ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T Y.3607 (01/2023)

SERIES Y: Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities

Big Data

Big data – Functional architecture for data provenance

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
Computing power networks	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.3599
BIG DATA	**Y.3600–Y.3799**
QUANTUM KEY DISTRIBUTION NETWORKS	Y.3800–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.3607

Big data – Functional architecture for data provenance

Summary

Recommendation ITU-T Y.3607 describes a functional architecture for big data provenance (BDP). To provide the functional architecture for big data provenance, the big data provenance functions are defined based on the functional requirements and logical components identified in Recommendation ITU-T Y.3602. This Recommendation also provides the relationship between the functional architecture of big data provenance and the big data reference architecture in Recommendation ITU-T Y.3605.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.3607	2023-01-13	13	11.1002/1000/15242

Keywords

Big data, data provenance, functional architecture.

* 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/software copyrights, 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 appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

© ITU 2023

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

Recommendation ITU-T Y.3607

Big data – Functional architecture for data provenance

1 Scope

This Recommendation provides a functional architecture for big data provenance. It specifies the following:

– Functions for supporting big data provenance;

– Functional architecture of big data provenance;

– Reference points among functions for big data provenance.

In addition, this Recommendation provides a relationship between the functional architecture of big data provenance (BDP) and the big data reference architecture of [ITU-T Y.3605] in appendices.

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.3602] Recommendation ITU-T Y.3602 (2022), *Big data – Functional requirements for data provenance.*

[ITU-T Y.3605] Recommendation ITU-T Y.3605 (2020), *Big data – Reference architecture*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 big data [b-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 big data provenance [ITU-T Y.3602]: 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).

3.1.3 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.

3.1.4 reference architecture [b-ISO/IEC 26550]: Core architecture that captures the high-level design of a software and systems product line including the architectural structure and texture

(e.g., common rules and constraints) that constrains all member products within a software and systems product line.

NOTE – Application architectures of the member products included in the product line reuses (possibly with modifications) the common parts and binds variable parts of the domain architecture. Application architectures of the member products may (but do not need to) provide variability.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

BD	Big Data
BDAS-FS	Big Data Analytics Support Functions
BDP	Big Data Provenance
BDSC-FS	Big Data System Connection Functions
BDSI-F	Big Data System Interface Function
BDSP	Big Data Service Provider
DB	Data Broker
DP	Data Provider
H/W	Hardware
JSON	JavaScript Object Notation
OS	Operating System
PIC-F	Provenance Information Composition Function
PII	Personally Identifiable Information
PIM-F	Provenance Information Monitor Function
PIP-FS	Provenance Information Processing Functions
PIPM-F	Provenance Information Policy Management Function
PIT-F	Provenance Information Transformation Function
PUE-F	Provenance Unit Extraction Function
PUM-F	Provenance Unit Management Function
PUP-FS	Provenance Unit Processing Function
PUV-F	Provenance Unit Validator Function
UIAS-F	User Interface for Analytic Support Function
WE-F	Workflow Explorer Function
WMB-F	Workflow Map Building Function
WM-F	Workflow Management Function
XML	extensible Markup Language

5 Conventions

Throughout this Recommendation, the term "big data system" is to be understood as a big data system that includes capabilities of big data provenance functions.

6 Relationship between BDP logical components and BDP functions

There are six big data provenance (BDP) logical components in [ITU-T Y.3602] including:

1) provenance model management;
2) provenance lifecycle management;
3) provenance sharing policy management;
4) personally identifiable information (PII) management;
5) analysis support; and
6) monitoring.

These BDP logical components are realized by the BDP functions. Table 6-1 shows the relationship between BDP logical components in [ITU-T Y.3602] and BDP functions in this Recommendation.

Table 6-1 – Relationship between BDP logical components and BDP functions

BDP logical components in [ITU-T Y.3602]	BDP functions in this Recommendation
Provenance model management	Provenance unit processing (see clause 7.1)
Provenance lifecycle management	Provenance unit processing (see clause 7.1)
Provenance sharing policy management	Provenance information processing (see clause 7.2)
Personally identifiable information (PII) management	Provenance information processing (see clause 7.2)
Analysis support	Big data analytics support (see clause 7.3), and big data system connection (see clause 7.4)
Monitoring	Provenance information processing (see clause 7.2), and big data system connection (see clause 7.4)

7 BDP functions

This clause provides the BDP functions which are derived from the BDP functional requirements in [ITU-T Y.3602]. The relationships between BDP functional requirements and BDP functions are described in Appendix I.

7.1 Provenance unit processing

The provenance unit processing functions (PUP-FS) include:

– provenance unit extraction (see clause 7.1.1);
– provenance unit validator (see clause 7.1.2);
– provenance unit management (see clause 7.1.3).

7.1.1 Provenance unit extraction

The provenance unit extraction function (PUE-F) extracts the provenance units from the big data (BD) system. This function:

– performs collecting the required metadata for the provenance unit by query message;

NOTE 1 – Collecting the required metadata for the provenance unit has different approaches according to the data types of the provenance unit.

– records the metadata of the original data instance in the data catalogue;

NOTE 2 – The metadata of the data instance includes responsible party information, time, size, and the locations of the original data instance.

– captures the processes of data, and input / output data in the temporal storage simultaneously when the data-processing occurs;

– monitors the computational environment for recording up-to-date computational environments in the provenance unit;

– monitors the responsible party for recording up-to-date responsible party in the provenance unit.

7.1.2 Provenance unit validator

The provenance unit validator function (PUV-F) validates the extracted provenance units with the policy and checks the provenance units for finding errors in the format or invalid information. This function:

– informs the invalid information in the provenance units to the big data system user;

NOTE 1 – The invalid information includes the incomplete data, vacant data, and un-allowed data.

NOTE 2 – The provenance unit validator can request to correct the current invalid information in the provenance unit to the responsible party of the data.

– transmits the dataset information to the big data analytics support functions;

– checks the existence of PII in the metadata of the dataset;

– removes or abstracts the PII if the PII permission is rejected by the policy.

