Deliverable

D2.11 Technical solutions on open, dynamic, high volume, cloud-based services

Deliverable information	
Work package	WP2
Lead	GFZ, INGV, KNMI on behalf of ORFEUS
Authors	Reinoud Sleeman, Javier Quinteros, Peter Danecek, Alberto Michelini, Paride Legovini, Luca Trani, Carlo Cauzzi
Reviewers	Ian Main as WP2 Leader
Approval	[Management Board]
Status	Final
Dissemination level	[Public]
Delivery deadline	[31.08.2021]
Submission date	[17.08.2021] – first submission; [19.08.2021] – resubmission after minor revisions
Intranet path	[DOCUMENTS/DELIVERABLES/File Name]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821115

Table of contents

Summary

One of the key challenges for seismological data centers is to be prepared for handling the increasing growth of high-volume data in an interconnected, distributed system. Current emerging techniques, methods and technologies in monitoring, data acquisition and processing have already started to pose new technical and organisational challenges to the seismological infrastructure in terms of a) data collection and storage, b) providing services for transparent and rapid access, c) efficient processing of huge amounts of data and d) quality assurance of data and services. In this Deliverable we selectively describe current state-of-the-art technical solutions to rapidly serve, access and process massive seismic datasets, including the current strategies provided by the European Integrated Data Archive (EIDA; http://www.orfeus-eu.org/data/eida/), the recommendations compiled by the EPOS-ORFEUS Competence Center (CC) within project EOSC-Hub (https://www.eosc-hub.eu/), emerging challenges to handle new exotic datasets like those generated by distributed acoustic sensing (DAS) systems, and initial experiences gained into Cloud services and distributed computing environments for data processing and interactive exploration at ORFEUS associated data centers.

1. The European Integrated data Archive EIDA: current strategy and lessons learned in project EOSC-Hub

The European Integrated Data Archive (EIDA; http://www.orfeus-eu.org/eida; Strollo et al., 2021) provides access to the seismic waveform data collected by seismological monitoring and academic agencies in the greater European region. EIDA is a core component of ORFEUS (Observatories and Research Facilities for European Seismology; http://www.orfeus-eu.org), the research infrastructure for seismological waveform data in Europe. ORFEUS services are integrated in the European Plate Observing System EPOS (https://www.epos-eu.org/) and its data portal (https://www.ics-c.epos-eu.org/). ORFEUS is a core founder of the EPOS Thematic Core Service for Seismology.

EIDA currently comprises 12 European data archives, called EIDA nodes (Figure 1), that are federated on both organisational and technical levels. Each node collects and disseminates seismic waveform (meta)data at national / regional level and provides access to seismic data through standard services, making available data from over 13,000 seismic stations (sometimes co-located with accelerometers, pressure sensors, and other sensors) from permanent and temporary networks. Access to data and products is *via* state-of-the-art information and communication technologies, with strong emphasis on federated web services that considerably improve seamless and automated user access to data. The five main webservices (fdsnws-dataselect, fdsnws-station, eidaws-wfcatalog, eidaws-routing, eidaws-federator; https://orfeus-eu.org/data/eida/webservices/) in EIDA are compliant with EPOS, implying they are successfully integrated in the EPOS Integrated Core Services (ICS) portal.

The EIDA infrastructure has been designed to scale up with data holdings, services and user requests since the inception in 2013. **With the growing demand for large volumes of data, the implementation of distributed archives has become a very attractive solution to minimize failures related to single access points. EIDA demonstrates that a federated approach is a viable solution to serve large amounts of data to the research community.** The present EIDA system is a modular scalable infrastructure based on standard interfaces. EIDA continues to grow in terms of archive volume, number of nodes, new exotic massive datasets, and an increasing need to have rapid access to massive data volumes is emerging. To prepare for these challenges, EIDA data centres are collaborating with the Incorporated Research Institutions for Seismology (IRIS) Data Management Center (https://ds.iris.edu/ds/nodes/dmc/) to prepare data management concepts for emerging technologies such as DAS, massive numbers of cheap sensors, and microsensors for the Internet of Things. On a global scale ORFEUS and the EIDA nodes are also involved in the FDSN (Federation of Digital Seismograph Networks; https://www.fdsn.org/) in efforts to a) standardize the developments of new services and b) in

the design of new data formats for other types of data (e.g., large-N deployments including cheap sensors, and DAS) and their data management (note that one single DAS experiment can generate high-sampling-rate data equivalent to those of up to 10^9 traditional seismic sensors). This **FDSN coordination is crucial to expand the success of the seismological community on a global scale in terms of standardisation and cooperation**. Standards approved by the global community within the FDSN allow seismological data centers to support the FAIR data principles (Findable, Accessible, Interoperable, and Repeatable; Wilkinson et al., 2016) and to properly manage acknowledgement of networks through the use of Digital Object Identifiers (DOIs) to identify seismic networks.

