Journal of Software Engineering and Applications, **2011, 4, 653-665**
doi:10.4236/jsea.2011.412077 Published Online December 2011 (http://www.SciRP.org/journal/jsea)

Strategy for Data Stream Processing Based on Measurement Metadata: An Outpatient Monitoring Scenario

Mario Diván[1,2], Luis Olsina[2], Silvia Gordillo[3]

[1]Law and Economy School, Universidad Nacional de La Pampa, Santa Rosa, Argentina; [2]Engineering School, National University of La Pampa, General Pico, Argentina; [3]LIFIA, Informatics School, National University of La Plata, La Plata, Argentina.
Email: mjdivan@eco.unlpam.edu.ar, olsinal@ing.unlpam.edu.ar, gordillo@lifia.info.unlp.edu.ar

Received October 25th, 2011; revised December 1st, 2011; accepted December 12th, 2011.

ABSTRACT

In this work we discuss SDSPbMM, an integrated Strategy for Data Stream Processing based on Measurement Metadata, applied to an outpatient monitoring scenario. The measures associated to the attributes of the patient (*entity*) *under monitoring, come from heterogeneous data sources as data streams, together with metadata associated with the formal definition of a measurement and evaluation project. Such metadata supports the patient analysis and monitoring in a more consistent way, facilitating for instance*: 1) *The early detection of problems typical of data such as missing values, outliers, among others*; *and* 2) *The risk anticipation by means of on-line classification models adapted to the patient. We also performed a simulation using a prototype developed for outpatient monitoring, in order to analyze empirically processing times and variable scalability, which shed light on the feasibility of applying the prototype to real situations. In addition, we analyze statistically the results of the simulation, in order to detect the components which incorporate more variability to the system.*

Keywords*: *Measurement, Data Stream Processing, C-INCAMI, Statistical Analysis

1. Introduction

Nowadays, there are applications which make customized processing of data sets, generated in a continuous way, in response to queries and/or to adjust their behavior depending on the arrival of new data [1]. Examples of these applications are namely for vital signs monitoring of patients; for behavioral tracking of financial markets; for air traffic monitoring, among others. In such applications, the arrival of a new data represents the arrival of a value (e.g. for a cardiac frequency, a foreign currency rate, etc.) associated to a syntactical behavior. Frequently, they only analyze the number (value) itself without formal and semantic support, disregarding not only the measurement metadata, but also the context in which the phenomenon happens. Therefore, in order to understand the meaning of arriving data and then act accordingly, such applications must necessarily incorporate a logic layer, *i.e*. procedures and metadata, which transform and/or interpret the data streams. Since a lack of clear separation of concerns between the syntactic and semantic aspects of those current applications, very often an expert (e.g., for the outpatient monitoring system, the expert can be a doctor responsible for the monitoring) should intervene in order to interpret the situation. So we argue that given the state-of-the-art of IT in metadata and semantic processing the intervenetion of experts should be minimized as long as the applications can perform the job.

Taking into account the semantic and formal basis for measurement and evaluation (M&E), the C-INCAMI (*Context-Information Need*, *Concept model*, *Attribute*, *Metric* and *Indicator*) conceptual framework establishes an ontology that includes the concepts and relationships necessary to specify data and metadata in any M&E project [2,3]. On the other hand, we have envisioned the need of integrating heterogeneous data streams with metadata based on the C-INCAMI framework, in order to allow a more consistent and richer analysis of data sets (measures). As result, the Strategy for Data Stream Processing based on Measurement Metadata (SDSPbMM) [4,5] was developed.

The main SDSPbMM aim is filling the gap among the integration of heterogeneous data sources; the incorporation and processing of metadata for attributes, contextual

properties, metrics (for measurement) and indicators (for evaluation); and on-line classifiers that support in a more robust way decision-making processes.

Thus, by using the SDSPbMM approach for the abovementioned applications, in this paper we present particularly the foundations for developing the outpatient monitoring scenario. We also performed a simulation using a prototype developed for this scenario, in order to analyze empirically processing times and variable scalability, which shed light on the feasibility of applying the prototype to real situations. For this end, statistical techniques such as descriptive analysis, correlation analysis and principal component analysis are used.

The contributions of this work is manifold: 1) *related to metrics*: the detection of deviations of metrics' values with respect to their formal definitions, identification of missing values and outliers; 2) *related to the set of measures*: the instant detection of correlations; the identification of variability factors of the system; and the detection of trends on data streams, considering also the contextual situation; and 3) *related to the empirical study*: the validation of the implemented prototype on a specific domain scenario, *i.e.* the outpatient monitoring, which allow us determining the feasibility to be applied in real situations.

The quoted contributions represent a step further with regard to our previous works [4,5], because now the enhanced prototype incorporate the online classifiers which support proactive decision making, and their interaction with statistical techniques.

Following this introduction, Section 2 points out the main motivation and provides an overview of the C-INCAMI framework and the SDSPbMM approach. Section 3 illustrates the outpatient monitoring scenario, and Section 4 discusses the planning of the study and the analysis of results related to the simulation. Section 5 analyzes the contributions of this research regarding related work and, finally, Section 6 draws the main conclusions and outlines future work.

2. Fundamentals of SDSPbMM

2.1. Motivation

The SDSPbMM [4] approach proposes a flexible framework in which co-operative processes and components are specialized for data stream management with the ultimate aim of having proactive decision making. In this sense, SDSPbMM allows the automation of data collection and adaptation processes supporting also the incurporation of heterogeneous data sources; the correction and analysis processes supporting the early detection of problems typical of data such as missing values, outliers, etc.; and online decision-making processes based on formal definitions of M&E projects, and current/updated classifiers (see **Figure 2**, which depicts a view of the SD SPbMM approach).

For instance, to deal with detection, correction and analysis processes, our proposal uses in online form, statistical techniques such as descriptive analysis, correlation analysis and principal component analysis. In addition to these techniques other statistical techniques are used to initially validate the work. In a nutshell, we performed a simulation using a prototype developed for outpatient monitoring scenario, in order to analyze empirically processing times and variable scalability, which shed light on the feasibility of applying the prototype in real situations.

Before going through the simulation and statistical analysis issues, it is necessary to illustrate the main aspects of the C-INCAMI framework, which is a key part to the SDSPbMM approach.

2.2. C-INCAMI Overview

C-INCAMI is a conceptual framework [2,3], which defines the concepts and their related components for the M&E area in software organizations. It provides a domain (ontological) model defining all the terms, properties, and relationships needed to design and implement M&E processes. It is an approach in which the requirements specification, M&E design, and analysis of results are designed to satisfy a specific information need in a given context. In C-INCAMI, concepts and relationships are meant to be used along all the M&E activities. This way, a common understanding of data and metadata is shared among the organization's projects lending to more consistent analysis and results across projects.

The SDSPbMM approach reuses totally the C-INCAMI conceptual base, in order to obtain a repeatable and consistent data stream processing where raw data usually is coming from data sources as sensors. While C-INCAMI was initially developed for software applications, the involved concepts such as metric, measurement method, scale, scale type, indicator, elementary function, decision criterion, etc., are semantically the same when applied to other domains, as it is the case when applied to the outpatient monitoring system for the healthcare domain.

The C-INCAMI framework is structured in six components, namely: 1) *M&E project definition*, 2) *Nonfunctional requirements specification*, 3) *Context specification*, 4) *Measurement design and implementation*, 5) *Evaluation design and implementation*, and 6) *Analysis and recommendation specification.*

The components are supported by ontological terms defined in [3].

