© Copyright Microsoft Corporation. All rights reserved.

FOR USE <u>ONLY</u> AS PART OF MICROSOFT VIRTUAL TRAINING DAYS PROGRAM. THESE MATERIALS ARE <u>NOT</u> AUTHORIZED FOR DISTRIBUTION, REPRODUCTION OR OTHER USE BY NON-MICROSOFT PARTIES.



Microsoft Azure Virtual Training Day: Data Fundamentals







Objetivos de aprendizagem

- Principais conceitos de dados
- Funções e serviços de dados



Objetivo de aprendizagem: principais conceitos de dados

O que são dados?

Valores usados para registrar informações, geralmente representando *entidades* com um ou mais *atributos*

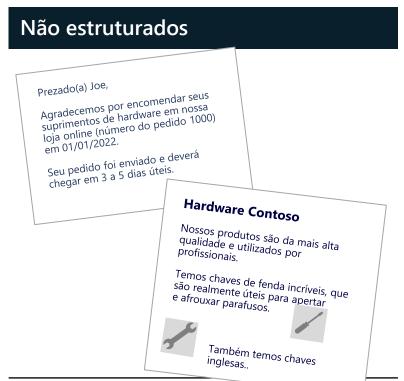
Estruturados

Cı	Customer				
ID	FirstName	LastName	Email	Address	
1	Joe	Jones	joe@litware.c om	1 Main St.	
2	Samir	Nadoy	samir@north wind.com	123 Elm Pl.	

Product		
ID	Name	Price
123	Martelo	2.99
162	Chave de Fenda	3,49
201	Chave Inglesa	4,25

Semi-estruturados

```
"firstName": "Joe",
    "lastName": "Jones",
    "address":
        "streetAddress": "1 Main
                                        "firstName": "Samir",
St.",
                                        "lastName": "Nadoy",
        "city": "New York",
                                        "address":
        "state": "NY",
        "postalCode": "10099"
                                             "streetAddress": "123 Elm
    },
    "contact":
                                             "unit": "500",
                                            "city": "Seattle",
                                            "state": "WA",
          "type": "home",
                                            "postalCode": "98999"
          "number": "555 123-1234'
                                         "contact":
          "type": "email",
          "address":
                                               "type": "email".
"joe@litware.com'
                                              "address":
                                     "samir@northwind.com"
```



Como os dados são armazenados?

Arquivos

Texto delimitado

```
FirstName, LastName, Email
Joe, Jones, joe@litware.com
Samir, Nadoy, samir@northwind.com
```

JavaScript Object Notation (JSON)

```
"customers":
  { "firstName": "Joe", "lastName": "Jones"},
   "firstName": "Samir", "lastName": "Nadoy"}
```

Extensible Markup Language (XML)

<Customer firstName="Joe" lastName="Jones"/>

Objeto binário grande (BLOB)

10110101101010110010...

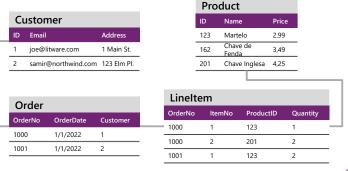
Formatos otimizados:

Avro, ORC, Parquet

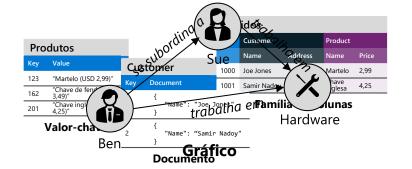
Bancos de dados

Relacionais





Não relacionais



Workloads de dados operacionais

Os dados são armazenados em um banco de dados otimizado para operações de *processamento transacional online* (OLTP) compatíveis com aplicações

Uma mistura de atividade de leitura e de gravação

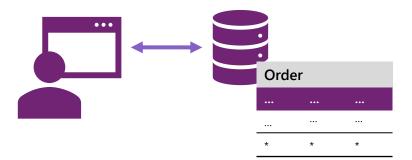
Por exemplo:

- Leia a tabela Product para exibir um catálogo
- Grave na tabela Order para registrar uma compra

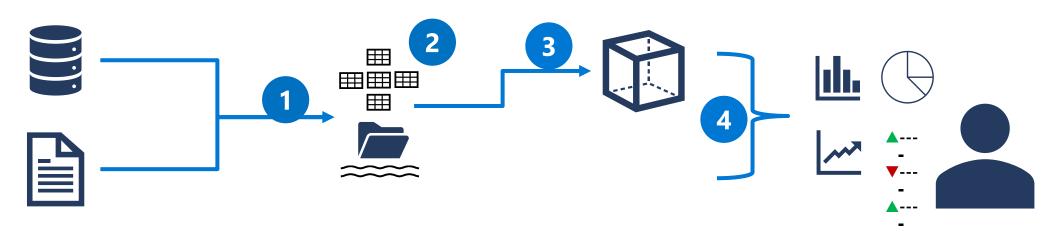
Os dados são armazenados usando transações

As transações se baseiam em "ACID":

- Atomicidade: cada transação é tratada como uma única unidade de trabalho, que tem sucesso total ou fracasso total
- Consistência: as transações só podem mover os dados no banco de dados de um estado válido para outro
- Isolamento: transações concorrentes não podem interferir uma na outra
- **Durabilidade**: quando uma transação tiver sido bem-sucedida, as alterações persistirão no banco de dados