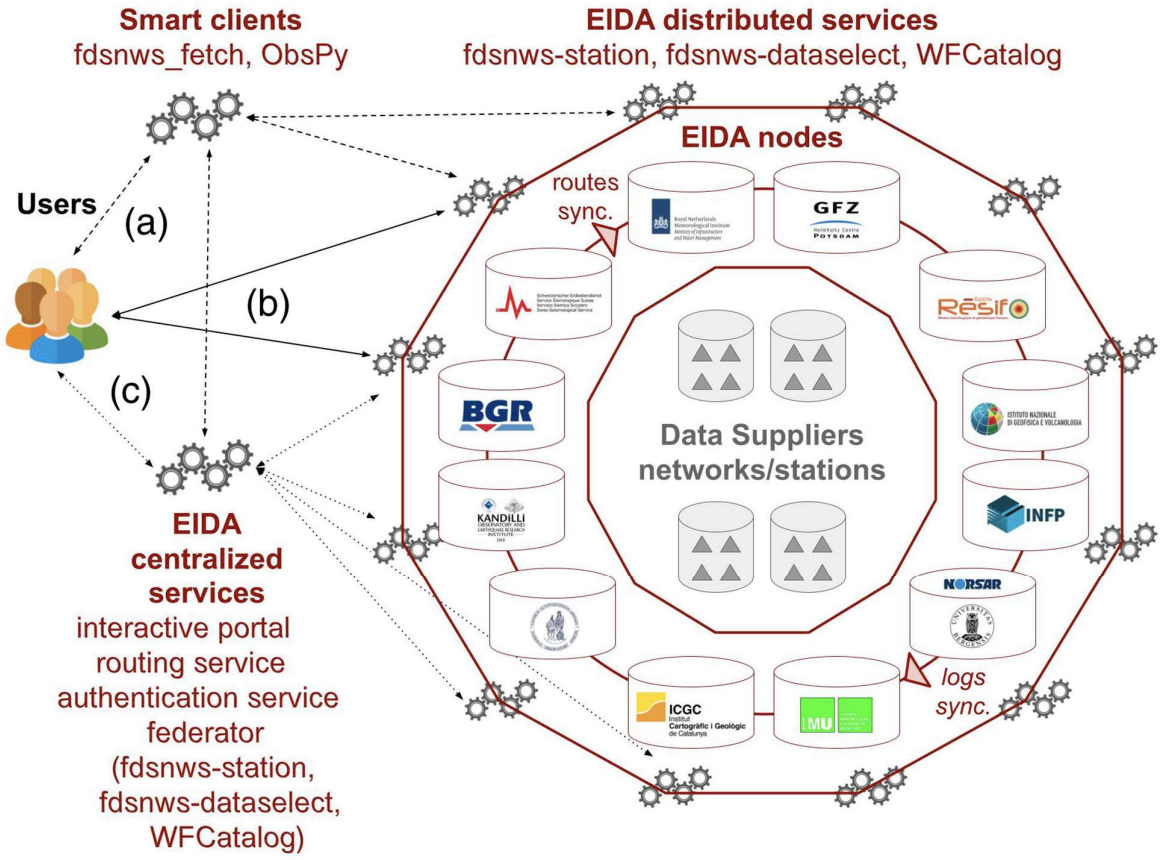

Figure 1 - Schematic view of the EIDA federated infrastructure and user's workflow (after Strollo et al., 2021). The data suppliers provide networks and stations distributed via the national or regional EIDA nodes. There are three way to access the federated archives: (a) the user sends a data request via smart client that will get routes from the central Routing Service then contact the necessary nodes and provide data back to the user (dashed line); (b) the user requests data directly to the nodes (solid line); (c) the user requests data via a centralized service that will act as a proxy requesting the actual data to the nodes and provide them to the user (dotted line).

