1. Covid-X Sandbox Services Description (1st Release)

1.1 Data Integration and Management Services

The COVID-X Sandbox (CovidSandbox) introduces a combination of different functions that collectively aim to enable seamless access to a set of healthcare data sources. The Sandbox is capitalising on the ELK Stack[1] components to deliver data management, harmonization/cataloguing, storage, indexing, search, visualization, querying and retrieval services. It is based on a highly modular architecture, as depicted in Figure 1, and each component within the architecture realizes and delivers one or more of the aggregated services:

- **APIs:** This layer includes a set of technological tools which implement the process of bringing together healthcare data from various data sources and ingesting into the Covid-X Sandbox. Different tools can be used to collect the data from each data source, depending on the method that has been established in order to access the source system. An API Gateway tool is utilized to manage the exposed APIs on top of the microservices components, for use by third party solutions.
- **Security:** The security module is responsible for applying the mechanisms and policy rules to enforce data privacy and security, as described in section 1.2.
- **Data Lake:** A repository for storing pools of data "as is", if and when required, and for data that are not of high volume. This means that the data are stored in a schema-less manner, allowing the combination of structured and unstructured data. The technologies that will be used for implementing the data lake focus more on data transactions rather than relations between data.
- **Data Integration/Harmonization:** This component is responsible for performing Extract, Transform and Load (ETL) operations. This process includes pulling data that are either stored in the data lake or directly extracted via the API layer from the source systems, converting them into a consistent format in compliance to the Semantic Reference Model and then loading the integrated data and/or standardized annotations into a centralised target repository. In order to achieve that, this component provides the mappings between the data sources and the common schema/data model, based on which data will be curated and enriched in order to be harmonized and easily thus further managed.
- **Dara Storage/Indexing/Search:** This component can be described as a centralised repository or data catalogue for storing the curated metadata and, per case, data, if needed/allowed, as well as a search engine that enables complex and fast queries. It is

[1] https://www.elastic.co/what-is/elk-stack

used for indexing and cataloguing the integrated data. In addition to that, it offers an extensive set of APIs for querying and retrieving the data, that may reside also in other database systems.

- **Visualization:** Responsible for creating customizable and interactive dashboards that will provide visualization capabilities on top of the indexed content.

