



DOCK DATA HACKATHON

www.dockdatahackathon.com.br

PATROCÍNIO:

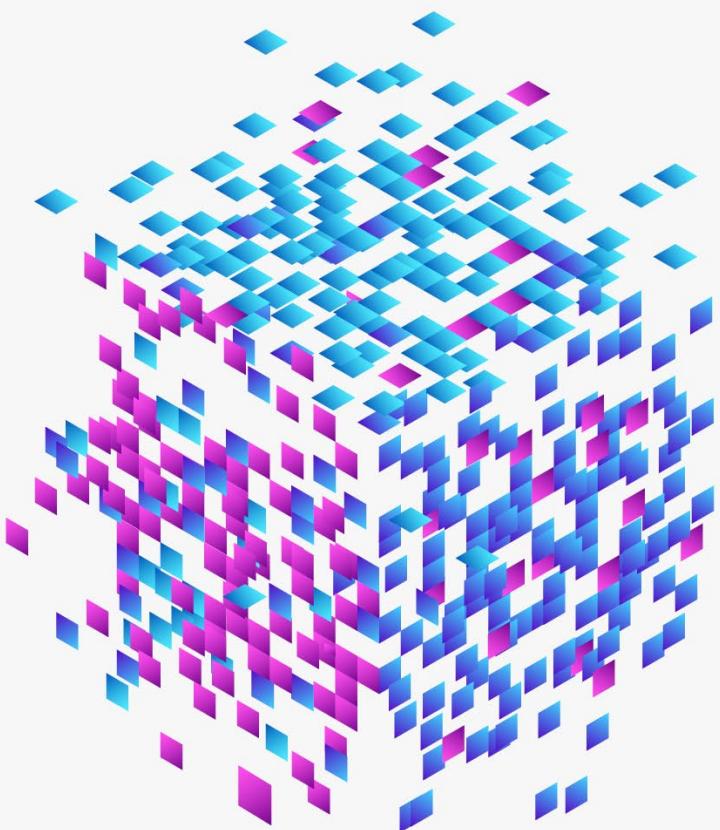


REALIZAÇÃO:

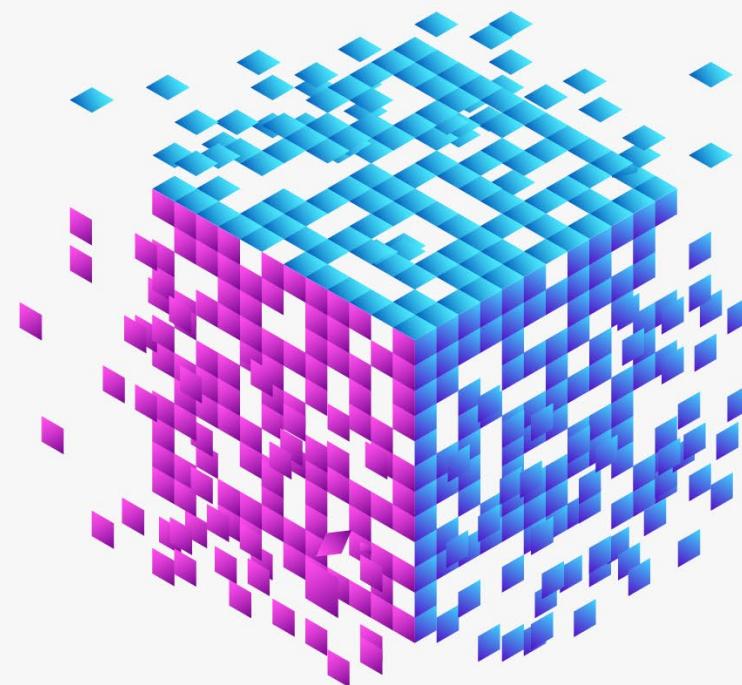
e.SethCloud

Azure
Academy
.com.br

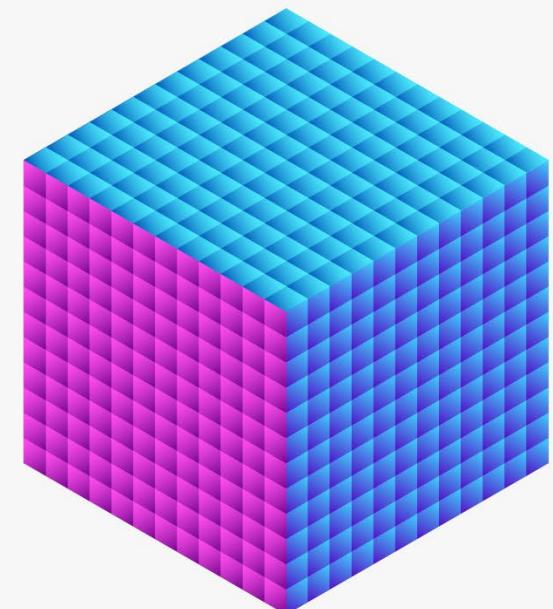
BIG DATA



ANALYTICS

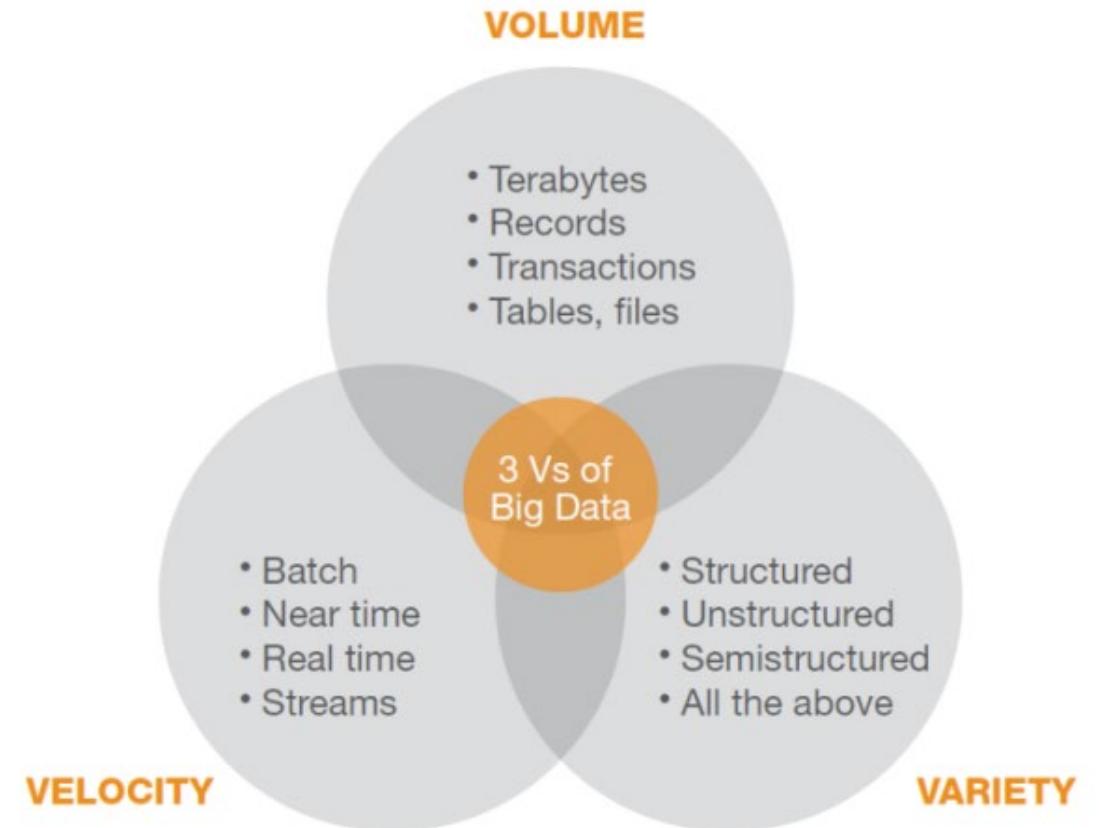
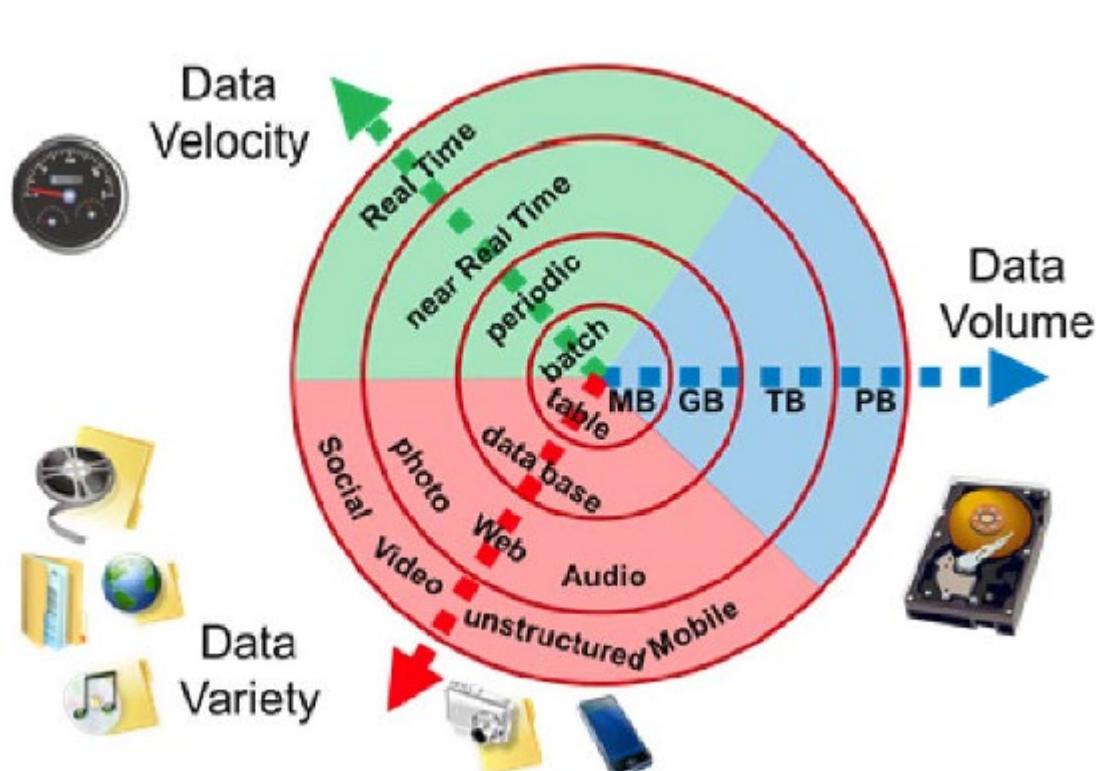


DECISIONS



BIG DATA

A ARQUITETURA EM 3Vs AJUDA EQUIPES A ATUAREM EM CAMADAS DE FORMA ORGANIZADA.



PATROCÍNIO:



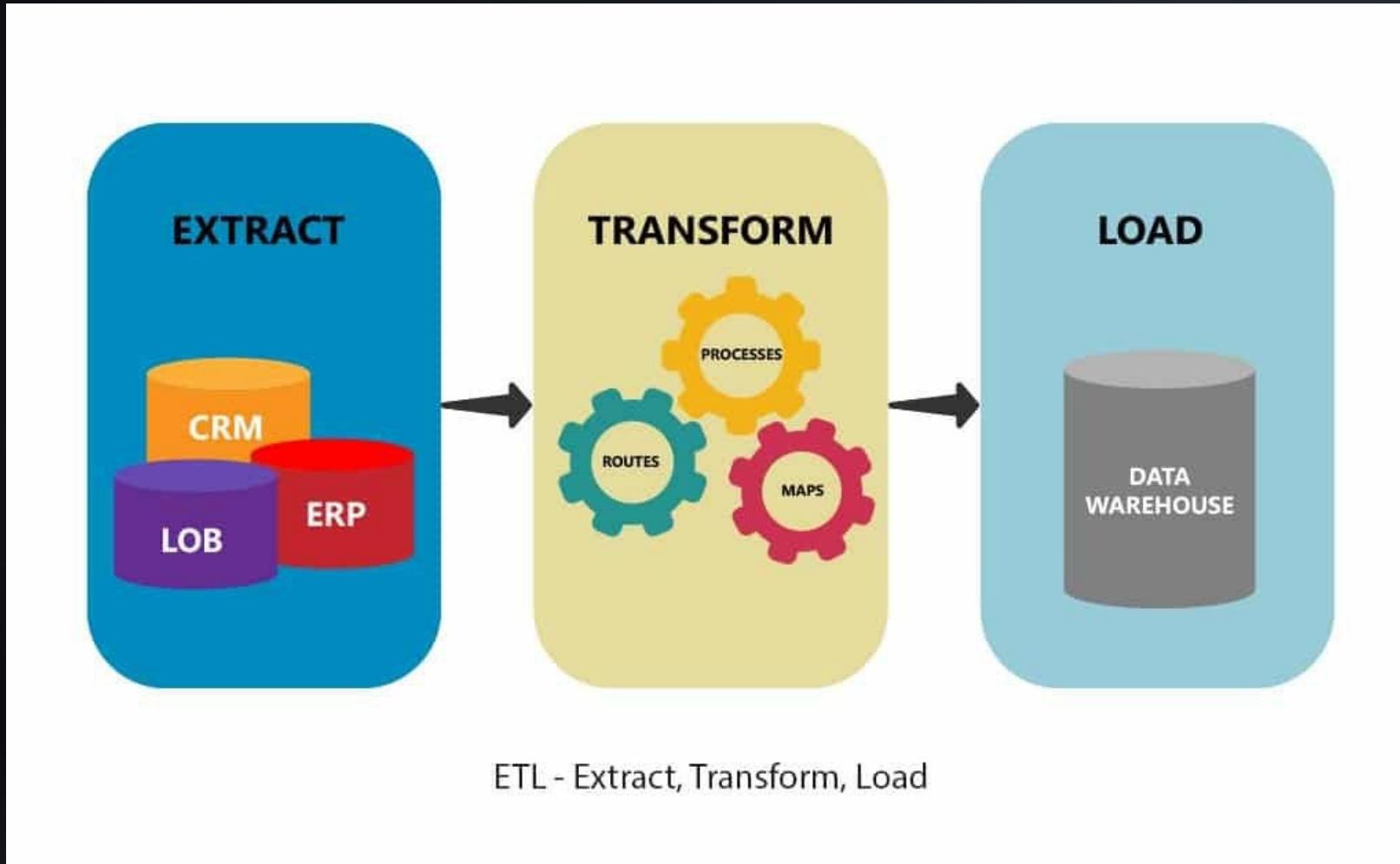
REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

ETL

ATUA NAS CAMADAS INICIAIS DO BIG DATA PARA CENTRALIZAR E NORMALIZAR OS DADOS.



PATROCÍNIO:

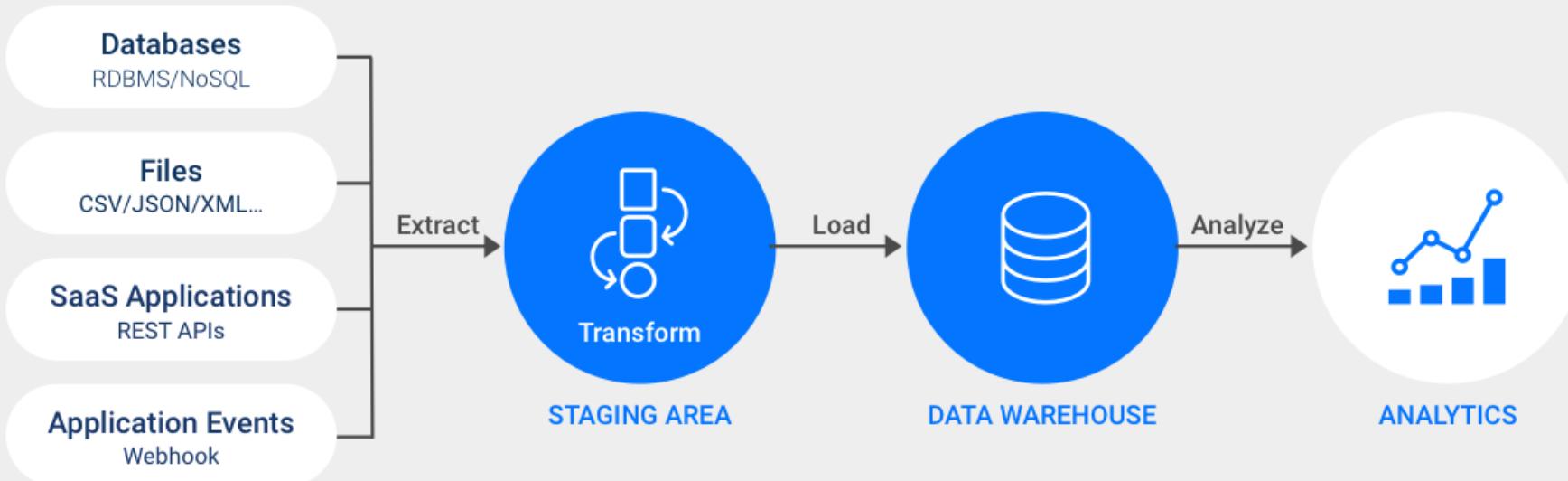


REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

ETL PROCESS



PATROCÍNIO:

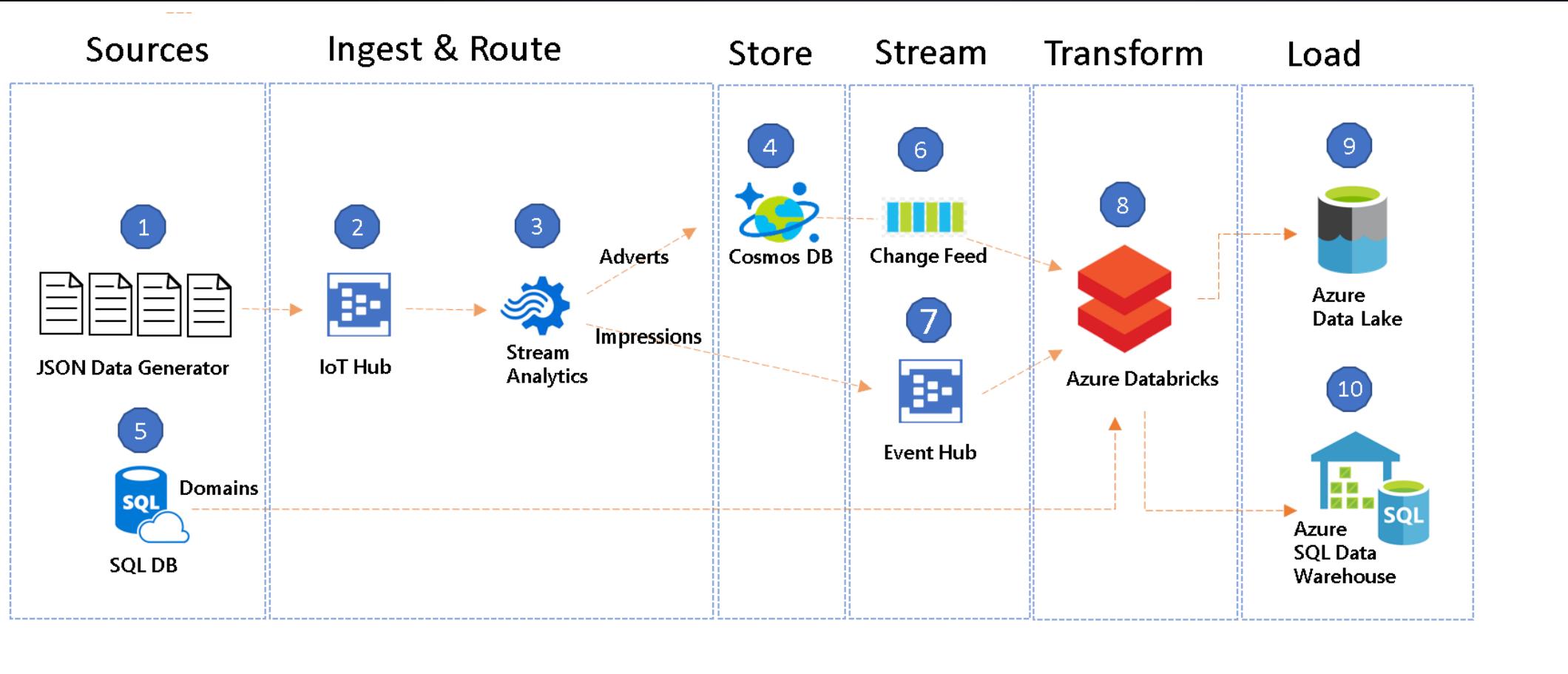


REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

ETL



PATROCÍNIO:



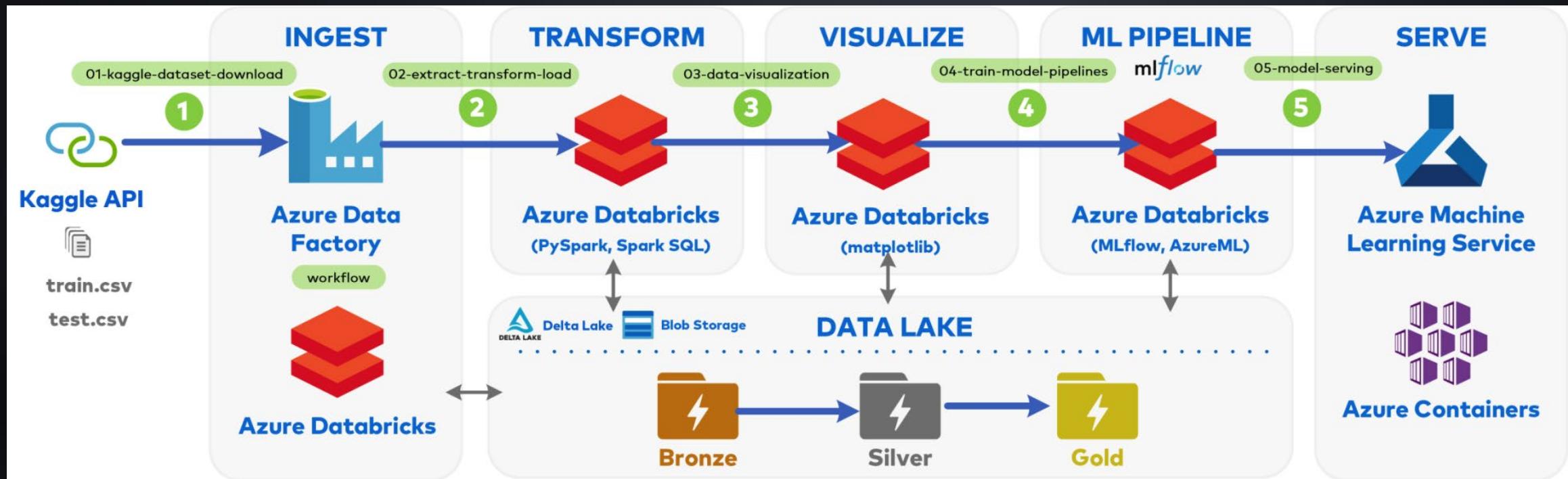
REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

CONCEITOS DE ESTEIRAS E ORQUESTRAÇÃO

Compreenda o potencial de cada serviço PaaS para compor esteiras de ingestão, tratamento, Machine learning e demandas analíticas.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

DOCK
DATA
HACKATHON

DATA LAKE

PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

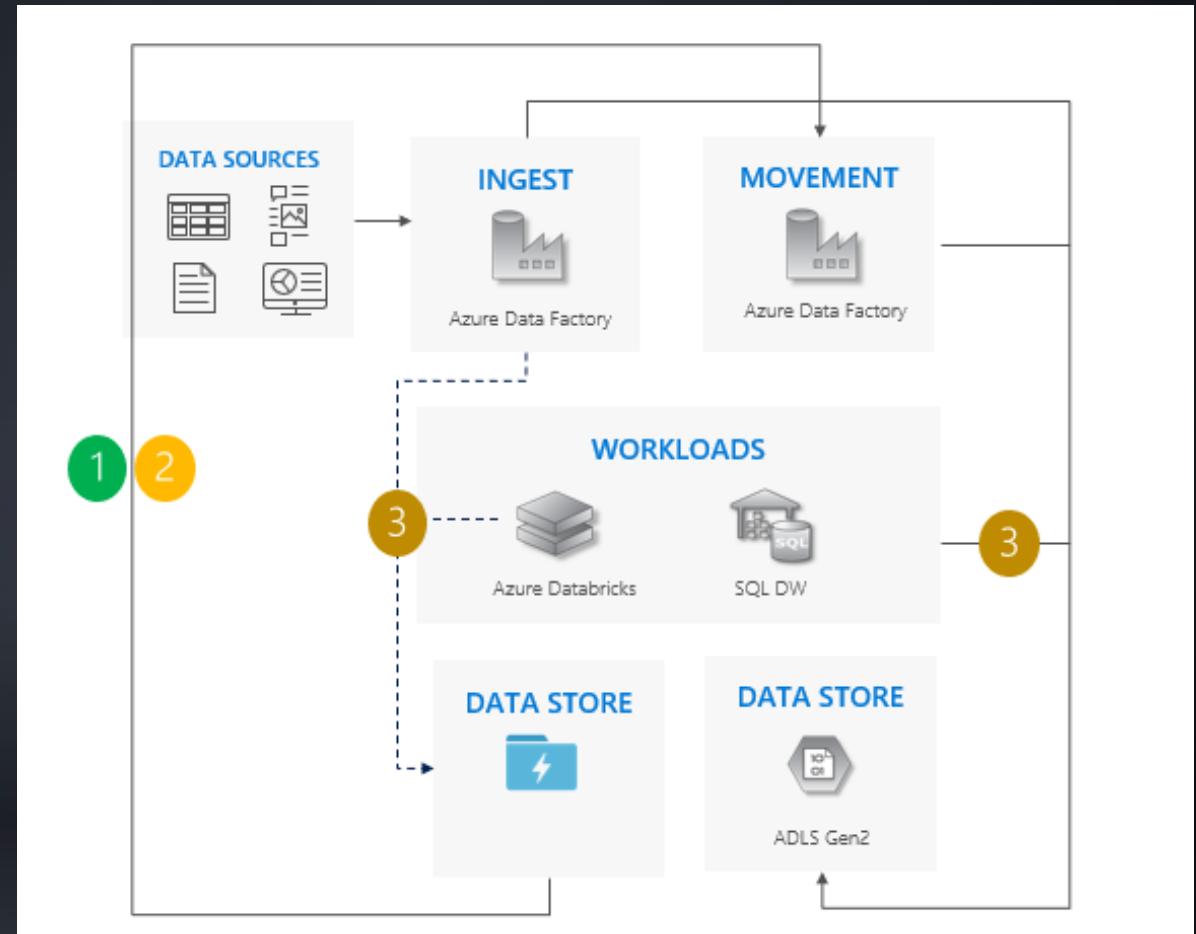
Azure
Academy
.com.br

DATA LAKE

- Storage preparado para BIG Data.
- No Azure pode ser ativado como serviço através da **Conta de Armazenamento**.
- Organizado por contêineres.
- Fácil integração aos demais serviços no Azure.

