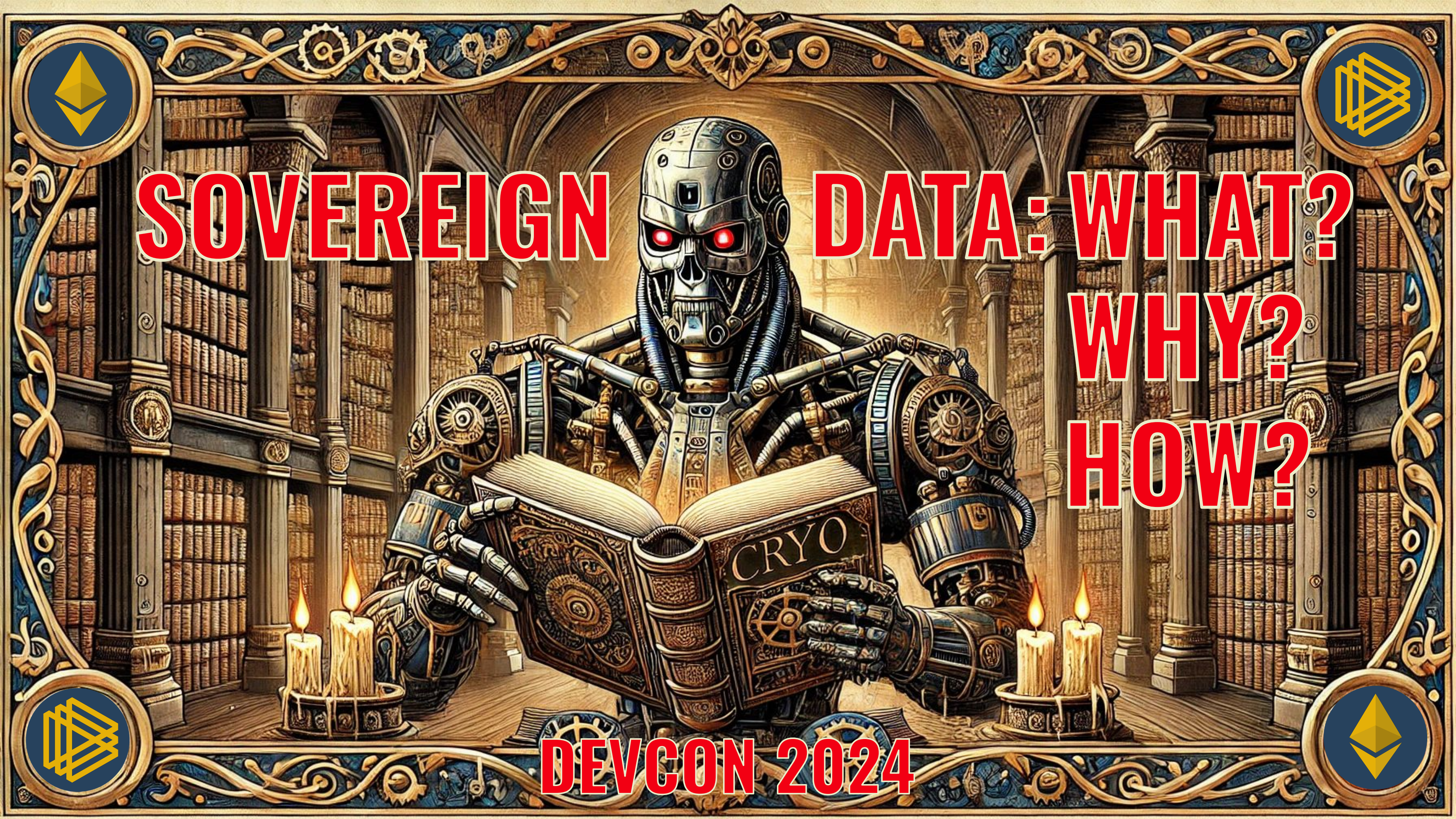
SOVEREIGN DATA: WHAT? WHY? HOW?



DEVCON 2024

To analyze Ethereum historical data, what do you need?

hardware? software? expertise?

