Arquitectura de Procesamiento centrada en Metadatos de Mediciones

Mario Diván, María de los Ángeles Martín

Facultad de Ingeniería, UNLPam, General Pico, La Pampa, Argentina.

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

Abstract. El enfoque integrado de procesamiento de flujos de datos es un gestor de flujos de datos sustentado en un marco de medición y evaluación, que incorpora comportamiento detectivo y predictivo mediante el empleo de las mediciones y metadatos asociados. Aquí presentamos la Arquitectura de Procesamiento centrada en Metadatos de Mediciones que evoluciona el esquema de procesamiento, extendiéndolo al ámbito de los repositorios Big Data, e incorporando servicios por suscripción a partir de las fuentes de datos. Adicionalmente, una Memoria Organizacional permite guiar el entrenamiento de los clasificadores como así también el proceso de toma de decisión en base al conocimiento previo documentado. Finalmente, un caso de aplicación sobre el RADAR de la Estación Experimental Anguil es presentado.

Keywords: Arquitectura, Medición, Grandes Datos, Flujo de Datos

1 Introducción

Actualmente, existen arquitecturas de procesamiento que permiten procesar datos generados en tiempo real, mediante topologías de procesamiento configurables tales como Apache Storm y Spark [1, 2]. Estas arquitecturas pueden definir dinámicamente la topología de procesamiento sobre los flujos de datos, ajustándose a diferentes necesidades de cómputo, y delegando la definición estructural y significado del dato en la lógica embebida dentro de la aplicación. En este tipo de aplicaciones se incorpora el Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones (*EIPFDcMM*) [3], el cual sustentado en el marco de medición y evaluación C-INCAMI (*Context-Information Need, Concept model, Attribute, Metric and Indicator*) [4, 5], 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*), 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 [1, 2, 6, 7], gracias a la incorporación de metadatos mediante C-INCAMI/MIS (*Measurement Interchange Schema*) [8], el EIPFDcMM es capaz de guiar el procesamiento de las medidas, analizando cada una en base al significado definido en el proyecto de M&E y dentro de su contexto de procedencia. Adicionalmente, ello permite incorporar un comportamiento detectivo y predictivo sobre las medidas contextualizadas, lo que posibilita un monitoreo activo sobre las entidades bajo análisis sustentado en una memoria organizacional [9] capaz de gestionar experiencias previas y soportar el proceso de toma de decisiones.

La Arquitectura de Procesamiento centrada en Metadatos de Mediciones (APcMM), se apoya en la definición de procesos en SPEM *(Software & Systems Process Engineering Metamodel)* [10] del EIPFDcMM para garantizar su comunicabilidad y extensibilidad [11, 12], y evoluciona la estrategia para soportar adicionalmente al procesamiento de flujos, su almacenamiento y gestión en repositorios de grandes datos en entornos de computación distribuida. Ello implica un nuevo esquema de procesamiento que requiere compatibilizar los contextos de flujos de datos y el de grandes datos, muy diferentes en términos de cómputo y utilización de recursos [13].

De este modo, la APcMM capitaliza la madurez y escalabilidad de tecnologías de cómputo distribuido y grandes datos tales como Apache Storm [1, 14], Apache Kafka [15, 16], Apache Hadoop [17, 18], Apache HBase [19, 20], entre otras, para montar sobre ellas la estrategia de procesamiento guiada por metadatos de mediciones, sustentada en los procesos formalmente definidos en [11, 12]. Así, el abordaje de una medida en APcMM, no solo se acota al arribo de un valor sintáctico, sino que la medida arriba acompañada por sus metadatos y los atributos que cuantifican su contexto de procedencia. Ello permite guiar el procesamiento, almacenamiento y provisión de la medida, a partir de la interpretación de sus respectivos significados dentro del proyecto de M&E.

Así, y como contribuciones específicas se plantea, *(i) relacionado con la gestión de medidas:* la posibilidad de homogenizar las medidas mediante C-INCAMI/MIS a nivel de procesamiento, almacenamiento y para la provisión de medidas a terceros, a través de mecanismos por suscripción, lo que permite mejorar la interoperabilidad del sistema tanto a nivel de flujos como de grandes datos*, ii) relacionado con la toma de decisiones:* ahora es posible gestionar la Memoria Organizacional sobre grandes repositorios, lo que aporta mayor experiencia y volumen para el entrenamiento de los clasificadores, a la vez que los conocimientos previos permitirán recomendar eventuales cursos de acción al proceso de toma de decisión*, y iii) relacionado con la arquitectura:* nuestra estrategia ahora se monta sobre tecnologías maduras de cómputo y almacenamiento distribuido [14, 15, 17, 19], lo que posibilita incorporar la capacidad de gestión de grandes volúmenes de datos, escalabilidad, junto con la posibilidad de procesar datos con altas tasas de arribo, y aprovisionar grandes volúmenes de datos sobre demanda a partir de estrategias como MapReduce.

