AGORA-EO: A UNIFIED ECOSYSTEM FOR EARTH OBSERVATION – A VISION FOR BOOSTING EO DATA LITERACY –

Arne de Wall[∗], *Björn Deiseroth*[∗], *Eleni Tzirita Zacharatou*[∗], *Jorge-Arnulfo Quiané-Ruiz*[∗,⋆],
Begüm Demir[∗], *Volker Markl*[∗,⋆]

[∗]Technische Universität Berlin [⋆]DFKI

ABSTRACT

Today's EO exploitation platforms are limited to the processing functionalities and datasets they provide, and there is no single platform that provides all datasets of interest. Thus, it is crucial to enable cross-platform (federated) analytics to make EO technology easily accessible to everyone. We envision AgoraEO, an EO ecosystem for sharing, finding, composing, and executing EO assets, such as datasets, algorithms, and tools. Making AgoraEO a reality is challenging for several reasons, the main ones being that the ecosystem must provide interactive response times and operate seamlessly over multiple exploitation platforms. In this paper, we discuss the different challenges that AgoraEO poses as well as our ideas to tackle them. We believe that having an open, unified EO ecosystem would foster innovation and boost EO data literacy for the entire population.

Index Terms— Big Data, Earth Observation Ecosystem, Federated Analytics

1. INTRODUCTION

The growing operational capability of global Earth Observation (EO) provides scientists and practitioners with a wealth of information that can be used for various EO applications. However, to take advantage of this type of data, users need a significant amount of expertise to validate their assumptions and gain insights. This includes scientific knowledge about the applicability of remote sensing (RS) data and technical expertise to manage and process vast amounts of EO data. Consequently, non-specialist users cannot easily exploit EO data for their use-cases.

Recently, the industry and research communities have proposed different approaches aiming to make EO data more accessible. Different EO exploitation platforms, such as the Google Earth Engine, Open Data Cube, and Sentinel Hub, offer a certain level of processing and data abstraction that considerably ease large-scale analyses [1]. This enables researchers and practitioners to create rapid insights without caring about the underlying infrastructure. Despite these benefits, all exploitation platforms rely on a heterogeneous set of technologies with varying sets of interfaces and data formats, making the cross-platform and federated use of these platforms difficult. For instance, it is not straightforward to take an analytics pipeline developed on a given platform and apply it on a different platform. Unfortunately, this results in vendor lock-in effects, hence, naturally leads to a few players controlling ("monopolizing") most EO technologies. Users often have to stick to one given platform. The openEO project [2] tries to counteract this problem by introducing a harmonized API that connects various clients (e.g., JavaScript, Python, R) to different EO processing backends (e.g., Google Earth Engine, Sentinel Hub). However, openEO operates on single processing backends only, restricting the use to the provided data and processing capabilities of the respective platform. The cross-platform and federated use of data and processing technology is not supported.

Moreover, we believe that EO technology must be accessible by *everyone* and hence enabling organizations, individuals, and EO exploitation platforms to interoperate is of crucial importance. We envision AgoraEO, a decentralized, open, and unified EO *ecosystem*, where one can share, find, compose, and efficiently execute cross-platform EO assets. An asset can be any technology one can use in an analytics pipeline to get value out of EO data. Assets are thus highly diverse, ranging from datasets (e.g., Sentinel data), processing tools (e.g., SNAP and GDAL), and machine learning (ML) algorithms (e.g., K-means and SVM) to computing power (e.g., Amazon AWS) and human expertise (e.g., a RS expert providing consultancy services).

Realizing the AgoraEO ecosystem vision would enable users to achieve a particular task at hand with the best existing assets, regardless of who owns the assets. More importantly, such an ecosystem would foster innovation by reducing the cost of getting new insights and improving EO data literacy for the population as a whole. Yet, we must tackle three main roadblocks to make AgoraEO a reality. First, the high heterogeneity of assets in the envisioned ecosystem renders their management (storing, querying, and composition) a challenging problem. We envision a unified data access layer that facilitates the exploration of EO assets in the ecosystem. Second, most EO analytics are exploratory in nature which requires lightning-fast query results. Thus, we plan to develop a fast EO data management engine that spans multiple exploitation platforms to provide a good user experience. Third, data assets are naturally scattered across multiple sites and platforms, making conventional (AI-based) analytics hard. We will devise a framework to support federated analytics.

In the remainder of this paper, we first illustrate how AgoraEO would empower users in Section 2. We then present our envisioned AgoraEO architecture in Section 3. We discuss our envisioned data management engine and the different open challenges to support EO data exploration in Section 4. We then discuss in Section 5 how we plan to support distributed AI-based analytics in EO as well as the open challenges to achieve so. Note that AgoraEO is an instance of the overall vision of Agora [3], whose goal is to provide the core infrastructure for building data-related ecosystems at scale.