NOTE 3 – The provenance unit validator abstracts the PII for preventing the exposure of personal information.

7.1.3 Provenance unit management

The provenance unit management function (PUM-F) supports the handling of the provenance units. This function:

– requests the continuous recording of provenance unit;

– manages the extraction algorithm which performs extracting the metadata from the dataset;

– performs the provenance data compressions without loss of information;

– provisions resources of temporal storage for the provenance unit data before completion of data-processing or workflow;

– manages storage for provenance units for backward tracing of data history.

7.2 Provenance information processing

The provenance information processing functions (PIP-FS) include:

– provenance information composition (see clause 7.2.1);

– provenance information transformation (see clause 7.2.2);

– provenance information monitor (see clause 7.2.3);

– provenance information policy management (see clause 7.2.4).

7.2.1 Provenance information composition

The provenance information composition function (PIC-F) generates and stores the provenance information with multiple provenance units. This function:

– composes the provenance units into a provenance information;

– registers the provenance information to the data storage when the user's request occurs;

NOTE – The registry of the provenance information can be used to effectively retrieve the frequently used provenance information. The retrieval algorithms are the implementation issues that are not covered in this Recommendation.

– transmits the required resources for storing provenance information to the big data system connection functions.

7.2.2 Provenance information transformation

The provenance information transformation function (PIT-F) transforms the provenance information into pre-defined formats. This function:

– encodes the provenance model into the standardized format;

NOTE 1 – The standardized format has a specified provenance model for the data provenance. For the encoding provenance unit, the metadata converts into the standardized format algorithmically and the decoding is the opposite process.

– decodes the provenance model format from an external big data system;

– transforms the provenance information into a common format.

NOTE 2 – The common format includes the programming languages such as the eXtensible markup language (XML), JavaScript object notation (JSON), etc.

7.2.3 Provenance information monitor

The provenance information monitor function (PIM-F) monitors the provenance information and provenance operation for maintaining the stability of the provenance information. This function:

– supports backward tracing for provenance information history;

– executes the provenance operation when the dataset deletion occurs.

NOTE – The provenance operation includes the 'keep', 'delete', and 'combine' operations for the provenance unit. The detailed operations are described in clause 7.3 of [ITU-T Y.3602].

7.2.4 Provenance information policy management

The provenance information policy management function (PIPM-F) manages the policies for sharing and monitoring provenance information. This function:

– establishes the PII policy, provenance information monitoring policy, and provenance information sharing policy;

– manages the registry to save the profiles of the provenance information policies;

– supplies the capabilities for adjusting policies for big data system users;

NOTE – Big data system administrators can design appropriate policies for each specific purpose of data analysis.

– transmits the provenance information sharing policy to the provenance information transformation function;

– transmits the PII and monitoring policies to the provenance information monitor function.

7.3 Big data analytics support

The big data analytics support functions include:

– workflow map building (see clause 7.3.1);

– workflow management (see clause 7.3.2);

– workflow explorer (see clause 7.3.3).

NOTE 1 – The requirements for workflow are described in [ITU-T Y.3602]. The workflow depicts the actual sequence of the data processes in the data.

NOTE 2 – The big data analytics support functions support building, managing, and exploring the workflow to help the big data system user.

7.3.1 Workflow map building

The workflow map building function (WMB-F) generates the workflow map. This function:

– composites the workflows into a workflow map in the form of a graph;

– combines the data processes produced from multiple different system configurations;

– lists up the provenance information entities by comparing workflows and removing duplicated common information;

– manages the workflows into an integrated graph.

7.3.2 Workflow management

The workflow management function (WM-F) supports big data systems to register workflows and handle workflow annotations. This function:

– maintains the resources for storing the workflow;

– prepares the registry for annotating at the workflow;

NOTE 1 – Big data system users can make a note or remark for the observations of workflow.

– supports the capability of writing free-text annotations and documentation;

– executes the specific workflow based on the user's request;

– offers the parameters and policies for exploring workflow maps;

– counts the usage frequency of data processes for finding frequently used data process and its usage patterns;

NOTE 2 – The usage frequency of workflow is used for recommendation of useful or best-matched workflow for data analysis.

– measures the majority of the usage of data processes;

– performs to reproduce the workflows and recommends the workflow through a recommendation algorithm designed for data analysis.

7.3.3 Workflow explorer

The workflow explorer function (WE-F) allows a big data system user to share and transform workflow. This function:

– executes the encapsulation of data processes into user-defined data processes;

– transforms the provenance workflow into an interchanging format (encoding / decoding);

NOTE 1 – Checking the syntax errors in the provenance information and workflow.

NOTE 2 – Reporting the errors to big data system users when the syntax errors are found.

– maps the equivalent data processes among different data analysis tools;

– finds the combinations of existing data processes by comparing the information of process steps;

NOTE 3 – The information on process steps includes the name of the data process, the format and structure of input and output data of this data process, the frequency of the data process, and the relationship among them.

– visualizes the workflow graph into supporting image format;

– offers retrieval parameters for searching specific workflows;

NOTE 4 – Workflow retrieval parameters could be combinations of input / output data types, or formats, data process, usage frequency, etc.

– measures the frequency of retrieval values.

NOTE 5 – The measurement of the frequency of retrieval values can be used for a quick response when the retrieval is requested.

7.4 Big data system connection

The big data system connection support functions include:

– Big data system interface (see clause 7.4.1);

– User interface for analytic support (see clause 7.4.2).

NOTE – The big data system connection functions utilize the functional component of access control in [ITU-T Y.3605].

7.4.1 Big data system interface

The big data system interface function (BDSI-F) provides the capabilities to access the big data system and connect with external big data systems. This function:

– configures the provenance lifecycle with user's preferences;

– monitors the data in the big data system for data provenance;

– maintains the connection with the external big data systems.

NOTE – The connection supports the exchange of workflow data and the provenance information among the big data systems.

