Dynamic Switching in the Measurements’ Collecting from Heterogeneous Data Sources

Mario Diván\textsuperscript{1,*}, María L. Sánchez Reynoso\textsuperscript{1}, and Mohd Helmy Abd Wahab\textsuperscript{2}

\textsuperscript{1} Economy School, National University of La Pampa, Argentina
\textsuperscript{2} Dept. of Computer Engineering, Universiti Tun Hussein Onn Malaysia, Malaysia
\textsuperscript{*} E-mail: mjdivan@eco.unlpam.edu.ar

Abstract. The Data Stream Processing Strategy (DSPS) focus on online data processing based on IoT environments. Because of the IoT devices are possibly heterogeneous between them, the data gathering is coordinated through a measurement adapter (MA), who is responsible for data transmission to the data processor. In this work, measures’ indirect transmission is incorporated through measurement adapters, promoting cooperative behaviour and the interaction between them. This extends DSPS allowing direct and indirect data collecting. The new states and process view are analysed in terms of the MA. An enveloping strategy is introduced for verifying the data integrity in the transmission path. A simulation was carried out around the generated overhead with this kind of strategy in the MAs. A mean-time of 5.781 ms is obtained as overhead, which makes it feasible of application, considering that in another way the data would be lost, and the MA becomes uncommunicated.

1. Introduction

The real-time decision making tends to advance on the integration with data origins for carrying out the data processing as soon as data arrive in order to be able to react at the moment in which an event happens [1]. The Data-driven decision making uses the available data (be them online or offline) for supporting a given decision based on previous experiences or knowledge, for which its data organization constitutes a critical aspect [2]. The data organization is an essential aspect from the structural viewpoint (e.g. data types, precision, etc.) but also from the semantic viewpoint (i.e. paying special attention to the data meaning and implicates) [3].

A measurement process has an associated aim, which allows defining the entity to be monitored jointly with its associated dimensions (i.e. attributes or characteristics), independently of the underlying method [4]. In this sense, the core aspect is the ability of the process of being repeatable, consistent, trustworthy, and their results comparable and traceable [5].

The Internet-of-Things (IoT) devices are commonly used as data sources in the Wireless Sensor Networks (WSN), and they are naturally heterogeneous because each one is oriented to monitor an entity’s specific aspect. Thus, measurement automatization involves different kinds of devices from collecting to processing. In order to support decision making, they require to be supported by the ability for providing recommendations based on previous experiences and knowledge [6].

The Data Stream Processing Strategy (DSPS) is a topology oriented to online data processing supported by Apache Storm, which organizes the necessary logic to process measurement guided by a measurement framework. The measurement framework allows reaching a common and shareable understanding related to the underlying concepts (e.g. precision, data source, etc), the concept to be
monitored jointly with the objective associated with the project. This is essential for the system reliability in order to warranty the process repeatability and extensibility [7].

Each measurement project in order to be automatized must be defined and loaded in the DSPS using a specific schema, which defines the entity to be monitored, its characteristics jointly with devices responsible for obtaining its measures [8]. The Measurement Adapter (MA) is a DSPS’s component which is responsible for data collecting from indicated IoT devices. Because each device could have different data format, MA unifies the data (i.e. measures) jointly with metadata (i.e. data indicating the belonging and meaning) under a stream organized by Measurement Interchange Schema (MIS) [9]. The data streams are continuously informed from each MA to the Gathering Function (GF) in DSPS, who is responsible by the data processing itself. However, when the MA loses the connectivity with the GF the measures are discarded, and they become in unrecoverable.

