Escopo Treinamento

Primeiro dia (Introdução)

1. Introdução a Data Lake;
2. Tecnologias utilizadas:
    a. Sistemas Distribuídos;
    b. Sistema Operacional Linux.
    c. Containers (Docker e Kubernetes);
    d. Linguagens de programação;
    e. Versionamento de código (Git e Gitlab)
3. Arquitetura de um Data Lake;
4. Camadas de um Data Lake;
    a. Aquisição de dados:
        i. Batch;
        ii. Streaming;
    b. Armazenamento;
    c. Processamento;
    d. Orquestração;
5. Serviços da AWS para:
    a. Aquisição de dados;
    b. Armazenamento;
    c. Processamento;
    d. Segurança;
    e. Catalogação de dados.
6. Laboratórios:
    a. Comandos básicos e essenciais para o Linux;
    b. Execução de contêineres no Docker;
    c. Criação de arquivos Dockerfile e Docker Compose para o Docker;
    d. Python com SQL para extração de dados;
    e. Execução de scripts Python no Docker;
    f. Controle de versão com Git e GitLab;
    g. Jupyter EMR.

Segundo dia (Ingestão)

1. Introdução ao armazenamento:
    a. Política de acesso, organização e configurações do Amazon S3;
2. Processamento de dados por Batch;
    a. Conceitos e Tecnologias;
    b. Ferramentas:
        i. Python, PySpark, SQL;
        ii. AWS: EC2 e EMR
3. Processamento de dados por Streaming:
    a. Conceitos e Tecnologias;
    b. Ferramentas:
        i. NiFi, Kafka, Debezium;
        ii. AWS: MSK, Kinesis;
2. Catalogação
    a. Glue Crawler;

    b. Glue e Spark;
4. Laboratórios:
    a. Criação de script em Python para processo de ETL por batch;
    b. Criação de script em PySpark para processo de ETL por batch;
    c. Execução do Kafka e sua utilização com Python;
    d. Execução e criação de fluxo no NiFi;
    e. Execução do Debezium;
    f. Armazenamento dos dados no S3;

Terceiro dia (Armazenamento)

1. Introdução a armazenamento de dados no mundo de Big Data;
    a. Ecossistema Hadoop;
    b. Banco de dados;
        i. Relacionais e não relacionais
    c. Armazenamento de objetos;
2. Armazenamento de dados na AWS:
    a. S3:
        i. Conceitos;
        ii. Estrutura;
        iii. Organização;
        iv. Segurança;
    b. DynamoDB:
        i. Conceitos;
        ii. Estrutura;
        iii. Armazenamento e acesso.
    c. RDS:
        i. Conceitos
        ii. Estrutura;
        iii. Segurança;
    d. EMR:
        i. Conceitos;
        ii. Armazenamento no HBase e Hive;
    e. Redshift:
        i. Conceitos;
        ii. Criação de clusters.
3. Laboratórios:
    a. Criação de Lambda com S3 e DynamoDB;
    b. Armazenamento de dados no Redshift com Airflow;
    c. Processamento de dados com Debezium e S3.

Quarto dia (Processamento)

1. Organização do Data Lake;
    a. Raw, Stage, Curated, Analytics e Sandbox;
2. Conceitos e estratégias de processamento de dados;
3. Modelagem de dados (Data Lake, DW, Data Mart);
4. Ferramentas para processamento e transformação de dados;
    a. Python e bibliotecas de Data Science;

    b. Spark (PySpark e Scala);
    c. NiFi;
    d. DBT;
    e. Dremio;
5. Processamento de dados na AWS:
    a. EMR;
    b. Lambda;
    c. Glue ETL;
6. Laboratórios:
    a. Criação de Pipeline com DBT, Athena e ECS;
    b. Criação de transformações no Spark;
    c. Produtização de Spark com EMR e Lambda.

Quinto dia (Orquestração)

1. Conceitos de Pipeline de dados;
2. Orquestrão de Pipeline de dados;
3. Ferramentas:
    a. Airflow;
    b. NiFI;
    c. AWS Lambda Step Functions;
    d. AWS CloudWatch Events
4. Laboratórios:
    a. Step Functions e Cloud Watch Events com DBT;
    b. Lambda com Spark e Terraform;
    c. Orquestração e monitoramento com Airflow.

Sexto dia (CI/CD e Estrutura do Inter)

1. Conceitos de DevOps;
2. Deploy em produção de um Data Lake;
3. Conceitos de Infraestrutura como código (IaC);
4. Ferramentas de IaC:
    a. Cloud Formation;
    b. Terraform.
5. CI/CD no Gitlab;
6. CI/CD Inter;
7. Laboratórios:
    a. Criação de ambiente com o iac do Inter.