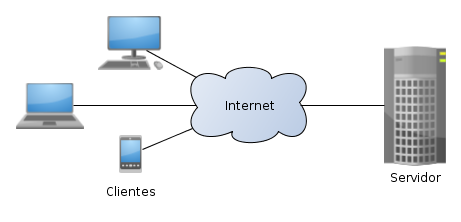
*Stack Tecnológico II*

**Computação em Nuvem**

Atualmente, o Cloud Computing é uma tendência em alta, especialmente no mundo da Análise ou Analítica de Dados. Como Data Scientists, é muito importante entendermos o que é o Cloud, suas características e quais são os diferentes serviços que podemos utilizar na nuvem para complementar nosso Stack Tecnológico.

**Arquitetura Cliente – Servidor**

****

Pontos chave da arquitetura:

* Cliente.
* Servidor.

**O que é o Cloud Computing?**

O Cloud Computing ou Computação em Nuvem é uma nova tecnologia que nos permite acessar softwares, processamentos de dados e armazenamentos de arquivos através de um sistema remoto.



Tipos de armazenamento em Nuvem disponíveis:

* Nuvem pública.
* Nuvem privada.

**Vantagens proporcionadas pelo Cloud Computing para Big Data**

Existem muitas vantagens e benefícios que podemos mencionar sobre a implementação de Cloud em nossas organizações, as mais relevantes são:

* Barateamento de custos.
* Operações ágeis.
* Capacidade de processamento.
* Acesso simultâneo.
* Segurança.

**Importante:** O Cloud Computing permite a aceleração e a gestão de processos computacionais, os tornando mais eficientes.

**Considerações importantes para serem analisadas**

**Vantagens In-House:**

* **Customização:** comnossa própria infraestrutura de redes e dados, podemos aproveitar uma maior customização de recursos com base nas necessidades da organização.
* **Conhecimento:** o conhecimento dos sistemas e dados reside unicamente in-house. Isso pode ser mais adequado para companhias com requerimentos de segurança muito específicos.
* **Sem custos mensais:** geralmente, os planos de deployment on-premise são estruturados em planos anuais ou multianuais, potencialmente eliminando a necessidade de custos mensais.

**Desvantagens In-House:**

* **Flexibilidade**: falta de flexibilidade para o uso de mais recursos quando for necessário devido a um pico de demanda.
* **Custos de Infraestrutura**: os custos de infraestrutura para montar um datacenter próprio são muito superiores se os compararmos com a aquisição de uma licença com algum provedor de serviços Cloud.
* **Pessoal altamente qualificado**: Se quisermos ter nossa infraestrutura In–House, deveremos contar com pessoal altamente qualificado em temas de Segurança e Infraestrutura Informática.

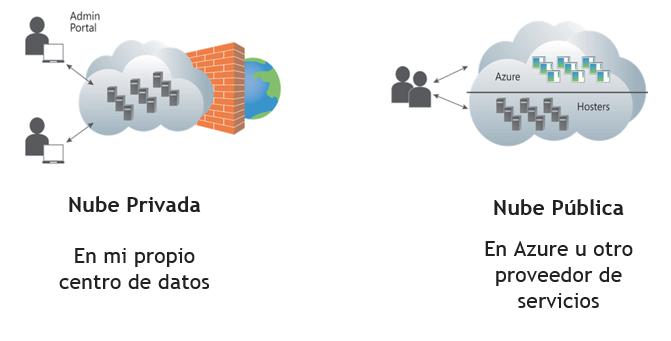
**Modelos de Serviços em Nuvem:**

* Infraestrutura como serviço (IaaS).
* Plataforma como serviço (PaaS).
* Software como serviço (SaaS).



**Tipos de Nuvem**

* Nuvem Privada.
* Nuvem Pública.