As main contributions, i) A central record in DSPS is incorporated for identifying each authorized MA for retransmitting data streams. The authorization is given by the domain’s experts in terms of the measurement project definition. ii) MA incorporates the ability for dynamically switching in the channel of measurements’ providing. In this way, when the MA lose the connectivity to GF, it could send the data stream through an authorized MA avoiding the data losing, iii) The Cincaminis library was updated and extended, incorporating an enveloped mechanism for keeping the message integrity. This is important to differentiate the data from the node coming to those coming from other measurement adapters. Now, each MA in DSPS is able to make available data to GF in two ways. On the one hand, the MA organizes the data stream and it is directly transported to GF. On the other hand, MA is able to use a set of authorized measurement adapters to reach the GF in an Indirect Way. Of course, the indirect way is limited by the data lifespan and the allowed number of jumps.

The work is structured into six sections. Section 2 describes related works. Section 3 describes the central record associated with measurement adapters jointly with states’ analysis of its new behaviour. Section 4 introduces the dynamic switching from the process viewpoint, appreciating both data producer and consumer behaviours. Section 5 describes the main changes on the cincaminis library for implementing the envelop mechanism jointly with simulation results which give proofs of its applicability. Lastly, a synthesis of conclusions and future work are presented.

2. Related Works
Song, Duan, and others [10] propose a model in which they discriminate between storage nodes and producer nodes in a WSN. The storage node has a limited capacity for stored intermediate data from the data producer nodes, while the data producer nodes could use alternatively the storage nodes for transmitting the scalar data to the data processor in an indirect way. They employee IoT devices for representing each data source, and propose a protocol named CNDTSN for implementing the indirect communication. As a similarity, our work defines the authorized nodes (based on a central record) which is able to store data coming from other nodes, which could be viewed as a data producer node. As a difference, our proposal is based on a measurement framework and the data interchanging jointly incorporates data (i.e. measures), complementary data (e.g. pictures), and metadata (i.e. data identifying the data meaning). In addition, each authorized node is able to verify the integrity of the intermediate data for avoiding storing inconsistent data.

Begum and Nandury [11] introduce a fault-tolerance approach which is able to transmit data based on the WSN topology. The approach outlines the WSN network from the graph theory, and it finds for a connected path from the data source to the target node. In this way, two algorithms are proposed: SCR and SCR-DTRA. The combination of previous ones allows looking for alternative paths for transmitting the data in case of failures. As a similarity, our approach looks for alternative nodes for sending data in an indirect way in front of failures. As a difference, our proposal looks the alternative nodes based on a central record which specifies each authorized node by the domain’s experts.

Kar, Roy, and Misra [12] present a scheme oriented to the Connectivity Reestablishment using Adjustable sensor nodes in the presence of Dumb nodes (CoRAD) in WSN. Basically, when some event provokes a miscommunication derived from some failure in a dumb node, the proposal indicates a set
of actions to reach a successful communication up to the dumb node can be reestablished. In this case, some intermediate nodes could be activated or not depending on the necessity in order to establish the connectivity. As a similarity, our proposal tries to reach an alternative path to send data avoiding their discarding. As a difference, all the nodes are active, and they could be simultaneously (depending on whether they are authorized or not) data producers and data consumers (i.e. intermediate nodes used for transporting data coming from a different data source).

Akber, Khan, and others focus on their proposal around the data collecting and transmission as key functions of the WSN. In that sense, they introduce a data collection technique based on the data volume using a mobile collector. The mobile collector uses information related to data volume to schedule the nodes’ path to reach, while the nodes without data will be ignored. As a similarity, an alternative for optimizing the data collecting is introduced in our proposal, discarding any data which do not verify the basic integrity. As a difference, our proposal understands all nodes as data producers, and eventually, some authorized nodes could be consumers for data transporting.

Some proposals tend to optimize the data transmission in WSN based on heterogeneous data sources incorporating an aggregation strategy in the middle [13]–[15]. That is to say, in place of transmitting whole the data, it is possible to reduce the data volume avoiding data repetition and focusing on the values’ changes. However, it implies that the aggregation should be made in IoT devices, which will demand a certain level of hardware requirement.

