Draft Recommendation ITU-T X.websec-7

Reference monitor for online analytics services

Summary

Recommendation ITU-T X.websec-7 provides specification of common reference monitoring mechanism to prevent a linkable or identifiable analysis.

Keywords

Reference Monitor, Online Analytics

CONTENTS

1 Scope 3

2 References 3

3 Definitions 3

3.1 Terms defined elsewhere 3

3.2 Terms defined in this Recommendation 3

4 Abbreviations and acronyms 3

5 Conventions 4

6 Overview of Reference Monitor 4

6.1 Basic Reference Monitor 5

6.2 Reference Monitor for Network based Analytics Application 6

7 Security threats and countermeasures in reference monitor 7

7.1 Security requirement for data protection in reference monitor 7

7.2 Security requirement for code protection in reference monitor 8

7.3 Security requirement for resource protection in reference monitor 8

8 Reference Monitor for analytics service 8

8.1 Reference Monitor for On-Line Analytical Processing (OLAP) 9

8.2 Reference Monitor for Network based Application System 10

9 Reference Monitor for video and audio analysis service 11

10 Reference Monitor for relational data analysis service 11

Draft Recommendation ITU-T X.websec-7

Reference Monitor for Online Analytics Services

1 Scope

This Recommendation provides specification of common reference monitoring and rule set to prevent an illegal or unintended data composition or analysis. It includes the operation of reference monitor and rule set for interoperability.

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.

<TBD>

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 access control [ITU-T X.800]: The prevention of unauthorized use of a resource, including the prevention of use of a resource in an unauthorized manner.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 reference monitor: it enforces an access control policy over subjects, which is various data sources and has an ability to perform operations on objects.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACLs	Access Control Lists
MLS	Multi Level Security
SMS	Short Message Service
RBAC	Role Based Access Control

<TBD>

5 Conventions

In this Recommendation:

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

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

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

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

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

6 Overview of Reference Monitor

In mobile application, user data is aggregated and analysed by service or contents provider in order to provide a better convenience service. It tries to understand user behaviour or opinions by analysis of user profile and social activities. However, many people do not know that the data can be used for another purpose such as marketing and advertisement even if the company take a user agreement. Therefore, analytics application should provide legitimated reference monitoring mechanism to prevent a linkable or identifiable data with those collected data.

Sometimes, personal information seep into the black market, it causes a lot of troubles such as a SMS spam or phishing. However the technical issue in the mobile application how to manage and protect a personalized device and personal assistance SW. For example, the voice recognition is used for the understanding of human command, but the next step of this technology is an understanding of communication voice over the telephone. These personalized services understand the trends or pattern of user behaviour (includes intention in the telecommunication message) and prepare the assist function such as calling a phone, send a message, add a new schedule and alarm appointments.

Those personalized applications provide a convenient assistant and increasing the human ability, includes knowledge retrieval, collective intelligence and long-term memory. But, if there is somebody who knows those data and also can identify who is owner, then this personal information can be abused.

To prevent those abuses, reference monitoring mechanisms are required for the analytics application such as Figure 6-1. The reference monitoring has a rule set for detection of abuses. In this case, the rule is made for the un-linkable and un-identifiable analysis. Hence, reference monitoring mechanisms are required for the analytics application and the rule of the referencing is interoperable between applications.

