Vista de Proceso del Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones

Mario Diván[1,2], Luis Olsina[2]

[1]Facultad de Ciencias Económicas y Jurídicas, UNLPam, Santa Rosa, La Pampa, Argentina
[2]Facultad de Ingeniería, UNLPam, General Pico, La Pampa, Argentina.

{mjdivan,olsinal}@ing.unlpam.edu.ar

Abstract. El enfoque integrado de procesamiento de flujos de datos centrado en metadatos de mediciones, es un gestor de flujos de datos sustentado en un marco de medición y evaluación, el cual incorpora comportamiento detectivo y predictivo, mediante el empleo de las mediciones y metadatos asociados. Este trabajo discute la formalización de los procesos asociados con el funcionamiento del gestor de flujos de datos, como así también la interacción entre ellos. La formalización de los procesos se realiza en base a SPEM, y se consideran las actividades comprendidas entre la configuración de las fuentes de datos involucradas en un proceso de medición y evaluación, hasta aquellas asociadas con la emisión de las alarmas detectivas o predictivas. Esto permite hacer comunicable el aspecto de procesos del enfoque, y adicionalmente, abre la posibilidad para medir y evaluar los propios procesos formalizados del enfoque, como medio para monitorear cuantitativamente su salud funcional en línea.

Keywords: Procesos, Medición, Flujo de Datos, C-INCAMI.

1 Introducción

Actualmente, existen aplicaciones que procesan un conjunto de datos a medida, generados en forma continua, a los efectos de responder a consultas y/o adecuar su comportamiento en función del propio arribo de los datos [1], como es el caso de las aplicaciones para el monitoreo de signos vitales de pacientes; del comportamiento de los mercados financieros; entre otras. En dicho tipo de aplicaciones, se enmarca el Enfoque Integrado de Procesamiento de Datos centrado en Metadatos de Mediciones (*EIPFDcMM*) [2], el cual sustentado en el marco de medición y evaluación C-INCAMI (*Context-Information Need, Concept model, Attribute, Metric and Indicator*) [3,4], incorpora metadatos al proceso de medición, promoviendo la repetitividad, comparabilidad y consistencia del mismo. Desde el punto de vista del sustento semántico y formal para la medición y evaluación (*M&E*), el marco conceptual C-

INCAMI establece una ontología que incluye los conceptos y relaciones necesarias para especificar los datos y metadatos de cualquier proyecto de M&E. Por otra parte, y a diferencia de otras estrategias de procesamiento de flujos de datos [5,6,7], gracias a la incorporación de metadatos, el EIPFDcMM es capaz de guiar el procesamiento de las medidas provenientes de fuentes de datos heterogéneas, analizando cada una de ellas dentro de su contexto de procedencia, como así también su significado dentro del proyecto de M&E en el que se definió. Adicionalmente, se incorpora en el enfoque el comportamiento detectivo y predictivo sobre las medidas contextualizadas, lo que permite realizar el monitoreo activo de las entidades bajo análisis.

De este modo, el abordaje de una medida en el EIPFDcMM, no se acota exclusivamente al arribo de un valor sintáctico, sino por el contrario, la medida arriba acompañada por los atributos que cuantifican su contexto de procedencia, y sus metadatos. Estos últimos, permiten guiar el procesamiento de la medida contextualizada, a partir de la interpretación de sus respectivos significados dentro del proyecto de M&E. Ahora bien, hasta este momento, tal procesamiento dentro del enfoque, era descrito en forma conceptual y narrativa, lo que dificultaba, o al menos dejaba poco claro al lector, de qué modo los metadatos se asociaban con las fuentes de datos que generan las medidas, en qué forma los mismos guiaban el procesamiento, y cómo en base a ellos, actuaban los mecanismos detectivos y predictivos. Así, y como contribuciones específicas se plantea, *(i) relacionado con la configuración de las fuentes de datos*: la formalización del proceso de configuración en base al lenguaje SPEM (*Software & Systems Process Engineering Metamodel*) [8], lo que permite comunicar formalmente el modo en que cada fuente de datos heterogénea es incorporada en la estrategia de procesamiento; *(ii) relacionado con la captación de medidas:* la formalización del proceso de adaptación y recolección de medidas en base a SPEM. Ello permite comunicar formalmente, el orden en que se da la recolección de medidas desde las fuentes de datos, cómo se gestionan localmente las medidas, hasta el momento de su posterior envío y procesamiento; *(iii) relacionado con el procesamiento de medidas:* la formalización del proceso de corrección y análisis en base a SPEM, permite comunicar el orden en el que los análisis son realizados, las actividades efectuadas en paralelo, como así también, su relación con el software estadístico R en cada instante; *(iv) relacionado con la toma de decisión sobre las medidas:* la formalización del proceso de toma de decisión en base a SPEM, facilita la interpretación sobre cómo las alarmas estadísticas son gestionadas, y adicionalmente, cómo se realiza la aplicación y toma de decisión en base al clasificador incremental, con el objetivo de prevenir riesgos. Estas contribuciones representan un importante avance con respecto al modelo de procesamiento presentado en [9,10], ya que ahora hay una descripción formal de los procesos asociados con el enfoque, lo que promueve su comunicación, y clarifica el orden en que sus actividades son llevadas a cabo.

El presente artículo se organiza en seis secciones. La sección 2 resume el marco C-INCAMI. La sección 3 sintetiza la idea y arquitectura del EIPFDcMM. La sección 4 esquematiza la vista global de los procesos, sus relaciones y plantea la formalización de los procesos de configuración, recolección y adaptación, corrección y análisis, y

toma de decisión. La sección 5 discute los trabajos relacionados, y por último, se resumen las conclusiones y trabajos a futuro.