The *M&E project definition* component (not shown in **Figure 1**), defines and relates a set of project concepts needed to deal with M&E activities, roles and artifacts. It allows defining the terms for a requirements project, and

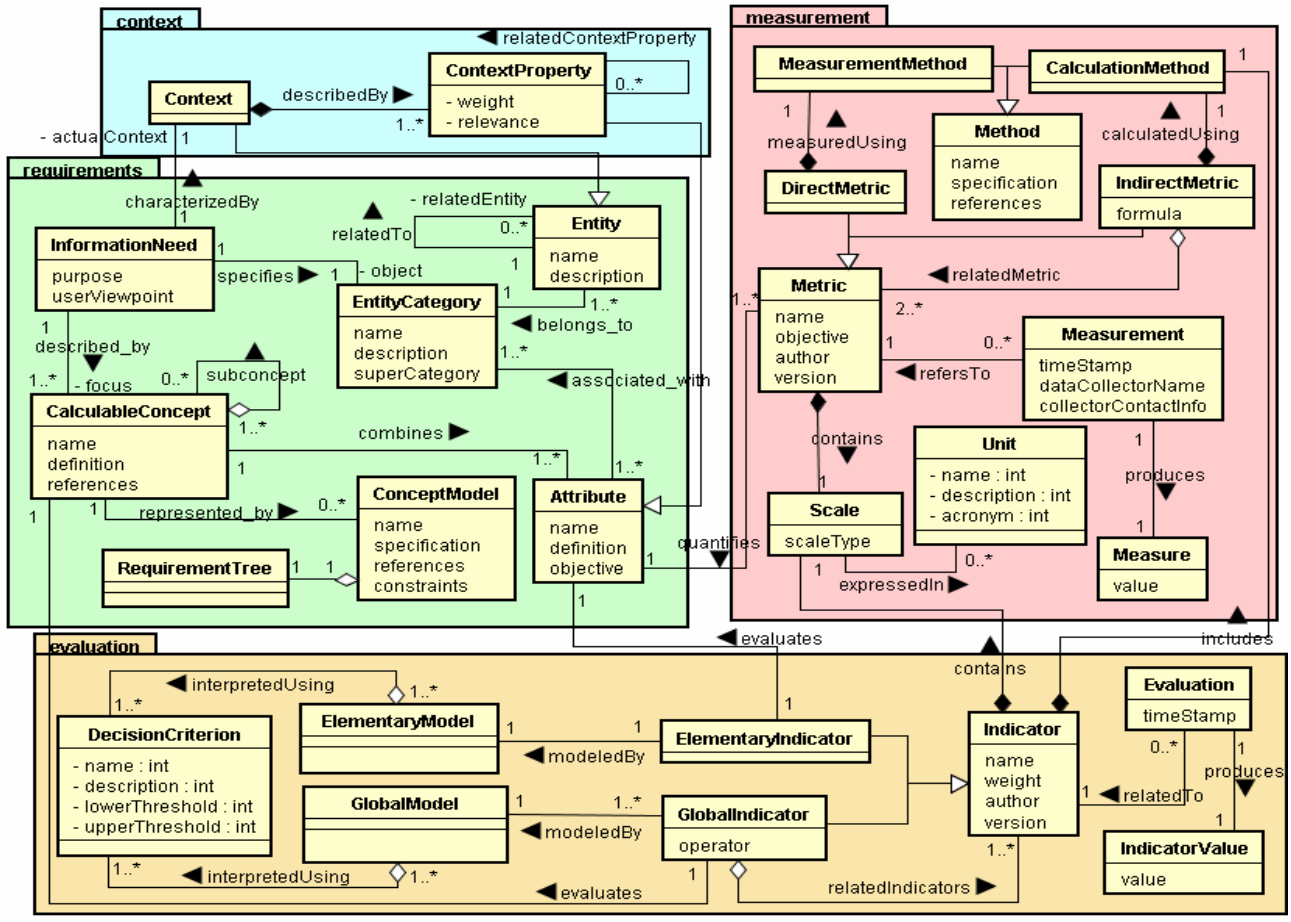


Figure 1. C-INCAMI main concepts and relationships for nonfunctional requirements specification, context specification, measurement design and implementation, and evaluation design and implementation components.

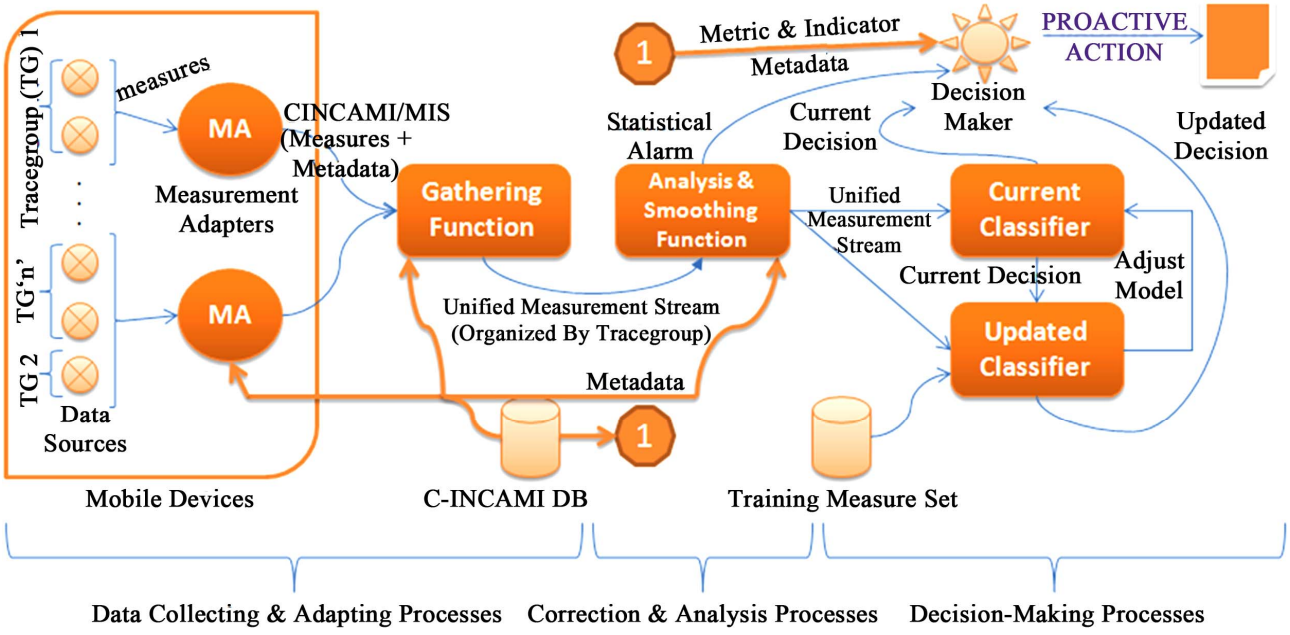


Figure 2. Conceptual schema for the strategy for data stream processing based on measurement metadata.

its associated measurement and evaluation sub-projects.

The *Nonfunctional requirements specification* component (requirements package in **Figure 1**) allows specifying the *Information Need* of any M&E project. The information need identifies the *purpose* (e.g. "understand", "predict", "monitor", etc.) and the *user viewpoint* (e.g. "patient", "final user", etc); in turn, it focuses on a *Calculable Concept*—e.g. software quality, quality of vital signs, etc. and specifies the *Entity Category* to evaluate —e.g. a resource, a product, etc. A calculable concept can be defined as an abstract relationship between attributes of an entity and a given information need. This can be represented by a *Concept Model* where the leaves of an instantiated model are *Attributes*. Attributes can be measured by metrics.

