ETSI GR PDL 009 V1.1.1 (2021-09)

Permissioned Distributed Ledger (PDL); Federated Data Management

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e8840d-c1cf-42b1-ac42-3e1bd4654d5c/etsi-gr-pdl-009-v1-1-1-2021-09

Disclaimer

The present document has been produced and approved by the Permissioned Distributed Ledger ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference

DGR/PDL-009_Fed_Data_Mgmt

Keywords

authentication, data preservation, data sharing, privacy, security

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ETSI GR PDL 009 V1.1.1 (2021-09)

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Notice of disclaimer & limitation of liability

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.
No recommendation as to products and services or vendors is made or should be implied.
No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.
In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

Copyright Notification

No part may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm except as authorized by written permission of ETSI.
The content of the PDF version shall not be modified without the written authorization of ETSI.
The copyright and the foregoing restriction extend to reproduction in all media.

© ETSI 2021.
All rights reserved.

Contents

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e8840d-c1cf-42b1-ac42-
3f1bdd66d45f/etsi-gr-pdl-009-v1-1-1-2021-09

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/06c0840d-e1c0-4201-ae42-
3e1bd4654d5c/etsi-gr-pdl-009-v1-1-1-2021-09

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Permissioned Distributed Ledger (PDL).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document will describe use case scenarios, functional architecture, key functional components mechanisms of leveraging PDL for federated data management (e.g. PDL for federated learning, the integration of PDL and the whole data pipeline).

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e8840d-c1cf-42b1-ac42-5c18d4654d5c/etsi-gr-pdl-009-v1-1-1-2021-09

[i.1] Industrial Internet Consortium: "The Industrial Internet of Things Volume G1: Reference Architecture (Version 1.9)", June 19, 2019.

NOTE: Available at https://www.iiconsortium.org/pdf/IIRA-v1.9.pdf.

[i.2] A. C. Yao: "Protocols for Secure Computations", 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982), Chicago, IL, USA, 1982, Pages 160-164.

[i.3] D. W. Archer, D. Bogdanov, Y. Lindell, L. Kamm, K. Nielsen, J. I. Pagter, N. P. Smart and R. N. Wright: "From Keys to Databases - Real-World Applications of Secure Multi-Party Computation", The Computer Journal, Volume 61, Issue 12, December 2018, Pages 1749-1771.

[i.4] C. Zhao, S. Zhao, M. Zhao, Z. Chen, C.-Z. Gao, H. Li and Y. Tan: "Secure Multi-Party Computation: Theory, practice and applications", Information Sciences, Volume 476, 2019, Pages 357-372.

[i.5] World Economic Forum White Paper: "Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data", July 2019.

NOTE: Available at http://www3.weforum.org/docs/WEF_Federated_Data_Systems_2019.pdf.

[i.6] World Economic Forum, Insight Report: "Sharing Sensitive Health Data in a Federated Data Consortium Model - An Eight-Step Guide", July 2020.

NOTE: Available at http://www3.weforum.org/docs/WEF_Sharing_Sensitive_Health_Data_2020.pdf.

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

federated data collection: data collection scenario where multiple data types are involved and/or multiple organizations jointly collect data of their interest, for instance, to improve data collection efficiency

federated data computing: data computing scenario where multiple organizations work together to solve a data computation task

NOTE: Examples of federated data computing include, but not limited to, federated learning, multi-party computation, and even decentralized Artificial Intelligence/Machine Learning (AI/ML).

federated data discovery and sharing: data discovery and sharing scenario where federated data is discovered by and shared among multiple organizations

federated data management: data management scenario where multiple organizations and/or multiple data types could get involved in each stage of the entire data pipeline or lifecycle and form data federation

NOTE: Examples of federated data management are federated data collection, federated data storing, federated data computing such as federated learning and multi-party computation, federated data sharing, etc.

federated data storing: data storing scenario where multiple organizations participate in storing data, likely, in distributed places

iTeh STANDARD PREVIEW
(standards.iteh.ai)

federated learning: distributed machine learning approach where multiple clients and a federated learning server jointly learn an AI model and provide data privacy protection

NOTE: A federated learning process generally works with a few steps:

ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e8840d-a1a6-43b1-ac42-3e1bd4654d5c/etsi-gr-pdl-009-v1-1-1-2021-09

1) training data are distributed and kept at federated learning clients;

2) a federated learning server coordinates all federated learning clients for them to perform local training and generate local and temporary model updates for each learning round;

3) the federated learning server receives model updates from federated learning clients and aggregate them together to generate a global model;

4) the global model will be sent to federated learning clients for them to perform next round of local training until the goal model converges to the one meeting the expected accuracy.

multi-party computation: secure computation protocol where multiple parties jointly compute a function and guarantees their data privacy

NOTE: In a multi-party computation:

1) multiple parties jointly compute a function over their individual data inputs to get a computation result;

2) each party knows the computation result; and

3) none of parties can learn other parties' data inputs but only knows the computation result.