2 Panorama de C-INCAMI

C-INCAMI es un marco conceptual [3,4] que define los módulos, conceptos y relaciones que intervienen en el área de M&E, para organizaciones de software. Se basa en un enfoque en el cual la especificación de requerimientos, la medición y evaluación de entidades y la posterior interpretación de los resultados están orientadas a satisfacer una necesidad de información particular. Está integrado por los siguientes componentes principales: 1) Gestión de Proyectos de M&E; 2) Especificación de Requerimientos no Funcionales; 3) Especificación del Contexto del Proyecto; 4) Diseño y Ejecución de la Medición: y 5) Diseño y Ejecución de la Evaluación. La mayoría de los componentes están soportados por los términos ontológicos definidos en [4]. En la Figura 1, se muestra un diagrama con los principales conceptos y relaciones para los componentes de requerimientos, contexto y medición.

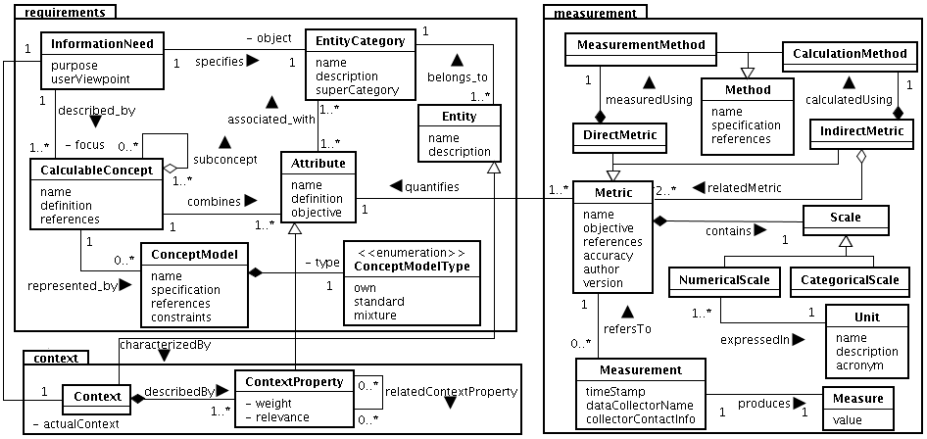

Fig. 1. Principales conceptos y relaciones de los componentes Especificación de Requerimientos no Funcionales, Especificación del Contexto y de la Medición

Los flujos de medidas que se informan desde las fuentes de datos al EIPFDcMM, se estructuran incorporando a las medidas, metadatos basados en C-INCAMI tales como la métrica a la que corresponde, el grupo de seguimiento asociado, el atributo de la entidad que se mide, entre otros. Dentro del flujo, se etiquetan conjuntamente con cada medida asociada al atributo, las medidas asociadas a cada propiedad de contexto. Gracias a la formalización del proyecto de M&E en base a C-INCAMI, el hecho de procesar el flujo etiquetado, permite la estructuración del contenido de un modo consistente y alineado con el objetivo del proyecto. Esta estructuración de las mediciones dentro del EIPFDcMM, mantiene el concepto con el que se asocia cada medida; por ejemplo, si es una medida de atributo o bien de propiedad contextual. De este modo, se

enriquece el análisis estadístico dado que es posible en forma directa, verificar la consistencia formal y sintáctica de cada medida contra su definición formal, en forma previa a avanzar con técnicas estadísticas más complejas. En trabajos anteriores [2,10], se introdujo el caso de monitoreo sobre pacientes transplantados ambulatorios, en donde los médicos de un centro de salud pretendían contar con un seguimiento continuo del paciente, a los efectos de evitar reacciones adversas y daños mayores. Así, la entidad bajo análisis se definió como el paciente transplantado ambulatorio, y se especificaron atributos tales como la frecuencia cardíaca, la temperature axilar, la presión arterial sitólica y diastólica. Adicionalmente, a los efectos de circunscribir los riesgos que el entorno podía representar en el paciente, se definieron como propiedades del contexto a la temperatura atmosférica, la presión ambiental, la humedad y el posicionamiento del paciente (latitutd y longitud). Una vez definidos los atributos y las propiedades de contexto, se definieron sus métricas asociadas (por ejemplo, *el valor de la temperatura axilar*), para proceder luego a su medición, evaluación y disparo de alarmas ante situaciones preventivas o de riesgo.

3 Panorama del EIPFDcMM

El EIPFDcMM es un gestor de flujos semi-estructurados de mediciones, enriquecidos con metadatos sustentados en C-INCAMI, especializado en proyectos de M&E, que incorpora comportamiento detectivo y predictivo en línea. Como puede apreciarse en la Figura 2, la idea que subyace al modelo en términos de procesamiento de flujos [10] es la siguiente.

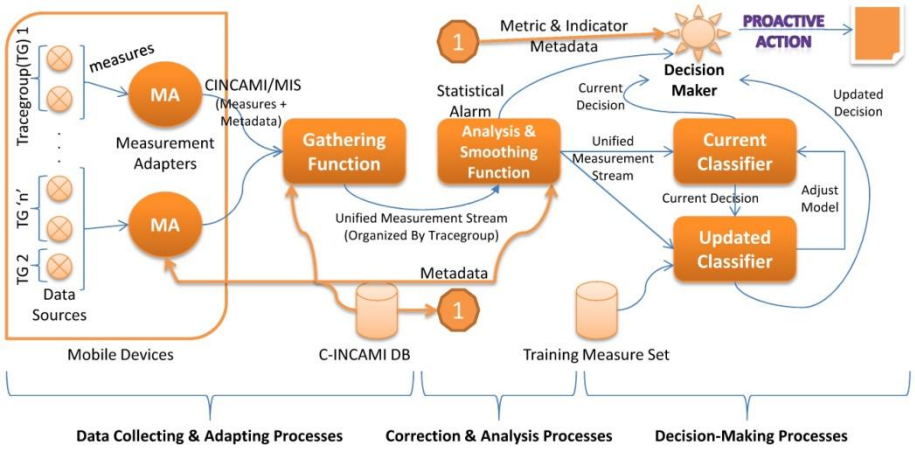

Fig. 2. Esquema Conceptual del EIPFDcMM

Las mediciones se generan en las fuentes de datos heterogéneas, las cuales abastecen a un módulo denominado adaptador de mediciones (*MA* en la Figura 2) generalmente embebido en dispositivos móviles por una cuestión de portabilidad y practicidad,