El artículo se organiza en seis secciones. La sección 2 resume el marco C-INCAMI y el esquema C-INCAMI/MIS. La sección 3 sintetiza la Memoria Organizacional empleada por la estrategia. La sección 4 plantea la nueva Arquitectura, su idea conceptual y la tecnología subyacente que permite el cómputo distribuido y su escalabili-

dad. La sección 5 presenta el caso de aplicación en el Radar de la Estación Experimental Agropecuaria (EEA) del INTA Anguil. La sección 6 discute los trabajos relacionados, y por último, se resumen las conclusiones y trabajos a futuro.

2 Panorama de C-INCAMI y C-INCAMI/MIS

C-INCAMI es un marco conceptual [4, 5] 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 [5].

Los flujos de medidas que se informan desde las fuentes de datos al APcMM, 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, si la medida es determinista o no con su probabilidad, entre otros.

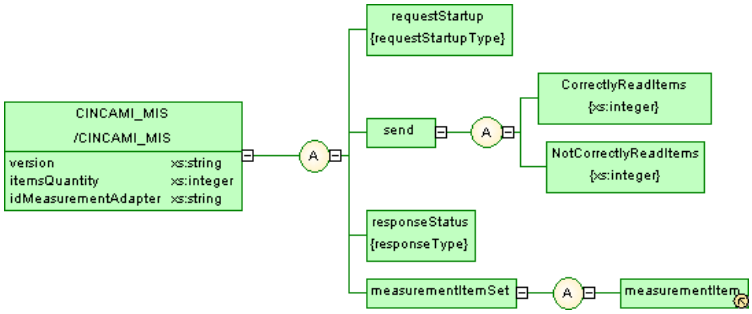

Fig. 1. Nivel Superior del Esquema C-INCAMI/MIS

En tal sentido, C-INCAMI/MIS (Ver Figura 1) es el esquema de intercambio de mediciones que permite dentro de un mismo flujo de datos etiquetar conjuntamente con cada medida asociada al atributo, las medidas vinculadas a cada propiedad de contexto. El conjunto de mediciones del flujo se organiza bajo la etiqueta denominada *measurementItemSet* de la Figura 1, e identificando bajo cada etiqueta *measurementItem,* a una medida con sus respectivas propiedades de contexto. Esto representa un aspecto importante en APcMM para la gestión de mediciones, ya que al disponer de fuentes heterogéneas de datos, es posible homogenizar las mediciones bajo un mismo esquema independientemente del origen mediante el *Adaptador de Mediciones* (Ver sección 4). Así, tanto el almacenamiento, como el procesamiento y los servicios a terceros, gestionarán siempre flujos C-INCAMI/MIS sin importar la fuente que los haya generado, lo que facilita su consulta, intercambio y extensibilidad.

3 Memoria Organizacional

Una vez que los flujos C-INCAMI/MIS son incorporados desde las fuentes de datos al repositorio persistente de grandes datos, es conveniente estructurar los mismos en una memoria organizacional, de manera que posteriormente pueda ser explotada y utilizada para la recomendación durante el proceso de toma de decisión.

El conocimiento aporta ventaja estratégica en materia de competitividad empresarial, en tal sentido, los sistemas de administración del conocimiento permiten administrar y almacenar el conocimiento organizacional, con el objetivo de ser utilizado para aprender, resolver problemas y como apoyo a la toma de decisiones [21]. Nuestra propuesta, es almacenar el conocimiento aportado por los flujos de datos y sus metadatos, en forma estructurada bajo una Memoria Organizacional Basada en Casos [22].

Un caso es una pieza contextualizada de conocimiento que representa una experiencia o hecho. Típicamente, un caso comprende: *a)El problema*: describe el estado del mundo cuando ocurrió el caso, y *b) La solución*: describe cómo se resolvió el problema, qué curso de acción se tomó y los resultados logrados o esperados.

El proceso de razonamiento basado en casos consiste en asignar valores a las variables características del problema, y encontrar a partir de casos históricos similares ocurridos en el mismo contexto, los valores adecuados para las instancias de la solución, a través de criterios de similitud de casos teniendo en cuenta los metadatos.

El hecho de contar con un repositorio de grandes volúmenes con mediciones y metadatos asociados, fomenta un rápido aprendizaje y calidad en las recomendaciones de la memoria organizacional. Un mayor detalle de las prestaciones de la Memoria Organizacional puede encontrarse en [10, 11], donde los procesos han sido formalizados mediante el metamodelo SPEM para promover su comunicabilidad y extensibilidad.