You don’t need anything proprietary

You don’t need a team

You don’t need a database

All you need is a laptop

DATA SOVEREIGNTY

WHAT
1

WHY
2

HOW
3

Data Sovereignty = Transparency + Control

The data is **yours**

The format uses **open standards**

The code is **open source**

The pipeline is **locally runnable**

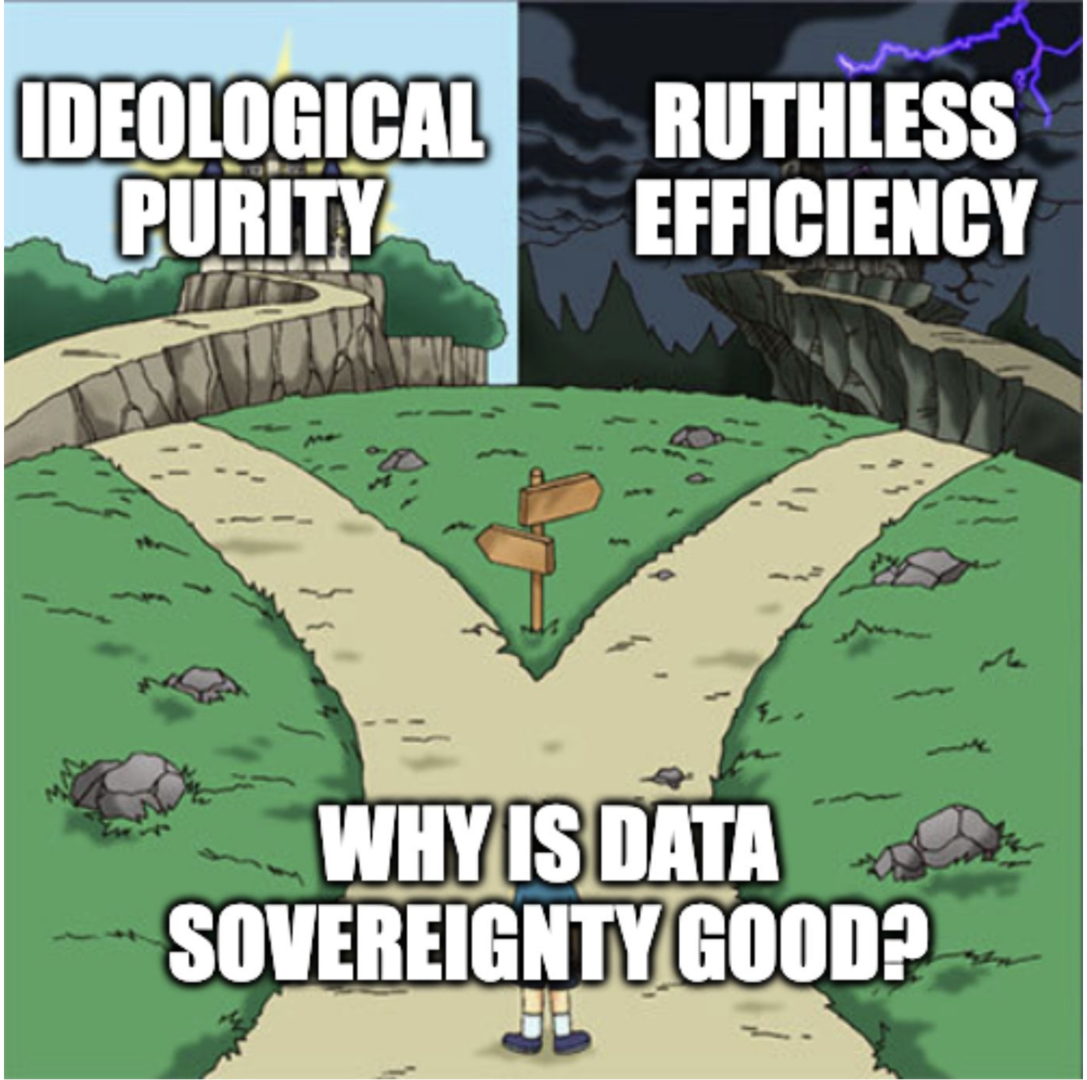
The results are **reproducible**

"Closed source is a non-starter"

"Information wants to be free"

"I shouldn't need to trust anyone"

"Vendor lock-in is bad"