For the *Context Specification* component (*context* package in **Figure 1**), one concept is *Context*, which represents the relevant state of the situation of the entity to be assessed with regard to the information need. We consider Context as a special kind of *Entity* in which related relevant entities are involved. To describe the context, attributes of the relevant entities are used—which are also Attributes called *Context Properties* (see [2] for details).

The *Measurement Design and Implementation* component (*measurement* package in **Figure 1**), includes the concepts and relationships intended to specify the measurement design and implementation. Regarding measurement design, a *Metric* provides a *Measurement* specification of how to quantify a particular attribute of an entity, using a particular *Method*, and how to represent its values, using a particular *Scale*. The properties of the measured values in the scale with regard to the allowed mathematical and statistical operations and analysis are given by the *scale Type*.

Two types of metrics are distinguished. *Direct Metric* is that for which values are obtained directly from measuring the corresponding entity's attribute, by using a *Measurement Method*. On the other hand, the *Indirect Metric* value is calculated from other direct metrics' values following a *function* specification and a particular *Calculation Method*. For measurement implementation, a *Measurement* specifies the activity by using a particular metric description in order to produce a *Measure* value. Other associated metadata is the *data collector name* and the *timestamp* in which the measurement was performed.

The *Evaluation Design and Implementation* component (evaluation package in **Figure 1**) includes the concepts and relationships intended to specify the evaluation design and implementation. It is worthy to mention that the selected metrics are useful for a measurement process as long as the selected indicators are useful for an evaluation process in order to interpret the stated information need. *Indicator* is the main term, and there are two types of indicators. First, *Elementary Indicator* that evaluates attributes combined in a concept model. Each elementary indicator has an *Elementary Model* that provides a mapping function from the metric's measures (the domain) to the indicator's scale (the range). The new scale is interpreted using agreed decision criteria, which help analyze the level of satisfaction reached by each elementary nonfunctional requirement, *i.e.* by each attribute. Second, *Partial/Global Indicator*, which evaluates mid-level and higher-level requirements, *i.e*. sub-characteristics and characteristics in a concept model. Different aggregation models (*Global Model*) can be used to perform evaluations. The global indicator's value ultimately represents the global degree of satisfaction in meeting the stated information need for a given purpose and user viewpoint. As for the implementation, an *Evaluation* represents the activity involving a single calculation, following a particular indicator specification—either elementary or global-, producing an *Indicator Value*.

The *Analysis and Recommendation Specification* component (not shown in **Figure 1**), includes concepts and relationships dealing with analysis design and implementation as well as conclusion and recommendation. Analysis and recommendation use information coming from each M&E project, which includes requirements, context, measurement and evaluation data and metadata. By processing all this information and by using different kinds of statistical techniques and visualization tools, stakeholders can analyze the assessed entities' strengths and weaknesses with regard to established information needs, and justify recommendations and ultimately decision making in a consistent way.

Considering the SDSPbMM strategy and its developed prototype, streams coming from data sources (*i.e*. usually sensors) are structured incorporating to measures the metadata based on C-INCAMI such as the entity being measured, the attribute and its corresponding metric, the trace group, among others. For a given data stream, not only measures associated to metrics of attributes are tagged but also measures associated accordingly to contextual properties as well.

Thanks to each M&E project specification is based on C-INCAMI, the processing of tagged data streams are then in alignment with the project objective and information need, allowing thus traceability and consistency by supporting a clear separation of concerns. For instance, for a given project—more than one can be running at the same time—it is easy to identify whether a measure is coming from an attribute or from a contextual property, and also its associated scale type and unit. Therefore, the statistical analysis is benefited because the verification of each measure for consistency against its formal (metric) definition can be performed.

2.3. SDSPbMM Overview

Data collecting and adapting processes deal with how to adapt different measurement devices to collect measures and then communicate them to correction and analysis processes. The main components (see **Figure 2**) are data sources, measurement adapters and the gathering function.

The underlying idea of the SDSPbMM approach [4] is depicted in **Figure 2**.

Briefly, the measurement stream is informed by each heterogeneous data source to the *measurement adapter* (MA). The MA incorporates the metadata (e.g. metric ID, context property ID, etc.) associated to each data source into the stream, in order to transmit measurements to the *gathering function* (GF). Such measurements are organized in GF by their metadata and then sent to the *Analysis & Smoothing Function* (ASF). ASF performs a set of statistical analysis on the stream in order to detect deviations or problems with data, considering its formal definition (as per C-INCAMI DB). In turn, the incremental classifiers (*i.e.* the *current and updated classifiers*) analyze the arriveing measurements and act accordingly triggering alarms in case a risk situation arises.

SDSPbMM is made up of the following processes: 1) *Data Collecting and Adapting Processes*; 2) *Correction and Analysis Processes*; and 3) *Decision-Making Processes*, which are summarized below.

2.3.1. Data Collecting and Adapting Processes

The data collecting and adapting processes deal with how to adapt different measurement devices to collect measures and then communicate them to correction and analysis processes. The main components (see **Figure 2**) are data sources, measurement adapters and the gathering function.

In short, measures are generated in the heterogeneous data sources, and sent continuously to the MA. MA can usually be embedded in mobile devices, but also can be embedded in any computing device associated to data sources. It incorporates the measured values join to the M&E project metadata respectively, sending in turn both to the GF.

GF introduces streams into a buffer (see **Figure 3**) organized by trace groups—a flexible way to group data sources established dynamically by the M&E project director. This organization allows consistent statistical analysis at trace group level, without representing an additional processing load. Within each trace group, as shown in **Figure 3**, the organization of measurements is tracked by metric. This fosters a consistent analysis among different attributes (e.g. axillary temperature, cardiac frequency, etc.), which are monitored by a given trace group for a particular patient. Also, homogeneous comparisons of attributes can be made for different trace groups (or patients).

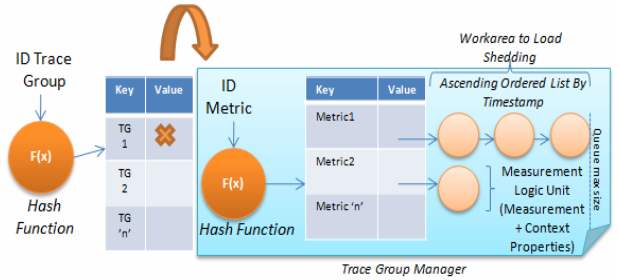


Figure 2. A view of the multilevel buffer.

Moreover, GF incorporates *load shedding techniques* [6], which allow managing the queue of services associated to measurements, thus mitigating overflow risks regardless of how they are grouped.

2.3.2. Correction and Analysis Processes

The correction process is based on statistical techniques where data and their associated metadata allow richer (semantic) analysis. The semantic lies in the formal definition of each M&E project regarding the C-INCAMI conceptual framework (introduced in sub Section 2.2).

It is important to remark that the formal definition of each project is made by experts. In this way, such a definition becomes a reference pattern in order to determine if a particular measure (value) is coherent and consistent with regard to its associated metric specification.

Once the measures are organized in the buffer, the SDSPbMM prototype applies *descriptive*, *correlation* and *principal components analysis*. These techniques allow detecting inconsistent situations, trends, correlations, and/or identifying system components that incorporate more variability. If some situation is detected in ASF (see **Figure 2**), a statistical alarm is triggered to the *decision maker* (*DM*) component in order to evaluate whether it is necessary to send an external alarm (via e-mail, SMS, etc.) for reporting on this situation to medical staff or not.

