Simulación de un Enfoque Integrado de Procesamiento de Flujos Aplicado a un Escenario de Pacientes

Mario Diván[1,2,3], Luis Olsina[2], Silvia Gordillo[4]

[1]Facultad de Ciencias Económicas y Jurídicas, UNLPam, Santa Rosa, La Pampa, Argentina
[2]Facultad de Ingeniería, UNLPam, General Pico, La Pampa, Argentina.
[3]Facultad Regional Córdoba, UTN, Córdoba, Argentina
[4] LIFIA, Facultad de Informática, UNLP, La Plata, Buenos Aires, Argentina.

{mjdivan,olsinal}@ing.unlpam.edu.ar
gordillo@lifia.info.unlp.edu.ar

Resumen. Este trabajo discute el proceso de análisis estadístico que se realiza sobre una simulación del escenario de pacientes trasplantados ambulatorios, cuyas mediciones son generadas en fuentes de datos heterogéneas y enviadas a través de flujo de datos (data streams), las cuales arriban para su procesamiento junto con los metadatos asociados a la definición formal de un proyecto de medición y evaluación. Esto permite guiar el análisis en forma consistente en busca de problemáticas típicas asociadas a los datos. Se prueba el prototipo generado para el proceso de análisis estadístico en un ambiente controlado, a los efectos de contrastar empíricamente los tiempos insumidos por el mismo y detectar las principales causas de variabilidad del sistema.

Palabras Clave: Medición, Flujo de Datos, C-INCAMI, Análisis Estadístico.

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; de tráfico aéreo; entre otras. En dichas aplicaciones, el arribo de un nuevo dato, representa la llegada de un valor (por ej: una frecuencia cardíaca, la cotización de una divisa, etc.) asociado a un comportamiento sintáctico, debido a que solo se analiza el número en sí mismo, careciendo a menudo de sustento semántico y formal no sólo con respecto a los metadatos de la medida, sino también al contexto en el que sucede el fenómeno.

Desde el punto de vista de sustento semántico y formal para la medición y evaluación (M&E), el marco conceptual C-INCAMI (*Context-Information Need*, *Concept model*, *Attribute*, *Metric* and *Indicator*) establece una ontología que incluye los conceptos y relaciones necesarios para especificar los datos y metadatos de

cualquier proyecto de M&E [2] [3]. Por otra parte, en el Enfoque Integrado de Procesamiento de Flujos de Datos (EIPFD) [4] [5] se ha planteado la necesidad de integrar los flujos de datos heterogéneos con metadatos basados en el marco C-INCAMI, permitiendo de este modo un análisis consistente de las mediciones, considerando su contexto de procedencia y su significado dentro de un proyecto de M&E. La finalidad específica, radica en promover un adecuado abordaje de la problemática típica de conjunto de datos (datasets) tales como outliers, valores faltantes, etc., que sustentarán en forma más robusta un proceso de toma de decisión.

A partir de esta estrategia de procesamiento de flujos de datos y considerando el tipo de aplicaciones mencionadas, el presente artículo plantea los fundamentos para llevar adelante una simulación sobre el escenario de pacientes trasplantados ambulatorios, resaltando el efecto de la incorporación de metadatos dentro del flujo, con respecto a la organización de las mediciones y su análisis estadístico. Adicionalmente, se realiza un análisis estadístico de los resultados surgidos de la simulación del escenario de aplicación, lo cual permite validar inicialmente nuestro prototipo. De este modo, el procesamiento on-line sobre el flujo de datos de la simulación, es realizado de un modo consistente, sustentado en información contextual y metadatos embebidos dentro de los propios flujos, lo que permitiría incrementar la robustez del análisis. Como contribuciones específicas se plantea, *(i) relacionado con métricas:* la detección de desviaciones con respecto a la definición formal, identificación de outliers y de ausencia de valor; *(ii) relacionado con el grupo de mediciones:* la detección instantánea de correlaciones, identificación de los factores de variabilidad del sistema y la detección de tendencia sobre el propio flujo de datos, considerando la situación contextual de la fuente generadora de las mediciones; *(iii) relacionado con la experimentación*: se valida inicialmente la aplicación del prototipo sobre un escenario de aplicación específico. Estas contribuciones representan un importante avance con respecto al modelo de procesamiento presentado en [4] [5], ya que ahora hay una descripción funcional de cómo llevar adelante las tareas involucradas en la integración de fuentes heterogéneas, la organización de sus mediciones, el intercambio de las mismas y un prototipo desarrollado hasta la funcionalidad de análisis estadístico que permite simular tiempos de procesamiento y respuesta.

