## Engenharia de dados com Azure

## Seção 1 – Informações

Engenharia de dados e Big data.

Prática aplicada para engenharia de dados.

## Seção 2 – Mercado de BI

### Aula 01 – Diferentes Profissões

**1. Analista de Negócios**

O Analista de Negócios atua como **ponte** entre as áreas **técnicas e o negócio**. Seu papel é entender as necessidades da empresa, traduzir essas demandas em requisitos para projetos de dados e garantir que as soluções entreguem valor real. Costuma trabalhar com levantamento de requisitos, análise de processos e validação de resultados junto aos stakeholders. É fundamental para garantir que os projetos de dados estejam alinhados com os objetivos estratégicos da organização.

**2. Desenvolvedor de BI (Business Intelligence)**

O Desenvolvedor de BI é responsável por criar soluções que transformam dados (dados do passado) brutos em informações úteis para tomada de decisão. Isso inclui o desenvolvimento de dashboards, relatórios, ETL (Extração, Transformação e Carga de dados) e integração de diferentes fontes de dados. Utiliza ferramentas como Power BI, Tableau, Qlik, entre outras. O foco está em facilitar o acesso à informação e apoiar gestores na análise de indicadores de desempenho.

**3. Cientista de Dados**

O Cientista de Dados é o profissional que explora grandes volumes de dados (Big Data) para gerar insights, construir modelos preditivos (análise preditiva) e apoiar decisões estratégicas. Utiliza técnicas de estatística, machine learning e programação para identificar padrões, prever tendências e automatizar processos analíticos. É uma função que exige forte conhecimento técnico e capacidade de comunicação para traduzir resultados complexos em ações práticas.

**4. Engenheiro de Dados**

O Engenheiro de Dados é responsável pela infraestrutura que permite o armazenamento, processamento e disponibilização dos dados. Atua na construção de pipelines de dados, arquitetura de bancos de dados, integração de sistemas e garantia da qualidade e segurança dos dados. É quem prepara o terreno para que analistas e cientistas possam trabalhar com dados confiáveis e acessíveis.

**5. Analista de Dados**

O Analista de Dados trabalha na análise exploratória e descritiva dos dados, identificando tendências, padrões e oportunidades de melhoria. Utiliza ferramentas estatísticas e de visualização para responder perguntas de negócio e apoiar decisões táticas e operacionais. É uma função que exige curiosidade, atenção aos detalhes e capacidade de comunicar resultados de forma clara.

**Resumindo**

* **Analista de Negócios:** Traduz necessidades do negócio em projetos de dados.
* **Desenvolvedor de BI:** Cria soluções para visualização e análise de dados.
* **Cientista de Dados:** Gera insights e modelos preditivos a partir de grandes volumes de dados.
* **Engenheiro de Dados:** Constrói e mantém a infraestrutura de dados.
* **Analista de Dados:** Analisa dados para apoiar decisões e identificar oportunidades.

### Aula 02 – Entendendo BI x Big Data

Business Intelligence (BI)

O BI refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte à gestão de negócios. O objetivo do BI é permitir uma fácil interpretação de grandes volumes de dados, transformando dados brutos em informações úteis para tomada de decisão. Ferramentas de BI, como Power BI, Tableau e Qlik, são usadas para criar dashboards, relatórios e análises que ajudam gestores a acompanhar indicadores de desempenho e identificar oportunidades de melhoria.

Resumo do BI:

* Foco em dados históricos e estruturados.
* Suporte à tomada de decisão tática e estratégica.
* Utiliza ferramentas de visualização e análise.
* Processos bem definidos e rotinas de análise.

Big Data

Já o Big Data é um termo da Tecnologia da Informação que trata sobre grandes conjuntos de dados que precisam ser processados e armazenados. O Big Data envolve dados em volume, variedade e velocidade muito superiores ao que o BI tradicional costuma lidar. Ele permite trabalhar com dados estruturados e não estruturados, vindos de diversas fontes (redes sociais, sensores, logs, etc.), e utiliza tecnologias como Hadoop, Spark e bancos NoSQL para processar e analisar esses dados em escala.

Resumo do Big Data:

* Foco em grandes volumes, variedade e velocidade de dados.
* Trabalha com dados estruturados e não estruturados.
* Utiliza tecnologias avançadas para processamento distribuído.
* Permite análises preditivas, identificação de padrões e insights em tempo real.

