CyberPeace Institute

CyberPeace Institute

Accelerator Program

A machine learning project in support of human rights

Why introduce machine learning into our data pipeline?

TO MAP THE THREATS & HARMS FROM THE USE OF SPYWARE AGAINST VULNERABLE COMMUNITIES & NGOS

Much of the data we use as part of ongoing research to trace and document cyber incidents relating to the misuse and abuse of commercial-grade spyware technologies comes in an unstructured form. With 100s, if not 1000s, of articles to process manual methods are no longer sufficient or efficient.

Our analysts are spending more time processing data than analyzing it!

By introducing machine learning we plan to automate much of the extraction of key information into a format usable by analysts.

Goals: What we seek to achieve

FASTER TIME TO ANALYSIS & THE IDENTIFICATION OF HIDDEN LINKS

AS DATA INCREASES

TIME TO ANALYSIS DECREASES

Deliverables: What are we building?

TOOLS TO FACILITATE AND SPEED UP THE ANALYSIS PROCESS

Graph-oriented Named Entity Recognition

Table Question Answering with BigQuery

Document Summarization

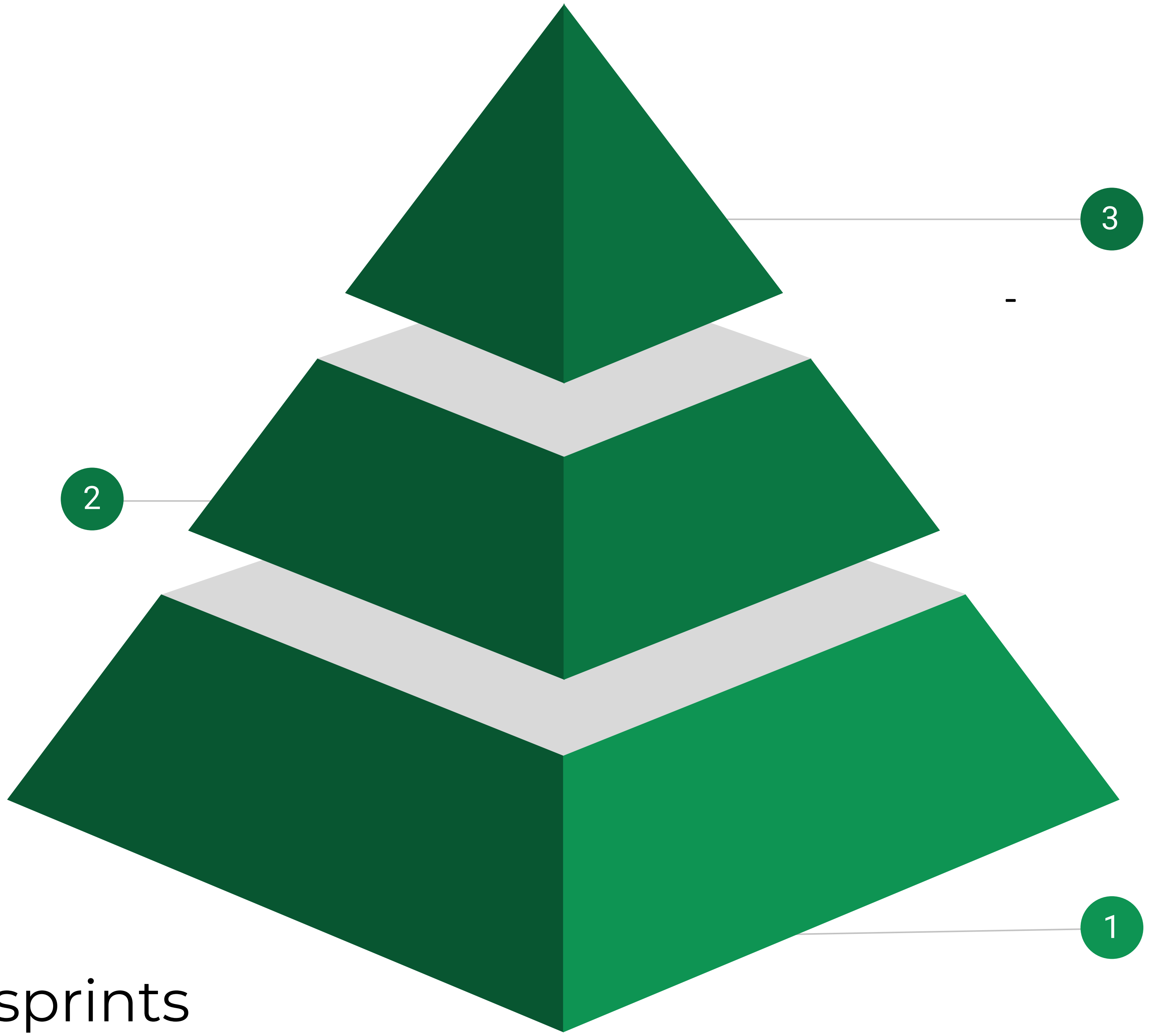
Document Question Answering

WITH OUTPUTS ACCESSIBLE TO AN ANALYST WITH LIMITED CODING SKILLS

Machine Learning Approach

PHASE 1

- Document-to-Text using Document AI from Google services
- NER using Google services only
- Pushing the structured entities to BigQuery via Apache Airflow



PHASE 2

- Labeling of the unstructured data and training of a custom NER model for ensembling
- Creating Knowledge Graphs using the NER ensemble
- Document Summarization using Google's open-source Pegasus model
- Document Question Answering with LayoutLM
- Table Question Answering on a fine tuned TAPAS model from Google

FOUNDATIONS

- Investigating the types of entities that would provide value within the data
- Configuring the data pipeline to feed into and extract from ML models
- Performing exploratory data analysis to understand the preprocessing requirements of the extracted raw data

→ Delivering in **2 week** sprints through the **agile** methodology

○ PJMF Deliverables

Project Timeline

	Month	April		May		June		July			August		September	
PHASE	Activity Week	03-16	17-30	01-14	15-28	29-11	12-25	26-09	10-23	24-06	07-20	21-03	04-17	18-01