El presente artículo se organiza en seis secciones. La sección 2 resume el marco C-INCAMI, sintetiza el EIPFD y presenta el escenario de aplicación para la simulación. La sección 3 plantea el modo en que los metadatos inciden en el análisis estadístico de los diferentes flujos de datos e ilustra la forma en que se lleva adelante dicho análisis. La sección 4 realiza la simulación del prototipo y el análisis de sus resultados. La sección 5 discute los trabajos relacionados y por último, se resumen las conclusiones y trabajos a futuro.

2. Procesamiento de Flujos de Mediciones: Fundamentos

2.1 Objetivo y Motivación

En el EIPFD [4] se plantearon las componentes de una arquitectura especializada en

la gestión de flujos de mediciones. En este sentido, la idea central del EIPFD es: automatizar los procesos de recolección permitiendo la incorporación de fuentes heterogéneas, analizar y detectar anomalías sobre los datos ante el propio arribo, y tomar decisiones on-line en base a la definición formal de un proyecto de M&E. En cuanto al análisis y detección de anomalías, este artículo propone un enfoque on-line con técnicas estadísticas como análisis descriptivo, correlación y componentes principales, las que se abordarán, conceptualmente, desde la óptica del proceso y empíricamente desde la simulación de carga de trabajo (secciones 3 y 4). En esta línea, previo analizar cualquier cuestión estadística del procesamiento de las mediciones, es necesario introducir brevemente el rol de C-INCAMI como marco formal de M&E, dar un panorama del EIPFD y presentar el escenario de aplicación sobre el que se desarrolla la simulación.

2.2 Panorama de C-INCAMI

C-INCAMI es un marco conceptual [2] [3] 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 [3]. En la Figura 1 se muestra un diagrama con los principales conceptos y relaciones para los componentes de requerimientos, contexto y medición.

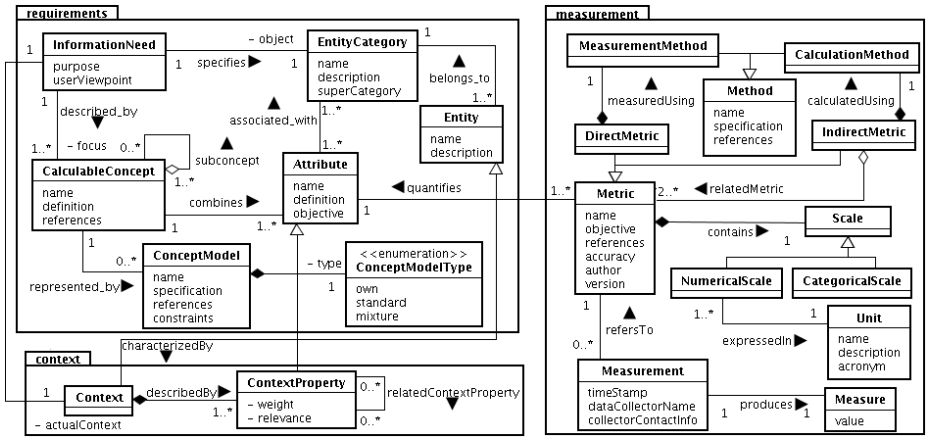


Figura 1. *Principales conceptos y relaciones de los componentes Especificación de Requerimientos no Funcionales, Especificación del Contexto y de la Medición.*

Los flujos de medidas que se informan desde las fuentes de datos al prototipo, 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 prototipo (tratado en la sección 3) 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 previo a avanzar con técnicas estadísticas más complejas.

2.3. Panorama del EIPFD

Como puede apreciarse en la Figura 2, la idea que subyace al modelo en términos de procesamiento de flujos [4] es la siguiente.

