SecureCloud

Joint EU-Brazil Research and Innovation Action
SECURE BIG DATA PROCESSING IN UNTRUSTED CLOUDS

http://www.securecloudproject.eu/

Data Management Plan
D6.2

Due date: 30 06 2016
Submission date: 09 06 2016

Start date of project: 1 January 2016

Document type: Deliverable
Work package: WP6

Editor: Giovanni Mazzeo, Francesco Tessitore (SYNC)

Contributing partners: All

Reviewers: André Martin (TUD)
Rodrigo Jardim Riella (LACTEC)

Dissemination Level

PU	Public	√
CO	Confidential, only for members of the consortium (including the Commission Services)	
CI	Classified, as referred to in Commission Decision 2001/844/EC	

SecureCloud has received funding from the European Union's Horizon 2020 research and innovation programme and was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under grant agreement No 690111.

Revision history:

Version	Date	Authors	Institution	Description
0.1	2016/05/20	Giovanni Mazzeo	SYNC	Initial Version
0.2	2016/05/24	Francesco Tessitore	SYNC	Second Version
0.4	2016/05/28	Michal Fischer	IEC	Inputs Provided
0.5	2016/05/29	Marcio Hamerschmidt	COPEL	Inputs Provided
1.0	2016/06/09	Luigi Romano	SYNC	Final Version

Tasks related to this deliverable:

Task No.	Task description	Partners involved°
T6.2	Data Management Plan	SYNC,TUD

°This task list may not be equivalent to the list of partners contributing as authors to the deliverable
*Task leader

Executive Summary

This deliverable provides the SecureCloud data management plan version 1. The Deliverable outlines how the research data collected or generated will be handled during and after the SecureCloud action, describes which standards and methodology for data collection and generation will be followed, and whether and how data will be shared. This document follows the template provided by the European Commission in the Participant Portal.

Contents

1 Building a DMP in the Context Of H2020

1.1 Purpose of the SecureCloud Data Management Plan (DMP)

SecureCloud is a Horizon 2020 project participating in the Open Research Data Pilot. This pilot is part of the Open Access to Scientific Publications and Research Data programme in H2020 [1]. The goal of the program is to foster access to data generated in H2020 projects. Open Access refers to a practice of giving online access to all scholarly disciplines information that is free of charge to the end-user. In this way data becomes re-usable, and the benefit of public investment in the research will be improved. The EC provided a document with guidelines for projects participants in the pilot. The guidelines address aspects like research data quality, sharing and security. According to the guidelines, projects participating will need to develop a DMP. The DMP describes the types of data that will be generated or gathered during the project, the standards that will be used, the ways how the data will be exploited and shared for verification or reuse, and how the data will be preserved. This document has been produced following these guidelines and aims to provide a consolidated plan for SecureCloud partners in the data management plan policy that the project will follow. The document is the first version of the DMP, delivered in M6 of the project.

1.2 Background Of The SecureCloud DMP

The SecureCloud DMP is written in reference to the Article 29.3 in the Model Grant Agreement called Open access to research data (research data management). Project participants must deposit their data in a research data repository and take measures to make the data available to third parties. The third parties should be able to access, mine, exploit, reproduce and disseminate the data. This should also help to validate the results presented in scientific publications. In addition Article 29.3 suggests that participants will have to provide information, via the repository, about tools and instruments needed for the validation of project outcomes. The DMP will be important for tracking all data produced during the SecureCloud project. Article 29 [2] states that project beneficiaries do not have to ensure access to parts of research data if such access would be lead to a risk for the projects goals. In such cases, the DMP must contain the reasons for not providing access.

2 SecureCloud Data Context

2.1 The Smart Grid Use Case

Traditionally, the term ”grid” is used for an electricity system that may support all or some of the following four operations: electricity generation, electricity transmission, electricity distribution, and electricity control. the SmartGrid uses two-way flows of electricity and information to create an automated and distributed advanced energy delivery network.

The SmartGrid can be regarded as an electric system that uses information, two-way, cyber-secure communication technologies, and computational intelligence in an integrated fashion across electricity generation, transmission, substations, distribution and consumption to achieve a system that is clean, safe, secure, reliable, resilient, efficient, and sustainable.

This description covers the entire spectrum of the energy system from the generation to the end points of consumption of the electricity. The ultimate SmartGrid is a vision. It is a loose integration of complementary components, subsystems, functions, and services under the pervasive control of highly intelligent management-and-control systems.