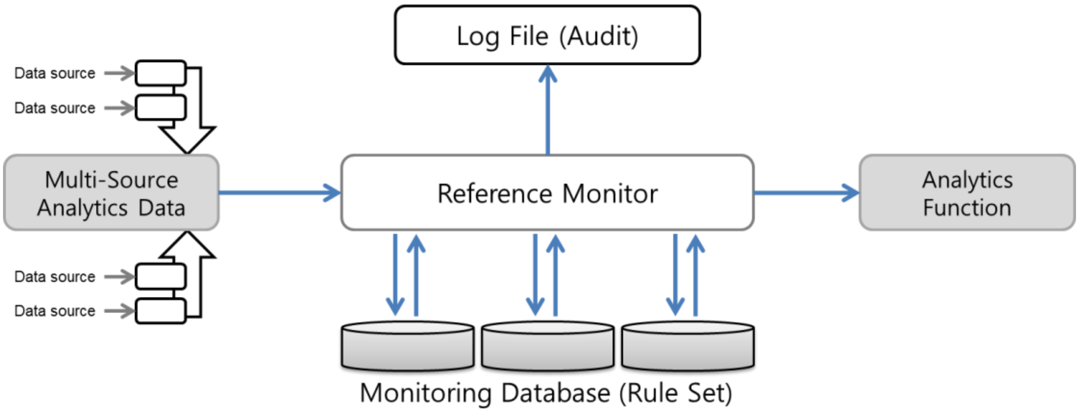


Figure 6-1 Rule based Reference Monitor

6.1 Basic Reference Monitor

Reference Monitor is used for the mechanism of access control between subject and object. In a distributed system, subject can be a person, processes or system, but otherwise object can be a computer resources such as memory, files, sockets, system functions or printers.

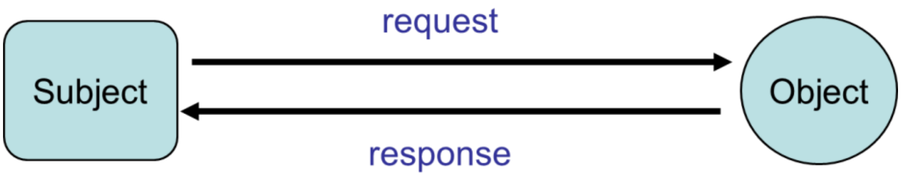


Figure 6-2 Access Control between Object and Subject

In the Figure 6-2, the subject sends a request to the object. The request is a getting privileges for the access of object. The privileges are generally “read”, “write” and “execution” (launch a function). In a distributed system, the system should provide a decision which operation is permit. The decision always based on the authorization result and the decision logic should be tamper-free.

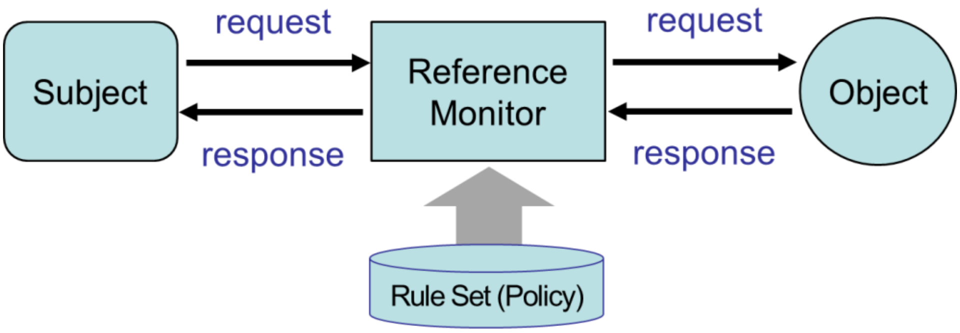


Figure 6-3 Basic Reference Monitor

The Figure 6-3 depicts the fundamental model of reference monitor. It has a role of mediator for subject and object and also it provides all decisions which is based on the rule set. The rule is defines which subject has an access permission for the object such as ACLs (access control lists) and MLS (Multi-Level Security). If we use an intermediate node (group), then make a permission graph based on the group permission and negative permissions. It is Role-Based Access Control (RBAC).

For the basic reference monitor, following three requirements should be satisfied;

1. Complete Mediation: all request and response are passing through the reference monitor and it is always working
2. Tamper-proof: The reference monitor and policy should not be changed by unauthorized object or subject
3. Verifiable: When we want to check the working of the reference monitor, the system should provide a test and verifiable functions

When we implements a reference monitor, inline method are widely used for maintenance advantage. In the figure 6-4, left depicts the reference monitor in subject. The subject request a decision to the internal reference monitor and the object performs the decision, but the rule set is located in the outside and reference monitor is standalone logic in the subject.

