Engenharia de Dados

18/02/2021

Programa de Aceleração Banco Inter

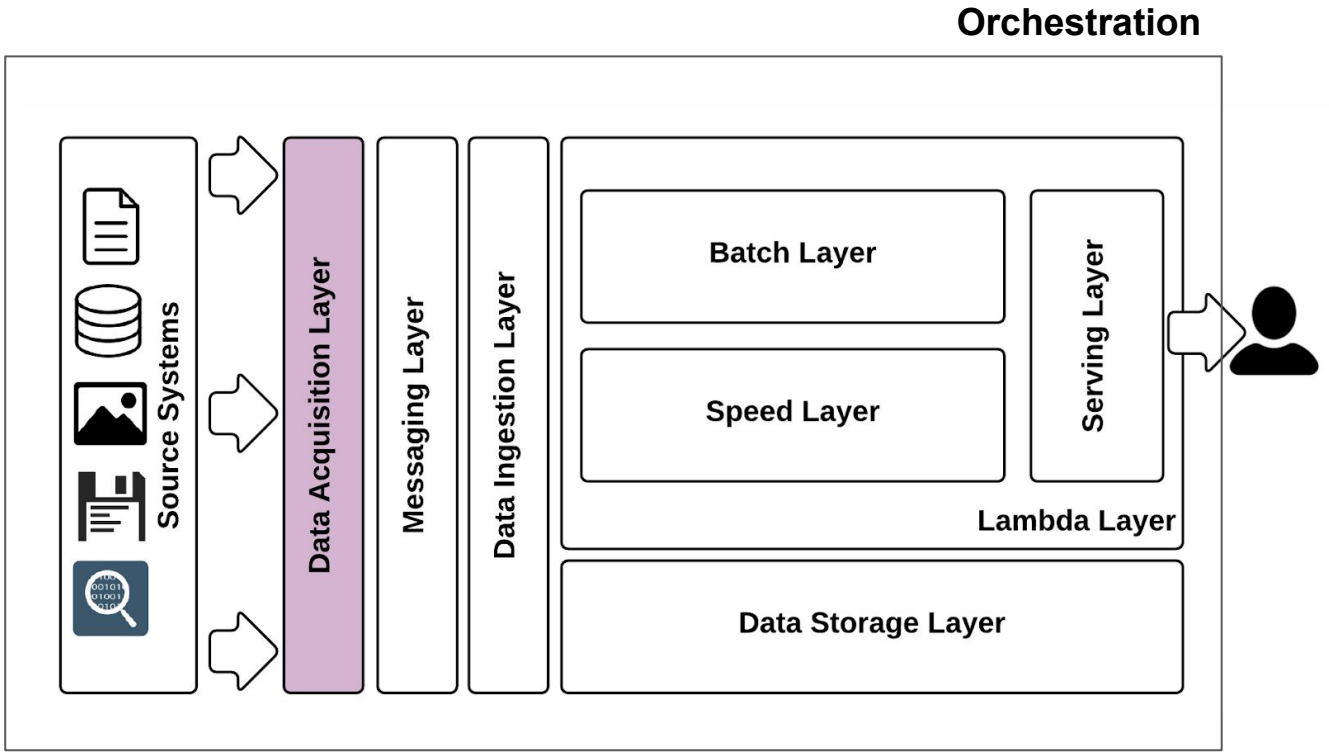
Orquestração para Pipelines de Dados

datasprints

Agenda

1. Camadas de um Data Lake;
2. Pipeline de dados;
3. DataOp's;
4. Ferramentas para Orquestração de Pipelines;
5. Orquestração de dados na AWS;
6. Boas práticas em Arquiteturas;
7. Case prático.

