**1. Definição do Projeto**

* Objetivo: Automatizar o download, processamento e armazenamento de dados em arquivos CSV.
* Tecnologias: Python, Pandas, MySQL, CSV, Agendamento de Tarefas (Apache Airflow).

**2. Estrutura do Projeto**

projeto\_auto\_extrator/  
│── data/ → Armazena os CSVs gerados  
│── logs/ → Armazena logs de execução  
│── scripts/ → Código-fonte do projeto   
│── tests/ → Testes unitários e de integração  
│── .env → Credenciais e configurações (usar variáveis de ambiente)  
│── main.py → Script principal  
│── requirements.txt → Dependências do projeto

**3. Desenvolvimento**

**3.1 Conexão com o Banco de Dados**

* Criar função para conectar ao MySQL.
  + ~~Executar query para extrair dados já existentes ou realizar atualizações de acordo com a lógica de negócios.~~

**3.2 Download dos Dados do Kaggle**

* Acessar o Kaggle pelo navegador.
* Baixar o dataset manualmente.
* Armazenar o arquivo na pasta do projeto (dados/).

**3.3 Manipulação dos Dados**

* Carregar os dados no Pandas.
* Aplicar transformações necessárias conforme os requisitos do projeto (renomeação de colunas etc.).
  + ~~Aplicar transformações necessárias conforme os requisitos do projeto (limpeza de dados, renomeação de colunas, conversão de tipos, etc.).~~
  + ~~Validação: Verificar schema, tratar dados nulos e duplicados.~~

**3.4 Armazenamento no Banco de Dados**

* Criar ou atualizar tabela no MySQL com os dados extraídos do Kaggle.
* Inserção ou Atualização: Decidir se deve inserir dados novos ou atualizar dados existentes com base na lógica de negócios.
  + ~~Validação de Schema: Verificar se os dados estão no formato correto antes de realizar a inserção no banco.~~

**3.5 Download e Novo Processamento**

* Baixar os dados do MySQL após a inserção.
* Aplicar novas transformações conforme necessário.

**3.6 Exportação Final**

* Salvar os dados processados como CSV na pasta dados/ com timestamp.
* Garantir que o CSV esteja livre de inconsistências (duplicatas e nulos).

**3.7 Automação**

* Criar script de agendamento para execução automática do processo:
  + Apache Airflow
    - ~~Windows: Task Scheduler~~
    - ~~Linux: cron job~~
* Monitoramento: Registrar logs detalhados e configurar alertas para eventuais falhas.

**3.8 Logging e Monitoramento**

* Implementar logs estruturados usando o módulo logging.
* Configurar notificações automáticas em caso de falhas.

**5. Segurança e Configurações**

* Uso de .env para credenciais (MySQL).
  + ~~Uso de .env para credenciais (API Key do Kaggle e MySQL).~~
* Segurança de arquivos de configuração: Proteger arquivos .env para que não sejam expostos a ambientes não autorizados.
* Tratamento de Erros: Implementar tratamento robusto com try/except para falhas de conexão com MySQL ou durante manipulação de dados.
  + ~~Tratamento de Erros: Implementar tratamento robusto com try/except para falhas de conexão com Kaggle, MySQL ou durante manipulação de dados.~~