The broader seismological community of EPOS was represented by ORFEUS in the construction of the European Open Science Cloud (EOSC) with the aim to explore sustainable solutions for the aforementioned challenges. The H2020 funded project EOSC-Hub allowed a few selected EIDA data centers (GFZ, INGV, ODC/KNMI, NOA) to perform a pilot study for building a Competence Centre (CC) for solid earth in EPOS, named EPOS-ORFEUS-CC. The aim of this pilot study was to foster the provisioning of data and services and their level of FAIRness. Within the pilot study EPOS-ORFEUS-CC explored the feasibility of creating a platform for seismological services through the integration of existing domain-specific infrastructure components with components offered by EOSC-Hub (www.eosc-hub.eu/catalogue), e.g., EGI, EUDAT, GÉANT. One of the outcomes of this project was the development of an Authentication/Authorization system (AAI) for EIDA, which is currently in production and that is being successfully adopted by the community

(https://geofon.gfz-potsdam.de/eas/). Other lessons learned from this project were valuable as input for the preliminary design of the EPOS ICS-D (Distributed Integrated Core Services) and also strengthened the synergy between EPOS and EOSC.

The pilot study the EPOS-ORFEUS-CC provided recommendations to:

1. **focus on distributed data storage on a Cloud[1];**
2. **use advanced community standard APIs (Application Programming Interfaces), based on the HTTP protocol[2].**

Current developments in EIDA are in-line with the recommendations from the EPOS-ORFEUS-CC. Alternative technologies like the distributed computing environments Jupyter and HPC Cloud may help to solve the scaling problems. However, these are more expensive and less suited to support the integration of the current distributed data services. The integration of such distributed computing environments is still not so straightforward within the current EOSC architecture, as the interoperability of the different deployments is not fully achieved: for instance, due to different implementations of Jupyter Notebooks; or similar, but configured in a way that cannot be integrated with other services like AAI, the storage system, or the processing resources available.

2. Big-Data management challenges identified by ORFEUS (GFZ, RESIF) & IRIS data centers

Current FDSN specifications regarding data formats and services were designed for different operational limitations and usage than the seismological data centers are facing now. With new and emerging technologies in data acquisition ahead, the increase in data will be huge and technical problems and data management challenges are expected. Quinteros et al. (2021) described these challenges and investigate solutions for data acquisition, archival, metadata, distribution/services and usage.

Data acquisition

For traditional seismic deployments, data are collected in the field (mainly for temporary deployments) or transmitted to the data centre (mainly permanent sites) for example though the SeedLink protocol originally developed for the SeisComP package (http://ds.iris.edu/ds/nodes/dmc/services/seedlink/). Those data are usually available directly as mini-SEED files, or require a relatively easy conversion from a local format into mini-SEED (http://www.fdsn.org/pdf/SEEDManual_V2.4.pdf). In the cases of large-N or DAS experiments, the volume of data acquired could be hundreds of terabytes in a relatively short time. For such deployment, real-time transmission of the data may not be practical, or even impossible. **One option is to process or decimate the data before real-time transmission and archive the raw data at a later stage. Otherwise, local storage would be the preferred option and handling to a datacentre must take place at a later stage through a physical device**.

Data storage

Today's FDSN standard format for seismic waveforms is miniSEED. The *de facto* standard procedure is to store data per sensor component in one file of approximately 24 hours. **For medium**

[1] The term "Cloud services" refers to a wide range of on-line services that are delivered on demand by third-party providers/vendors to customers through the Internet. Cloud services comprise infrastructure, platforms and software. Current public Cloud providers are for example Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, Oracle Cloud and Alibaba Cloud.

[2] These APIs – that include, e.g., standard web services - are well integrated and widely used by the community (FDSN, EPOS-ORFEUS infrastructures). They remain the preferred way to serve the community.

and large volume datasets, the HDF5 format is commonly used, in particular in the HPC[3] environment. PH5 (https://zenodo.org/record/1284569) and ASDF (Krischer et al., 2016) are examples of this. A redesign of PH5 is expected in 2021 and could be of use for large-N and DAS data. Currently, **the main disadvantage of HDF5 is that it cannot be streamed on the fly. Possibly the THREDDS data server** (https://www.unidata.ucar.edu/software/tds/) **or Highly Scalable Data Service (HSDS) distribution services could develop into that direction.** The challenge is to optimize the trade-offs between data compression, rapid data extraction, and seamless streaming. An overview of these and other formats is given in Quinteros et al. (2021). The miniSEED format will have its next generation in order to overcome the difficulties of the miniSEED v2 format (i.e., it was designed for independent, single time series, usually very small chunk sizes, and not for processing). It is expected that a proposal will be made to the FDSN later this year. At the same time, **new emerging formats like Zarr** (https://zarr.readthedocs.io), **that allows dynamic data chunking, may be of interest** although this has not been tested in practice by the community.

