



UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS
“Facultad 3”

TESIS PRESENTADA EN OPCIÓN AL TÍTULO DE
MÁSTER EN INFORMÁTICA AVANZADA

TÉCNICA BASADA EN CELDAS DE PROCESOS PARA DIAGNOSTICAR
PROCESOS DE NEGOCIO

Autor: Ing. Dina Yaksilik Torres Sakipova

Tutor: DrC. Arturo Orellana García

La Habana, Julio, 2018

AGRADECIMIENTOS

A mis padres, que son la luz de mis ojos, las personas que más quiero y de quienes vivo eternamente orgullosa y agradecida.

A mi esposo, por ser una persona maravillosa que me quiere y me apoya siempre.

A mi tutor, por su ayuda y consejos con todo lo concerniente a la investigación y a la maestría en general.

A Mailén, por apoyarme con el tiempo necesario para poder terminar la maestría.

Al colectivo de profesores de la maestría, por aportar a nuestra formación con tanta metodología y conocimientos.

A todos los que de una forma u otra me apoyaron en la realización de este trabajo.

DEDICATORIA

Le dedico este trabajo a mis padres, que son todo para mí y que sé que lo recibirán con la mayor alegría del mundo.

DECLARACIÓN JURADA DE AUTORÍA

Declaro por este medio que yo, Dina Yaksilik Torres Sakipova, con carné de identidad 85093014178, soy el autor principal del trabajo final de maestría Técnica basada en celdas de procesos para diagnosticar procesos de negocio, desarrollada como parte de la Maestría en Informática Avanzada y que autorizo a la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo.

Y para que así conste, firmo la presente declaración jurada de autoría en La Habana a los días del mes de del año 2018.

Firma

RESUMEN

La minería de procesos se encarga de descubrir, monitorear y mejorar los procesos reales que se llevan a cabo en una empresa. Para ello extrae conocimiento de los registros de eventos que manejan los sistemas informáticos actuales y genera modelos que describen los procesos de negocio a partir de sus técnicas. Un novedoso enfoque dentro de su área son los cubos de proceso, formados por celdas de proceso que permiten visualizar modelos de proceso a partir de diferentes dimensiones. La comparación de celdas es un área de investigación, debido a la dificultad de ajustar varios resultados en una misma pantalla y extraer información relevante de los mismos. La presente investigación se basa en el enfoque de los cubos de proceso para comparar celdas de proceso y contribuir a su diagnóstico. Propone una nueva técnica desarrollada para el marco de trabajo ProM, que fragmenta los registros de eventos en celdas y los visualiza de forma gráfica y a nivel de datos. La instanciación de la técnica en una herramienta informática permitió aplicar registros de eventos reales y analizar los resultados. La valoración por expertos y potenciales usuarios, así como las ventajas de la propuesta sobre otras alternativas identificadas, permitió validar la pertinencia y aplicabilidad de la técnica propuesta y concluir que la misma contribuye al diagnóstico de procesos de negocio.

Palabras Clave: comparación de procesos, cubos de proceso, diagnóstico de procesos, minería de procesos, operaciones OLAP.

ABSTRACT

Process mining is in charge of discovering, monitoring and improving the real processes that take place in a company. To do this, it draws knowledge from the event *logs* that manage current computer systems and generates models that describe business processes through their techniques. A novel focus in the area is the process cubes, formed by process cells that allow to visualize process models starting from different dimensions. Cells comparison is a research field, due to the difficulty of adjusting several results in the same screen and to extract outstanding information from there. This investigation is based on the process cubes approach to compare process cells and to contribute to their diagnosis. It proposes a new technique, developed for the ProM framework, that divide event logs into cells and visualize them in a graphic and data level way. The development of the technique in a computer tool allowed to apply real event logs and analyze the results. The criterion issued by experts and potential users, as well as the advantages of the proposal over other identified alternatives, allowed to validate the pertinence and applicability of the proposed technique and to conclude that it contributes to the diagnosis of business processes.

Keywords: OLAP operations, process comparison, process cube, process diagnostic, process mining.

ÍNDICE

| | |
|--|----|
| INTRODUCCIÓN | 4 |
| CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN..... | 10 |
| 1.1 Gestión de Procesos de Negocio | 10 |
| 1.2 Minería de procesos..... | 11 |
| 1.2.1 Tipos de procesos..... | 12 |
| 1.2.2 Tipos de minería de procesos | 12 |
| 1.2.3 Aplicaciones de la minería de procesos | 13 |
| 1.2.4 Limitaciones de la minería de procesos | 14 |
| 1.3 Cubos OLAP..... | 16 |
| 1.3.1 Limitaciones de los cubos OLAP | 16 |
| 1.4 Cubos de proceso | 17 |
| 1.4.1 Operaciones en los cubos de proceso | 18 |
| 1.4.2 Investigaciones referentes a los cubos de proceso | 20 |
| 1.4.3 Resultado del análisis de las investigaciones realizadas sobre los cubos de proceso ... | 23 |
| 1.5 Ambiente tecnológico para la solución | 24 |
| 1.6 Conclusiones del capítulo..... | 25 |
| CAPÍTULO 2: PROPUESTA DE SOLUCIÓN | 26 |
| 2.1 Conceptualización de la solución | 26 |
| 2.2 Estructura general de la propuesta de solución..... | 28 |
| 2.3 Descripción operacional de la propuesta de solución | 29 |
| 2.4 Implementación..... | 36 |
| 2.5 Conclusiones del capítulo..... | 41 |
| CAPÍTULO 3: VALIDACIÓN DE LA SOLUCIÓN | 42 |
| 3.1 Diseño de la validación | 42 |
| 3.2 Validación de la propuesta de solución | 42 |
| 3.2.1 Estudio de casos..... | 42 |
| 3.2.1.1 Aplicación de la técnica en un proceso real en una oficina de tráfico | 42 |
| 3.2.1.2 Aplicación de la técnica en un proceso real en un hospital..... | 48 |
| 3.2.1.3 Análisis de los resultados del estudio de casos múltiples..... | 52 |
| 3.2.2 Criterio de expertos aplicando escalamiento de Likert..... | 55 |
| 3.2.3 Evaluación de la satisfacción de la técnica desarrollada por potenciales usuarios | 57 |
| 3.3 Conclusiones del capítulo..... | 60 |
| CONCLUSIONES GENERALES | 61 |

RECOMENDACIONES 62
BIBLIOGRAFÍA 63
ANEXOS 69

INTRODUCCIÓN

La Gestión de Procesos de Negocio (*Business Process Management*, BPM) es la disciplina que combina conocimiento sobre tecnología de información y ciencias de gestión y lo aplica en conjunto a los procesos de negocio operacionales [1] y [2]. Esta se compone de siete fases, siendo el diagnóstico la que comprende la ejecución de los procesos, detección de anomalías y cuellos de botella, búsqueda de patrones interesantes, entre otros. El diagnóstico ayuda a tener una visión general del proceso, de sus aspectos más significativos y de las técnicas que pueden ser más útiles para su análisis [3].

Los mercados demandan la mejora continua de la administración y la toma de decisiones, para lo cual necesitan aprovechar mejor sus datos e información para realizar un buen diagnóstico de sus procesos. A su vez, los procesos de negocio generan un gran volumen de datos. Estos no suelen proporcionar beneficios directamente, sino que de ellos se puede extraer información que ayude a comprender fenómenos o a tomar decisiones. Por tanto, se necesitan soluciones capaces de lidiar con la gran cantidad de datos que se almacenan hoy a una velocidad superior a la que se pueden analizar.

Existen mecanismos para analizar la información almacenada en Bases de Datos (BD) que propician la resolución de problemas de planificación, gestión y operación de sistemas. Por un lado se encuentran la Minería de Datos (MD) y el Procesamiento Analítico En Línea (*Online Analytical Processing*, OLAP), ambos orientados al análisis de datos.

La MD cuenta con métodos matemáticos y técnicas de software para el análisis inteligente de datos y búsqueda de regularidades en los mismos a partir de BD generalmente grandes. Combina técnicas semiautomáticas de inteligencia artificial, análisis estadístico, BD y visualización gráfica, para obtener información que no esté representada explícitamente en los datos. Descubre relaciones, tendencias, comportamientos atípicos, patrones y trayectorias ocultas, con el propósito de soportar los procesos de toma de decisiones con mayor conocimiento [1].

A su vez, la tecnología OLAP es un método para ayudar a la toma de decisiones en situaciones donde un conjunto de datos de difícil medición, tales como ventas o ganancias, necesiten ser analizados en diferentes niveles de agregación estadística. Adopta el paradigma multidimensional para proveer un acceso rápido a datos cuando el análisis se realiza desde diferentes perspectivas [4]. En estos modelos los datos son vistos como cubos que consisten en categorías descriptivas (dimensiones) y valores cuantitativos (medidas). Los datos son clasificados en diferentes dimensiones, que pueden ser vistas unas con otras en diferentes combinaciones para obtener diferentes análisis de los datos que contienen.

Por otro lado, otra forma de analizar los procesos, que no necesariamente depende de BD y que es orientada a procesos, es a partir de la Minería de Procesos (MP). Esta es un área joven que se encarga de descubrir, monitorear y mejorar los procesos reales, extrayendo conocimiento de los registros de

eventos (*logs* por su nombre en inglés) que manejan los sistemas informáticos actuales [1]. Generalmente descubre modelos de procesos, a partir de los cuales permite un estudio con un enfoque más analítico, basado en hechos, a diferencia de la modelación de procesos usando herramientas que existen, pero que pueden no ser correctamente utilizadas.

Es importante diferenciar la MP del modelado estándar que se realiza cuando se desea describir procesos con ayuda de una herramienta de modelado y el conocimiento y la experiencia de una persona calificada. El modelado estándar presenta las actividades como están establecidas, como deben funcionar, y está determinado por el conocimiento que tenga la persona, tanto de los procesos como de la herramienta con la que modela. Sin embargo, la MP representa las actividades como ocurrieron y se registraron en los *logs*.

Es frecuente que exista una brecha entre cómo se supone que deben desarrollarse los procesos y cómo realmente lo hacen [5]. Normalmente las personas involucradas en los procesos comerciales consideran que conocen bien cómo estos funcionan. La MP usa los datos históricos del proceso para confirmar o refutar esta creencia [6]. Su uso permite descubrir los procesos, realizar un análisis de conformidad (investigando si la realidad se ajusta al modelo dado y viceversa) y una extensión (aumentando un modelo existente con información adicional extraída desde los *logs*).

Estos *logs* contienen información sobre qué actividades se realizaron, quién las realizó, cuándo, etc. A través de ellos se puede identificar la información de las actividades ejecutadas durante el proceso. Cuando los *logs* muestran que los procesos se caracterizan por una manera controlada de ejecutarse y que poseen una clara estructura, dichos procesos son estructurados y se denominan “procesos lasaña” [7]. Son frecuentemente menos interesantes debido a que es fácil aplicarle técnicas de MP y, debido a que están organizados, las mejoras son pequeñas.

Sin embargo, en muchas ocasiones los *logs* contienen actividades ejecutadas con diversos órdenes, donde varía el momento en el que se ejecutaron, las actividades precedentes y posteriores cambian, cambian los autores, o se evidencian relaciones o variaciones entre las actividades. Estos procesos poco estructurados, denominados *spaguetti* [7], generan modelos poco comprensibles que no aportan información a partir de un análisis subjetivo, debido al alto grado de dificultad que asume comprenderlos.

Actualmente no todos los negocios poseen sus procesos bien estructurados. De hecho, estos tienden a ser menos estructurados de lo que se espera [8]. Esto influye en los modelos de procesos de negocio como artefacto de salida del descubrimiento como tipo de minería. Las técnicas empleadas para descubrir los modelos, de manera independiente, no manejan apropiadamente la mayoría de las situaciones desafiantes: el ruido, la incompletitud, las tareas escondidas o duplicadas, la visualización de resultados, entre otros [9]. Esto provoca que el resultado sea un modelo inadecuado que no

representa la realidad o que resulta difícil interpretarlo. El resultado de minar los eventos de procesos poco estructurados es usualmente un modelo poco estructurado y difícil de entender, que resulta inútil para los analistas de procesos.

El problema fundamental no radica en que las técnicas existentes producen resultados incorrectos, sino que el modelo resultante muestra todos los detalles sin brindar una abstracción conveniente, lo que implica poco o ningún entendimiento del mismo. Esto se debe a que los mismos procesos son realmente como *spaguetti*, de modo que el modelo refleja la realidad.

Los desafíos 10 y 11 de la MP [1] son aumentar la usabilidad y la comprensión. Sin embargo, las técnicas padecen a menudo de incompreensión. Por lo tanto, es importante explorar otras vías de solución a la dificultad visual que supone emplear estas técnicas y analizar sus modelos descubiertos.

Algunas dificultades que poseen las técnicas de MP son:

- No son capaces de proporcionar un alto nivel de visibilidad en el proceso, abstrayéndose de detalles indeseados.
- Los modelos de procesos generados generalmente son estáticos, es decir, no brindan prácticamente ninguna posibilidad de que el usuario pueda interactuar y modificar las vistas del modelo. En ellos muchas veces se observa que no queda claro el flujo principal de actividades del proceso o que los nodos resultan conectados en ambas direcciones, lo que le dificulta al analista de procesos definir causa y efecto.
- Generalmente poseen dificultad para lidiar con el ruido, la precisión (que es una de las principales causantes de los modelos *spaguetti* grandes y complejos) y la singularidad (que defiende la idea de que un solo modelo es suficiente para representar toda la realidad).
- No manejan múltiples variantes del proceso y / o colecciones heterogéneas de casos.
- Dificilmente producen modelos globales capaces de relacionar y analizar diferentes grupos y variantes de procesos [10].

La dificultad en la comprensión de los modelos de procesos obtenidos con la MP incide directamente en el diagnóstico de procesos de negocio, es decir, en la detección de anomalías, fraudes, procesos incompletos, desviaciones, en el análisis de similitudes o diferencias entre subprocesos, temas que influyen en la calidad de las decisiones y en el tiempo de análisis de los procesos de las empresas.

Por ello se estudian diversas vías para mejorar la comprensibilidad de los modelos de procesos descubiertos. Lograr representar un modelo en diferentes vistas permite mejorar su comprensión. Una vista de un modelo de proceso de negocio muestra una representación parcial de acuerdo a alguna propiedad de forma similar a las vistas de tablas en BD [11].

A partir de los cubos OLAP y como nuevas formas de visualización de procesos que sean más adecuadas para representar y analizar procesos no estructurados, se han realizado algunas propuestas aproximándose a la MP multidimensional. La idea de los cubos de proceso (CP) surge en [7], donde se describe que la misma está relacionada con el enfoque de “divide y vencerás” en la MP, donde grandes *logs* son particionados para formar registros más pequeños para mejorar el desempeño. Este enfoque posee limitantes o dificultades que repercuten en el objetivo de la MP: identificar y extraer rápidamente información importante para apoyar la auditoría y la toma de decisiones para elevar los niveles de competencia del negocio.

A partir del análisis documental realizado, la experiencia de la autora de este trabajo en el área y las investigaciones previas del Grupo de Investigación de Minería de Procesos de la UCI, se identificaron insuficiencias vinculadas a la problemática existente:

- La escasa información que pueden brindar los modelos de procesos generados a partir de técnicas de MP, cuando los procesos son poco estructurados, afecta el análisis, gestión y control de los mismos, y, por consiguiente, se desaprovecha la oportunidad de generar información valiosa basada en datos reales almacenados.
- La comparación que se puede realizar a partir de las técnicas de MP basadas en CP, se limita a un análisis visual de los modelos de procesos generados, dependiendo del conocimiento y experiencia de la persona que realiza el análisis, lo cual puede introducir errores en los resultados o afectar el diagnóstico del proceso.
- Las técnicas de MP basadas en CP, en su estado actual, no tienen lo suficientemente en cuenta los datos almacenados para extraer información que permita realizar un análisis mejor fundamentado de la ejecución de los procesos de negocio, lo que limita la búsqueda de comportamientos similares o diferentes en los procesos.
- Dichas técnicas carecen de flexibilidad con respecto a las posibilidades de filtrado de registros y el cambio de dimensiones en el cubo, lo que provoca que el diagnóstico del proceso sea limitado.

A partir de la situación descrita anteriormente se identifica el siguiente **problema de investigación**: ¿Cómo comparar celdas de proceso de manera que contribuya al diagnóstico de procesos de negocio?

El **objeto de estudio** se centró en las técnicas de minería de procesos enfocadas en cubos de proceso. Para darle solución al problema se plantea como **objetivo general**: Desarrollar una técnica de minería de procesos para comparar celdas de proceso en el marco de trabajo ProM, que contribuya al diagnóstico de procesos de negocio.

Para darle cumplimiento al objetivo general, se plantean los siguientes **objetivos específicos**:

- Elaborar el marco teórico referencial de la investigación, relacionado con la minería de procesos y los cubos de proceso.
- Diagnosticar el estado actual del análisis de procesos de negocio basado en el enfoque de cubos de proceso.
- Desarrollar una técnica de minería de procesos instanciada en un complemento para la herramienta ProM.
- Validar los resultados alcanzados a partir de los métodos definidos en la investigación.

El **campo de acción** se delimita a las celdas de proceso para la comparación de procesos de negocio. La **hipótesis** formulada plantea que el desarrollo de una técnica de minería de procesos para comparar celdas de proceso en el marco de trabajo ProM, contribuirá a diagnosticar procesos de negocio.

La definición conceptual de la **variable independiente** técnica de minería de procesos para comparar celdas de proceso de negocio es: mecanismo que propicia el análisis de la información contenida en celdas de cubos de proceso para establecer relaciones, diferencias, semejanzas o aspectos interesantes en el diagnóstico de su ejecución.

La definición conceptual de la **variable dependiente** contribuir al diagnóstico de procesos de negocio es: aportar al análisis de procesos de negocio a partir de la capacidad de identificar comportamientos comunes, patrones infrecuentes, procesos incompletos o información adicional interesante.

Métodos teóricos empleados:

Análítico-sintético: para descubrir, analizar y organizar la información referente a la MP y sus técnicas de descubrimiento. Utilizado además para descomponer el problema de investigación en elementos, profundizar en su estudio y luego sintetizarlos en la solución propuesta.

Histórico-lógico: para comprender la evolución y desarrollo de los temas vinculados a la problemática, con el objetivo de determinar los elementos viables en la solución.

Hipotético-deductivo: para obtener un grupo de conocimientos, analizando los datos generales válidos para llegar a una conclusión particular, específicamente sobre la forma de organizar la información poco variable en el tiempo en un sistema informático.

Sistémico: para tratar el problema como una unidad, un todo, determinando sus componentes y relaciones con el objetivo de determinar la dinámica y el funcionamiento de la solución.

Modelación: para contextualizar los elementos involucrados en la visualización de procesos. Considerado para representar las vistas alternativas de la solución. Mediante este método se crean abstracciones para explicar la realidad de los elementos abordados.

Métodos empíricos empleados:

Criterio de expertos empleando el escalamiento de Likert: para obtener la valoración de expertos sobre la técnica desarrollada en la investigación.

Técnica de ladov: para evaluar el aporte práctico de la técnica propuesta, observable en el nivel de satisfacción de potenciales usuarios con la técnica propuesta.

Estudio de caso: para analizar una unidad, comprobar si se cumple la hipótesis y validar los resultados obtenidos.

Encuesta: para obtener mediciones cuantitativas de los elementos abordados en la investigación a partir de cuestionarios elaborados previamente.

La **novedad** del presente trabajo de investigación se expresa en sus aportes prácticos mediante el desarrollo de una técnica para la comparación de celdas de proceso de negocio en el marco de trabajo ProM, basada en el enfoque de cubos de proceso.

Estructura del trabajo

Capítulo 1. Fundamentación teórica de la investigación: se elabora el marco teórico de la investigación referente a la minería de procesos, sus tipos, perspectivas, limitaciones y aplicaciones. Se realiza un estudio de los cubos de proceso, analizando las principales propuestas encontradas en la literatura sobre su aplicación y limitaciones. Por último, se caracteriza la tecnología utilizada para la solución.

Capítulo 2. Propuesta de solución: se define el modelo conceptual de la solución. Se describen las etapas que conforman la técnica desarrollada con sus entradas y salidas. Se explica la estructura del componente desarrollado y se presentan interfaces del mismo.

Capítulo 3. Validación de la solución: se realiza la validación de la propuesta de solución. Para ello se aplica el escalamiento de Likert para obtener el criterio de expertos y la técnica de ladov a potenciales usuarios para evaluar la propuesta. Además, se realiza un estudio de casos con *logs* y técnicas reales para validar la hipótesis de la investigación.

Por último se establecen las conclusiones, se emiten recomendaciones para extender la propuesta y abrir nuevas investigaciones, se relacionan las referencias bibliográficas de la memoria escrita y los anexos.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN

Introducción

En el presente capítulo se realiza un análisis referente a la minería de procesos y los conceptos asociados a la investigación. Se describen los cubos OLAP y de procesos, así como sus operaciones. Además, se analizan las investigaciones referentes a aplicaciones o aportes novedosos sobre cubos de proceso. Como resultado de la posición crítica asumida durante la consulta bibliográfica, se identifican las dificultades presentes en dichas investigaciones y se selecciona la comparación entre celdas como motivo de la presente investigación.

1.1 Gestión de Procesos de Negocio

La BPM es la disciplina que combina conocimiento sobre tecnología de información y ciencias de gestión y lo aplica en conjunto a los procesos de negocio operacionales. Su ciclo de vida abarca las siete fases de un proceso de negocio. En la fase de (re)diseño se crea un nuevo modelo de proceso o se adapta un modelo de proceso existente. En la fase de análisis se analiza un modelo candidato y sus alternativas. Después de la fase de (re)diseño, se implementa el modelo (fase de implementación) o se (re)configura un sistema existente (fase de (re)configuración). En la fase de ejecución se ejecuta el modelo diseñado. En esta fase se monitorea el proceso. En la fase de ajuste se pueden realizar pequeños ajustes sin rediseñar el proceso. En la fase de diagnóstico se analiza el proceso ejecutado y la salida de esta fase puede conducir al rediseño del proceso nuevamente [1].

La fase de diagnóstico comprende la ejecución de los procesos, la detección de anomalías y cuellos de botella, la búsqueda de patrones interesantes, entre otros. Ayuda a tener una visión general del proceso, de los aspectos más significativos del mismo y de las técnicas que pueden ser más útiles en su análisis posterior [3].

Los sistemas de información orientados a procesos posibilitan gestionar los procesos de negocio de una organización. Son conocidos como sistemas PAIS (*Process Aware Information Systems*) y facilitan la colaboración, proporcionando medios para la comunicación y el intercambio de datos, y que hacen que la ejecución de los procesos empresariales sea observable. Cada ejecución de una tarea en el proceso es notada por el PAIS y puede ser almacenada en un *log*. Así, el sistema puede documentar la ejecución del proceso para una auditoría o para análisis posteriores.

Dichos sistemas permiten generar *logs* automáticamente para ser usados por la MP para descubrir modelos de procesos. Son cada vez más usados por organizaciones para dar soporte a sus negocios [12]. Sin embargo, en su mayoría solo existe una noción implícita del proceso. Es decir, la mayoría de los procesos se encuentran codificados en aplicaciones informáticas y el conocimiento solo lo poseen las personas que las usan, lo que se traduce en una pobre gestión de soporte. Si no hay un modelo de

proceso que refleje la realidad, es difícil lograr la conformidad y el rendimiento de manera precisa y unificada [13].

1.2 Minería de procesos

Una disciplina que permite realizar el diagnóstico de procesos es la MP. En [14] se presenta una relación de definiciones sugeridas por diversos autores que han evolucionado al incluir los análisis de los *logs* y las técnicas, algoritmos y herramientas basadas en principios de MD, Inteligencia de Negocios y BPM [15], [16] y [17].

La MP es valiosa para la mayoría de las fases de BPM mostradas en la Figura 1, especialmente para el diagnóstico. Es un área joven que se encarga de descubrir, monitorear y mejorar los procesos reales extrayendo conocimiento de los *logs* que manejan los sistemas informáticos actuales [1]. Consiste en extraer modelos de procesos a partir de *logs* almacenados que contienen información sobre qué actividades se realizaron, quién las realizó, cuándo, etc. Un *log* es un registro cronológico de las actividades de un sistema informático que son almacenadas en un archivo en el sistema [18]. A través de ellos se puede identificar la información de las actividades ejecutadas durante el proceso [19]. A partir de los modelos generados con la MP es posible realizar un estudio más analítico basado en hechos, a diferencia de la modelación de procesos usando herramientas que existen, pero que no son correctamente utilizadas.

