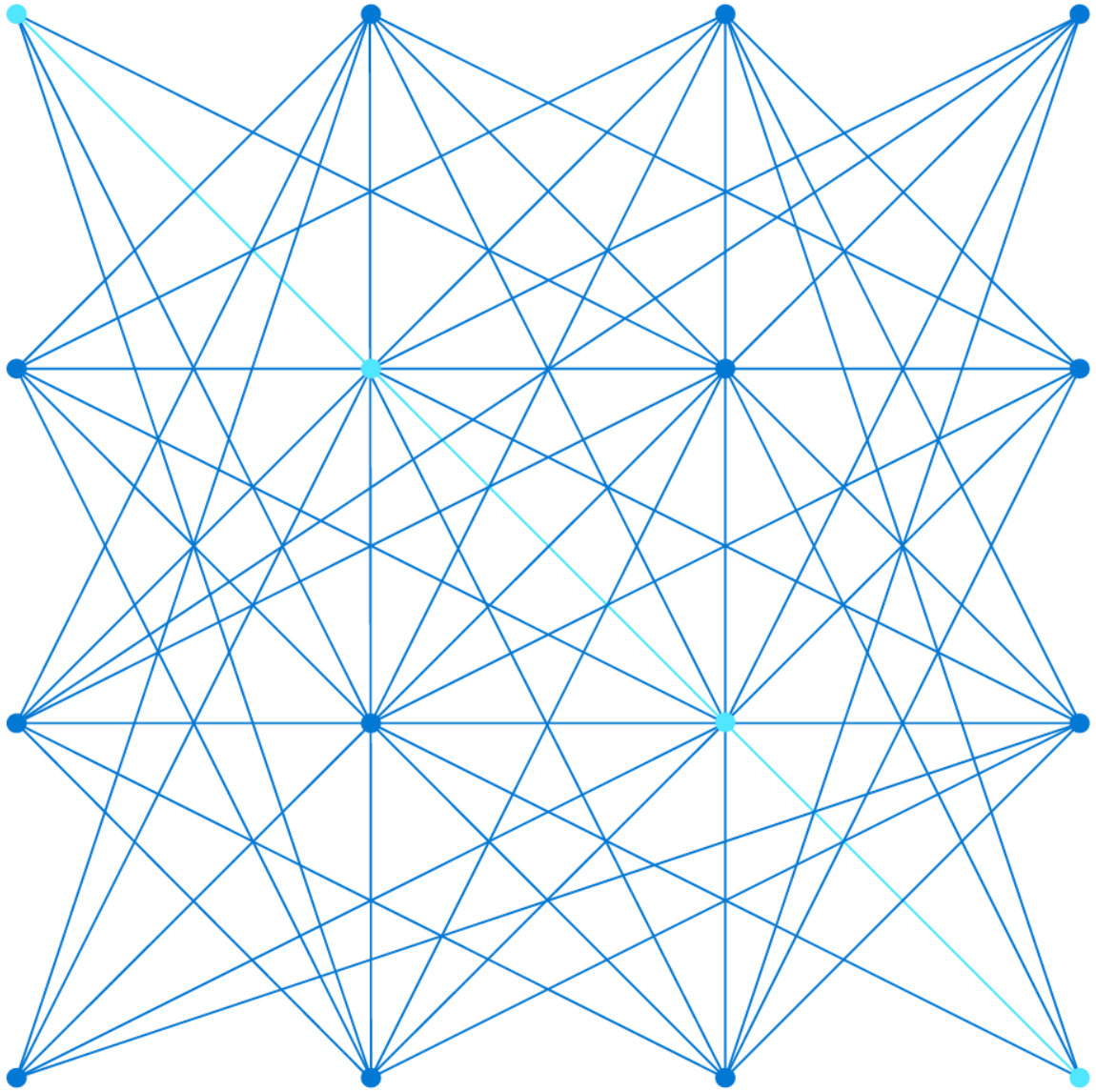
Microsoft Azure

Módulo 4: Explorar conceitos básicos da análise de dados



Agenda

Data warehousing em grande escala

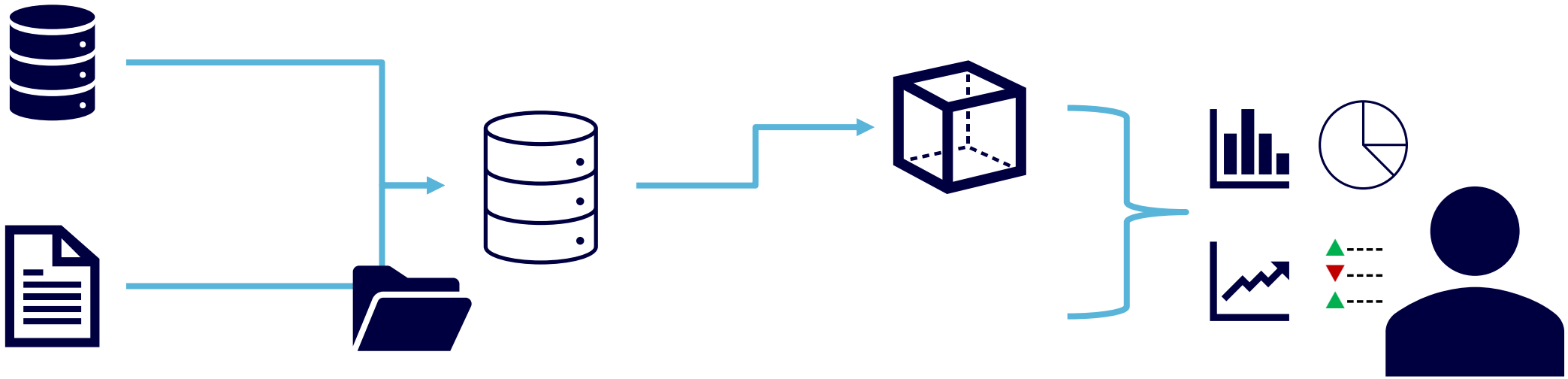
Streaming e análise em tempo real

Visualização de dados

Lição 1: Data warehousing em grande escala

O que é data warehousing em grande escala?

Processamento e ingestão de dados	Armazenamento de dados analíticos	Modelo de dados analíticos	Visualização de dados



