

[Gonçalves&all18]. i^* is the widely used GPML by the research community [Gonçalves&all18]. It supports extension to model requirements for specific domains such as the Internet of Things, big data, security, etc. In the literature, big data security requirements are generally modeling as quality requirements [Arruda&Madhavji18] despite its important impact on decision making and service trust. Therefore, security requirements should be mapped to functional requirements to avoid the security impact that can big data outsourcing projects suffer from.

This paper is about big data security requirements and their modeling. It presents the big data life cycle phases, main Cloud services security threats, main issues and security requirements of outsourcing big data in a Cloud environment, and enhanced security solutions. To model security requirements for the Cloud-assisted big data, a conceptual extension to i^* model is proposed and a general model for big data security requirements are generating using our i^* extension. An Electronic Healthcare Record (EHR) example is proposed to illustrate big data security requirement modeling using our proposed i^* extension.

2 State of the art

In this section, we describe briefly the big data life cycle, security requirement engineering, Cloud services threats, and big data security requirements that are related to our work.

2.1 Big Data Life Cycle

The big data life cycle can be divided into five phases as shown in Fig1. In each phase, multiple steps are involved. Data generation is the first phase of big data cycle where data are generated continuously from various distributed data sources. Such data sources include social media, connected devices, sensors, etc. These data are then collected, transmitted, and pre-processed. These three steps form the second big data life cycle. In the collection step, various collection tools and techniques are used such as Log files, sensory data using sensors nodes, etc. The collected raw data are then transferring to distributed data storage infrastructures. Since data are collecting from various sources, the datasets may contain redundant data, uncertain data, noise, etc. They must pre-process to ensure the accuracy of data and trustworthiness of processing results. The pre-processing phase comprises integration, cleaning, and deduplication of datasets. In the integration step, various data are combined and transformed into the same format to provide unique data view to the user. The cleaning step addresses incomplete, meaningless, and uncertain data to ensure data consistency and deduplication eliminates the redundant data to save storage space and gain in processing performance. Data storage phase refers to the process of storing and managing datasets in distributed storage infrastructures based on various data management techniques. Various analytical methods are applied to the stored data to derive meaningful information in the analysis phase [Chen&all14]. The result of the analysis phase will be visualized to end-user using various tools and interfaces.

In order to secure big data in the Cloud, it is important to identify security requirements within these phases. Therefore, a security requirement engineering process must be carried out with requirement engineering steps for big data services projects in each big data lifecycle phase.

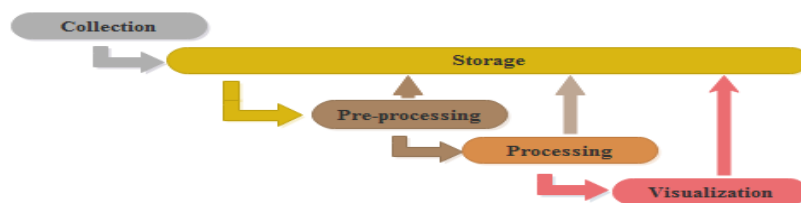


Figure 1 Big Data Life Cycle

2.2 Security Requirement Engineering

Recently, an increasing part of enterprises and organizations use big data to extract insights and improve their decisions. Sensitive data are outsourcing for storage and processing in an external environment such as the Cloud to benefit from its platform, software, and infrastructure distributed as services. Cloud offers various services that can deal with big data characteristics. These services are developed taking into consideration the requirements of consumers and clients such as large storage, management of unstructured data, etc. in the requirement engineering process. However, security

requirement selection and modeling are needed in the early stage as functional requirements to ensure the privacy and security of consumers and client’s data, and the quality or correctness of the processing and mining results. In the literature, security requirement engineering proposals are divided into approaches that use an off-the-shelf framework such as UML, i*, and KAOS to model the security requirements and approaches that enhance existing requirement engineering frameworks with security constructs [Giorgini&all05]. Requirement modeling is carried out in the elicitation phase using a modeling framework. In this paper, we use the i* because it is the very used modeling language by academic research as it is extensible [Gonçalves&all18]. It is a goal-oriented requirement engineering method that has been proposed in [Yu&Eric95]. i* focuses on actors and provides an answer to “WHO” and “WHY” questions to model requirements at the early and late phase of software development based on the dependency relationships among actors. It has two models basing on dependencies between actors: Strategic Dependency (SD) model and Strategic Rationale (SR) model. SR model represents a network model of dependencies between actors. There are nodes that represent actors and links between them where each link map out one dependency between two actors to accomplish a goal. An actor can be an agent or a role. SR model used to describe how each actor will achieve the goal [Yu&Eric95]. There are four types of elements in this language: goal, softgoal, resource, and task with its dependencies.