2. MOTIVATING APPLICATION

Consider a Business Analyst (BA) who works for an insurance company that holds assets, i.e., insured objects, in a spatial database. Now, assume she wants to estimate the amount of flood-damaged properties in a particular city during an ongoing flood event. To do so, she first searches in AgoraEO for a time-series of Synthetic Aperture Radar (SAR) images (e.g., Sentinel-1) that matches the desired spatial and temporal constraints of the event. Next, she takes an asset that computes the water masks and the maximum extent layer for the stack of SAR images based on a threshold approach. She repeatedly modifies the threshold parameter for this asset to optimize the generated output. Finally, she searches for permanent water bodies in Germany, finds the corresponding dataset, which is in vector format, and subtracts the water bodies from the previously generated raster product to receive a flood mask. In the next phase, assume that the BA wants to apply and automate the workflow for determining flood-damaged properties on a larger geographical scale (e.g., Europe). As manually fine-tuning the thresholding approach does not scale in practice and is difficult to automate for Sentinel-1 data, she decides to train a more sophisticated deep neural network asset to extract water masks from SAR imagery. She again searches AgoraEO for assets that provide water bodies and SAR imagery in Europe. For the SAR imagery, she finds several regional and national scale data assets providing SAR as Analysis-Ready-Data (e.g., data cubes). To train a more robust and less biased model, she would like to train a neural network for semantic segmentation of the water masks over all available data. Thus, she wants to train a model collaboratively among a diverse set of available data and combine the training results into a common model.

To enable the above motivating example, it is important to (i) leverage all relevant assets in the ecosystem, regardless of the platform that contains them, and (ii) provide federated analytics capabilities as the required assets might potentially be located on different platforms. Next, we present a set of ideas for supporting the above BA use-case.

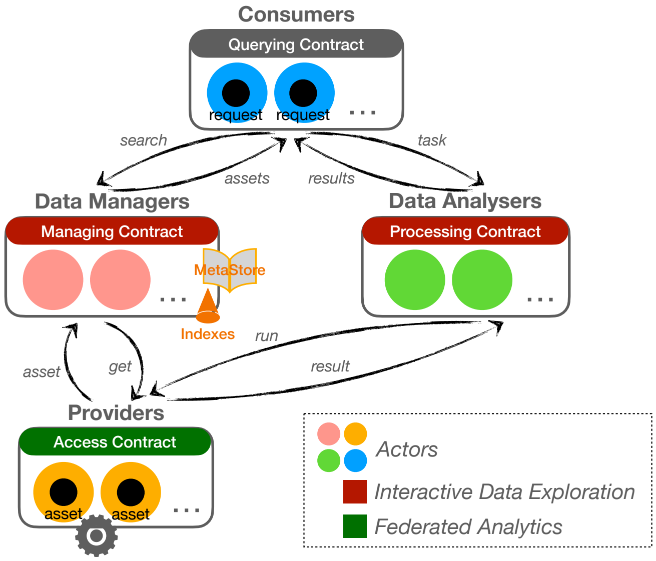

Fig. 1. Actor model-based architecture of AgoraEO.

3. AGORA-EO

We build AgoraEO around assets, i.e., diverse data-related units of production allowing users to exploit the value of data, such as datasets, algorithms, analytical pipelines, and processing systems. AgoraEO's key idea is that an ecosystem should offer both the resources and the infrastructure required to run any data-driven EO application. Users of the ecosystem form a community that creates, exchanges, and uses assets. Through this open and unified EO ecosystem, anyone offering an EO asset can benefit from shared revenues.

We envision an architecture for the AgoraEO ecosystem that follows the actor model [4] to achieve a plug-and-play behavior. We believe that this is fundamental for the success of such an envisioned ecosystem. Figure 1 shows the architecture of AgoraEO. As we observe, AgoraEO is mainly composed of components with the following roles: a *consumer* uses assets in the ecosystem; a *provider* offers assets in the ecosystem; a *data manager* manages (searches and composes) assets in the ecosystem; and a *data analyzer* performs analytics on a set of assets. AgoraEO has an actor for each of these roles. An actor thus defines the behavior (the capabilities), via a contract (interface), that a user of the ecosystem can have. In other words, AgoraEO users can play any of the above roles as long as they comply with the actors' contract. The advantage of this architecture is twofold: (i) it prevents lock-in effects in the ecosystem by enabling actors to interact among them seamlessly; and (ii) its simplicity allows AgoraEO to operate in a fully decentralized and scalable manner.