3. A Nodes’ Unified Record in DSPS: Describing the Measurement Adapter’s States

Each measurement project is defined by the domain’s experts who define both the entity to monitor and the kind of IoT devices to use for collecting [8]. Each expert determines the point in which the measurement adapter jointly with the IoT sensors will be physically installed in order to reach the project’s aim. In this sense, they know the network cards and the rest of the characteristics of each measurement adapter, which is useful from the security and data storing viewpoint. For that reason, DSPS incorporates a nodes’ unified record in which the measurement adapters are outlined in order to control in a central way and share it among all the involved measurement adapters.

Table 1. Main fields of the Nodes’ Unified Record.

| Field   | Description                                                                 |
|---------|-----------------------------------------------------------------------------|
| MA_ID   | ID related to the measurement adapter                                       |
| Footprint | A hash computed using MA_ID, ANC, ADS, GML, and the Role.                      |
| ANC     | Acronym for Authorized Network Cards. It is a Comma-separated list integrated by mac addresses authorized to transmit measures to the data processor. |
| ADS     | Acronym for Authorized Data Sources. It is a list with the identification (ID) of IoT sensors authorized to transmit data through this measurement adapter. |
| Role    | It defines the current behaviour of the measurement adapter.                 |
| GML     | A synthetic Geography Markup Language document indicating the positioning    |

Table 1 synthesizes the main fields related to the Nodes’ Unified Record (NUR), which its aim is identifying each data source or intermediary able to spread out in the field for collecting data. NUR is said unified because this record located at DSPS is shared among all the measurement projects. However, each MA is informed just with the record’s rows associated with its project. The “footprint” field will allow to detect if a record has been modified (e.g. when a new network card has been incorporated without authorization in the MA, the associated footprint will change). In case of some record has been modified, other MA will answer the cooperation request based on its local footprint. The “role” field describes the current behavior associated with the MA. It could assume one of the
follows values i) Data Collector: The MA will inform just the data collected from its authorized data sources; ii) Gateway: The MA has its functionality limited to transmit measures on behalf another measurement adapter; iii) Cooperative: The MA will inform measures coming from other measurement adapters in addition to its sensors; and iv) Blocked: It is not authorized to interact to the data processor. This kind of changes and the new role of measurement adapters in DSPS implies an updating in their states.

**Figure 1.** Updated State Machine Diagram for the Measurement Adapter in DSPS.

The “loading” state is associated with the loading of the measurement project definition (i.e. the entity under monitoring, entity’s attributes, the IoT devices which quantify the entity’s attributes, etc.). Once the project is loaded, the MA is ready to receive measures, or even, reload the measurement project if it has been updated. During the “ready” state, the MA is able to answer control requests (e.g. cooperation requests from other measurement adapters, integrity verification from DSPS, etc.) and collect data from IoT devices. When the local buffer is full, MA will transit to “Transmitting to the Gathering Function” State and transmit the data to the data processor. After that, the buffer is released, and the MA comes back to the “ready” state. In the case of the Gathering Function (GF) does not answer during the transmission intent, automatically the transition will go to the “Looking for Complementary MA” state for trying to reach an alternative. On the one hand, in case of no alternative, the data are discarded, and MA will come back to the “ready” state. On the other hand, when to exist at least one authorized MA able to collaborate, the transmission will follow the established order based on collaborators' answers.

**4. Dynamic Switching of the Measurement Adapter: A Process View**

The measurement adapter in DSPS changes its original behaviour related to a data collector for becoming additionally in a gateway. That is to say, the data collector could send its own measures but also measures coming from other adapters for avoiding data losing. In this sense, only the authorized adapters are able to expose this kind of cooperative behaviour. Figure 2 uses the Business Process Model Notation (BPMN)[16], [17] for describing the process view related to measurement adapters. On the one hand, when a measurement adapter starts the first task is related to load the measurement project definition and associate it with the linked IoT devices. On the other hand, the MA ends its behaviour when it receives the shutdown message from the Gathering Function -GF- (i.e. a component from DSPS). The original behaviour was related to the transmitting of measures.

