metodologia Agile, adaptada para lidar com os requisitos complexos e em constante mudança de projetos de Big Data. 5. ( ) A metodologia Design Thinking é adequada para projetos de Big Data devido à sua ênfase na entrega contínua de incrementos funcionais e na adaptação a mudanças de requisitos.**

Questão 5:

a.A metodologia Agile é mais adequada para projetos de Big Data devido à sua ênfase na colaboração entre equipes de TI, análise de dados e negócios.

b.A metodologia LEAN é aplicada em projetos de Big Data para lidar especificamente com desafios de escalabilidade, gerenciamento de dados em larga escala e processamento distribuído.

c.A metodologia CRISP-DM é uma abordagem centrada no ser humano, focada em compreender as necessidades dos usuários finais e criar soluções orientadas para o cliente.

d.A metodologia Design Thinking é adequada para projetos de Big Data devido à sua ênfase na entrega contínua de incrementos funcionais e na adaptação a mudanças de requisitos.

e.A metodologia CRISP-BigData é uma variação da CRISP-DM, adaptada especificamente para projetos de Big Data, considerando as características típicas desse contexto, como escalabilidade e processamento distribuído.

 **Qual das seguintes características não faz parte das principais "Vs" do Big Data, conforme mencionado no texto?**
Questão 6:

a.Valor

b.Variedade

c.Velocidade

d.Validade

e.Volume

 **O componente central do Hadoop é o Hadoop Distributed File System (HDFS), um sistema de arquivos distribuído que divide os dados em blocos e os replica em vários nós para garantir a disponibilidade e a redundância dos dados. Outra parte crucial do Hadoop é o _____, um modelo de programação para processamento paralelo e distribuído. Ele permite que os usuários definam tarefas de mapeamento e redução para processar e analisar os dados em paralelo.**
Questão 7:

a.Flink

b.Hive

c.HBase

d.Pig

e.Spark

 **Complete as lacunas no contexto do texto sobre Processamento e Análise de Dados. 1. O _____ é uma abordagem comum para lidar com o Big Data devido ao tamanho exorbitante dos conjuntos de dados. Tecnologias como Apache Hadoop e Apache Spark permitem dividir tarefas complexas em tarefas menores e executá-las em paralelo em diversos nós de um cluster, garantindo velocidade e eficiência no processamento. 2. A _____ de dados é a etapa inicial do processo, em que os dados são coletados e armazenados em um sistema de gerenciamento, como o Hadoop Distributed File System (HDFS) ou bancos de dados NoSQL. Ferramentas como o Apache Kafka são frequentemente utilizadas para a ingestão de dados em tempo real. 3. Antes da análise propriamente dita, é comum realizar o _____ dos dados. Nesta etapa, os dados são limpos, tratados, transformados e, se necessário, agregados para torná-los adequados para a análise. 4. A _____ de dados é uma fase inicial da análise, na qual os dados são investigados em busca de padrões, tendências e insights iniciais. Gráficos, estatísticas descritivas e outras técnicas podem ser usados para entender melhor os dados. 5. Além da análise exploratória, várias técnicas de _____ podem ser aplicadas, dependendo dos objetivos específicos da organização. Isso inclui análise de regressão, aprendizado de máquina, análise de redes, mineração de texto, entre outras.**

Questão 8:

a.(1) Processamento Exploratório, (2) Análise de Dados, (3) Ingestão de Dados, (4) Pré-Processamento, (5) Visualização de Dados

b.(1) Processamento em Tempo Real, (2) Análise de Dados, (3) Visualização de Dados, (4) Análise Avançada, (5) Ingestão de Dados

c.(1) Análise Exploratória, (2) Pré-Processamento, (3) Ingestão de Dados, (4) Análise de Dados, (5) Processamento Distribuído

d.(1) Análise Avançada, (2) Pré-Processamento, (3) Processamento Distribuído, (4) Ingestão de Dados, (5) Análise Exploratória

e.(1) Processamento Distribuído, (2) Análise Avançada, (3) Visualização de Dados, (4) Análise Exploratória, (5) Ingestão de Dados

