

Programa de Aceleração Banco Inter

Engenheiro de Dados, Ferramentas e Data Lake



data **sprints**
we enjoy sprinting

28jan2021

www.datasprints.com.br



AGENDA

1. Introdução e Contextualização;
2. Introdução a Data Lake;
3. Arquitetura de Data Lakes;
4. Tecnologias Essenciais em Engenharia de Dados;
5. Camadas de um Data Lake;
6. Data Lake na AWS.



Introdução

- Papel de um Data Engineer;
- Tipos de Data Engineers:
 - Analyser;
 - Builder;
 - Coder.



Principais Ferramentas

- Sistemas Distribuídos;
- Sistema Operacional Linux;
- Containers;
- Linguagens de Programação;
- DevOp's.



Linux

- Distribuições: CentOS, Debian, RedHat;
- Estrutura de arquivos e pastas;
- Comandos Básicos;
- Comandos para Debug.
- SSH, SFTP, SCP;
- Site de estudo: <https://www.tutorialspoint.com/unix>

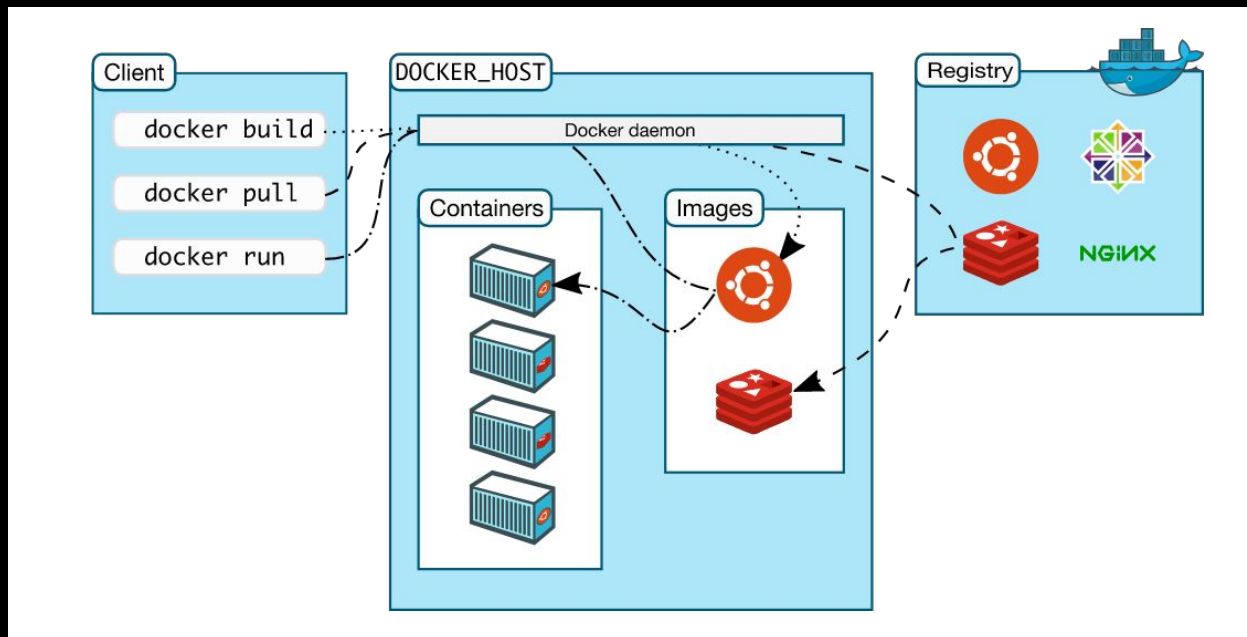


Docker e Kubernetes

- Docker:
 - Ferramenta para Containers;
 - Instalação em multiplataformas;
 - Repositório de imagens - DockerHub;
 - Fácil de iniciar:
 - `docker run -p 8888:8888 jupyter/datascience-notebook`

