Semantic-driven Secured Data Access in Distributed IoT Systems

Riste Stojanov, Sasho Gramatikov, Ognen Popovski and Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje, R. Macedonia
{riste.stojanov@, sasho.gramatikov@, ognen.popovski@students., dimitar.trajanov@}finki.ukim.mk

***Abstract*—Internet of Things (IoT) is everywhere and expanding. While we all enjoy the benefits and the convenience of the IoT services in our everyday life, very few of us are aware of the security risks we are exposed to when IoT acquired data is not properly handled. The sensitivity of this data requires increased precaution in its protection and ability to validate the access control correctness in design time.**

In the IoT domain, the main focus is real time contextual actuation based on a small amount of up-to-date data. Although the raw IoT data has no deeper meaning, when a semantic abstraction is added, it becomes suitable for reasoning, fusing and actuation. With the semantics in hands, it is easier to integrate the distributed IoT devices.

In this paper we evaluate the Linked Data Authorization (LDA) platform for semantic data access control in the IoT context. The LDA platform provides contextual protection of semantic data using flexible security policies that can be validated in design time. We demonstrate that this platform performs well when protecting decent amount of fresh data with respect to the context.

***Keywords*—IoT, Data Security, Semantic Web, Evaluation**

I. Introduction

Throughout the history, the science and technology in their essence are based on four scientific paradigms, each one of them revolutionary in its time, superior to its ancestor and innovative in the tools and methods used [1]. As a result of the rapid development of the computer systems, the storage devices, and network technologies, which became cheap and easily affordable resource, huge amount of data has been generated, stored and exchanged on daily bases. All these advances gave rise to the fourth paradigm, known as data exploration or e-Science [1] that unifies the first three paradigms of empirical observations, theory, and computation and simulation. The data-driven science made it possible for the computers to generate models and programs, giving them the ability to learn from large data sets. What made this trend even more important in the last decade is the Internet of Things (IoT) which connected diverse devices equipped with sensors and embedded software helping them acquire and exchange data. The services and the convenience of the IoT become widely accepted and its presence has been constantly growing. It is estimated that approximately 30 billion devices will become connected by the year 2020 [1]. The growing popularity of the IoT imposes rapid growth of the amount of generated and exchanged data. The amount of generated traffic by these devices will reach 600 ZB per year by 2020, 275 times more than the estimated traffic exchanged between the data centers and end users/devices [2].

Although the storage and networking technologies support this trend, the main concern regarding the IoT data is its protection. The data generated from the IoT devices represents the individuals and their surrounding environment. These information is highly sensitive and can harm the owners if exposed to wrong entities. Moreover, the fact that 60% of the devices in IoT are owned by ordinary people [2] gives the security context even more critical importance. Another alarming fact is that a great part of the industry generated IoT data, also known as dark data [3], is stored but not deeply examined since it is only used for regulating some process. This data can become easy target for the attackers and impose a security risk when the companies are unaware that the data even exists.

The unawareness and incapability of the data owners adds a risk of its exposure, in addition to the implicit vulnerability caused by its very distributed and dynamic nature. Therefore, it is crucial to focus on the access rights that will protect the IoT generated data, from both malicious attackers, and from the uninformed data owners. The access rights are technically represented as security policies enforced by the authorization systems. The security policy format should overcome the heterogeneity of the devices and the data they generate, regarding precision, measurement unit and different serialization formats. The large number of IoT devices that should be protected requires ability to control the access for multiple devices at once, based on the data nature, context and origin.

User's authorization in distributed ubiquitous environments is one of the challenges where improvements are needed. The authorization is usually declared with policies that are enforced by an access control module implementation. Even though there are standards for policy definition, such as XACML [4], most of the real systems have separate policy formats and enforcement modules where the authentication is their integration point. The separate authorization definition is mainly due to the lack of integration in multi-domain scenarios. Multiple research groups are targeting this problem, but it is not trivial to determinate the most suitable solution in a given context. Semantic Web technologies can address many

This research was partially supported by the Faculty of Computer Science and Engineering at Ss Cyril and Methodius University in Skopje, R. Macedonia.

of the challenges that the IoT access control is facing with today opening many opportunities, such as new approaches of authentication using security policies based on a natural language and interoperability using a common ontology [5].