- Orquestração de *ETL* (extração, transformação e carregamento) e *ELT* (extração, carregamento e transformação)
- Processamento distribuído para limpar e reestruturar dados em escala
- Processamento de dados em lote e em tempo real

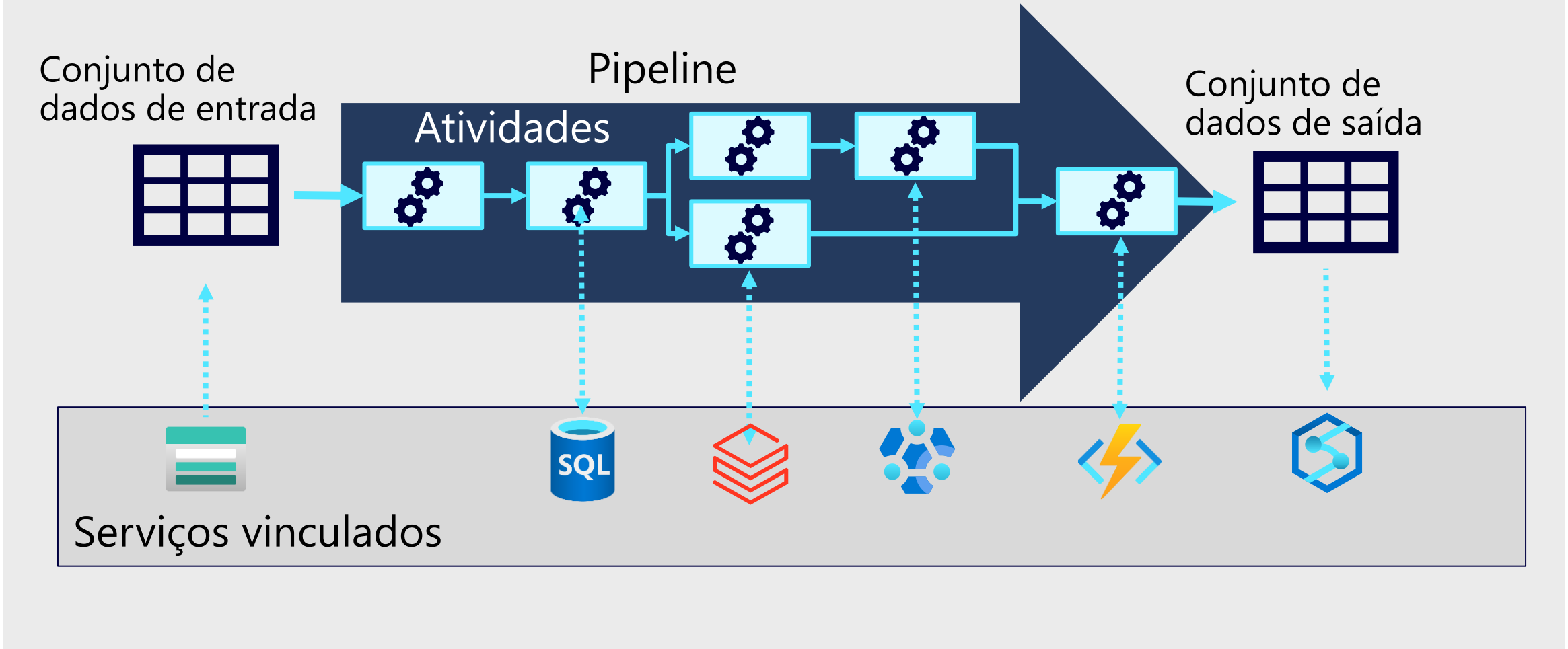
- Armazenamento de dados relacionais desnormalizado em um *data warehouse*
- Armazenamento de arquivos semiestruturados em um *data lake*

