

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Y.3602

(12/2018)

SERIES Y: GLOBAL INFORMATION
INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS,
NEXT-GENERATION NETWORKS, INTERNET OF
THINGS AND SMART CITIES

Cloud Computing

**Big data – Functional requirements for data
provenance**

Recommendation ITU-T Y.3602

ITU-T



ITU-T Y-SERIES RECOMMENDATIONS

GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

GLOBAL INFORMATION INFRASTRUCTURE

General	Y.100–Y.199
Services, applications and middleware	Y.200–Y.299
Network aspects	Y.300–Y.399
Interfaces and protocols	Y.400–Y.499
Numbering, addressing and naming	Y.500–Y.599
Operation, administration and maintenance	Y.600–Y.699
Security	Y.700–Y.799
Performances	Y.800–Y.899

INTERNET PROTOCOL ASPECTS

General	Y.1000–Y.1099
Services and applications	Y.1100–Y.1199
Architecture, access, network capabilities and resource management	Y.1200–Y.1299
Transport	Y.1300–Y.1399
Interworking	Y.1400–Y.1499
Quality of service and network performance	Y.1500–Y.1599
Signalling	Y.1600–Y.1699
Operation, administration and maintenance	Y.1700–Y.1799
Charging	Y.1800–Y.1899
IPTV over NGN	Y.1900–Y.1999

NEXT GENERATION NETWORKS

Frameworks and functional architecture models	Y.2000–Y.2099
Quality of Service and performance	Y.2100–Y.2199
Service aspects: Service capabilities and service architecture	Y.2200–Y.2249
Service aspects: Interoperability of services and networks in NGN	Y.2250–Y.2299
Enhancements to NGN	Y.2300–Y.2399
Network management	Y.2400–Y.2499
Network control architectures and protocols	Y.2500–Y.2599
Packet-based Networks	Y.2600–Y.2699
Security	Y.2700–Y.2799
Generalized mobility	Y.2800–Y.2899
Carrier grade open environment	Y.2900–Y.2999

FUTURE NETWORKS

	Y.3000–Y.3499
--	---------------

CLOUD COMPUTING

Y.3500–Y.3999

INTERNET OF THINGS AND SMART CITIES AND COMMUNITIES

General	Y.4000–Y.4049
Definitions and terminologies	Y.4050–Y.4099
Requirements and use cases	Y.4100–Y.4249
Infrastructure, connectivity and networks	Y.4250–Y.4399
Frameworks, architectures and protocols	Y.4400–Y.4549
Services, applications, computation and data processing	Y.4550–Y.4699
Management, control and performance	Y.4700–Y.4799
Identification and security	Y.4800–Y.4899
Evaluation and assessment	Y.4900–Y.4999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T Y.3602

Big data – Functional requirements for data provenance

Summary

Recommendation ITU-T Y.3602 describes a model and operations for big data provenance. Also, this Recommendation provides the functional requirements for big data service provider (BDSP) to manage big data provenance. The reliability of data is an important factor in determining the reliability of the analysis result. Data provenance aims to ensure the reliability of data by providing transparency of the historical path of the data. In a big data environment, complex data processing and migration due to the big data lifecycle and data distribution cause various difficulties in managing data provenance.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.3602	2018-12-14	13	11.1002/1000/13817

Keywords

Big data, data provenance, provenance model, provenance operation, provenance requirements, use case.

* To access the Recommendation, type the URL <http://handle.itu.int/> in the address field of your web browser, followed by the Recommendation's unique ID. For example, <http://handle.itu.int/11.1002/1000/11830-en>.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2019

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

	Page
1 Scope.....	1
2 References.....	1
3 Definitions	1
3.1 Terms defined elsewhere	1
3.2 Terms defined in this Recommendation.....	1
4 Abbreviations and acronyms	1
5 Conventions	2
6 Introduction to data provenance	2
6.1 General concept of data provenance.....	2
6.2 Data provenance in big data ecosystem.....	3
7 Overview of big data provenance	4
7.1 Data provenance in big data ecosystem.....	4
7.2 Conceptual model of big data provenance information	5
7.3 Operations of provenance information.....	7
7.4 Logical components for big data provenance management	10
8 Functional requirements of big data provenance.....	11
8.1 Provenance lifecycle requirements.....	11
8.2 Analysis support requirements	12
8.3 Monitoring requirements	12
8.4 Policy management requirements.....	12
9 Security considerations	13
Appendix I – Use cases of big data provenance	14
Bibliography.....	23

Recommendation ITU-T Y.3602

Big data – Functional requirements for data provenance

1 Scope

This Recommendation specifies the functional requirements for data provenance in a big data ecosystem as defined in [ITU-T Y.3600]. This Recommendation introduces data provenance as well as data provenance in big data ecosystem, and provides a conceptual model, operations, logical components, and functional requirements for big data provenance. The functional requirements provided in this Recommendation are derived from use cases.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.3600] Recommendation ITU-T Y.3600 (2015), *Big data – Cloud computing based requirements and capabilities*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 big data [ITU-T Y.3600]: A paradigm for enabling the collection, storage, management, analysis and visualization, potentially under real-time constraints, of extensive datasets with heterogeneous characteristics.

NOTE – Examples of datasets characteristics include high-volume, high-velocity, high-variety, etc.

3.1.2 provenance [b-ITU-T X.1255]: Information pertaining to any source of information including the party or parties involved in generating it, introducing it and/or vouching for it.

3.2 Terms defined in this Recommendation

This Recommendation defines the following term:

3.2.1 big data provenance: Information that records the historical path of data according to the data lifecycle operations in a big data ecosystem.

NOTE 1 – Data lifecycle operations include data generation, transmission, storage, use, and deletion.

NOTE 2 – Data provenance information provides details about the source of data, such as the person responsible for the provision of data, functions applied to data, and information about the computing environment for data processing (e.g., operating system, description of the hardware, locale settings and time zone).

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

BD Big Data

BDC	Big Data service Customer
BDSP	Big Data Service Provider
DB	Data Broker
DP	Data Provider
DS	Data Supplier
H/W	Hardware
OS	Operating System
PI	Provenance Information
PII	Personally Identifiable Information
URI	Uniform Resource Identifier

5 Conventions

In this Recommendation:

The keywords "**is required to**" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this document is to be claimed.

The keywords "**is recommended**" indicate a requirement which is recommended but which is not absolutely required. Thus this requirement need not be present to claim conformance.

The keywords "**can optionally**" indicate an optional requirement which is permissible, without implying any sense of being recommended. This term is not intended to imply that the vendor's implementation must provide the option and the feature can be optionally enabled by the network operator/service provider. Rather, it means the vendor may optionally provide the feature and still claim conformance with the specification.

In the body of this document and its annexes, the words shall, shall not, should, and may sometimes appear, in which case they are to be interpreted, respectively, as is required to, is prohibited from, is recommended, and can optionally. The appearance of such phrases or keywords in an appendix or in material explicitly marked as informative are to be interpreted as having no normative intent.

6 Introduction to data provenance

6.1 General concept of data provenance

The reliability of data used is an important factor to determine the trustworthiness of a data analysis outcome. Indeed, data can be manipulated and transformed according to the intent of the analyst and distorted in order to extract the desired result. In this sense, the data provenance aims to ensure the reliability of data and analysis results by providing transparency of the historical path of the data.

Provenance is information pertaining to any source of information, including the party or parties involved in generating it, introducing it and/or vouching for it. In the field of data management, data provenance is information about the origin and creation process of data with:

- data product;
NOTE 1 – A data product is the output data production for distribution (open or sell) purpose.
- process that enable the creation of data;
NOTE 2 – A process is described by the applied functions on data source, intermediate outputs and their order.
- metadata recording process of workflow, annotations, notes about processes; and,

- information that helps determine derivation history of a data product, starting from its original sources.

