International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Y.3605

(09/2020)

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

Big Data

Big data – Reference architecture

Recommendation ITU-T Y.3605

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.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.3605

Big data – Reference architecture

Summary

Recommendation ITU-T Y.3605 defines a big data reference architecture (BDRA) that can serve as a fundamental reference point for big data standardization, and which provides an overall framework for the basic concepts and principles of big data. The Recommendation provides a description of reference architecture concepts, two distinct viewpoints including user view and functional view, and also cross cutting aspects. Furthermore, the Recommendation addresses layering framework, functional components within framework and detailed functional descriptions for big data.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.3605	2020-09-29	13	11.1002/1000/14406

Keywords

Big data, cross cutting, functional component, functional view, reference architecture, user view.

* 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/.

Table of Contents

Recommendation ITU-T Y.3605

Big data – Reference architecture

1 Scope

This Recommendation specifies the big data reference architecture (BDRA). The Recommendation provides a description of reference architecture concepts, user view, functional view and cross cutting aspects.

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.3502] Recommendation ITU-T Y.3502 (2014) | ISO/IEC 17789:2014, *Cloud computing – reference architecture*.

[ITU-T Y.3519] Recommendation ITU-T Y.3519 (2018), *Cloud computing – functional architecture of big data as a service*.

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

[ITU-T Y.3601] Recommendation ITU-T Y.3601 (2018), *Big data – framework and requirements for data exchange*.

[ITU-T Y.3603] Recommendation ITU-T Y.3603 (2019), *Big data – requirements and conceptual model of metadata for data catalogue*.

[ITU-T Y.3604] Recommendation ITU-T Y.3604 (2020), *Big data – overview and requirements for data preservation*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 activity [ITU-T Y.3502]: A specified pursuit or set of tasks.

3.1.2 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.

3.1.3 big data provenance [b-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 – The provenance information provides the details about data source, such as a 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).

3.1.4 data catalogue [ITU-T Y.3601]: A listing of all metadata which a data broker makes available.

3.1.5 functional component [ITU-T Y.3502]: A functional building block needed to engage in an activity backed by an implementation.

3.1.6 metadata [b-ITU-T H.752]: Structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment and management of the described entities.

3.1.7 party [b-ITU-T Y.3500]: Natural person or legal person, whether or not incorporated, or a group of either.

3.1.8 personally identifiable information [b-ITU-T X.1361]: Any information that a) can be used to identify the PII principal to whom such information relates, or b) is or might be directly or indirectly linked to a PII principal.

NOTE – To determine whether a PII principal is identifiable, account should be taken of all the means which can reasonably be used by the privacy stakeholder holding the data, or by any other party, to identify that natural person.

3.1.9 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.

3.1.10 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 reuse (possibly with modifications) the common parts and bind variable parts of the domain architecture. Application architectures of the member products may (but do not need to) provide variability.

3.1.11 role [ITU-T Y.3502]: A set of activities that serves a common purpose.

3.1.12 sub-role [ITU-T Y.3502]: A subset of the activities of a given role.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

BDC Big Data Service Customer

BDRA Big Data Reference Architecture

BDSP Big Data Service Provider

ETL Extract, Transform, Load

HTML Hypertext Markup Language

NoSQL Not Only SQL

PII Personally Identifiable Information

SLA Service Level Agreement

XML Extensible Markup Language

5 Conventions

The following conventions apply:

Diagrams are used throughout this Recommendation to help illustrate the big data reference architecture (BDRA). Figure 5-1 provides the conventions used regarding the content of the diagrams.

NOTE – In Figure 5-1, "Aspect" is to be understood as referring to "Cross-cutting aspect".

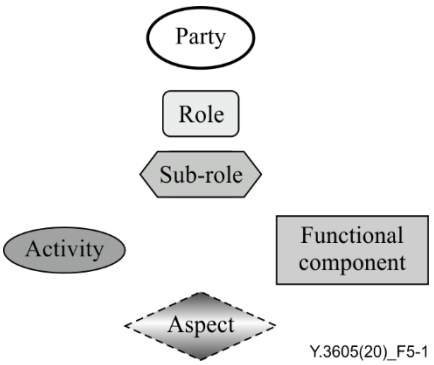

Figure 5-1 – Legend to the diagrams used throughout this Recommendation [ITU-T Y.3502]

6 Reference architecture concepts

This clause provides an overview of the architectural concepts that are used in this Recommendation.

6.1 BDRA architectural views

Big data systems can be described using a viewpoint approach.

Four distinct viewpoints are used in the BDRA (see Figure 6-1):

- user view;
- functional view;
- implementation view; and
- deployment view.

