Smallpond

Lightweight Analytics at Scale

deepseek

$WHOAMI

Valery C. Briz

Find me in social networks as **@valerybriz**

Founder of Python Guatemala, previously Co-Leader for multiple PyLadies chapters.

What is Smallpond?

Smallpond is a lightweight, distributed data processing framework built by DeepSeek.

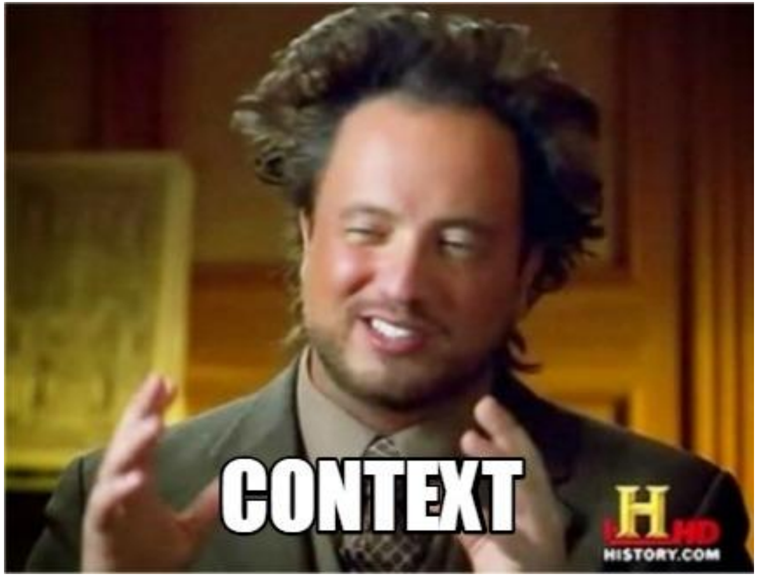
It takes DuckDB and gives it the power to scale across multiple machines.

In other words, smallpond helps DuckDB handle bigger data by spreading the work across a distributed storage and compute system.

Why it’s interesting?

- **Distributed Analytics**: Lets **DuckDB** crunch datasets larger than memory by splitting them up and processing in parallel.
- **Open Source + Efficient**: Paired with **3FS**, it can deliver serious performance without the enterprise price tag.
- **Manual Partitioning:** You control how data is divided; smallpond takes care of distributing it across nodes for parallel execution using **Ray**.

Now let's go deeper with some



DuckDB

Is an in-process analytical database, which basically means it runs inside your application, no separate server needed.

You can add it just like any other library in your favorite programming language.

You can use SQL to query the data

and it is compatible with other tools like Pandas.

3FS (Fire-Flyer File System)

Is a high-performance parallel file system built by DeepSeek. It’s designed specifically for AI and high-performance computing (HPC) workloads.

3FS uses SSDs and RDMA networking to deliver massive throughput and ultra-low latency.

It’s the high-speed, distributed storage layer that powers smallpond’s

incredible performance.

Ray

Ray is a distributed computing framework that makes it simple to scale Python applications from a laptop to a cluster.

It handles:

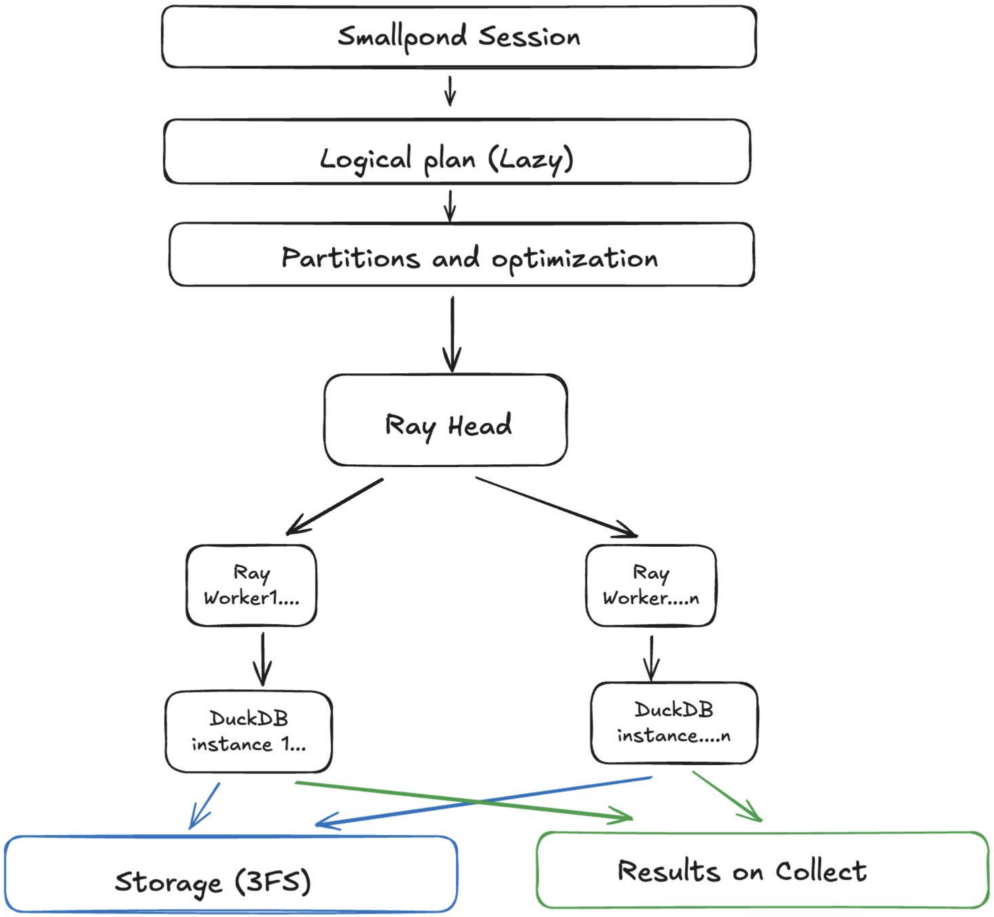
- **Parallelism** running many tasks simultaneously
- **Cluster management** distributing work across CPUs and GPUs
- **Fault tolerance** restarting tasks automatically if nodes fail
- **Unified APIs** letting you run data, AI, or model serving workloads in one system