More importantly, AgoraEO will provide the following two main features. First, it envisions an interactive data exploration interface, i.e., users must get the results of their analytics almost instantly. Providing this interface is key for boosting human productivity and democratizing big EO analytics by making the techniques available to a broader community. Yet, achieving fast response times is challenging because data analyses typically involve computationally-intensive operations, such as ML. Besides accelerating the data analysis itself, this also calls for effective and fast content-based retrieval of EO data. Doing so requires scalable solutions that can precisely characterize the high-level spatial, spectral, and even temporal information content of satellite images while providing fast data ingestion and retrieval. Second, the ecosystem must analyze (or learn from) data that are highly scattered around the world. Moving large amounts of EO data into a single place for analysis is typically not feasible. The challenge resides in how to run analytics in-situ without compromising the quality of the results at a global scale.

4. AGORA-EO: INTERACTIVE EXPLORATION

The increasing number of EO satellites combined with the ever-increasing acquisition rates lead to the tremendous growth of EO data archives. Furthermore, researchers and practitioners increasingly combine EO data with a wealth of spatial data that originate from other sources such as social media [5]. As the amount and diversity of data grow, extracting useful knowledge from them is becoming even more cumbersome, especially for non-expert users. Today's EO platforms provide limited exploration capabilities, failing to meet the needs of users searching for relevant data in this data deluge. While they usually allow searching by geographical extent, acquisition time, or type of sensor, they do not support searching by the semantic content of satellite images, which is crucial for many applications, such as mapping burnt forests or flooded residential areas. Furthermore, EO platforms lack efficient support for operations that combine vector data (e.g., administrative region boundaries) with satellite raster images. As a result, users need to perform several iterations of data retrieval and analysis in a hit-and-trial process to decide what is interesting and useful. This onerous process hinders exploratory analyses, often preventing users from making unexpected discoveries by interacting with the data.

Enabling data exploration is challenging for three main reasons. First, data exploration requires interactivity [1]. The system must respond fast to user requests, as delays reduce the rate at which users make observations, and draw generalizations and hypotheses. However, combining raster and vector data, as often required for the analysis, is a computationally-intensive operation. Existing approaches either vectorize the raster data or rasterize the vector data. While vector-based approaches involve expensive point-in-polygon tests, raster-based ones suffer from the computational overhead of the rasterization step. To overcome this challenge, we leverage the insight that minor errors can be tolerated in exploratory analyses. Allowing approximate results enables reducing the costly point-in-polygon tests and trading precision for response time [6]. After looking at the big picture, users can progressively refine the results if they wish to perform a more detailed analysis. Furthermore, we plan to exploit modern GPUs for rasterizing vector data as rasterization is an important component of the graphics rendering pipeline that is natively supported by GPUs.

Second, the responsiveness of the system is also challenged by the complexity of the applied ML models. To alleviate this challenge, we envision an optimizer that learns both from user interactions and explored data. We also plan to enable the optimizer to perform database optimizations, such as operator reordering, across complex pipelines that combine geospatial operators with ML tasks thereby reducing the number of expensive computations. For instance, recall the motivating application of Section 2 and assume that the BA wants to investigate historical flood-related insurance claims. This task can largely benefit from smart optimizations. The optimizer can create a plan that (i) first filters the SAR imagery with the set of vector polygons that correspond to the locations of the insured objects, (ii) then applies a quick filter to eliminate the tiles where there is no water present using the simple threshold-based asset with a rough threshold, and finally (iii) refines the output with a more sophisticated and computationally-intensive deep neural network asset.

Lastly, content-based image retrieval is an open challenge. Knowledge discovery systems with "query-by-example" or even just ”by-description" functionality do not exist yet. Thus, we need novel techniques that can precisely characterize the spectral, spatial, and even temporal information content of satellite images. Considering the volume and the velocity of the produced data, indexing strategies that embed high-dimensional features into an efficient sparse representation (e.g., binary hash codes) while having low latency for ingestion and retrieval are of particular importance.

AgoraEO aims to tackle the aforementioned limitations and support *interactive data exploration*. To that end, AgoraEO's interactive data exploration layer (cf. Figure 1) will incorporate the following key enabling features: (i) fast algorithms for combining vector and raster data, (ii) optimizations for complex processing pipelines that mix geospatial analytics on both raster and vector data with ML models, and (iii) efficient content-based image retrieval.

5. AGORA-EO: FEDERATED LEARNING

AgoraEO envisions an ecosystem that operates in a fully decentralized fashion. In particular, this idea fits the current situation found in the EO domain, where datasets are highly scattered across multiple sites and where no single processing platform exists that serves all possible datasets of interest. In contrast, modern AI-based solutions increasingly leverage data fusion from various sources (e.g., RS imagery, weather data, social media). However, standard ML algorithms require centralizing the data before training. This can become problematic in an EO setting since EO and other georeferenced data sources are often particularly large in volume. Additionally, centralizing datasets might be impossible due to data constraints, such as privacy regulations, or legal issues.

