Olá, aluno(a)!
Seja bem-vindo(a) à aula interativa!

Você entrará na reunião com a câmera e o microfone desligados.

Sua presença será computada através da enquete.
Fique atento(a) e não deixe de respondê-la!

Tecnologias de Big Data - Processamento de dados massivos

Segunda Aula Interativa

Prof. Dr. Pedro Calais

O que o engenheiro de dados faz?

- Coleta dados
- Gerencia dados
- Disponibiliza dados em uma formato usável para analistas e cientistas de dados
- Prepara dados para analistas e cientistas de dados

Data engineering is all about the movement, manipulation, and management of data.

—Lewis Gavin5

Spark: o melhor amigo do engenheiro de dados!

SPONSORED

Spark: A Data Engineer's Best Friend

BrandPost Sponsored by HPE | Learn More

By Don Wake | OCT 13, 2021 1:13 PM PDT

Data engineering, as a separate category of expertise in the world of data science, did not occur in a vacuum. The role of the data engineer originated and evolved as the number of data sources and data products ballooned over the years. Therefore, the

https://www.cio.com/article/189412/spark-a-data-engineer-s-best-friend.html

Spark: o melhor amigo do engenheiro de dados!

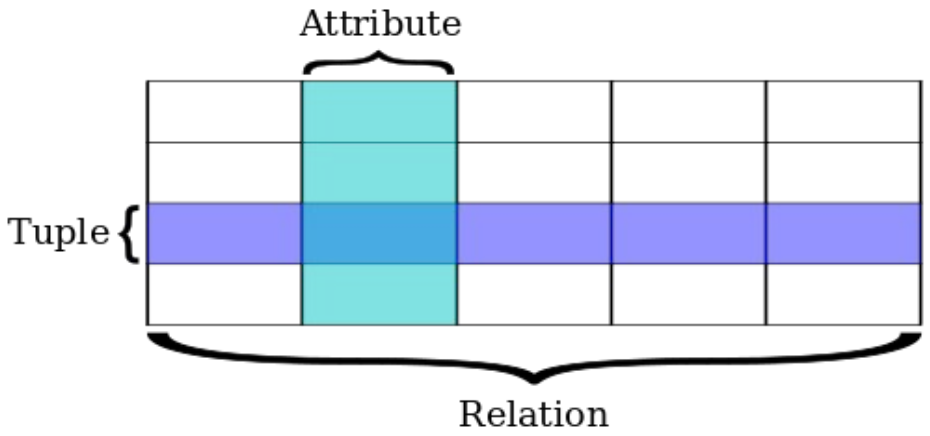
SPONSORED

Spark: A Data Engineer's Best Friend

- Write programs that access, transform, and store the data. Many common programming languages have APIs to integrate Spark code directly, and Spark offers many powerful functions for performing complex ETL-style data cleaning and transformation functions. Spark also includes a high-level API that allows users to seamlessly write queries in SQL.

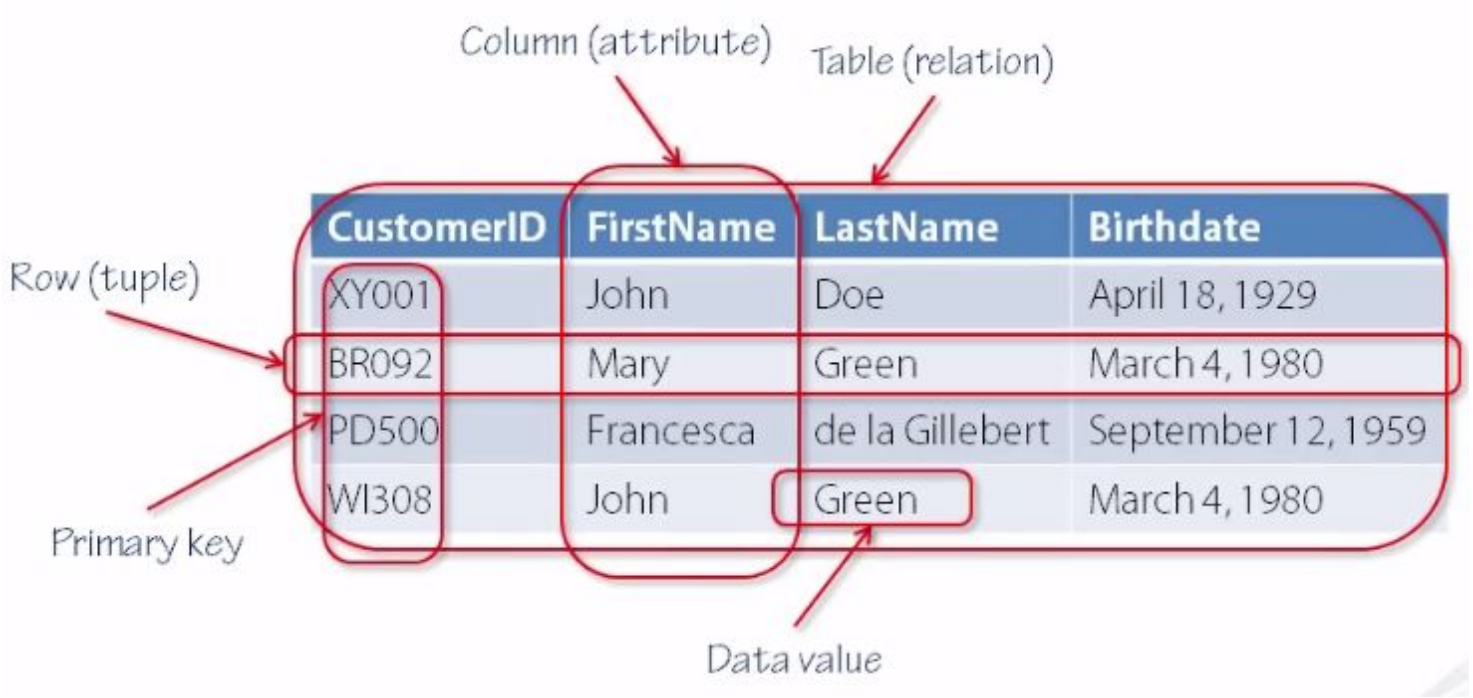
O que é SQL?

