D4.1 Data Management Platform Architecture

Deliverable ID:	D 4.1
Deliverable Title:	Data Management Platform Architecture
Revision #:	1.4
Dissemination Level:	Public
Responsible beneficiary:	VTT
Contributing beneficiaries:	All
Contractual date of delivery:	30.09.2017
Actual submission date:	27.10.2017

Table of Content

1. Introduction

1.1 Definitions and abbreviations

DMP (Data Management Platform) - is a platform for managing data from multiple sources for storing, processing, analyzing and provisioning. In this project, the platform is built over Azure which is Microsoft cloud platform developed in the PaaS model.

Architecture - is both the process and the product of planning, designing, and constructing data solutions. It contains following elements: information technology specifications, models and guidelines, using a variety of Information Technology notations. Architecture of DMP is covered further in this document by describing all aspects of application Azure services and processes to fulfill project objectives.

AI (Artificial Intelligence) - is apparently intelligent behavior by machines, rather than the natural intelligence (NI) of humans and other animals. In computer science AI research is defined as the study of "intelligent agents".

ETL (Extract, Transform, Load) - refers to a process in database usage and especially in data warehousing. Data extraction is where data is extracted from homogeneous or heterogeneous data sources; data transformation where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis; data loading where the data is loaded into the final target database.

GDPR (General Data Protection Regulation) – A 2016 formed regulation by European Union that targets to strengthen and unify Data Protection for individuals in EU.

HSM (Hardware Security Module) - is a physical computing device that safeguards and manages digital keys for strong authentication and provides crypto processing.

HTTPs (Hyper Text Transport Protocol) - also called HTTP over Transport Layer Security, is a communications protocol for secure communication over a computer network which is widely used on the Internet.

IaaS (Infrastructure as a service) - refers to online services that provide high-level APIs used to dereference various low-level details of underlying network infrastructure like physical computing resources, location, data partitioning, scaling, security, backup etc.

IoT (Internet of Things) - is the network of physical devices, vehicles, and other items embedded with electronics, software, sensors, actuators, and network connectivity which enable these objects to collect and exchange data.

ML (Machine Learning) – field of computer science that utilizes algorithms to enable learning form data and predicting based on the trained algorithms.

PaaS (Platform as a Service) - is a category of cloud computing services that provides a platform allowing customers to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app.

REST API (Representational State Transfer Application Protocol Interface) - is a way of providing interoperability between computer systems on the Internet. REST-compliant Web services allow requesting systems to access and manipulate textual representations of Web resources using a uniform and predefined set of stateless operations.

SQL (Structured Query Language) - domain-specific language used in programming and designed for managing data held in a relational database management system.

SSL/ TLS (Secure Sockets Layer / Transport Layer Security) – SSL is a predecessor of TLS, both frequently referred to as "SSL", are cryptographic protocols that provide communications security over a computer network.

VPN (Virtual Private Network) - extends a private network across a public network, and enables users to send and receive data across shared or public networks as if their computing devices were directly connected to the private network.

BYOL (Bring Your Own License) - it enables companies to leverage the power of cloud computing in order to use their existing licenses to deploy and track their license use across the entire company.

1.2 Introduction

This deliverable is about the Establish Data Management Cloud Platform Architecture. A strong architectural foundation is important for any project. Good understanding of the common issues and possible solutions are essential to guarantee good performance and provide good Data Security. Industry recognized best practices and Azure PaaS services ensure that all necessary threats are considered.

The aim of the Data Management platform is to provide a secure and reliable platform for data management purposes. The platform is intended to gather, store, process and utilize data provided by ESTABLISH country pilots and possible applications or custom Web sites.

Establish Data Modelling is built on Microsoft Azure Cloud Platform[1]. Azure is the current Cloud Platform leader achieving global scale with 42 announced regions – more than any other cloud provider. Compared to other platform providers on the market, it is characterized by more comprehensive compliance coverage, with more than 60 compliance offerings. Azure was also the first major cloud provider to contractually commit to the requirements of the General Data Protection Regulation (GDPR). In-order-to protect solutions, Azure embeds security, privacy and compliance into its development methodology, and has been recognized as the most trusted cloud for government institutions, earning a FedRAMP High authorization that covers 18 Azure services.