2.3.3. Decision-Making Processes

Once the statistical analysis was performed, the unified streams are communicated to the *current classifier* (CC) component, which classifies measurements to decide whether they correspond to a risk situation or not and report accordingly such decision to DM. Simultaneously, CC is regenerated by incorporating the unified streams to the training measure set, and then producing a new model named *Updated Classifier* (UC) in **Figure 2**.

Later, the UC classifies the unified streams and produces an updated decision notifying to DM. Ultimately, DM evaluates if both decisions (from CC and UC) correspond to a risk situation and its probability of occurrence.

Finally, regardless the selected decision made by DM, the UC becomes the CC replacing the previous one (see the adjust model arrow in **Figure 2**), only if an impro-

vement in the classification capacity according to the adjustment model based on ROC (*Receiver Operating Characteristic*) [7] curves exists.

2.3.4. Contribution of Metadata to the Measurement Process

In this subsection the added value of metadata for data interoperability, consistency and processability is addressed.

Recall that measures are sent from heterogeneous data sources to the GF component through MA. When MA receives data streams from each data source, incorporates metadata accordingly to a common stream-independentlty that measures come from several data sources- and transmit it by means of the C-INCAMI/MIS (*Measurement Interchange Schema*) [4] schema to the GF component. Thus, previous to sending measures, each data source must configure just once each metric that quantifies each attribute (e.g. the cardiac frequency attribute) of the entity under assessment (e.g. an outpatient), and the including contextual properties (e.g. environmental temperature) of the situation. This allows MA be aware of how such metadata should be embedded into the stream.

Hence, CINCAMI/MIS is a schema—based on the C-INCAMI conceptual base as discussed in subsection 2.2-, which cope with interoperability issues in the provision of data from heterogeneous devices, and their further organization.

In **Figure 4** an annotated schema of a C-INCAMI/MIS stream is presented. For each sent stream, MA incorporates to the raw data—e.g. the value 80—the structure of C-INCAMI/MIS schema, indicating the correspondence of each measure with each attribute and contextual property. For instance, in the message of **Figure 4**, IDEntity = 1 represents the *outpatient* entity, IDMetric = 2 the metric *value of cardiac frequency*, and IDProperty = 5 the metric value of *environmental humidity percentage*, in the patient location—representing a contextual property. Thus, the metadata in the message clearly includes a set of information which allows keeping a link between a measure value and the origin of data to identify the data source, the metric and entity ID, among others. This information allows increasing the consistency in the processing model for each M&E project definition.

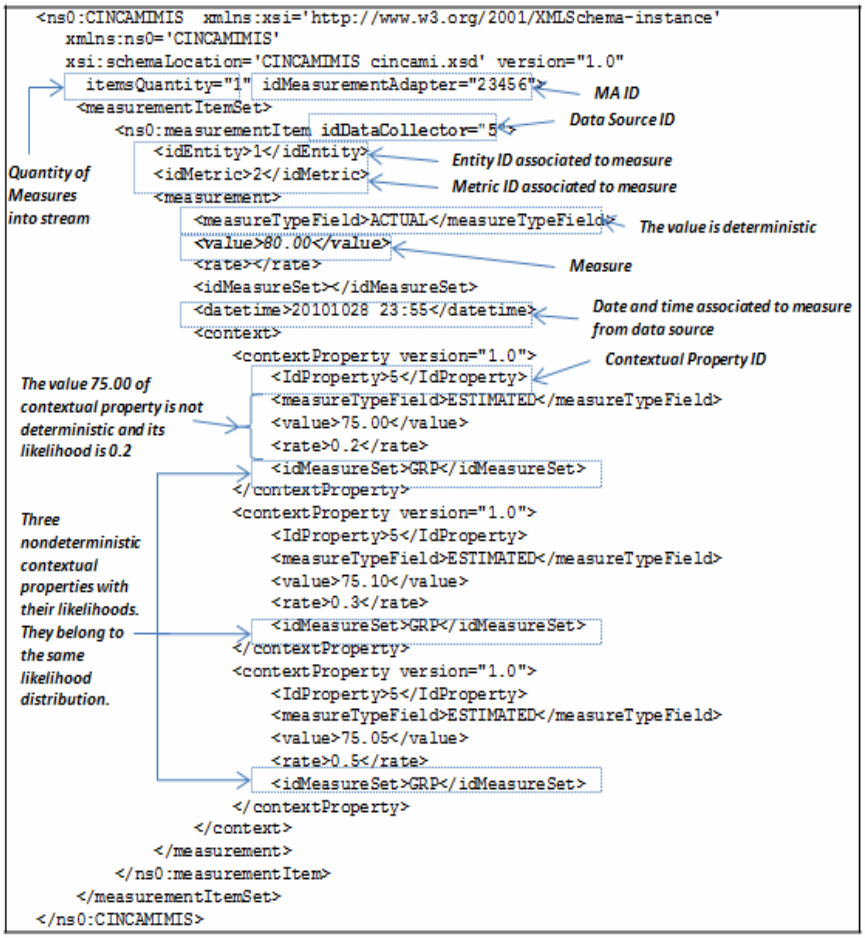


Figure 3. Annotated XML (Extended Markup Language) schema of a C-INCAMI/MIS stream.

Let's suppose, for example, a value of 80 associated to a cardiac frequency of an outpatient arrived; then, the following basic questions can be raised: What does it represent? Which unit of measure does it have? Which mathematical and statistical properties have the value regarng the scale type? Is it good or bad? What is good and what is bad, *i.e*. what are the decision criteria? Could any software process the measure?

Therefore, if the stream metadata were not available, many questions as those could not be answered in a consistent way. Even more, the processability of measures can be hampered and the analysis be skewed.

3. Outpatient Monitoring Scenario

In this section, we illustrate the formal definition of a M&E project for outpatient monitoring, and some aspects of its implementation as well. In the M&E project definition, the knowledge of experts (e.g. doctors) is a valuable asset.

3.1. Definition

The present scenario aims at illustrating the SDSPbMM approach. The underlying hypothesis is doctors of a healthcare centre could avoid adverse reactions and major damage in the health of patients (particularly, outpatients) if they had a continuous monitoring over them. That is to say, doctors should have a mechanism by which can be informed about unexpected variations and/or inconsistencies in health indicators defined by them (as experts). So, the core idea is that there exists some proactive mechanism based on health metrics and indicators that produces an on-line report (alarm) for each risk situation associated to the outpatient under monitoring.

Considering C-INCAMI, the information need is "*to monitor the principal vital signs of an outpatient when he/she is given the medical discharge from the healthcare centre*". The entity under analysis is the *outpatient*. According to medical experts, the *corporal temperature*, the *systolic arterial pressure* (maximum), the *diastolic arterial pressure* (minimum) and the *cardiac frequency* represent the relevant attributes of the outpatient vital signs to monitor. They also consider as necessary monitoring the *environmental temperature*, *environmental pressure*, *humidity*, and *patient position* (*i.e*. latitude and longitude) contextual properties. The definition of the information need, the entity, its associated attributes and the context are part of the "*Nonfunctional requirements specification*" and "*Context specification*" components as discussed in sub-Section 2.2.

The quantification of attributes and contextual properties is performed by metrics as shown in the *Measurement Design and Implementation* component in **Figure 1**. For monitoring purposes, the metrics that quantify the cited attributes, were selected from the C-INCAMI DB repository; likewise the metrics that quantify the cited contextual properties. **Figure 5** shows the specification of the metric for *environmental temperature* contextual property.