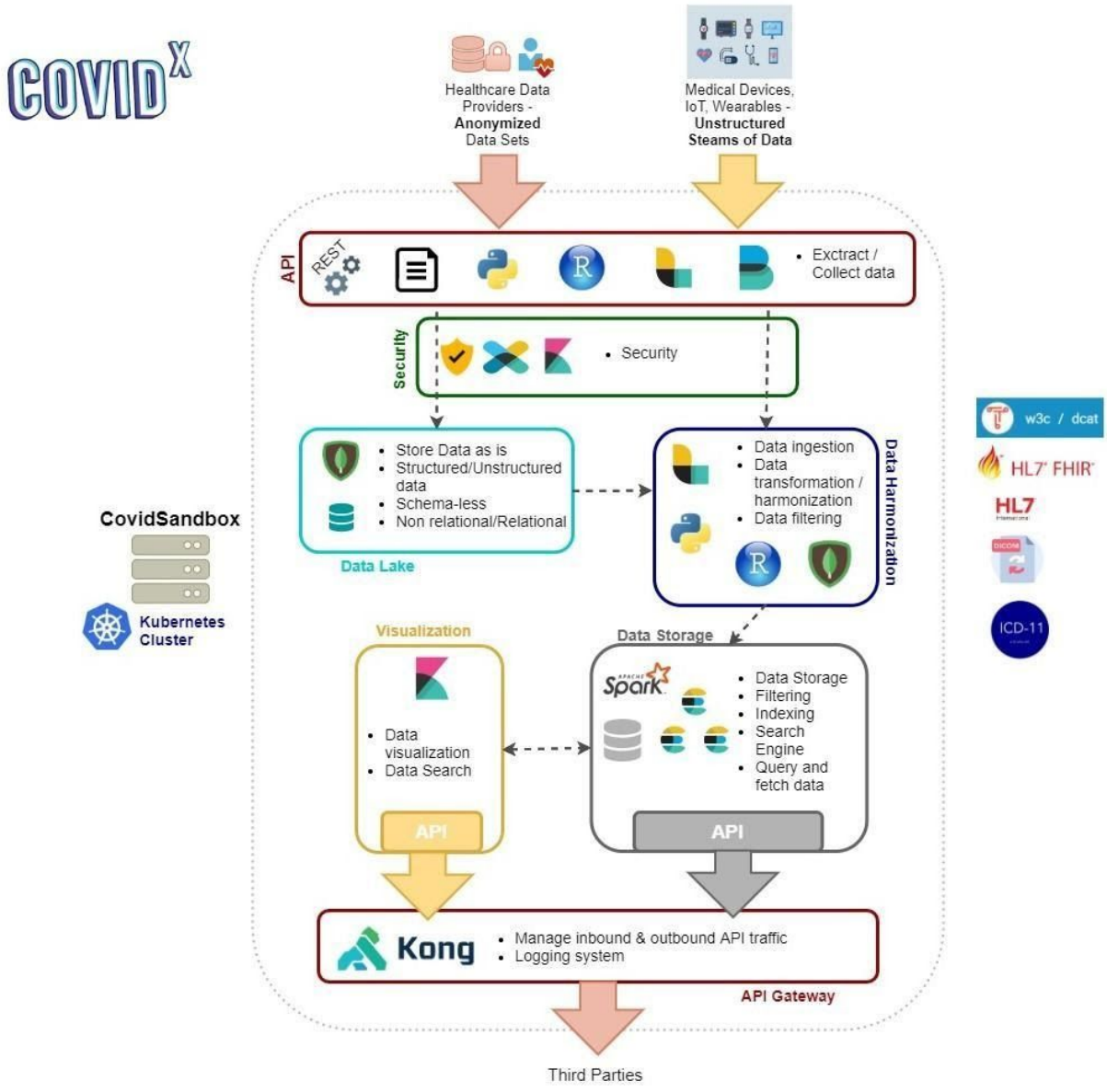

Figure 1: CovidSandbox architecture - 1st Release

1.2 Security Services of the COVID-X Sandbox

The Covid-X Sandbox tools and services adopt the three core security principles which are *confidentiality*, *integrity* and *availability*, allowing protection from multiple sources. Security measures will be implemented for every *Covid-X Sandbox* deployment, including the container and orchestrator environments. However, the implementation of all security measures will follow the same principles.

The main scope of the security services that will be offered by the *Covid-X Sandbox*, is the protection of medical data that will be ingested in it. Primarily, all data is required to be fully anonymized by the data owners/controllers prior to being ingested into the *Covid-X Sandbox*, according to the *Anonymization Guideline* proposed by The COVID-X project. Consequently, data *anonymization* constitutes the first step towards data *privacy* and *security*. The next steps that will ensure that data ingested in the *Covid-X Sandbox* will be fully protected are: secure *authentication* and *authorization mechanisms*, *Role Based Access Control (RBAC)*, *encrypted data transfer*, *monitoring* and *auditable trace of activities*. All of the aforementioned steps will be briefly explained.

Security steps

1. *Covid-X Sandbox* will be available to a limited and well-defined number of parties.
2. *Covid-X Sandbox* deployment will use specific *network protocols* to analyze its network *traffic*.
3. *Data transfer* and communication of the *Covid-X Sandbox* with external applications, devices and also between its own components will be secured by *encryption protocols* such as *HTTPs* and secure connections via *VPN* (Virtual Private Network).
4. *Covid-X Sandbox* will utilize virtualization tools which include isolation, encapsulation and partitioning properties and promote security. These will operate on certified resources and will be up-to-date, avoiding any vulnerabilities of older versions.
5. All activities will be recorded using *audit logging* tools which will assist in the traceability and accountability of user actions inside the *Covid-X Sandbox.* Each activity log will be processed by an *alert system* that detects abnormal behavior. This measure focuses on containing any possible insider threat. Security logs that contain sensitive information will be backed up. Additionally, customizable dashboards specifically used for *activity monitoring* for users and *Covid-X Sandbox* components will be supported.
6. A *novel SIEM tool* implemented by 8BELLS will be integrated in order to track abnormal behavior in the functionality of the proposed processes.

7. Moreover, access to the *Covid-X Sandbox* will be controlled with *authentication* and *authorization* tools that support identity validation, along with *access control lists*. More precisely, the System Owner will be responsible for creating *Covid-X Sandbox* users. Each user can be assigned with a role that entitles him/her with specific privileges and can perform the appropriate actions, therefore avoiding misuse of *Covid-X Sandbox* tools and data. Such role-based access control mechanisms will follow a minimum user access and functionalities policy, avoiding scenarios where users can escalate their rights inside the *Covid-X Sandbox* or consume excessive system resources.

2. APIs for third parties

The Third Party APIs, in particular, will be focused on two main operations: data integration and data accessing, including data visualization through Kibana. A complete API REST specification will be provided once the platform matures and is finally deployed. This specification will be made by means of a specific swagger that will include the related documentation and test-case generation. The details of these elements can be seen in the following sections.

2.1 Data ingestion API REST for data providers

The initial sample API specification provided below will cover the CRUD (create, read, update and delete) operations in the form of POST, GET, PUT and DELETE calls and will include a data provider identification.