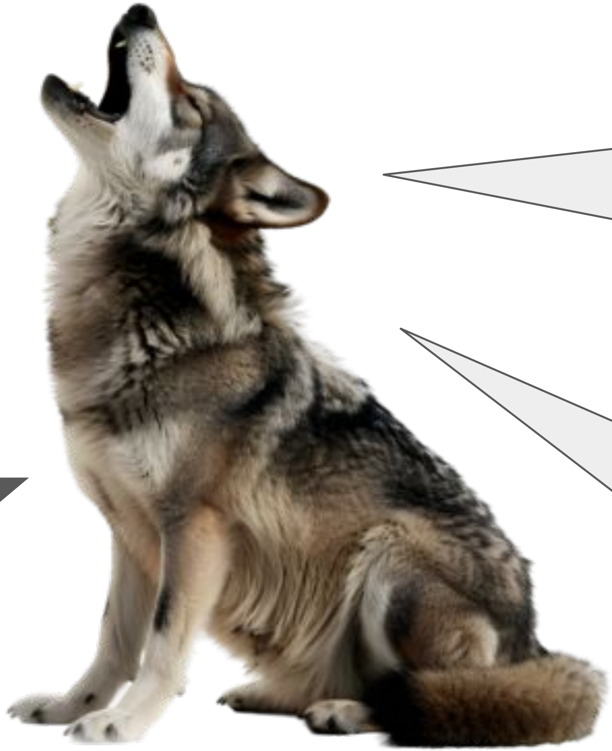
"A rich ecosystem of open source tools evolves faster and better"

"This data can be trusted for billion dollar decisions"

"I get to run 'bigquery' on my laptop"

Data sovereignty amplifies the lone wolf data workflow

operational simplicity enables an individual to do a team’s work



minimizing friction allows exploring data more quickly and deeply

local-first enables a rich ecosystem of OSS tools

most crypto data ppl

Data sovereignty makes EIP’s better

How to achieve sovereignty?

