How I built a high-performance Cosmos indexer



Fabien Penso

DEVMOS 2024

@fabienpenso
https://pen.so

01

Time Spent writing indexers

2 years full-time, 3500h

02

Project size

77,000 lines of Rust

03

Performance

Indexing time from genesis:

Kujira in 1h, Osmosis in 6h, Stargaze in 3h

0	4
	_

Usage

Stargaze: > 15M requests per day

05

Indexed chains

Stargaze, Osmosis, Neutron, Noble, dYdX, Kujira, ...

06

Infrastructure

Stargaze: 4 redundant servers

Kujira, Osmosis, Neutron, Noble, dYdX: 1 single server

Thank you @fabienpenso for Constellations @IBCMuffins

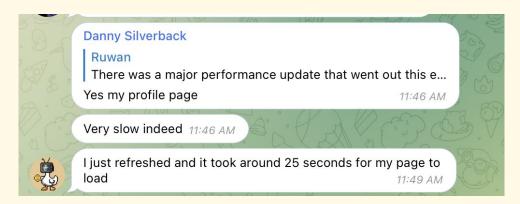
Constellations indexer [...] singlehandedly sped up development [...] threefold. @josefleventon

... and many using the public API

Stargaze: Benefits from no indexer to using an indexer?

No Indexer

- Front-end fetch lots of data directly from the chain
- Some pages took > 30 seconds. Stargaze launchpad failed when too much content was loaded. Not sustainable.



With Indexer

Page loaded < 500ms

Every chain should have an indexer

Users expect web3 products with web2 performance

Blockchain launch

Smart Contracts

A lot of blockchain launches are all and only about smart contracts, but that's a tiny 10% of the whole shebang. The remaining 90% comes from the frontend, the backend, and the infrastructure.

The significance of these components is often overlooked, yet they're crucial for developing a high-quality product. It requires approximately two years to reach the level of Stargaze, accounting for all these elements.

Infrastructure

Indexer

Frontend

Prepare for Chaos

How to build an indexer?

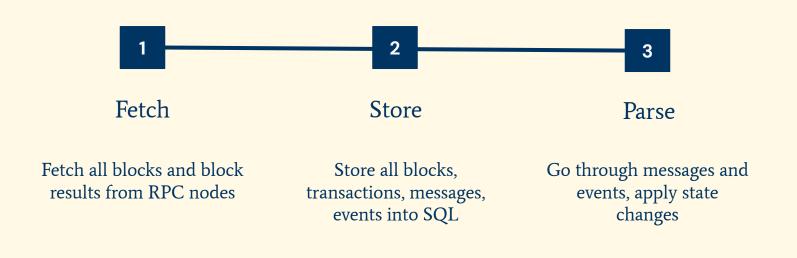
Rust

- ✓ Memory safety (multi-thread)
- ✓ Most loved language 8 years in a row (Stack Overflow)
- ✓ Performance
- ✓ Functional programming

But that's not enough, lots of time spent tweaking performance, making things faster, adding caching layers and instrumentation

Make it work Make it robust Make it fast

Naive way to write indexers



Configuration

```
⊙ 1 example.toml
config > chains > m example.toml
 14 [[chains.osmosis_mainnet.nodes]]
 13 rpc_endpoint = "https://r-osmosis--NOTIONAL_TOKEN.gw.notionalapi.net"
 12 rest_endpoint = "https://a-osmosis--NOTIONAL_TOKEN.gw.notionalapi.net"
 11 websocket_endpoint = "wss://r-osmosis--NOTIONAL_TOKEN.gw.notionalapi.net/websocket"
  10 current = true
  8 [[chains.osmosis_mainnet.nodes]]
  7 rpc_endpoint = "https://r-osmosis-archive-sub1--NOTIONAL_TOKEN.gw.notionalapi.net"
  6 rest_endpoint = "https://a-osmosis-archive-sub1--NOTIONAL_TOKEN.gw.notionalapi.net"
  4 [[chains.osmosis_mainnet.nodes]]
   3 rpc_endpoint = "https://r-osmosis-archive-sub2--NOTIONAL_TOKEN.gw.notionalapi.net"
  2 rest_endpoint = "https://a-osmosis-archive-sub2--NOTIONAL_TOKEN.qw.notionalapi.net"
    [[chains.osmosis_mainnet.nodes]]
  1 rpc_endpoint = "https://r-osmosis-archive-sub3--NOTIONAL_TOKEN.gw.notionalapi.net"
  2 rest_endpoint = "https://a-osmosis-archive-sub3--NOTIONAL_TOKEN.gw.notionalapi.net"
  / main 1[+]
                                                                                                          88%/17 < studio
```





