Grupo de Estudos de Pipelines de Dados

Objetivo e Estrutura dos Encontros

- **Objetivo:** Realizar encontros quinzenais (a cada 15 dias) para explorar e discutir pipelines de dados, focando em ferramentas, técnicas e boas práticas.
- **Duração:** Cada encontro terá entre 1 hora e 1 hora e meia.
- **Estrutura do Encontro:**
    - Apresentação inicial dos participantes.
    - Apresentação de um tema ou ferramenta técnica relacionada a pipelines de dados.
    - Demonstração prática do uso da ferramenta ou conceito apresentado.
    - Bate-papo interativo para tirar dúvidas e discutir o tema.
    - Construção coletiva de exercícios práticos sobre pipelines de dados.

Atividades práticas do Grupo de Estudo

Esse documento explica os fluxos que iremos criar nesse grupo de estudo .

FluxoFormulario de Cadastra GU.pdf

Apresentação do Gu Bigdata IA

Temos aqui uma apresentação realizada no grupo de usuário que fala da estruturação de um datalake e como estrutura pipeline de dados num ambiente agnóstico usando hadoop. Assista esse vídeo e venha participar do nosso grupo de estudos

Construindo-Datalakes-Agnosticos-de-Baixo-Custo-com-Hadoop-e-Outras-Tecnologia…

Colocar o aqui os slides e o vídeo do YouTube e colocar os slides em PDF para a pessoa navegar deve estar aberto

O que é uma PipeLine de Dados ?

Uma **pipeline de dados** de um jeito simples e fácil de entender, imagine que você está preparando uma refeição, como um bolo ou um jantar. Vamos pensar no processo passo a passo:

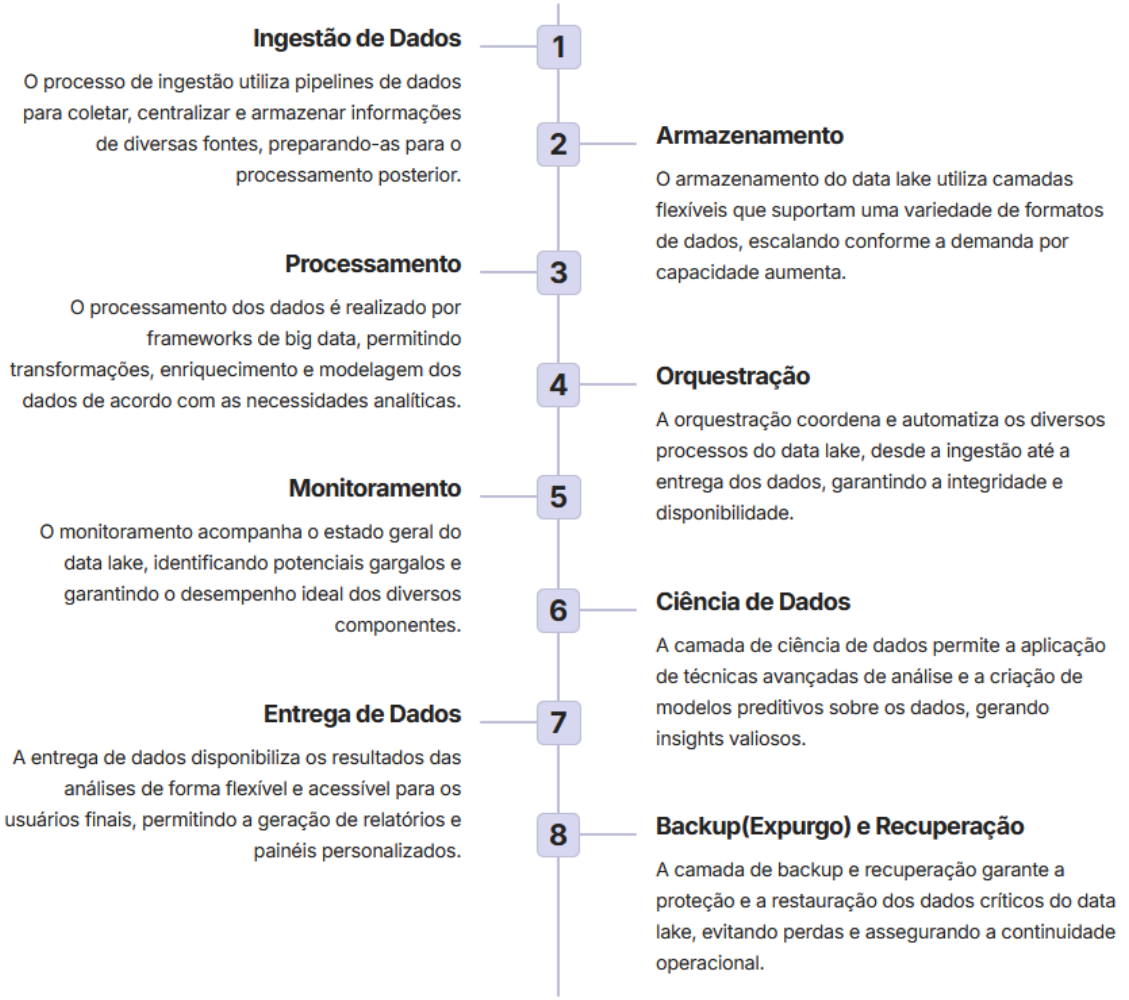
1. **Coletar os ingredientes**: Você vai ao mercado e pega tudo o que precisa, como farinha, ovos e açúcar. Isso é como os **dados brutos** — as informações que vêm de algum lugar, como números, textos ou registros de um site, por exemplo.
2. **Limpar e preparar**: Antes de usar, você lava os vegetais ou separa o que não precisa, como cascas. Nos dados, isso é parecido com **limpar e organizar** as informações, tirando erros ou coisas que não servem.
3. **Cozinhar ou processar**: Você mistura os ingredientes, coloca no forno e transforma tudo em algo gostoso. Na pipeline, os dados são **processados** — ou seja, calculados, organizados ou transformados para ficarem mais úteis.
4. **Servir a refeição**: No final, você coloca o prato na mesa para as pessoas comerem. Nos dados, isso é como **apresentar as informações** de um jeito claro, como um gráfico ou relatório, para que alguém possa usá-las para tomar decisões.