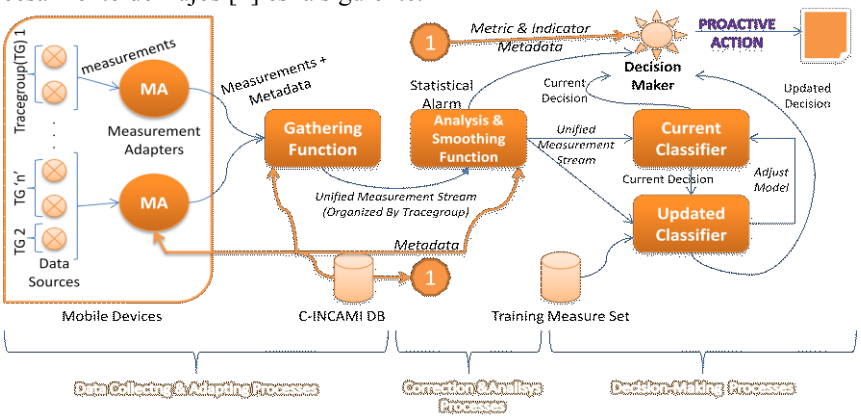


Figura 2. *Esquema Conceptual del Modelo Integrado de Procesamiento de Flujos de Datos*

Las mediciones se generan en las fuentes de datos heterogéneas, las cuales abastecen a un módulo denominado *adaptador de mediciones* (*MA en 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* [6] 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*) [7].

2.4 Un Escenario de Aplicación

El caso de aplicación que se presenta en esta sub-sección tiene por objetivo ilustrar el EIPFD centrado en Metadatos de Mediciones [8]. La idea subyacente es que los médicos del centro de salud puedan evitar reacciones adversas y daños mayores en la salud del paciente (trasplantado ambulatorio) en la medida en que puedan disponer de un seguimiento continuo del mismo. Es decir, que puedan disponer de un mecanismo por el cual les informe ante variaciones no previstas y/o inconsistencias en indicadores de salud definidos por ellos (como expertos) para un tipo de trasplante realizado en particular. En definitiva, la idea central es que exista proactivamente algún mecanismo que, basado en las métricas e indicadores de salud definidas por los especialistas para un tipo de trasplante (y potencialmente para segmentos de edades), informe sobre situaciones que pudieran afectar la salud del paciente bajo monitoreo.

Considerando C-INCAMI, la necesidad de información es "*monitorear los principales signos vitales en un paciente trasplantado al momento en que se le da el alta desde el centro médico*". La entidad bajo análisis es el *paciente trasplantado ambulatorio*. Según los expertos, la *temperatura corporal*, la *presión arterial sistólica* (máxima), la *presión arterial diastólica* (mínima) y la *frecuencia cardiaca* representan los atributos de los signos vitales relevantes a monitorear en este tipo de paciente. Además, los expertos señalan que es necesario monitorear la *temperatura ambiental*, la *presión ambiental*, la *humedad* y la *posición del paciente* (latitud y longitud) como parte de las propiedades de contexto. La necesidad de información junto con la definición de la entidad, sus atributos y contexto, forman parte de la

"*Definición y Especificación de Requerimientos no Funcionales*" y de la "*Definición del Contexto del Proyecto*" (ver sub-sección 2.2).

La cuantificación de los atributos se realiza por medio de las métricas conforme al componente *Diseño y Ejecución de la Medición* (ver Figura 1). Para el monitoreo, se desea disponer de las métricas que cuantifiquen a los atributos citados, a saber: la presión arterial sistólica, presión arterial diastólica, temperatura corporal y frecuencia cardiaca. En cuanto a las propiedades de contexto, se desea disponer de un monitoreo sobre la temperatura ambiental, la presión ambiental, la humedad y la posición del paciente (latitud y longitud).