aunque podría embeberse en cualquier dispositivo de cómputo con asociación a fuentes de datos. MA incorpora junto a los valores medidos, los metadatos del proyecto de medición y los informa a una función de reunión central (*Gathering Function –GF*). GF incorpora los flujos de mediciones en un buffer organizado por grupos de seguimiento –modo dinámico de agrupar a las fuentes de datos definido por el director del proyecto de M&E- con el objeto de permitir análisis estadísticos consistentes a nivel de grupo de seguimiento o bien por región geográfica donde se localicen las fuentes de datos, sin que ello implique una carga adicional de procesamiento. Adicionalmente, GF incorpora técnicas de load shedding [11] que permiten gestionar la cola de servicios asociada a las mediciones, mitigando los riesgos de desborde independientemente el modo en que se agrupen.

Una vez que las mediciones se encuentran organizadas en el buffer, se aplica análisis descriptivo, de correlación y componentes principales *(Analysis & Smoothing Function –ASF-*) guiados por sus propios metadatos, a los efectos de detectar situaciones inconsistentes con respecto a su definición formal, tendencias, correlaciones y/o identificar las componentes del sistema que más aportan en términos de variabilidad. De detectarse alguna situación en ASF, se dispara una alarma estadística al tomador de decisiones (*Decision Maker -DM*) para que evalúe si corresponde o no disparar la alarma externa (vía, e-mail, SMS, etc) que informe al personal responsable de monitoreo sobre la situación. En paralelo los nuevos flujos de mediciones son comunicados al clasificador vigente (*Current Classifier –CC-*), quien deberá clasificar las nuevas mediciones si corresponden o no a una situación de riesgo e informar dicha decisión al DM. Simultáneamente, se reconstruye el CC incorporando las nuevas mediciones al conjunto de entrenamiento y produciendo con ellas un nuevo modelo (*Updated Classifier -UC*). El UC clasificará las nuevas mediciones y producirá una decisión actualizada que también será comunicada al DM. El DM determinará si las decisiones indicadas por los clasificadores (*CC y UC*) corresponden a una situación de riesgo y en cuyo caso con qué probabilidad de ocurrencia, actuando en consecuencia según lo definido en el umbral mínimo de probabilidad de ocurrencia definido por el director del proyecto. Finalmente, independientemente de las decisiones adoptadas, el UC se torna en CC sustituyendo al anterior, en la medida que exista una mejora en su capacidad de clasificación según el modelo de ajuste basado en curvas ROC (*Receiver Operating Characteristic*) [12].

4 Formalización de los procesos del EIPFDcMM

4.1 Vista global de los procesos

El EIPFDcMM contiene tres procesos centrales y uno de soporte. Los procesos centrales se asocian con la recolección y adaptación de medidas, su posterior corrección y análisis, y finalmente, la toma de decisión en base a ellas (ver Figura 3). Por otro lado, el único proceso de soporte, se refiere a la configuración de las fuentes de datos con respecto a un proyecto de medición y evaluación dado.

Como puede apreciarse en la Figura 3, todos los procesos centrales dependen directa o indirectamente del proceso de configuración. El proceso de configuración,

para un proyecto de M&E dado, tiene por objetivo establecer la correspondencia entre las fuentes de datos asociadas con un MA, y las métricas respectivas vinculadas con un atributo de la entidad bajo análisis, o bien, con sus propiedades contextuales.

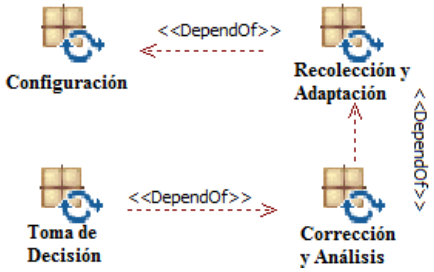

Fig. 3. Dependencia entre procesos del EIPFDcMM usando SPEM

Luego, el proceso de recolección y adaptación, recolectará las medidas desde las fuentes de datos, les incorporará los metadatos necesarios para su gestión, los informará conjuntamente a través del flujo C-INCAMI basado en el esquema de intercambio de mediciones, hasta poder ubicar los mismos dentro del buffer central. El esquema de intercambio de mediciones, se denomina CINCAMI/MIS [10] por ***M**easurement **I**nterchange **S**chema*, y sintéticamente, se trata de un flujo semi-estructurado basado en XML y C-INCAMI, el cual incorpora etiquetado específico para discriminar los aspectos semánticos de cada medida.

A medida que los datos y los metadatos arriban al buffer central, el proceso de corrección y análisis tomará instantáneas del buffer, liberará sus recursos, y establecerá así ventanas temporales para el procesamiento. Sobre dichas ventanas, se aplicarán una serie de análisis estadísticos, y se contrastarán los resultados con respecto a la definición formal del proyecto de M&E. De existir desvíos, se informan como alarmas al proceso de toma de decision, para que proceda según lo definido.

Por otro lado, el proceso de toma de decisión tiene por objetivo anticipar situaciones de riesgo a través de los clasificadores, como así también y ante la eventual existencia de alarmas estadísticas, analizar y dar curso a estas cuando correspondan. A seguir, describimos en cada sub-sección sendos procesos.

4.2 Proceso de configuración

Todos los procesos dependen directa o indirectamente del proceso de configuración, lo cual es lógico por cuanto es quien establece la correspondencia entre el dispositivo fisico de medición, representado lógicamente por la fuente de datos, y las métricas asociadas con la entidad bajo análisis, o bien, sus propiedades contextuales.