 **Marque se as afirmações a seguir são verdadeiras (V) ou falsas (F) com base no tópico sobre Processamento de Consultas. 1. ( ) Indexação é uma técnica que não tem impacto no desempenho das consultas em grandes conjuntos de dados. 2. ( ) Otimização de consultas não envolve técnicas como particionamento de dados e execução paralela de consultas. 3. ( ) A técnica de sharding é utilizada para distribuir os dados em vários nós ou servidores, permitindo escalabilidade horizontal. 4. ( ) Replicação de dados não é relevante para garantir maior disponibilidade e tolerância a falhas em sistemas de Big Data. 5. ( ) Processamento de consultas em tempo real não é uma característica importante em projetos de Big Data.**
Questão 9:

a.V V F V F

b.V F V V F

c.V V V F V

d.F F V F F

e.F F F V V

 **Complete as lacunas no contexto do texto sobre Processamento e Análise de Dados. 1. A análise exploratória de dados é uma fase inicial da análise, na qual os dados são investigados em busca de padrões, tendências e insights _____. Gráficos, estatísticas descritivas e outras técnicas podem ser usados para entender melhor os dados. 2. O processamento em tempo real é necessário para tomar decisões imediatas com base em dados em constante _____. Para isso, são utilizadas tecnologias como Apache Flink, Spark Streaming e outras ferramentas de processamento em tempo real. 3. Além da análise exploratória, várias técnicas de análise mais _____ podem ser aplicadas, dependendo dos objetivos específicos da organização. Isso inclui análise de regressão, aprendizado de máquina, análise de redes, mineração de texto, entre outras. 4. A ingestão de dados é a etapa inicial do processo, em que os dados são coletados e armazenados em um sistema de gerenciamento, como o Hadoop Distributed File System (HDFS) ou bancos de dados _____. Ferramentas como o Apache Kafka são frequentemente utilizadas para a ingestão de dados em tempo real. 5. O _____ é uma abordagem comum para lidar com o Big Data devido ao tamanho exorbitante dos conjuntos de dados. Tecnologias como Apache Hadoop e Apache Spark permitem dividir tarefas complexas em tarefas menores e executá-las em paralelo em diversos nós de um cluster, garantindo velocidade e eficiência no processamento.**

Questão 10:

a.(1) iniciais, (2) constante, (3) avançadas, (4) NoSQL, (5) Processamento Exploratório

b.(1) em tempo real, (2) processamento, (3) distribuída, (4) NoSQL, (5) exploratória

c.(1) iniciais, (2) Processamento Distribuído, (3) avançadas, (4) distribuída, (5) processamento

d.(1) iniciais, (2) mudança, (3) distribuída, (4) NoSQL, (5) Processamento Distribuído

e.(1) iniciais, (2) distribuição, (3) avançadas, (4) NoSQL, (5) Processamento

 **Marque se as afirmações a seguir são verdadeiras (V) ou falsas (F) com base no tópico sobre Processamento Distribuído. 1. ( ) O processamento distribuído envolve o uso de várias máquinas interconectadas para processar grandes volumes de dados de forma simultânea e coordenada. 2. ( ) Uma das vantagens do processamento distribuído é a incapacidade de lidar com o aumento do volume de dados sem perder desempenho. 3. ( ) A redundância e a tolerância a falhas são desafios enfrentados pelo processamento distribuído, pois os nós não conseguem assumir responsabilidades uns dos outros. 4. ( ) A execução de tarefas de forma paralela em diversos nós não resulta em maior velocidade e desempenho, sendo desvantajosa. 5. ( ) O processamento distribuído pode economizar custos ao permitir o uso de clusters de computação mais acessíveis em vez de servidores caros.**

Questão 11:

a.V V V F F

b.V F V V F

c.F V V F V

d.V V F F V

e.F F V V F

 **Por que o processamento em tempo real é crucial em ferramentas de Big Data?**

Questão 12:

a.Para permitir análise de dados à medida que são gerados

b.Para expandir a infraestrutura conforme a demanda

c.Para criar gráficos e dashboards interativos

d.Para realizar análise avançada com machine learning

e.Para garantir a tolerância a falhas

 **Marque se as afirmações a seguir são verdadeiras (V) ou falsas (F) com base no tópico sobre Banco de Dados NoSQL. 1. ( ) Bancos de dados NoSQL são projetados apenas para armazenar dados estruturados em cenários de Big Data. 2. ( ) Escalabilidade horizontal é uma característica que não está presente nos bancos de dados NoSQL. 3. ( ) A replicação de dados nos bancos de dados NoSQL não contribui para maior disponibilidade e tolerância a falhas. 4. ( ) Os bancos de dados NoSQL geralmente oferecem alto desempenho apenas para operações de leitura. 5. ( ) Bancos de dados NoSQL são restritos a um único modelo de dados, não permitindo diferentes tipos de categorias.**
Questão 13:

a.F F F V F

b.V F F V V

c.V F F V F

d.V V V F F

e.F V F F F

 **Qual é uma das principais vantagens do Apache Kafka em relação à ingestão e processamento de dados?**
Questão 14:

a.Integração com o Apache Pig.

b.Processamento em lote de grandes volumes de dados.

c.Armazenamento de dados em formato JSON.

d.Lidar com grandes volumes de dados e alto throughput com baixa latência.

e.Capacidade de correlacionar eventos de diferentes fontes.

