📘 Guia de Estudos - Lakehouse na Azure

Criado por Aprender Dados

🎓 **Pronto para começar sua jornada?**
Faça sua assinatura e libere todos os cursos!

---

1. Setup do Ambiente

📹 Vídeo no YouTube 🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-09:02**: Introdução ao projeto Adventure Works.
  - Contextualização do treinamento e objetivos.
  - Importância da especialização na Azure para o mercado de trabalho.
- **14:50-57:45**: Configuração do ambiente na Azure.
  - Criação de assinatura na Azure.
  - Configuração dos recursos principais (Databricks, ADLS, ADF, Azure SQL).
- **57:45-1:24:24**: Integração e validação de recursos.
  - Explicação da arquitetura de dados.
  - Testes iniciais e solução de erros comuns.

Destaque:

- Aprender a configurar um ambiente completo na Azure é o primeiro passo para dominar arquiteturas modernas de dados.

---

2. Extraindo Dados com o Azure Data Factory (ADF)

📹 Vídeo no YouTube 🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-05:15**: Introdução ao ADF e seu papel no projeto.
  - Benefícios do Azure Data Factory para orquestração de dados.
- **09:27-30:05**: Configuração básica do pipeline de ingestão.
  - Configurando **Linked Services** e **Datasets**.
  - Pipeline simples para ingestão de dados do Azure SQL ao ADLS.
- **35:45-50:35**: Soluções para erros comuns.
  - Correção de permissões e ajustes em conectividade.
- **50:35-1:01:15**: Planejamento para ingestão em escala.
  - Uso de parâmetros para automação.

Destaque:

- O ADF simplifica a automação de fluxos de dados, essencial para projetos de grande escala.

---

3. Extração com Metadados no Azure Data Factory - Parte 1

📹 Vídeo no YouTube 🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-09:26**: Introdução ao conceito de ingestão com metadados.
  - Eliminação de configurações manuais em pipelines.
- **12:30-36:45**: Implementação prática.
  - Criação de tabelas de controle para ingestão dinâmica.
  - Configuração de loops funcionais no ADF.
- **45:38-59:45**: Ajustes e validação.
  - Testes de ingestão e solução de problemas.
  - Organização inicial do Data Lake.

Destaque:

- Ingestão com metadados melhora a escalabilidade e reduz erros operacionais.

---

4. Extração com Metadados no Azure Data Factory - Parte 2

📹 Vídeo no YouTube 🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-09:02**: Revisão da tabela de controle e loops no ADF.
  - Limitações no uso de Google Sheets como fonte de dados.
- **09:02-25:46**: Migração para banco de dados relacional.
  - Criação de tabelas de controle no Azure SQL.
  - Configuração de filtros para ingestão seletiva.
- **28:34-40:25**: Validação e testes.
  - Execução de ingestões completas e análise de logs.

Destaque:

- Tabelas de controle otimizam a ingestão, permitindo maior flexibilidade e governança.

---

5. Integrando o ADLS ao Databricks

📹 Vídeo no YouTube 🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-09:05**: Integração do Azure Data Lake Storage (ADLS) com o Databricks.
  - Configuração de **App Registration** e permissões.

    - Uso de **Databricks Secrets** para autenticação segura.
- **13:50-23:50**: Criação de tabelas Delta no Databricks.
    - Demonstração prática de leitura e gravação de dados.
- **26:15-31:45**: Boas práticas.
    - Organização do Data Lake em containers **Bronze, Prata e Ouro**.
    - Validação final da integração.

Destaque:

- A integração ADLS + Databricks é o coração da arquitetura **Lakehouse**. Ela combina o armazenamento escalável do ADLS com o processamento eficiente do Databricks.

---

6. Databricks Secrets e Azure Key Vault

📹 Vídeo no YouTube 🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-09:30**: Configuração de Key Vault e integração com o Databricks.
    - Criação e configuração de **Key Vault** na Azure.
    - Permissões para o Key Vault e sua integração com o Databricks.
- **15:25-30:35**: Implementação de segredos no Databricks.
    - Uso de **scopes** para armazenar credenciais sensíveis.
    - Testes de integração para acessar o Data Lake.
- **40:20-55:00**: Aplicação prática.
    - Uso de segredos protegidos em pipelines.
    - Boas práticas para segurança em projetos de dados.

Destaque:

- Garantir a segurança das credenciais é essencial em projetos de dados. O Key Vault centraliza e protege informações sensíveis.

---

7. Camada Bronze - Ingestão de Dados com PySpark

📹 Vídeo no YouTube 🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-10:20**: Introdução à camada Bronze.
    - Diferenças entre tabelas **Managed** e **External** no Databricks.
- **15:30-25:40**: Ingestão de dados no formato Delta.
    - Automação da ingestão com loops e lógica **Read-Transform-Save**.
- **40:15-55:30**: Otimização de pipelines.
    - Estratégias de paralelismo para grandes volumes de dados.
    - Uso de **Current Timestamp** para registro de ingestão.

Destaque:

- A camada Bronze é fundamental para organizar os dados crus, criando uma base sólida para as próximas transformações.