The increase of data also asks for new strategies towards the choice of physical storage, and the policies for access. For example, **data that are rarely accessed, can be kept in slower, cheaper media, while data that are requested frequently on media with rapid access. Managing how data are tiered ideally would be dynamic, depending on past access patterns. Public Cloud services** like Amazon Web Services (AWS) and Google Cloud Platform (GCP) **are an emerging alternative** for on-premises storage and are being tested by some data centers (e.g., the ORFEUS Data Center ODC, the Southern California Earthquake Center; Yu et al., 2021). Usually, their scale of operation exceeds the scale that can be handled by individual organizations. Using public Cloud providers is typically associated with the following advantages: a) the on-demand scalability of the infrastructure, b) a decrease in costs for maintenance and investments, and c) the flexibility to start or to shutdown services on-the-fly. In contrast, increased operational costs, the need for new knowledge and the dependency on the Cloud platform may hinder the usage of such public Cloud services. Care also must be taken to minimise the use of platform-specific code and to have an exit strategy. Also, a user survey conducted recently (Quinteros et al., 2021) showed that **data creators and big-data users do not consider this a feasible solution, mainly because of the hidden costs** considering read/write operations even for users. **Overall, the suitability of using public Cloud services depends on factors like a) scope and requirements (e.g., amount of data, reliability of service), b) overall benefits (not just the infrastructure, but also ways of working like DevOps[4]); c) risks and legal aspects (e.g., costs, lock-in dependency, GDPR, open data policies); and d) trends in IT (moving to public Cloud services seems to be the future anyway).**

The choice of storage system may also be directed by the need of direct access for processing next to the existing services for data extraction. Direct access adds new challenges to both infrastructure and management of the processes related to response times, rates and permissions. Depending on the purpose of the services it makes sense that for a (distributed) archive the access for processing should not interfere with those for standard requests. **The use of High Performance Computing (HPC) Cloud systems that bring together data storage and processing services must be explored by our community** to assess the feasibility of such platforms for supporting new processing services.

[3] High Performance Computing (HPC) is the ability to process data and perform complex calculations at high speeds. HPC solutions have three main components: CPU, network and storage. In a high performance computing architecture, CPUs are networked together into a cluster. Software and algorithms are run simultaneously in the cluster that is networked to the data storage. Together, these components operate seamlessly to complete a diverse set of tasks. Cloud HPC is a relatively new technology that may be seen as a descendant of grid computing, in which the required high-speed Internet connection potentially may be a challenge.

[4] https://aws.amazon.com/devops/what-is-devops/

Metadata

The current metadata standard in seismology, StationXML (https://www.fdsn.org/xml/station/), is comprehensive and widely used for today's seismic data sets. However, new emerging data types such as DAS cannot be handled by StationXML easily (or at all) for many reasons. For instance, the concept of stream identification (by network, station, location and channel codes) is completely different, as well as the way to describe the instrumental response of the sensing devices. Defining and implementing DAS metadata, and identifying data itself, is an outstanding difficult challenge to be tackled by our community.

Data distribution

Current standard webservices used by the seismological community (e.g., https://www.orfeus-eu.org/data/eida/webservices/) are designed for relatively small requests of data (several MBs per request). The size of data that is required today most likely is much larger. Together with the potentially large datasets from large-N or DAS deployments, a growth towards requests for much larger datasets, or entire datasets, is expected. Distribution of huge amounts of data would be a challenge for the complete infrastructure (e.g., bandwidth, local storage, processing). Even reducing the amount of data, for example by smart selection procedures or slicing the requests in smaller parts, would still challenge the infrastructure. Moreover, the current technology (web services) is - although very successful for relatively small requests - less reliable for larger requests due to connection timeouts. **The IRIS's ROVER tool** (https://iris-edu.github.io/rover/) **solves the issue of dropped connections by using a smart client to manage the synchronous data transfer in miniSEED. The final dataset is then merged on the client side**. Although the client is rather complex, it has shown to be a robust and efficient technique to transfer large datasets by a single request. **An alternative may be the extension of the existing web services to allow asynchronous data transfer.**

Data usage

To provide services for access to and processing of large datasets **non-FDSN standards are already used by researchers**, making the need to deploy new services providing these formats eminent. Such **an extension of standardization in the FDSN would leverage the community efforts to develop new processing tools, dedicated to large datasets**. **Notable in this sense if the being developed software package "dastools"** (Quinteros et al., 2021) for automatic standardization of data, and semi-automatic standardization of metadata, mapping to current community standards), which contributes to the portfolio of RISE-associated products for handling large datasets. The challenge for the federated infrastructure of data archives is how to bring together the infrastructure to increase efficient access to data and processing (e.g., EOSC).