Foundations	Definition of project road map	■ ○ map												
	Initial design of infrastructure & data pipeline	■												
	Data exploration & discovery [entity identification]	■	■											
	Data exploration & discovery [processing requirements]	■	■											
	Configure data pipeline to feed & extract from ML pipeline		■											
Phase 1	ML process [document-to-text]			■										
	User Testing & Feedback loop [doc-to-text]			■										
	ML process [NER via Google services]				■									
	User Testing & Feedback loop [NER]				■									
	ETL structured entities to data lake					■								
Phase 2	Labelling unstructured data						■ ○ blog							
	Training custom NER for ensembling						■							
	Creating knowledge graphs using NER							■						
	User Testing & Feedback loop [knowledge graphs]							■						
	Document summarization								■ ○ repo					
	User Testing & Feedback loop [document summarization]								■					
	Document Question Answering									■				
	Table Question Answering									■				
	User Testing & Feedback loop [Question ^ Answering]										■			
	Data visualization & analysis											■	■	
	Deploy into production												■	■

The Project Team

The CyberPeace Institute prides itself on the diversity of professional backgrounds and expertise whilst also empowering the younger generation to take part in our projects through internship programs.

Machine Learning Intern

ATA

Chief Information Security Officer

FLORENT

Senior Software Engineer

SIEGFRIED

Full stack web developer

NIKOLAOS

Chief Research & Analysis Officer

EMMA

Senior Intelligence Officer

IAN

Success Outcomes

PHASE 1

- Integrating a simple ML pipeline into the pre-existing data pipeline
- Having moderate success in recognizing entities in documents
- Successfully testing different use cases.

Success Outcomes

PHASE 2

- Extracting Knowledge Graphs out of unstructured documents
- Adding summarization and labeling functionalities to increase analysis efficiency
- Enabling the filtering of details by incorporating Document Question Answering capabilities
- Facilitating analysts to interact with knowledge graphs and extract information through natural language via Table Question Answering