However smart technologies improve the observability and/or the controllability of the power system. Thereby Smart Grid technologies help to convert the power grid from a static infrastructure to be operated as designed, to a flexible, living infrastructure operated proactively. The Smart Grid is integrating the electrical and information technologies in between any point of generation and any point of consumption. Examples:

- Smart metering could significantly improve knowledge of what is happening in the distribution grid, which nowadays is operated rather blindly.
- The controllability of the distribution grid is improved by load control and automated distribution switches.
- Common to most of the Smart Grid technologies is an increased use of communication and IT technologies, including an increased interaction and integration of formerly separated systems.

European Technology Platform Smart Grid defines smart grid as follows: A SmartGrid is an electricity network that can intelligently integrate the actions of all users connected to it generators, consumers and those that do both in order to efficiently deliver sustainable, economic and secure electricity supplies. A SmartGrid employs innovative products and services together with intelligent monitoring, control, communication, and self-healing technologies to:

- better facilitate the connection and operation of generators of all sizes and technologies;
- allow consumers to play a part in optimizing the operation of the system;
- provide consumers with greater information and choice of supply;
- significantly reduce the environmental impact of the whole electricity supply system;
- deliver enhanced levels of reliability and security of supply.

Smart Grid deployment must include not only technology, market and commercial considerations, environmental impact, regulatory framework, standardization usage, ICT (Information and Communication Technology) and migration strategy but also societal requirements and governmental edicts.

2.1.1 Existing Standards

The IEC 62357 Reference Architecture [?] addresses the communication requirements of the application in the power utility domain. Its scope is the convergence of data models, services and protocols for efficient and future-proof system integration for all applications. This framework comprises communication standards including semantic data models, services and protocols for the abovementioned intersystem and subsystem.

ABNT NBR 14522

The Brazilian Standard is ABNT NBR 14522, Data Exchange for electricity metering systems. It defines the standard for the exchange of information in the electricity metering system in order to achieve compatibility between systems and electricity metering equipment from different sources.

This pattern consists of the following items:

- conventional communication reader-meter,
- directional communication reader-meter,
- synchronous remote communication,
- user exits,
- communication-computer reader,
- public format,
- expanded public format,
- FK7 format,
- operational program format.

Load format parameters:

- 1/2 magnetic tape format,
- display codes,
- standardized readings.

2.1.2 Data Model

The Power supply companies today face the urgent task of optimizing their core processes. This is the only way that they can survive in this competitive environment. The vital step here is to combine the large number of autonomous IT systems into a homogeneous IT landscape. However, conventional network control systems can only be integrated with considerable effort because they do not use uniform data standards. Network control systems with a standardized data format for source data based on the standardized Common Information Model (CIM), in accordance with IEC 61970, offer the best basis for IT integration. The CIM defines a common language and data modeling with the objective of simplifying the exchange of information between the participating systems and applications via direct interfaces. The CIM was adopted by IEC TC 57 and fast-tracked for international standardization. The standardized CIM data model offers a very large number of advantages for power suppliers and manufacturers:

- Simple data exchange for companies that are near each other.
- Standardized CIM data remains stable, and data model expansions are simple to implement. It results to be simpler, faster and less risky upgrading the energy management systems.
- The CIM application program interface creates an open application interface. The aim is to use this to interconnect the application packages of all kinds of different suppliers using Plug and Play to create an EMS.

The CIM forms the basis for the definition of important standard interfaces to other IT systems. The working group in IEC TC 57 plays a leading role in the further development and international standardization of IEC 61970 and the CIM. Working group WG14 (IEC 61968 Standards) in the TC 57 is responsible for standardization of interfaces between systems, especially for the power distribution area. Standardization in the outstation area is defined in IEC 61850. With the extension of document IEC 61850 for communication to the control centre, there are overlaps in the object model between IEC 61970 and IEC 61850. The CIM data model describes the electrical network, the connected electrical components, the additional elements and the data needed for network operation as well as the relations between these elements. The Unified Modeling Language (UML), a standardized, objectoriented method that is supported by various software tools, is used as the descriptive language. CIM is used primarily to define a common language for exchanging information via direct interfaces or an integration bus and for accessing data from various sources. The CIM model is subdivided into packages such as basic elements, topology, generation, load model, measurement values and protection. The sole purpose of these packages is to make the model more transparent. Relations between classes may extend beyond the boundaries of packages.

The ABNT NBR 14522 data definition can be found in deliverable 5.1.