In this paper we consider the semantically based LDA platform III as a solution for contextual protection of data access and evaluate its performance for use in IoT systems. The rest of the paper is structured as follows. In Section II, we give a short overview of the current solutions on data security. Then, in section III we describe the LDA platform as a solution for access control in heterogeneous distributed systems. Afterwards, in Section IV, we present the environment for measurement of the performance of the LDA platform and evaluate the obtained results. Eventually, we conclude our work in Section V.

II. Related Work

The vast variety of protocols used by the IoT devices requires many work hours to make the different protocol devices communicate flawlessly. Furthermore, the different data format, units and precision make their integration even harder. The raw sensory data does not have any deeper meaning for the humans, but when the abstraction is added to the sensory data, it becomes more suitable for the reasoning process used to produce perceptions. The Semantic Web technologies [6] provide a solid ground for abstraction of real world processes and knowledge, and this is already accepted in the IoT community. The Semantic Web technologies define standards for machine-readable and technology agnostic representation of real world concepts [7]. The RDF standard [7] models the knowledge using graph structure composed of multiple quads of the form: $< Subject, Predicate, Object, Graph >$, where $Subject$ is the concept that is being described, $Predicate$ is an attribute or a feature of the $Subject$, $Object$ is the object assigned as a value to the $Predicate$, and $Graph$ is a logical group of the triples $< Subject, Predicate, Object >$ that enable a logical organization of the semantic triples within a dataset.

The Semantic Sensor Networks (SSN) ontology [8], [9] is one of the most influential semantic achievements in the IoT domain. It is defined using the Ontology Web Language (OWL) [10] and provides abstraction of the IoT devices, their properties and observations. Even though this ontology does not model the different measurement units, it allows integration with other domain ontologies for this purpose [8], [11], [12]. This ontology is mainly used to annotate IoT acquired data streams [11], [13]. On the other hand, the work in [12], [14], [15] represents the devices as sensor services using SemSOS ontology [15]. The impact of the different semantic formats regarding CPU cycles, power consumption, and packet size is analyzed in [16], where the authors conclude that the Entity Notation is the most optimal for semantic data representation in resource-constrained environments. However, even though the semantically represented data introduces some performance drawbacks, it provides data abstraction and easier combination of the raw sensory data, leading towards smarter and better observations.

In order to control who can access the data, the entities must be authenticated. There are multiple solutions for authentication in distributed ubiquitous environments, such as single-sign-on services [17], WebID [18] [19] and OAuth [20]. There are also frameworks that enable integration and combination [21] of these services. The access control of the authenticated users is handled by the authorization process.

The general trend for authorization in IoT relies on securing the communication channel with Transport Layer Security (TLS) or Datagram TLS (DTLS) [22], [23] which do not provide an option for contextual and partial data protection. The necessity for context-aware access control is already considered in [24], [25], where the authors emphasize the importance of emergency security policies, but there is no explicit policy format presented. The work in [26], [27] define a secure view, read, aggregate and join operations for IoT data streams, without decentralized policy management. In [28], data owner embeds the policies in the generated stream and lets the stream processors or brokers decide whom to distribute the information. The work in [29], [30] analyses securing machine-to-machine communication for cloud managed IoT devices with the use of an extended Information Flow Control model [31].

In [11], the data and the device discovery information are represented in semantic format, enabling the same policy to protect the device discovery. A comprehensive policy model will enable easy maintenance of the policies [32]. One example is the architecture defined in [11], which allows policies to be stored and retrieved by each gateway using the SPARQL endpoints. Even though there is considerate work that includes the streaming data in the semantic web [33], [34], there are still challenges that need to be addressed regarding access control over semantic streams.

III. LDA platform overview

In our previous work [32], we defined the Linked Data Authorization (LDA) platform, which defines a flexible and maintainable policy language and rules for its enforcement. This platform is intended to protect a semantically annotated data stored in multiple locations. The semantic data standards are chosen since they provide the Resource Definition Framework (RDF) [7] for representation of heterogeneous data from multiple devices and the Ontology Web Language (OWL) [10] to define a structure to the data.

The LDA platform offers the data owners to define security policies for controlling the access to their data, also referred to as guarded data. The guarded data is represented in RDF. The protection rules are defined as security policies using an extension of the widely accepted semantic web query language SPARQL [35]. The native form of the language is used to define which portion of the data is protected, while its extension is used to define the query operation and the dataset the policy is activated for.