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AI	Artificial Intelligence
ETSI	European Telecommunications Standards Institute
FDDSS	Federated Data Discovery and Sharing Service
FDM	Federated Data Management
FDS	Federated Discovery Service
FL	Federated Learning
FPP	FDM-PDL Proxy
GDPR	General Data Protection Regulation
IIC	Industrial Internet Consortium
IIoT	Industrial Internet of Things
IoT	Internet of Things
IT	Information Technology
LMS	Ledger Messaging Service
LSS	Ledger Storage Service
ML	Machine Learning
MPC	Multi-Party Computation
MSG	Message
PDL	Permissioned Distributed Ledger
TXN	Transaction

4 Use Cases for Federated Data Management

iTeh STANDARD PREVIEW
(standards.iteh.ai)

4.1 Introduction of Use Cases

ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e884d4-c1cf-42b1-ac42-2e1b44654d5c/etsi-gr-pdl-009-v1.1.1-2021-09

This clause describes some selected federated data management use cases or scenarios, which could be benefited from the use of Permissioned Distributed Ledger (PDL) technology and/or introduce new requirements to PDL technology. As illustrated in Figure 4.1-1, a general data pipeline in federated data management could consist of a set of relatively sequential stages such as data collection, data storing, data computing, data sharing, and data visualization. For each stage, multiple organizations could participate and work together. Each organization could have their own data, for example, generated from ubiquitous devices deployed for different applications such as connected vehicles. In general, a data pipeline (e.g. data pipeline A and data pipeline B) starts with data collection from devices, but it could complete in different places in the networking system. For example, data pipeline B in Figure 4.1-1 stops in edge networks leveraging edge servers for data storing, data computing and data visualization, while data pipeline A ends in the cloud. This clause will not cover the entire data pipeline but focus more on the stages and corresponding scenarios, which are more relevant to PDL technology.

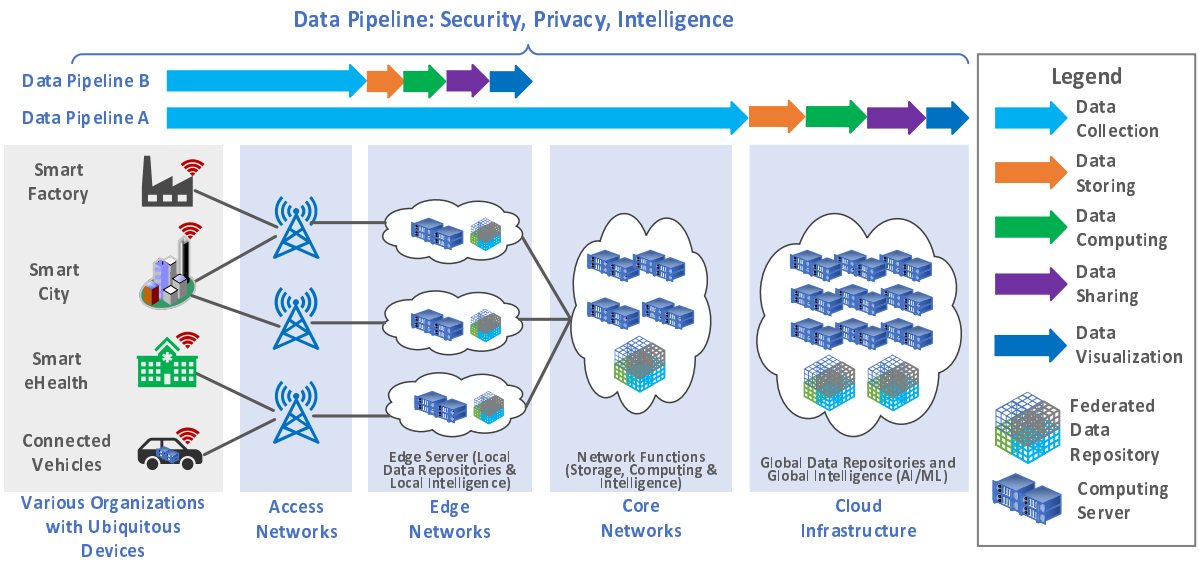


Figure 4.1-1: General Data Pipeline in Federated Data Management

4.2 Federated Data Collection

Our daily lives are surrounded by a variety of sensors and devices. Internet of Things (IoT) technology enables us to leverage these sensors/devices to monitor and measure the physical world in a real-time manner. In many data-driven IoT applications, the first and most important stage is data collection. During data collection, the system can collect data from different devices such as consumer equipment, personal devices, cameras, and wearable health devices; data can also be collected from commercial equipment including security monitoring systems, traffic monitoring equipment, production lines, logistics and supply chain systems, etc. These devices generate different types of data and could belong to and be owned by multiple organizations; the resulted data collection that contains multiple data types and/or relies on multiple organizations is referred to as federated data collection.

