Data Engineering

Filipe Nascimento
fgvn@cesar.school

C.E.S.A.R

C . E . S . A . R

Pessoas impulsionando inovação.

Inovação impulsionando negócios.

Everton Dias
etgdb@cesar.org.br

Janaína Branco
jcb@cesar.org.br

PROGRAMA

Neste curso iremos abordar desde os fundamentos do que é ser um ENGENHEIRO DE DADOS, conhecer técnicas, ferramentas e compartilhar casos de uso e projetos.

4 SEMANAS

2880 MINUTOS

48 HORAS

172800 SEGUNDOS

SOBRE MIM

Filipe Nascimento

36 anos
casado
4 gatos
+15 anos na área de tecnologia
Apaixonado por Tecnologia, Música,Gatos
Cultura Nerd (RPG, Star Wars) e
Futebol Americano.

linkedin.com/in/filipegvnscm

CONCEITOS

O QUE É ENGENHARIA DE DADOS ?

O QUE É ENGENHARIA DE DADOS ?

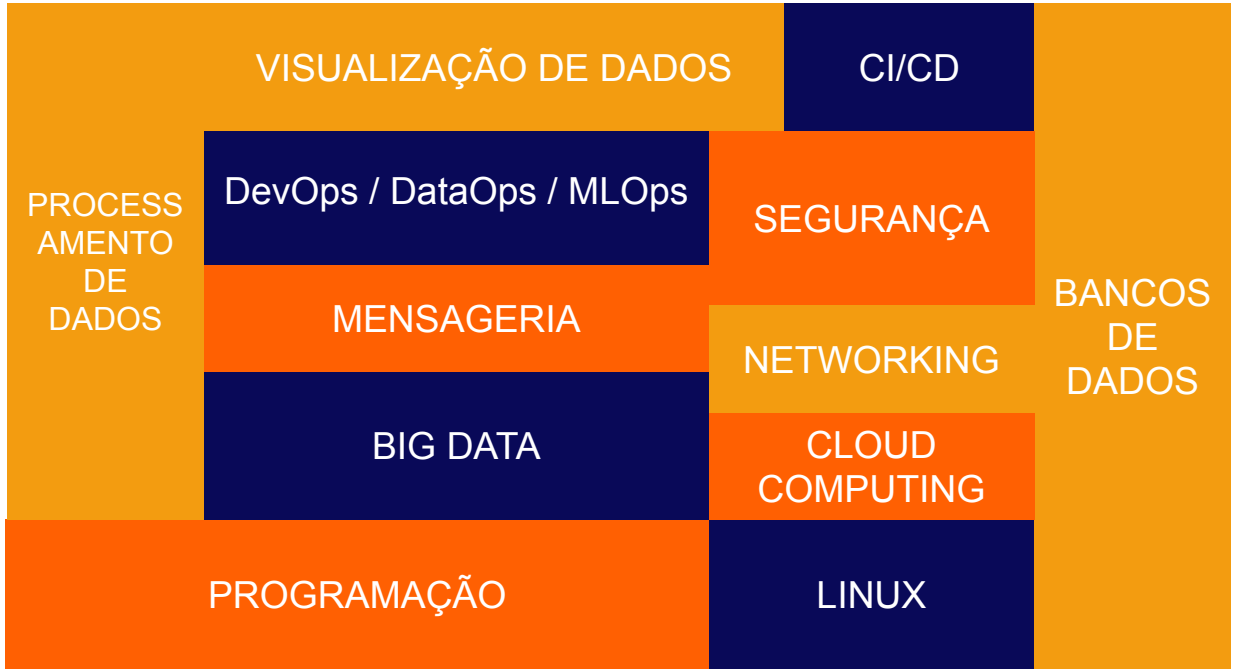
Segmento profissional que atua com as mais complexas atividades relacionadas aos dados de uma empresa, seja na transformação, guarda, transporte e disponibilização destes aos times de ciência de dados, analistas e áreas de negócio em geral.

Dentre estes casos podemos citar:

- Projetar rotinas de extração de dados e ingestão em um Data Lake;
- Escrever micro-aplicações para realizar transformações nos dados;
- Projetar e disponibilizar catálogo e repositório de dados para os times de negócio...