7.4.2 User interface for analytic support

The user interface for the analytic support function provides the user interface for utilizing the workflow and the provenance information. This function:

– provides the user interface for analysing the workflow with the provenance information;

– supports the capability to make the annotations for the workflow;

NOTE – The annotations for workflow are utilized for the purpose of future analysis by providing information such as the sharing objectives of the workflow.

– collects the provenance information when the user request occurs.

8 BDP functional architecture

Figure 8-1 shows the BDP functional architecture.

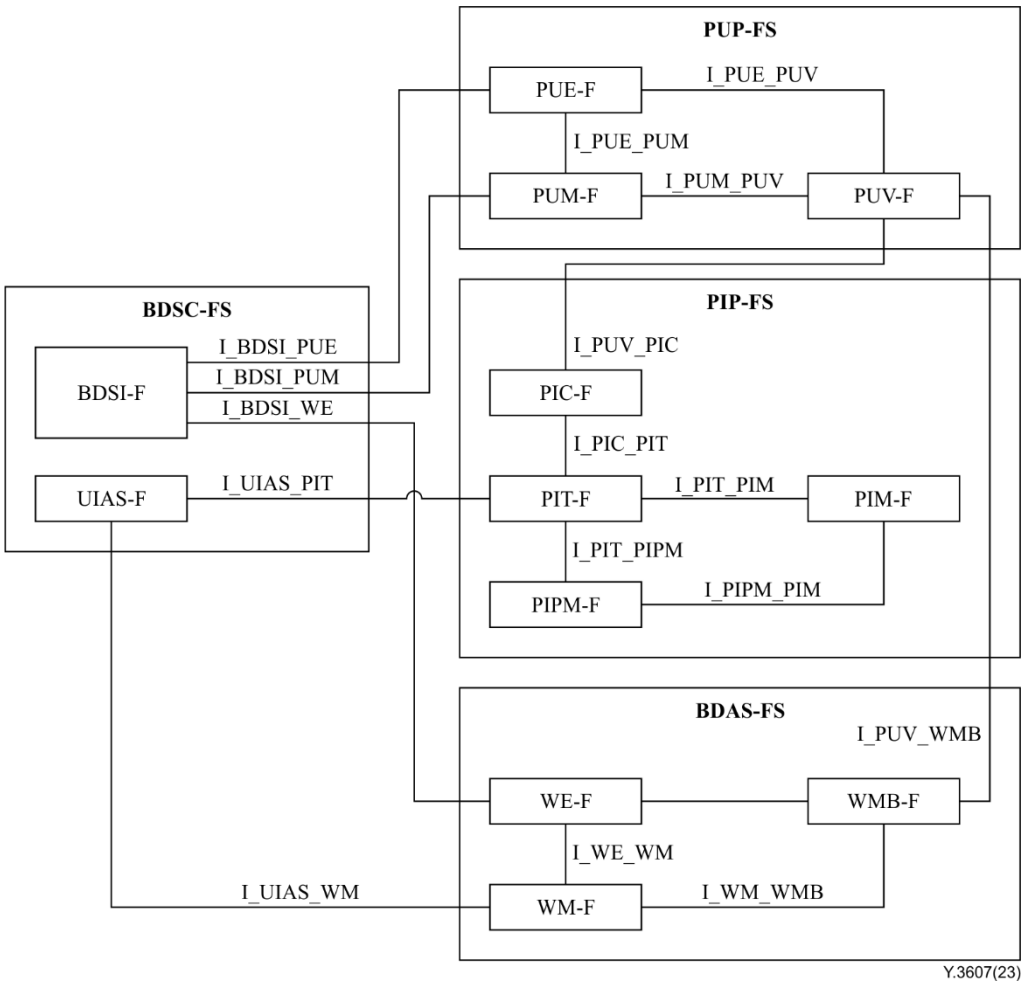

Figure 8-1 – BDP functional architecture

8.1 Provenance unit processing functions

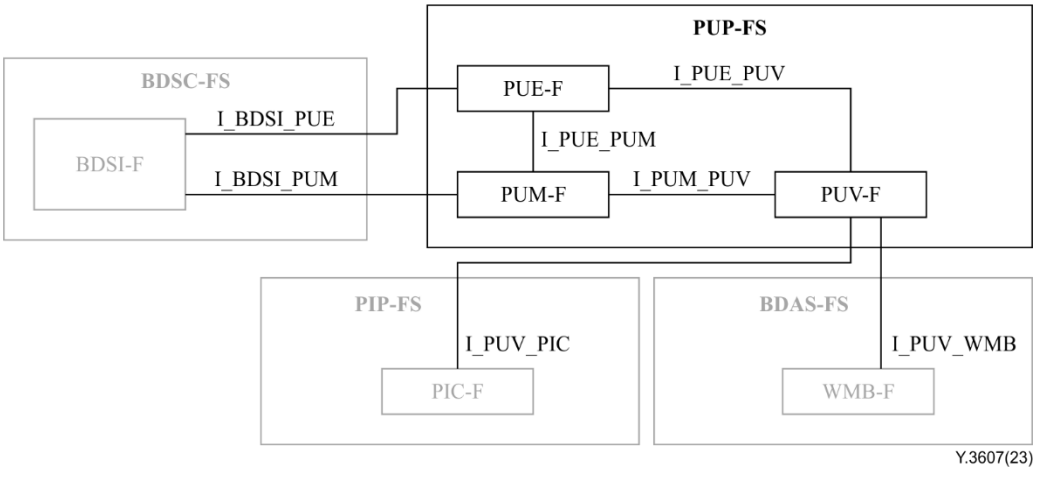

Figure 8-2 – PUP-FS and related reference points

The provenance unit processing functions (PUP-FS) have three functions including the provenance unit extraction function (PUE-F), provenance unit validator function (PUV-F), and provenance unit management function (PUM-F) as shown in Figure 8-2. The PUP-FS interfaces with the big data system connection functions (BDSC-FS), provenance information processing functions (PIP-FS), and big data analytics support functions (BDAS-FS).

