*Stack Tecnológico I*

Como bem sabemos, um(a) cientista de dados é um(a) profissional dedicado(a) exclusivamente à análise e interpretação de grandes bancos de dados. Para isso, deve aprender a utilizar múltiplas ferramentas, as quais vamos explorar ao longo desta seção! Vamos começar falando sobre o mundo dos Bancos de Dados.

**Bancos de Dados**

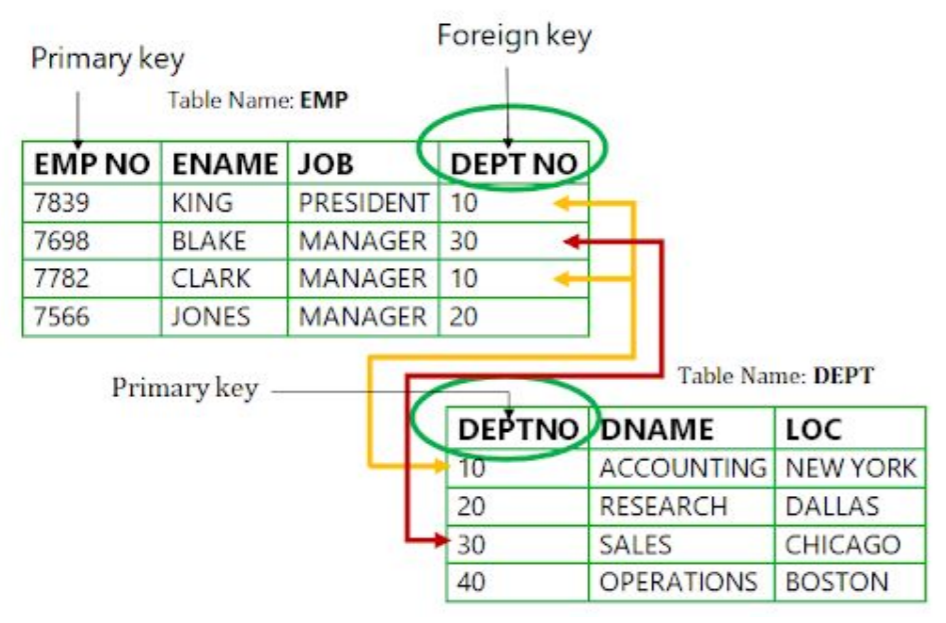
Como já comentado em seções anteriores do curso, um Banco de Dados (ou base de dados, DB – Database), é um conjunto de dados que pertencem a um mesmo contexto e permanecem sistematicamente armazenados para seu uso posterior.

Em termos gerais, podemos dividir o mundo dos Bancos de Dados em dois grandes tipos:

* Bancos de Dados Relacionais.
* Bancos de Dados não Relacionais – NoSQL.

**Bancos de Dados Relacionais**

Representam o Modelo de Dados relacional proposto por Edward Frank Codd em 1970, que podemos observar nesta imagem:



Alguns dos Motores de Bancos de Dados Relacionais mais populares são:

**Microsoft SQL Server:** é um Sistema de Gestão de Bancos de Dados Relacionais (RDBMS) que abrange uma ampla variedade de aplicações de processamento de transações, inteligência empresarial e análise em contextos corporativos de informática.



**MySQL:** é um Sistema de Gestão de Bancos de Dados Relacionais desenvolvido sobre uma licença dual: Licença Pública Geral/Licença Comercial pela Oracle Corporation. Além disso, em geral, é um dos bancos de dados mais populares, junto ao Oracle e o SQL Server.



**PostgreSQL:** é um Sistema de Gestão de Bancos de Dados Relacionais multiplataforma, open-source e orientado a objetos. Foi desenvolvido em 1996 pela comunidade de SGBD POSGRES.



**Outras**: Oracle Database, IMB DB2, Access, SQL Cloud e mais!

**Tecnologias de Bancos de Dados não Relacionais**

Nos últimos anos, a quantidade de dados digitais gerados no mundo têm se multiplicado. As redes sociais e a internet têm sido muitíssimo importantes para esse processo.