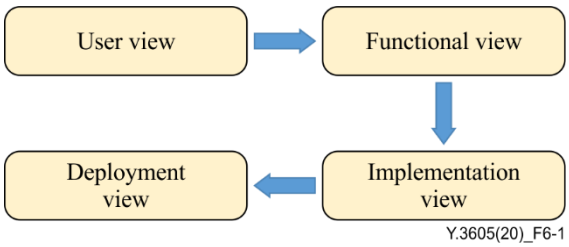

Figure 6-1 – Transformations between architectural views [ITU-T Y.3502]

Table 6-1 provides a description of each of these views.

Table 6-1 – BDRA views

BDRA view	Description of the BDRA view	Scope
User view	The ecosystem, the parties, the roles, the sub-roles and the big data activities	Within scope
Functional view	The functions necessary for the support of big data activities	Within scope
Implementation view	The functions necessary for the implementation of a big data service within service parts and/or infrastructure parts	Out of scope
Deployment view	How the functions of a big data service are technically implemented within already existing infrastructure elements or within new elements to be introduced in this infrastructure	Out of scope
NOTE – Details of the user view is defined in clause 7. Clause 8 addresses details of the functional view. The implementation and deployment views are related to technology and vendor-specific big data implementations and actual deployments, and are out of the scope of this Recommendation.		

Figure 6-2 shows the transition from the user view to the functional view.

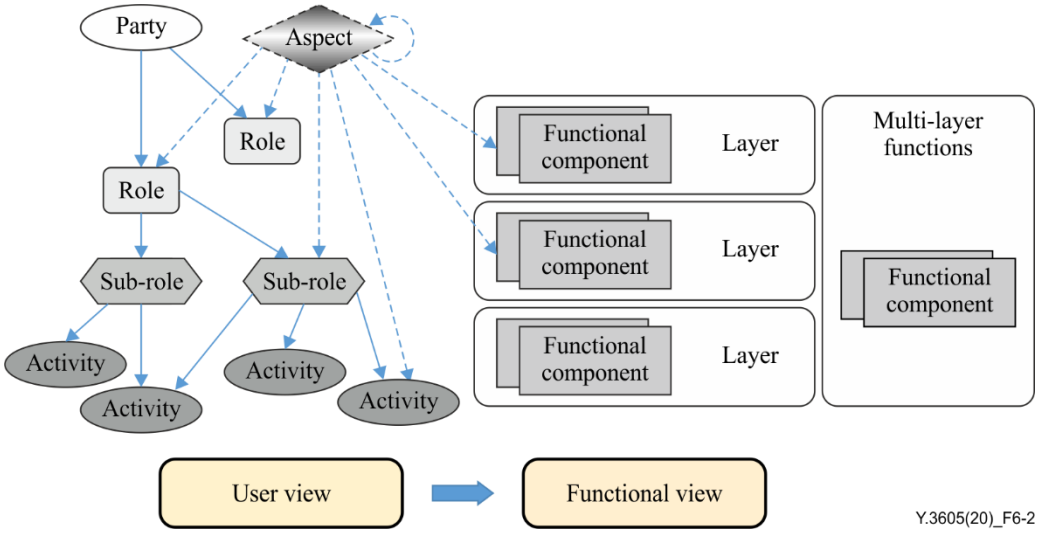

Figure 6-2 – Transition from user view to functional view [ITU-T Y.3502]

Cross-cutting aspects are behaviours or capabilities which need to be coordinated across roles and implemented consistently in a big data system. Cross-cutting aspects can be shared and can impact multiple roles, activities and functional components. Cross-cutting aspects apply to multiple individual roles or functional components. An example of a cross-cutting aspect is data quality. A description of the cross-cutting aspects is provided in clause 9.

6.2 Relationship between the user view and the functional view

Figure 6-3 illustrates how the user view provides the set of activities that are represented within the functional view (and realized using the technologies of the implementation view).

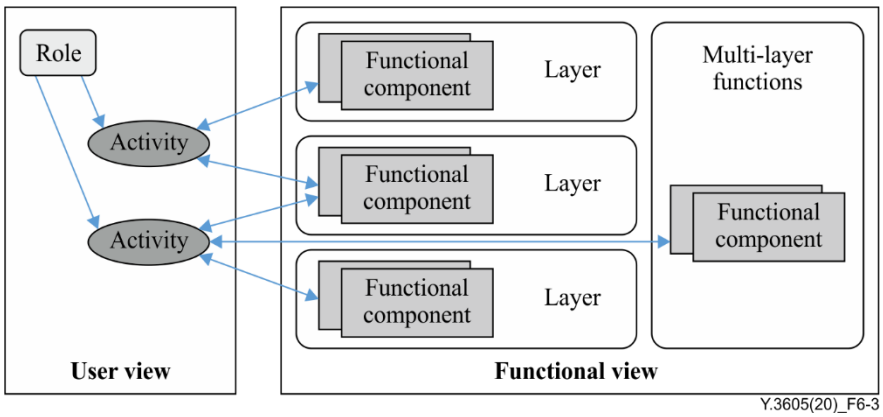

Figure 6-3 – From user view to functional view [ITU-T Y.3502]