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ETSI GR PDL 009 V1.1.1 2021-09
https://standards.iteh.ai/catalog/standards/sist/...-etsi-gr-pdl-009-v1.1.1-2021-09

Figure 4.2-1 shows an Industrial Internet of Things (IIoT) use case, which includes a few processes such as smart manufacturing, smart logistics, and customer experience monitoring. Multiple scenarios could be involved in each process. For example, the smart manufacturing process could cover product quality control, storage management, onsite energy management, equipment maintenance, etc. All those processes and scenarios need to be monitored in real-time to ensure overall product delivery and product quality. As a result, a large amount of production, logistics and customer experience data are generated at all times and need to be collected. However, in a real-world production environment, manufacturing equipment and Information Technology (IT) systems usually involve multiple manufacturers; in the meantime, a complete manufacturing process could involve different departments or even different companies/organizations. Similarly, during the smart logistics process, products will be transported from factories to customers through multiple intermediate transit places, where multiple organizations are involved as well. All of these facts demonstrate that data collection in IIoT is a complex system and needs multi-party collaboration, which is referred to as federated data collection. Please note that Industrial Internet Consortium (IIC) [i.1] defines more IIoT use cases, which are not limited to Figure 4.2-1. In these use cases being considered, data security could be needed; as a result, data at rest and/or data in transit could be encrypted when there is a risk of data leakage.

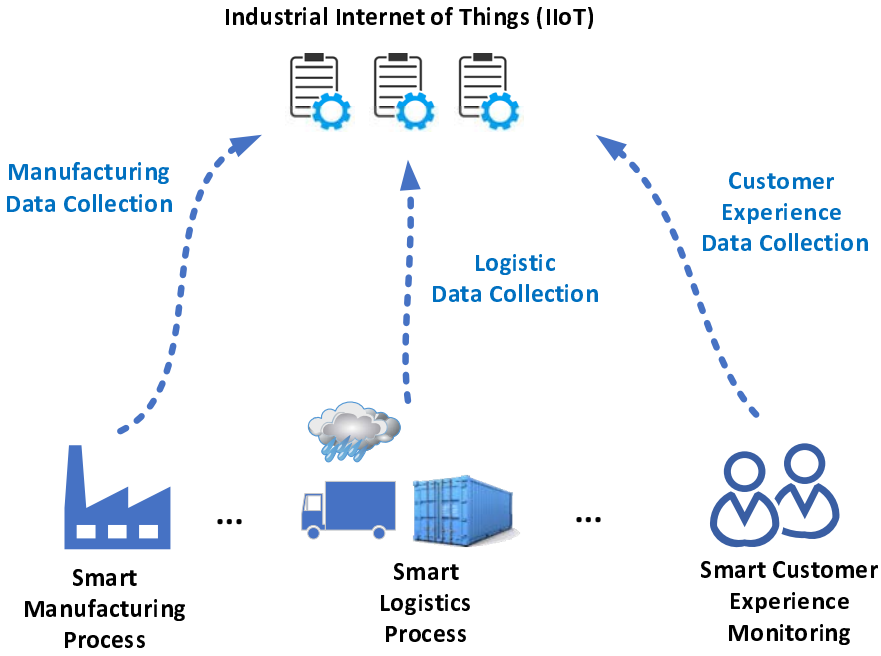


Figure 4.2-1: Federated Data Collection for Industrial Internet of Things (IIoT)

4.3 Federated Learning

iTeh STANDARD PREVIEW
(standards.iteh.ai)

Traditional Machine Learning (ML) technology is usually centralized, in the sense that:

1) training data is usually collected to be stored at a centralized location such as a centralized database; and

2) learning process is performed at a centralized location such as clouds as well. However, traditional ML could cause data leakage issues, since training data is maintained at a location, different than its original place and likely losing data privacy protection.

ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e8e8a0d-c1ef-47b1-ac42-3e1bd4654d5c/etsi-gr-pdl-009-v1-1-1-2021-09

As a distributed ML technology and a type of federated data computing, Federated Learning (FL) was to implement a distributed ML model training process by multiple FL participants while still ensuring data privacy, security and legal compliance. Using FL-based mobile keyboard prediction as an example, FL usually consists the following steps:

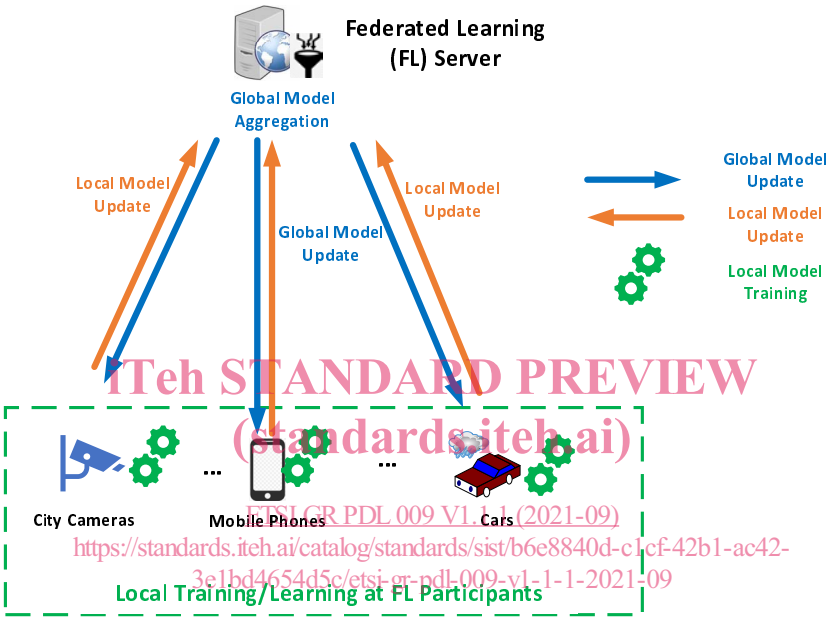
- Step 1: Mobile phones as FL participants participating in an FL task first download initial training model (i.e. the initial global model) from an FL Server.
- Step 2: Each mobile phone conducts the local training over its local data to train the model and generate its local model (or model update).
- Step 3: After the local model is trained, the mobile phone uploads the encrypted local model update (i.e. gradients) to the FL sever.
- Step 4: The FL server aggregates all local model updates collected from multiple mobile phones to obtain a new/updated global model. The updated global model will be then further sent to each mobile phone for the next round of training (Similar to Step 1).
- Overall, steps 1-4 will be executed for multiple rounds to improve the global model with expected quality and/or other requirements.

From the above process, it can be seen that FL can make full use of the data and computing power of the FL participants. Multiple parties (i.e. participants) can collaborate to build a more robust ML model without sharing/moving their data. This is very important for ML tasks when a strict data law/supervision is enforced. For example, the General Data Protection Regulation (GDPR) in Europe puts forward strict requirements on the storage, use, and transfer of users' private data. Therefore, FL can be used to solve key issues such as data ownership, data privacy, and data access rights in this environment.

Consider a general use case of smart city and smart transportation as shown in Figure 4.3-1:

- In smart city applications, many cameras will be deployed on streets and generate continuous data or data streams. These urban camera data can be used to train an ML model for urban environmental monitoring and predicting. However, uploading all camera data to cloud could be cumbersome or unrealistic. Accordingly, FL is a more feasible and efficient method.

- Similarly, in smart transportation applications, there will be a large number of vehicles driving on the road, and each vehicle will generate massive real-time driving data. These data can be trained to generate many ML models (e.g. to predict which road sections or during which time periods vehicles are most likely to have poor driving behaviour/performance). However, these data are not only large in quantity, but also contain personal privacy information; as a result, it is unwise or inefficient to upload these data to a cloud for centralized processing/training as in traditional ML. FL can be applied in this use case such that a global ML model can be jointly trained by vehicles without uploading driving data from vehicles to cloud.



iTeh STANDARD PREVIEW
(standards.iteh.ai)
ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e8840d-c1cf-42b1-ac42-3e1bd4654d5c/etsi-gr-pdl-009-v1-1-1-2021-09

Figure 4.3-1: Federated Learning in Smart City and Smart Transportation

4.4 Multi-Party Computation Use Case

Secure Multi-Party Computation (MPC) was originally introduced in [i.2] in the form of "The Millionaire's Problem". Since then, many advances have been made in both MPC theories and practical MPC deployments [i.3], [i.4].