2.1.3 Protocols

Communication technology has continued to develop rapidly over the past few years and the TCP/IP protocol has also become the established network protocol standard in the power supply sector. The modern communication standards as part of the IEC 62357 reference architecture (e.g. IEC 61850) are based on TCP/IP and provide full technological benefits for the user. The protocol used by Copel is defined in ABNT NBR 14522.

IEC 61850 Communication networks and systems in substations

Since being published in 2004, the IEC 61850 communication standard has gained more and more relevance in the field of substation automation. It provides an effective response to the needs of the open, deregulated energy market, which requires both reliable networks and extremely flexible technology flexible enough to adapt to the substation challenges of the next twenty years. IEC 61850 has not only taken over the drive of the communication technology of the office networking sector, but it has also adopted the best possible protocols and configurations for high functionality and reliable data transmission. Industrial Ethernet, which has been hardened for substation purposes and provides a speed of 100 Mbit/s, offers enough bandwidth to ensure reliable information exchange between IEDs (Intelligent Electronic Devices), as well as reliable communication from an IED to a substation controller. The definition of an effective process bus offers a standardized way to digitally connect conventional as well as intelligent CTs and VTs to relays. More than just a protocol, IEC 61850 also provides benefits in the areas of engineering and maintenance, especially with respect to combining devices from different vendors.

Telecontrol IEC 60870-5

IEC 60870-5 provides a communication profile for sending basic telecontrol messages between two systems, which uses permanent directly connected data circuits between the systems. The IEC Technical Committee 57 (Working Group 03) have developed a protocol standard for Telecontrol, Teleprotection, and associated telecommunications for electric power systems. The result of this work is IEC 60870-5, Telecontrol equipment and systems. Five documents specify the base IEC 60870-5:

- IEC 60870-5-1, Transmission frame formats
- IEC 60870-5-2, Link transmission procedures
- IEC 60870-5-3, General structure of application data
- IEC 60870-5-4, Definition and coding of application information elements
- IEC 60870-5-5, Basic application functions

IEC TC 57 has also generated companion standards:

- IEC 60870-5-101, Transmission Protocols, companion standard for basic telecontrol tasks
- IEC 60870-5-102, Companion standard for the transmission of integrated totals in electric power systems (this standard is not widely used)
- IEC 60870-5-103, Transmission protocols, Companion standard for the informative interface of protection equipment
- IEC 60870-5-104, Transmission Protocols, Network access for IEC 60870-5-101 using standard transport profiles

2.2 Smart Grids in SecureCloud

The SecureCloud project will consider use cases in the area of smart grids. Smart grid applications offer the opportunities to consider many of the requirements that a sensitive big data applications may have when executing in the cloud. First, smart grid applications consider a growing volume of data. Smart meters and sensors for monitoring distribution and transmission grids are being deployed and are capable of continuously collecting and transmitting data. Adequate use of this data enables energy distributors not only to optimise their infrastructure, but also to reduce the environmental impact of supplying power to a given load or region. Second, these promising data analysis opportunities require having access to detailed information about energy consumption.

In the first use case we consider, smart meters collect detailed power consumption data from a residential or industrial consumer. Collecting data at granularities of minutes, or even seconds, enables for sophisticated applications that prevent power theft, detect power quality issues, in order to calculate and prevent penalties for fault duration, among other applications for triggering adaptation actions that increase efficiency or robustness of the power grid. Currently, these applications are deployed on dedicated servers maintained by utilities and system integrators. Therefore, they cannot be exploited for all customers because this would require a large data storage and processing infrastructure. Using cloud computing can help to provide such an infrastructure. Nevertheless, once this data is under control of an energy provider, an adversary who compromises this providers infrastructure, e.g., a malicious employee or an oppressive government, could gain access to them. The data therefore needs to be processed in a secure fashion and never be readable in a non-encrypted form outside the secure container.

2.2.1 COPEL

Copel *Companhia Paranaense de Energia*, the largest company of the State of Paraná, was founded on October 26, 1954 with ownership control held by the State of Paraná. The Company went public in April 1994 (BM&FBovespa) and, in 1997, it was the first company of the Brazilian electricity sector to be listed at the New York Stock Exchange. As from June 2002, the brand is also present at the European Economic Community, having been listed at Latibex the Latin American arm of the Madrid Stock Exchange. As of May 7, 2008, Copels shares were ranked at Level 1 of So Paulo Stock Exchange (BM&FBovespa) Corporate Governance.