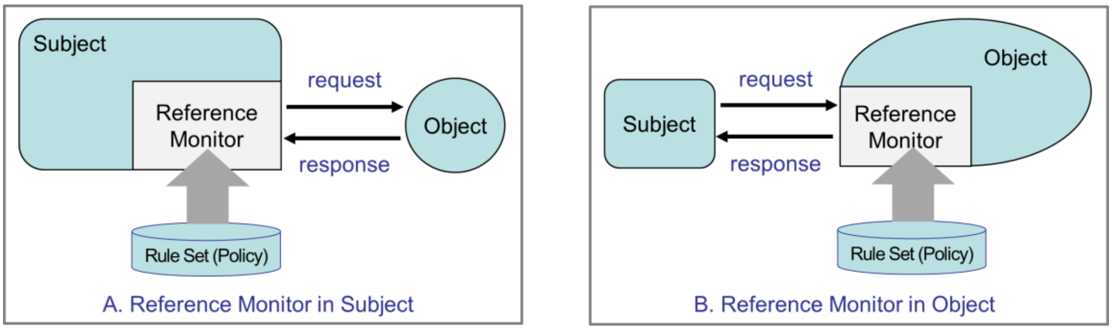


Figure 6-4 Inline Reference Monitor

Reference monitor in object has a little difference, because the object should provide an isolated function or logic to the subject. The subject send a request message to get a function of object. For example, the subject is mobile phone application and object is face recognition server. The mobile phone (subject) take a picture and send a request message to the server (object). The face recognition server check the privilege of request owner (mobile phone) and make a decision which level of service provides.

6.2 Reference Monitor for Network based Analytics Application

Analytic mechanisms are performed in the personal device or networked storage such as personal cloud or personalized server. In the personal devices, all the data is stored in the personal device and there are no communication channels for the analysis, the reference monitor within devices just controls or prevents the data leakage or covert channel. However, the networked application or network based analysis application share personalized data with the network server. Therefore we have to control two point of data flow, first one is communication channel and second one is server system.

The communication is normally encrypted or well protected because the data itself is important and the service operator has a duty of protection. But how many communication channels are used and who can access this data, is provided by the reference monitor.

Network based application system [Figure 2], many analytics functions are performed in the server system. And the personalized profiles are distributed to the mobile devices. Then the device provides personalized services or assistant services.

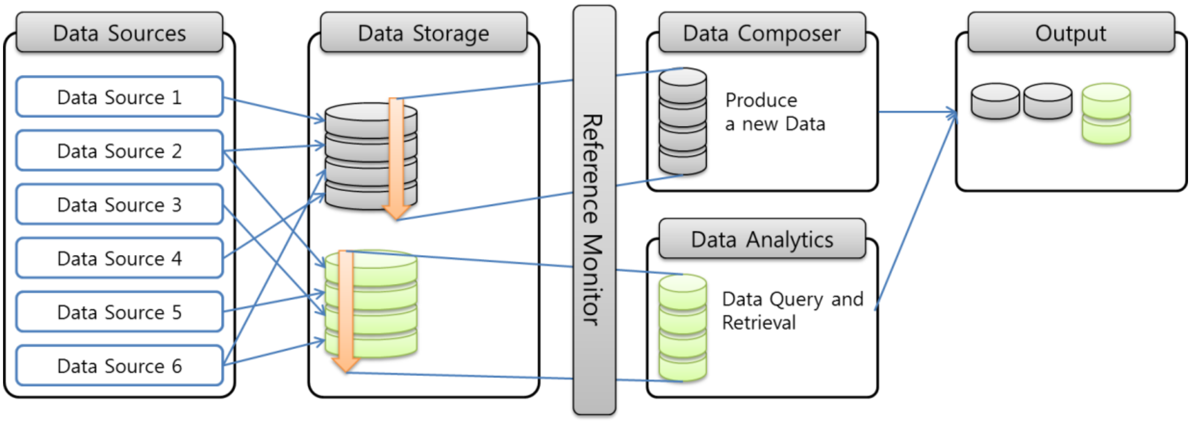


Figure 6-5 Reference Monitor for Network based Analytics Application

In the analytics application, several kind data from diverse sources are aggregated to the data storage and the stored data will be used for the analysis. Raw data in the data source is delivered to the analysis server and the server do the first level filtering in the aggregated data. Data composing and multiple data analysis are performed in the server. Therefore reference monitor should located in the analysis server. Even though the edge analysis application do the first level data filtering in the local machine, they just used their own data, not used the other side data. Therefore, there is no illegal or unintentional data leakage.

