Questão 1/10 - Big Data

GraphX é o componente do Spark para computação de grafos em sistemas distribuídos de larga escala. Ele foi desenvolvido através de um projeto de pesquisa como forma de unificar o processamento de grafos e de sistemas paralelos até se tornar uma parte integral do projeto Spark.   
  
Avalie as seguintes afirmações a respeito do GraphX:  
  
I – A API GraphX implementa abstrações de arestas e vértices baseados em extensões de RDDs

II – A computação de grafos é importante em casos onde as relações entre os dados são mais importantes que os dados em si.

III – Grafos são uma forma muito eficiente de representar dados de tabelas esparsas.

IV – O algoritmo *Connected Components* é um tipo de algoritmo paralelo de grafos que é capaz de medir a coesão entre os seus dados (componentes).

V – O algoritmo *PageRank* desenvolvido pelo Google é capaz de classificar a importância de sites através da quantidade de referências apontadas para um site.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Todas, exceto IV  Você assinalou essa alternativa (A)  **Você acertou!**  Justificativa: O algoritmo Connected Components busca encontrar todos os componentes de um grafo que estão conectados entre si onde cada vértice (ou componente) dentro de um grupo pode ser alcançado a partir de qualquer outro vértice do grupo. Além disso não deve haver nenhum caminho, ou seja, arestas, entre dois grupos distintos. Diferente do que alega INCORRETAMENTE a afirmação IV. Todas as outras afirmações (I, II, III e V) estão CORRETAS. |
|  | B | I, II e III |
|  | C | II, III e IV |
|  | D | IV e V |
|  | E | Todas |

Questão 2/10 - Big Data

Analisar os requisitos de um projeto é um dos pontos mais importantes no desenvolvimento de aplicações, uma vez que tais requisitos devem refletir todas as necessidades e críticas dos usuários.  
  
A respeito dos requisitos de aplicações de Big Data, avalie as seguintes afirmações:  
  
I – Deve-se utilizar uma solução com tecnologias de Big Data para qualquer tamanho de dados, uma vez que as vantagens da aplicação de tais tecnologias são vantajosas em diversos cenários.

II – Pode ser muito caro mover os dados armazenado em um determinado sistema de armazenamento para outro lugar, uma vez que o local de armazenamento foi definido. Por isso deve-se tentar prever cenários futuros da melhor forma possível.

III – Escalar a complexidade de um sistema para atender uma demanda além do previsto é uma boa estratégia que pode auxiliar na antecipação de custos de manutenção futuros. É sempre uma boa estratégia, avaliar qual é o máximo de infraestrutura necessária para operar a aplicação.

IV – Em muitos casos pode ser necessário manter os dados armazenados por um tempo maior por motivos de auditoria ou por necessidades jurídicas. Por isso é importante a definição de uma política de retenção de dados.

V – É importante avaliar a relação do sistema com o teorema CAP. É sempre necessário saber se devemos priorizar a disponibilidade ou a consistência dos dados.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, II e V |
|  | B | I e III  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: Não devemos utilizar uma tecnologia de big data, como o armazenamento distribuído se o volume de dados previsto não é grande o suficiente. Pois corre-se o risco de ter uma solução demasiadamente complexa desnecessariamente. Assim a afirmação I está INCORRETA. Escalar a complexidade de um sistema para atender uma demanda além do previsto não é uma boa estratégia. Você pode aumentar os custos de manutenção sem que haja demanda para tal. Além disso, várias das tecnologias em big data são intercambiáveis e, dessa forma, podem ser substituídas facilmente. Tente sempre avaliar qual é o mínimo de infraestrutura necessária para operar a aplicação. Portanto a afirmação III é INCORRETA. As demais afirmações (II, IV e V) estão CORRETAS. |
|  | C | Nenhuma |
|  | D | II, III e IV |
|  | E | II, IV e V |

Questão 3/10 - Big Data

A API GraphX oferece uma forma de armazenar grafos na forma de tabelas e utilizar operações de tabelas para expressar operações de grafos.  
  
A respeito da API do GraphX avalie as seguintes afirmações:  
  
I – Um grafo de propriedades em GraphX é uma abstração capaz de definir um multigrafo dirigido com objetos definidos para cada vértice e aresta.

II – Um grafo de propriedade pode ser definido como um par de conjuntos de RDDs que codificam as características de vértices e arestas.

III – Grafos de propriedades são distribuídos, tolerantes a falhas e seus valores e estruturas podem ser modificados facilmente.

IV – O construtor da classe Edge recebe os parâmetros que correspondem aos identificadores dos vértices de origem e destino.

V – O membro *vertices* da classe Graph retorna um *VertexRDD[(String,String)]* que é uma extensão do objeto *RDD[(VertexId,(String,String)]*. Por outro lado, o membro *edge* retorna um *EdgeRDD* que contém um objeto *Edge[String]*.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, II e III |
|  | B | I, III e IV |
|  | C | Apenas III  Você assinalou essa alternativa (C)  **Você acertou!**  Justificativa: Diferentemente do que diz INCORRETAMENTE a afirmação III, grafos de propriedade não podem ser modificados, pois da mesma forma que RDDs, grafos de propriedades são imutáveis, distribuídos e tolerantes a falhas. Qualquer mudança em valores ou estrutura do grafo produz um novo grafo. Todas as outras afirmações (I, II, IV e V) estão CORRETAS. |
|  | D | Todas |
|  | E | II, IV e V |

Questão 4/10 - Big Data

Avalie as seguintes afirmações sobre os componentes do Hadoop: Impala e Accumulo:  
  
I – Impala é um motor de consultas SQL capaz de realizar consultas de baixa latência em HDFS ou Hive.

II – O Impapa foi projetado para consultas analíticas em Haddop utilizando SQL ou ferramentas de *business inteligence* (BI).

III – Accumulo é um sistema de armazenamento distribuído baseado em chave-valor que implementa segurança e nível de célula.

IV – Todos os dados armazenados no Accumulo devem possuir os mesmos requisitos de segurança em uma mesma tabela.

V – O Accumulo permite realizar operações em pares de chave-valor assim que são inseridos.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, II e III |
|  | B | Todas |
|  | C | I, II e IV |
|  | D | II, III e V  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: Entre as principais características do Impala destaca-se a possibilidade de realizar consultas SQL de baixa latência de dados armazenados em HDFS e **HBase**. Dessa forma a afirmação I encontra-se INCORRETA. Uma das principais características do Accumulo é a segurança em nível de célula, onde cada par chave-valor possui o seu próprio rótulo de segurança que tem a capacidade de limitar os resultados de uma consulta baseado nas autorizações de acesso do usuário. Portanto a afirmação IV está INCORRETA. As afirmações II, III e V estão CORRETAS. |
|  | E | Apenas III |

Questão 5/10 - Big Data

Para que o valor dos dados contidos em um Data Lake não seja perdido, se faz necessário criar uma solução de Data Lake que inclua as noções de gerenciamento, acessibilidade e governança.  
  
Avalie as seguintes afirmações a respeito do gerenciamento de Data Lake:  
  
I – Um Data Lake oferece o mesmo nível de complexidade que outras estratégias de armazenamento como Data Mart e Data Warehouse.

II – Uma das características que podem ser agregadas aos dados armazenados em uma estratégia de Data Lake utilizando bons esquemas de governança é a transparência no uso dos dados.

III – Entre as informações que se podem obter dos dados em um Data Lake podemos destacar os metadados técnicos que fornecem informações sobre a qualidade, perfil, origem, e a sua linhagem.

IV – Os metadados técnicos de um Data Lake podem fornecer informações a respeito da forma ou estrutura dos dados tais como tamanho, tipo de dado ou esquema.

V – Os metadados de um Data Lake permitem buscar, localizar e aprender sobre os dados armazenados.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | II, IV e V  Você assinalou essa alternativa (A)  **Você acertou!**  Justificativa: A estratégia de armazenamento de dados Data Lake pode oferecer uma redução de complexidade, que as estratégias Data Warehouse e Data Mart não são capazes de oferecer. Dessa forma entendemos que a afirmação I está INCORRETA. Metadados técnicos são os metadados que fornecem informações a respeito da forma ou estrutura dos dados tais como: tamanho, tipo de dado, esquema. Portanto a afirmação III está INCORRETA. As demais afirmações (II, IV e V) estão CORRETAS. |
|  | B | Todas, exceto IV |
|  | C | I, II e V |
|  | D | III, IV e V |
|  | E | Todas |

Questão 6/10 - Big Data

O funcionamento estrutural de um fluxo de aprendizado de máquina pode ser descrito através das seguintes etapas: Utilizamos um conjunto de dados de teste e de treino para escolher os modelos de aprendizado de máquina mais adequados, treiná-los, e avaliá-los com o objetivo de alcançar um modelo que seja capaz de realizar predições úteis a partir de dados do mundo real.

Avalie as seguintes afirmações a respeito do fluxo de aprendizado de máquina:

I – Na etapa de obtenção de dados são adquiridos os dados que serão utilizados tanto para treinamento dos modelos quanto para avaliar a eficiência dos modelos.

II – Os problemas de dados faltantes, dados com ruído ou dados inconsistentes devem ser resolvidos na etapa de obtenção dos dados.

III – Os dados devem ser formatados da maneira mais adequada durante a etapa de preparação dos dados.

IV – A etapa de treinamento é quando utilizamos os dados de treino na expectativa de que o modelo aprenda a processar os dados da forma esperada.

V – Durante a etapa de avaliação o desempenho do modelo é avaliado e caso não seja considerado suficiente, o modelo passa por um novo ciclo de ajuste de parâmetros, treinamento e avaliação.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, IV e V |
|  | B | Apenas II  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: Diferentemente do que diz INCORRETAMENTE a afirmação II, os problemas de dados faltantes, dados com ruído ou dados inconsistentes devem ser resolvidos na etapa de Preparação dos Dados. As afirmações I, III, IV e V estão CORRETAS. |
|  | C | II e III |
|  | D | Todas, exceto IV |
|  | E | Nenhuma |

Questão 7/10 - Big Data

O Kafka consiste e um uma plataforma distribuída e, dessa forma, é executado através de um cluster. Um Cluster Kafka armazena fluxos de eventos que são divididos em categorias denominadas tópicos.  
  
A respeito dos conceitos de eventos e tópicos em Kafka avalie as seguintes afirmações:  
  
I – Cada evento é constituído de um par chave-valor.

II – Os servidores Kafka são capazes de armazenar a indicação de qual a última mensagem que cada cliente inscrito em um tópico recebeu.

III – As aplicações cliente podem se inscrever em um ou mais tópicos.

IV – O Kafka é capaz de implementar sistemas de mensagens baseados em publicação-inscrição, mas não é capaz de entregar a implementação de uma fila de mensagens, uma vez que os dados dos eventos são armazenados de forma permanente e permanecem disponíveis.

V – O Kafka permite a implementação de um sistema de armazenamento a partir do armazenamento dos dados em disco. Outra característica é que os dados armazenados no Kafka são replicados para garantir o princípio da tolerância a falhas.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | II, III e V  Você assinalou essa alternativa (A)  **Você acertou!**  Justificativa: Ao contrário do que afirma INCORRETAMENTE a afirmação I, cada evento consiste em uma chave, um valor e um *timestamp*. A afirmação IV está INCORRETA, pois o Kafka é capaz de implementar sistemas de mensagens que utilizam tanto o modelo de fila de mensagens quanto o modelo de publicação-inscrição. Além disso, os dados enviados para o Kafka são armazenados em disco e replicados para garantir o princípio da tolerância a falhas. Todas as outras afirmações (II, III e V) estão CORRETAS. |
|  | B | I, II e III |
|  | C | Todas |
|  | D | I, III e IV |
|  | E | Todas, exceto IV |

Questão 8/10 - Big Data

O aprendizado de máquina é parte do campo da Inteligência Artificial que se refere ao estudo de modelos estatísticos para resolver problemas específicos com padrões e inferências.  
  
Sobre as diferentes categorias de aprendizado de máquina avalie as seguintes afirmações:  
  
I – Os algoritmos de classificação e de regressão fazem parte da categoria de algoritmos de aprendizado supervisionado.

II – Os algoritmos de aprendizado supervisionado buscam identificar estruturas inerentes ao conjunto de dados.

III – O que diferencia os algoritmos de aprendizado por reforço dos algoritmos de aprendizado não-supervisionado é que no aprendizado por reforço os dados nunca são explicitamente relacionados com os seus rótulos.

IV – Os algoritmos de aprendizado por reforço se baseiam nos conceitos de reforço e punição para maximizar parâmetros através da tomada de ações.

V – Os algoritmos de clustering fazem parte da categoria de aprendizagem não-supervisionados.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, II, e III |
|  | B | Todas |
|  | C | I, III e IV |
|  | D | I, IV e V  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: Os algoritmos de aprendizado supervisionado utilizam conjuntos de dados que contém ambos dados de entrada e os resultados esperados, ou seja, os dados de entrada são rotulados. Os algoritmos que buscam identificar estruturas inerentes ao conjunto de dados são os de aprendizado não-supervisionado. Portanto a afirmação II está INCORRETA. Os algoritmos de aprendizado não-supervisionado também não relacionam explicitamente os dados aos rótulos. A afirmação III INCORRETAMENTE confunde os algoritmos de aprendizado supervisionado com os algoritmos de aprendizado não-supervisionado. As demais afirmações (I, IV e V) estão CORRETAS. |
|  | E | Todas, exceto III |

Questão 9/10 - Big Data

Recomendações se baseiam nas previsões de preenchimento das avaliações desconhecidas. Existem duas principais abordagens para isso: recomendações baseadas em conteúdo (*content-based recommendations*) e filtragem colaborativa (*collaborative filtering*).  
  