Acesse:

[Data Lake Store para análise de Big Data | Microsoft Azure](#)



PATROCÍNIO:



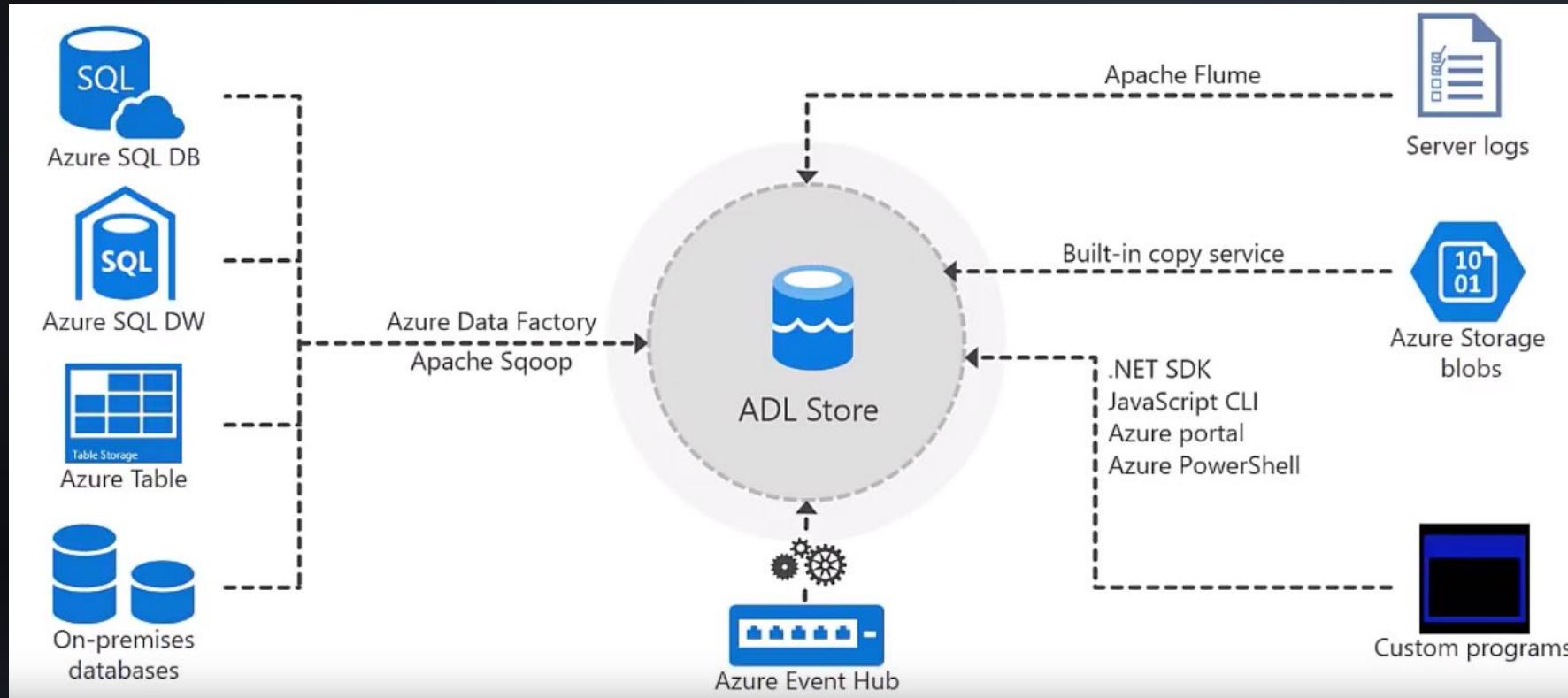
REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

DATA LAKE PAAS NO AZURE = CONTA DE ARMAZENAMENTO

- Storage suporta N formatos de arquivos.
- Permite armazenar arquivos de petabytes de tamanho e trilhões de objetos.
- Implantação rápida com pagamento por armazenamento.



PATROCÍNIO:



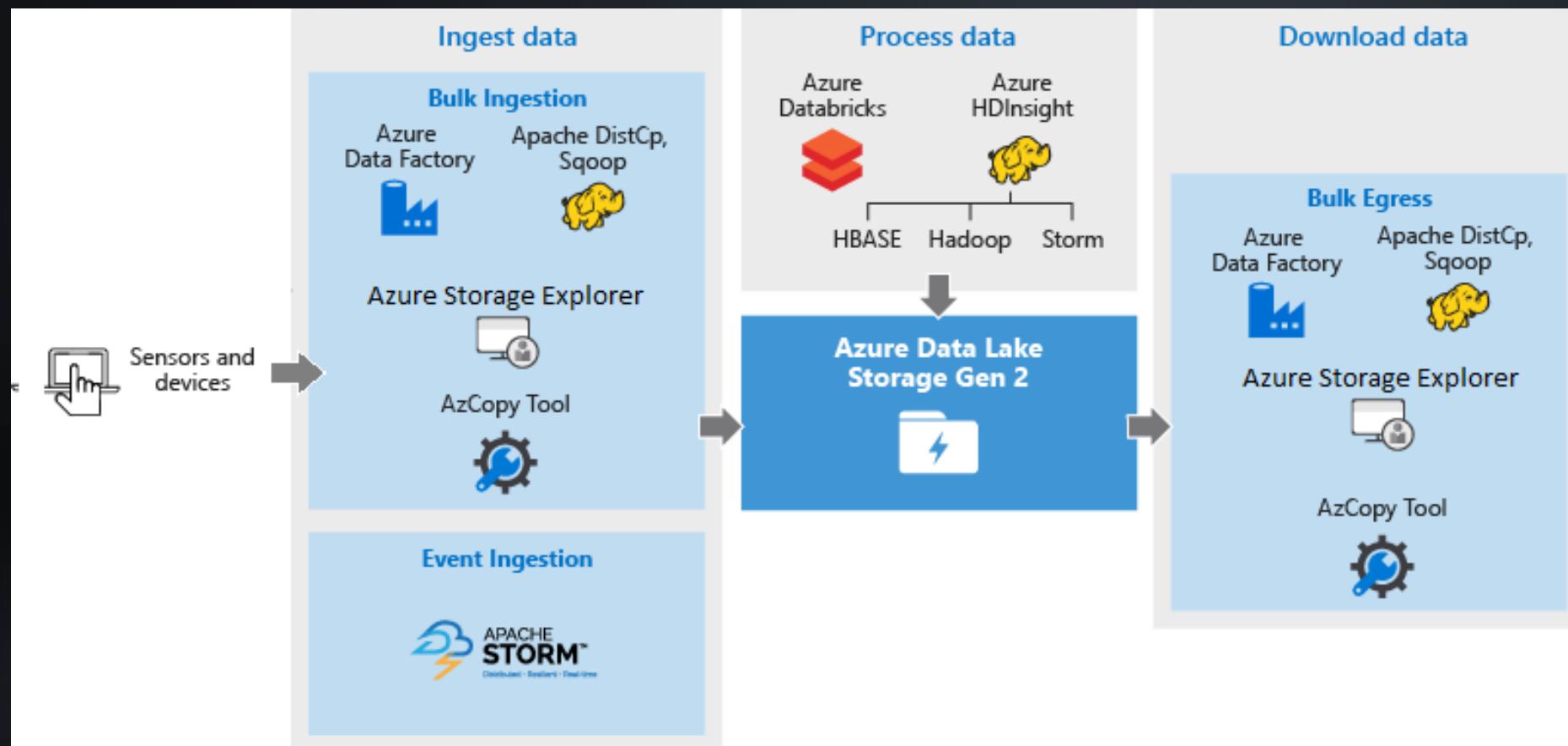
REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

STORAGE / STAGING

Staging área é o armazenamento intermediário / temporário em um processo de orquestração de dados. Pode ser utilizado para centralizar dados de diversas fontes após a ingestão e permitir que um serviço de processamento evolua o processo a partir dele.



REALIZAÇÃO:

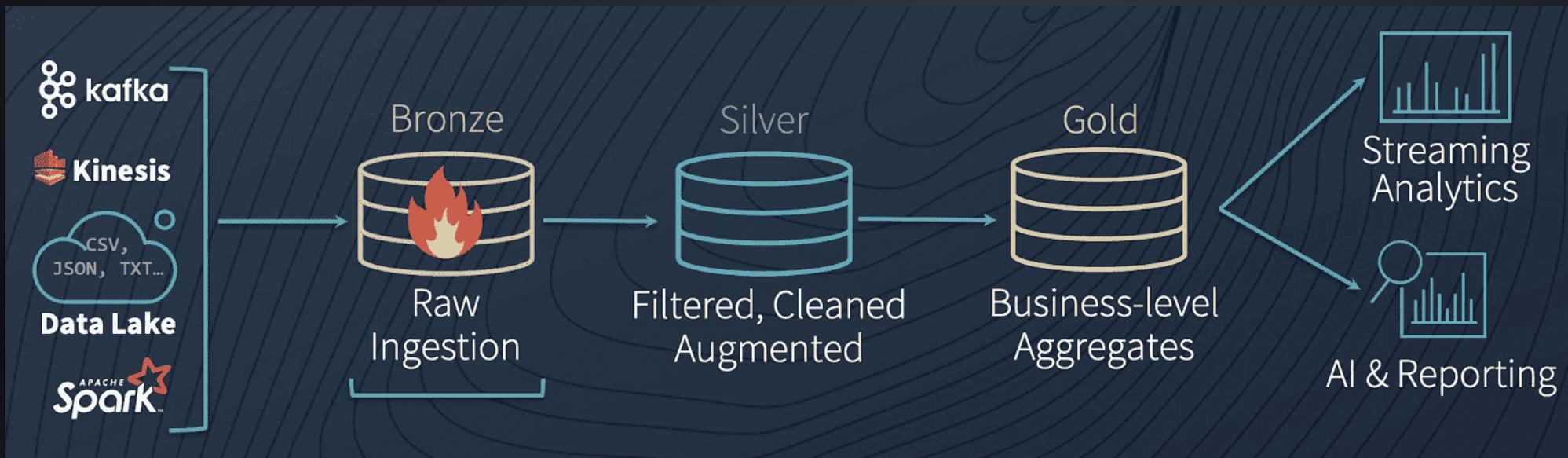
e.SethCloud

Azure
Academy
.com.br

ORQUESTRAÇÃO E CONVERSÃO DE FORMATOS

A orquestração de dados possui objetivos como:

- CENTRALIZAR DADOS DE DIVERSAS FONTES.
- CONVERSÃO DE N FORMATOS EM UM ÚNICO PADRONIZADO.
- ORGANIZAR OS DADOS PARA TRABALHO DO ENGENHEIRO, CIENTISTA E ANALISTA.
- PREPARO PRÉ PROCESSAMENTO.
- MIGRAÇÕES.
- DESCARTE E TRATAMENTO.



PATROCÍNIO:



REALIZAÇÃO:

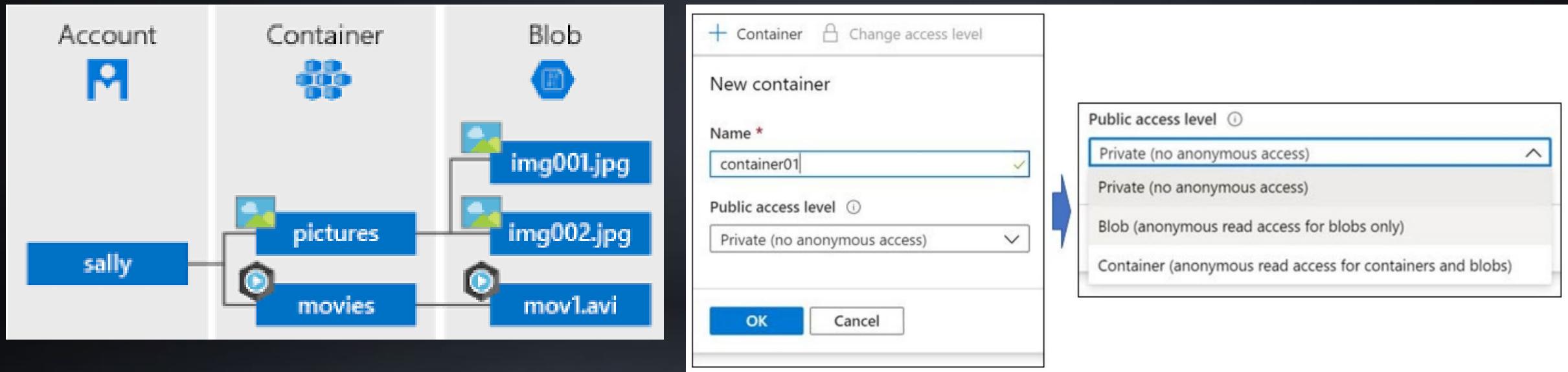
e.SethCloud

Azure
Academy

www.azureacademy.com.br

CONTAINERS X BLOBS

- Arquivos são blobs armazenados em containers.
- Permissões por contêineres.
- Possibilidade de integração aos demais serviços no Azure ou soluções híbridas.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

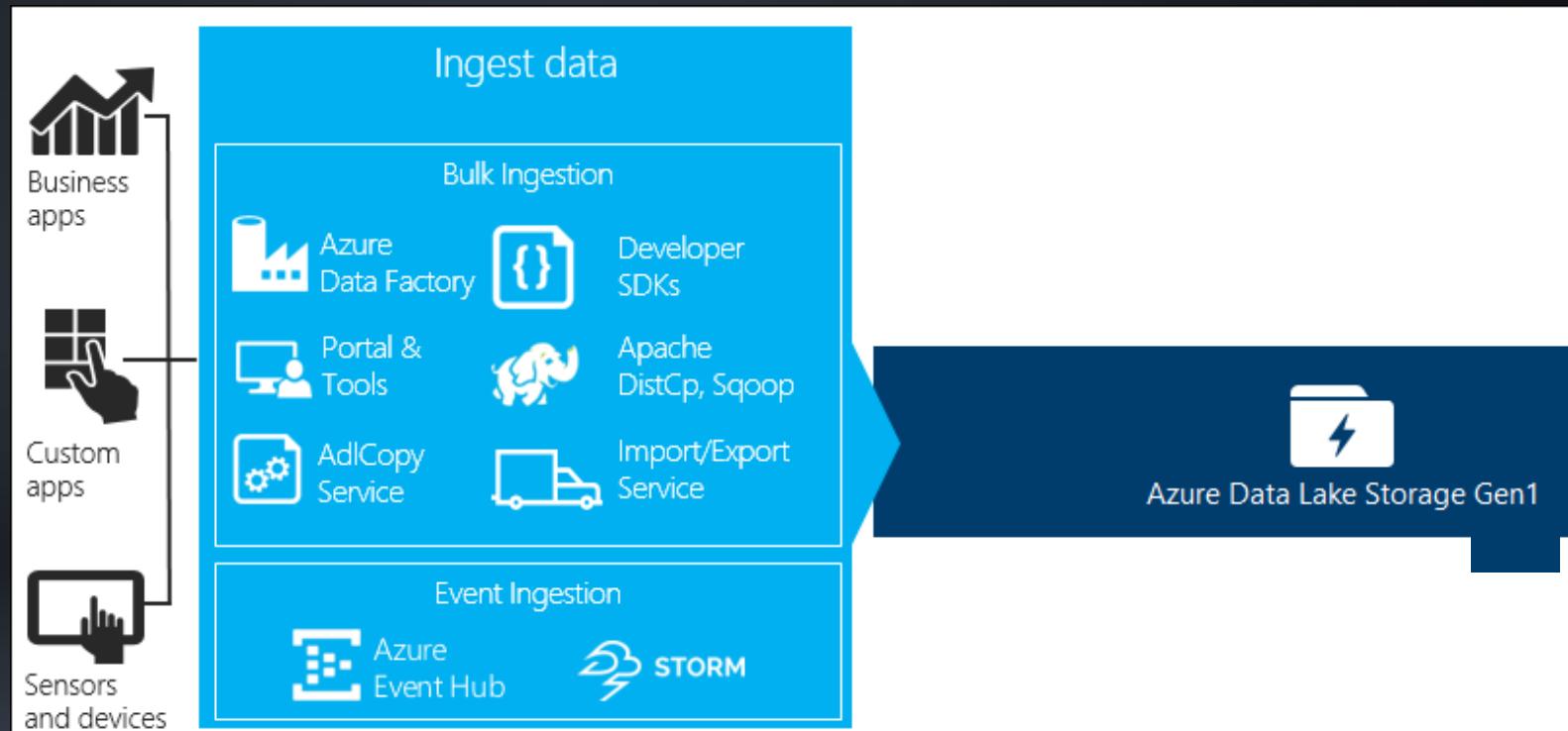
Azure
Academy

www.azureacademy.com.br

DATA LAKE STORAGE GEN 2

- Sem limite de storage.
- Adaptado para BIG Data.
- Arquitetura HDFS.
- Alto desempenho e seguro (ACL, AD, Firewalls, Criptografia).

O Azure Data Lake Storage Gen2 permite gerenciar e acessar dados como você faria com um Sistema de Arquivos Distribuído do Hadoop (HDFS). Possui driver disponível em ambientes Apache Hadoop, incluindo Azure HDInsight, Azure Databricks e Azure Synapse Analytics para acessar dados armazenados no Data Lake Storage Gen2



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

FORMATOS

CSV é mais rápido para escrever, o JSON o mais fácil de ser entendido por um ser humano e o Parquet o mais rápido de ler.

CSV é o padrão de fato para muitos dados; fácil de compreender para usuários e computadores e tornou-se mais acessível através do Excel. Muitos sistemas comprehendem CSV.

JSON é o padrão para comunicação na web. APIs e sites estão se comunicando constantemente usando JSON por causa de suas propriedades de usabilidade, como esquemas personalizáveis.

Parquet é otimizado para o paradigma Write Once Read Many (WORM). É lento para escrever, mas incrivelmente rápido para ler, especialmente quando você está acessando apenas um subconjunto das colunas totais. Para casos de uso que requerem operação em linhas inteiras de dados, um formato como CSV, JSON ou mesmo AVRO deve ser usado.