After the set of metrics and contextual properties for outpatient monitoring has been selected, the corresponding elementary indicators for interpretation purposes (as discussed in sub-Section 2.2) have also to be selected by experts. In this way, they have included the following elementary indicators: *level of corporal temperature*, *level of pressure*, *level of cardiac frequency* and *level of difference between the corporal and the environmental temperature*. The concepts related to indicators are part of the *Evaluation Design and Implementation* component (see **Figure 1**).

Figure 6 shows the specification of the *level of corporal temperature* elementary indicator. For example, the different acceptability levels with their interpretations are shown, among other metadata. Besides, considering that ranges of the acceptability levels (shown in **Figure 6**) are in a categorical scale (*i.e*. an ordinal scale type), then the target variable for the mining function (classification) is also categorical. So, the classifiers both CC and UC, act relying on the values of the given indicators and their acceptability levels.

3.2. Implementation Issues

Once all the above project information was established, it is necessary for implementation issues to choose a concrete architecture to deploy the system. **Figure 7** depicts an abridged deployment view for the outpatient monitoring system considering the SDSPbMM approach.

Let's suppose we install and set up the MA in a mobile device—the outpatient device-, which will work in conjunction with sensors as shown in **Figure 7**. Therefore,

Context ID	**CtxOutpatient**	
Context Property	Environmental Temperature	
Metric	Value of the environmental temperature of the outpatient	
ID	VTAPT	
Type of Metric	Direct	
Scale	**Type of Scale**	**Interval**
	The Value Domain	**Numeric, Continuous, Real⁺**
	Unit	**Centigrade Degree (°C)**
Method	**Name**	**Temperature Sensing Method**
	Measurement Method	Objective
	Specification	Embedded inside the instrument
	Instrument	
	Name	**Digital Thermometer**
	Version	**TCH305003**
	Provider	**Technidea S.A.**
	Description: The device allows measuring the internal and external temperature and the environmental humidity. Ranges: • Temperature from 10° to +60° • Humidity from 10% to 99%	

Figure 5. Metric definition for the environmental temperature contextual property.

Indicator Code	ETEMP
Indicator Name	Level of corporal temperature
Domain	
(hyperpyrexia, very high fever, fever, normal, hypothermia risk, hypothermia)	
Scale	
Type of Scale	Categorical
The Value Domain	Ordinal
Unit	-

Calculation Method (Elementary Model)

Name	Temperature Analysis

Specification

Axillary Temperature (Direct Metric)	Indicator Value
>= 41.50	Hyperpyrexia
[38.30 , 41.50)	Very high fever
[37.50 , 38.30)	Fever
[36.50 , 37.50)	Normal
[35.00 , 36.50)	Hypothermia risk
< 35.00	Hypothermia (While the temperature is descending there is risk of plain encephalogram).

References:

1. Loscalzo, J., Fauci, A., Braunwald, E., Dennis L., Hauser, S., Longo, D. (2008). "Harrison's principles of internal medicine". McGraw-Hill Medical. pp. Chapter 17, Fever versus hyperthermia. ISBN 0-07-146633-9.
2. Marx, John (2006). "Rosen's emergency medicine: concepts and clinical practice". Mosby/Elsevier. p. 2239. ISBN 9780323028455.

Decision Criteria

Indicator Value	Level of Acceptability
Hyperpyrexia	**Not Acceptable**
Very high fever	**Not Acceptable**
Fever	**Poorly Acceptable**
Normal	**Acceptable**

Figure 6. Details of the level of corporal temperature elementary indicator specification.

while the data collecting and adapting processes are implemented in a mobile device by the MA, the gathering function and other processes can reside in the healthcare center computer. The MA component, using web services, informs the measures (streams) to the gathering function (GF) in an asynchronous and continuous way. MA takes the measures from sensors—the data sources- and incurporates the associated metric metadata for attributes and contextual properties accordingly. For instance, it incurporates the contextual property ID for the *environmental temperature* (VTAPT, in **Figure 5**) joint to the value to transmit; and so for every attribute and contextual property. Note that data (values) and metadata are transmitted through the C-INCAMI/MIS schema to the gathering function (GF), as discussed in sub-Section 2.3.4.

When the gathering function receives measures from several outpatients under monitoring it arranges them, for instance, by patient (*i.e.* the trace group) and transmits them to the analysis and correction processes. As discussed in subSection 2.3.2, ASF mainly solves typical problems of data such as missing values, noises, among others. For example, and thanks to metadata, if ASF receives for the *Value of axillary temperature* metric a zero value, by the metric definition the processing model identifies an error because the scale is numeric (in interval scale type),

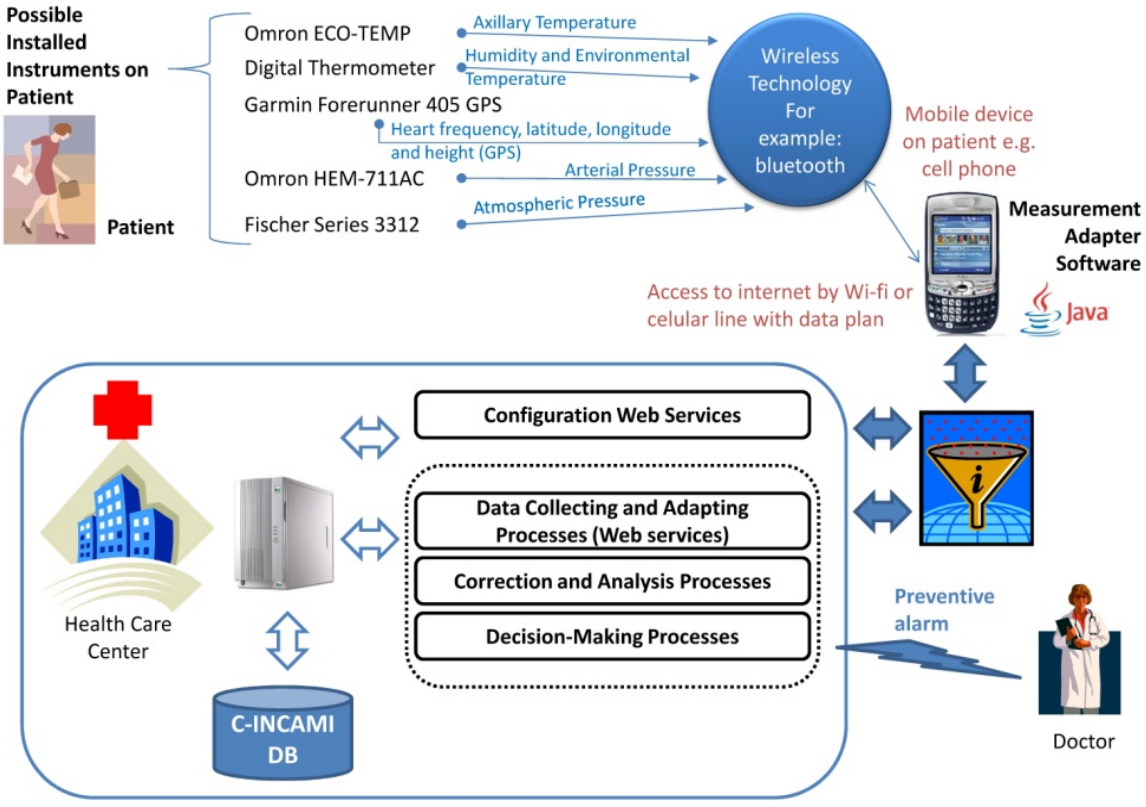


Figure 7. A deployment view for the Outpatient Monitoring System.

continuous, and defined in the interval of positive real numbers.