Avalie as seguintes avaliações sobre as duas principais abordagens de recomendações:  
  
I – O sistema de recomendações baseadas em conteúdo possuem como estratégia a recomendação de itens mais parecidos aos que o usuário avaliou.

II – Antes de recomendar um novo item, o sistema de recomendações baseadas em conteúdo precisa que algum usuário avalie o novo item.

III – O cálculo de similaridade entre os usuários é utilizado por sistemas de filtragem colaborativa para medir a chance de um par de usuários avaliar determinado item de forma parecida.

IV – Métodos de filtragem colaborativa pode ser aplicado apenas para medir a semelhança entre usuários.

V – É possível utilizar métodos de diferentes abordagens de recomendação de uma forma híbrida para tentar melhorar a qualidade das predições.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Todas, exceto V |
|  | B | I e III |
|  | C | Nenhuma |
|  | D | II e IV  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: Na afirmação II, é incorreto dizer que o sistema de recomendações baseadas em conteúdo precisa que algum usuário avalie o novo item, pois este sistema se baseia nas características do item e não nas avaliações de usuários. A afirmação IV diz que os métodos de filtragem colaborativa podem ser aplicados apenas para medir a semelhança entre usuários, porém podemos aplicar o mesmo princípio para as avaliações entre itens (ou *Item-Item Collaborative Filtering*). Portanto tal afirmação está INCORRETA. As demais afirmações (I, III e V) estão corretas. |
|  | E | II e V |

Questão 10/10 - Big Data

Uma vez que conhecemos os conceitos que envolvem a arquitetura big data assim como as aplicações que implementam suas características, podemos pensar em como combinar todas essas tecnologias e ideias para o desenvolvimento de um produto visando atender a alguma necessidade existente.  
  
Avalie as seguintes afirmações a respeito do desenvolvimento e design de soluções Big Data:  
  
I – O desenvolvimento de aplicações Big Data possui grandes desafios em como combinar todas as suas tecnologias e ideias para desenvolver novos produtos. Porém a integração com tecnologias tradicionais é uma questão trivial atualmente.

II – Em muitos casos as soluções de Big Data têm como objetivo otimizar processos de negócios, adquirir vantagens competitivas ou otimizar operações. Dessa forma podem representar muitas vantagens para uma empresa.

III – O entendimento das necessidades que nossa aplicação busca atender é uma das maneiras mais importantes de se obter as informações necessárias para guiar o desenvolvimento

IV – A documentação de requisitos é capaz de evitar desperdícios e garantir que a implementação está alinhada com as necessidades do usuário.

V – Working Backwards é uma abordagem de desenvolvimento de produtos que prioriza o entendimento dos requisitos e o feedback do usuário antes mesmo do início do desenvolvimento.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | II, III e IV |
|  | B | I, IV e V |
|  | C | I, II e IV |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D)  **Você acertou!**  Todas exceto I  Justificativa: Existem desafios no que se trata da complexidade em integrar sistemas de big data com os sistemas tradicionais existentes. Ao contrário do que alega a afirmação I que está INCORRETA. Todas as outras afirmações (II, III, IV e V) estão CORRETAS. |
|  | E | Todas |

Questão 1/10 - Big Data

O Kafka consiste e um uma plataforma distribuída e, dessa forma, é executado através de um cluster. Um Cluster Kafka armazena fluxos de eventos que são divididos em categorias denominadas tópicos.  
  
A respeito dos conceitos de eventos e tópicos em Kafka avalie as seguintes afirmações:  
  
I – Cada evento é constituído de um par chave-valor.

II – Os servidores Kafka são capazes de armazenar a indicação de qual a última mensagem que cada cliente inscrito em um tópico recebeu.

III – As aplicações cliente podem se inscrever em um ou mais tópicos.

IV – O Kafka é capaz de implementar sistemas de mensagens baseados em publicação-inscrição, mas não é capaz de entregar a implementação de uma fila de mensagens, uma vez que os dados dos eventos são armazenados de forma permanente e permanecem disponíveis.

V – O Kafka permite a implementação de um sistema de armazenamento a partir do armazenamento dos dados em disco. Outra característica é que os dados armazenados no Kafka são replicados para garantir o princípio da tolerância a falhas.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | II, III e V  Justificativa: Ao contrário do que afirma INCORRETAMENTE a afirmação I, cada evento consiste em uma chave, um valor e um *timestamp*. A afirmação IV está INCORRETA, pois o Kafka é capaz de implementar sistemas de mensagens que utilizam tanto o modelo de fila de mensagens quanto o modelo de publicação-inscrição. Além disso, os dados enviados para o Kafka são armazenados em disco e replicados para garantir o princípio da tolerância a falhas. Todas as outras afirmações (II, III e V) estão CORRETAS. |
|  | B | I, II e III |
|  | C | Todas |
|  | D | I, III e IV |
|  | E | Todas, exceto IV  Você assinalou essa alternativa (E) |

Questão 2/10 - Big Data

Uma vez que conhecemos os conceitos que envolvem a arquitetura big data assim como as aplicações que implementam suas características, podemos pensar em como combinar todas essas tecnologias e ideias para o desenvolvimento de um produto visando atender a alguma necessidade existente.  
  
Avalie as seguintes afirmações a respeito do desenvolvimento e design de soluções Big Data:  
  
I – O desenvolvimento de aplicações Big Data possui grandes desafios em como combinar todas as suas tecnologias e ideias para desenvolver novos produtos. Porém a integração com tecnologias tradicionais é uma questão trivial atualmente.

II – Em muitos casos as soluções de Big Data têm como objetivo otimizar processos de negócios, adquirir vantagens competitivas ou otimizar operações. Dessa forma podem representar muitas vantagens para uma empresa.

III – O entendimento das necessidades que nossa aplicação busca atender é uma das maneiras mais importantes de se obter as informações necessárias para guiar o desenvolvimento

IV – A documentação de requisitos é capaz de evitar desperdícios e garantir que a implementação está alinhada com as necessidades do usuário.

V – Working Backwards é uma abordagem de desenvolvimento de produtos que prioriza o entendimento dos requisitos e o feedback do usuário antes mesmo do início do desenvolvimento.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | II, III e IV |
|  | B | I, IV e V |
|  | C | I, II e IV |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D)  **Você acertou!**  Todas exceto I  Justificativa: Existem desafios no que se trata da complexidade em integrar sistemas de big data com os sistemas tradicionais existentes. Ao contrário do que alega a afirmação I que está INCORRETA. Todas as outras afirmações (II, III, IV e V) estão CORRETAS. |
|  | E | Todas |

Questão 3/10 - Big Data

Analisar os requisitos de um projeto é um dos pontos mais importantes no desenvolvimento de aplicações, uma vez que tais requisitos devem refletir todas as necessidades e críticas dos usuários.  
  
A respeito dos requisitos de aplicações de Big Data, avalie as seguintes afirmações:  
  
I – Deve-se utilizar uma solução com tecnologias de Big Data para qualquer tamanho de dados, uma vez que as vantagens da aplicação de tais tecnologias são vantajosas em diversos cenários.

II – Pode ser muito caro mover os dados armazenado em um determinado sistema de armazenamento para outro lugar, uma vez que o local de armazenamento foi definido. Por isso deve-se tentar prever cenários futuros da melhor forma possível.

III – Escalar a complexidade de um sistema para atender uma demanda além do previsto é uma boa estratégia que pode auxiliar na antecipação de custos de manutenção futuros. É sempre uma boa estratégia, avaliar qual é o máximo de infraestrutura necessária para operar a aplicação.

IV – Em muitos casos pode ser necessário manter os dados armazenados por um tempo maior por motivos de auditoria ou por necessidades jurídicas. Por isso é importante a definição de uma política de retenção de dados.

V – É importante avaliar a relação do sistema com o teorema CAP. É sempre necessário saber se devemos priorizar a disponibilidade ou a consistência dos dados.

Dentre essas afirmações estão INCORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I, II e V |
|  | B | I e III  Justificativa: Não devemos utilizar uma tecnologia de big data, como o armazenamento distribuído se o volume de dados previsto não é grande o suficiente. Pois corre-se o risco de ter uma solução demasiadamente complexa desnecessariamente. Assim a afirmação I está INCORRETA. Escalar a complexidade de um sistema para atender uma demanda além do previsto não é uma boa estratégia. Você pode aumentar os custos de manutenção sem que haja demanda para tal. Além disso, várias das tecnologias em big data são intercambiáveis e, dessa forma, podem ser substituídas facilmente. Tente sempre avaliar qual é o mínimo de infraestrutura necessária para operar a aplicação. Portanto a afirmação III é INCORRETA. As demais afirmações (II, IV e V) estão CORRETAS. |
|  | C | Nenhuma  Você assinalou essa alternativa (C) |
|  | D | II, III e IV |
|  | E | II, IV e V |

Questão 4/10 - Big Data

Para que o valor dos dados contidos em um Data Lake não seja perdido, se faz necessário criar uma solução de Data Lake que inclua as noções de gerenciamento, acessibilidade e governança.  
  
Avalie as seguintes afirmações a respeito do gerenciamento de Data Lake:  
  
I – Um Data Lake oferece o mesmo nível de complexidade que outras estratégias de armazenamento como Data Mart e Data Warehouse.

II – Uma das características que podem ser agregadas aos dados armazenados em uma estratégia de Data Lake utilizando bons esquemas de governança é a transparência no uso dos dados.

III – Entre as informações que se podem obter dos dados em um Data Lake podemos destacar os metadados técnicos que fornecem informações sobre a qualidade, perfil, origem, e a sua linhagem.

IV – Os metadados técnicos de um Data Lake podem fornecer informações a respeito da forma ou estrutura dos dados tais como tamanho, tipo de dado ou esquema.

V – Os metadados de um Data Lake permitem buscar, localizar e aprender sobre os dados armazenados.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | II, IV e V  Justificativa: A estratégia de armazenamento de dados Data Lake pode oferecer uma redução de complexidade, que as estratégias Data Warehouse e Data Mart não são capazes de oferecer. Dessa forma entendemos que a afirmação I está INCORRETA. Metadados técnicos são os metadados que fornecem informações a respeito da forma ou estrutura dos dados tais como: tamanho, tipo de dado, esquema. Portanto a afirmação III está INCORRETA. As demais afirmações (II, IV e V) estão CORRETAS. |
|  | B | Todas, exceto IV |
|  | C | I, II e V |
|  | D | III, IV e V |
|  | E | Todas  Você assinalou essa alternativa (E) |

Questão 5/10 - Big Data

Recomendações se baseiam nas previsões de preenchimento das avaliações desconhecidas. Existem duas principais abordagens para isso: recomendações baseadas em conteúdo (*content-based recommendations*) e filtragem colaborativa (*collaborative filtering*).  
  
Avalie as seguintes avaliações sobre as duas principais abordagens de recomendações:  
  
I – O sistema de recomendações baseadas em conteúdo possuem como estratégia a recomendação de itens mais parecidos aos que o usuário avaliou.

II – Antes de recomendar um novo item, o sistema de recomendações baseadas em conteúdo precisa que algum usuário avalie o novo item.

III – O cálculo de similaridade entre os usuários é utilizado por sistemas de filtragem colaborativa para medir a chance de um par de usuários avaliar determinado item de forma parecida.

IV – Métodos de filtragem colaborativa pode ser aplicado apenas para medir a semelhança entre usuários.

V – É possível utilizar métodos de diferentes abordagens de recomendação de uma forma híbrida para tentar melhorar a qualidade das predições.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Todas, exceto V |
|  | B | I e III |
|  | C | Nenhuma |
|  | D | II e IV  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: Na afirmação II, é incorreto dizer que o sistema de recomendações baseadas em conteúdo precisa que algum usuário avalie o novo item, pois este sistema se baseia nas características do item e não nas avaliações de usuários. A afirmação IV diz que os métodos de filtragem colaborativa podem ser aplicados apenas para medir a semelhança entre usuários, porém podemos aplicar o mesmo princípio para as avaliações entre itens (ou *Item-Item Collaborative Filtering*). Portanto tal afirmação está INCORRETA. As demais afirmações (I, III e V) estão corretas. |
|  | E | II e V |

Questão 6/10 - Big Data

Avalie as seguintes afirmações sobre os componentes do Hadoop: Impala e Accumulo:  
  
I – Impala é um motor de consultas SQL capaz de realizar consultas de baixa latência em HDFS ou Hive.

II – O Impapa foi projetado para consultas analíticas em Haddop utilizando SQL ou ferramentas de *business inteligence* (BI).

III – Accumulo é um sistema de armazenamento distribuído baseado em chave-valor que implementa segurança e nível de célula.

IV – Todos os dados armazenados no Accumulo devem possuir os mesmos requisitos de segurança em uma mesma tabela.

V – O Accumulo permite realizar operações em pares de chave-valor assim que são inseridos.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I, II e III  Você assinalou essa alternativa (A) |
|  | B | Todas |
|  | C | I, II e IV |
|  | D | II, III e V  Justificativa: Entre as principais características do Impala destaca-se a possibilidade de realizar consultas SQL de baixa latência de dados armazenados em HDFS e **HBase**. Dessa forma a afirmação I encontra-se INCORRETA. Uma das principais características do Accumulo é a segurança em nível de célula, onde cada par chave-valor possui o seu próprio rótulo de segurança que tem a capacidade de limitar os resultados de uma consulta baseado nas autorizações de acesso do usuário. Portanto a afirmação IV está INCORRETA. As afirmações II, III e V estão CORRETAS. |
|  | E | Apenas III |

Questão 7/10 - Big Data

O aprendizado de máquina é parte do campo da Inteligência Artificial que se refere ao estudo de modelos estatísticos para resolver problemas específicos com padrões e inferências.  
  