- *Standard Query Language*
- Criada na década de 70
- Criada para gerir dados armazenados em um sistema de banco de dados relacional (SGBD)
- relações: **conjuntos de tabelas** contendo linhas e colunas



O que é SQL?

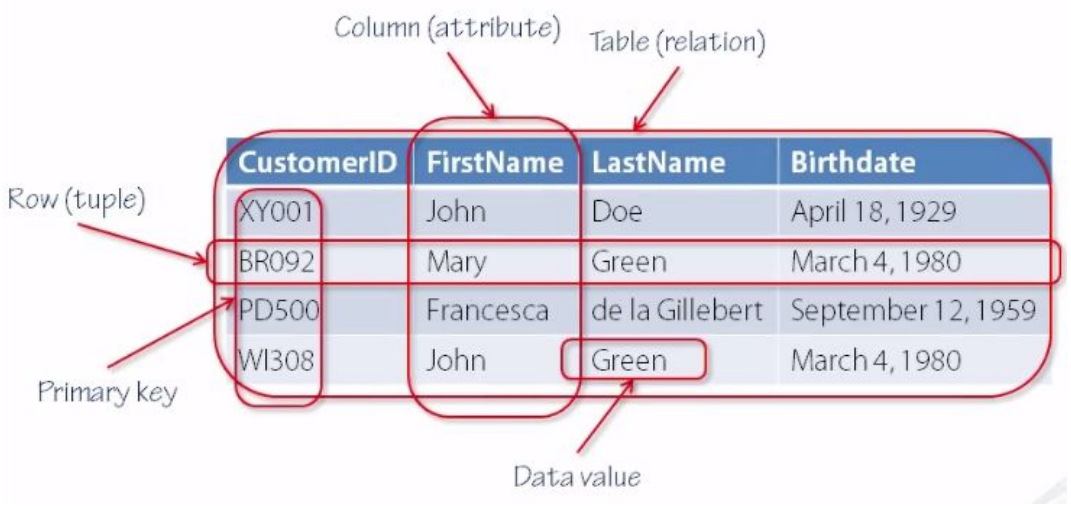
- Relação “Customer”:



CustomerID	FirstName	LastName	Birthdate
XY001	John	Doe	April 18, 1929
BR092	Mary	Green	March 4, 1980
PD500	Francesca	de la Gillebert	September 12, 1959
WI308	John	Green	March 4, 1980