Manipular dados Parquet:

[Parquet Files - Spark 3.1.2 Documentation \(apache.org\)](#)

Manipular dados Json:

[JSON Files - Spark 3.1.2 Documentation \(apache.org\)](#)

Spark Format Showdown		File Format		
		CSV	JSON	Parquet
A t t r i b u t e	Columnar	No	No	Yes
	Compressable	Yes	Yes	Yes
	Splittable	Yes*	Yes**	Yes
	Human Readable	Yes	Yes	No
	Nestable	No	Yes	Yes
	Complex Data Structures	No	Yes	Yes
	Default Schema: Named columns	Manual	Automatic (full read)	Automatic (instant)
	Default Schema: Data Types	Manual (full read)	Automatic (full read)	Automatic (instant)

PATROCÍNIO:



REALIZAÇÃO:

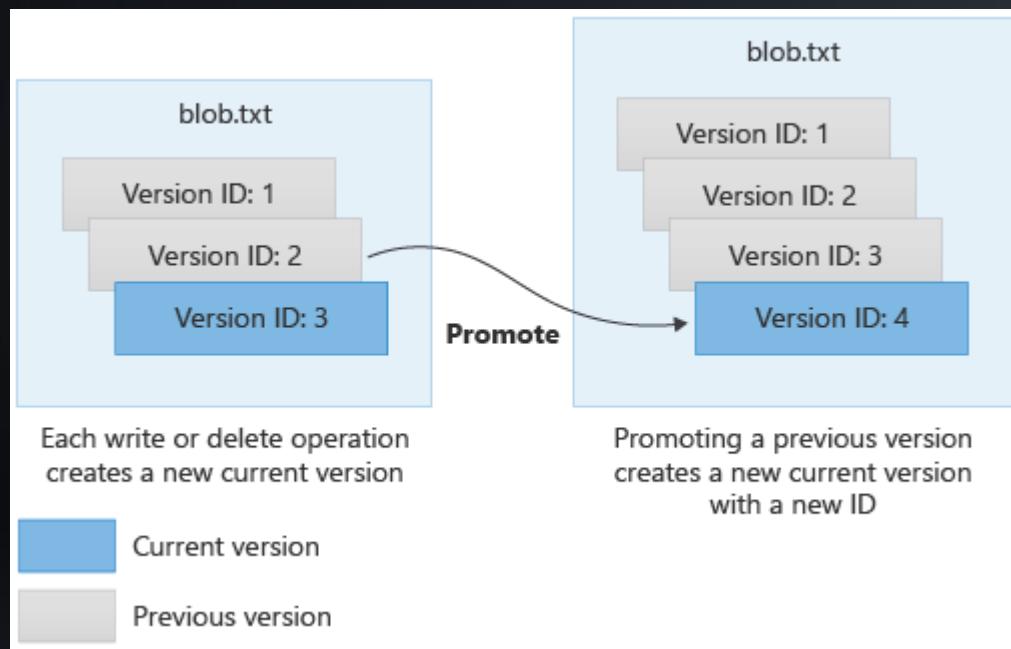
e.SethCloud

Azure
Academy

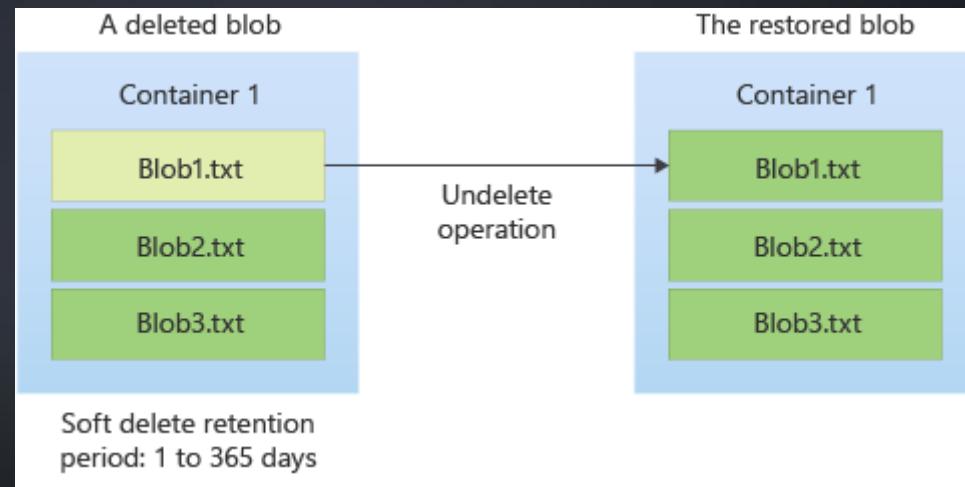
www.azureacademy.com.br

Controle de versão e Exclusão reversível

Quando o controle de versão do blob está habilitado para uma conta de armazenamento, o armazenamento do Azure cria automaticamente uma nova versão de um blob cada vez que o blob é modificado ou excluído.



Quando a exclusão reversível de BLOBs está habilitada em uma conta de armazenamento, você pode recuperar objetos depois que eles tiverem sido excluídos, dentro do período de retenção de dados especificado.



PATROCÍNIO:



REALIZAÇÃO:

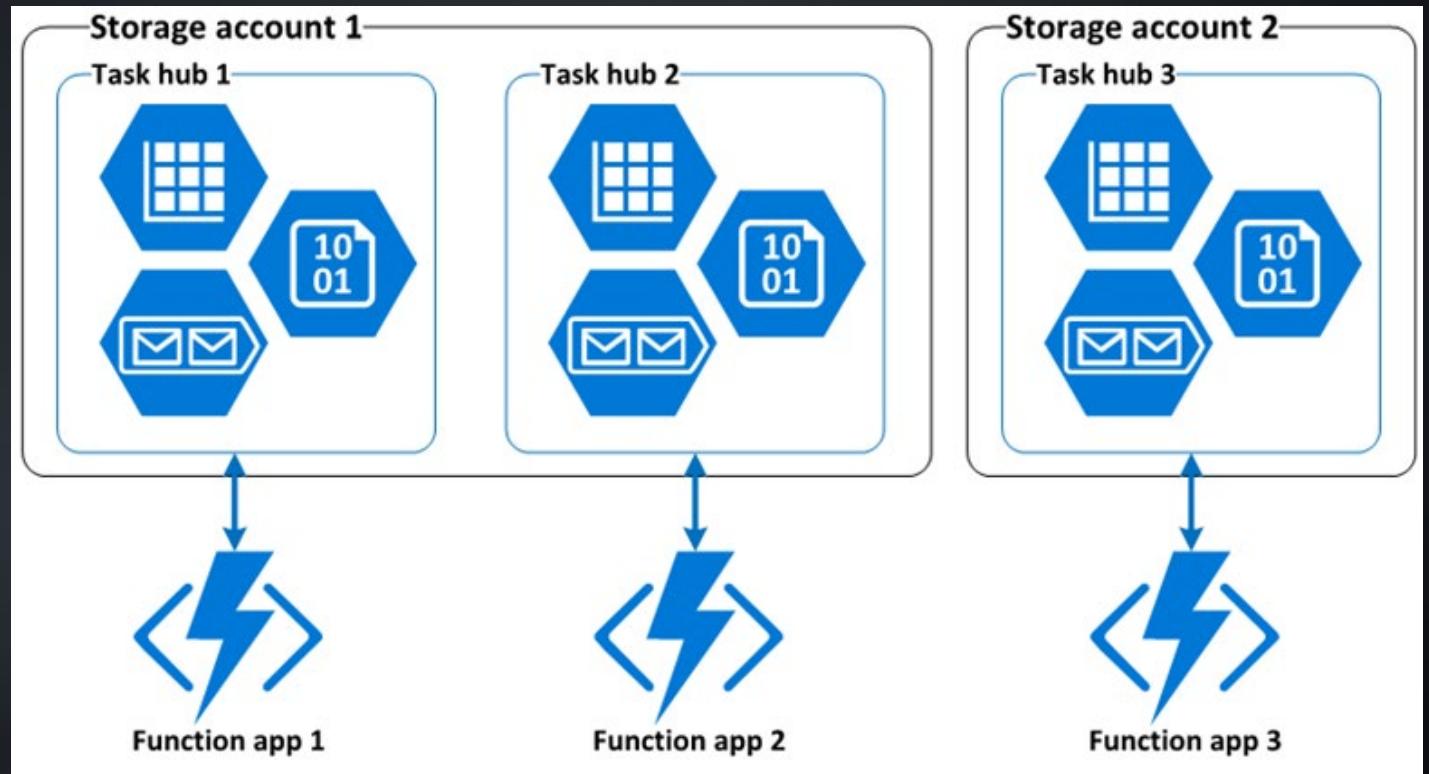
e.SethCloud

Azure
Academy

www.azureacademy.com.br

EVENTOS

Você pode associar **funções** (micro serviços) aos eventos da Storage Account para, por exemplo, tratar dados assim que forem armazenados no container, gerar miniaturas, descartar arquivos em formatos não permitidos, percorrer e validar arquivos CSV, Json e muitas outras possibilidades. As funções podem ser programadas em Python, C# e outras linguagens.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br



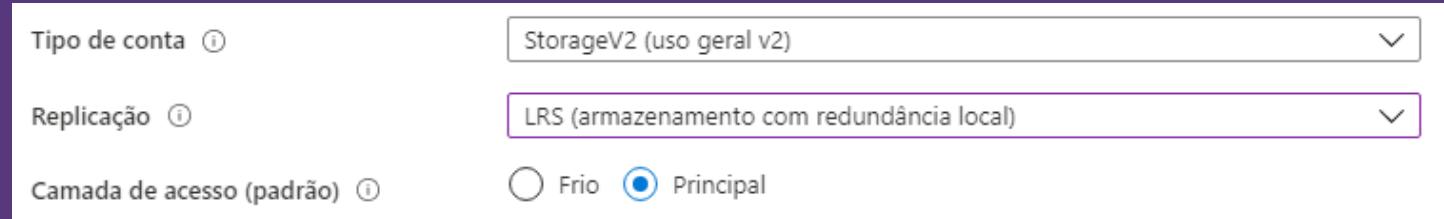
Laboratório

Preparando o Data Lake no Azure.

LAB - DATA LAKE STORAGE Gen2

Siga as etapas a seguir para instalar o Data Lake Storage Gen2:

1. Acesse o Marketplace do Azure e instale uma nova Conta de armazenamento (Storage Account).



2. Na guia Avançado, habilite o 'Data Lake Storage Gen2'.



3. Após a instalação, acesse o serviço e crie os containers para o seu projeto. Exemplos: Raw, Silver e Gold.

DOCK
DATA
HACKATHON

AZURE DATA FACTORY

PATROCÍNIO:



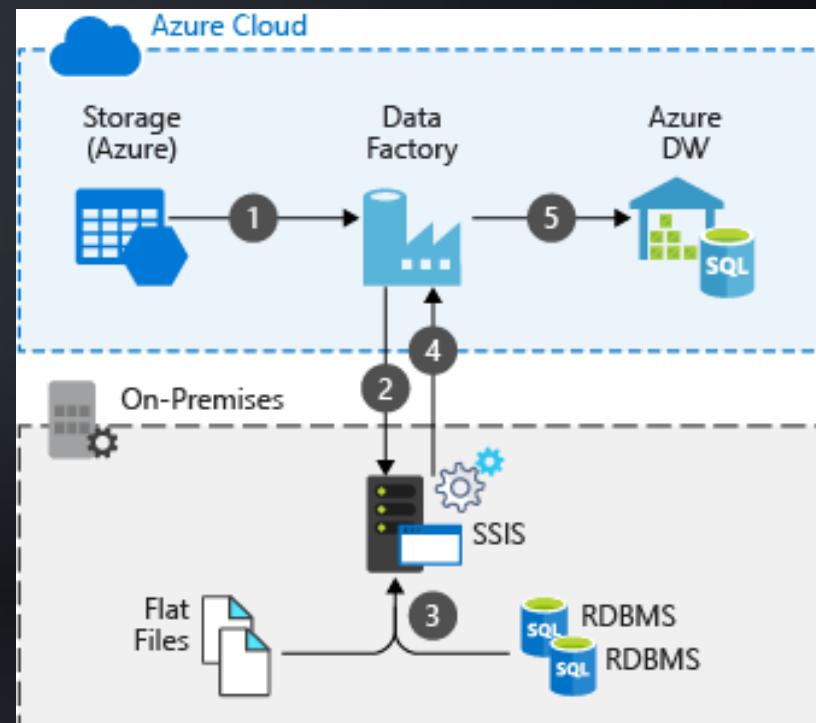
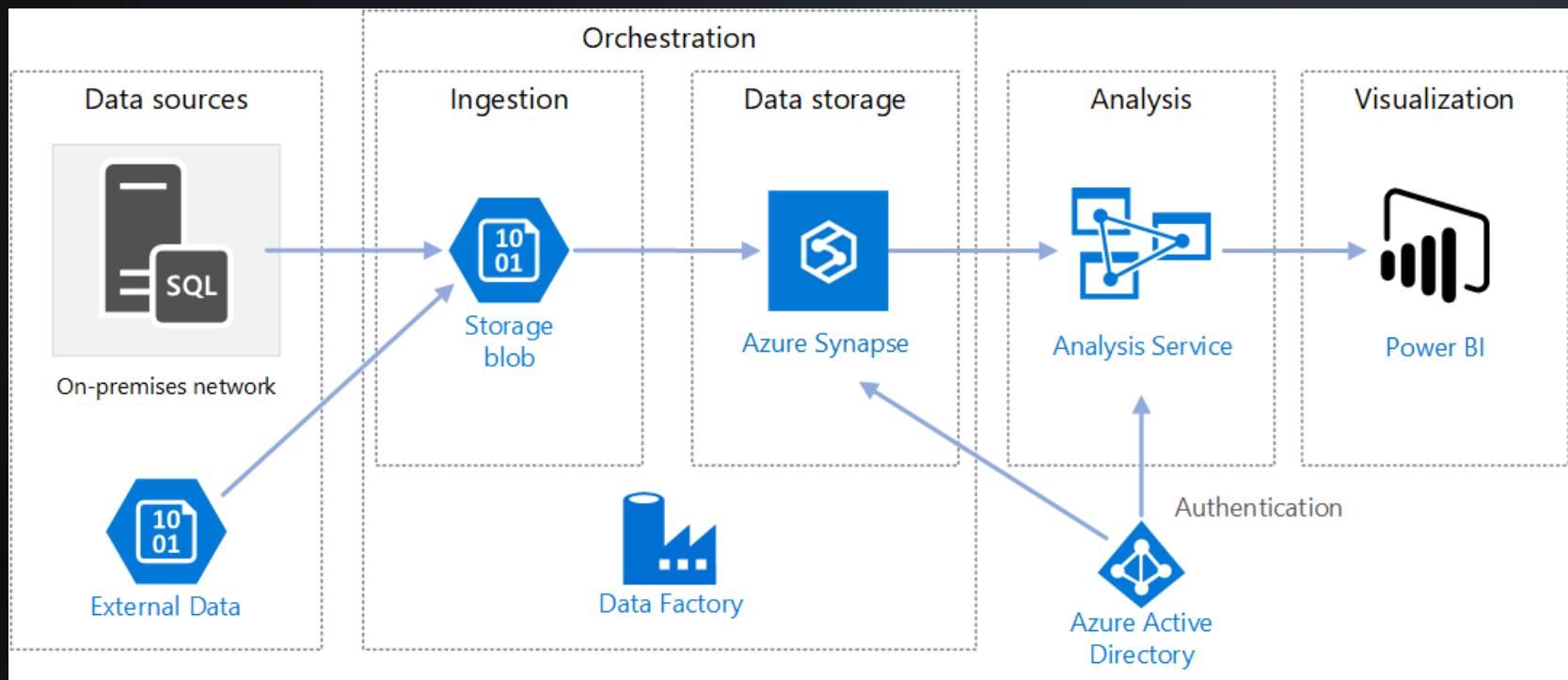
REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

AZURE DATA FACTORY

- Orquestra dados brutos e etapas de integração com outros serviços.
- Fontes e Destinos relacionais, não relacionais e outros armazenamentos.
- Suporte híbrido de ETL.
- **ETL livre de código em nuvem como serviço.**
- Trabalha com Gatilhos (batch, API, etc)



PATROCÍNIO:



REALIZAÇÃO:

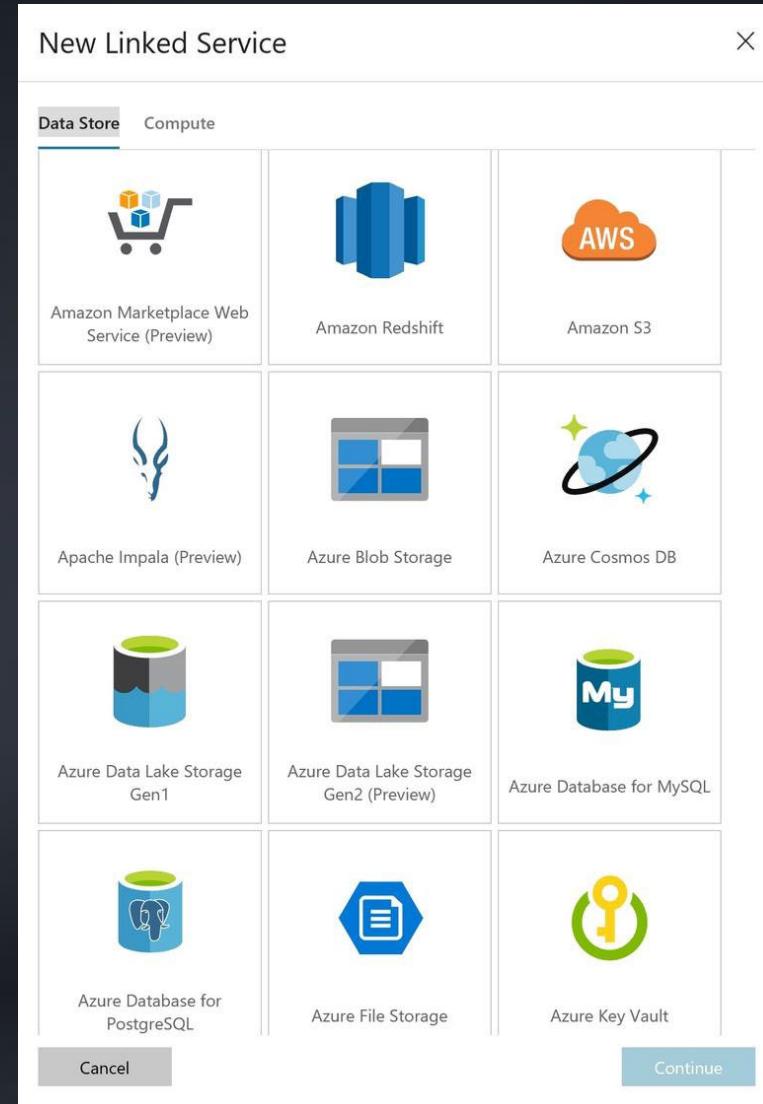
e.SethCloud

Azure
Academy

www.azureacademy.com.br

Oportunidades

- Chamadas para Procedures;
- Chamadas para Runbooks;
- Chamadas com triggers externas;
- Integração de Azure Functions no pipeline;
- Guia Integration Routine em Connections para chamadas on-premises;
- Conversão de formatos.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

Conceitos

Linked services (Connections):

Estabelece a conexão com os drivers das fontes/destinos de dados externos tais como Discos, SQL Server, Cosmos DB, etc. Estas conexões serão consumidas na esteira do Pipeline.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, there's a sidebar with options: Data Factory, Author (highlighted with a red box), Monitor, and Manage. The main area is titled 'Linked services' with the sub-instruction 'Linked service defines the connection information to a data store or compute.' Below this, there's a 'New' button and a search/filter bar. A message says 'Showing 0 - 0 of 0 items'. At the bottom right, there's a 'Create linked service' button highlighted with a red box.

Datasets:

Select de registros da tabela da Connection – SQL. Selecionar um container dentro da storage account (Connection).

O Dataset depende da Connection. Enquanto a connection se conecta ao driver, o dataset navega nos registros/discos selecionando ou setando os dados como fonte ou destino.

The screenshot shows the Microsoft Azure Data Factory 'Author' interface. On the left, the sidebar has 'Author' (highlighted with a red box) and 'Manage'. The main area shows 'Factory Resources' with a 'Pipelines' section containing 'pipeline1' (marked with a blue dot). Under 'Activities', there's a 'Datasets' section (highlighted with a red box) which is currently empty. To the right, there are buttons for 'New dataset' (highlighted with a red box) and 'New folder'.

PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

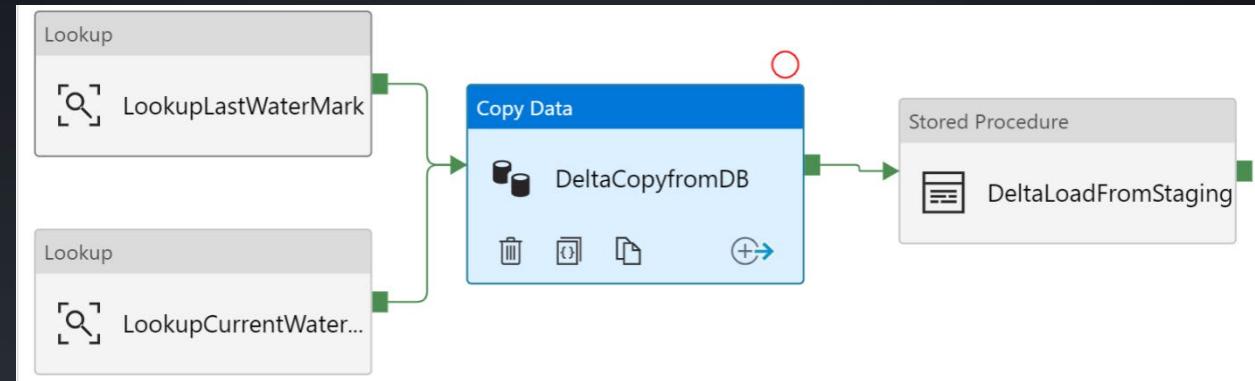
Azure
Academy

www.azureacademy.com.br

Conceitos

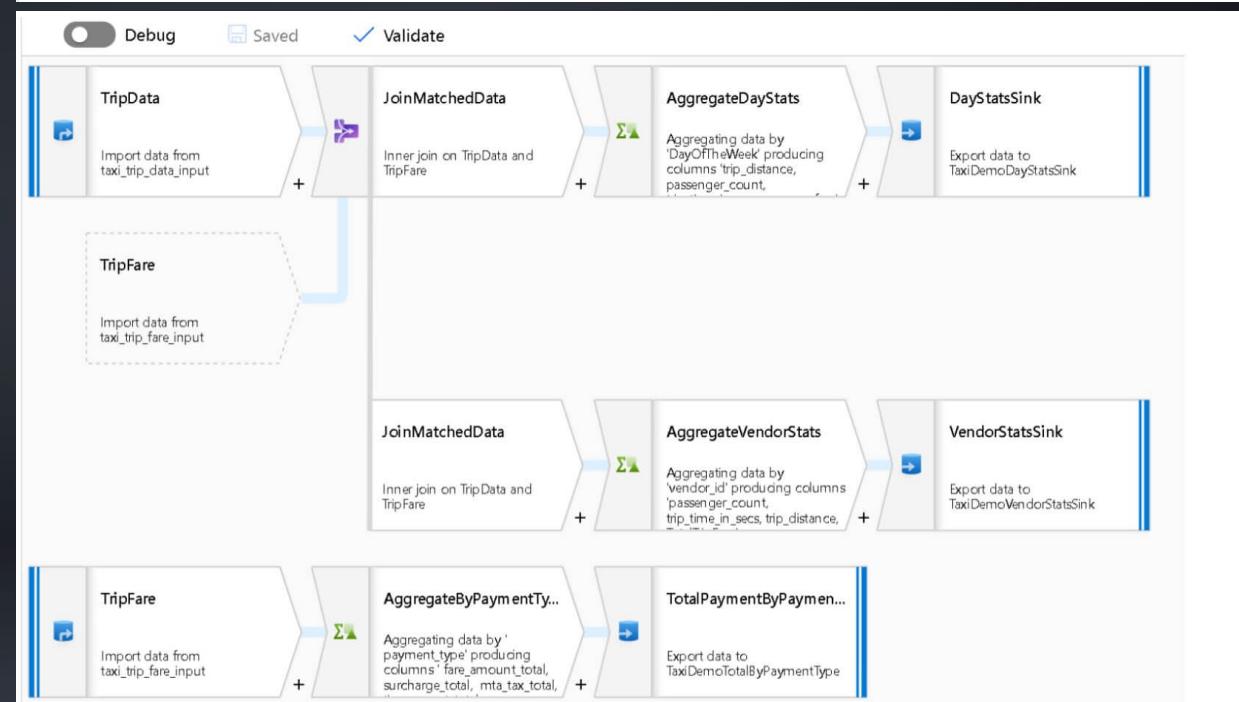
Pipelines

Esteiras com ações agrupadas de cópia, processamento, gravação e outros procedimentos.



Dataflows

Lógica de transformação de dados para data mining, agregação e isolamento.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br



Laboratório

Importação incremental com Delta

Assuntos:

- Lookup.
- Querys personalizadas.
- Flow de atividades.
- Parâmetros.