El proceso comienza leyendo la configuración local del MA (ver Figura 4), lo que permite conocer si existe una configuración previa que se desee actualizar, o bien, si se trata de una nueva configuración. Luego, se verifica la disponibilidad del servidor, el cual permite el acceso a la definición de todos los proyectos de M&E vigentes, estructurados mediante C-INCAMI. A seguir, y a través de servicios web, se solicita

que se informen los proyectos de M&E activos, a los efectos de escoger aquel en el que se desean configurar las fuentes de datos. Seleccionado el proyecto de M&E, puede configurarse una métrica asociada con un atributo de la entidad bajo análisis, o bien, una métrica asociada con una propiedad contextual.

Si se desea configurar una métrica asociada con una entidad bajo análisis, se solicitan las entidades definidas y asociadas al proyecto de M&E seleccionado, para poder escoger una de ellas. Una vez que la entidad bajo análsis ha sido escogida, se indagará sobre aquellos atributos definidos para ella, y se elegirá uno. Con el atributo seleccionado, se requerirán las métricas que hayan sido definidas para el mismo, y se escogerá una. Finalmente, se asociará la métrica seleccionada con la fuente de datos vinculada al MA, sí y sólo sí la precisión requerida por la métrica es satisfecha por la fuente de datos.

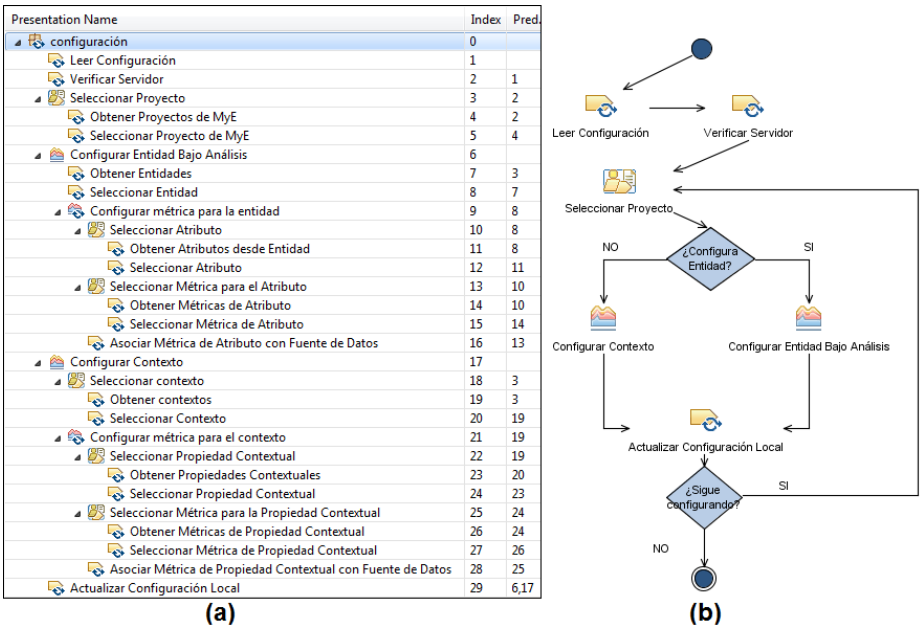

Fig. 4. Proceso de configuración: (a) WBS (b) Diagrama de actividad con notación SPEM

Por otro lado, si se desea configurar una métrica asociada con una propiedad contextual, se solicitan los contextos definidos para el proyecto de M&E seleccionado, seleccionando uno de ellos. A seguir, se obtienen las propiedades de contexto definidas para el contexto seleccionado, y se elije una de ellas. A partir de la propiedad contextual seleccionada, se requieren las métricas definidas para la misma, y se escoge una. Finalmente, se asociará la métrica seleccionada con la fuente de datos vinculada al MA, sí y sólo sí la precisión requerida por la métrica es satisfecha por la fuente de datos.

Sea que se haya definido una métrica asociada con un atributo de la entidad bajo análisis, o bien, vinculada con una propiedad contextual, la definición se actualiza

localmente, y se brinda la posibilidad de seguir configurando. En este último sentido, nótese que en caso de continuar configurando, puede seleccionar otro proyecto de M&E, lo que pone en contraste que un MA puede estar informando luego, medidas de varios proyectos de M&E en forma simultánea.

El proceso de configuración solo se lleva adelante para el start-up o inicialización del MA con respecto a sus fuentes de datos, o bien, para actualizar su configuración, luego de ello, dicho proceso no se vuelve a instanciar.

4.3 Proceso de recolección y adaptación

Una vez que las fuentes de datos asociadas con un MA han sido configuradas con respecto a uno o más proyectos de M&E a través del proceso de configuración, se conoce exactamente qué está representando cada medida (determinística o no), como así también si se asocia a un atributo de entidad, o bien, a un contexto defninido para el proyecto.

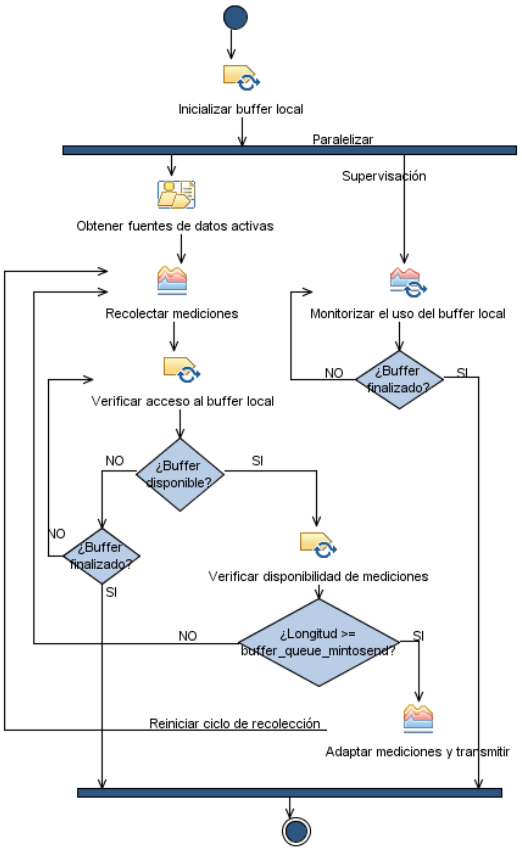

Fig. 5. Diagrama de actividad con notación SPEM, para el proceso de recolección y adaptación

