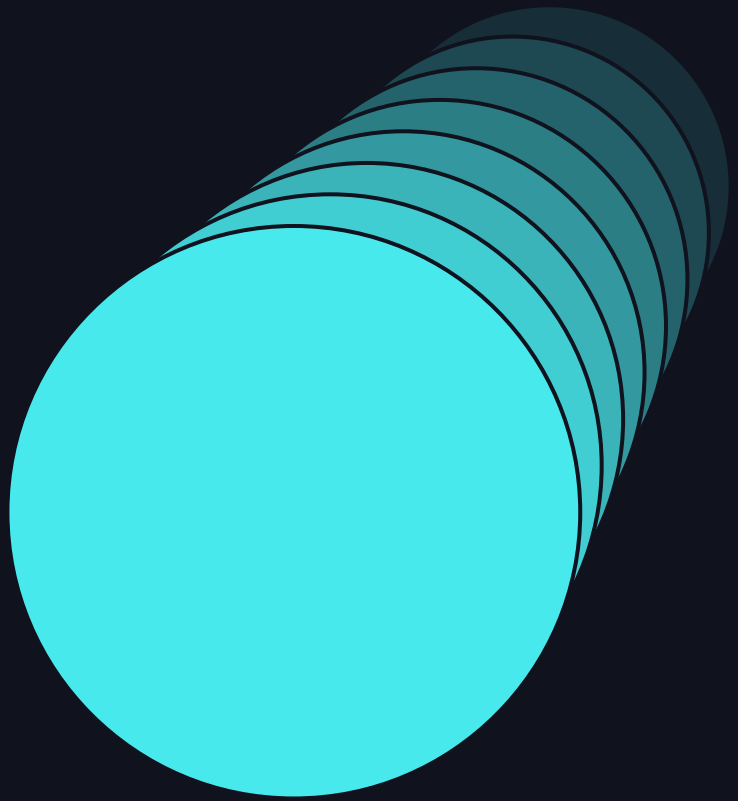
DATA+AI SUMMIT BY databricks

Project Alexandria: A Digital Library



John Darrington
June 2024

Who am I?

- Software architect working with large, distributed systems and petabytes of scientific data
- Father of two and avid 3-D printing supporter
- GitHub: https://github.com/dnoberon – Personal Site: https://notyourlanguage.com (does not represent INL or it's opinions)

What is Idaho National Laboratory?

- 75-year-old nuclear energy laboratory
- Was the provider to the first city run on nuclear power in the USA
- Innovator in integrated and clean energy spaces
- 6,000+ employees and a campus the size of Rhode Island
- $1.6B Budget

How did the Idaho National Laboratory find Databricks?

Search For Efficient Data Storage

- Needs to store petabytes of scientific, structured data
- In-house data lake compatibility
- Easy integration with outside services
- Small footprint

What is Project Alexandria?

A Flexible Data Management Platform

- Platform funded and developed by the NNSA
- Built for modular management and search of projects and ventures in the non-proliferation space
- Collaboration across 7+ Department of Energy National Laboratories and Organizations

Project Alexandria Core Platforms

lakeFS

databricks

Ingest

Open-Source Metadata and Data Collection

- A data ingestion platform tailored to capture metadata from the user is a sophisticated system designed to efficiently collect large volumes of data while simultaneously capturing comprehensive metadata associated with each data item.
- This platform employs advanced technologies such Delta Lake to store information in a structured manner on cloud storage versus traditional databases.

Ingest is currently in pre-release state for open-sourcing – it will be available on https://inlsoftware.inl.gov/ soon.

LakeFS

Git for your data

lakeFS

- Manage data with git-like operations; rollback, branch, and more are supported with this powerful data versioning system
- Alexandria connected LakeFS to Azure Data Lake Gen2 for enhanced data storage capabilities
- Serves as source of truth *and* staging area for project/ventures data in the Alexandria platform's storage layer

https://lakefs.io/

DataHub

An interconnected metadata catalog

- Extensible data catalog platform that allows for complex relationships and information storage on data and data sets
- Developed by LinkedIn, completely open-source
- Integrations with Azure Serverless Functions for a more comprehensive experience than what is offered out of the box