2.3 Cloud Services Security Threats

With the development of Cloud computing technologies, organizations and enterprises big data problems are transformed to services by using the Cloud delivered resources such as computing and storage. However, Cloud computing technologies proposed for big data are often developing without taking into consideration the security concerns [Madhvaraj&Manjaiah16]. According to Cloud Security Alliance (CSA) [Alliance13], there are nine critical threats to Cloud security that must contemplate when proposing new architectures in Cloud. These security threats and their relevance are presenting in Fig .2. Therefore, any security proposal approaches for big data architectures deployed in a Cloud environment must consider Cloud security threats and Cloud security requirements. These threats are briefly describing as follows [Alliance13, Mahajan15]:

Data Breaches : The multi-tenancy allows the Cloud to share the same infrastructure to multiple users. Thus, any breaches of this infrastructure will expose all data users.

Data Loss : Enterprises and organizations such as healthcare organization outsource their data to the Cloud to improve low cost and high efficiency and availability [Belle&all15]. However, the Cloud services and infrastructures are in exposure to attacks such as phishing attacks, spoofing attacks, malicious insider attacks, that makes user sensitive data into loss risk.

Insecure Interface and APIs: APIs are used by the Cloud users to manage their data and communicate with the Cloud service provider. Therefore, any security vulnerabilities in these interfaces can lead to unauthorized access to the user application.

Malicious Insiders : Malicious insider relies on a user that have rights to access data, manage them, or to access Cloud resources but abuse his position for its personal gain, and corrupt or reap sensitive data what makes it an attractive opportunity for adversaries.

Service Traffic Hijacking : is a kind of threats where a Cloud user account is hijacking, or his credentials are accessing and reusing by an attacker or unauthorized user. Then, this attacker can manipulate user data, access and steal confidential information, return faulty data to the client or redirect them to wrong sites.

Misuse of Cloud Services: The abuse or misuse of Cloud services threats relies on a malicious client of Cloud services that may avail from the computing power of Cloud servers to distribute malicious software, abuse virtual machines, or stage a Distributed Denial of Service (DDoS) attacks.

Shared Technology Vulnerabilities : Shared and multi-tenancy of Cloud software, infrastructure, and platforms with no isolation may allow a Cloud user to impact other tenants.

Denial of Service : Attackers in the case of denial of service attacks will attempt the availability of resources and prevent the Cloud user from being access to their data or services by allocating all the available resources or cause slow access to services by using many Cloud resources.

Insufficient Due Diligence : Various companies and organizations outsource their business into the Cloud without taking into consideration the Cloud risk and without a complete understanding of Cloud computing service. Thus, these organizations will be found with various issues.

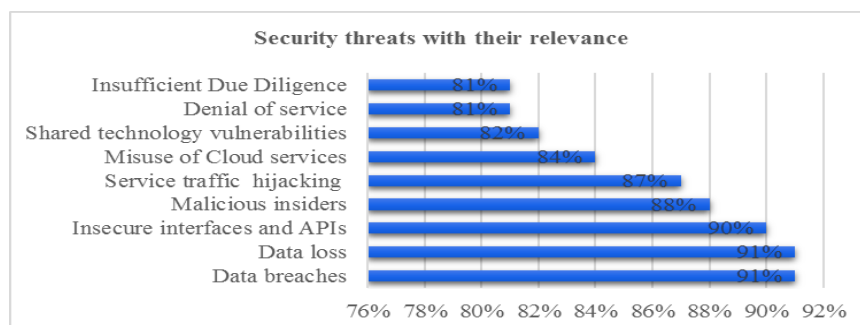


Figure 2 Cloud Security Threats with their Relevance [Alliance13]

2.4 Big Data Security Requirements

Big Data Vs make their storage and analysis using traditional systems and technologies complex. These Vs bring up various challenges and issues where security presents the most important. The Cloud Security Alliance (CSA) has identified the main security and privacy challenges that must be considered in big data application [CSA13]. Table 1 describes the main big data challenges and issues created by Volume, Variety, Velocity, and Veracity.