Sobre as diferentes categorias de aprendizado de máquina avalie as seguintes afirmações:  
  
I – Os algoritmos de classificação e de regressão fazem parte da categoria de algoritmos de aprendizado supervisionado.

II – Os algoritmos de aprendizado supervisionado buscam identificar estruturas inerentes ao conjunto de dados.

III – O que diferencia os algoritmos de aprendizado por reforço dos algoritmos de aprendizado não-supervisionado é que no aprendizado por reforço os dados nunca são explicitamente relacionados com os seus rótulos.

IV – Os algoritmos de aprendizado por reforço se baseiam nos conceitos de reforço e punição para maximizar parâmetros através da tomada de ações.

V – Os algoritmos de clustering fazem parte da categoria de aprendizagem não-supervisionados.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I, II, e III |
|  | B | Todas  Você assinalou essa alternativa (B) |
|  | C | I, III e IV |
|  | D | I, IV e V  Justificativa: Os algoritmos de aprendizado supervisionado utilizam conjuntos de dados que contém ambos dados de entrada e os resultados esperados, ou seja, os dados de entrada são rotulados. Os algoritmos que buscam identificar estruturas inerentes ao conjunto de dados são os de aprendizado não-supervisionado. Portanto a afirmação II está INCORRETA. Os algoritmos de aprendizado não-supervisionado também não relacionam explicitamente os dados aos rótulos. A afirmação III INCORRETAMENTE confunde os algoritmos de aprendizado supervisionado com os algoritmos de aprendizado não-supervisionado. As demais afirmações (I, IV e V) estão CORRETAS. |
|  | E | Todas, exceto III |

Questão 8/10 - Big Data

GraphX é o componente do Spark para computação de grafos em sistemas distribuídos de larga escala. Ele foi desenvolvido através de um projeto de pesquisa como forma de unificar o processamento de grafos e de sistemas paralelos até se tornar uma parte integral do projeto Spark.   
  
Avalie as seguintes afirmações a respeito do GraphX:  
  
I – A API GraphX implementa abstrações de arestas e vértices baseados em extensões de RDDs

II – A computação de grafos é importante em casos onde as relações entre os dados são mais importantes que os dados em si.

III – Grafos são uma forma muito eficiente de representar dados de tabelas esparsas.

IV – O algoritmo *Connected Components* é um tipo de algoritmo paralelo de grafos que é capaz de medir a coesão entre os seus dados (componentes).

V – O algoritmo *PageRank* desenvolvido pelo Google é capaz de classificar a importância de sites através da quantidade de referências apontadas para um site.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | Todas, exceto IV  Justificativa: O algoritmo Connected Components busca encontrar todos os componentes de um grafo que estão conectados entre si onde cada vértice (ou componente) dentro de um grupo pode ser alcançado a partir de qualquer outro vértice do grupo. Além disso não deve haver nenhum caminho, ou seja, arestas, entre dois grupos distintos. Diferente do que alega INCORRETAMENTE a afirmação IV. Todas as outras afirmações (I, II, III e V) estão CORRETAS. |
|  | B | I, II e III |
|  | C | II, III e IV |
|  | D | IV e V  Você assinalou essa alternativa (D) |
|  | E | Todas |

Questão 9/10 - Big Data

A API GraphX oferece uma forma de armazenar grafos na forma de tabelas e utilizar operações de tabelas para expressar operações de grafos.  
  
A respeito da API do GraphX avalie as seguintes afirmações:  
  
I – Um grafo de propriedades em GraphX é uma abstração capaz de definir um multigrafo dirigido com objetos definidos para cada vértice e aresta.

II – Um grafo de propriedade pode ser definido como um par de conjuntos de RDDs que codificam as características de vértices e arestas.

III – Grafos de propriedades são distribuídos, tolerantes a falhas e seus valores e estruturas podem ser modificados facilmente.

IV – O construtor da classe Edge recebe os parâmetros que correspondem aos identificadores dos vértices de origem e destino.

V – O membro *vertices* da classe Graph retorna um *VertexRDD[(String,String)]* que é uma extensão do objeto *RDD[(VertexId,(String,String)]*. Por outro lado, o membro *edge* retorna um *EdgeRDD* que contém um objeto *Edge[String]*.

Dentre essas afirmações estão INCORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I, II e III |
|  | B | I, III e IV  Você assinalou essa alternativa (B) |
|  | C | Apenas III  Justificativa: Diferentemente do que diz INCORRETAMENTE a afirmação III, grafos de propriedade não podem ser modificados, pois da mesma forma que RDDs, grafos de propriedades são imutáveis, distribuídos e tolerantes a falhas. Qualquer mudança em valores ou estrutura do grafo produz um novo grafo. Todas as outras afirmações (I, II, IV e V) estão CORRETAS. |
|  | D | Todas |
|  | E | II, IV e V |

Questão 10/10 - Big Data

O funcionamento estrutural de um fluxo de aprendizado de máquina pode ser descrito através das seguintes etapas: Utilizamos um conjunto de dados de teste e de treino para escolher os modelos de aprendizado de máquina mais adequados, treiná-los, e avaliá-los com o objetivo de alcançar um modelo que seja capaz de realizar predições úteis a partir de dados do mundo real.

Avalie as seguintes afirmações a respeito do fluxo de aprendizado de máquina:

I – Na etapa de obtenção de dados são adquiridos os dados que serão utilizados tanto para treinamento dos modelos quanto para avaliar a eficiência dos modelos.

II – Os problemas de dados faltantes, dados com ruído ou dados inconsistentes devem ser resolvidos na etapa de obtenção dos dados.

III – Os dados devem ser formatados da maneira mais adequada durante a etapa de preparação dos dados.

IV – A etapa de treinamento é quando utilizamos os dados de treino na expectativa de que o modelo aprenda a processar os dados da forma esperada.

V – Durante a etapa de avaliação o desempenho do modelo é avaliado e caso não seja considerado suficiente, o modelo passa por um novo ciclo de ajuste de parâmetros, treinamento e avaliação.

Dentre essas afirmações estão INCORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I, IV e V |
|  | B | Apenas II  Justificativa: Diferentemente do que diz INCORRETAMENTE a afirmação II, os problemas de dados faltantes, dados com ruído ou dados inconsistentes devem ser resolvidos na etapa de Preparação dos Dados. As afirmações I, III, IV e V estão CORRETAS. |
|  | C | II e III |
|  | D | Todas, exceto IV |
|  | E | Nenhuma  Você assinalou essa alternativa (E) |

Questão 1/10 - Big Data

A respeito do ecossistema Hadoop avalie as seguintes afirmações:  
  
I – Pig é uma API de alto nível para realizar consultas SQL. Dessa forma é possível processar diretamente os dados contidos no HDFS.

II – Spark é um componente capaz de substituir o MapReduce para realizar o processamento distribuído no ecossistema Hadoop.

III – Grafos Acíclicos Dirigidos são uma estratégia de otimização de processamento utilizada por alguns componentes como o Tez

IV – Sqoop, Flume e Kafka são aplicações capazes de obter dados de fontes externas ao Hadoop para o HDFS

V – Zookeeper é um componente capaz de coordenar as aplicações do cluster. Dessa forma, ele é amplamente utilizado por diversos outros componentes para melhorar a sua confiabilidade e desempenho.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I e III |
|  | B | Todas, exceto I |
|  | C | I  Você assinalou essa alternativa (C)  **Você acertou!**  Justificativa: O Pig não processa diretamente os dados do HDFS, mas utiliza operações Map Reduce para tal. Dessa forma a afirmação I está INCORRETA. As demais afirmações estão corretas. |
|  | D | II, IV e V |
|  | E | Nenhuma |

Questão 2/10 - Big Data

O Hadoop é um framework amplamente utilizado no desenvolvimento de sistemas de processamento de Big Data. A sua grande contribuição foi aliar um sistema de arquivos distribuído a um modelo de programação distribuída. Com isso, avalie as seguintes afirmações:

I – O Hadoop Distributed File System é o sistema do Hadoop responsável por gerenciar os arquivos armazenados. Para garantir que não ocorra perda de dados no caso de falhas, tal sistema realiza o backup dos arquivos de cada um dos servidores do cluster em um servidor dedicado para backup.

II – A natureza distribuída do Hadoop se relaciona diretamente com a escalabilidade de sua capacidade de processamento e com a capacidade de garantir seu funcionamento mesmo que alguns dos componentes de sua infraestrutura falhem.

III – O Hadoop Map Reduce é o sistema responsável por gerenciar o processamento de forma distribuída. Tal sistema é capaz de dividir a execução da aplicação em pequenas tarefas que são processadas pelos servidores do cluster.

IV – As etapas de processamento do Map Reduce são: Map, Shuffle e Reduce.

V – Devido às suas características distribuídas o Haddop é capaz de realizar o balanceamento de carga entre os servidores do cluster, o que permite que todos os recursos operem de forma balanceada.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | II, III e V |
|  | B | Todas |
|  | C | I e IV |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: As afirmações II, III, IV e V corretamente informam as características do framework Hadoop. Porém a afirmação I erra ao afirmar que o Hadoop possui um servidor dedicado ao backup. No Hadoop os dados são armazenados em pequenos blocos que são replicados por todo o cluster, garantindo assim que os dados não sejam perdidos em caso de falhas no cluster. |
|  | E | III e V |

Questão 3/10 - Big Data

Uma estratégia de coleta e armazenamento de dados é essencial quando nos referimos a soluções de Big Data. Nesse sentido, temos que estar atentos à origem dos dados.

Dessa forma considere as seguintes afirmações:

I – Os dados em posse do proprietário do projeto são considerados dados internos.

II – Os maiores desafios a respeito da coleta de dados de sensores é a dificuldade de comunicação entre os equipamentos.

III – Não são considerados dados externos os dados obtidos de mídias sociais através de protocolo REST.

IV – O monitoramento em tempo real de ações e análises preditivas de dados quantificados são utilizados em processos de dataficação

V – O registro de log de diversos tipos são normalmente considerados dados internos.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | II, III e IV |
|  | B | Todas exceto III |
|  | C | I, IV e V  Justificativa: As afirmações CORRETAS (I, IV e V) se referem às estratégias de obtenção de dados. Porém a afirmação II (INCORRETA) alega incorretamente que a dificuldade de comunicação imponha alguma dificuldade na obtenção de dados de sensores. E a afirmação III (INCORRETA) exclui erroneamente os dados vindos de mídias sociais através de protocolo REST, uma vez que estes são considerados como dados externos, pois são obtidos de uma fonte de outra companhia. |
|  | D | II e III  Você assinalou essa alternativa (D) |
|  | E | I e V |

Questão 4/10 - Big Data

Muitas aplicações necessitam que os dados sejam processados à medida que são recebidos pela aplicação e que o resultado de tal processamento esteja disponível tão rapidamente quanto. A esse tipo de operação damos o nome de processamento de fluxo em tempo real.  
  
A cerca de tais aplicações, avalie as seguintes afirmações:  
  
I – O conceito de baixa latência define que os dados devem ser processados em um tempo igual ou menor que o tempo em que novos dados chegam ao fluxo.

II – Para ser considerado consistente a solução não pode tolerar imperfeições e inconsistências

III – Dados significativos podem ser perdidos caso a aplicação tenha problemas na coleta, transmissão ou processamento dos dados.

IV – Spark é uma aplicação que utiliza o Map Reduce para processar dados em tempo real.

V – Spark possui seu próprio sistema de arquivos, o SparkSQL que implementa um banco de dados otimizado para realizar consultas em tempo real.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I e II |
|  | B | Todas, exceto III |
|  | C | IV e V |
|  | D | Todas |
|  | E | II, IV e V  Você assinalou essa alternativa (E)  **Você acertou!**  Justificativa: As afirmações I e III informam corretamente a respeito dos conceitos de baixa latência e alta disponibilidade respectivamente. Na afirmação II o conceito de consistência está invertido, é exatamente o contrário. A solução deve tolerar imperfeições e inconsistências para ser considerada consistente. Enquanto que as afirmações IV e V erram em definir o framework de processamento distribuído do Spark (IV) e o armazenamento de dados no Spark (V). No caso, o Spark implementa um framework próprio de processamento distribuído de dados, sua principal vantagem é, justamente, ser mais eficiente que o próprio Hadoop Map Reduce. E a respeito do SparkSQL é o framework de acesso aos dados armazenados em qualquer sistema de arquivos suportado pelo Spark (inclusive o HDFS) utilizando a linguagem SQL. |

Questão 5/10 - Big Data

Avalie as seguintes afirmações a respeito das operações definidas pelo MapReduce:

I – A função Mapper tem por objetivo mapear conjuntos de valores necessários pela aplicação a determinadas chaves que são geradas automaticamente pelo MapReduce.

II – A função Mapper tem por objetivo extrair e organizar os dados necessários para o processamento.

III – A função Reducer recebe o conjunto de dados no formato chave-valor que foi definido pela função Mapper.

IV – Na função Reducer, cada par chave-valor constitui uma tarefa a ser processada pelo servidor que estiver mais próximo dos dados.