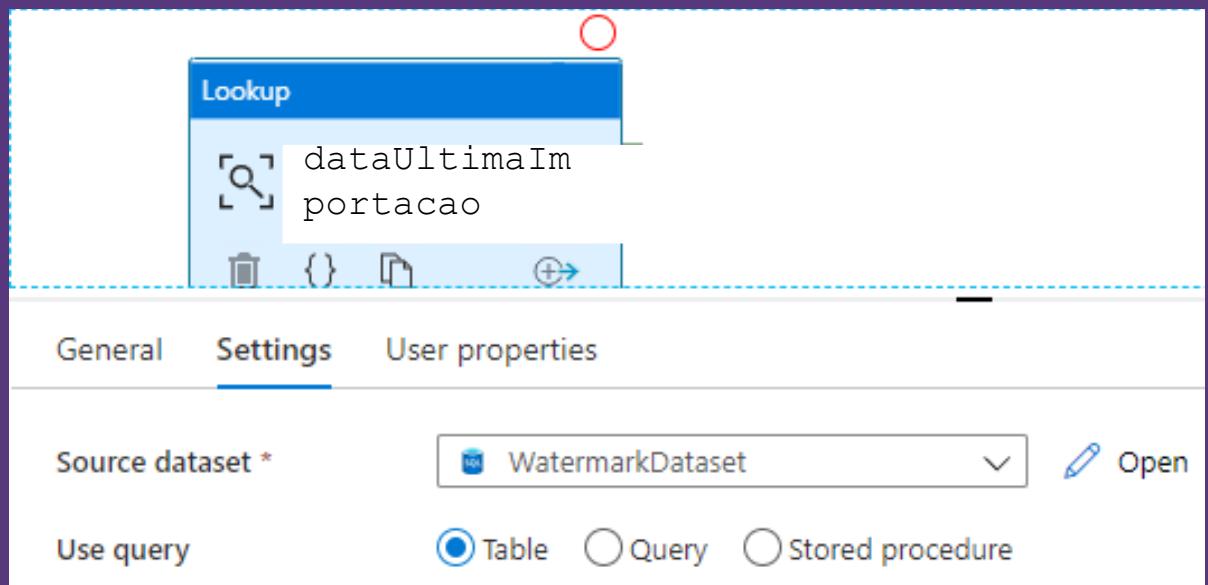
Tecnologias:

- SQL do Azure
- Storage Account - Data Lake Storage

Lab – Importação incremental com Delta

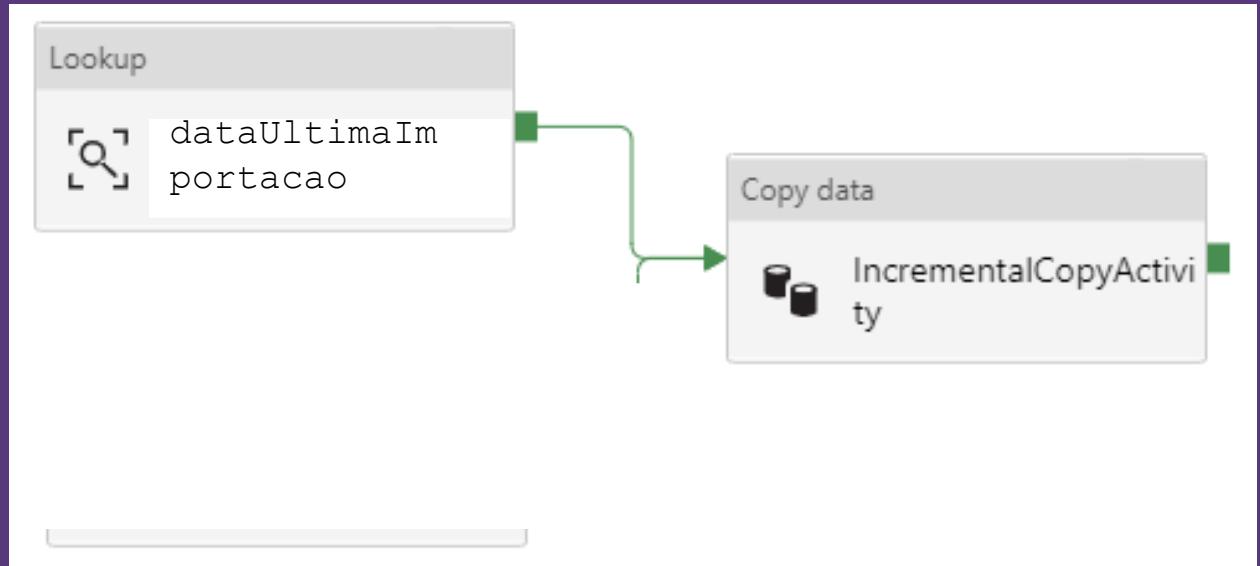
No Azure:

1. Instale um SQL do Azure com banco de dados de amostra e autorize o seu IP no Firewall do SQL.
2. Na sequencia, acesse o editor de consultas. Execute os comandos do arquivo 'lab2_SQL.txt' disponível no Portal do Aluno.
3. Abra o Data Factory e crie um novo Pipeline.
4. Insira a atividade **Lookup**. Configure o nome para **dataUltimaImportacao**.
5. Acesse a guia **Settings** e crie um novo **dataset** com o nome **datasetUltimaImportacao** apontando para a tabela **watermarktable** do SQL.



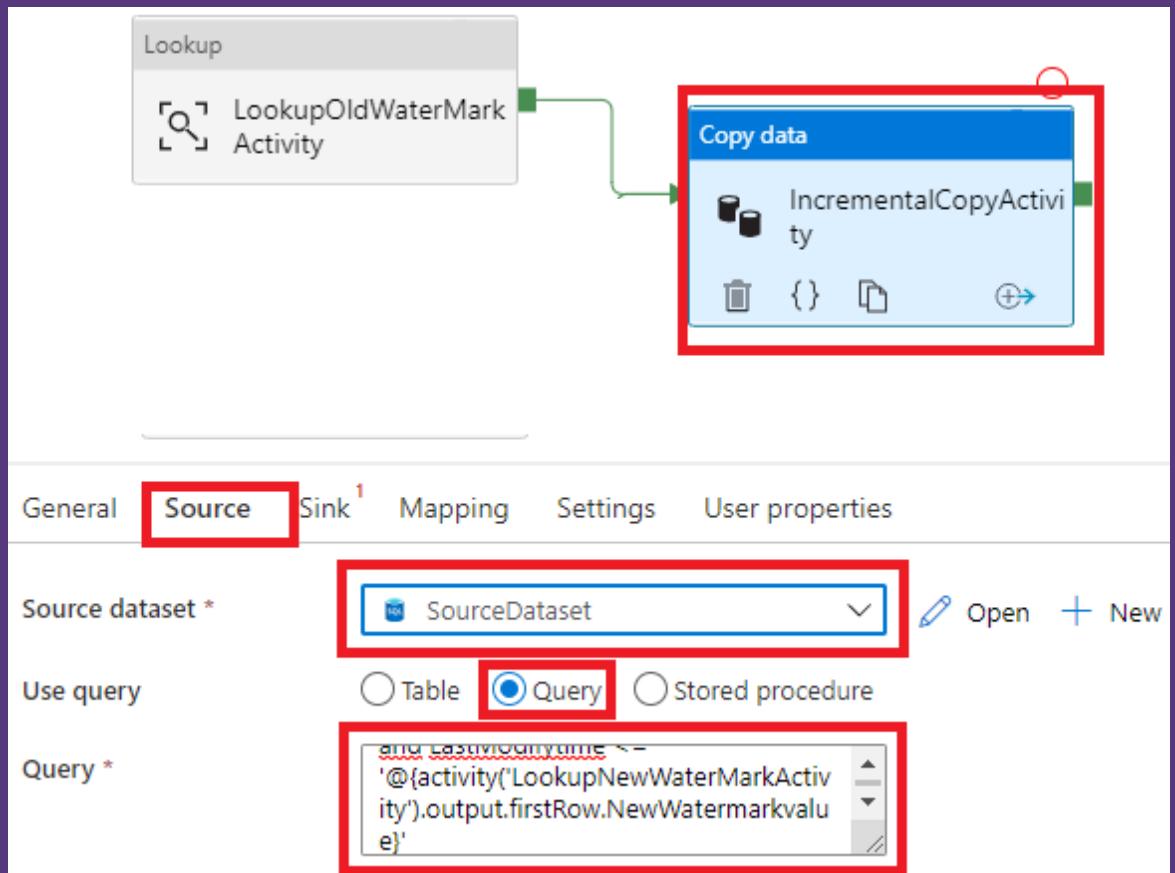
Lab – Importação incremental com Delta

1. Insira uma atividade de Copy. Configure o nome para IncrementalCopyActivity e ligue o fluxo conforme a imagem:



Lab – Importação incremental com Delta

1. Na atividade Copy, crie um dataset apontando para a tabela `data_source_table`.
2. Personalize a Query, inserindo um critério de seleção de dados somente para aqueles que ainda não foram importados.

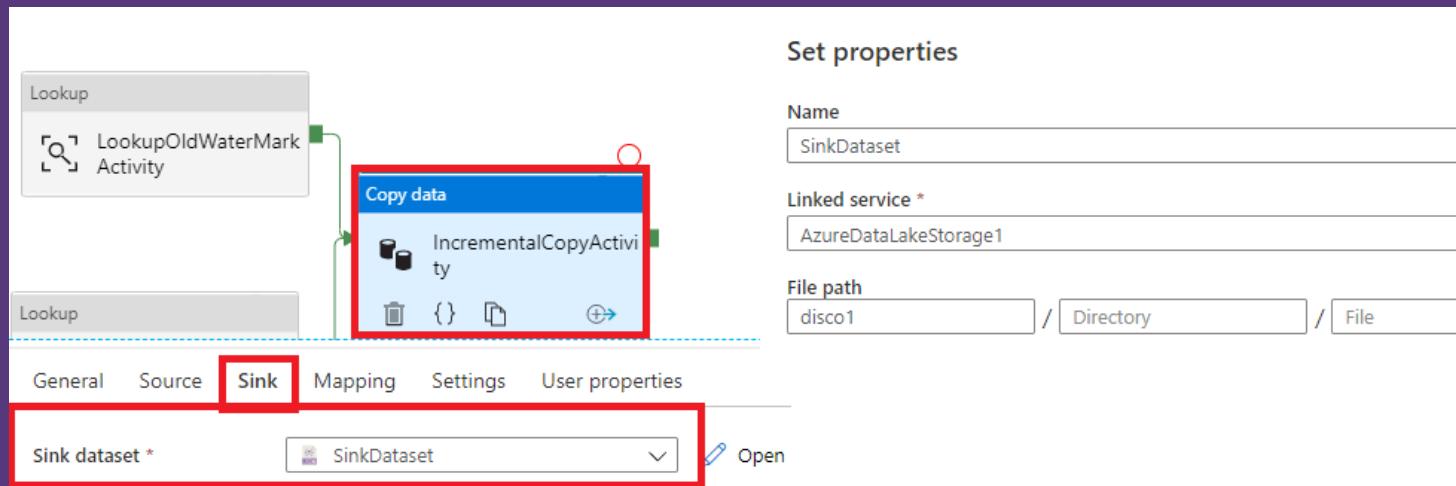


```
select * from  
data_source_table where  
LastModifytime >  
'{@{activity('dataUltimaImportacao').output.firstRow.WatermarkValue}}'
```

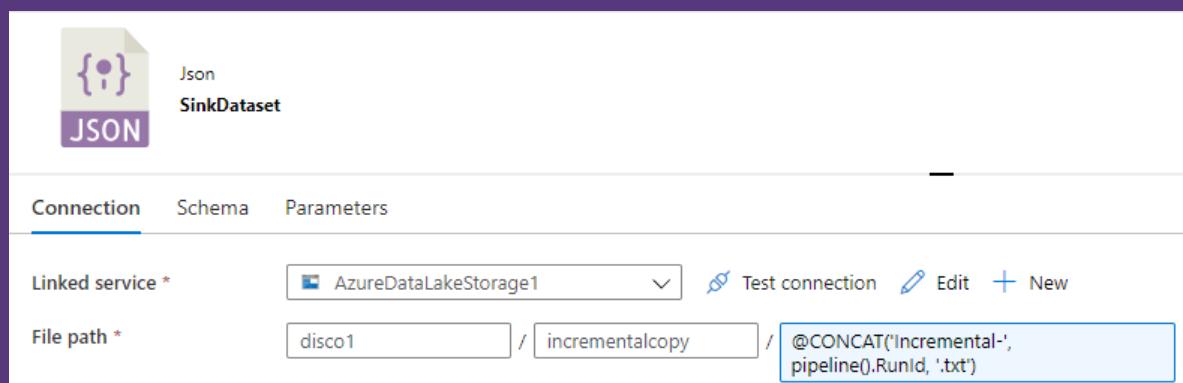


Lab – Importação incremental com Delta

1. Na atividade Copy, selecione a guia **Sink** e crie um novo dataset apontando para um container de blob da storage account criada no lab anterior.



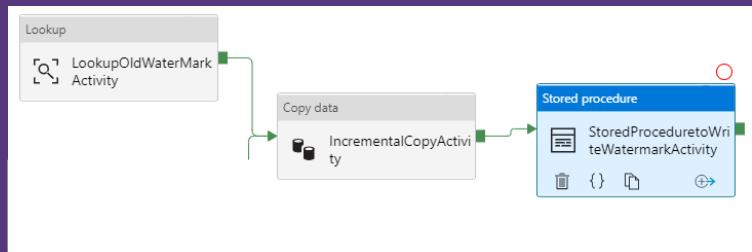
2. Após Concluir a criação do Dataset, selecione a opção Open e personalize o dataset conforme imagem a seguir. Este procedimento criará arquivos com os dados da extração a cada execução do pipeline.



```
@CONCAT('Incremental-',
pipeline().RunId, '.txt')
```

Lab – Importação incremental com Delta

1. De volta ao pipeline, insira uma atividade **Stored procedure** com o nome **StoredProceduretoWriteWatermarkActivity** e configure:



2. Configure a guia **Settings** conforme a imagem a seguir. Note a utilização dos parâmetros que serão herdados das atividades do lookup:

General **Settings** User properties

Linked service * **AzureSqlDatabase1**

Stored procedure name * **[dbo].[usp_write_watermark]**

▲ Stored procedure parameters ⓘ

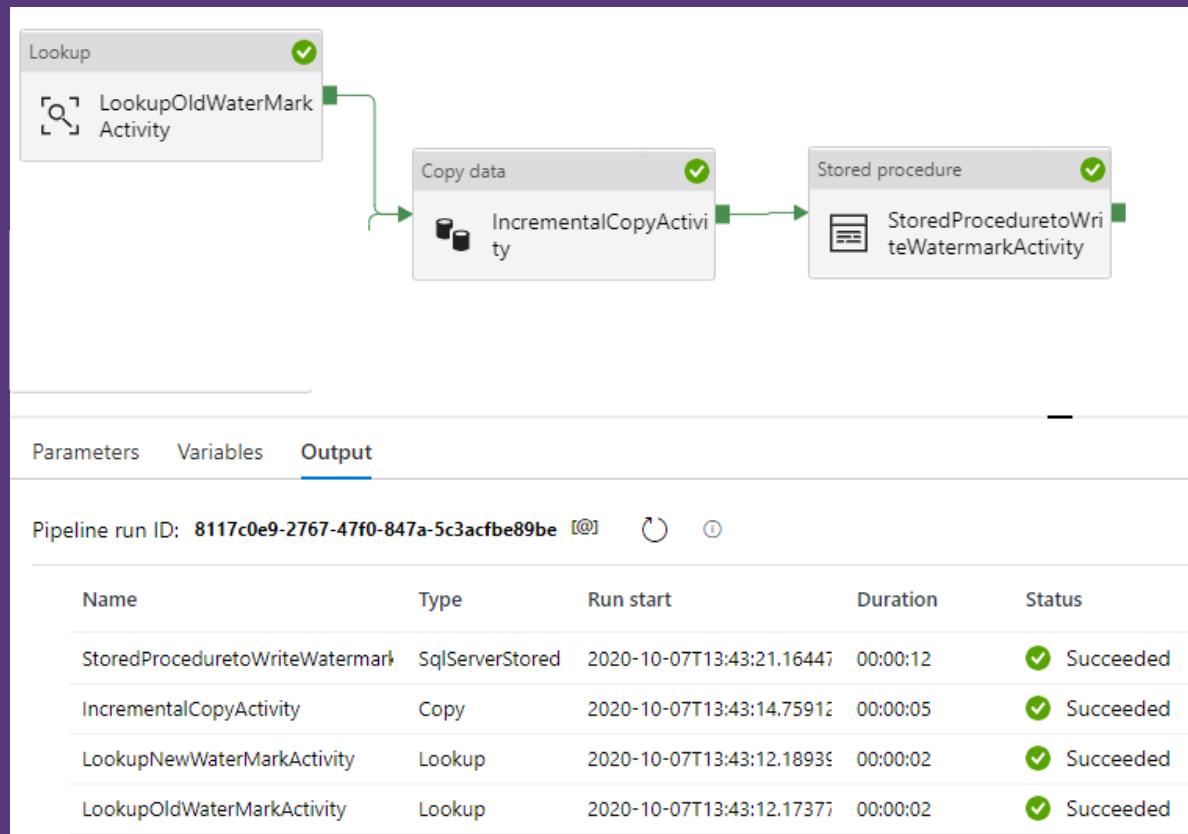
Import

NAME	TYPE	VALUE
TableName	String	@{activity('LookupOldWaterMarkActivity...')}

```
@{activity('dataUltimaImportacao').output.firstRow.TableName}
```

Lab – Importação incremental com Delta

1. Ative o modo Debug do Pipeline.
2. Clique em Validate para verificar se todas as configurações estão consistentes.
3. Execute o Pipeline e verifique os arquivos criados na Storage Account.
4. Aproveite para verificar se a tabela do SQL Server foi atualizada.



Conceitos

Transformação e limpeza de dados com DataFlows



Permite utilizar connections de diversas arquiteturas simultaneamente controlando visualmente o processo de transformação de dados.

PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br



Laboratório

DataFlows

Assuntos:

- DataFlows
- Funções
- Agregações
- Filtros
- Expressões

Tecnologias:

- Storage Account - Data Lake Storage

Lab – DataFlow

1. Crie uma nova storage account do tipo V2 e crie um container de Blob chamado **arquivos**.
2. Faça o upload para o container, do arquivo **moviesDB.csv** disponível no Portal do Aluno.
3. No Data Factory, verifique se já existe uma connection para a storage account. Caso não, crie uma nova.
4. Crie um **Dataflow**. Configure o Source conforme as imagens a seguir e clique na opção para criar um novo dataset.

The screenshot displays the Azure Data Flow 'New dataset' configuration interface across three panels:

- Source settings:** Shows the 'Output stream name' set to 'MoviesDb', 'Source type' set to 'Dataset', and a 'Dataset' dropdown with a red box around it and a 'New' button highlighted with a red box.
- Select a data store:** A grid of data stores including Azure Blob Storage, Azure Data Lake Storage Gen1, Azure Data Lake Storage Gen2 (highlighted with a blue box), Azure SQL Data Warehouse, Azure SQL Database, and Amazon Marketplace Web Service.
- Select format:** A grid of data formats including Parquet, DelimitedText (highlighted with a blue box), JSON, Avro, ORC, and Binary.

Below these panels is the 'Set properties' section:

- Name:** DelimitedText1
- Linked service:** AzureDataLakeStorage1
- File path:** arquivos/moviesDb.csv (highlighted with a red box)
- First row as header:**
- Import schema:** From connection/store From sample file None

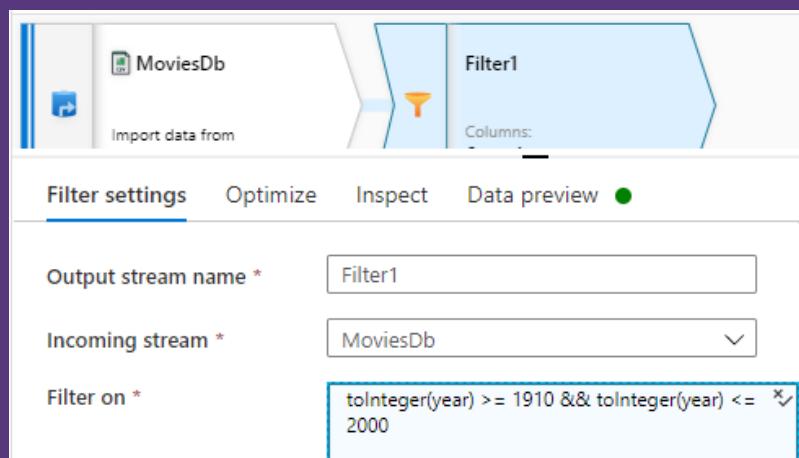
Lab – DataFlow

1. Acione o Data Preview para conferir a conexão.



Data preview					
Number of rows		+ INSERT 100	* UPDATE 0	- DELETE 0	* UPSERT 0
⟳ Refresh Typecast Modify Map drifted Statistics Remove					
↑↓	movie	abc	title	abc	genres
+	108583		Fawlty Towers (1975)		Comedy
+	32898		Trip to the Moon, A (Voyage dan...	Action Adventure Fantasy Sci-Fi	1902
+	7065		Birth of a Nation, The	Drama War	1915
+	7243		Intolerance: Love's Struggle Thr...	Drama	1915
+	62383		20,000 Leagues Under the Sea	Action Adventure Sci-Fi	1915
+	8511		Immigrant, The	Comedy	1917

2. Caso não tenha importado o título das colunas, edite o dataset e importe o schema.
3. Insira uma atividade 'Filter' e configure o filtro com a expressão: `tolnteger(year) >= 1910 && tolnteger(year) <= 2000`



Para descobrir quais filmes são Comedies, você pode usar a `rlike()` função para localizar o padrão 'comédia' nos gêneros de coluna. Union a expressão RLIKE com a comparação de anos para obter:
`tolnteger(year) >= 1910 && tolnteger(year) <= 2000 && rlike(genres, 'Comedy')`

Lab – DataFlow

1. Adicione uma atividade do tipo ‘agregação’. Configure conforme as imagens:

The screenshot displays two side-by-side configurations of an Azure Data Flow pipeline. Both configurations start with an 'Import data from DelimitedText1' source connected to a 'Filter1' activity, which is then connected to an 'Aggregate1' activity.

Left Configuration (Group by):

- Aggregate settings:** Output stream name: Aggregate1; Incoming stream: Filter1.
- Group by:** Selected.
- Columns:** abc year (selected) and year (Name as).

Right Configuration (Aggregates):

- Aggregate settings:** Output stream name: Aggregate1; Incoming stream: Filter1.
- Group by:** Aggregates (selected).
- Grouped by:** year.
- Add:** Add, Clone, Delete, Open expression builder.
- Expression:** AverageComedyRating (Column) and avg(toInteger(Rating)) (Expression).

Text at the bottom: avg (toInteger (Rating))

Mais sobre agregação: <https://docs.microsoft.com/pt-br/azure/data-factory/data-flow-aggregate>

Lab – DataFlow

1. Adicione uma atividade do tipo ‘Sink’. Configure um novo dataset:

The screenshot shows the 'Sink' configuration pane for a Data Flow job. The pipeline consists of the following stages: 'MoviesDb' (Import data from DelimitedText1), 'Filter1' (Filtering rows using expressions on columns 'year'), 'Aggregate1' (Aggregating data by 'year' producing columns 'AverageComedyRating'), and 'sink1'. The 'sink1' stage is currently selected. The configuration options are as follows:

- Output stream name ***: sink1
- Incoming stream ***: Aggregate1
- Sink type ***: Dataset

The screenshot shows the 'New dataset' configuration pane. It displays a grid of data store icons and names. The 'Azure' tab is selected. The visible data stores are:

Icon	Name
Azure Blob Storage	Azure Blob Storage
Azure Data Lake Storage Gen1	Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2	Azure Data Lake Storage Gen2
Azure SQL Data Warehouse	Azure SQL Data Warehouse
Azure SQL Database	Azure SQL Database
Amazon Marketplace Web Service	Amazon Marketplace Web Service

The screenshot shows the 'Select format' configuration pane. It displays a grid of data format icons and names. The 'DelimitedText' format is selected and highlighted with a blue border.

