UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION

HERRAMIENTA ETL PARA LOGS DE PROCESOS

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL EN COMPUTACION

JORGE FRANCISCO TORO VALDIVIA

PROFESORA GUÍA:
MARIA CECILIA BASTARRICA PIÑEYRO

MIEMBROS DE LA COMISIÓN:
ROMAIN ROBBES
JOSÉ ALBERTO PINO URTUBIA

Este trabajo ha sido parcialmente financiado por Proyecto Gems

SANTIAGO DE CHILE
2016

Resumen

Hoy en día las empresas nacen como respuesta a distintos problemas o desafíos entregando diversas de soluciones, las cuales se llevan a cabo ejecutando una serie de procesos. Estos pueden estar definidos formalmente o se van construyendo de forma natural a partir de la forma en que las personas abordan los problemas. Por esto, se hace fundamental poder realizar distintos análisis sobre dichos procesos, en especial descubrir cómo se llevan a cabo en forma real. Estos análisis nos permiten optimizar, mejorar, transformar estos procesos permitiendo a las empresas y grupos de trabajo evolucionar.

Muchos de estos análisis, y los programas que los implementan, suelen ser de alta complejidad y tienden a tomar datos de los procesos en formatos específicos. Esto ocasiona que los registros de datos, también conocidos como logs de procesos, deban adaptarse a dichos formatos y por consiguiente a las herramientas y/o personas que los generan. Otra opción, es realizar un desarrollo sobre los programas de análisis para que puedan aceptar un nuevo formato, lo que tiene un costo proporcional a la complejidad del software.

Es por esto que se propuso una nueva opción: desarrollar una herramienta ETL intermediaria que nos permita transformar logs desde distintos formatos a un formato utilizado por programas de descubrimientos de procesos u otros tipos de análisis relacionados.

Se logró desarrollar una aplicación que, a través de una GUI, permite hacer distintas transformaciones desde bases de datos y planillas Excel al formato XES, el cual es usado por varios programas de descubrimiento de procesos. Esta aplicación permite al usuario configurar sus conversiones eligiendo los datos de origen así como su tipo.

Este programa fue validado a través de distintas pruebas, utilizando datos reales y un programa que realiza descubrimientos de procesos sobre archivos XES. Este desarrollo ha permitido abarcar una amplia gama de formatos sin tener que modificar datos ni los otros programas que los utilizaban, permitiendo que diversos procesos puedan ser estudiados y mejorados.

Agradecimientos

En primer lugar agradecer a Dios por la vida y la vocación de servicio.

Gracias a mi familia que siempre estuvo presente apoyándome.

A mis amigos y comunidad.

A mi profesora guía Cecilia Bastarrica por ayudarme en todo con este trabajo de tesis.

A Fabian Rojas por ayudar con distintas pruebas de usuario final y dar feedback tanto en funcionalidad y usabilidad del software.

Al proyecto GEMS.

Por últimos gracias por los iconos diseñados por Freepik.

Tabla de contenido

1. INTROUCCIÓN

1.1 Contexto del Problema

Hoy en día las empresas nacen como respuesta a distintos problemas o desafíos entregando una variedad de soluciones, las cuales se destacan por distintos factores como calidad, precio, tiempos de respuesta, etc. Para que las empresas puedan llegar a construir estas soluciones ejecutan una serie de procesos, siendo estos desde analizar la realidad de una situación a tratar, hasta entregar la solución desarrollada.

Estos procesos pueden estar definidos formalmente o se van construyendo de forma natural a partir de la forma en que las personas abordan los problemas a enfrentar. De una u otra manera ir optimizando, mejorando, transformado estos procesos es una de las formas en como las empresas y grupos de trabajos van evolucionando. Realizando estos cambios de forma consciente y sistemática se logran evoluciones más favorables. Desde este hecho es que toma gran relevancia poder descubrir y definir los procesos ejecutados por las personas. Incluso en los casos donde existe una definición formal de lo que se quiere hacer, muchas veces la gente realiza tareas dentro de los proyectos de forma distinta o realizan tareas que no fueron consideradas a la hora de definir el proceso.