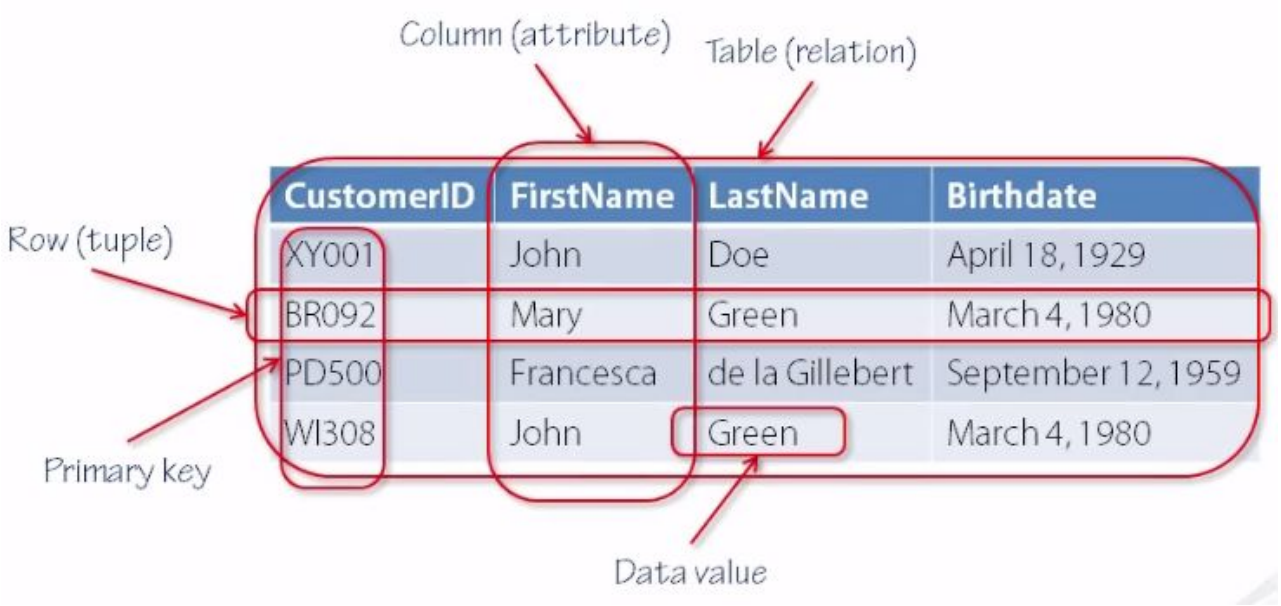
O que é SQL?

- Permite ao programador manipular os dados em um banco de dados relacional:
  - **adicionar** registros
  - **remover** registros
  - **atualizar** registros
  - **ler** registros [Spark SQL entra aqui!]



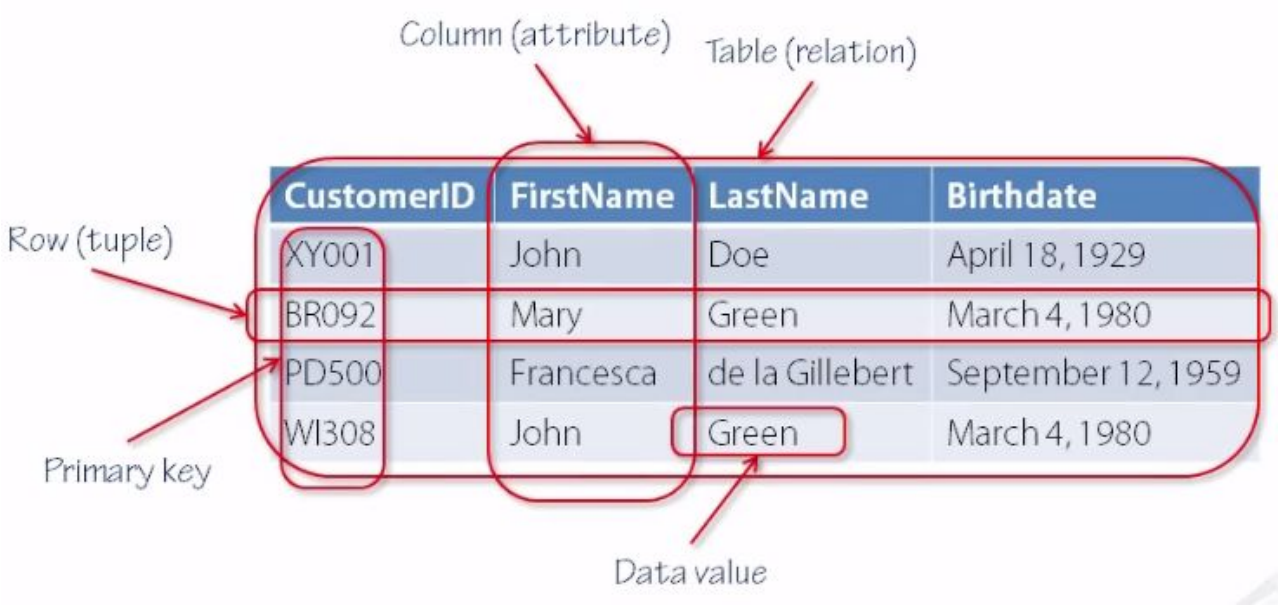
O que é SQL?