The LDA platform enables enforcement of contextual policies through the Intent Provider component, which injects the contextual information (obtained from the request) in a separate graph referred to as Intent. This component is pluggable and extensible, which enables the use of dynamic contextual evidences.

The Intent Provider module intercepts the request and creates an Intent from the provided evidences. Then, it passes the Intent to the Enforcement module to build a temporal dataset containing the allowed data for the Intent. In this process, the Enforcement module combines the security policies and creates a query for temporal dataset creation. Eventually, it submits the query to the underlying semantic database, which, in our case, is the TDB database. Then, the originally requested query is executed against the temporal dataset.

IV. Evaluation

As expected, every authorization system introduces certain authorization overhead. The main overhead introduced in the LDA platform is the temporal dataset creation. The policy activation and combination steps are carried out completely in memory, while the dataset construction is carried out by the underlaying storage engine, which usually uses multiple I/O disk operations. Therefore, the main focus of the evaluation is to determine the dependency of the execution time with respect to the dataset size, the allowed data size and the type of the queries used to construct the allowed data. For the purpose of the LDA platform performance evaluation, we generated groups of incremental test datasets containing quads following the sensory data ontology [**?**]. The initial dataset DS_0 contains only one Sensor associated to its owner User and no Observations. Each following dataset is generated by adding quads for a new Sensor with multiple Observations. The first group of datasets has increments of 100, 200, ..., 600 observations, the next group 1000, 2000, ..., 6000 and so on, until we reach a total of 24 datasets. With this progression the largest dataset DS_{24} reaches the size of approximately $2.3 * 10^6$ instances i.e. $9.2 * 10^6$ quads.

We also generated four different groups of test SPARQL queries that we run against the test datasets. The first group of test query returns all the observations for a given sensor (simple queries), the next group additionally filters the observations that fulfill certain condition (filter queries), the third group is formed of queries that filter the observations that fulfill certain condition with respect to aggregated value (aggregate queries), and the last group combines the result of multiple sensors using the set union and minus operations (composite queries). All these queries return incremental result sizes, since the observations for the sensors were generated incrementally in the datasets.

In Figure 1 we show the average execution time obtained from evaluating the simple query group against all datasets. From the results we can observe that the execution time does not depend on the dataset size, but it does depend only on the size of the results returned by the queries. It is important to emphasize that we only show the average execution times for those queries that return results. Therefore, the length of each plot line shortens as the number of results increases, resulting in a single point for the query that returns $6 * 10^5$ results, that can be obtained only from the largest dataset DS_{24}.

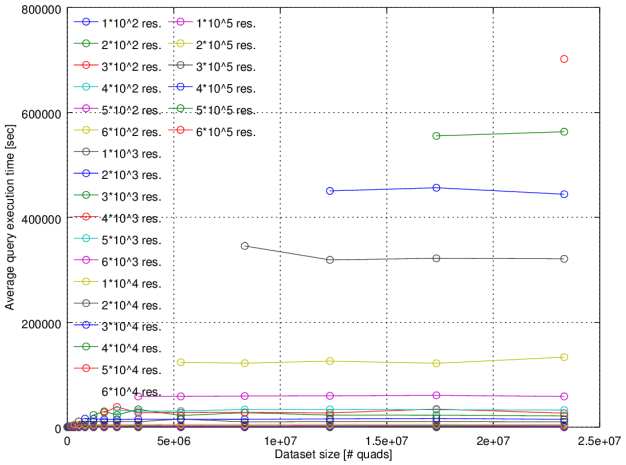

Fig. 1. Execution time dependency on dataset size for different simple queries

Similar results are obtained for all other evaluated query groups, but we do not display due to space limitation. However, we did sublime the dependency of the average execution time of the queries on the result size for all query groups in Figure 2. From the figure we can conclude that all query groups have near linear dependency on the result size. Notably, the aggregate query group is the slowest one since it processes all the observations in order to obtain the aggregation.