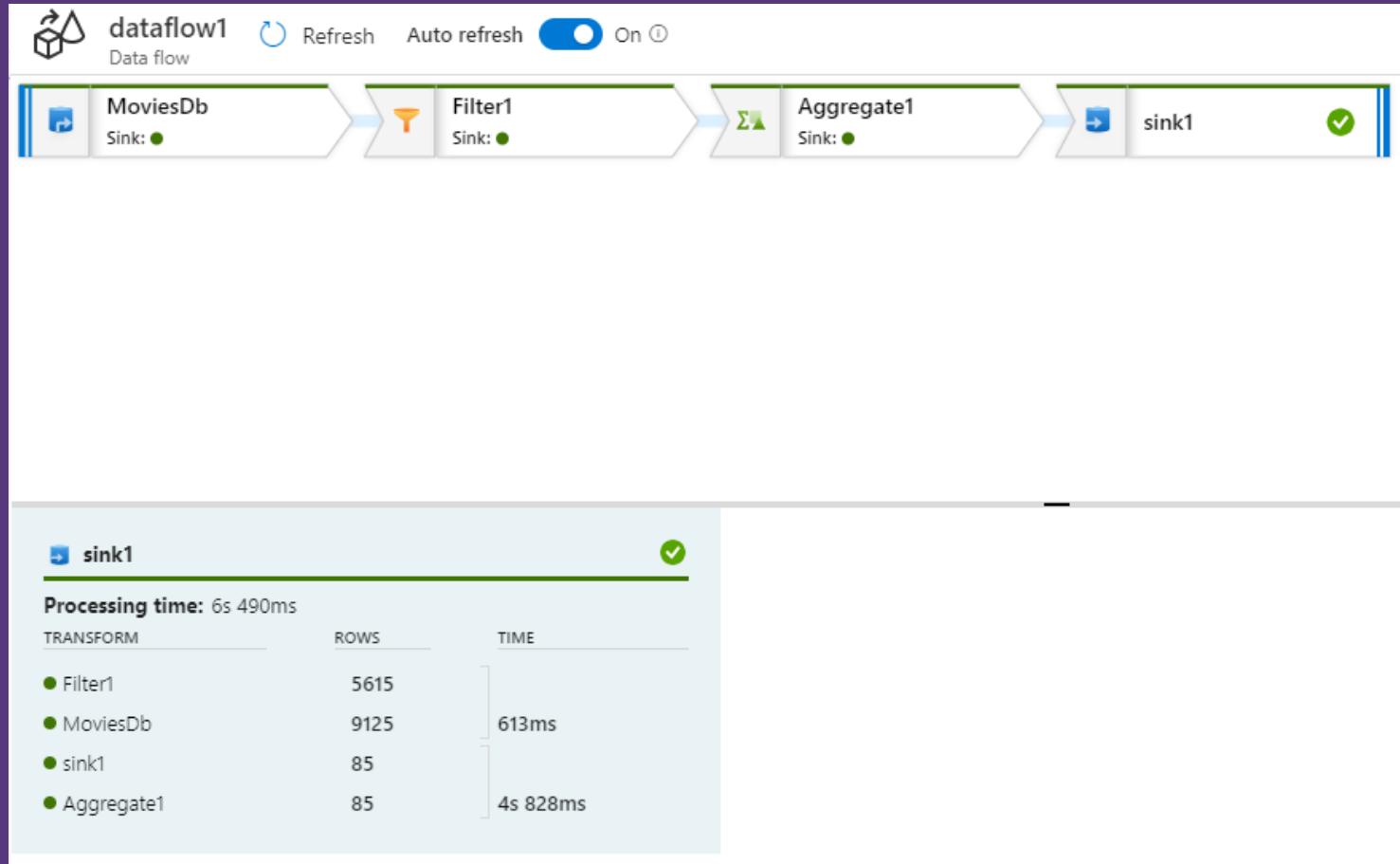
Icon	Format
Parquet	Parquet
CSV	DelimitedText
JSON	Json
Avro	Avro
ORC	ORC
Binary	Binary

The screenshot shows the 'Set properties' configuration pane for a 'DelimitedText' dataset. The fields are as follows:

- Name**: DelimitedText2
- Linked service ***: AzureDataLakeStorage1
- File path**: arquivos (highlighted with a red box)
- First row as header**: checked
- Import schema**:
 - From connection/store
 - From sample file
 - None (highlighted with a red box)

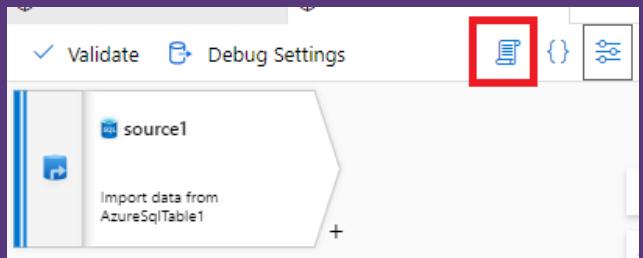
Lab – DataFlow

1. Retorne ao Pipeline e insira a atividade do DataFlow.
2. Teste a execução e confira os resultados.



Lab – DataFlow – Remover duplicidades ou nulos

1. Crie um novo Dataflow e conecte o Source a um dataset do SQL do Azure com uma tabela de exemplo.
2. Em seguida, clique no botão 'scripts' conforme imagem:



3. Insira as linhas a seguir para procurar e remover duplicidades:

```
source1 aggregate(groupBy(mycols = sha2(256,columns())),
    each(match(true()), $$ = first($$))) ~> DistinctRows
```

4. Teste os resultados utilizando o Data Preview.

5. O Código a seguir procura e elimina valores nulos:

```
source1 split(contains(array(columns()),isNull(#item)),
    disjoint: false) ~> LookForNULLs@(hasNULLs, noNULLs)
```

Consulte outros scripts de transformação de dados:

<https://docs.microsoft.com/pt-br/azure/data-factory/data-flow-script#distinct-row-using-all-columns>

Conceitos

Metadata:

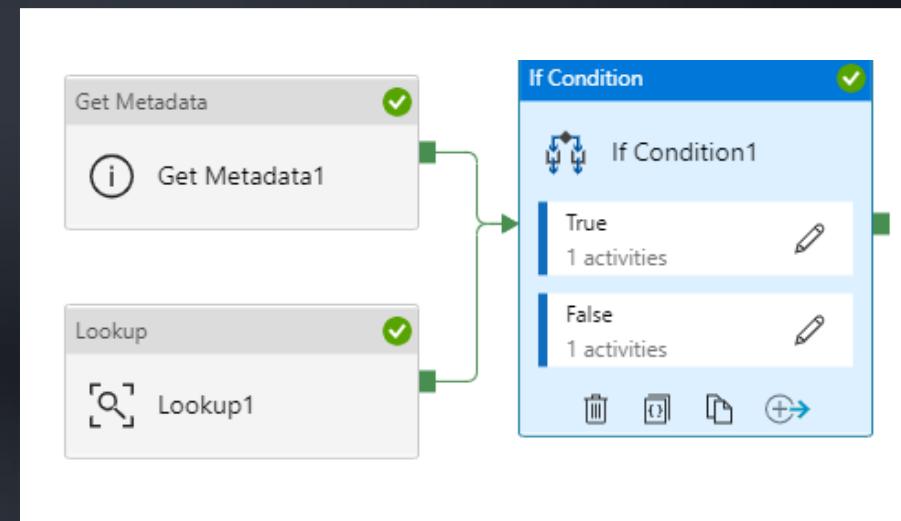
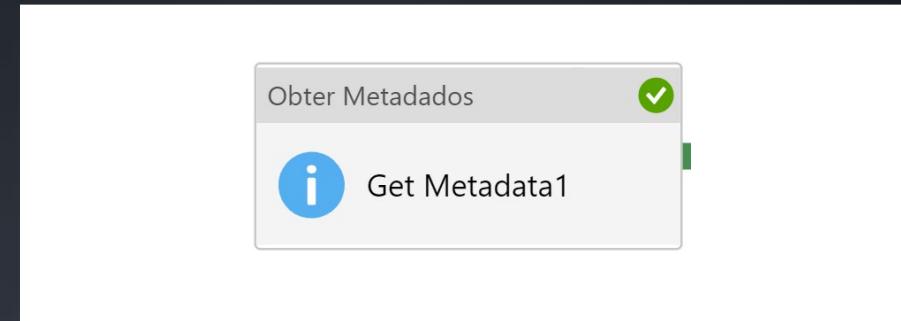
Permite validar a estrutura de fontes e destinos de dados antes ou depois de movimentações. Utilize a propriedade **ESTRUTURA** para retornar uma lista de nomes de coluna e tipos de coluna.

Consulte outros tipos de retornos para metadados:

<https://docs.microsoft.com/pt-br/azure/data-factory/control-flow-get-metadata-activity>

Estrutura condicional - IF:

Permite avaliar condições lógicas e resultados de variáveis agrupando atividades para TRUE e FALSE.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

DOCK
DATA
HACKATHON

TRANSFORMAR DADOS COM PYTHON

PATROCÍNIO:



Microsoft

REALIZAÇÃO:

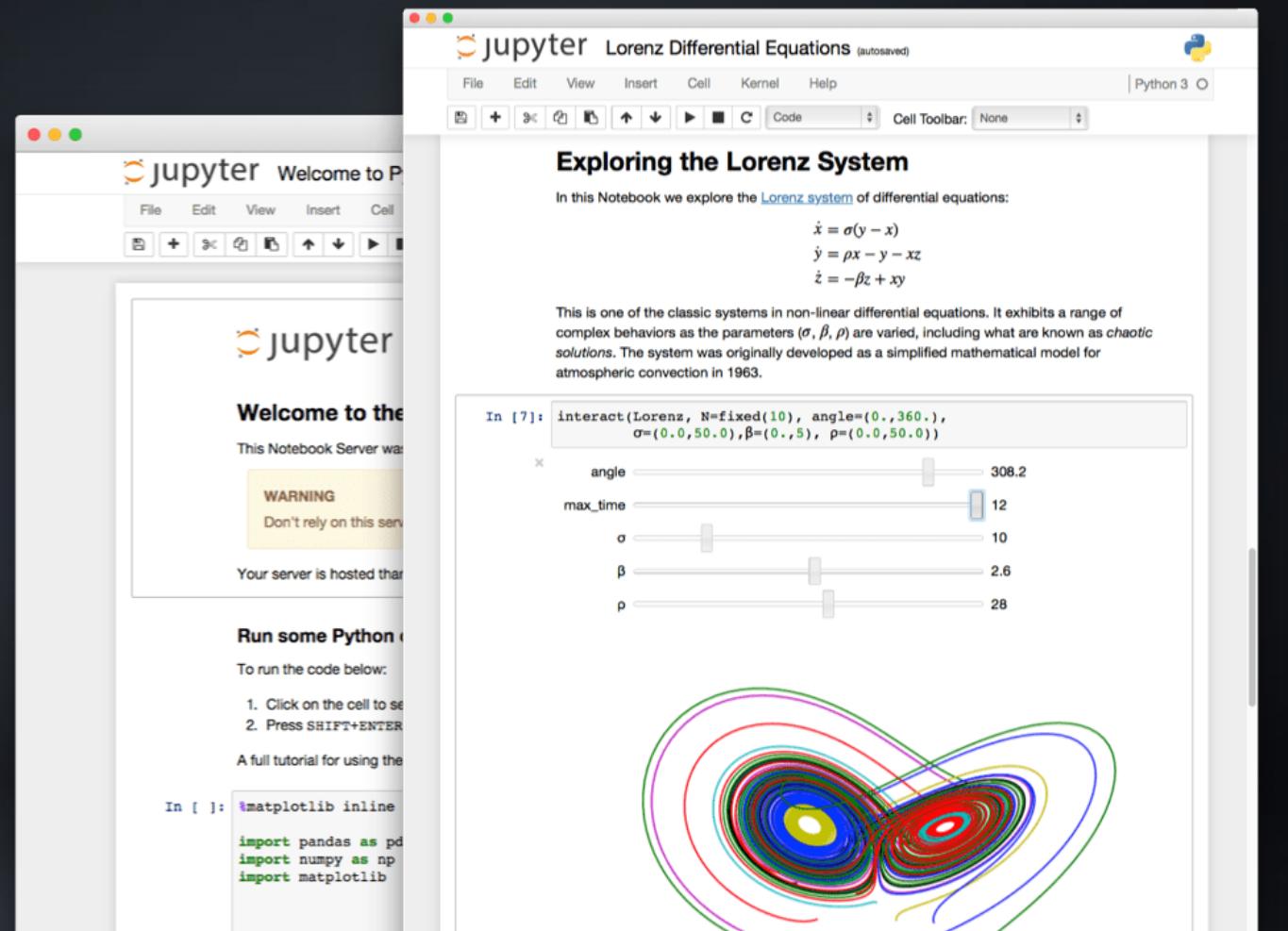
e.SethCloud

Azure
Academy
.com.br

Jupyter Notebook

O Jupyter Notebook é um aplicativo de código aberto que permite criar e compartilhar documentos que contêm código ativo, equações, visualizações e texto narrativo. Os usos incluem: limpeza e transformação de dados, simulação numérica, modelagem estatística, visualização de dados, aprendizado de máquina e muito mais.

<https://jupyter.org/>



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

Conheça mais Python

Possibilidade de baixar o Python e rodar localmente ou utilizar em Plataformas como Databricks, ML Studio e outras.

<https://www.python.org/downloads>

Compreenda como utilizar as principais bibliotecas que oferecem recursos desde a organização de datasets a modernos visualizadores de resultados, além de bibliotecas para cálculo de matrizes e Processamento de Linguagem Natural.

[Numpy](#) - Cálculo matemáticos para Arrays Multidimensionais.

[Pandas](#) - Manipulação e Análise de Dados.

[Scikit Learn](#) - Poderosa biblioteca para Machine Learning.

[Matplotlib](#) - Plotagem de gráficos.

[Plotly](#) - Criação de visualizações interativas.

[Keras](#) - Rede neural e Deep Learning.

[NLTK](#) - Processamento de Linguagem Natural (NLP).

[Scrapy](#) - Biblioteca para Web Crawling.

PATROCÍNIO:



REALIZAÇÃO:

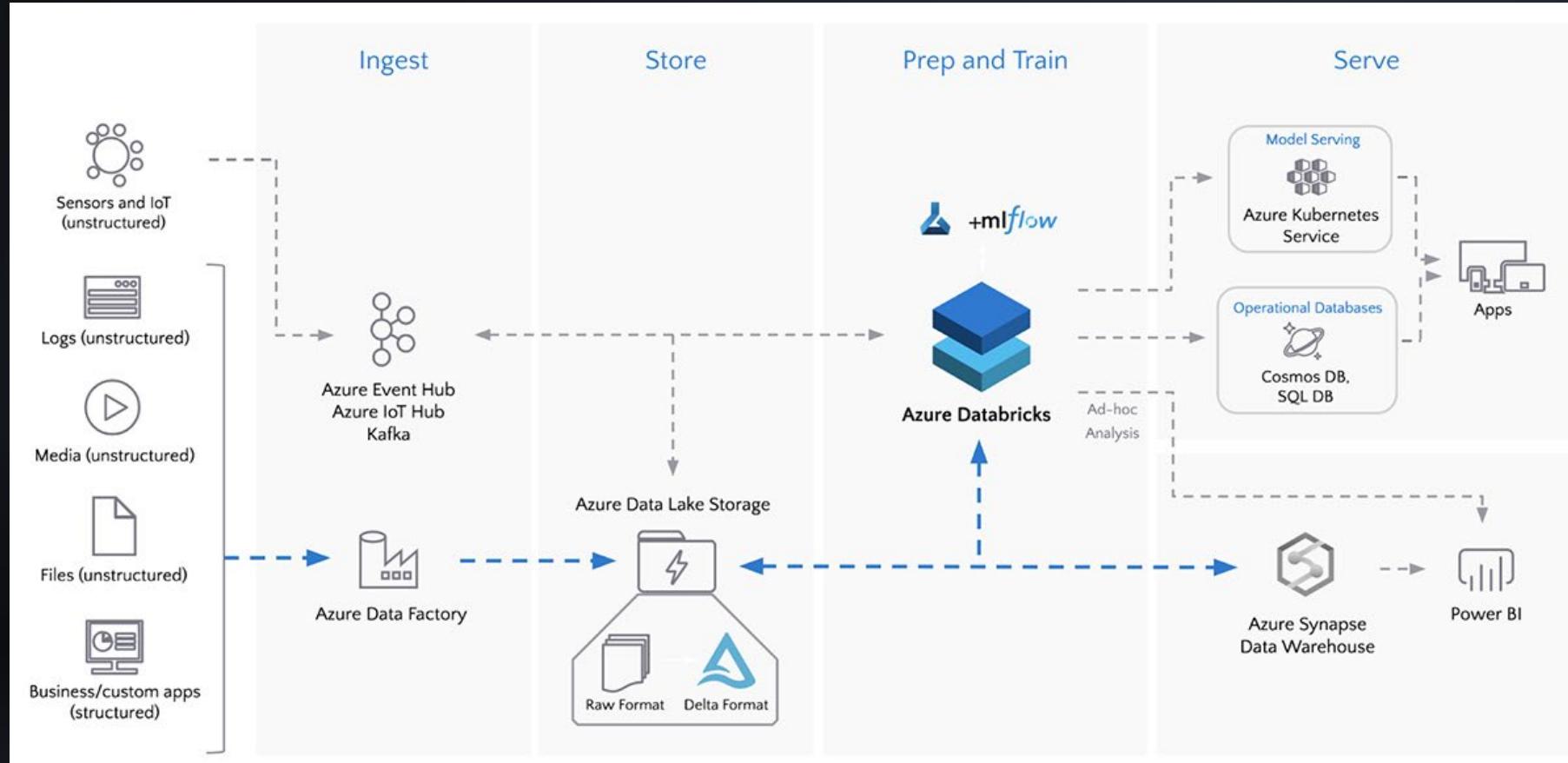
e.SethCloud

Azure
Academy

www.azureacademy.com.br

Serviços compatíveis com Jupyter Notebooks

- Funciona com Clusters: Interativos integrados à notebooks ou Jobs.
- Apache Spark possui foco em Big Data e tem o objetivo de processar grandes volumes de dados.



PATROCÍNIO:



REALIZAÇÃO:

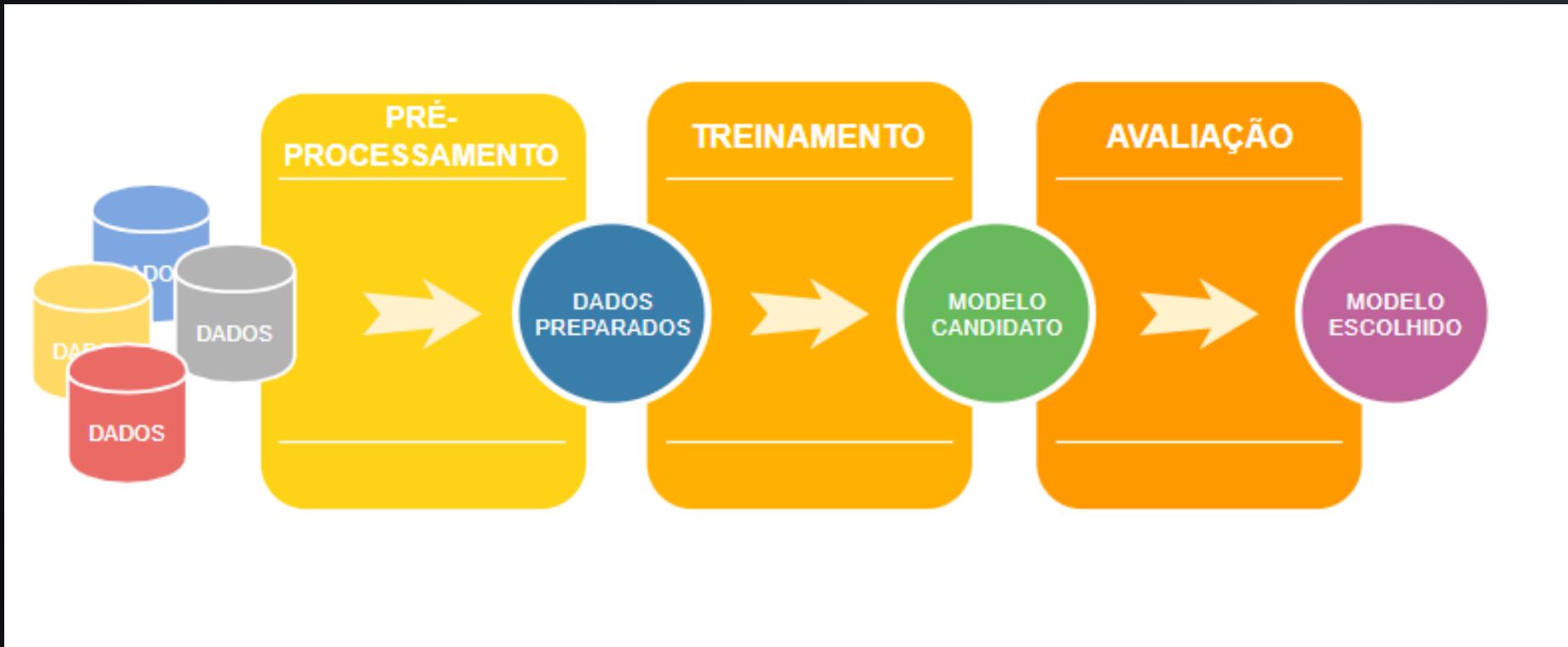
e.SethCloud

Azure
Academy

www.azureacademy.com.br

Arquitetura

Etapas de treinamento ensinam e aprimoram o algoritmo de acordo com os objetivos da detecção e classificação.



