**1. Extração do Dados**

Vamos usar o processo em BATCH através do ETL ( Extract Transform e Load) / ELT (Extract Load e Transform) a depender da regra de negócio. Por serem dados que não sofre muitas variações no decorrer do tempo esse processo seria no meu ver o mais adequado.  
Para a extração, podemos usar um script python simples com algumas bibliotecas combinadas como: Requests, Pandas, ZipFile , etc.  
**2. Camada Bronze / Camada Raw.**  
Nessa camada, os dados são armazenados conforme extraídos da fonte. Assim temos como fazer auditoria e rastrear sua origem ou quaisquer desvio.

Como exemplo, trago o formato .zip e .csv , são formatos muito comuns em qualquer ambiente de dados.

Passaremos os arquivos da Landing Zone para a Camada Bronze ainda usando o MinIO . Utilizaremos python para essa tarefa. Como outras opções cito o Airbytes , apache Hop , Apache NIFI.

**3. Camada Silver / Dados tratados**

Aqui serão higienizados, padronizados, removido duplicidades enriquecimentos com outras fontes.Pandas para ler os dados, aplicar as transformações necessárias. Poderia fazer diretamente no Pyspark, mas acredito que pelo volume de informações o Pandas nos atende perfeitamente.

Ainda na pegada do MinIO , com o python. Como Opção podemos usar o Apache NIFI ou DBT .

**4. Camada Gold / Dados Analíticos**

Nessa última camada, vamos sumarizar as informações, modelar e otimizar para uma análise eficiente. Aqui teremos tabelas com métricas de negócio aplicadas e KPIs.Aqui podemos usar o MinIO ainda como Datalake para armazenar na GOLD\_ZONE os arquivos em formato Parquet prontos para serem consumidos.

**5. Orquestração do Processo**

Nesse passo vamos implementar uma ferramenta que gerencie, agende e monitore todo nosso pipeline, desde a extração até a disponibilização para o usuário. Pensando em algo mais robusto, usaremos o AIRFLOW para essa tarefa.

**6.Disponibilização para Consumo.**

Os dados da camada Gold são disponibilizados para diversos públicos, não só para os mais avançados como Analista de Dados, Cientistas. Más também para públicos como o Analista de Negócio, Gerentes ou até mesmo aplicações consumindo na camada Gold.

Como formas de acesso podemos ter :  
Power BI , Looker , Tableau , Metabase , APIs : FastApi.