STANDARDIZATION EFFORTS ACROSS SPACE AGENCIES: APPLICATIONS AND ANALYSIS READY DATA DISCOVERY IN THE CLOUD

Ingo Simonis

Open Geospatial Consortium (OGC), London, UK

ABSTRACT

The advancements in Big data processing pipelines have consolidated the "processes to the cloud" paradigm over the last decade. This represents a shift away from downloads and local processing. Recent publications have addressed commercialization elements such as quoting and billing across clouds and predict the development of markets pretty similar to application markets we know from the mobile phone sector. At the same time, Analysis Ready Data becomes available at many places. This paper addresses the question how to discover all these applications, Analysis Ready Data, and cloud resources in an interoperable way, meaning that the process is not different from resource to resource.

Index Terms— Applications, Cloud, Big Data, Analysis Ready Data, Standards

1. INTRODUCTION

The advancements in Big data processing pipelines have consolidated the "processes to the cloud" paradigm over the last decade, which represents a shift away from previously established architectures that favored downloads and local processing. It is not only the enormously growing amount of available data that makes - despite all improvements in network capacities - the traditional architectural approaches impractical these days. It is, with growing importance, the success of artificial intelligence and machine learning technologies that has changed the way data is often processed these days. Progress in cloud technologies and increased processing resources now allow to process large quantities of e.g. satellite scenes or climate change ensemble data instead of being constraint to relatively small number of files. This development can be seen across domains and is not limited to spatial data [1, 2]. It applies similarly to e.g. the health analytics, genome research, or material design [3, 5, 4].

With the successfully established new architectures come new commercialization opportunities. Whereas before the market was somewhat constraint to sales and distribution of data on one side and desktop applications for data analysis on the other, the market is now broadened. Application developers can develop applications that address any step of data processing pipelines and offer these applications for sale. Consumers can request the ad-hoc deployment of these applications and executed them with selected data sets effortlessly. These new opportunities, together with standardization approaches to make quoting and billing processes more interoperable, have been described in recent publications [6].

At the same time, 'Analysis Ready Data' is floating around as a buzzword, with most definitions having in common the fact that the data is the product of some processing that qualifies it for direct knowledge generation and fact display. This paper addresses the question how to discover all these applications, Analysis Ready Data, and cloud resources in an interoperable way, meaning that the process is not different from resource to resource.

The remainder of this paper is structured as follows. First, we will describe recent standardization efforts that underpin the level of maturity that the 'application to the data' architectures have reached. Next, we discuss the discovery issues for Analysis Ready Data, applications, and cloud processing resources, before we introduce first ideas to address these. The paper concludes with the description of a large research and development effort executed within the Open Geospatial Consortium (OGC) to further address the discovery aspect, which is currently inadequately addressed across data and infrastructure providers.

2. CLOUD ARCHITECTURES

Following the idea of a generic set of Earth Observation Exploitation Platforms that build a type of transparent and permeable platform ecosystem, the European Space Agency (ESA) has made available a number of Thematic Exploitation Platforms (TEPs) as well as Mission Exploitation Platforms (MEPs) on cloud platforms over the last couple of years. At the same time, driven by rapidly rising data volumes, NASA's Earth Observing System Data and Information System (EOSDIS) is migrating to a cloud computing based archive over the next few years with the main aim to provide the data in an environment where end users can bring their analysis to the data rather than attempting to download and manage ever-increasing volumes [7].

Both NASA's and ESA's approach is based on open services. These services allow the efficient combination of essential capabilities, such as data capturing and cleaning,