4 Arquitectura de Procesamiento centrada en Metadatos de Medición

La APcMM [11, 12, 23] es una estrategia de procesamiento de flujos de datos especializada en proyectos M&E y sustentada en C-INCAMI [4, 5], la cual incorpora comportamiento detectivo y predictivo en línea, a la vez que permite el aprovisionamiento a terceros de los flujos mediante suscripción, y el almacenamiento de las medidas en grandes repositorios para responder consultas de datos ad-hoc.

Sintéticamente y como puede apreciarse en la figura 2, la idea conceptual de procesamiento consiste en que los flujos de medidas provienen desde fluentes de datos heterogéneas (por ejemplo, un radar) estructurados bajo el esquema C-INCAMI/MIS. Cada flujo C-INCAMI/MIS es generado a partir de la fuente de datos por un adaptador de mediciones (MA en figura 2) que establece la correspondencia entre la medida, sus metadatos y las propiedades de contexto en base al proyecto de M&E definido. Así, cada flujo C-INCAMI/MIS es enviado desde el MA a la función de reunión informando las medidas, sus propiedades de contexto y sus metadatos. De este modo, la función de reunión: a) incorpora el flujo en el repositorio de grandes datos, b) provee el flujo en tiempo real a los terceros suscriptos al servicio, y c) provee una copia del

flujo a la función de análisis y suavización. Esta última, realiza diversos análisis estadísticos (por ejemplo, análisis de correlación) sobre las métricas del flujo, permitiendo almacenar una instantánea de la situación de la entidad bajo análisis en memoria, disparar alarmas en caso de desvíos respecto de los establecido en el proyecto de M&E al tomador de decisiones (Decision Maker –DM- en Figura 2), y suavizar el flujo en base a la configuración del proyecto de M&E (por ejemplo, filtrar valores atípicos). De este modo, los flujos suavizados se informan al clasificador actual, quien: *a)* Toma una decisión al instante (D^t), y *b)* En paralelo, se actualiza incrementalmente, generando un nuevo clasificador actualizado, y toma una nueva decisión (D^{t+1}). Si algunas de las decisiones D^t o D^{t+1} se corresponden con una situación de eventual riesgo según lo definido en el proyecto de M&E, se dispara una alarma al DM. Ambos modelos, el clasificador actual y el actualizado, son comparados en línea contrastando su área bajo la curva ROC (acrónimo de *Receiver Operating Characteristic*) [24], y aquel que mayor área bajo la curva posea se tornará en el nuevo clasificador actual. De este modo, el clasificador no sólo aprende desde el conjunto de entrenamiento dado por la memoria organizacional del repositorio en el momento cero, sino que va ajustando incrementalmente su comportamiento a partir de los datos tratados estadísticamente en línea para ajustarse a nuevas situaciones, retroalimentado a su vez a la memoria organizacional. Un mayor detalle puede encontrarse en [11, 12], donde los procesos han sido formalizados mediante el metamodelo SPEM para promover su comunicabilidad y extensibilidad.

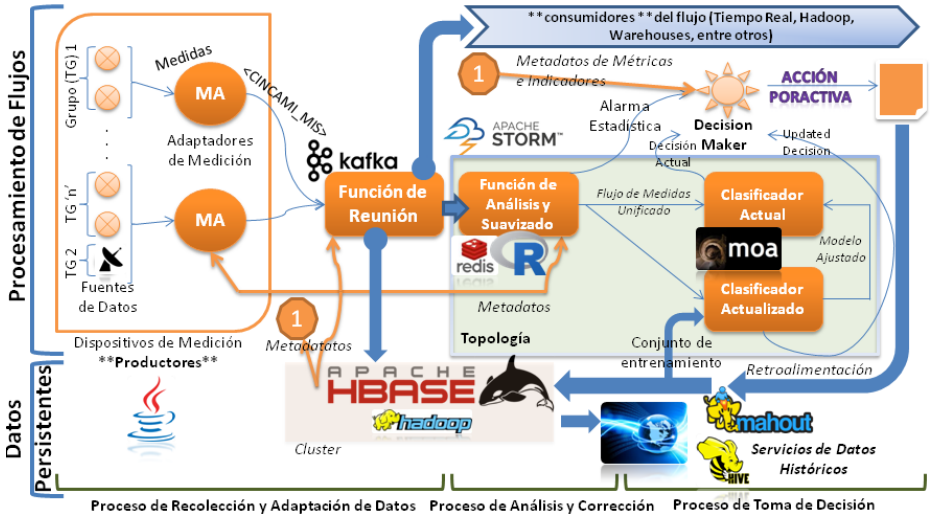

Fig. 2. Arquitectura de Procesamiento centrada en Metadatos de Mediciones con Big Data

Desde el punto de vista técnico, la APcMM evoluciona la solución propuesta en el EIPFDcMM, ya que ahora es posible la gestión de repositorios de grandes datos en entornos de computación distribuida junto la provisión de datos mediante servicios por suscripción, en forma adicional al procesamiento de flujos original. Así y a los efectos de promover la extensibilidad, dinamismo y difusión de la arquitectura, se ha priorizado el empleo de tecnologías de código abierto, maduras y escalables.