MinIO

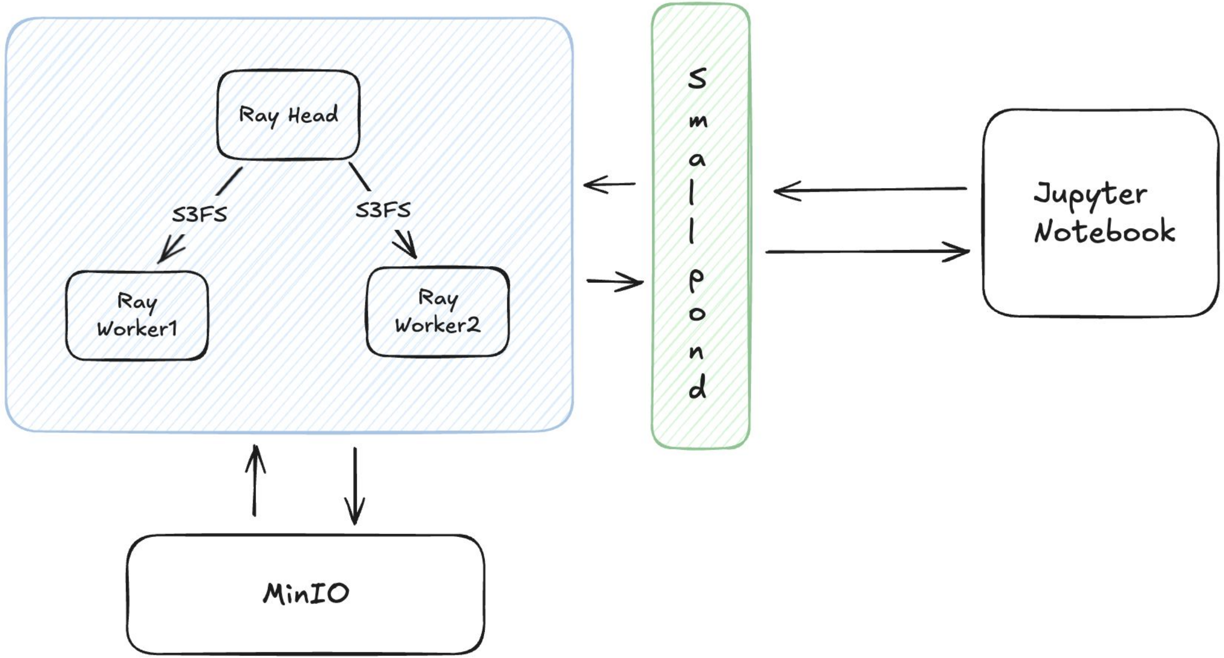
Is an open-source, high-performance object storage system built to store unstructured data like images, videos, logs, backups, ML models, etc, at scale.

It speaks the Amazon S3 API, meaning it can act as a drop-in replacement for AWS S3, but runs anywhere.

How Smallpond works



Our infrastructure



Disclaimer

Smallpond’s documentation is nearly nonexistent and there’s very small info about how people is using it but the good news is that it is based on all this tools that have better documentation so that we can understand how it works.