Although all the values of metrics and contextual properties from monitored outpatients are simultaneously received and analyzed, let's consider for a while, for illustration purpose, that the system only receives data for the *axillary temperature* attribute and the *environmental temperature* contextual property from one outpatient, and that also the system visualizes them. As depicted in **Figure 8**, the lower and upper limit defined for the *level of* (*axillary temperature*) *corporal temperature* indicator together with the evolution of the *environmental temperature* and the *axillary temperature* can be tracked.

The measures and, ultimately, the acceptability level achieved by the *level of corporal temperature* elementary indicator (see **Figure 8**) indicate a normal situation for the patient. Nevertheless, the on-line decision-making process, apart from analyzing for attributes the level of acceptability met also analyzes the interaction with contextual properties and their values. This analysis allows detecting a situation like that exposed in **Figures 9(a)** and **(b)**.

At first glance, what seemed to be normal and evident, it was probably not because in a proactive form the processing model has detected a correlation between axillary temperature and environmental temperature as shown in **Figure 9(b)**. This could cause the triggering of a preventive alarm from the healthcare centre to doctors, because the increment on the environmental temperature can drag in turn the increment in the corporal temperature, and therefore this situation can be associated to a gradual raise in the risk probability for the outpatient.

4. Scenario Simulation

4.1. Goal

The developed prototype for the SDSPbMM approach implements functionalities (see **Figure 2**) ranging from the formal definition of the M&E project including the C-INCAMI repository with metadata, the integration of heterogeneous data sources, trace groups and MA, to classifiers for on-line decision-making process. In addition, it implements the C-INCAMI/MIS schema for the interchange of measures in an interoperable way, and the multilevel buffer based on metadata (see **Figure 3**).

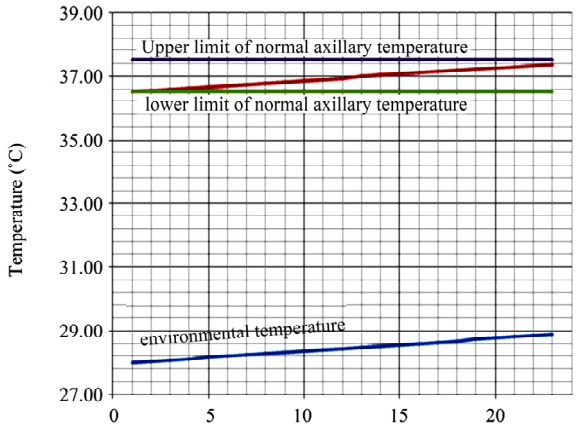
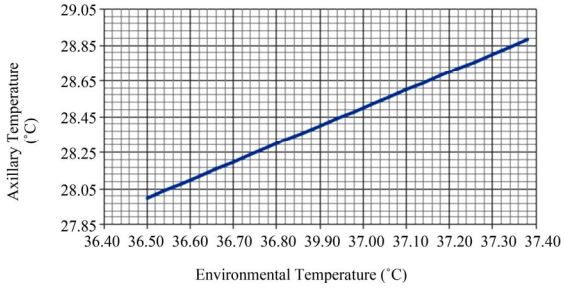


Figure 8. Visualization of the evolution of axillary temperature versus environmental temperature measures.



(a)

Correlation Coefficients

Pearson's Correlation: coefficients/likelihoods

	Axillary Temp	Environmental Temp
Axillary Temp	1.00	0.00
Environmental Temp	1.00	0.00

(b)

Figure 9. (a) Correlation Analysis for the axillary temperature versus the environmental temperature; (b) Correlation Matrix.

The prototype has been implemented in JAVA, using R [8] as statistical calculus engine, and the *CRAN* (*Comprehensive R Archive Network*) Rserve mechanism to access TCP/IP from the streaming application to the R engine, without requiring persistence and prioritizing the direct communication.

The simulation goal is to determine the processing times involved in the outpatient scenario and the variable scalability. This simulation can allow us analyzing the feasibility of applying the prototype to real situations. Furthermore, we discuss statistically the results of the simulation, in order to detect the components which incorporate more variability to the system.

4.2. Simulation Planning and Execution

The simulation has been performed from the illustrated scenario in Section 3. The measurement data have been generated in a pseudo-random way considering two parameters: *quantity of metrics* (in a simulation each metric corresponds to a variable), and *quantity of measurements by variable*. Each patient has 8 associated metrics pertaining to attributes and contextual properties as commented in sub-Section 3.1.

The simulation discretely varied the quantity of variables (metrics) into the data stream from 3 to 99, and the quantity of measurements by variable from 100 to 1000.

The idea of discretely vary the quantity of metrics instead of doing it as a multiple of 8—*i.e.* based on the 8 ones associated per each patient-, lies in analyzing the prototype behavior in presence of missing values and the progressive reincorporation of measures to the stream.

The prototype, R and Rserve were running in a PC equipped with AMD Athlon × 2 64 bits processor, 3 GB of RAM, and Windows Vista Home Premium as operating system.

For the simulation, the following variables which are the target of measurement considering the stream as the entity under analysis have been defined, namely:

- **Startup**: the necessary time (in ms) to start up the functions of the analysis
- **AnDesc**: the necessary time (ms) to make the descriptive analysis on the complete data stream
- **Cor**: the necessary time (ms) to make the correlation analysis by trace group inside the complete data stream
- **Pca**: the necessary time (ms) to make the principal component analysis by trace group inside the complete data stream
- **Total**: the necessary time (ms) to make all the analysis on the complete data stream

The simulation parameters used for the statistical analysis of results are represented by *qVar*, to indicate the quantity of variables of the data stream; and by *meds*, to indicate the quantity of measures by variable of the data stream. From now onwards, in order to simplify the reading of the statistical analysis, the parameters *qVar* and *meds* will be directly referred to as variables, and *Startup*, *Andesc*, *Cor*, *Pca* and *Total* will be called variables as well.

From the simulation process standpoint, we have obtained 1390 measurements over the overall processing time in relation to the evolution of the quantity of variables and measurements. This allows us to statistically arrive at verifiable results that help us consequently validating the prototype in a controlled environment.

4.3. Analysis of Results

The chart in **Figure 10(b)** clearly shows us how the evolution of quantity of variables affects significantly the overall processing time of data streams, incrementing according to the values shown in chart **(a)**. Here, we can observe that the increment in the processing time produced by the increase of measurements is extremely low in compareson with the one produced by the increase of variables. This latter aspect indicates that the load shedding mechanism really achieved the goal of avoiding overflows without affecting the time of stream processing against the variation of the stream volume. While the incorporation of new variables does influence because besides the stream volume by adding a new variable, there exists also the interaction with the preexistent variables, being this the cause

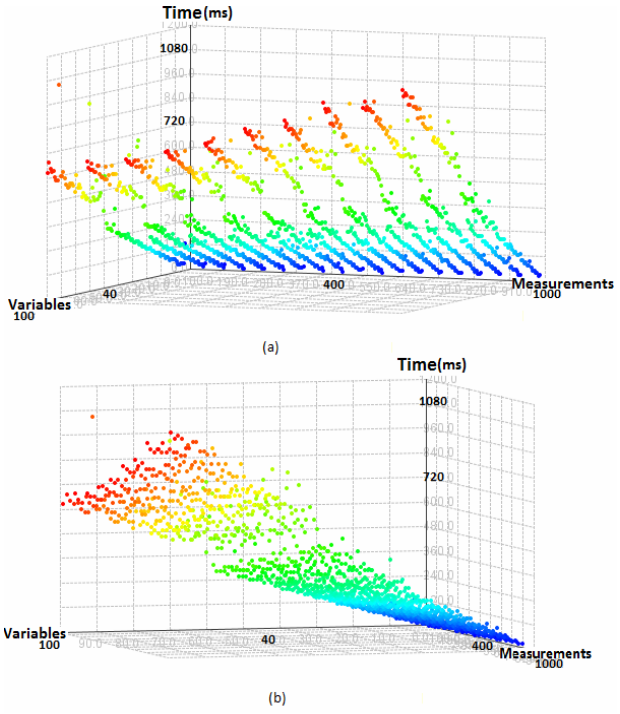


Figure 10. Two views of the evolution of overall processing time (ms) against the evolution of the quantity of variables and measurements.

and main difference in terms of the processing time with respect to the increase produced by measurements.