Workloads de dados analíticos



- 1. Os dados operacionais são extraídos, transformados e carregados (ETL) em um *data lake* para análise
- 2. Os dados são carregados em um esquema de tabelas, geralmente com base em Spark data lakehouse com abstrações tabulares, em arquivos no data lake, ou em um data warehouse com um mecanismo SQL completamente relacional
- Os dados nas tabelas podem ser agregados e carregados em um modelo de processamento online analítico (OLAP) ou cubo
- 4. Os arquivos no data lake, as tabelas relacionais e o modelo analítico podem ser consultados para produzir *relatórios* e *painéis*



Objetivo de aprendizagem: funções e serviços de dados

Funções de profissionais de dados



Provisionamento, configuração e gerenciamento de bancos de dados

Segurança do banco de dados e acesso do usuário

Backups e resiliência de bancos de dados

Monitoramento e otimização de performance de bancos de dados



Pipelines de integração de dados e processos de ETL

Limpeza e transformação de dados

Esquemas de armazenamento de dados analíticos e cargas de dados



Modelagem analítica

Relatórios e resumos de dados

Visualização de dados

Serviços de nuvem da Microsoft para dados

Workloads de dados operacionais



SQL Azure

 Família de serviços de bancos de dados relacionais baseados no SQL Server



Banco Open Source no Azure

Maria DB, MySQL, PostgreSQL



Azure Cosmos DB

• Banco de dados não relacional e vetorial altamente escalável



Armazenamento do Azure

- Armazenamento de arquivos, blobs e tabelas
- Namespace hierárquico para armazenamento do data lake

Workloads de dados analíticos

Software como serviço (SaaS)



Microsoft Fabric

Análise integrada de ponta a ponta:

- Ingestão de dados e ETL
- Data lakehouse
- Data warehouse
- Ciência de dados e ML
- Análise em tempo real
- Visualização de dados
- Gerenciamento e governança de dados

Microsoft Purview



Solução para governança e descoberta de dados em toda a empresa:

- Crie um mapa de seus dados e rastreie a linhagem de dados em várias fontes de dados.
- Imponha a governança de dados em toda a empresa e garanta a integridade dos dados.

Plataforma como Serviço (PaaS)



Azure Databricks

 Análise e processamento de dados do Apache Spark Lakehouse

outros...





Objetivos de aprendizagem

- Explorar os conceitos dos dados relacionais
- Explorar os serviços do Azure para dados relacionais



Objetivo de aprendizagem: explorar os conceitos dos dados relacionais

Tabelas relacionais

- · Os dados são armazenados em tabelas
- As tabelas consistem em linhas e colunas
- · Todas as linhas têm as mesmas colunas
- · Cada coluna recebe um tipo de dados

Customer						
ID	FirstName	Middle	LastName	Email	Address	City
1	Joe	David	Jones	joe@litware.com	1 Main St.	Seattle
2	Samir		Nadoy	samir@northwind.com	123 Elm Pl.	New York

Product			
ID	Name	Price	
123	Martelo	2.99	
162	Chave de Fenda	3,49	
201	Chave Inglesa	4,25	

Order			
OrderDate	Customer		
1/1/2022	1		
1/1/2022	2		
	1/1/2022		

LineItem				
OrderNo	ItemNo	ProductID	Quantity	
1000	1	123	1	
1000	2	201	2	
1001	1	123	2	

Normalização

Sales Data				
OrderNo	OrderDate	Customer	Product	Quantity
1000	1/1/2022	Joe Jones, 1 Main St, Seattle	Martelo (USD 2,99)	1
1000	1/1/2022	Joe Jones- 1 Main St, Seattle	Chave de fenda (USD 3,49)	2
1001	1/1/2022	Samir Nadoy, 123 Elm Pl, New York	Martelo (USD 2,99)	2

- · Separe cada entidade em sua própria tabela
- · Separe cada atributo discreto em sua própria coluna
- · Identifique exclusivamente cada instância de entidade (linha) usando uma chave primária
- · Use as colunas de *chave estrangeira* para vincular entidades relativas



Customer				
ID	FirstName	LastName	Address	City
1	Joe	Jones	1 Main St.	Seattle
2	Samir	Nadoy	123 Elm Pl.	New York

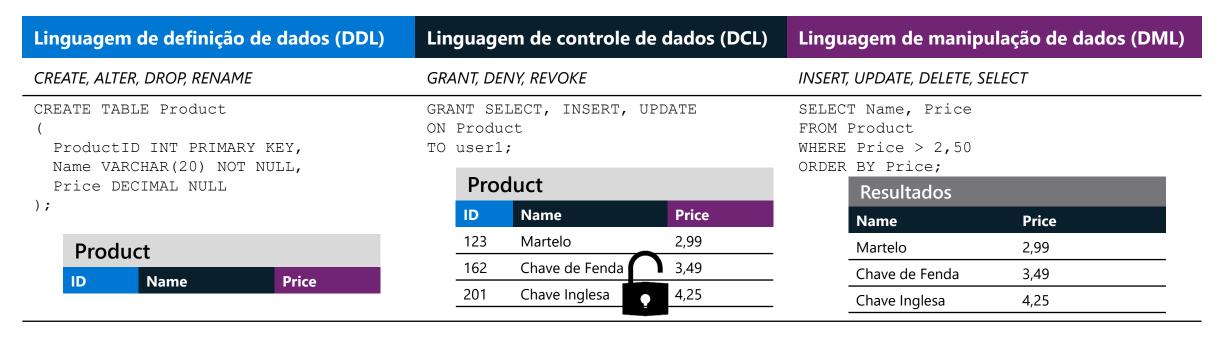
Order				
OrderNo	OrderDate	Customer		
1000	1/1/2022	1		
1001	1/1/2022	2		
		-		

LineItem				
OrderNo	ItemNo	ProductID	Quantity	
1000	1	123	1	
1000	2	201	2	
1001	1	123	2	