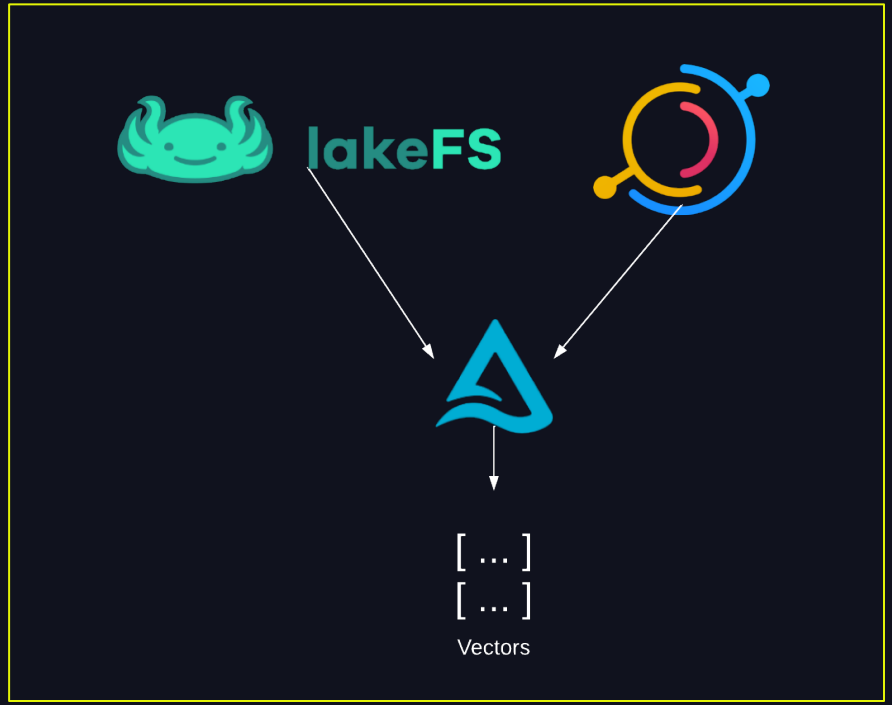
https://datahubproject.io/

Now that we have the data, what do we do with it?

Vector Management

Making the Data Useful

- Combine LakeFS managed metadata/data with Datahub's graph
- Vectors are generated for use in specific model training
- Databricks Vector Search used to maintain index and run queries
- Currently a CLI tool, but plans to develop various RAG LLM chatbots



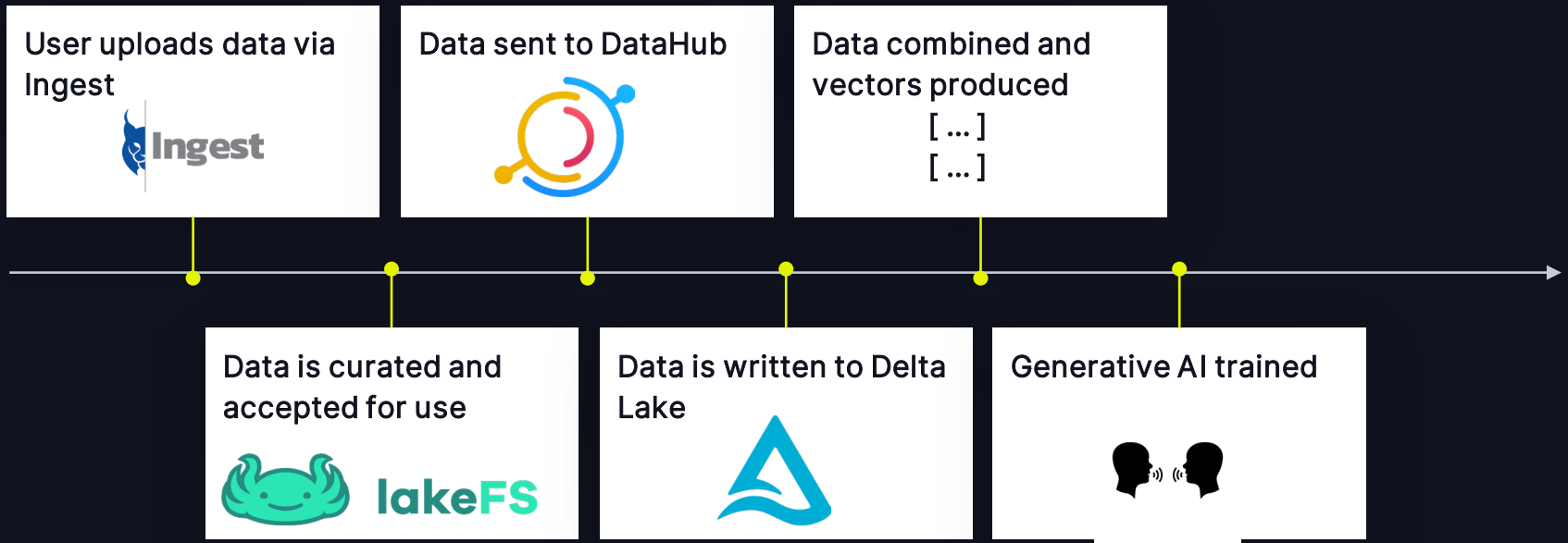
Using the Data cont.

Getting even more out of vectors

- Can run pre-filtering queries on the data to make targeted sets of vectors / manage classified data
- Vectors can now be used to train on user-provided data and metadata without involving them in the training process directly
- Vectors can be reused on classified networks without Databricks thanks to open-source Delta Lake storage drivers

Data Use Timeline

A brief recap



Landmines and Hidden Obstacles

- Unintentional model poisoning
- Feedback loop necessary to produce filtered result sets for models
  - Hard to acquire researcher feedback
- DRY (Don’t repeat yourself) harder to achieve with siloed projects
- Data stagnation
- Classified data = Classified models

GenAI allows for faster time-to-conclusion on complex data than ever before

Questions?