The new behaviour of the measurement adapter described in Figure 2 incorporates: i) Integrity Verification: The Gathering Function (GF) verifies the integrity of the Nodes’ Unified Record (NUR) of each MA for avoiding identity substitution; ii) Logistic Information: GF is able to update the local copy of the NUR for extending the cooperative possibilities; iii) Cooperation Availability: The measurement adapters could interact between them without intervention of GF, requesting the possibility of pass-through in case of some failure related to the communication with GF. The cooperation is optional for each measurement adapter, and each one will answer depending on the

---

| State Machine Diagram for the Measurement Adapter in DSPS |
|------------------------------------------------------------|
| **Loading** → **Rated** |
| **Shut Down** → **Rated** |
| **Transmitting to the Complementary MA** |
| **Looking for Complementary MA** |
| **Transmitting to the Gathering Function** |
| **Looking for Complementary MA** |

**Legend:**
- MA: Measurement Adapter
- GF: Gathering Function
available resources and authorizations (See “Collecting and Adapting” sub-process in Figure 2); iv) Cooperation Request: When some MA accept the cooperation request, it is able to apply selective filtering on data based on the data lifespan and the maximum number of allowed jumps. It pretends to avoid consuming resources for unnecessary data; v) Envelope Mechanism: The data transmitted through a different MA will use an envelope for verifying the data integrity. If data integrity is not satisfied when the target MA receives the data, then they will be discarded.

![BPMN Process View of the Measurement Adapter Behavior](image)

**Figure 2.** A BPMN Process View of the Measurement Adapter Behavior.

The buffer reserved percentage depends on the current role of the MA (See Table 1). When the role is “cooperative”, then only a region of its local buffer is defined as shared with other adapters. However, when the role is “gateway”, all the local buffer is applied to support to other adapters. Finally, when the MA has the role of “data collector” or “Blocked” there is no reserve of the local buffer. In the first case, the buffer is oriented to support measures coming from directly associated IoT devices, while the second case is a passive adapter which does not have permissions for transmitting measures to the GF.

5. Enveloping the Measurement Interchange Schema

Because each IoT sensor could have its own way to provide data and the devices’ gathering possibly implies that a set of heterogeneous devices are informing at the same time to the measurement adapter, a data homogenization is necessary before communicating all the measures to the data processor. In this sense, the cincami/mis library[9] is an open-source library which allows communicating all the measures jointly with the metadata using the measurement project definition [8].

In order to support the cooperative behavior between measurement adapters and the indirect transmission, a quick envelope was incorporated to the mentioned library. Basically, the new tags associated with the envelope are: i) Origin: It identifies to the measurement adapter responsible for obtaining the measures from the IoT devices; ii) originTimestamp: It represents the timestamp in which the original measurement adapter derives the measures to the first derived MA; iii) Lifespan: the lifespan expressed in seconds and accounted from the originTimestamp; iv) Jumps: The number of jumps made from the originTimestamp; v) knownMA: The measurement adapters whom have been visited in the path; vi) originalMessage: The CINCAMI/MIS message which is being retransmitted through the envelope; vii) Fingerprint: A MD5 fingerprint [18] related to the original message.
The fingerprint tag allows verifying the data integrity related to the native message (see Figure 2). In case the original message does not satisfy the integrity test, the data are filtered. Then, the jumps tag is used for determining whether it is possible to receive the original message or not depending on the MA’s local policy associated with the maximum number of allowed jumps. Once the integrity and jumps have been verified, the local MA will verify the data lifespan using the originTimestamp and Lifespan tags. The underlying idea is really to be cooperative with useful data and not expired. Moreover, the knownMA tag is used along with the “cooperation availability” message for avoiding requesting cooperation two times to the same measurement adapter. That is to say, once a measurement adapter has cooperated, the requester will not request cooperation again for the same data, it is preferable to discard them and pay attention to the most recent data.