Product			
ID	Name	Price	
123	Martelo	2.99	
162	Chave de Fenda	3,49	
201	Chave Inglesa	4,25	

Structured Query Language (SQL)

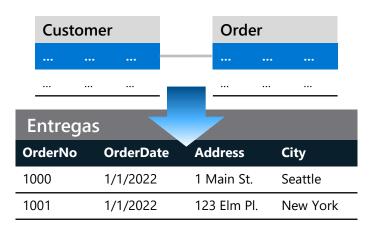
- · O SQL é uma linguagem padrão para uso com bancos de dados relacionais
- Os padrões são mantidos por ANSI e ISO
- · A maioria dos sistemas de RDBMS é compatível com extensões proprietárias do SQL padrão



Outros objetos de banco de dados comuns

Visualizações

Consultas SQL predefinidas que se comportam como tabelas virtuais



Procedimentos armazenados

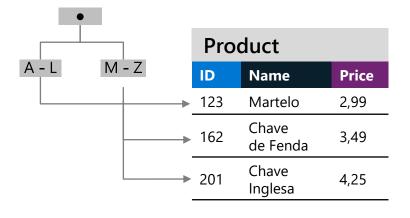
Instruções SQL predefinidas que podem incluir parâmetros

Product		
ID	Name	Price
201	Chave inglesa Chave inglesa	4,25

Índices

Estruturas baseadas em árvores que melhoram a performance da consulta

CREATE INDEX idx_ProductName
ON Product(Name);





Objetivo de aprendizagem: explorar os serviços do Azure para dados relacionais

SQL Azure



Família de serviços de bancos de dados de nuvem baseados no SQL Server



SQL Server em VMs do Azure

- Compatibilidade garantida com o SOL Server na infraestrutura local
- O cliente gerencia tudo: upgrades do sistema operacional, atualizações de software, backups, replicação
- Paque pelos custos de execução e licenciamento de software da VM do servidor, não por banco de dados
- Ótimo para nuvem híbrida ou migração de configurações complexas de bancos de dados na infraestrutura local



Instância Gerenciada SQL do Azure

- Compatibilidade quase total com o SOL Server na infraestrutura local
- Backups automáticos, patches de software, monitoramento de banco de dados e outras tarefas de manutenção
- Use uma única instância com vários bancos de dados ou várias instâncias em um pool com recursos compartilhados
- Ótimo para migrar a maioria dos bancos de dados SQL Server para a nuvem



Banco de Dados SQL do Azure

- Compatibilidade de funcionalidade de banco de dados principal com o SQL Server
- Backups automáticos, patches de software, monitoramento de banco de dados e outras tarefas de manutenção
- Banco de dados individual ou pool elástico para compartilhar de forma dinâmica recursos entre vários bancos de dados
- Ótimo para aplicações novas baseadas em nuvem

laaS

PaaS

Serviços de Banco de dados do Azure para Open Source

Soluções gerenciadas do Azure para RDBMS Open Source comuns



Banco de Dados do Azure para PostgreSQL

- Serviço de banco de dados no Microsoft Cloud baseado no mecanismo de banco de dados do PostgreSQL Community Edition
- Armazenamento de objetos e relacional híbrido



Banco de Dados do Azure para MySQL

- Implementação de PaaS do MySQL na nuvem do Azure, com base na MySQL Community Edition
- Comumente usada em arquiteturas de aplicações Linux, Apache, MySQL, PHP (LAMP)