The figure 6-5 depicts the generic analysis processing with diverse data sources. All aggregated data is stored in the data storage all together, if the analysis server run a distributed and clustering system, then the storage system used distributed file system and nobody know where data is located and what data is stored. In the system, only analysis processor, which is main process of data analysis, can handle the data and accesses. The reference monitor between storage and data analysis processor is right location and do monitoring. The other case is high level data translation. Originally collected data is coming from diverse source and diverse format, therefore all aggregated data is translated for analysis and that time new translated data can be produced and also the data can make an illegal or unintentional data leakage.

7 Security threats and countermeasures in reference monitor

Reference monitor has two assets to protect from outside components, it is security policy and reference monitor program itself. The security policy is periodically updated and it includes how to monitoring the access privilege and data usages. The other one is program code for reference monitor, which is performing an access control in the system. In traditional system, reference monitor is located within operating system or hardware logic such as Trusted Platform Module (TPM), however the reference monitor for analytics application is not guaranteed that the program of reference monitor itself is well protected.

Therefore, security threats are summarized with data protection, code protection and execution environment protection such as resource protection.

7.1 Security requirement for data protection in reference monitor

In reference monitor, some data is very important and should be protected because the host machine can be a malicious and internal data can be revealed. Therefore analytics application should provide a way to protect their secure and important data such as private key file, security policy file, and personal profile. Security requirements related to data protection in reference monitor are as following;

- Reference monitor provide secure storages or secure ways to store data.

- Reference monitor provide a secure update mechanism for the security policy.
- Reference monitor provide a way to check the integrity of security policy.
- Reference monitor provide a client function for real-time key distribution method.

7.2 Security requirement for code protection in reference monitor

If the reference monitor program itself is cracked or revealed the monitoring results in the host server or analysis server, then the reference monitor does not needed. However the protection of program code has three approaches. First one is marshalling or scrambling the program code, second one is testing and hiding program code and last one is code encryption

- Code marshaling and scrambling: marshaling and scrambling is changing the instruction and program header file to hide the internal code. The marshalled program is testing the execution environment before the demarshalling. If the execution environment is not secure, then the program is stopped and exit. However, if the testing results are safe, then the original instruction code are executed.
- Code testing and hiding program code: comparing the code marshaling, this approach also testing the execution environment and it is decides launch a secure code or not. Changing an original header of the security program to a first header and transmitting a packed program including the encrypted code blocks and the changed first header to the server.
- Code encryption: last one is simple encryption all program code and if the user has a privilege, then program code are decryption

7.3 Security requirement for resource protection in reference monitor

Communication channel, memory protection, process isolation, instruction protection

<TBD>

8 Reference Monitor for analytics service

Architecture of online reference monitor is based on the virtual machine or cloud based analysis system. Commonly, online analytics services are sharing the aggregated data to detect a specific situation or event, but the combination of shared data can make a data leakage, which are prohibited data to outside of the data provider. Therefore reference monitor is designed for the controlling and monitoring the usage of data and its combination.

The following figure is illustrated 3 phases for analytics, the first de-identification, the second data combination and the last reference monitor check.