De este modo, las fuentes continuan implementando la interface *DataSource* original, a través de la cual se definen las responsabilidades que una fuente del EIPFDcMM debe satisfacer para aprovisionar datos a la arquitectura, pero ahora en APcMM se constituyen adicionalmente en *productoras* en términos de Apache Kafka [15] como puede apreciarse en la Figura 2. Así, los datos enviados por los productores (por ejemplo, un radar), serán procesados por un servicio de suscripción dentro del cluster de procesamiento de mensajería, bajo el concepto de función de reunión, donde lógicamente todas las medidas de la misma entidad bajo análisis, son agrupadas para informarse en forma conjunta mediante esta misma tecnología a los consumidores. Así, los consumidores podrán procesar en tiempo real el flujo C-INCAMI/MIS reunido a partir de Kafka, entendiéndose por tales a: 1) Los suscriptores que consumen en tiempo real el flujo de medidas a partir de Apache Kafka, 2) La topología de procesamiento de flujos de datos sustentada en Apache Storm [14] que continuará con el procesamiento en tiempo real, y 3) Apache HBase [19, 20] como repositorio de grandes datos que almacena las medidas para su uso posterior.

Así, la estrategia interna de procesamiento de flujos de datos se monta ahora sobre Apache Storm [14], y consume los flujos en forma continua desde la función de reunión a partir de Apache Kafka, como puede apreciarse en el recuadro de la Figura 2. Esto último, aporta flexibilidad, escalabilidad y dinamismo respecto de la configuración de las *topologías*[1] de procesamiento de datos, ya que tanto la función de suavización como los clasificadores se corresponden con *Bolts*1 que pueden ser reorganizados en forma ágil y simple dentro de la misma. Adicionalmente, dentro de la topología de procesamiento ejecutada sobre Storm, APcMM continua empleando R [25] para los cómputos estadísticos de la función de análisis y suavizado, empleando a partir de ahora Redis [26] como base de datos NoSQL en memoria para: *i)* Gestión de cache, *ii)* La utilización de resultados intermedios desde R, y *iii)* El almacenamiento de las instantáneas sobre el ultimo estado conocido de cada entidad bajo análisis. Finalmente, se utiliza dentro de la topología de procesamiento los clasificadores del marco Massive Online Analysis (MOA) [27] que permiten actualizaciones incrementales a la vez que están nativamente preparados para minería de flujos.

Por otro lado y en relación a la gestión persistente de medidas, la Arquitectura emplea para el almacenamiento y procesamiento de grandes datos un cluster Apache Hadoop [17] para promover el procesamiento distribuido, con una base de datos columnar Apache HBase [19, 20] que permite el escalamiento monolítico y el acceso aleatorio a las mediciones asociadas con un Proyecto de M&E dado. A partir de este repositorio de medidas provenientes desde diferentes orígenes, se emplea Apache Hive [28] para soportar consultas ad-hoc sobre entornos de cómputo distribuido, al igual que Apache Mahout [29] para poder llevar adelante diferentes análisis de agrupamiento y clasificación que permitan detectar nuevos patrones de comportamientos respecto del objetivo del Proyecto de M&E. Esto posibilita que la Memoria Organiza-

[1] Se entiende por *Topología* en Apache Storm a un conjunto de fuentes de datos (denominadas Spout) que proveen datos a uno o más componentes vinculados (denominados Bolt), a partir de los cuales se realiza alguna síntesis, transformación o disgregación del flujo de datos original a los efectos de ser consumido por un usuario final, o bien, constituir la entrada de uno o más componentes (otros Bolts) [1].

cional tome ventajas respecto de las capacidades de paralelismo, distribución y escalabilidad que ofrece este nuevo contexto de procesamiento, ya que el motor de razonamiento basado en casos de la memoria organizacional, utiliza programas tipo MapReduce [18] en su sistema de recomendación, para estructurar naturalmente cada conjunto (hechos, solución) como <clave, valor>.

Así, la arquitectura técnicamente no solo se orienta al cómputo distribuido, su escalabilidad y extensibilidad en base a productos maduros con el objetivo de poder enfrentar repositorios de grandes datos, sino que también ahora se monta sobre Apache Storm, para dotar a la topología de procesamiento de *dinamismo*, facilitando su versionado e interoperabilidad ante eventuales cambios de requerimientos.