Using RPC nodes from **Rhino** (https://rhinostake.com/), **Notional** (github.com/notional-labs) or provided by the chain. Managing nodes can be a full-time work... Waiting for weeks for an archive node from a chain.

Async with Tokio

```
⊙ 1 block_fetcher_example.rs
crates > block_fetcher > src > @ block_fetcher_example.rs > @ fetch_blocks
       ync fn fetch_blocks<I>(ctx: Arc<AppContext>, block_range: I) -> Result<(), crate::Error>
                                                                                                    ■ file not included i
        I: DoubleEndedIterator<Item = u32>,
        let mut set = JoinSet::new();
        for block_height in block_range {
            // Semaphore to limit how many blocks we fetch at the same time
             let permit = ctx
                 .semaphores
                 .fetcher_limited_tasks
                 .clone()
                 .acquire_owned()
            // To avoid having too many tasks in memory
             while set.len() > ctx.config.parser.concurrency {
                set.join_next().await;
            set.spawn(async move {
                 let _permit = permit;
                 fetch_block(block_height, true).await?;
                Ok::<_, anyhow::Error>(())
        // Wait for all tasks to finish
        while (set.join_next().await).is_some() {}
        Ok(())
```

Multi-thread with Rayon

```
⊙ 1 rayon.rs
src > bin > @ rayon.rs > \text{\text{$\text{$\text{$}}} fetch_blocks}
   9 use anyhow::Result;
   8 use rayon::prelude::*;
    use std::sync::Arc;
   5 fn fetch_blocks<I>(ctx: Arc<AppContext>, block_range: I) -> Result<()>
   4 where
         I: DoubleEndedIterator<Item = u32>,
         let pool = rayon::ThreadPoolBuilder::new()
              .num_threads(ctx.config.parser.concurrency)
 10
              .build()
              .unwrap();
         pool.install(|| {
              block_range.par_iter().for_each(|&block_height| {
                  fetch_block(block_height, true).unwrap();
         0k(())
  11
```

Encountered issues

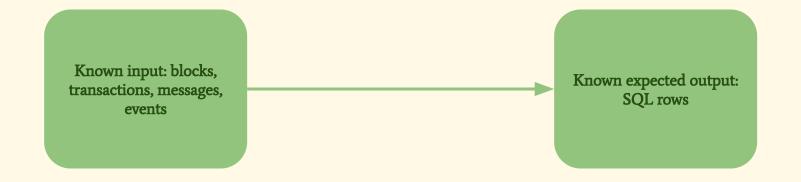
- ✓ Very slow archive nodes
- ✓ 502 timeouts, 429 rate limiting, nodes down for hours
- ✓ Secp256k1 public keys with invalid SEC1 tags are accepted (cosmos-sdk issue #20406) by the go SDK, refused by CosmRS
- ✓ Invalid txs_results returned for legacy ABCI responses (CometBFT issue #3002), preventing fetching some dYdX/Sei block results
- ✓ Blank validator keys

✓ Some Osmosis block results can be > 280MB, and node is failing

You're dealing with on-chain and off-chain data, you can't trust any of it. Malformed user submitted UTF8 strings, renamed smart contracts events, null bytes breaking Postgres, wrong smart contract address, invalid base64. It's the wild west.

Must build for errors and resilience.