Table 1: Issues and Security Challenges by Big Data Vs

V	Issue	Security challenge	Security challenge identifier
Volume	Distributed storage	Cryptographically enforced data	C1
	Distributed processing nodes	centric security	
	Analytical skills	Secure transaction logs	C2
			Real-time security monitoring
Variety	Data access	Secure computation	C4
		Granular access control	C5
		Security for non-relational data stores	C6
Velocity	Information sharing	Data provenance	C7
		Granular audit	C8
		Scalable privacy-preserving data	C9
		mining and analytics	
Veracity	Fine database performance	End point input validation/filtering	C10
	Quality improvement	Cryptographically enforced data centric security	

For each security requirement, there are possible enhanced techniques that can be used. These techniques are desired solutions that must be improved in the real-world big data application where taking into consideration the efficiency and performance of the application. Table 2 presents the enhanced security techniques for each big data security challenges and possible threats. The identified threats need to be modeling in the security requirement elicitation phase in order to correctly specify the enhanced techniques that satisfy the security requirement goals.

3 Literature Review

In recent years, agents and goal-based modeling-oriented approaches are emerged in the security requirements engendering. However, few are considered big data security requirements modeling. In the following, we discuss some previous works in the literature focusing in the proposed models for big data security requirements.

In [Jutla&al13], the authors proposed privacy extensions to UML use case diagrams representing needed big data privacy services to help software engineers to learn about privacy requirements in the analysis phase. These extensions are implemented in Visual Studio as MS Visio extensions. A prototype in healthcare domain is presented to show how to use these extensions to model the privacy requirements and improve their utility. This proposal helps to include the user privacy choice during creation and generation of the big data application. However, it is not useful for including user security requirements after software creation.

In [Eunjang&al18], the authors propose an extended reference model called IRIS-Reference model (RM) for requirements of big data analytics by adding security and privacy as non-functional requirements. It contributes to provide theoretical foundations of big data analytics requirements in terms of efficient management of three big data Vs: Volume, Value, and Velocity.

In [Alshboul&al15], a security threat model is presented by integrating big data life cycle, security threats and attacks using a block representation. This model can conduct research in big data security and provide a clear foundation for research to secure big data. However, the model is not implemented, and no use case is presented by the paper.

In [Moreno&al18], authors proposed a security reference architecture for big data using UML models trying to ease security big data implementations and allowing to apply security patterns in order to secure final big data systems. This work presents clearly specified relationships between components.

All cited works consider big data security requirement as enhanced features in the creation process of the big data applications and services. However, security requirement modeling for end-users such as enterprises and organizations that intend to outsource its big data storage and processing on existing Cloud-big data services is needed. Furthermore, threats need to be modeled within security requirements to improve elicitation phase. In this paper, we propose a conceptual extension to i^* modeling language in order to support big data security requirement modeling for end-users.

Table 2: Possible Threats and Security Techniques for Big Data Security Challenges

Security challenge identifier	Possible threat	Enhanced security technique
C1	Infer sensitive information Data loss, Data breaches	Encryption techniques supporting search over encrypted data Attribute-based encryption Signatures schemes Fully-Homomorphic Encryption
C2	Collusion	Policy-based encryption
C3	Data poisoning Exploiting software vulnerabilities	Trusted hardware Tamper-proof software Cryptographic protocols
C4	Infer sensitive computation result Unsecure computing nodes	Cryptographic protocols Homomorphic encryption Trusted hardware
C5	Man-in-the-middle attacks Traffic hijacking	Authentication Authorization Leveled granularity of access
C6	Man-in-the-middle attacks	Encrypt data at rest Secure communication
C7	Insider threats Inconsistent data	Authentication, Authorization, System-based Access control Cryptography-based access control Integrity verification schemes
C8	Unauthorized access to audit information	Periodic audit Integrity of audit information
C9	Data leakage Insider threat Exploiting vulnerabilities of software/hardware Untrusted nodes	Differential-privacy based techniques Policy-based access control to data storage Encryption schemes, authorization techniques Access control
C10	Data poisoning Man-in-the-middle attacks Compromised data Fake devices	Secure log techniques Cryptographic protocols Tamper-proof software

4 Big Data Security Requirements Modeling

In this section, the security requirements of big data application in Cloud were modeling using i^* model. In our model, we use the same components of the big data reference architecture proposed by the National Institute of Standards and Technology (NIST) which have received the consensus of the scientific community [NIST17]. Thus, the actors who appear in our model are the NIST components. These components are:

System orchestrator: It is the component that identifies and integrates the different requirements into the ecosystem.