The Company directly serves 4,391,313 consuming units, across 395 cities and 1,113 locations (districts, villages and settlements), located in the State of Paraná. This network consists of 3.5 million homes, 89 thousand plants, 373 thousand commercial establishments and 369 thousand rural properties. The staff is composed of 8,653 employees.

Copels structure comprises the operation of:

- An own generating complex composed of 20 power plants (18 hydroelectric plants, 1 thermal plant and 1 wind plant), whose installed capacity totals 4,754 MW;
- The transmission system totaling 2,302 km of lines and 33 substations (all of them automated);
- The distribution system, which consists of 192,508 km of lines and network of up to 230 kV enough to spin four times round the Earth through the Equator line and 362 substations (100% automated);

- The optical telecommunication system Paranás Infoway), which has 9,793 km of OPGW cables installed between the main ring and urban radials (self-sustained cables), totaling 18,212 km and reached 41,153 clients distributed reaching 399 cities of the State of Paraná and 3 cities of the State of Santa Catarina.

COPEL as a large utility will contribute to the requirements collection from an end-users point-of-view and will provide access to real-world Smart Meter measurements.

The data provided by Copel are:

- **Historic Data of consumers:** Ref. 3.1.1 Smart meter data covers consumption data (i.e. energy usage as well as historical consumption), production data.

2.2.2 Israel Electric Corporation IEC

Israel Electric Corporation is the main supplier of electrical power in Israel. IEC builds, maintains and operates power generation stations, sub-stations, as well as the transmission and distribution networks.

The company is the sole integrated electric utility in the State of Israel and generates, transmits and distributes substantially all the electricity used in the State of Israel. The State of Israel owns approximately 99.85% of the Company.

Since its establishment and up to today, IEC builds infrastructure, generates, transmits, and supplies electricity to 2.6 million customers. The Companys main activities take place within the State of Israel. It generates, transmits, distributes, and supplies most of the electricity used in the Israeli economy according to licenses granted by virtue of the Electricity Sector Law, 1996. In addition, the Company acts as administrator of the countrys electricity system.

The data provided by Israel Electric Corporation IEC concern 3.2.1:

- Distribution Management System
- Transmission System
- Smart Home

3 Data Description

3.1 COPEL

Copel collects data from Group A (high voltage customers) for at least 13 months.
The data is useful to application developers. The data analysis or the application can feed a research paper.

3.1.1 Data Set Description

Data set reference and name

Consumers Group A: Historic Data.

Standards and metadata

a) Data Capture Methods:

- Data is exported by an Oracle Database;
- There is no standard or metodology associated with the kind of data that Copel exports;
- Copel can name folders and files according to the customers' number or the meters; Additionally, all data are obtained based in ABNT NBR 14522 standard and all captured data have an associated time stamp. The data definition can be found in deliverable 5.1.
- The dataset will be exported only once.

b) Metadata:

- Copel are not able to define/describe metadata issues. Maybe some examples can help to better understand this need.

Data sharing

Data sharing criteria were not defined yet. It will depend on the anonymizing process in order to not expose customers personal information.

- Data consists in historic values of voltage, current, consumption, power usage, power factor and geographic location. The exporting format is CSV and XLSX. Copel expects no more than 100 GB of data;
- Data must be anonimized before sharing.

COPEL will adopt the Zenodo repository (appendix A) to make the data available to the research community.

Archiving and preservation (including storage and backup)

Copel recommends to fit with the application requirements.

3.2 Israel Electric Corporation IEC

3.2.1 Data Set Description

The data will be generated by 3 simulators in the IEC lab:

- Energy generation simulator (EGS)
- Energy transmission and distribution simulator (ETDS)
- Smart homes simulator (SHS)

The simulated data contains the momentary state of each sub-system during the simulator execution and simulated IED's (Intelligent Electrical Devices that are usually installed un the grid) data. The method for data capture is: when a value is changed (updated), it is written to the data base. Note: The user has to know that he will use mostly simulated data that is generated by simulated processes and not data that is generated by real field equipment.

Standards and metadata

The data of the above mentioned simulators is stored in an SQL server by SCADA system according to SCADA standards.
Then procedure that runs automatically every hour (or other time slot that will be decided), extracts the data to text files.

The folders in the repository are in hierarchic structure:

a) IEC
    a) EGS
        a) CONFIGURATION-ID
            a) MODEL-DESCRIPTION OR DATE

b) ETDS

c) SHS

Note that CONFIGURATION-ID is updated due to change of configuration either in the subsystem itself or change of configuration in a subsystem which has relationships with it.
The data of the different files in the IEC dataset will be synchronized by the date of generation which appears in the file name.