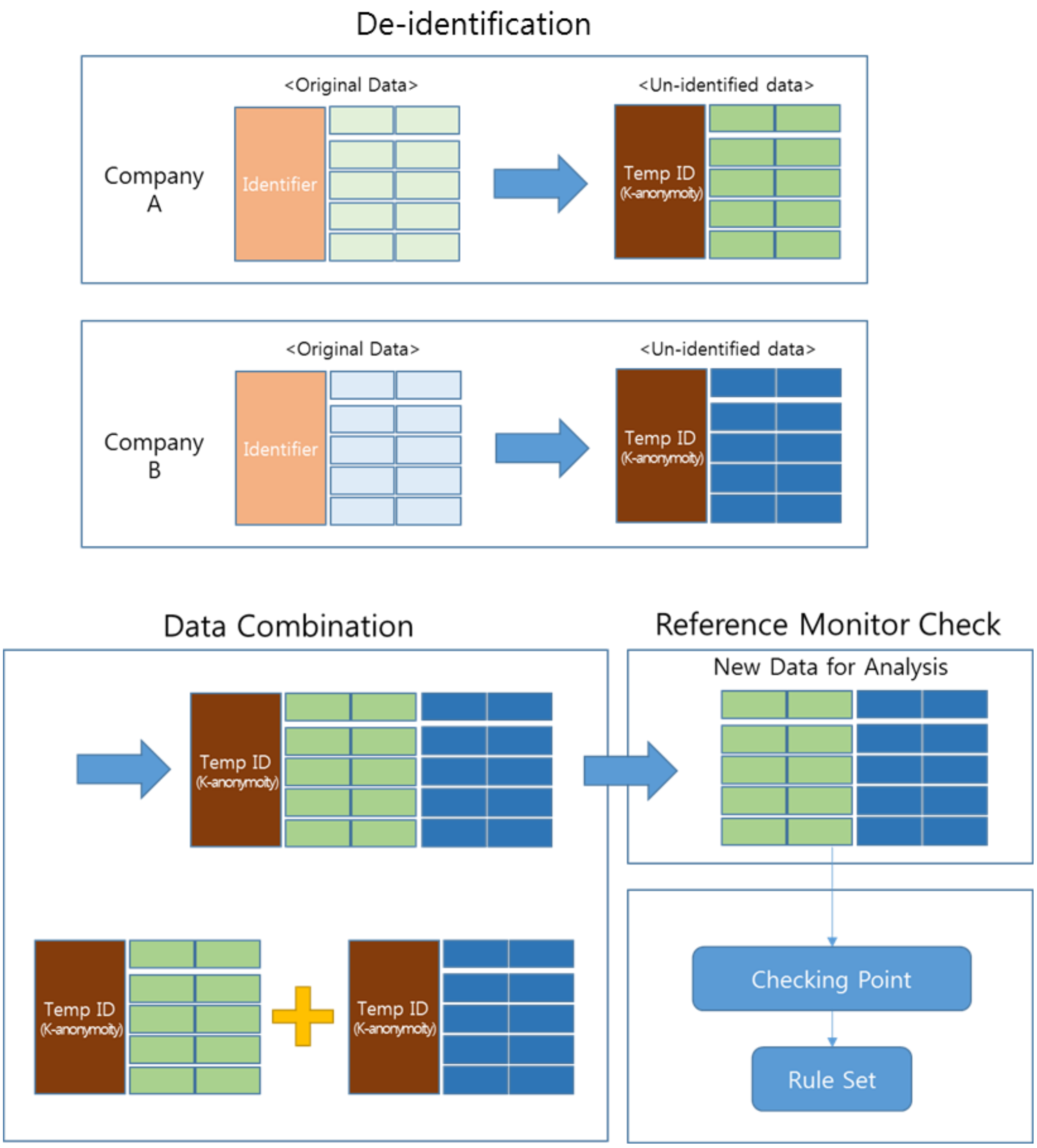


Figure 8-1 Data combination and Reference Monitor

When two companies or data sources want to use combined data for analytics, new data is generated with two de-identified data set and reference monitor check the status and violation of rule set. If the new data is violated, then the reference monitor reports to the analytics service about the combination of two data sources is not allowed

In an analytics service, reference monitor is designed with inline reference model and the location of reference monitor is server which performs analytics services. However, the rule set is located in the separated place and the contents are automatically updated.

In this clause, we will provide a common reference for the online validation check mechanism in the analytics services.

<TBD>

8.1 Reference Monitor for On-Line Analytical Processing (OLAP)

Public analytics service or batch analysis is performed for the understanding of preference or trends of user’s common behavior.