Project Components

CyberPeace Institute

DATA

Structured Data - for learning

Unstructured Data - to process and analyze

MACHINE LEARNING

NER - Custom models, GCP Services

Labeling - Custom models, Open-source

Summarization - Open-source

Question Answering - Hugging Face

CORE INFRASTRUCTURE

Cloud Based Infrastructure

EC : Kubernetes

Storage - GCP Technology

Storage - AWS + GCP

ETL - Airflow, Cloud Dataflow

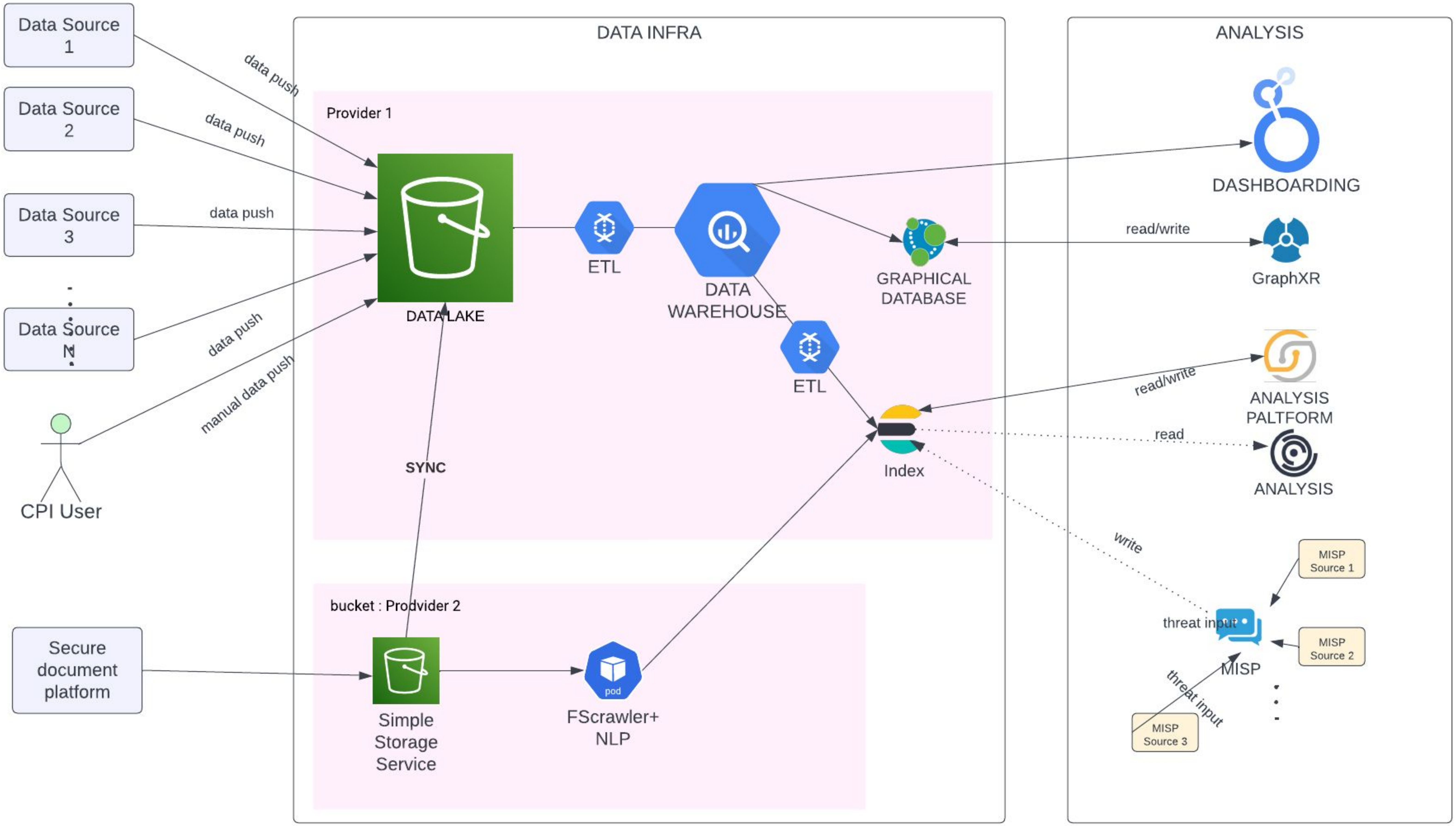
END USER TOOLS