 **Qual é a vantagem da aplicação da metodologia Design Thinking em projetos de Big Data, de acordo com o texto? 1. ( ) A metodologia Design Thinking permite um foco exclusivo na escalabilidade e no processamento distribuído, facilitando a adaptação a requisitos em constante mudança. 2. ( ) O Design Thinking é eficaz em projetos de Big Data devido à sua abordagem colaborativa entre equipes de TI, análise de dados e negócios. 3. ( ) A metodologia Design Thinking enfatiza a entrega contínua de incrementos funcionais, permitindo ajustes contínuos e experimentação rápida. 4. ( ) O Design Thinking é centrado no ser humano e é útil para garantir que as análises e insights obtidos estejam alinhados com as necessidades dos usuários finais. 5. ( ) A metodologia Design Thinking é apropriada para projetos de Big Data porque se concentra exclusivamente na compreensão dos dados e na preparação dos mesmos.**
Questão 15:

a.O Design Thinking é centrado no ser humano e é útil para garantir que as análises e insights obtidos estejam alinhados com as necessidades dos usuários finais.

b.A metodologia Design Thinking enfatiza a entrega contínua de incrementos funcionais, permitindo ajustes contínuos e experimentação rápida.

c.A metodologia Design Thinking permite um foco exclusivo na escalabilidade e no processamento distribuído, facilitando a adaptação a requisitos em constante mudança.

d.A metodologia Design Thinking é apropriada para projetos de Big Data porque se concentra exclusivamente na compreensão dos dados e na preparação dos mesmos.

e.O Design Thinking é eficaz em projetos de Big Data devido à sua abordagem colaborativa entre equipes de TI, análise de dados e negócios.

 **Qual é a característica fundamental do Apache Flink que o distingue de outras ferramentas de processamento de dados no ecossistema do Big Data?**

Questão 16:

a.Suporte a agregações complexas.

b.Processamento em lote com baixa latência.

c.Capacidade de processamento de dados distribuídos.

d.Armazenamento eficiente de dados em memória.

e.Integração com o Apache Kafka.

 **Qual das seguintes afirmações sobre o desenvolvimento de projetos Big Data está correta, com base no texto? 1. ( ) A metodologia Agile não é adequada para projetos de Big Data, uma vez que não lida bem com a natureza complexa e em constante mudança dos dados. 2. ( ) A metodologia CRISP-DM é a única abordagem possível para projetos de Big Data, pois sua estrutura cíclica garante resultados significativos em todas as fases. 3. ( ) A metodologia LEAN é preferencialmente aplicada em projetos de Big Data devido ao seu foco na colaboração entre equipes e na iteração constante. 4. ( ) A adaptação do processo de desenvolvimento é desnecessária em projetos de Big Data, já que os requisitos são sempre estáveis e previsíveis. 5. ( ) A natureza complexa e em constante mudança dos dados em projetos Big Data requer flexibilidade e abordagens iterativas para alcançar resultados bem-sucedidos.**

Questão 17:

a.A natureza complexa e em constante mudança dos dados em projetos Big Data requer flexibilidade e abordagens iterativas para alcançar resultados bem-sucedidos.

b.A metodologia CRISP-DM é a única abordagem possível para projetos de Big Data, pois sua estrutura cíclica garante resultados significativos em todas as fases.

c.A metodologia LEAN é preferencialmente aplicada em projetos de Big Data devido ao seu foco na colaboração entre equipes e na iteração constante.

d.A adaptação do processo de desenvolvimento é desnecessária em projetos de Big Data, já que os requisitos são sempre estáveis e previsíveis.

e.A metodologia Agile não é adequada para projetos de Big Data, uma vez que não lida bem com a natureza complexa e em constante mudança dos dados.