Es importante diferenciar el resultado de la MP y el modelado estándar que se realiza cuando se desea describir procesos con ayuda de una herramienta de modelado y el conocimiento y la experiencia de una persona calificada. Este modelado presenta las actividades como están establecidas, como deben funcionar y está determinado por el conocimiento que tenga la persona, tanto de los procesos como de la herramienta con la que modela. Sin embargo, la MP presenta las actividades como ocurrieron y se archivaron en los *logs*.

La MP permite reconstruir, analizar y mejorar los procesos comerciales utilizando datos de eventos grabados de sistemas transaccionales [20]. Puede verse como una concretización de la Inteligencia de Negocios [21]. Sólo sus técnicas [22] y [23] relacionan directamente los datos del evento con los procesos de negocio de extremo a extremo. Los enfoques existentes de modelado de procesos de negocio que generan considerables modelos de proceso suelen estar desconectados de los procesos reales y los sistemas de información. Las técnicas de análisis orientadas a datos (por ejemplo, la MD y el aprendizaje automático) suelen centrarse en problemas de clasificación simple, agrupamiento, regresión o problemas de reglas de aprendizaje [24].

Los resultados del análisis con MP pueden ser usados para mejorar el rendimiento de los procesos o la conformidad con las reglas y regulaciones de una organización. Por tanto, la MP provee un puente entre la MD y la disciplina de BPM [24].

1.2.1 Tipos de procesos

Procesos estructurados

Cuando los *logs* muestran que los procesos se caracterizan por una manera de ejecutarse bastante controlada y que poseen una clara estructura, dichos procesos son estructurados y se denominan “procesos lasaña” según [25], [7], [26], [27], [1], [3] y [8]. Son frecuentemente menos interesantes debido a que es fácil aplicarle técnicas de MP y, debido a que están organizados, las mejoras son pequeñas.

Procesos poco estructurados

En muchas ocasiones los *logs* contienen actividades ejecutadas con diversos órdenes, donde varía mucho el momento en el que se ejecutaron, las actividades precedentes y posteriores cambian, cambian los autores, o se evidencian muchas relaciones o variaciones entre las actividades. Estos procesos poco estructurados, denominados *spaguetti* según [7], solo permiten aplicársele un subconjunto de las técnicas de MP. Generan modelos poco comprensibles que no aportan información debido al alto grado de dificultad que asume comprenderlos [25], [26], [27], [1], [3] y [8].

Actualmente los procesos tienden a ser menos estructurados de lo que se espera [8]. Esto influye en el modelo de proceso descubierto. Si se trata de construir un modelo para casos muy diferentes, entonces el modelo es probablemente demasiado complicado.

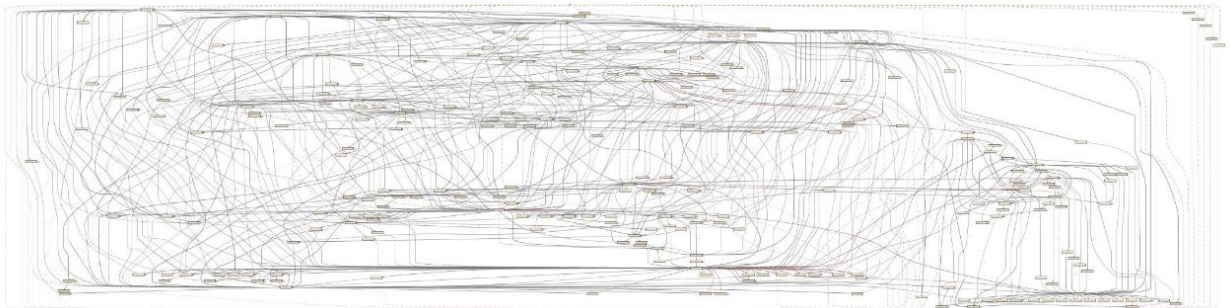


Figura 1 Modelo de proceso de un proceso poco estructurado. Fuente: [28].

En entornos flexibles como la salud [29], el flujo de los procesos de negocio generalmente depende de muchos factores. Como resultado, muchos casos siguen un camino único a través del proceso. Esta variabilidad causa problemas para las técnicas de MP existentes. Muchas descubren modelos *spaguetti*. Es problemático encontrar comportamiento común o desviaciones, ya que la mayoría de los casos son únicos y lo que es normal o excepcional puede depender de muchos factores y, en la práctica, a menudo es borroso [30]. En estos casos, realizar el diagnóstico a través de modelos es un proceso embarazoso, tedioso y a veces no es factible.

1.2.2 Tipos de minería de procesos

La literatura clasifica a la MP en tres tipos.

El **descubrimiento** produce un modelo de proceso a partir de un log, sin ninguna otra información a priori. Durante el descubrimiento, los usuarios investigan el modelo de proceso real y desglosan los datos del proceso para identificar patrones no deseados y fuentes de ineficiencias [20] y [31]. Es el tipo de MP más estudiado y utilizado según el estudio realizado por [6] del 2014 al 2017. Sorprende a muchas organizaciones al observar que con estas técnicas es posible descubrir procesos reales solamente con los *logs* que almacenan el comportamiento [32] y [1].

La **verificación**, monitoreo o chequeo de conformidad, compara un modelo de proceso existente con un *log* del mismo proceso, permitiendo chequear si la realidad, tal como está almacenada en el *log*, es equivalente al modelo y viceversa. Permite identificar las desviaciones, inconsistencias y anomalías en el proceso ejecutado, a partir de la comparación entre un modelo a priori basado en perspectivas teóricas y los *logs* [3], [32], [1] y [33]. Es, después del descubrimiento, el tipo de MP más investigado según el estudio realizado por [6].

Por último, la **mejora** consiste en extender o mejorar un modelo de proceso existente [1], [3], [8] y [34]. Mientras que la verificación de conformidad chequea la alineación entre el modelo y el log, la mejora persigue extender o cambiar un modelo a priori [32] al enriquecerlos con nuevos aspectos y perspectivas ilustradas por el modelo descubierto a posteriori.

1.2.3 Aplicaciones de la minería de procesos

La mayoría de los negocios poseen una limitada información sobre lo que realmente ocurre en su organización. Esto impide el enriquecimiento de criterios para la toma de decisiones necesarias en el descubrimiento de nuevas reglas de negocio, mejora continua, monitoreo o un control de los procesos ya establecidos [35]. En la práctica existe a menudo una distancia entre lo que está predefinido o se supone que debe ocurrir y lo que realmente ocurre.

La MP puede resultar en algunas de las siguientes acciones de mejora [7]:

- Rediseño: cambio permanente al software o modelo. Por ejemplo, ejecutar actividades secuenciales en paralelo.
- Ajuste: ajustar el proceso sin cambios en el software o modelo. Por ejemplo, asignar más recursos al proceso.
- Intervención: intervenir en el proceso para casos o recursos particulares. Por ejemplo, disciplinando a trabajadores que cometieron violaciones a regulaciones.
- Soporte: soportar el proceso operacional. Por ejemplo, recomendar acciones con menor costo esperado [36].

Con la MP las organizaciones, entre otras cosas, pueden: [37], [14] y [38]:

- Descubrir el modelo de ejecución real del proceso a partir de un *log* y una técnica de descubrimiento [39]. El modelo describe la situación real y no se basa en la documentación de cómo se debería ejecutar el proceso o en la percepción de las personas.
- Determinar si el proceso cumple con la reglamentación y procedimientos documentados, a través de la comparación del modelo real con los procedimientos documentados, pueden consultarse en [40] y [41], así como la detección y mitigación de posibles fuentes de no conformidad y fraudes en [42].
- Analizar la interacción del personal a través del análisis de la red social del proceso [43]. El chequeo de conformidad aplicado al análisis de redes sociales y producción de sistemas en dúos puede consultarse en [44] y en [45], así como el análisis de redes sociales asociadas a la perspectiva organizacional en [46].
- Descubrir bucles o cuellos de botella para mejorar el nivel de desempeño del proceso, puede consultarse en [47].
- Monitorear la productividad del personal y los tiempos de ciclo por actividad, lo que elimina la necesidad de realizar estudios de tiempos y movimientos de forma “manual”.
- Analizar el ambiente de trabajo, identificando problemas y proponiendo soluciones puede consultarse en [48] y [44].
- Navegar en los procesos. Consiste en explorar, predecir y recomendar para predecir el futuro de un caso particular y guiar al usuario en la selección de acciones adecuadas [49].
- Auditar los procesos. Se destacan las propuestas [18] y el marco de trabajo *Auditing 2.0* [45]. Plantean que la MP permite analizar todos los datos y no solo una muestra, implementar el modelo de riesgo de la auditoría y realizar análisis que no son posibles con las herramientas convencionales. Sus resultados constituyen un referente avalado según [50].
- Predecir procesos. Consiste en predecir resultados de instancias de proceso basados en los logs de casos similares. Es una mezcla de MP y MD. Se han realizado investigaciones [51] y [52] para predecir diferentes métricas, procesando tiempo, errores y retrasos.

Otros ejemplos de aplicaciones exitosas de la MP son [53], [54], [42], [55], [56], [57], [58], [59] y [60]. En [38] se exponen los factores críticos que se deben tener en cuenta para su aplicación en un contexto organizacional.

1.2.4 Limitaciones de la minería de procesos

Los *logs* a menudo no se generan específicamente para la MP, por lo que su granularidad puede ser de bajo nivel. Es vital para la aplicación exitosa de técnicas de descubrimiento de procesos tener *logs*

en un nivel apropiado de abstracción. Se han explorado varios métodos que abordan el reto de abstraer eventos de bajo nivel a eventos de nivel superior [61]. Sin embargo, las técnicas de descubrimiento de procesos, cuando la granularidad del *log* de entrada es demasiado baja, pueden resultar en un modelo de proceso con una o más propiedades no deseadas.

Los algoritmos tradicionales de descubrimiento de procesos tienen limitaciones para tratar procesos poco estructurados. Visualizar el comportamiento de procesos poco estructurados como los sistemas tradicionales de cuidado de la salud [56] resulta complejo. Los modelos resultantes son difíciles de comprender o incluso erróneos [8]. Estos modelos constituyen una de las principales dificultades de la MP [7]. El problema del efecto *spaguetti* es que crea un proceso largo o variable y difícil de entender debido a características inherentes del problema a resolver [62]. Estos temas y otros retos investigativos fueron expuestos en [38].

Muchos problemas vienen asociados con la búsqueda de la precisión. Con el fin de mostrar todo el comportamiento del proceso, un gran número de arcos es introducido en el modelo, generando modelos largos y complejos.

Resumiendo, algunos de los problemas [7], [1] y [27] en la MP, agudizados por la heterogeneidad de las fuentes de los *logs*, son:

- *Ruido*: los datos registrados pueden ser incorrectos o incompletos, creando problemas cuando los datos son minados. El *log* contiene comportamiento raro o infrecuente que no representa el típico comportamiento de los procesos. Uno de los factores críticos que determinan la calidad de los resultados de la MP es la calidad de los datos con los que se hace el análisis. Por lo tanto, es muy importante depurar y limpiar los datos [63], para lo cual son de gran utilidad las herramientas estadísticas para la identificación de los casos con datos atípicos e involucrar al personal de la organización para decidir sobre su inclusión, imputación o eliminación [38].
- *Distintas perspectivas*: los eventos del proceso pueden añadirse con información adicional con propósitos específicos.
- *Visualización de resultados*: pueden presentarse de formas gráficas en paneles de gestión.
- *Procesos concurrentes*: procesos que ocurren a la misma vez.
- *Singularidad*: la intención de conservar todo el comportamiento del proceso en un solo modelo. Esto obliga a observar toda la información a la vez, complicando su comprensión.
- *Inmutabilidad*: la mayoría de los algoritmos, una vez cargado y analizado el *log*, no permiten al usuario modificar parámetros del resultado obtenido para mejorarlo, por lo que el usuario debe comenzar desde el principio definiendo otros parámetros al inicio y esperar el resultado. Esto impide

usar la experiencia y conocimientos para mejorar el modelo resultante realizando ajustes manuales, de manera que no se quede detenido en un resultado indeseado.

1.3 Cubos OLAP

Las herramientas OLAP permiten navegar rápidamente por los datos, pero no se genera información en el proceso. Se llaman sistemas OLAP a aquellos sistemas que deben soportar requerimientos complejos de análisis, analizar datos desde diferentes perspectivas y soportar análisis complejos contra un gran volumen de datos.

OLAP es un método para ayudar a la toma de decisiones en situaciones donde un conjunto de datos de difícil medición, tales como ventas o ganancias, necesiten ser analizados en diferentes niveles de agregación estadística. Adopta el paradigma multidimensional como medio para proveer un acceso rápido a datos cuando el análisis se realiza desde diferentes perspectivas [4].

En estos modelos los datos son vistos como cubos que consisten en categorías descriptivas (dimensiones) y valores cuantitativos (medidas). Los datos son clasificados en diferentes dimensiones y pueden ser vistas unas con otras en distintas combinaciones para obtener diferentes análisis de los datos que contienen. El modelo multidimensional de datos simplifica a los usuarios acciones como: formular consultas complejas, arreglar datos en un reporte, cambiar datos resumidos a datos detallados, entre otros.

Las operaciones básicas de un cubo OLAP son:

- *Roll-Up y Drill-Down*: permiten definir el nivel de granularidad para analizar los datos. La primera puede involucrar acumulaciones simples o agrupaciones complejas que incluyen datos interrelacionados. La segunda permite moverse en la dirección contraria y presentar automáticamente datos detallados que abarcan datos consolidados.
- *Slice y Dice*: se refiere a la capacidad de visualizar la BD desde diferentes puntos de vista, es decir, permiten navegar entre las dimensiones.

Las técnicas OLAP organizan los datos típicamente bajo medidas numéricas. Su aplicación en datos no numéricos está siendo explorada, dado que algunas aplicaciones del tema pudieran ser los grafos y secuencias de eventos complejos [64].

1.3.1 Limitaciones de los cubos OLAP

Los cubos OLAP están optimizados para el análisis de datos, más bien para almacenar datos originados desde múltiples fuentes para evitar la redundancia. Por tal motivo, se basa mayormente en datos históricos y no en datos instantáneos, los cuales son complicados de analizar, ordenar, agrupar o comparar [10]. Sin embargo, poder analizar datos y explorar el orden en el procesamiento de tareas es

un requisito importante. Desafortunadamente, la mayoría de las herramientas existentes no se enfocan en la MP, u ofrecen sólo funcionalidad básica de análisis de registros basada en algunos valores agregados y estadísticos. Para un análisis de tipo OLAP tales funcionalidades son insuficientes.

1.4 Cubos de proceso

Dado que se necesitan nuevas formas de visualización de procesos que sean más adecuadas para representar y analizar procesos, en especial los no estructurados, se han realizado algunas propuestas, entre las que se destacan los CP.

Un CP es una estructura multidimensional construida a partir de *logs* de manera que facilite un análisis de MP más exhaustivo. Está compuesto por un conjunto de celdas de proceso que no permite la superposición de eventos en sus celdas y de cada una se puede tener un modelo de proceso descubierto a partir del *log* exclusivo de la propia celda [65].

Los autores de [66] plantean dos diferencias significativas entre cubos OLAP y de procesos: el resumen (*summarizability* en inglés) y la representación. La primera asegura la corrección de las operaciones de resumen sobre la base de datos OLAP [67]. Estas bases de datos se denominan generalmente "bases de datos resumidas" y tienen un modelo de datos basado en una o más medidas definidas sobre el producto cruzado de dimensiones. Una dimensión en una base de datos de resumen es resumible en relación con una medida, si una estadística de resumen (suma, promedio, etc.) aplicada sobre la dimensión produce resultados correctos [68]. Esta propiedad no se garantiza en el CP debido a la naturaleza orientada a proceso de los datos de eventos que usa el mismo. Cada evento se relaciona a una o más trazas y la relevancia de un evento como dato está dada mayormente por sus relaciones con otros eventos dentro de esas trazas. No se puede simplemente unir o separar celdas de CP para resumir. Debido a que las celdas se encuentran ordenadas, cualquier pequeño cambio en ese orden puede cambiar la representación completa de la celda donde se encuentra el evento. La segunda se refiere a la confiabilidad de OLAP en la agregación de hechos para reducir el conjunto de valores a un simple valor que puede ser representado de muchas maneras. Los CP manejan representación de datos mucho más compleja. Las celdas de CP se asocian a modelos de procesos y no solo a los datos de los eventos, y ambas se encuentran directamente relacionadas.

Además, Mamaliga (2013) plantea que su principal diferencia con un cubo OLAP radica en las características de sus celdas [4]. A diferencia de los cubos OLAP, no existe una medida real del interés de cuantificar una operación de negocio. Mientras las estructuras OLAP están diseñadas para análisis de operaciones de negocio, los CP optan por el análisis de procesos. Por tanto, cada dimensión de análisis está compuesta por atributos de eventos. El contenido de una celda cambia de números reales en el cubo OLAP a eventos en el CP. Mientras en OLAP, las dimensiones de análisis son usadas para poblar el cubo, en los CP los eventos de un *log* son usados para crear dichas dimensiones.

Una dimensión describe los eventos de una perspectiva a través de cualquier combinación de sus atributos (los atributos ciudad y país pueden describir un lugar), donde los atributos describen la perspectiva desde un mayor o menor nivel de detalle (ciudad describe un lugar en un nivel más detallado que país). El usuario puede definirla como desee. Todas las dimensiones son independientes entre sí, es decir, no poseen atributos en común [66].

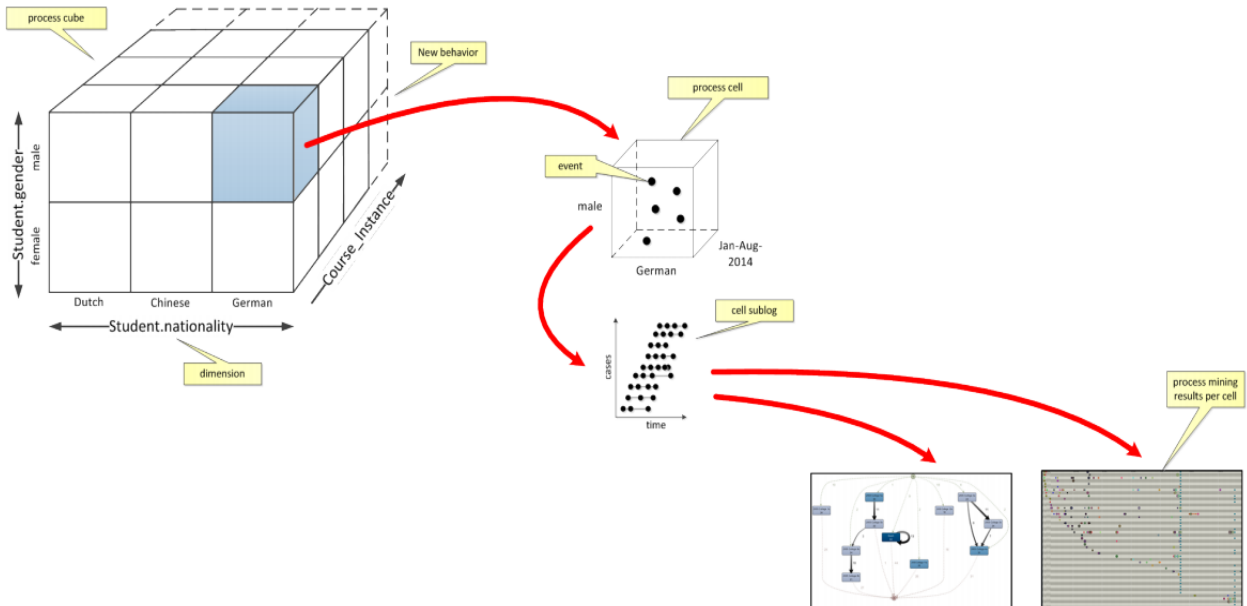


Figura 2. Cubo de proceso. Fuente: [20].

En la figura anterior se muestra un CP cuya dimensión vertical es el género de los estudiantes, la horizontal muestra su nacionalidad y la de profundidad muestra el curso escolar dado en tiempo. De esta forma se pueden entender las diferencias, por ejemplo, entre los estudiantes de género femenino y masculino, de diferentes nacionalidades o cursos.

Además, cada dimensión puede tener una estructura jerárquica natural (por ejemplo, un año está compuesto por meses y un país por regiones) que pueden explotarse para la agregación, refinamiento y selección de datos de eventos.

El CP proporciona a los usuarios la oportunidad de analizar y explorar los procesos de forma interactiva sobre la base de una visión multidimensional de los datos de eventos. En principio, también pueden usarse para descomponer problemas de MP difíciles en problemas más pequeños usando las técnicas descritas en [69], [70] y [71].

1.4.1 Operaciones en los cubos de proceso

El CP trabaja con los operadores *slice*, *dice*, *roll-up*, y *drill-down*. Sus celdas almacenan instancias del proceso y a ellas se les aplican los algoritmos de descubrimiento [72].

La operación *slice* produce un nuevo cubo “rebanado” al permitir al analista filtrar o escoger un valor específico de atributo en una de las dimensiones. De esta forma se remueve esa dimensión de la parte visible del cubo (es un caso especial de la operación *dice*) y se obtienen los datos del atributo seleccionado en la dimensión. Por ende, si se desea, hipotéticamente, obtener los datos de ventas de una entidad en “La Habana”, se puede cortar el cubo para esa ubicación. Es decir, la dimensión de ubicación se elimina del cubo y solo se consideran las ventas de las tiendas en La Habana, de manera que se realiza una selección sobre la dimensión “lugar”.

La operación *dice* produce un subcubo. Permite elegir valores específicos de varias dimensiones. En este caso, a diferencia de *slice*, no se remueve ninguna dimensión, sino que se consideran solo los valores seleccionados. Por ejemplo, para cortar las ventas del cubo para las ubicaciones "La Habana" y "Matanzas", no se eliminan las dimensiones, sino que se consideran las ventas de esas ciudades. Ambas operaciones se muestran en la siguiente figura.

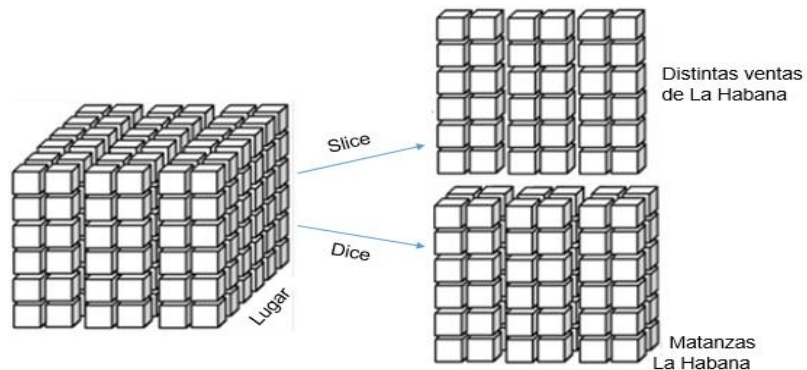


Figura 3. Operaciones *slice* y *dice*. Fuente: Elaboración propia.

El objetivo de las operaciones *roll-up* y *drill-down* es mostrar los mismos datos, pero con más o menos detalle [66]. No eliminan dimensiones, solo les cambian su nivel de granularidad.

Por ejemplo, antes de aplicar *drill-down* a la dimensión tiempo, esta estaría formada de la siguiente manera: Dimensión (tiempo) = {T2015; T2016; T2017} y después de aplicar dicha operación al 2016 quedaría: Dimensión (tiempo) = {T2015; T-Ene-2016; T-Feb-2016;...; T-Dec- 2016; T2017}. Esta operación equivale a dividir (*split* en inglés) las dimensiones.

La operación *roll-up* es lo contrario, equivale a agrupar (*merge* en inglés) (ver siguiente figura).

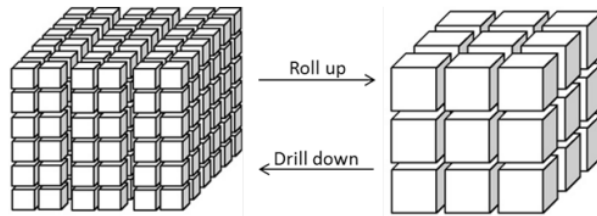


Figura 4. Operaciones *roll-up* y *drill-down*. Fuente: [66].