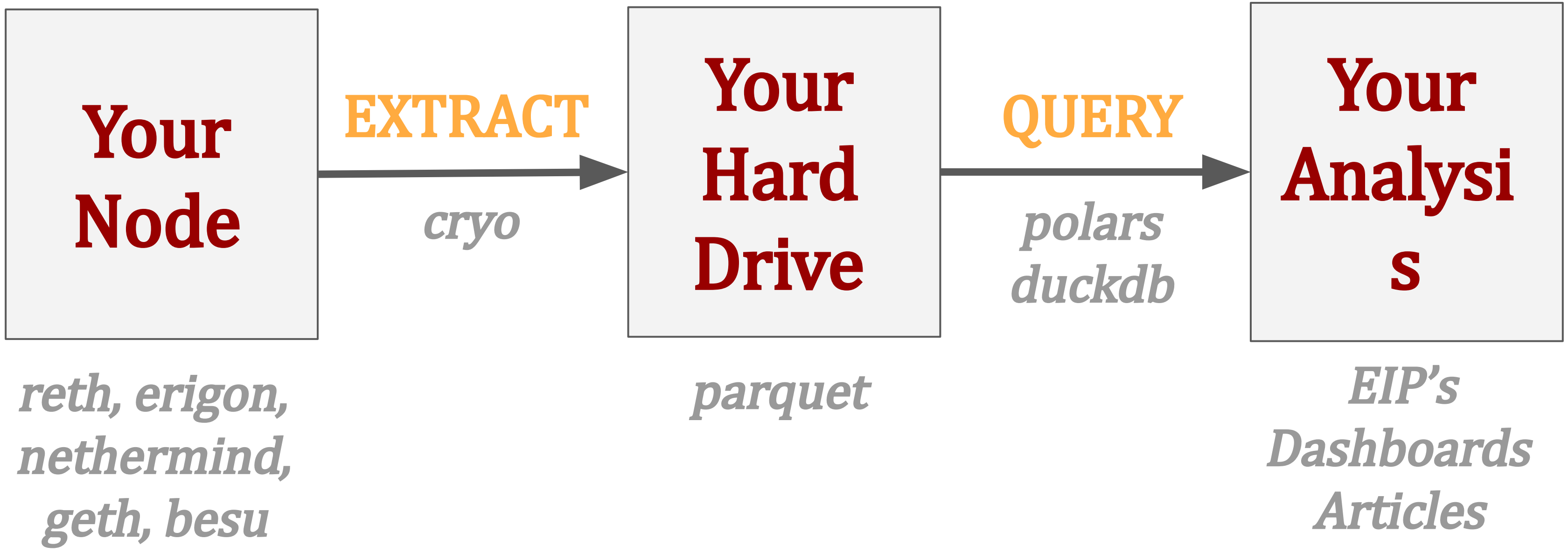
Modern Data Engineering

Advances in tooling and architectures

Advances in open standards

Advances in efficiency

Sovereign Data Workflow



Your Node

reth, erigon,
nethermind,
geth, besu

EXTRACT

cryo

Your
Hard
Drive

parquet

QUERY

polars
duckdb

Your
Analysi
s

EIP’s
Dashboards
Articles

every step of this process can run on your laptop or in the cloud or wherever you want

cryo

is a tool for collecting EVM datasets

36 total cryo datasets available

```
cryo DATASET_NAME --blocks START:END
```

(cli syntax)

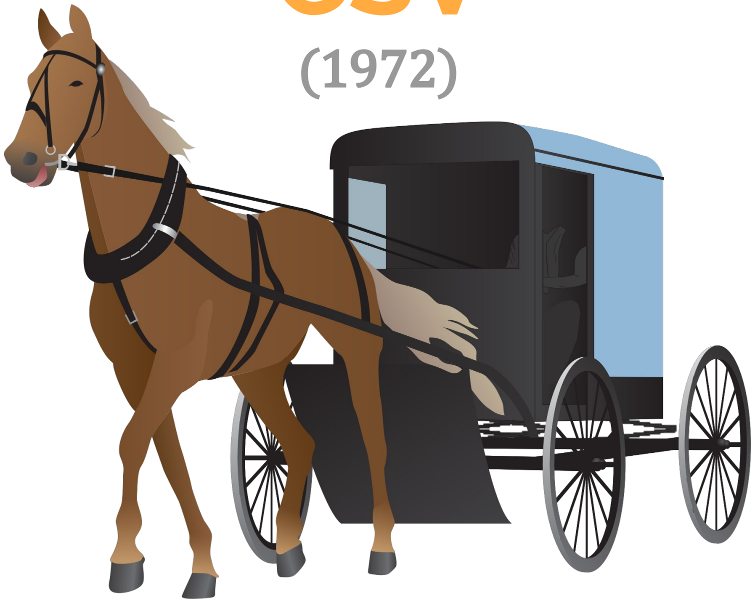
```
df = cryo.collect('DATASET_NAME', ...)
```

(python syntax)

[Demo: extract data using cryo]

CSV

(1972)



☑ Human-readable

☑ Legacy Ecosystem

vs

Parquet

(2013)

☑ Compression

☑ Indices

☑ Queries & Subsets

☑ Modern Ecosystem

Parquet datasets by the numbers

Size of various mainnet datasets extracted using cryo

36 total cryo datasets available

Blocks
929 MB

TXs
539 GiB

Logs
170 GiB

Contracts
15 GiB

ERC20 Transfers
97 GiB

ERC721 Transfers
12 GiB

Call Traces
756 GiB

State Diffs
348 GiB

varies up or down depending on: what data fields you want + partition size + compression scheme

[Demos: Querying and Processing Files]

[Demos: Querying and Processing 0]