Data provider: This component provides the required data to the big data application system by using interfaces between data sources and big data application.

Big data application provider: It provides a set of services along the big data life cycle to maintain the requirements specified by the system orchestrator.

Big data framework provider: This component presents the environment implementation of the big data application by providing platforms, infrastructures, and processing nodes.





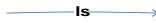
Data consumer: It is the component that interacts within the end-user by providing interfaces.

The proposed modeling language presented by i* provides a notation to model actors as agent or role. Actors achieved goals basing on its own resources furnished and tasks to be performed or on other actor goals, resources, and tasks. However, i* notation does not provide explicit modeling for concepts such as threat, malicious actor, and characteristics that can have a resource and impacts goal achieving or task performing. Therefore, we propose some extensions to the i* concept notations to provide conceptual modeling and identifying big data security requirements in a Cloud environment.

4.1 Concepts added to i*

In this subsection, we present the needs of i* extension to model the big data security requirements. The big data security requirement elicitation requires not only to identify the security requirements but also to consider big data characteristics. Therefore, big data security requirement engineering needs to respond to the ‘how to’ while taking big data characteristics into account. In addition, i* models do not provide a case where an actor is malicious and want to achieve a malicious goal. To address big data characteristics and possible threats while modeling the security requirement of Cloud-assisted big data systems projects, we choose to add malicious actor, malicious goal, threat, and characteristic concepts. Table 3 presents the added concepts and their notations.

Table 3: Extension Concepts and Notations

Concept	Notation	Description
Malicious actor		This element is used to describe an actor that can make malicious threats and tasks
Malicious goal		This element is used to describe a security threat goal that intends to be achieved by a malicious actor and that interferes with the achievement of the normal actor goals.
Threat		This element indicates that the task to be performed can lead to achieve a threat goal. It can be an intentional our unintentional.
Characteristic		This element is used to presents characteristic that can describe a resource
Characterization link		This element presents a contribution link to describe a resource

4.2 Big data security requirement modeling using our i* extension

Figure 3 presents the SD model. In this model, the main relationship between actors is defined. In addition to the main requirement of our proposed model (security), there is a goal named “Execution”. This goal is integrated in order to respect the general modeling of big data applications as it has other requirements that can be integrated within security requirements. In this figure, actors need the goal “execution” of big data application while depending on security goal that must be achieved by system orchestrator and does not specify what is the main characteristics of the data to be secured such as its Volume and its format. Therefore, the goal will be failed if the security approach choosing by the system orchestrator does not support data characteristics. The “Execution” goal is described in SR model where concepts added to i* are inserted and the security requirements C4, C5, and C7 are presented as shown in Fig 4. In this figure, we have presented only three challenges, but it can be extended to present the whole security requirements.

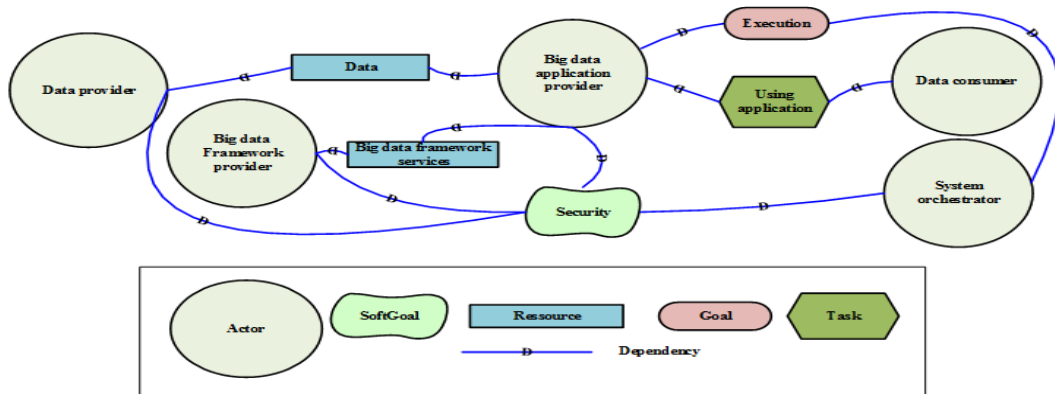


Figure 3 Strategic Dependency Model for Big Data Application Security Requirements