Looking from the project's requirements' and deliverables perspective, Azure in the only Cloud Platform that enables building intelligent solutions at scale (Big Data) using cognitive APIs, bots, IoT and Machine Learning capabilities as PaaS With complaining these capabilities with powerful GPU-based computing, Azure accelerates deep learning, enables high-performance computing simulations and conducts real-time data analytics, which is crucial when dealing with sensor data. Also, to support local pilots, Azure offers hybrid consistency everywhere – in application development, management and security, identity management across the data platform. This helps reduce the risk and cost of a hybrid cloud environment by enabling a common set of skills and offering portability of applications and workloads.

Azure is a great helper in getting more value out of investments by reducing costs of solutions. It enables to utilize services on pay-as-you-go basis, and is the only cloud provider that enables hourly based billing. What is more, it supports BYOL model and a vast variety of open source applications, which can be implemented at desired pricing tiers to increase even more economics of the project.

This deliverable will cover the reference architecture of the Establish Data Management platform in the next chapter. In scope for this deliverable is to give a high-level description of ESTABLISH reference architecture that can be utilized on the project. The report contains a description about the current

[1] *Microsoft Azure site 2017 [Online]. https://azure.microsoft.com/en-us/overview/what-is-azure.*

available Azure PaaS services. A More detailed on-premise solution of the country pilots and a general project architecture will be described in WP3 deliverable of architecture.

The second chapter introduces the ESTABLISH Azure Reference architecture, gives a quick introduction to the PaaS services included in the architecture, and describes Data Management platform in relation to the whole ESTABLISH architecture. In the third chapter, there is discussion about relevant issues related to Data Management and security issues. The issues are addressed with the help of industry recognized best practices. Also, a solution provided by Azure is introduced. The fourth chapter is about the remaining issues related to data management e.g. metadata management. Finally, the last chapter goes over the conclusions of the report.

2. ESTABLISH Data Management Platform Architecture

2.1 Data Management Platform in relation to Whole ESTABLISH Architecture

The purpose of the Data Management Platform is to provide a solution to gather, store and leverage the data produced by the country pilots of the project. The data is gathered trough suitable application interface solutions (e.g. REST API), using ETL logic, and/or tools designed for unstructured data import. The Data is then stored to the Cloud Platform Data Warehouse. The Data Management Platform in relation to whole ESTABLISH architecture is presented in Figure 1.

Data Analytics procedures will be used to process data and gather further insight from data in the cloud environment, if necessary. The data stored to the Cloud Platform can then be used for various different services and/or use cases.

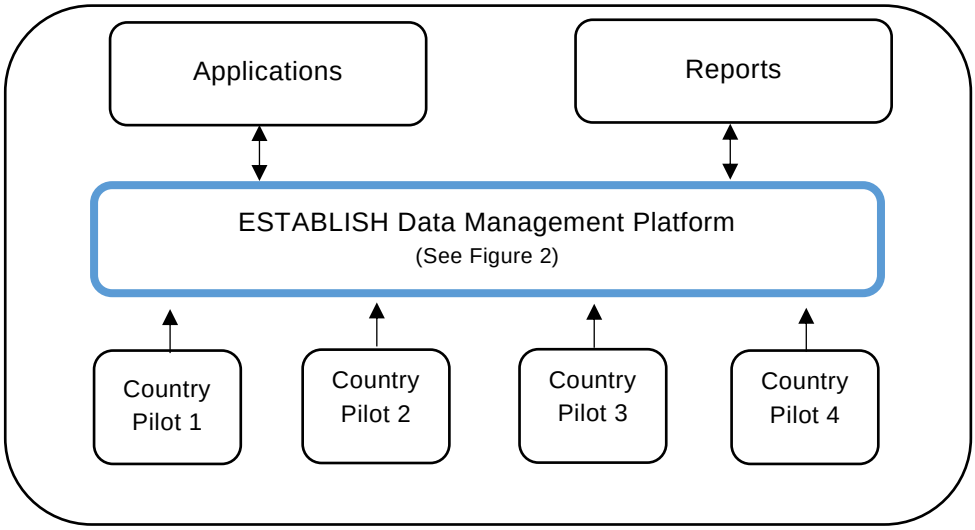

Figure 1. Data Management Platform in relation to Whole ESTABLISH Architecture.

2.2 ESTABLISH Microsoft Azure Reference Architecture

ESTABLISH reference architecture is built on Microsoft Azure PaaS services, see Figure 2. PaaS is a cloud computing model which provides a platform allowing customers to develop, manage and utilize different solutions and services without the complex building and maintaining of the infrastructure.