**Diferenças principais**

| **Aspecto** | **BI** | **Big Data** |
| --- | --- | --- |
| **Volume de dados** | **Moderado** | **Muito grande** |
| **Tipo de dados** | **Estruturados** | **Estruturados e não estruturados** |
| **Ferramentas** | **Power BI, Tableau, Qlik** | **Hadoop, Spark, NoSQL** |
| **Objetivo** | **Apoio à decisão, análise histórica** | **Insights, predição, análise em tempo real** |
| **Complexidade** | **Menor** | **Maior** |

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Caminho 1 – Armazenamento >> Processamento em lote

Caminho 2 – Busca de dados em tempo real >> Processamento em tempo real

### Aula 03 – Ferramentas usadas pelo Eng. de Dados | BigData

Ferramentas de Big Data da Azure

**Data Factory**

- Papilines

- Orquestradores

- ETL

**Azure Synapse Analytics**

- Armazenamento de dados analíticos e

- Grande escala

**Data Store**

- Data Lake

- Blobs

- Azure SQL (Banco de dados nas nuvens)

Entrando no portal.azure.com

Teams

O conteúdo gerado por IA pode estar incorreto.

Uma das ferramentas usada pelo engenheiro de dados é ARMAZENAMENTO:

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Outra ferramenta no Azure é o BANCO DE DADOS:

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

E outra ferramenta é ANÁLISE DE DADOS:

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

## Seção 3 – Arquitetura de um DW moderno

### Aula 01 – Mapa do curso

Duas partes nesse curso serão ministrados:

Análise de dados buscando dados do passado e analisando os dados para tomada de decisão.

Busca de dados para análise preditiva e tomada de decisões com grande quantidade de dados.

As tecnologias usadas são:

- Data Factory - Pipelines - Data Flow - Data lake

- SQL database - Pool SQL - Apache Spark Pool

### Aula 02 – Arquitetura moderna de um DW

Interface gráfica do usuário, Aplicativo, Teams

O conteúdo gerado por IA pode estar incorreto.

A primeira coisa a ser feita é ‘Ingerir’ os dados. É fazer com que os dados cheguem ao Data Lake. Tecnologia utilizada: **DATA FACTORY.**

Antigamente pagava-se os dados e os colocavam no DW, mas essa prática limitava as possibilidades de trabalho.

Os dados no **Data Lake** não vão ser trabalhados ou tratados. Aqui armazena-se os dados brutos.

Para colocar os dados no **DW** é preciso tratar e limpar os dados através da tecnologia **DATA FLOW.**

Uma vez os dados tratados e armazenados no DW, agora sim pode ser usado em tecnologias de análise, por exemplo, Power BI.

Texto

O conteúdo gerado por IA pode estar incorreto.

1° Passo – Criar o Data Lake. Local onde se guarda as informações brutas.

### Aula 03 – Prática | Ferramentas de um DW

1° Passo: Criar o DATA LAKE

No portal da Azure ir em >> Uma imagem contendo Ícone

O conteúdo gerado por IA pode estar incorreto.

>> Análises

>> Data Factorie Texto

O conteúdo gerado por IA pode estar incorreto.

**O que é o Data Factory da Azure?**

O **Azure Data Factory** é uma ferramenta de orquestração e integração de dados na nuvem. **Ela permite criar, gerenciar e automatizar pipelines de dados**, ou seja, fluxos que buscam, transformam, limpam e movem dados entre diferentes fontes e destinos. O objetivo é facilitar o processo de ETL (Extração, Transformação e Carga de dados), integrando sistemas diversos e preparando os dados para análise ou armazenamento.

**Principais funções do Data Factory:**

* **Ingestão de dados:** Buscar e coletar dados de várias fontes (bancos de dados, arquivos, APIs, etc.).
* **Orquestração:** Controlar o fluxo dos dados, definindo etapas e dependências entre tarefas.
* **Transformação:** Limpar, transformar e preparar os dados para uso, utilizando atividades como **Data Flow**.
* **Carga:** Enviar os dados tratados para destinos como Data Lake, Azure SQL ou outros sistemas de armazenamento.
* **Automação:** Agendar e monitorar pipelines para garantir que os dados estejam sempre atualizados.

No contexto do engenheiro de dados, **o Data Factory é essencial para garantir que os dados cheguem ao Data Lake**, sejam tratados e estejam prontos para análise em ferramentas como Power BI ou para armazenamento em bancos como Azure SQL.