1.4.2 Investigaciones referentes a los cubos de proceso

La propuesta de Aalst (2013) define tres etapas para conformar un CP [10]:

- Base de Eventos (*Event Base*, EB): *log* que abarca todo o la unión de una colección de *logs* relacionados. Es el conjunto de atributos que se podrían usar como dimensiones del cubo.
- Estructura del CP (*Process Cube Structure*, PCS): define las dimensiones (esquema) del cubo.
- Vista de CP (*Process Cube View*, PCV): define las dimensiones visibles y los eventos seleccionados. Se define a través del PCS y el EB.

En la propuesta de Mamaliga (2013) se desarrolló un prototipo inicial utilizando el marco de trabajo ProM [73] y [74] y el conjunto de herramientas Palo OLAP (cliente JPalo y servidor Palo MOLAP) por sus capacidades OLAP. La aplicación crea *sublogs* por celda y visualiza modelos de procesos descubiertos usando algunos complementos del ProM. Funciona a partir de la creación de la PCS con Palo OLAP Server, luego aplica las operaciones OLAP y finalmente materializa las celdas [4].

La materialización consiste en extraer un *sublog* basado en los datos del CP o partes de él (celdas de proceso o grupos de celdas). Los *logs* resultantes sirven de entrada a diferentes algoritmos de MP. El resultado es la visualización de un conjunto de modelos de proceso.

La propuesta de Gupta and Sureka (2014) propone una metodología para comparar celdas segmentada en tres etapas [75]:

- *Extracción y transformación de los datos*: su objetivo es extraer los datos de diferentes repositorios para integrarlos en una EB que contiene dos tipos de atributos: de caso (propiedades asociadas al caso que se mantienen constantes para todos los eventos pertenecientes al mismo caso) y de eventos (propiedades que describen a cada evento).
- *Visualización de los CP y operaciones OLAP*: para la visualización se aplican técnicas como Agrupamiento, Jerarquías y Selección de Sub-Conjuntos y luego se pueden aplicar las operaciones OLAP para obtener las celdas deseadas.
- *Aplicación de la MP*: a las celdas obtenidas se le aplican las diferentes técnicas de MP para obtener los modelos de las celdas resultantes y poder compararlos.

Dicha propuesta permite solamente obtener representaciones gráficas de las celdas obtenidas, pero no informaciones sobre los valores de los atributos de dichas celdas, de manera que no se pueden comparar esos datos entre las celdas obtenidas.

Por su parte, Arias et al. (2015) proponen un marco de trabajo flexible para asignar dinámicamente los recursos más apropiados para ejecutar los subprocesos [76]. Los resultados son experimentales y proponen aplicarlo a datos reales para validar la técnica. Esta línea de investigación constituye un reto en la BPM, debido a que una errónea asignación puede conducir al uso incorrecto del recurso, altos costos o pobre rendimiento del proceso. No es objetivo de la presente investigación porque no abarca la comparación entre celdas.

La propuesta de Hompes et al. (2015) introduce una técnica para la clusterización de trazas capaz de encontrar variantes de proceso y desviaciones de comportamiento [77]. Con ello se pueden crear modelos de procesos más estructurados, analizarlos y mejorarlos, de manera que las desviaciones positivas, por ejemplo, puedan ser más generalizadas hacia otros casos, y las negativas puedan ser detectadas y modificar los procesos para evitar comportamientos indeseados en el futuro. Aunque esta técnica se enfoca en segmentar las trazas, similar en cierta forma a la idea de los CP, la vía para realizarlo dista del objetivo de los CP.

En la investigación de Bolt and Aalst (2015) se implementa el complemento llamado Process Mining Cube (PMC), que posee dos funcionalidades: división de *logs* (creación de celdas) y generación de resultados (convertir cada celda en un modelo de proceso) [66]. Primero se crea la BE, luego la PCS. Se chequea si estos son compatibles. Si lo son, se crea la PCV, que no cambia una vez que ha sido definida. Al aplicarle las operaciones OLAP solo se cambia la manera de visualizarse el cubo y su contenido, ya que la PCV define la parte visible de la PCS. Una vez seleccionada la parte de la PCS y el conjunto de celdas definido como la parte visible del cubo, se le agregan los eventos a esas celdas para que puedan ser usadas por los algoritmos de la MP.

Los autores de esta propuesta afirman que aún quedan pendientes temas por resolver en cuanto a los CP, destacando la comparación de celdas y el cambio de tendencia (*concept drift* en inglés). Esta última es una característica distintiva del flujo de datos: que son variables en el tiempo, opuesto a los sistemas de BD tradicionales, donde solo guardan el estado actual [78]. Esto significa que el proceso puede cambiar mientras está siendo analizado.

La propuesta de Aalst et al. (2015) aplica los CP para analizar datos de estudiantes de la Universidad Tecnológica de Eindhoven, con el objetivo de establecer comparaciones en cuanto a procedencia, comportamiento, género y resultados docentes [79]. Como resultado muestran los modelos generados de las celdas seleccionadas. Dichos modelos no presentan un orden o alineación en las actividades,

de manera que un nodo (actividad) puede aparecer en una posición distinta en cada modelo, lo que dificulta ver, por ejemplo, semejanzas y diferencias.

Los autores Vogelgesang and Appelrath (2015) introducen el PMCube Explorer, una herramienta de la MP multidimensional que permite el análisis de procesos a partir de varias vistas [80]. Su funcionamiento se rige por los siguientes pasos:

- *Log* multidimensional: organiza los atributos de eventos y de casos en diferentes niveles, tal que, mientras los CP almacenan un conjunto de eventos en cada celda, PMCube almacena casos en el primer nivel en sus celdas. En el segundo nivel cada caso tiene una secuencia de eventos (trazas) que forman un cubo distinto.
- Selección de los datos con operaciones OLAP: esta estructura de cubos anidados permite definir operaciones de filtrado y agregación, usando consultas OLAP, para extraer subconjuntos de eventos personalizados.
- MP: a los mismos se le aplican técnicas de MP para descubrir un modelo de proceso.
- Consolidación: consiste en seleccionar automáticamente los modelos más relevantes, especialmente cuando se seleccionó un conjunto de eventos muy grande y el análisis del resultado se torna complicado.
- Visualización: finalmente, los resultados son visualizados.

Esta herramienta permite destacar las diferencias entre dos modelos descubiertos, pero fue implementada en C# usando Microsoft .NET, que son tecnologías que no contribuyen al libre desarrollo de la MP.

También Bolt et al. (2015) aplican técnicas de MP para analizar el rendimiento de los estudiantes de la Universidad Tecnológica de Eindhoven [81]. Para relacionar el comportamiento con el rendimiento de los mismos usan la herramienta PMC de [66], que les permite analizar los datos según cursos y grados. El objetivo de la propuesta es mostrar cómo se pueden combinar los CP y el flujo de trabajo analítico en el campo de la MP, por tanto, no proponen elementos novedosos en cuanto a los CP, sino que utilizan una herramienta y exponen los resultados.

Los autores Vogelgesang et al. (2015) analizaron los requerimientos para la aplicación de la MP multidimensional, limitantes y retos pendientes [82]. Plantean que entre las principales limitantes se encuentran el alto esfuerzo para la integración de datos (por ejemplo, definir la jerarquía de la dimensión), la optimización del desempeño (debido a temas de materialización), el manejo del cambio de tendencia y la comparación de celdas, que expresan continúa siendo el mayor reto de la MP multidimensional, pues cuando el número de celdas aumenta se dificulta mucho su análisis.

En la investigación de Schumacher et al. (2017) se aplican los CP para analizar los datos concernientes a miles de jugadores en Nueva Zelanda durante los años 2013-2014 y ayudar a caracterizarlos psicológicamente [83]. Para ello se auxilia además de la MD y de técnicas estadísticas confirmatorias [84]. Esta propuesta se enfoca en la multidisciplinariedad y no en la comparación entre celdas.

Por último, los autores Bayerl and Granitzer (2015 y 2017) proponen algoritmos para fusionar CP [85] y [86]. Se basan en propiedades sintácticas y semánticas para evaluar similitudes. Abarcan las jerarquías estructurales, enlace de dimensiones y similitudes de palabras, para lo cual usan aprendizaje automático. Esta propuesta tampoco responde al objetivo de la presente investigación.

1.4.3 Resultado del análisis de las investigaciones realizadas sobre los cubos de proceso

En el análisis bibliográfico realizado se observaron algunas limitaciones que poseen los CP [4] y [10]:

- Problema de la explosión del número de celdas (dispersión). Un CP que excede un número de dimensiones, con un gran número de elementos por dimensión y con muchos valores de celdas faltantes, conduce a problemas de dispersión y altos tiempos de ejecución.
- Dificultad de visualizar estructuras hipercubo. La multidimensionalidad no es la forma natural en la que las personas ven y resulta difícil proporcionar visualizaciones multidimensionales en más de tres dimensiones.
- Carencia de flexibilidad con respecto a las posibilidades de filtrado de registros y el cambio de dimensiones en el cubo. El usuario selecciona atributos para formar las dimensiones del CP, mientras que el resto de los atributos y otra información de registro se almacenan en BD relacionales. Esto limita las posibilidades de filtrado de logs. Además, cambiar una dimensión del cubo implica la creación de un nuevo CP, seleccionando todas las dimensiones de nuevo.
- Dificultad para visualizar múltiples resultados de celdas. El ajuste de múltiples resultados en una sola pantalla puede perjudicar la visualización de los resultados y, por lo tanto, impedir la comparación entre las celdas (dimensionalidad).
- Comparar y visualizar diferentes celdas [10], [4], [75], [82] y [66]. A diferencia de los valores numéricos mostrados en cubos OLAP, es necesario visualizar modelos que no se pueden reducir a números. Los autores Aalst et al. (2015) enuncian, entre otras ideas, la opción de descubrir los modelos de procesos de varias celdas “todo en uno”, utilizando colores y notas para diferenciar los elementos de cada uno [79]. La comparación visual de estos modelos sigue siendo un problema de investigación en el área.

Al realizar un análisis de las investigaciones del epígrafe anterior, se observa que [82] es teórica. Analiza los requerimientos para aplicar CP, limitantes y retos pendientes, entre los que destaca la comparación de celdas. La propuesta [77] no se relaciona directamente con los CP, sino que constituye

una técnica para segmentar trazas, persiguiendo al final el mismo objetivo: obtener modelos a partir de segmentos de trazas, de manera que sean más comprensibles. Las investigaciones [83], [79] y [81] son resultados de aplicar CP, por tanto tampoco constituyen referentes a analizar. Asimismo, [10] no desarrolla ninguna herramienta, sino que presenta teóricamente la noción de CP de la herramienta desarrollada en [4], que posee una pobre visualización [79]. En [76] se propone un marco de trabajo de modo experimental que debe validarse, mientras que en [75] y [66] se presentan herramientas para generar CP. En [85] y [86] se propone fusionar CP, lo que no se corresponde con el tema abordado en la investigación.

Solo [80] permite destacar diferencias entre modelos descubiertos, pero posee dos inconvenientes. El primero es que almacena casos en el primer nivel en sus celdas, mientras que los CP almacenan conjuntos de eventos, lo que demanda un tratamiento diferente. El segundo es que es software propietario y no desarrollado para la comunicad de investigadores de la MP.

En general, todas las propuestas que generan herramientas para conformar CP permiten generar modelos de proceso a partir de celdas materializadas. Sin embargo, ninguna permite obtener información extra que contribuya al diagnóstico de procesos.

1.5 Ambiente tecnológico para la solución

ProM posee una arquitectura orientada a componentes que soporta la adición de complementos. De esta manera constituye un entorno tecnológico para desarrollar nuevos algoritmos y técnicas de MP. Estos complementos implementan algoritmos para obtener y almacenar los modelos, importar objetos, analizar y convertir los resultados. Es una herramienta libre, de código abierto, multiplataforma y está implementada con Java. Es compatible con los estándares de *logs* MXML (*Mining eXtensible Markup Language*) y XES (*eXtensible Event Stream*) [87]. Proporciona un marco de trabajo común que permite reutilizar funcionalidades, simplifica el desarrollo de interfaces de usuario y ofrece un gran conjunto de funcionalidades disponibles en un solo lugar, además de una amplia variedad de tipos de modelo. Un complemento se puede aplicar a cualquier subconjunto de objetos en ese grupo como dato de entrada, y puede transferir los objetos que crea de regreso al grupo. Esto permite a los complementos aprovechar la funcionalidad de otros complementos, ejecutándose en secuencia.

Posee más de 1500 técnicas [88], métricas, parámetros de configuración y procedimientos que resultan de alta complejidad en general. Su aplicación es limitada para profesionales no expertos [89]. Sin embargo, se ha convertido en la herramienta por excelencia en la MP, es la herramienta académica de MP más usada y libremente descargada [62].

Se utilizará Java como lenguaje de programación, característico en las técnicas desarrolladas para el ProM [90] y [91].

Como gestor de BD se usará MySQL, que es el sistema gestor de BD más popular. Entre sus características se encuentran que es uno de los gestores con mejor rendimiento debido a su velocidad, consume pocos recursos, es altamente confiable, estable, fácil de usar, multiplataforma, gratuito, de código abierto y soporta grandes volúmenes de datos [92], [93]. Es un sistema ampliamente probado con alto éxito.

Eclipse se usará como entorno de desarrollo integrado, ya que es una potente plataforma de programación, desarrollo y compilación de aplicaciones Java. Es de código abierto y multiplataforma. Además, permite realizar pruebas unitarias con JUnit [94] y brinda aserciones para probar los resultados esperados y corredores de prueba [95].

Palo [96] es un motor MOLAP que permite crear y almacenar datos, así como generar informes. Se basa en componentes de código abierto y trabaja fundamentalmente con grandes volúmenes de datos, permitiendo a varios usuarios compartir un almacenamiento centralizado de dichos datos [97]. Palo Suite está compuesto por Palo para Excel, Palo ETL Server, Palo Web y Palo OLAP Server. Este último, usado en [4], presenta tecnología “en memoria” que organiza los datos en la BD de Palo en forma cubo, dimensiones, elementos y atributos de elementos.

1.6 Conclusiones del capítulo

En este capítulo se abordaron los elementos teóricos que propician el sustento teórico metodológico de la investigación. Como resultado del estudio realizado se concluye lo siguiente:

- La MP constituye una vía para apoyar el diagnóstico de procesos de negocio. Propicia un análisis basado en evidencias reales su ejecución, a diferencia de la modelación de procesos donde se describe cómo el proceso debe funcionar.
- Los cubos OLAP se basan mayormente en el almacenamiento y análisis de datos en su mayoría históricos, a diferencia de los CP, que se basan en eventos y análisis de procesos.
- El enfoque de CP es novedoso y permite analizar y explorar los procesos basado en una visión multidimensional de los eventos. Del estudio de las dificultades que presentan sus técnicas, para esta investigación se decide abordar la comparación entre celdas de proceso.
- La tecnología seleccionada permite el desarrollo de la solución (MySQL y Palo para la gestión de BD, Eclipse como entorno de desarrollo integrado, Java como lenguaje de programación y ProM como marco de trabajo).

CAPÍTULO 2: PROPUESTA DE SOLUCIÓN

Introducción

En este capítulo se presenta la propuesta de solución de la investigación. Para ello se define y describe el funcionamiento de sus componentes. Se divide la solución en cuatro etapas, de las cuales se detallan sus entradas, salidas y funcionalidades. Por último, se presentan algunas interfaces de la solución desarrollada, que permite conformar un CP a partir de un *log*, aplicarle las operaciones *slice*, *dice*, *roll-up* y *drill-down* para obtener celdas de proceso y visualizar modelos de procesos y tablas de datos sobre los mismos.

2.1 Conceptualización de la solución

En la siguiente figura se muestra el modelo conceptual referente a la presente investigación y a continuación se describen sus conceptos y relaciones. En lo adelante se presentan las definiciones relacionadas con la propuesta de solución, con la notación D# (ejemplo: D1).

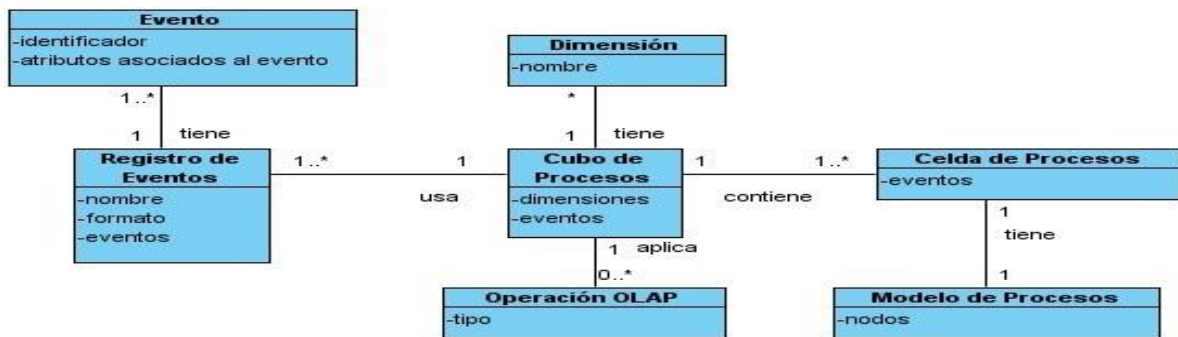


Figura 5 Modelo conceptual referente a la presente investigación. Fuente: Elaboración propia.

Evento: actividad (paso bien definido) que se relaciona con un caso particular o instancia de proceso [50]. Cada evento posee atributos definidos, por ejemplo: identificador y tiempo.

Registro de Eventos (*log*): registro cronológico de actividades (eventos) de un sistema informático que son guardadas en un archivo en el sistema [18]. Constituye la entrada de las técnicas de descubrimiento de la MP. Debe estar en formato XES.

Dimensión: atributo del *log* que representa una dimensión del cubo.

Cubo de proceso: posibilita un análisis efectivo de los modelos de grandes procesos desde perspectivas diferentes y con múltiples granularidades [98].

Operación OLAP: operación que retorna un subcubo [4].

- **Dice:** selecciona un subconjunto de miembros en determinadas dimensiones.
- **Slice:** selecciona un miembro de alguna de sus dimensiones.

- **Roll-up:** consolida algunos elementos de una dimensión en un solo elemento correspondiente a su nivel jerárquico superior.
- **Drill-down:** refina un miembro de una dimensión en un conjunto de miembros correspondientes a un nivel jerárquico inferior.

Celda de Proceso: subcubo obtenido al aplicar operaciones OLAP a las dimensiones del CP.

Modelo de proceso: representación de un esquema teórico simplificado a partir de técnicas y herramientas especializadas, de las evidencias reales de ejecución de un sistema o de una realidad compleja, con el fin de contribuir a su comprensión y al análisis de su comportamiento [99].

Para definir la técnica propuesta, se analizaron las propuestas de [66], [10], [4] y [75].

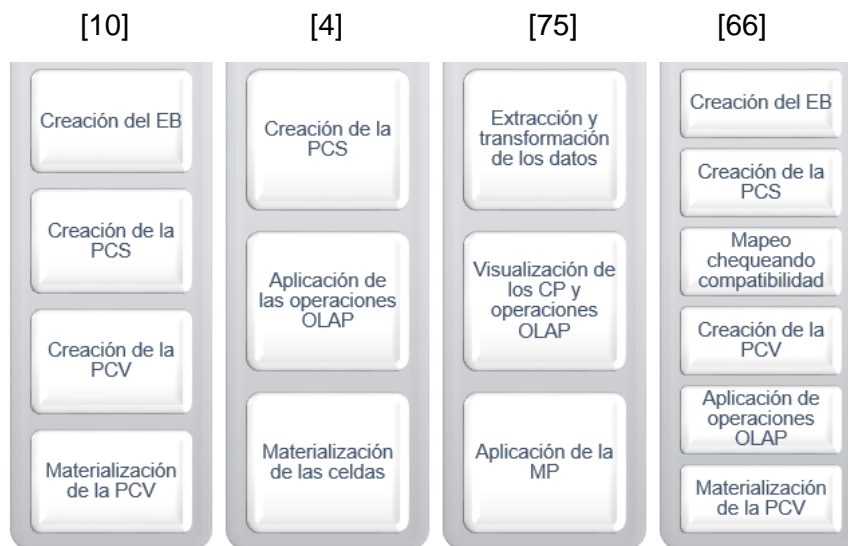


Figura 6 Etapas para conformar CP según varios autores. Fuente: Elaboración propia.

En general, las propuestas analizadas primero preparan los datos, luego definen las dimensiones que conformarán el CP y finalmente lo generan. En [66] se define el chequeo de compatibilidad como una etapa, aunque en [10] también realizan esta actividad una vez terminada la segunda etapa. Ambas investigaciones definen formalmente qué son el EB, PCS y PCV:

EB: se crea a partir de un *log*. Es el conjunto de atributos que se podrán utilizar como dimensiones del cubo.

PCS: se caracteriza por el conjunto de dimensiones definidas, donde cada dimensión tiene su propia jerarquía.

PCV: define las dimensiones visibles y los eventos seleccionados a partir de la EB y PCS.

La propuesta de solución, que parte del análisis anterior, tiene como objetivo visualizar modelos de celdas de proceso y mostrar datos de las mismas, siendo esta última una nueva funcionalidad, de modo que permitan compararlas en cuanto a atributos seleccionados. Sus **funcionalidades** son:

1. Importar *logs* en formato *.xes*.
2. Construir el CP con el *log*.
3. Seleccionar atributos del registro para filtrar las dimensiones del cubo a formar.
4. Ejecutar las operaciones *slice* y *sice* sobre el CP para obtener un nuevo cubo.
5. Ejecutar las operaciones *drill-down* y *roll-up* sobre el CP detallar análisis.
6. Visualizar modelos de procesos a partir de las celdas de los CP en ventanas paralelas usando el algoritmo *Alpha Miner*, que visualiza el resultado en forma de redes de Petri.
7. Visualizar tabla con datos de celdas del CP teniendo en cuenta varios indicadores (por ejemplo: tiempo, costo y recurso).

2.2 Estructura general de la propuesta de solución

En la siguiente figura se muestra la estructura de la propuesta de solución. Una descripción de sus componentes, funciones y sus relaciones se expone a continuación.

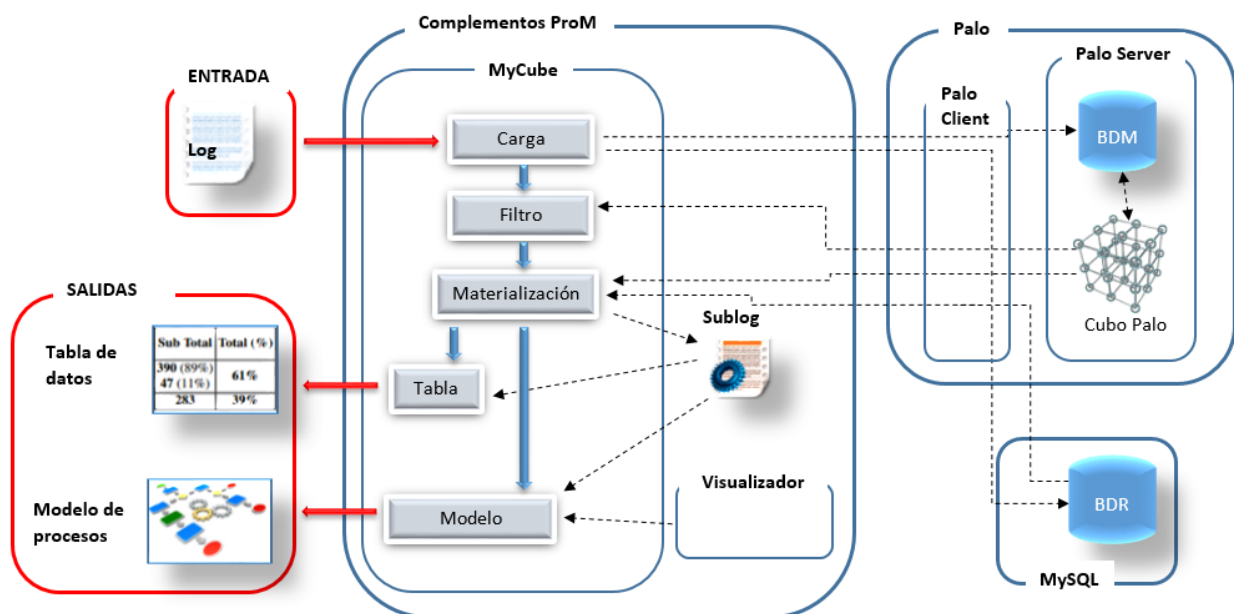


Figura 7 Estructura de la propuesta de solución. Fuente: Elaboración propia.