<TBD>

8.1.1 Code protection for On-Line Analytical Processing (OLAP)

Program code of reference monitor enforces a rule-based access control or monitoring the processing unit based on security policy. If the program code can be changed in the server or terminal devices, the testing and monitoring results also not be trusted. There are many reference systems to care the integrity and confidentiality of reference monitor program code.

In On-Line Analytical Processing program code should be encrypted and check the execution environment which is safe or not. Therefore the program code has three phases for the integrity and confidentiality.

First phase is program packing with encryption key.

- Original reference monitor program has several part of code blocks. Code distribution server analysing the reference monitor program for the program transmitted to a client in a secure way
- Each blocks are encrypted by using an encryption key, the selection of key depends on the code analysis result. The code is divided into normal code, secure code and environment check code. The corresponding key is received from a key distribution server through a network.
- Reference monitor program header also changed for the first run of checking program, which are check the execution environment and decide which codes are executed. It is depends on the checking result. If the system is safe and they had a proper authority, then the original program codes are decrypted and the second location value of header is setting the original reference monitor program code. Otherwise the second location value of header indicates dummy code.
- Finally, the packed program including the encrypted code blocks and the changed header are transmitted to the client.

Second phase is unpacking of reference monitor program with decryption key.

- Receive a packed program from a server connected to the client through a network, the packed program including an encrypted code and a changed first header.
- Restore an original header from the changed first header of the packed program.
- Unpack the packed program by using a decryption key received from a key distribution server, the decryption key corresponding to an encryption key used for the encrypted code.
- Execute a reference monitor code corresponding to the packed program when the unpacking is finished.

Third phase is checking the execution environment and decrypt secure code.

- Whether there is a possibility of hacking of the received packed program, and controlling the restoring to be executed only when there is no possibility of hacking.
- The checking determines whether an environment in which the received packed program is executed is exposed to hacking and whether a hacking tool exists.
- The reference monitor program is executed only when the client has legal use authority.

8.2 Reference Monitor for Network based Application System

Private analysis service is performed for the understanding of user preference.

<TBD>

8.2.1 Analysis in the server system

All user data is stored in the server system, and the preference analysis is also performed in the server.

<TBD>

8.2.2 Analysis in the terminal device

User data is selectively stored in the server system, but the preference analysis is performed in the terminal device.

<TBD>

9 Reference Monitor for video and audio analysis service

Video and audio data is not regulated and the contents are recognized after analysis, however, in the sharing time, what kinds of contents will be shared is not determined yet. In this clause, we are additionally analyzing the video and audio content, which includes personal and important data.

<TBD>

10 Reference Monitor for relational data analysis service

In analytics process, relational data (for example, social relation, community, group messaging) can make a new data type and cause a new data.

<TBD>

Bibliography

<TBD>

Annex A

<knowledge metadata for analytics application>

(This annex forms an integral part of this Recommendation.)

Visual Knowledge is a set of text based data and it provides a representation facility for machine, some motion and objects in video/image are translated into the metadata, which is visual knowledge. The object and activity in movie is just a data pattern and so many variations are possible to understand but if we had a formal form and translate tool, then we can make an analytics function for those data (unstructured data) such as image, video, chart and tables.

In visual data, we can approach two types of categories;

- Syntactic: physical components and what is appeared in the image.
- Semantics: understand situation and geometric relations in the image