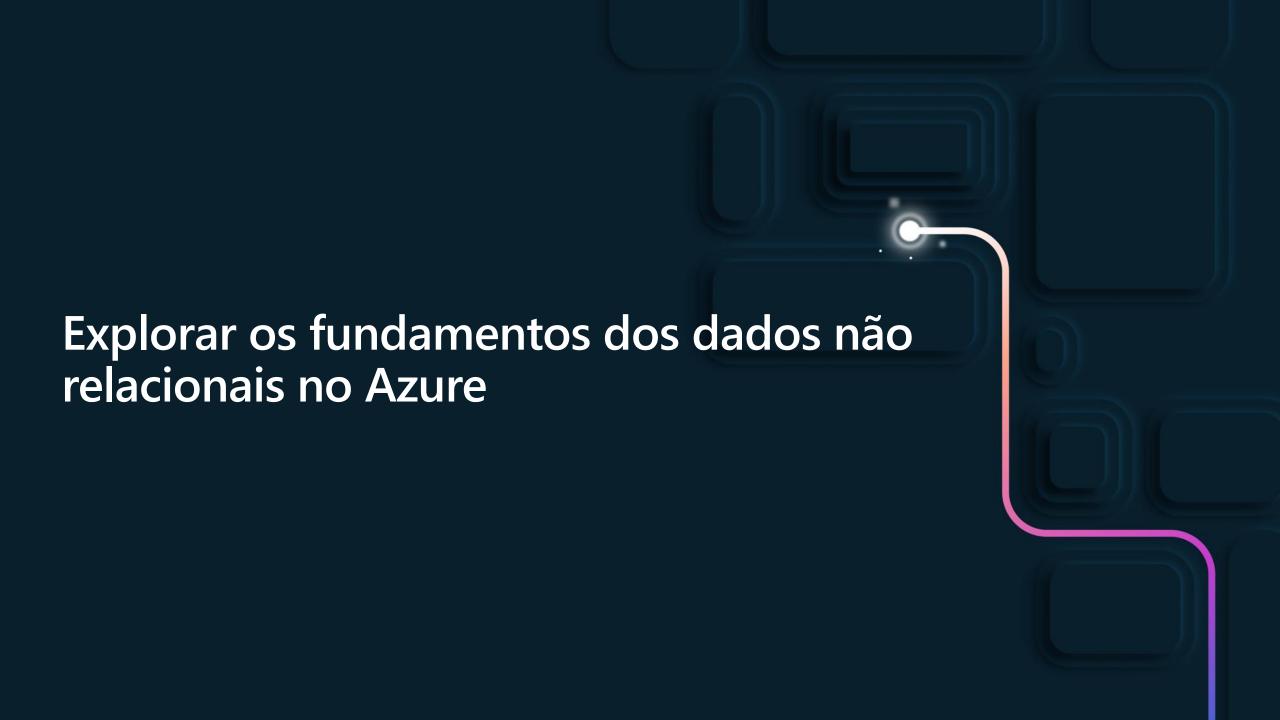


Banco de dados do Azure para MariaDB

 Uma implementação do sistema de gerenciamento de banco de dados MariaDB Community Edition adaptado para ser executado no Azure

Demonstração

 Laboratório: provisionar serviços de bancos de dados relacionais do Azure





Objetivos de aprendizagem

- Fundamentos do Armazenamento do Azure
- Fundamentos do Azure Cosmos DB



Objetivo de aprendizagem: fundamentos do Armazenamento do Azure

Armazenamento de Blobs do Azure

Armazenamento de dados como objetos binários grandes (BLOBs)

Blobs de blocos

- · Objetos binários grandes, discretos que mudam com pouca frequência
- Os blobs podem ter até 4,7 TB, compostos por blocos de até 100 MB
 - Um blob pode conter até 50.000 blocos

Blobs de página

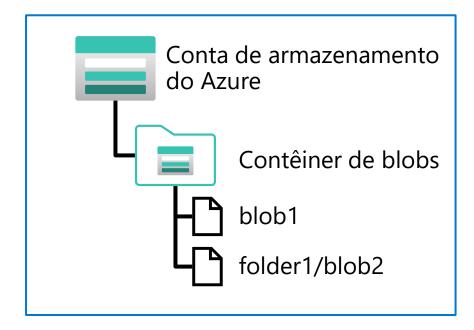
- Usados como armazenamento de disco virtual para VMs
- Os blobs podem ter até 8 TB, compostos de páginas de tamanho fixo de 512 bytes

Blobs de acréscimo

- Blobs de blocos que são usados para otimizar as operações de acréscimo
- Tamanho máximo de pouco mais de 195 GB (cada bloco pode ter até 4 MB)

Camadas de armazenamento por blob

- Hot: maior custo, menor latência
- Cool: custo mais baixo, latência mais alta
- Archive: menor custo, maior latência



Os blobs podem ser organizados em diretórios virtuais, mas cada caminho é considerado um único blob em um namespace plano (não há suporte a operações no nível da pasta)

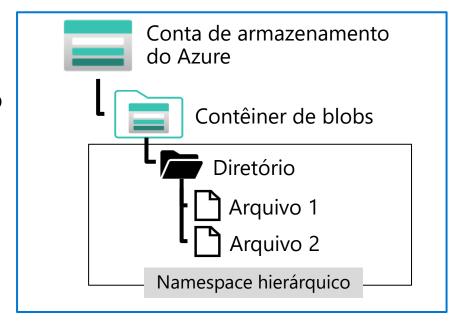
Azure Data Lake Store Gen 2

Sistema de arquivos distribuídos criado no Armazenamento de Blobs

- Combina o Azure Data Lake Store Gen 1 com o Armazenamento de Blobs do Azure para análise e armazenamento de arquivos em grande escala
- Habilita o gerenciamento e o controle de acesso em nível de arquivo e diretório
- Compatível com sistemas analíticos comuns em grande escala

Habilitado em uma conta de armazenamento do Azure por meio da opção *Namespace hierárquico*

- Definir durante a criação de conta
- Atualizar a conta de armazenamento existente
- Processo de atualização unidirecional

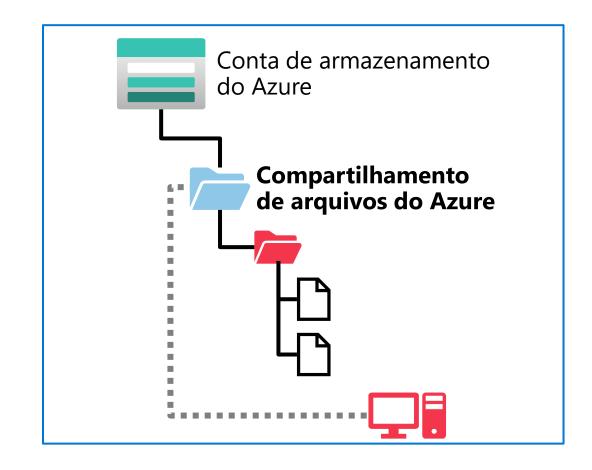


O sistema de arquivos inclui diretórios e arquivos e é compatível com sistemas de análise de dados em grande escala, como Databricks

Azure Files

Compartilhamentos de arquivos na nuvem que podem ser acessados de qualquer lugar com uma conexão de Internet