BD relacional de MySQL (BDR a partir de ahora): se usa para almacenar datos de los *logs* que pueden no ser relevantes para el procesamiento multidimensional.

Palo: ofrece la posibilidad de realizar las operaciones OLAP. Se divide en:

- Palo Server: genera (1) BD multidimensional en memoria (BDM a partir de ahora) para almacenar los datos relevantes para el procesamiento multidimensional, como el caso de las jerarquías. De esta manera aumenta la velocidad, dado que la BD en memoria es más rápida que la almacenada

en disco. Además, crea (2) un cubo Palo a partir de los datos residentes en la BDM para soportar las operaciones OLAP.

- Palo Client: permite realizar las operaciones OLAP sobre los *logs*.

ProM: utiliza al algoritmo *Alpha Miner* para generar los modelos de proceso una vez que se tiene materializada la celda de proceso. Contiene la técnica desarrollada MyCube, cuyos componentes son:

- Carga: procesa y carga los *logs* en la BDM y en la BDR.
- Filtro: permite seccionar el cubo a partir de dimensiones y operaciones OLAP seleccionadas.
- Materialización: usa los fragmentos filtrados del cubo para descargar sus *logs* correspondientes desde ambas BDs. Se genera un *log* materializado que sirve como entrada a los complementos del ProM.
- Modelo: muestra los modelos generados en una interfaz para facilitar la comparación.
- Tabla: genera la tabla con los datos numéricos sobre las celdas seleccionadas.
- Visualizador: permite vincularse al complemento *Alpha Miner* del ProM para generar un modelo de proceso.

2.3 Descripción operacional de la propuesta de solución

La entrada del complemento es un *log* que se carga en la BDR y en la BDM. Con los datos residentes en la BDM se crea el cubo Palo. A partir del cubo es posible seleccionar dimensiones para formar un subcubo y realizar operaciones OLAP. Estas operaciones se realizan con el objetivo de analizar el cubo desde diferentes perspectivas. Según la operación y selección de información a filtrar, se obtienen los datos de los eventos seleccionados desde las BDs y se crea un *log* con esos datos. Dicho registro constituye la entrada para el ProM donde, a partir de un complemento, se obtiene un resultado visual.

Precondición: el *log* debe tener el formato XES.

Entradas y salidas

Entrada: un registro de eventos en formato XES.

Salidas: (1) dos modelos de procesos en una misma ventana, que permiten visualizar y comparar la ejecución de los mismos. (2) Una tabla generada a partir de los fragmentos de *logs* obtenidos durante el filtrado, que muestra datos para poder analizar y comparar los procesos.

Objetivo y función principal: la herramienta permite seccionar un *log* a conveniencia del usuario, para generar modelos de procesos y tablas de datos que permiten analizar y comparar los mismos, brindando soporte al diagnóstico de procesos y a la toma de decisiones.

Descripción de la propuesta de solución

La técnica desarrollada y materializada en el componente se compone de cuatro etapas en su funcionamiento. A continuación, se describe cada etapa teniendo en cuenta sus entradas, salidas, descripción general, esquema y algunas definiciones.

1. Cargar datos

Entrada: *log* con formato XES.

Salidas: EB (BDM, BDR y cubo Palo).

Descripción: la etapa consta de tres pasos.

1. Primeramente, el componente Carga recibe el *log* con formato XES. Almacena en la BDR los datos de los eventos, incluyendo las propiedades que no sean dimensiones del cubo Palo que contendrá toda la información original. De esta manera crea una copia de respaldo para realizar diversas operaciones. Para la técnica se considera un evento y un *log* como se definen en [13]:

D1. *Evento:* Sean C y U el universo de características y de posibles valores respectivamente. Un evento e es una asignación de valores a características, es decir, $e \in C \rightarrow U$.

D2. *Traza, Registro de Eventos:* Una traza $t \in E^*$ es una secuencia de eventos. Sea $T = E^*$ el universo de trazas. Un registro de eventos L es un conjunto múltiple de trazas, es decir, $L \in B(T)$, donde $B(T)$ denota el conjunto de todos los posibles conjuntos múltiples sobre T .

2. El componente Carga almacena el *log* también en la BDM, que contendrá los datos transformados para facilitar el proceso.

3. El Palo Server se encarga de crear el cubo Palo con los datos residentes en la BDM. Para ello, las trazas y eventos se añaden a un árbol jerárquico junto con sus atributos. De la raíz parten todos los enlaces del árbol. El primer nivel contiene las propiedades de los casos y eventos, el siguiente nivel contiene los valores de dichas propiedades. A partir del árbol se crea el cubo Palo seleccionando todos los atributos del *log* como dimensiones (D4). Para disminuir su tamaño, en lugar de cargar todos los datos de los eventos, a cada celda se le asigna un identificador único, siendo esa la información almacenada en el cubo. El Palo Server permite trabajar con jerarquías (D3) que pueden emplearse para reducir el número de miembros por dimensión. El tiempo, por ejemplo, que puede expandirse mucho debido a que cada evento puede tener un tiempo diferente y, a su vez, este se puede dividir en año, mes, día, etc., se divide y se almacena cada valor de la división en la BDM, mientras que el tiempo real con toda su información, se almacena en la BDR.

D3. *Jerarquía:* Grafo no dirigido, en el cual dos nodos cualesquiera están conectados por un camino simple, donde se cumple que para cada nodo $h \in H_i$, dos hijos cualesquiera $h_1, h_2 \in \text{hijos}(h)$, $\text{todasHojas}(h_1) \cap \text{todasHojas}(h_2) = \emptyset$ [4].

D4. *Dimensión*: Sea $CD = D_1 \times D_2 \times \dots \times D_n$. Para cualquier $1 \leq i \leq n$, $D_i \subseteq H_i$ es un conjunto de elementos dimensiones [4], donde CH define la jerarquía.

D5. *Cubo de Proceso* [4]: Sean:

D_i , $1 \leq i \leq n$ un conjunto de elementos dimensiones, donde n es el número de dimensiones,

H_i , $1 \leq i \leq n$ un conjunto de elementos jerárquicos,

$CD = D_1 \times D_2 \times \dots \times D_n$ las dimensiones del cubo,

$CH = H_1 \times H_2 \times \dots \times H_n$ las jerarquías del cubo,

hijos : $H_i \rightarrow P(H_i)$, donde hijos(h) es la función que devuelve los hijos de $h \in H_i$,

todasHojas : $H_i \rightarrow P(H_i)$, donde todasHojas(h) es la función que devuelve todas las hojas de $h \in H_i$,

$h \in H_i$, $h_1, h_2 \in \text{hijos}(h)$, $\text{todasHojas}(h_1) \cap \text{todasHojas}(h_2) = \emptyset$,

$CS = (CD, CH)$ la estructura del CP,

$CE : CH \rightarrow P(E)$ la función de los eventos de la celda, $CE(h_1, h_2, \dots, h_n) = \{e \in E \mid (d_1, d_2, \dots, d_n) = CC(e), d_i \in \text{todasHojas}(h_i), 1 \leq i \leq n\}$, para $(h_1, h_2, \dots, h_n) \in CH$.

Un CP se define como $PC = (CS, CE)$ [4].

D6. *Celda de Proceso*: Sea PC, $PC = (CS, CA)$. La celda de proceso es $\text{slice}_{1,v_1} (\text{slice}_{2,v_2} \dots (\text{slice}_{n-1,v_{n-1}} (\text{slice}_{n,v_n}(PC))) \dots) = PC'$. Cada celda en el CP corresponde a un conjunto de eventos retornados por la función de los eventos de la celda CE [4].

Los tres artefactos generados en esta etapa constituyen la EB (D7) y son generados a través del componente Carga. Además, constituyen la entrada para el componente Filtro.

D7. *Base de Eventos (EB)*: Una EB = (E, P, π) define un conjunto de eventos E, un conjunto de propiedades de eventos P y una función $\pi \in P \rightarrow (E \rightarrow U_v)$, siendo U_v el universo de todos los valores de atributos. Para cualquier propiedad $p \in P$, π_p es una función parcial de mapeo de eventos a valores. Si $\pi_p(e) = v$, entonces el evento $e \in E$ tiene una propiedad $p \in P$ y el valor de dicha propiedad es $v \in U_v$. Si $e \notin \text{dom}(\pi_p)$, entonces el evento e no tiene una propiedad y se denota por $\pi_p(e) = \perp$ para indicarlo [79].

Para lograr un mayor entendimiento, si un log se transforma en una EB, entonces el log E consiste en todos los diferentes elementos de la columna que define el ID del evento de la tabla. El conjunto de propiedades P es el conjunto de los encabezados de las columnas de la tabla. La función π toma el valor de la celda surgida al combinar la fila (evento) y la columna (propiedad).

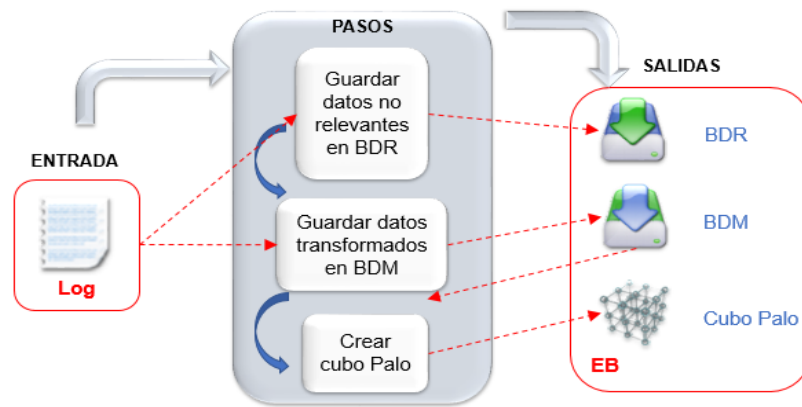


Figura 8 Primera etapa: Cargar datos. Fuente: Elaboración propia.

2. Filtrar

Entrada: EB.

Salida: PCV.

Descripción: el filtrado consta de tres pasos.

1. El componente Filtro recibe la EB como entrada. Selecciona las dimensiones que haya escogido el usuario a partir de los atributos que se poseen. Estos datos se obtienen del cubo Palo creado por Palo Server. Cualquier caso o evento puede ser una dimensión. Una vez realizado este paso, se tiene conformada la PCS.

D8. *Estructura de Cubo de Proceso:* tripleta PCS = (D, tipo, jer), donde:

– D es el conjunto de dimensiones.

– $\text{tipo} \in D \rightarrow U_S$ función que define el posible conjunto de valores para cada dimensión, por ejemplo, $\text{tipo}(\text{edad}) = \{0, 1, 2, \dots, 120\}$ para $\text{edad} \in D$.

– $\text{jer} \in D \rightarrow U_H$ define una jerarquía para cada dimensión tal que para cualquier $d \in D$: $\text{tipo}(d) = U_{\text{jer}(d)}$ [79].

Para esclarecer la definición se define un ejemplo de PCS = (D, tipo, jer) teniendo en cuenta el log aplicado en el primer estudio de caso de la investigación.

– $D = \{\text{id_multa}, \text{tipo_vehículo}, \text{edad}, \text{actividad}, \text{tiempo}, \dots\}$ dimensiones.

– tipo es una función para obtener el conjunto de posibles valores de cada dimensión:

- $\text{tipo}(\text{id_multa}) = \{C11016, C11029, \dots\}$ es el conjunto de identificadores de los casos,
- $\text{tipo}(\text{tipo_vehículo}) = \{A, B, C\}$ es el conjunto de clases de vehículos,
- $\text{tipo}(\text{edad}) = \{18, 19, 20, 21, \dots\}$ es el conjunto de posibles edades,
- $\text{tipo}(\text{actividad}) = \{\text{crear multa}, \text{adicionar sanción}, \dots\}$ es el conjunto de actividades,

- o $\text{tipo(tiempo)} = \{2000-04-08T00:00:00+02:00, \dots\}$ es el conjunto de marcas de tiempo posibles.
- jer es una función para definir la jerarquía de una dimensión:
 - o $\text{jer(id_multa)} = \{\{C11016, C11029, \dots\}, \{C11016\}, \{C11029, \dots\}\}$. El primer elemento es la raíz de la jerarquía. El resto de los valores individuales son las hojas de la jerarquía.
 - o $\text{jer(edad)} = \{\{18, 19, \dots, 78, 79, \dots\}, \{18, 19, \dots, 25\}, \{26\}, \dots\}$. El primer subconjunto representa todas las posibles edades. El segundo representa a los jóvenes y el último representa a una edad específica, constituyendo la hoja de la jerarquía (ver siguiente figura).
 - o jer(tiempo) permite dividir el tiempo en años, meses, semanas, días, etc.

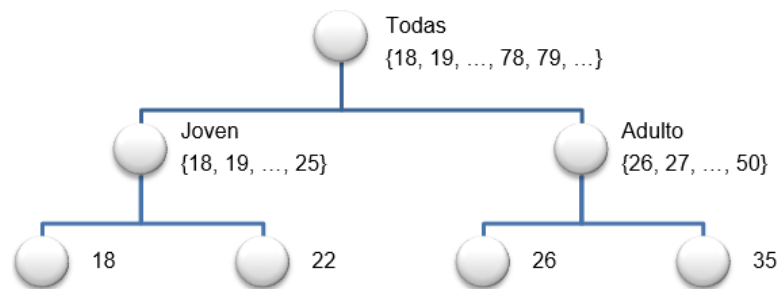


Figura 9 Jerarquía de edad en tres niveles. Fuente: Elaboración propia.

2. Se chequea si la EB y la PCS son compatibles. Esto permite confirmar que las dimensiones corresponden a propiedades y los valores de las propiedades de los eventos son del tipo correcto.
3. Se aplican las operaciones OLAP seleccionadas. Con este paso se crea la PCV. Una PCV usa ambos elementos, EB y PCS, y constituye una vista concreta que puede modificarse utilizando las operaciones OLAP definidas. Una vez que la estructura del CP es definida, esta no cambia. Al aplicarle operaciones OLAP solo se cambia la manera en que se visualiza el contenido del cubo.

D9. *Vista de Cubo de Proceso* [79]: Es un par $\text{PCV} = (D_{\text{sel}}, \text{sel})$ tal que:

- $D_{\text{sel}} \subseteq D$ son las dimensiones seleccionadas.
- $\text{sel} \in D \rightarrow U_H$ es una función que selecciona la parte de la jerarquía considera por dimensión. La función sel es aquella donde para cualquier $d \in D$: $\text{sel}(d) \subseteq \text{jer}(d)$.

Para esclarecer la definición se presenta un ejemplo de $\text{PCV} = (D_{\text{sel}}, \text{sel})$ teniendo en cuenta el primer *log* aplicado en la investigación (ver siguiente figura). $D_{\text{sel}} = \{\text{edad}, \text{tipo_vehículo}, \text{actividad}\}$ son las dimensiones seleccionadas. $\text{sel}(\text{edad}) = \{\{18, 19, \dots, 25\}, \{26, 27, \dots, 50\}, \dots\}$, $\text{sel}(\text{tipo_vehículo}) = \{\{A\}, \{B\}, \{C\}\}$ y $\text{sel}(\text{actividad}) = \{\{\text{crear multa}\}, \{\text{adicionar sanción}\}, \dots\}$.

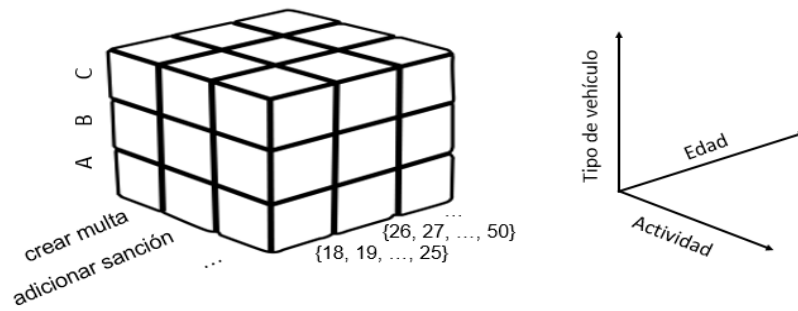


Figura 10 Vista del PCV con tres dimensiones. Fuente: Elaboración propia.

D10. *Dice*: Sean $PCS = (D, \text{tipo}, \text{jer})$ y $PCV = (D_{\text{sel}}, \text{sel})$. Sea $\text{res} \in D_{\text{sel}} \not\rightarrow U_H$ una restricción para cualquier $d \in \text{dom}(\text{res})$: $\text{res}(d) \subseteq \text{sel}(d)$. $\text{dice}_{\text{res}}(PCV) = (D_{\text{sel}}, \text{sel}')$ con $\text{sel}'(d) = \text{res}(d)$ para $d \in \text{dom}(\text{res})$ y $\text{sel}'(d) = \text{sel}(d)$ para $d \in D \setminus \text{dom}(\text{res})$.

Aplicando la operación *dice* no se remueven dimensiones, sino que se restringen conjuntos de valores para una o varias dimensiones.

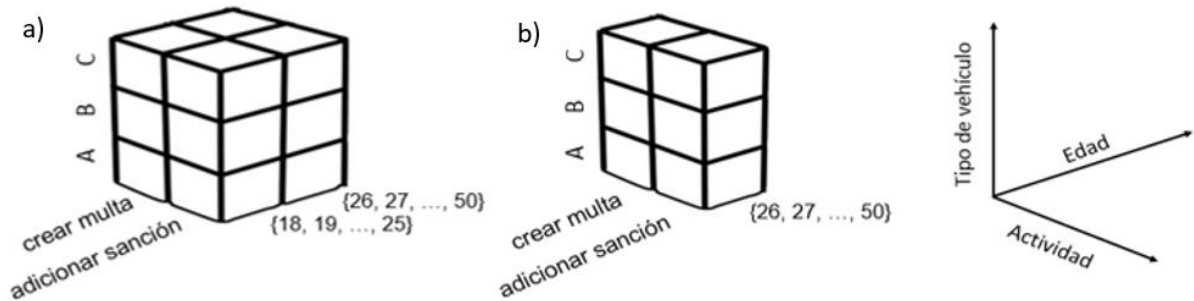


Figura 11 Aplicación de las operaciones a) *dice* y b) *slice*. Fuente: Elaboración propia.

D11. *Slice*: Sean $PCS = (D, \text{tipo}, \text{jer})$ y $PCV = (D_{\text{sel}}, \text{sel})$. Para cualquier $d \in D_{\text{sel}}$ and $H \in \text{sel}(d)$: $\text{slice}_{d,H}(PCV) = (D'_{\text{sel}}, \text{sel}')$ con $D'_{\text{sel}} = D_{\text{sel}} \setminus \{d\}$, $\text{sel}'(d) = \{H\}$, y $\text{sel}'(d') = \text{sel}(d')$ para $d' \in D \setminus \{d\}$.

Sin embargo, la operación *slice* elimina la dimensión “d” seleccionada y se muestran solo los eventos considerados dentro de la jerarquía “H” seleccionada.

La figura anterior muestra el resultado de aplicar las operaciones anteriores al cubo de la Figura 10. En el inciso a) se aplicó un *dice* basado en las dimensiones “actividad” y “edad”. En el inciso b) se aplicó un *slice* basado en la clasificación de “adulto” de la dimensión “edad” ($\text{slice}_{\text{edad},\{26, 27, \dots, 50\}}$).

D12. *Roll-up*: Sea OC , $OC = (CS, CA)$ y $v \in H_k$, donde $1 \leq k \leq n$. La operación *roll-up* es $\text{rollup}_{k,v}(OC) = OC'$, donde $OC' = (CS', CA)$ con $CS' = (CD', CH)$, y $CD' = D_1 \times \dots \times D_{k-1} \times (D_k \setminus \text{hijos}(v)) \cup \{v\} \times \dots \times D_n$.

D13. *Drill-down*: Sea OC , $OC = (CS, CA)$ and $v \in D_k$, donde $1 \leq k \leq n$. La operación *drill-down* es $\text{drilldown}_{k,v}(OC) = OC'$, donde $OC' = (CS', CA)$ con $CS' = (CD', CH)$, y $CD' = D_1 \times \dots \times D_{k-1} \times (D_k \setminus \{v\}) \cup \text{hijos}(v) \times \dots \times D_n$.

El filtrado solo extrae la parte del cubo Palo deseada, no lo modifica. Por consiguiente, se mantiene el cubo Palo con la información original para si se desean realizar otros análisis posteriores. Finalmente, el componente Filtro genera la PCV como salida y entrada del componente Materialización.

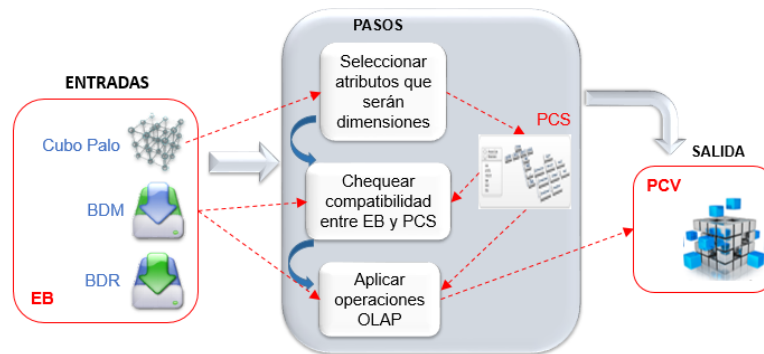


Figura 12 Segunda etapa: Filtrar. Fuente: Elaboración propia.

3. Obtener subcubo

Entradas: PCV, BDR.

Salida: *sublog* materializado.

Descripción: el objetivo de esta etapa es generar el *sublog* materializado.

- El componente Materialización recibe la PCV para materializar las celdas, es decir, crear un *sublog* que contenga solamente los datos enmarcados en los filtros realizados. Para ello se extraen los datos almacenados en la BDR del *log* completo y los de la PCV. De la PCV se obtienen los atributos y de la BDR los valores de los mismos. Estos datos se utilizan junto con los del cubo Palo para crear una lista de eventos, solo con los eventos deseados, que contenga atributos del *log*, de trazas y de eventos. Se comprueba que no haya elementos repetidos, teniendo en cuenta que una traza puede corresponder a varios eventos. Finalmente se obtiene el *sublog* deseado, que sirve de entrada para los componentes Tabla y Modelo.

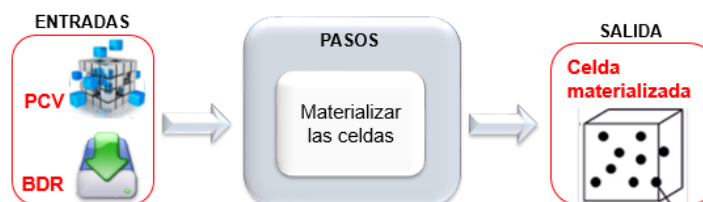


Figura 13 Tercera etapa. Fuente: Elaboración propia.

4. Visualizar

Entrada: *sublog* materializado.

Salidas: modelos de procesos y tabla de datos.

Descripción general: el objetivo de esta etapa es generar los modelos de proceso de las celdas seleccionadas y las tablas con datos que brinden más información para el análisis y diagnóstico de los procesos.

- Se generan los modelos de proceso. Para ello el componente Modelo se integra con el ProM a través del componente Visualizador, para poder acceder a la técnica *Alpha Minery* y utilizarla con los *sublogs* materializados que recibe del componente Materialización como entrada. Se descartaron otros algoritmos porque: (1) necesitan muchas ventanas de configuración (el objetivo es visualizar de forma rápida la información de la celda, realizar un análisis “superficial” al pequeño segmento de *logs* seleccionado, a diferencia de análisis profundos que se pueden realizar con otros complementos y objetivos), (2) poseen componentes visuales con tamaños propensos a generar modelos de proceso grandes, y se necesitan pequeños para poder representar dos modelos en una misma ventana, (3) poseen otros propósitos como filtrar, por ejemplo, que es el objetivo de los CP. El resultado consiste en una interfaz con los modelos de proceso descubiertos.
- Se genera las tablas de datos. Para ello, a partir del *sublog* materializado recibido por el componente Materialización, el componente Tabla se encarga de generar tablas con datos numéricos de las celdas seleccionadas. Muestra operaciones como sumas, promedios y ocurrencias de los atributos de los *sublogs* analizados.