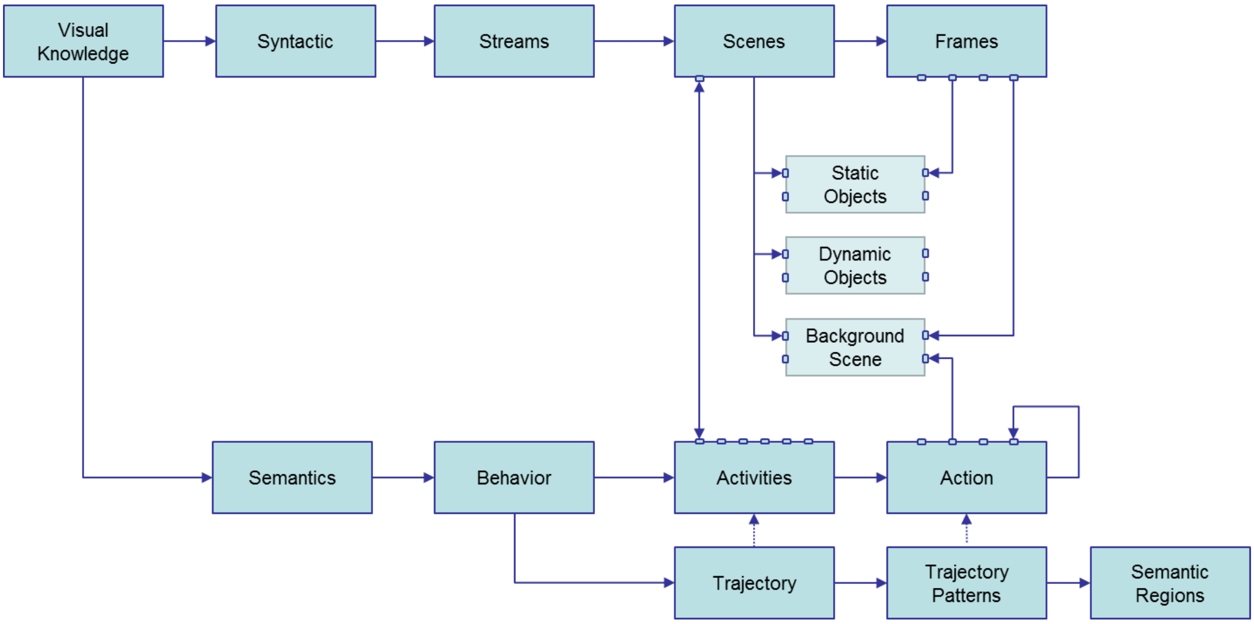


Figure A-1 – Structure of Visual Knowledge

Video/image had three components, which are object and object activity and background.

Syntactically video has stream, scene and frame. scene and frames are used for analysis.

- Object: in the image, several objects are existed and individual objects has a relation with other objects and doing action such as carry, kick, push and run.
- Object activity: some of objects are not moved such as candle, cup, television, audio and umbrella. It is static object. Otherwise dynamic object can move to other location such as car, bicycle, airplane and motorbike. These moving action makes a activities with a set of action.
- Background: it means “where picture is taken?”. Location and location categories such as zoo, park, aprtment, department store and gas station.

[Editor’s note: if anybody know idea or standard for visual knowledge representation, we will discuss about it in the next meeting. And it is used just for reference]

Appendix A

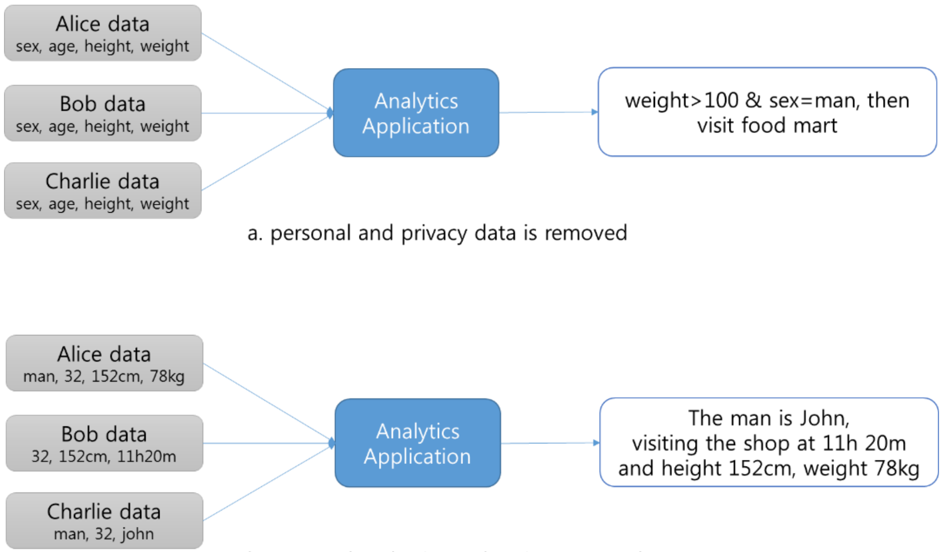
<Online Analytics Application and privacy>

(This appendix is not part of this Recommendation.)