- SELECT * FROM Customers;



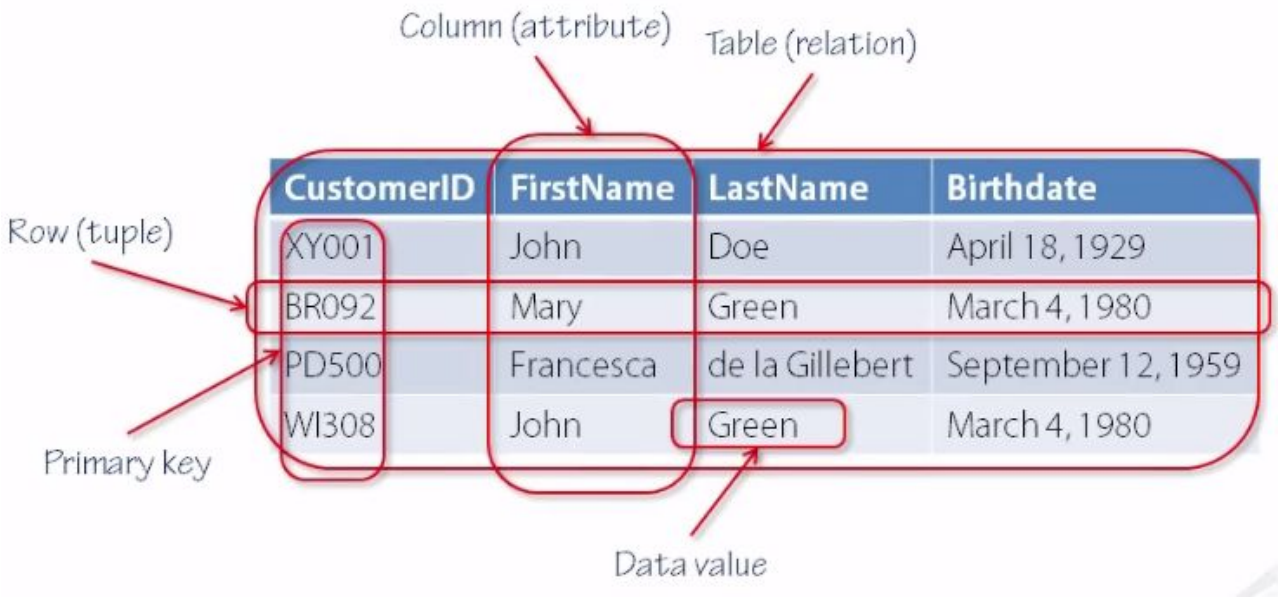
O que é SQL?

- SELECT FirstName FROM Customers;



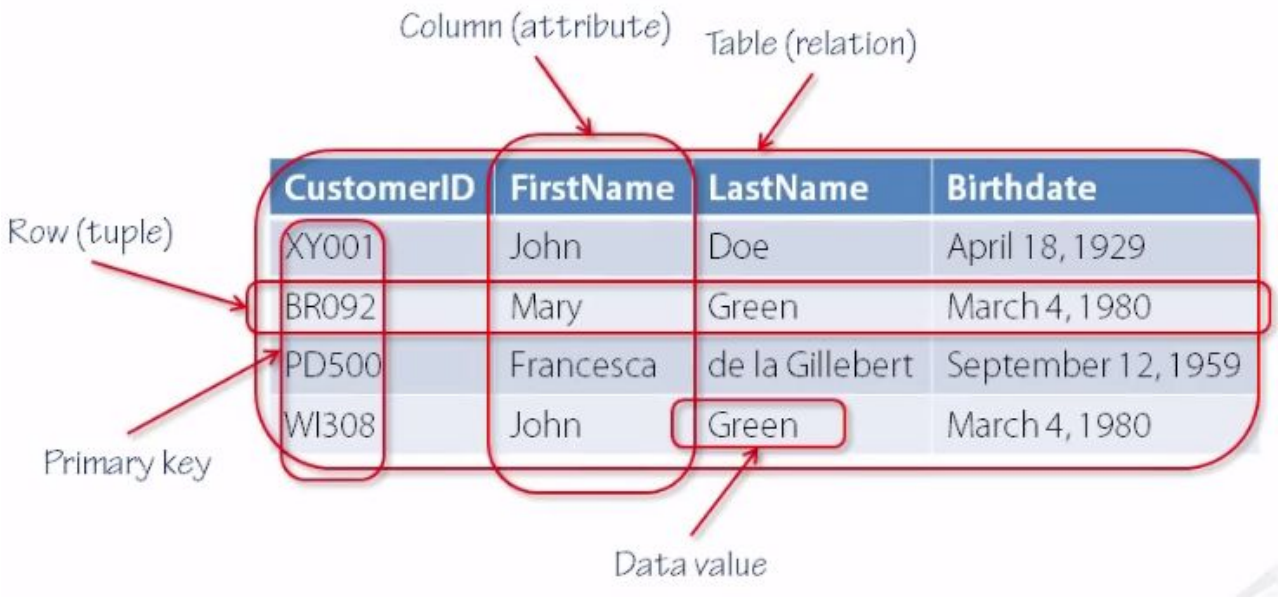
O que é SQL?

```
SELECT FirstName FROM Customers WHERE LastName = 'Green';
```