The “Using application” task presented between Data consumer and big data application provider presents a task that can be changed according to the purpose use of the big data application. In this figure, actors depend on the system orchestrator for the goal of their resource security. Big data framework provider depends on the system orchestrator to achieve the goal of “provide a secure cluster” to the Big data framework service. The big data framework provider depends on the system orchestrator to “run the function” using data provided by data provider in a secure manner. In the same time, a malicious insider user attempt to achieve a malicious goal “threat exploit” basing on intentional/unintentional threat to break user’s data. Basing on data characteristics and possible insider threat, system orchestrator needs to provide the security technique that leads big data application provider to run processing functions on data while ensuring their security from malicious insider users.

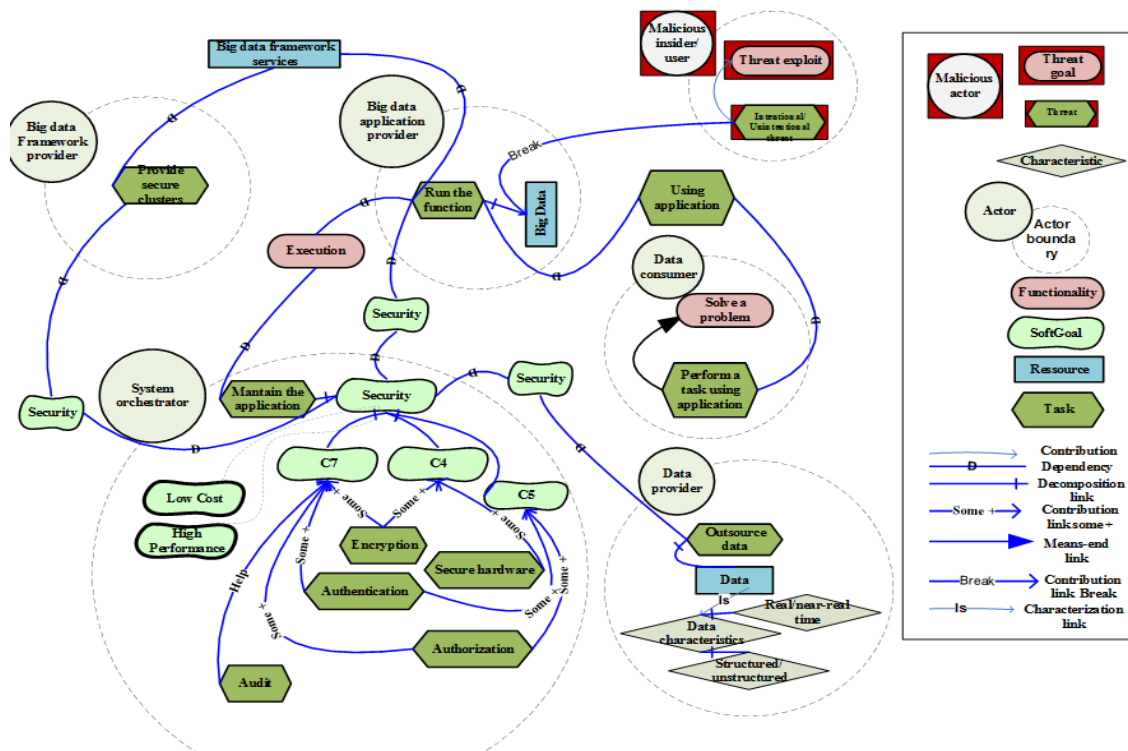


Figure 4 Strategic Rationale Model for Big Data Application Security Requirements

Basing on the modeling sequence of extending i* language presented in [Mussabacher&all14], we extend the recent publishing meta-model presented in [Dalpiaz&all16] to show our extension to the i* model. Figure 5 presents our extensions in red. In this meta-model, we add malicious actor, characteristic, and threat concepts. Malicious actor presents an actor that wants to achieve a malicious goal “threat” in the system. We add characteristic to describe specific concepts that characterize resource to be used to achieve a goal. For example, big data collected from social network are used in the mining process to extract useful information, but they include sensitive data to be secured from unauthorized access. These data are unstructured, generated with high velocity, and with high volume. Unstructured, high velocity and high volume are characteristics of big data to be used and to have secured. Quality differs from the characteristic concept as it presents the desired level of goal achievement such as performance, limited time, etc.