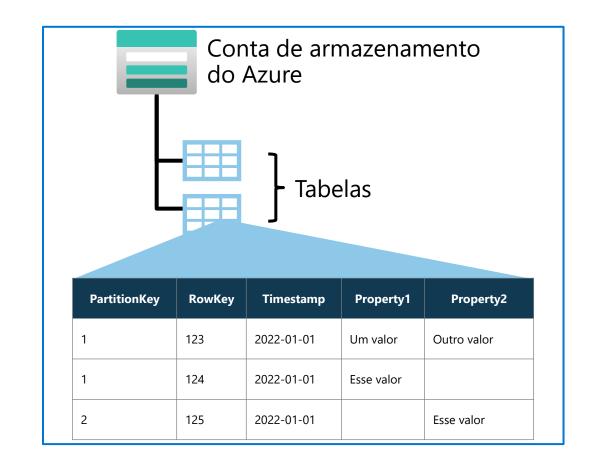
- Suporte para protocolos comuns de compartilhamento de arquivos:
 - Bloco de mensagens de servidor (SMB)
 - O Network File System (NFS) requer camada premium
- Os dados são replicados para redundância e criptografados em repouso



Armazenamento de Tabelas do Azure

Armazenamento Valor-chave para dados de aplicações

- As tabelas consistem em colunas chave e valor
 - Chaves de partição e de linha
 - Colunas de propriedade personalizada para valores de dados
 - Uma coluna Timestamp is adicionada automaticamente para registrar alterações de dados
- As linhas são agrupadas em partições para melhorar a performance
- As colunas de propriedade são atribuídas a um tipo de dados e podem conter qualquer valor desse tipo
- As linhas não precisam incluir as mesmas colunas de propriedade



Demonstração

• Laboratório: explorar o armazenamento do Azure

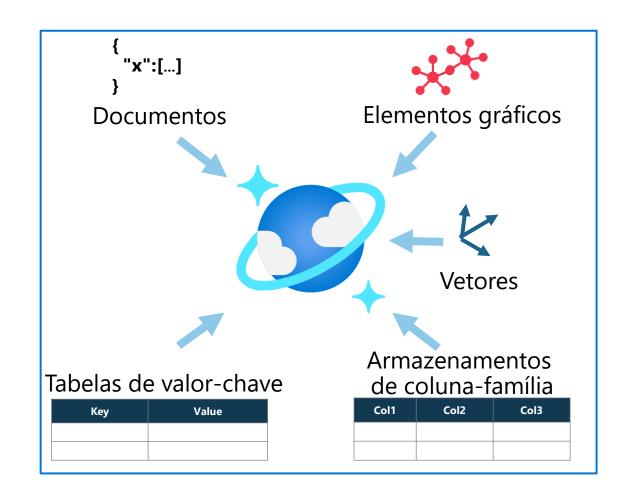


Objetivo de aprendizagem: fundamentos do Azure Cosmos DB

O que é o Azure Cosmos DB?

Banco de dados vetorial e NoSQL totalmente gerenciado para aplicações modernas

- Compatível com várias APIs para desenvolvimento de aplicações
- Acesso em tempo real com performance rápida de leitura e gravação
- Habilite gravações de várias regiões para replicar os dados globalmente. Habilite usuários em regiões especificadas para trabalhar com uma réplica local



APIs do Azure Cosmos DB

Azure Cosmos DB for NoSQL

• API nativa para Cosmos DB

SELECT *
FROM customers c
WHERE c.id =
"joe@litware.com"

```
{
    "id": "joe@litware.com",
    "name": "Joe Jones",
    "address": {
        "street": "1 Main St.",
        "city": "Seattle"
    }
}
```

Azure Cosmos DB for Table

- API de armazenamento de valor-chave
- Compatível com o armazenamento de tabelas do Azure

PartitionKey	RowKey	Name
1	123	Joe Jones
1	124	Samir Nadoy

Azure Cosmos DB for MongoDB

Compatibilidade com MongoDB

db.products.find({ id: 123})

```
{
    "id": 123,
    "name": "Hammer",
    "price": 2.99}
}
```

Azure Cosmos DB for Apache Cassandra

 Compatibilidade com o Apache Cassandra

id	name	dept	manager
1	Sue Smith	Hardware	
2	Ben Chan	Hardware	Sue Smith

Azure Cosmos DB for PostgreSQL

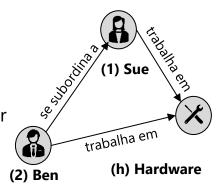
Compatibilidade com PostgreSQL

id	name	dept	manager
1	Sue Smith	Hardware	Joe Jones
2	Ben Chan	Hardware	Sue Smith