In both dispersion charts, **(a)** and **(b)**, each point is represented with a color that is associated with the quantity of variables. This allows us to identify regions in the graph in a graceful way and to compare them from both perspectives. In chart **(b)**, we can observe that the overall processing time keeps a linear relation according to the quantity of variables. Considering such a situation, and on the basis that the statistical analyzer (ASF in **Figure 2**) performs a series of analysis on the data stream, we have studied the incidence of each analysis in the overall processing time, in order to detect which of them are more critical in temporary terms.

The Pearson's correlation matrix shown in **Figure 11(a)** would confirm, firstly, the linear relationship indicated between the quantity of variables (*qVar*) and the overall processing time of the data stream (*Total*) given the coefficient value of 0.95. Secondly, it can be concluded that the overall processing time would keep a strong linear relationship with respect to the time of the descriptive analysis with a coefficient of 0.99, followed up by the time of *Pca* with 0.9, and *Cor* with 0.89 respectively.

The resulting matrixes of the principal component analysis—shown in **Figures 11(b)** and **(c)** reveal which of the variables provide more variability to the system. Thus, the first autovalue (row 1, **Figure 11(b)**) explains the 66% of

Correlation coefficients

Pearson correlation: Coefficients\probabilities

	qVar	meds	Startup	AnDesc	Cor	Pca	Total
qVar	1.00	1.00	0.15	0.00	0.00	0.00	0.00
meds	8.9E-05	1.00	0.11	5.8E-06	0.00	0.00	0.00
Startup	−0.04	−0.04	1.00	0.28	0.30	0.68	0.41
AnDesc	0.97	0.12	−0.03	1.00	0.00	0.00	0.00
Cor	0.79	0.40	−0.03	0.85	1.00	0.00	0.00
Pca	0.77	0.45	−0.01	0.85	0.89	1.00	0.00
Total	0.95	0.21	−0.02	0.99	0.89	0.90	1.00

(a)

Principal component analysis

Standardized data

Eigenvalues

Lambda	Value	Proportion	Cum. prop.
1	4.62	0.66	0.66
2	1.11	0.16	0.82
3	0.99	0.14	0.96
4	0.14	0.02	0.98
5	0.10	0.01	1.00
6	0.02	3.3E-03	1.00
7	3.3E-05	4.7E-06	1.00

(b)

Eigenvectors

Variables	e1	e2	e3
qVar	0.43	−0.31	−0.08
meds	0.14	0.87	0.18
Startup	−0.02	−0.22	0.98
AnDesc	0.45	−0.19	−0.04
Cor	0.44	0.14	0.04
Pca	0.44	0.18	0.07
Total	0.46	−0.11	−0.01

(c)

Figure 11. (a) Pearson's correlation matrix, (b) Matrix of autovalues, and (c) Matrix of autovectors associated to principal component analysis (PCA).

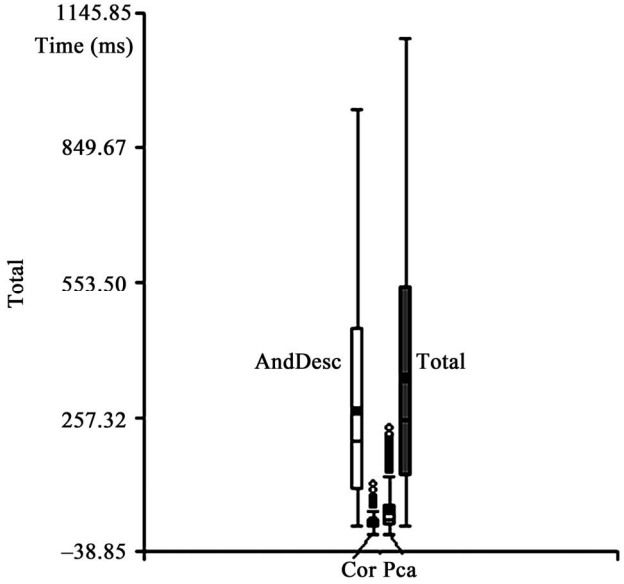


Figure 12. Boxplot of the AnDesc, Cor, PCA and total variables.

the variability of the system. Also, if we look at its composition in the matrix of autovectors (col. e1, **Figure 11 (c)**), the variables that more contribute—in absolute terms-are *AnDesc*, *Cor*, *Pca* y *qVar*.

Therefore, if we want to replace the seven cited variables with the three new variables (e1 to e3), we would be explaining the 96% of the variability of the system, where the main variables in terms of contribution are associated to *AnDesc*, *Cor*, *Pca* y *qVar*. The system is only affected in a 16% by the evolution of measurements and in a 14% by the startup time. This is an important aspect to remark because the only external variable to the prototype, *i.e.* the volume of measurements arrival, which cannot be controlled by it, represents just a 16% and by no way constituted an overflow situation in the queue of services.

Lastly, taking into account the four variables that more contributed to the system variability, three of them are part of the overall processing time. In this way, and using the box plot of **Figure 12**, we can corroborate that the most influential variable, in terms of the magnitude to the overall processing time, is *AnDesc*. In addition, note that the biggest resulting time to process 99 variables (metrics) with 1000 measures each (*i.e.* in total 99,000 measures by stream) was 1092 ms.

This outcome represents a satisfactory applicability threshold for the prototype, especially taking into account the basic hardware used. So, in our humble opinion, this could easily meet the response time requirements for the outpatient monitoring scenario.

5. Related Work and Discussion

There are many researches oriented to data stream processing from the syntactical point of view, in which the continuous query over data streams is made in terms of attributes and their associated values using CQL (*Continuous Query Language*) [9]. This approach has been implemented in several projects such as Aurora & Borealis [10], STREAM [11], and TelegraphCQ [12], among others. Our approach (and prototype) includes the capability to incorporate metadata based on an M&E framework, which allows guiding the organization of data streams in the buffer; making possible the consistent and comparable analysis from the statistical standpoint; triggering alarms in a proactive way using several statistical analysis or from taken decisions stemming from classifiers.

MavStream [13] is a prototype for a data stream management system, which has the capability of processing complex events as an intrinsic aspect for data stream processing. In this sense, our prototype supports the on-line data stream analysis with the incorporation of metadata to measures (data), handling not only measures values coming from attributes of the assessed entity but also those coming from contextual properties related to the situation of the entity. In addition, the SDSPbMM prototype can process measures with nondeterministic results, and perform analysis by trace group (or an overall analysis), which in practical scenarios such as is the case for monitoring of outpatients [4], represent crucial features.

Nile [14] is a data stream management system based on a conceptual framework for detection and tracking of phenomena or situations supported by deterministic mea-

sures. Our prototype unlike Nile, allows the incorporation of heterogeneous data sources embracing not only deterministic but also nondeterministic measures. On the other hand, Singh *et al.* [15] introduce a system architecture for a formal framework of data mining oriented to the situation presented in [16]. This system is used in medical wireless applications and shows how the architecture can be applied to several medical areas such as diabetes treatment and risk monitoring of heart disease. In our humble opinion, this system neglects central issues for assuring repeatability and interoperability because it lacks a clear specification of metrics both for entity attributes and contextual properties, indicators, scales and scale types, among other metadata.