A simulation was carried out on the updated library, which ran on a MacBook Pro with macOS Mojave (Version 10.14.16), 16GB 2133MHz LPDDR3 of RAM, a processor 2.9GHz Intel Core i7, Radeon Pro 560 4GB and Intel HD Graphics 630 1536 MB. The updated library is freely available on GitHub (github.com/mjdivan/cincamimis). The simulation had two parts, it consisted in analyzing the consumed time by the envelop strategy varying the number of measures and transmitting over 5 minutes in a continuous way. On the one hand, the number of measures by message was varied from 100 to 5000 for analyzing the consumed times related to the envelop strategy when the data volume is changing. On the other hand, a message with 500 measures was transmitted in a continuous way for 5 minutes, measuring the consumed time related to the envelop strategy. Thus, the idea is to analyze the consumed time from the data volume of the original message but also from the continuous transmission proper of the data stream context. In both perspectives, the measured time is integrated by the consumed time for obtaining the MD5 fingerprint, generating the envelope message to JSON data format, and compressing/decompressing the enveloped message.

![Figure 3. Evolution of the Consumed Time, varying the number of measures per message](image)

From the simulation results, the evolution of the transmission and reception consumed time (See Figure 3) while the measures are varied in each message shows a linear behavior which is logical. However, figure 4 shows the behavior over 5 minutes of continuous transmission and reception between measurement adapters. The arithmetic mean of the transmission total time is 2.75 ms with a standard deviation of 0.487 ms, a median of 2.565 and a trimmed mean (95%) of 2.566 ms. Thus, the effect of outliers on the arithmetic mean is evidenced.
Figure 4. Evolution of the Transmission and Reception Total Time in the Measurement Adapter over five minutes with a message containing 500 measures.

Something similar happens with the reception total time in which the arithmetic mean is 3.031 ms with a standard deviation of 0.328 ms, a median of 2.961 ms and a trimmed mean (95%) of 2.961 ms. Along 5 minutes, 2187 messages were processed and measured, around 7.29 messages per second (500 measures per each message). Thus, results would indicate that the enveloping strategy is feasible, being possible its applicability without a meaningful overhead in the MA.

6. Conclusion
In this work, a Nodes’ Unified Record was introduced as an adjustment method along with the measurement adapters spread out in the topology related to the measurement project. This allows that MA becomes in a data collector, but also in a gateway when it is necessary. In this way, the direct and indirect data transmission is supported and incorporated in DSPS. The states’ changes updated in the MA were introduced, jointly with the process view of the updated behavior in the data collecting. For indirect transmission, an enveloping strategy was presented in order to verify the integrity of the transmitted data and limit its transmission by mean of the number of jumps or lifespan. The CINCAMI/MIS library was updated and released for supporting the enveloping strategy. In addition, a simulation was carried out for analyzing the overhead related to the enveloping. The simulation results indicated that around 2.75 ms was the mean-time associated with the transmission total time, which is integrated by the consumed time for obtaining the MD5 fingerprint, generating the envelope message to JSON data format, and compressing the enveloped message. The reception total time was around 3.031 ms, which indicates that in 5.781 ms is the overhead mean total time related to the cooperation between measurement adapters. This mean-time is feasible time considering the ability for dynamic switching and its usefulness, mainly thinking that in another way the data would be lost, and MA becomes passive because of lack of communication. As future work, the behavior of load shedding will be reviewed in terms of location (currently in the GF) jointly with its applicable policies.