É importante mencionar que, com o surgimento dos bancos de dados relacionais, as empresas encontraram o aliado perfeito para cobrir suas necessidades de armazenamento, disponibilidade, backups e gestão de dados. No entanto, devido às tendências atuais em relação às quantidades gigantescas de dados, esse tipo de sistema tem começado a experimentar dificuldades técnicas.

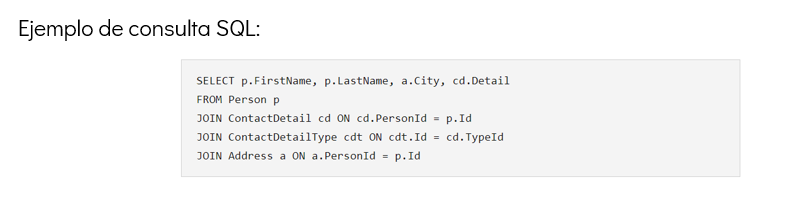
Os tipos de dados massivos recebem o nome de BigData, e o tipo de tecnologia que surgiu para solucionar muitos desses problemas é conhecida como NoSQL.

O termo NoSQL foi inicialmente utilizado no ano 1998 para denominar um banco de dados relacional que não utilizava a linguagem SQL para funcionar. A partir desse acontecimento, o termo foi resgatado em 2009 em bate-papos de defensores dos bancos de dados não relacionais.

Essencialmente, os Bancos de Dados NoSQL buscam:

* Evitar complexidade desnecessária.
* Conseguir um alto rendimento.
* Escalabilidade horizontal e hardware de baixo custo.
* Transformar a famosa frase: “One size fit’s it all” .

Exemplo de consulta SQL:





Alguns dos Bancos de Dados NoSQL mais populares são:

**MongoDB:** é um banco de dados orientado a documentos, ou seja, em vez de armazenar os dados em registros, os armazena em documentos. Esses documentos são armazenados em BSON, uma representação binária do JSON.



**Apache Cassandra:** é um software NoSQL distribuído e baseado em um modelo de armazenamento de «chave-valor», com código aberto e é escrito em Java. Permite utilizar grandes volumes de dados de forma distribuída.



**Redis:** é um motor de banco de dados em memória baseado no armazenamento em tabelas de hashes (chave/valor), mas que, opcionalmente, pode ser usado como um banco de dados durável ou persistente.



**Neo4j:** é um banco de dados open-source orientado a grafos e escrito em Java. Com esse tipo de banco de dados NoSQL, é possível armazenar informação em formato de nodos e relacionais.



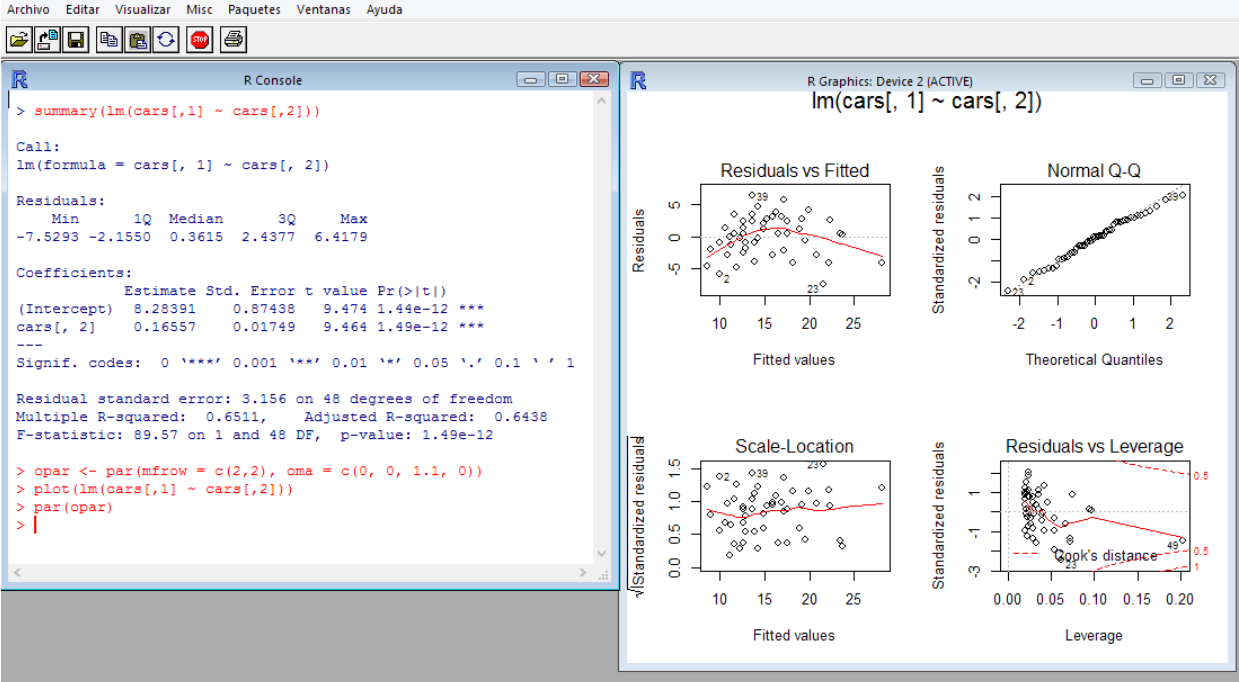
**Outras**: Hbase, CouchDB, NoSQL Cloud e mais!