This Recommendation specifies the roles and activities view in clause 7 and the functional view, including architectural functional components in clause 8. This Recommendation describes in this clause the logical relationships of the roles and cloud computing activities to the functional components.

Standards can be relevant to some of these relationships. Standards associated with a relationship can be used to (i) specify degrees of information flow or other types of interoperability; and/or (ii) ensure specified degrees of quality (e.g., security or service level).

Logical relationships defined in these reference architecture concepts are a significant part of specifying the BDRA and its behaviour. The relationship describes matters such as the required information flows between the functional components in the BDRA.

NOTE – Refer to more detailed common view of roles, activities and functional components in [ITU-T Y.3502]

7 User view

This clause describes an environment, called big data ecosystem, with roles and sub-roles. It also defines necessary activities for roles providing and consuming big data services as well as relationships between roles.

The big data ecosystem includes the following roles:

- data provider;
- big data service provider;
- big data service customer.

The big data ecosystem is shown in Figure 7-1.

This ecosystem is defined in [ITU-T Y.3601], which is extended from the big data ecosystem defined in [ITU-T Y.3600].

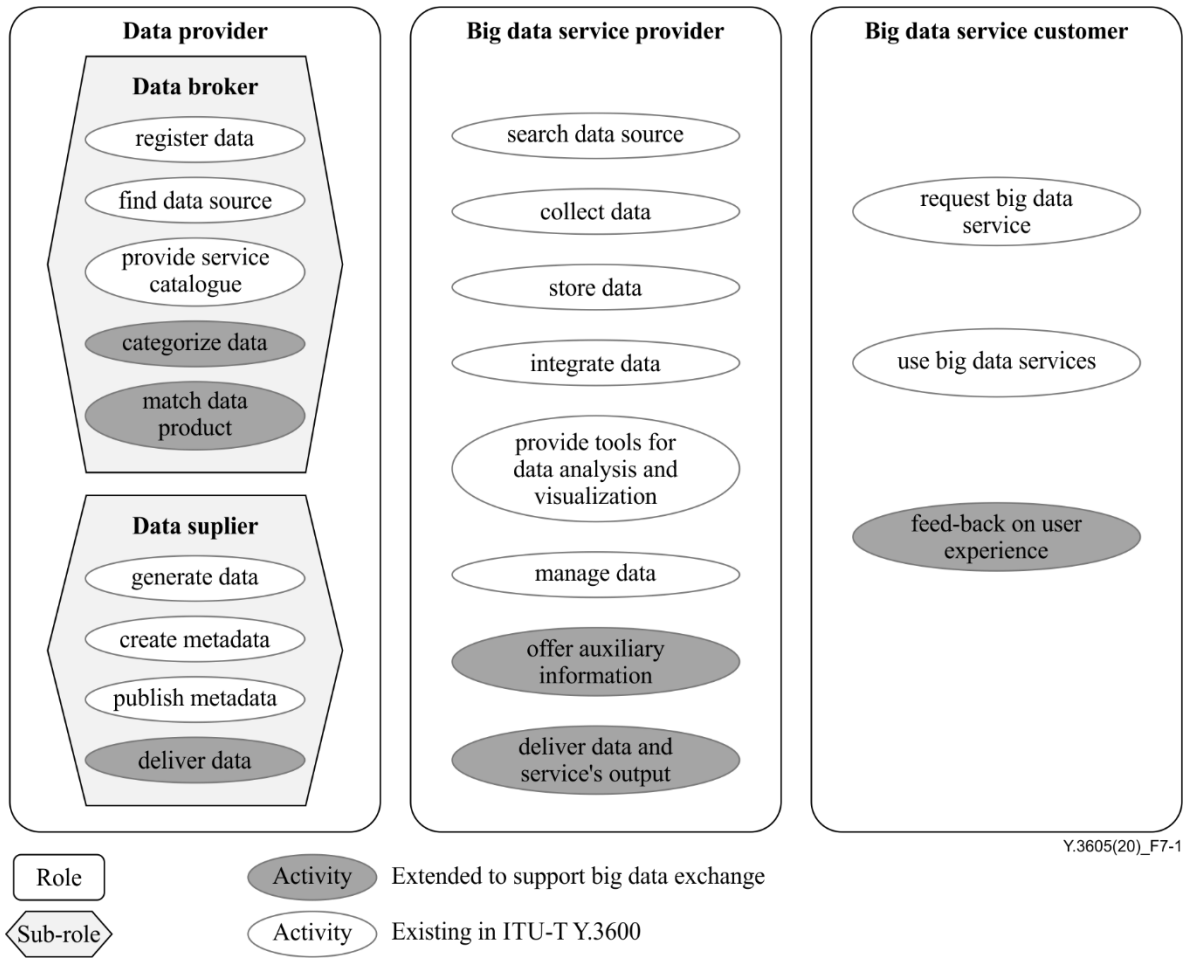

Figure 7-1 – User view of big data [ITU-T Y.3601]

Details of each sub-roles and activities shown in Figure 7-1 are provided in clause 6.2 of [ITU-T Y.3600] and clause 7.2 of [ITU-T Y.3601].

8 Functional view

8.1 Layering framework for big data

The layering framework used in the big data reference architecture has four layers, plus a set of functions which span across the layers. The four layers are:

- Access layer;
- Application layer;
- Processing layer;
- Data source layer.

The functions which span the layers are called the multi-layer functions.

The layering framework is shown in Figure 8-1.

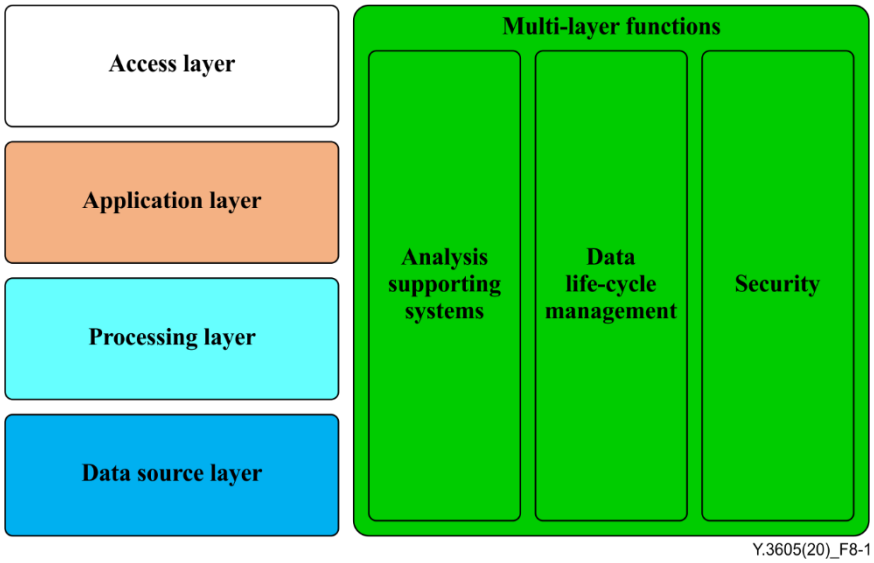

Figure 8-1 – Big data layering framework

Each of the layers in the framework is described in the following sub-clauses.

8.1.1 Access layer