Using Azure's PaaS Services enables various types of Data Importation, Data Analysis and ways of Utilizing data for ESTABLISH Data Management Platform. Different services can be easily included to expand the architecture with desired capabilities. Most of the architectures components are easily scalable, and is based on pay-as-you-go logic. This means that the architecture can store huge amounts of data if and withstand numerous users of ESTABLISH services, if required.

The ESTABLISH reference Architecture is composed utilizing the Bigdatapump Azure 2020 reference architecture as a basis for Data Management architecture planning. This reference architecture is composed by Bigdatapump Cloud Architects according to our experiences gathered from multiple enterprise level Azure ramp-up projects and possible project needs gathered from WP2. The idea is to choose and include the right PaaS services according to the project needs defined in the future.

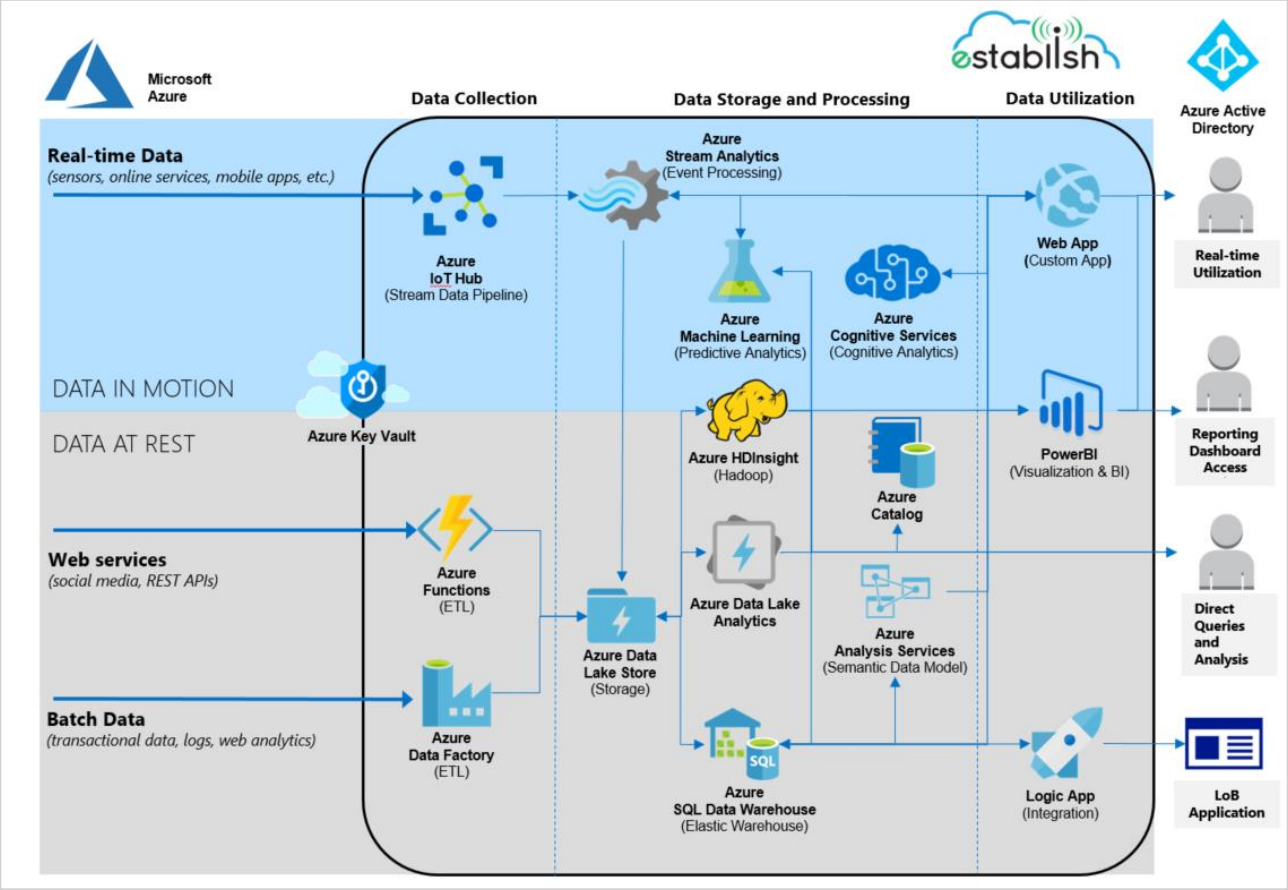

Figure 2. ESTABLISH Data Management Azure Reference Architecture.