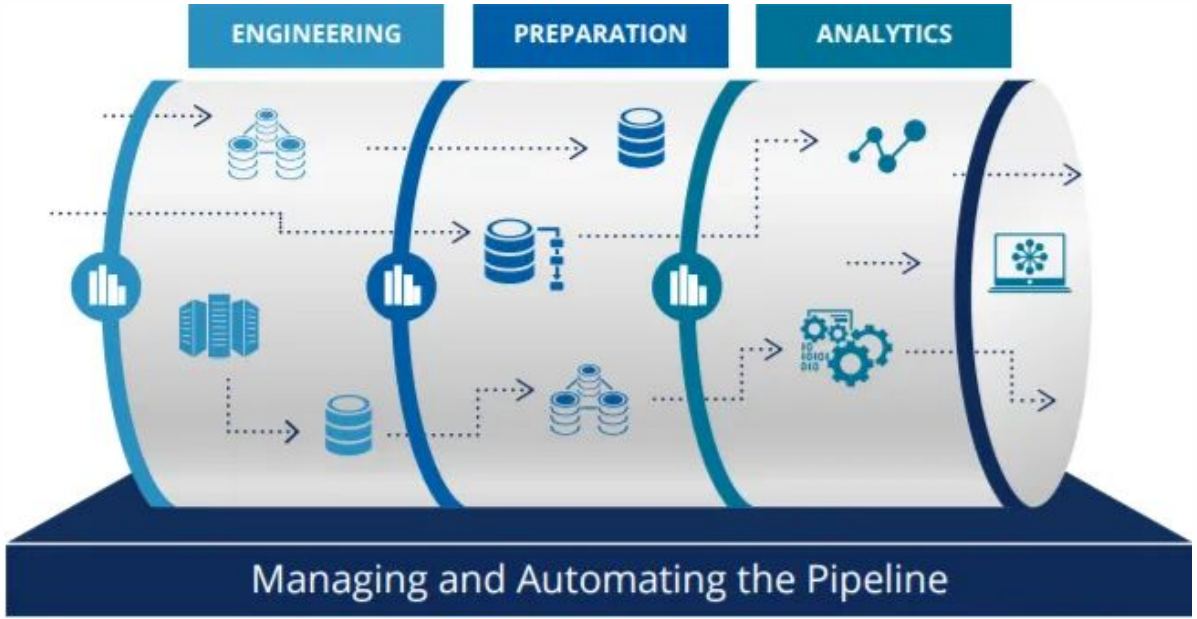
Camadas de um Data Lake



Pipeline de Dados

- Todo o caminho dos dados dentro da arquitetura:
  - Da fonte até o usuário final.
- Inclui:
  - Captura;
  - Ingestão;
  - Processamento;
  - Análise;
  - Orquestração.

Pipeline de Dados



https://airflow-tutorial.readthedocs.io/en/latest/pipelines.html

DataOp's: Introdução

- Uma adequação de DevOp's para o mundo dos dados;
- Metodologia;
- Junção da área de dados com métodos ágeis;
- Monitoramento e controle Pipeline de dados;
- Produtizar soluções de dados.

DataOp's: Conhecimentos

- Orquestração de Pipeline:
- Testes Automatizados;
- Implantação de software.

Ferramentas para Orquestração

datasprints

Orquestração de Dados na AWS

- Lambda:
  - Serverless;
  - Integração nativa com diversos serviços;
  - Trigger e destinos diversos.

- Step Functions:
  - Serverless;
  - Orquestração de funções;
  - Sequenciamento e carga de trabalhos.