Graph analytics - GraphXR, Maltego

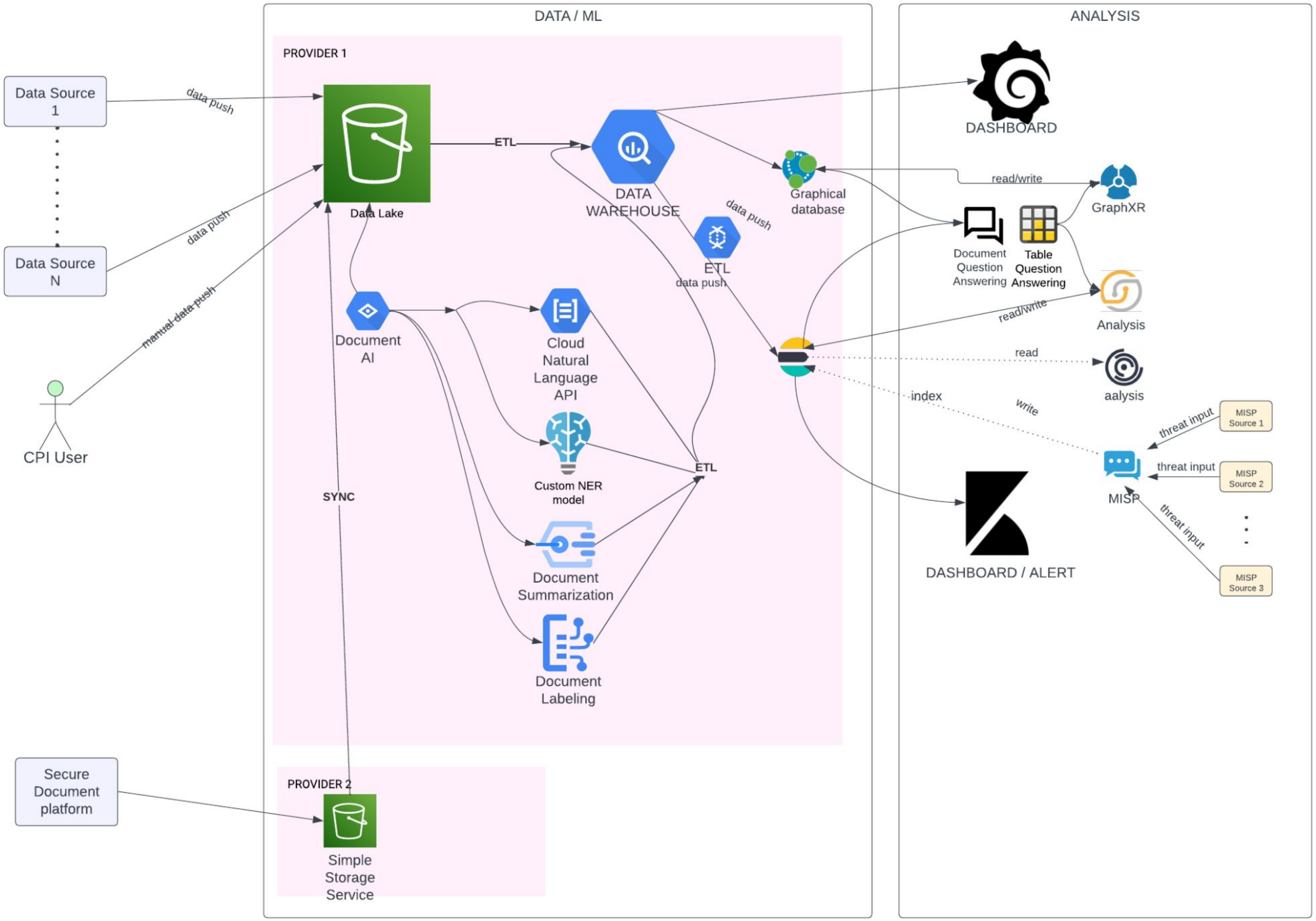
Dashboarding - Kibana / Looker / Grafana

Search & query - BigQuery

Our Infrastructure | today



Our Infrastructure | in 6 months



Our Data

Today

Structured Data

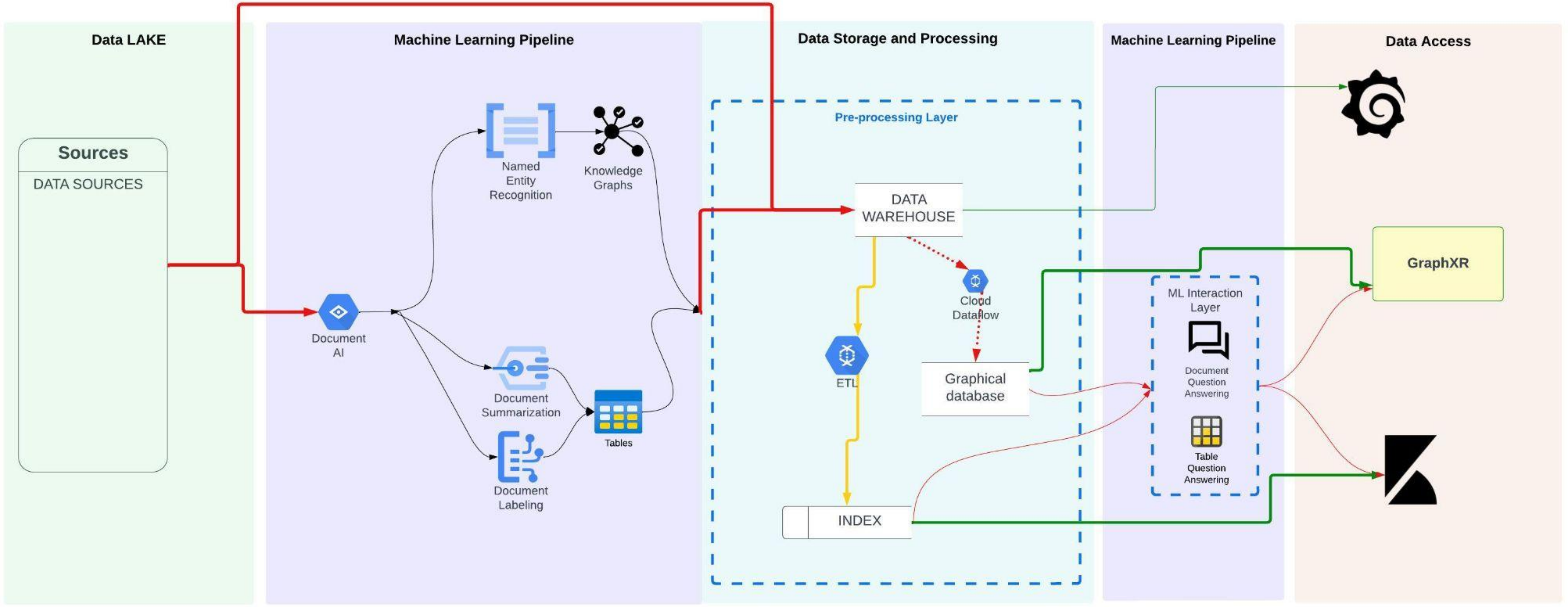
- Cyber incidents / events
- Entity-based information
  - Organizations [e.g. companies registration information]
  - Persons
  - Locations
- Exports or extracts from databases
- ....