V – A função Shuffle and sort transforma um conjunto de dados em formato chave-valor onde os valores devem ser embaralhados ou ordenados de acordo com a sua chave.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | II e IV  Justificativa: Durante a operação Mapper, as chaves são definidas pelo programador e, não são geradas pelo MapReduce como INCORRETAMENTE diz a afirmação I. Antes da operação Reduce, ocorre a operação Shuffle And Sort que reorganiza os dados que foram mapeados. Portanto o conjunto de dados recebidos pela operação Reduce NÃO é o mesmo definido pela Mapper, como informa INCORRETAMENTE a afirmação III. O que ocorre na operação Shuffle and Sort é que as chaves dos pares chave-valor que foram geradas pelo Mapper são verificadas e os valores dos pares que possuem a mesma chave são agregados em uma única list. Diferente do comportamento apontado INCORRETAMENTE pela afirmação V. As demais afirmações (II e IV) estão CORRETAS. |
|  | B | II, III e IV  Você assinalou essa alternativa (B) |
|  | C | I, IV e V |
|  | D | Todas, exceto I |
|  | E | I, III e IV |

Questão 6/10 - Big Data

O MongoDB é um banco de dados NoSQL distribuído e multiplataforma desenvolvido pela MongoDB Inc.  
  
Avalie as seguintes afirmações a respeito do MongoDB:  
  
I – É um banco de dados. Eventualmente consistente, ou seja, que prioriza a disponibilidade e a tolerância a falhas no teorema CAP.

II – O MongoDB é um banco de dados orientado a documentos com uma estrutura muito semelhante ao formato JSON.

III – Dados embutidos, ou desnormalizados são aqueles que armazenam pedaços de informação relacionada em um mesmo documento

IV – Dados referenciados ou normalizados são aqueles se fazem referência entre si e são úteis para evitar duplicações de dados desnecessárias.

V – O conjunto de réplicas do MongoDB melhoram a consistência do banco.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Nenhuma |
|  | B | I e V  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: O MongoDB é um banco que, em relação ao teorema CAP, privilegia a consistência dos dados e a tolerância a falhas. E, portanto, é um banco que pode estar eventualmente indisponível. Com isso, a afirmação I encontra-se INCORRETA. Os conjuntos de replicação oferecem redundância e melhoram a disponibilidade do banco. Dessa forma, a afirmação V está INCORRETA. |
|  | C | III e IV |
|  | D | Apenas II |
|  | E | II, IV e V |

Questão 7/10 - Big Data

Avalie as seguintes afirmações sobre as operações de leitura e escrita no HDFS:

I – Durante a operação de leitura no HDFS, o Name Node informa ao cliente quais são os Data Nodes mais eficientes que possuem os blocos de dados desejados.

II – O cliente deve requisitar a operação de escrita ao Name Node que informa ao cliente quais são os Name Nodes mais eficientes. Em seguida o cliente deve comunicar com esses Name Nodes para armazenar os dados do arquivo dividido em blocos de 128MB.

III – A operação de escrita de dados deve ser realizada pelo servidor cliente que distribui o arquivo em blocos aos Name Nodes mais eficientes do cluster.

IV –  A operação de escrita de dados deve ser realizada pelo servidor cliente que requisita a criação de um arquivo ao Name Node. Em seguida o cliente comunica com um único Data Node que irá distribuir os dados entre os outros Data Nodes. Por fim o Data Node retorna à localização dos blocos de dados ao cliente. E o cliente repassa a informação para ser armazenada no Name Node

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | III e IV |
|  | B | I e IV |
|  | C | I e III |
|  | D | II e III  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: A operação de escrita de dados no HDFS começa com o client node requisitando a criação do arquivo ao Name Node. E seguida o Client Node comunica com um único Data Node que redistribui blocos do arquivo para outros Data Nodes. A localização dos blocos é retornada para o Client Node que informa ao Name Node e, com isso, conclui a escrita. Portando, o Name Node não informa nenhum Data Node durante a escrita, como INCORRETAMENTE dizem as afirmações II e III. As demais afirmações (I e IV) estão CORRETAS. |
|  | E | II e IV |

Questão 8/10 - Big Data

“Apenas recentemente a capacidade de armazenamento e processamento se tornaram suficientes para permitir que dados antes ignorados fossem analisados. Entretanto, além dos componentes tecnológicos, o analista de dados deve ser capaz de identificar quais dados se deve utilizar, como integrá-los, quais as perguntas que serão úteis para a tomada de decisão, e qual a melhor maneira de apresentar os resultados obtidos da análise. ”  
  
Sobre a fase de análise de dados avalie as seguintes afirmações:  
  
I – Durante a fase de análise de dados, o cientista de dados deve preparar os dados brutos para que possam ser utilizados na produção de informações úteis e padrões ocultos em conjuntos massivos de dados.

II –  A partir do conjunto de dados inicial podemos aplicar análises estatísticas utilizando modelos estatísticos ou ainda aplicar modelos de predições baseados em aprendizado de máquina de forma que os dados sejam refinados.

III – Realizar análise exploratória e verificar a qualidade dos dados são atividades da análise de dados que fazem parte da fase de preparação dos dados

IV –  A fase de modelagem dos dados é composta pelas seguintes tarefas: Selecionar técnicas de modelagem, projetar testes, definir e construir o modelo de dados, seus parâmetros e sua descrição, validar o modelo e definir os parâmetros a serem revisados, avaliar os resultados do modelo; revisar processos; e determinar os passos seguintes.

V – O aprendizado de máquina é uma das atividades necessárias na automatização da construção de modelos analíticos que é uma das tarefas da fase de compreensão dos dados.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I e II  Justificativa: As afirmações I e II (CORRETAS) acertam sobre o papel do cientista de dados e a aplicação de modelos estatísticos sobre os dados que estão sendo recebidos. A afirmação III indica um conjunto de atividades que fazem parte da compreensão dos dados e não da preparação dos dados. A afirmação IV confunde e mistura tarefas da modelagem de dados e da avaliação do modelo. A Automatização da construção de modelos analíticos NÃO é uma das tarefas da compreensão dos dados. |
|  | B | Todas |
|  | C | IV e V |
|  | D | Todas, exceto I |
|  | E | II, IV e V  Você assinalou essa alternativa (E) |

Questão 9/10 - Big Data

“Zookeeper é um serviço open source distribuído de coordenação para aplicações distribuídas. Ele é capaz de manter dados de forma consistente por todo o cluster”. A respeito do Zookeeper avalie as seguintes afirmações:

I – O Zookeeper fornece um conjunto de primitivas que permitem a implementação de serviços de sincronização, manutenção de configuração, gerenciamento de grupos e nomeação.

II – Em caso de falha do nó mestre, o próprio sistema elege um novo nó mestre que assume imediatamente as informações de execução do nó que falhou. Em caso de falha de um dos nós trabalhadores, o Zookeeper redistribui o trabalho para outro nó adequadamente.

III – Znodes são arquivos permanentes que visam manter, de forma consistente, as informações a respeito de nós e tarefas de uma aplicação.

IV – Quando um nó deixa de atualizar o seu registro, ele deixa de existir no sistema de arquivos, tornando-se indisponível para o Zookeeper.

V – Cada nó da aplicação possui um cliente Zookeeper que monitora as informações que devem ser registradas. Os clientes Zookeeper nos nós são supervisionados por um conjunto de servidores que replicam seus dados entre si para garantir que seja possível continuar operando normalmente mesmo em caso de falha.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Todas |
|  | B | I, IV e V  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: Em caso de falha de um dos nós trabalhadores, o Zookeeper pede para a aplicação redistribuir o trabalho para outro nó adequadamente. O Zookeeper não redistribui diretamente o trabalho, ou seja, a afirmação II encontra-se INCORRETA. Os Znodes podem ser efêmeros, como os que armazenam o estado de um nó ou permanentes, como os que armazenam as tarefas dos nós trabalhadores. Portanto a afirmação III é INCORRETA. As demais afirmações estão CORRETAS. |
|  | C | IV e V |
|  | D | Todas exceto II |
|  | E | II, III e IV |

Questão 10/10 - Big Data

Os bancos de dados NoSQL podem ser definidos em alguns modelos de acordo com a forma que armazenam seus dados. A respeito dos modelos de bancos de dados NoSQL existentes avalie as seguintes afirmações:  
  
I – Os bancos de dados de esquema chave-valor possuem chaves com um modelo discretamente ordenado, dessa forma os dados são mantidos em ordenação lexicográfica

II – Os documentos dos bancos de dados baseados em documentos são estruturas de dados que devem possuir a característica de encapsular e codificar os dados

III – Os bancos de dados baseados em documentos permitem a buscar documentos baseado em seus conteúdos

IV – Os bancos de dados baseados em colunas são muito utilizados para a representação de dados esparsos

V – Os bancos de dados baseados em grafos utilizam um tipo de armazenamento de dados específico chamado de adjacência livre de índices

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | Todas  Você assinalou essa alternativa (A) |
|  | B | II, IV e V |
|  | C | Todas, exceto V  Justificativa: A adjacência livre de índices é um modelo de processamento adotado por bancos de dados orientados a grafos. Alguns bancos de dados baseados em grafos utilizam um tipo de armazenamento especificamente projetado para estruturas de grafos. Porém podem utilizar bancos relacionais, baseados em documentos ou colunas, como camada de armazenamento dentro de uma estrutura lógica de grafos. Dessa forma a afirmação V encontra-se INCORRETA. As demais afirmações (I, II, III e IV) estão CORRETAS. |
|  | D | I, II e IV |
|  | E | III e V |

Questão 1/10 - Big Data

“Apenas recentemente a capacidade de armazenamento e processamento se tornaram suficientes para permitir que dados antes ignorados fossem analisados. Entretanto, além dos componentes tecnológicos, o analista de dados deve ser capaz de identificar quais dados se deve utilizar, como integrá-los, quais as perguntas que serão úteis para a tomada de decisão, e qual a melhor maneira de apresentar os resultados obtidos da análise. ”  
  
Sobre a fase de análise de dados avalie as seguintes afirmações:  
  
I – Durante a fase de análise de dados, o cientista de dados deve preparar os dados brutos para que possam ser utilizados na produção de informações úteis e padrões ocultos em conjuntos massivos de dados.

II –  A partir do conjunto de dados inicial podemos aplicar análises estatísticas utilizando modelos estatísticos ou ainda aplicar modelos de predições baseados em aprendizado de máquina de forma que os dados sejam refinados.

III – Realizar análise exploratória e verificar a qualidade dos dados são atividades da análise de dados que fazem parte da fase de preparação dos dados

IV –  A fase de modelagem dos dados é composta pelas seguintes tarefas: Selecionar técnicas de modelagem, projetar testes, definir e construir o modelo de dados, seus parâmetros e sua descrição, validar o modelo e definir os parâmetros a serem revisados, avaliar os resultados do modelo; revisar processos; e determinar os passos seguintes.

V – O aprendizado de máquina é uma das atividades necessárias na automatização da construção de modelos analíticos que é uma das tarefas da fase de compreensão dos dados.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I e II  Justificativa: As afirmações I e II (CORRETAS) acertam sobre o papel do cientista de dados e a aplicação de modelos estatísticos sobre os dados que estão sendo recebidos. A afirmação III indica um conjunto de atividades que fazem parte da compreensão dos dados e não da preparação dos dados. A afirmação IV confunde e mistura tarefas da modelagem de dados e da avaliação do modelo. A Automatização da construção de modelos analíticos NÃO é uma das tarefas da compreensão dos dados. |
|  | B | Todas |
|  | C | IV e V |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D) |
|  | E | II, IV e V |

Questão 2/10 - Big Data

Avalie as seguintes afirmações sobre as operações de leitura e escrita no HDFS:

I – Durante a operação de leitura no HDFS, o Name Node informa ao cliente quais são os Data Nodes mais eficientes que possuem os blocos de dados desejados.

II – O cliente deve requisitar a operação de escrita ao Name Node que informa ao cliente quais são os Name Nodes mais eficientes. Em seguida o cliente deve comunicar com esses Name Nodes para armazenar os dados do arquivo dividido em blocos de 128MB.

III – A operação de escrita de dados deve ser realizada pelo servidor cliente que distribui o arquivo em blocos aos Name Nodes mais eficientes do cluster.

IV –  A operação de escrita de dados deve ser realizada pelo servidor cliente que requisita a criação de um arquivo ao Name Node. Em seguida o cliente comunica com um único Data Node que irá distribuir os dados entre os outros Data Nodes. Por fim o Data Node retorna à localização dos blocos de dados ao cliente. E o cliente repassa a informação para ser armazenada no Name Node

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | III e IV |
|  | B | I e IV |
|  | C | I e III |
|  | D | II e III  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: A operação de escrita de dados no HDFS começa com o client node requisitando a criação do arquivo ao Name Node. E seguida o Client Node comunica com um único Data Node que redistribui blocos do arquivo para outros Data Nodes. A localização dos blocos é retornada para o Client Node que informa ao Name Node e, com isso, conclui a escrita. Portando, o Name Node não informa nenhum Data Node durante a escrita, como INCORRETAMENTE dizem as afirmações II e III. As demais afirmações (I e IV) estão CORRETAS. |
|  | E | II e IV |

Questão 3/10 - Big Data

O Hadoop é um framework amplamente utilizado no desenvolvimento de sistemas de processamento de Big Data. A sua grande contribuição foi aliar um sistema de arquivos distribuído a um modelo de programação distribuída. Com isso, avalie as seguintes afirmações:

I – O Hadoop Distributed File System é o sistema do Hadoop responsável por gerenciar os arquivos armazenados. Para garantir que não ocorra perda de dados no caso de falhas, tal sistema realiza o backup dos arquivos de cada um dos servidores do cluster em um servidor dedicado para backup.

II – A natureza distribuída do Hadoop se relaciona diretamente com a escalabilidade de sua capacidade de processamento e com a capacidade de garantir seu funcionamento mesmo que alguns dos componentes de sua infraestrutura falhem.

III – O Hadoop Map Reduce é o sistema responsável por gerenciar o processamento de forma distribuída. Tal sistema é capaz de dividir a execução da aplicação em pequenas tarefas que são processadas pelos servidores do cluster.