The access layer provides a common interface for access to the capabilities available in the other layers of the reference architecture.

The access layer is responsible for authenticating the request through the use of user credentials and for validating the authorization of the user to use particular capabilities. The access layer is also responsible for handling encryption and checking for request integrity, where required.

The access layer passes on validated requests to the components in the other layers of the reference architecture. This layer handles big data service customer's service consumption requests to access big data services and resources provided by the big data service provider.

8.1.2 Application layer

The application layer is responsible for finding insights from big data and reporting the insights using various visualization tools. This layer contains the implementation for performing big data analysis.

8.1.3 Processing layer

The processing layer prepares data for further processing such as data analysis. This layer provides data cleaning, data aggregation, data transformation, data extraction and data loading.

8.1.4 Data source layer

The data source layer is where data are collected and stored. This layer supports registration and storage methods that are appropriate for the type of data collected. It also contains and supports a variety of storage needs for data preparation, data processing and data analysis.

8.1.5 Multi-layer functions

The multi-layer functions include a series of functional components that interact with functional components of the above four other layers to provide supporting capabilities including and not limited to:

− analysis supporting systems capabilities (data catalogue, workflow management, monitoring and reporting);

– data life-cycle management capabilities (data provenance management, data preservation management, data licence management);

– security capabilities (privacy management, authentication management, authorization management).

8.2 Functional components

This clause describes the big data reference architecture in terms of the common set of big data functional components. Figure 8-2 presents a high-level overview of the big data reference architecture functional components organized by means of the layering framework.