2.3 ESTABLISH Azure Reference Architecture's PaaS services

The project can utilize various PaaS services for the architecture. The possibilities provided by the ESTABLISH architecture cover different target needs - from small app development basis to huge enterprise level ramp-up projects with complicated AI capabilities included.

PaaS has three main features [2][3][4]. First there is what is called the stack – layers of software to allow applications to run on. They consist of frameworks, services and libraries that the data engineers can use to build the solutions made for the project. The second one of the main features comparison to IaaS is the lack of virtual machine related work. This is possible because of deployment machinery that instantiates virtual servers and provisions them with instances of the of the stack. It is possible just to simply deploy the solution according to the specs of your choosing and let the PaaS deployment machinery set the service up and running.

[2] *Fielding, R. T., 2017. Representational State Transfer (REST), Architectural Styles and the Design of Network-based Software Architectures. [Online].*
Available at: http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
[3] *Hanna, C. & Mak, D., 2017. Learn about Language Understanding Intelligent Service (LUIS). [Online]*
Available at: https://docs.microsoft.com/en-us/azure/cognitive-services/luis/home
[4] *Nik, J. e. a., 2017. Computer Vision API Version 1.0.. [Online]*
Available at: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/home

The third component is the choosing between complexity and flexibility. The latter one can be accomplished by choosing more PaaS services with different features. All the different PaaS elements bring a different functionality and possibilities with them.

There are plenty of good reasons to choose PaaS as the basis for architecture. It enables faster deployment of solutions, as the need for complex settings get simpler and allows focus on the solution – for example a machine learning algorithms. The PaaS also enables very flexible scalability to react to different situations. This means that for many solutions e.g. Data Warehouse, can be scaled up or down based on need. This combined with automated security measures, such as vulnerability scanning, enable maximized up time for the solutions.

PaaS takes advantage of the latest technologies developed by the service provider. Azure provides the proven quality of Microsoft developed solutions that can enable better and more secure features compared to building infrastructure from the start. The following are descriptions of the ESTABLISH reference architecture, and can be found on numerous places on internet e.g. Microsoft Azure site[5].

2.3.1 Data collection related ESTABLISH Azure reference architecture PaaS services

Azure IOT Hub is designed to enable bi-directional real-time connectivity with organizations IoT resources. Bi-directionality means that the service allows two-way connectivity between the platform and the sensors. The service allows importation of huge amounts of data in an instant.

Azure Functions allows you to build and run serverless apps with little effort. Easily Scalable service that allows you to implement different language based (e.g. C#) services with ease. The service can also be used to import data from different web services.

Azure Data Factory is a PaaS service designed for full data integration operations and ELT process management in the cloud environment. The service is fully scalable, and can either retrieve or bring data from different sources according to configurations.

2.3.2 Data Processing and Storage related ESTABLISH Azure reference architecture PaaS services

Azure Stream Analytics allows insight gathering from huge amounts of stream or data on rest, without the usual problems of running servers. The PaaS utilizes a SQL-kind of query language.

Azure Machine Learning allows implementing of Machine Learning algorithms in seconds. The service enables simultaneous use and cooperation of both R and Python language.

Azure Cognitive Services is a service designed to use advanced AI to leverage business communication. The service can be used to analyze images used in communication with clients to get the best insights.

Azure HDInsight provides enterprise level Bigdata analysis tools. This allows the formation optimized clusters for different open source services: Hadoop, Spark, Hive, HBase, Storm, Kafka, and Microsoft R Server.

Azure Analytics Services is a PaaS designed to help get insight from data and visualize it. Use e.g. semantic analysis to gather further knowledge.

Azure Data Lake Analytics enables running simultaneous transforming and analysis of huge amounts of unstructured data. Use R, Python or U-SQL all at once.

Azure Data Lake Store is a storage PaaS for huge amounts of unstructured data.

[5] *Microsoft Azure site 2017 [Online]. https://azure.microsoft.com/en-us/.*

Azure Data Warehouse is a secure place to store organizations data. The PaaS provides storage and instant scalability for huge amounts of data.

Azure Catalog Store and manage meta data. The service provides ways to make data management easy and effortless.

2.3.3 Security related PaaS services for ESTABLISH Reference Architecture

Azure Active Directory allows to deploy and maintain multifactor authentication. Only the right people have access to the right data.

Azure Key Vault secures encryption keys with Microsoft Azure PaaS services.

Azure Security Center allows security monitoring and implementation of policies. Uses advanced analytics to solve security issues and secure data. It features services such as vulnerability scans and alerts.