Test Driven Development



Indexers are the perfect use-case for TDD

Test Driven Development

```
⊙ 1 live_actions.rs
tests > ® live_actions.rs > ♡ parse_settle_auctions
    #[test_context::test(stargaze_testnet)]
   1 async fn parse_settle_auctions() -> Result<(), anyhow::Error> {
         // Parse settle auction
         let block = 5533383;
         index_blocks(test_ctx.app_ctx.clone(), &vec![block]).await?;
         let events = entity::event::Entity::find()
             .settle_auctions()
             .all(test_ctx.db())
             .await?;
         assert_that!(events).has_length(1);
  11
  12
         Ok(())
    # main
                                                    13%/312 < studio
                                         < ® rust
```

Protobuf files

Cosmos-sdk chains are using protobuf for on-chain messages. But parsing historical messages isn't as easy as you'd think.

I had to dig in full git history to retrieve deleted protobuf files and fields, and merge all needed within my private repository.

Protobuf files

Buf Schema Registry		$\bigcirc cosmos.vesting.v1beta1.MsgCreateClawbackVestingAccount$	Ų	Sign in	Sign up	
Filter by type	No searcl	No results found				
Repositories Plugins Organizations Users						

- ✓ Missing/removed fields
- ✓ Missing files
- ✓ Linked to a buf.build project in buf.yaml, but not pushed and not available

Protobuf files

Do's

- Copy and save proto files into your repository
- Write your own protos to Rust structs into a specific crate (using prost, prost-build)
- Might need to search older deleted fields from proto files

Don'ts

- Don't link to existing repovia git submodule
- Don't rely on buf.build, or only to copy current existing files
- Don't think using existing proto files is fine, fields get deleted and replaced with reserved *later*

Storing in SQL

```
src > controllers > event_controller > @ store.rs > @ save_all_models_in_sql
   57 async fn save_all_models_in_sql(
         ctx: Arc<AppContext>,
          mut messages: Vec<entity::message::Model>,
          let split = 2000;
          let may_panic = async {
             let txn = ctx.db().begin().await?;
 1585
             while !blocks.is empty() {
                  let remaining = blocks.split off(std::cmp::min(blocks.len(), split));
                  let saving blocks = blocks
                      .into iter()
                      .map(|b| b.into())
                      .collect::<Vec<entity::block::ActiveModel>>();
                  if let Err(error) = entity::block::Entity::insert_many(saving blocks)
                      .on conflict(
                          sea_orm::sea_query::OnConflict::columns(vec![
                              entity::block::Column::BlockHeight,
                          .do nothing()
                          .to_owned(),
                      .exec_without_returning(&txn)
                      error!("Failed to save blocks: {:?}", error);
                      return Err(error);
                  blocks = remaining:
```

Decoding IBC packets



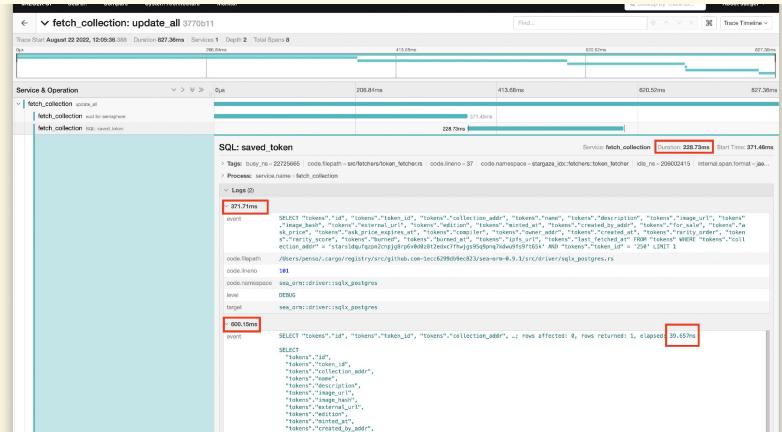
Processing stored txs

- Create NFT models (collections, nfts) and apply historical messages (owners, sales, ...)
- Create IBC related models (clients, connections, channels, denoms)
- Create Stargaze Names models
- Set invalid events, invalidated by later new events (bid invalidated by a sale)
- Create validator related models
- Fetch off-chain data