<Data analytics>

Data is used for the analysis of current status and predict rear future with the accumulated data, it is based on the stochastic and pattern of data. Computer program can predict the next step based on the current data status and accumulated data patterns which is training or learning mechanism.

In the analytics application, many information is generated based on the private/individuals information such as “sex, age, height, weight, home land, favorite food/restaurant”. When we predict the shopping path of an entered man in a shopping mall, we want what is interest of the man and which item is good for the man. It is based on his personal information and we analysis the pattern as an accumulated data. In the training and learning phase, the private information is derived into the generic data excluding private and personal data because the analysis mechanisms are used to make knowledge of those data (Figure A-1.a).



a. personal and privacy data is removed

b. personal and privacy data is generated

Figure A-1 privacy issues in analytics application

Otherwise, some application generates personal and privacy related data such as Figure A-1.b.

Appendix B

<Identifiers>

(This appendix is not part of this Recommendation.)

Identifiers are used to identify object which owns data including person and company. There are two type of identifiers, first one is directly representation of target object and second one is indirectly representation of target object. Indirect means that after some reasoning, we can guess which is owner of data. In this appendix, we are listing the possible identifiers and we are considering those identifiers for the target of protection in reference monitor.

< Identifiers for data analytics >

#	Name of Identifiers	Description
1	Social Number	Officially identification numbers including social security number and personal number.
2	Phone numbers	Personal phone number, which is private and not shared number
3	Electronic mail addresses	Public or private email address to send or receive a message from others
4	Medical record numbers	In medical service, identification number for the management of patient or outpatient
5	Health plan beneficiary number	To serve a health care including hospital, medical, and other health care services
6	Account numbers	It is used for bank account management and transfer money to others
7	Credit card number	Bank credit cards including magnetic, IC and internet based account services
8	Certificate/license number	Automobile license number and related certificates
9	Vehicle identifiers/Serial number	The vehicle is owned by someone and the owner is equal to the vehicle identifiers
10	Device identifiers/Serial number	Personal devices such as mobile phone and electronic watch
11	Internet Protocol (IP)	If the IP and related address is assigned to a person, then it is a identifier of the owner
12	Biometric identifiers (Finger/Voice Prints)	A specific biometric information is belonging to personal and not changed
13	Face images	Face images can be used for identification and the images itself say who is.

Identifiers are directly used for who you are and semi-identifiers are used for the identification with some combination.

< semi - identifiers - attribute value >

#	Name	Attributes
1	Personal attributes	Gender, Age, Nationality, Born city, Marriage, Military service, Religion, Habit, Clubs.
2	Physical attributes	Blood type/pressure, Height, Weight, Waist size, Eye colour, medial record
3	Credit attributes	Tax, Credit level, Donation record, Insurance record, income
4	Career attribute	School, collage, University and major, Work experience, profession and affiliations
5	Electrical attribute	Internet access information (Visit date, time and related log), mobile phone and GPS data
6	Family attributes	Related information for Father, Mother, Wife/Husband, Son and Daughter

Some of collected attribute values can make an identifier, so there are many restriction to use these data in analytics application.

Appendix C

<De-Identification methods>

(This appendix is not part of this Recommendation.)

De-Identification methods remove identifiers in data and we cannot guess who is owner of data. In this appendix, we summarized well known methods and we are considering those methods for de-identification and re-identification.

< List of de-identification methods >

#	Methods	Description
1	Pseudonymization	[TBD] Heuristic Pseudonymization K-anonymity Encryption Swapping
2	Aggregation	[TBD] Aggregation Micro Aggregation Rounding Rearrangement
3	Data Reduction	[TBD] Reducing Variables Reducing Partial Variables Reducing Records Trivial Anonymization
4	Data Suppression	[TBD] Data Suppression Random Rounding Data Range Subdivide Level Controlling Controlled Rounding
5	Data Masking	[TBD] Adding Random Noise Blank and Impute