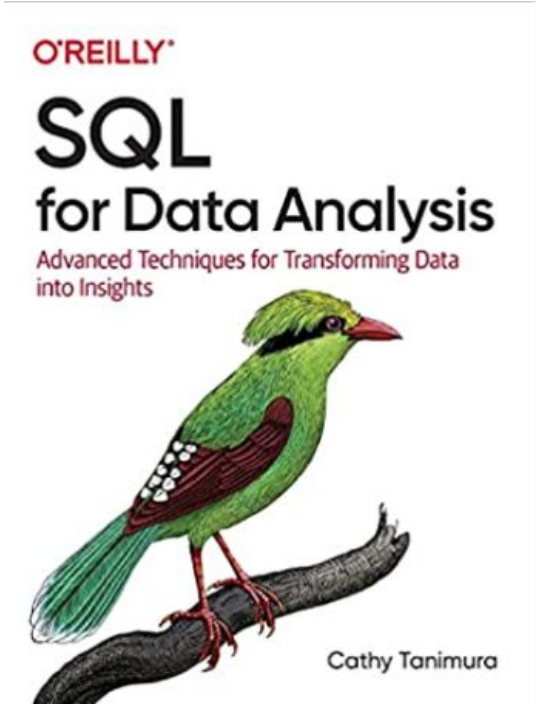
O que é SQL?

```
SELECT FirstName FROM Customers WHERE Age > 50;
```



O que é SQL?

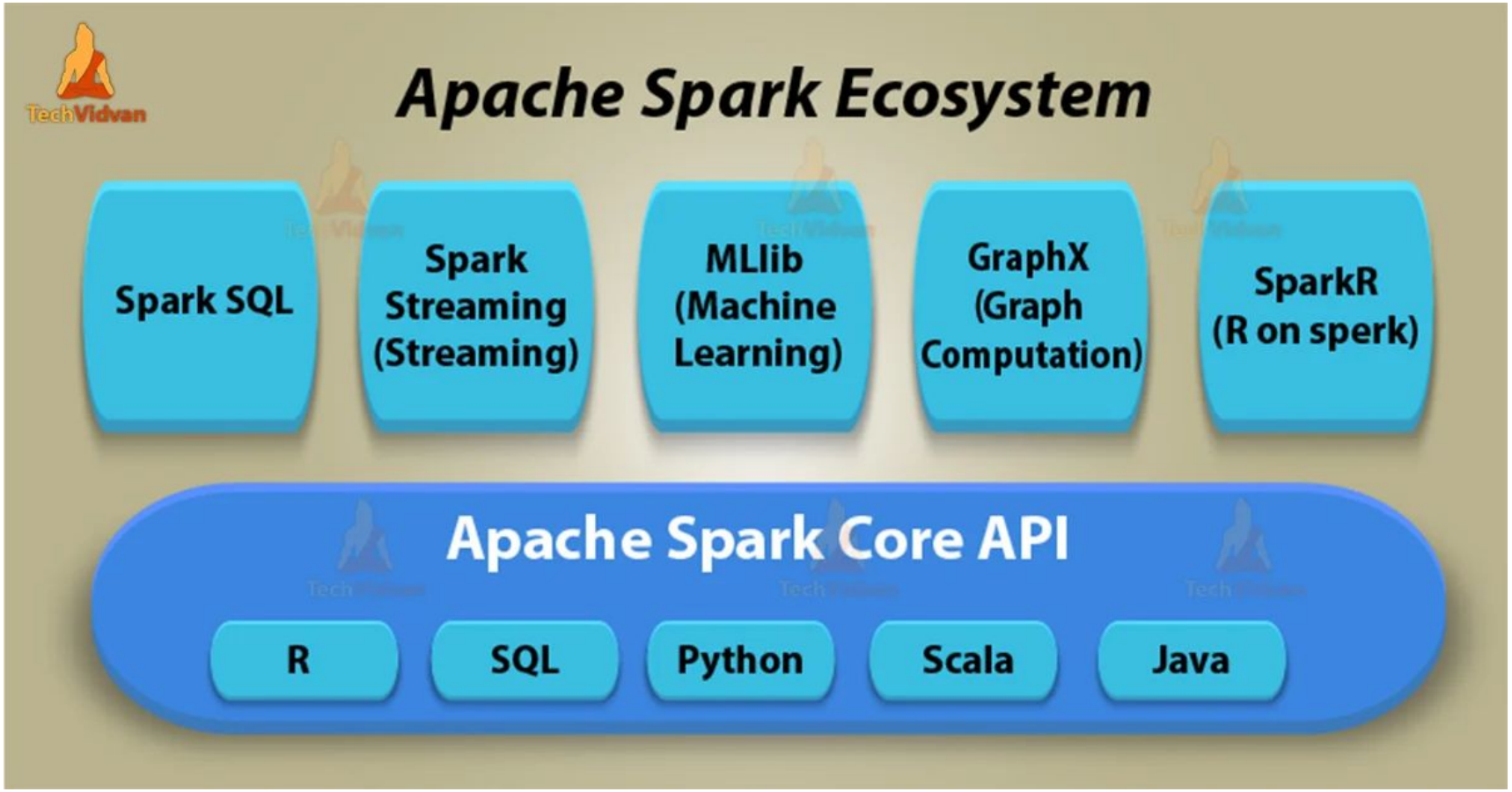
- Muitos dados são armazenados em bancos de dados relacionais!
- Ferramenta padrão para cientistas e engenheiros de dados