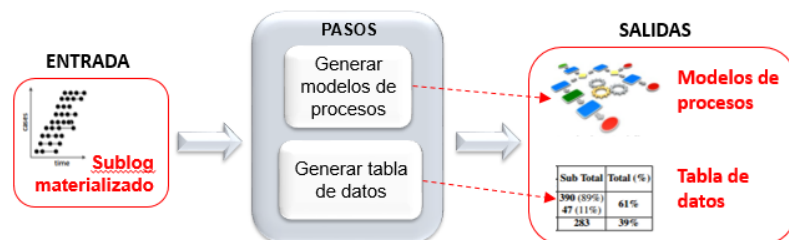


Figura 14 Cuarta etapa. Fuente: Elaboración propia.

2.4 Implementación

La técnica se desarrolló para ser añadida al ProM. Los pasos a seguir para cargarla como un paquete del mismo fueron:

- Exportar el proyecto de eclipse como archivo .jar.
- Copiar en C:\Users\...\ProM64\packages\MyCube el archivo .jar.
- Abrir el archivo packages.xml y poner en paquetes instalados `<package name="..." version="..." os="..." url="C:\Users\dina\ProM64\packages\tesis\mycube.jar" desc="mycube for Windows 32bit" org="SourceForge" auto="false" hasPlugins="true" license="unknown" author="Dina, LpSolve, peno32" maintainer="Dina, LpSolve, peno32" logo="">` `</package>`.
- Se deben escribir las dependencias del complemento, en caso de que las tenga.

- La url debe apuntar al propio complemento para que luego se pueda instalar desde el administrador de paquetes que se encuentra en la carpeta del ProM.
- Ejecutar el ProM.

Su desarrollo se organizó en cuatro paquetes que colaboran entre sí. El paquete MyCube posee 25 clases y contiene la lógica del negocio que crea y maneja los CP. El paquete ProM posee cuatro clases y maneja, conecta y muestra la aplicación como complemento del ProM. El paquete Connection posee 32 clases y maneja las conexiones del complemento. El paquete Business posee seis clases y contiene la lógica del negocio relacionada con la comparación en el CP. Dentro de él están las clases que contienen los algoritmos de la técnica creada.

El algoritmo para generar las tablas es el Compare(), localizado en la clase CompareController. Recibe como entrada el *sublog* materializado. A partir de atributos que el usuario selecciona para comparar, se obtiene el conjunto de atributos seleccionados, denotado por AS, y los fijos, denotados por AF, que pueden ser ocurrencias o recursos. Se verifica que existen atributos seleccionados $AS_i \in AS$ que cumplan con $AS_i \in \mathbb{R}$, es decir, que sean numéricos. Para cada uno se realiza la operación seleccionada por el usuario, por ejemplo, total o promedio, denotada por OS. Luego, en AF se verifica cuál seleccionó el usuario y se realiza la operación que seleccione el usuario, denotada por OAF: hallar la frecuencia relativa o la absoluta.

La frecuencia absoluta f_i es el número de veces que aparece un determinado valor en un estudio estadístico. La frecuencia relativa es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos [100]:

$$n_i = f_i / N \quad \text{siendo } n_i \text{ la frecuencia relativa, } f_i \text{ la frecuencia absoluta y } N \text{ el número total del dato.}$$

Tabla 1 Algoritmo para generar las tablas de datos.

| Algoritmo |
|---|
| Entradas: Log |
| Salidas: Datos de AS y AF |
| <p>Procedimiento: Compare</p> <pre> si existe(Atributos) entonces AS = Extraer_Atributos_Seleccionados() AF = Extraer_Atributos_Fijos() si $AS \in \mathbb{R}$ entonces para cada $AS_i \in AS$: //para cada atributo seleccionado si OS ← "Total" entonces Calcular_Total_Atributo(AS_i) // calcula la suma de todos los valores del atributo fin si sino si OS ← "Promedio" entonces </pre> |

```

        Calcular_Promedio_Atributo(ASi) // calcula el promedio de todos los valores
        del atributo
    fin sino si
fin para
para cada AFi ∈ AF: //para cada atributo fijo seleccionado
    si AFi ← "Ocurrencias" entonces
        si OAF ← "Frecuencia_Absoluta" entonces
            Contar_Ocurrencias(AFi) // contar todas las ocurrencias del atributo fijo
        fin si
        sino si OAF ← "Frecuencia_Relativa" entonces
            Calcular_Porciento_Ocurrencias(AFi) // calcular el porciento de
            ocurrencias del atributo fijo
        fin sino si
    fin si
    sino si AFi ← "Recursos" entonces
        si OAF ← "Frecuencia_Absoluta" entonces
            Contar_Ocurrencias(AF_Recursosi) // calcular el porciento de
            ocurrencias del atributo fijo
        fin si
        sino si OAF ← "Frecuencia_Relativa" entonces
            Calcular_Porciento_Ocurrencias(AF_Recursosi) // calcular el porciento
            de ocurrencias del atributo fijo
        fin sino si
    fin sino si
fin para
Mostrar_Datos(AS,AF)
fin si
    sino Error
fin

```

Una traza del primer *log* utilizado para comprobar la funcionalidad de la técnica se muestra a continuación:

<trace>

```

<string key="concept:name" value="A100"/>
<event>
    <float key="amount" value="35.0"/>
    <string key="org:resource" value="561"/>
    <string key="dismissal" value="NIL"/>
    <string key="concept:name" value="Create Fine"/>
    <string key="vehicleClass" value="A"/>
    <float key="totalPaymentAmount" value="0.0"/>
    <string key="lifecycle:transition" value="complete"/>
    <date key="time:timestamp" value="2006-08-02T00:00:00+02:00"/>
    <int key="article" value="157"/>
    <int key="points" value="0"/>
</event>

```

```

<event>
  <string key="concept:name" value="Send Fine"/>
  <string key="lifecycle:transition" value="complete"/>
  <float key="expense" value="11.0"/>
  <date key="time:timestamp" value="2006-12-12T00:00:00+01:00"/>
</event>
<event>
  <string key="notificationType" value="P"/>
  <string key="concept:name" value="Insert Fine Notification"/>
  <string key="lifecycle:transition" value="complete"/>
  <string key="lastSent" value="P"/>
  <date key="time:timestamp" value="2007-01-15T00:00:00+01:00"/>
</event>
<event>
  <float key="amount" value="71.5"/>
  <string key="concept:name" value="Add penalty"/>
  <string key="lifecycle:transition" value="complete"/>
  <date key="time:timestamp" value="2007-03-16T00:00:00+01:00"/>
</event>
</trace>

```

Para obtener la interfaz que contiene los modelos de proceso, así como la tabla con datos numéricos de las celdas seleccionadas, se debe acceder a la técnica desde el ProM.

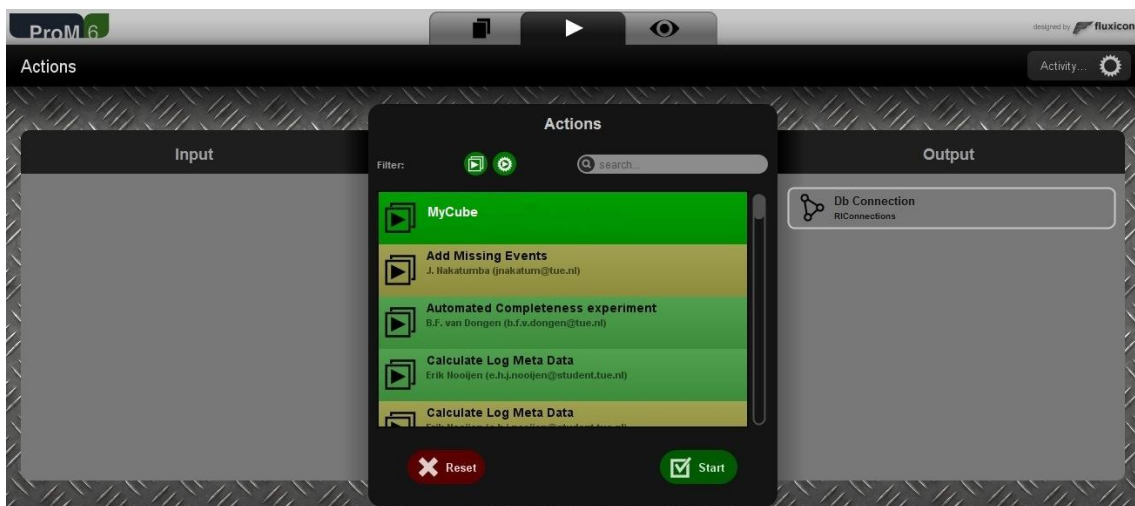


Figura 15 Interfaz de ProM para seleccionar el complemento. Fuente: Elaboración propia.

El usuario puede seleccionar las dimensiones que tendrá el cubo a través del menú Cube para conformar el CP. En la siguiente imagen se muestra la interfaz donde se manejan las dimensiones.

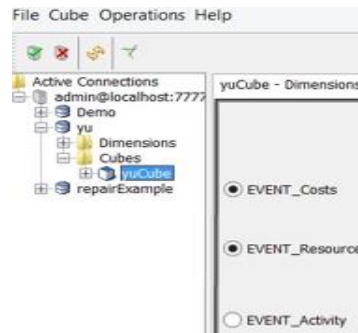


Figura 16 Interfaz para conformar el CP. Fuente: Elaboración propia.

El menú Operations (Operaciones) permite aplicarle las operaciones OLAP al CP para crear el subcubo deseado. Luego se pueden seleccionar los elementos por los que se desea comparar y conformar las celdas de proceso y la tabla de datos. Los resultados se obtienen como muestran las siguientes dos figuras.

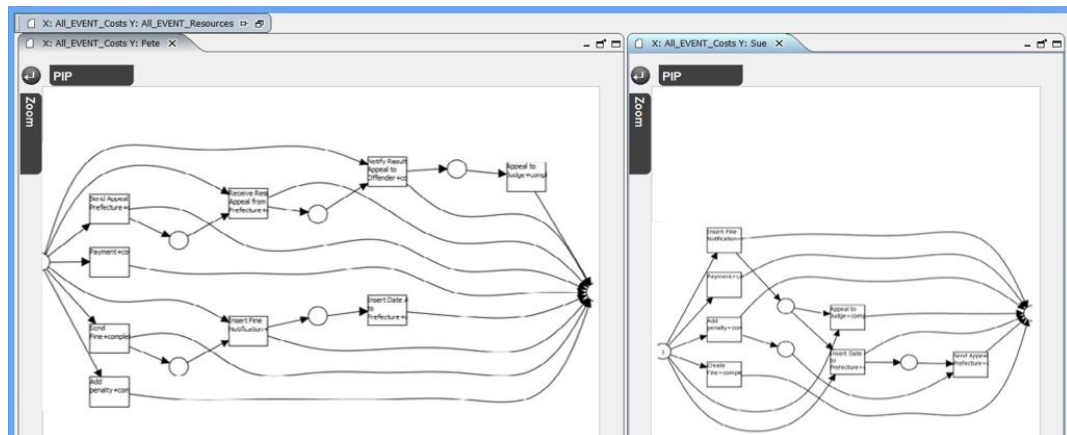


Figura 17 Interfaz con los modelos descubiertos. Fuente: Elaboración propia.

| Activity | Sum | | Average | | Median | |
|---------------------------------------|---------|-------|---------|------|--------|------|
| | Log1 | Log2 | Log1 | Log2 | Log1 | Log2 |
| Send Fine | - | 0.0 | - | 0.0 | - | 0.0 |
| Create Fine | 880.0 | - | 22.0 | - | 22.0 | - |
| Add penalty | 25579.0 | 286.0 | 68.03 | 71.5 | 71.5 | 71.5 |
| Insert Fine Notification | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Send for Credit Collection | - | - | - | - | - | - |
| Payment | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Insert Date Appeal to Prefecture | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Send Appeal to Prefecture | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Appeal to Judge | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Notify Result Appeal to Offender | - | 0.0 | - | 0.0 | - | 0.0 |
| Receive Result Appeal from Prefecture | - | 0.0 | - | 0.0 | - | 0.0 |

Figura 18 Interfaz con la tabla de datos. Fuente: Elaboración propia.

2.5 Conclusiones del capítulo

- Se conceptualizó y caracterizó la arquitectura de la propuesta de solución, conformada por cinco componentes que permiten realizar un análisis multidimensional a un *log*, a partir de celdas de proceso.
- Se desarrolló una técnica de MP basada en el enfoque de CP, definida por cuatro etapas para obtener celdas de proceso, a partir de las cuales se generan modelos de procesos desde la técnica *Alpha Miner* del ProM y tablas de datos numéricos.
- La técnica contribuye a la comparación de celdas de proceso y, por consiguiente, al diagnóstico de procesos. Su aporte fundamental es la información que proveen las tablas y que se encuentra ausente en los modelos, que permite identificar patrones infrecuentes, comportamientos comunes, procesos incompletos y mostrar datos adicionales sobre la ejecución de los procesos.
- La instanciación de la técnica en un complemento para el ProM, favorece su integración con otras técnicas de MP y la generalización de sus resultados.

CAPÍTULO 3: VALIDACIÓN DE LA SOLUCIÓN

Introducción

En el presente capítulo se valida la técnica desarrollada a partir de los métodos científicos definidos en el marco teórico de la investigación. Para ello se aplican el escalamiento de Likert para obtener el criterio de expertos y la técnica de ladov para conocer el nivel de satisfacción de potenciales usuarios. Esto se realiza a través de encuestas elaboradas que pueden consultarse en los anexos. Se realiza un estudio de casos con *logs* y técnicas reales, del cual se analizan sus resultados para validar la hipótesis de la investigación.

3.1 Diseño de la validación

La validación se diseñó mediante los siguientes métodos para validar la hipótesis de la investigación:

- Estudio de casos para analizar *logs*, comprobar si se cumple la hipótesis y validar los resultados obtenidos.
- Criterio de expertos empleando el escalamiento de Likert para validar la técnica y sus componentes.
- Satisfacción de potenciales usuarios mediante la técnica de ladov para evaluar el nivel de satisfacción de potenciales usuarios con la técnica propuesta.

3.2 Validación de la propuesta de solución

3.2.1 Estudio de casos

Se llevó a cabo un estudio de casos múltiples [101] para comprobar el cumplimiento de la hipótesis científica que guía la presente investigación. A partir de la aplicación de la técnica desarrollada a dos *logs* reales, de una oficina de tráfico y de un hospital, se observó el comportamiento de la variable dependiente contribuir al diagnóstico de procesos de negocio.

3.2.1.1 Aplicación de la técnica en un proceso real en una oficina de tráfico

Se usó un *log* real de un departamento de control de tráfico que representa al proceso “gestionar multas”. Está compuesto por 54 casos y 207 eventos, con una media de 4 eventos por caso. La actividad “crear multa” es la que inicia el proceso y “enviar al registro de créditos” la que lo finaliza.

Para percibir la diferencia entre los resultados, inicialmente se le aplicó la técnica *Alpha Miner* del ProM al *log*, para descubrir el modelo que representa al proceso completo. De esta forma, al aplicar la técnica propuesta en esta investigación, se puede apreciar su aporte en la comparación entre celdas y el diagnóstico del proceso.

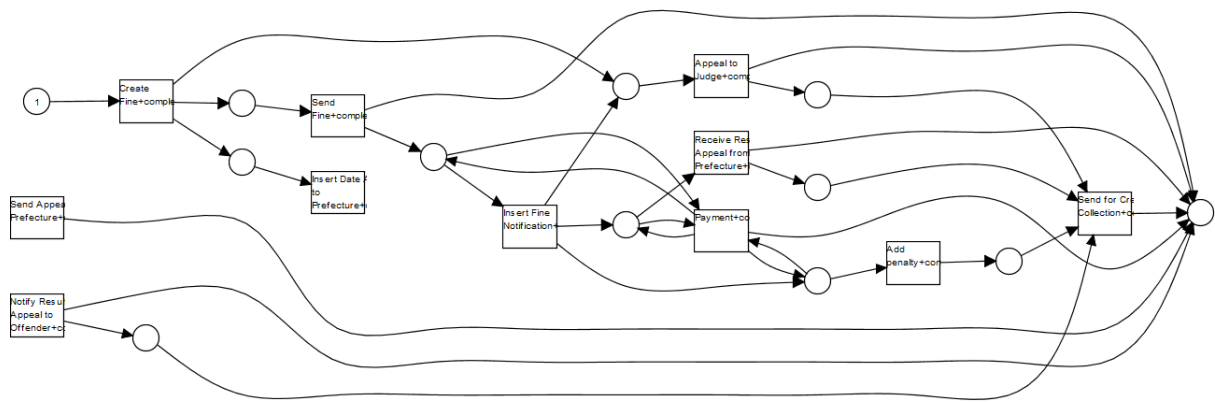


Figura 19 Modelo del proceso “gestionar multas” con *Alpha Miner*. Fuente: Elaboración propia.

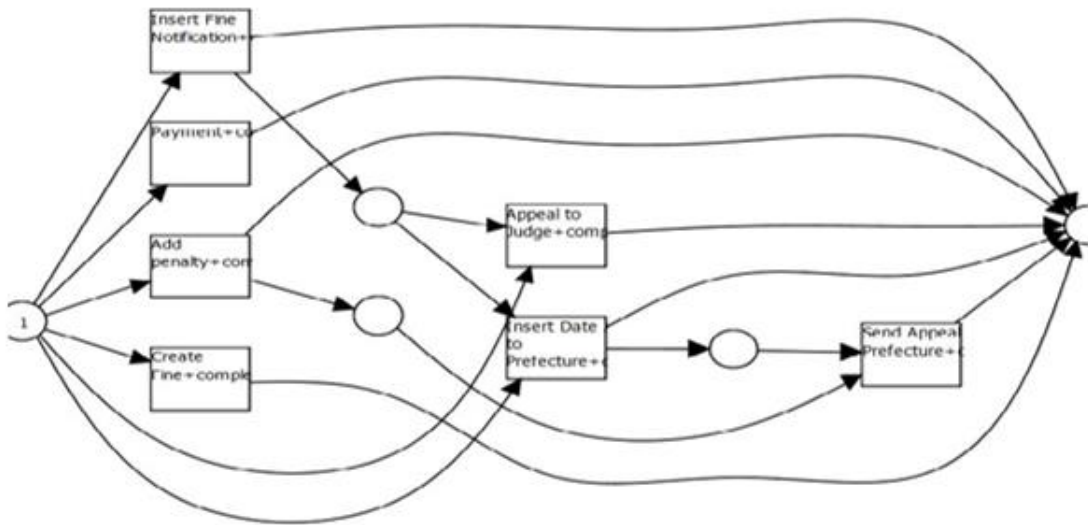
El modelo permite comprender que el proceso está formado por las actividades: crear multa, enviar multa, insertar notificación de multa, apelar al juez, insertar fecha a la prefectura, enviar apelación a la prefectura, notificar resultado de la apelación al solicitante, recibir resultado de apelación de la prefectura, añadir penalización, pagar y enviar al registro de créditos.

La actividad “crear multa” es la que inicia el proceso, pero se observa que este finaliza con casi todas las actividades, lo que indica que no se terminan de ejecutar todas las instancias, es decir, hay procesos incompletos. Se observan también, actividades que se repiten o que no siguen una misma secuencia, relacionadas con la actividad de pago.

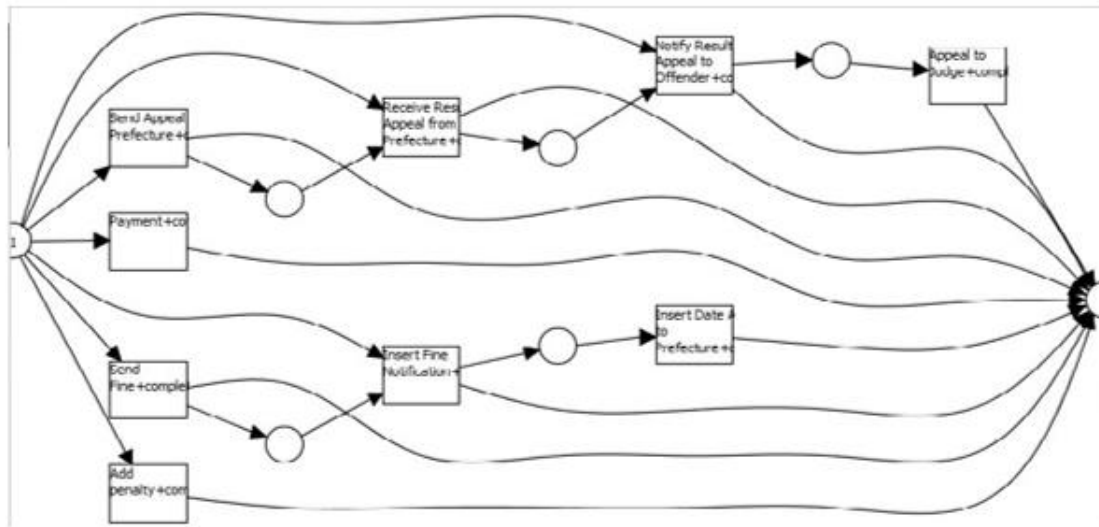
El instrumento para el estudio de caso lo constituyó el componente MyCube. Se seleccionaron las dimensiones “tiempo”, “recurso” y “actividad”, para formar el CP. Se le aplicó la operación *slice* ($slice_{tiempo,H}(PCV) = (2017, \text{meses})$), para considerar las actividades ocurridas durante los meses del año 2017. Se le aplicó la operación *dice* ($dice_{res}(PCV) = (\text{tiempo}, \{\text{Enero, Febrero}\}, \{\text{Julio, Agosto}\})$) para obtener dos subcubos, uno con los valores Enero y Febrero (E-F) del atributo “tiempo” y el otro con los valores Julio y Agosto (J-A). Esta misma operación se realizó para seleccionar los períodos analizados (Mayo-Junio, Noviembre-Diciembre, 2017).

A partir del *log* resultante al aplicar las operaciones anteriores, se obtuvieron los modelos de procesos y la tabla de datos asociados a dichos modelos, que se pueden observar en la siguiente figura.

En el inciso a), se puede apreciar que en el período E-F de 2017 existen procesos incompletos. Teniendo en cuenta que el proceso debe comenzar con la actividad “crear multa”, se observa que hay instancias del proceso que comienzan con actividades más avanzadas en el orden en que se debe ejecutar el proceso, específicamente con las actividades “insertar notificación de multa”, “añadir penalización”, “pago”, “apelar al juez” e “insertar fecha a la prefectura”. Asimismo, todas las instancias reflejadas en el modelo terminan con otras actividades que no constituyen la actividad final.



a)



b)

Figura 20 Modelo correspondiente al período a) E-F y b) J-A. Fuente: Elaboración propia.

El inciso b) refleja el período J-A de 2017. Se observa una mayor variedad de actividades relacionadas con las apelaciones. No se inicia ningún nuevo proceso, sino que se continúan procesos iniciados anteriormente, como los procesos cuyas multas fueron apeladas anteriormente. En este período tampoco se terminan las instancias del proceso.

Una vez consultados los modelos, se procedió a analizar la tabla resultante con los datos asociados a los mismos. El objetivo fue comparar las celdas del proceso seleccionadas y observar si se obtiene información relevante para lograr un mejor diagnóstico.

| Event Data | | |
|---------------------------------------|-------------|------|
| Activity | Occurrences | |
| | Log2 | Log4 |
| Insert Fine Notification | 406.0 | 51.0 |
| Send for Credit Collection | - | - |
| Payment | 197.0 | 45.0 |
| Insert Date Appeal to Prefecture | 40.0 | 1.0 |
| Send Appeal to Prefecture | 10.0 | 4.0 |
| Appeal to Judge | 4.0 | 1.0 |
| Create Fine | 40.0 | - |
| Add penalty | 376.0 | 4.0 |
| Send Fine | - | 13.0 |
| Notify Result Appeal to Offender | - | 9.0 |
| Receive Result Appeal from Prefecture | - | 7.0 |

Figura 21 Datos de los períodos E-F y J-A. Fuente: Elaboración propia.

En la tabla anterior, el período E-F está representado por “Log2” y J-A por “Log4”. La columna *Occurrences* (ocurrencias) informa la cantidad de veces que se ejecutó cada actividad en cada *log* seleccionado. La actividad “enviar al registro de créditos” no tiene ocurrencias, por lo que se puede observar un comportamiento común: en esos períodos no se concluyen instancias del proceso. Las dos últimas actividades de la tabla anterior, relacionadas con las apelaciones, no poseen ocurrencias durante inicios de año, indicando que en ese período no se realizaron actividades en ese sentido. Se aprecia que la actividad “añadir penalización” se realiza con mucha mayor frecuencia durante E-F que en J-A. Debido a que una regla del negocio es que un chofer tiene dos meses hábiles para pagar una multa, se entiende que las penalizaciones son de multas de los meses anteriores a E-F y no del período actual. Por tanto, se observa que una mayor cantidad de choferes no pagaron las multas creadas en fin de año.