IV – As etapas de processamento do Map Reduce são: Map, Shuffle e Reduce.

V – Devido às suas características distribuídas o Haddop é capaz de realizar o balanceamento de carga entre os servidores do cluster, o que permite que todos os recursos operem de forma balanceada.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | II, III e V |
|  | B | Todas |
|  | C | I e IV |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: As afirmações II, III, IV e V corretamente informam as características do framework Hadoop. Porém a afirmação I erra ao afirmar que o Hadoop possui um servidor dedicado ao backup. No Hadoop os dados são armazenados em pequenos blocos que são replicados por todo o cluster, garantindo assim que os dados não sejam perdidos em caso de falhas no cluster. |
|  | E | III e V |

Questão 4/10 - Big Data

A respeito do ecossistema Hadoop avalie as seguintes afirmações:  
  
I – Pig é uma API de alto nível para realizar consultas SQL. Dessa forma é possível processar diretamente os dados contidos no HDFS.

II – Spark é um componente capaz de substituir o MapReduce para realizar o processamento distribuído no ecossistema Hadoop.

III – Grafos Acíclicos Dirigidos são uma estratégia de otimização de processamento utilizada por alguns componentes como o Tez

IV – Sqoop, Flume e Kafka são aplicações capazes de obter dados de fontes externas ao Hadoop para o HDFS

V – Zookeeper é um componente capaz de coordenar as aplicações do cluster. Dessa forma, ele é amplamente utilizado por diversos outros componentes para melhorar a sua confiabilidade e desempenho.

Dentre essas afirmações estão INCORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I e III  Você assinalou essa alternativa (A) |
|  | B | Todas, exceto I |
|  | C | I  Justificativa: O Pig não processa diretamente os dados do HDFS, mas utiliza operações Map Reduce para tal. Dessa forma a afirmação I está INCORRETA. As demais afirmações estão corretas. |
|  | D | II, IV e V |
|  | E | Nenhuma |

Questão 5/10 - Big Data

Muitas aplicações necessitam que os dados sejam processados à medida que são recebidos pela aplicação e que o resultado de tal processamento esteja disponível tão rapidamente quanto. A esse tipo de operação damos o nome de processamento de fluxo em tempo real.  
  
A cerca de tais aplicações, avalie as seguintes afirmações:  
  
I – O conceito de baixa latência define que os dados devem ser processados em um tempo igual ou menor que o tempo em que novos dados chegam ao fluxo.

II – Para ser considerado consistente a solução não pode tolerar imperfeições e inconsistências

III – Dados significativos podem ser perdidos caso a aplicação tenha problemas na coleta, transmissão ou processamento dos dados.

IV – Spark é uma aplicação que utiliza o Map Reduce para processar dados em tempo real.

V – Spark possui seu próprio sistema de arquivos, o SparkSQL que implementa um banco de dados otimizado para realizar consultas em tempo real.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I e II |
|  | B | Todas, exceto III |
|  | C | IV e V |
|  | D | Todas |
|  | E | II, IV e V  Você assinalou essa alternativa (E)  **Você acertou!**  Justificativa: As afirmações I e III informam corretamente a respeito dos conceitos de baixa latência e alta disponibilidade respectivamente. Na afirmação II o conceito de consistência está invertido, é exatamente o contrário. A solução deve tolerar imperfeições e inconsistências para ser considerada consistente. Enquanto que as afirmações IV e V erram em definir o framework de processamento distribuído do Spark (IV) e o armazenamento de dados no Spark (V). No caso, o Spark implementa um framework próprio de processamento distribuído de dados, sua principal vantagem é, justamente, ser mais eficiente que o próprio Hadoop Map Reduce. E a respeito do SparkSQL é o framework de acesso aos dados armazenados em qualquer sistema de arquivos suportado pelo Spark (inclusive o HDFS) utilizando a linguagem SQL. |

Questão 6/10 - Big Data

Avalie as seguintes afirmações a respeito das operações definidas pelo MapReduce:

I – A função Mapper tem por objetivo mapear conjuntos de valores necessários pela aplicação a determinadas chaves que são geradas automaticamente pelo MapReduce.

II – A função Mapper tem por objetivo extrair e organizar os dados necessários para o processamento.

III – A função Reducer recebe o conjunto de dados no formato chave-valor que foi definido pela função Mapper.

IV – Na função Reducer, cada par chave-valor constitui uma tarefa a ser processada pelo servidor que estiver mais próximo dos dados.

V – A função Shuffle and sort transforma um conjunto de dados em formato chave-valor onde os valores devem ser embaralhados ou ordenados de acordo com a sua chave.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | II e IV  Justificativa: Durante a operação Mapper, as chaves são definidas pelo programador e, não são geradas pelo MapReduce como INCORRETAMENTE diz a afirmação I. Antes da operação Reduce, ocorre a operação Shuffle And Sort que reorganiza os dados que foram mapeados. Portanto o conjunto de dados recebidos pela operação Reduce NÃO é o mesmo definido pela Mapper, como informa INCORRETAMENTE a afirmação III. O que ocorre na operação Shuffle and Sort é que as chaves dos pares chave-valor que foram geradas pelo Mapper são verificadas e os valores dos pares que possuem a mesma chave são agregados em uma única list. Diferente do comportamento apontado INCORRETAMENTE pela afirmação V. As demais afirmações (II e IV) estão CORRETAS. |
|  | B | II, III e IV |
|  | C | I, IV e V |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D) |
|  | E | I, III e IV |

Questão 7/10 - Big Data

O Oozie é o componente do Hadoop que realiza a orquestração de trabalhos. Avalie as seguintes afirmações sobre o Oozie:

I – O Ozzie utiliza de grafos cíclicos direcionados para implementar o encadeamento de tarefas de modo que tarefas que não possuem relação de dependência podem executar paralelamente.

II – Os fluxos de trabalho são especificados em notação YAML.

III – O Oozie permite a que sejam definidos coordenadores onde se pode configurar controles de execução, horário de início e frequência.

IV – O Oozie pode ser executado apenas através de gerenciadores de específicos em algumas interfaces.

V – Os fluxos de trabalho representam grafos acíclicos direcionados de tarefas capazes de executar de forma encadeada e automática.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, II e IV  Você assinalou essa alternativa (A)  **Você acertou!**  Justificativa: Os fluxos de trabalho do Oozie são grafos acíclicos direcionados que encadeiam diversas tarefas. Por isso a afirmação I encontra-se INCORRETA. Os fluxos de trabalho são escritos em XML e não em YAML como alega a afirmação II. Além de poder ser executado através de gerenciadores em algumas interfaces, também pode ser executado através da linha de comando, portanto a afirmação IV está INCORRETA. As demais afirmações (III e V) sobre o Oozie estão CORRETAS. |
|  | B | I e II |
|  | C | III e V |
|  | D | Todas, exceto II |
|  | E | II, IV e V |

Questão 8/10 - Big Data

“Zookeeper é um serviço open source distribuído de coordenação para aplicações distribuídas. Ele é capaz de manter dados de forma consistente por todo o cluster”. A respeito do Zookeeper avalie as seguintes afirmações:

I – O Zookeeper fornece um conjunto de primitivas que permitem a implementação de serviços de sincronização, manutenção de configuração, gerenciamento de grupos e nomeação.

II – Em caso de falha do nó mestre, o próprio sistema elege um novo nó mestre que assume imediatamente as informações de execução do nó que falhou. Em caso de falha de um dos nós trabalhadores, o Zookeeper redistribui o trabalho para outro nó adequadamente.

III – Znodes são arquivos permanentes que visam manter, de forma consistente, as informações a respeito de nós e tarefas de uma aplicação.

IV – Quando um nó deixa de atualizar o seu registro, ele deixa de existir no sistema de arquivos, tornando-se indisponível para o Zookeeper.

V – Cada nó da aplicação possui um cliente Zookeeper que monitora as informações que devem ser registradas. Os clientes Zookeeper nos nós são supervisionados por um conjunto de servidores que replicam seus dados entre si para garantir que seja possível continuar operando normalmente mesmo em caso de falha.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | Todas  Você assinalou essa alternativa (A) |
|  | B | I, IV e V  Justificativa: Em caso de falha de um dos nós trabalhadores, o Zookeeper pede para a aplicação redistribuir o trabalho para outro nó adequadamente. O Zookeeper não redistribui diretamente o trabalho, ou seja, a afirmação II encontra-se INCORRETA. Os Znodes podem ser efêmeros, como os que armazenam o estado de um nó ou permanentes, como os que armazenam as tarefas dos nós trabalhadores. Portanto a afirmação III é INCORRETA. As demais afirmações estão CORRETAS. |
|  | C | IV e V |
|  | D | Todas exceto II |
|  | E | II, III e IV |

Questão 9/10 - Big Data

O sistema de arquivos HDFS possui dois tipos de estruturas principais: Name Nodes e Data Nodes. Sobre a arquitetura do HDFS avalie as seguintes afirmações:  
  
I –  Name Node possui um registro de edição que contém a informação dos arquivos criados. Além disso, é o tipo de estrutura responsável por armazenar a localização dos blocos de cada um dos arquivos armazenados no HDFS. Dessa forma é possível rastrear quais os blocos de quais arquivos estão armazenados em cada Data Node.

II – Não é possível possuir mais de um Name Node no HDFS. Dessa forma se o Name Node se torna indisponível não há como obter os metadados contido nele.

III – Data Node é a estrutura que responsável por armazenar os blocos de dados. Cada Data Node se comunica apenas com o Name Node, de forma que todas as operações de escrita e leitura de arquivos é mediada pelo Name Node.

IV – Para melhorar a disponibilidade do HDFS é possível definir um Name Node secundário que funciona apenas como uma cópia do Name Node principal e entra em funcionamento quando o Name Node principal se torna indisponível. Outra possibilidade é a utilização do Zookeeper para auxiliar na recuperação de um Name Node inacessível.

V – Data Node é a estrutura responsável por armazenar os blocos de dados. Uma vez que o Cliente consulta o Name Node para buscar a localização dos blocos de dados que deseja consultar, deve se comunicar diretamente com o Data Node correspondente para realizar o acesso aos dados.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, III e IV |
|  | B | I, IV e V  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: É possível utilizar zero ou mais Name Nodes secundários que replicam os dados do Name Node em operação (principal). No caso de falha do Name Node principal um dos Name Nodes secundários podem substituí-lo. Portanto a afirmação II encontra-se INCORRETA. O Name Node armazena os registros de localização e edição dos blocos dos arquivos e comunica-se apenas com o Client Node que busca tais informações para acessar diretamente os Data Nodes para escritas e leituras de arquivos. Portanto a afirmação III encontra-se INCORRETA. As demais afirmações trazem informações CORRETAS sobre o HDFS. |
|  | C | I, II e V |
|  | D | Todas, exceto II |
|  | E | II, III e V |

Questão 10/10 - Big Data

“Bancos de Dados relacionais originalmente foram projetados baseados em uma técnica para o armazenamento de dados conhecida por *schema on write*. Essa estratégia define que as partes dos dados precisam se ajustar a um padrão ou um plano no momento da escrita (…) o Hive inverte a lógica e faz uso do conceito contrário: *scheme on read*. Ou seja, a estrutura dos dados, também conhecida por esquema, só é definida durante a leitura dos dados”. Sobre a arquitetura do Hive avalie as seguintes afirmações:

I – *Driver* é o componente do Hive que recebe os comandos e consultas, compila a entrada, otimiza a computação necessária e executa os passos com tarefas MapReduce ou Tez.

II – O Hive utiliza Mappers e Reducers genéricos que são sequenciados por planos de trabalho escritos em XML.

III – A estratégia *Scheme on Read* é considerada muito restritiva, uma vez que uma quantidade considerável de esforços é investida ajustando os dados à estrutura definida.

IV – Os metadados que definem os esquemas e particionamentos do Hive são armazenados em um banco de dados relacional. Normalmente Derby ou MySQL.

V – É possível realizar consultas utilizando o Hive através de uma interface de linha de comando (CLI) ou através da Hive Web Interface (HWI).

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Todas |
|  | B | I, II e III |
|  | C | Todas, exceto III  Você assinalou essa alternativa (C)  **Você acertou!**  Justificativa: A estratégia de armazenamento em banco de dados que é considerada restritiva é a Scheme on Write, uma vez que exige que o esquema do banco seja conhecido de antemão, ou seja, antes que os dados sejam escritos. Portanto a afirmação III é INCORRETA. As demais afirmações (I, II, IV e V) estão CORRETAS. |
|  | D | III e IV |
|  | E | II, IV e V |

Questão 1/10 - Big Data

Em relação à seguinte afirmação:

“O crescente volume de dados gerados a todo o momento indica com relativa clareza que este é, justamente, o atributo mais significativo quando falamos em Big Data.”

Podemos afirmar que:

I – O conceito de Big Data pode ser definido única e exclusivamente pelo volume de dados processados.

II – A relação entre dado volume de dados e a escalabilidade, eficiência, custo e complexidade de processá-los é o que define se o volume de dados necessita da utilização de soluções de Big Data.

III – É amplamente definido que para ser considerada uma solução de Big Data, tal implementação deve ser capaz de processar um volume de pelo menos 40 Gigabytes de dados.

IV – Dados de sensores devem ser necessariamente processados por soluções de Big Data devido ao seu volume que pode alcançar Terabytes.

V – As soluções de Big Data devem ser aplicadas nos casos em que a tecnologia tradicional não é capaz de processar um grande volume de dados com a eficiência, escalabilidade, custo e complexidade necessária.