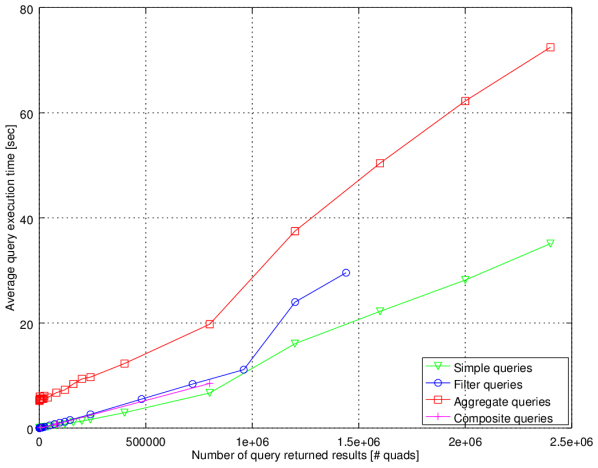

Fig. 2. Execution time dependency on query result size for different groups of queries for dataset DS_{24}

Despite the apparently unacceptable average execution times for more than $0.5 * 10^6$ result quads, the main focus in the IoT domain is real time contextual actuation based on a small amount of fresh data. The sensitivity of this data

requires increased precaution in its protection and ability to validate the correctness of the security policies in design time, which is the strong side of the LDA platform. In the IoT context, we face with small amount of contextual data where proper protection is crucial, hence, under these constrains, the LDA platform proves to provide acceptable performance. Additionally, it is noteworthy to mention that the largest query result set corresponds to a temperature measurement for 4.5 years obtained on a minute interval.

V. Conclusion

In this paper we emphasize that the Semantic data annotation can unify and align the raw sensory data originating from the IoT devices. On one hand, this data is usually personal and highly sensitive, requiring strong protection, while on the other hand, it needs to be shared with an actuation systems in order to be beneficial for its owners and the wider community.

Despite its significance, there is not much of a progress in the field of protecting semantically annotated IoT data. The LDA platform tackles this problem, but it infers some performance issue. In order to determine its applicability in the IoT domain, in this paper, we conducted a thorough evaluation of its performance for different dataset sizes, query types and query result sizes. The results from our analysis show that the LDA platform authorization performance does not depend on the dataset size, but only on the quantity of the allowed data. Considering the nature of the IoT domain, characterized by actuation based on small amounts of recent contextual data, the LDA platform proves to offer flexible, maintainable and testable protection with acceptable performance.

References

[1] T. Hey, S. Tansley, K. M. Tolle, *et al.*, *The fourth paradigm: data-intensive scientific discovery*, vol. 1. Microsoft research Redmond, WA, 2009.

[2] C. V. Networking, "Cisco global cloud index: Forecast and methodology, 2015-2020. white paper," *Cisco Public, San Jose*, 2016.

[3] D. Trajanov, V. Zdraveski, S. Riste, and K. Ljupco, "Dark data in internet of things (iot): Challenges and opportunities," in *7th Small Systems Simulation Symposium*, pp. 1–8, 2018.

[4] S. Godik, A. Anderson, B. Parducci, P. Humenn, and S. Vajjhala, "Oasis extensible access control 2 markup language (xacml) 3," tech. rep., Tech. rep., OASIS, 2002.

[5] R. Stojanov, V. Zdraveski, and D. Trajanov, "Challenges and opportunities in applying semantics to improve access control in the field of internet of things," *Electronics Journal*, vol. 21, no. 2, pp. 66–75, 2017.

[6] T. Berners-Lee, J. Hendler, O. Lassila, *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[7] G. Klyne and J. J. Carroll, "Resource description framework (rdf): Concepts and abstract syntax," W3C recommendation, W3C, feb 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[8] L. Lefort, C. Henson, K. Taylor, P. Barnaghi, M. Compton, O. Corcho, R. Garcia-Castro, J. Graybeal, A. Herzog, K. Janowicz, *et al.*, "Semantic sensor network xg final report," *W3C Incubator Group Report*, vol. 28, 2011.

[9] M. Compton, P. Barnaghi, L. Bermudez, R. GarcíA-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, *et al.*, "The ssn ontology of the w3c semantic sensor network incubator group," *Web semantics: science, services and agents on the World Wide Web*, vol. 17, pp. 25–32, 2012.

[10] D. L. McGuinness, F. Van Harmelen, *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[11] P. Barnaghi, W. Wang, L. Dong, and C. Wang, "A linked-data model for semantic sensor streams," pp. 468–475, 2013.

[12] W. Wang, S. De, G. Cassar, and K. Moessner, "Knowledge representation in the internet of things: semantic modelling and its applications," *automatika*, vol. 54, no. 4, pp. 388–400, 2013.

[13] F. Ganz, P. Barnaghi, F. Carrez, and K. Moessner, "Context-aware management for sensor networks," p. 6, 2011.