Para buscar una explicación a lo que sucede en E-F se analizó el modelo de Noviembre-Diciembre (N-D) del año precedente, donde se debieron generar las multas no pagadas que generaron penalizaciones en E-F.

En el modelo de N-D de la siguiente figura se observan dos casos: se crearon multas que fueron pagadas y multas que no fueron pagadas al terminar el período.

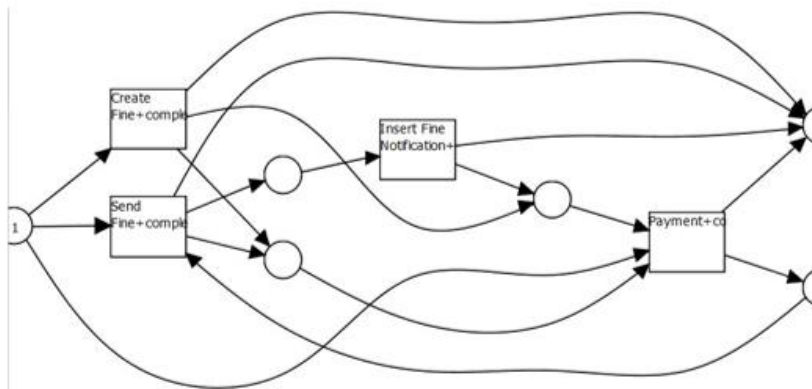


Figura 22 Modelo correspondiente al período N-D. Fuente: Elaboración propia.

Para buscar más información se procedió a analizar la tabla siguiente resultante con los datos asociados al período.

| EVENT_amount | | | | |
|---------------------------------------|---------|---------|---------|-------|
| Activity | Sum | | Average | |
| | Log1 | Log2 | Log1 | Log2 |
| Insert Fine Notification | 0.0 | 0.0 | 0.0 | 0.0 |
| Send for Credit Collection | - | - | - | - |
| Payment | 0.0 | 0.0 | 0.0 | 0.0 |
| Insert Date Appeal to Prefecture | - | 0.0 | - | 0.0 |
| Send Appeal to Prefecture | - | 0.0 | - | 0.0 |
| Appeal to Judge | - | 0.0 | - | 0.0 |
| Create Fine | 38544.0 | 880.0 | 33.69 | 22.0 |
| Add penalty | - | 25579.0 | - | 68.03 |
| Send Fine | 0.0 | - | 0.0 | - |
| Notify Result Appeal to Offender | - | - | - | - |
| Receive Result Appeal from Prefecture | - | - | - | - |

Figura 23 Datos de los períodos N-D y E-F. Fuente: Elaboración propia.

En esta imagen el período N-D está representado por “Log1” y E-F por “Log2”. Analizando los datos asociados a este modelo, la suma de las multas creadas en fin de año es casi 44 veces superior a la de inicios de año. Esta información permitió constatar que en el período N-D se crearon muchas multas que no fueron pagadas en el período debido, por lo que en el período E-F se aplicaron muchas penalizaciones a los choferes que no pagaron en el tiempo establecido. Además, la columna *Average* (promedio) permitió conocer que el monto promedio de las multas en el período N-D fue superior al del período E-F. Esto permite asociar el aumento de cantidad y monto de las multas a una mayor cantidad de infracciones y más peligrosas o serias durante fin de año, por parte de los choferes.

Retomando el análisis del período J-A, no se crearon instancias nuevas (multas), sino que se continuaron o terminaron instancias previamente iniciadas, concentrándose mayormente en las apelaciones y penalizaciones. Para buscar más información, se realizó una comparación de dicho período con respecto al año 2017 completo y con el período Mayo-Junio (M-J).

En la siguiente figura “Log0” representa las ocurrencias de las actividades durante todo el año 2017, “Log4” representa las ocurrencias del período J-A y “Log 3” las de M-J. La actividad “crear multa” durante M-J y J-A no tiene ocurrencias. Dicho dato indica un comportamiento común donde en esos períodos no se crean multas, a pesar de que en todo el año en general, se crearon 3458 multas. Otro comportamiento común que se evidencia en estos períodos es que no se registran ocurrencias de la actividad “enviar al registro de créditos”, lo que significa que no finalizan instancias del proceso en dichos períodos. Según el negocio, pudiera analizarse por qué en estos períodos no se terminan instancias del proceso, para examinar si son desviaciones del proceso.

| Event Data | | | |
|---------------------------------------|-------------|-------|------|
| Activity | Occurrences | | |
| | Log0 | Log3 | Log4 |
| Insert Fine Notification | 1195.0 | 4.0 | 51.0 |
| Send for Credit Collection | 782.0 | - | - |
| Payment | 2109.0 | 116.0 | 45.0 |
| Insert Date Appeal to Prefecture | 79.0 | 15.0 | 1.0 |
| Send Appeal to Prefecture | 143.0 | 20.0 | 4.0 |
| Appeal to Judge | 6.0 | - | 1.0 |
| Create Fine | 3458.0 | - | - |
| Add penalty | 1190.0 | 288.0 | 4.0 |
| Send Fine | 2019.0 | 69.0 | 13.0 |
| Notify Result Appeal to Offender | 15.0 | 1.0 | 9.0 |
| Receive Result Appeal from Prefecture | 15.0 | 5.0 | 7.0 |

Figura 24 Datos de los períodos J-A y Enero-Diciembre. Fuente: Elaboración propia.

En J-A se realizan 45 pagos de 2190 de todo el año, cuando el promedio bimensual es de 365. Esto representa un 2% de los pagos realizados en el año, pudiendo significar un patrón infrecuente.

La actividad “crear multa” durante M-J y J-A no tiene ocurrencias. Dicho dato indica un comportamiento común donde en esos períodos no se crean multas, a pesar de que el promedio bimensual es de 576. Asimismo, indica que se debe revisar el negocio, pues pudiera significar una desviación del proceso.

En el período J-A se continúan y terminan apelaciones iniciadas anteriormente. Se realiza el 60% de las notificaciones del año a los solicitantes sobre los resultados de las apelaciones. Incluso, casi la mitad de los resultados de apelaciones se realizan en ese período. Además, el 47% recibe el resultado de la apelación desde la prefectura. Estos datos significan que este período es el más activo, donde se procesan y terminan la mayoría de las apelaciones del año.

A su vez, en el año ocurren 1190 penalizaciones. En el período J-A solo ocurren 4, lo que representa el 0.34%, mientras que la cantidad de E-F fue superior, teniendo en cuenta que el promedio bimensual es de 198.

A modo de resumen se pudo observar que:

- A partir de los modelos de proceso obtenidos no era posible conocer, con mayor profundidad, el funcionamiento del proceso.
- En el período J-A se añaden muy pocas penalizaciones, por lo que se debe valorar si se corresponde con el negocio.

- A finales de año se crearon muchas multas que no fueron pagadas. Los motivos deben analizarse, pues puede deberse a que los choferes cometieron muchas infracciones al manejar o a que en esa época se realiza mejor el procedimiento establecido.
- A inicios de año se penalizó a muchos choferes por esa infracción.
- A mediados de año la oficina se centró en las actividades de apelaciones.
- El proceso, en los períodos de inicio y mediado de año, tiene marcadas diferencias que reflejan comportamientos diferentes.

3.2.1.2 Aplicación de la técnica en un proceso real en un hospital

Este *log* real contiene eventos de casos de infecciones complicadas de un hospital. Cada caso representa un camino a través del hospital. El *log* representa a un proceso con 1050 casos y 15214 eventos, con una media de 14 eventos por caso. Contiene 16 actividades diferentes, 39 atributos de los datos (responsable, resultados de pruebas, etc.) y recoge los eventos desde noviembre de 2013 hasta junio de 2015. La actividad inicial es “ER Registration” y la final es “Return ER”.

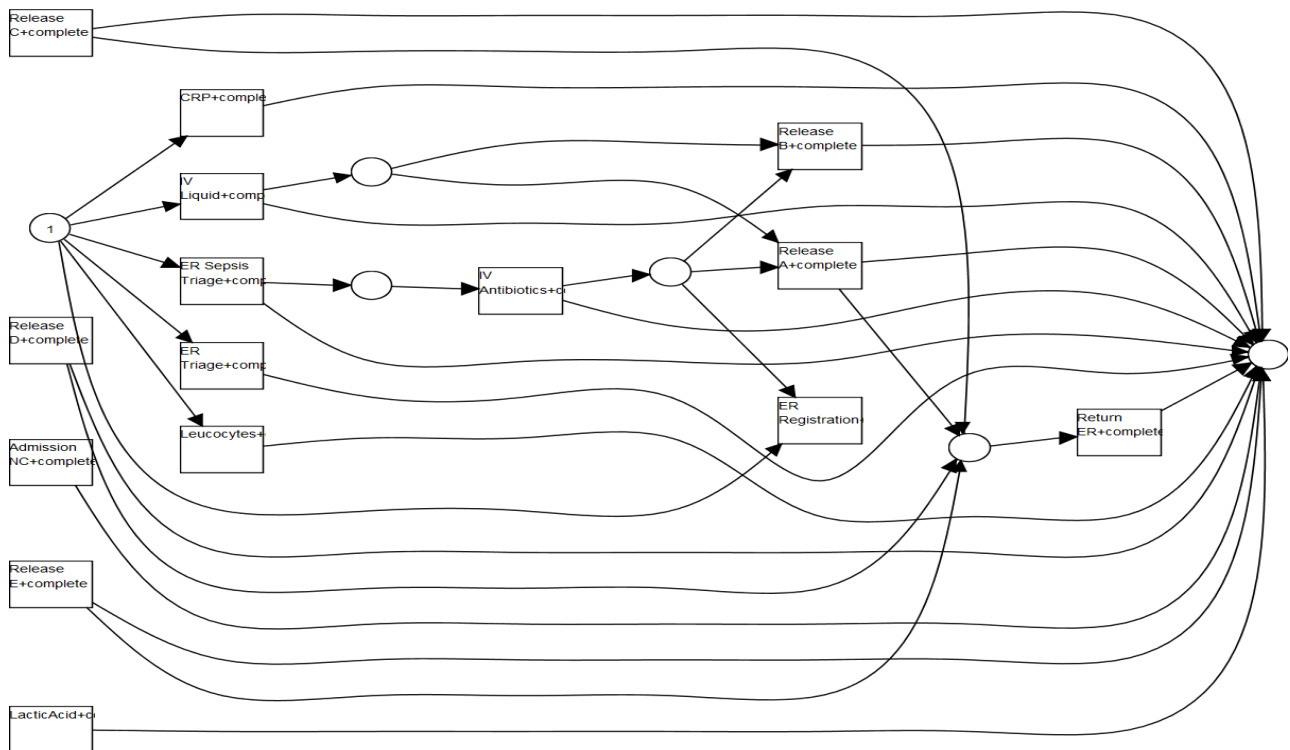


Figura 25 Modelo del proceso completo descubierto con el Alpha Miner. Fuente: Elaboración propia. Inicialmente se descubrió el modelo del proceso con el *Alpha Miner* para observar el proceso completo, que se observa en la figura anterior. No se distingue una actividad inicial o final, sino que se observan varias en ambos casos, lo que indica que hay instancias incompletas. El proceso debe comenzar con la actividad “ER Registration”, sin embargo, se observan varias actividades de inicio. Esto debe ser

analizado para determinar si pudieran ser violaciones o desviaciones, o si constituyen procesos incompletos.

El instrumento para el estudio lo constituyó el componente MyCube. Se seleccionaron las dimensiones “tiempo”, “edad” y “actividad”, para formar el CP. Se le aplicó la operación *slice* ($slice_{\text{tiempo,H}}(\text{PCV}) = (2014, \text{meses})$), para considerar las actividades ocurridas durante los meses del año 2014. Se le aplicó la operación *dice* ($dice_{\text{res}}(\text{PCV}) = (\text{tiempo}, \{\text{Enero, Febrero}\}, \{\text{Julio, Agosto}\}, 2014)$) para obtener tres celdas, una con los valores Enero y Febrero (E-F) del atributo “tiempo”, otra con los valores Julio y Agosto (J-A), y otra con los valores del año 2014 completo. El objetivo es comparar el proceso a inicios y mediados del año. Esta misma operación se realizó para seleccionar los demás atributos analizados.

Los modelos obtenidos de los períodos seleccionados no poseen diferencias significativas con respecto al general mostrado en la Figura 25, por lo que no se consideró necesario presentarlos en este documento. Luego se procedió a analizar la tabla resultante con los datos asociados a los mismos.

| Event Data | | | |
|------------------|-------------|-------|--------|
| Activity | Occurrences | | |
| | E-F | J-A | 2014 |
| ER Sepsis Triage | 132.0 | 128.0 | 784.0 |
| ER Registration | 129.0 | 127.0 | 790.0 |
| ER Triage | 145.0 | 141.0 | 793.0 |
| Leucocytes | 436.0 | 284.0 | 2123.0 |
| CRP | 422.0 | 301.0 | 2089.0 |
| IV Liquid | 119.0 | 92.0 | 569.0 |
| Admission NC | 141.0 | 143.0 | 842.0 |
| IV Antibiotics | 103.0 | 99.0 | 611.0 |
| LacticAcid | 150.0 | 146.0 | 896.0 |
| Release A | 76.0 | 59.0 | 387.0 |
| Release D | 2.0 | 1.0 | 15.0 |
| Release C | 1.0 | 1.0 | 14.0 |
| Release E | 1.0 | - | 4.0 |
| Return ER | 49.0 | 11.0 | 146.0 |
| Admission IC | 4.0 | - | 68.0 |
| Release B | - | 3.0 | 25.0 |

Figura 26 Ocurrencias de las actividades de los períodos seleccionados. Fuente: Elaboración propia.

En la figura anterior se puede observar que, en ambos conjuntos de meses, las actividades que más se realizan son “Leucocytes” y “CRP”. Ambas tienen mayor número de ocurrencias que la media bimensual de 353 y 348 respectivamente, lo que muestra un comportamiento común. A su vez, son las actividades que más poseen variación entre ambos períodos. Este detalle resulta interesante, pues pudiera representar que (1) que esas actividades se realizan menos en J-A debido a que las personas se enferman menos o sus síntomas en esa época no ameritan realizar esas actividades o que (2) hay otras razones para que esas actividades se realicen en menor cantidad. Por ejemplo, menor cantidad de personal laborando (tendría que coincidir que el personal fuera el encargado de estas actividades y si fuera personal en general, debería existir una variación similar en otras actividades), equipos rotos,

o desviaciones en el proceso. De ahí que deba analizarse los motivos por los que entre inicios y mediados de año estas dos actividades varían tanto en cuanto a su ocurrencia.

También se puede observar que a inicios de año no se realizó la actividad “Release B” y a mediados de año no se realizaron las actividades “Release E” ni “Admission IC”. Según el negocio, pudieran ser desviaciones o comportamientos poco frecuentes.

En J-A se termina aproximadamente la mitad de las instancias que se terminan bimensualmente, identificadas por la actividad “Return ER” y cuyo promedio bimensual de ocurrencias es 24. Sin embargo, en E-F se termina el doble de las instancias del promedio, a la vez que la cifra constituye la tercera parte de las instancias terminadas en el año. Este comportamiento debe analizarse según el negocio, pues pudiera indicar desviaciones en el proceso. El resto de las actividades se comportan de manera estable.

Otro detalle interesante resulta que la mayoría de las actividades se realizan cientos o miles de veces, pero solo hay 146 casos cerrados durante todo el año. Esto pudiera significar que (1) es un comportamiento común, es decir, que no todos los casos atendidos deban cerrar con esa actividad, (2) existen desviaciones en el proceso o (3) pudieran estar ejecutándose reiteradas veces estas actividades, ya sea por errores, por caducidad según el tiempo de validez de los exámenes, u otros factores a considerar según el negocio.

Para continuar el análisis, luego se seleccionaron las dimensiones “edad”, “tiempo” e “infección severa”, para los mismos períodos de tiempo analizados previamente.

| Event Data | | | | | |
|------------|-------------|-------|--------|---------------|-------|
| Age | Occurrences | | | Occurrences % | |
| | E-F | J-A | 2014 | E-F | J-A |
| 0-10 | 83.0 | 61.0 | 319.0 | 26.02 | 19.12 |
| 11-20 | 82.0 | 73.0 | 465.0 | 17.63 | 15.70 |
| 21-30 | 55.0 | 54.0 | 308.0 | 17.86 | 17.53 |
| 31-40 | 58.0 | 11.0 | 310.0 | 18.71 | 3.55 |
| 41-50 | 49.0 | 47.0 | 301.0 | 16.28 | 15.61 |
| 51-60 | 57.0 | 50.0 | 282.0 | 20.21 | 17.73 |
| 61-70 | 122.0 | 75.0 | 526.0 | 23.19 | 14.26 |
| 71-80 | 66.0 | 56.0 | 343.0 | 19.24 | 16.33 |
| 81-90 | 2.0 | 1.0 | 25.0 | 8.00 | 4.00 |
| 91-99 | 1.0 | – | 4.0 | 25.00 | – |
| Total | 575.0 | 428.0 | 2883.0 | 19.94 | 14.85 |

Figura 27 Ocurrencias de infecciones severas por edades. Fuente: Elaboración propia.

Según los datos de la figura anterior, los niños de 11-20 años y los ancianos de 61-70 años son los que más sufren de infecciones severas en el año. Un dato interesante resulta que sufren esta situación con mayor cantidad a inicios de año con respecto a mediados de año.

Además, existen diferencias con respecto a inicios de año o al año completo. Se observa un marcado contraste en las personas de 31-40 años que sufrieron esta complicación de salud a inicios y mediados de año. Asimismo, a inicios de año se registraron aproximadamente 150 casos más de “infección severa” que a mediados de año, lo que indica que se debe revisar las causas de este comportamiento. El resto de las actividades básicamente tuvieron comportamientos comunes en los períodos analizados. Para observar el comportamiento de algunos atributos durante el año, se seleccionaron las dimensiones “síntomas”, “tiempo” y “casoID”.

| Event Data | | | | | |
|------------------|-------------|--------|--------|---------------|-------|
| Symptoms | Occurrences | | | Occurrences % | |
| | E-F | J-A | 2014 | E-F | J-A |
| Severe infection | 575.0 | 428.0 | 2883.0 | 19.94 | 14.85 |
| DisfuncOrg | 2.0 | 2.0 | 36.0 | 5.56 | 5.56 |
| Hypotensie | 31.0 | 49.0 | 233.0 | 13.30 | 21.03 |
| Hypoxie | 38.0 | 23.0 | 167.0 | 22.75 | 13.77 |
| Transfusion | 124.0 | 50.0 | 744.0 | 16.67 | 6.72 |
| Oligurie | 6.00 | 125.00 | 682.0 | 0.88 | 18.33 |
| Total | 776.0 | 677.0 | 4745.0 | 79.11 | 80.25 |

Figura 28 Ocurrencias de síntomas en los períodos seleccionados. Fuente: Elaboración propia.

En la figura anterior se observa que hay un aumento de personas con los síntomas “Hypotensie” y “Oligurie” durante mediados de año. Este último aumentó más de 20 veces, mientras que la actividad “Transfusion” disminuyó casi 3 veces, por lo que resultan comportamientos interesantes pues pudieran ser poco comunes o desviaciones que deben ser analizados.

A modo de resumen se pudo observar que:

- Las actividades que más se realizan y más diferencias en su ocurrencia tienen entre inicios y mediados de año son “Leucocytes” y “CRP”.
- Las actividades “Release B”, “Release E” y “Admission IC” no tuvieron ocurrencias en algunos períodos del año.
- En E-F se terminan el doble de los casos del promedio y en J-A la mitad.
- Solo hay 146 casos cerrados durante todo el año a pesar de que la mayoría de las actividades se ejecutan hasta miles de veces.
- Las personas de 11-20 años y 61-70 años son las que más sufren infecciones severas en el año y las de 31-40 años sufrieron la mayor diferencia de ocurrencias a inicios y mediados de año.
- Hay un aumento de personas con los síntomas “Hypotensie” y “Oligurie” y una disminución de “Transfusion” durante mediados de año.

3.2.1.3 Análisis de los resultados del estudio de casos múltiples

Ambos casos demuestran que la técnica contribuye a la comparación de celdas de proceso y, por tanto, al diagnóstico de procesos, siendo capaz de proveer datos para detectar comportamientos comunes, patrones infrecuentes, desviaciones y procesos incompletos.

El análisis realizado fue posible por 2 elementos fundamentales:

1. El CP permite dividir, segmentar o filtrar los *logs* según desee el usuario. Esto resulta una ventaja con respecto a las demás técnicas de MP, pues no lo permiten.
2. La técnica desarrollada en la presente investigación brinda datos numéricos, presentes o disponibles en los *logs* y ausentes en las salidas de otras técnicas relacionadas con los CP, que permiten comparar celdas diferentes con el objetivo de realizar un mejor análisis del funcionamiento del proceso, y de esta forma, tributar a su diagnóstico.

Es importante destacar que de no usar técnicas de CP, el procedimiento a seguir para intentar obtener similares resultados constaría de varios momentos: (1) dividir el *log* para obtener las trazas comprendidas en el período deseado con la técnica *LTL Checker*, (2) aplicar la técnica *Alpha Miner* para descubrir modelos de procesos y (3) observar las ocurrencias de cada actividad con la técnica *Log Visualizer*.

Para dividir el *log* según las trazas comprendidas en el período E-F, se deben ejecutar los siguientes pasos:

1. Importar el *log* y seleccionar el complemento *LTL Checker*.
2. Seleccionar las fórmulas “eventually_before_time_T” y “eventually_after_time_T”.
3. Introducir las fechas 2017/01/01 y 2017/02/28 como parámetros.
4. Chequear la fórmula.
5. Seleccionar Exportar-> “Instancias correctas” para exportar el *log* en el que estos criterios se satisfacen.
6. Seleccionar Exportar-> “Instancias incorrectas” para exportar el *log* en el que no se satisfacen los criterios.

Con los pasos anteriores se obtiene el *log* dividido por los criterios de fecha seleccionados. Luego es posible descubrir el modelo de proceso con *Alpha Miner* y obtener las ocurrencias de las actividades con los siguientes pasos:

7. Importar el *log* del paso 5.
8. Descubrir el modelo de proceso usando *Alpha Miner*.

9. Visualizar el *log* usando la técnica *Log Visualizer*, que ofrece las ocurrencias de cada actividad.
10. Importar el *log* del paso 6.
11. Descubrir el modelo de proceso usando *Alpha Miner*.
12. Visualizar el *log* usando la técnica *Log Visualizer*.

En caso de seleccionar una técnica que muestre las ocurrencias, como *Inductive Visual Miner*, se obvian los pasos 9 y 12.

Las desventajas de esta variante consisten en que (1) para analizar varios períodos, como se hizo en el estudio de casos realizado en los epígrafes anteriores, se deben reproducir todos los pasos recién expuestos para cada período y (2) solo se puede filtrar el log según fórmulas ya predefinidas en la técnica LTL Checker. Esto último ocasiona que no siempre se puedan obtener los resultados deseados, limitando la capacidad de diagnosticar procesos de negocio.

Con la técnica propuesta, sin embargo, los pasos para su ejecución fueron:

1. Seleccionar los atributos del *log* que conformarán las dimensiones del CP.
2. Aplicar operaciones OLAP a los períodos que se quieran analizar.
3. Descubrir los modelos de los subcubos donde se muestran las actividades de los períodos seleccionados.
4. Visualizar los datos de cada modelo mediante la opción *Compare*.