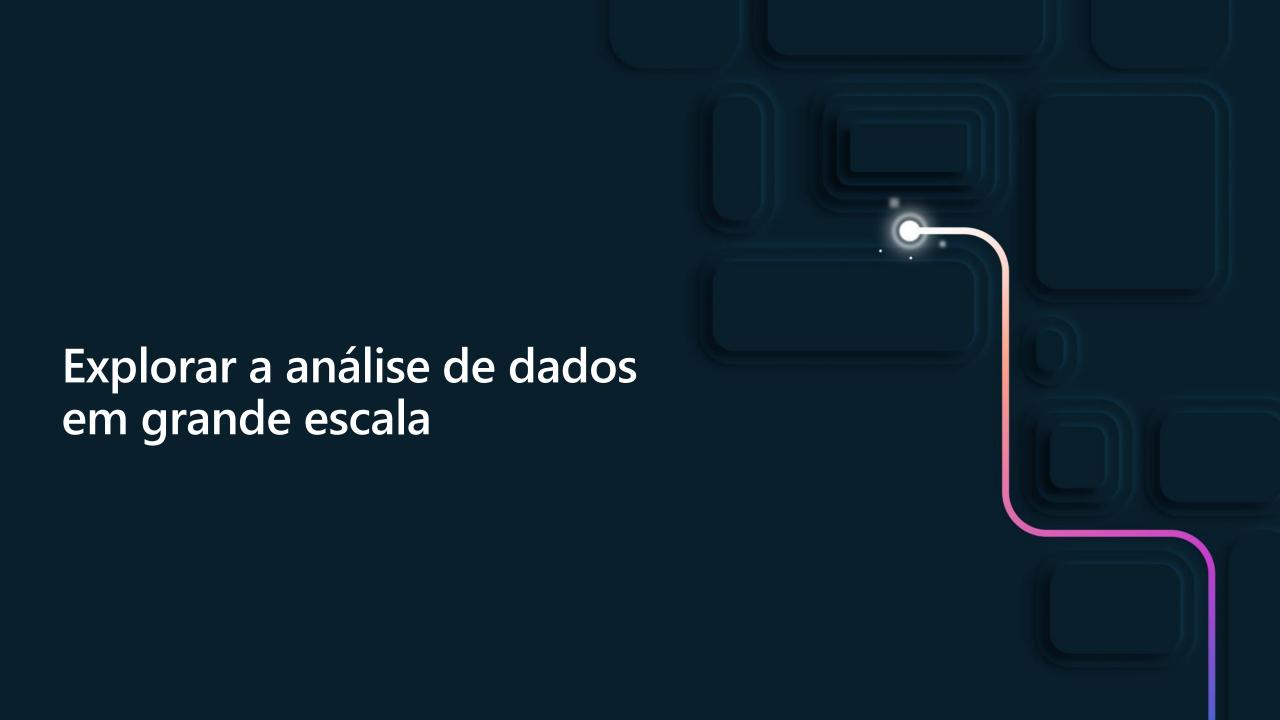
 Usado para trabalhar com dados gráficos

 vértices são conectados por meio de relacionamentos (bordas)



Demonstração

• Laboratório: Explorar o Azure Cosmos DB





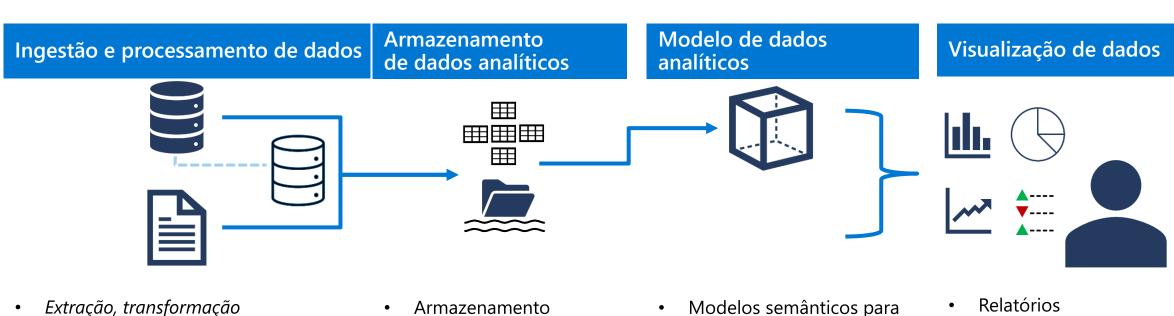
Objetivos de aprendizagem

• Análise de dados em grande escala



Objetivo de aprendizagem: análise de dados em grande escala

Elementos de uma solução de análise de dados em grande escala



- Extração, transformação e carregamento (ETL) ou orquestração (ELT) para mover dados
- Espelhamento de banco de dados para replicar dados operacionais para análise
- Processamento distribuído para limpar e reestruturar dados em grande escala
- Processamento de dados em tempo real e em lotes

- Armazenamento de arquivos flexível e escalonável em um data lake
- Tabelas relacionais em um data lakehouse ou data warehouse

- Modelos semânticos para entidades analíticas
- Geralmente, a forma de agregados cubes que resume valores numéricos em uma ou mais dimensões
- Gráficos
- Painéis

Processamento de dados em análises em grande escala



Banco de dados relacionais

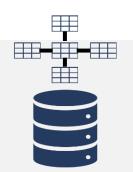
- Modelo bem estabelecido para processamento e armazenamento de dados relacionais
- Suporte abrangente à linguagem SQL para consulta e manipulação de dados



Apache Spark

- Plataforma Open Source para processamento de dados distribuídos e escaláveis
- Código de processamento de dados em várias linguagens (Python, Scala, Java, SQL,..)

Arquiteturas de armazenamento de dados analíticos



Data Warehouse

- Os dados são armazenados em um banco de dados relacional e consultados usando um mecanismo de consulta SQL
- As tabelas são desnormalizadas para otimização de consulta
 - Normalmente, como um esquema star ou snowflake de fatos numéricos que podem ser agregados por dimensões





Data Lakehouse

- Os arquivos de dados são armazenados em um sistema de arquivos distribuídos (um data lake) e normalmente processados usando o Apache Spark
- Metadados são usados para definir tabelas que fornecem uma interface SQL relacional para os dados do arquivo
 - Geralmente, um formato Delta Parquet é usado para fornecer funcionalidade de banco de dados transacional