2.3.4 Data Utilization related ESTABLISH Azure reference architecture services

Power BI is an interactive business intelligence tool for visualizing data. The service can be used to build reports, graphs and visuals to communicate the insights of the data.

Logic App allows to quickly connect and integrate various solutions to your cloud environment. Add the features of apps to the Cloud environment with ease.

Custom Web sites is a way to build unique interfaces that can utilize the data of the cloud platform.

2.4 Data Management Platforms relation to country pilots and relevant processes & principles

This chapter covers possible scenarios of the ESTABLISH Data Management platform to work with pilots' data, and introduces the needed process and documentation requirements. The country pilots differ from each other, and more detailed descriptions of the pilot architectures, their data management processes etc. are presented by deliverables of WP3. Documentation of the different connections with the Data Management Platform will be created and published in the later software development phase of the project. As the final solution and data sources is unknown, Finnish pilot will be used as an example with the knowledge available at the time of this report.

2.4.1 Planning phase

All development work should be done with respect to ESTABLISH project use cases and pilots. As demonstrated in Bigdatapump's best practice process for creating value from data (see Figure 3), the use cases should be the starting point of the planning in country pilots, as well as in the Data Management platform. After the understanding of use cases and their requirements, data analysis should be conducted. The process is then followed by data preparation and modelling before evaluation deployment and iteration phase.

Iteration and incremental development allows corrective and further development after evaluation phase. The process involves the planning and building of architectural solutions. As mentioned before, this document will concentrate to discuss matters and parts of the processes relevant from the ESTABLSIH Data Management platform perspective.

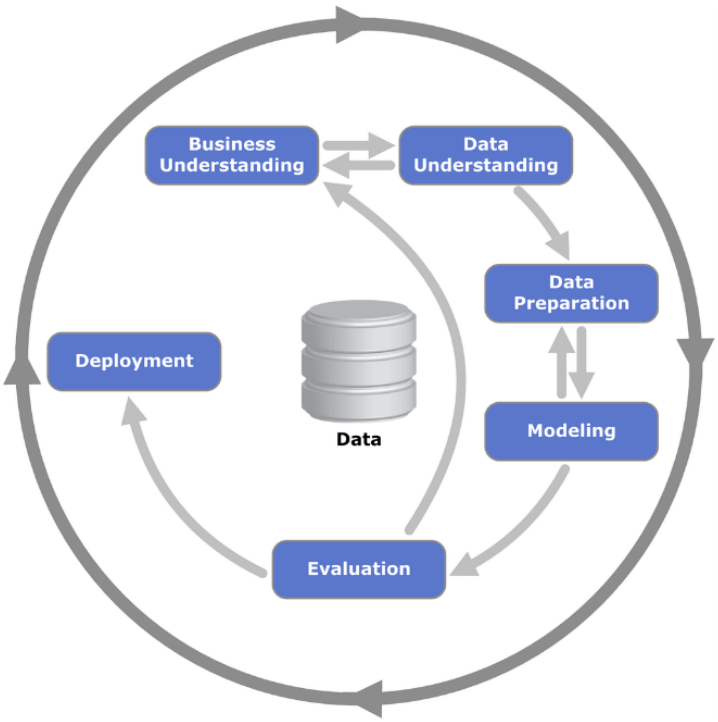

Figure 3. Bigdatapump process of Creating Value from Data.

2.4.2 Connecting to Data Management Platform: Processes, Procedures & Documentation

Establish project's data will be derived from numerous data sources provided by country pilots. Thus, there is a need to introduce data governance tools and techniques that help organizing wide catalogue of project data.

Processes and responsibilities are aligned with the requirements mentioned in chapter 3 of this document. They are then written down as principles and communicated to the relevant parties.

The country pilots should collect and register metadata. The functionality and the service available on Azure (Data Catalog) is described in chapter 3.1 of this document. If Azure Data Catalog is not running at the time, the data should be stored securely by the Data Management responsible of the pilot. The registered meta data should include at minimum:

- Name - name of a source,
- Friendly Name – a name of a source that can be easily read by human,
- Description – a brief description of a data source,
- Expert – a mail to a person with expert knowledge on data,
- Tags – unique set of phrases by which a source will be searched for in the catalogue, and
- Connection Info – a set of attributes essential to establish connection to a source.

A good practice for planning govern processes is that a data owner is named responsible for entering meta data to a catalog when a new data source is added to the platform architecture. He is also responsible of ensuring that the data form, source and security compliance to regulations. There are defined in more detail in chapter 3.4 of this document and in GDPR.

