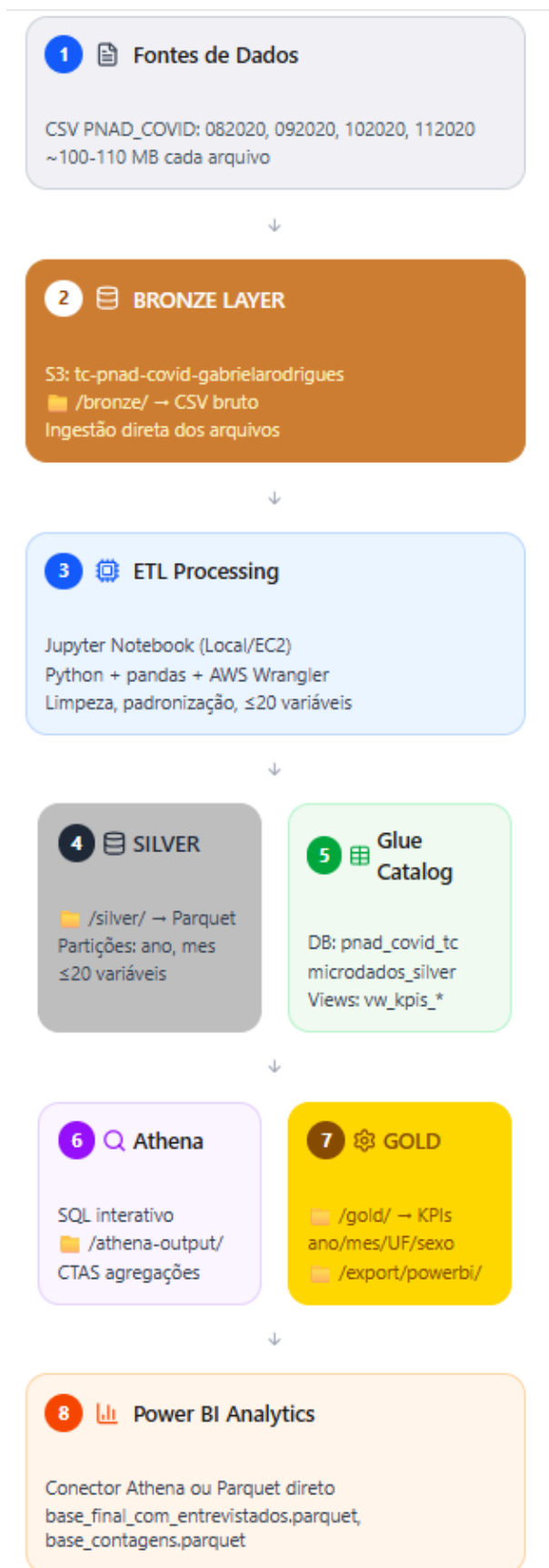











Diagrama da Arquitetura Data Lake

Fluxo numerado com camadas Bronze, Silver e Gold



Tecnologias & Responsabilidades

Mapeamento completo dos serviços e suas funções

	Camada	Serviço	Função
	Ingestão	S3 (Console/CLI)	Armazenar CSV no Bronze
	Processamento	Jupyter + pandas/AWS Wrangler	Limpeza/feature engineering, escrita Silver (Parquet)
	Catálogo	AWS Glue Data Catalog	Esquemas, partições, database pnad_covid_tc
	Consulta	Amazon Athena	SQL, views (vw_kpis_*), CTAS, saída em /athena-output/
	Camada Gold	Athena/ETL	KPIs agregados por UF/sexo/tempo
	BI	Power BI	Conector Athena ou ingestão de Parquet de /export/powerbi/
	Segurança	IAM, SSE-KMS, Bucket Key, Block Public Access	Acesso mínimo, criptografia e governança
	Observabilidade	CloudTrail/CloudWatch	Auditoria de chamadas e métricas
	Custos	S3 Storage + Athena Scan	Otimizar com Parquet, partições e filtros de coluna

Decisões de Design:

- **Parquet + partições** para reduzir custo Athena
- **≤20 variáveis no Silver** para performance
- **Export BI em Parquet** para consumo rápido
- **Lifecycle rules** limpam /athena-output/

\$ Otimização de Custos

- **Athena:** Pago por TB → Parquet reduz ~80% vs CSV
- **S3:** Lifecycle para IA/Glacier em dados antigos
- **Partições:** Filtram apenas meses necessários