Endpoint	Description
POST/{dataprovider}/datatype	This call will enable the creation of a new type of data to be filled with the info to be ingested, according to the elastic search structure. The allowed types are: - Plain text: txt, pdf. - Structured/semi-structured data: json, csv, xls, xml, html. Image data: jpeg, png, nifty, CR, DX, CT (the info to be ingested will be the metadata and the url of the source)
GET/{dataprovider}/datatype	This call will provide the list of data types from a specific data provider
GET/datatype/{id}	This call will allow the access to specific data by the id
PUT/{dataprovider}/datatype/{id}	This call will allow the modification of the data type given by its id
DELETE/datatype/{id}	This call will allow the removal of specific data by the id

2.2 Data access API for third parties

In a similar manner, the purpose of the API described below will be to allow third parties to access and visualize the authorized data for their testbeds.

Endpoint	Description

GET/{thirdparty}	This call will list the data each third party is allowed to access, according to the elastic search structure
GET/{thirdparty}/{data_id}	This call will allow each third party to access specific data by the id. The data available to be accessed can be seen in Section 4.
GET/kibana/{thirdparty}/dashboard	This call will allow to access the Kibana dashboard for the third party allowed data
GET/kibana/{thirdparty}/dataaccess	This call will allow to use the Kibana API for data monitoring and information accessing

3. Local deployment guidelines and integration/testing approach

3.1 Sandbox Deployment

COVID-X Sandbox supports a centralized cloud-based deployment, capitalizing on virtualization and orchestration technologies. In addition to that, a local deployment on premises of the data providers/clinical partners can be adopted as a solution in order to tackle the closed anonymized data access/distribution issues and ensure GDPR and national regulations compliance. In this deployment scenario, data providers must provide a local infrastructure that will be able to host a local instance of the Sandbox.

3.2 CI/CD Stack Description

COVID-X Sandbox offers a Continuous Integration (CI) and Continuous Deployment (CD) stack that provides the technical support to third parties to integrate the different solutions into the Sandbox and perform system testing activities. The CI process will enable the extensive testing of each software component/module, as well as the integrated Sandbox platform as a whole. Automated, per component tests and combined integration tests performed on each new build (version) of a component shall be executed and upon success, the CD process will update the integrated platform by means of deployment of the new versions of the modules.

The Sandbox CI/CD stack is implemented as a collection of open-source tools. A Source Code Management component, such as Gitlab, will manage the software version control and the development process, acting as a source code repository. The Test-Driven Development approach in an automated manner will be supported by a CI server, responsible for pulling the source code, which is then compiled, built and tested. If the unit and integration tests are successful, the produced component artifacts are stored in a centralized registry. From there, they can be pulled and deployed on a server via a script which automates the entire process.

4. Available Data sets and Variables

Table: High-level clinical data characteristics

	SERMAS	ICH	KI

Hospitalization			
Administrative data (length-of-stay, Service, internal transfers, cause of discharge…)	X	X	X
Diagnosis and procedures at discharge	ICD10	ICD9	X
Treatment prescribed and administered during admission	X	X	X
Labwork (complete blood cell count, chemistry, coagulation testing, D-Dimer, IL-6…)	Non protocolized, performed by medical decision based on the clinical manifestations	X	X
Imaging test (chest X-ray, CT scan…)	Non protocolized, performed by medical decision based on the clinical manifestations	Non protocolized, performed by medical decision based on the clinical manifestations	
Clinical notes	Unannotated in Spanish	Unannotated in Italian	
Emergency Department			
Administrative data (length-of-stay, Service, internal transfers, cause of discharge…)	X	X	
COVID-19 related clinical manifestations	X	X	
Clinical scores (SOFA, NEWS…)	X		
Vital Signs (blood pressure, temperature…)	X	X	X
Diagnosis and procedures at discharge	X	X	
Treatment prescribed and administered during admission	X	X	X
Labwork (complete blood cell count, chemistry, coagulation testing, D-Dimer, IL-6…)	Non protocolized, performed by medical decision	X	X

	based on the clinical manifestations		
Imaging test (chest X-ray, CT scan…)	Non protocolized, performed by medical decision based on the clinical manifestations	CT at admission in the emergency department. In addition, other imagine test non protocolized, performed by medical decision based on the clinical manifestations	X
Clinical notes	Unannotated in Spanish	Unannotated in Italian	
Intensive Care Unit			
Administrative data (length-of-stay, Service, internal transfers, cause of discharge…)	X	X	X
Clinical scores (SOFA, NEWS…)	X		
Vital Signs (blood pressure, temperature…)	X		
Treatment prescribed and administered during admission	X	X	X
Labwork (complete blood cell count, chemistry, coagulation testing, D-Dimer, IL-6…)	Non protocolized, performed by medical decision based on the clinical manifestations	X	X
Imaging test (chest X-ray, CT scan…)	Non protocolized, performed by medical decision based on the clinical manifestations	Non protocolized, performed by medical decision based on the clinical manifestations	
Clinical notes	Unannotated in Spanish	Unannotated in Italian	