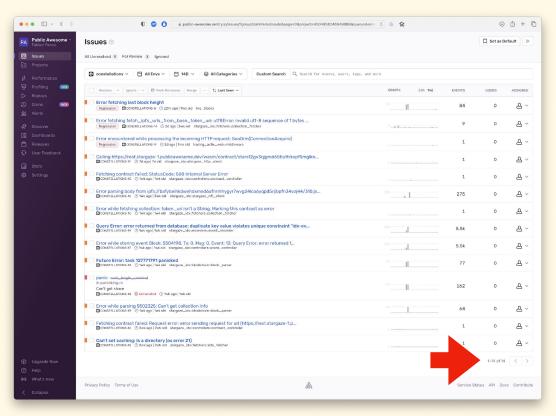
tracing + opentelemetry

```
⊙ 1 parse.rs
src > controllers > contract_controller > @ parse.rs > @ update_contracts_between_blocks
36 #[tracing::instrument(skip_all)]
⊙ ♭ main
```

Jaeger dashboard



Sentry dashboard



Indexing speed

Look at parsing performance based on data throughput, not block count. I had different speed for different chains and found out I had the same data throughput.

2024-06-01T15:18:13.904101Z		[1020001] [0.98s]	10002 blk at 1	0255/sec,	7196 txs at	7378/sec,	12008 msgs at 12312/s	c, 2311 events at 2370/sec, 189.34 MB at 194.13 MB/sec
2024-06-01T15:56:11.180524Z	INFO	[11230001] [1.69s]	9996 blk at	5903/sec.	6161 txs at	3638/sec.	14563 msgs at 8600/	ec, 15755 events at 9304/sec, 257.45 MB at 152.03 MB/sec
2024-06-01T15:56:12.856578Z		[11240001] [1.68s]			4213 txs at			ec, 12001 events at 7160/sec, 253.37 MB at 151.17 MB/sec
2024-06-01T15:56:14.378092Z		[11250001] [1.52s		6578/sec,	5667 txs at			ec, 15708 events at 10324/sec, 253.25 MB at 166.45 MB/sec
2024-06-01T15:56:15.947895Z	INFO	[11260001] [1.57s]	9993 blk at	6366/sec,	4202 txs at	2677/sec,	9078 msgs at 5783/	ec, 15646 events at 9967/sec, 254.49 MB at 162.12 MB/sec
2024-06-01T15:56:17.520986Z		[11270001] [1.57s]	10004 blk at	6359/sec,	5619 txs at	3572/sec,	10494 msgs at 6671/	ec, 16008 events at 10176/sec, 262.56 MB at 166.90 MB/sec
2024-06-01T15:56:19.442579Z		[11280001] [1.92s]	10010 blk at	5209/sec,	6385 txs at	3323/sec,	12701 msgs at 6610/	ec, 20018 events at 10417/sec, 262.16 MB at 136.43 MB/sec
2024-06-01T15:56:21.029600Z		[11290001] [1.59s]		6295/sec,	6908 txs at	4353/sec,	13252 msgs at 8350/	ec, 20048 events at 12632/sec, 270.84 MB at 170.66 MB/sec
2024-06-01T15:56:22.665028Z		[11300001] [1.64s]			5287 txs at			ec, 16047 events at 9812/sec, 264.56 MB at 161.77 MB/sec
2024-06-01T15:56:24.366842Z		[11310001] [1.70s]					15173 msgs at 8916/	
2024-06-01T15:56:26.290736Z		[11320001] [1.92s]					16231 msgs at 8437/	
2024-06-01T15:56:28.359491Z		[11330001] [2.07s]					20921 msgs at 10113/	
2024-06-01T15:56:30.212042Z		[11340001] [1.85s]					16807 msgs at 9072/	
2024-06-01T15:56:32.257848Z		[11350001] [2.05s]					18446 msgs at 9017/	
2024-06-01T15:56:34.441282Z		[11360001] [2.18s]					18973 msgs at 8690/	
2024-06-01T15:56:36.465269Z		[11370001] [2.02s]					19769 msgs at 9767/	
2024-06-01T15:56:38.746407Z		[11380001] [2.28s]					24519 msgs at 10749/	
2024-06-01T15:56:40.933164Z		[11390001] [2.19s]					18064 msgs at 8261/	
2024-06-01T15:56:43.176581Z		[11400001] [2.24s]						ec, 31211 events at 13912/sec, 291.63 MB at 130.00 MB/sec
2024-06-01T15:56:45.237934Z		[11410001] [2.06s]						ec, 23774 events at 11533/sec, 302.87 MB at 146.93 MB/sec
2024-06-01T15:56:47.073616Z 2024-06-01T15:56:49.243425Z		[11420001] [1.84s]					14399 msgs at 7844/	
2024-06-01715:56:49.2434252 2024-06-01715:56:51.388819Z		[11430001] [2.17s]					11971 msgs at 5517/s	
2024-06-01T15:56:51.3868192 2024-06-01T15:56:53.415389Z		[11440001] [2.15s] [11450001] [2.03s					14299 msgs at 6665/ 13089 msgs at 6459/	
2024-06-01115:56:53.4153892 2024-06-01T15:56:55.579500Z		[11460001] [2.035] [11460001] [2.16s]					12738 msgs at 5886/	
2024-06-01T15:56:57.777336Z		[11470001] [2.105] [11470001] [2.20s]					14147 msgs at 6437/	
2024-06-01T15:56:59.632248Z		[11480001] [1.85s]						ec, 16112 events at 8686/sec, 273.74 MB at 147.58 MB/sec
2024-06-01T15:57:01.518456Z		[11490001] [1.89s]					11803 msgs at 6258/	
2024-06-01T15:57:03.208182Z		[11500001] [1.69s]			5134 txs at			
2024-06-01T15:57:04.873601Z		[11510001] [1.67s]					14026 msgs at 8422/	
2024-06-01T15:57:06.502106Z		[11520001] [1.63s]					12315 msqs at 7562/	
2024-06-01T15:57:08.227562Z		[11530001] [1.73s						ec, 16349 events at 9475/sec, 275.66 MB at 159.76 MB/sec
2024-06-01T15:57:10.000242Z	INFO	[11540001] [1.77s]	10012 blk at	5648/sec,				ec, 19611 events at 11063/sec, 285.73 MB at 161.19 MB/sec
2024-06-01T15:57:11.996282Z		[11550001] [2.00s]	10039 blk at	5029/sec,			12454 msgs at 6239/	
2024-06-01T15:57:13.731602Z		[11560001] [1.74s]						ec, 20041 events at 11549/sec, 283.75 MB at 163.52 MB/sec
2024-06-01T15:57:15.512674Z		[11570001] [1.78s]	10026 blk at	5629/sec,	7918 txs at	4446/sec,	16382 msgs at 9198/	ec, 23934 events at 13438/sec, 285.88 MB at 160.51 MB/sec
2024-06-01T15:57:17.260093Z		[11580001] [1.75s]						ec, 24156 events at 13824/sec, 276.27 MB at 158.10 MB/sec
2024-06-01T15:57:19.187426Z	INFO	[11590001] [1.93s]	9994 blk at	5185/sec,	9234 txs at	4791/sec,	16815 msgs at 8725/	ec, 24448 events at 12685/sec, 278.31 MB at 144.40 MB/sec