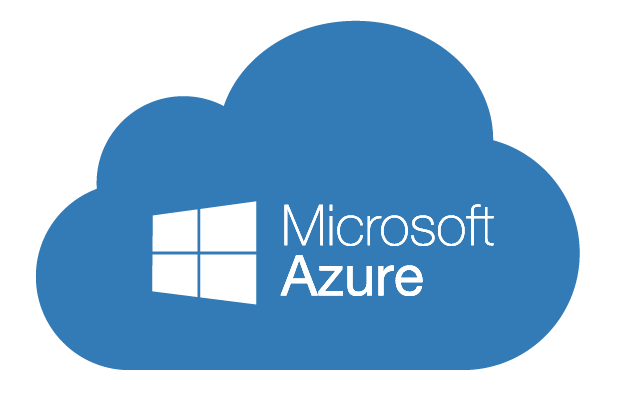


**Provedores de Serviços em Nuvem**

A Amazon Web Services, uma das plataformas mais conhecidas de serviços em nuvem, fornece uma variedade de serviços de infraestrutura, como armazenamento, redes, bancos de dados, serviços de aplicações, potência de computação, comunicação, inteligência artificial, serviços móveis, segurança, identidade e conformidade, entre outros. Qual é sua principal vantagem? Sem dúvidas, o fato de ser uma plataforma consagrada no mercado e o serviço de atendimento ao cliente que oferece.



O Microsoft Azure fornece uma série de serviços na nuvem, incluindo serviços vinculados à computação, analítica, armazenamento e redes. Os usuários podem escolher entre esses serviços para desenvolver e escalar novas aplicações ou executar aplicações existentes. Qual é sua principal vantagem? A rapidez.



O Google Cloud Platform, também conhecido pela sigla GCP, é uma plataforma que contém diversos serviços que funcionam na mesma infraestrutura utilizada internamente pelo Google em serviços como o Youtube e o Google Search. Qual é sua principal vantagem? A segurança.



**Agora, qual escolher?**

É importante destacar que as diversas soluções mencionadas anteriormente oferecem uma série de vantagens, semelhanças e certas diferenças. A escolha de uma ou outra tecnologia deverá ser feita considerando as necessidades computacionais da organização e com acompanhamento de uma assessoria feita por um profissional capacitado na área. Dito isso, uma boa recomendação, antes de tomar uma decisão definitiva sobre o assunto, é obter uma amostra gratuita de cada uma das ferramentas para experimentar todas elas.



**Big Data**

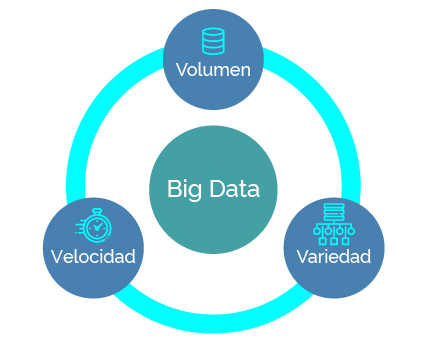
*Fundamentos Gerais*

*A Big Data é uma coleção de grandes volumes de dados complexos e muito difíceis de processar utilizando ferramentas de gestão e processamento de dados tradicionais.*

Impulsores do Big Data: sem dúvidas, existem vários fatores propagadores do Big Data, podemos mencionar:

* IoT.
* Redes Sociais.
* A internet.
* TICs.
* Cloud Computing.

**Os 3 V’s do Big Data**

****

**Principais Ferramentas de Big Data**

*Hadoop*

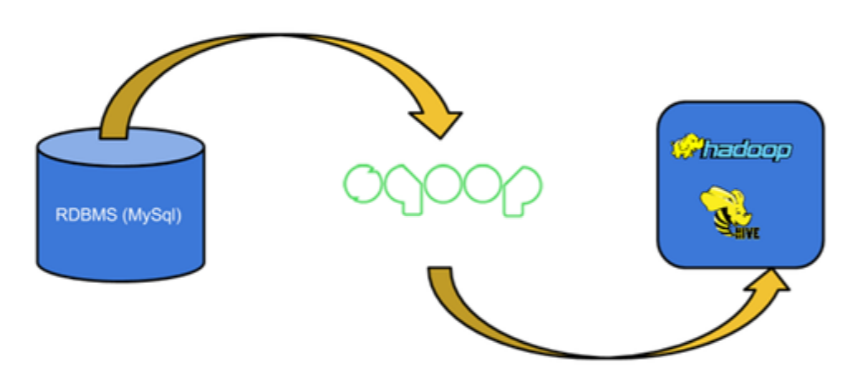
Sem dúvidas, o primeiro software utilizado para Big Data foi o Apache Hadoop. Mas o que é o Hadoop? É um framework implementado em Java que permite o armazenamento e processamento distribuído de grandes conjuntos de dados estruturados, semiestruturados e não estruturados. Foi projetado para trabalhar em clusters e tem uma alta tolerância a falhas.