Some of the metadata could be created manually and some not. For example the model-description file is a free text file describing the model and the data. The metadata (that describes the data items), which are the columns names will be generated automatically.

Data sharing

IEC has started to store data in a repository for some of the simulators.

The repository will be available on a server in the IEC lab network. This server can be accessed also from the internet by IEC provided access rights to the lab. The files could be copied and processed by each partner in the project.
There will not be any restrictions on using the data.

IEC will publish the dataset in a disciplinary repository. Analysis will be performed using freely available Open Source Software tools. IEC will adopt the Zenodo repository (appendix A) to make the data available to the research community.

Archiving and preservation (including storage and backup)

The data of the simulators is stored in SQL data base. It will be extracted from the data base to text files in CSV format.

Currently our estimation of the volume of the data is 10 MB/day. This amount could increase as some of the simulators are in development process and will produce more data.

IEC preserves data for a log-term period of at least 7 years. Approximated end volume of dataset is 100 GB. Associated cost in preparing the dataset to be ready for archiving will be covered by project itself.

Test dataset IEC will be generated in a test laboratory by IEC team. Dataset could be useful to other research groups working on similar research questions in the area of energy generation and distribution. IEC plans to make our dataset publicly accessible to the all project needs.

Example of Data Set Description

Electricity Chain Systems	Data set and reference name	Data set description	Standards and Metadata	Data Sharing	Archiving and preservation
Distribution Management System (DMS) operational mode normal and under attack	Modbus/TCP normal	network behavior of normal DMS operation	Standard protocol TCP/IP	pcap file available on file server	Files saved on external storage for a year
	Modbus/TCP attack	network behavior of DMS under attack	Standard protocol TCP/IP	pcap file available on file server	Files saved on external storage for a year
	Attacker action	Standard protocol TCP/IP Specific attacker activity	Standard protocol TCP/IP Specific attacker activity	pcap file available on file server	Files saved on external storage for a year
Transmission System	IEC-60870-5-104 IEC-61850-9 normal	traffic between national communication center and substation in normal operation	IEC-60870-5-104 IEC-61850-9	pcap file available on file server	Files saved on external storage for a year
	IEC-60870-5-104 IEC-61850-9 attack	traffic between national communication center and substation under attack	IEC-60870-5-104 IEC-61850-9	pcap file available on file server	Files saved one xternal storage for a year
	Attacker action	network attacker activity	network specific attacker activity	pcap file available on file server	Files saved on external storage for a year
Smart Home	Zigbee normal	common set to actuator and sensors normal	Zigbee and Modbus/TCP	pcap file available on file server	Files saved on external storage for a year
	Zigbee attack	common set to actuator and sensors normal	Zigbee and Modbus/TCP	pcap file available on file server	Files saved on external storage for a year
	Attacker action	network attacker activity	Specific attacker activity	pcap file available on file server	Files saved on external storage for a year

Table 3.1: Example of Data Set Description

Appendices

A Zenodo Repository

The SecureCloud project has chosen Zenodo repository for data sharing. Zenodo builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share, preserve and showcase multidisciplinary research results (data and publications) that are not part of the existing institutional or subject-based repositories of the research communities.
Zenodo enables researchers, scientists, EU projects and institutions to:

- easily share the long tail of small research results in a wide variety of formats including text, spreadsheets,, audio, video, and images across all fields of science.
- display their research results and receive credit by making the research results citable and integrating them into existing reporting lines to funding agencies like the European Commission.
- easily access and reuse shared research results.

Deliverables:

- An open digital repository for everyone and everything not served by a dedicated service; the so called long tail of research results.
- Integration with OpenAIRE infrastructure and assured inclusion in OpenAIRE corpus.
- Easy upload and semi-automatic metadata completion by communication with existing online services such as DropBox for upload, Mendeley/ORCID/CrossRef/OpenAIRE for upload and pre-filling metadata.
- Easy access to research results via an innovative viewing option, open APIs, integration with existing online services, and the preservation of community independent data formats.
- A safe and trusted service by combining community based curation with short- and long-term archival and digital preservation strategies in accordance with best practices.
- Persistent identifiers, Digital Object Identifiers (DOIs), for sharing research results.
- Service hosting according to industry best practices in CERNs professional data centres.
- An easy way to link research results with other results and products, funding sources, institutions, and licenses.

Bibliography

[1] Guidelines on Data Management in Horizon 2020 Version 2.1, 15 February 2016, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 Version 2.1, 15 February 2016, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Bibliography