De este modo, mediante el proceso de recolección y adaptación, se obtendrán las medidas a partir de las fuentes de datos, se incorporarán los metadatos en formato XML conjuntamente con las medidas utilizando C-INCAMI/MIS, y se informarán al buffer central o de reunion, de manera que pueda analizarse holísticamente junto con otras medidas que afecten al proyecto de M&E.

Cada MA contiene un buffer local, el cual va completándose con las medidas recolectadas desde las fuentes de datos, hasta su transmisión.

Como se representa en la Figura 5, una vez que el buffer local es inicializado, se paraleliza la carga de trabajo del MA, en a) la supervisación del buffer, y b) la recolección, adaptación y transmisión de medidas. La supervisación del buffer, tiene que ver con medir el estado de los recursos del mismo, liberar espacios de memoria asociados con medidas informadas, y comprimir la misma a los efectos de optimizar el uso de recursos, principalmente considerando que el MA se ejecuta generalmente en dispositivos móviles.

Por otro lado, la recolección comienza identificando las fuentes de datos activas asociadas al MA, es decir, a aquellas que responden al pedido de verificación de existencia (similar al 'ping' en las redes). A partir de allí, se le solicita a las fuentes de datos activas que informen sus medidas, incorporando las mismas dentro del buffer local, y organizándolas de acuerdo a la métrica con la que se asocien. En tal sentido, debe destacarse que gracias al proceso de configuración, se conoce la semántica vinculada con la métrica, el atributo o propiedad de contexto con el que se asocia, como así también, el proyecto de M&E en el que está definida.

Luego de que las medidas han sido incorporadas al buffer local, se verifica el acceso al mismo, es decir, si este está disponible para consulta, o bien, se encuentra en mantenimiento en dicho instante, por ejemplo, en tareas de compresión de memoria. De estar disponible, se verifica la disponibilidad de medidas por cada métrica gestionada en el MA. Si alguna métrica contiene una cantidad de medidas igual o superior al parámetro *buffer_queue_mintosend*, se comienza la adaptación y envío al buffer central (ver *Gathering Function* en Figura 2), de lo contrario, se continúan las tareas de recolección. De este modo, el objetivo del parámetro *buffer_queue_mintosend*, es establecer una valor mínimo de referencia que regule los envíos de medidas, a los efectos de que los tiempos de establecimiento de la comunicación, sean lo suficientemente pequeños con respecto al tiempo total de transmisión de las medidas. Así, cuando la cantidad de medidas es tal que se justifica su transmisión, se toma una instantánea del buffer local, se genera el flujo C-INCAMI/MIS incorporando datos y metadatos en forma conjunta, se transmiten e incorporan las mediciones al buffer central, culminado lo cual, se libera la instantánea del buffer local al MA, para incrementar la disponibilidad de recursos. El proceso de recolección y envío continuará, mientras que el buffer no se finalice, es decir, mientras sus recursos no sean desafectados por la terminación de la aplicación.

4.4 Proceso de corrección y análisis

El proceso de corrección y análisis procesa mediante ventanas temporales los datos arribados al buffer central (o de reunión), analizando estadísticamente los mismos, y contrastándolos contra su correspondiente definición formal en el proyecto de M&E.