3. Cloud strategy at the ORFEUS Data Centre ODC

The ODC has been hosted at KNMI since 1993. The ODC operations and maintenance of services are undertaken by the Research and Development department of Seismology and Acoustics at KNMI (https://www.knmi.nl/research/seismology-acoustics#). Hence, the underlying IT infrastructure of ODC is integrated in the KNMI infrastructure. At KNMI/ODC a strategic choice has been made to use the public Cloud provider Amazon Web Services (AWS) to modernize the computer infrastructure in computing power, data storage and other services. The main reasons for this choice are the large growth of the number of services and the scalability offered by AWS. The increase of applications and services require stable platforms that can host containers (e.g., Dockers) and complex research environments. The large scale of operations by AWS enables an organization like KNMI to operate more efficiently when outsourcing these services than maintaining an in-house infrastructure. However, the choice of service provider for, and the implementation of a specific application often de-

pends on requirements on data storage, bandwidth, costs and performance, and are therefore always in consultation with (KNMI) specialists on IT architecture and services. **The leading direction at KNMI in the migration process for outsourcing the infrastructure is the 'Cloud first' principle**, i.e., adopting the latest Cloud-based technologies to cut back on costs, whilst delivering better service to the users. Currently, applications at KNMI/ODC are implemented at different platforms:

- **Government Data Center**: this is the central hub in the KNMI network infrastructure. It is maintained by Equinix with 2 facilities in Amsterdam. Together with the service organization SSC-Campus, Equinix offers an environment specifically for governmental services. Currently all seismological services at KNMI/ODC are operated here using virtual machines.

- **Public Cloud Amazon Web Services (AWS)**: AWS will host operational software and related data for KNMI. This Cloud service environment is geographically located in three data centers in Ireland. All seismological waveform data for ODC and KNMI are hosted at AWS. Amazon S3 (Simple Storage Service) is storage for the Internet and works with buckets and objects. A *bucket* being a container for objects, while an *object* is a file and any metadata that describes that file. Amazon's S3 API is the *de facto* standard for object storage APIs. The S3 API is an HTTP/S REST API where all operations are via HTTP PUT, POST, GET, DELETE, and HEAD requests. Beyond the basic object operations provided by S3, there are advanced APIs for versioning, multi-part upload and access control. KNMI/ODC adapted SeiscomP in order to directly use S3 API in the most optimal way to read an SDS structure hosted in an object storage facility. Applications run by KNMI R&DSA (department R&D Seismology and Acoustics https://www.knmi.nl/research/seismology-acoustics) and ODC are progressively being migrated to AWS. This process requires individuals with Cloud expertise and skills to transform applications and services to the AWS environment effectively. Examples of this migration process currently involve the automated publication of earthquake locations provided by SeisComP (SC), the application of a Coincidence Trigger algorithm for tectonic earthquakes in the Netherlands using ObsPy and Python (Figure 2), and an automated backup procedure using iRods to backup all SC SDS waveform data at SURF.

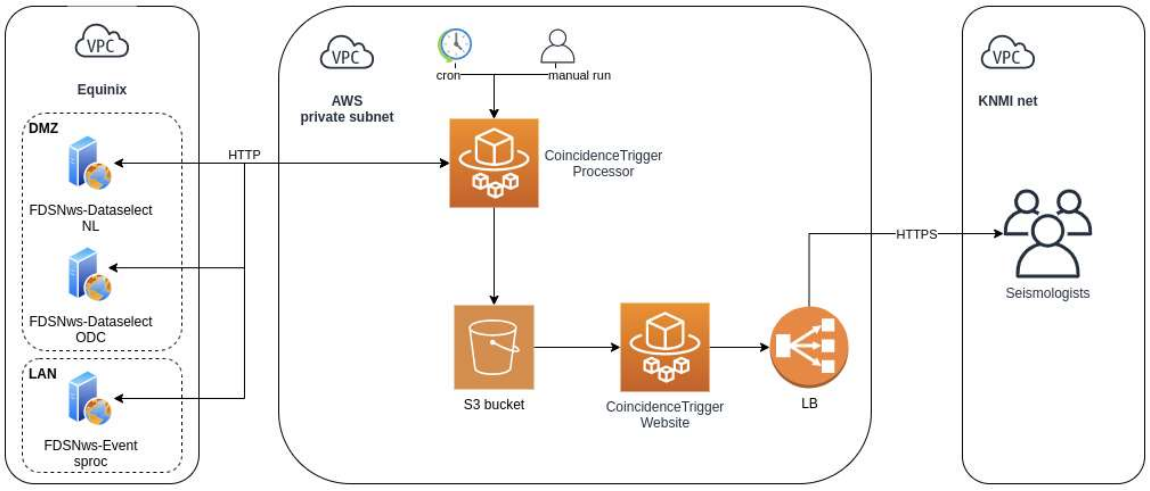

Figure 2 - High Level Diagram (HLD) of the implementation of ObsPy's Coincidence Trigger at KNMI using the AWS platform.

