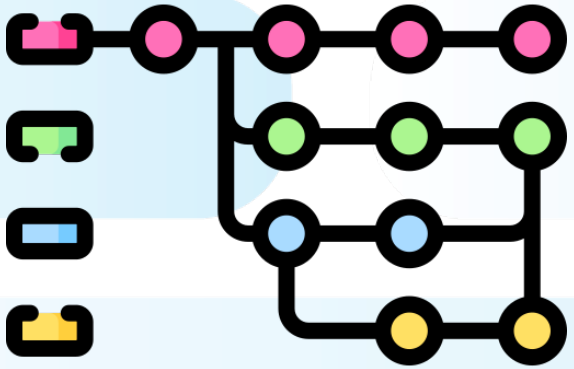
Versionamiento de datos

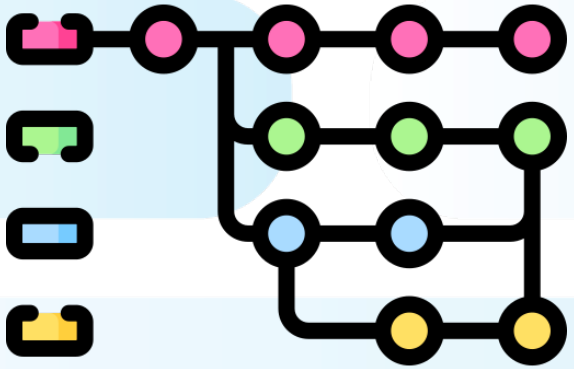
Versionamiento de código

- Podemos usar Git para versionar código
- Usar Git no es un buen lugar para almacenar datasets
- ¿Qué alternativas existen?



Versionamiento de datos

- Es un área relativamente nueva
- Es difícil porque se habla de enormes cantidades de datos



¿Una propuesta?

2023-01-01.csv

2023-01-02.csv

2023-01-03.csv

2023-01-04.csv

2023-01-05.csv

2023-01-06.csv

2023-01-07.csv

2023-01-08.csv

2023-01-09.csv

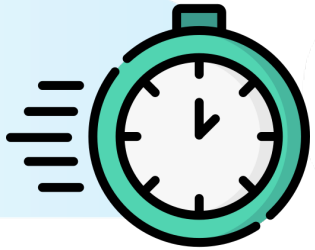
2023-01-10.csv

2023-01-11.csv

2023-01-12.csv

¿Por qué versionar los datasets?

- Velocidad
- Reproducibilidad
- Gobernanza



Duplicación total

2023-01-01.csv

2023-01-02.csv

2023-01-03.csv

2023-01-04.csv

2023-01-05.csv

2023-01-06.csv

2023-01-07.csv

2023-01-08.csv

2023-01-09.csv

2023-01-10.csv

2023-01-11.csv

2023-01-12.csv

Metadatos

f	price	valid_from
	65.3	2023-01-01
	34.2	2023-01-01
	6.1	2023-01-01
	16.7	2023-01-02
	6.1	2023-01-02

Herramientas dedicadas

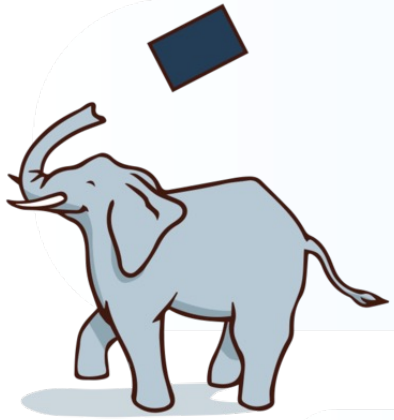
- Minimizando la cantidad de espacio que ocupa
- Permite interactuar fácilmente con diferentes versiones
- Funciona a cualquier escala

Herramientas dedicadas

DELTA LAKE

lakeFS

DVC



Pachyderm