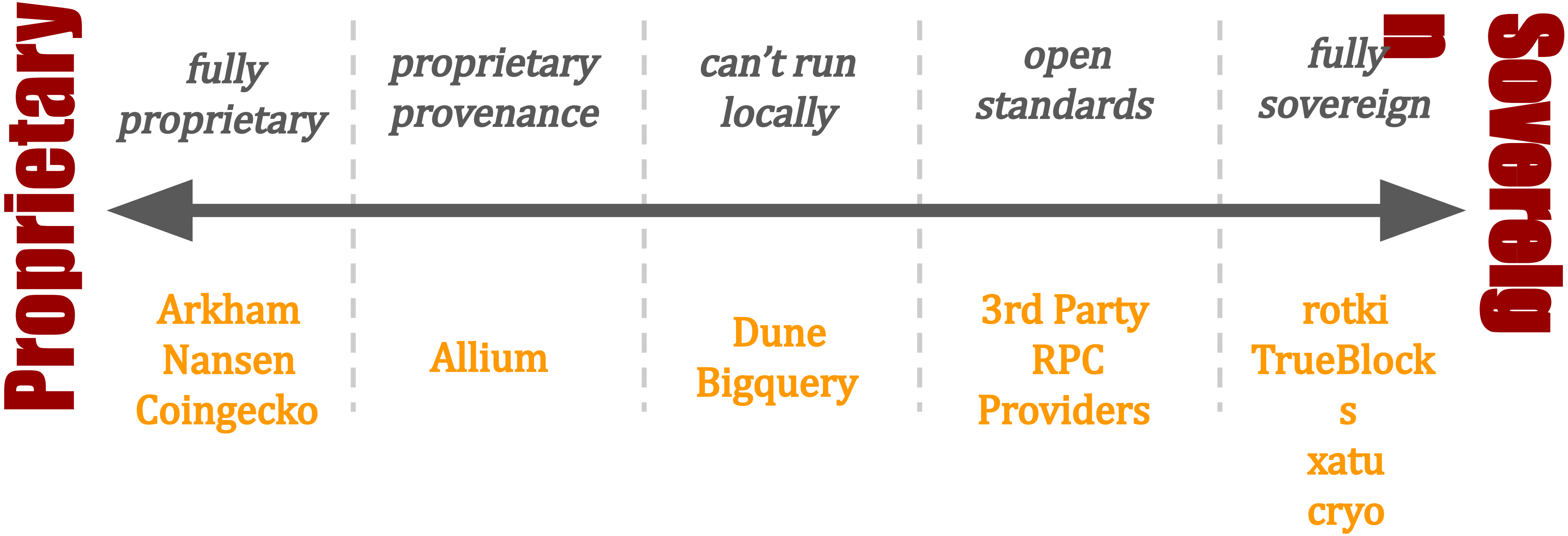
[Demos: Querying and Processing 1]

[Demos: Querying and Processing 2]

[Demos: Querying and Processing 3]

Data Sovereignty is a spectrum

all of these tools are useful!