- **SURF**: a cooperative association of Dutch educational and research institutions. **SURF Research Cloud** (https://researchclouddocs.readthedocs.io/en/latest/about.html) provides a virtual research environment for direct access to computing and data services from SURF.

SURF typically combines computing power and data storage which makes it ideal for scientific purposes. As mentioned, seismic waveform data in SDS are being archived at AWS and being backed up automatically at SURF through iRODS.

4. Big data processing framework and interactive exploration: status and outlook of SeiSpark at INGV

INGV, a node of the EIDA federation and core participant of ORFEUS, currently holds more the 100 TB of seismological waveform data. The infrastructure for archiving and distribution of these data is well-established, functional and robust. However, it is getting increasingly more challenging to make good use of all this data assets, because the currently established workflows require users to download the dataset of interest (usually a rather small portion of the archive), and to process them on local resources. This approach reaches its limits where/when a more significant portion of the data holdings should be processed, both for the user and for the data center. The user would need to ensure adequate local resources for storage, I/O and computation, while the data center gets challenged by the immense amount of data requests for which the provided services and infrastructure were not constructed. **The SeiSpark project was started at INGV with the goal to** add a) significant computational resources and b) an adequate framework to the seismological data archive, **creating a "computational archive" where storage resources and computational resources converge. It should follow the paradigm of data locality, avoiding unnecessary data transfers on every level as much as possible.** The processing framework leverages on existing solutions from the Big Data ecosystem and combines them with the popular open-source scientific Python framework which is well established in seismological research. The key design criteria of SeiSpark are:

- **use of well-known (and modern) Python based seismological software** for processing, analysis and visualization;
- **leveraging on existing open-source solutions** to limit maintenance effort and facilitate long-term sustainability;
- enabling access (or pre-stage) to very large seismological waveform datasets, ideally a considerable portion of our archives, if not the whole data holdings of the INGV EIDA node;
- **scalability**: the framework must be future proof to cope with the aforementioned challenges and allow for scalability of the infrastructure, in-house or in the Cloud.

With this background, a pilot infrastructure implementation was carried out based on Apache Spark (https://spark.apache.org), a unified analytics engine for large-scale data processing, to a) evaluate the overall concept, b) verify the feasibility and efficiency of such an infrastructure, c) explore this potential, d) experiment with different software setups and e) explore various processing strategies and algorithms. It also served to guide the technological choices for the upcoming production setup of this infrastructure. In his final setup it implemented the following software stack:

- Apache Hadoop (https://hadoop.apache.org/);
- Apache HDFS (https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html), a persistent distributed location-aware file system;
- Apache YARN (https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html), a resource management, job scheduling and monitoring system;
- Apache Spark (https://spark.apache.org/), the principal analytics engine based on the Resilient Distributed Datasets (RDD) abstraction;
- a complete customised Python 3 (https://www.python.org/downloads/) environment;
- Jupyter (https://jupyter.org/), the interactive notebook environment;
- ObsPy (https://github.com/obspy/obspy/wiki), the seismological Python processing and analysis framework;
- various plotting and visualisation tools for interactive plots;

- JupyterHub (https://jupyter.org/hub);
- A multi-user environment.

A representative selection of waveform data from the seismological archive was replicated (staged) to the (computable) HDFS file system, as a persistent copy. We implemented reading access to native seismic waveform data, various basic access patterns and fundamental processing algorithms, in order to extract various information and to provide results in a concise and graphical format. Basic performance testing and evaluation was performed: **in spite of a startup overhead from the resource manager of up to two minutes, remarkably improved performance was observed with processing up to 12 times faster than traditional jobs on local disk**. This is expected to further improve with migration to modern hardware.