[14] Z. Song, A. A. Cárdenas, and R. Masuoka, "Semantic middleware for the internet of things," pp. 1–8, 2010.

[15] C. A. Henson, J. K. Pschorr, A. P. Sheth, and K. Thirunarayan, "Semsos: Semantic sensor observation service," pp. 44–53, 2009.

[16] X. Su, J. Riekki, J. K. Nurminen, J. Nieminen, and M. Koskimies, "Adding semantics to internet of things," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 8, pp. 1844–1860, 2015.

[17] A. Armando, R. Carbone, L. Compagna, J. Cuellar, and L. Tobarra, "Formal analysis of saml 2.0 web browser single sign-on: breaking the saml-based single sign-on for google apps," in *Proceedings of the 6th ACM workshop on Formal methods in security engineering*, pp. 1–10, ACM, 2008.

[18] M. Sporny, T. Inkster, H. Story, B. Harbulot, and R. Bachmann-Gmür, "Webid 1.0: Web identification and discovery," *Editor's draft, W3C*, 2011.

[19] H. Story, B. Harbulot, I. Jacobi, and M. Jones, "Foaf+ ssl: Restful authentication for the social web," in *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web (SPOT2009)*, 2009.

[20] D. Hardt, "The oauth 2.0 authorization framework," 2012.

[21] C. Scarioni, *Pro Spring Security*. Apress, 2013.

[22] A. Niruntasukrat, C. Issariyapat, P. Pongpaibool, K. Meesublak, P. Aiumsupucgul, and A. Panya, "Authorization mechanism for mqtt-based internet of things," in *Communications Workshops (ICC), 2016 IEEE International Conference on*, pp. 290–295, IEEE, 2016.

[23] P. Fremantle and B. Aziz, "Oauthing: privacy-enhancing federation for the internet of things," 2016.

[24] R. Neisse, G. Steri, I. N. Fovino, and G. Baldini, "Seckit: a model-based security toolkit for the internet of things," *Computers & Security*, vol. 54, pp. 60–76, 2015.

[25] R. M. Savola and H. Abie, "Metrics-driven security objective decomposition for an e-health application with adaptive security management," in *Proceedings of the International Workshop on Adaptive Security*, p. 6, ACM, 2013.

[26] S. Papadopoulos, Y. Yang, and D. Papadias, "Cads: Continuous authentication on data streams," in *Proceedings of the 33rd international conference on Very large data bases*, pp. 135–146, VLDB Endowment, 2007.

[27] B. Carminati, E. Ferrari, J. Cao, and K. L. Tan, "A framework to enforce access control over data streams," *ACM Transactions on Information and System Security (TISSEC)*, vol. 13, no. 3, p. 28, 2010.

[28] R. V. Nehme, E. A. Rundensteiner, and E. Bertino, "A security punctuation framework for enforcing access control on streaming data," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pp. 406–415, IEEE, 2008.

[29] J. Bacon, D. Eyers, T. F.-M. Pasquier, J. Singh, I. Papagiannis, and P. Pietzuch, "Information flow control for secure cloud computing," *IEEE Transactions on Network and Service Management*, vol. 11, no. 1, pp. 76–89, 2014.

[30] J. Singh, T. F.-M. Pasquier, J. Bacon, and D. Eyers, "Integrating messaging middleware and information flow control," in *Cloud Engineering (IC2E), 2015 IEEE International Conference on*, pp. 54–59, IEEE, 2015.

[31] D. E. Denning, "A lattice model of secure information flow," *Communications of the ACM*, vol. 19, no. 5, pp. 236–243, 1976.

[32] R. Stojanov, S. Gramatikov, I. Mishkovski, and D. Trajanov, "Linked data authorization platform," *IEEE Access*, vol. 6, pp. 1189–1213, 2018.

[33] M. Koubarakis and K. Kyzirakos, "Modeling and querying metadata in the semantic sensor web: The model strdf and the query language stsparql," in *Extended Semantic Web Conference*, pp. 425–439, Springer, 2010.

[34] D. F. Barbieri, D. Braga, S. Ceri, E. Della Valle, and M. Grossniklaus, "C-sparql: Sparql for continuous querying," in *Proceedings of the 18th international conference on World wide web*, pp. 1061–1062, ACM, 2009.

[35] E. Prud'hommeaux and A. Seaborne, "SPARQL query language for RDF," W3C recommendation, W3C, jan 2008. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.