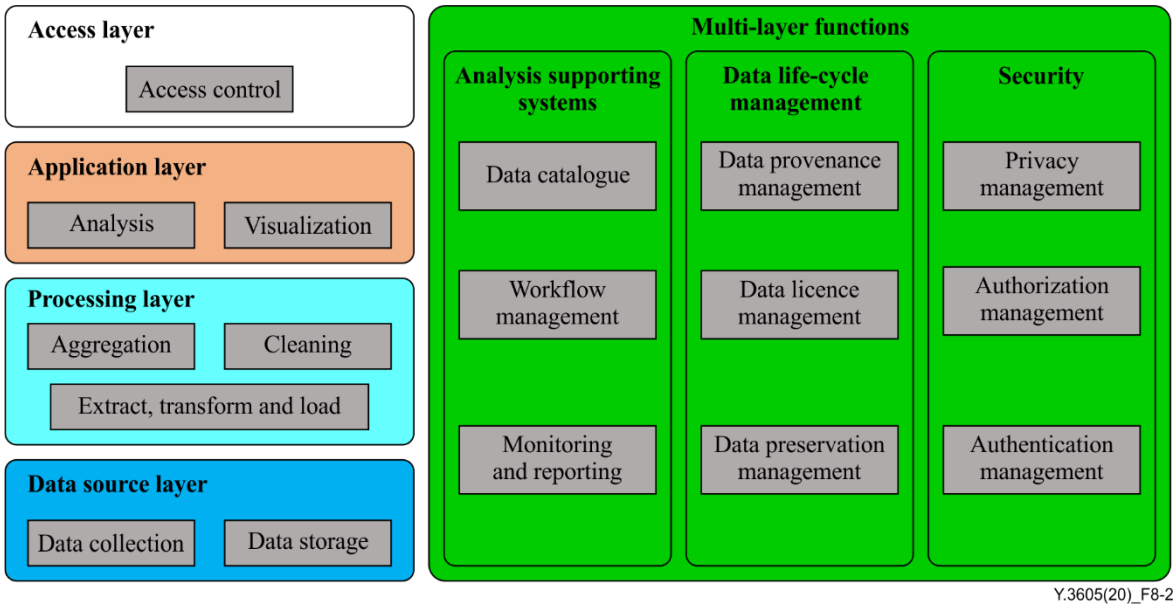

Figure 8-2 – Big data functional components

8.2.1 Access layer functional components

8.2.1.1 Access control

The access control functional component limits users to the use of particular big data services. Principally, access control involves the authentication of a user through the presentation and validation of credentials, followed by the authorization of this authenticated user to use specific big data services.

This functional component provides:

– accepting big data service customer's service consumption requests to access big data services and resources of the big data service provider;

– authenticating the request through the use of user credentials and for validating the authorization of the user to use particular capabilities;

– passing validated requests to the components in the other layers of the reference architecture.

Access control restricts users to use certain services. In principle, access control includes authenticating users through the presentation and validation of credentials and authorizing users who are authorized to use certain services.

8.2.2 Application layer functional components

8.2.2.1 Analysis

The analysis functional component is responsible for extracting useful information or valuable insights from big data. This functional component provides support for multiple data analysis methods. Besides, this functional component also supports customization of specific analysis methods.

This functional component provides:

- registration of data analysis methods which are used for data analysis. Typical data analysis methods are classification analysis, clustering analysis, association analysis, regression analysis, customized analysis;

  NOTE 1 – Classification analysis supports decision tree, support vector machine, neural networks and other algorithms, to identify to which set of categories data belongs.

  NOTE 2 – Clustering analysis supports k – means, k – centre point, overlapping clustering, fuzzy clustering, etc., to classify data into different classes or clusters according to their similarity.

  NOTE 3 – Association analysis supports some specific algorithms to find associations between stored data. Examples of association algorithms include Apriori algorithm and Frequent Pattern Growth algorithm. Apriori algorithm and Frequent Pattern Growth algorithm are two classical association analysis algorithms which can mine the associations through the frequency of data appearing together in the dataset.

  NOTE 4 – Regression analysis supports linear regression and logistic regression and other algorithms, to estimate the relationships among data.

  NOTE 5 – Customization of analysis supports the customization of detail data analysis method according to the customer's specific requirements.

- setting up procedures which enable the analysis using the registered analysis methods in the analysis function registry;

- executing analysis process according to the procedures.

  NOTE 6 – This functional component is same as the analysis functional component in [ITU-T Y.3519].

8.2.2.2 Visualization

The visualization functional component makes data more intuitive and easier to understand for users by using various data visualization tools.