PATROCÍNIO:



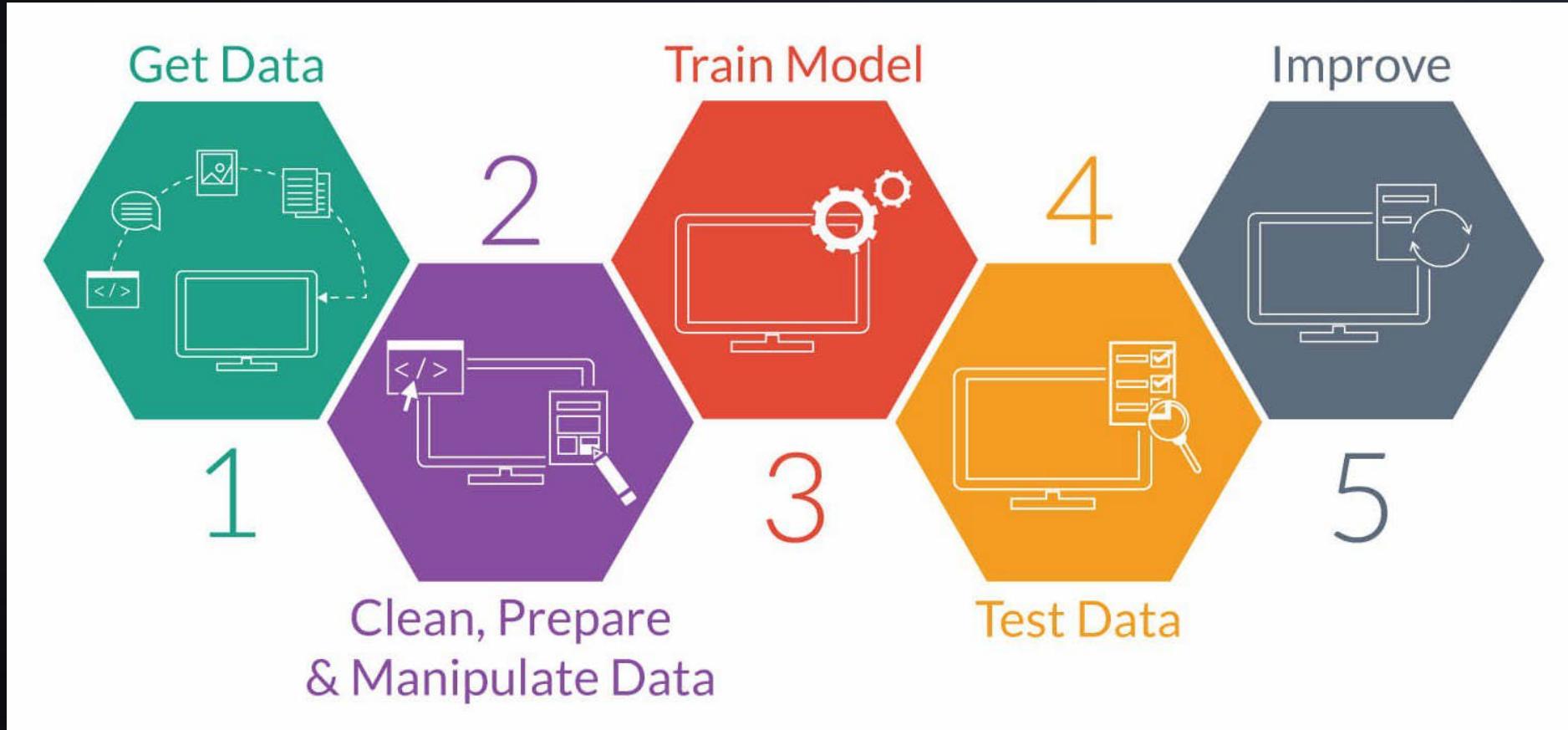
REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

Machine Learning



PATROCÍNIO:



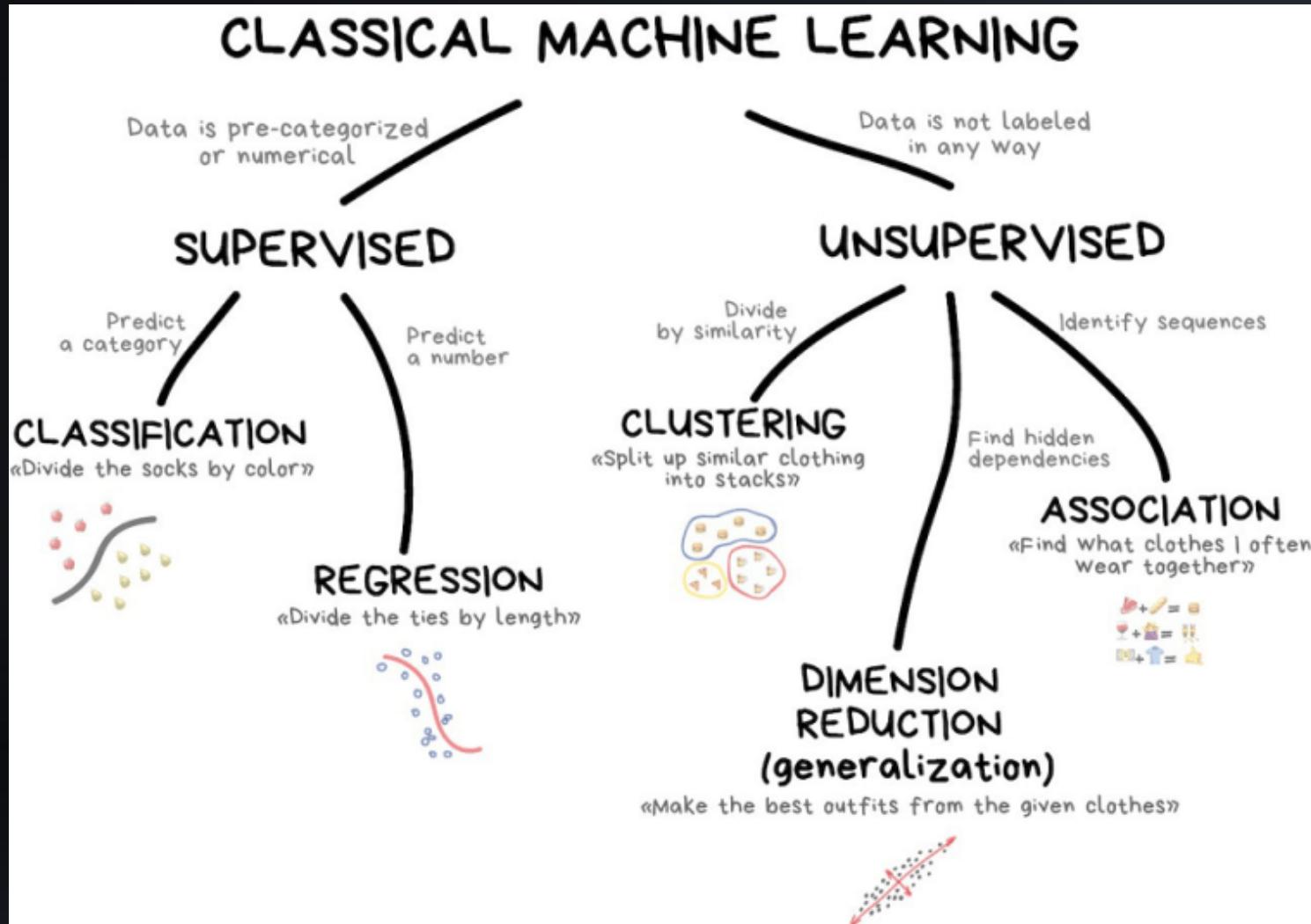
REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

MACHINE LEARNING



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

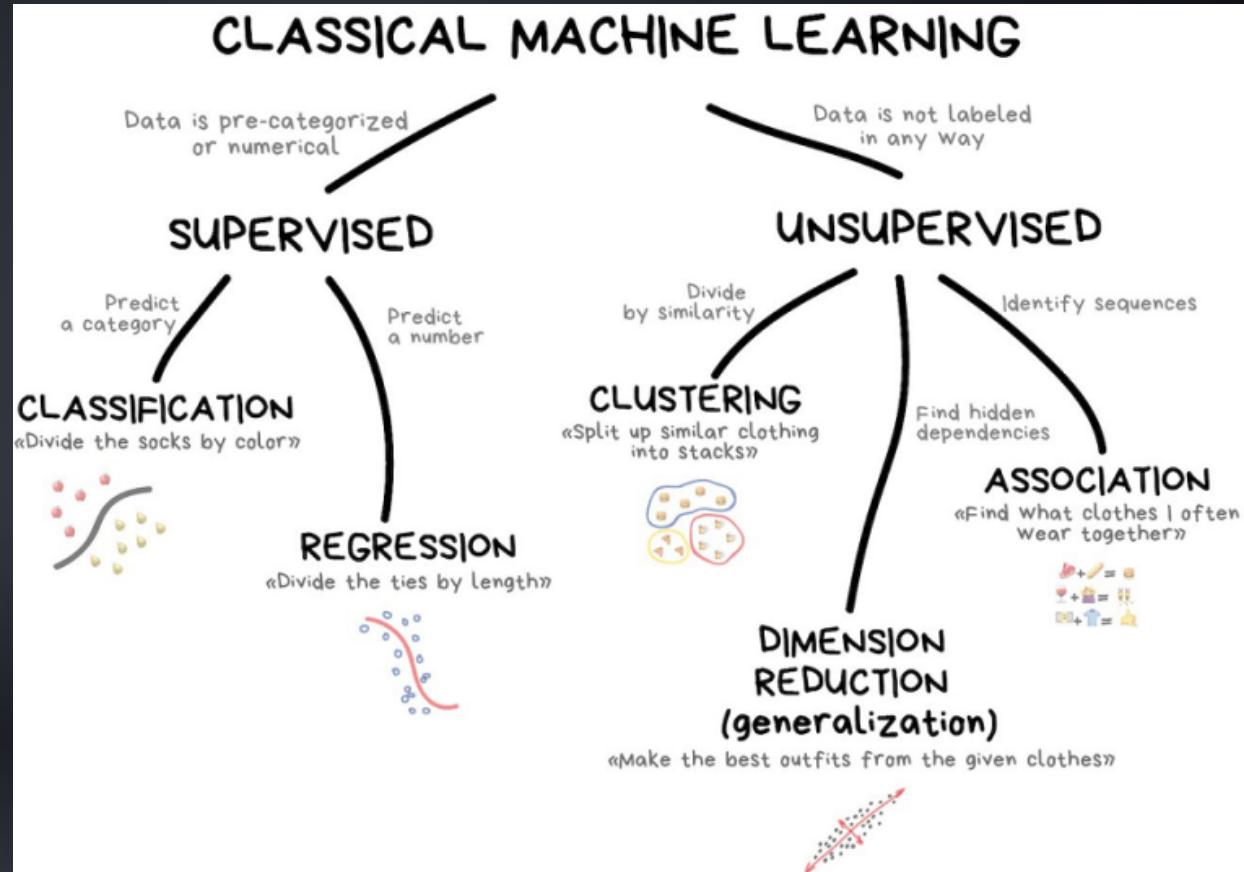
Modelos de aprendizado

Supervisionado:

- Tem um gabarito a ser seguido.
- Você determina o modelo.
- Possui definição.

Não supervisionado:

- Busca padrões para sugerir algo, por exemplo: um produto similar no e-commerce.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

Modelos de aprendizado

Supervisionado – Regressão:

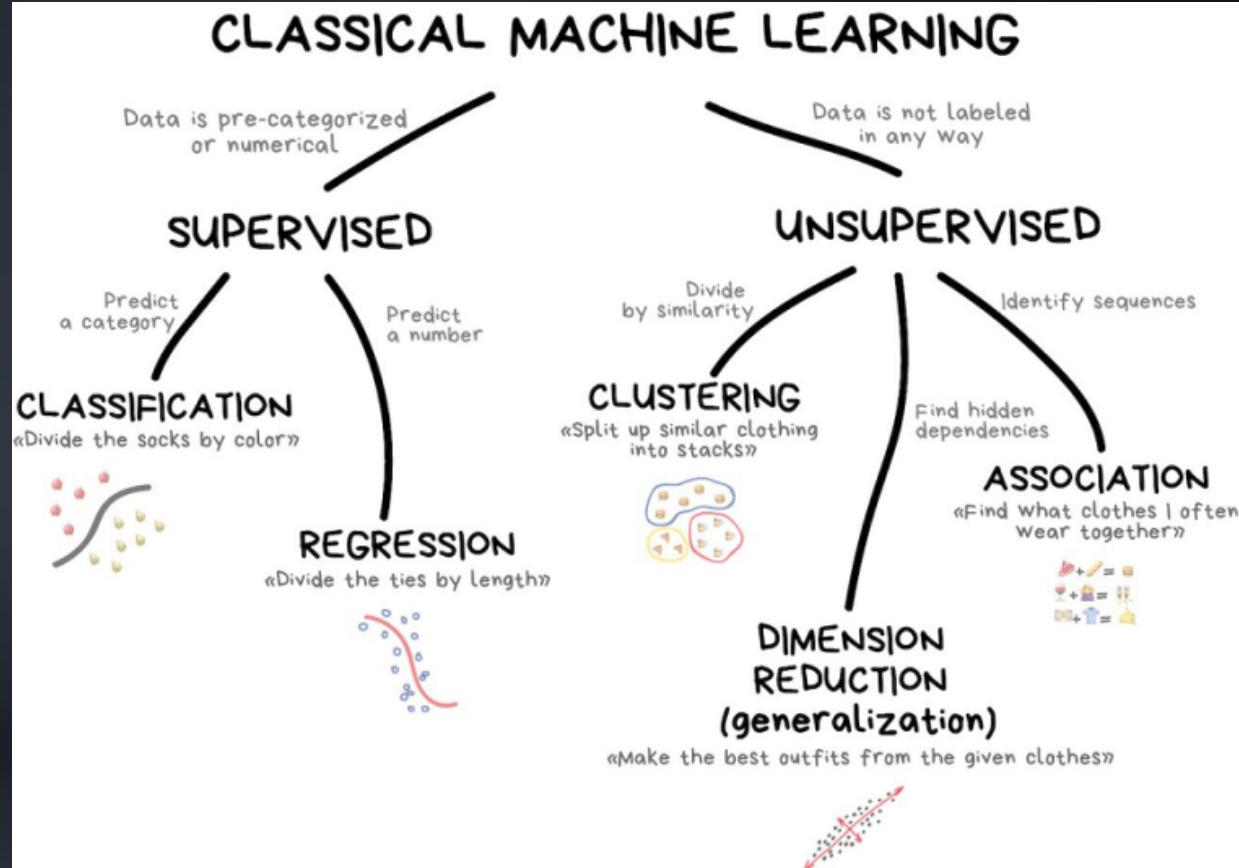
- Ex: Previsão do tempo.
- Recebe características na entrada e desenvolve a previsão na saída.

Supervisionado – Classificação:

- Tenta prever uma classe ao invés de um número.
- Separa e classifica através de superfícies.

Não Supervisionado – Clusterização:

- Agrupa os dados de forma similar. Trabalha com posições.



PATROCÍNIO:



REALIZAÇÃO:

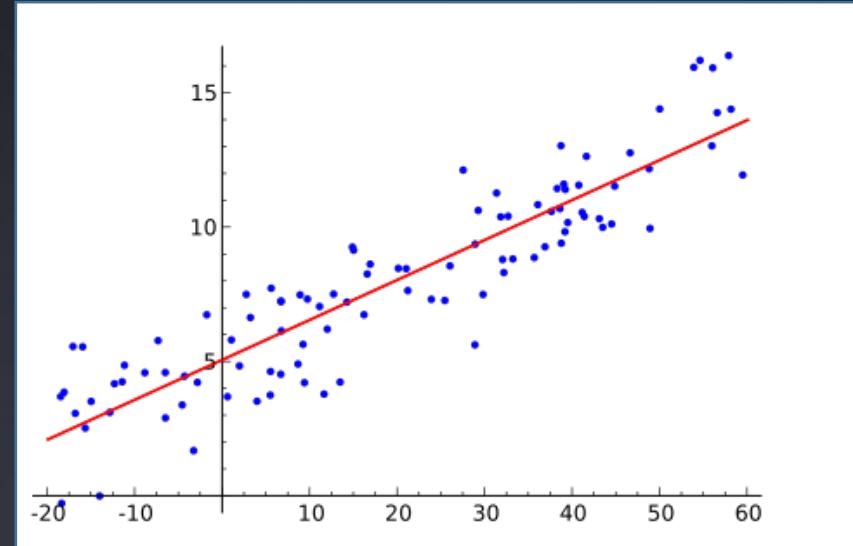
e.SethCloud

Azure
Academy

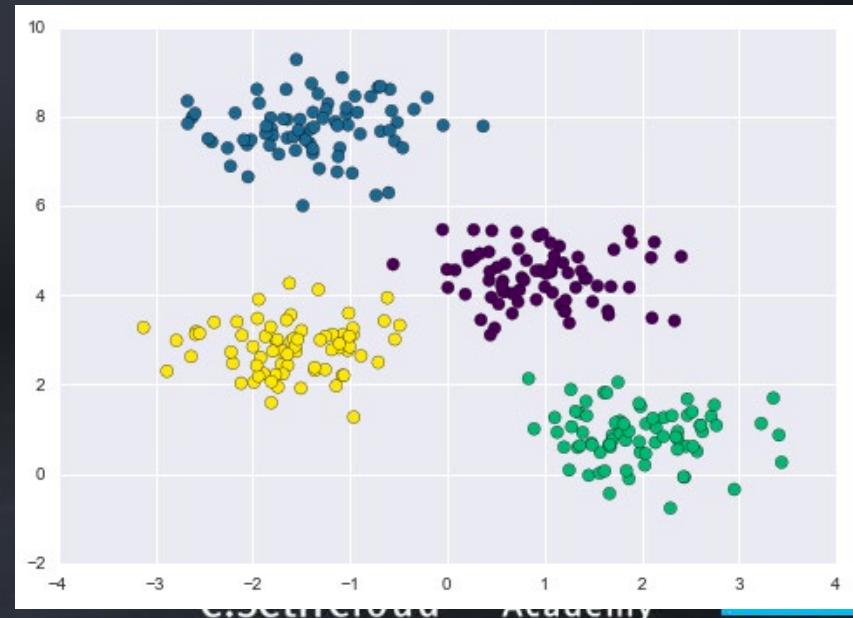
www.azureacademy.com.br

ML: Aprendizagem supervisionada x não supervisionada

A aprendizagem supervisionada visa prever o valor de algum resultado com base em uma ou mais medidas de entrada. Em um dataset com dados de perfis de casas x renda percapta, se precisarmos prever o preço de casa, a saída é o preço de uma casa e a entrada são recursos como o número de quartos



A aprendizagem não supervisionada descreve associações e padrões em dados sem um resultado conhecido. Um exemplo disso seria o agrupamento de dados de clientes para encontrar os segmentos de clientes. Nesse caso, nenhuma saída conhecida é usada como entrada. Em vez disso, o objetivo é descobrir como os dados são organizados em segmentos naturais ou clusters.

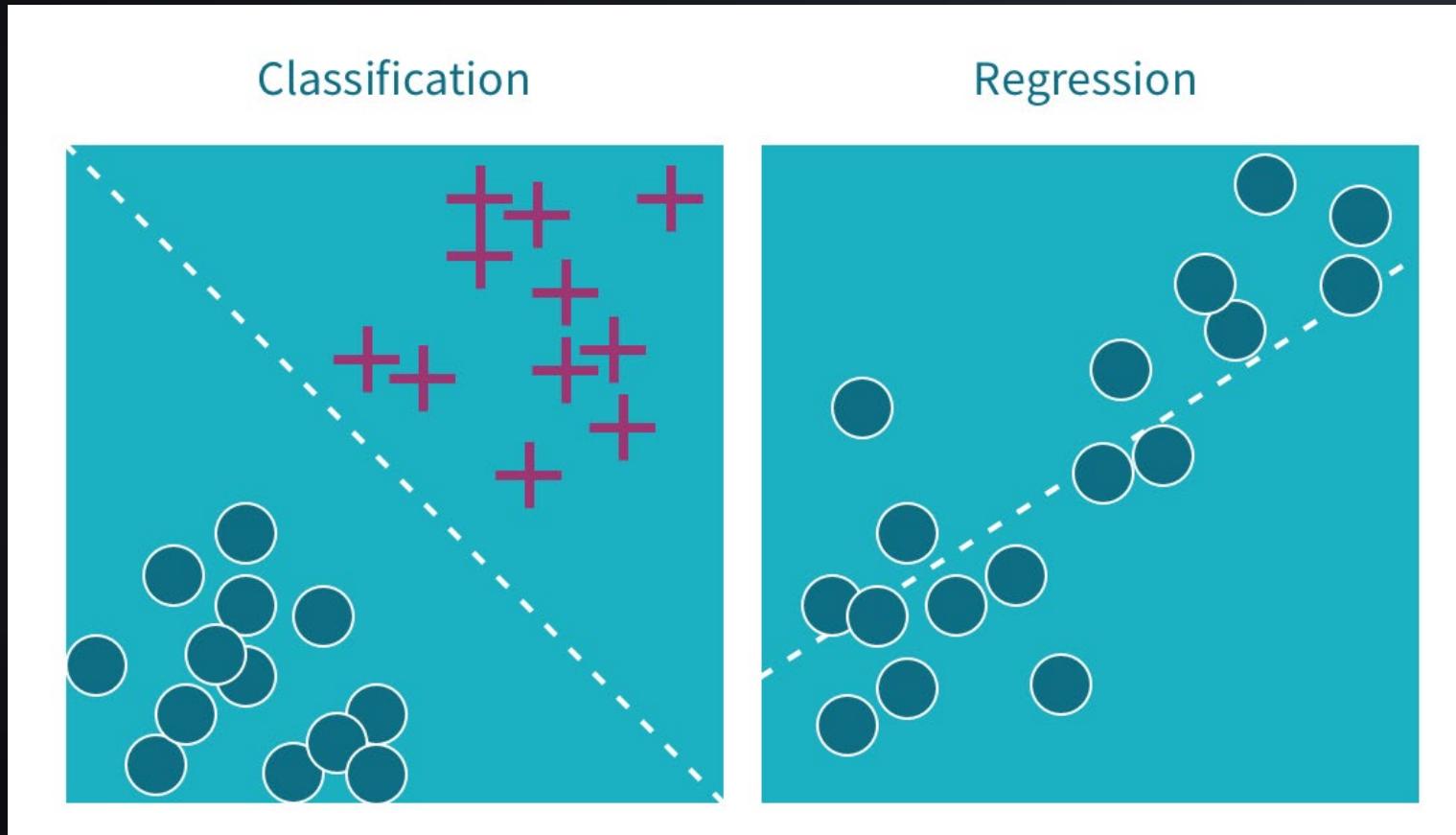


PATROCÍNIO:



Microsoft

ML: Algoritmos supervisionados x não supervisionados



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

DOCK
DATA
HACKATHON

SYNAPSE ANALYTICS

PATROCÍNIO:

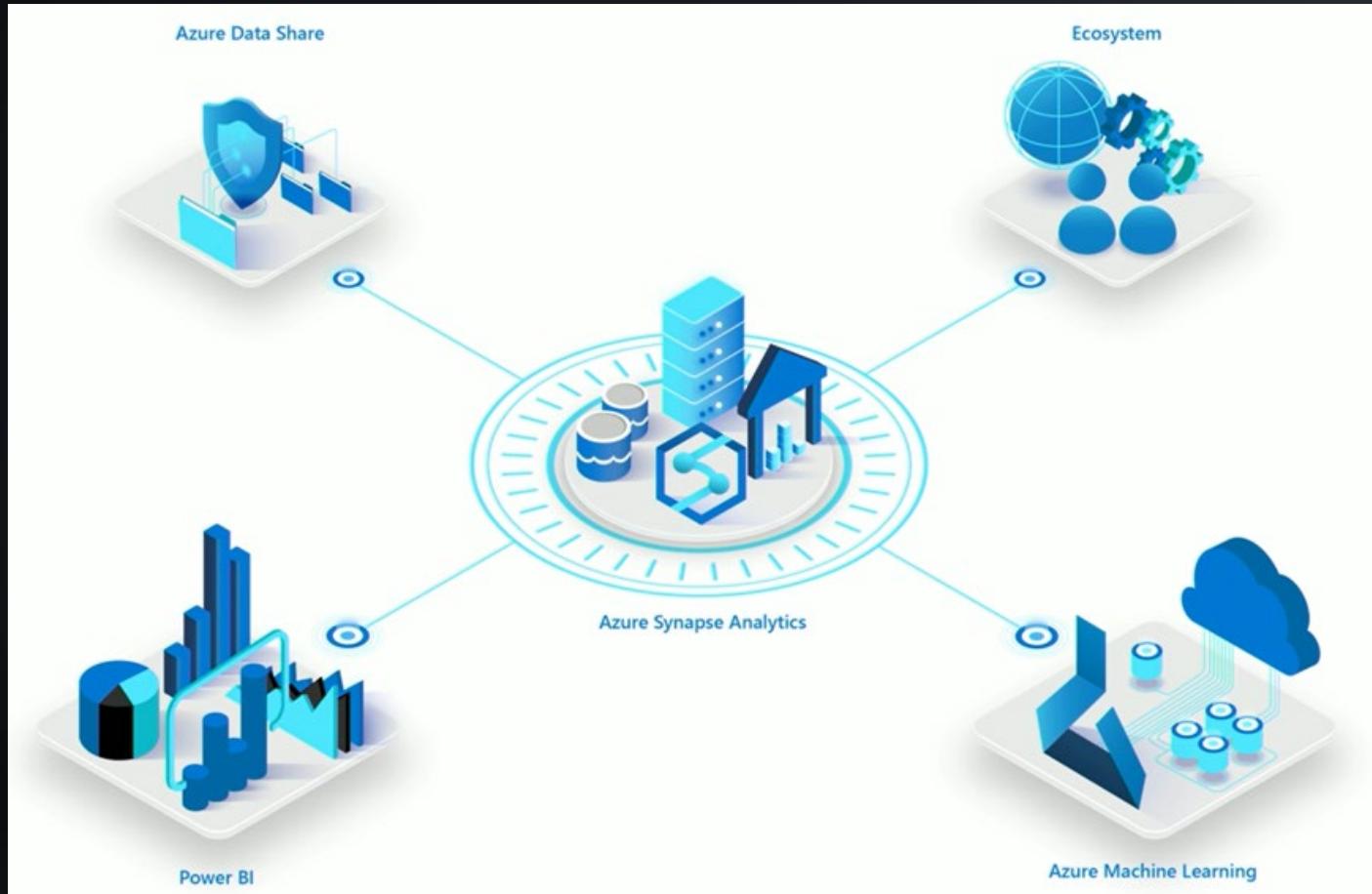


Microsoft

REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br



- Ecosistema completo de Ingestão, Tratamento, Análise e Machine Learning.
- Integrado ao Pool SQL DW.
- Permite rodar notebooks internos.