**Linguagens de Data Science**

A **R** é um ambiente e linguagem de programação orientada a objetos, com foco em análise estatística. É uma das linguagens de programação mais utilizadas em pesquisas científicas. O R fornece uma ampla variedade de técnicas estatísticas (modelos lineares e não lineares, testes estatísticos clássicos, análises de séries temporais, classificação, agrupamento, etc.) e geração de gráficos, além de ser altamente extensível.



**Como o R é exibido?**



Um pouco retrô, não é mesmo?

Realmente, a interface gráfica da R não é muito atrativa ou intuitiva. Por isso, para nossas aulas práticas, trabalharemos com o RStudio, o IDE da R.

**O que é um IDE?**

Um Ambiente de Desenvolvimento Integrado ou Ambiente de Desenvolvimento Interativo (IDE, Integrated Development Environment, em inglês) é um programa que fornece serviços integrais para facilitar a atuação do(a) desenvolvedor(a) ou programador(a) no desenvolvimento de softwares ou estruturas computacionais. No caso do RStudio em particular, ele nos disponibiliza uma interface muito mais amigável e conveniente para trabalhar com R.



**R STUDIO**

Ao utilizar o RStudio, teremos várias vantagens. As mais importantes são:

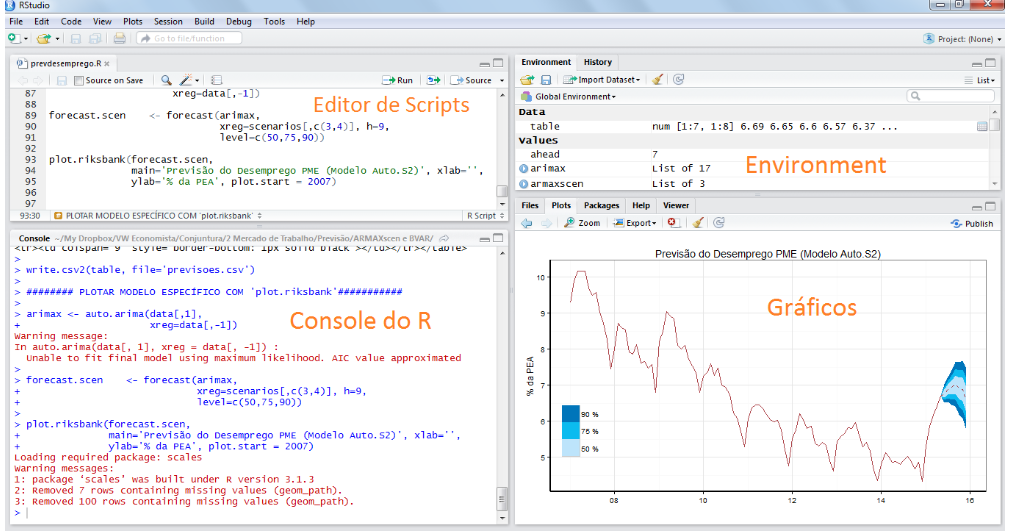
1. Recurso de Autocompletar.