Para el escenario actual, los expertos ya han consensuado el conjunto de métricas y propiedades contextuales a monitorear. Ahora, deben establecer los indicadores elementales, a los efectos de sentar la base para la interpretación de los atributos y conceptos calculables. De este modo, han definido los siguientes indicadores elementales: el nivel de temperatura corporal, el nivel de presión, el nivel de la frecuencia cardiaca y el nivel de diferencia de la temperatura corporal versus la temperatura ambiental. Estos indicadores integran el componente *Diseño y Ejecución de la Evaluación* (documentado en [3])

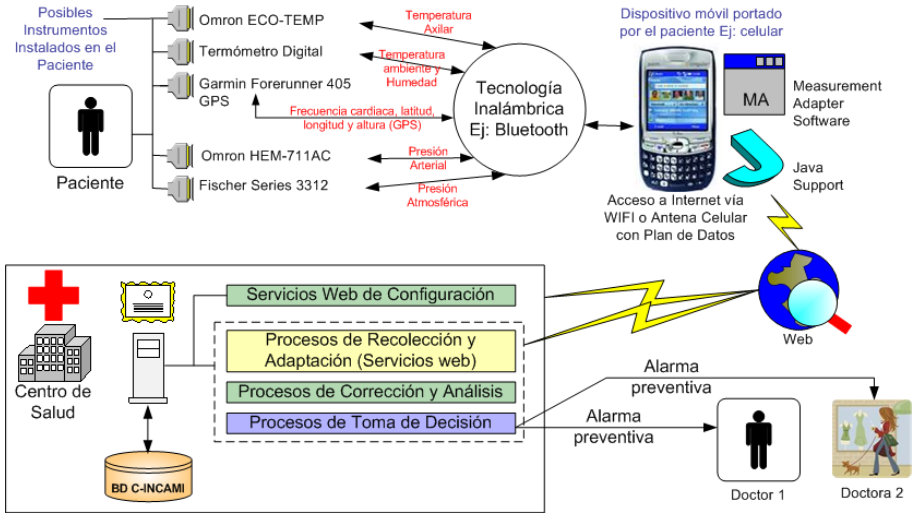


Figura 3. *Esquema de aplicación del EIPFD centrado en Metadatos de Mediciones a Pacientes Trasplantados Ambulatorios*

Una vez establecida la necesidad de información, el ente a monitorear, los atributos a medir, las propiedades contextuales, las métricas que los cuantifican y los indicadores elementales que los interpretan conforme a los criterios dados por los expertos, se estará en condiciones de instalar y configurar el MA en un dispositivo móvil –el del paciente-, el cual trabajará en forma conjunta con los sensores tal y como se expone en la figura 3.

3 Organización de las Mediciones y su Análisis Estadístico

Las mediciones vienen asociadas con metadatos que contienen su definición formal, como así también las propiedades contextuales relacionadas al ámbito del valor medido [8]. Estos metadatos permiten estructurar el buffer de mediciones (dentro de *Gathering Function en Figura 2*) en dos niveles. El primer nivel agrupa las mediciones por grupo de seguimiento (por ejemplo, cada paciente trasplantado ambulatorio), mientras que dentro de cada grupo de seguimiento, en el segundo nivel, las mediciones se estructuran por cada métrica (ver Figura 4), la cual mide un determinado atributo para una entidad dada (por ej., la presión arterial sistólica, presión arterial diastólica, temperatura corporal y frecuencia cardiaca en el paciente trasplantado ambulatorio). Así, la contribución de los metadatos a la organización, comparación y análisis de las mediciones es sustancial, por lo que mantiene agrupadas las unidades lógicas de medición a nivel de métrica sin perder relación con el grupo de seguimiento al que pertenece. A su vez, mantiene la relación de cada medición con su situación contextual (por ejemplo, la temperatura ambiental, la presión ambiental, la humedad y la posición del paciente trasplantado ambulatorio), posibilitando la comparación entre grupos de seguimiento. De este modo, los metadatos permiten guiar las técnicas de load shedding, dado que ante una situación de potencial desborde, es factible priorizar automáticamente las métricas críticas y aplicar descarte selectivamente.