Data provenance is useful for:

- managing derivation history of a data product starting from its original sources;
- ascertaining quality of data based on ancestral data and derivation;
- tracking back sources of errors;
- allowing automated re-enactment of derivations to update a data;
- providing attribution of data sources.

Provenance information (PI) is composed of a set of data flows, and each flow contains information of processes (f), data sources (d) and responsible parties (p). In this sense, PI is notated as:

$$PI = \{(f, p), (d, p)\}$$

A data flow is divided into a directly associated flow and subordinately associated flow. For example, in Figure 6-1, the provenance information (PI) about *Data d* is composed by a set of:

- directly associated flow: $PI = \{(f2, pC), (Data\ c, pC)\}$;
- subordinately associated flow: $PI (Data\ c) = \{(f1, pC), ((Data\ a, pA), (Data\ b, pB))\}$

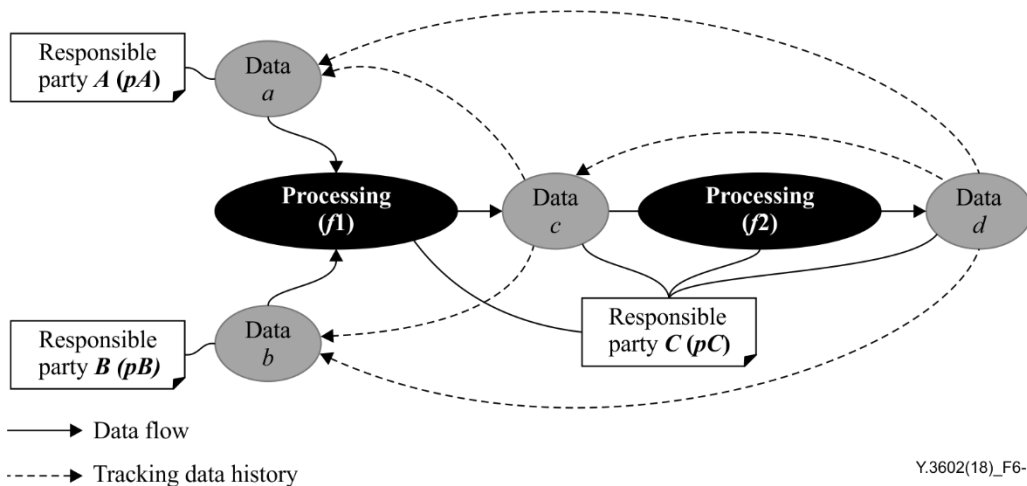


Figure 6-1 – An example of data provenance information

6.2 Data provenance in big data ecosystem

In a big data environment, complex data processing and migration due to the big data lifecycle operations (e.g., data generation, transmission, storage, use, and deletion) and data distribution cause various difficulties in managing the data provenance. According to the big data ecosystem described in [ITU-T Y.3600], big data provenance needs to treat:

- huge volumes of non-structured, semi-structured and structured data;
- functions description for various types and formats of data;
- data traceability across multi-application domains.

NOTE 1 – Application domain is an area of knowledge or activity applied for one specific economic, commercial, social or administrative scope [b-ITU-T Y.4100].

NOTE 2 – Transport application domain, health application domain and government application domain are examples of application domains.

In addition, big data computing environment causes several challenges for data provenance such as:

- **efficient storing mechanism for provenance data:** The size of provenance data can be larger than the original data, causing storage overhead;

- **minimize provenance collection overhead:** In a distributed system environment, consideration of the recording provenance and computation cost together is important;
- **reproduce an execution from provenance for big data applications:** In some case of big data execution, the environment information (e.g., hardware (H/W) information and parameter configuration of big data engines) is an important factor.

The application area of big data provenance and its benefits are:

- **collaborative big data analysis:** Big data provenance allows collaboration of big data analysis among multiple domains or applications by data sources information and their process steps;
- **reuse of data processing:** Generally, a big data analysis has complex process steps. Thus, a well-defined analysis model which can be derived from provenance information is helpful for a similar case of big data processing;
NOTE 3 – In data processing system, data processing means a course of events occurring according to an intended purpose of effect.
- **automating big data analysis process:** Provenance gives a context in which to use the data, and allows automated validation and revision of derived data when the base data is updated;
- **audit and protect intellectual property:** Provenance gives a lineage of data, and it allows auditing and tracing of digital rights on mash-up data.

7 Overview of big data provenance

This clause presents an overview of big data provenance. This clause describes data provenance in a big data ecosystem, a conceptual model, provenance operations, and logical components for big data provenance.

7.1 Data provenance in big data ecosystem

According to [ITU-T Y.3600], a big data service provider (BDSP) supports data provenance as a part of data management by managing information about the origin and generation process methods of data, including the party or parties involved in the generation, introduction and/or mash-up processes for data.

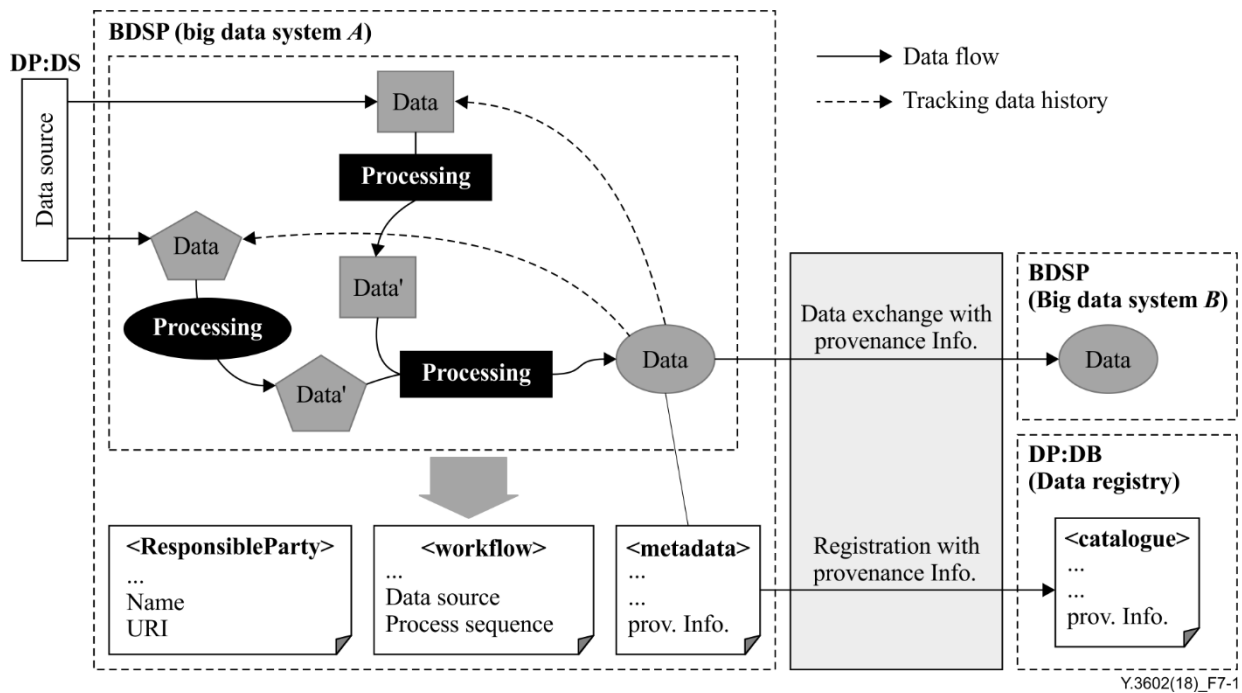


Figure 7-1 – Using data provenance in big data ecosystem

Figure 7-1 shows the use of data provenance in big data ecosystem:

- when data is imported from an outside data source (data provider (DP):data supplier (DS)) and stored, BDSP (big data system A) generates the metadata based on importing context (e.g., responsible party information, time, size) and these metadata entities are used for provenance information;
- BDSP A monitors and stores a process of data mash-up or analysis as a form of provenance information to ensure the reliability of data quality and reproducibility of analysis result;
- when BDSP A exports data to BDSP B or registers data catalogue to a data registry in data market (DP:data broker (DB)), BDSP A delivers provenance information.