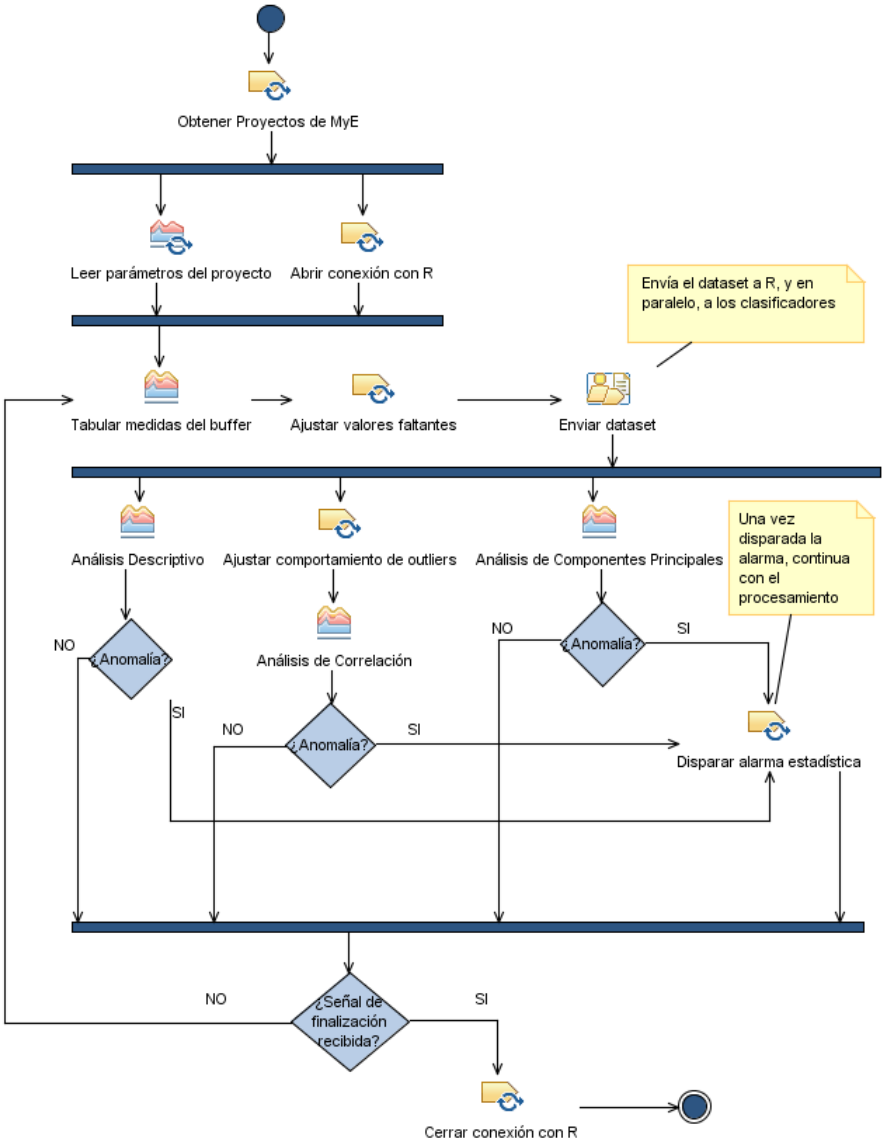

Fig. 6. Diagrama de actividad con notación SPEM, para el proceso de corrección y análisis

Inicialmente, se obtiene la definición de los proyectos de M&E relacionados con los datos en el buffer, tal como se muestra en la Figura 6. Luego, se leen los parámetros definidos para los diferentes análisis (por ejemplo, nivel de confianza en el análisis de correlación, para el cálculo del intervalo de confianza del coeficiente 'r'), y se abre la conexión con R [13]. A seguir, se toma la instantánea del buffer, expresando en forma tabular tanto métricas de atributos, como de propiedades de contexto. Se aplica luego al conjunto de datos (*dataset*), el tratamiento de valores faltantes definido en los parámetros, y se remite en paralelo a R, como así también a los clasificadores. A partir de allí, tres análisis diferentes se ejecutan en paralelo, a) el análisis descriptivo, b) el análisis de correlación, y c) el análisis de componentes principales. Si en alguno de los mencionados análisis se detecta una eventual anomalía, se dispara una alarma, la cual será retomada por el proceso de toma de decisión. Cuando los análisis culminan, de no existir señal de finalización, se vuelve a tomar otra instantánea desde el buffer central y reiterar el ciclo de análisis. Solo cuando la señal de finalización es recibida, se cierran las conexiones con R. No obstante, nótese que aún cuando se haya recibido la señal de finalización, todos los análisis deben culminar antes de proceder al cierre de la conexión con R.

Para el caso del análisis de correlación, es posible ajustar el comportamiento de los outliers, a los efectos de minimizar su incidencia en dicho análisis, ya que por ejemplo, un outlier podría generar un falso positivo en una asociación entre métricas.

4.5 Proceso de toma de decisión

Al iniciarse el proceso de toma de decisión (ver Figura 7), por un lado, se leen los datos remitidos desde la ventana de procesamiento tabular, generada en el proceso de recolección y análisis, con el objetivo de construir, actualizar y aplicar los clasificadores incrementales a las medidas de la entidad bajo análisis, incorporando un comportamiento predictivo. Por otro lado, se analiza cada alarma estadística, a los efectos de determinar su procedencia, incorporando así, un análisis detectivo en línea, basado en la semantica de la métrica.

El proceso verificará continuamente, si existen alarmas estadísticas pendientes de análisis. De existir alarmas, éstas son analizadas individualmente, dado que en la definición del proyecto de M&E, es possible que dos métricas se encuentren correlacionadas. Por ejemplo, las métricas indirectas se obtienen a partir de otras métricas directas, lo cual generará posiblemente una correlación entre ellas. Tal correlación, será descartada por el EIPFDcMM indicando que se trata de una falsa alarma, ya que la misma reside en la definición del proyecto, y no en un hecho emanado de la medición con respecto a la entidad bajo análisis. Ahora bien, cuando se corrobora que la alarma surge a partir de mediciones sobre atributos y-o propiedades contextuales, sin interferir en forma alguna las cuestiones de definición del proyecto, éstas son informadas a través de los medios que hayan sido indicados en el proyecto.

En paralelo, se leen los datos arribados mediante la ventana de procesamiento tabular, y cuando no existiere clasificador, se construye el mismo basado en el algoritmo *Adaptive-Size Hoeffding Tree* (*ASHT*) [14,15]. Luego, se verifica la integridad del clasificador a los efectos de su aplicación. Posteriormente, se paraleliza el trabajo en

dos, a saber: a) se aplica el clasificador vigente a las nuevas medidas, para obtener una clasificación en tiempo 't', y b) se actualiza el clasificador vigente con los nuevos datos, y luego se aplica el mismo para obtener una clasificación en 't+1'. A continuación, se calcula la curva ROC [12] para ambos clasificadores, determinando a través de ella, con qué clasificador se continuará. A seguir, se analizan las dos clasificaciones obtenidas, determinando si se corresponden con situaciones de riesgo. De ser así, se envía la alarma para su distribución, y se continua con el análisis. En caso contrario, simplemente se continúa con el análisis.