Então, uma **pipeline de dados** é como uma linha de produção ou uma receita: ela pega informações bagunçadas e cruas (como os ingredientes no mercado), organiza, limpa e transforma tudo isso passo a passo, até que esteja pronto para ser usado de um jeito prático e compreensível.

Por exemplo, pense em um site que quer saber quantas pessoas o visitam por dia. A pipeline coleta os dados de quem entrou, limpa qualquer erro, calcula o total de visitas e, no final, mostra esses números em uma tabela simples. É esse processo completo que chamamos de pipeline de dados.

Camadas de uma Pipeline de Dados

Abaixo, explicamos resumidamente cada camada de uma pipeline de dados e listamos ferramentas open-source e comerciais associadas a elas.



Camada	Descrição	Ferramentas Open-Source	Ferramentas Comerciais
Ingestão (Gestão)	Coleta e importa dados de diversas fontes.	Apache NiFi, Logstash, Fluentd	AWS Glue, Google Cloud Dataflow, Azure Data Factory
Armazenamento	Armazena dados em formatos estruturados ou não estruturados.	Apache Hadoop HDFS, Apache Cassandra, MongoDB	Amazon S3, Google Cloud Storage, Azure Blob Storage
Processamento	Transforma, limpa e agrega dados para análise.	Apache Spark, Apache Flink, Apache Beam	AWS EMR, Google Cloud Dataproc, Azure HDInsight

Machine Learning / Inteligência Artificial	Aplica algoritmos para análise avançada e previsões.	TensorFlow, PyTorch, Scikit-learn	AWS SageMaker, Google AI Platform, Azure ML
Entrega	Apresenta dados processados ou insights de forma acessível.	Apache Superset, Metabase, Grafana	Tableau, Power BI, Looker
Orquestração	Coordena e gerencia a execução das etapas da pipeline.	Apache Airflow, Luigi, Prefect	AWS Step Functions, Google Cloud Composer, Azure Logic Apps
Monitoramento e Logging	Acompanha performance e registra eventos para análise.	Prometheus, ELK Stack	Datadog, New Relic, Splunk

---

Agenda de Encontros - Terça-feira – 20h

Próximos Encontros:

1. **25/03 – Introdução a Pipelines de Dados:** Conceitos básicos e arquiteturas comuns.

- **08/04 – Ferramentas de Ingestão:** Usando Apache NiFi.
- **22/04 – Armazenamento de Dados:** Quando optar por SQL ou NoSQL?
- **06/05 – Batch vs. Streaming:** Diferenças, casos de uso com SnowFlake.
- **20/05 – Orquestração com Airflow:** Criando e gerenciando workflows
- **03/06 – Entrega de dado:** Conheça o NOSQL ClickHouse.
- **17/06 – Visualização de Dados:** Boas práticas e ferramentas como Power BI
- **01/17 – Terça-feira – 20h**

Participe do Nosso Grupo!

Venha fazer parte do **Grupo de Estudos de Pipelines de Dados**! Nossos encontros quinzenais são uma oportunidade única para aprender, trocar experiências e construir pipelines de dados na prática, junto com uma comunidade apaixonada por tecnologia. Seja você iniciante ou experiente, há espaço para todos.

Inscreva-se agora para receber informações sobre os próximos encontros e acesso às gravações das apresentações. Vamos juntos explorar o universo dos dados e transformar conhecimento em ação!