NOTE – When a BDSP exports or registers data with its provenance information, BDSP manages the level of detail through simplification of provenance information based on their own data or service policy.

7.2 Conceptual model of big data provenance information

Big data provenance information is an extension of the general data provenance concept, which is described in clause 6.1.

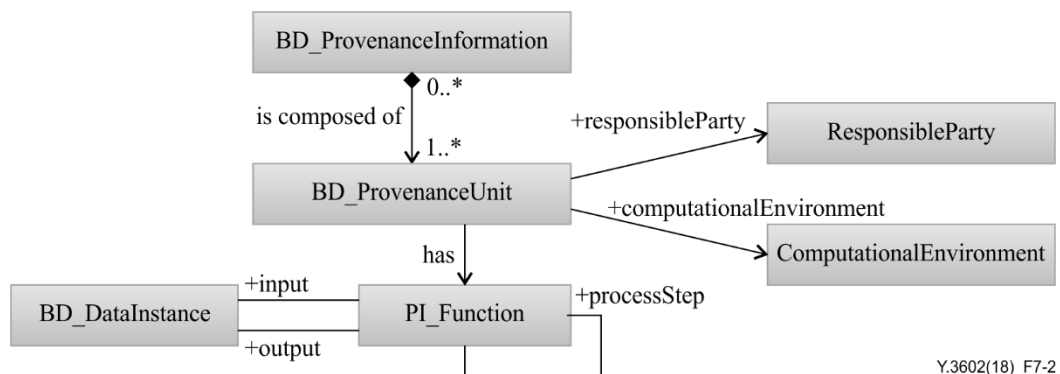


Figure 7-2 – Conceptual model for big data provenance information

Figure 7-2 shows a high-level conceptual model for big data provenance information. Big data provenance information (*BD_ProvenanceInformation*) is an aggregated set of big data provenance units (*BD_ProvenanceUnit*) which records a history of the most recent changes to the data.

Big data provenance unit is a minimum set of big data provenance information. It provides information about data ownership or authority (*ResponsibleParty*), data processing environment (*ComputationalEnvironment*), and a sequence of functions (*PI_Function*) with input and output data (*BD_DataInstance*) which are involved in data mash-up or analysis.

NOTE 1 – A workflow depicts the actual sequence of the functions to describe a data processing. In the big data provenance information model, a workflow can be derived by the association *+processStep* which describes the sequence of *PI_functions*.

NOTE 2 – *BD_DataInstance* is a metadata composed of identifiable information (e.g., access information, type and format of data, data size, date, personally identifiable information (PII)) for a data instance.

Figure 7-3 illustrates an example of capturing the provenance unit on Data C.

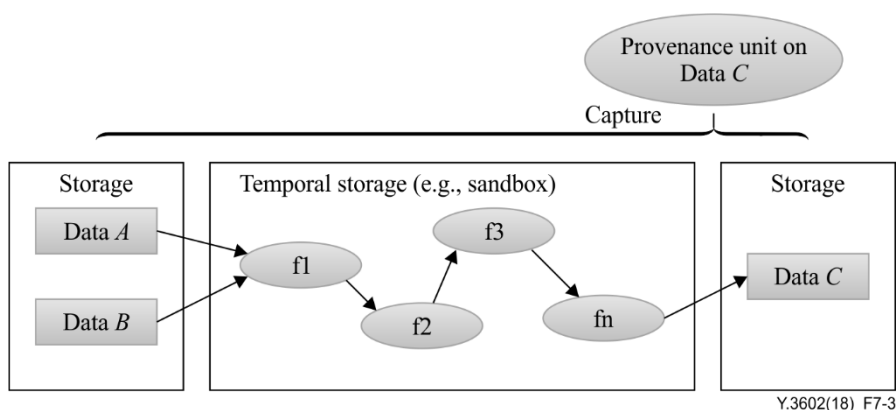
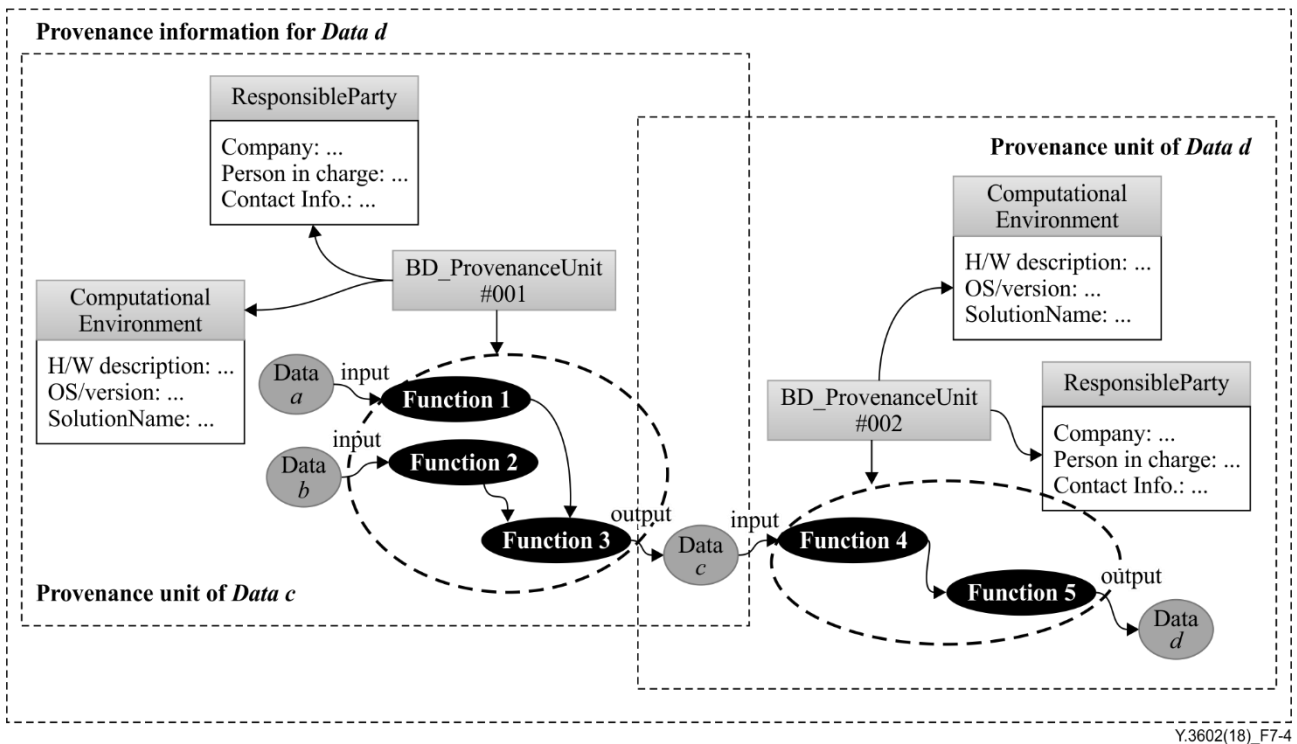


Figure 7-3 – An example of capturing provenance unit on data instance

The capturing of the provenance unit occurs simultaneously when Data C is stored in the data storage. Even though all functions have input and output data, the information about the first input data and output data are captured in a provenance unit with a process steps described by a sequence of functions.