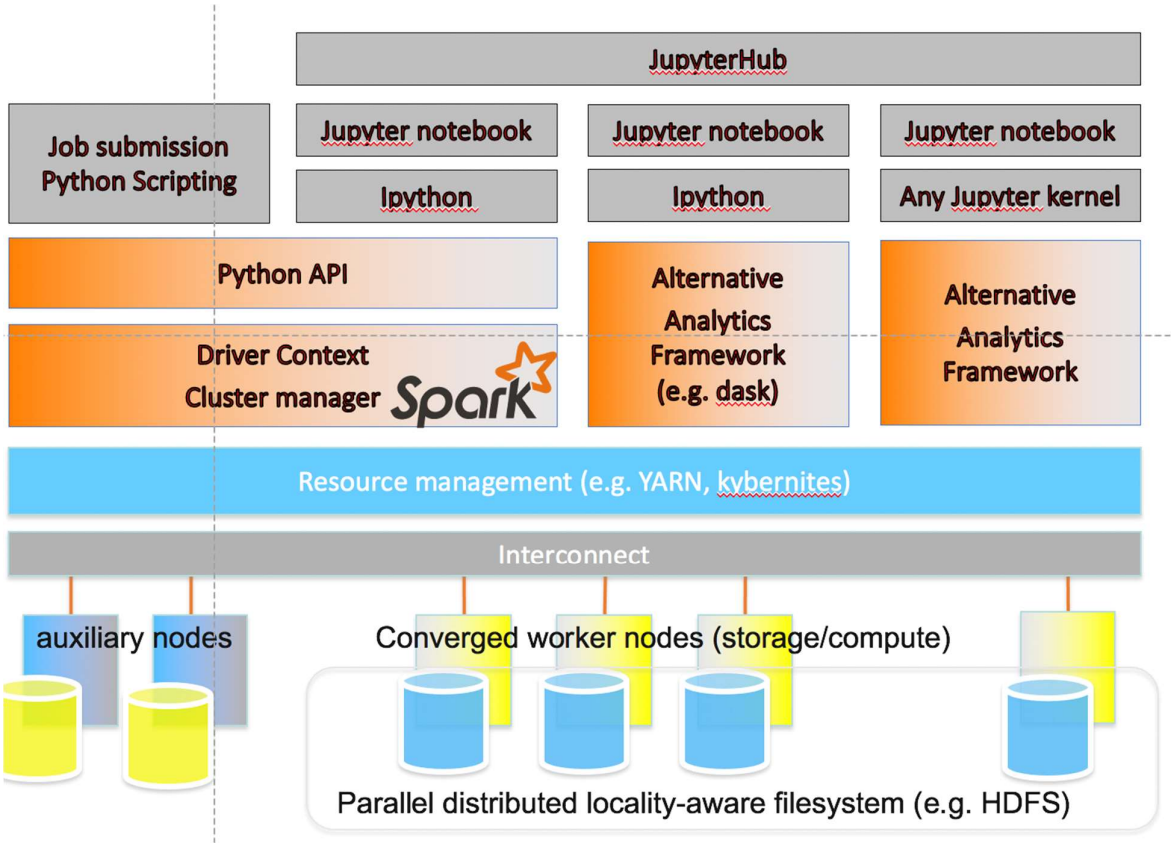

Figure 3 – SeiSpark concept.

Within the framework and timescale of project RISE, INGV plans to migrate the pilot infrastructure described above to a target infrastructure with the following hardware.

- Number of nodes:
  - 8 converged worker nodes +
  - 3 auxiliary nodes for resource management & associated tasks;

- High performance low latency interconnect: 100 Gbit/s
- Separated management network
- Total memory RAM aggregated across the worker nodes: 4 TB
- Large scale storage available across the converged worker nodes:
  - minimum 900 TB raw storage
  - minimum 300 TB effective usable for data storage

- Memory bandwidth aggregated across the cluster: minimum 3200 GB/s

- Number of cores per node: minimum 48

The processing platform will largely replicate the software stack which was developed and deployed on the pilot infrastructure, starting from its latest iteration. However, we plan to increase flexibility and manageability by exploiting recent technological developments in virtualisation, containerisation, supervisor and orchestration. In particular, a container-based deployment strategy and management of applications will facilitate the lifetime management of this relative dynamically developing field and also provide the opportunity to experiment with alternative existing and upcoming data analytics frameworks. This should ensure sustainability over the longer lifetime of the new storage-compute infrastructure.

Concerning the scaling strategy, the question is how a modest scale infrastructure could be embedded into a larger context (e.g., Cloud) and which role larger scale infrastructures should play. The concern is that any additional stage in operation is rather costly and adds to the overall processing time of the whole workflow. Also, keeping large datasets pre-staged in a persistent workspace would imply significant costs. The following strategies are being considered (see also previous Section2): (a) to keep only the presumably most relevant subset of data persistently pre-staged, e.g., most recent data, rotating time window, high quality stations only; (b) to stage in for a limited duration of a project only the specific subset of data; (c) to use the data-locale infrastructure to perform first data reducing processing steps and transfer the intermediate products.