Para que serve o Hadoop?

* Armazenar arquivos de forma distribuída.
* Processar arquivos de forma distribuída.

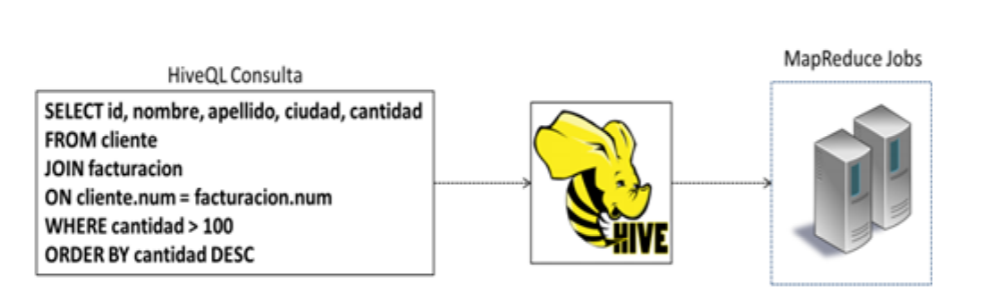
*Apache Sqoop*

É uma ferramenta que está dentro do projeto Apache e faz parte do ecossistema Hadoop. O Sqoop nos permite transferir dados entre um RDBMS, como o MySQL ou Oracle, e o HDFS.



*Hive*

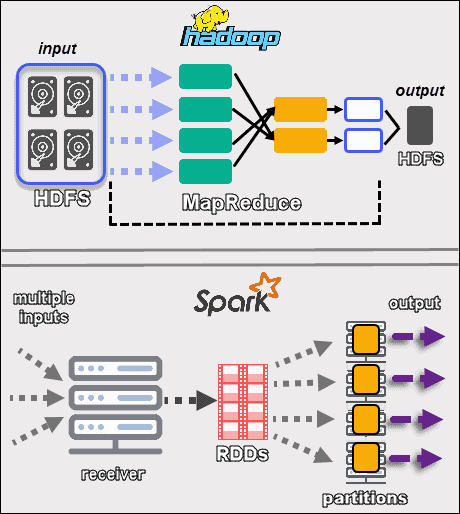
O que é o Hive? É um software que, usando SQL, facilita a leitura, escrita e gerenciamento de grandes conjuntos de dados que estejam em um armazenamento distribuído. Outro conceito que é importante mencionar é o MapReduce (MR). Basicamente, o MR é um modelo de programação utilizado em Hadoop que facilita o trabalho com grandes volumes de dados distribuídos através de múltiplas máquinas dentro de um cluster.



**Spark**

Apache Spark é um framework de propósito geral para o processamento de dados em um cluster. Certamente, ele tem suas próprias características, como:

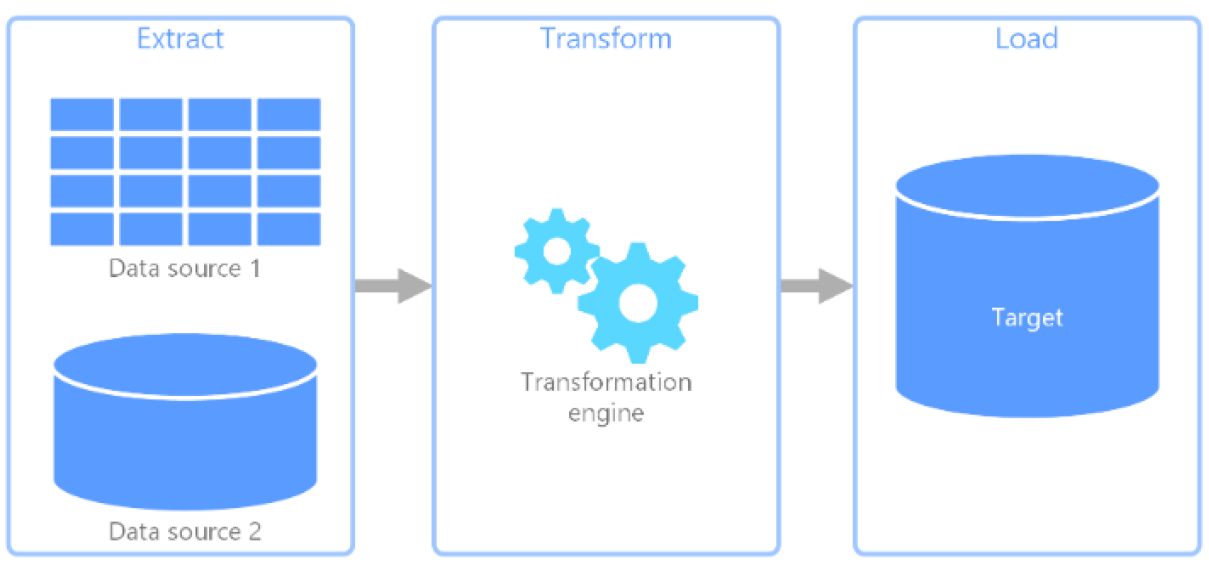
* Projeto de código aberto.
* Framework de processamento unificado (Batch e Streaming).
* Escalável e tolerante a falhas.
* Potente motor de processamento de dados massivos em memória (In-Memory Processing) sobre um cluster (O Spark está integrado ao Hadoop, mas não depende dele).

