**Lista de Exercícios - Arquitetura Online para FCD**

Curso: Ciência de Dados

Data: //2018

Nome:

Matrícula:

1 - O que são “fluxos contínuos de dados”? Cite suas principais características. Forneça 6 exemplos de aplicações e dispositivos que geram dados possíveis de serem utilizados como fluxos contínuos de dados.

2 - Por que os modelos clássicos de Machine Learning são considerados obsoletos quando queremos tratar dados que chegam via *stream*. Qual é o funcionamento básico desses modelos? Em que ponto eles não atendem a esse novo formato de dados?

3 - Duas grandes novidades de mercado geraram e ainda geram um grande aumento na quantidade de aplicações e dados baseados em fluxos contínuos de dados. A primeira é o fenômeno da IoT (*Internet of Things*), também chamado de Internet das Coisas. A segunda novidade é a chegada do protocolo IPv6. Como esses fenômenos contribuíram para esse aumento da quantidade de aplicações?

4 - Uma grande contribuição para o tratamento dos dados em tempo real foi a evolução das ferramentas de Big Data. As mais atuais são muito escaláveis e podem tratar grandes volumes de dados em tempo real. Pensando em uma arquitetura para tratamento de dados em tempo real, quais ferramentas você pode utilizar para fazer: a ingestão de dados, o processamento dos dados e o armazenamento dos dados tratados (pense em NoSQL)? Explique o objetivo de cada uma das ferramentas informadas.

5 - Baseado nas ferramentas informadas na questão 4, pense em um pipeline básico para processamento de fluxos contínuos de dados e descreva o papel de cada ferramenta, explicando como seria esse fluxo, desde a ingestão dos dados até a persistência dos dados processados.

6 - Imagine uma grande empresa que possui diversas aplicações. Ela precisa que todas elas consigam se comunicar. Foi feito um grande levantamento de requisitos para atender a essa demanda e diversas soluções baseadas em ETL (Extract, Transform, Load - exemplo de ferramentas são o Talend e o Pentaho) precisam ser desenvolvidas. Um dos desenvolvedores informou aos arquitetos que existe uma ferramenta da Apache que poderia solucionar esse problema com excelência e que é muito escalável. Os arquitetos não acreditaram tanto no desenvolvedor e resolveram pesquisar mais sobre a ferramenta. Após a pesquisa uma nova solução utilizando o que foi proposto pelo desenvolvedor foi feita e aprovada para a solução. Baseado nessa situação, qual foi a ferramenta informada pelo desenvolvedor?

1. Apache Flume.
2. Apache Kafka.
3. Apache Storm.
4. Apache Kylin.

7 - Em uma empresa de vendas online foram extraídos todos os dados de avaliações dos produtos. Baseado nesses dados um modelo para Sistemas de Recomendações foi criado e foram geradas 10 recomendações para cada cliente, baseado nesse modelo. As recomendações geradas foram armazenadas em um banco de dados NoSQL (MongoDB, Cassandra, HBase, BigTable, etc). Essas recomendações são geradas novamente todos os dias ao final do dia, baseado nos novos e antigos dados de vendas, que aumentam a cada venda gerada. Após a geração dessas recomendações foi necessário criar um fluxo online com o objetivo de obter os itens recomendados para um dado usuário que acesse a página da empresa. Para isso um tópico foi criado no Apache Kafka, que recebe o CPF do usuário logado e um ID da sessão do usuário. Esse tópico é consumido por uma ferramenta de processamento de dados em tempo real e baseado no CPF a lista de recomendações é consultada no banco de dados e retornada para um serviço ou para um tópico de resposta, juntamente com o identificador da sessão que necessita desses dados. Qual das opções abaixo contém uma ferramenta que poderia ser utilizada para o processamento desses dados em tempo real?

1. Apache Flume.
2. Apache Kafka.
3. Apache Storm.
4. Apache Kylin.

8 - Baseado na resposta anterior e no conteúdo de sala de aula, existe uma outra ferramenta de processamento distribuído que poderia ser utilizada para solucionar o problema da empresa de vendas. Qual seria essa ferramenta? Explique a diferença entre ela e a ferramenta escolhida na questão anterior.

9 - Atualmente as empresas que trabalham com Big Data em aplicações de tempo real utilizam soluções com infraestrutura em Cloud. Explique porque essas soluções atendem muito bem este tipo de demanda e possui um custo bem reduzido em comparação a uma infraestrutura criada toda internamente, para resolver o mesmo problema.

10 - Existe uma arquitetura denominada Flafka por diversas empresas. Ela possui a combinação de duas ferramentas muito utilizadas na etapa de ingestão de dados. Quais são essas duas ferramentas? Com relação à tolerância a falha, porque essa combinação de ferramentas é interessante? Qual o papel de cada uma das duas ferramentas nessa arquitetura?

11 - O que é o Apache Flume? Como ele funciona?

12 - O que é o Apache Kafka? Como ele funciona? Qual empresa é responsável pela sua criação?