**Exemplo prático do uso do Data Factory:**

1. **Buscar/Ingerir:** O Data Factory coleta dados brutos de diferentes fontes.
2. **Guardar:** Os dados são armazenados no Data Lake.
3. **Limpar:** Utiliza **Data Flow** para tratar e limpar os dados.
4. **DW:** Os dados tratados são enviados para o Azure SQL (Data Warehouse).
5. **Relatórios:** Os dados podem ser analisados em ferramentas como Power BI.

Texto

O conteúdo gerado por IA pode estar incorreto.Para iniciar a criação do Data Lake preciso criar o mecanismo que vai gerenciar tudo isso: **Data Factory.**

Portal Azure

>> Mais serviços

>> Análises

>> Data Factory (DF)

**Obs.: Todo DF, Data Lake (DL) e Banco SQL da Azure precisam ser guardados em um container chamado Grupo de Recursos.**

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Criação do Data Factory no Azure:

Tela de celular com aplicativo aberto

O conteúdo gerado por IA pode estar incorreto.

Obs.: Se apagar o Grupo de Recursos (Container), todos os recursos lá guardados serão também apagados.

Texto

O conteúdo gerado por IA pode estar incorreto.

Mesmo procedimento para criar Banco de dados e DL.

Toda vez que criar um BD é preciso armazená-lo no servidor.

Após criar o servidor é possível acessar esse servidor e o BD pela Microsoft SQL Server Management Studio para fazer as consultas de SQL.

## Seção 4 – Armazenamento na Nuvem

### Aula 01 – Armazenamento de dados no Azure

Há três tipos de dados para se trabalhar:  
Dados estruturados / Não estruturados / Semiestruturados

**Tipos de Dados**

**1. Dados Estruturados**

São dados organizados em um formato fixo, geralmente tabelas, com linhas e colunas bem definidas. Exemplos comuns **incluem bancos de dados relacionais (como SQL)**, onde cada campo tem um tipo e uma posição específica. Esses dados são facilmente pesquisáveis e processados por sistemas tradicionais de BI (Business Intelligence).

**Exemplo:**

* Planilhas do Excel
* Tabelas SQL

**2. Dados Não Estruturados**

São dados que não seguem um modelo ou estrutura fixa. Eles podem ser textos livres, imagens, vídeos, áudios, e-mails, posts em redes sociais, etc. Esses dados exigem técnicas avançadas para extração de informações, pois não possuem um padrão claro de organização. No Azure, esses dados são armazenados no **BLOB STORAGE** ou **DATA LAKE.**

**Exemplo:**

* E-mails
* Documentos de texto
* Fotos e vídeos

**3. Dados Semiestruturados**

Ficam entre os dois anteriores: possuem alguma organização, mas não tão rígida quanto os dados estruturados. Normalmente, usam marcadores ou tags para separar elementos, como arquivos JSON, XML ou logs de sistemas. Eles facilitam a identificação de informações, mas não têm a rigidez de um banco de dados relacional. Armazenado em **AZURE COSMO DB**, no Azure.

**Exemplo:**

* Arquivos JSON
* Arquivos XML
* Logs de aplicações

### Aula 02 – Blob Storage e Data Lake

São locais de armazenamento de dados não estruturados.

**Blob Storage no Azure**

O **Blob Storage** é um serviço de armazenamento de objetos na nuvem do Azure. Ele é ideal para guardar grandes volumes de dados não estruturados, como documentos, imagens, vídeos, backups, arquivos de log, entre outros.

* **Blob** significa “Binary Large **Object**”, ou seja, qualquer arquivo digital grande.
* No contexto de engenharia de dados, o Blob Storage é utilizado para armazenar dados que não seguem uma estrutura fixa, como textos livres, fotos, vídeos, etc.
* É altamente escalável, seguro e permite o acesso rápido aos dados para diversas aplicações.
* O Blob Storage foi projetado para ser um repositório de dados. Não foi projetado para análise dos dados.
* Não é possível criar sub containers para armazenamentos.

**Data Lake no Azure**

O **Data Lake** é uma solução de armazenamento voltada para dados em grande escala, sejam eles estruturados, não estruturados ou semiestruturados.