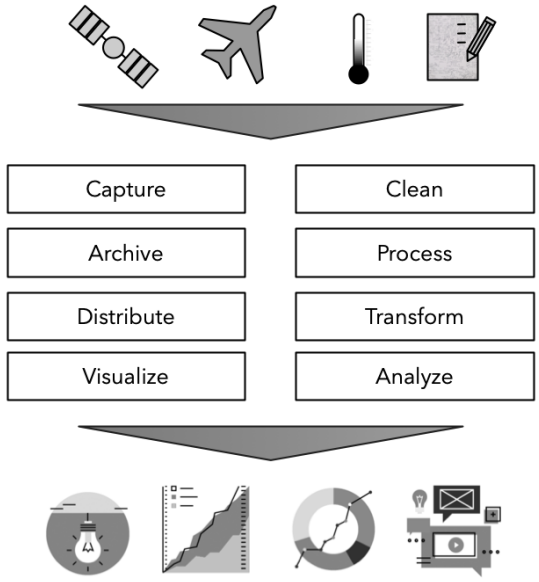

Fig. 1. Cloud environment services for custom-made data processing and analytics

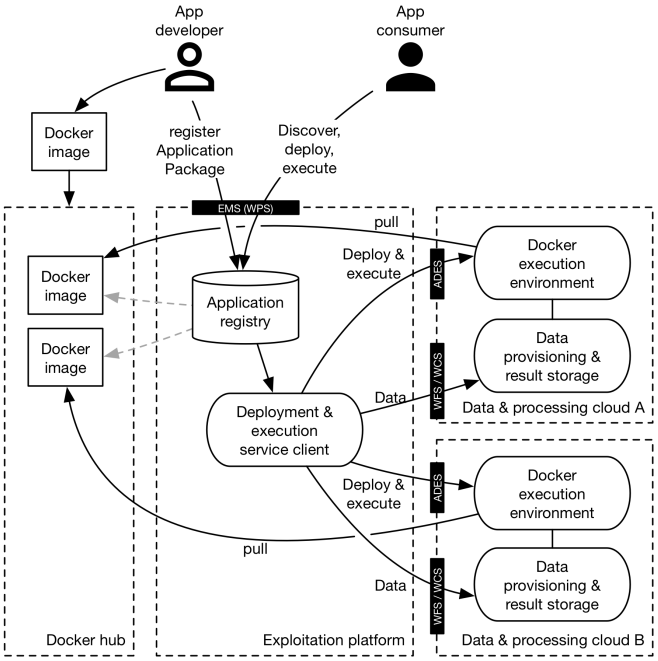

Fig. 2. Rapid application deployment and execution architecture; source: [6]

archiving, processing, re-distribution, transformation, visualization, or data analysis. As illustrated in figure 1, these processes can be chained and put into any sequence a consumer deems relevant if provided as individual services. The setup allows the integration of any type of sensing device (figure 1, top), and the provisioning of tailored products for the consumers at the bottom, that may be scientists, private industry, public sector institutions, or the general public.

As introduced in [9] and described in detail in [10, 11, 12], a corresponding infrastructure has been built over the last two years within the ESA environment. These efforts, all executed as OGC Innovation Program initiatives (Testbed-13 and Testbed-14), resulted in the development of models and specifications related to packaging, deployment, and execution of applications in cloud environments that expose standardized interfaces (Web Processing Service, WPS). Though the initial activities were built on ESAs cloud platform environment, the developed approach is agnostic to the underlying cloud platform as long as a dedicated standardized interface (e.g. a Web Processing Service, WPS), a container execution environment (e.g. Docker), and data access mechanisms (ideally supporting dynamic mounting into and out of containers) are provided. Figure 2 illustrates the various components that have been developed. App developers make their products available as Docker images that are stored on Docker hubs and described according to the OGC Application Package specification. Application Package descriptions are stored in the application registry, where they are discovered by application consumers. The latter can request the ad-hoc deployment of applications close to the data. In such a case, the corresponding Docker image is pulled from the hub, dynamically deployed and executed in a Docker execution environment, and final results are provided back to the consumer.

3. APPLICATION, SERVICE, AND DATA DISCOVERY

In order to allow end-users to exploit applications and already deployed processing capacities that serve or can produce analysis ready data, facilities must be provided for users to discover the particular components, services, and data sets. In the case of ad-hoc deployable applications, detailed descriptions and invocation instructions are essential. In this discovery and invocation context, a number of research questions need to be addressed.