In the Finnish pilot, an architecture encompasses Event Hub to collect streaming data and Azure Blob Storage for cold data. The PaaS was chosen because the pilot contains Stream data originating from sensors. Configured data sources are listed and gathered in Azure SQL Warehouse by the project member in charge of data source configuration. The responsibilities are communicated in the pilot by writing Data Management principles for parties involved in the project before starting the configuration phase. The principles are then discussed with all Pilot parties to ensure common understanding of the procedures.2.4.3 Connecting to the data management platform

The ESTABLISH Platform, as mentioned in point 1.2 of this document, is built upon Azure PaaS services which provide great selection of endpoints. Thus, basing on implementation and character of pilot data (on rest or streaming) ingesting and analytical architecture may differ.

An analysis of the data sources is needed to determine the best way to connect the country pilot platform with the ESTABLISH Data Management platform. This includes listing the form, quality, frequency and quantity of the data and other relevant features. The right PaaS service is then chosen considering the needs of the use case and taking into account the features of the data sources. The chosen PaaS is then configurated to enable data collection to the Data Management Platform. The configuration is done with respect to the Data Management Principles and processes mentioned earlier in this chapter.

The Finnish pilot will bring relational data most probably in CSV format to the ESTABLISH Data Management Platform. These tables are information gathered from sensors enriched with data from enquiry collected from end users in the country architect platform. This country platform architecture will be described in more detail by deliverable of WP3. Data will be imported by using Azure Data Factory.

2.4.4 Processing and storing data

The collected data that is collected from the country pilots is stored to the Azure Data Management Platform using PaaS services designed for data storage. The form, quantity and use case of the data will determine the right service to use. For example, large quantities of unstructured data can be stored to Azure Data Lake in the ESTABLISH reference architecture.

The data can be processed before and/or after storing to produce more insight. data source analysis & data modelling is a crucial part for the processing the data. The need for preprocessing of the data is determined analyzing the need to access original data later during the project. If only the processed data holds value, then there is no need to store the unprocessed data and processed data. The need to aggregate data beforehand has reduced as the storage services have evolved and become more cost friendly. With more cheap prices of storage, organizations can more cost efficiently preserve data in case they want to do research to find insight from it later.

The quantity and form of the data gathered from the Finnish Pilot is such, that Azure Data Warehouse is the PaaS service most probably selected to store the data in the ESTABLISH Data Management Platform reference architecture. The original data will be stored in the Country Pilot platform before analyzing it using a R Modell deployed in Machine Learning Studio Azure PaaS service. Insights gathered from the country pilot will then be collected to Establish Data Management platform using a gateway PaaS service, such as Azure Data Factory.

Processing involves getting new insight of data. This is commonly done with the help of Data Analytics, which is covered in WP5 of ESTAPLISH project. The use case requirements, determine what kind of data models & analytic capabilities are needed. The right PaaS services are then chosen to enable the usage of right kind of capabilities.

3. Security related questions and Best Practices regarding Cloud Architecture

The best practices are based on project experiences discussed in with ESTABLISH project partners and industry [1] best practices.

3.1 Multi-factor Authentication and Role Based Access Control

Access to Data can be restricted in various ways. The possibilities for security breaches are often caused by human errors. Because of this, strong authentication procedures are needed. A multi-factor authentication leverages multiple pieces of confirmation to ensure the identification of the user. Often this is accomplished by combining the traditional User ID, password and a PIN code sent to a preannounced contact detail.

Multi-Factor authentication is a good way to make sure, that unauthorized personnel cannot gain access data, if a person forgets to log out of their account, or if the credentials have been stolen. It should be taken to notice, that a PIN code sent to user will probably be received through a device carried around with the person. If the mobile phone e.g. is stolen with the credentials, there is a possibility of a data breach without the use of other security measures. The phone should also be encrypted to support the security measures. Another possible solution for the problem described earlier is to use Privileged Access Workstations. The method lowers the risk of data breach in case of an endangered work station. Access Management should be centralized in organizations.

Role based access control is a way to control data access. EU regulation governs that a person should not have access to third party personal information, if they do not need it regarding their duties. This is also an industry best practice, that can be implemented with role based access to data. This means, that there are different roles within an organization to determine who has access to the data. An important part of platform security is ensuring correct way of managing resources. It is vital, that only the right personnel can create and assign resources within the platform.

Azure Active Access is a tool that enables multi-factor identification and role based access management for ESTABLISH Data Management Platform. Access can be easily defined trough defining different user groups that have access to different resources in Azure subscription. Responsibilities for managing technical aspects of access and access control must be assigned in the project to a party with sufficient area knowledge and technical expertise. The project Data Management can also utilize Azure Resource Manager, which is a service dedicated towards creating, deploying and managing resources in a secure and efficient way.

3.2 Data Encryption and Encryption Key Management

Data Encryption is a security related activity where data is translated into another form, or code to make sure that only the right people with the right decryption keys can access the data. This means that the data is turned into secured cyphertext using algorithms. The encryption algorithms also create a decryption key, to allow data access. Data Encryption is currently one of the most effective ways to enforce data security.

Microsoft Databse Sercurity Best Practises for the Vigilant Database Administrator and developer 2010. [Online] Avialable: Https://technet.microsoft.com/en-us/security/gg483744.aspxhardware security modules (HSMs), and promotes separation of management of keys and data to help meet regulatory compliances.

ESTABLISH Data Management Platform utilizes Azure Data Encryption in rest. Data is encoded with symmetric encryption keys. This service is hosted by Microsoft and does not require any additional efforts – it is transparent from user's perspective. This way the data is secured, when it's not being used, and can quickly be retrieved, if needed.

Azure Data Encryption can facilitate Azure Key Vault which is Azure's cloud-based external key management system to allow the user to take control over their encryption keys and control who can access them and when. Key Vault provides central key management, leverages tightly monitored

When data is being transferred between components, locations or programs, such as over the network, across a service bus (from on-premises to cloud and vice-versa, including hybrid connections such as ExpressRoute), or during an input/output process, it is thought of as being in-motion. Since data will be moving back and forth from many locations, ESTABLISH Data Management Platform uses SSL/TLS/HTTPs protocols and different VPN configurations to exchange data across.

3.3 Active Monitoring

Good security solutions monitor various activities for unwanted activities and incidents. Security operations settings should take into consideration the industries' best practises and organisations must have an up-to-date security policies and procedures. Monitoring consists of actively keeping track of platform security & infrastructure e.g. of logging and monitoring & alarming of unauthorized access.

An important part of having an up-to-date security is to actively update the latest stack, and improve security related activities and processes. Making sure that organization personnel is aware and acts the correct way in different situations is equally crucial. Security organization should be responsible for automated monitoring and employees should not be able to effect on what is being monitored on their devices.

With the help of Azures Security Center, Company can automate vital scanning modes and set alarms to notify of events. The reports created help evaluate the current state of security and take corrective action. Azure guidelines help to make sure the best practices are correctly in place to make platform security more reliant.

Additionally, particular components are equipped with features that help detect suspicious activities, potential vulnerabilities and limit access from the internet.

3.4 Compliance with Regulation

ESTABLISH Data Management Platform will follow the guidelines and regulations set by governments and authorities. From architectural perspective, there are a couple of common questions to solve that rise from regulatory circumstances.

Possible restrictions of personal data. Many laws and regulations govern the use and storing of data that is classified personal. Often the personal data classifications include features such as: health related information or social security related information. Working in a multinational project means, that project policies and processes related to data transfer must make sure, that personal data and governing regulations are identified and taken into consideration before data transfer or storage. The policies & processes should be in placed both on global and national pilot project level.

Physical location of data. This means that ESTABLISH projects physical Data Center locations must be defined according to regulation. As an example, EU's General Data Protection Regulation governs, that

no personal data is allowed to be transferred outside of EU. This matter must also be taken into consideration in the policies and processes of the project.

4. Data management in ESTABLISH

4.1 Data Management

Data Management practices [1] can be divided into four different categories: data collection, data storages, access rights and infrastructure. The precious chapters have already tackled some of the previous subjects. In this chapter, we focus on data storing related issues.

Data will be brought to ESTABLISH cloud platform with the help of PaaS data collection related services, such as Azure Data Factory Gateway service. The way that data is brought on platform depends on the data source e.g. Extract Transform & Load -logic. Data engineers will analyze and help county pilots to configurate and, if necessary, integrate selected data sources, which will be registered to ESTABLISH Data Catalog.

The relational data will be stored in scalable Data Warehouse and the unrelated data in Azure Data Lake storage PaaS. The limits for storage and data sizes will be decided on cost based evaluation together with project cooperates.

From the different storage solutions, data can be accessed by authorized parties and used to provide insight regarding the project targets.

4.2 Meta Data Management

Meta Data Management refers to the methods built around creating, controlling, defining and managing data about other data.[1] When data amounts grow, it requires more to keep up with what kind of data your organization has and where it can be found.

Good Metadata management enables organizations to spend less effort on finding the information, and to focus more on extracting value from it. This means making the metadata easy to find and access for relevant user groups. The process of metadata registration is important to be set correctly so that it benefits the use of metadata.

Azures Data Catalog is a service that drives ESTABSLISH Data Management Platform towards this target. You can register data assets with data catalogue and make sure it can be leveraged. Data Catalogue allows you to take advantage of more trivial data sets and share knowledge about the data resources. There is a good change for issues, if the knowledge is limited only to people working straight with the data. The Data Catalog allows to share details, but also name an expert, who has the most insight about the information, if there are more details needed. This way knowledge sharing is made easier within the organization.

4.3 Storing Audit Trail information on Cloud Environment

It is a useful practice to store Audit Trail information on the cloud environment. ESTABLSIH Azure based Data Management platform collects data from activities, which can be used to trace back actions made in the Cloud environment. This data can be securely stored on the cloud environment and retrieved later for use if necessary.

An example of possible use case is that an unwanted incident is detected. The causer of the incident can be traced using e.g. HDInsight. This way the organization can make sure that the right person can get feedback and instructions, if possible to avoid such an incident in the future.

[1] *Boston College Libraries 2017 [Online]. https://libguides.bc.edu/dataplan/best_practices.*

4.4 Documentation

Best practices suggest adapting standards for documentation purposes, so that information is documented on a satisfactory level. The purpose of documentation is to collect and store data about solutions in a way, that can be easily communicated. For these purposes standards and scopes are important to define. For ESTABLISH project, these standards are especially important regarding ways, that different pilots execute documentation.

Documentation process is carried out through Itea Portal reporting, in the project standard template. WP4 will concentrate on documenting the ESTABLISH Data Management Platform Azure architecture in such a way, that a Microsoft Azure solutions certified and experienced professional can re-produce the result of the platform development. Documentation will become more accurately on the later phase of the project, as the Country Pilots reach a more mature phase.

4.5 Life Cycle Management and Maintainability

It's important forward to plan forward about life cycle related issues. How long will the development phase take time? What is the time, that the service will be running? What is the appropriate time to keep data in storage, before it comes un-relevant? When will the services of the project end? If the services and solutions are not shut down after the project, a proper life cycle plan including should be included to the project plan including data life cycle principles.

It is good data governance not to store un-relevant data that can't be used in the future. For research project determining what is relevant data is a hard job, because the relevance might emerge later. The ESTABLISH Azure Reference Architecture allows to set a time limitation, after which the stored data is erased automatically. This should possibility should be taken to consideration when planning for project & Data life cycle.

After an IT-related development project has finished, the development team either takes over the responsibility of maintenance, or usually hands it over to a different service team. There is also a possibility that the project does not go to production, in which case the data should be erased safely from the Data Management Platform. The same action should be taken, when the project is shut down. ESTABLSIH Data Management platform allows safe ways to erase data permanently.

If, the project goes to production, an assigned service can take on the responsibility of maintaining the platform build for ESTABLSIH project. Possible scope of the maintenance team effort is planned together with the cooperatives. Service window is set for 30 days regarding urgent matters and 90 days for un-urgent matters. The service team will have capability to handle all Platform related maintenance work, except those provided by country pilots. Further developments are planned together with the business representatives. The maintenance team can do most of the development work, but is also able to rely on the help of the actual developers, if needed. A discussion will be held about the necessity for a service desk. First suggestion is to use a ticket based notification system. The service team will not be responsible to solve matters outside of the Azure environment.

All of the possibilities described earlier must be taken to account in the planning phase.

5. Conclusions

This report has described the Architecture of ESTABLISH Data Management Platform. The platform will be build using Microsoft Azure Cloud PaaS services. These services will provide the architecture with many beneficial attributes e.g. automated vulnerability scanning's and flexibility. A general description of the whole ESTABLISH project Architecture is presented in deliverable 3.1.

There are important factors related to data management platform development - one of the most important is data security. There are many security related factors that must be considered e.g. role based access management. Other Data Management related issues are also important, from the perspective of being able to use project data. Good Meta Data Management, for example, will ensure that the organization can find and utilize the data.

After the development, it is important to make sure, that the developed platform maintains good performance and tight security trough maintenance and further development.