Datamule

Making SEC data cheap and easy to use

SEC data is expensive and hard to use

- Datasets derived from SEC filings cost $20-$200k/year.
- Incumbents put strict restrictions on sharing and redistribution that hinders innovation.
- Data is messy, requiring expensive engineer time to make it usable.

We make this easy and cheap

- Sell alternative datasets 10-1000x cheaper.
- Permissive license, enabling customers to easily integrate our data into their GenAI products.
- Open source platform that allows customers to tweak our data to their needs.

Traction

- Main open source package has 427 GitHub stars and 250k downloads.
- $32k in LOI in the last two weeks. This is from companies emailing us, not via outbound.
- Several established players have already tried to acquire us.

Our competitive edge

- We’ve figured out how to iterate quickly.
  - A dataset that took an incumbent data vendor three months to make, we made in ten hours.
- We’ve figured out how to distribute data cheaply.
  - We distribute the entire SEC archive for ~$20/month in storage and distribution costs.
- We can scale.
  - Datasets that cost others a million dollars to create, we create with a dollar.

How we make money

- We sell datasets similar to our competitors, but many times cheaper while enjoying a great margin.
- We sell datasets for which there is no competing offering at a premium.
- Customers can either prepay by loading credits into their account wallet, or by subscribing to a monthly plan with credits that expire if unused.

Future growth

- As a byproduct of making SEC data easy to use, we have also developed high performance generalized parsing algorithms, such as doc2dict.
- These algorithms are many times more efficient than existing methods that utilize GenAI. For example, doc2dict can process 5,000 pages per second. This is 30,000x faster than the typical GenAI method.
- As GenAI adoption spreads, we expect to become a provider of efficient data cleaning APIs. The value of this is unclear, but extremely promising.

Team

John Friedman

- PhD at UCLA.
- Built data pipelines at Berkeley and MIT for economic research.
- Really likes data.

What we need

- Money to help us grow faster.
- Advice to figure out how much we need.
- The questions we don’t know that we should be asking.

Contact

Email: johnfriedman@datamule.xyz

LinkedIn: https://www.linkedin.com/in/johngfriedman/

GitHub: https://github.com/john-friedman