5 Caso de Aplicación: RADAR de la EEA INTA Anguil

La Estación Experimental Agropecuaria (EEA) INTA Anguil tiene instalado un Radar Meteorologico (RM) marca Gematronik modelo Meteor 600C (Ver figura 3.a) que genera un flujo de datos estimado de 17gb diarios, lo que representa un desafío para su almacenamiento, gestión y posterior servicio al público, principalmente considerando la importancia que los datos poseen para la región productiva de influencia.

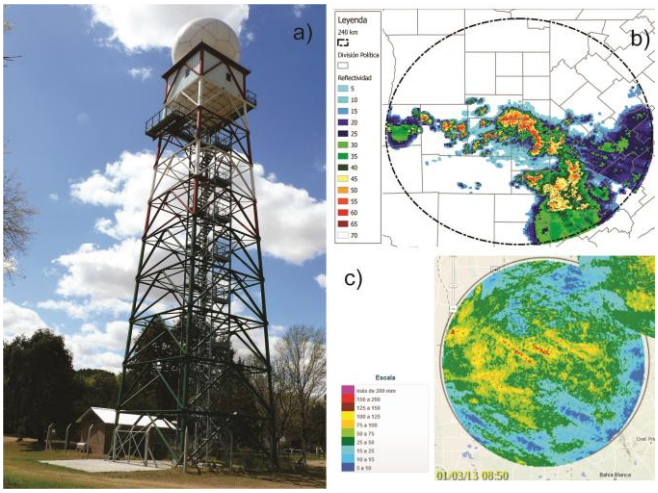

Fig. 3. a) Infraestructura del RM instalado en la EEA Anguil, b) Imágen de reflectividad de la primera elevación (0,5°) del 15-01-2011, 23:40hs generado con Software de INTA, c) PAC (*Precipitation Accumulation*) de Febrero de 2013 generado con Rainbow 5 de Gematronic

El radar posee sistema doppler y es de doble polarización (DP). Opera en banda C a una frecuencia de 5,64 Ghz y longitud de onda de 5,4 cm [30]. La antena permite un giro en el sentido horizontal (azimut) y puede elevarse en ángulo vertical hasta 45º. Este RM está configurado para completar una serie de giros a 360º que se repite para 12 ángulos de elevación, entre 0,5º de base y 15,1º de tope, en rangos de 120 km, 240 km y 480 km [30], un ejemplo para la primera elevación puede verse en la figura 3(b).

La frecuencia de un escaneo completo está programada cada 10 minutos, totalizando 144 adquisiciones diarias normalmente. Cada adquisición se realiza en forma volumétrica, con una unidad de muestreo de 1km^2 y1°, almacenando *hoy* cada variable, o cómputo derivado, en archivos separados denominados volúmenes. Dentro de las variables que permite recolectar el RM se encuentra el factor de reflectividad (Z), la reflectividad diferencial (ZDR), el coeficiente de correlación polarimétrica (RhoHV), el desplazamiento de fase diferencial (PhiDP), el desplazamiento de fase diferencial específica (KDP), la velocidad radial (V) y la anchura del espectro (W) [23, 31].

En tal sentido debe considerarse que tan solo un RM produce alrededor de 17 GB diarios, lo que arroja aproximadamente un volumen de 6,2 TB por año. Este volumen de datos representa un desafío tanto para su almacenamiento como para su procesamiento y aprovisionamiento en línea a terceros. A la fecha, ante un requerimiento en particular a la EEA de INTA Anguil, los datos deben ser procesados artesanalmente y ad-hoc a partir de los archivos físicos del RM por personal especializado, lo que conlleva en forma implícita a demoras en la respuesta y riesgos propios derivados del procesamiento manual.

En [12, 23] se plantearon los ajustes necesarios sobre la estrategia de procesamiento, a los efectos de incorporar la posibilidad de brindar servicios a terceros y el aprovisionamiento ad-hoc de grándes volúmenes, pero aún no se había definido qué tecnología utilizar para lograr escalabilidad, dinamismo, paralelismo y cómputo distribuido a partir del prototipo original documentado en [8, 22].

De este modo, cada RM es una fuente de datos heterogénea que informará el flujo de medidas mediante C-INCAMI/MIS. Para ello, una pequeña aplicación a instalar en el RM implementa el adaptador de mediciones de APcMM y el Productor de Apache Kafka para un tópico particular, leerá directamente desde el buffer de generación en memoria del RM, informando en tiempo real el flujo. La función de reunión de APcMM actuará como consumidor del flujo C-INCAMI/MIS, efectuando en paralelo: a) La replicación del flujo a terceros mediante suscripción, b) el volcado masivo del flujo en la tabla de medidas específica para el proyecto de M&E dentro del repositorio HBase, y c) La replicación del flujo reunido a la función de análisis y suavización. Éste último, en términos de Apache Storm, sería un *Spout*1 de la topología de procesamiento, cuya estructura queda definida por C-INCAMI/MIS. Luego, tanto el tomador de decisiones, como la función de análisis y suavización, los clasificadores, el consumo y la retroalimentación de la memoria organizacional son *Bolts*1 de la topología Apache Storm, que pueden ser ajustados ante eventuales cambios de los requerimientos de procesamiento, versionándose las topologías e incluso permitiendo ejecutarlas en paralelo y en forma independiente una de otra [1].