2. Reconhecimento de sintaxe de programação.

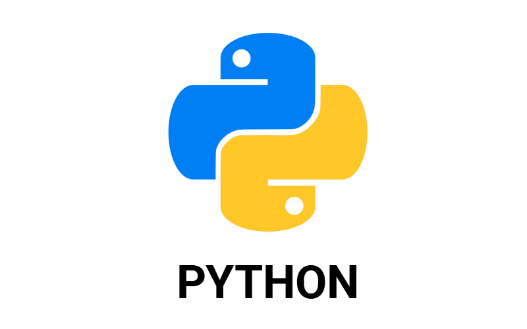
3. Depurador de erros (debug).

4. Manual de usuário e ajuda online.

Como o RSTUDIO é exibido?



O Python é uma linguagem de programação poderosa e fácil de aprender. Pode ser classificada como uma linguagem interpretada (executa as instruções conforme as lê) e de alto nível. O Python foi criado no final dos anos 80 por um programador holandês chamado Guido van Rossum, que até hoje continua sendo o líder de desenvolvimento da linguagem.





**Ferramentas de Visualização de Dados**



O Power BI é um conjunto de ferramentas e serviços de Inteligência de Negócios (Business Intelligence) que permite a conexão a diferentes origens de dados para serem analisados, visualizados e compartilhados com toda a organização e clientes. Além disso, é composto por várias aplicações e serviços (versão desktop, mobile e o servidor).



O Tableau é uma potente ferramenta de visualização de dados utilizada na área da Inteligência de Negócios (Business Intelligence). A essência do Tableau é simples e, ao mesmo tempo, muito relevante: ajudar as pessoas e empresas a visualizar e compreender todos os seus dados.



Tableau está disponível em 3 versões principais:

* Desktop (Tableau Desktop).
* Servidor (Tableau Server).
* Online (Tableau Online).

Além dessas três ferramentas principais, o Tableau integra outras ferramentas adicionais para proporcionar uma experiência a mais completa possível aos usuários:

* Tableau Mobile.
* Tableau Public.
* Tableau Prep.

QlikView é uma ferramenta de Inteligência de Negócios (Business Intelligence) que permite coletar dados de diferentes origens baseados em ERP, CRM, data warehouses, bancos de dados SQL, dados de Excel, etc.



O Google Data Studio é uma ferramenta de visualização de dados e criação de painéis de informação (dashboards). O objetivo da ferramenta é possibilitar uma análise dos dados de forma visual, obtendo resultados imediatamente e de maneira mais fácil.



O QuickSight é um serviço rápido de análise de negócios baseado em nuvem que facilita a criação de painéis de informação interativos (dashboard). Também permite a realização de análises ad-hoc e a obtenção rápida de informações de negócio baseadas em dados.



A IBM Cognos Analytics é uma plataforma de Inteligência de Negócios (Business Intelligence) integrada baseada em web. Fornece um conjunto de ferramentas para relatórios, análises, painéis de informação (dashboards) e acompanhamento de eventos e métricas.



A MicroStrategy é uma organização que fornece softwares de relatórios, análises e monitoramento integrados, o que permite que empresas analisem dados armazenados da própria empresa e de todas as áreas necessárias para tomar melhores decisões.



Outras ferramentas open-source para Visualização de Dados são:

* Shiny de R ou RMarkdown.
* Flourish em sua versão FREE.
* Apache Superset, etc.



É importante mencionar que existem muitas ferramentas projetadas para a gestão de comunicação, versionamento de dados e trabalho em equipe que precisamos nos familiarizar como Cientistas de Dados. Algumas das mais importantes são:

* Slack.
* Trello.
* G - Suite ou M365.
* Mural.
* Jamboard.
* Whatsapp ou Telegram.
* GitHub, GitLab
* Zoom, Meet, Teams, Jitsi, etc.