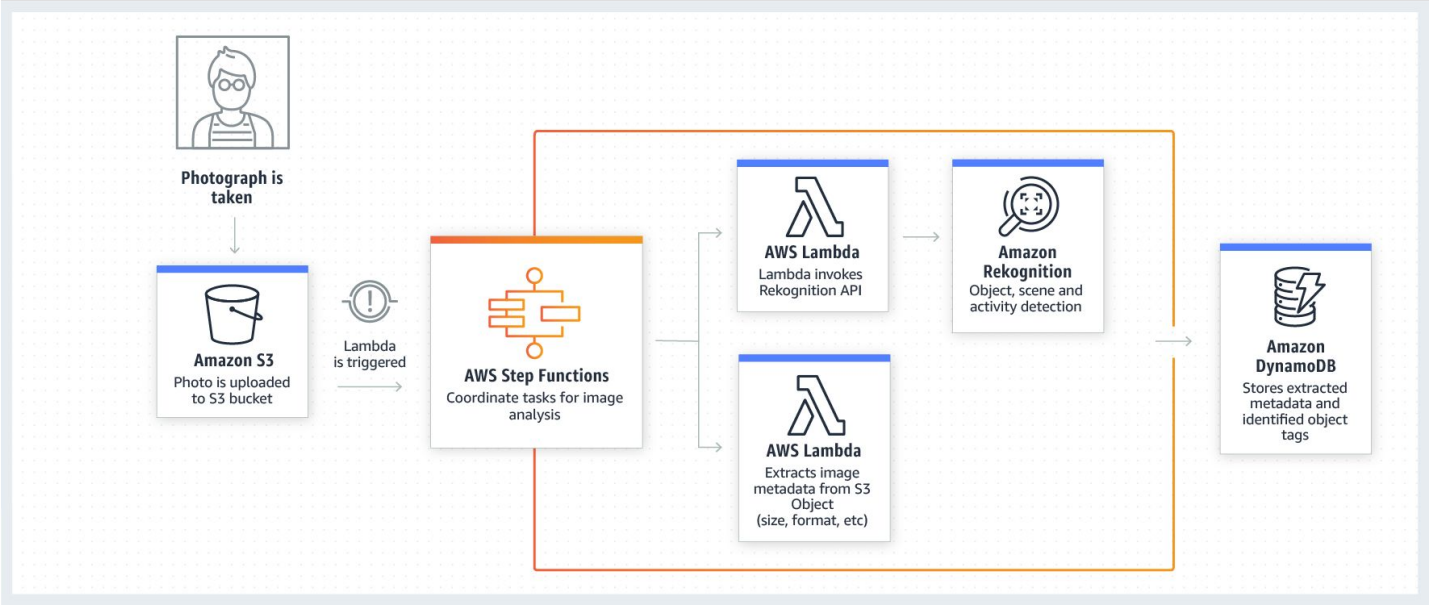
Orquestração de Dados na AWS

- CloudWatch:
  - Events (trigger e cron);
  - Logs;
  - Monitoramento em geral.

- MWAA:
  - Apache Airflow gerenciado;
  - Integração fácil com diversos serviços AWS.

Arquiteturas com Orquestração

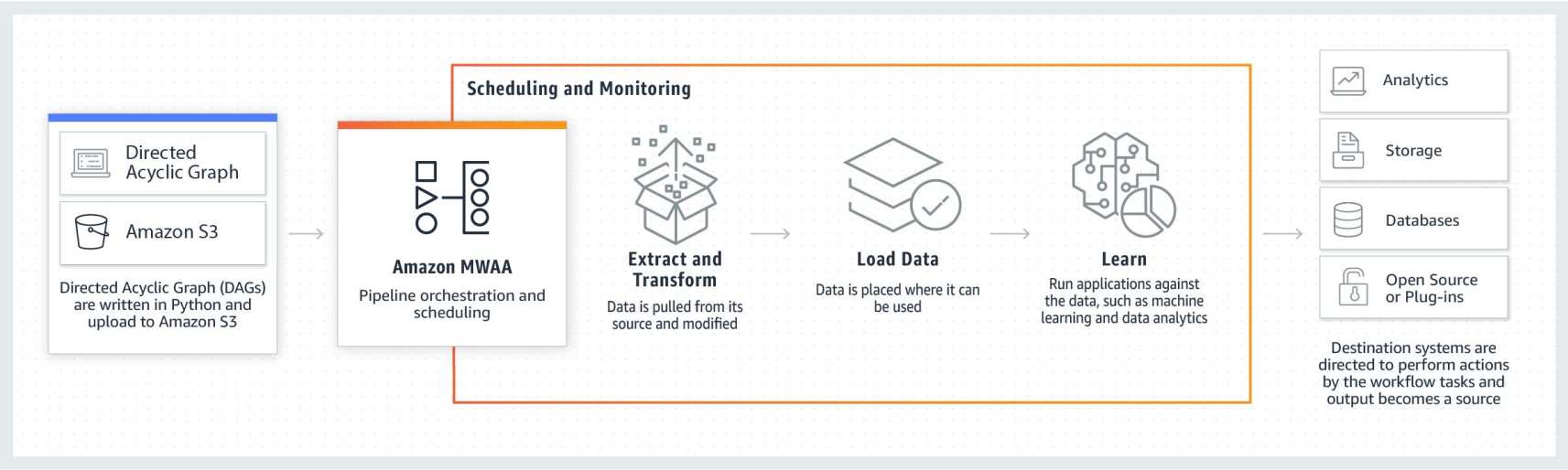
- Step Functions:



https://aws.amazon.com/pt/step-functions/use-cases/

Arquiteturas com Orquestração

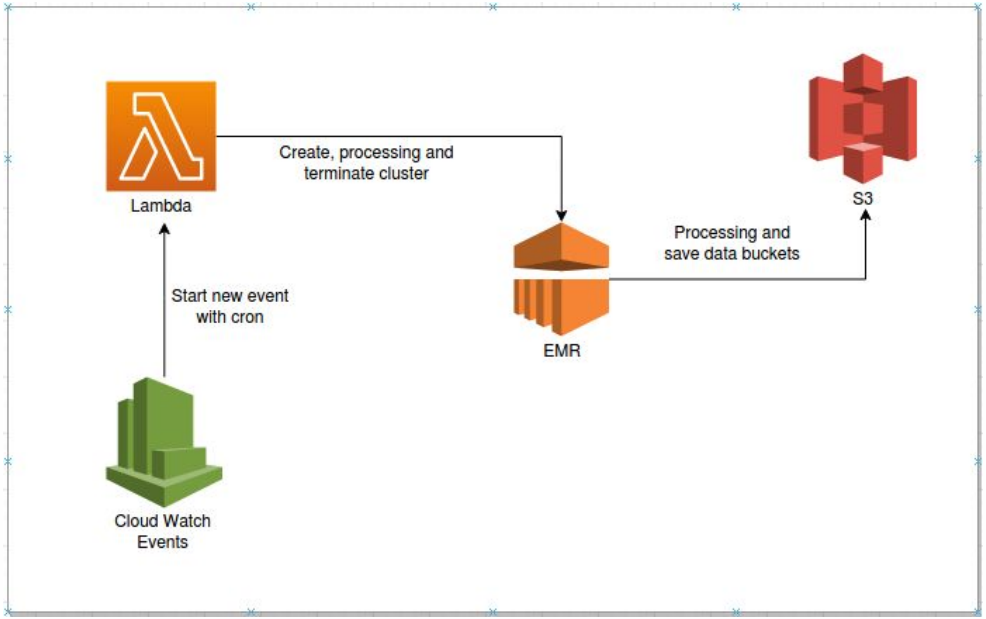
- Manage Workflows Apache Airflow:



https://aws.amazon.com/pt/managed-workflows-for-apache-airflow/

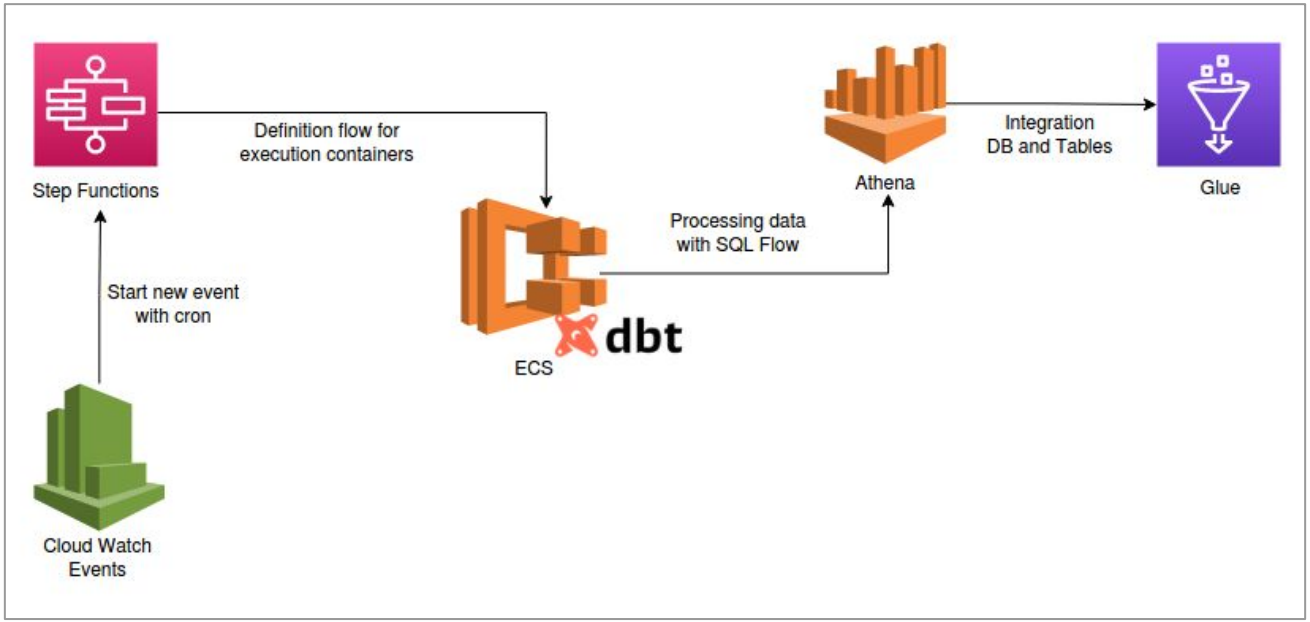
Arquiteturas com Orquestração

- Cloud Watch Events and Lambda:



Case Prático

- Deploy do DBT com ECS e Step Functions:



Atividades propostas

1. Produtizar uma pipeline com o PySpark;
2. Adicionar CI/CD no Projeto do DBT mostrado na Aula.

Obrigado!

datasprints