* O Data Lake permite guardar dados brutos, sem a necessidade de tratamento ou organização prévia.
* É uma solução de armazenamento projetada para análise de big data corporativa.
* É muito utilizado em projetos de Big Data, pois facilita o armazenamento de dados vindos de diversas fontes (bancos, sensores, redes sociais, logs, etc.).
* No Azure, o Data Lake é integrado com outras ferramentas, como o Data Factory, que faz a ingestão dos dados, e o Data Flow, que trata e limpa os dados antes de enviá-los para análise ou para um Data Warehouse (DW).
* Já no Data Lake, diferente do que acontece no Blob Storage, **há um sistema de arquivos hierárquicos.**
* É possível a análise de dados no Data Lake.

**Diferenças e Usos**

* **Blob Storage:**
  + Foco em dados não estruturados.
  + Armazenamento simples e direto.
  + Usado para guardar arquivos, imagens, vídeos, etc.
* **Data Lake:**
  + Foco em grandes volumes de dados de diferentes tipos.
  + Armazena dados brutos para posterior processamento.
  + Facilita projetos de Big Data e análises avançadas.

### Aula 03 – Prática / Ferramentas de Armazenamento

Aula prática para criação de Grupo de recurso, BD e integração com SQL Server.

**1° Passo** – Primeiramente criar o Grupo de Recursos DevZenitEduc.

**2° Passo** – Criar ‘Conta de Armazenamento’ para criar o Data Lake.

Conta de armazenamento: **devzenithdeca**

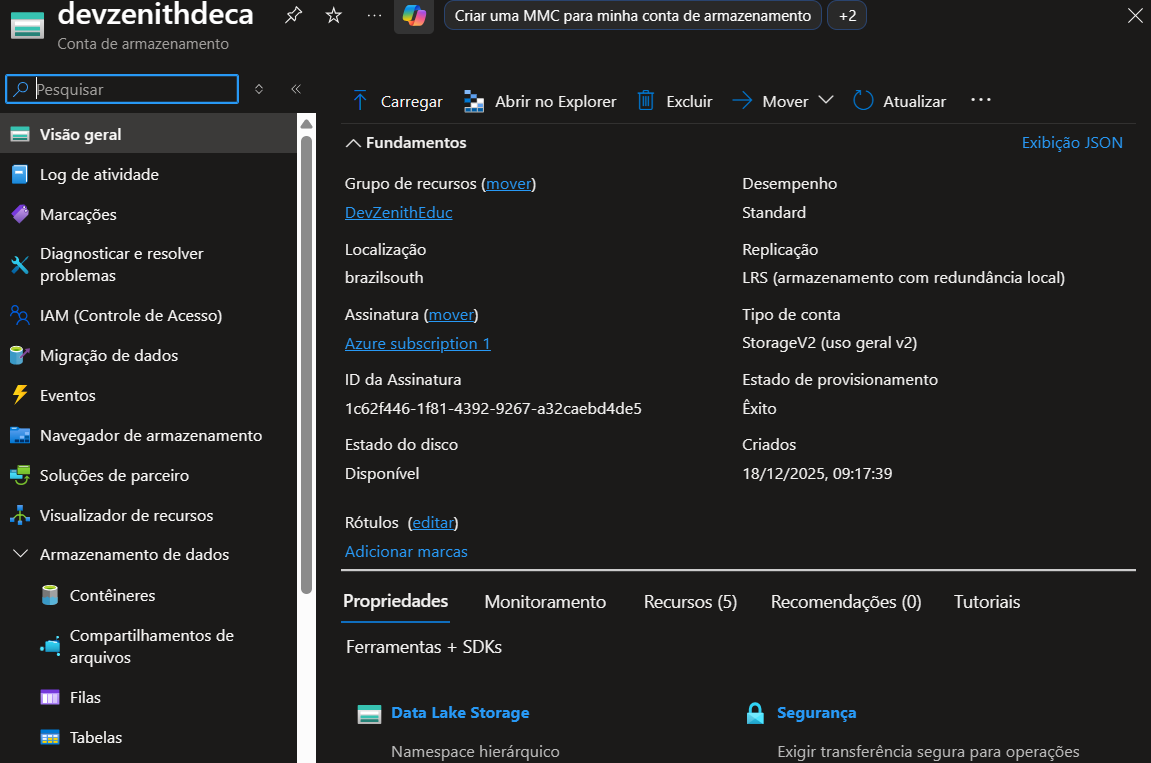
**3° Passo** – Na aba ‘Avançado’ marcar a opção ‘Habilitar namespace hierárquico’. Essa marcação habilita a hierarquia de pasta do Data Lake. É o que diferencia do Blobs Storage.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.