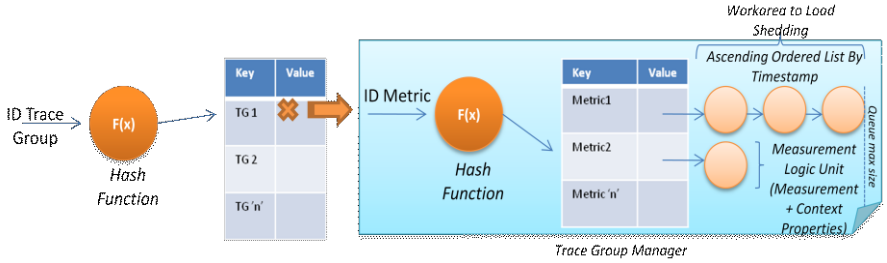


Figura 4. *Esquema Conceptual del Buffer Multinivel*

El proceso de análisis estadístico (*ASF*) toma los datos a partir del buffer organizado mediante metadatos (ver Figura 4) y aplicará análisis multivariado a nivel del grupo de seguimiento junto con análisis descriptivo sobre cada métrica de cada grupo de seguimiento, permitiendo la posterior comparación entre grupos.

A grandes rasgos el *analizador estadístico*, es una pieza de software de nuestro prototipo, responsable de llevar adelante el proceso de análisis estadístico en el EIPFD e implementar ASF. Así, ASF recorre cada uno de los grupos de seguimiento presentes en el buffer (situado dentro de *GF en Figura 2*) y aplica on-line dos técnicas multivariadas: Análisis de Componentes Principales y Análisis de Correlación [9]. El análisis de componentes principales se emplea para reducir la dimensionalidad de los problemas y el análisis de correlación, verificará si eventualmente existen correlaciones del tipo lineal entre las métricas que conforman el grupo de seguimiento. El grupo de seguimiento, puede ser entendido, por ejemplo, como un paciente trasplantado ambulatorio en particular, en donde cada métrica hace referencia a un atributo que se desea monitorear (por ej. la frecuencia cardíaca, la que

es medida e informada continuamente). Así, el análisis de componentes principales buscará identificar qué métricas (variables) incorporan mayor variabilidad al paciente (grupo de seguimiento, sistema) y el análisis de correlación intentará identificar potenciales relaciones entre las métricas monitoreadas sobre el paciente, a los efectos de detectar situaciones de arrastre, por ejemplo, la temperatura ambiental (métrica de propiedad contextual) con respecto a la frecuencia cardíaca (métrica de atributo).

Una vez que ASF culmina el análisis de un grupo de seguimiento y previo a avanzar sobre otro grupo, analiza descriptivamente cada métrica que lo compone a los efectos de detectar anomalías, ruido, outliers y/o tendencias basándose en la definición formal de la métrica o propiedad contextual, generando en paralelo sinopsis [10]. De detectarse alguna situación, ésta se informa al tomador de decisiones para que en base a lo definido en el proyecto de M&E actúe proactivamente.

4. Análisis de Resultados de la Simulación

El prototipo asociado al EIPFD implementa la funcionalidad que va desde la integración de las fuentes de datos heterogéneas, *grupos de seguimiento* y MA, hasta el ASF, incluyendo la definición formal del proyecto de medición y el repositorio de metadatos C-INCAMI (ver Figura 2). Además, implementa C-INCAMI/MIS para el intercambio de las mediciones y el buffer multinivel basado en metadatos (ver Figura 4). La implementación del prototipo se ha desarrollado en JAVA, empleando R [11] como motor de cálculo estadístico y el *CRAN* (*Comprehensive R Archive Network*) RServe para permitir el acceso TCP/IP desde la aplicación de streaming a R, sin requerir persistencia alguna y primando la comunicación directa.

La simulación se desarrolló, a partir del escenario de aplicación de la sección 2.4, generando los datos de las mediciones en forma pseudo-aleatoria, considerando dos parámetros: cantidad de métricas (cada métrica en la simulación se corresponde con una variable) y cantidad de mediciones por variable. La simulación varió en forma discreta el parámetro de la cantidad de variables en el flujo de datos de 3 a 99 y el parámetro del volumen de mediciones por variable de 100 a 1000. El prototipo, R y Rserve se ejecutaron sobre una PC con procesador AMD Athlon x2 64bits con 3GB de RAM y un sistema operativo Windows Vista Home Premium.