PATROCÍNIO:



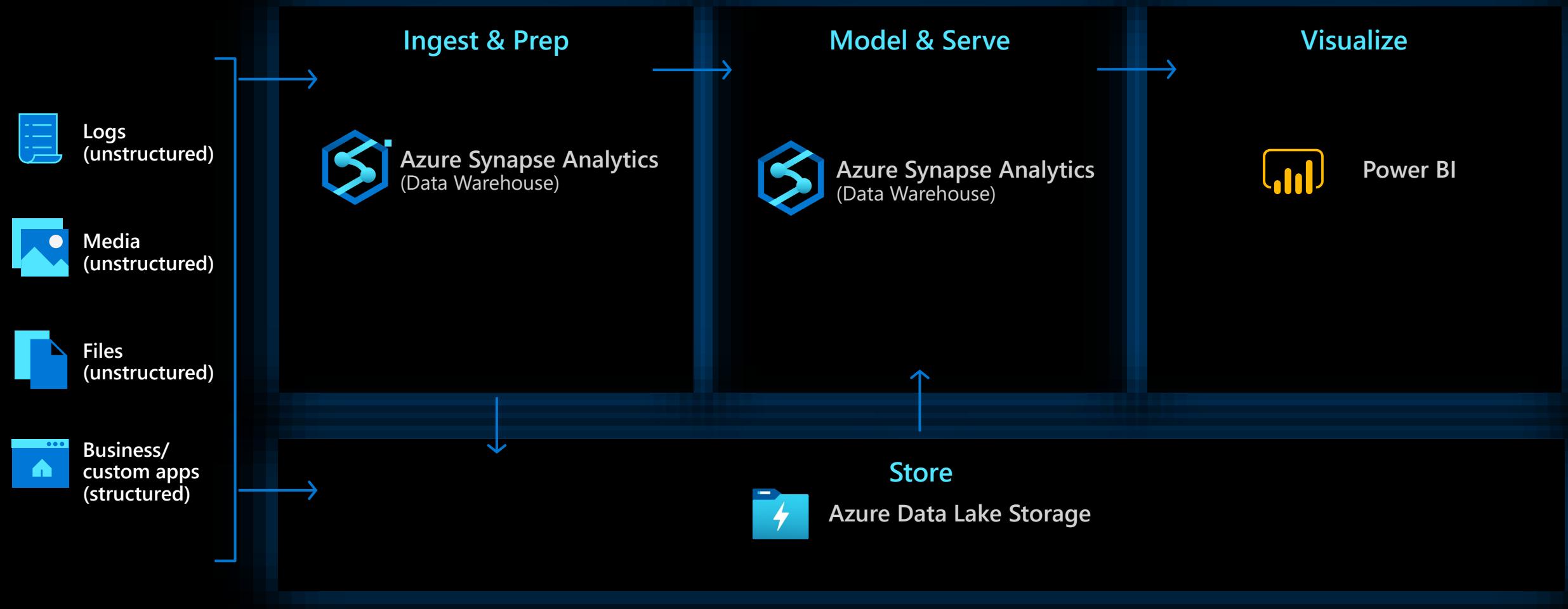
REALIZAÇÃO:

e.SethCloud

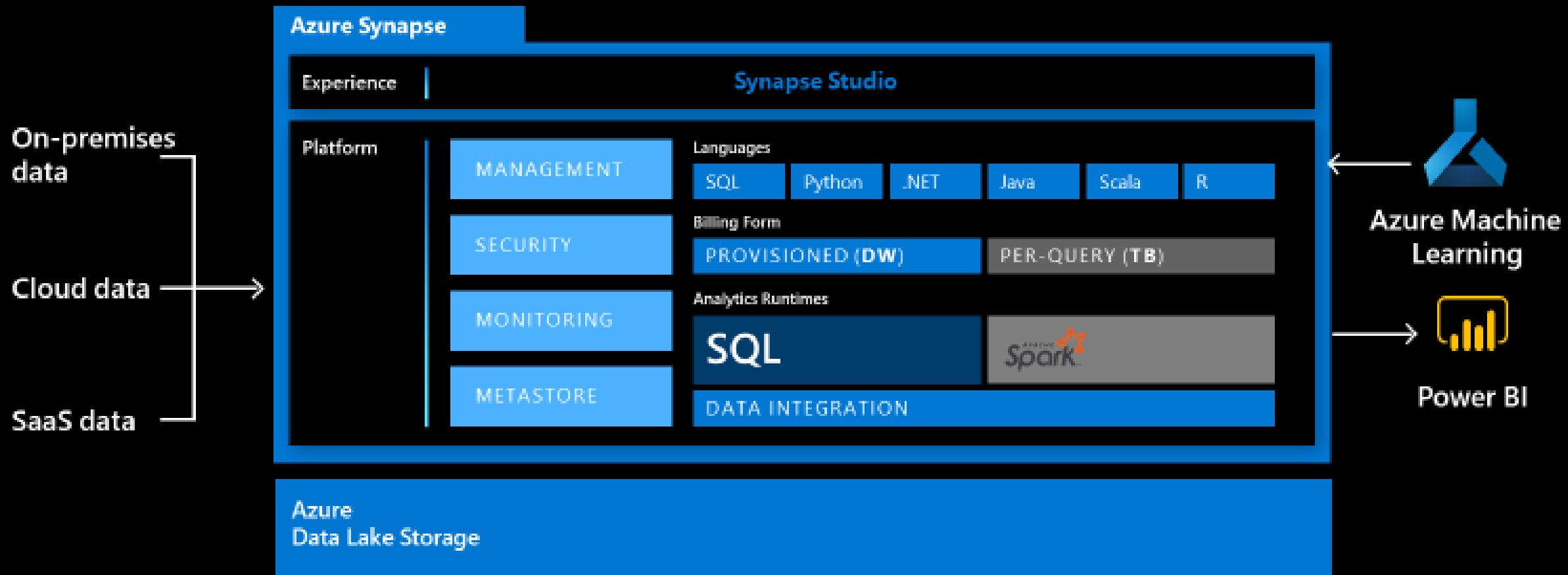
Azure
Academy

www.azureacademy.com.br

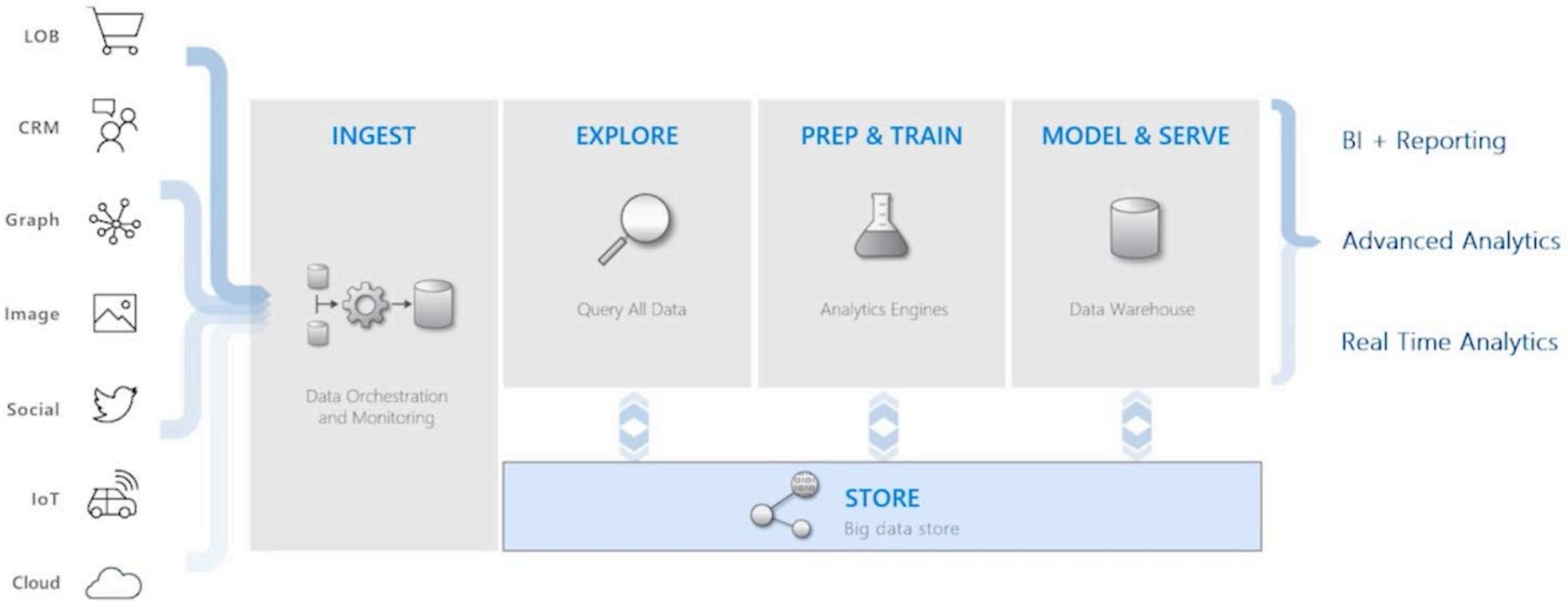
ETL + Analytics



Azure Synapse Analytics

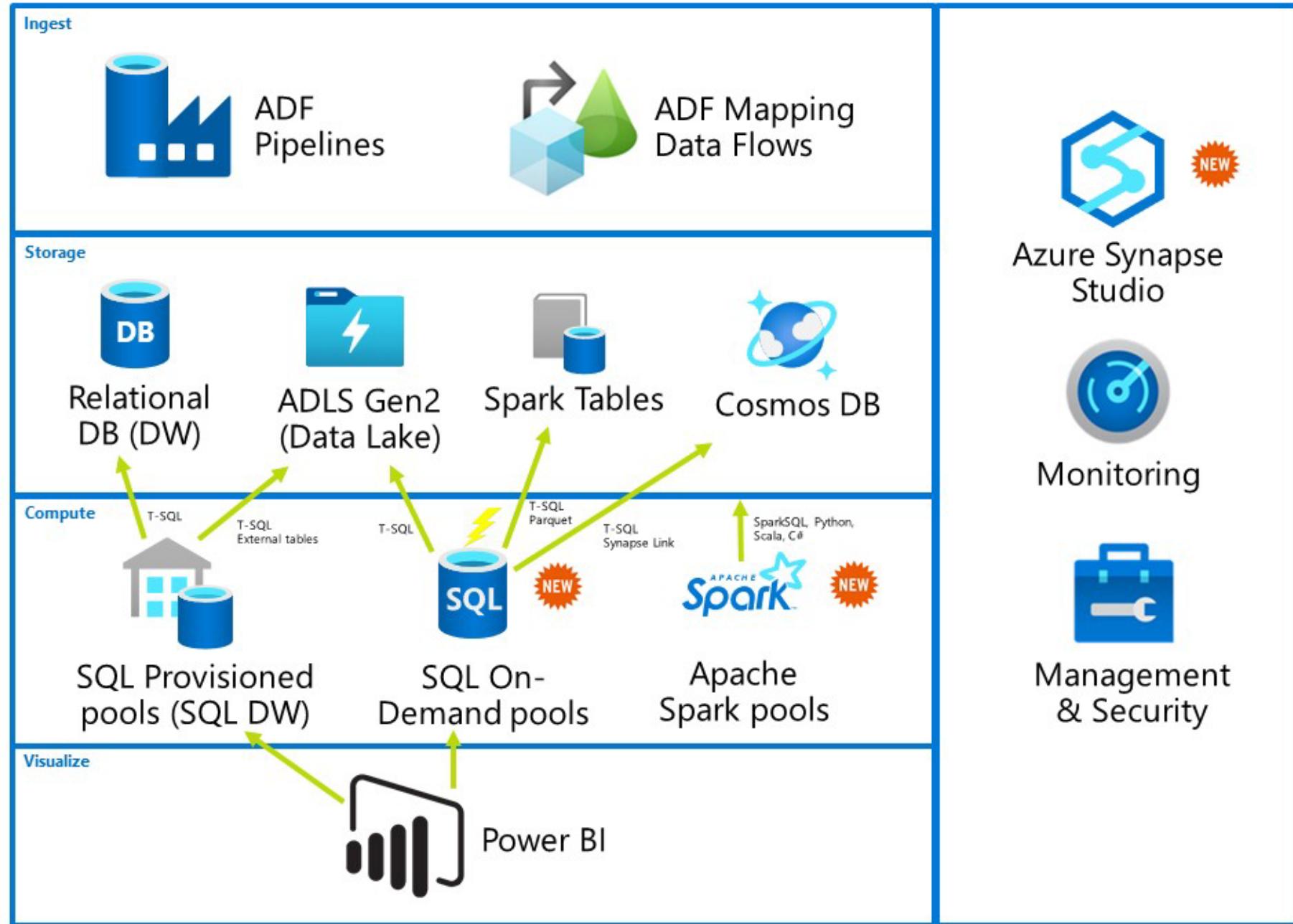


Modern Data Warehousing





Azure Synapse
Analytics
(workspaces)



DOCK
DATA
HACKATHON

DATABRICKS

PATROCÍNIO:



Microsoft

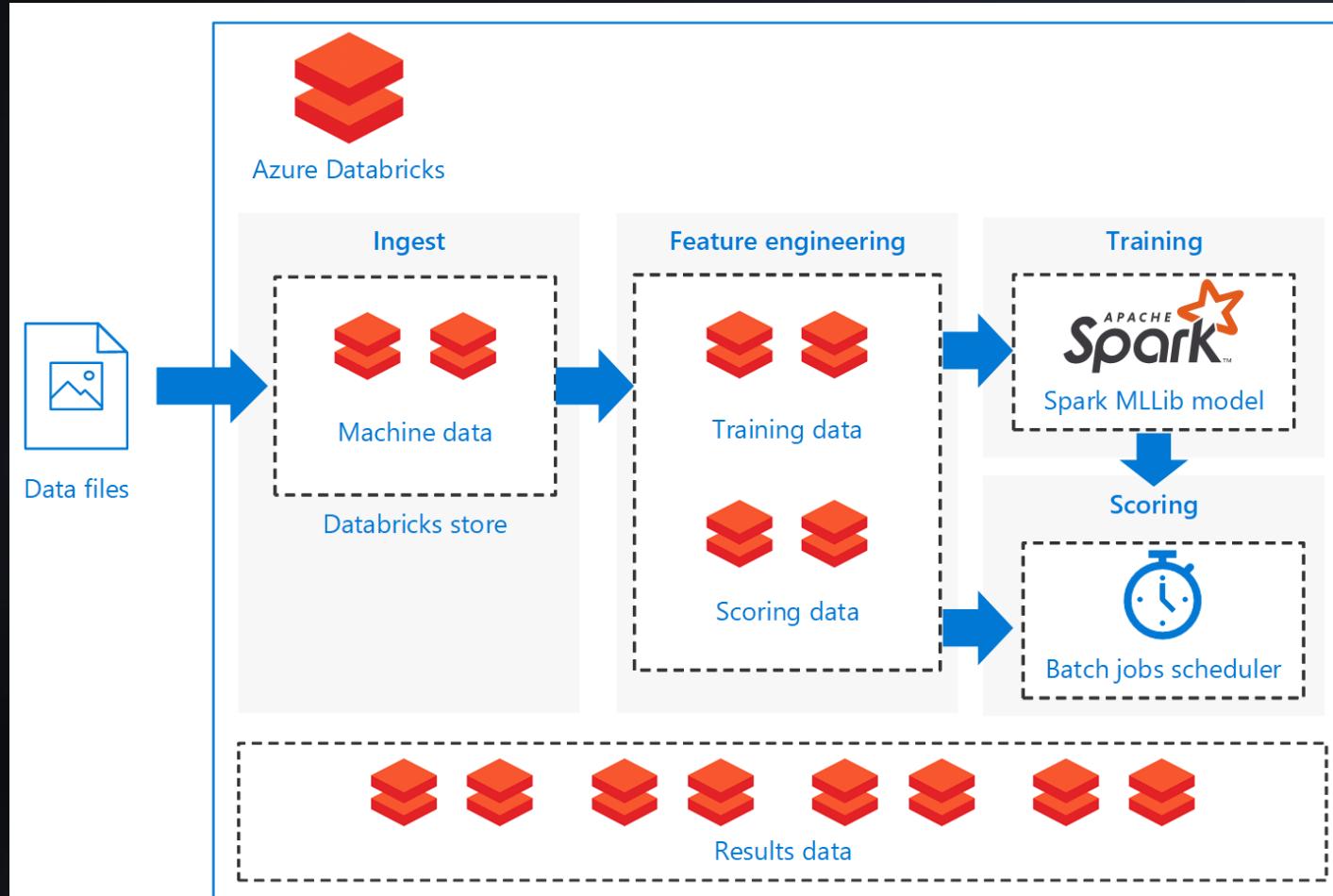
REALIZAÇÃO:

e.SethCloud

Azure
Academy
.com.br

DATABRICKS

Plataforma focada em orquestração e análise de dados em camada BIG Data.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

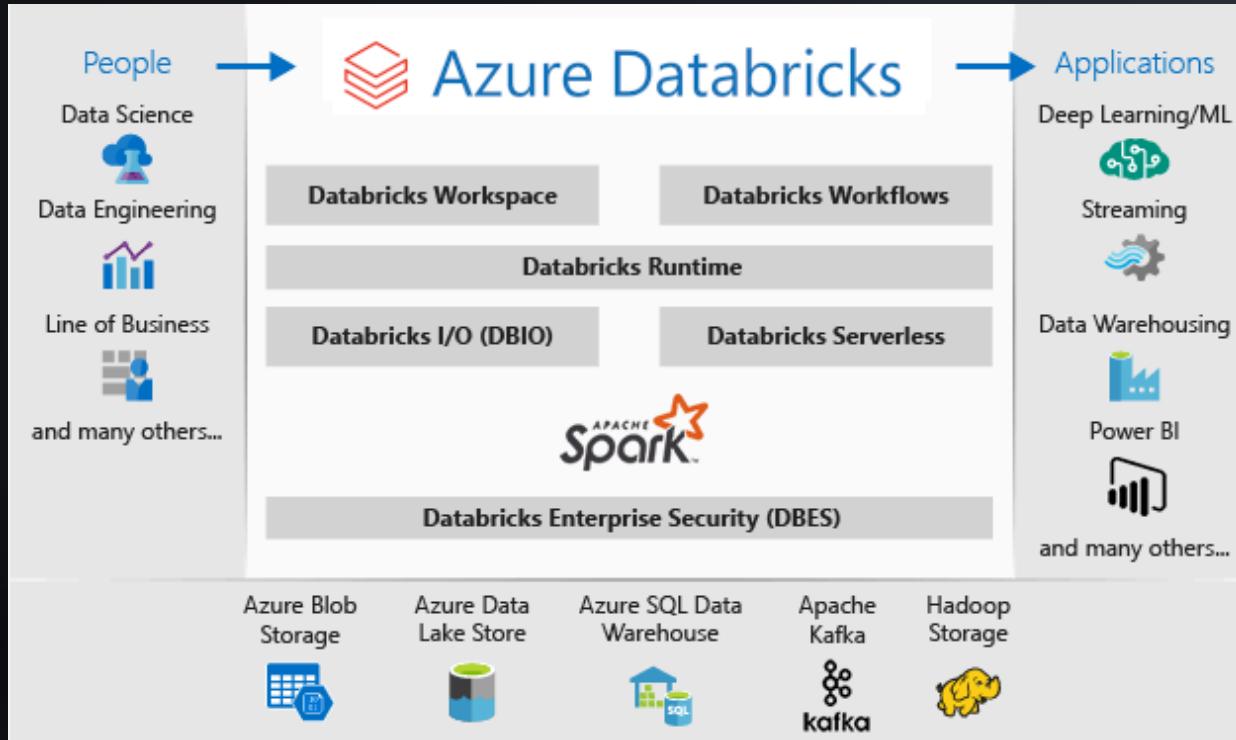
Azure
Academy

www.azureacademy.com.br

AZURE DATABRICKS - WORKSPACE

O Workspace do Azure Databricks é uma plataforma de análise baseada no Apache Spark.

No Azure, oferece facilidades colaborativa, desde a instalação a partir de poucos cliques até a colaboração entre engenheiros de dados, cientistas de dados e engenheiros de Machine Learning.



PATROCÍNIO:



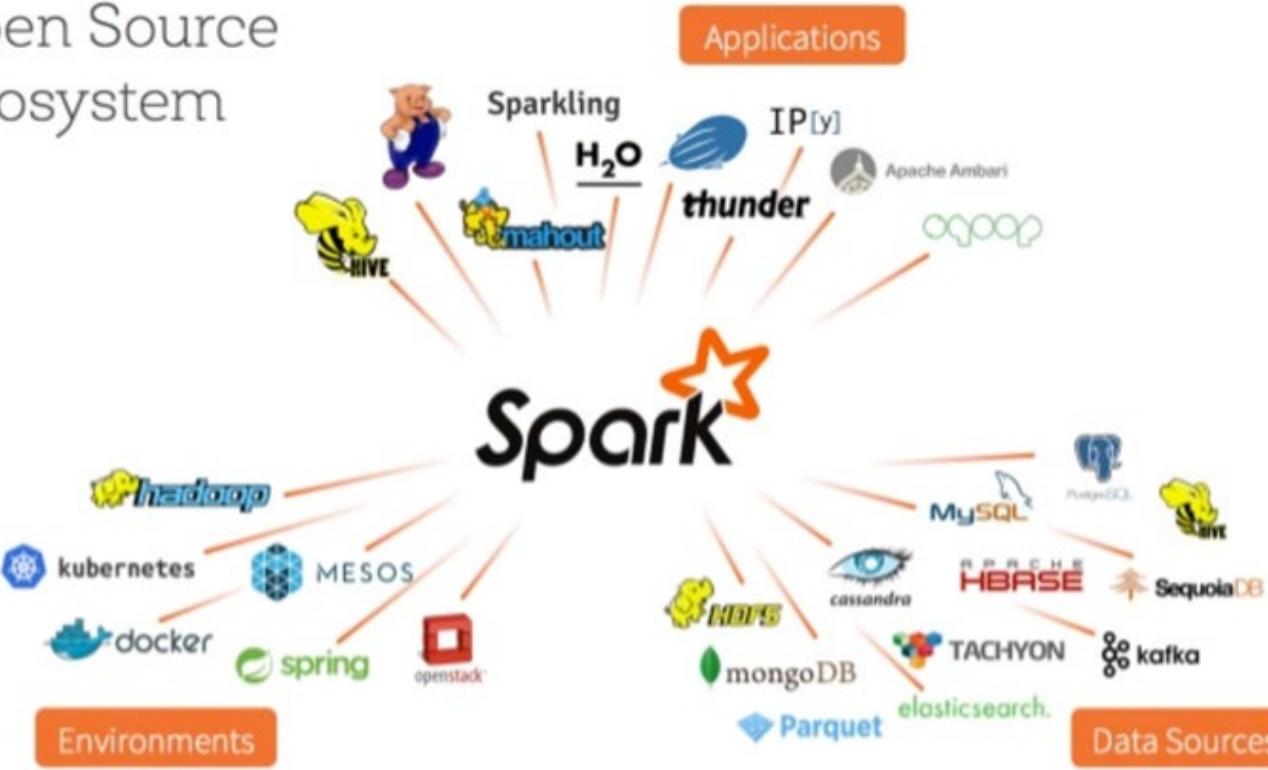
REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

Open Source Ecosystem



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

Conceitos

Pandas - DataFrame:

O Pandas é uma biblioteca do Python utilizada para análise e manipulação de dados e permite trabalhar de forma rápida e eficiente com arquivos em diversos formatos.