Como foi criado um Data Lake, há a hierarquização dos containers no armazenamento da Conta de armazenamento:





**4° Passo** – Criar um container dentro da ‘Conta de armazenamento’

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

**5° Passo** – Uma vez criado o container, podemos criar as pastas e as hierarquizações para o projeto. Nome do container criado: **devzenith1**

Tela de computador com texto preto sobre fundo branco

O conteúdo gerado por IA pode estar incorreto.

Finalizada a criação do Data Lake.

## Seção 5 – Azure Data Factory

### Aula 1 – Pré-requisitos para as sessões 5 e 6

O Azure Data Factory é uma das ferramentas do Azure e trás grandes feitos para o Analista de Dados e BI.

O Data Factory >> Busca dados (Extração de dados).

Data Flow (que está dentro do DF) >> Transforma e limpa os dados.

O Data Factory também carrega os dados (Load).

>> Tudo isso é o ingest << Extrair e carregar o dado já tratado.

PASSO IMPORTANTE

Na estrutura de ETL dos dados é preciso compreender a dinâmica total do cominho percorrido pelos dados.

Extração (E) >> Entrada dos dados. É preciso ter um **conector** na entrada para as entradas de dados.

Load (L)>> Saída dos dados. Aqui também precisa de um **conector** para a respectiva saída.

O Azure Data Factory, em sua essência, trabalha e funciona com conexões.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

### Aula 2 – Conhecendo o Azure Data Factory (ADF)

Diagrama

O conteúdo gerado por IA pode estar incorreto.

É um serviço de **integração de dados** e **ETL** baseado em **nuvem** que permite criar fluxos de trabalho orientados a dados para **orquestrar a movimentação de dados** e transformá-los em escala.

Ferramentas:

* Pipelines
* Activities
* Datasets
* Linked Services
* Data Flows
* Integration Runtimes

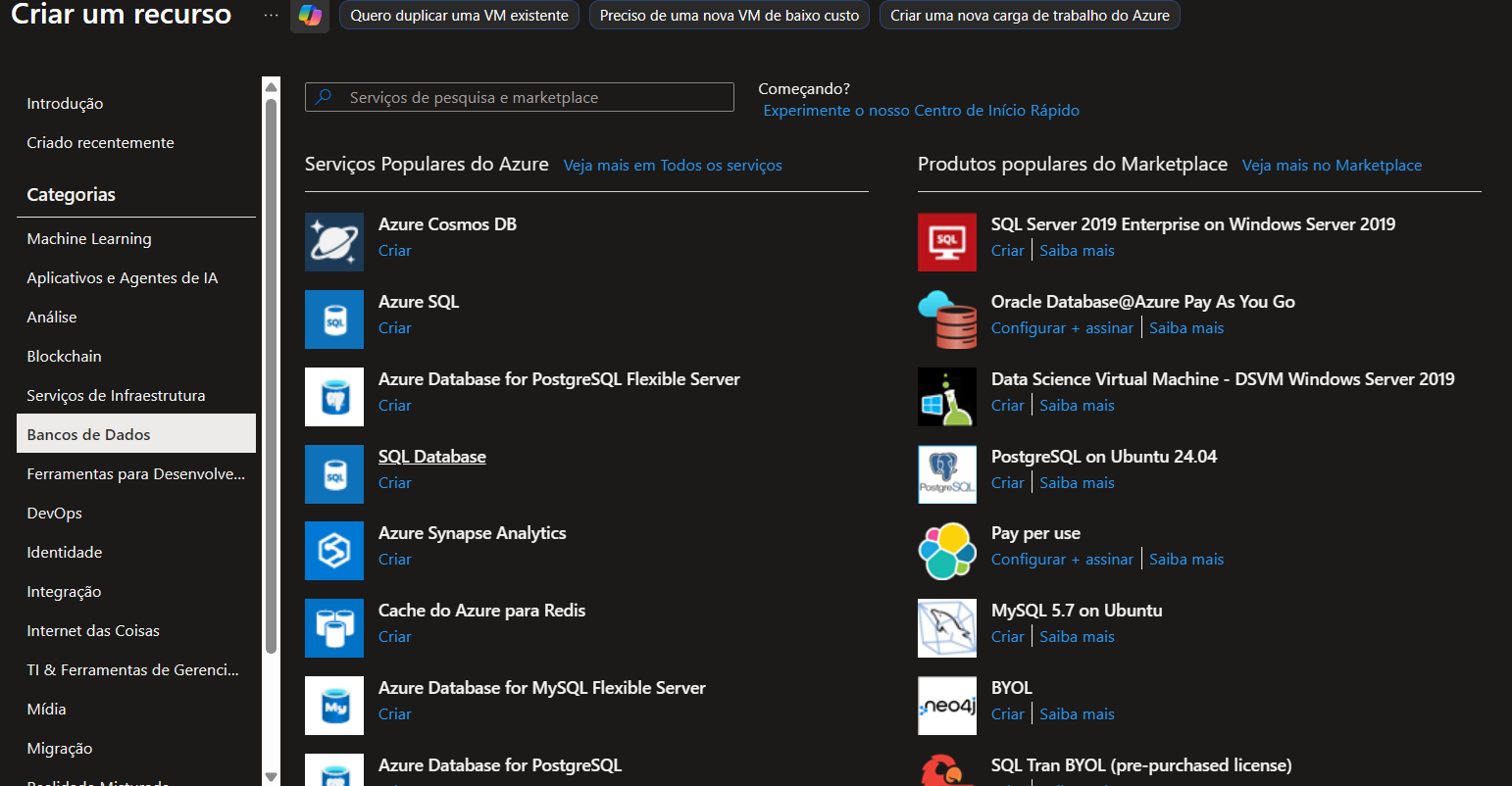
Interface gráfica do usuário, Aplicativo, Teams