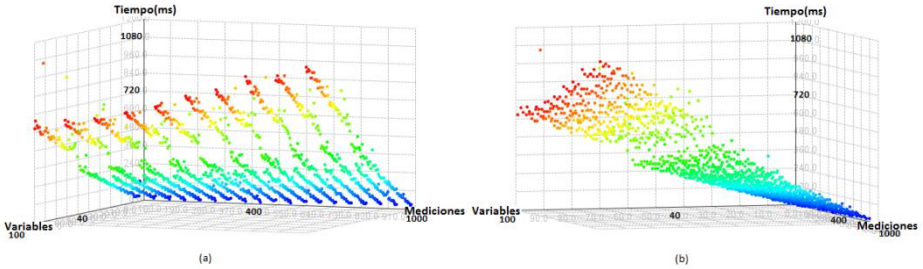


Figura 5. *Comparativa de la evolución del tiempo de procesamiento total (ms) frente a la evolución de la cantidad de variables y mediciones*

Del proceso de simulación, se han obtenido 1390 mediciones sobre los tiempos totales de procesamiento en base a la evolución de la cantidad de variables y

mediciones, lo que permite estadísticamente confluir a resultados comprobables que permitieron validar el prototipo en un contexto controlado. La gráfica de la Figura 5b muestra claramente cómo la evolución de la cantidad de variables afecta notablemente el tiempo de procesamiento total del flujo de datos, incrementándolo con respecto a lo expuesto por la gráfica (a); aquí se observa que el incremento en el tiempo de procesamiento que se produce debido al aumento de las mediciones es ínfimo en comparación al referido por las variables. Este último aspecto indica que los mecanismos de load shedding realmente consiguen su objetivo de evitar desbordes y no alterar el tiempo de procesamiento del flujo frente a la variación del volumen del mismo, mientras que la incorporación de variables influye dado que además del volumen del dato que se incorpora por la variable extra, se introduce la interacción con las variables preexistentes que es la causa y diferencia principal en términos de tiempo de procesamiento, con respecto al incremento producido por las mediciones.

Ambas partes de la gráfica de superficie (a) y (b) representan cada punto con un color el cual se asocia a la cantidad de variables. A partir de la simulación, se definieron las siguientes variables que han sido objeto de medición considerando al flujo, como la entidad bajo análisis en cada una de las 1390 mediciones indicadas:

- **Startup**: *Tiempo en ms necesario para inicializar las funciones de análisis*
- **AnDesc**: *Tiempo en ms necesario para efectuar el análisis descriptivo sobre el flujo completo*
- **Cor**: *Tiempo en ms necesario para efectuar el análisis de correlación por grupo de seguimiento dentro del flujo completo*
- **Pca**: *Tiempo en ms necesario para efectuar el análisis de componentes principales por grupo de seguimiento dentro del flujo completo*
- **Total**: *Tiempo total en ms necesario para efectuar todos los análisis sobre el flujo completo*

Los parámetros de la simulación, a los efectos del análisis estadístico de los resultados, se representan como *qVar* para indicar la cantidad de variables del flujo y *meds* para indicar la cantidad de mediciones por variable en el flujo. En adelante y a los efectos de simplificar la lectura del análisis estadístico, los parámetros qvar y meds se referenciarán directamente como variables, al igual que se lo hace con *startup*, *andesc*, *cor*, *pca* y *total*.

La matriz de correlación de Pearson expuesta en la Figura 6(a), en primer lugar confirmaría la relación lineal indicada entre la cantidad de variables (*qvar*) y el tiempo total de procesamiento del flujo (*total*) dado la presencia de un coeficiente de 0.95. En segundo lugar, se puede concluir que el tiempo total de procesamiento guardaría una fuerte relación lineal con respecto al tiempo del análisis descriptivo con un coeficiente 0.99, siguiéndole en ese orden el tiempo de *pca* con 0.9 y *cor* con 0.89. Las matrices resultantes del análisis de componentes principales expuestas en la Figura 6 b) y c), revelan cuáles de las variables aportan mayor variabilidad al sistema. En este sentido, el primer autovalor (fila 1, Figura 6(b)) explica el 66% de la variabilidad del sistema y si se observa su composición en la matriz de autovectores (col. e1, Figura 6(c)), las variables que más contribuyen en términos absolutos son *AnDesc, Cor, Pca* y *qVar*.

