1. Arquitetura e Configuração:

Descreva a arquitetura geral de uma implementação típica do Databricks. Como você configuraria um ambiente Databricks para otimizar o desempenho e a escalabilidade?

2. Apache Spark e Databricks:

Como o Databricks se integra ao Apache Spark? Quais são as principais vantagens do uso do Databricks em comparação com uma instalação padrão do Apache Spark?

3. Notebooks e Linguagens de Programação:

Explique como você usaria Notebooks no Databricks para criar e executar código. Além disso, como o Databricks suporta várias linguagens de programação, e como você decidiria qual linguagem usar em um projeto específico?

4. Integração de Fontes de Dados:

Como o Databricks facilita a integração com diferentes fontes de dados, como Data Lakes, bancos de dados relacionais e fontes externas? Você pode fornecer um exemplo prático de como lidar com essas integrações?

5. Machine Learning no Databricks:

Descreva a abordagem que você seguiria para desenvolver e treinar modelos de machine learning no Databricks. Quais são as principais ferramentas e bibliotecas que você usaria para esse fim?

6. Segurança e Controle de Acesso:

Como o Databricks aborda questões de segurança e controle de acesso? Quais são as práticas recomendadas para garantir a proteção dos dados e ambientes de desenvolvimento?

7. Desafios e Soluções:

Pergunta: Conte-nos sobre um desafio específico que você enfrentou ao trabalhar com Databricks e como o resolveu. Qual foi a solução implementada e quais foram os resultados alcançados?

Conjunto de Dados do Kaggle:

Escolha um conjunto de dados do Kaggle relacionado a vendas. Certifique-se de que o conjunto de dados inclui informações como datas, produtos, quantidades vendidas, etc.

Projeto de Engenharia de Dados:

Ingestão e Carregamento de Dados:

Carregue o conjunto de dados no Databricks.

Explore o esquema dos dados e faça ajustes conforme necessário.

Transformações de Dados:

Realize transformações necessárias, como tratamento de valores nulos, conversões de tipos, etc.

Adicione uma coluna calculada, por exemplo, o valor total de cada transação.

Agregue os dados para obter estatísticas de vendas, por exemplo, o total de vendas por produto ou por categoria.

Introduza uma regra mais complexa, como identificar padrões de comportamento de compra ao longo do tempo ou criar categorias personalizadas de produtos com base em determinados critérios.

Saída em Parquet e Delta:

Grave os dados transformados e agregados em um formato Parquet para persistência eficiente.

Grave os mesmos dados em formato Delta Lake para aproveitar as funcionalidades de versionamento e transações ACID.

Exploração Adicional (Opcional):

Execute consultas exploratórias para entender melhor os dados e validar as transformações.

Crie visualizações ou relatórios para comunicar insights.

Agende o notebook para execução automática em intervalos regulares para garantir a atualização contínua dos dados.