Tutorial básico de SQL

https://www.w3schools.com/sql/

ecossistema do Spark



Você já conhece: a API de Dataframes

```
employees
  .join(dept, employees("deptId") === dept("id"))
  .where(employees("gender") === "female")
  .groupBy(dept("id"), dept("name"))
  .agg(count("name"))
```

Suporte à linguagem SQL

- Em algumas situações, o SQL “puro” é mais conveniente.

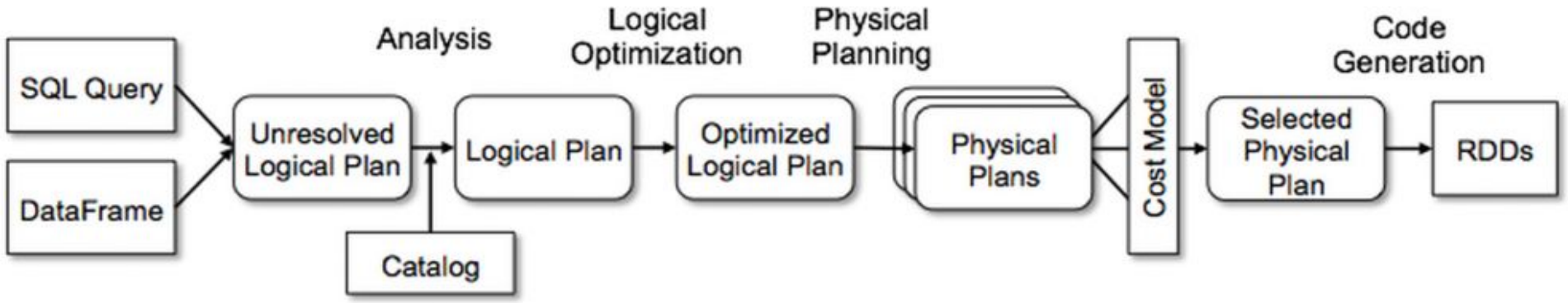
```
users.where(users("age") < 21)
     .registerTempTable("young")
ctx.sql("SELECT count(*), avg(age) FROM young")
```

Benefícios do Spark SQL

- Integração:
  - Spark SQL “mistura” consultas SQL em programas Spark.
  - Integrar consultas SQL com *analytics* complexos.

Benefícios do Spark SQL

- Otimização de consultas:
  - Plano lógico e físico para cada consulta.
  - *Catalyst Optimizer:*



Artigo “oficial”

Spark SQL: Relational Data Processing in Spark

Michael Armbrust[†], Reynold S. Xin[†], Cheng Lian[†], Yin Huai[†], Davies Liu[†], Joseph K. Bradley[†], Xiangrui Meng[†], Tomer Kaftan[‡], Michael J. Franklin[†‡], Ali Ghodsi[†], Matei Zaharia[†*]

[†]Databricks Inc. [*]MIT CSAIL [‡]AMPLab, UC Berkeley

Artigo “oficial”