This functional component provides:

- presenting data with multiple styles such as statistical graphics, forms, diagrams, charts and reports;

- reporting tools that can be configured by user.

NOTE – This functional component is same as the analysis functional component in [ITU-T Y.3519].

8.2.3 Processing layer functional components

The processing layer functional components include:

- Aggregation functional component (see clause 8.2.3.1);

- Cleaning functional component (see clause 8.2.3.2);

- Extract, transform and load functional component (see clause 8.2.3.3).

8.2.3.1 Aggregation

The aggregation functional component provides a set of capabilities for gathering raw data and expressing in a summary form. Data aggregation enables the gathering of data through various

criteria, such as average, peak, maximum, sum and count. There are two types of data aggregation, namely time aggregation and spatial aggregation.

NOTE 1 – Time aggregation: all data points for a single resource over a specified time period.

NOTE 2 – Spatial aggregation: all data points for a group of resources over a specified time period.

This functional component provides:

- schema mappings which define how data are converted between the schema of an external data source and the integration session schema;
- support of various aggregation method such as time aggregation and spatial aggregation.

8.2.3.2 Cleaning

The cleaning functional component provides a set of capabilities for filtering data, smoothing noise data, and identifying and removing outliers to improve data quality. Data collected from various sources may affect the accuracy of the analysis results. The cleaning functional component provides better data quality, which helps organizations to make their data available for analysis phase.

NOTE 1 – Outlier refers to abnormal data in the dataset. If it is not trimmed out, data quality may be damaged.

This functional component provides:

- data filtering, which is the process of selecting a small portion of data set and using that subset for viewing or analysis;
- data cleansing, which is the process of detecting and modifying (or removing) corrupted or inaccurate data in a dataset. Data cleansing also refers to identifying incomplete or unrelated parts of data and then replacing, modifying or deleting the data;
- data smoothing, which is the process of removing noise from a data set. When datasets are utilized, data can be manipulated to eliminate or reduce volatile or other types of noise;
- identifying and removing outliers, which is the process to discover outliers and deleting them.

NOTE 2 – Outliers can be discovered by detection functions such as model-base method or graphical method.

8.2.3.3 Extract, transform and load

The extract, transform and load (ETL) functional component provides procedures for pulling data out of one or more datasets and placing them in a target system. During the ETL process, data are extracted from the source dataset and converted to a format that can be analysed and loaded in another storage system.

This functional component provides:

- extraction, which is the process of retrieving the representative features from a large number of datasets for further processing such as data analysis;
- transformation, which is the process of converting the format of the extracted data into the required form so that it can be placed in another dataset;
- loading, which is the process of writing the data into the target dataset.

8.2.4 Data source layer functional components

The data source layer functional components include:

- Data collection functional component (see clause 8.2.4.1);
- Data storage functional component (see clause 8.2.4.2).

8.2.4.1 Data collection

The data collection functional component performs data collection based on various data collection configurations.

This functional component provides:

- setting up various data collection configurations, such as data amount, traffic volume, collection period, collection method;

  NOTE 1 – Examples of collection methods include crawling, rich site summary collecting, log /sensor collecting.

  NOTE 2 – Rich site summary is used to aggregate syndicated web content, such as online newspapers, blogs, podcasts and video blogs in one location.

  NOTE 3 – Crawling is used to gather data from the web, especially the web indexing.

  NOTE 4 – Log collecting is used to collect data from log files generated by web servers.

- gathering data based on established configurations of data collection. The collected data is stored in an appropriate storage according to the data type.

  NOTE 5 – This functional component is same as the data collection functional component in [ITU-T Y.3519]

8.2.4.2 Data storage

The data storage functional component provides a set of capabilities for storing data. This functional component also provides different types storage for different data types and different database types while storing data.

This functional component provides:

- provisioning storage considering the various types of data storage, database, and different types of data such as structured data, unstructured data, and semi-structured data

  NOTE 1 – Data storage type includes block storage, file storage and object storage.

  NOTE 2 – Database includes relational database, not only SQL (NoSQL) database.

  NOTE 3 – Unstructured data can include mass data, such as log files, video, audio data, email, Web pages, data generated on social-media sites. Semi-structured data can include data stored in extensible markup language (XML), hypertext markup language (HTML) and other format documents. Structured data can include record data persistent in databases (see [ITU-T Y.3600]).

- allocating the appropriate storage when a storage usage request comes in;

- releasing storage when the storage usage is terminated;

  NOTE 4 – Data storage functional component interworks with the data collection functional component (see clause 8.2.4.1) to identify the characteristics of the data such as data type, data volume and so on.