NOTE 3 – In the Figure 7-3, Data A and B are input data, and Data C is output data.

Figure 7-4 shows an example of big data provenance information with a graph model. The provenance information of Data d consists of provenance unit of Data c and provenance unit of Data d by aggregating two units.

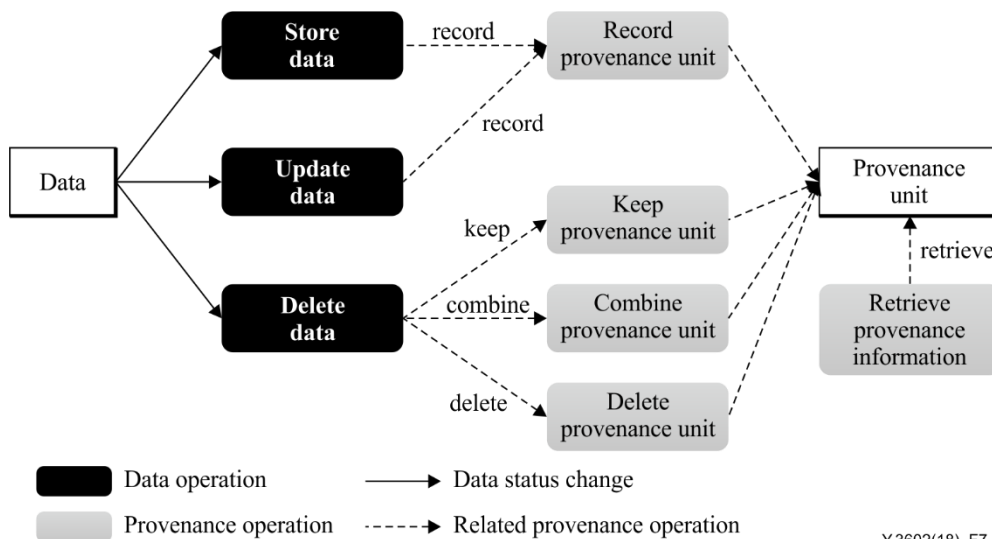


Y.3602(18)_F7-4

Figure 7-4 – An example of big data provenance information

7.3 Operations of provenance information

According to the change of the data state such as storing data, updating data and deleting data, a provenance unit is recorded, kept, combined or deleted. Figure 7-5 shows the relationship between data state change and provenance operations.



Y.3602(18)_F7-5

Figure 7-5 – Provenance operations according to data state change

Operations to manage provenance information are:

- record provenance unit (see clause 7.3.1);
- keep provenance unit (see clause 7.3.2);
- combine provenance unit (see clause 7.3.2);
- delete provenance unit (see clause 7.3.2);

- retrieve provenance information (see clause 7.3.3).

7.3.1 Provenance information in case of storing and updating data

Figure 7-6 shows provenance units recording according to the data storing and updating.

- **provenance unit recording:** When Data 1 are stored, BDSP records Provenance unit 1. Data 1 is updated to Data 2 and 3, BDSP records Provenance unit 2 and Provenance unit 3 sequentially.

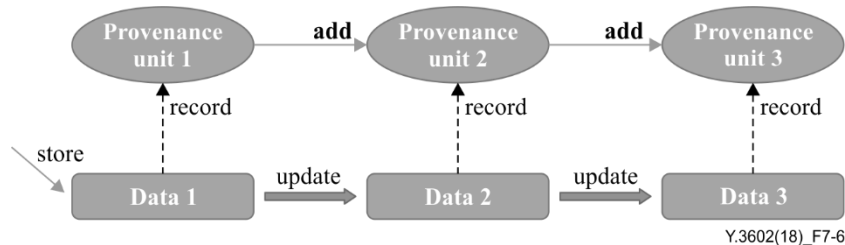


Figure 7-6 – Record provenance units

7.3.2 provenance information in case of deleting data

In case of deleting data, the provenance management system provided by BDSP acts in the three ways described in clauses 7.3.2.1 to 7.3.2.3.

7.3.2.1 Keep provenance unit

When data are deleted from storage, the provenance management system keeps its provenance unit to support traceability of data within process steps. Figure 7-7 shows an example where provenance units are kept after the deletion of data.

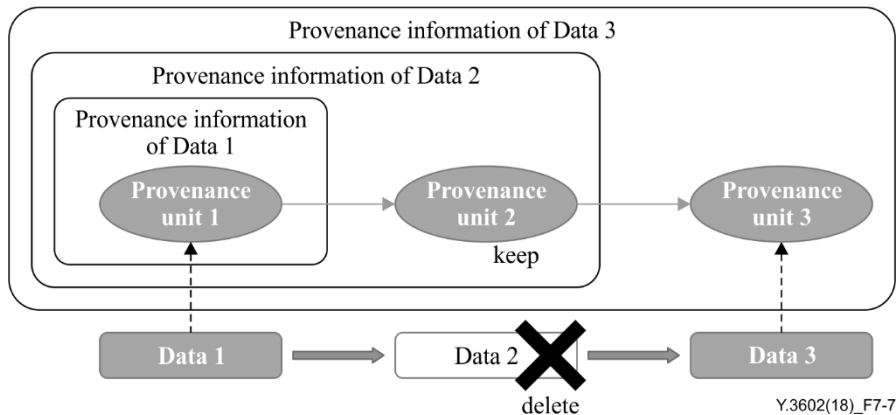


Figure 7-7 – Keep provenance unit to support data traceability

7.3.2.2 Combine provenance units

When data are deleted from storage, the provenance management system combines the data's provenance unit with the forward nearest provenance unit within process steps. This is described in the example given in Figure 7-8.

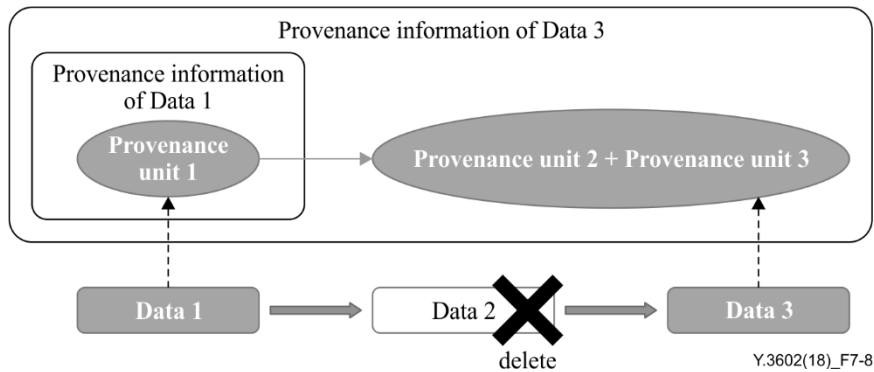


Figure 7-8 – Combine provenance units to support data traceability

NOTE – In Figure 7-8, the forward nearest provenance unit of provenance unit 2 is provenance unit 3.

7.3.2.3 Delete provenance unit

When data are deleted from storage, and if the data instance is placed at the right end node of process steps, the provenance management system deletes the provenance unit together with the data. This is illustrated in the example given in Figure 7-9.