QUAIS SÃO AS HABILIDADES NECESSÁRIAS PARA SER UM ENGENHEIRO DE DADOS ?

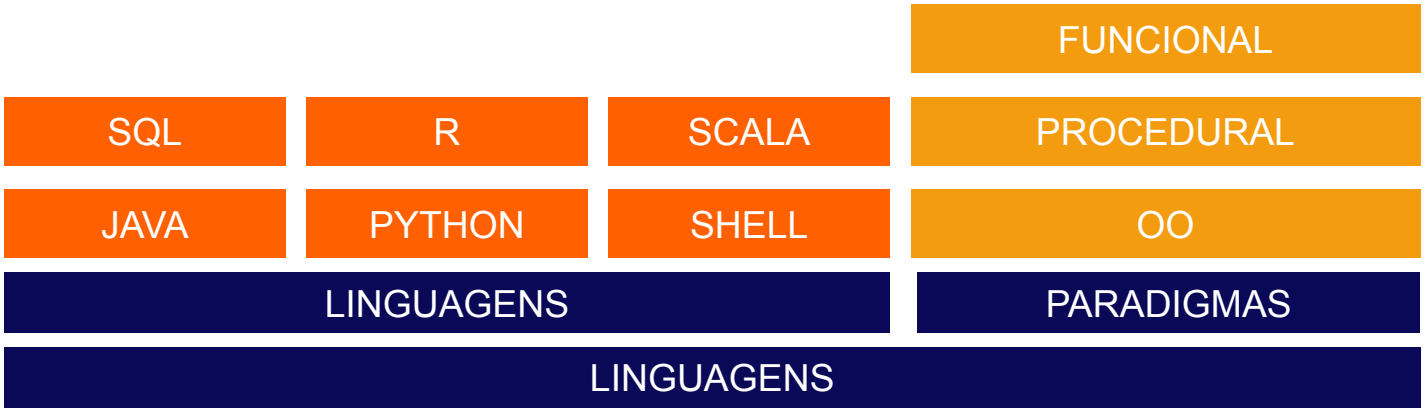
ROADMAP SKILLS



ROADMAP SKILLS

PROGRAMAÇÃO

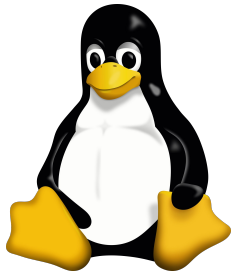
Extremamente importante para os engenheiros de dados é saber programar e de preferência ter em sua caixa de ferramentas mais do que apenas uma linguagem como também conhecer diferentes paradigmas e algoritmos.



ROADMAP SKILLS

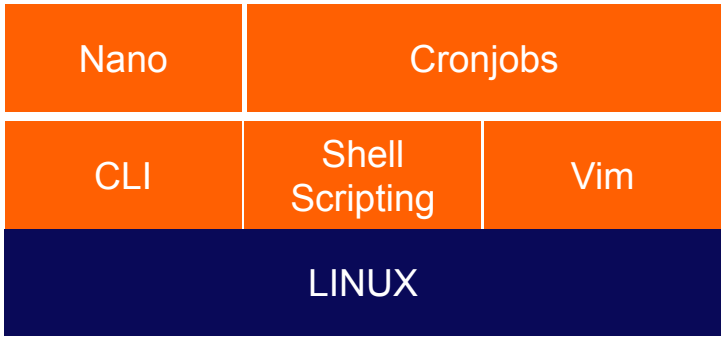
LINUX

Fundamental na vida do engenheiro de dados, principalmente por ter sua ampla utilização na computação distribuída e utilização massiva em projetos envolvendo big data.