- storing data on various storage. It supports storage mirroring and provides data fragmentation to distribute and store data on distributed storage systems. This provides the ability to update data;

  NOTE 5 – Distributed storage system stores data on multiple independent storages. It adopts the scalable system structure, and uses multiple storage servers which are used to share the storage load.

  NOTE 6 – Storage mirroring is the replication of logical storage volumes onto separate physical disks.

- data indexing which are stored together with data to improve the speed of data retrieval operation;

- removing duplicated and redundant data while storing data.

  NOTE 7 – This functional component is same as the data storage functional component in [ITU-T Y.3519] expect last function "removing duplicated and redundant data while storing data".

8.2.5 Multi-layer functions

8.2.5.1 Analysis supporting system

8.2.5.1.1 Data catalogue

The data catalogue functional component provides a listing of all metadata which a data broker makes available [ITU-T Y.3601]. The data catalogue functional component is responsible for interfaces for data searching, metadata mapping and tools for generating data catalogue. The data catalogue acts as an inventory of available data, allowing the big data service customer (BDC) to search and find data. The BDC is able to view the metadata of all available data sets, and to select data sets.

This functional component provides:

- data searching interfaces;
- metadata mapping;
- tools for generating data catalogue.

NOTE 1 – The tools have ability to generate a data catalogue using information about the creation of the data sets and informal description of the data sets.

NOTE 2 – There are requirements and conceptual models of the data catalogue in [ITU-T Y.3603].

8.2.5.1.2 Workflow management

The workflow management functional component is responsible for defining, creating and managing workflow execution. Workflow is all or part of a process in which data, information or tasks are passed from one part to another part for action in accordance with a set of procedural rules. In data processes, workflows are widely used across the data collection, data storage and analysis processes. In the case of data analysis, workflow can be designed in conjunction with a series of tasks for loading, refining, consolidating and analysing data.

This functional component provides:

- defining and creating workflow;
- managing the execution of workflow.

8.2.5.1.3 Monitoring and reporting

The monitoring and reporting functional component is responsible for monitoring the big data services with service implementations to ensure that they meet agreed service level agreement (SLA) targets, including the reporting and resolution of problems with those services.

The monitoring and reporting functional component provides:

- monitoring the use of big data services by the BDC such as the service access and service implementation;
- reporting results or problems derived from data collection, processing, analysis and visualization;
- reporting results of the utilization of each data.

8.2.5.2 Data life-cycle management

8.2.5.2.1 Data provenance management

The data provenance management functional component is responsible for managing provenance lifecycle and policy related to data provenance. It supports analysis supporting (control workflow, data analysis automation, etc.) and monitoring for change of computational environments.

This functional component provides:

- recording provenance information when the state of data changes, that is, when it is stored, when it is changed, or when it is deleted;
- keeping, combining and deleting provenance information when data is deleted from storage;
- aggregating provenance information according to aggregate big data.

8.2.5.2.2 Data licence management

The data licence management functional component is responsible for notifying copyright, and managing licence information of data. Consideration and management of data licences are required during processing data or generating data.

This functional component provides:

- setting forth in the data licence agreement which are generated by the data creator or data seller;
- provision of the data licence agreement defining the licensed data including usage manner and frequency limits with how the data is provided and updated.

NOTE – Terms defined in data licence agreement may include the use of encryption and a secure delivery mechanism, designated communications technology platforms, and specific hardware or software configuration requirements.

8.2.5.2.3 Data preservation management

The data preservation management functional component is responsible for managing operation for big data preservation such as selecting and preparing the data to be preserved, preserving and retrieving the preserved data under policy rules which should be referenced and obeyed. Big data preservation is a combination of policies, strategies and actions developed by a big data service provider (BDSP) to ensure that digital information of continuing value remains accessible and usable within a big data ecosystem. Refer to [ITU-T Y.3604].

This functional component provides:

– managing data preservation operation in accordance with various preservation policy rules;

NOTE 1 – Examples of preservation policy rules are setting preservation selection rules, preservation period rules and preservation storage rules.

NOTE 2 – The preservation storage rules specify which storage tier these data should be stored in and in which conditions they should be migrated to other storage tiers.

NOTE 3 – The preservation period rules specify how long the data should be preserved in big data service provider (BDSP).

– verifying privileges of the preserved data requestor to access the preserved data;

– setting preservation selection rules, preservation period rules and preservation storage rules.

8.2.5.3 Security

8.2.5.3.1 Privacy management

The privacy management functional component provides a set of capabilities for managing personal information in data. This functional component aims to avoid data being collected, stored by or disclosed to whom it may not be appropriate.