“

Docker - Arquitetura



<https://docs.docker.com/get-started/overview/>

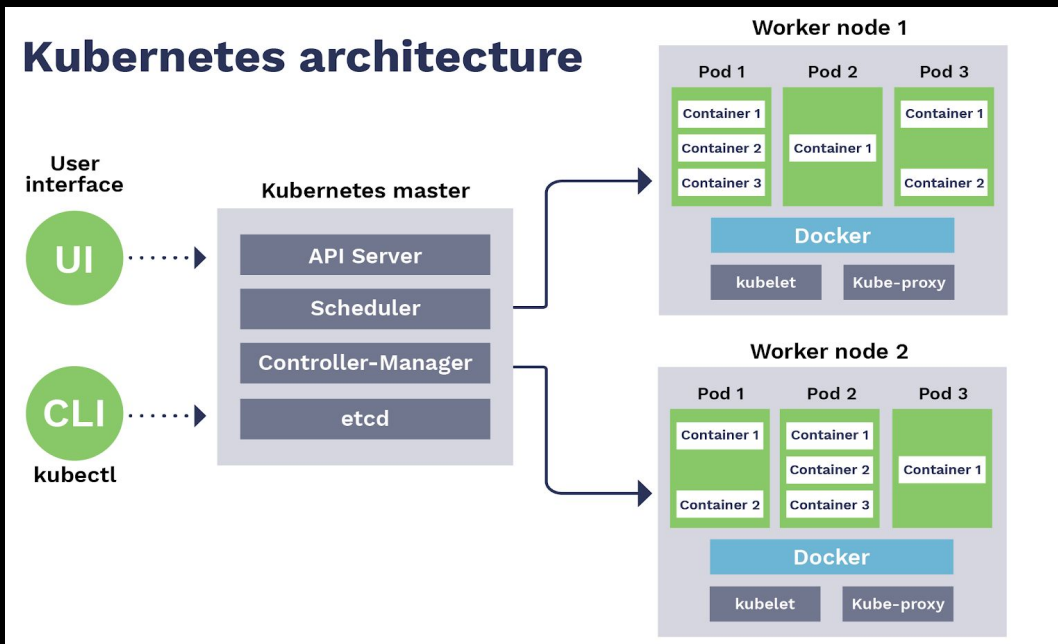


Kubernetes

- Cluster de Containers;
- Auto Scaling e Manager;
- Configurável por Arquivos;
- Ferramentas:
 - Amazon EKS;
 - OpenShift;
 - Google Kubernetes;
 - Kubectl.

“

Kubernetes - Arquitetura



<https://dzone.com/articles/how-kubernetes-works>



Linguagens de Programação

- Essencial para automatização de processos;
- Construção de ETL e ELT;
- Principais para Engenharia de dados;
 - Python;
 - SQL;
 - Java;
 - PySpark e Scala (Spark);



DevOp's

- Integração entre Desenvolvimento e Operação;
- Automatização de ambientes;
 - Deploy, Testes, Replicação e Qualidade;
- Algumas ferramentas:
 - Git;
 - Docker e Kubernetes;
 - CI/CD (Jenkins, GitLab...)
 - IaC (Infra as code);



Git e GitLab

- **Git:**
 - Ferramenta para controle de versão;
 - Principais comandos: <https://www.tutorialspoint.com/git>
- **GitLab:**
 - Repositório para arquivos;
 - Organização de projetos, segurança, CI/CD...
 - Alternativas: GitHub, BitBucket...