5 Evaluation

To evaluate the model, a use case study in a particular domain is needed. The first step of evaluation is to identify the main security requirements for the big data application according to the security level of the data and processing. Then, the generic Strategic Dependency model and Strategic Rationale model can be generated and validated by the big data application provider. In this section, our i* extension has been evaluated through an illustrative example.

5.1 Electronic Healthcare Record (EHR)

Large amounts of healthcare data are generated continuously, stored, and analyzed by healthcare providers to improve efficiency and to reduce healthcare cost [Belle&all15]. In our example, the big data EHR system consists of the following actors:

- **Patient:** is a data provider who needs healthcare.
- **Medical center :** is a big data framework provider that collects, stores, and analyzes EHR for data consumers.
- **Care provider or doctor:** is the data consumer that receives information about the patient in order to provide him the appropriate treatment.
- **Healthcare Cloud computing provider:** is a big data application provider that presents storage and processing services for a medical center.
- **System orchestrator:** It is the component that identifies and integrates the different requirements into the ecosystem provided to the medical center.

Figure 6 presents a hybrid SD/SR model that illustrates the security requirements for outsourcing big data EHR to a Cloud environment basing on the NIST reference architecture. In this model, we present only one threat, but it can be extended to present more.

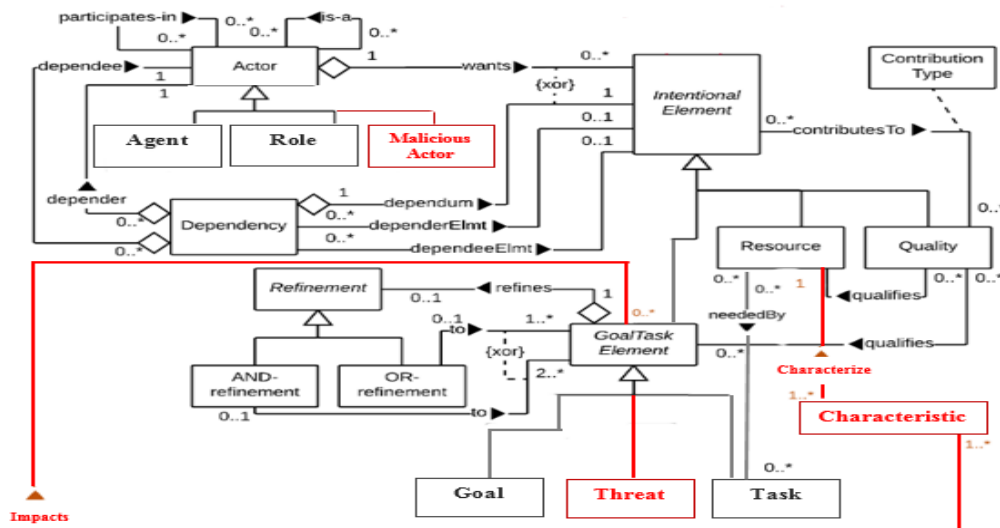


Figure 5 Our Extended i* Meta-Model

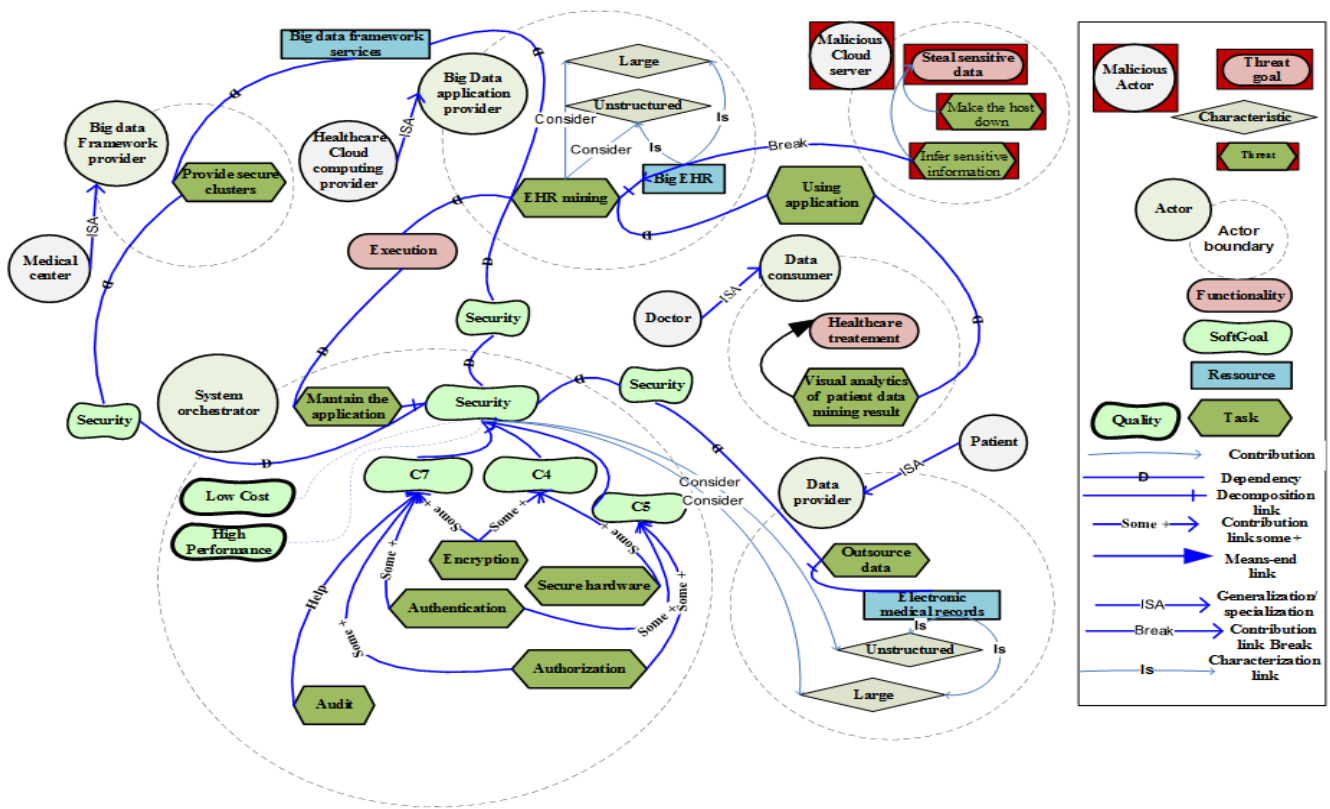


Figure 6 Strategic Dependency/Rationale Model for EHR Big Data Security Requirements Example

5.2 Discussion

Figure 6 presents the following elements :

The model presents a medical center that collects big EHR from patients, analyze them to help doctor for deciding to patients an appropriate treatment. By our extension, we understand that EHR are unstructured and with large size. Therefore, system orchestrator must achieve “security” goal by choosing security techniques that support these characteristics with high performance and low cost. In addition, it must consider that a malicious insider in the Cloud-based big data framework wants