En una prueba realizada mediante el método de la observación, la diferencia en cuanto a cantidad de pasos y tiempo empleados por la misma persona entre ambas variantes, se muestra a continuación:

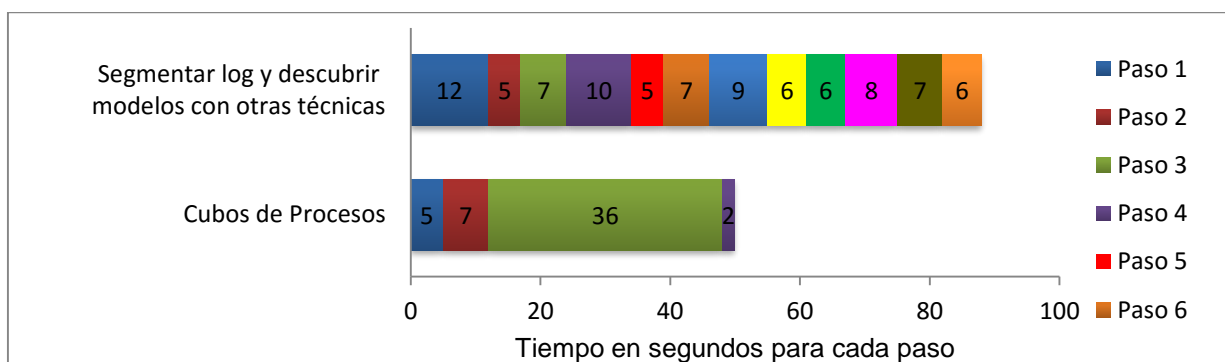


Figura 29 Tiempo de ejecución de ambas propuestas. Fuente: Elaboración propia.

En cuanto a las técnicas de CP identificadas, estas no visualizan resultados a nivel de datos que permitan revelar detalles sobre la ejecución de los procesos de negocio. La inspección visual de los modelos es insuficiente para descubrir patrones. Esta situación se agrava cuando aumenta el número

de trazas y eventos diferentes, pues la complejidad de los modelos obtenidos en estos casos resulta mayor. El análisis solo se basa en el conocimiento del analista.

La técnica propuesta posee varias ventajas:

- Permite filtrar el *log* según el criterio deseado, al igual que las técnicas de CP identificadas, lo que hace a las técnicas de CP superiores en este sentido a las técnicas de MP en general.
- Los resultados se obtienen con menos pasos, tiempo y cantidad de técnicas involucradas que con las técnicas de MP en general.
- No es necesario reproducir todos los pasos desde el inicio para analizar varios períodos, pues teniendo la PCS, solo se deben aplicar operaciones OLAP para obtener diferentes vistas.
- Se obtienen datos de la ejecución del proceso que no brindan las técnicas identificadas de CP y que apoyan su diagnóstico.

Por tanto, la técnica MyCube resulta novedosa y pertinente, ya que aumenta la posibilidad de comparar celdas del proceso y, por ende, de diagnosticar el proceso desde múltiples variantes y a partir de datos asociados a su ejecución.

Operacionalizar las variables permite constatar el cumplimiento de la hipótesis científica de la investigación. Estas variables se tienen en cuenta en el diseño de cada método de validación de la investigación.

Variable independiente: técnica de minería de procesos para comparar celdas de proceso de negocio.

Variable dependiente: contribuir al diagnóstico de procesos de negocio.

Tabla 2 Indicadores para medir la variable dependiente.

| Variable dependiente | Valor | Indicadores | Valor |
|--|---|--|-------|
| Contribuir al diagnóstico de procesos de negocio | Alto: si (DE) && (II) && (DP) && (CC) = Sí | Identificar instancias de procesos incompletas (II) | Sí/No |
| | Medio: si (DE) = No && ((II) && (DP) && (CC) = Sí) | Descubrir patrones infrecuentes (DP) | Sí/No |
| | | Identificar comportamientos comunes (CC) | Sí/No |
| | Bajo: otra combinación | Analizar datos y estadísticas generados en el flujo del proceso (DE) | Sí/No |

Específicamente, la técnica MyCube responde positivamente a la operacionalización de la variable dependiente “contribuir al diagnóstico de procesos de negocio”. La misma alcanza un nivel “Alto”, debido a que sus criterios de medida resultan evaluados positivamente.

3.2.2 Criterio de expertos aplicando escalamiento de Likert

Para conocer la opinión de los expertos sobre la contribución de la técnica desarrollada para comparar celdas de proceso y contribuir al diagnóstico de procesos, fue aplicado el escalamiento de Likert [102], [103], [104] y [105].

Proceso de selección de expertos

Se realizó una valoración inicial de 27 posibles expertos con años de experiencia en la investigación y el trabajo con la MP y el análisis de procesos de negocio, pertenecientes a la Universidad de las Ciencias Informáticas y a la empresa Desoft. De ellos, 23 estuvieron dispuestos a participar y colaborar con la investigación.



Figura 30 Coeficiente de competencia de los candidatos a expertos encuestados. Fuente: Elaboración propia.

Las preguntas diseñadas y el procedimiento empleado para determinar el nivel de competencia de cada candidato a experto, mediante el cálculo de su coeficiente de competencia, pueden ser consultados en el Anexo 1. Los resultados arrojados se muestran en la figura anterior. De los 23 posibles expertos, 14 poseen un nivel de competencia alto, 7 medio y 2 bajo, por lo que 21 poseen un nivel de competencia adecuado para emitir criterios. A su vez, el 14% posee el grado científico de Doctor en Ciencias y el 24% de Máster en Ciencias. El promedio de años de experiencia es de 10 años.

Aplicación del escalamiento de Likert

Su escala psicométrica fue aplicada a través de un cuestionario para conocer el nivel de acuerdo o desacuerdo con las características, funcionalidades y pertinencia de la técnica desarrollada en la presente investigación. Las preguntas están enfocadas en obtener las valoraciones de los expertos en función de los indicadores definidos.

1. ¿Cree que la funcionalidad de la técnica contribuye a su aplicación para diagnosticar procesos de negocio?
2. ¿Considera que resulta útil la técnica desarrollada para identificar comportamientos comunes, instancias de procesos incompletas, desviaciones y patrones infrecuentes?
3. ¿Considera pertinente aplicar la técnica desarrollada para el diagnóstico de procesos de negocio?

4. ¿Cómo valora la complejidad funcional de la técnica desarrollada?
5. ¿Considera usted que la posibilidad de consultar las tablas de datos que aporta la técnica como novedad, favorece el diagnóstico de procesos de negocio?
6. ¿Cómo valora la aplicabilidad de la técnica desarrollada para la comparación de celdas de proceso?
7. ¿Considera que la actualidad de la técnica propuesta contribuye al avance en la investigación y aplicación de los cubos de proceso para el diagnóstico de procesos de negocio?

Cada experto expresó su valoración de cada indicador mediante la siguiente escala. Se representan por 5- Muy de acuerdo (MA), 4- De acuerdo (DA), 3- Ni de acuerdo ni en desacuerdo (Sí-No), 2- En desacuerdo (ED) y 1- Completamente en desacuerdo (CD). Sus valoraciones se pueden consultar en el Anexo 2.

Para el procesamiento de los resultados se calcula la frecuencia de cada categoría de la escala de Likert definida en la encuesta realizada y se calculan los porcentajes de concordancia de cada categoría de acuerdo a las formulaciones propuestas por el autor. Luego se calcula en un índice porcentual (IP), que integra en un solo valor la aceptación del grupo de evaluadores sobre las características del modelo. Esto se realiza mediante la siguiente fórmula:

$$IP = \frac{5 (\% \text{ de MA}) + 4 (\% \text{ de DA}) + 3 (\% \text{ de SI-NO}) + 2 (\% \text{ de ED}) + 1 (\% \text{ de CD})}{5}$$

La siguiente figura muestra que el índice porcentual relacionado con la valoración de los expertos, sobre los aspectos planteados, es superior a 89 en todos los casos. Este resultado evidencia la alta valoración que los expertos le otorgan a la técnica desarrollada.

La confiabilidad del instrumento fue validada por el coeficiente Alpha de Cronbach. Mientras más cerca de 1 esté α , más alto es el grado de confiabilidad (se puede definir como la estabilidad o consistencia de los resultados obtenidos) [106]. El Anexo 3 muestra los detalles y las bases para su aplicación fueron consultadas en [99]. El resultado obtenido fue de 0.75, lo cual demuestra que los datos procesados son confiables.

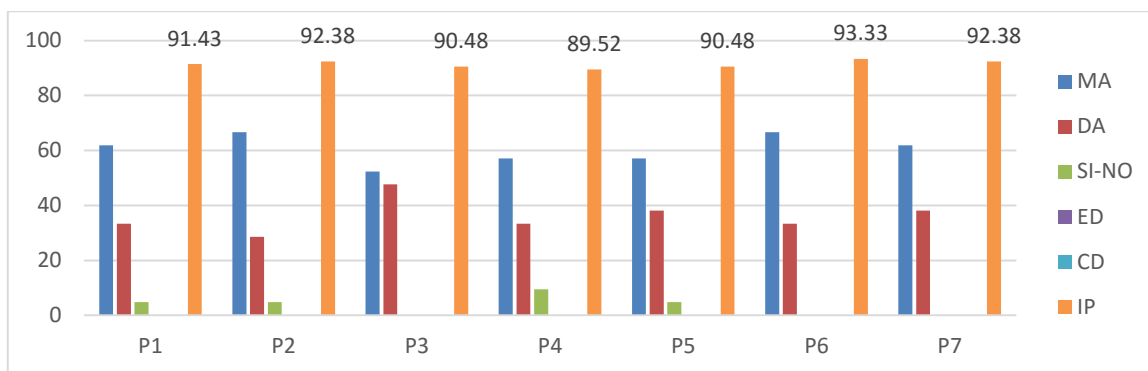


Figura 31 Valoración de los expertos sobre la técnica. Fuente: Elaboración propia.

Los resultados obtenidos a través del escalamiento de Likert evidencian que los elementos teóricos, las características y las funcionalidades de la técnica, poseen una alta valoración por parte de los expertos. Durante el proceso se obtuvieron criterios favorables para el uso y aplicación de la técnica desarrollada para contribuir al diagnóstico de procesos de negocio.

3.2.3 Evaluación de la satisfacción de la técnica desarrollada por potenciales usuarios

La técnica de ladov permite conocer el grado de satisfacción de los implicados en el proceso objeto de análisis. Ha sido aplicada para valorar la satisfacción en múltiples campos y como parte de diagnósticos y validaciones en diferentes investigaciones [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119] y [104]. En esta investigación se aplicó para medir la satisfacción de potenciales usuarios con relación al modelo.

Para su aplicación se realiza una encuesta que permite conocer el grado de satisfacción de los usuarios con respecto a la técnica desarrollada para contribuir al diagnóstico de procesos de negocio a partir de la comparación de celdas de proceso. Para ello se elabora un cuestionario conformado por tres preguntas cerradas y dos abiertas. Los criterios que se utilizan se fundamentan en las relaciones que se establecen entre dichas preguntas y cuya relación el sujeto desconoce. Las preguntas cerradas se muestran en el "Cuadro lógico de ladov", que establece la relación entre las mismas. La primera pregunta evalúa la necesidad de una innovación en el área investigada, la segunda evalúa la pertinencia del resultado de dicha innovación y la tercera evalúa el grado de satisfacción de los usuarios con dicho resultado. Las preguntas de carácter abierto ayudan a conocer las causas que influyen en los diferentes niveles de satisfacción alcanzados. Las respuestas expresan recomendaciones que deben considerarse para mejorar la propuesta presentada.

Tabla 3 Cuadro lógico de ladov aplicado a la investigación.

| | |
|--|---|
| ¿Le satisface la técnica desarrollada en esta investigación como herramienta para comparar y | ¿Considera que solo con los modelos de procesos y sin datos asociados a la ejecución de los mismos, se puede contribuir a su diagnóstico? |
|--|---|

| | | | | | | | | | |
|---|--|-------|----|-------|-------|----|----|-------|----|
| diagnosticar procesos a partir de cubos de proceso? | No | | | No sé | | | Sí | | |
| | ¿Si usted necesitara comparar celdas para diagnosticar procesos, emplearía la técnica desarrollada en la presente investigación? | | | | | | | | |
| | Sí | No sé | No | Sí | No sé | No | Sí | No sé | No |
| Me gusta mucho | 1 | 2 | 6 | 2 | 2 | 6 | 6 | 6 | 6 |
| No me gusta mucho | 2 | 2 | 3 | 2 | 3 | 3 | 6 | 3 | 6 |
| Me da lo mismo | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Me disgusta más de lo que me gusta | 6 | 3 | 6 | 3 | 4 | 4 | 3 | 4 | 4 |
| No me gusta nada | 6 | 6 | 6 | 6 | 4 | 4 | 6 | 4 | 5 |
| No sé qué decir | 2 | 3 | 6 | 3 | 3 | 3 | 6 | 3 | 4 |

El número resultante de la interrelación de las tres preguntas indica la posición de cada cual en la escala de satisfacción siguiente: 1- Clara satisfacción, 2- Más satisfecho que insatisfecho, 3- No definida, 4- Más insatisfecho que satisfecho, 5- Clara insatisfacción y 6- Contradictoria.

Dicha posición es necesaria para establecer la escala numérica para calcular el Índice de Satisfacción Grupal (ISG). La escala es la siguiente:

- Clara satisfacción: +1
- Más satisfecho que insatisfecho: 0,5
- No definido y contradictorio: 0
- Más insatisfecho que satisfecho: - 0,5
- Máxima insatisfacción: -1

Con estos elementos es posible calcular el ISG mediante la fórmula:

$$ISG = \frac{A(+1) + B(+0.5) + C(0) + D(-0.5) + E(-1)}{N}$$

El ISG permite conocer en qué categoría se encuentra el resultado de la encuesta:

- Insatisfacción: Entre -1 y -0.5
- Contradicción: Entre -0.49 y 0.49
- Satisfacción: Entre 0.5 y 1

Aplicación de la técnica de ladov en la investigación

Para medir la satisfacción se tomó una muestra de 23 potenciales usuarios de la rama de la informática y la ingeniería industrial, teniendo en cuenta los años de experiencia en la investigación y el trabajo con la MP y con el análisis de procesos de negocio.

Los resultados obtenidos a partir del análisis de los resultados de la encuesta fueron:

Tabla 4 Resultados de la encuesta de satisfacción.

| Nivel de Satisfacción | Cantidad | Por Ciento |
|---------------------------------|----------|------------|
| Clara satisfacción | 16 | 69.57% |
| Más satisfecho que insatisfecho | 6 | 26.09% |
| No definido o contradictorio | 1 | 4.35% |
| Más insatisfecho que satisfecho | 0 | 0.00% |
| Máxima insatisfacción | 0 | 0.00% |

Aplicando la fórmula para calcular el ISG se obtuvo:

$$\text{ISG} = (16 * (+1) + 6 * (+0.5) + 1 * (0) + 0 * (-0.5) + 0 * (-1)) / 23 = \mathbf{0.826}$$

El valor de ISG obtenido indica una clara **satisfacción** de potenciales usuarios con respecto a la técnica desarrollada en esta investigación.

Asimismo, las preguntas abiertas permiten profundizar en las causas que originan los diferentes niveles de satisfacción y se obtienen sugerencias de utilidad para la presente y futuras investigaciones. Las preguntas abiertas realizadas fueron:

- ¿Qué elementos considera positivos de la técnica desarrollada?
- ¿Qué elementos considera negativos de la técnica desarrollada?
- ¿Qué sugerencias propone para el desarrollo de la técnica desarrollada?

La valoración de los encuestados sobre las preguntas abiertas refleja que la técnica desarrollada es útil para la comparación de celdas de proceso, su análisis y diagnóstico. Se considera que las tablas de datos, además de los modelos, resultan clave para la detección de desviaciones o anomalías en los procesos. Además, se obtuvieron las siguientes recomendaciones que constituyen oportunidades de mejora para la técnica propuesta:

- Incorporar la posibilidad de descubrir modelos de procesos con otras técnicas además de *Alpha Miner*.
- Destacar diferencias en los modelos descubiertos a partir de las celdas de proceso.

3.3 Conclusiones del capítulo

- La implementación e instanciación de la técnica desarrollada en un complemento para el ProM, facilitó su aplicación en estudio de casos múltiples.
- El estudio de casos múltiples permitió validar la capacidad de la técnica para contribuir al diagnóstico de procesos de negocio a partir de la comparación de celdas, así como su pertinencia y ventajas sobre otras alternativas. La tabla de datos que genera permitió obtener información relevante que no se aprecia mediante el análisis de los modelos de proceso.
- La valoración de potenciales usuarios y criterios de expertos evidenció que existe una alta satisfacción con la técnica desarrollada.

CONCLUSIONES GENERALES

La investigación realizada permite llegar a las siguientes conclusiones:

- La minería de procesos constituye una vía para apoyar el diagnóstico de los procesos. Propicia un análisis basado en evidencias reales su ejecución, a diferencia de la modelación de procesos donde se describe cómo el proceso debe funcionar.
- Los cubos de proceso permiten analizar y explorar los procesos desde una perspectiva multidimensional de los eventos. Las investigaciones realizadas sobre este enfoque describen como una problemática vigente la dificultad para comparar celdas de proceso.
- La técnica desarrollada, basada en cubos de proceso e instanciada en un complemento del ProM, permite comparar celdas de proceso a partir de la visualización simultánea de modelos de celdas de proceso y de tablas con datos sobre la ejecución del proceso. Su aporte fundamental es la información adicional que brindan las tablas (ausente en los modelos), que permite identificar patrones infrecuentes, comportamientos comunes, procesos incompletos y desviaciones, contribuyendo al diagnóstico de procesos de negocio.
- El proceso de validación de la solución permitió constatar que:
 - Los resultados obtenidos al aplicar la técnica sobre un registro de eventos real, permiten afirmar que la solución propuesta contribuye al diagnóstico de procesos de negocio y posee ventajas con respecto a otras alternativas existentes.
 - La valoración de potenciales usuarios y criterios de expertos evidenció que existe una alta satisfacción con la técnica desarrollada.

RECOMENDACIONES

- Generalizar el uso de algoritmos base para aumentar la cantidad de análisis a realizar.
- Desarrollar un algoritmo que permita destacar diferencias y semejanzas en los modelos descubiertos a partir de las celdas de proceso.

BIBLIOGRAFÍA

1. Aalst, W., et al. *Process mining manifesto*. in *Business process management workshops*. 2012: Springer.
2. Van Der Aalst, W.M., *Business process management: a comprehensive survey*. ISRN Software Engineering, 2013. **2013**: p. 1-37.
3. Yzquierdo-Herrera, R., et al., *Diagnóstico de proceso basado en el descubrimiento de subprocesos*. Ingeniería Industrial, 2012. **33**(2): p. 133-141.
4. Mamaliga, T., *Realizing a process cube allowing for the comparison of event data*. 2013, Eindhoven University of Technology: Eindhoven.
5. Mekhala, *A review paper on Process Mining*. International Journal of Engineering and Techniques (IJET), 2015. **1**(4): p. 11-17.
6. Milani, F.P. and F.M. Maggi, *Process Mining in Industry*. 2017, Institute of Computer Science.
7. Aalst, W., *Process mining: Discovery, conformance and enhancement of business processes*. 2011, Springer-Verlag, Berlin.
8. Bose, R.J.C. and W.M. van der Aalst, *Process diagnostics using trace alignment: opportunities, issues, and challenges*. Information Systems, 2012. **37**(2): p. 117-141.
9. Rebuge, Á. and D.R. Ferreira, *Business process analysis in healthcare environments: A methodology based on process mining*. Information systems, 2012. **37**(2): p. 99-116.
10. van der Aalst, W.M. *Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining*. in *Asia-Pacific Conference on Business Process Management*. 2013: Springer.
11. Pérez-Castillo, R., et al., *Modernización Software mediante Descubrimiento de Vistas de Procesos de Negocio*. XVI Jornadas de Ingeniería del Software y Bases de Datos (JISBD'11)-Track Desarrollo Software Dirigido por Modelos (DSDM), 2011.
12. Kalenkova, A.A., et al., *Process mining using BPMN: relating event logs and process models*. Software & Systems Modeling, 2017. **16**(4): p. 1019-1048.
13. de Leoni, M., W.M. van der Aalst, and M. Dees, *A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs*. Information systems, 2016. **56**: p. 235-257.
14. Velazquez-Solis, P.E., et al., *Evidencia Empírica de la Minería de Procesos en la Implantación de CMMI-DEV*. ReCIBE, 2016. **5**(3).
15. Lazarte, I., C.A. Acosta Parra, and G.D. Vilallonga. *Mejora y gestión de procesos de negocio inter-organizacionales aplicando técnicas de minería de procesos*. in *XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina)*. 2016.
16. Arias, M. and E. Rojas, *Guía para gestionar procesos de negocio a través de minería de procesos*. InterSedes: Revista de las Sedes Regionales, 2016. **17**(36).
17. Gamba Ramírez, A.F. and E. Guaidía Flórez, *Aporte de la Minería de Procesos a la Gestión Moderna de Proyectos*. 2016.
18. Jans, M., M. Alles, and M. Vasarhelyi, *The case for process mining in auditing: Sources of value added and areas of application*. International Journal of Accounting Information Systems, 2013. **14**(1): p. 1-20.
19. Sedrakyan, G., J. De Weerd, and M. Snoeck, *Process-mining enabled feedback: "tell me what I did wrong" vs. "tell me how to do it right"*. Computers in human behavior, 2016. **57**: p. 352-376.
20. van der Aalst, W.M., *Process mining: data science in action*. 2016: Springer.
21. GIRALDO, J.C., J. JIMÉNEZ, and M.S. TABARES, *Modelo para optimizar el proceso de gestión de negocio combinando minería de procesos con inteligencia de negocios desde almacenes de datos*. Revista ESPACIOS, 2017. **38**(02).
22. Vasallo, A.D., et al., *ALGORITMOS PARA EL DESCUBRIMIENTO DE PROCESOS. ESTADO DEL ARTE*. CIENCIA DIGITAL, 2017. **1**(2): p. 6-13.
23. Lee, W.L.J., et al., *Replay using Recomposition: Alignment-Based Conformance Checking in the Large*. Applied Sciences, 2017. **7**(10): p. 1023.
24. van Eck, M.L., et al. *PM²: A Process Mining Project Methodology*. in *International Conference on Advanced Information Systems Engineering*. 2015: Springer.

25. Bose, R.J.C. and W. van der Aalst. *Trace alignment in process mining: opportunities for process diagnostics*. in *International Conference on Business Process Management*. 2010: Springer.
26. Weijters, A. and J. Ribeiro. *Flexible heuristics miner (FHM)*. in *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium*. 2011: IEEE.
27. ZarehFarkhady, R. and S.H. Aali, *A Two Phase Approach for Process Mining in Incomplete and noisy Logs*. *IJCSI International Journal of Computer Science Issues*, 2012. **9**.
28. Epure, E.V., et al. *Devising News Recommendation Strategies with Process Mining Support*. in *Atelier interdisciplinaire sur les systèmes de recommandation/Interdisciplinary Workshop on Recommender Systems*. 2017.
29. García, A.O., D.P. Alfonso, and V.E. Sentí, *Análisis de procesos hospitalarios desde la perspectiva de tiempo utilizando minería de procesos*. *Revista Cubana de Ciencias Informáticas*, 2018. **12**(1): p. 130-146.
30. Hompes, B., et al. *Discovering deviating cases and process variants using trace clustering*. in *Proceedings of the 27th Benelux Conference on Artificial Intelligence (BNAIC), November*. 2015.
31. VAN DER AALST, W., *Process mining in the large*. *Process Mining*. Springer Berlin Heidelberg., 2016: p. 353-385.
32. Van der Aalst, W.M., *Extracting event data from databases to unleash process mining*, in *BPM-Driving innovation in a digital world*. 2015, Springer. p. 105-128.
33. Adriansyah, A., B.F. van Dongen, and W.M. van der Aalst. *Towards robust conformance checking*. in *International Conference on Business Process Management BPM 2010 Workshops, Proceedings of the 6th Workshop on Business Process Intelligence (BPI2010). Lecture Notes in Business Information Processing*. 2010. Berlin: Springer.
34. Contreras Muñoz, N.E., *Control y seguimiento de atención de incidencias utilizando minería de procesos*. 2016.
35. Arturo, O.G., L.A. Osvaldo Ulises, and P.A. Damián. *Generador de Registros de Eventos para el análisis de procesos en el Sistema de Información Hospitalaria xavia HIS*. in *Convención Salud 2015*. 2015.
36. van der Heijden, T., *Process mining project methodology: Developing a general approach to apply process mining in practice*, in *Operations Management and Logistics*. 2012, TUE. School of Industrial Engineering: Netherlands.
37. Mayorga, S.A. and C.P. Rodriguez, *Metodología para la aplicación de minería de procesos*.
38. Aguirre Mayorga, H.S. and N. Rincón García, *Minería de procesos: desarrollo, aplicaciones y factores críticos*. *Cuadernos de Administración*, 2015. **28**(50): p. 137-157.
39. Bravo, F.P., D.P. Alfonso, and W.P. García, *Aplicación de la minería de procesos en el Hotel Club Amigo Atlántico Guardalavaca/Application of the process mining in Hotel Club Amigo Atlántico Guardalavaca*. *Retos Turísticos*, 2017. **15**(3).
40. Van der Aalst, W., A. Adriansyah, and B. van Dongen, *Replaying history on process models for conformance checking and performance analysis*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2012. **2**(2): p. 182-192.
41. Rozinat, A.A., *Process mining: conformance and extension*. 2010, Technische Universiteit Eindhoven: The Netherlands.
42. Jans, M., et al., *A business process mining application for internal transaction fraud mitigation*. *Expert Systems with Applications*, 2011. **38**(10): p. 13351-13359.
43. Conca Reckmann, T.I., *Minería de procesos como herramienta para el análisis de la colaboración multidisciplinaria para el tratamiento de pacientes con DMT2 en atención primaria*. 2017.
44. Çela, O. *Cross-Company Collaboration Analyzed Through Process Mining*. in *Actes du 8 e Forum Jeunes Chercheurs du congrès INFORSID*. 2016.
45. van Aalst, W.M., et al., *Auditing 2.0: Using process mining to support tomorrow's auditor*. *Computer*, 2010. **43**(3).
46. Mohamed, A.A.F.G., *Process Mining application considering the organizational perspective using Social Network Analysis*. 2016.