8.1.1 Provenance unit extraction function

The PUE-F monitors the big data system and gathers the information for the provenance unit through the I_BDSI_PUE. The PUE-F connects the PUV-F for sending the extracted provenance units through the I_PUE_PUV. The configurations for provenance units are set up from PUM-F through the I_PUE_PUM.

8.1.2 Provenance unit validator function

The PUV-F manages the PII by following the policy configured by the PUM-F through the I_PUM_PUV. The valid provenance units are transferred to the provenance information composition function (PIC-F) through the I_PUV_PIC. PUV-F sends the dataset information in provenance units to the WMB-F through the I_PUV_WMB.

8.1.3 Provenance unit management function

The PUM-F configures the policy for PII and the provenance unit lifecycle. These policies are applied to the PUE-F and PUV-F through the I_PUE_PUM and I_PUM_PUV separately.

8.2 Provenance information processing functions

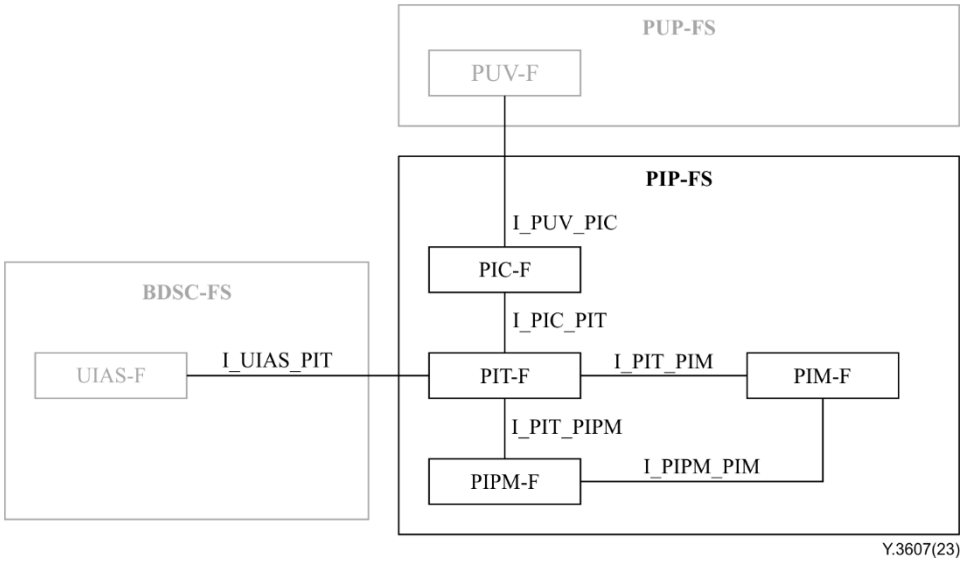

Figure 8-3 – PIP-FS and related reference points

The provenance information processing functions (PIP-FS) have four functions including the provenance information composition function (PIC-F), provenance information transformation function (PIT-F), provenance information monitor function (PIM-F), and provenance information policy management function (PIPM-F) as shown in Figure 8-3. The PIP-FS interfaces with the provenance unit processing functions (PUP-FS) and the big data system connection functions (BDSC-FS).

8.2.1 Provenance information composition function

The PIC-F collects the provenance units from the PUV-F through the I_PUV_PIC to combine the provenance information. The PIC-F stores the combined provenance information in the registry for provenance information and sends it to the PIT-F through the I_PIC_PIT.

8.2.2 Provenance information transformation function

The PIT-F encodes and decodes the provenance information by following the provenance policy of the PIPM-F. The provenance policy is configured through the I_PIT_PIPM. The PIT-F corrects the error inside the provenance information with the monitoring information from the PIM-F through I_PIT_PIM. The PIT-F exports the provenance information when the user request occurs from the user interface for analytic support function (UIAS-F) through the I_UIAS_PIT.

8.2.3 Provenance information monitor function

The PIM-F monitors the provenance information through the I_PIT_PIM. The PIM-F sends the invalid information to the PIT-F. The monitoring policy is configured by PIPM-F through the I_PIPM_PIM.

8.2.4 Provenance information policy management function

The PIPM-F applies the provenance sharing and monitoring policy to the PIT-F and PIM-F. The reference points of I_PIT_PIPM and the I_PIPM_PIM are used to configure these policies.

8.3 Big data analytics support functions

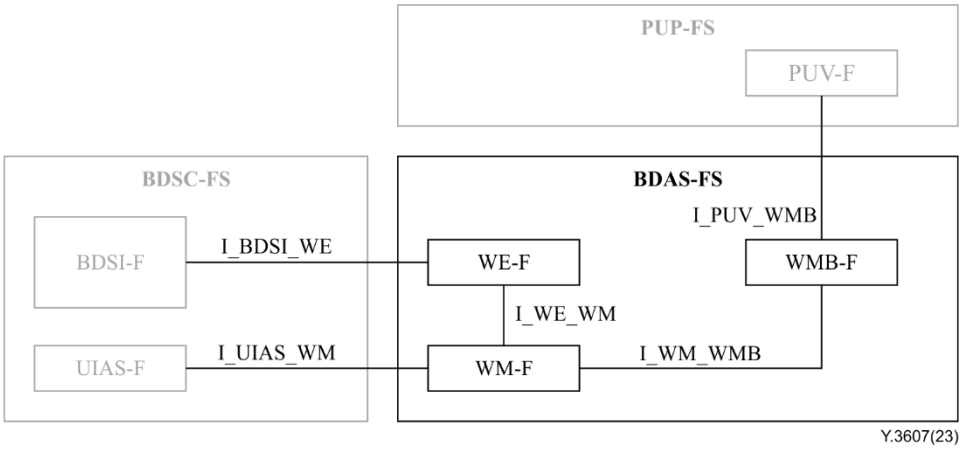

Figure 8-4 – BDAS-FS and related reference points

The big data analytics support functions (BDAS-FS) have three functions including workflow explorer function (WE-F), workflow management function (WM-F), and workflow map building function (WMB-F) as shown in Figure 8-4. The BDAS-FS interfaces with the provenance unit management functions (PUP-FS) and big data system connection functions (BDSC-FS).