 **Marque se as afirmações a seguir são verdadeiras (V) ou falsas (F) com base no tópico sobre Clusterização. 1. ( ) A clusterização é uma técnica que não tem utilidade em projetos de Big Data devido ao volume de dados ser sempre pequeno. 2. ( ) O K-means é um algoritmo de clusterização que não pode ser executado de forma paralela. 3. ( ) A aprendizagem de máquina distribuída não oferece suporte a técnicas de clusterização, limitando suas aplicações em Big Data. 4. ( ) Algoritmos baseados em grafos, como o algoritmo de propagação de afinidade, podem ser aplicados de forma distribuída usando o Apache Spark GraphX. 5. ( ) A clusterização é útil principalmente para casos de uso que envolvem análise de fluxos de dados em tempo real.**
Questão 18:

a.V F F V F

b.F F F V F

c.V F V V F

d.V V V F F

e.F V V F V

 **Qual é a função da Integração de Dados em ferramentas de Big Data?**
Questão 19:

a.Facilitar a análise de informações de diferentes fontes

b.Criar gráficos e dashboards interativos

c.Fornecer processamento em tempo real

d.Realizar análise avançada com machine learning

e.Garantir a expansão da infraestrutura

 **O Hive é amplamente utilizado em aplicações que envolvem análise de big data, processamento de logs, geração de relatórios e business intelligence. Sua linguagem de consulta SQL-like e a capacidade de trabalhar com o Hadoop tornam-no uma escolha popular para organizações que buscam extrair informações valiosas de grandes conjuntos de dados. No entanto, é importante notar que o Hive é projetado para processamento em lote, o que pode resultar em _____ mais altas em comparação com ferramentas de processamento em tempo real, como o Apache Spark. Portanto, a escolha entre o Hive e outras ferramentas dependerá das necessidades específicas do projeto e dos requisitos de desempenho de cada caso de uso.**
Questão 20:

a.redundâncias

b.performances

c.taxas de transferência

d.latências

e.velocidades

 **Marque se as afirmações a seguir são verdadeiras (V) ou falsas (F) com base no tópico sobre Armazenamento. 1. ( ) Sistemas de Arquivos Distribuídos não são usados em projetos de Big Data, pois são ineficientes para gerenciar grandes volumes de dados. 2. ( ) Bancos de Dados NoSQL não oferecem escalabilidade e flexibilidade para armazenar dados não estruturados ou semiestruturados. 3. ( ) O armazenamento em nuvem é uma opção que não oferece flexibilidade ou escalabilidade para lidar com grandes volumes de dados. 4. ( ) Armazenamento em Memória é uma tecnologia que não contribui para o aumento do desempenho em operações de leitura e gravação. 5. ( ) A escolha da tecnologia de armazenamento é irrelevante na construção de uma arquitetura de Big Data escalável.**
Questão 21:

a.V F F V V

b.F V V F F

c.V V F V F

d.F F V V V

e.F F F F F

 **Qual das seguintes características não é comum em ferramentas de Big Data?**
Questão 22:

a.Processamento em Tempo Real

b.Processamento Distribuído

c.Armazenamento Distribuído

d.Análise Simples

e.Escalabilidade

 **Complete as lacunas no contexto do texto sobre Processamento e Análise de Dados. 1. A ingestão de dados é a etapa inicial do processo, em que os dados são coletados e armazenados em um sistema de _____, como o Hadoop Distributed File System (HDFS) ou bancos de dados NoSQL. Ferramentas como o Apache Kafka são frequentemente utilizadas para a ingestão de dados em tempo real. 2. O processamento em tempo real é necessário para tomar decisões imediatas com base em dados em _____. Para isso, são utilizadas tecnologias como Apache Flink, Spark Streaming e outras ferramentas de processamento em tempo real. 3. Além da análise exploratória, várias técnicas de análise mais _____ podem ser aplicadas, dependendo dos objetivos específicos da organização. Isso inclui análise de regressão, aprendizado de máquina, análise de redes, mineração de texto, entre outras. 4. A análise exploratória de dados é uma fase inicial da análise, na qual os dados são investigados em busca de padrões, tendências e insights _____. Gráficos, estatísticas descritivas e outras técnicas podem ser usados para entender melhor os dados. 5. O _____ é uma abordagem comum para lidar com o Big Data devido ao tamanho exorbitante dos conjuntos de dados. Tecnologias como Apache Hadoop e Apache Spark permitem dividir tarefas complexas em tarefas menores e executá-las em paralelo em diversos nós de um cluster, garantindo velocidade e eficiência no processamento.**

Questão 23:

a.(1) Análise Avançada, (2) Processamento Distribuído, (3) distribuição, (4) iniciais, (5) Processamento Avançado

b.(1) Análise de Dados, (2) mudança, (3) avançadas, (4) exploratória, (5) distribuída

c.(1) distribuída, (2) constante, (3) avançadas, (4) iniciais, (5) Processamento Distribuído

d.(1) gerenciamento, (2) constante, (3) distribuída, (4) iniciais, (5) Processamento Exploratório

e.(1) distribuição, (2) Processamento Distribuído, (3) iniciais, (4) exploratória, (5) distribuída