In a general setting of MPC, there are n parties. Each party *Pi* hosts its own input data *xi*. They want to jointly compute a function to get a result: *result=f(x1, x2, …, xn)* with the requirement that no party can know or deduce input data hosted by other parties. In other words, all parties only know the function and the computed result. Figure 4.4-1 shows such a general MPC structure as an example of MPC use cases, which consists of the following procedures:

- Step 1: Parties encrypt their input data.

- Step 2: Parties exchange their encrypted input data.

- Step 3: One (or multiple) party computes the function over received encrypted input data from other parties to generate a temporary result.

- Step 4: The temporary result is sent to other parties.

- Step 5: One (or multiple) party computes the function over the temporary result to generate the final result.

- Step 6: The final result is sent to other parties.

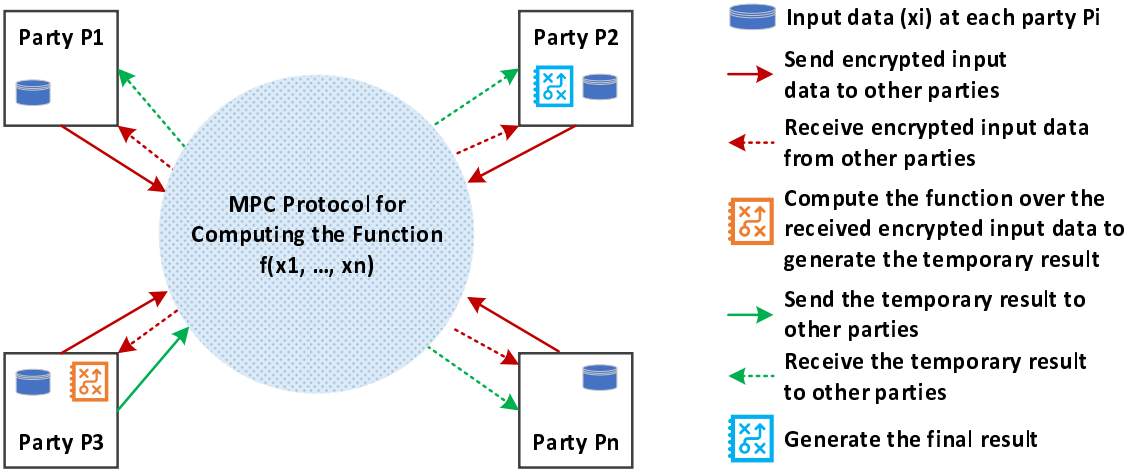


Figure 4.4-1: General Multi-Party Computation

4.5 Federated Data Discovery and Sharing

As a critical stage of federated data management pipeline, federated data discovery and sharing refers to the process, where data discovery cannot be solely served by a single organization, but served by multiple organizations. In other words, a federated data discovery request will trigger data lookup operations on data maintained locally by different organizations, and discovery results from each organization will be combined or aggregated as the final result for the federated data discovery request.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

Figure 4.5-1 illustrates a federated data discovery scenario, where a user (e.g. a researcher) can discover data (e.g. genomic data) from multiple organizations (e.g. hospitals). In other words, the user's data discovery request will not be served by a single organization, but served by multiple, unnecessarily trusted, organizations [i.5], [i.6]. This scenario consists of the following steps:

ETSI GR PDL 009 V1.1.1 (2021-09)
https://standards.iteh.ai/catalog/standards/sist/b6e8840d-c1cf-42b1-ac42-3e1bd4b54a8a/etsi-gr-pdl-009-v1.1.1-2021-09

- Step 1: A user (e.g. a doctor or a researcher) issues an initial data discovery request to a Federated Data Discovery and Sharing Service ("FDDSS"), which is a logical function and has access to data maintained locally at different organizations. It is assumed that the user knows the address of FDDSS (e.g. through pre-configuration or provisioning).

- Step 2: FDDSS could simply forward the initial data discovery request to organizations (e.g. Organization-1, Organization-2 and Organization-3); alternatively, it could transform the initial data discovery request to multiple transformed data discovery requests and forward each transformed data discovery request to a different organization. Within this step, FDDSS could first authenticate and authorize if the user has the right to leverage the discovery service. Then, FDDSS could enforce certain access control rules limiting data discovery based on access criteria. As an example, access control rules could specify the list of data types or items that are not discoverable.

- Step 3: Each organization receives a separate data discovery request from FDDSS. The organization will authenticate and authorize the data discovery request, look up the data maintained locally against any discovery criteria contained in the data discovery request, and generate discovery result. If any data cannot be discovered (e.g. due to confidential or privacy considerations), the organization could reject the data discovery request and/or exclude such data from the discovery result.

- Step 4: FDDSS receives data discovery results from multiple organizations, aggregates these results, generates an aggregated result, and forwards the aggregated result to the user.