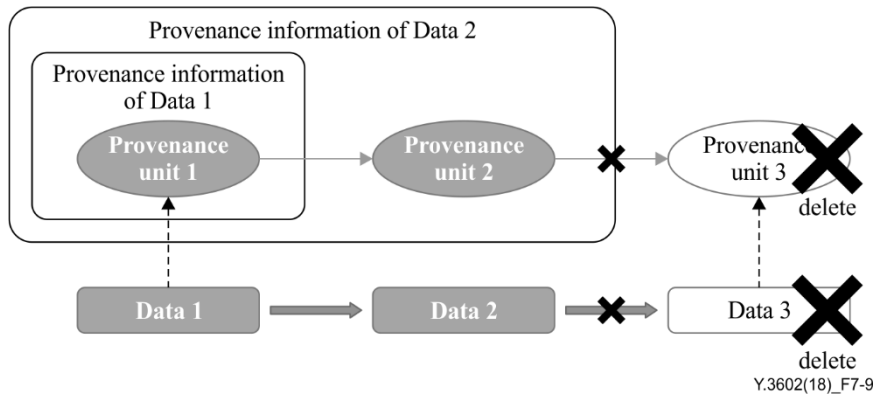


Figure 7-9 – Delete provenance units

NOTE – In Figure 7-9, the right end node of process steps is Data 3.

7.3.3 retrieval of provenance information

Figure 7-10 shows an example where a provenance unit is retrieved. As shown in the figure, when the provenance information of Data 3 is requested from an application (step 1), the BDSP traces the history of data based on each provenance unit (step 2). The BDSP aggregates the identified provenance units (step 3), and provides responses to the provenance information of Data 3 (step 4).

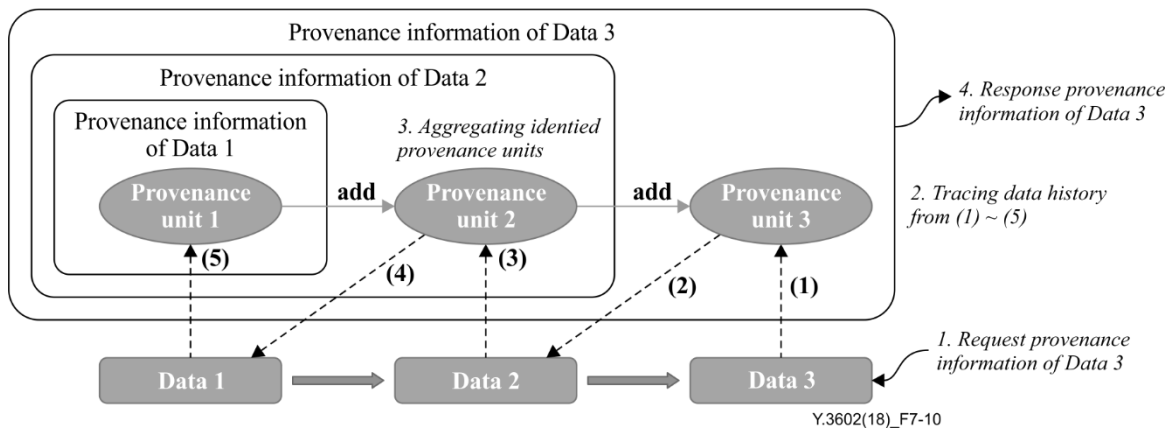


Figure 7-10 – Retrieval of provenance information

7.4 Logical components for big data provenance management

The configuration of logical components for managing big data provenance consists of provenance model management, provenance lifecycle management, analysis support, monitoring, provenance sharing policy management and personally identifiable information management, as shown in Figure 7-11.

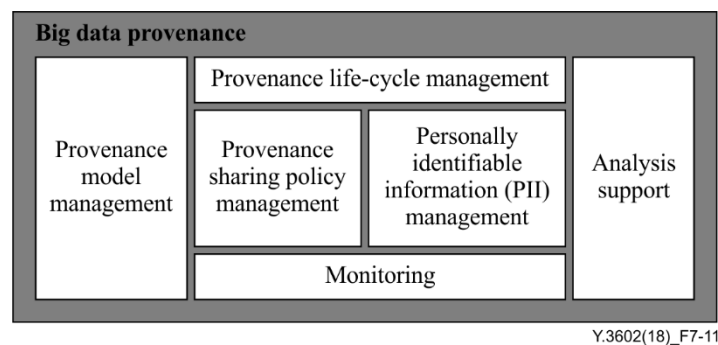


Figure 7-11 – Configuration of logical components for big data provenance

The logical components shown in Figure 7-11 are as follows:

- **provenance model management.** This logical component manages provenance information compatibility among different BDSPs. This logical component validates the provenance information transmitted from outside based on the big data provenance model (see clause 7.2). The valid provenance information is then encoded as a common model and delivered to provenance lifecycle management component to store it;
- **provenance lifecycle management.** This component performs the recording and deletion of provenance information according to store, update and delete data (see clauses 7.3.1 and 7.3.2). This logical component supports retrieving provenance information (see clause 7.3.3);
- **analysis support.** This logical component extracts the workflows from provenance information, and stores them. From the stored workflows, this logical component retrieves the candidate analysis workflows based on the information of BDSP's data analysis functions and data. For the request of provenance information or workflow from the different system (e.g., external BDSP), this logical component may check the adaptability of the computational environment, and map to an equivalent function for that system. This logical component also supports automating data analysis process based on update of data, adding user annotation on provenance information, and managing the relationship between BDSP's functions and data;

NOTE – Based on the relationship between functions and data in provenance information, it is possible to query the list of available data with functions, and the list of functions applicable to the data.

- **provenance sharing policy management.** This logical component manages multiple sharing policies on provenance information. When exporting a provenance information, a BDSP checks the sharing policy and may simplify it before sending to another BDSP;
- **personally identifiable information (PII) management.** This logical component checks whether data instance contains PII when recording a provenance unit. This logical component also requests a protection mechanism to BDSP on provenance information that includes PII;
- **monitoring.** This logical component monitors changes in value about computational environment and responsible party in provenance information. When changes are detected, this logical component updates them.

8 Functional requirements of big data provenance

8.1 Provenance lifecycle requirements

Provenance lifecycle requirements include:

- **(provenance model description)** It is required that BDSP supports the model for big data provenance information;

NOTE 1 – Big data provenance information model includes function name and its uses, computational environment, data type and format of input and output data, input parameters, responsible party information, etc.

NOTE 2 – Example of computational environment information is OS, H/W description, locale settings, and time zone, etc.

- **(common format for exchange)** It is recommended that BDSP supports encoding and decoding a provenance information in a common format for use on different systems;

NOTE 3 – In this Recommendation, the meaning of encoding is the process of converting provenance information into a specialized format. Decoding is the opposite process.

- **(provenance recoding initiation)** It is required that BDSP records provenance unit when data is stored;

NOTE 4 – The information contained in the metadata (from DP:DB or generated by BDSP) can be used for recoding provenance unit.

- **(storing provenance unit)** It is required that BDSP supports a cost-efficient storing mechanism for provenance units;

NOTE 5 – In case of recording provenance information of streaming data, for the efficient storage usage, it is needed to designate a predetermined period of time to record provenance unit, rather than recording it every time data are stored. Data compression techniques can also be considered.

- **(storing provenance information)** BDSP can optionally support pre-storing provenance information prior to request time to reduce retrieval time;

- **(searching provenance unit)** It is required that BDSP supports searching a provenance unit;

- **(Combining provenance units)** It is required that BDSP supports combining of provenance units;

NOTE 6 – In case of deleting data, a provenance unit needed to combine (see clause 7.3.2).

- **(retrieving provenance information)** It is required that BDSP supports provenance unit aggregation to retrieve a provenance information;

- **(deleting provenance unit)** It is required that BDSP provides a provenance unit deletion mechanism.