Alguns tópicos sobre:

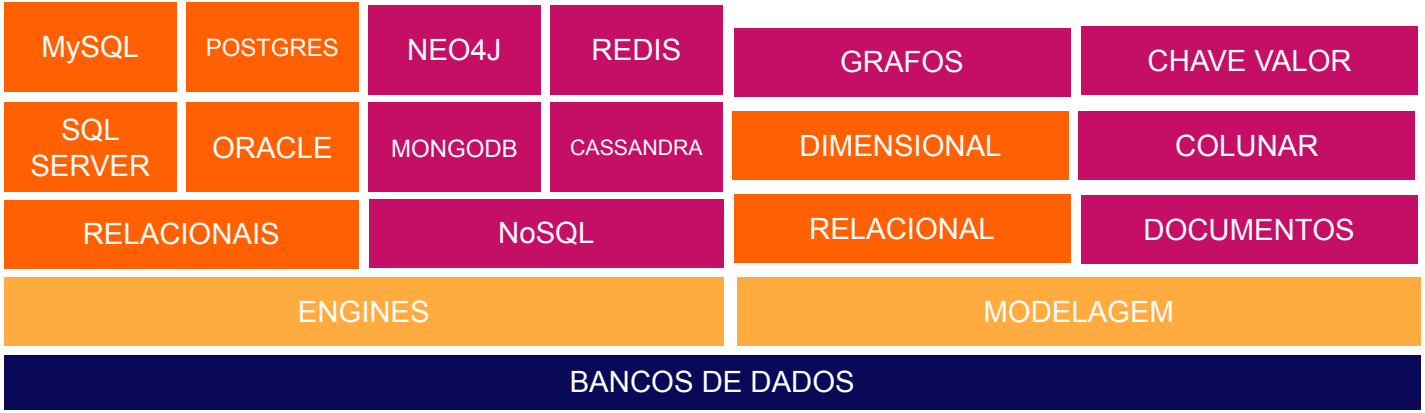
- Anunciado oficialmente em 1991 por Linus Torvalds - seu criador;
- GNU GPL;
- Desenvolvido no MINIX;
- ~600 distribuições pelo mundo;
- Bash - Bourne Again Shell um poderoso interpre tador de linguagem de comando.



ROADMAP SKILLS

BANCOS DE DADOS

Surgindo a partir das necessidades dos escritórios em armazenar dados em meados da década de 60, a notável evolução dos bancos de dados e a sua ocupação quase que central no mundo dos negócios, guardando bens valiosos e estratégicos para as organizações.



ROADMAP SKILLS

BIG DATA

Big Data é comumente definida como um segmento da engenharia de dados que lida com altos volumes, velocidade crescente e alta variedade - são os 3 V´s iniciais, porém temos muito mais como o V de veracidade e o V de valor.

Soluções para Big Data surgiram a partir do momento em que houve a necessidade de lidar com esse cenário de crescimento vertiginoso dos dados e seu uso, porém sem que conseguissem facilmente ser atendidos pelas engines de bancos de dados e processamento que já existiam. Neste segmento estudaremos conceitos e tecnologias dentro do contexto de Big Data como o Hadoop.

ROADMAP SKILLS

BIG DATA

VOLUME
O tamanho do dado em si.

VALOR
O quanto o dado representa para o negócio, não faz sentido utilizar big data sem que esse entregue valor.

VERACIDADE
O dado precisa representar a verdade de maneira acurada.

VELOCIDADE
A velocidade com o qual o dado é produzido.

VARIEDADE
Possuir diferentes tipos de dados de diferentes fontes.

ROADMAP SKILLS

CLOUD COMPUTING

Um fenômeno que cresceu bastante nos últimos anos e tem ocupado espaço significativo sem demonstrar nenhuma sombra de recuo é o uso da computação em nuvem.
Define-se como computação em nuvem a utilização de recursos computacionais dos mais variados em um ambiente externo, sem no entanto preocupar-se com o provisionamento e gestão do hardware.

Podemos classificar a computação em nuvem quanto aos modelos de serviços:

- SaaS -Software as a Service;
- IaaS - Infrastructure as a Service;
- PaaS - Platform as a Service

e quanto a sua implantação:

- Pública:
  - AWS, Azure, GCP;
- Privada
  - Openshift, Rancher;

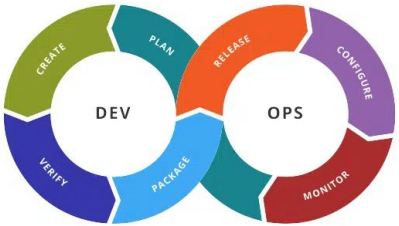
ROADMAP SKILLS

DevOps, DataOps e MLOps

DevOps - Vem da união das palavras Development e Operations, sendo mencionado pela primeira vez em 2009 por uma dupla do Flickr. Propõe um modelo de união de times, compartilhamento e aprendizado como cultura, entregando com velocidade sem perder a qualidade - Full Cycle.