Por otro lado, la consulta de medidas y-o de la memoria organizacional almacenada en Apache HBase para los diferentes proyectos de M&E, será provista como servicio por suscripción a partir de Apache Hive. Esto permite, por un lado, el intercambio de datos sin intervención humana, lo que promueve su comunicabilidad y extensibilidad; y por otro, posibilita a los desarrolladores las consultas ad-hoc mediante Hive-QL (acrónimo de *Query Language*) [28], un lenguaje fácil de aprender ya que posee una estructura similar al tradicional SQL (acrónimo de *Structured Query Language*). De este modo y a partir de tales repositorios, se posibilitarán diferentes análisis de patro-

nes ad-hoc sobre todas las mediciones históricas de los proyectos de INTA EEA Anguil empleando Apache Mahout [29], lo que permitirá no solo retroalimentar a la APcMM, sino también ajustar las definiciones de los proyectos de M&E, como por ejemplo, lo referido a la emisión de alarmas y/o recomendaciones en vivo.

6 Trabajos relacionados y Discusión

Existen trabajos recientes que enfocan el procesamiento de flujos de datos desde una óptica sintáctica, donde el modelo de datos del flujo se basa en una estructura clave-valor e incorporan técnicas para la gestión adaptativa de tasas de arribo, para poder abordar el tratamiento de volúmenes de datos explosivos en los flujos [32]. Nuestra arquitectura incorpora la capacidad de introducir metadatos basados en un marco formal de M&E, que guian la organización de las medidas (por ejemplo, mediante instantáneas en memoria y el último estado conocido de la entidad bajo análisis), facilitando análisis consistentes y comparables desde el punto de vista estadístico, con la posibilidad de disparar alarmas basada en la interpretación de los criterios de decisión de los indicadores que se obtienen a partir de los datos. No obstante, Lee y otros [32] plantean una interesante limitación de Apache Storm para abordar los flujos de datos explosivos, que será motivo de estudio y verificación en la implementación del caso de estudio del RM en INTA EEA Anguil. Adicionalmente, nuestra propuesta cuenta con los procesos formalizados mediante SPEM, lo que promueve una especificación bien establecida, comunicable y extensible.

Themis [33] es un sistema de procesamiento de flujos federado para despliegue multi-sitio y recursos limitados, ejecuta las consultas sobre flujos en modo global, brindando al usuario retroalimentación permanente respecto de la calidad de procesamiento experimentada para su consulta. Incorpora técnicas de balance y descarte selectivo distribuido sobre los flujos de datos, los cuales se estructuran bajo un modelo de flujo relacional. 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 casos de aplicación como el RM representan aspectos cruciales.

SECRET [34] 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 a) se focaliza en el procesamiento de flujos, b) incorpora metadatos a los efectos de guiar dicho procesamiento, y c) cuenta con procesos formalmente especificados usando SPEM.

7 Conclusiones y Trabajo Futuro

En el presente artículo hemos presentado la APcMM como evolución del EIPFDcMM, incorporando la posibilidad de gestionar grandes repositorios y la provisión de datos a terceros mediante servicios por suscripción, sobre entornos de cómputo distribuido y basado en tecnologías maduras, open source, ampliamente usadas en entornos productivos (por ejemplo, Apache Kafka en LinkedIn), a los efectos de permitir la extensibilidad, dinamismo, comunicabilidad y escalabilidad de la arquitectura.