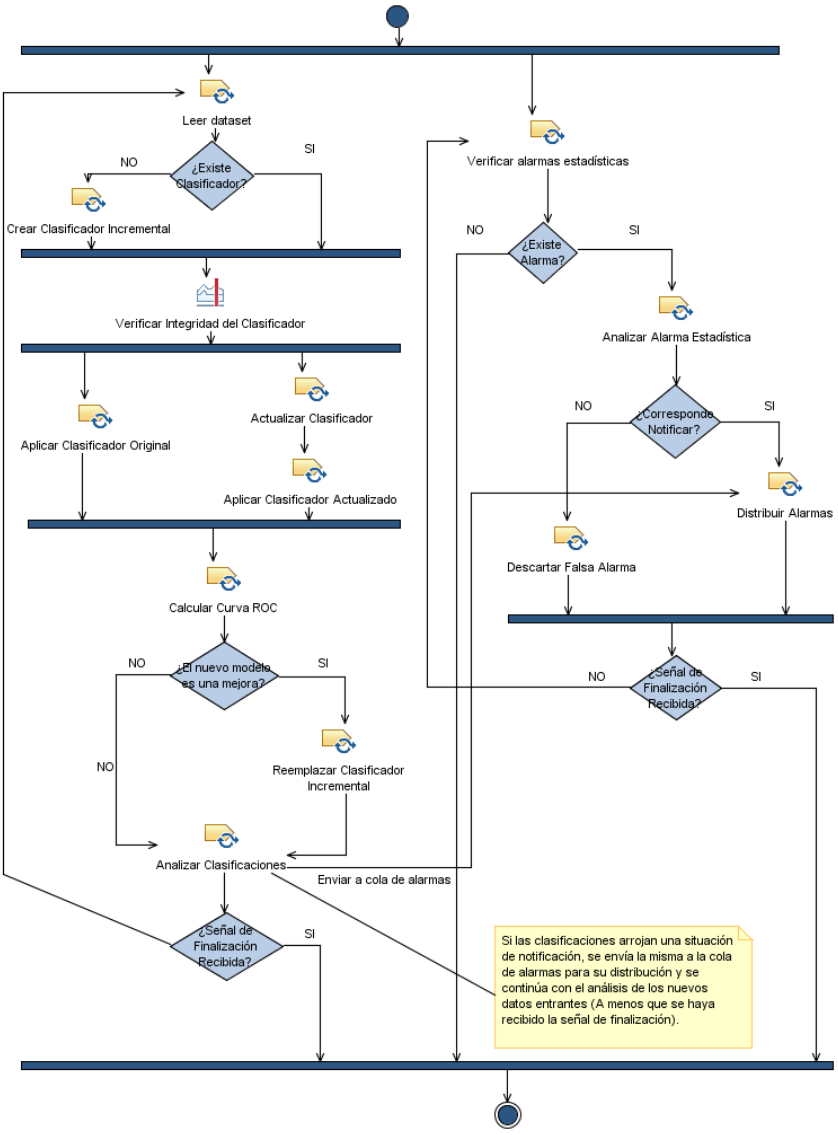

Fig. 7. Diagrama de actividad con notación SPEM, para el proceso de toma de decisión

El proceso culmina ante la recepción de la señal de finalización, no obstante, notar que la misma no puede interferir con el análisis de las alarmas, o bien, con el proceso de actualización del clasificador, solo cuando ambos culminaron, el proceso finaliza.

5 Trabajos relacionados y Discusión

Existen trabajos que enfocan el procesamiento de flujos de datos desde una óptica sintáctica, donde se permite el modelado del flujo de procesamiento y la consulta continua sobre el flujo, es realizada en términos de atributos y sus valores asociados mediante CQL (*Continuous Query Language)* [16,17]. Nuestra estrategia, incorpora la capacidad de introducir metadatos basados en un marco formal de M&E, que guían la organización de las medidas (datos) en el buffer, facilitando análisis consistentes y comparables desde el punto de vista estadístico, con la posibilidad de disparar alarmas en forma proactiva a partir de los diferentes análisis estadísticos o bien de la decisión a la que arriben los clasificadores. Adicionalmente, cuenta con los procesos formalizados mediante SPEM, lo que promueve una especificación bien establecida, comunicable y extensible.

MavStream [18] es un prototipo de sistema de gestión de flujos de datos que incorpora la capacidad de procesamiento de eventos complejos [19], como aspecto natural del procesamiento de flujos. En este sentido, nuestro prototipo soporta el análisis del flujo on-line, la generación de alarmas en forma proactiva con sustento estadístico y adicionalmente, gracias a la incorporación de los metadatos enlazados a las medidas, soporta el manejo de propiedades contextuales, procesamiento de mediciones cuyos resultados son probabilísticos y la capacidad de análisis global o por grupo de seguimiento, lo que en escenarios de uso como el de paciente trasplantado ambulatorio [2] representan aspectos cruciales.

SECRET [20] es un modelo descriptivo que permite a los usuarios analizar y comprender el comportamiento de los sistemas de procesamiento de flujos (*SPE, stream processing engines*), a partir de consultas basadas en ventanas. Este modelo, aborda la problemática sobre la diversidad semántica de procesamiento, existente entre las diferentes propuestas de SPE, sean académicas o comerciales. Nuestra estrategia se diferencia básicamente, por cuanto se focaliza a) en el procesamiento de flujos, b) incorpora metadatos a los efectos de guiar dicho procesamiento, y c) cuenta con procesos formalmente especificados usando SPEM.

6 Conclusiones y Trabajo Futuro

En el presente artículo hemos discutido los procesos asociados con la estrategia integrada de procesamiento de flujos de datos centrado en metadatos de mediciones. EIPFDcMM, permite el empleo de metadatos basados en un marco conceptual de M&E e incorporados en forma conjunta con las medidas, lo cual otorga consistencia y comparabilidad al análisis estadístico. En tal sentido, se ha discutido cómo a través de procesos formalizados mediante SPEM, es posible configurar una fuente de datos

heterogenea dentro de la estrategia de procesamiento, de qué modo las medidas son recolectadas desde las fuentes de datos y son transmitidas, qué rol juega el análisis estadístico en base a dichas medidas, y en qué orden se dan los distintos análisis. Además, se define el rol que juegan en la toma de decisión, los clasificadores y los análisis estadísticos. Esto representa una contribución, por cuanto hasta el momento, los procesos no se encontraban formalizados, y en consecuencia, se tornaba dificultoso la comunicabilidad de la estrategia desde el punto de vista del procesamiento. Para este fin, hemos empleado el lenguaje SPEM, el cual se encuentra ampliamente difundido y estandarizado.