O conteúdo gerado por IA pode estar incorreto.

Caminho total dos dados. Da extração à Apresentação.

### Aula 3 – Prática 1 – Criando o BD Azure SQL

Criação do Banco de Dados para estocagem dos dados.





Criar um Banco de dados SQL Database.

Servidor criado para o BD: devzenithserver (devzenithserver.database.windows.net)

Login do administrador do Servidor: DevZenith8522407

Senha: padrão Windows

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Primeiro foi criado do servidor. Logo após criado o BD dentro do servidor.

Feito o teste de integração com o SQL Server Local com o BD criado no Azure

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

A conexão também é válida através do Azure data Studio (ADS)

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

### Aula 4 – Prática 2 – Criando ADF

Criação data Factory.  
Criação simples e conhecendo o ambiente do DF.

### Aula 5 – Componentes do Azure Data Factory

Ferramentas que compõem o Data Factory

* Pipelines
* Activities
* Datasets
* Linked Services
* Data Flows

Linha do tempo

O conteúdo gerado por IA pode estar incorreto.

As ferramentas do DF estão interligadas.

Papiline

É um conjunto de atividades lógicas cuja atividade é pegar um dado do BD SQL e colocar em um CSV, por exemplo.

Permite também orquestração, monitoramento, criação e consumo de **conjunto de dados**(Dataset). Em uma atividade pode-se criar também arquivos e dados.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

### Aula 6 – Prática 3 – Pipelines, Linked Services, Dataset e Activite

Criação de uma Pipeline para transferir um dado que está na web para dentro de um BD na Azure.

1° Passo – Abir o Estúdio do Azure Data Factory:

Tela de celular com publicação numa rede social

O conteúdo gerado por IA pode estar incorreto.



Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Um arquivo CSV está na WEB e o objetivo é buscá-lo e colocar os dados em uma tabela que está dentro de um BD o qual está em um servidor SQL Azure.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Serviço Vinculado e Dataset da origem.

Serviço Vinculado e Dataset de destino.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

1° Passo - Criar o Serviço Vinculado da Origem

Gerenciar

>> Serviços Vinculados

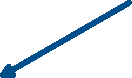
É o serviço vinculado de origem onde estão os dados de coleta.

2° Passo – Criar o conjunto de dados relacionados à origem

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.



O conjunto de dados (Dataset) pode ser criado diretamente na Pipeline:  
Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.



Após configurar o Conjunto de dados vai parecer do lado esquerdo a tabela solicitada:  
Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.Quando aparecer, basta arrastar para dentro da pasta criada (DevZenitCD).



Na criação da papiline vai aparecer a opção ‘Visualizar Dados’Clicando aqui, vai aparecer o conteúdo da tabela solicitada:

Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

3° Passo – Criar o Serviço vinculado para o Destino

4° Passo – Criar Dataset para o destino

### Aula 7 – Script tbMovies (BDMovies)

## Seção 6 – Data Flow

## Seção 7 – Big Data

## Seção 8 – Data Warehouse Cloud

## Seção 9 – Pool de SQL

## Seção 10 – Apache Spark

## Seção 11 – Considerações Finais