47. Aguirre, S., C. Parra, and J. Alvarado. *Combination of process mining and simulation techniques for business process redesign: a methodological approach*. in *International Symposium on Data-Driven Process Discovery and Analysis*. 2012: Springer.
48. Przybytek, P.M.a.M.R., *Minimizing total completion time in flowshop with availability constraint on the first machine*. Federated Conference on Computer Science and Information Systems (FedCSIS), 2016: p. 1175-1178.
49. Llatas, C.F., et al., *Minería de procesos interactiva: Aproximando el Big Data a la práctica clínica*. I+ S: informática y salud, 2017(124): p. 44-50.
50. Yzquierdo-Herrera, R., *Minería de proceso como herramienta para la auditoria*. Revista Ciencia de la Información, 2013. **44**(2).
51. Verenich, I., et al. *Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring*. in *International Conference on Business Process Management*. 2015: Springer.
52. Leontjeva, A., et al. *Complex symbolic sequence encodings for predictive monitoring of business processes*. in *International Conference on Business Process Management*. 2015: Springer.
53. Mukala, P., et al. *Learning Analytics on Coursera Event Data: A Process Mining Approach*. in *SIMPDA*. 2015.
54. Calderón-Ruiz, G.E. and H.R. Velarde-Bedregal, *La minería de procesos, una tecnología para automatizar la gestión de procesos de negocio*. Ingenio Magno, 2016. **7**(1): p. 102-115.
55. Cerone, A. *Process mining as a modelling tool: beyond the domain of business process management*. in *International Conference on Software Engineering and Formal Methods*. 2015: Springer.
56. Mans, R., W. van der Aalsta, and R.J. Vanwerschb, *Process Mining in Healthcare*. Springer Briefs in Business Process Management. Springer

International Publishing: Cham, Germany, 2015.

57. Rojas, E., et al., *Question-driven methodology for analyzing emergency room processes using process mining*. Applied Sciences, 2017. **7**(3): p. 302.
58. Rey, N.H., *Aplicación de técnicas de process mining a la secuencia de logs de eventos de un proceso de negocio*. Tecnología Investigación y Academia, 2017. **5**(1): p. 19-26.
59. Norambuena, B.K. and V.V. Zepeda, *Minería de procesos de software: una revisión de experiencias de aplicación*. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, 2017(21): p. 51-66.
60. Crespo Afonso, E., et al., *Virtual reality and process mining applied to operator training in complex assembly tasks*. 2018.
61. Tax, N., et al., *Event abstraction for process mining using supervised learning techniques*. arXiv preprint arXiv:1606.07283, 2016.
62. Fernandez-Llatas, C., et al., *Process mining methodology for health process tracking using real-time indoor location systems*. Sensors, 2015. **15**(12): p. 29821-29840.
63. Suriadi, S., et al., *Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs*. Information systems, 2017. **64**: p. 132-150.
64. Liu, M., et al. *E-cube: Multi-dimensional event sequence processing using concept and pattern hierarchies*. in *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. 2010: IEEE.
65. van der Aalst, W.M.P., *Mining Process Cubes from Event Data (PROCUBE), project proposal (under review)*. 2012.
66. Bolt, A. and W.M. van der Aalst. *Multidimensional process mining using process cubes*. in *International Conference on Enterprise, Business-Process and Information Systems Modeling*. 2015: Springer.
67. Niemi, T., et al., *Detecting summarizability in OLAP*. Data & Knowledge Engineering. Elsevier, 2014. **89**: p. 1-20.
68. Shoshani, A. *OLAP and statistical databases: Similarities and differences*. in *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. 1997: ACM.
69. van der Aalst, W.M. *Distributed process discovery and conformance checking*. in *International Conference on Fundamental Approaches to Software Engineering*. 2012. Berlin: Springer-Verlag.

70. Van Der Aalst, W.M. *A general divide and conquer approach for process mining*. in *Federated Conference on Computer Science and Information Systems (FedCSIS), 2013*. 2013: IEEE.
71. Van der Aalst, W.M., *Decomposing Petri nets for process mining: A generic approach*. *Distributed and Parallel Databases*, 2013. **31**(4): p. 471-507.
72. Koncilia, C., H. Pichler, and R. Wrembel. *A generic data warehouse architecture for analyzing workflow logs*. in *East European Conference on Advances in Databases and Information Systems*. 2015: Springer.
73. Promtools.org. 2018 23 Jun. 2018]; Available from: <http://www.promtools.org/doku.php>.
74. Processmining.org. *Process Mining*. 2018 23 Jun. 2018]; Available from: <http://www.processmining.org/prom/start>.
75. Gupta, M. and A. Sureka. *Process cube for software defect resolution*. in *Software Engineering Conference (APSEC), 2014 21st Asia-Pacific*. 2014: IEEE.
76. Arias, M., et al. *A framework for recommending resource allocation based on process mining*. in *International Conference on Business Process Management*. 2015: Springer.
77. Hompes, B., et al. *Discovering deviating cases and process variants using trace clustering*. in *Proceedings of the 27th Benelux Conference on Artificial Intelligence (BNAIC)*. 2015.
78. Kamber, M., J. Han, and J. Pei, *Data mining: Concepts and techniques*. 2012: Elsevier.
79. van der Aalst, W.M., S. Guo, and P. Gorissen. *Comparative process mining in education: An approach based on process cubes*. in *International Symposium on Data-Driven Process Discovery and Analysis*. 2015: Springer.
80. Vogelgesang, T. and H.-J. Appelrath. *Multidimensional Process Mining with PMCube Explorer*. in *BPM (Demos)*. 2015.
81. Bolt, A., et al. *Exploiting Process Cubes, Analytic Workflows and Process Mining for Business Process Reporting: A Case Study in Education*. in *SIMPDA*. 2015.
82. Vogelgesang, T., et al. *Multidimensional process mining: questions, requirements, and limitations*. in *Proceedings of the CAiSE*. 2016.
83. Schumacher, C., et al., *Using Data-driven and Process Mining Techniques for Identifying and Characterizing Problem Gamblers in New Zealand*. 2017.
84. SURIADI, S., et al, *Characterizing problem gamblers in New Zealand: A novel expression of process cubes*. 2016.
85. Bayerl, S. and M. Granitzer. *Bacon: Linked Data Integration Based on the RDF Data Cube Vocabulary*. in *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*. 2015: ACM.
86. Bayerl, S. and M. Granitzer. *Discovering, Ranking and Merging RDF Data Cubes*. in *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*. 2017: IEEE.
87. Günther, C. and E. Verbeek, *XES-Standard Definition (2012)*. www.xes-standard.org. Acessado em, 2013. **22**.
88. Grigorova, K., E. Malysheva, and S. Bobrovskiy. *Application of Data Mining and Process Mining approaches for improving e-Learning Processes*. in *Информационные технологии и нанотехнологии (ИТНТ-2017)*. 2017.
89. García, A.O., D. Pérez Alfonso, and V. Estrada Sentí, *Revisión de los principales modelos para aplicar técnicas de Minería de Procesos (Review of models for applying process mining techniques)*. 2016.
90. Barnes, D.J., M. Kölling, and J. Gosling, *Objects first with Java: A practical introduction using Bluej*. 2017: Pearson.
91. Keith, M., M. Schincariol, and M. Nardone, *Introduction*, in *Pro JPA 2 in Java EE 8*. 2018, Springer. p. 1-24.
92. MySQL. *Sitio oficial de MySQL*. 2018 [cited 2018 10-02-2018]; Oracle Corporation:[Available from: <https://www.mysql.com/it/why-mysql/white-papers/10-razones-para-elegir-mysql-para-las-aplicaciones-web-de-la-proxima-generacion/>].
93. Zhang, P., *Practical Guide for Oracle SQL, T-SQL and MySQL*. 2017.
94. Eclipse. *Sitio oficial de Eclipse*. 2018 10-02-2018]; Eclipse:[Available from: <http://www.eclipse.org/ide/>].
95. JUnit. *Sitio ificial de JUnit*. 2018 07-02-2018]; JUnit:[Available from: <https://junit.org>].
96. AG, J. *Sitio oficial de Jedox*. 2016 10-12-2016]; Available from: <https://www.jedox.com>.

97. Taniar, D. and L. Chen, *Integrations of Data Warehousing, Data Mining and Database Technologies*. 2011: Information Science Reference.
98. Benatallah, B. and H.R. Motahari-Nezhad, *Scalable graph-based OLAP analytics over process execution data*. Distributed and Parallel Databases, 2016. **34**(3): p. 379-423.
99. García, A.O., *MODELO PARA LA DETECCIÓN DE VARIABILIDAD EN PROCESOS HOSPITALARIOS UTILIZANDO TÉCNICAS DE MINERÍA DE PROCESOS*. 2016, Universidad de las Ciencias Informáticas.
100. Fernández, R.L. and D.P. Urquiza, *Reseña histórica del concepto de Probabilidades*. Gestión Ingenio y Sociedad, 2017. **2**(1): p. 67-73.
101. Hernández Sampieri, R., C. Fernández Collado, and P. Baptista Lucio, *Metodología de la investigación. Sexta Edición. Editorial Mc Graw Hill. México. 2014*• Hernández, R. Metodología de la Investigación. 6a Edición, Mc Graw Hill, México, 2014.
102. Weiner, A.B., et al., *The Influence of decision aids on prostate cancer screening preferences: A randomized survey study*. The Journal of urology, 2018.
103. Ivanov, O., V. Ivanova, and A. Saltan, *Likert-scale questionnaires as an educational tool in teaching discrete mathematics*. International Journal of Mathematical Education in Science and Technology, 2018: p. 1-9.
104. Garcia, A.O., V.E. Senti, and D.P. Alfonso, *MVD: A Model for the Detection of Variability in Hospital Processes Using Process Mining*. IEEE Latin America Transactions, 2018. **16**(3): p. 966-974.
105. Ramírez Pérez, J.F., et al., *Knowledge management and analysis model for the selection of surgical work teams in health information systems applying organizational intelligence techniques*. Revista Cubana de Información en Ciencias de la Salud (ACIMED), 2017. **28**(1): p. 42-59.
106. Alonso, J.A.G. and M.P. Santacruz, *Cálculo e interpretación del Alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario, con dos posibles escalas tipo Likert*. Publicando, 2015. **2**(1): p. 66-77.
107. Flores, I.G. and V.M. Miguel, *Una contribución a la gestión de la información de ciencia, tecnología e innovación/A contribution to the management of information science, technology and innovation*. Vivat Academia, 2017(140): p. 55-63.
108. Collado, N.V. and A.L. Padrón, *El trabajo docente-metodológico en el Departamento Carrera de Contabilidad y Finanzas de la Universidad Agraria de La Habana: estrategia metodológica para su perfeccionamiento/Teaching and Methodological work into the Accounting and Finance degree at the Agrarian University of Havana: A Methodological strategy for its Improvement*. Revista Cubana de Educación Superior, 2017. **3**(3): p. 64-72.
109. Herrera, Y.R., et al., *Plan de acciones dirigido a las estrategias de autoaprendizaje en estudiantes de primer semestre de Enfermería*. Educación Médica Superior, 2018. **31**(4).
110. Martínez, L.A.C., M.I. León, and A.D.E. Brito, *Valoración de la aplicación del sistema de tareas docentes para la formación de la habilidad toma de decisiones médicas en la asignatura Medicina Interna*. Revista Finlay, 2017. **7**(1): p. 37-49.
111. Ponce, D.K.P., C.V.P. Alvarez, and J. Elizabeth, *Viabilidad Del Modelo De Gestion De La Innovacion Para Microempresas Familiares Del Sector Alimenticio*. Global Journal of Management And Business Research, 2018.
112. Hernández, P. and T. Pérez, *Criterios para verificar y validar mecanismos en el desarrollo de videojuegos*. Revista Antioqueña de las Ciencias Computacionales, 2017. **7**(1).
113. Ortiz, D.M.L., et al., *MODELO CONCEPTUAL DE INNOVACIÓN PARA LA COOPERACIÓN CIENTÍFICA*. Latin American Journal of Business Management, 2017. **8**(1).
114. Barroso, W.Q., A.M.P. Núñez, and Y.C.C. Girón, *LA EVALUACIÓN DE LA CONDICIÓN FÍSICA EN LA ENSEÑANZA PRIMARIA EN DIFERENTES CONDICIONES SOCIOAMBIENTALES*. Revista Magazine de las Ciencias. ISSN 2528-8091, 2017. **2**(1): p. 13-30.
115. Mar Cornelio, O. and J. González Gulín, *MODELO PARA LA EVALUACIÓN DE HABILIDADES EN INGENIERÍA AUTOMÁTICA*. 3C TIC, 2018. **7**(1).

116. Kasavuve, M.K. and W.G. Hernández, *Estrategia metodológica para contribuir a la evaluación desarrolladora de los conocimientos matemáticos en la formación de profesores de matemática de la escuela de formación de profesores*. Revista Publicando, 2017. **3**(9): p. 150-172.
117. Nugzar, T., *Impact of food labeling on consumers buying decision (Georgian case)*. International Journal of Innovative Technologies in Economy, 2018(1 (13)).
118. Todua, N. and T. Gogitidze, *Georgian Farmers' Attitudes Towards Genetically Modified Crops*. Economics World, 2017. **5**(4): p. 362-369.
119. Iadov, V., *Work Satisfaction: An Analysis of Empirical Generalizations and an Attempt at Their Theoretical Interpretation*, in *Revival: Soviet Work Attitudes (1979)*. 2017, Routledge. p. 43-59.

ANEXOS

Anexo 1. Preguntas realizadas a los posibles expertos para determinar su coeficiente de competencia.

1. Defina el grado de conocimiento que usted posee sobre análisis y diagnóstico de procesos de negocio. Defina el valor en escala ascendente de 0 a 10 ____.
2. Para los siguientes indicadores defina el grado de influencia de cada uno, seleccionando A (alto), M (medio) o B (bajo):
 - a. __ Estudios teóricos realizados por usted.
 - b. __ Experiencia adquirida durante su vida profesional.
 - c. __ Conocimiento de investigaciones o publicaciones nacionales e internacionales.
 - d. __ Conocimiento propio sobre el estado del tema de investigación.
 - e. __ Actualización en cursos de postgrado, diplomados, maestrías, doctorado, etc.
 - f. __ Intuición.

El cálculo del coeficiente de competencia se realiza de la forma siguiente:

$K_{comp} = \frac{1}{2} (K_c + K_a)$ siendo:

K_{comp} : Coeficiente de competencia.

K_c : Coeficiente de conocimiento o información que tiene el experto acerca del problema. Es el valor asignado por el propio experto multiplicado por 0.1.

K_a : Coeficiente de argumentación o fundamentación de los criterios del experto, obtenido como resultado de la suma de los puntos de acuerdo a la siguiente tabla patrón:

Tabla 5 Fuentes de argumentación del conocimiento de los expertos.

| No | Fuentes de argumentación | Alto (A) | Medio (M) | Bajo (B) |
|----|---|----------|-----------|----------|
| 1 | Estudios teóricos realizados por usted. | 0,30 | 0,20 | 0,10 |
| 2 | Experiencia adquirida durante su vida profesional. | 0,50 | 0,37 | 0,30 |
| 3 | Conocimiento de investigaciones o publicaciones nacionales e internacionales. | 0,05 | 0,04 | 0,03 |
| 4 | Conocimiento propio sobre el estado del tema de investigación. | 0,05 | 0,04 | 0,03 |
| 5 | Actualización en cursos de postgrado, diplomados, maestrías, doctorado, etc. | 0,05 | 0,04 | 0,03 |
| 6 | Intuición | 0,05 | 0,03 | 0,02 |

| | | | | |
|--|--------------|-------------|-------------|-------------|
| | Total | 1,00 | 0,70 | 0,50 |
|--|--------------|-------------|-------------|-------------|

Si $K_{comp} > 0,7$ entonces la competencia del experto es Alta (A).

Si $0,5 < K_{comp} = < 0,7$ entonces la competencia del experto es Media (M).

Si $K_{comp} = < 0,5$ entonces la competencia del experto es Baja (B).

Tabla 6 Resultados de la encuesta aplicada a los candidatos a expertos para determinar su nivel de competencia.

| Expertos | Kc | 1 | 2 | 3 | 4 | 5 | 6 | Ka | Kcomp | valor |
|----------|-----|-----|------|------|------|------|------|------|-------|-------|
| 1 | 0.8 | 0.2 | 0.5 | 0.04 | 0.05 | 0.05 | 0.05 | 0.89 | 0.85 | alto |
| 2 | 0.4 | 0.1 | 0.37 | 0.03 | 0.03 | 0.03 | 0.02 | 0.58 | 0.49 | bajo |
| 3 | 0.7 | 0.2 | 0.5 | 0.04 | 0.04 | 0.05 | 0.05 | 0.88 | 0.79 | alto |
| 4 | 0.5 | 0.1 | 0.3 | 0.04 | 0.04 | 0.04 | 0.02 | 0.54 | 0.52 | medio |
| 5 | 0.7 | 0.3 | 0.5 | 0.04 | 0.04 | 0.05 | 0.05 | 0.98 | 0.84 | alto |
| 6 | 0.7 | 0.2 | 0.37 | 0.04 | 0.05 | 0.05 | 0.05 | 0.76 | 0.73 | alto |
| 7 | 0.8 | 0.2 | 0.5 | 0.05 | 0.05 | 0.05 | 0.05 | 0.90 | 0.85 | alto |
| 8 | 0.9 | 0.3 | 0.5 | 0.05 | 0.05 | 0.05 | 0.05 | 1.00 | 0.95 | alto |
| 9 | 0.7 | 0.2 | 0.5 | 0.04 | 0.04 | 0.05 | 0.05 | 0.88 | 0.79 | alto |
| 10 | 0.6 | 0.2 | 0.37 | 0.04 | 0.04 | 0.04 | 0.03 | 0.72 | 0.66 | medio |
| 11 | 0.9 | 0.3 | 0.5 | 0.05 | 0.04 | 0.05 | 0.05 | 0.99 | 0.95 | alto |
| 12 | 0.5 | 0.2 | 0.37 | 0.03 | 0.03 | 0.04 | 0.02 | 0.69 | 0.60 | medio |
| 13 | 0.7 | 0.2 | 0.37 | 0.04 | 0.05 | 0.04 | 0.05 | 0.75 | 0.73 | alto |
| 14 | 0.8 | 0.2 | 0.5 | 0.04 | 0.05 | 0.05 | 0.05 | 0.89 | 0.85 | alto |
| 15 | 0.5 | 0.2 | 0.3 | 0.04 | 0.04 | 0.04 | 0.03 | 0.65 | 0.58 | medio |
| 16 | 0.6 | 0.2 | 0.37 | 0.04 | 0.04 | 0.04 | 0.03 | 0.72 | 0.66 | medio |
| 17 | 0.6 | 0.3 | 0.5 | 0.03 | 0.03 | 0.03 | 0.02 | 0.91 | 0.76 | alto |
| 18 | 0.5 | 0.1 | 0.37 | 0.03 | 0.03 | 0.04 | 0.03 | 0.60 | 0.55 | medio |
| 19 | 0.6 | 0.2 | 0.37 | 0.04 | 0.04 | 0.04 | 0.03 | 0.72 | 0.66 | medio |
| 20 | 0.4 | 0.1 | 0.37 | 0.03 | 0.03 | 0.03 | 0.02 | 0.58 | 0.49 | bajo |
| 21 | 0.7 | 0.2 | 0.37 | 0.05 | 0.04 | 0.04 | 0.03 | 0.73 | 0.72 | alto |
| 22 | 0.8 | 0.2 | 0.37 | 0.03 | 0.03 | 0.03 | 0.03 | 0.69 | 0.75 | alto |
| 23 | 0.7 | 0.3 | 0.5 | 0.04 | 0.03 | 0.04 | 0.03 | 0.94 | 0.82 | alto |

Anexo 2. Valoración de los expertos sobre cada formulación.

| Experto | Indicador | | | | | | |
|---------|-----------|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 4 | 3 | 4 | 4 | 4 | 4 | 4 |
| 2 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 3 | 5 | 5 | 4 | 4 | 5 | 5 | 5 |
| 4 | 5 | 4 | 5 | 3 | 4 | 5 | 5 |

| | | | | | | | |
|----|---|---|---|---|---|---|---|
| 5 | 4 | 5 | 5 | 5 | 4 | 4 | 5 |
| 6 | 4 | 5 | 5 | 5 | 5 | 5 | 5 |
| 7 | 5 | 4 | 4 | 4 | 4 | 5 | 4 |
| 8 | 5 | 4 | 5 | 5 | 5 | 4 | 4 |
| 9 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 10 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| 11 | 3 | 4 | 4 | 3 | 4 | 4 | 4 |
| 12 | 5 | 5 | 5 | 5 | 4 | 5 | 4 |
| 13 | 5 | 5 | 4 | 5 | 4 | 5 | 5 |
| 14 | 4 | 4 | 4 | 4 | 5 | 4 | 4 |
| 15 | 4 | 5 | 5 | 5 | 4 | 5 | 5 |
| 16 | 5 | 5 | 4 | 4 | 5 | 5 | 4 |
| 17 | 4 | 5 | 5 | 4 | 5 | 4 | 5 |
| 18 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 19 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| 20 | 5 | 4 | 4 | 4 | 3 | 4 | 5 |
| 21 | 4 | 5 | 5 | 5 | 5 | 5 | 4 |

Anexo 3. Determinación de la confiabilidad de los datos obtenidos en la encuesta a expertos.

Los resultados fueron validados a partir del coeficiente Alpha de Cronbach.

La fórmula del coeficiente Alpha de Cronbach requiere de una sola aplicación del instrumento y se basa en la medición de la respuesta del sujeto con respecto a los elementos del instrumento.

$$\alpha = \frac{K}{K-1} \left[1 - \frac{\sum S_i^2}{S_T^2} \right]$$

K: El número de ítems.

$\sum S_i^2$: Sumatoria de varianzas de los elementos (ítems).

S_T^2 : Varianza de la suma de los elementos (ítems).

α : Coeficiente Alpha de Cronbach.

Se procedió a calcular la varianza de la población de los 21 expertos a los que se le aplicó la encuesta. Para ello se utilizó Microsoft Excel. Las respuestas fueron codificadas a los valores 1, 2, 3, 4 y 5. Los valores obtenidos para cada ítem fueron:

Tabla 7 Varianza de los resultados de la encuesta por preguntas.

| Ítems | Varianza de la población |
|------------|--------------------------|
| Pregunta 1 | 0.34 |
| Pregunta 2 | 0.33 |
| Pregunta 3 | 0.25 |
| Pregunta 4 | 0.44 |
| Pregunta 5 | 0.34 |
| Pregunta 6 | 0.22 |
| Pregunta 7 | 0.24 |

La sumatoria de varianzas de los ítems ($\sum S_i^2$) dio como resultado 2.16 y la varianza de la suma de los ítems (S_T^2) resultó 6.10. Al aplicar la fórmula anterior y conociendo que el valor del número de ítems (K) es 7, el coeficiente Alpha de Cronbach resultó 0, 75.

$$\alpha = \frac{7}{7 - 1} \left[1 - \frac{2,16}{6.10} \right]$$

$\alpha = 0.75$