- Modelos semânticos para entidades analíticas
- Geralmente na forma de *cubos* agregados que resumem valores numéricos em uma ou mais *dimensões*

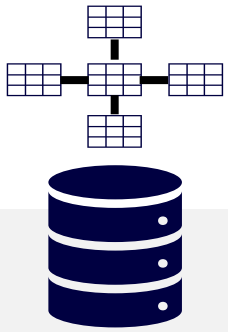
- Relatórios
- Gráficos
- Painéis

Pipelines de processamento e ingestão de dados

Crie pipelines no **Azure Data Factory** ou no **Azure Synapse Analytics**

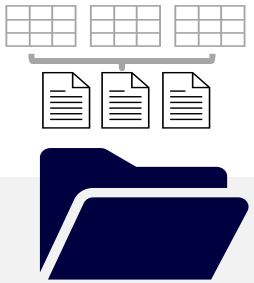


Armazenamentos de dados analíticos



data warehouse

- Armazenamento de banco de dados relacional e mecanismo de consulta em grande escala
- Os dados são *desnormalizados* para otimização de consulta
  - Normalmente, em um esquema *floco de neve* ou *estrela* de *fatos* numéricos que podem ser agregados por *dimensões*



Data Lake

- Os arquivos de dados são armazenados em um sistema de arquivos distribuído
- As camadas de armazenamento tabular podem ser usadas para abstrair arquivos e fornecer uma interface relacional.
  - Use tabelas externas do *PolyBase* ou crie um *banco de dados lake* no Azure Synapse Analytics
  - Use tabelas de banco de dados e pontos de extremidade SQL no Azure Databricks
  - Use o Spark *Delta Lake* para adicionar semântica de armazenamento relacional e criar um *data lakehouse* no Azure Synapse Analytics, no Azure Databricks e no Azure HDInsight

Escolha um serviço de armazenamento de dados analítico

Azure Synapse Analytics

- Solução unificada para data warehouse relacional e análise de data lake
- Processamento e consulta escalonável por meio de vários runtimes de análise
  - SQL do Synapse
  - Apache Spark
  - Synapse Data Explorer
- Experiência interativa no Azure Synapse Studio
- Integração de pipeline interna para ingestão e processamento de dados

Use para uma solução analítica unificada de grande escala no Azure

Azure Databricks

- Implementação baseada no Azure da plataforma de análise de nuvem Databricks
- Consulta escalonável de Spark e SQL para análise de data lake
- Experiência interativa no workspace do Azure Databricks
- Use o Azure Data Factory para implementar pipelines de processamento e ingestão de dados

Use para aproveitar habilidades do Databricks e para portabilidade na nuvem

Azure HDInsight

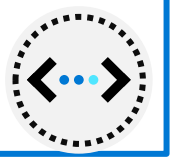
- Implementação baseada no Azure de estruturas comuns de "Big Data" do Apache criadas em um data lake
  - Hadoop – Consultar arquivos de data lake usando tabelas do Hive
  - Spark – Usar APIs do Spark para consultar dados e abstrair o armazenamento de arquivos subjacente como tabelas
  - Kafka – Processamento de eventos em tempo real
  - Storm – Processamento de fluxo
  - HBase – Armazenamento de dados NoSQL

Use quando precisar dar suporte a várias plataformas de código aberto

Laboratório: Explorar o Azure Synapse Analytics

Neste exercício, você provisionará um workspace do Azure Synapse Analytics e o usará para ingerir e processar dados

1. Inicie a máquina virtual para este laboratório
   ou vá para a página do exercício em https://aka.ms/dp900-synapse-lab
2. Siga as instruções para concluir o exercício no Microsoft Learn
   Use a assinatura do Azure fornecida para este laboratório



Lição 1: Verificação de conhecimentos

Quais serviços do Azure você pode usar para criar um pipeline para ingestão e processamento de dados?

- ❑ Banco de Dados SQL do Azure e Azure Cosmos DB
- ☑ Azure Synapse Analytics e Azure Data Factory
- ❑ Azure HDInsight e Azure Databricks

O que você precisa definir para implementar um pipeline que lê dados do Armazenamento de Blobs do Azure?

- ☑ Um serviço vinculado para sua conta de Armazenamento de Blobs do Azure
- ❑ Um pool de SQL dedicado em seu workspace do Azure Synapse Analytics
- ❑ Um cluster do Azure HDInsight em sua assinatura

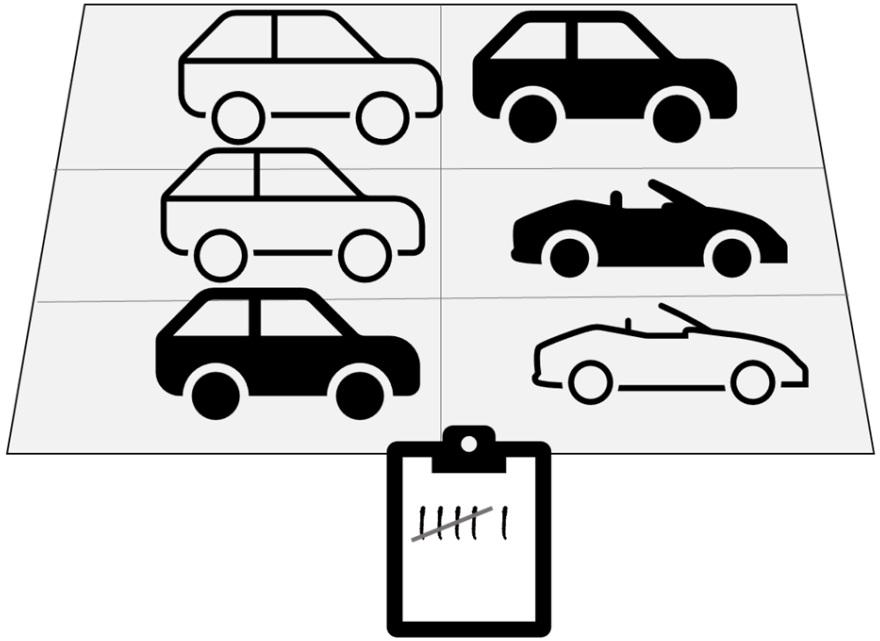
Qual mecanismo de processamento distribuído de código aberto é incluído no Azure Synapse Analytics?

- ❑ Apache Hadoop
- ☑ Apache Spark
- ❑ Apache Storm

Lição 2: Streaming e análise em tempo real

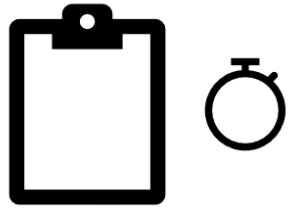
Processamento em lotes e de fluxo

Processamento em lotes



Os dados são coletados e processados em intervalos regulares

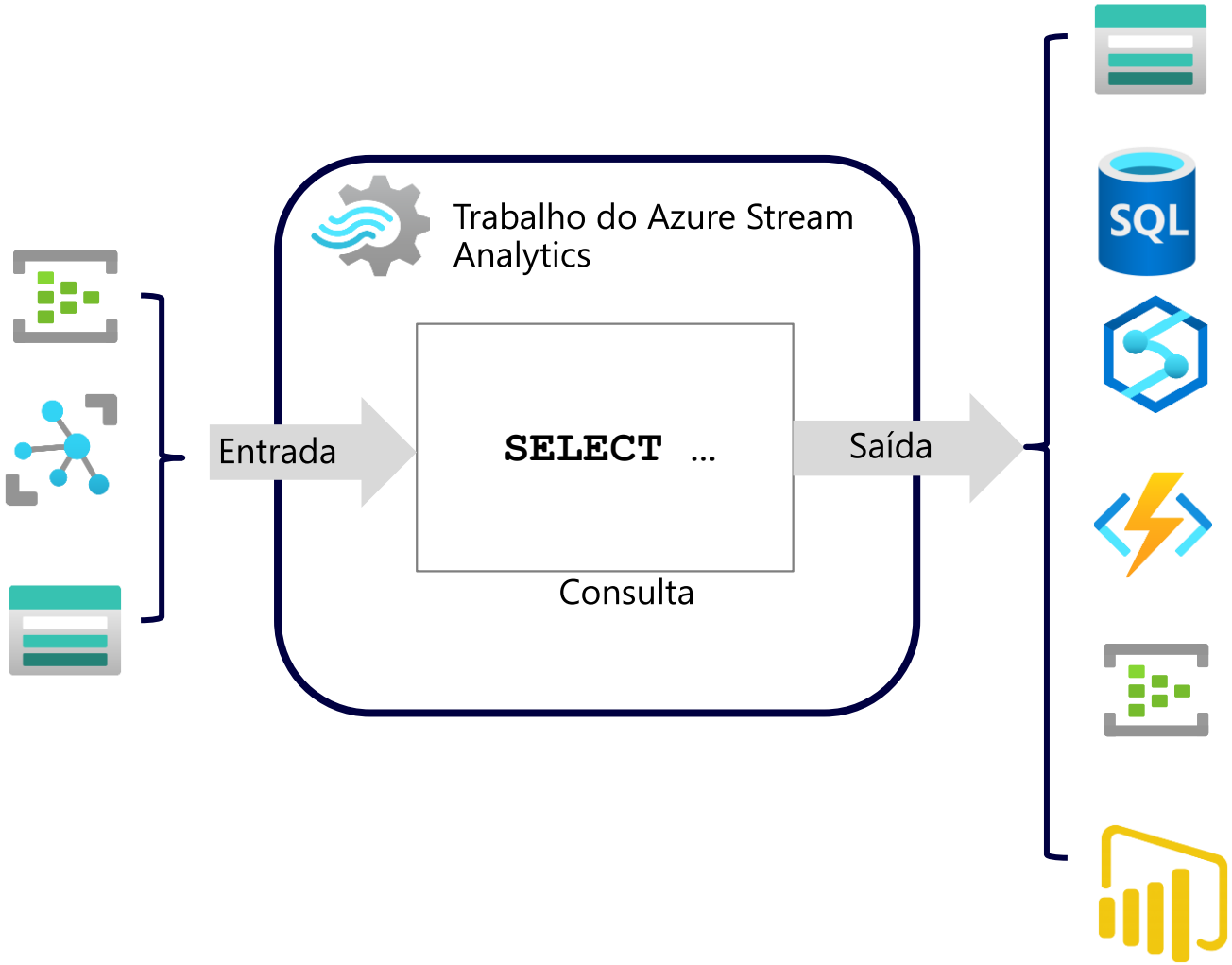
Processamento de fluxo



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

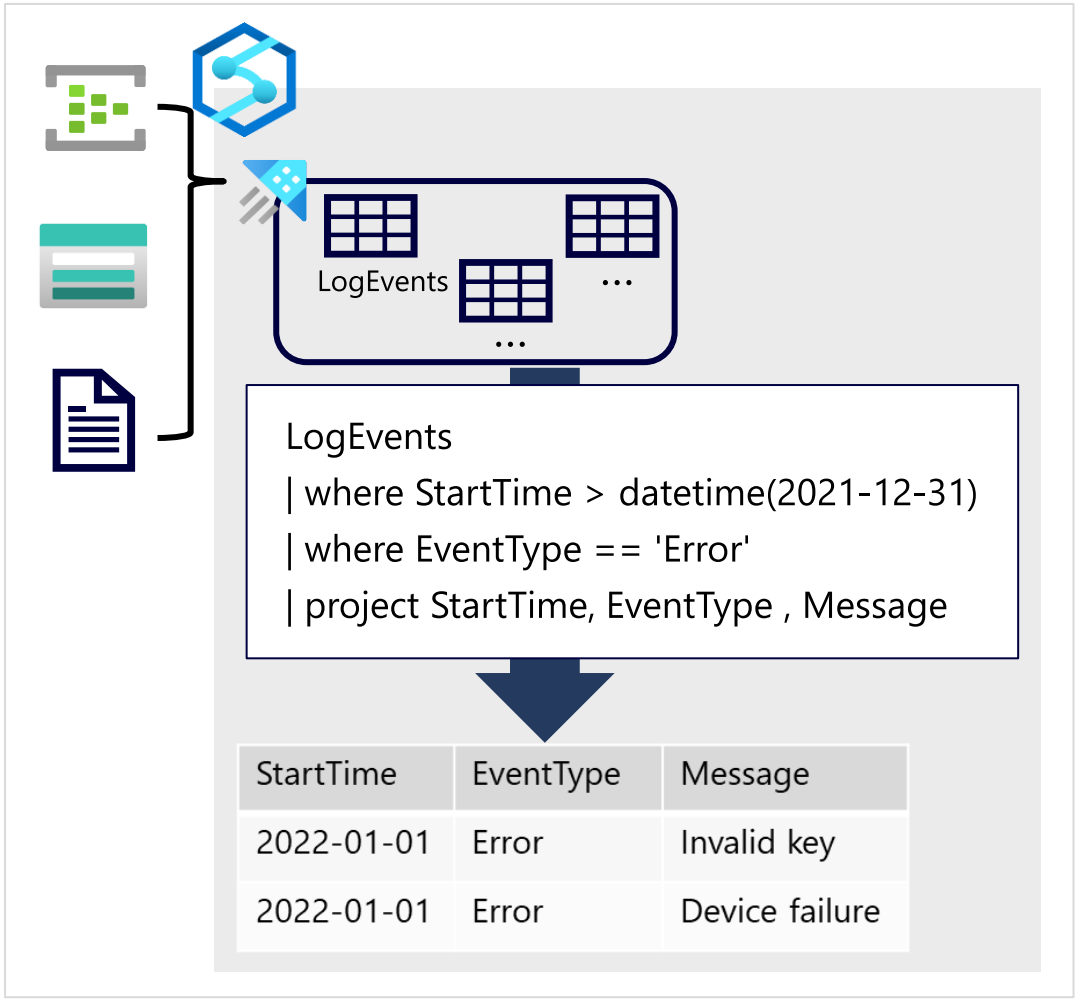
Processamento de dados em tempo real com o Azure Stream Analytics

- Crie um *trabalho* individual ou um *cluster* do Azure Stream Analytics
- Faça a ingestão de dados de uma *entrada*, como:
  - Hubs de eventos do Azure
  - Hub IoT do Azure
  - Armazenamento de Blobs do Azure
  - ...
- Processar dados com uma *consulta* perpétua
- Enviar resultados para uma *saída*, como:
  - Armazenamento do Blobs do Azure
  - Banco de Dados SQL do Azure
  - Azure Synapse Analytics
  - Azure Function
  - Hubs de eventos do Azure
  - Power BI
  - ...



Análise de log e telemetria em tempo real com o Azure Data Explorer

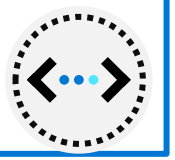
- Alta taxa de transferência, serviço escalonável para dados em lotes e de streaming
  - **Serviço** dedicado do Azure Data Explorer
  - Runtime do **Data Explorer do Azure Synapse** no Azure Synapse Analytics
- Os dados são ingeridos de fontes de streaming e em lotes em tabelas em um banco de dados
- As tabelas podem ser consultadas usando *KQL* (Linguagem de Consulta Kusto):
  - Sintaxe intuitiva para consultas somente leitura
  - Otimizado para dados brutos de telemetria e série temporal



Laboratório: Analisar dados de streaming

Neste laboratório, você usará o Azure Stream Analytics para processar um fluxo de dados em tempo real

1. Inicie a máquina virtual para este laboratório
   ou vá para a página do exercício em https://aka.ms/dp900-stream-lab
2. Siga as instruções para concluir o exercício no Microsoft Learn
   Use a assinatura do Azure fornecida para este laboratório e um cloud shell no portal do Azure



Lição 2: Verificação de conhecimentos

Qual a definição de *processamento de fluxo* está correta?

- [x] Os dados são processados continuamente à medida que novos registros de dados chegam
- [ ] Os dados são coletados em um armazenamento temporário e todos os registros são processados em conjunto como um lote
- [ ] Os dados estão incompletos e não podem ser analisados

Qual serviço você usaria para capturar continuamente dados de um Hub IoT, agregá-los em períodos temporais e armazenar os resultados no Banco de Dados SQL do Azure?

- [ ] Azure Cosmos DB
- [x] Stream Analytics do Azure
- [ ] Armazenamento do Azure

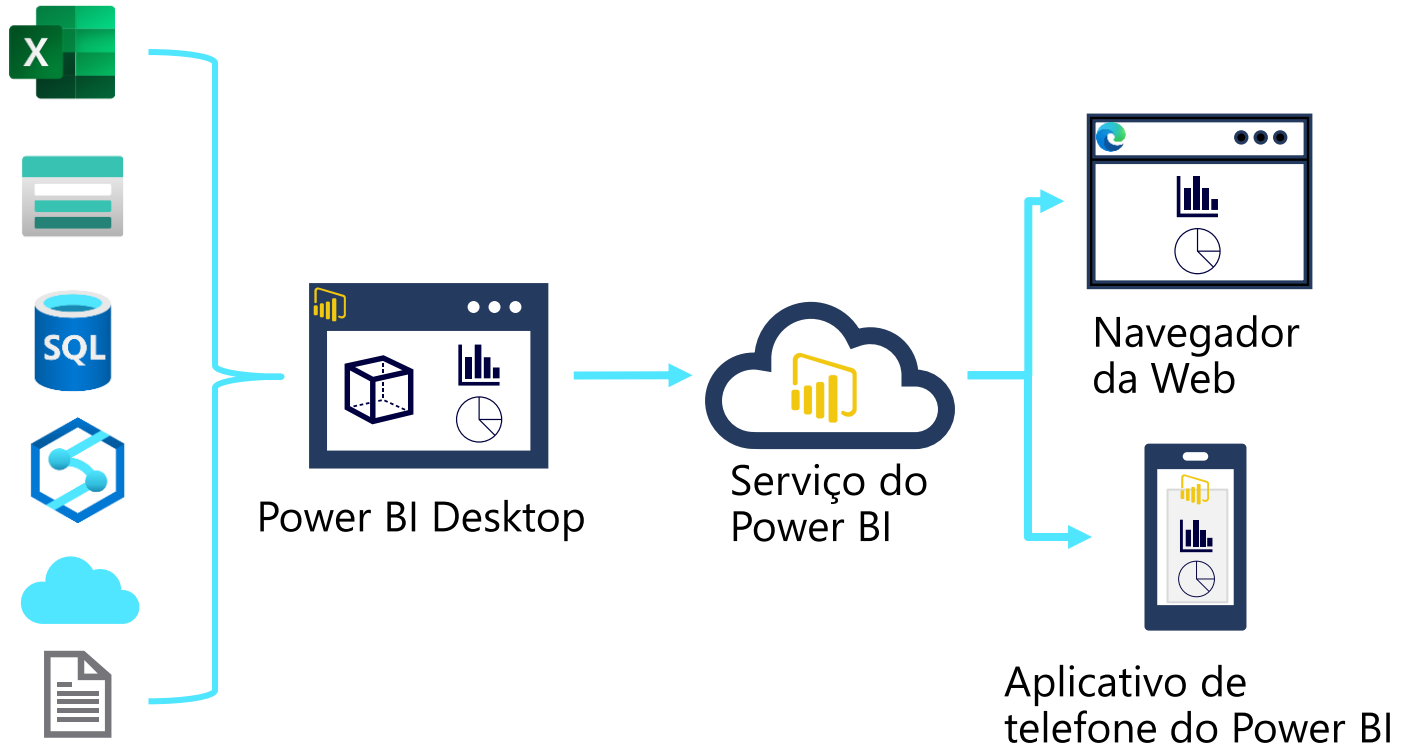
Qual linguagem você usaria para consultar dados de log em tempo real no Azure Synapse Data Explorer?

- [ ] SQL
- [ ] Python
- [x] KQL

Lição 3: Visualização de dados

Introdução à visualização de dados com o Power BI

- Começar com o Power BI Desktop
  - Importar dados de uma ou mais fontes
  - Definir um modelo de dados
  - Criar visualizações em um relatório
- Publicar no serviço do Power BI
  - Agendar atualização de dados
  - Criar dashboards e aplicativos
  - Compartilhar com outros usuários
- Interagir com relatórios publicados
  - Navegador da Web
  - Aplicativo de telefone do Power BI



Modelagem de dados analíticos

Cliente *(dimensão)*

Chave	Nome	Endereço	City
1	Joe	1 Main St.	Seattle
2	Samir	123 Elm Pl.	Nova Iorque
3	Alice	2 High St.	Seattle

Produto *(dimensão)*

Chave	Nome	Categoria
1	Martelo	Ferramentas
2	Screwdriver	Ferramentas
3	Chave inglesa	Ferramentas
4	Bolts	Hardware

Vendas *(fato)*

Chave	TimeKey	ProductKey	CustomerKey	Quantidade	Receita
1	01012022	1	1	1	2,99
2	01012022	2	1	2	6.98
3	02012022	1	2	2	5,98

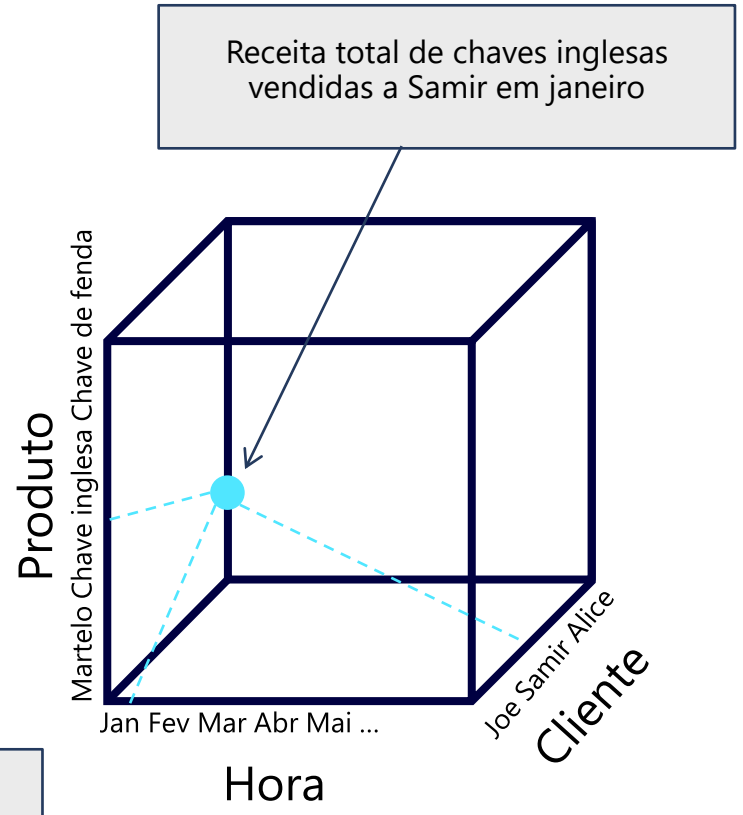
Medidas

Tempo *(dimensão)*

Chave	Ano	Month	Dia	WeekDay
01012022	2022	Jan	1	Sat
02012022	2022	Jan	2	Sun

Hierarquia

∑



O modelo agrega medidas em cada nível de hierarquia

Ano	Month	Dia	Receita
2022			8221.48
	Jan		574.86
		1	9.97
		2	5,98
		...	...

Visualizações de dados comuns em relatórios

Tabelas e texto

Product Sales

Name	Quantity
Bolts	2
Hammer	4
Nails	1
Screwdriver	2
Screws	2
Wrench	4
Total	**15**

$302.91

Revenue

Gráfico de barras ou de colunas

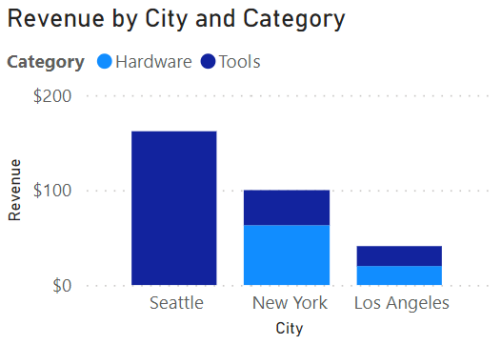


Gráfico de Linhas

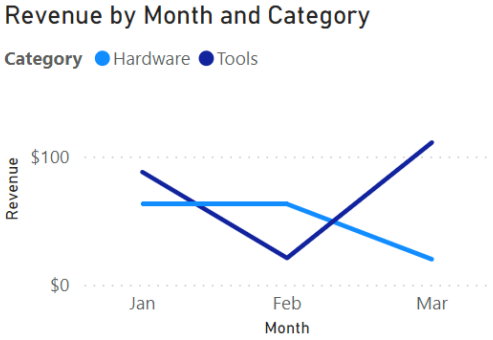


Gráfico de pizza

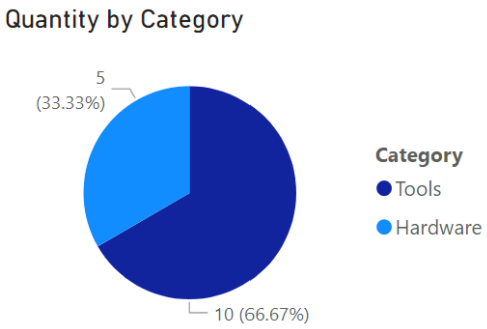
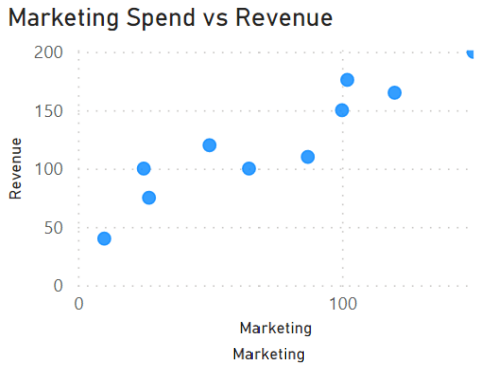
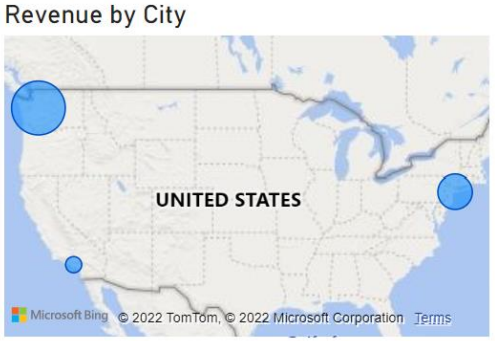


Gráfico de dispersão



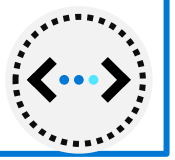
Mapeamento



Laboratório: Visualizar os dados com o Power BI

Neste laboratório, você usará o Power BI Desktop para criar um modelo de dados e um relatório

1. Inicie a máquina virtual para este laboratório
   ou vá para a página do exercício em https://aka.ms/dp900-pbi-lab
2. Siga as instruções para concluir o exercício no Microsoft Learn
   Use a assinatura do Azure fornecida para este laboratório



Lição 3: Verificação de conhecimentos

Qual ferramenta você deve usar para importar dados de várias fontes de dados e criar um relatório?

- [x] Power BI Desktop
- [ ] Aplicativo de telefone do Power BI
- [ ] Fábrica de dados do Azure

O que você deve definir em seu modelo de dados para permitir a análise de drill up/down?

- [ ] Uma medida
- [x] Uma hierarquia
- [ ] Uma relação

Qual tipo de visualização você deve usar para analisar as tarifas de aprovação para vários exames ao longo do tempo?

- [ ] Um gráfico de pizza
- [ ] Um gráfico de dispersão
- [x] Um gráfico de linhas

Mais aprendizado

Para revisar o que você aprendeu e fazer laboratórios adicionais, examine os módulos do Microsoft Learn para este curso:

- Explorar os principais conceitos de dados https://aka.ms/ExploreDataConcepts-ptb
- Explorar dados relacionais no Azure https://aka.ms/ExploreRelationalData-ptb
- Explorar dados não relacionais no Azure https://aka.ms/ExploreNonRelationalData-ptb
- Explorar a análise de dados no Azure https://aka.ms/ExploreNonRelationalData-ptb

Microsoft Azure