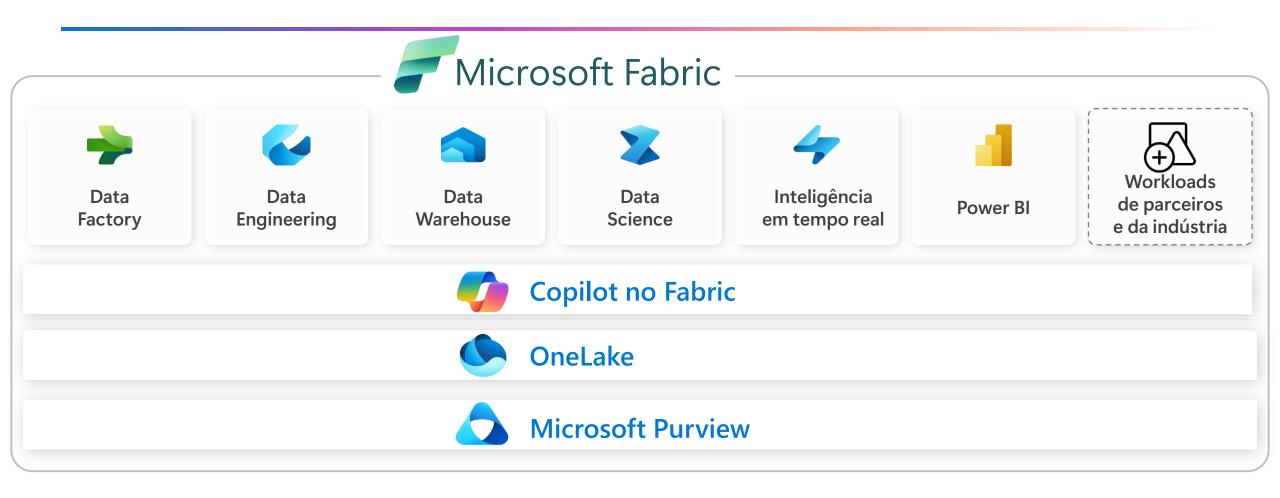
Análise de dados PaaS com o Azure Databricks

Azure Databricks

- Implementação baseada no Azure da solução analítica de nuvem do Databricks
- Consulta escalável do Spark e do SQL para data lake analytics
- Experiência interativa no espaço de trabalho do Azure Databricks
- Use o Azure Data Factory para implementar pipelines de processamento e ingestão de dados

Use para aproveitar as habilidades do Databricks e para portabilidade para a nuvem

Análise de dados de SaaS com o Microsoft Fabric



Demonstração

 Laboratório: explore análise de dados no Microsoft Fabric





Objetivos de aprendizagem

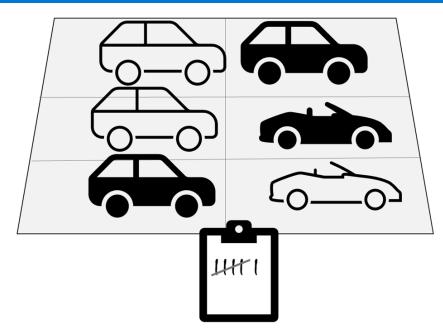
 Explore a análise em tempo real e de streaming



Objetivo de aprendizagem: explore a análise em tempo real e de streaming

Processamento em lote x streaming

Processamento em lote



Os dados são coletados e processados em intervalos regulares

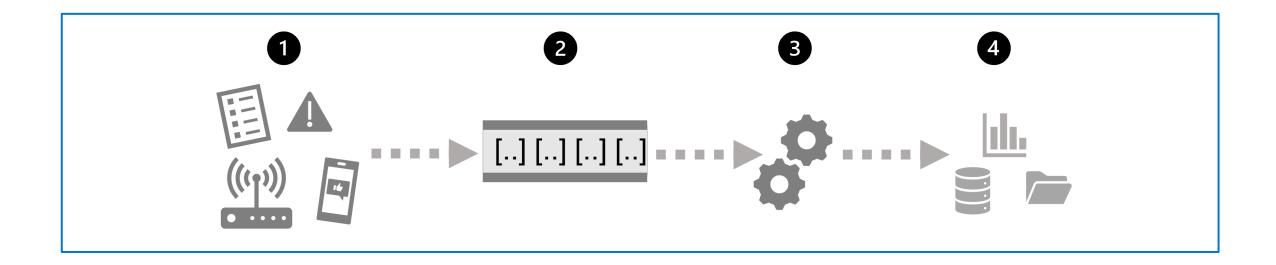
Processamento de streaming



Os dados são processados (quase) em tempo real à medida que chegam

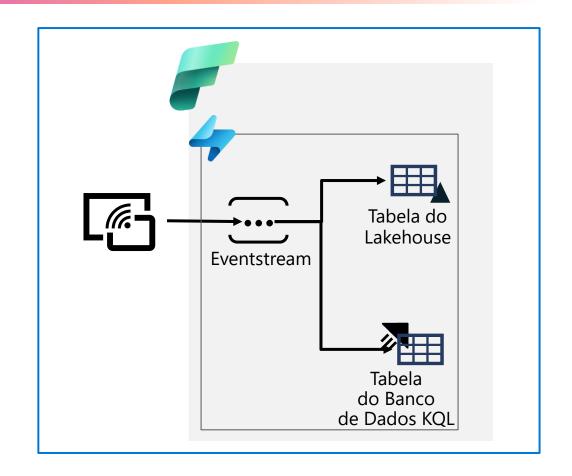
Elementos comuns do processamento de streaming

- 1. Um evento gera alguns dados.
- 2. Os dados gerados são capturados em uma fonte de fluxo para processamento.
- 3. Os dados do evento são processados.
- 4. Os resultados da operação de processamento de fluxo são gravados em uma saída (ou sink).