---

8. Camada Bronze - Automação entre ADF e Databricks

📹 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-05:30**: Introdução à aula e objetivos.
  - Revisão do progresso e metas para integração.
- **12:10-24:50**: Criação de tabelas de controle.
  - Gestão de ingestão e extração.
  - Melhorias nos pipelines da camada Bronze.
- **32:30-50:05**: Integração prática.
  - Configuração de pipelines entre ADF e Databricks.
  - Uso de parâmetros para personalização de ingestões.
- **1:00:20-1:30:45**: Automação completa.
  - Pipelines otimizados e execução paralela de tarefas.
  - Finalização e validação.

Destaque:

- A integração entre ADF e Databricks permite automação e escalabilidade no fluxo de dados.

---

9. Camada Prata - O que é e como fazer

📹 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-09:02**: Introdução à camada Prata.
  - Benefícios de granularidade e deduplicação de dados.
- **15:30-30:15**: Transformações na camada Prata.
  - Redução de complexidade dos dados crus.
  - Regras de validação e qualidade de dados.
- **40:25-1:05:20**: Implementação prática.
  - Criação de tabelas Prata no Databricks.
  - Aplicação de filtros e regras de negócio.

Destaque:

- A camada Prata é essencial para preparar os dados para análises mais avançadas.

---

10. Camada Prata - Framework com PySpark

📹 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-12:30**: Introdução ao framework de automação.
  - Conceitos e objetivos do framework.
- **18:45-50:25**: Implementação prática.
  - Criação de funções reutilizáveis com PySpark.
  - Transformações básicas para qualidade de dados.
- **1:02:30-1:28:15**: Testes e otimizações.
  - Automação de transformações para múltiplas tabelas.
  - Planejamento para a camada Ouro.

Destaque:

- Frameworks reutilizáveis aumentam a eficiência e reduzem o esforço manual.

---

11. Camada Prata - Implementação do Framework

📹 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-18:32**: Revisão e objetivos.
  - Complementação do framework com PySpark.
- **25:50-40:15**: Implementação de transformações avançadas.
  - Deduplicação e qualidade de dados na camada Prata.
- **46:10-52:13**: Finalização e validação.
  - Configuração de pipelines automatizadas com validações robustas.

Destaque:

- Frameworks bem implementados garantem escalabilidade e qualidade nos projetos de dados.

---

12. Camada Prata - Finalizando o Framework com IA

📹 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-25:50**: Otimização com IA.

    - Aprimorando regras de qualidade de dados com técnicas avançadas.
- **31:26-46:10**: Automação e testes unitários.
    - Implementação de testes para validação de transformações realizadas.
- **1:00:15-1:05:20**: Planejamento para a camada Ouro.
    - Preparação para análises avançadas.

Destaque:

- A combinação de IA e frameworks potencializa o desempenho e a confiabilidade.

---

13. Camada Ouro - Planejando as Tabelas

📽️ Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-09:27**: Planejamento inicial.
    - Organização do esquema e definição de tabelas.
- **15:30-35:40**: Tabelas de fato e dimensões.
    - Relacionamentos e otimizações para relatórios.
- **40:25-46:10**: Fluxo de dados entre camadas.
    - Configuração de pipelines para a camada Ouro.

Destaque:

- Um planejamento bem estruturado é essencial para garantir eficiência e qualidade nos dados analíticos.

---

14. Camada Ouro - Implementando em SQL

📽️ Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-25:50**: Criação de tabelas com SQL.
    - Joins, validação e limpeza de dados.
- **35:40-55:00**: Implementação prática.
    - Configuração de tabelas Ouro no Databricks.
    - Testes e ajustes finais.

Destaque:

- SQL é uma ferramenta poderosa para modelagem e validação de dados na camada Ouro.

---

15. Camada Ouro - Implementando em PySpark e GitHub

🎥 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-18:32**: Uso de PySpark para transformações.
  - Modularidade e reuso de código.
- **25:50-52:13**: Integração com GitHub.
  - Versionamento e organização de projetos.
- **1:10:20-1:15:50**: Finalização.
  - Planejamento para otimizações futuras.

Destaque:

- A integração com GitHub facilita a colaboração e o controle de versão em projetos de dados.

---

16. Camada Ouro - Finalização com PySpark e GPT

🎥 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-40:15**: Integração entre camadas Prata e Ouro.
  - Refinamento e otimização de tabelas.
- **46:10-1:28:15**: Discussão sobre IA em projetos de dados.
  - Reflexão sobre pipelines inteligentes e futuras implementações.

Destaque:

- A IA está transformando a forma como gerenciamos e otimizamos dados.

---

17. Debugando o Projeto

🎥 Vídeo no YouTube
🔗 Aula na Plataforma

Tópicos abordados:

- **00:00-25:50**: Identificação de erros.
  - Debugging de pipelines e mensagens de erro.
- **31:26-52:13**: Correção de problemas comuns.
  - Ajustes em joins e schemas inconsistentes.
- **1:00:15-1:05:20**: Otimizações finais.
  - Ferramentas de profiling e validação.

Destaque:

- Técnicas de debugging são fundamentais para resolver problemas e garantir a qualidade do projeto.