Dentre essas afirmações estão CORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | I, II e III |
|  | B | Apenas II e V  Você assinalou essa alternativa (B) |
|  | C | Nenhuma |
|  | D | Apenas I |
|  | E | III, IV e V |

Questão 2/10 - Big Data

O Hadoop é um framework amplamente utilizado no desenvolvimento de sistemas de processamento de Big Data. A sua grande contribuição foi aliar um sistema de arquivos distribuído a um modelo de programação distribuída. Com isso, avalie as seguintes afirmações:

I – O Hadoop Distributed File System é o sistema do Hadoop responsável por gerenciar os arquivos armazenados. Para garantir que não ocorra perda de dados no caso de falhas, tal sistema realiza o backup dos arquivos de cada um dos servidores do cluster em um servidor dedicado para backup.

II – A natureza distribuída do Hadoop se relaciona diretamente com a escalabilidade de sua capacidade de processamento e com a capacidade de garantir seu funcionamento mesmo que alguns dos componentes de sua infraestrutura falhem.

III – O Hadoop Map Reduce é o sistema responsável por gerenciar o processamento de forma distribuída. Tal sistema é capaz de dividir a execução da aplicação em pequenas tarefas que são processadas pelos servidores do cluster.

IV – As etapas de processamento do Map Reduce são: Map, Shuffle e Reduce.

V – Devido às suas características distribuídas o Haddop é capaz de realizar o balanceamento de carga entre os servidores do cluster, o que permite que todos os recursos operem de forma balanceada.

Dentre essas afirmações estão CORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | II, III e V |
|  | B | Todas |
|  | C | I e IV |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D) |
|  | E | III e V |

Questão 3/10 - Big Data

Avalie as seguintes afirmações sobre as operações de leitura e escrita no HDFS:

I – Durante a operação de leitura no HDFS, o Name Node informa ao cliente quais são os Data Nodes mais eficientes que possuem os blocos de dados desejados.

II – O cliente deve requisitar a operação de escrita ao Name Node que informa ao cliente quais são os Name Nodes mais eficientes. Em seguida o cliente deve comunicar com esses Name Nodes para armazenar os dados do arquivo dividido em blocos de 128MB.

III – A operação de escrita de dados deve ser realizada pelo servidor cliente que distribui o arquivo em blocos aos Name Nodes mais eficientes do cluster.

IV –  A operação de escrita de dados deve ser realizada pelo servidor cliente que requisita a criação de um arquivo ao Name Node. Em seguida o cliente comunica com um único Data Node que irá distribuir os dados entre os outros Data Nodes. Por fim o Data Node retorna à localização dos blocos de dados ao cliente. E o cliente repassa a informação para ser armazenada no Name Node

Dentre essas afirmações estão INCORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | III e IV |
|  | B | I e IV |
|  | C | I e III |
|  | D | II e III  Você assinalou essa alternativa (D) |
|  | E | II e IV |

Questão 4/10 - Big Data

O MongoDB é um banco de dados NoSQL distribuído e multiplataforma desenvolvido pela MongoDB Inc.  
  
Avalie as seguintes afirmações a respeito do MongoDB:  
  
I – É um banco de dados. Eventualmente consistente, ou seja, que prioriza a disponibilidade e a tolerância a falhas no teorema CAP.

II – O MongoDB é um banco de dados orientado a documentos com uma estrutura muito semelhante ao formato JSON.

III – Dados embutidos, ou desnormalizados são aqueles que armazenam pedaços de informação relacionada em um mesmo documento

IV – Dados referenciados ou normalizados são aqueles se fazem referência entre si e são úteis para evitar duplicações de dados desnecessárias.

V – O conjunto de réplicas do MongoDB melhoram a consistência do banco.

Dentre essas afirmações estão INCORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | Nenhuma |
|  | B | I e V  Você assinalou essa alternativa (B) |
|  | C | III e IV |
|  | D | Apenas II |
|  | E | II, IV e V |

Questão 5/10 - Big Data

Ao criar uma aplicação capaz de processar e analisar seus dados de forma a extrair informações precisas que possam gerar o valor esperado o analista de dados deve estar atento aos seguintes fatores:

I – Quão variados eles precisam ser

II – Qual o seu nível de atualização

III – Qual o seu nível de popularidade

IV – A relevância dos dados

V – A velocidade necessária para processá-los

Estão corretas as afirmações:

|  |  |  |
| --- | --- | --- |
|  | A | Todas as alternativas |
|  | B | Apenas I e II |
|  | C | Todas, exceto I e III |
|  | D | Apenas II, III e V |
|  | E | Todas exceto III  Você assinalou essa alternativa (E) |

Questão 6/10 - Big Data

O Oozie é o componente do Hadoop que realiza a orquestração de trabalhos. Avalie as seguintes afirmações sobre o Oozie:

I – O Ozzie utiliza de grafos cíclicos direcionados para implementar o encadeamento de tarefas de modo que tarefas que não possuem relação de dependência podem executar paralelamente.

II – Os fluxos de trabalho são especificados em notação YAML.

III – O Oozie permite a que sejam definidos coordenadores onde se pode configurar controles de execução, horário de início e frequência.

IV – O Oozie pode ser executado apenas através de gerenciadores de específicos em algumas interfaces.

V – Os fluxos de trabalho representam grafos acíclicos direcionados de tarefas capazes de executar de forma encadeada e automática.

Dentre essas afirmações estão INCORRETAS:

Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I, II e IV |
|  | B | I e II |
|  | C | III e V |
|  | D | Todas, exceto II |
|  | E | II, IV e V  Você assinalou essa alternativa (E) |

Questão 7/10 - Big Data

Muitas aplicações necessitam que os dados sejam processados à medida que são recebidos pela aplicação e que o resultado de tal processamento esteja disponível tão rapidamente quanto. A esse tipo de operação damos o nome de processamento de fluxo em tempo real.  
  
A cerca de tais aplicações, avalie as seguintes afirmações:  
  
I – O conceito de baixa latência define que os dados devem ser processados em um tempo igual ou menor que o tempo em que novos dados chegam ao fluxo.

II – Para ser considerado consistente a solução não pode tolerar imperfeições e inconsistências

III – Dados significativos podem ser perdidos caso a aplicação tenha problemas na coleta, transmissão ou processamento dos dados.

IV – Spark é uma aplicação que utiliza o Map Reduce para processar dados em tempo real.

V – Spark possui seu próprio sistema de arquivos, o SparkSQL que implementa um banco de dados otimizado para realizar consultas em tempo real.

Dentre essas afirmações estão INCORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | I e II |
|  | B | Todas, exceto III |
|  | C | IV e V |
|  | D | Todas |
|  | E | II, IV e V  Você assinalou essa alternativa (E) |

Questão 8/10 - Big Data

“Zookeeper é um serviço open source distribuído de coordenação para aplicações distribuídas. Ele é capaz de manter dados de forma consistente por todo o cluster”. A respeito do Zookeeper avalie as seguintes afirmações:

I – O Zookeeper fornece um conjunto de primitivas que permitem a implementação de serviços de sincronização, manutenção de configuração, gerenciamento de grupos e nomeação.

II – Em caso de falha do nó mestre, o próprio sistema elege um novo nó mestre que assume imediatamente as informações de execução do nó que falhou. Em caso de falha de um dos nós trabalhadores, o Zookeeper redistribui o trabalho para outro nó adequadamente.

III – Znodes são arquivos permanentes que visam manter, de forma consistente, as informações a respeito de nós e tarefas de uma aplicação.

IV – Quando um nó deixa de atualizar o seu registro, ele deixa de existir no sistema de arquivos, tornando-se indisponível para o Zookeeper.

V – Cada nó da aplicação possui um cliente Zookeeper que monitora as informações que devem ser registradas. Os clientes Zookeeper nos nós são supervisionados por um conjunto de servidores que replicam seus dados entre si para garantir que seja possível continuar operando normalmente mesmo em caso de falha.

Dentre essas afirmações estão CORRETAS:

Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | Todas |
|  | B | I, IV e V |
|  | C | IV e V |
|  | D | Todas exceto II  Você assinalou essa alternativa (D) |
|  | E | II, III e IV |

Questão 9/10 - Big Data

“Apenas recentemente a capacidade de armazenamento e processamento se tornaram suficientes para permitir que dados antes ignorados fossem analisados. Entretanto, além dos componentes tecnológicos, o analista de dados deve ser capaz de identificar quais dados se deve utilizar, como integrá-los, quais as perguntas que serão úteis para a tomada de decisão, e qual a melhor maneira de apresentar os resultados obtidos da análise. ”  
  
Sobre a fase de análise de dados avalie as seguintes afirmações:  
  
I – Durante a fase de análise de dados, o cientista de dados deve preparar os dados brutos para que possam ser utilizados na produção de informações úteis e padrões ocultos em conjuntos massivos de dados.

II –  A partir do conjunto de dados inicial podemos aplicar análises estatísticas utilizando modelos estatísticos ou ainda aplicar modelos de predições baseados em aprendizado de máquina de forma que os dados sejam refinados.

III – Realizar análise exploratória e verificar a qualidade dos dados são atividades da análise de dados que fazem parte da fase de preparação dos dados

IV –  A fase de modelagem dos dados é composta pelas seguintes tarefas: Selecionar técnicas de modelagem, projetar testes, definir e construir o modelo de dados, seus parâmetros e sua descrição, validar o modelo e definir os parâmetros a serem revisados, avaliar os resultados do modelo; revisar processos; e determinar os passos seguintes.

V – O aprendizado de máquina é uma das atividades necessárias na automatização da construção de modelos analíticos que é uma das tarefas da fase de compreensão dos dados.

Dentre essas afirmações estão CORRETAS:

Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I e II |
|  | B | Todas  Você assinalou essa alternativa (B) |
|  | C | IV e V |
|  | D | Todas, exceto I |
|  | E | II, IV e V |

Questão 10/10 - Big Data

O sistema de arquivos HDFS possui dois tipos de estruturas principais: Name Nodes e Data Nodes. Sobre a arquitetura do HDFS avalie as seguintes afirmações:  
  
I –  Name Node possui um registro de edição que contém a informação dos arquivos criados. Além disso, é o tipo de estrutura responsável por armazenar a localização dos blocos de cada um dos arquivos armazenados no HDFS. Dessa forma é possível rastrear quais os blocos de quais arquivos estão armazenados em cada Data Node.

II – Não é possível possuir mais de um Name Node no HDFS. Dessa forma se o Name Node se torna indisponível não há como obter os metadados contido nele.

III – Data Node é a estrutura que responsável por armazenar os blocos de dados. Cada Data Node se comunica apenas com o Name Node, de forma que todas as operações de escrita e leitura de arquivos é mediada pelo Name Node.

IV – Para melhorar a disponibilidade do HDFS é possível definir um Name Node secundário que funciona apenas como uma cópia do Name Node principal e entra em funcionamento quando o Name Node principal se torna indisponível. Outra possibilidade é a utilização do Zookeeper para auxiliar na recuperação de um Name Node inacessível.

V – Data Node é a estrutura responsável por armazenar os blocos de dados. Uma vez que o Cliente consulta o Name Node para buscar a localização dos blocos de dados que deseja consultar, deve se comunicar diretamente com o Data Node correspondente para realizar o acesso aos dados.

Dentre essas afirmações estão CORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | I, III e IV |
|  | B | I, IV e V  Você assinalou essa alternativa (B) |
|  | C | I, II e V |
|  | D | Todas, exceto II |
|  | E | II, III e V |

Questão 1/10 - Big Data

Em relação à seguinte afirmação:

“O crescente volume de dados gerados a todo o momento indica com relativa clareza que este é, justamente, o atributo mais significativo quando falamos em Big Data.”

Podemos afirmar que:

I – O conceito de Big Data pode ser definido única e exclusivamente pelo volume de dados processados.

II – A relação entre dado volume de dados e a escalabilidade, eficiência, custo e complexidade de processá-los é o que define se o volume de dados necessita da utilização de soluções de Big Data.

III – É amplamente definido que para ser considerada uma solução de Big Data, tal implementação deve ser capaz de processar um volume de pelo menos 40 Gigabytes de dados.

IV – Dados de sensores devem ser necessariamente processados por soluções de Big Data devido ao seu volume que pode alcançar Terabytes.

V – As soluções de Big Data devem ser aplicadas nos casos em que a tecnologia tradicional não é capaz de processar um grande volume de dados com a eficiência, escalabilidade, custo e complexidade necessária.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, II e III |
|  | B | Apenas II e V  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: O conceito de Big Data não pode ser definido apenas pelo tamanho do volume de dados e nem é possível determinar um valor específico em que o tamanho dos dados é suficientemente grande. Ao invés disso, devemos relacionar o volume de dados com a escalabilidade, eficiência, custo e complexidade de forma a superar as limitações das tecnologias tradicionais. |
|  | C | Nenhuma |
|  | D | Apenas I |
|  | E | III, IV e V |

Questão 2/10 - Big Data

O Hadoop é um framework amplamente utilizado no desenvolvimento de sistemas de processamento de Big Data. A sua grande contribuição foi aliar um sistema de arquivos distribuído a um modelo de programação distribuída. Com isso, avalie as seguintes afirmações:

I – O Hadoop Distributed File System é o sistema do Hadoop responsável por gerenciar os arquivos armazenados. Para garantir que não ocorra perda de dados no caso de falhas, tal sistema realiza o backup dos arquivos de cada um dos servidores do cluster em um servidor dedicado para backup.

II – A natureza distribuída do Hadoop se relaciona diretamente com a escalabilidade de sua capacidade de processamento e com a capacidade de garantir seu funcionamento mesmo que alguns dos componentes de sua infraestrutura falhem.

III – O Hadoop Map Reduce é o sistema responsável por gerenciar o processamento de forma distribuída. Tal sistema é capaz de dividir a execução da aplicação em pequenas tarefas que são processadas pelos servidores do cluster.