to “infer sensitive information” from analyzed data. As a result, the orchestrator can decide to select more appropriate encryption scheme that ensure the processing of data in encrypting formats (such as homomorphic encryption) or the use of secure hardware to store or process data (such as trusted platform module TPM), basing on overall response time that the medical center usually used to response doctors and cost that can support. The cost and performance are not considered in this paper because it relies on the main functional requirement modeling process of service outsourcing. Also, access control methods can be used to ensure that only trusted, authenticated, and authorized Cloud server basing on a trust evaluation of the cluster Cloud servers are used by big data framework provider.

6 Conclusion

Today, big data-based projects and Cloud-based platforms and tools are investing in enterprises businesses process hoping to achieve competitive advantages. Various issues and challenges are created by big data characteristics. However, big data security and privacy present the main concern. In this paper, we have presented the big data life cycle, main security threats of Cloud services, the main issues and security challenges that big data characteristics create. For each challenge and possible threats, we have identified the enhanced techniques that can be adapted to secure big data in the Cloud environment. A model to big data security requirement is proposed. To model these security requirements, we

propose a conceptual extension to i^* notation. This extension considers big data characteristics and presence of malicious goal by malicious actors in Cloud-based big data storage and processing to ensure proper elicitation of security requirements. The big data security requirements are modeled using our extended i^* . The Strategic Dependency model is created to present the identified security requirements based on NIST reference big data architecture components. The Strategic Dependency model has been developed to a Strategic Rationale model where notation added by our extension is used. A meta-model for our i^* extension is presented. To illustrate our model, we present an EHR big data example.

As a complement to this work, the rest steps of i^* extension modeling sequence will be addressed, and rest phases of security requirement engineering process will be completed.