Análise em tempo real no Microsoft Fabric

- Suporte para ingestão contínua de dados de várias fontes
- Capture dados de streaming em um eventstream
- Grave dados em tempo real em uma tabela em um Lakehouse ou em um banco de dados KQL
- Consulte dados em tempo real usando SQL ou KQL
- Crie visualizações em tempo real



Análise de dados com o Apache Spark

O Apache Spark é uma estrutura de processamento distribuído para análise de dados em grande escala. Você pode usar o Spark no Microsoft Azure nos seguintes serviços:

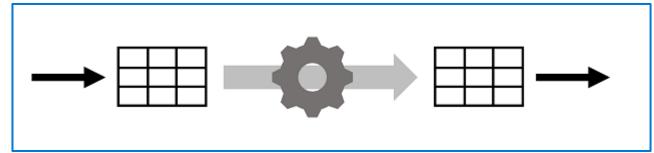
- Microsoft Fabric
- Azure Databricks

Spark Structured Streaming

A biblioteca Spark Structured Streaming, que fornece uma API (Application Programming Interface, Interface de programação de aplicações) para ingerir, processar e emitir resultados a partir de fluxos perpétuos de dados.

Delta Lake

O Delta Lake pode ser usado no Spark para definir tabelas relacionais para processamento em lote e de streaming.



Demonstração

 Laboratório: explorar análise em tempo real no Microsoft Fabric





Objetivos de aprendizagem

Explorar os fundamentos da visualização de dados



Objetivo de aprendizagem: explorar os fundamentos da visualização de dados

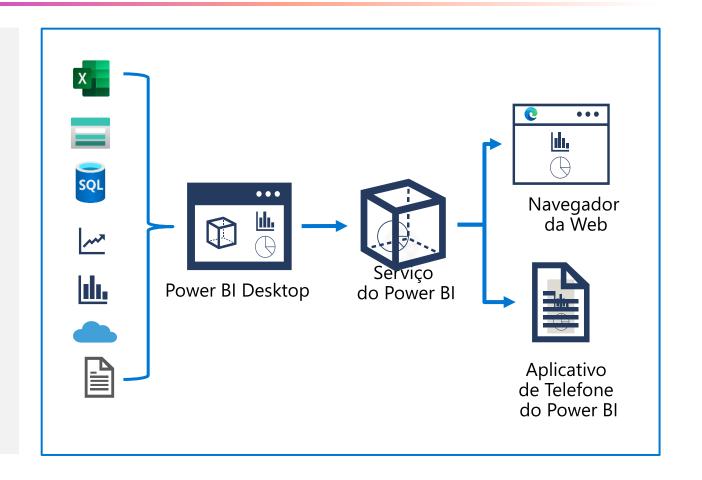
Introdução à visualização de dados com o Power BI

Comece com o Power BI Desktop

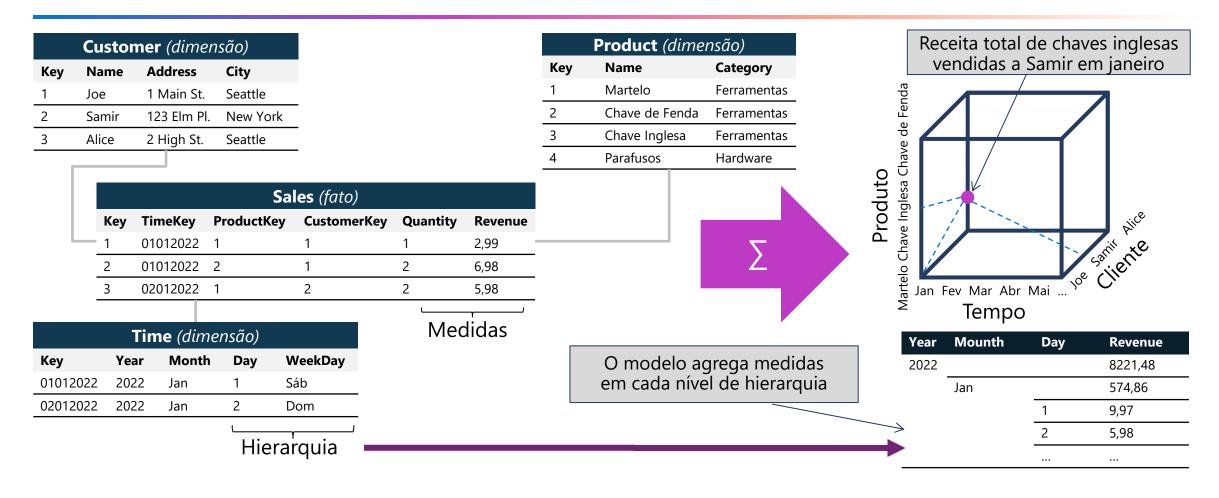
- o Importe dados de uma ou mais fontes
- Defina um modelo de dados
- Crie visualizações em um relatório

Publique no Serviço do Power BI

- Agende a atualização de dados
- Crie painéis e aplicativos
- Compartilhe com outros usuários
- Interaja com relatórios publicados
 - Navegador da Web
 - Aplicativo de Telefone do Power BI



Modelagem de dados analítica



Visualizações de dados comuns em relatórios

Tabelas e texto

Product Sales

Name	Quantity
Bolts	7
Hammer	4
Nails	
Screwdriver	2
Screws	â
Wrench	4
Total	1!

\$302.91

Revenue

Gráfico de barras ou colunas

Revenue by City and Category



Gráfico de linhas

Revenue by Month and Category

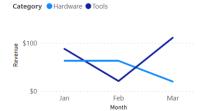


Gráfico de pizza

Quantity by Category

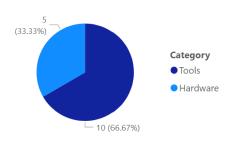
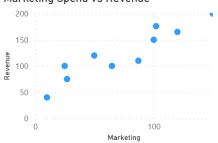


Gráfico de dispersão

Marketing Spend vs Revenue



Mapear

Revenue by City



Demonstração

• Laboratório: visualize dados com o Power BI