Lastly, Huang *et al*. [17] present an approach based on self-managed reports for tracking of patients. Such reports are made up of a set of questionnaire items with numeric (scale) responses, which are filled in by patients at home. The patient's responses feed a classification model based on neural networks in order to progressively improve the selection of questionnaire items incorporated in the reports. Hence, they argue this decrease the patients' response time and allow identifying those aspects that will foster an improvement in their quality of life. Likewise happens in Singh *et al*. [15,16] proposal; this approach says nothing about how to define metrics, indicators, scales, and so on.

Our strategy and its prototype support data stream processing in alignment with a conceptual base, *i.e*. the metric and indicator ontology [3], which guarantees not only syntactic but also semantic processing, in addition to interoperability and consistency.

6. Conclusions and Future Work

In this work, we have discussed how the presence of metadata based on the C-INCAMI M&E framework linked to measures in data streams, allows the organization of measurements which foster the consistency for statistical analysis, since they specify not only the formal components of data but also the associated context. Hence, it is possible to perform particular analysis at trace group or at a more general level, comparing values of metrics among different trace groups in order to identify, for example, deviations of measures against their formal definition; the main system variability factors, as well as relations among variables.

In the outpatient monitoring scenario introduced in Section 3, we have shown in **Figures 8** and **9** the relationships between the data/metadata of metrics—*i.e*. metrics that quantify contextual properties and attributes- and the statistical analysis, considering our data stream processing strategy. In this sense, even when the measures seemed to have normal values, the data and metadata of metrics in conjunction with the correlation analysis allowed identifying a drag situation and then triggering alarms to prevent it. Moreover, such metadata allowed identifying variability factors and detecting trends in a consistent way considering the contextual situation as well.

Using the developed prototype which implements the SDSPbMM strategy, we have demonstrated from the statistical analysis viewpoint of the simulation outcomes that the prototype is more susceptible, in terms of processing time, to the increase of the quantity of variables than to the increase of the quantity of measurements by variable. Using the principal components analysis technique, we have proved that the aspects that more contribute to the system variability are those associated to the *Andesc*, *Cor*, *Pca* and *qVar* variables, being *Andesc* the one that defines the biggest proportion of the final processing time of data streams.

Taking into account that the implemented prototype ran in a system which is totally accessible in the market, we could establish as a benchmark that to process 99,000 measurements (99 variables and 1000 measures/variable), the biggest time spent was 1092 ms. This is an important starting point since now we can consistently evaluate several application scenarios against this benchmark. On the other hand, the effectiveness of the load shedding mechanism in the multilevel buffer was also proved statistically, coming up that the evolution of the quantity of measurements does not compromise the prototype operation and the final processing time of the data stream has not been affected.

Although the present work is just a simulation of the outpatient monitoring scenario by means of a prototypical software application, we have initially proved that it can scale up to a real scenario. As a future work, we also plan to experimentally test our data stream processing strategy enriched with context, measurement and evaluation metadata on several scenarios in order to statistically validate the initial benchmark obtained for the outpatient monitoring scenario.

7. Acknowledgements

This research is supported by the PICT 2188 project from the Science and Technology Agency and by the 09/F052 project from the UNLPam, Argentina.

REFERENCES

[1] J. Namit, J. Gehrke and H. Balakrishan, "Towards a Streaming SQL Standard," *Proceedings of the VLDB Endowment*, Vol. 1, No. 2, 2008, pp. 1379-1390.

[2] H. Molina and L. Olsina, "Towards the Support of Contextual Information to a Measurement and Evaluation Framework," *International Conference on Quality of Information and Communications Technology*, Lisbon, 12-14

September 2007, pp. 154-166.

[3] L. Olsina, F. Papa and H. Molina, "How to Measure and Evaluate Web Applications in a Consistent Way," In: G. Rossi, O. Pastor, D. Schwabe and L. Olsina, Eds., *Web Engineering*: *Modelling and Implementing Web Applications*, Springer Book, London, 2008, pp. 385-420.

[4] M. Diván and L. Olsina, "Integrated Strategy for the Data Stream Processing: A Scenario of Use," *Proceeding of Iberoamerican Conference in "Software Engineering"*, Medellín, 2009, pp. 374-387.

[5] M. Diván, L. Olsina and S. Gordillo, "Data Stream Processing Enriched with Measurement Metadata: A Statistical Analysis," *Proceeding of Iberoamerican Conference in "Software Engineering"* (CIbSE), Rio de Janeiro, 2011, p. 29.

[6] M. Wei, W. Rundensteiner and M. Mani, "Utility-Driven Load Shedding for XML Stream Processing," *International Conference on World Wide Web*, Beijing, 21-25 April 2008, pp. 855-864.

[7] C. Marrocco, R. Duin and F. Tortorella, "Maximizing the Area under the ROC Curve by Pairwise Feature Combination," *ACM Pattern Recognition*, Vol. 41, No. 6, 2008, pp. 1961-1974. doi:10.1016/j.patcog.2007.11.017

[8] R. Software Foundation, "The R Foundation for Statistical Computing," 2010. http://www.r-project.org/foundation/

[9] S. Babu and J. Widom, "Continuous Queries over Data Streams," *ACM SIGMOD Record*, Vol. 30, No. 3, 2001, pp. 109-120. doi:10.1145/603867.603884

[10] D. Abadi, Y. Ahmad, M. Balazinska, U. Cetintemel, M. Cherniack, J. Hwang, W. Lindner, A. Maskey, A. Rasin, E. Ryvkina, N. Tatbul, Y. Xing and S. Zdonik, "The Design of the Borealis Stream Processing Engine," *Conference on Innovative Data Systems Research* (CIDR), Asilomar, 2005, pp. 277-289.

[11] The Stream Group, "STREAM: The Stanford Stream Data Manager," Stanford, 2003.

[12] S. Krishnamurthy, S. Chandrasekaran, O. Cooper, A. Deshpande, M. Franklin, J. Hellerstein, W. Hong, S. Madden, F. Reiss and M. Shah, "Telegraph CQ: An Architectural Status Report," *IEEE Data Engineering Bulletin*, Vol. 26, No. 2, 2003, pp. 11-18.

[13] S. Chakravarthy and Q. Jiang, "Stream Data Processing: A Quality of Service Perspective," Springer Book, New York, 2009.

[14] M. Ali, W. Aref, R. Bose, A. Elmagarmid, A. Helal, I. Kamel and M. Mokbel, "NILE-PDT: A Phenomenon Detection and Tracking Framework for Data Stream Management Systems," *Very Large Database*, Trondheim, 2005, pp. 1295-1298.

[15] S. Singh, P. Vajirkar and Y. Lee, "Context-Aware Data Mining Framework for Wireless Medical Application," *LNCS of Springer*, Vol. 2736, 2003, pp. 381-391.

[16] S. Singh, P. Vajirkar and Y. Lee, "Context-Based Data Mining Using Ontologies," *LNCS of Springer*, Vol. 2813. 2003, pp. 405-418.

[17] Y. Huang, H. Zheng, C. Nugent, P. McCullagh, N. Black, K. Vowles and L. McCracken, "Feature Selection and Classification in Supporting Report Based Self Management for People with Chronic Pain," *IEEE Transactions on Information Technology in Biomedicine*, Vol. 15, No. 1, 2011, pp. 54-61.