CPU Usage while indexing

```
• • •
                                                                               Tasks: 355, 1230 thr; 16 running
                                                                  8.83G/31.0G
                                                                               Load average: 12.77 7.25 3.57
                                                                               Uptime: 111 days(!), 19:07:32
   PID USER
               PRI NI VIRT RES SHR S CPU%-MEM% TIME+ Command
 459242 penso
                20 0 3608M 2050M 43388 S 1507 3.2 20:32.11 target/release/parse_contracts
1459305 penso
                20 0 3608M 2050M 43388 R 97.5 3.2 1:17.09 target/release/parse_contracts
1459306 penso
                20 0 3608M 2050M 43388 R 96.8 3.2 1:17.10 target/release/parse contracts
1459313 penso
                    0 3608M 2050M 43388 R 96.2 3.2 1:17.06 target/release/parse_contracts
1459314 penso
                20 0 3608M 2050M 43388 R 96.2 3.2 1:17.00 target/release/parse_contracts
1459315 penso
                20 0 3608M 2050M 43388 R 96.2 3.2 1:17.03 target/release/parse contracts
1459307 penso
                20 0 3608M 2050M 43388 R 95.5 3.2 1:16.96 target/release/parse contracts
1459300 penso
                20 0 3608M 2050M 43388 R 94.9 3.2 1:16.93 target/release/parse contracts
1459302 penso
                   0 3608M 2050M 43388 R 94.9 3.2 1:16.82 target/release/parse contracts
1459308 penso
                20 0 3608M 2050M 43388 R 94.2 3.2 1:17.00 target/release/parse contracts
1459309 penso
                20 0 3608M 2050M 43388 R 94.2 3.2 1:17.03 target/release/parse_contracts
                20  0 3608M 2050M 43388 R 94.2 3.2 1:17.10 target/release/parse_contracts
1459310 penso
1459301 penso
                   0 3608M 2050M 43388 R 93.5 3.2 1:16.81 target/release/parse_contracts
1459311 penso
                20 0 3608M 2050M 43388 R 92.2 3.2 1:16.92 target/release/parse_contracts
1459312 penso
                20 0 3608M 2050M 43388 R 92.2 3.2 1:16.95 target/release/parse contracts
1459303 penso
                20 0 3608M 2050M 43388 R 90.2 3.2 1:16.81 target/release/parse contracts
1459304 penso
                20 0 3608M 2050M 43388 R 90.2 3.2 1:16.94 target/release/parse_contracts
1461110 penso
                20 0 2685M 51000 39792 S 31.0 0.1 0:00.47 target/release/fetch_blocks
1376682 penso
                20 0 17364 11600 2192 R 2.0 0.0 10h33:22 http
1430060 root
                20 0 2067M 17060 2196 S 0.7 0.0 0:02.12 /usr/bin/docker-proxy -proto tcp -host-ip 127.0.0.1 -host-port 5432 -container-ip 172.20.0.2 -container-port 5432
1855468 root 20 0 390M 23532 6132 $ 0.7 0.0 30:01.07 /usr/bin/python3 /usr/bin/fail2ban-server -xf start F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice 4F9Kill F10Quit
                                                                                                                                          RAM 13GB/62GB ✓ main
```