Coeficientes de correlación

Correlación de Pearson: Coeficientes\probabilidades

	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)

Autovalores

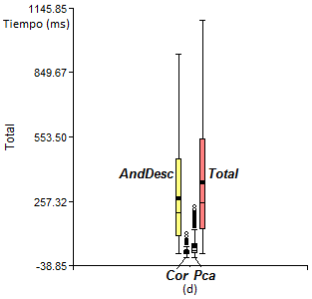
Lambda	Valor	Proporción	Prop Acum
1	4.62	0.66	0.66
2	1.11	0.16	0.82
3	0.99	0.14	0.96
4	0.14	0.02	0.98
5	0.10	0.01	1.00
6	0.02	3.3E-03	1.00
7	3.3E-05	4.7E-06	1.00

(b)

Autovectores

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)



(d)

Figura 6. (a) *Matriz de Correlación de Pearson,* **(b)** *Matriz de Autovalores y* **(c)** *Matriz de Autovectores asociadas al Análisis de Componentes Principales (PCA) y* **(d)** *Boxplot de las variables AnDesc, Cor, PCA y Total*

De este modo, si se deseara reemplazar las siete variables enunciadas por las nuevas tres variables (e1 a e3), se estaría explicando el 96% de la variabilidad del sistema, donde las principales variables en término de aporte están asociadas con AnDesc, Cor, Pca y qVar. El sistema sólo es afectado en un 16% por la evolución de las mediciones y en un 14% por el tiempo de inicialización, lo que implica un punto importante de resaltar dado que la única variable externa al prototipo que no puede ser controlada por el mismo, el volumen de arribo de las mediciones, sólo representa un 16% y en ningún caso representó una situación de desborde en la cola de servicios.

De las cuatro variables que más variabilidad aportan al sistema, tres de ellas componen parte del tiempo total de procesamiento por lo que mediante el box plot de la Figura 6, puede corroborarse que la variable más influyente a la magnitud del tiempo total de procesamiento es *AnDesc*. Adicionalmente, debe destacarse que el mayor tiempo obtenido para procesar 99 variables con 1000 mediciones (99000 mediciones en total por flujo) fue de 1092 ms, es decir 1,092 segundos, sobre un hardware básico y totalmente accesible en el mercado, lo que permite establecer un umbral de aplicabilidad del prototipo, que satisface holgadamente los requisitos en cuanto a tiempo de respuesta en el escenario de paciente trasplantado ambulatorio.

5. Discusión

Existen trabajos que enfocan el procesamiento de flujos de datos desde una óptica sintáctica, donde la consulta continua sobre el flujo es realizada en términos de atributos y sus valores asociados mediante CQL [12]. Este enfoque ha sido implementado por proyectos tales como Aurora & Borealis [13], STREAM [14] y TelegraphCQ [15]. Nuestro prototipo incorpora la capacidad de introducir metadatos basados en un marco formal de M&E, que guían la organización de las mediciones en el buffer, posibilitando 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. MavStream [10] es un prototipo de sistema de gestión de flujos de datos que

incorpora la capacidad de procesamiento de eventos complejos 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 en forma conjunta a las mediciones, soporta el manejo de propiedades contextuales a la medición, procesamiento de mediciones cuyos resultados son probabilistas y la capacidad de análisis global o por grupo de seguimiento, lo que en escenarios de uso como el de paciente trasplantado ambulatorio [4] representan aspectos cruciales.

Nile [16] es un sistema de gestión de flujos de datos basado en un maco conceptual para detección y seguimiento de fenómenos sustentado en medidas deterministas. Nuestro prototipo, a diferencia de Nile, admite incorporar fuentes de datos heterogéneas, las cuales mediante el MA (Figura 2) introducen metadatos específicos del proyecto de medición sobre el flujo de medidas, lo que permite llevar adelante un análisis estadístico consistente, considerando medidas no sólo deterministas sino también probabilistas y con la capacidad de actuar proactivamente frente a la detección de tendencias, inconsistencias de los datos con respecto a la definición de su métrica, entre otras causas surgidas del análisis estadístico.