Unstructured Data

- Media articles
- Blog posts / online articles
- Social media profiles [e.g. LinkedIn]
- Lengthy reports
- Exports from online databases
- Leaked documents or emails
- ...

Our Data Pipeline

In 6 months



How do we achieve this

Though we still exploring the technologies and algorithms we will use; we currently expect to experiment with the following three pillars of Machine Learning.

1

Knowledge Extraction

Extract all points-of-interest from unstructured text.

2

Knowledge Organization

Create data structures so as to create structured outputs from unstructured data and potentially including structured data as well.

3

Knowledge Search

Be able to search the organized knowledge and automate obtaining points-of-interest from the knowledge base.

Knowledge Extraction

EXTRACTING POINTS OF INTEREST

This pillar of the work deals with extracting points of interest from unstructured data namely reports in the form of pdfs, online articles and message forums.

This takes the form of three different problems:

- Named Entity recognition
- Document Labeling
- Document Summarization

Here machine learning can do all the tasks. A lot of work has been done in these areas of work and should be explored comfortably within the organization.

Knowledge Organization

CREATE EXPLORABLE DATA

This pillar takes the data from the knowledge extraction pillar, combines it with structured datasets and organizes it into a data structure that is explorable. This can take the form of knowledge graphs that seems to be the obvious fit.

This work requires some experience before being taken up. If achieved, however, this should lead to significant improvements in the workflow of analysts and make their work much more efficient.

Knowledge Exploration

KNOWLEDGE GRAPHS

This pillar of work comes after knowledge graphs have been created. This doesn't need to, however. The whole area of information retrieval subfield in natural language processing is made for such cases. If the knowledge graph area is explored then it can be integrated with bayesian inference methods and LLMs for search that makes it extremely robust as compared to what exists today.

Project Risks & Mitigations

CyberPeace Institute

KNOWLEDGE & EXPERTISE

- Academic level experience in ML with limited professional experience.
  - Lean on PJMF & cohort expertise for support
- Introducing custom models into our data pipeline will be dependent on expert guidance from outside PJMF
  - Actively use the Slack channel to ask for help
- With limited experience in the team putting ML into a professional environment and data pipeline, our proposed phases and ML components may turn out to be overly ambitious given then project timeframe.
  - Revisit the project roadmap on a regular basis and update it accordingly

DATA & INFRA CHALLENGES

- Dataset may be too small to provide sufficient data for ML.
  - Acquire additional datasets using allocated budget.
- Variety of data sources could cause challenges when preprocessing the data and summarizing files.
  - Invest sufficient type at the Foundation stage to explore the existing data.
- Testing the accuracy of results from the chatbot will be difficult.
  - Involve different team members & receive recommendations from the cohort.
- We are yet to define our tolerance for false positives / false negatives with possible implications on delivery timeframe.
  - Be flexible as part of the PoC and consider stricter rules as the project matures.
- Latency issue connected to the runtime of ML model inference.
  - Executing the ML pipeline once fresh data is included to ensure that the extracted information is prepared for analysis instantly