ABSTRACT

Spark SQL is a new module in Apache Spark that integrates relational processing with Spark's functional programming API. Built on our experience with Shark, Spark SQL lets Spark programmers leverage the benefits of relational processing (*e.g.*, declarative queries and optimized storage), and lets SQL users call complex analytics libraries in Spark (*e.g.*, machine learning). Compared to previous systems, Spark SQL makes two main additions. First, it offers much tighter integration between relational and procedural processing, through a declarative DataFrame API that integrates with procedural Spark code. Second, it includes a highly extensible optimizer, Catalyst, built using features of the Scala programming language, that makes it easy to add composable rules, control code generation, and define extension points. Using Catalyst, we have built a variety of features (*e.g.*, schema inference for JSON, machine learning types, and query federation to external databases) tailored for the complex needs of modern data analysis. We see Spark SQL as an evolution of both SQL-on-Spark and of Spark itself, offering richer APIs and optimizations while keeping the benefits of the Spark programming model.

Vamos escrever um pouco de código?

Dados em fluxo contínuo

fluxos contínuos de dados

With Spark, data engineers can:

- Connect to different data sources in different locations, including cloud sources such as Amazon S3, databases, Hadoop file systems, data streams, web services, and flat files.

Dados em batch vs Dados em stream

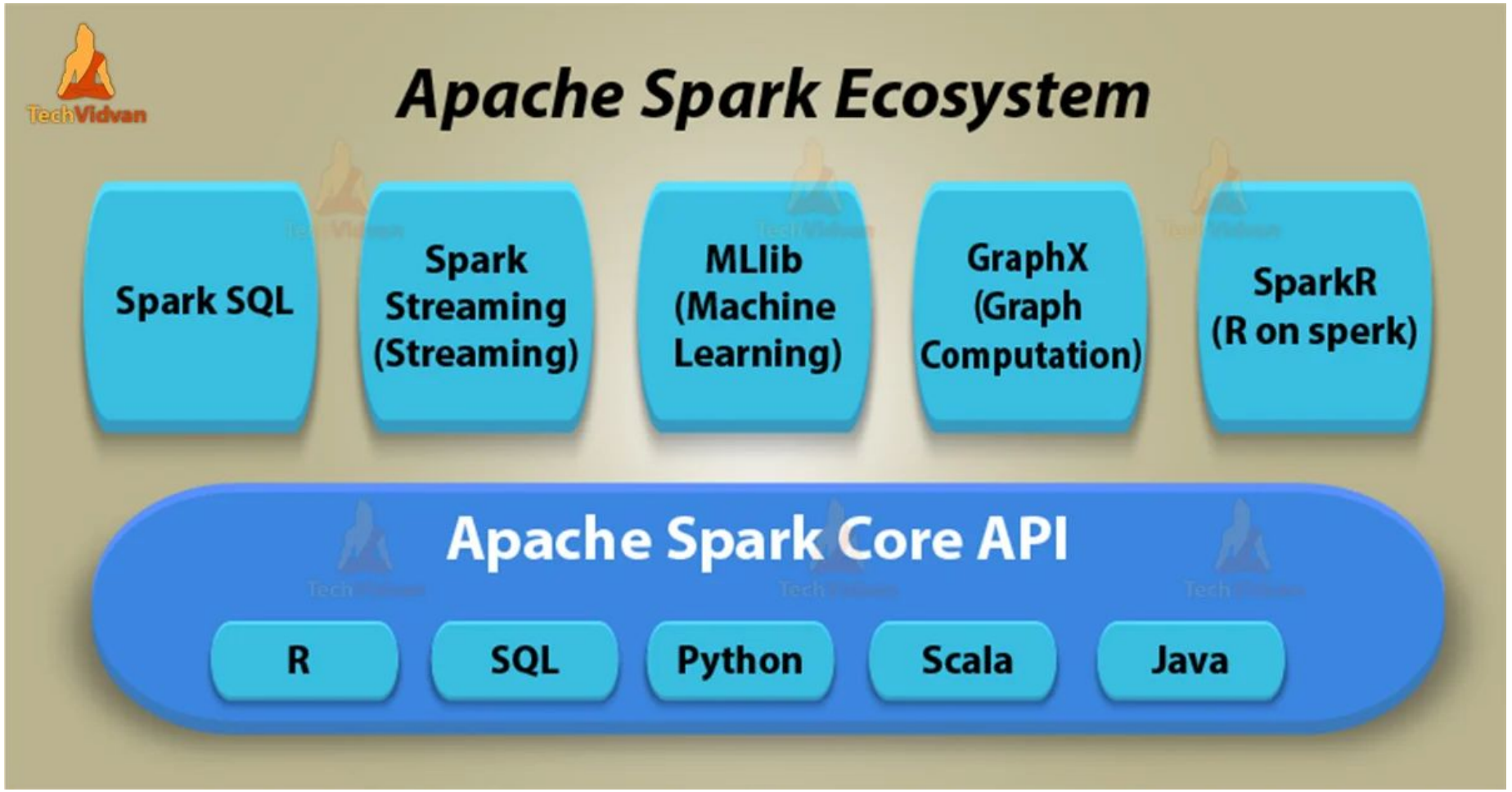
- Dados em batch:
  - São coletados ao longo do tempo.
  - Depois de coletados, dados são enviados para processamento.
  - Processamento pode ser demorado.
- Dados em stream:
  - São produzidos continuamente.
  - São processados um a um.
  - Processamento deve ser rápido e o resultado é necessário imediatamente.