First, it needs to be analyzed how an application and application data catalog can be established without inviting yet another catalog specification. In the context of geospatial data, we already see an overload of specifications that allow resource discovery, with OGC CSW (Catalog Service Web), OGC OpenSearch, STAC (SpatioTemporal Asset Catalog specification), Web Feature Service (WFS 3.0), DCAT and GeoDCAT-AP, the Semantic Resource Information Model (SRIM), application-based catalogs such as Google Play Store or Apples App Store, or the Digital Object Interface Protocol just to name the most important ones. Key to establishing a meaningful discovery solution is solid understanding of the differences and commonalities of the various catalog models, their limitations and key characteristics.

Once a base catalog technology has been identified, the

question how to link apps and data needs to be addressed next. Consumers need to have some guidance on which application works on what type of data or can be applied to which type of data in principle and without comparing apples with oranges.

Analyzing existing catalog instances, it can be observed that catalogs are often pointing to the access interface, but not necessarily to the data itself. Thus, there is a gap between the link provided by the catalog and the actual data that needs to be bridged; ideally without adding any extra burden on the consumers.

If the data and applications are discoverable and linked to each other and the gap between catalog entries and actual data sets has been successfully addressed, other aspects come into play, such as e.g. the quality of any given application, or even more complex, the quality of any given application processing a particular set of data.

4. CONCLUSIONS

Substantial progress has been made to allow application developers to make their products available to consumers for ad-hoc deployment and execution in cloud environments. Simultaneously, standardization efforts produced a set of specifications that allow making any type of processing capacity available as a service. Both, services and applications can be deployed close to the physical location of the data. The key challenge now is discovery of the various elements, i.e. processing capacities, applications, and data. The Open Geospatial Consortium has started another large initiative, Testbed-15, to address these topics. Though first results will be available end of the year, it is certainly acknowledged that this is a heavy, multi-year endeavour that requires continuous rapid prototyping and research along the way.

5. REFERENCES

[1] Li, D., Wang, S., Yuan, H., and Li, D. (2016). Software and applications of spatial data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(3), 84-114.

[2] Al Nuaimi, E., Al Neyadi, H., Mohamed, N., and Al-Jaroodi, J. (2015). Applications of big data to smart cities. Journal of Internet Services and Applications, 6(1), 1-15.

[3] Warren, E. (2016) Strengthening research through data sharing. N. Engl. J. Med., 375375, 401-403.

[4] Lengauer, C., Bouge, L., Silva, F., Li, Z., Li, K., Gesing, S., and WilkinsDiehr, N. (2015). Araport: an application platform for data discovery. Concurrency and Computation: Practice & Experience, 27(16), 4412-4422.

[5] Jones, D., Ferris, K., Muellerleile, J., Hyatt, R. (2009). Generic materials property data storage and retrieval for the semiconducting materials knowledge base. Proceedings of SPIE, 7449(1), 74491R-74491R-10.

[6] Simonis (2019). Quoting and Billing: Commercialization of Big Data Analytics. In: Proceedings of Big Data from Space, BIDS'2019.

[7] Lynnes, Christopher and Ramachandran, Rahul (2018). Generalizing a Data Analysis Pipeline in the Cloud to Handle Diverse Use Cases in NASA's EOSDIS. In: Proceedings of IGARSS'2018, 422-425, doi = 10.1109/IGARSS.2018.8519178

[8] Simonis, I. (2019). OGC Testbed-15 Technical Architecture. Open Geospatial Consortium.

[9] Simonis, I. (2018). Container-based Architecture to Optimize the Integration of Microservices into Cloud-Based Data-Intensive Application Scenarios. In: Proceedings of ECSA18, September 2428, 2018, Madrid, Spain

[10] OGC (2019). OGC Testbed-14: ADES & EMS Results and Best Practices Engineering Report. OGC document 18-050r1. Open Geospatial Consortium

[11] OGC (2019). OGC Testbed-14: Application Package ER. OGC document 18-049r1. Open Geospatial Consortium

[12] OGC (2019). OGC Testbed-14: Authorisation Authentication and Billing ER. OGC document 18-057. Open Geospatial Consortium