No Hadd

**No Hadoop ficamos limitados ao Java, enquanto que no Spark, podemos utilizar R, Python Scala o próprio Java entre outros.**

**ETL**

Extrair, transformar e carregar (ETL, Extract, Transform, Load) é o processo de compilação de dados a partir de um número ilimitado de fontes, sua posterior organização e seu carregamento em um repositório.



**Principais Ferramentas de ETL – On Premise**

O Oracle Data Integrator (ODI) foi lançado ao mercado em 2006. É uma plataforma de integração completa de ETL que cobre os requisitos de integração de dados. Além disso, também oferece um ambiente gráfico para construir, gerenciar e manter processos de integração de dados em sistemas de Inteligência de Negócios (Business Intelligence).



Sem dúvidas, o Pentaho Data Integration (PDI) é a ferramenta open-source mais utilizada e conhecida atualmente. A plataforma Pentaho nos permite implementar com grande eficiência os processos de extração, transformação e carregamento de dados. Além disso, nos possibilita fazer relatórios e análises sobre os dados. Também é possível adquirir uma licença da ferramenta para ampliar ainda mais suas capacidades.



O Talend oferece uma variedade de produtos para realizar a integração de dados e os processos de ETL. O Talend Open Studio for Data Integration é open-source e realmente muito potente, funcionando bem com integrações para Google Big Query e Redshift. Dito isso, existem múltiplas versões pagas que combinam diversas funcionalidades e ferramentas.



Criado pela empresa líder no setor, a solução SAS Data Integration disponibiliza uma ferramenta de design visual que simplifica significativamente a construção, execução e manutenção dos processos de integração de dados empresariais. Também conta com uma potente interface que é fácil de usar e intuitiva para o usuário (user-friendly).



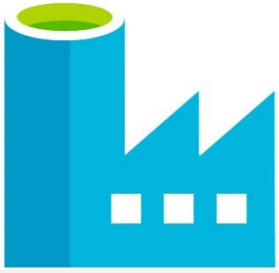
O SQL Server Integration Services, também conhecido pelas siglas – SSIS, é um componente do Microsoft SQL Server utilizado para a extração, transformação e carregamento de dados (ETL). Adicionalmente, é uma ferramenta muito popular e com muito potencial, sendo uma das mais reconhecidas e utilizadas atualmente no mercado, principalmente no setor bancário.



**Principais Ferramentas de ETL – On Cloud**

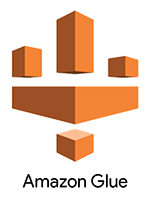
**Azure Data Factory:**

É um serviço serverless de ETL que nos permite realizar processos de *data integration*, *data transformation* e *data load*. Ao mesmo tempo, oferece uma Interface Gráfica muito intuitiva para a aplicação e criação de ETLs.



**AWS Glue:**

Serviço administrado pelo Cloud AWS que nos permite fazer processos ETL de forma visual para extrair ou esvaziar dados no AWS.



**Data Flow:**

Nos permite desenvolver ETLs desde o zero utilizando um pipeline ou fluxos de dados. Também nos ajuda a realizar a ingestão de informação em tempo real ou em Batch, de acordo com as necessidades da empresa.