Proprietary

fully proprietary

Arkham
Nansen
Coingecko

proprietary provenance

Allium

can't run locally

Dune
Bigquery

open standards

3rd Party
RPC
Providers

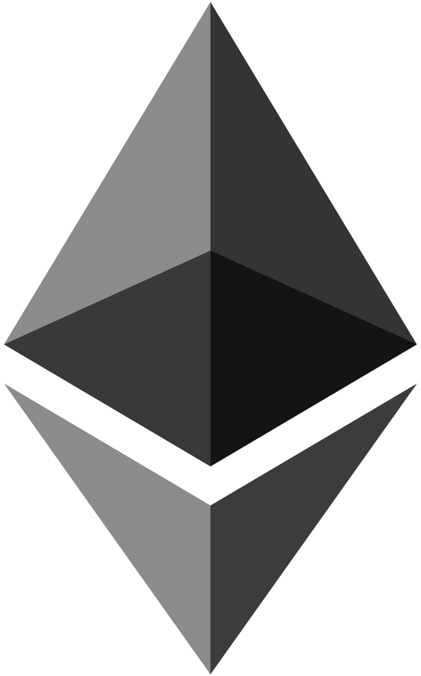
fully sovereign

rotki
TrueBlock
s
xatu
cryo

Sovereig
n

Data Sovereignty is OPOE

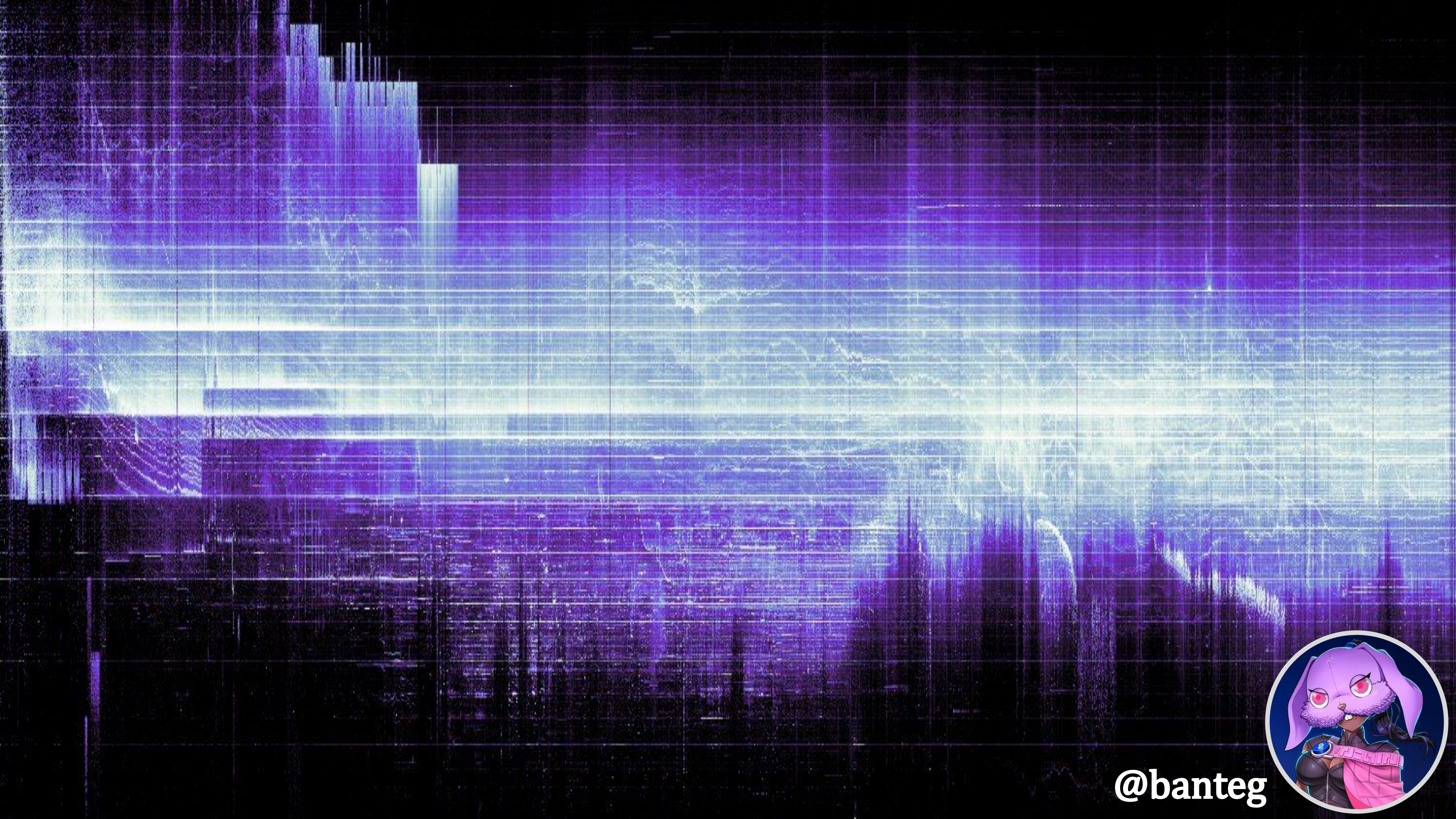
Only Possible On Ethereum



- ✅ introspection is prioritized ❌
- ✅ tooling is mature ❌
- ✅ scale is tractable ❌
- ✅ sovereignty ≈ decentralization ❌