NOTE 7 – In case of deleting data, BDSPP acts with three mechanisms on the provenance unit (keep, combine, delete) based on the context (see clause 7.3.2).

NOTE 8 – The BDSPP can maintain the associated provenance unit even if the data are deleted, which is subject to management policy.

8.2 Analysis support requirements

Analysis support requirements include:

- **(extracting workflow)** It is required that BDSPP provides extraction of workflow information from a provenance information;

- **(storing workflow)** It is recommended that BDSPP supports storing workflow;

NOTE 1 – The workflow is stored in forms of graph, which is organized with the usage frequency of the analysis functions and sequential relationship among them.

- **(retrieving workflow)** It is recommended that BDSPP supports workflow retrieval;

- **(providing data list on function)** It is recommended that BDSPP provides a list of data related to a given function recorded in given workflow;

- **(providing function list on data)** It is recommended that BDSPP provides a list of functions related to a given data recorded in given workflow;

- **(data analysis automation)** It is recommended that BDSPP supports analysis automation based on workflow;

- **(user annotation)** BDSPP can optionally support annotation on provenance information;

- **(equivalent function for process steps)** It is recommended that BDSPP provides an equivalent function mapping for reusing provenance information coming from a different system;

NOTE 2 – For the equivalent function mapping, the name of the function, the format and structure of input and output data of this function, the frequency of analysis functions and the relationship among them can be used.

NOTE 3 – The results of the equivalent function mapping can be the same function with different names or a combination of functions that provide the same output.

- **(adaptability of computational environment)** It is recommended that BDSPP provides diagnose computational environment to reuse provenance information which came from the different system.

8.3 Monitoring requirements

Monitoring requirements include:

- **(monitoring computational environment)** It is required that BDSPP monitors the change of computational environment;

- **(monitoring responsible party)** It is required that BDSPP monitors the change of responsible party;

- **(applying the monitoring result)** It is required that BDSPP reflects the monitoring results to the recorded provenance unit.

NOTE – The monitoring results include the change of computational environment and responsible party.

8.4 Policy management requirements

Policy management requirements include:

- **(verifying PII)** It is required that BDSPP provides verifying PII in a data instance when recording a provenance unit;

NOTE 1 – Verification of PII follows BDSP's policy on PII.

NOTE 2 – In a provenance unit, data instance information (*BD_DataInstance*) includes information about whether PII is contained or not (see clause 7.2).

- **(protecting PII)** It is required that BDSP provides a protection mechanism for a PII in data;

NOTE 3 – When a PII is included in data sources, BDSP decide omit it or not based on the user's access authority.

- **(simplifying provenance information)** It is recommended that BDSP supports simplifying provenance information based on a sharing policy;

NOTE 4 – Methods of provenance information simplification include multiple level of detail and encoding formats, etc.

- **(sharing level of provenance)** It is required that BDSP supports sharing policy according to the different levels of provenance.

NOTE 5 – The provenance level decides traceability of data, and it is determined by the sharing policy. Provenance information contains process step with the applied functions, intermediate data, and responsible party information. For the transfer the provenance information, the provenance information can be simplified according to the sharing policy.

9 Security considerations

Relevant security requirements of [b-ITU-T Y.2201], [b-ITU-T Y.2701] and applicable X, Y and M series of ITU-T Recommendations need to be taken into consideration, including access control, authentication, data confidentiality, data retention policy, network security, data integrity, availability and protection of personal information.

Appendix I

Use cases of big data provenance

(This appendix does not form an integral part of this Recommendation.)

Table I.1 – Use case – Initiating provenance information

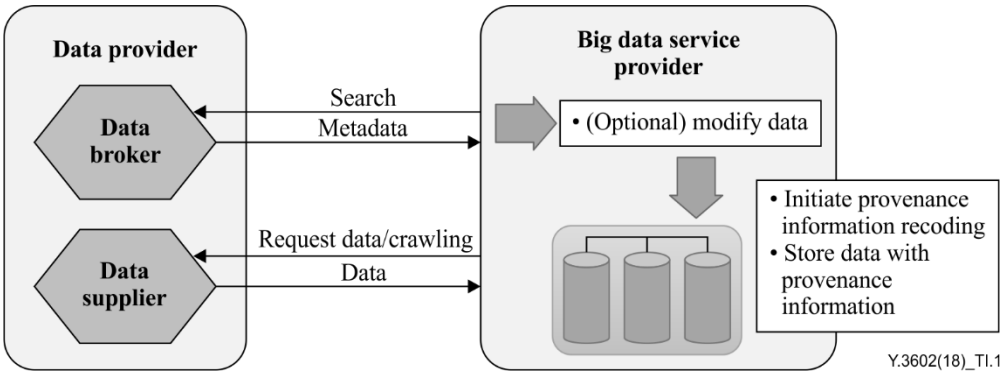
Title	Initiating provenance information record
Description	A BDSP requests data from a data provider (DP), and gets the data. At this time, BDSP stores data as it is or modifies the original data to fit for their own database. During this process, BDSP start is to record provenance information which includes the origin of data described in metadata, functional processes which were applied to modify data, and stores them to distributed database.
Roles/sub-roles	DP:DS DP:DB BDSP
Figure (optional)	 <p style="text-align: right; font-size: small;">Y.3602(18)_T1.1</p>
Pre-conditions (optional)	DP:DS published metadata to DP:DB BDSP searches data from DP:DB and request data to DP:DS or crawls data from DP:DS
Post-conditions (optional)	
Derived requirement	<ul style="list-style-type: none"> – Provenance recoding initiation (see clause 8.1); – Provenance model description (see clause 8.1); – Storing provenance unit (see clause 8.1); – User annotation (see clause 8.2); – Sharing level of provenance (see clause 8.4).

Table I.2 – Use case – Updating data and managing provenance information

Title	Updating data and managing provenance information in big data system
Description	<p><Adding provenance information based on the change of data source status></p> <ul style="list-style-type: none"> – A BDSP uses external data source regularly. When DP:DS update the status of data (e.g., schema version upgrade, change of responsible person's information), BDSP manages the change of source information, and checks existing data is still available or not; – BDSP searches the existing provenance information and adds the changing information; <p><Managing provenance information caused by deleting or preserving data from local data storage></p> <ul style="list-style-type: none"> – BDSP deletes or retires the stored data for storage efficiency or the other management issues. BDSP monitors the provenance information and decides to delete it or not.
Roles/Sub roles	<p>DP:DS BDSP</p>
Figure (optional)	<p>The diagram illustrates the interaction between a Data supplier (represented by a hexagon) and a Big data service provider (represented by a rounded rectangle containing three cylinders). The Data supplier sends a 'Request data regularly' message to the Big data service provider. The provider returns 'Data ver.1' and 'Data ver.2' to the supplier. The Big data service provider also performs internal actions: 'Managing data source information', 'Search provenance', 'Validate availability existing data', and 'Delete/preserve data'. The diagram is labeled 'Y.3602(18)_TI.2'.</p>
Pre-conditions (optional)	BDSP stored extensible markup language (XML) data with external uniform resource identifier (URI) from DP:DS.
Post-conditions (optional)	
Derived requirement	<ul style="list-style-type: none"> – Searching provenance unit (see clause 8.1); – Combining provenance units (see clause 8.1); – Deleting provenance unit (see clause 8.1); – Monitoring computational environment (see clause 8.3); – Monitoring responsible party (see clause 8.3); – Applying monitoring result (see clause 8.3).