IV – As etapas de processamento do Map Reduce são: Map, Shuffle e Reduce.

V – Devido às suas características distribuídas o Haddop é capaz de realizar o balanceamento de carga entre os servidores do cluster, o que permite que todos os recursos operem de forma balanceada.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | II, III e V |
|  | B | Todas |
|  | C | I e IV |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: As afirmações II, III, IV e V corretamente informam as características do framework Hadoop. Porém a afirmação I erra ao afirmar que o Hadoop possui um servidor dedicado ao backup. No Hadoop os dados são armazenados em pequenos blocos que são replicados por todo o cluster, garantindo assim que os dados não sejam perdidos em caso de falhas no cluster. |
|  | E | III e V |

Questão 3/10 - Big Data

Avalie as seguintes afirmações sobre as operações de leitura e escrita no HDFS:

I – Durante a operação de leitura no HDFS, o Name Node informa ao cliente quais são os Data Nodes mais eficientes que possuem os blocos de dados desejados.

II – O cliente deve requisitar a operação de escrita ao Name Node que informa ao cliente quais são os Name Nodes mais eficientes. Em seguida o cliente deve comunicar com esses Name Nodes para armazenar os dados do arquivo dividido em blocos de 128MB.

III – A operação de escrita de dados deve ser realizada pelo servidor cliente que distribui o arquivo em blocos aos Name Nodes mais eficientes do cluster.

IV –  A operação de escrita de dados deve ser realizada pelo servidor cliente que requisita a criação de um arquivo ao Name Node. Em seguida o cliente comunica com um único Data Node que irá distribuir os dados entre os outros Data Nodes. Por fim o Data Node retorna à localização dos blocos de dados ao cliente. E o cliente repassa a informação para ser armazenada no Name Node

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | III e IV |
|  | B | I e IV |
|  | C | I e III |
|  | D | II e III  Você assinalou essa alternativa (D)  **Você acertou!**  Justificativa: A operação de escrita de dados no HDFS começa com o client node requisitando a criação do arquivo ao Name Node. E seguida o Client Node comunica com um único Data Node que redistribui blocos do arquivo para outros Data Nodes. A localização dos blocos é retornada para o Client Node que informa ao Name Node e, com isso, conclui a escrita. Portando, o Name Node não informa nenhum Data Node durante a escrita, como INCORRETAMENTE dizem as afirmações II e III. As demais afirmações (I e IV) estão CORRETAS. |
|  | E | II e IV |

Questão 4/10 - Big Data

O MongoDB é um banco de dados NoSQL distribuído e multiplataforma desenvolvido pela MongoDB Inc.  
  
Avalie as seguintes afirmações a respeito do MongoDB:  
  
I – É um banco de dados. Eventualmente consistente, ou seja, que prioriza a disponibilidade e a tolerância a falhas no teorema CAP.

II – O MongoDB é um banco de dados orientado a documentos com uma estrutura muito semelhante ao formato JSON.

III – Dados embutidos, ou desnormalizados são aqueles que armazenam pedaços de informação relacionada em um mesmo documento

IV – Dados referenciados ou normalizados são aqueles se fazem referência entre si e são úteis para evitar duplicações de dados desnecessárias.

V – O conjunto de réplicas do MongoDB melhoram a consistência do banco.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Nenhuma |
|  | B | I e V  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: O MongoDB é um banco que, em relação ao teorema CAP, privilegia a consistência dos dados e a tolerância a falhas. E, portanto, é um banco que pode estar eventualmente indisponível. Com isso, a afirmação I encontra-se INCORRETA. Os conjuntos de replicação oferecem redundância e melhoram a disponibilidade do banco. Dessa forma, a afirmação V está INCORRETA. |
|  | C | III e IV |
|  | D | Apenas II |
|  | E | II, IV e V |

Questão 5/10 - Big Data

Ao criar uma aplicação capaz de processar e analisar seus dados de forma a extrair informações precisas que possam gerar o valor esperado o analista de dados deve estar atento aos seguintes fatores:

I – Quão variados eles precisam ser

II – Qual o seu nível de atualização

III – Qual o seu nível de popularidade

IV – A relevância dos dados

V – A velocidade necessária para processá-los

Estão corretas as afirmações:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | Todas as alternativas |
|  | B | Apenas I e II |
|  | C | Todas, exceto I e III |
|  | D | Apenas II, III e V |
|  | E | Todas exceto III  Você assinalou essa alternativa (E)  **Você acertou!**  Justificativa: O nível de popularidade dos dados não é um fator que altera como eles serão processados pela aplicação. As demais opções referem exatamente a algum aspecto abordado pelos Big Data V’s. |

Questão 6/10 - Big Data

O Oozie é o componente do Hadoop que realiza a orquestração de trabalhos. Avalie as seguintes afirmações sobre o Oozie:

I – O Ozzie utiliza de grafos cíclicos direcionados para implementar o encadeamento de tarefas de modo que tarefas que não possuem relação de dependência podem executar paralelamente.

II – Os fluxos de trabalho são especificados em notação YAML.

III – O Oozie permite a que sejam definidos coordenadores onde se pode configurar controles de execução, horário de início e frequência.

IV – O Oozie pode ser executado apenas através de gerenciadores de específicos em algumas interfaces.

V – Os fluxos de trabalho representam grafos acíclicos direcionados de tarefas capazes de executar de forma encadeada e automática.

Dentre essas afirmações estão INCORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I, II e IV  Justificativa: Os fluxos de trabalho do Oozie são grafos acíclicos direcionados que encadeiam diversas tarefas. Por isso a afirmação I encontra-se INCORRETA. Os fluxos de trabalho são escritos em XML e não em YAML como alega a afirmação II. Além de poder ser executado através de gerenciadores em algumas interfaces, também pode ser executado através da linha de comando, portanto a afirmação IV está INCORRETA. As demais afirmações (III e V) sobre o Oozie estão CORRETAS. |
|  | B | I e II |
|  | C | III e V |
|  | D | Todas, exceto II |
|  | E | II, IV e V  Você assinalou essa alternativa (E) |

Questão 7/10 - Big Data

Muitas aplicações necessitam que os dados sejam processados à medida que são recebidos pela aplicação e que o resultado de tal processamento esteja disponível tão rapidamente quanto. A esse tipo de operação damos o nome de processamento de fluxo em tempo real.  
  
A cerca de tais aplicações, avalie as seguintes afirmações:  
  
I – O conceito de baixa latência define que os dados devem ser processados em um tempo igual ou menor que o tempo em que novos dados chegam ao fluxo.

II – Para ser considerado consistente a solução não pode tolerar imperfeições e inconsistências

III – Dados significativos podem ser perdidos caso a aplicação tenha problemas na coleta, transmissão ou processamento dos dados.

IV – Spark é uma aplicação que utiliza o Map Reduce para processar dados em tempo real.

V – Spark possui seu próprio sistema de arquivos, o SparkSQL que implementa um banco de dados otimizado para realizar consultas em tempo real.

Dentre essas afirmações estão INCORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I e II |
|  | B | Todas, exceto III |
|  | C | IV e V |
|  | D | Todas |
|  | E | II, IV e V  Você assinalou essa alternativa (E)  **Você acertou!**  Justificativa: As afirmações I e III informam corretamente a respeito dos conceitos de baixa latência e alta disponibilidade respectivamente. Na afirmação II o conceito de consistência está invertido, é exatamente o contrário. A solução deve tolerar imperfeições e inconsistências para ser considerada consistente. Enquanto que as afirmações IV e V erram em definir o framework de processamento distribuído do Spark (IV) e o armazenamento de dados no Spark (V). No caso, o Spark implementa um framework próprio de processamento distribuído de dados, sua principal vantagem é, justamente, ser mais eficiente que o próprio Hadoop Map Reduce. E a respeito do SparkSQL é o framework de acesso aos dados armazenados em qualquer sistema de arquivos suportado pelo Spark (inclusive o HDFS) utilizando a linguagem SQL. |

Questão 8/10 - Big Data

“Zookeeper é um serviço open source distribuído de coordenação para aplicações distribuídas. Ele é capaz de manter dados de forma consistente por todo o cluster”. A respeito do Zookeeper avalie as seguintes afirmações:

I – O Zookeeper fornece um conjunto de primitivas que permitem a implementação de serviços de sincronização, manutenção de configuração, gerenciamento de grupos e nomeação.

II – Em caso de falha do nó mestre, o próprio sistema elege um novo nó mestre que assume imediatamente as informações de execução do nó que falhou. Em caso de falha de um dos nós trabalhadores, o Zookeeper redistribui o trabalho para outro nó adequadamente.

III – Znodes são arquivos permanentes que visam manter, de forma consistente, as informações a respeito de nós e tarefas de uma aplicação.

IV – Quando um nó deixa de atualizar o seu registro, ele deixa de existir no sistema de arquivos, tornando-se indisponível para o Zookeeper.

V – Cada nó da aplicação possui um cliente Zookeeper que monitora as informações que devem ser registradas. Os clientes Zookeeper nos nós são supervisionados por um conjunto de servidores que replicam seus dados entre si para garantir que seja possível continuar operando normalmente mesmo em caso de falha.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | Todas |
|  | B | I, IV e V  Justificativa: Em caso de falha de um dos nós trabalhadores, o Zookeeper pede para a aplicação redistribuir o trabalho para outro nó adequadamente. O Zookeeper não redistribui diretamente o trabalho, ou seja, a afirmação II encontra-se INCORRETA. Os Znodes podem ser efêmeros, como os que armazenam o estado de um nó ou permanentes, como os que armazenam as tarefas dos nós trabalhadores. Portanto a afirmação III é INCORRETA. As demais afirmações estão CORRETAS. |
|  | C | IV e V |
|  | D | Todas exceto II  Você assinalou essa alternativa (D) |
|  | E | II, III e IV |

Questão 9/10 - Big Data

“Apenas recentemente a capacidade de armazenamento e processamento se tornaram suficientes para permitir que dados antes ignorados fossem analisados. Entretanto, além dos componentes tecnológicos, o analista de dados deve ser capaz de identificar quais dados se deve utilizar, como integrá-los, quais as perguntas que serão úteis para a tomada de decisão, e qual a melhor maneira de apresentar os resultados obtidos da análise. ”  
  
Sobre a fase de análise de dados avalie as seguintes afirmações:  
  
I – Durante a fase de análise de dados, o cientista de dados deve preparar os dados brutos para que possam ser utilizados na produção de informações úteis e padrões ocultos em conjuntos massivos de dados.

II –  A partir do conjunto de dados inicial podemos aplicar análises estatísticas utilizando modelos estatísticos ou ainda aplicar modelos de predições baseados em aprendizado de máquina de forma que os dados sejam refinados.

III – Realizar análise exploratória e verificar a qualidade dos dados são atividades da análise de dados que fazem parte da fase de preparação dos dados

IV –  A fase de modelagem dos dados é composta pelas seguintes tarefas: Selecionar técnicas de modelagem, projetar testes, definir e construir o modelo de dados, seus parâmetros e sua descrição, validar o modelo e definir os parâmetros a serem revisados, avaliar os resultados do modelo; revisar processos; e determinar os passos seguintes.

V – O aprendizado de máquina é uma das atividades necessárias na automatização da construção de modelos analíticos que é uma das tarefas da fase de compreensão dos dados.

Dentre essas afirmações estão CORRETAS:

Nota: 0.0Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I e II  Justificativa: As afirmações I e II (CORRETAS) acertam sobre o papel do cientista de dados e a aplicação de modelos estatísticos sobre os dados que estão sendo recebidos. A afirmação III indica um conjunto de atividades que fazem parte da compreensão dos dados e não da preparação dos dados. A afirmação IV confunde e mistura tarefas da modelagem de dados e da avaliação do modelo. A Automatização da construção de modelos analíticos NÃO é uma das tarefas da compreensão dos dados. |
|  | B | Todas  Você assinalou essa alternativa (B) |
|  | C | IV e V |
|  | D | Todas, exceto I |
|  | E | II, IV e V |

Questão 10/10 - Big Data

O sistema de arquivos HDFS possui dois tipos de estruturas principais: Name Nodes e Data Nodes. Sobre a arquitetura do HDFS avalie as seguintes afirmações:  
  
I –  Name Node possui um registro de edição que contém a informação dos arquivos criados. Além disso, é o tipo de estrutura responsável por armazenar a localização dos blocos de cada um dos arquivos armazenados no HDFS. Dessa forma é possível rastrear quais os blocos de quais arquivos estão armazenados em cada Data Node.

II – Não é possível possuir mais de um Name Node no HDFS. Dessa forma se o Name Node se torna indisponível não há como obter os metadados contido nele.

III – Data Node é a estrutura que responsável por armazenar os blocos de dados. Cada Data Node se comunica apenas com o Name Node, de forma que todas as operações de escrita e leitura de arquivos é mediada pelo Name Node.

IV – Para melhorar a disponibilidade do HDFS é possível definir um Name Node secundário que funciona apenas como uma cópia do Name Node principal e entra em funcionamento quando o Name Node principal se torna indisponível. Outra possibilidade é a utilização do Zookeeper para auxiliar na recuperação de um Name Node inacessível.

V – Data Node é a estrutura responsável por armazenar os blocos de dados. Uma vez que o Cliente consulta o Name Node para buscar a localização dos blocos de dados que deseja consultar, deve se comunicar diretamente com o Data Node correspondente para realizar o acesso aos dados.

Dentre essas afirmações estão CORRETAS:

Nota: 10.0

|  |  |  |
| --- | --- | --- |
|  | A | I, III e IV |
|  | B | I, IV e V  Você assinalou essa alternativa (B)  **Você acertou!**  Justificativa: É possível utilizar zero ou mais Name Nodes secundários que replicam os dados do Name Node em operação (principal). No caso de falha do Name Node principal um dos Name Nodes secundários podem substituí-lo. Portanto a afirmação II encontra-se INCORRETA. O Name Node armazena os registros de localização e edição dos blocos dos arquivos e comunica-se apenas com o Client Node que busca tais informações para acessar diretamente os Data Nodes para escritas e leituras de arquivos. Portanto a afirmação III encontra-se INCORRETA. As demais afirmações trazem informações CORRETAS sobre o HDFS. |
|  | C | I, II e V |
|  | D | Todas, exceto II |
|  | E | II, III e V |

Questão 1/10 - Big Data

“Apenas recentemente a capacidade de armazenamento e processamento se tornaram suficientes para permitir que dados antes ignorados fossem analisados. Entretanto, além dos componentes tecnológicos, o analista de dados deve ser capaz de identificar quais dados se deve utilizar, como integrá-los, quais as perguntas que serão úteis para a tomada de decisão, e qual a melhor maneira de apresentar os resultados obtidos da análise. ”  
  
Sobre a fase de análise de dados avalie as seguintes afirmações:  
  
I – Durante a fase de análise de dados, o cientista de dados deve preparar os dados brutos para que possam ser utilizados na produção de informações úteis e padrões ocultos em conjuntos massivos de dados.

II –  A partir do conjunto de dados inicial podemos aplicar análises estatísticas utilizando modelos estatísticos ou ainda aplicar modelos de predições baseados em aprendizado de máquina de forma que os dados sejam refinados.

III – Realizar análise exploratória e verificar a qualidade dos dados são atividades da análise de dados que fazem parte da fase de preparação dos dados

IV –  A fase de modelagem dos dados é composta pelas seguintes tarefas: Selecionar técnicas de modelagem, projetar testes, definir e construir o modelo de dados, seus parâmetros e sua descrição, validar o modelo e definir os parâmetros a serem revisados, avaliar os resultados do modelo; revisar processos; e determinar os passos seguintes.

V – O aprendizado de máquina é uma das atividades necessárias na automatização da construção de modelos analíticos que é uma das tarefas da fase de compreensão dos dados.

Dentre essas afirmações estão CORRETAS:

Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I e II |
|  | B | Todas |
|  | C | IV e V |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D) |
|  | E | II, IV e V |

Questão 2/10 - Big Data

Avalie as seguintes afirmações sobre as operações de leitura e escrita no HDFS:

I – Durante a operação de leitura no HDFS, o Name Node informa ao cliente quais são os Data Nodes mais eficientes que possuem os blocos de dados desejados.

II – O cliente deve requisitar a operação de escrita ao Name Node que informa ao cliente quais são os Name Nodes mais eficientes. Em seguida o cliente deve comunicar com esses Name Nodes para armazenar os dados do arquivo dividido em blocos de 128MB.

III – A operação de escrita de dados deve ser realizada pelo servidor cliente que distribui o arquivo em blocos aos Name Nodes mais eficientes do cluster.

IV –  A operação de escrita de dados deve ser realizada pelo servidor cliente que requisita a criação de um arquivo ao Name Node. Em seguida o cliente comunica com um único Data Node que irá distribuir os dados entre os outros Data Nodes. Por fim o Data Node retorna à localização dos blocos de dados ao cliente. E o cliente repassa a informação para ser armazenada no Name Node

Dentre essas afirmações estão INCORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | III e IV |
|  | B | I e IV |
|  | C | I e III |
|  | D | II e III  Você assinalou essa alternativa (D) |
|  | E | II e IV |

Questão 3/10 - Big Data

O Hadoop é um framework amplamente utilizado no desenvolvimento de sistemas de processamento de Big Data. A sua grande contribuição foi aliar um sistema de arquivos distribuído a um modelo de programação distribuída. Com isso, avalie as seguintes afirmações:

I – O Hadoop Distributed File System é o sistema do Hadoop responsável por gerenciar os arquivos armazenados. Para garantir que não ocorra perda de dados no caso de falhas, tal sistema realiza o backup dos arquivos de cada um dos servidores do cluster em um servidor dedicado para backup.

II – A natureza distribuída do Hadoop se relaciona diretamente com a escalabilidade de sua capacidade de processamento e com a capacidade de garantir seu funcionamento mesmo que alguns dos componentes de sua infraestrutura falhem.

III – O Hadoop Map Reduce é o sistema responsável por gerenciar o processamento de forma distribuída. Tal sistema é capaz de dividir a execução da aplicação em pequenas tarefas que são processadas pelos servidores do cluster.

IV – As etapas de processamento do Map Reduce são: Map, Shuffle e Reduce.

V – Devido às suas características distribuídas o Haddop é capaz de realizar o balanceamento de carga entre os servidores do cluster, o que permite que todos os recursos operem de forma balanceada.

Dentre essas afirmações estão CORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | II, III e V |
|  | B | Todas |
|  | C | I e IV |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D) |
|  | E | III e V |

Questão 4/10 - Big Data

A respeito do ecossistema Hadoop avalie as seguintes afirmações:  
  
I – Pig é uma API de alto nível para realizar consultas SQL. Dessa forma é possível processar diretamente os dados contidos no HDFS.

II – Spark é um componente capaz de substituir o MapReduce para realizar o processamento distribuído no ecossistema Hadoop.

III – Grafos Acíclicos Dirigidos são uma estratégia de otimização de processamento utilizada por alguns componentes como o Tez

IV – Sqoop, Flume e Kafka são aplicações capazes de obter dados de fontes externas ao Hadoop para o HDFS

V – Zookeeper é um componente capaz de coordenar as aplicações do cluster. Dessa forma, ele é amplamente utilizado por diversos outros componentes para melhorar a sua confiabilidade e desempenho.

Dentre essas afirmações estão INCORRETAS:

Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | I e III  Você assinalou essa alternativa (A) |
|  | B | Todas, exceto I |
|  | C | I |
|  | D | II, IV e V |
|  | E | Nenhuma |

Questão 5/10 - Big Data

Muitas aplicações necessitam que os dados sejam processados à medida que são recebidos pela aplicação e que o resultado de tal processamento esteja disponível tão rapidamente quanto. A esse tipo de operação damos o nome de processamento de fluxo em tempo real.  
  
A cerca de tais aplicações, avalie as seguintes afirmações:  
  
I – O conceito de baixa latência define que os dados devem ser processados em um tempo igual ou menor que o tempo em que novos dados chegam ao fluxo.

II – Para ser considerado consistente a solução não pode tolerar imperfeições e inconsistências

III – Dados significativos podem ser perdidos caso a aplicação tenha problemas na coleta, transmissão ou processamento dos dados.

IV – Spark é uma aplicação que utiliza o Map Reduce para processar dados em tempo real.

V – Spark possui seu próprio sistema de arquivos, o SparkSQL que implementa um banco de dados otimizado para realizar consultas em tempo real.

Dentre essas afirmações estão INCORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | I e II |
|  | B | Todas, exceto III |
|  | C | IV e V |
|  | D | Todas |
|  | E | II, IV e V  Você assinalou essa alternativa (E) |

Questão 6/10 - Big Data

Avalie as seguintes afirmações a respeito das operações definidas pelo MapReduce:

I – A função Mapper tem por objetivo mapear conjuntos de valores necessários pela aplicação a determinadas chaves que são geradas automaticamente pelo MapReduce.

II – A função Mapper tem por objetivo extrair e organizar os dados necessários para o processamento.

III – A função Reducer recebe o conjunto de dados no formato chave-valor que foi definido pela função Mapper.

IV – Na função Reducer, cada par chave-valor constitui uma tarefa a ser processada pelo servidor que estiver mais próximo dos dados.

V – A função Shuffle and sort transforma um conjunto de dados em formato chave-valor onde os valores devem ser embaralhados ou ordenados de acordo com a sua chave.

Dentre essas afirmações estão CORRETAS:

Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | II e IV |
|  | B | II, III e IV |
|  | C | I, IV e V |
|  | D | Todas, exceto I  Você assinalou essa alternativa (D) |
|  | E | I, III e IV |

Questão 7/10 - Big Data

O Oozie é o componente do Hadoop que realiza a orquestração de trabalhos. Avalie as seguintes afirmações sobre o Oozie:

I – O Ozzie utiliza de grafos cíclicos direcionados para implementar o encadeamento de tarefas de modo que tarefas que não possuem relação de dependência podem executar paralelamente.

II – Os fluxos de trabalho são especificados em notação YAML.

III – O Oozie permite a que sejam definidos coordenadores onde se pode configurar controles de execução, horário de início e frequência.

IV – O Oozie pode ser executado apenas através de gerenciadores de específicos em algumas interfaces.

V – Os fluxos de trabalho representam grafos acíclicos direcionados de tarefas capazes de executar de forma encadeada e automática.

Dentre essas afirmações estão INCORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | I, II e IV  Você assinalou essa alternativa (A) |
|  | B | I e II |
|  | C | III e V |
|  | D | Todas, exceto II |
|  | E | II, IV e V |

Questão 8/10 - Big Data

“Zookeeper é um serviço open source distribuído de coordenação para aplicações distribuídas. Ele é capaz de manter dados de forma consistente por todo o cluster”. A respeito do Zookeeper avalie as seguintes afirmações:

I – O Zookeeper fornece um conjunto de primitivas que permitem a implementação de serviços de sincronização, manutenção de configuração, gerenciamento de grupos e nomeação.

II – Em caso de falha do nó mestre, o próprio sistema elege um novo nó mestre que assume imediatamente as informações de execução do nó que falhou. Em caso de falha de um dos nós trabalhadores, o Zookeeper redistribui o trabalho para outro nó adequadamente.

III – Znodes são arquivos permanentes que visam manter, de forma consistente, as informações a respeito de nós e tarefas de uma aplicação.

IV – Quando um nó deixa de atualizar o seu registro, ele deixa de existir no sistema de arquivos, tornando-se indisponível para o Zookeeper.

V – Cada nó da aplicação possui um cliente Zookeeper que monitora as informações que devem ser registradas. Os clientes Zookeeper nos nós são supervisionados por um conjunto de servidores que replicam seus dados entre si para garantir que seja possível continuar operando normalmente mesmo em caso de falha.

Dentre essas afirmações estão CORRETAS:

Você não pontuou essa questão

|  |  |  |
| --- | --- | --- |
|  | A | Todas  Você assinalou essa alternativa (A) |
|  | B | I, IV e V |
|  | C | IV e V |
|  | D | Todas exceto II |
|  | E | II, III e IV |

Questão 9/10 - Big Data

O sistema de arquivos HDFS possui dois tipos de estruturas principais: Name Nodes e Data Nodes. Sobre a arquitetura do HDFS avalie as seguintes afirmações:  
  
I –  Name Node possui um registro de edição que contém a informação dos arquivos criados. Além disso, é o tipo de estrutura responsável por armazenar a localização dos blocos de cada um dos arquivos armazenados no HDFS. Dessa forma é possível rastrear quais os blocos de quais arquivos estão armazenados em cada Data Node.

II – Não é possível possuir mais de um Name Node no HDFS. Dessa forma se o Name Node se torna indisponível não há como obter os metadados contido nele.

III – Data Node é a estrutura que responsável por armazenar os blocos de dados. Cada Data Node se comunica apenas com o Name Node, de forma que todas as operações de escrita e leitura de arquivos é mediada pelo Name Node.

IV – Para melhorar a disponibilidade do HDFS é possível definir um Name Node secundário que funciona apenas como uma cópia do Name Node principal e entra em funcionamento quando o Name Node principal se torna indisponível. Outra possibilidade é a utilização do Zookeeper para auxiliar na recuperação de um Name Node inacessível.

V – Data Node é a estrutura responsável por armazenar os blocos de dados. Uma vez que o Cliente consulta o Name Node para buscar a localização dos blocos de dados que deseja consultar, deve se comunicar diretamente com o Data Node correspondente para realizar o acesso aos dados.

Dentre essas afirmações estão CORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | I, III e IV |
|  | B | I, IV e V  Você assinalou essa alternativa (B) |
|  | C | I, II e V |
|  | D | Todas, exceto II |
|  | E | II, III e V |

Questão 10/10 - Big Data

“Bancos de Dados relacionais originalmente foram projetados baseados em uma técnica para o armazenamento de dados conhecida por *schema on write*. Essa estratégia define que as partes dos dados precisam se ajustar a um padrão ou um plano no momento da escrita (…) o Hive inverte a lógica e faz uso do conceito contrário: *scheme on read*. Ou seja, a estrutura dos dados, também conhecida por esquema, só é definida durante a leitura dos dados”. Sobre a arquitetura do Hive avalie as seguintes afirmações:

I – *Driver* é o componente do Hive que recebe os comandos e consultas, compila a entrada, otimiza a computação necessária e executa os passos com tarefas MapReduce ou Tez.

II – O Hive utiliza Mappers e Reducers genéricos que são sequenciados por planos de trabalho escritos em XML.

III – A estratégia *Scheme on Read* é considerada muito restritiva, uma vez que uma quantidade considerável de esforços é investida ajustando os dados à estrutura definida.

IV – Os metadados que definem os esquemas e particionamentos do Hive são armazenados em um banco de dados relacional. Normalmente Derby ou MySQL.

V – É possível realizar consultas utilizando o Hive através de uma interface de linha de comando (CLI) ou através da Hive Web Interface (HWI).

Dentre essas afirmações estão CORRETAS:

|  |  |  |
| --- | --- | --- |
|  | A | Todas |
|  | B | I, II e III |
|  | C | Todas, exceto III  Você assinalou essa alternativa (C) |
|  | D | III e IV |
|  | E | II, IV e V |