THAT'S IT

I'm @notnotstorm
on



@banteg

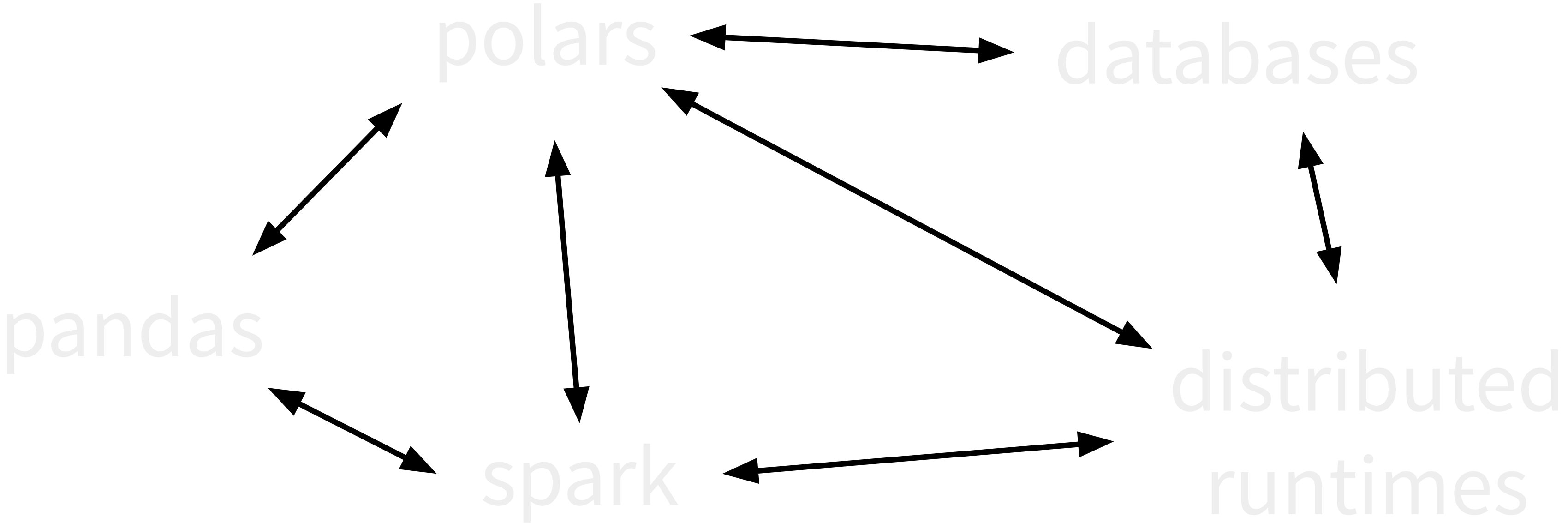
TODOs

Modern Data Engineering Trends

1. Standardized IPC
2. Modern Storage Formats
3. Separate storage vs compute

Data eng in crypto is a decade behind web2

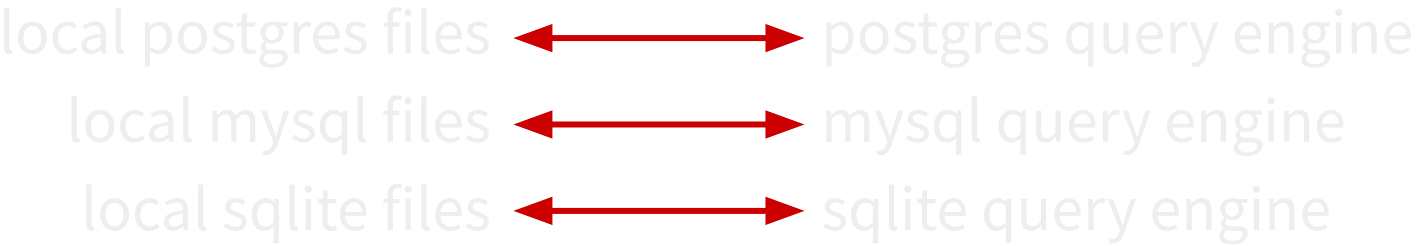
Arrow IPC: zero-copy data sharing



NO SERDE!

Storage vs Compute

the old way



local postgres files

local mysql files

local sqlite files

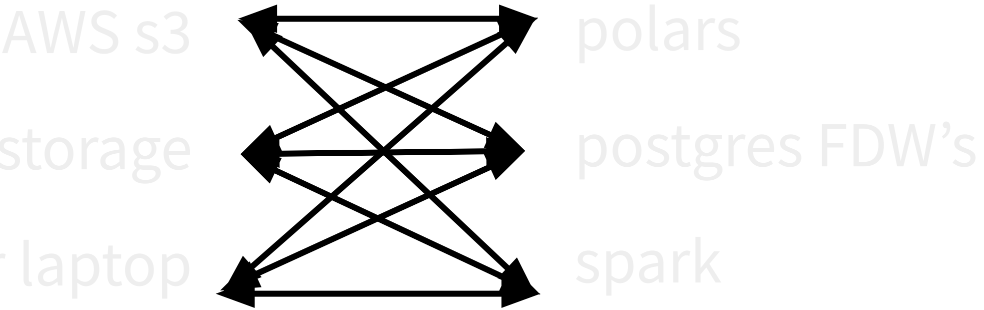
postgres query engine

mysql query engine

sqlite query engine

every engine locked to specific backend + format

NOW



AWS s3

GCP storage

your laptop

polars

postgres FDW's

spark

use any of these... **...with any of these**