8.3.1 Workflow map building function

The WMB-F builds the workflow map by collecting the dataset information from the PUV-F through the I_PUV_WMB. The WMB-F sends the built workflow mapping table to the WM-F through the I_WM_WMB.

8.3.2 Workflow management function

The WM-F sets the registry to store the workflow from the WMB-F and WE-F. The WM-F supports big data system users to analyse information from the workflow map. The WM-F sends the information to the UIAS-F for supporting big data analysis. The WM-F provides the functionality to annotate the analysis information for the workflow through the I_UIAS_WM.

8.3.3 Workflow explorer function

The WE-F provides the parameter for retrieving provenance information to the big data system interface function (BDSI-F) through the I_BDSI_WE. The WE-F exchanges the workflow map with the WM-F through the I_WE_WM.

8.4 Big data system connection functions

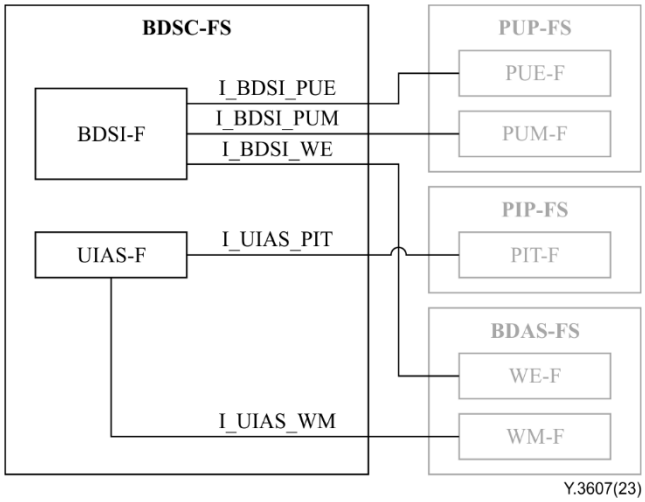

Figure 8-5 – BDSC-FS and related reference points

The big data system connection functions (BDSC-FS) have two functions including the big data system interface function (BDSI-F) and the user interface for analytic support function (UIAS-F) as shown in Figure 8-5. The BDSC-FS interfaces with the provenance unit processing functions (PUP-FS), provenance information processing functions (PIP-FS), and big data analytics support functions (BDAS-FS).

8.4.1 Big data system interface function

The BDSI-F supports the interface for executing the provenance functions to the PUP-FS, the PIP-FS, and the BDAS-FS. The BDSI-F has three reference points the I_BDSI_PUE, the I_BDSI_PUM, and the I_BDSI_WE to provide the big data system connection.

8.4.2 User interface for analytic support function

The UIAS-F provides the user interface to utilize the analysis support. For this function, the two reference points which are the I_UIAS_PIT and I_UIAS_WM are supported by the PIT-F and WM-F.

9 Reference points

This clause describes the reference points among BDP functions.

9.1 Reference points between PUP-FS and BDSC-FS

The reference points between PUP-FS and BDSC-FS are as follows:

I_BDSI_PUE reference point between BDSI-F and PUE-F. The BDSI-F sends the metadata for provenance information to the PUE-F through this reference point.

I_BDSI_PUM reference point between BDSI-F and PUM-F. The BDSI-F sends the provenance unit policy to the PUM-F through this reference point.

9.2 Reference point between PUP-FS and PIP-FS

The reference point between PUP-FS and PIP-FS is as follows:

I_PUV_PIC reference point between PUV-F and PIC-F. The PUV-F sends the valid provenance unit by PUV-F to PIC-F through this reference point.

9.3 Reference point between PUP-FS and BDAS-FS

The reference point between PUP-FS and BDAS-FS is as follows:

I_PUV_WMB reference point between PUV-F and WMB-F. The PUV-F sends the dataset information inside the provenance unit to the WMB-F through this reference point.

9.4 Reference point between PIP-FS and BDSC-FS

The reference point between PIP-FS and BDSC-FS is as follows:

I_UIAS_PIT reference point between UIAS-F and PIT-F. The PIT-F sends the exported and encoded provenance information to the UIAS-F through this reference point.

9.5 Reference points between BDAS-FS and BDSC-FS

The reference points between BDAS-FS and BDSC-FS are as follows:

I_BDSI_WE reference point between BDSI-F and WE-F. The WE-F sends the workflow data to the BDSI-F through this reference point.

I_UIAS_WM reference point between UIAS-F and WM-F. The WM-F sends the requested workflow to the UIAS-F and receives the user's annotation gathered by the UIAS-F through this reference point.

9.6 Reference points within PUP-FS

The reference points within PUP-FS are as follows:

I_PUE_PUV reference point between PUE-F and PUV-F. The PUE-F sends the extracted provenance unit to the PUV-F through this reference point.

I_PUE_PUM reference point between PUE-F and PUM-F. The PUM-F sends the configuration of the provenance unit to the PUE-F through this reference point.

I_PUM_PUV reference point between PUM-F and PUV-F. The PUM-F sends the PII policy to the PUV-F through this reference point.

9.7 Reference points within PIP-FS

The reference points within PIP-FS are as follows:

I_PIC_PIT reference point between PIC-F and PIT-F. The PIC-F sends the composed provenance information to the PIT-F through this reference point.

I_PIT_PIM reference point between PIT-F and PIM-F. The PIM-F monitors the transformed provenance information in the PIT-F through this reference point.

I_PIT_PIPM reference point between PIT-F and PIPM-F. The provenance sharing policy is configured by PIPM-F through this reference point.

I_PIPM_PIM reference point between PIPM-F and PIM-F. The provenance monitoring policy is configured by PIPM-F through this reference point.

9.8 Reference points within BDAS-FS

The reference points within BDAS-FS are as follows:

I_WE_WM reference point between WE-F and WM-F. The workflow map is exchanged through this reference point.