5. Conclusions and outlook

The last decade had witnessed significant technological developments and new research practices that require adaptation and improvement in the way seismological data centers manage waveform data, metadata, and the associated services and products. Notable examples are: (i) the emerging use of Distributed Acoustic Sensing (DAS) systems for seismological studies; (ii) the increased availability of low-cost seismic sensors ranging from portable field instrumentation to MEMs accelerometers for structural monitoring; (iii) the initial experiences in using the seismic sensors embedded in smartphones and smart household appliances for earthquake early warning and rapid response applications. This means that massive datasets – possibly up to 10^6-10^9 times larger than those generated by traditional sensors – are soon being routinely produced, and that seismological data centers need to implement technical solutions for the management of such huge amount of data, from acquisition to access and usage by stakeholders. Users are expecting to be able to access and process the data efficiently, without having to download and store locally the datasets they are interested in. This prompts datacenters to provide new processing, possibly Cloud-based solutions, developed in close collaboration with High Performance Computing (HPC) facilities and scientific computing centers. Selected European datacenters, all contributing to the European Integrated Data Archive (EIDA) within ORFEUS (http://www.orfeus-eu.org/data/eida/), started to gain experience to address the above challenges within the EC-funded project EOSC-Hub (https://www.eosc-hub.eu/) and are consequently aligning their data management strategies to the recommendations compiled and the lessons learned therein. In particular, the ORFEUS Data Center in the Netherlands recently switched to a 'Cloud first' approach for data storage while retaining the use of standardised webservices as the preferred way to serve the users. Standardized webservices and traditional data management strategies are still suitable technical options to serve large-N (i.e., comprising a large number of sensors) datasets. New solutions are definitely needed for DAS experiments concerning metadata curation, data acquisition, archival, distribution and use (e.g., dastools; Quinteros, 2021), as extensively documented in a pilot study (Quinteros et al., 2021) conducted by ORFEUS/EIDA associated datacenters – GFZ & RESIF - and the IRIS Data Management Center. Spurred by the need to test a seismological "computational" archive where storage resources and computational resources converge, INGV - also a core node of the ORFEUS/EIDA infrastructure – has prototyped and will soon deliver a framework (SeiSpark) for Big Data processing and interactive explorations centered on the analytics engine Apache Spark. It is expected that experiences like SeiSpark will be replicated at scientific computing centers in

the future, yet a distributed infrastructure for combining service access and direct access to data for HPC purposes seems not yet ready. **All the aforementioned experiences, documented in the previous sections of this Deliverable, are representative of the state-of-the-art and possibly the avant-garde on the topic at hand. Crucial for successful future standardised developments and implementations is the involvement of and coordination with several additional datacenters worldwide, within the framework of the Federation of Digital Seismograph Networks** (FDSN; https://www.fdsn.org/), **as well as the encouragement of international collaborations among scientists, datacenter operators and managers** (e.g., https://www.erasmus.gr/UsersFiles/microsite1193/Documents/SESSIONS_ABSTRACTS3.pdf).

6. References

The references highlighted with bold font are an output of project RISE.

- Krischer, L., J. Smith, W. Lei, M. Lefebvre, Y. Ruan, E. Sales de Andrade, N. Podhorszki, E. Bozdağ, and J. Tromp (2016). An adaptable seismic data format, Geophys. J. Int. 207, no. 2, November 2016, 1003–1011, doi: 10.1093/gji/ggw319.

- **Quinteros, J. (2021). dastools - Tools to work with data generated by DAS systems, Potsdam: GFZ Data Services. doi:10.5880/GFZ.2.4.2021.001.**

- **Quinteros, J., J. A. Carter, J. Schaeffer, C. Trabant, and H. A. Pedersen (2021). Exploring Approaches for Large Data in Seismology: User and Data Repository Perspectives, Seismol. Res. Lett. 92, 1531–1540, doi: 10.1785/0220200390.**

- Strollo, A., D. Cambaz, J. Clinton, P. Danecek, C. P. Evangelidis, A. Marmureanu, L. Ottemöller, H. Pedersen, R. Sleeman, K. Stammler, et al. (2021). EIDA: The European Integrated Data Archive and Service Infrastructure within ORFEUS, Seismol. Res. Lett. 92, 1788–1795, doi: 10.1785/0220200413.

- Wilkinson, M., M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016). The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3, Article number: 160018, doi: 10.1038/sdata.2016.18.

- Yu, E., A. Bhaskaran, S.-L. Chen, Z. E. Ross, E. Hauksson, and R. W. Clayton (2021). Southern California Earthquake Data Now Available in the AWS Cloud, Seismol. Res. Lett., doi: 10.1785/0220210039.