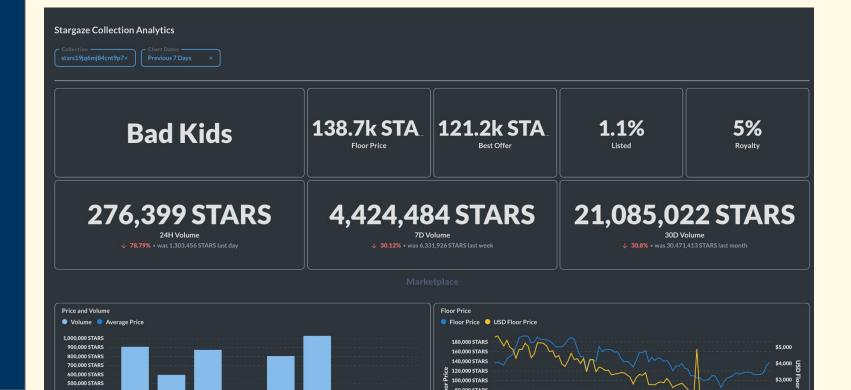
198 errors on 7M requests

```
13 - HTTP Status Codes

Hits h% Vis. v% Tx. Amount Data

7071388 99.96% 12754 99.92% 6.71 GiB 2xx Success
1868 0.03% 193 1.51% 960.0 B 3xx Redirection
941 0.01% 0 0.00% 90.55 KiB 4xx Client Errors
198 0.00% 29 0.23% 77.69 KiB 5xx Server Errors
```

Metabase





An excellent score falls in 1.00-0.94, a good score ranks from 0.93-0.85, a fair score hits 0.84-0.70 and a poor one between 0.69 and 0.49. Any lower number is unacceptable.

Source: TechTarget



The p75 threshold is the value at which 25% of transaction durations are greater than the threshold



ᄎ Stargaze: DEVMOS 2024 After-Party ᄎ

https://lu.ma/Stargaze_DEVMOS-2024

Thank you

For further discussions, reach out to @fabienpenso or devmos@pen.so