En relación al proceso de medición, la APcMM permite el empleo de metadatos basados en un marco conceptual de M&E, que incorporado en forma conjunta con las medidas, otorga consistencia y comparabilidad al análisis estadístico, como así también al análisis de los datos históricos. De este modo, los flujos de datos obtenidos a partir de fuentes heterogéneas, los datos procesados y/o provistos en tiempo real por suscripción, o bien aquellos almacenados en grandes repositorios, se estructuran bajo el esquema C-INCAMI/MIS, con el objetivo de mejorar la interoperabilidad del sistema. Luego, la posibilidad de gestionar grandes repositorios de datos en APcMM mediante tecnología madura y escalable permite: a) Acceder, comparar y analizar el volumen de datos históricos de las mediciones para cada proyecto de M&E en Apache HBase, sin requerir intervención humana, b) Organizar y estructurar una Memoria Organizacional que capitalice la experiencia previa como caso-solución (clave-valor) para el posterior entrenamiento de los clasificadores, como así también para poder recomendar cursos probables de acción al tomador de decisiones, c) Retroalimentar la Memoria Organizacional a partir de los clasificadores en línea, a medida que se actualizan en forma incremental ante nuevas situaciones, d) Soportar el proceso de toma de decisiones sobre los proyectos de M&E activos, a partir de estructuras de Data Warehousing implementadas sobre Apache Hive a partir de los datos originales dentro de Apache HBase, y e) Fomentar el análisis y descubrimiento de nuevos patrones sobre las entidades bajo análisis en los proyectos de M&E, a partir de procesos por lote sobre Apache Hadoop empleando Apache Mahout.

Así, la APcMM plantea el aprovisionamiento y consumo de datos en tiempo real como productor y consumidor (por suscripción) mediante Apache Kafka, la implementación de topologías ajustables de procesamiento distribuido mediante Apache Storm, un cluster Apache Hadoop para el procesamiento por lote y empleando un sistema de archivos distribuidos para gestionar grandes volúmenes de datos, Apache HBase para el acceso aleatorio a las mediciones y la definición de los proyectos de M&E, Apache Hive para la consulta de datos ad-hoc mediante Hive-QL, Apache Mahout para el análisis de patrones, Redis para la gestión de cache en memoria, MOA para implementar clasificadores incrementales sobre los flujos y R como motor de cálculo estadístico. De este modo, nuestra estrategia se monta sobre tecnologías de cómputo maduras y almacenamiento distribuido, lo que posibilita la capacidad de gestión de grandes volúmenes de datos, favorece la escalabilidad, posibilita el procesamiento de medidas con altas tasas de arribo, y aprovisionar grandes volúmenes de datos sobre demanda a partir de estrategias como MapReduce.

Un caso de aplicación sobre el RM de la EEA INTA Anguil ha sido presentado. Allí se contrasta la gestión de archivos actual respecto de las ventajas que se podrán

obtener a partir de la implementación de APcMM, permitiendo el consumo de datos en tiempo real por suscripción, consultas ad-hoc y capitalización de la experiencia mediante la Memoria Organizacional.

Como trabajo a futuro, se avanzará en la implementación de la APcMM sobre el RM de la EEA INTA Anguil a partir del Convenio de Cooperación Tecnica suscripto entra la Facultad de Ingeniería y dicho organismo. En tal sentido, la idea es documentar la experiencia respecto del comportamiento del RM (volúmenes por período, fluctuación de la tasa de arribo, etc.), y retroalimentar la APcMM a partir de dicho caso.

Reconocimientos. Esta investigación está soportada por el proyecto PICTO 2011-0277 de la Agencia de Ciencia y Tecnología, el proyecto 09/F068 de la Facultad de Ingenieria de la UNLPam y parcialmente por el Convenio de Cooperación Técnica Facultad de Ingeniería (UNLPam) – INTA EEA Anguil 2015-2017.

Referencias

1. Jain, A & Nalya, A. (2014) “Learning Storm. Create real-time stream processing applications with Apache Storm”. Packt Publishing Ltd. Birmingham.
2. Frampton, M. (2015) “Mastering Apache Spark”. Packt Publishing Ltd. Birmingham.
3. Diván, M, Olsina, L & Gordillo, S. (2011) “Strategy for Data Stream Processing Based on Measurement Metadata: An Outpatient Monitoring Scenario”. Journal of Software Engineering and Applications. Vol. 4, N° 12, pp. 653-665.
4. Molina, H & Olsina, L (2007) “Towards the Support of Contextual Information to a Measurement and Evaluation Framework”. In QUATIC, Lisboa, Portugal.
5. Olsina, L, Papa, F & Molina H (2007) “How to Measure and Evaluate Web Applications in a Consistent Way”. in Ch. 13 in Web Engineering, Springer. pp. 385–420.
6. Cugola, G. & Margara, A. (2012) “Processing flows of information: From data stream to complex event processing”. Journal of ACM Computing Surveys. Vol. 44, N° 3, Article N° 15.
7. Bockermann, C & Blom, H (2012) “Processing Data Streams with The RapidMiner Streams Plugin”. Dortmund, Germany.
8. Diván, M (2011) “Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones”. PhD Tesis. Facultad de Informática. Universidad Nacional de La Plata, La Plata, BA, Argentina.
9. Diván, M, Martín, M & Olsina, L (2013) “Hacia la Retroalimentación del Procesamiento de Flujos de Datos Sustentado en Memoria Organizacional”. 1er Congreso Nacional de Ingeniería Informática/Sistemas de Información. Córdoba, Argentina.
10. Object Management Group (2008) “Software Process Engineering Meta-Model Specification (SPEM)”. Versión 2.0.
11. Diván, M & Olsina, L (2014) “Process View for a Data Stream Processing Strategy based on Measurement Metadata”. Electronic Journal of Informatics and Operations Research (SADIO). Vol. 13, N° 1, pp. 16-34.
12. Diván, M & Martín, M (2015) “Estrategia de Procesamiento de Flujos de Datos Sustentada en Big Data y Memoria Organizacional”. 3er Congreso Nacional de Ingeniería Informática/Sistemas de Información. CABA, Argentina.
13. Guller, M (2015) “Big Data Analytics with Spark. A Practitioner's Guide to Using Spark for Large Scale Data Analysis”. Apress. New York.
14. Apache Software Foundation (2016) “Apache Storm” [Online]. Último acceso a http://storm.apache.org el 4 de mayo de 2016.