References
[1] J. Louis and P. S. Dunston, “Integrating IoT into operational workflows for real-time and automated decision-making in repetitive construction operations,” Autom. Constr., vol. 94, pp. 317–327, Oct. 2018.
[2] Q. Long, “Data-driven decision making for supply chain networks with agent-based computational experiment,” Knowledge-Based Syst., vol. 141, pp. 55–66, Feb. 2018.
[3] S. Wang, J. Wan, D. Li, and C. Liu, “Knowledge Reasoning with Semantic Data for Real-Time Data Processing in Smart Factory,” Sensors, vol. 18, no. 2, p. 471, Feb. 2018.
[4] A. Maul, L. Mari, D. Torres Irribarra, and M. Wilson, “The quality of measurement results in terms of the structural features of the measurement process,” Measurement, vol. 116, pp. 611–620, Feb. 2018.
[5] M. Noman, T. Tahir, and G. Rasool, “A systematic mapping study on success factors of implementing measurement processes in SMEs,” in 2018 International Conference on Advancements in Computational Sciences (ICACS), 2018, pp. 1–8.
[6] M. J. Diván and M. L. Sánchez Reynoso, “An Architecture for the Real-Time Data Stream Monitoring in IoT,” in Multimedia Big Data Computing for IoT Applications, S. Tanwar, S. Tyagi, and N. Kumar, Eds. Springer Nature Singapore, 2020, pp. 59–100.
[7] M. J. Diván and M. L. Sánchez Reynoso, “Real-Time Measurement and Evaluation as System Reliability Driver,” in System Reliability Management, A. Anand and M. Ram, Eds. Boca Raton: Taylor & Francis, a CRC title, part of the Taylor & CRC Press, 2018, pp. 161–188.
[8] M. Diván, M.; Sánchez Reynoso, “Fostering the Interoperability of the Measurement and Evaluation Project Definitions in PAbMM,” in 2018 7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2018, 2018, pp. 228–234.
[9] M. Diván and M. de los Ángeles Martín, “Towards a Consistent Measurement Stream Processing from Heterogeneous Data Sources,” Int. J. Electr. Comput. Eng., vol. 7, no. 6, pp. 3164–3175, Dec. 2017.
[10] Y. Song, P. Duan, L. Liu, and H. Ma, “CNDTSN,” in Proceedings of the ACM Turing Celebration Conference - China on - ACM TURC '19, 2019, pp. 1–5.
[11] B. A. Begum and S. V Nandury, “Component-based Self-Healing Algorithm with Dynamic Range Allocation for Fault-Tolerance in WSN,” in Proceedings of the 7th International Conference on Computer and Communication Technology - ICCCT’2017, 2017, pp. 58–65.
[12] P. Kar, A. Roy, and S. Misra, “Connectivity Reestablishment in Self-Organizing Sensor Networks with Dumb Nodes,” ACM Trans. Auton. Adapt. Syst., vol. 10, no. 4, pp. 1–30, Jan. 2016.
[13] A. Das and S. Chakraborty, “Context Aware Data Collection Framework for Critical Infrastructure Monitoring System,” in Proceedings of the 10th on Wireless of the Students, by the Students, and for the Students Workshop - S3 ’18, 2018, pp. 11–13.
[14] B. Olivieri and M. Endler, “DADCA,” in Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems - MSWiM’17, 2017, pp. 129–136.
[15] E. G. Prathima, H. Laxmikant, S. A. Naveen, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, “DAMS,” in Proceedings of the 5th International Conference on Communications and Broadband Networking - ICCBN’17, 2017, pp. 6–11.
[16] M. von Rosing, S. White, F. Cummins, and H. de Man, “Business Process Model and Notation—BPMN,” in The Complete Business Process Handbook, Elsevier, 2015, pp. 433–457.
[17] C. Arevalo, M. J. Escalona, I. Ramos, and M. Domínguez-Muñoz, “A metamodel to integrate business processes time perspective in BPMN 2.0,” Inf. Softw. Technol., vol. 77, pp. 17–33, Sep. 2016.
[18] Z. E. Rasjid, B. Soewito, G. Witjaksono, and E. Abdurachman, “A review of collisions in cryptographic hash function used in digital forensic tools,” Procedia Comput. Sci., vol. 116, pp. 381–392, 2017.