6. Conclusiones y Trabajo Futuro

El artículo ha discutido cómo la presencia de los metadatos basados en un marco formal de M&E e incorporados en forma conjunta con las mediciones, permiten una organización de las mismas que incorpora consistencia en el análisis estadístico, de modo que identifica las componentes formales de los datos y su contexto asociado. De este modo, es factible realizar análisis particulares a nivel de grupo de seguimiento o bien a nivel general, comparando métricas entre diferentes grupos de seguimiento, con el objeto de identificar desviaciones de las medidas con respecto a su definición formal, los principales factores de variabilidad del sistema como así también la interacción lineal entre variables.

Se ha probado a partir del análisis estadístico de los resultados de la simulación, que el prototipo que implementa el EIPFD es más susceptible al incremento de la cantidad de variables que al incremento de la cantidad de mediciones por variable en términos de tiempo de procesamiento. Mediante el Análisis de Componentes Principales se ha comprobado que las componentes de ASF que más aportan a la variabilidad del sistema, son las asociadas a las variables *andesc*, *cor*, *pca* y *qvar*, siendo *andesc* quien define la mayor proporción del tiempo final de procesamiento del flujo de datos. Considerando un entorno de prueba implementado mediante un hardware accesible en el mercado, se pudo establecer como patrón de comparación a los efectos de poder evaluar consistentemente los ámbitos de aplicación, que para procesar 99000 mediciones (99 variables y 1000 mediciones/variable) el tiempo máximo arrojado ha sido 1092ms. Como corolario del análisis estadístico, ha podido comprobarse la efectividad de los mecanismos de load shedding dentro del buffer multinivel, dado que la evolución en la cantidad de mediciones no ha comprometido el funcionamiento del prototipo ni tampoco afectado en gran proporción al tiempo final de procesamiento del flujo de mediciones.

Como trabajo a futuro, se pretende someter experimentalmente a nuestra estrategia de procesamiento a diferentes escenarios, a los efectos de validar estadísticamente los resultados iniciales obtenidos para el escenario del paciente trasplantado ambulatorio.

Reconocimientos. Esta investigación está soportada por los proyectos PICT 2188 de la Agencia de Ciencia y Tecnología y 09/F052 por la UNLPam, Argentina.

Referencias

[1] Gehrke J., Balakrishan H., Namit J. "Towards a Streaming SQL Standard" in *VLDB*, Auckland, New Zealand, 2008.

[2] 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.

[3] Olsina L., Papa F., Molina H. "How to Measure and Evaluate Web Applications in a Consistent Way," in *Ch. 13 in Web Engineering* Springer Book HCIS, 2008, pp. 385–420.

[4] Diván, M., Olsina, L. "Enfoque Integrado para el Procesamiento de Flujos de Datos: Un Escenario de Uso," in *CIBSE*, pp. 374-387, 2009

[5] Diván, M., Olsina, L., Gordillo, S. "Procesamiento de Flujos de Datos Enriquecidos con Metadatos de Mediciones," in *CIBSE*, 2011

[6] 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.

[7] Duin R., Tortorella F., Marrocco C. "Maximizing the area under the ROC curve by pairwise feature combination," *ACM Pattern Recognition*, pp. 1961-1974, 2008.

[8] Diván M., Olsina L. "Especificando Fuentes de Datos en el Esquema Integrado de Procesamiento de Flujos," in *CACIC*, San Salvador de Jujuy, Argentina, 2009.

[9] Johnson, D. *Métodos Multivariados Aplicados al Análisis de datos*. México: Thomson Editores, 2000.

[10] Jiang Q., Chakravarthy S. *Stream Data Processing: A Quality of Service Perspective*. Springer, 2009.

[11] R Software Foundation, *R Software*. Vienna, Austria: The R Foundation for Statistical Computing, 2010.

[12] Widom J., Babu S. "Continuous Queries over Data Streams," *ACM SIGMOD Record*, pp. 109-120, 2001.

[13] 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.

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

[15] 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.

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