Federated Learning (FL) is a way to address this problem of associated governance, privacy concerns, and technical challenges. The goal of FL is to realize collaborative learning among distributed clients without centralizing the data [7]. In other words, it allows to learn locally on each client close to the data and aggregate the resulting local models to a shared global model. We believe FL can also be quite beneficial for AgoraEO and the EO community in general. For instance, recall the motivating application of section 2 and assume that the BA wants to train a robust model across different assets. FL can support the use-case by enabling the cross-asset model training close to the data without exchanging the data. This reduces data transfer costs to a minimum and models can generalize across EO datasets, leading to more accurate and less biased models. FL also allows AgoraEO to unlock restricted datasets for answering EO-related questions. Note that many public authorities and private industries remain reluctant to share their data with third parties for confidentiality, privacy, and legal reasons, but often also due to its economic value.

Applying FL to geo-distributed EO data is challenging because FL relies on the general assumption that the underlying data are independent and identically distributed, which is rarely the case in real-world EO data scenarios. Furthermore, in RS, the data naturally tends to be particularly heterogeneous – not only because of the variety of different sensors, modalities, and characteristics of RS missions but even within the same product. For example, the same class can have different representations worldwide (e.g., urban areas can look quite different), or the distribution of labels across different local datasets can vary (i.e., classes such as forests do not exist in certain areas). Many prominent FL algorithms disappoint under these conditions, leading to a drop in the overall performance or divergence in collaborative learning [8]. Despite these challenges and the fact that FL is currently in an early stage of research, we consider it a powerful learning alternative that fits the vision of AgoraEO's decentralized ecosystem.

We plan to devise a framework that enables users to run analytics over multiple processing and exploitation platforms in a unified manner. In particular, we plan to devise new algorithms for efficient decentralized training of large-scale and robust models across different data asset providers. For this, we will leverage our previous experience in developing Rheem [9], a cross-platform system for seamlessly running analytics over multiple data processing platforms.

6. CONCLUSION

We presented AgoraEO, an open and unified asset ecosystem for EO. AgoraEO provides the technical infrastructure for supporting ecosystems where one can offer, discover, combine, and efficiently execute EO-related assets to get data-driven insights. We presented the actor model-based architecture of AgoraEO, allowing for high scalability and a plug-and-play behaviour. Furthermore, we discussed two envisioned features of AgoraEO, interactive data exploration and federated learning for EO, and highlighted the challenges that we need to overcome to enable them.

7. ACKNOWLEDGEMENT

This work was funded by the German Ministry for Education and Research as BIFOLD – Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A).

REFERENCES

[1] M. Sudmanns, D. Tiede, S. Lang, H. Bergstedt, G. Trost, H. Augustin, A. Baraldi, and T. Blaschke, "Big Earth data: disruptive changes in Earth observation data management and analysis?" *Int. J. Digit. Earth*, vol. 13, no. 7, pp. 832–850, 2020.

[2] M. Schramm, E. Pebesma, M. Milenković, L. Foresta, J. Dries, A. Jacob, W. Wagner, M. Mohr, M. Neteler, M. Kadunc *et al.*, "The openeo api–harmonising the use of earth observation cloud services using virtual data cube functionalities," *Remote Sensing*, vol. 13, no. 6, p. 1125, 2021.

[3] J. Traub, Z. Kaoudi, J.-A. Quiané-Ruiz, and V. Markl, "Agora: Bringing together datasets, algorithms, models and more in a unified ecosystem [vision]," *SIGMOD Record*, vol. 49, no. 4, 2021.

[4] C. Hewitt, P. B. Bishop, and R. Steiger, "A Universal Modular ACTOR Formalism for Artificial Intelligence," in *IJCAI*, 1973, pp. 235–245.

[5] B. Bischke, D. Borth, C. Schulze, and A. Dengel, "Contextual Enrichment of Remote-Sensed Events with Social Media Streams," in *ACM MM*, 2016, pp. 1077–1081.

[6] E. Tzirita Zacharatou, A. Kipf, I. Sabek, V. Pandey, H. Doraiswamy, and V. Markl, "The case for distance-bounded spatial approximations," in *Proc. CIDR*, 2021.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *PMLR AISTATS*, 2017, pp. 1273–1282.

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated Learning with Non-IID Data," *arXiv:1806.00582 [cs.LG]*, 2018.

[9] D. Agrawal, S. Chawla, B. Contreras-Rojas, A. K. Elmagarmid, Y. Idris *et al.*, "RHEEM: Enabling Cross-Platform Data Processing - May The Big Data Be With You!" *PVLDB*, vol. 11, no. 11, pp. 1414–1427, 2018.