References

- [Alliance13] Alliance, C. The notorious nine cloud computing top threats. Cloud Security Alliance, Tech. Rep (2013).
- [Alshboul&all15] Y.Allshboul & Wang, R.K.YongNepali. Big Data LifeCycle: Threats and Security Model. In *Twenty-first Americas Conference on Information Systems*, 1-7, (2015).
- [Arruda&Madhavji18] D.Arruda, N.H.Madhavji. (2018) State of Requirements Engineering Research in the Context of Big Data Applications. In: Kamsties E., Horkoff J., Dalpiaz F. (eds) Requirements Engineering: Foundation for Software Quality. REFSQ 2018. Lecture Notes in Computer Science, vol 10753. Springer, Cham, (2018), pp 307–323.
- [Bahsoon&all17] R.Bahsoon, N.Ali, M. Heisel, B.Maxim, I.Mistik: Introduction. Software Architecture for Cloud and Big Data: An Open Quest for the Architecturally Significant Requirements, Software Architecture for Big Data and the Cloud, Morgan Kaufmann, (2017), Pages 1-10.
- [Belle&all15] A.Belle, R.Thiagarajan, S.M.Soroushmehr, F.Navidi, A.D.Beard, K.Najarian: Big data analytics in healthcare. *BioMed. Res. Int.* **10**, 1–16 (2015)
- [Chen&all14] M.Chen , S.Mao , Y.Zhang , Victor C. M. Leung, Big Data: Related Technologies, Challenges and Future Prospects, Springer Publishing Company, Incorporated, (2014)
- [CSA13] Cloud Security Alliance (CSA), Expanded Top Ten Big Data Security and Privacy Challenges by CSA Big Data Working Group (2013)
- [Dalpiaz&all16] F.Dalpiaz, X.Franch, J.Horkoff: istar 2.0 language guide. CORR(2016)
- [Eunjang&all18] P.Eunjang, S.Vijayan, P.Sooyoung (2018). A reference Model for Big Data Analytics.
- [Gonçalves&all18] E.Gonçalves ; J.Castro ; J.Araújo; T.Heineck: A Systematic Literature Review of iStar extensions. In: *Journal of Systems and Software Bd.* 137 (2018), pp. 1–33.
- [Giorgini&all05] P. Giorgini, F. Massacci, J. Mylopoulos, N. Zannone, "Modeling Security Requirements through Ownership Permission and Delegation", *Proc. 13th IEEE Int'l Conf. Requirements Eng.*, (2005), pp. 167-176.
- [Jutla&all13] D.Jutla, P.Bodorik, S.Ali: Engineering privacy for big data apps with the unified modeling language. In: 2013 Proceedings of the IEEE International Congress on Big Data, BigData (2013), pp. 38–45
- [Kimball&Ross13] R. Kimball, M.Ross: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd edn, (2013).
- [Madhvaraj&Manjaiah16] M.Madhvaraj., D.H.Manjaiah: Data security in Hadoop distributed file system. In: *International Conference on Emerging Technological Trends (ICETT)*, Kollam, pp. 1–5 (2016)
- [Mahajan15] A.Mahajan, "The Malicious Insiders Threat in the Cloud", *International Journal of Engineering Research and General Science*, vol. 3, no. 2, March-April (2015).
- [Mark&all12] A.Mark, Beyer, L.Douglas: The Importance of Big Data: A Definition, Gartner, Analytics Report G00235055, (2012).
- [Moreno&all18] J.Moreno, M.A.Serrano, E.F.Medina, E.B.Fernandez. Towards a Security Reference Architecture for Big Data, DOLAP, (2018).
- [Mussbacher&all14] Mussbacher, G., Amyot, D., Breu, R., Bruel, J., Cheng, B., Collet, P., Combemale, B., France, R., Heldal, R., Hill, J., Kienzle, J., Schöttle, M., Steimann, F., Stikkolorum, D., Whittle, J. The relevance of model-driven engineering thirty years from now. *Model-Driven Engineering Languages and Systems*. Springer International Publishing, (2014). 183-200.
- [NIST17] NIST NBD-WG. 2017. NIST Big Data Reference Architecture. (2017). https://bigdatawg.nist.gov/_uploadfiles/M0639_v1_9796711131.docx
- [Yu&Eric95] Yu, Eric: Modeling Strategic Relationship for Process Reengineering. PhD Thesis, Graduate Department of Computer Science, University of Toronto, (1995).