I_WM_WMB reference point between WM-F and WMB-F. The data for the workflow mapping table is registered through this reference point.

10 Security considerations

Security considerations within the big data system are addressed in [b-ITU-T X.1750], [b-ITU-T X.1751] and [b-ITU-T X.1752]. [b-ITU-T X.1750] and [b-ITU-T X.1751] describe the threats and challenges for security in the big data system. [b-ITU-T X.1751] analyses security vulnerabilities and guidelines of big data lifecycle management and includes the security aspects in data collection, data transmission, data storage, data usage, data sharing and data destruction.

Security requirements for personally identifiable information (PII) management are addressed in [b-ITU-T X.1058]. [b-ITU-T X.1058] provides the implementation guidance for the protection of PII and other information for the protection of PII such as selecting PII controls, developing organization specific guidelines, and the lifecycle of PII.

Appendix I

Relationships among BDP functional requirements, BDP logical components, and BDP functions

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

Table I.1 – Relationships among big data provenance (BDP) functional requirements, BDP logical components, and BDP functions in this Recommendation

BDP functional requirements in [ITU-T Y.3602]		Related BDP logical components in [ITU-T Y.3602]	BDP functions in this Recommendation
Provenance lifecycle requirements (Clause 8.1)	**(Provenance model description)** It is required that big data service provider (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 – Examples of computational environment information are operating system (OS), hardware (H/W) description, locale settings, time zone, etc.	Provenance model management	Provenance unit extraction (clause 7.1.1), Provenance unit validator (clause 7.1.2), Provenance information policy management (clause 7.2.4)
	(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 specialised format. Decoding is the opposite process.	Provenance model management	Provenance information composition (clause 7.2.1), Provenance information transformation (clause 7.2.2), Provenance information monitor (clause 7.2.3)
	(Provenance recording initiation) It is required that BDSP records the provenance unit when data is stored; NOTE 4 – The information contained in the metadata (from data provider (DP):data broker (DB) or generated by BDSP) can be used for recording the provenance unit.	Provenance sharing policy management	Provenance unit extraction (clause 7.1.1), Provenance unit management (clause 7.1.3)

Table I.1 – Relationships among big data provenance (BDP) functional requirements, BDP logical components, and BDP functions in this Recommendation

BDP functional requirements in [ITU-T Y.3602]		Related BDP logical components in [ITU-T Y.3602]	BDP functions in this Recommendation
	(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 efficient storage usage, it is needed to designate a predetermined period of time to record provenance unit, rather than recording it every time data is stored. Data compression techniques can also be considered.	Provenance life-cycle management	Provenance unit extraction (clause 7.1.1), Provenance unit validator (clause 7.1.2), Provenance unit management (clause 7.1.3)
	(Storing provenance information) BDSP can optionally support pre-storing provenance information prior to the request time to reduce retrieval time.	Provenance lifecycle management	Provenance information composition (clause 7.2.1), Provenance information transformation (clause 7.2.2), Provenance information policy management (clause 7.2.4)
	(Searching provenance unit) It is required that BDSP supports searching a provenance unit.	Provenance lifecycle management	Provenance unit management (clause 7.1.3)
	(Combining provenance units) It is required that BDSP supports the combining of provenance units; NOTE 6 – In case of deleting data, a provenance unit is needed to combine (see clause 7.3.2).	Provenance lifecycle management	Provenance unit validator (clause 7.1.2), Provenance unit management (clause 7.1.3)
	(Retrieving provenance information) It is required that BDSP supports the provenance unit aggregation to retrieve a provenance information.	Provenance life-cycle management	Provenance information monitor (clause 7.2.3), Provenance information policy management (clause 7.2.4)

Table I.1 – Relationships among big data provenance (BDP) functional requirements, BDP logical components, and BDP functions in this Recommendation

BDP functional requirements in [ITU-T Y.3602]		Related BDP logical components in [ITU-T Y.3602]	BDP functions in this Recommendation
	(Deleting provenance unit) It is required that BDSP provides a provenance unit deletion mechanism; NOTE 7 – In case of deleting data, BDSP acts with three mechanisms on the provenance unit (keep, combine and delete) based on the context (see clause 7.3.2). NOTE 8 – The BDSP can maintain the associated provenance unit even if the data are deleted, which is subject to management policy.	Provenance lifecycle management	Provenance unit validator (clause 7.1.2), Provenance unit management (clause 7.1.3)
Analysis support requirements (Clause 8.2)	**(Extracting workflow)** It is required that BDSP provides the extraction of workflow information from a provenance information.	Analysis support	Workflow map building (clause 7.3.1), Workflow explorer (clause 7.3.3)
	(Storing workflow) It is recommended that BDSP supports storing workflow; NOTE 1 – The workflow is stored in forms of a graph, which is organized with the usage frequency of the analysis functions and the sequential relationship among them.	Analysis support	Workflow map building (clause 7.3.1), Workflow management (clause 7.3.2)
	(Retrieving workflow) It is recommended that BDSP supports workflow retrieval.	Analysis support	Workflow explorer (clause 7.3.3)
	(Providing data list on function) It is recommended that BDSP provides a list of data related to a given function recorded in a given workflow.	Analysis support	Workflow explorer (clause 7.3.3)
	(Providing function list on data) It is recommended that BDSP provides a list of functions related to a given data recorded in a given workflow.	Analysis support	Workflow map building (clause 7.3.1), Workflow explorer (clause 7.3.3)

Table I.1 – Relationships among big data provenance (BDP) functional requirements, BDP logical components, and BDP functions in this Recommendation

BDP functional requirements in [ITU-T Y.3602]		Related BDP logical components in [ITU-T Y.3602]	BDP functions in this Recommendation
	(Data analysis automation) It is recommended that BDSP supports analysis automation based on workflow.	Analysis support, monitoring	Workflow map building (clause 7.3.1), Workflow management (clause 7.3.2), Big data system interface (clause 7.4.1)
	(User annotation) BDSP can optionally support annotation on provenance information.	Analysis support	Workflow management (clause 7.3.2), User interface for analytic support (clause 7.4.2)
	(Equivalent function for process steps) It is recommended that BDSP 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 the 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.	Analysis support	Workflow map building (clause 7.3.1), Workflow management (clause 7.3.2)
	(Adaptability of computational environment) It is recommended that BDSP provides diagnose computational environment to reuse the provenance information which came from a different system.	Analysis support	Provenance unit extraction (clause 7.1.1), Provenance information composition (clause 7.2.1), Big data system interface (clause 7.4.1)

Table I.1 – Relationships among big data provenance (BDP) functional requirements, BDP logical components, and BDP functions in this Recommendation

BDP functional requirements in [ITU-T Y.3602]		Related BDP logical components in [ITU-T Y.3602]	BDP functions in this Recommendation
Monitoring requirements (Clause 8.3)	**(Monitoring computational environment)** It is required that BDSP monitors the change in the computational environment.	Monitoring	Provenance unit extraction (clause 7.1.1)
	(Monitoring responsible party) It is required that BDSP monitors the change of responsible party.	Monitoring	Provenance unit extraction (clause 7.1.1)
	(Applying the monitoring result) It is required that BDSP reflects the monitoring results to the recorded provenance unit; NOTE – The monitoring results include the change of the computational environment and the responsible party.	Monitoring	Provenance unit extraction (clause 7.1.1), Big data system interface (clause 7.4.1)
Policy management requirements (Clause 8.4)	**(Verifying PII)** It is required that BDSP 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).	Personally identifiable information (PII) management	Provenance unit validator (clause 7.1.2), Provenance information policy management (clause 7.2.4)
	(Protecting PII) It is required that BDSP provides a protection mechanism for a PII in data; NOTE 3 – When a PII is included in the data sources, BDSP decides to omit it or not based on the user's access authority.	Personally identifiable information (PII) management	Provenance unit validator (clause 7.1.2), Provenance information policy management (clause 7.2.4)
	(Simplifying provenance information) It is recommended that BDSP supports simplifying the provenance information based on a sharing policy; NOTE 4 – Methods of provenance information simplification include multiple levels of detail and encoding formats, etc.	Provenance sharing policy management	Provenance information transformation (clause 7.2.2), Provenance information policy management (clause 7.2.4)

Table I.1 – Relationships among big data provenance (BDP) functional requirements, BDP logical components, and BDP functions in this Recommendation

BDP functional requirements in [ITU-T Y.3602]		Related BDP logical components in [ITU-T Y.3602]	BDP functions in this Recommendation
	(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 the traceability of data, and it is determined by the sharing policy. Provenance information contains process steps with the applied functions, intermediate data, and responsible party information. For the transfer of provenance information, the provenance information can be simplified according to the sharing policy.	Provenance sharing policy management	Provenance information policy management (clause 7.2.4.)

Appendix II

The overall relationships among BDP functions

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

Appendix II depicts the overall relationships among the BDP functions which are described in clauses 7.1 through 7.4.

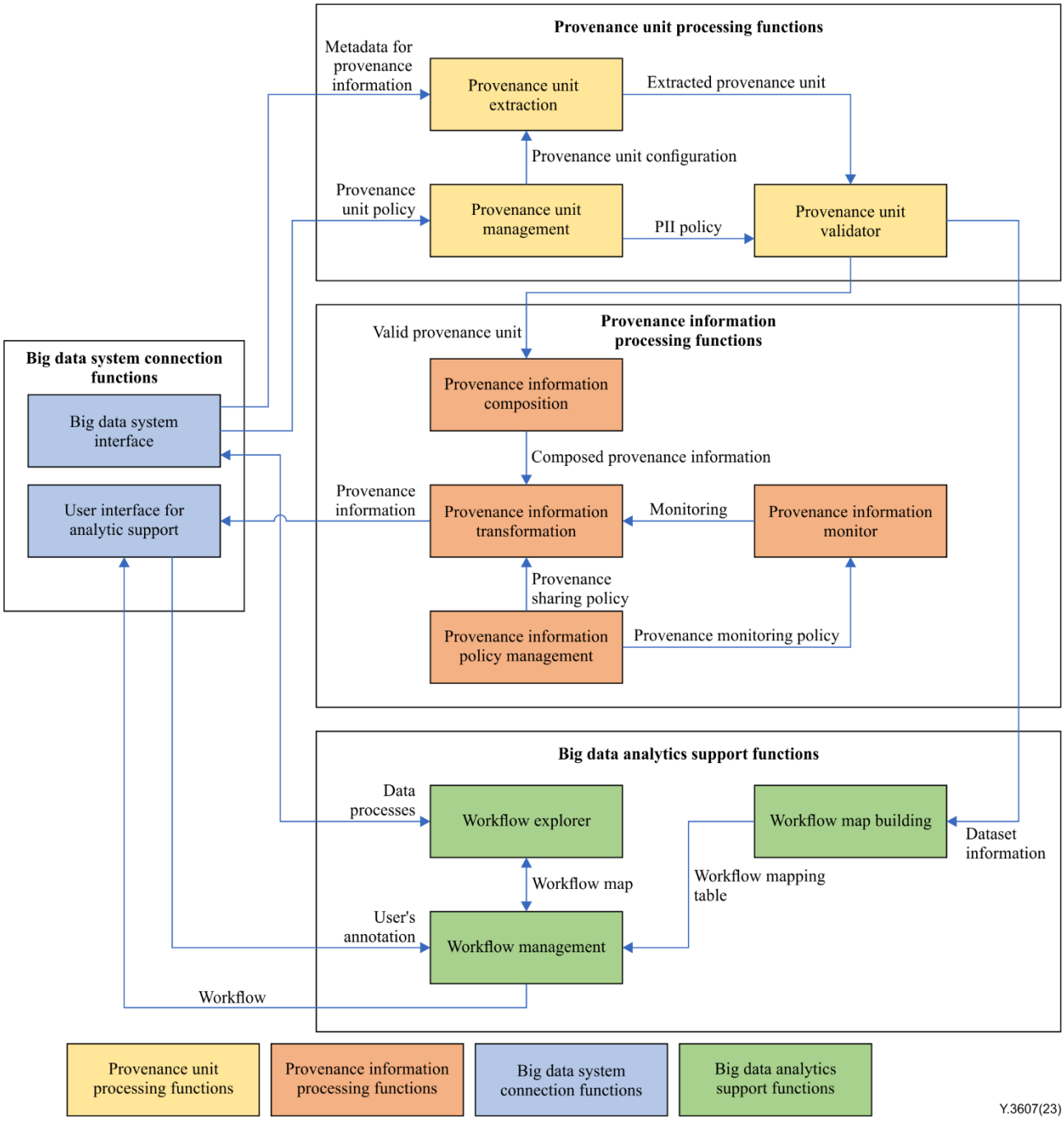

Figure II.1 – Relationships among BDP functions

Appendix III

Relationships between the BDP functions and functional components of big data architecture in [ITU-T Y.3605]

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

Appendix III shows the mapping between the BDP functions in clause 7 and the functional components in [ITU-T Y.3605].

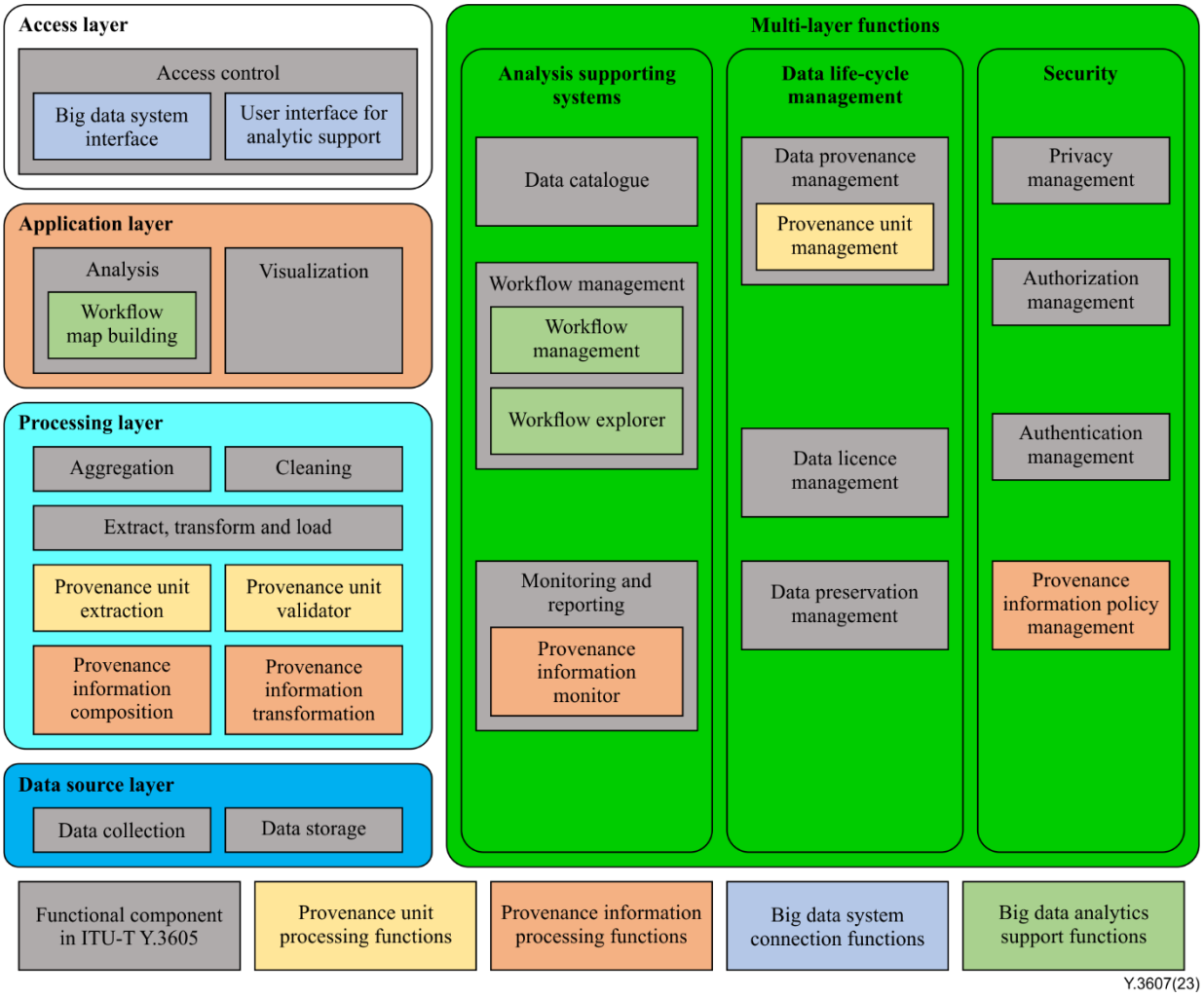

Figure III.1 – Mapping between BDP functions and functional components of big data reference architecture

Bibliography

[b-ITU-T X.1058] Recommendation ITU-T X.1058 (2017), *Information technology – Security techniques – Code of practice for personally identifiable information protection.*

[b-ITU-T X.1255] Recommendation ITU-T X.1255 (2013), *Framework for discovery of identity management information.*

[b-ITU-T X.1750] Recommendation ITU-T X.1750 (2020), *Guidelines on security of big data as a service for big data service providers.*

[b-ITU-T X.1751] Recommendation ITU-T X.1751 (2020), *Security guidelines for big data lifecycle management by telecommunication operators.*

[b-ITU-T X.1752] Recommendation ITU-T X.1752 (2022), *Security guidelines for big data infrastructure and platform.*

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

[b-ISO/IEC 26550] ISO/IEC 26550:2015, *Software and systems engineering – Reference model for product line engineering and management.*
<https://www.iso.org/standard/69529.html>

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

Published in Switzerland
Geneva, 2023