Exemplos de dados em *stream*

- Finanças: preços de ações, identificar oportunidades de arbitragem.
- Monitoramento de logs e detecção de fraudes / hacking / DDOS.
- Sites de e-commerce: Clickstream.

Exemplos de dados em *stream*

ecossistema do Spark



Spark Streaming

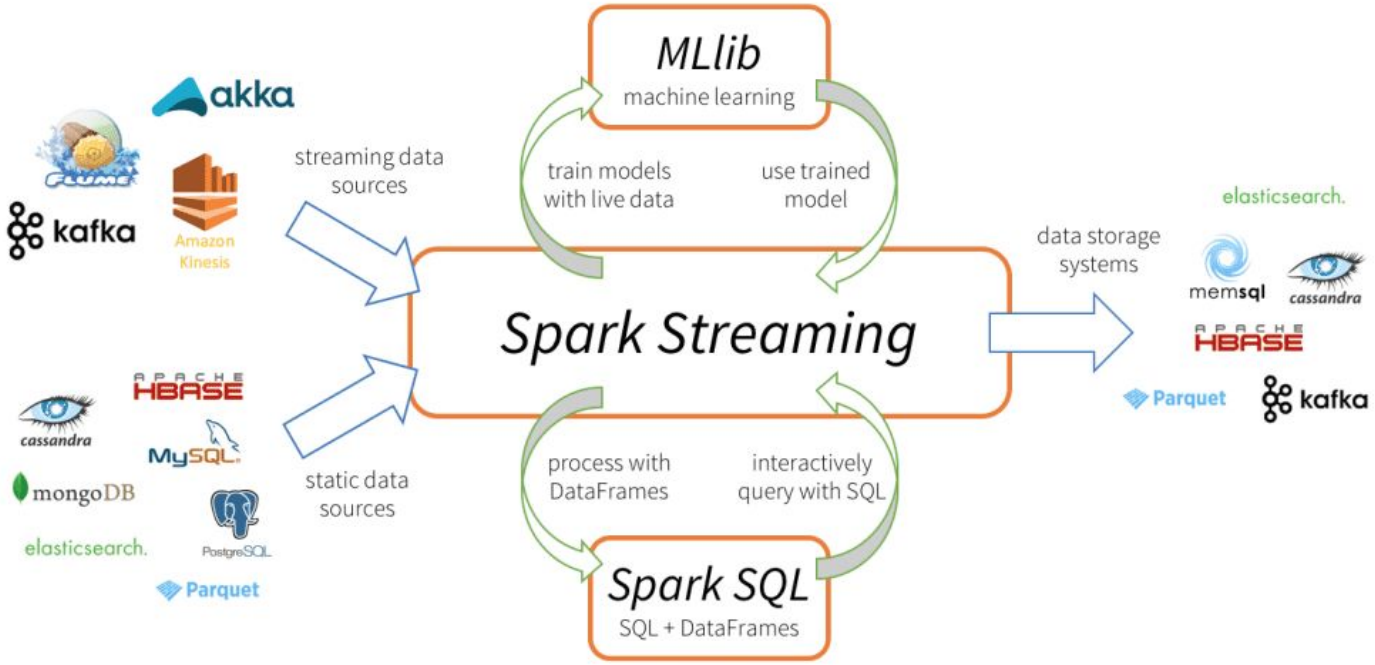
- Recuperação rápida de falhas e lentidões.
- Balanceamento de carga e otimização dos recursos.
- Combinação de dados em stream e dados estáticos.
- Integração com Spark SQL, GraphX e MLLib.

Spark Streaming

Spark Streaming

Spark Streaming

- Extensão da API *core* do Spark, que permite que Cientistas e Engenheiros de Dados lidem com dados em *streams* oriundos de várias fontes.



Vamos escrever um pouco de código?