This functional component provides:

- methods to protect the privacy of confidential data and sensitive data. For example, this function supports the data desensitization to protect the sensitive data;

NOTE 1 – Confidential data refers to provide for protection of data from unauthorized disclosure (see [b-ITU-T X.509]).

NOTE 2 – Sensitive data refers to personally identifiable information or other sensitive information which is collected, stored, used and finally destroyed or deleted.

- managing personally identifiable information (PII) to check whether a data instance contains PII when recording a provenance unit.

  NOTE 3 – This functional component is based on the privacy management functional component in [ITU-T Y.3519] and [b-ITU-T Y.3602].

8.2.5.3.2 Authentication management

The authentication management functional component provides the capabilities to manage identification so that only authenticated users shall access specific capabilities or data.

8.2.5.3.3 Authorization management

The authorization management functional component provides capabilities for controlling and enforcing permissions that allow users to access specific capabilities or data.

9 Cross cutting aspects

Cross-cutting aspects can be shared and can impact multiple roles, activities and functional components, as described in [ITU-T Y.3600] and [ITU-T Y.3519].

9.1 Data redundancy

Data redundancy refers to the repeated occurrence of the same data in the system. For example, in a relational database, data redundancy mainly refers to the repeated storage of the same data in the relational database, including the repetition of tables, attributes, tuples and attribute values. Necessary data redundancy can improve the anti-interference ability of data, thus preventing data loss and error. For example, redundantly encoding data by adding several bits based on the length of the original binary code, to prevent key data loss and errors. For more details, see clause 8.1 in [ITU-T Y.3519].

On the other hand, data redundancy should be minimized to improve storage space utilization, but in some cases, data redundancy should also be increased appropriately. Data compression and de-duplication are two key technologies to reduce data redundancy. For more details, see clause 8.1 in [ITU-T Y.3519].

9.2 Performance

Performance includes a set of non-functional facets relating to the operation of a big data service such as:

– data usability: this means that the data user (BDSP or BDC) can access data with enough metadata to understand its meaning;

– findability of data: this means that the data user (BDSP or BDC) can retrieve data with enough metadata according to different requirements;

– availability of the big data service;

– response time to complete big data service requests;

– transaction rate at which big data service requests are executed;

– latency for big data service requests;

– data throughput rate (input and output);

– number of concurrent big data service requests (scalability).

– When running an application in a service, the same performance aspects apply to the behaviour of the application running in a big data service provider's environment.

9.3 Data quality

Data quality can be assessed differently depending on the environment and target in which the data is used. Therefore, it is difficult to quantitatively evaluate data quality through some criteria. However, in some cases, high data quality means that the data used is accurate, usable and valued for its intended purpose.

Data quality is defined through several dimensions such as accuracy, completeness, consistency, timeliness, validity and uniqueness. More specific dimensions are defined depending on data usage context.

NOTE 1 – Accuracy refers to the degree to which data reflects a real object or event.

NOTE 2 – Uniqueness refers to point out that there should be no data duplicates reported.

NOTE 3 – Timeliness refers to the degree to which data represent reality from the required point in time.

NOTE 4 – Validity refers that data are valid if it conforms to the syntax (format, type, range) of its definition.

NOTE 5 – Completeness refers to whether all available data is present.

NOTE 6 – Consistency refers to providing the same data for the same object even if the data is displayed in different reports.

10 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. Security guidelines for the analysis of big data security vulnerabilities and big data lifecycle management are provided in [b-ITU-T X.1751]. The security roles and responsibilities of components for big data services are identified, and the security framework for platforms, applications, analytics, interfaces and big data infrastructure are provided in [b-ITU-T X.1750].

Bibliography

[b-ITU-T H.752] Recommendation ITU-T H.752 (2015), *Multimedia content provisioning interface for IPTV services*.

[b-ITU-T X.509] Recommendation ITU-T X.509 (2019), *Information technology – open systems interconnection – the directory: public-key and attribute certificate frameworks.*

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

[b-ITU-T X.1361] Recommendation ITU-T X.1361 (2018), *Security framework for the internet of things based on the gateway model.*

[b-ITU-T X.1750] Draft 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] Draft Recommendation ITU-T X.1751 (2020), *Security guidelines on big data lifecycle management for telecommunication operators.*

[b-ITU-T Y.2201] Recommendation ITU-T Y.2201 (2011), *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.3500] Recommendation ITU-T Y.3500 (2014), *Cloud computing – overview and vocabulary.*

[b-ITU-T Y.3602] Recommendation ITU-T Y.3602 (2018), *Big data – functional requirements for data provenance*.

[b-ISO/IEC 26550] ISO/IEC 26550:2015, *Software and systems engineering – reference model for product line engineering and management.*

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

Printed in Switzerland
Geneva, 2020