Table I.3 – Use case – Sharing and aggregating provenance information

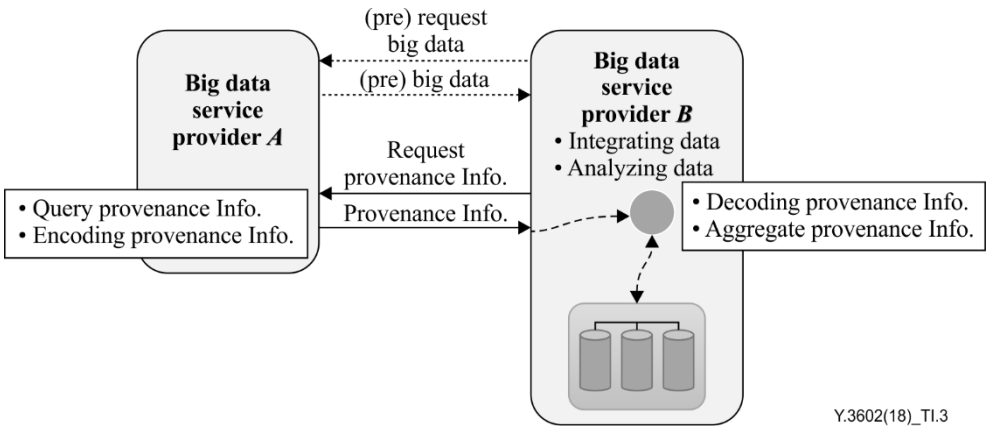
Title	Sharing and aggregating provenance information
Description	<p>Two collaborating BDSPs share provenance information with each other. When BDSP <i>B</i> uses data from BDSP <i>A</i>:</p> <ul style="list-style-type: none"> – BDSP <i>B</i> requests and receives data from BDSP <i>A</i>; – BDSP <i>B</i> initiating provenance information recoding with storing data from BDSP <i>A</i>; <p>If BDSP <i>B</i> needs more information (e.g., data history) of received data from</p> <ul style="list-style-type: none"> – BDSP <i>A</i>, then requests the provenance information about received data to BDSP <i>A</i>.; – BDSP <i>B</i> aggregates the delivered provenance information with the local one that was created when the data from BDSP <i>A</i> was stored.
Roles/sub-roles	– BDSP
	 <p style="text-align: right;">Y.3602(18)_TI.3</p>
Pre-conditions (optional)	BDSP <i>B</i> requested big data to BDSP <i>A</i> and received the data.
Post-conditions (optional)	BDSP <i>B</i> stored the aggregated provenance information.
Derived requirement	<ul style="list-style-type: none"> – Searching provenance unit (see clause 8.1); – Common format for exchange (see clause 8.1); – Retrieve provenance information (see clause 8.1); – Simplifying provenance information (see clause 8.4).

Table I.4 – Use case – Reuse data processing methods

Title	Reusing and automating data processing methods with big data provenance information
Description	<p>A data analyst is preparing an experiment based on existing big data analysis results with a different data source. To do this, the data analyst uses the provenance functions provided by the BDSP to extract the data analysis process and apply it.</p> <ul style="list-style-type: none"> – data analyst reviews analysis results; – data analyst selects an analysis result to reuse its processing methods; – data analyst extracts the data processing methods and related data from the provenance information of the analysis result and modifies them to fit for the new experiment; – data analyst applies the data processing method. <p>Data analyst is using analysis automation based on data updates by using the provenance information. Data analyst sets up the periodic analysis based on the updated data according to the update period of the data.</p>
Roles/Sub roles	<ul style="list-style-type: none"> – DP:DS – BDSP – big data service customer (BDC)
Figure (optional)	<p style="text-align: right;">Y.3602(18)_TI.4-1</p> <p style="text-align: center;"><Reusing data processing methods></p> <p style="text-align: center;">Y.3602(18)_TI.4-2</p> <p style="text-align: center;"><Automating big data analysis process></p>

Table I.4 – Use case – Reuse data processing methods

Pre-conditions (optional)	BDSP stores provenance information about the analysed result data.
Post-conditions (optional)	
Derived requirement	<ul style="list-style-type: none"> – Extracting workflow (see clause 8.2); – Retrieving workflow (see clause 8.2); – Data analysis automation (see clause 8.2); – Providing data list on function (see clause 8.2); – Providing function list on data (see clause 8.2).

Table I.5 – Use case – Managing personal information

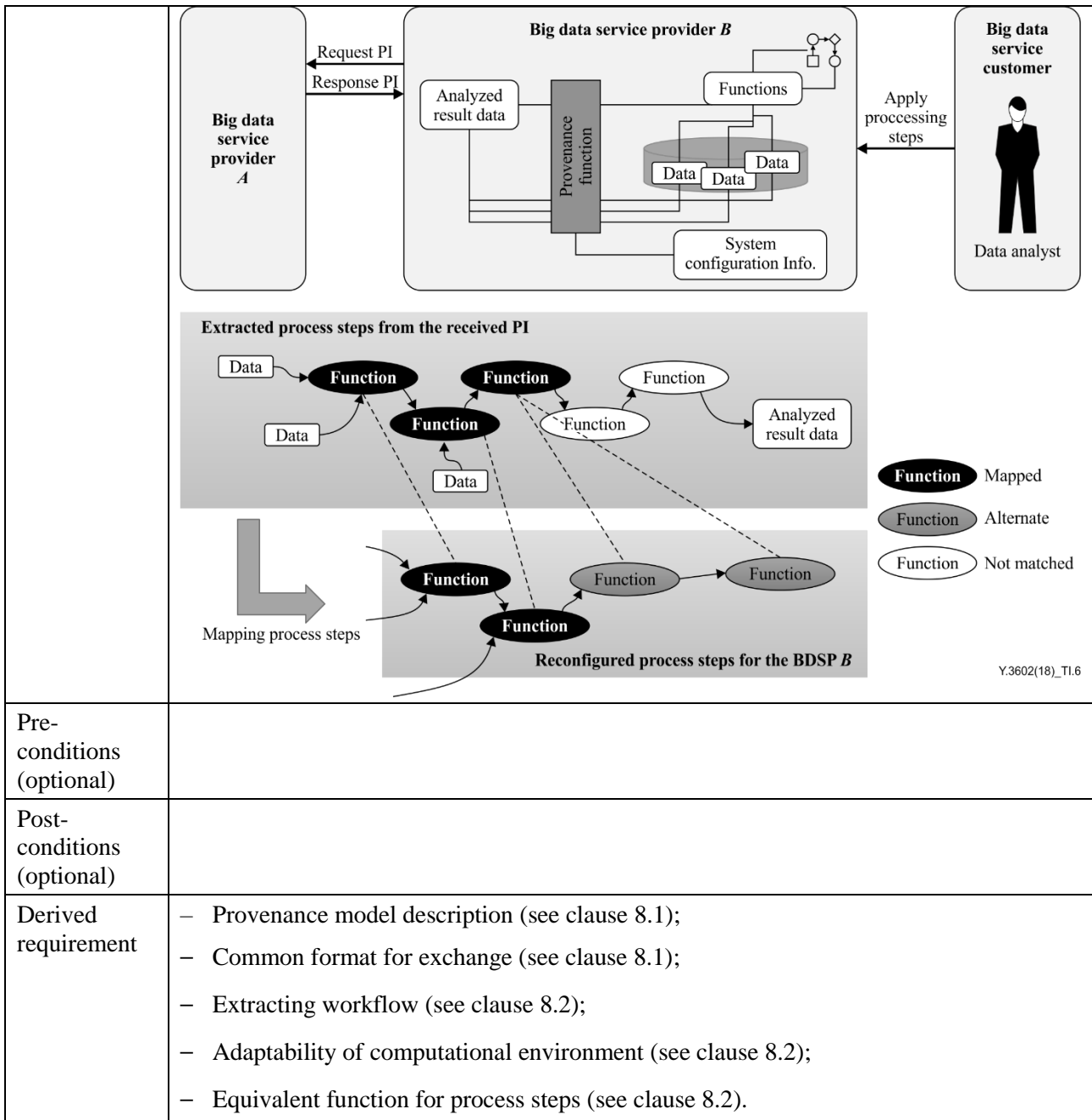
Title	Managing personal information
Description	<p>A data analyst (at a BDC) requests the provenance information on 'data 1' to BDSP. At this time, BDSP:</p> <ul style="list-style-type: none"> – traces history of 'data 1'; – checks whether each data includes personal information and find out 'data 3' includes personal information; – checks BDC's access right on 'data 3'. <p>If BDC has the access right on 'data 3', BDSP aggregates provenance units and responds it. If not, BDSP responds the provenance information excluding provenance unit 3 or abstracts the provenance information based on a predefined policy and responds it.</p>
Roles/Sub roles	<ul style="list-style-type: none"> – BDC – BDSP
Figure (optional)	

Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirement	<ul style="list-style-type: none"> – Retrieve provenance information (see clause 8.1) – Verifying PII (see clause 8.4) – Protecting PII (see clause 8.4) – Sharing level for provenance (see clause 8.4) – Simplifying provenance information (see clause 8.4)

Table I.6 – Use case – Reuse provenance information from the different analysis system

Title	Reusing provenance information from the different analysis system
Description	<p>A data analyst is preparing an experiment using the provenance information received from BDSP A with his/her own data. To this end, a data analyst uses the provenance functions (provided by the BDSP A) to extract the data analysis workflow and reconfigure it to fit BDSP B's analysis environment.</p> <ul style="list-style-type: none"> – data analyst requests the provenance information to BDSP B; – BDSP B decodes the provenance information; – BDSP B extracts the workflow from the provenance information; – BDSP B checks the adaptability of workflow and converts it to be available; <ul style="list-style-type: none"> A. BDSP B checks the adaptability of the computational environment of BDSP A. B. BDSP B maps the process steps extracted from the provenance information and the functions supported by the BDSP B. C. When the functions are not mapped correctly, BDSP B examines for the alternate functions, and data analyst selects the functions from them. – data analyst applies the process steps to his/her own data.
Roles/sub roles	<ul style="list-style-type: none"> – BDSP – BDC
Figure (optional)	

Table I.6 – Use case – Reuse provenance information from the different analysis system



Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirement	<ul style="list-style-type: none"> – Provenance model description (see clause 8.1); – Common format for exchange (see clause 8.1); – Extracting workflow (see clause 8.2); – Adaptability of computational environment (see clause 8.2); – Equivalent function for process steps (see clause 8.2).

Table I.7 – Use case –Scientific workflow provenance

Title	Provenance information collection and query of scientific workflow
Description	<p>Scientific workflow is a typical application system, facilitating e-Science. Scientists model, design, execute, debug, re-configure and re-run their analysis. Provenance information in the scientific workflow system is very useful for a scientist to interpret their workflow results and for other scientists to establish trust in the experimental result.</p> <p>The BDSP (scientific workflow system) initiates provenance information automatically, and stores the provenance information in database. BDC (scientists) need to retrieve the provenance information to confirm the source of the scientific data</p>

Table I.7 – Use case –Scientific workflow provenance

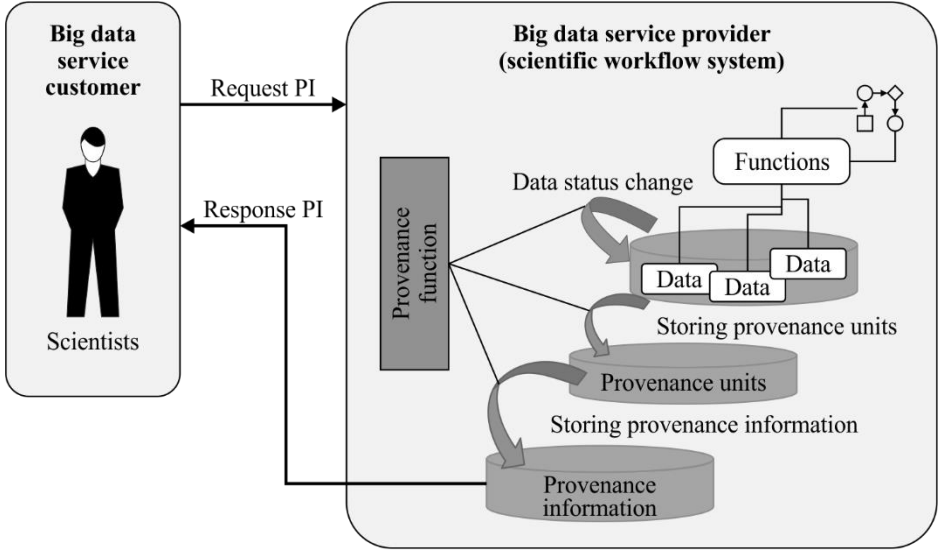
	on experiment or analysis process. The BDSP stores provenance information automatically when data status is changed to support for frequent retrieving.
Roles/sub-roles	<ul style="list-style-type: none"> – BDSP – BDC
Figure (optional)	 <p style="text-align: right; font-size: small;">Y.3602(18)_TI.7</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirement	<ul style="list-style-type: none"> – Retrieve provenance information (see clause 8.1) – Storing provenance information (see clause 8.1)

Table I.8 – Use case –Extracting analysis workflow from the accumulated provenance information

Title	Extracting analysis workflow from the accumulated provenance information
Description	<p>BDSP A collects the provenance information from the different BDSPs to accumulate the analysis workflows and reuse them. BDSP C want to find analysis methods that can be applied to its system functions and data through BDSP A.</p> <p>Accordingly:</p> <ul style="list-style-type: none"> – BDSP A collects provenance information from BDSP B; – BDSP A extracts the workflows from provenance information; – BDSP A stores the workflows with an integrated graph which organized by the usage frequency of the analysis functions and sequential relationship among them; – BDSP C requests workflow with a list of own functions and data as well as the information of OS, H/W description, locale settings; – BDSP A retrieves workflow based on the information came from BDSP C;

Table I.8 – Use case –Extracting analysis workflow from the accumulated provenance information

	<ul style="list-style-type: none"> – BDSP A sends the list of candidate workflows to BDSP C; – BDSP C selects the workflow that satisfies the analysis purpose; – BDSP A reconstructs the workflow in a form that it can run on the BDSP C and send it to BDSP C; – BDSP C uses the workflow.
Roles/sub-roles	– BDSP
Figure (optional)	<p style="text-align: right; font-size: small;">Y:3602(18)_T1.8</p>
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirement	<ul style="list-style-type: none"> – Provenance model description (see clause 8.1) – Extracting workflow (see clause 8.2) – Retrieving workflow(see clause 8.2) – Storing workflow (see clause 8.2) – Equivalent function for process steps (see clause 8.2) – Adaptability of computational environment (see clause 8.2)

Bibliography

- [b-ITU-T X.1255] Recommendation ITU-T X.1255 (2013), *Framework for discovery of identity management information*.
- [b-ITU-T Y.2201] Recommendation ITU-T Y.2201 (2009), *Requirements and capabilities for ITU-T NGN*.
- [b-ITU-T Y.2701] Recommendation ITU-T Y.2701 (2007), *Security requirements for NGN release 1*.
- [b-ITU-T Y.4100] Recommendation ITU-T Y.4100/Y.2066 (2014), *Common requirements of the Internet of things*.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems