**1. Como você utiliza o Delta Lake no Azure Databricks para garantir a integridade dos dados?**

O Delta Lake oferece funcionalidades como controle de versão, gerenciamento de transações e um esquema evolutivo. Isso garante a integradade dos dados, permitindo que o você realize operações como atualizações e exclusões de uma forma segura, além de manter um historico das alterações. Com o Delta Lake, voçê também pode utilizar o time travel para consultar dados em versões anteriores, ajudando na recuperação de dados.

**2. Quais são as vantagens do uso do Spark em comparação com outras tecnologias de processamento de dados?**

A utilização do Spark oferece diversas vantagens, incluindo a velocidade de processamento em memória, a otimização na execução dos jobs e o paralelismo de processamento. Ele também se destaca pela versatilidade em trabalhar com dados estruturados, semiestruturados e não estruturados, além de oferecer suporte a arquivos em formatos como JSON, Parquet e Avro. Outras vantagens incluem a escalabilidade, o processamento em tempo real e uma comunidade ativa, com uma vasta quantidade de documentação disponível.

**3. Descreva um caso em que você precisou sincronizar dados entre diferentes sistemas.**

Atualmente, trabalho com a ingestão de dados provenientes de diversos sistemas, incluindo bancos SQL, Oracle, SAP e APIs. Esses dados são inicialmente armazenados na camada Bronze, onde permanecem brutos, sem tratamento ou transformação. Na camada Silver, os dados são processados e transformados de acordo com as necessidades do negócio, até que sejam refinados na camada Gold. Os dados das APIs são armazenados na camada Bronze e, em seguida, passam por um processo de ELT conforme a demanda específica.

Exemplo: A conexão com um banco SQL é realizada através do Federation no Databricks, criando um schema com os dados do banco. Após esse acesso, é desenvolvido um pipeline de ingestão que armazena as tabelas necessárias na camada Bronze, conforme a solicitação. Em seguida, é criado um workflow para disponibilizar os dados, que passam por tratamento, transformação e, se necessário, joins com outras fontes nas camadas Silver e Gold. Todos os pipelines são agendados com tasks configuradas com ou sem dependências dentro do workflow. Por fim, a conexão é feita através de um Cluster DW, enviando os dados para o Power BI ou outra ferramenta de visualização.

**4. Desenhe uma arquitetura de dados comentada para uma empresa que utiliza Azure e Databricks, incluindo armazenamento, processamento e análise.**

**A screenshot of a computer

Description automatically generated**

**5. Como você garante a escalabilidade e a robustez da arquitetura de dados?**

A utilização do Spark para o processamento de dados, juntamente com o Delta Lake, permite a criação de camadas bem definidas, garantindo controle de versionamento eficaz. Implementamos governança e gerenciamento de dados rigorosos, além de testes para assegurar a qualidade. As ingestações são programadas de acordo com a demanda do negócio, e as execuções de processos são integradas em workflows, com monitoramento contínuo do ambiente para garantir eficiência e desempenho.

**6. Como você implementa a criptografia de dados em repouso e em trânsito?**

Criptografia em Repouso - Armazenamento seguro utilizando serviços de armazenamento que oferecem criptografia, como o Azure Blob Storage ou o Amazon S3. Criptografia de bancos de dados. Gerenciamento de chaves, utilizando o Azure Key Vault ou o AWS Key Management Service. Criptografia de arquivos, aplicando algoritmos robustos para criptografar arquivos sensíveis antes do armazenamento, aumentando a segurança.

Criptografia em Trânsito – Implementação de protocolos de segurança como HTTPS e TLS. Estabelecimento de conexões seguras com o uso de VPN. Proteção dos dados durante integrações com APIs, utilizando tokens de autenticação e criptografia para assegurar a comunicação segura. Implementação de auditoria e monitoramento contínuo para verificar se os dados em trânsito estão adequadamente protegidos.

**7. Como você gerencia a qualidade dos dados em um pipeline de dados?**

Versionamento de código, passando por aprovadores. Testes de códigos, fluxo para deploy e ambiente de desenvolvimento e produção separados.

**8. Qual a importância do FinOps para a engenharia de dados?**

O FinOps é fundamental para a engenharia de dados, pois traz práticas de gerenciamento financeiro que ajudam a otimizar custos e maximizar o valor dos investimentos em dados. Algumas características do FinOps incluem o gerenciamento de custos, planejamento e orçamento, aceleração de inovações, responsabilidade e cultura de custos, além de desempenho e escalabilidade.

**9. Como o DevOps ajuda o engenheiro de dados?**

O DevOps auxilia o engenheiro de dados por meio da implementação de versionamento eficaz, que garante rastreabilidade e controle sobre mudanças no código e nos dados. Além disso, facilita o deploy de aplicações, permitindo atualizações rápidas e seguras. A criação de pipelines CI/CD automatiza o processo de desenvolvimento e implantação, aumentando a eficiência e reduzindo o tempo de entrega. Isso resulta em um fluxo de trabalho mais ágil, melhor colaboração entre equipes e um ambiente de dados mais robusto e confiável.

**10. Como iniciamos um projeto de pipeline de dados?**

Definição de objetivos do projeto – Identificação das fontes de dados – Escolha da arquitetura – Seleção de ferramentas e tecnologias – Planejamento do pipeline – Implementação de segurança e governança – Desenvolvimento e testes – Automatização – Monitoramento – Documentação e treinamento, caso necessário.

**11. Como realizar CI/CD em um pipeline de dados?**

Versionamento de código, configuração de ambiente, automatização de testes, criação de pipelines CI/CD, implementação de monitoramento, automatização de deploy e documentação.

**12. Quais ferramentas de orquestração você já trabalhou?**

Workflow Databricks, Airflow e Airbyte.

**13. Quais suas motivações para ser um engenheiro de dados?**

O próprio desafio técnico, que estimula o estudo e a aprendizagem contínua. A oportunidade de aprender novas tecnologias e aplicá-las no dia a dia. A possibilidade de impactar o negócio, seja otimizando processos, realizando descobertas e apontando insights valiosos. Um bom mercado de trabalho, com salários competitivos. A oportunidade de conviver com pessoas que impactam meu dia a dia, promovendo ensino, aprendizado e compartilhamento de ideias.