DataFrame é uma estrutura bidimensional de dados, como uma planilha. Contém colunas e linhas.

	Name	Team	Number
index	values	values	values
0	Avery Bradley	Boston Celtics	0.0
1	John Holland	Boston Celtics	30.0
2	Jonas Jerebko	Boston Celtics	8.0
3	Jordan Mickey	Boston Celtics	NaN
4	Terry Rozier	Boston Celtics	12.0
5	Jared Sullinger	Boston Celtics	7.0
6	Evan Turner	Boston Celtics	11.0

Artigo selecionado sobre manipulação de dados com Pandas:

<https://medium.com/data-hackers/uma-introdu%C3%A7%C3%A3o-simples-ao-pandas-1e15eea37fa1>

PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br



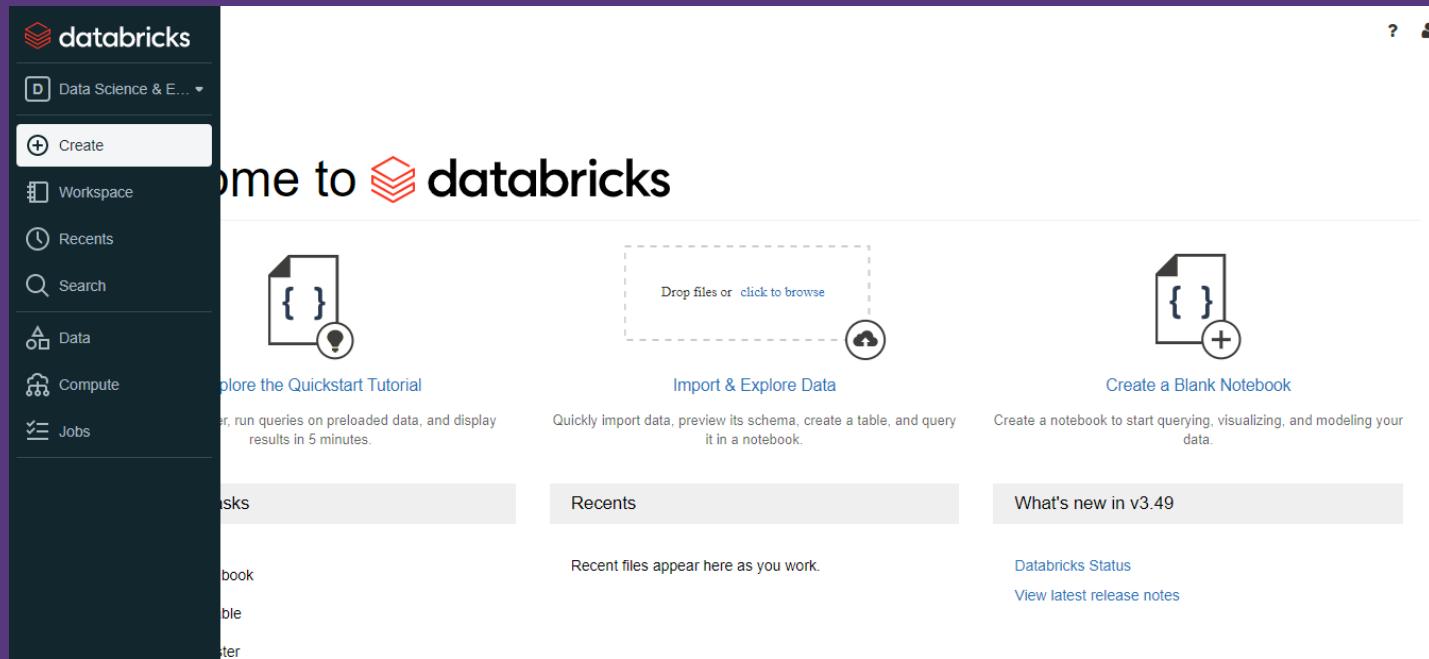
Laboratório

INGESTÃO, CONVERSÃO E SAÍDA PARA SQL SERVER

PROCESSAMENTO MULTI FORMATOS DATA LAKE

Lab – Workspace

1. Acesse o marketplace do Azure e instale o Azure Databricks na versão Premium 14 dias Free.
2. Como alternativa, você pode ativar uma conta de estudos através do link:
<https://community.cloud.databricks.com/>
3. Após a instalação, acesse o workspace do Databricks.



4. Acesse a guia Computação e crie um Cluster.

LAB – INGESTÃO, CONVERSÃO E SAÍDA SQL

DATA LAKE GEN 2 (Fonte)

1. Crie um container com o nome STAGING.
2. Faça o upload do arquivo ‘moviesdb.csv’ disponível no Portal do Aluno para o container.

SQL DO AZURE – PAAS (Destino)

1. Abra o editor de consultas no SQL do Azure ou estabeleça uma conexão através do SQL Management Studio e execute o script ‘lab2_tabela.txt’ para criar uma nova tabela de destino.



LAB – INGESTÃO, CONVERSÃO E SAÍDA SQL

DATABRICKS

1. Crie um novo Notebook do tipo Python. Script do lab disponível em 'databricks_notebook02.txt'.

Defina as variáveis que serão utilizadas na conexão com o Data Lake do Azure, substituindo o token pelo valor da sua conta de armazenamento do Azure:

```
blob_account_name = "storageteste123a"  
blob_container_name = "staging"  
blob_relative_path = "/"  
blob_sas_token = r"TOKEN_AQUI"  
arquivo = "moviesdb.csv"
```

Na sequência, estabeleça a conexão:

```
wasbs_path = 'wasbs://%s@%s.blob.core.windows.net/%s' % (blob_container_name, blob_account_name, blob_relative_path)  
spark.conf.set('fs.azure.sas.%s.%s.blob.core.windows.net' % (blob_container_name, blob_account_name), blob_sas_token)  
print('Remote blob path: ' + wasbs_path)
```



LAB – INGESTÃO, CONVERSÃO E SAÍDA SQL

DATABRICKS

1. Crie um Dataframe para ler todos os arquivos em formato CSV disponíveis no Data Lake, configurando a primeira linha como nome de colunas:

```
df = spark.read.format("csv").option("header", "true").option("mode",  
"DROPMALFORMED").load(wasbs_path)  
df.createOrReplaceTempView('source')
```

Verifique o resultado em tela:

```
display(spark.sql('SELECT * FROM source'))
```



LAB – INGESTÃO, CONVERSÃO E SAÍDA SQL

DATABRICKS – DATA MUNING

Nesta etapa, você poderia realizar tratamentos de dados, contagens, descartes e outras manipulações. Em nosso exemplo, avançaremos para a gravação dos dados em formato final no SQL do Azure:

Defina as variáveis de conexão para o driver do SQL do Azure, substituindo os valores:

```
from pyspark.sql import *
import pandas as pd
```

```
jdbcHostname = "SERVER_BANCO_AQUI"
jdbcPort = 1433
jdbcDatabase = "NOME_BANCO_AQUI"
username = "USUARIO_BANCO_AQUI"
password = "SENHA_BANCO_AQUI"
```

Estabeleça a conexão com o SQL do Azure:

```
jdbcUrl = "jdbc:sqlserver://{}:{};database={}".format(jdbcHostname, jdbcPort, jdbcDatabase)
print(jdbcUrl)
connectionProperties = {
    "user" : username,
    "password" : password,
    "driver" : "com.microsoft.sqlserver.jdbc.SQLServerDriver"}
```



LAB – INGESTÃO, CONVERSÃO E SAÍDA SQL

DATABRICKS

Faça um teste com um SELECT tradicional para verificar se o driver está conectado:

```
pushdown_query = "(select * from SalesLt.Product) Product"
df2 = spark.read.jdbc(url=jdbcUrl, table=pushdown_query, properties=connectionProperties)
display(df2)
```

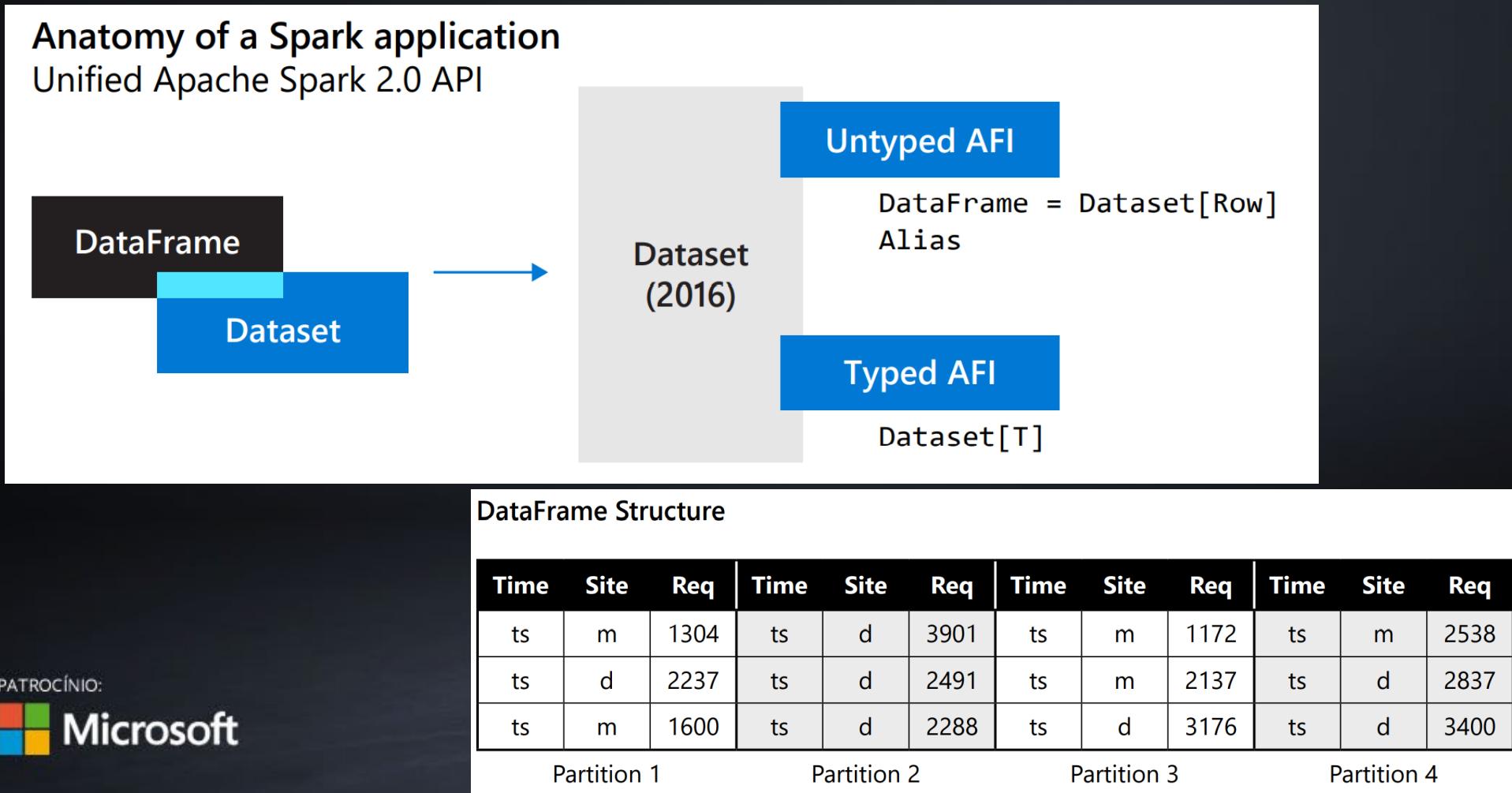
Ao final, utilize o comando a seguir para ler o conteúdo dos arquivos CSVs em formato SQL e realizar a importação dos dados para o SQL do Azure na tabela movies2 criada no início deste lab:

```
mydf = sqlContext.read.csv(wasbs_path,header=True)
myfinaldf = DataFrameWriter(mydf)
myfinaldf.jdbc(url=jdbcUrl, table= "movies2", mode ="overwrite", properties = connectionProperties)
```



DATAFRAMES

DataFrames são colunas de dados no Spark e impõem uma estrutura e esquema. Esta organização dita como processar os dados, expressar um cálculo ou emitir uma consulta. Por exemplo, seus dados podem ser distribuídos em quatro Partições RDD, cada partição com três colunas nomeadas: "Time," "Site," e "Req." Esta organização fornece uma maneira natural e intuitiva de acessar dados.



DATAFRAMES – MANIPULAR DADOS

- Juntar diversos dataframes em um único:

```
unionDF = df1.union(df2)
```

- Gravar a saída em disco ou banco de dados:

```
unionDF.write.parquet("/tmp/databricks-df-example.parquet")
```

- Filtrar dados com objetivos de Análise Exploratória ou Descarte:

```
filterDF = flattenDF.filter(flattenDF.firstName == "xiangrui").sort(flattenDF.lastName)
```

- Data Muning: substituir valores nulos por '--':

```
nonNullDF = flattenDF.fillna("--")
```

- Procurar por valores nulos em colunas específicas:

```
filterNonNullDF = flattenDF.filter(col("firstName").isNull() |  
col("lastName").isNull()).sort("email")
```

PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

DATAFRAMES – MANIPULAR DADOS

- **Group by, contagem e validação de dados com agregação:**

```
from pyspark.sql.functions import countDistinct
```

```
countDistinctDF = nonNullDF.select("firstName", "lastName")\
    .groupBy("firstName")\
    .agg(countDistinct("lastName").alias("distinct_last_names"))
```

- **Utilizar Spark SQL para definir comandos SQL na consulta do Dataframe:**

```
nonNullDF.createOrReplaceTempView("databricks_df_example")
```

```
countDistinctDF_sql = spark.sql("""
    SELECT firstName, count(distinct lastName) AS distinct_last_names
    FROM databricks_df_example GROUP BY firstName""")
```

PATROCÍNIO:



REALIZAÇÃO:

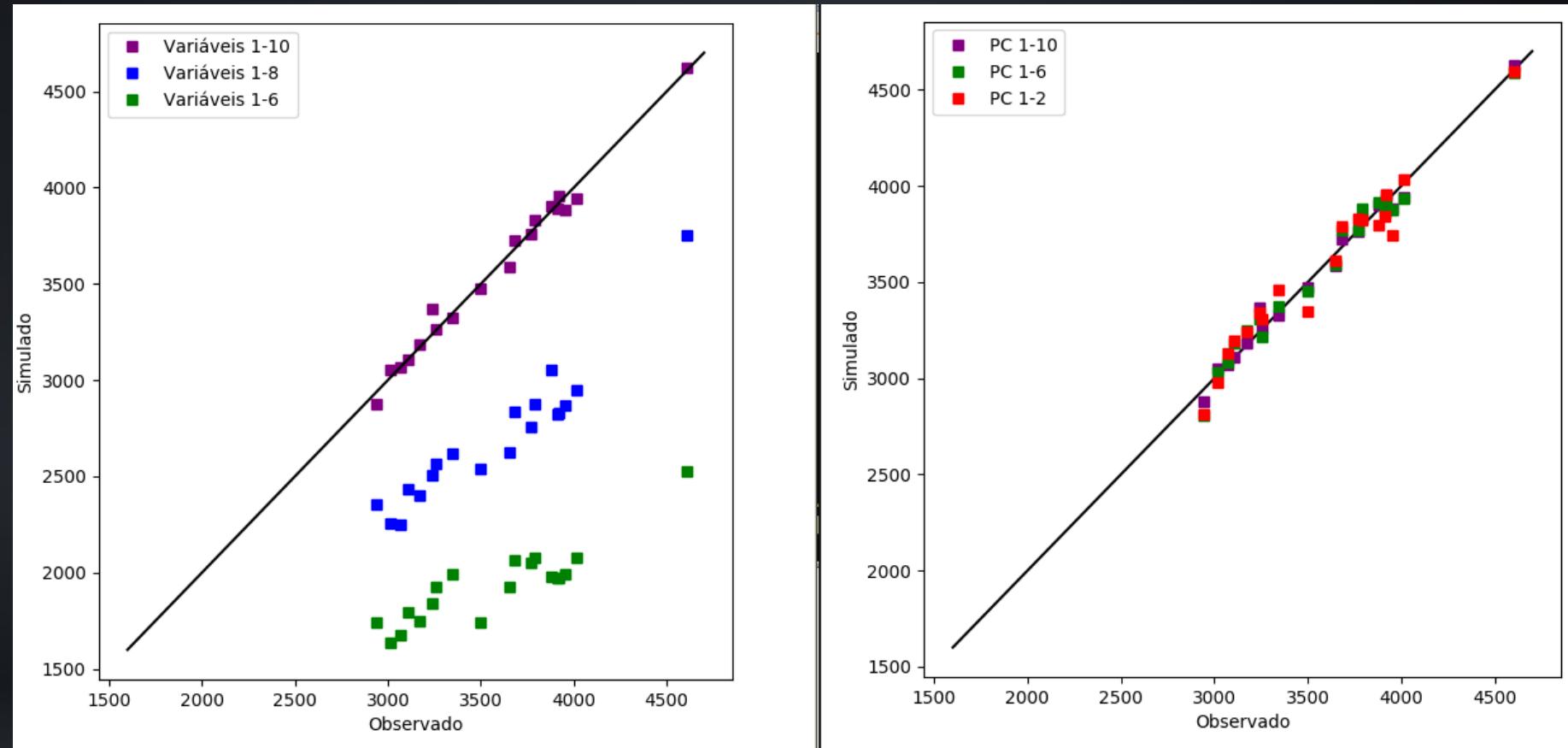
e.SethCloud

Azure
Academy

www.azureacademy.com.br

Análise Exploratória

Ajuda a identificar se os dados estão prontos qualitativamente ou se necessitam de ajustes.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br

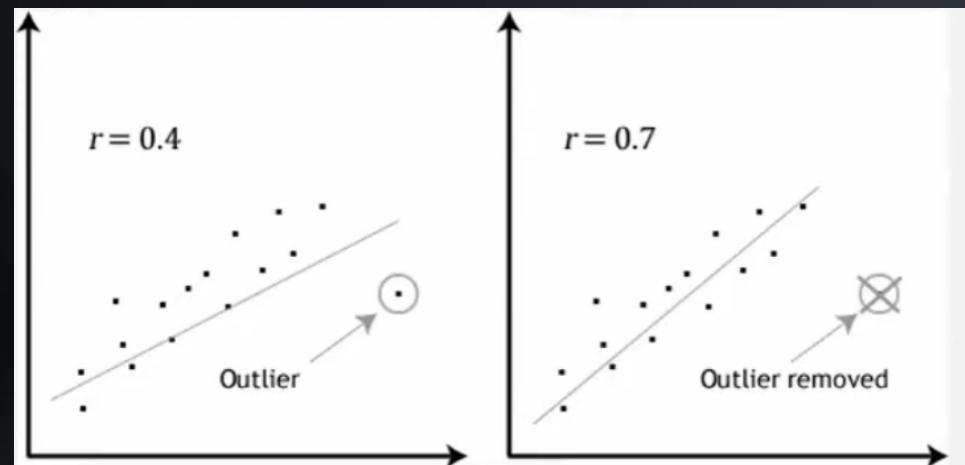
Análise Exploratória

Análise Exploratória de dados:

- Mínimo, Máximo, Média.

Valor discrepante:

- Outliers, pontos fora da curva. Calcular o desvio padrão.



PATROCÍNIO:



REALIZAÇÃO:

e.SethCloud

Azure
Academy

www.azureacademy.com.br



Laboratório

DATAFRAME

CRIAÇÃO, MANIPULAÇÃO E GRAVAÇÃO NO
STORAGE INTERNO DO DATABRICKS

LAB – DATAFRAMES

Notebook disponível no Portal do Aluno, Downloads, databricks_notebook03.txt

Siga as etapas para:

- Criar 02 dataframes com dados definidos no código ou carregados dinamicamente.
- Unir 02 dataframes em um único.
- Gravar a saída em storage interno do Databricks com formato parquet.
- Realizar análise exploratória.
- Filtros, count, somas e Agregações de dados.
- Data Mining para limpeza e normalização.
- Utilizar SQL para transacionamento.
- Verificar estatísticas dos dados.
- Exportar análise em formato gráfico.



LABS EXTRAS:

Notebook com PyTorch para redes neurais e análise de imagens:

[MLflow: Train with PyTorch - Databricks \(microsoft.com\)](#)

Registrar e disponibilizar um modelo para consumo como API:

[Modelo de MLflow servindo em Azure Databricks de Azure Databricks - Workspace | Microsoft Docs](#)

Streaming com EventHub do Azure:

[https://docs.microsoft.com/pt-br/azure/databricks/scenarios/databricks-stream-from-eventhubs](#)

Processando dados NoSQL Cosmos DB:

[https://docs.microsoft.com/pt-br/azure/databricks/scenarios/service-endpoint-cosmosdb](#)

Interação com SQL e Docker:

[https://docs.microsoft.com/pt-br/azure/databricks/scenarios/vnet-injection-sql-server](#)

Guia da linguagem Python:

[https://docs.microsoft.com/pt-br/azure/databricks/languages/python](#)

Labs adicionais

Extração e ETL com Databricks:

<https://docs.microsoft.com/pt-br/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse>

Machine Learning – Classificação com Databricks:

https://docs.microsoft.com/pt-br/azure/databricks/_static/notebooks/binary-classification.html

Machine Learning – Árvores de decisão:

https://docs.microsoft.com/pt-br/azure/databricks/_static/notebooks/decision-trees.html





Azure Academy

Rubens Guimarães
 [/rubensguimaraes](https://www.linkedin.com/in/rubensguimaraes)

2021 GLOBAL AWARDS:



MICROSOFT AZURE



ART. INTELLIGENCE



www.AzureAcademy.com.br

PATROCÍNIO E APOIO:



e.Seth Cloud

MAIS DE 100 TURMAS FORMADAS EM 11 PAÍSES.

BRASIL
EUA
CANADÁ
PORTUGAL
ANGOLA

MOÇAMBIQUE
ESPAÑA
FRANÇA
REINO UNIDO
AUSTRÁLIA

BÉLGICA



Azure Academy



@azure-Academy



@azureacademyoficial



@Azure_Academy



@azureacademyBR



AzureAcademy



www.AzureAcademy.com.br