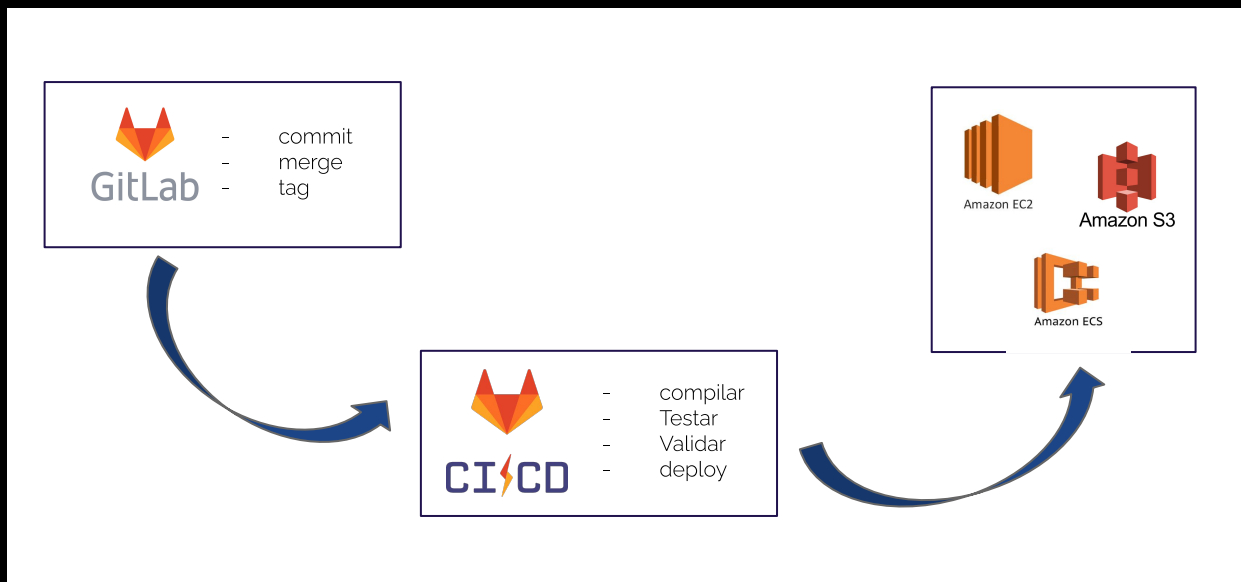


CI/CD GitLab

- Integração contínua de código (CI);
- Entrega contínua de código (CD);
- Automatização de uma parte do processo de DevOp's:

“

CI/CD GitLab - Arquitetura



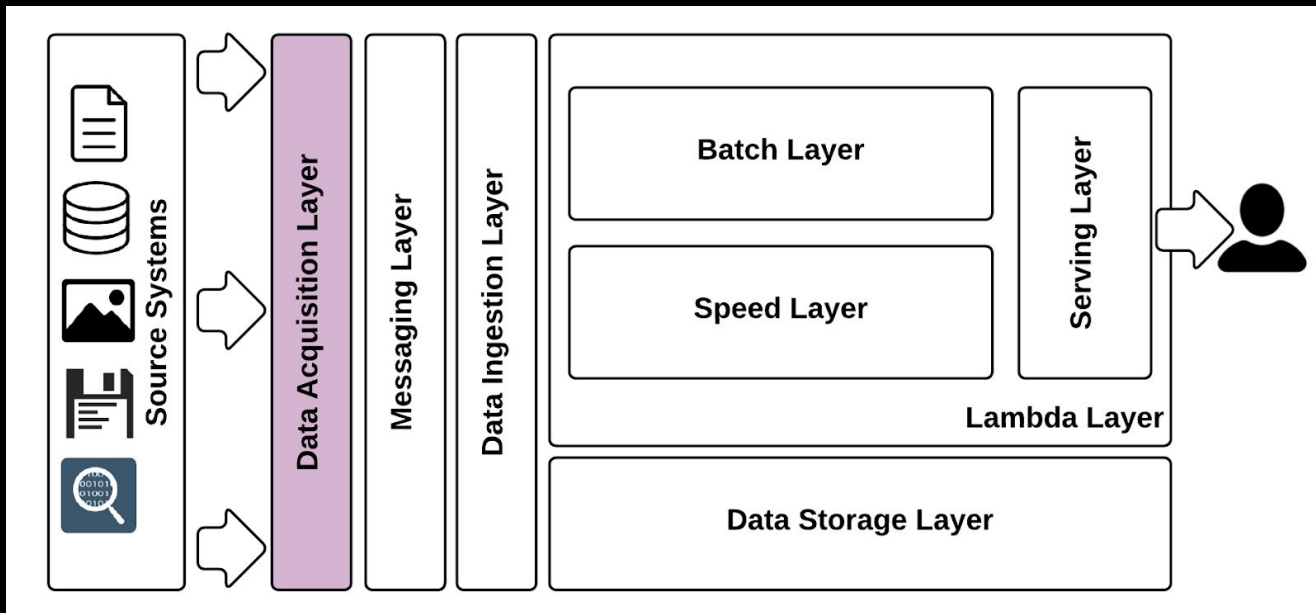


Data Lake

- Repositório central de Big Data;
- Organização Arquitetural;
- Muito além do armazenamento;
 - Processamento;
 - Análise dos dados;
 - Segurança;
 - Governança..

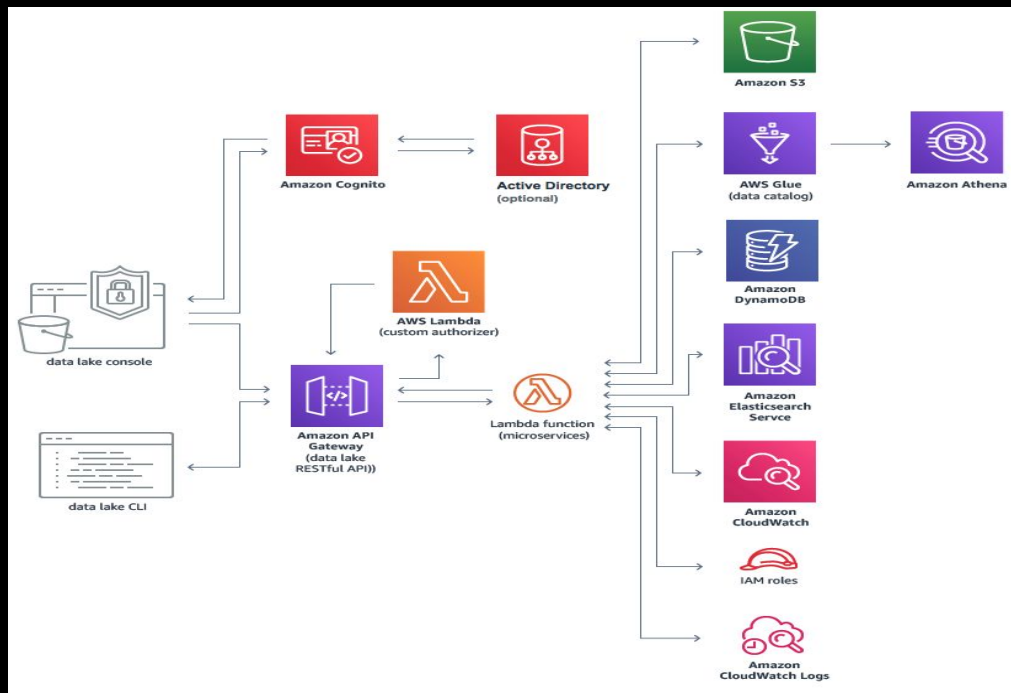
“

Data Lake - Camadas



https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781787281349/5/ch05lv1sec42/context-in-data-lake-data-acquisition

“Data Lake na AWS



<https://docs.docker.com/get-started/overview/>

“Atividades

- Executar uma stack de aplicações com Docker
 - Ex: Kafka, ELK, Airflow....
- Adicionar de preferência em um repositório no github.

Obrigado!



data **sprints**

we enjoy sprinting

www.datasprints.com.br