Dado que los procesos han sido formalizados mediante el metamodelo SPEM, la estrategia de procesamiento se torna comunicable, y por consiguiente se promueve la extensibilidad de la misma.

Uno de los principales aspectos en cualquier proceso de medición, reside en la comparabilidad de sus medidas a lo largo del tiempo. En tal sentido, nuestra estrategia, *permite incrementar la confiabilidad en el procesamiento con respecto al proyecto de M&E, haciendo consistente el cómputo sobre las medidas, y promoviendo así, la interoperabilidad con respecto a las diferentes fuentes de datos y-o los destinatarios que deseen emplear tal información,* gracias a: a) encontrarse sustentada en un marco formal de M&E como C-INCAMI, b) contar con una ontología subyacente de M&E, c) guiar su procesamiento de datos en base a los metadatos asociados con la entidad bajo análisis, incluyendo también a su contexto, y finalmente, d) contar con los procesos formalizados que facilitan la comunicabilidad y extensibilidad de la estrategia. Este último aspecto, representa la principal contribución del presente artículo.

Como trabajo a futuro, se analizarán otras estrategias orientadas a complementar la acción preventiva dentro del proceso de toma de decisión. En tal sentido, la idea es orientarse a aquellas estrategias no supervisadas, que no requieren entrenamiento previo, tales como las técnicas de clustering, con el objetivo de complementar los actuales árboles de clasificación incrementales del tipo ASHT.

Reconocimientos. Esta investigación está soportada por los proyectos PICTO 2011-0277 y PAE-PICT 2188 de la Agencia de Ciencia y Tecnología, CD 066/12 y 09/F047 por la UNLPam, Argentina.

Referencias

1. Gehrke J., Balakrishan J. & Namit H., "Towards a Streaming SQL Standard," Proceedings of the VLDB Endowment, vol. 1, no. 2, pp. 1379-1390, August 2008.
2. Diván M., Olsina L., & Gordillo S., "Strategy for Data Stream Processing Based on Measurement Metadata: An Outpatient Monitoring Scenario," Journal of Software Engineering and Applications, vol. 4, no. 12, pp. 653-665, December 2011.
3. Molina H. & Olsina L. "Towards the Support of Contextual Information to a Measurement and Evaluation Framework," in *QUATIC*, IEEE CS Press, Lisboa, Portugal, pp. 154–163, 2007.
4. Olsina L., Papa F, & Molina H., "How to Measure and Evaluate Web Applications in a Consistent Way," in Ch. 13 in Web Engineering.: Springer, 2007, pp. 385–420.

5. Aref M., Bose W., Elmagarmid R., Helal A., Kamel A., Mokbel I. & Ali M., "NILE-PDT: A Phenomenon Detection and Tracking Framework for Data Stream Management Systems," in VLDB, Trondheim, Norway, 2005, pp. 1295-1298.
6. Chandrasekaran S., Cooper O., Deshpande A., Franklin M., Hellerstein J., Hong W., Madden S., Reiss F., Shah M. & Krishnamurthy S."TelegraphCQ: An Architectural Status Report," *IEEE Data Engineering Bulletin*, Vol. 26, 2003.
7. Ahmad Y., Balazinska M., Cetintemel U., Cherniack M.. Hwang J., Lindner W., Maskey A., Rasin A., Ryvkina E., Tatbul N., Xing Y., Zdonik S. & Abadi D."The Design of the Borealis Stream Processing Engine," in *Conference on Innovative Data Systems Research (CIDR)*, Asilomar, CA, pp. 277-289, 2005.
8. SPEM, "Software Process Engineering Meta-Model Specification," Object Management Group (OMG), Ver.2.0, 2008.
9. Diván, M., Olsina, L. & Gordillo, S. "Procesamiento de Flujos de Datos Enriquecidos con Metadatos de Mediciones," in *CIBSE*, 2011
10. Diván, M. "Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones," UNLP, La Plata, PhD Thesis 2011.
11. Rundensteiner W., Mani M. & Wei M. "Utility-driven Load Shedding for XML Stream Processing," in *International World Wide Web*, Beijing, China, pp. 855-864, 2008.
12. Duin R., Tortorella F. & Marrocco C. "Maximizing the area under the ROC curve by pairwise feature combination," *ACM Pattern Recognition*, pp. 1961-1974, 2008.
13. R Core Team, *"R: A Language and Environment for Statistical Computing"*. R Foundation for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing, 2012
14. Bifet A., Holmes G., Pfahringer B., Kirkby R., & Gavaldà R., "New Ensemble Methods For Evolving Data Streams," in ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). International Conference on Knowledge Discovery and Data Mining, Paris (France), 2009, pp. 139-148.
15. Bifet A., Holmes G., Kirkby R., & Pfahringer B., "MOA: Massive Online Analysis," Journal of Machine Learning Research, vol. XI, pp. 1601-1604, 2010.
16. Widom J. & Babu S. "Continuous Queries over Data Streams," *ACM SIGMOD Record*, pp. 109-120, 2001.
17. Bockermann C. & Blom H., "Processing Data Streams with The RapidMiner Streams Plugin," Technical University of Dortmund, Dortmund, Germany, Report 2012.
18. Jiang Q. & Chakravarthy S. *Stream Data Processing: A Quality of Service Perspective*. Springer, 2009.
19. Cugola G. & Margara A., "Processing flows of information: From data stream to complex event processing," Journal of ACM Computing Surveys, vol. 44, no. 3, p. Article No. 15, June 2012.
20. Botan, I., Derakhshan, R., Dindar, N., Haas, L., Miller, R. & Tatbul, N., "SECRET: a model for analysis of the execution semantics of stream processing systems," In proc. of VLDB Endowment, vol. 3, no. 1-2, pp. 232-243, September 2010.