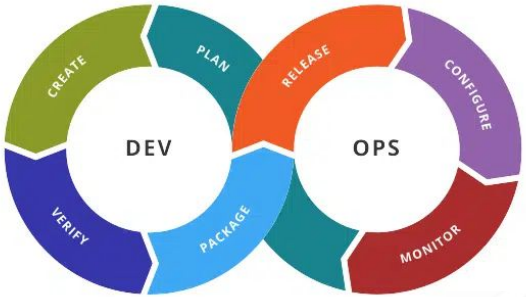
DataOps - Definido como um conjunto de práticas, tecnologias e processos que combinam automação de dados, métodos ágeis aplicados a engenharia de software visando agregar qualidade, velocidade e colaboração ao processo de Analytics.

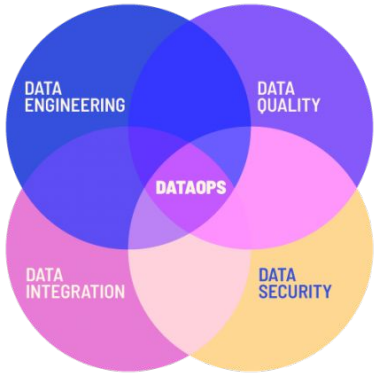
MLOps - Numa definição curta, é um conjunto de práticas visando manter e entregar modelos de machine learning em produção de maneira segura e eficiente.

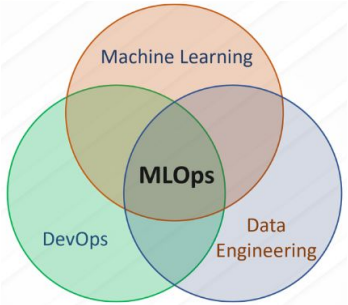


ROADMAP SKILLS

DevOps, DataOps e MLOps







ROADMAP SKILLS

NETWORKING

Conhecimento em redes é fundamental para um bom engenheiro de dados, conhecer protocolos de rede, serviços e camadas.

Comumente você terá que lidar com configuração de VPC em ambientes de nuvem, lidar com protocolos de rede para integrar-se a bancos de dados e garantir baixa latência quando necessário.

Buscar dados ou disponibilizar via SFTP, conexões com servidores via SSH serão parte diária de sua rotina.

ROADMAP SKILLS

SEGURANÇA

Em todos os aspectos que estaremos trabalhando a face da segurança é uma constante, seja com o dado repousado em um banco de dados ou sendo trafegado entre nós de um cluster por exemplo.
Dentro de cada tema abordaremos um aspecto de segurança como por exemplo:

- Criptografia de dados;
- Comunicação via protocolos seguros;
- Controle de acessos em bancos de dados;
- etc...

ROADMAP SKILLS

MENSAGERIA

Mensageria está fortemente relacionada a um conjunto de tecnologias e técnicas de modelagem que visam garantir o processamento de uma alta demanda de solicitações, mesmo se tivermos uma baixa vazão.

Estudaremos conceitos como Eventos, Filas e Tópicos.

ROADMAP SKILLS

PROCESSAMENTO DE DADOS

Não serão poucas vezes em que irão surgir necessidades de construção de pipelines seja para integração dos dados e ingestão em um Data Warehouse (ETL) quanto para um Data Lake (ELT).

Abordaremos um conjunto de técnicas para processamentos em batch (Hive, Spark), streaming (Spark Streaming), trazendo casos de processamento local e distribuído.

Também trabalharemos aspectos de dados estruturados e não-estruturados.

ROADMAP SKILLS

CRONOGRAMA

SEMANA 01

Conceitos
Linux
Container
Cloud
Computing

SEMANA 02

Bancos de Dados
SQL
NoSQL
Big Data

SEMANA 03

Big Data Cont.
Processamento de Dados
Hive
Spark

SEMANA 04

Projetos
Segurança
Governança de Dados

KAHOOT

e amanhã ...

VAMOS PARA O LADO LINUX DA FORÇA!