15. Apache Software Foundation (2016) “Apache Kafka” [Online]. Último acceso a http://kafka.apache.org el 4 de mayo de 2016.
16. Garg, N (2013) “Apache Kafka. Set up Apache Kafka clusters and develop custom message producers and consumers using practical, hands-on examples”. Packt Publishing Ltd. Birmingham.
17. Apache Software Foundation (2016) “Apache Hadoop” [Online]. Último acceso a http://hadoop.apache.org el 4 de mayo de 2016.
18. Holmes, A (2015) “Hadoop in Practice”. Segunda edición. Manning. New York.
19. Apache Software Foundation (2016) “Apache HBase” [Online]. Último acceso a http://hbase.apache.org el 4 de mayo de 2016.
20. Spaggiari, J & O'Dell, K (2015) “Architecting HBase Applications (Early Release)”. O’Reilly Media. New York.
21. Martin, M & Olsina, L (2009) “Added Value of Ontologies for Modeling an Organizational Memory”. In Building Organizational Memories: Will You Know What You Knew? Editor Girard, J. IGI Global. USA. Pp. 127-147.
22. Martín, M (2011) “Memoria Organizacional Basada en Ontologías y Casos para un Sistema de Recomendación en Aseguramiento de la Calidad”. PhD Tesis. Facultad de Informática. Universidad Nacional de La Plata, La Plata, BA, Argentina.
23. Diván, M, Bellini, Y, Martín, M, Belmonte, M, Lafuente, G & Caldera, J (2015) “Towards a Data Processing Architecture for the Weather Radar of the INTA Anguil”. In proc. of International Workshop on Data Mining with Industrial Applications, Asunción, Paraguay.
24. Marroco, C, Duin, R & Tortorella, F (2008) “Maximizing the area under the ROC curve by pairwise feature combination”. ACM Pattern Recognition. Pp. 1961-1974.
25. R Core Team (2016) “R: A Language and Environment for Statistical Computing”. Viena.
26. Da Silva, M & Tavares, H (2015) “Redis Essentials”. Packt Publishing Ltd. Birmingham.
27. Bifet, A, Holmes, G, Kirkby, R & Pfahringer, B (2010) “MOA: Massive Online Analysis”. Journal of Machine Learning Research. Vol. XI, pp. 1601-1604.
28. Rutherglen, J, Wampler, D & Capriolo, E (2012) “Programming Hive”. O’Reilly Media Inc. California, USA.
29. Gupta, A (2015) “Learning Apache Mahout Classification”. Packt Publishing Ltd. Birmingham.
30. Hartmann, T, Tamburrino, M & Bareilles, S (2010) "Preliminar Analysis of data obtained from the INTA radar network for the study of the precipitations in the pampeana region (in spanish)". In proc. of 39º Jornadas Argentinas de Informáticas - 2º Congreso Argentino de Agroinformática, CABA.
31. Gematronik (2007) “Instruction Manual. Rainbow 5”. Gematronik GmbH. Neuss.
32. Lee, M, Lee, M, Hur, S & Kim, I (2016) “Load Adaptive and Fault Tolerant Distributed Stream Processing System for Explosive Stream Data”. Journal of Transactions on Advanced Communications Technology. Vol. 5, N° 1, pp. 745-751.
33. Kalyvianaki, E, Fiscato, M, Salonidis, T & Pietzuch, P (2016) “THEMIS: Fairness in Federated Stream Processing under Overload”. In proc. of ACM SIGMOD. San Franciso, CA, USA.
34. Botan, I, Derakhshan, R, Dindar, N, Haas, L, Miller, R & Tatbul, N (2010) “SECRET: a model for analysis of the execution semantics of stream processing systems”. In proc. of VLDB Endowment. Vol. 3, N° 1-2, pp. 232-243.