Registrar las distintas acciones y tareas realizadas al ejecutar un proceso, de forma manual o automáticamente, permite el análisis de éste y de sus resultados. Además, hoy las tecnologías y herramientas computacionales se han vuelto indispensables a la hora de llevar a cabo cualquier tarea, por lo que también se encuentran de forma transversal en el desarrollo de los procesos. En la mayoría de estas herramientas está presente la capacidad para registrar información en un log.

Considerando esto, es que una forma de realizar análisis sobre los procesos, es aprovechar los mismos sistemas computacionales usados para llevarlo a cabo. Para esto es fundamental que los logs guarden la información requerida para los análisis. En la mayoría de los casos las herramientas que generan los logs de procesos son distintas a las que realizan los análisis de procesos. Uno de estos análisis es conocido como "descubrimiento de proceso" y actualmente corresponde a toda una área de estudio.

Los procedimientos desarrollados para obtener información desde los datos de un log no son triviales y estos pueden aumentar su dificultad debido a la gran cantidad de datos que posee cada registro, especialmente si esto se hace de forma manual y sin una herramienta adecuada. Se debe considerar que los formatos de log pueden variar desde un sistema a otro, e incluso

dentro de los distintos módulos de un mismo programa y/o empresa, ya que estos son definidos por los desarrolladores y/o las herramientas que utilizan.

Este último punto hace que los procesos estén registrados de diversas maneras, lo que dificulta la conversación entre los logs y las herramientas que los analizarán.

A partir de los datos obtenidos desde un log de proceso se pueden aplicar distintas operaciones para obtener información valiosa. En específico, en el área de descubrimiento de procesos existen distintas aproximaciones y estudios para poder, a partir del log, definir el proceso y comenzar un desarrollo sobre él. Incluso en los casos donde ya existe un proceso definido, el descubrimiento de éste permite realizar distintas comparaciones, evaluaciones y mejoras.

Respecto a otros tipos de análisis, por ejemplo, si un proceso definido tiene costos apropiadamente estimados, ya sea en materiales, tiempo, gente requerida, etc., a través de los logs podemos obtener los costos reales y podemos optimizar un proceso desde simplemente mejorar futuras estimaciones hasta tomar medidas para alcanzar las estimaciones hechas.

Estos análisis sobre los procesos suelen ser de alta complejidad y tienden a tomar un log en un formato específico, ya sea por las decisiones tomadas considerando los distintos requerimientos y restricciones de un problema, o la dificultad al momento de tomar un formato específico, entre otras razones. Esto provoca que los análisis desarrollados tengan un límite en cuanto a la información que utilizan. Esto genera una complicación que tiene mayor importancia según se consideren mayor cantidad de formatos a analizar y el nivel de generalidad de estos mismos. De haber más de un formato entre los logs de distintas etapas de un proceso, puede que los análisis no abarquen la totalidad de éste o peor aún, que nos entreguen información insuficiente o incorrecta.

Por eso surge la necesidad de potenciar una aplicación que realice descubrimiento de procesos, a través de una herramienta de software intermedia que amplíe el espectro de formatos con los que se pueda trabajar. Esto implica transformar los distintos formatos de logs a uno específico, el cual sea usado por aplicaciones de descubrimiento de proceso y así reutilizar lo que ya se tiene. Con esto aumentamos el alcance de soluciones y herramientas ya existentes que realicen descubrimiento de procesos, de forma que una sola aplicación se puede enfocar en el análisis específico sin tener que preocuparse del formato de origen dando la posibilidad de distribuir el trabajo de mejor forma.

Para acotar el alcance de este proyecto es que se ha decidido restringir el formato de salida para los logs de proceso para el estandar "extensible event stream" (XES). Esto ha sido considerando que este formato se desarrolló enfocado en ser usado para registro de eventos y principalmente debido a que ya es usado por varias aplicaciones de análisis y descubrimiento de procesos.

En la figura 1 se muestra cómo el transformador de formatos propuesto nos permite abarcar diversos formatos de logs, utilizando un único módulo de descubrimiento de procesos.

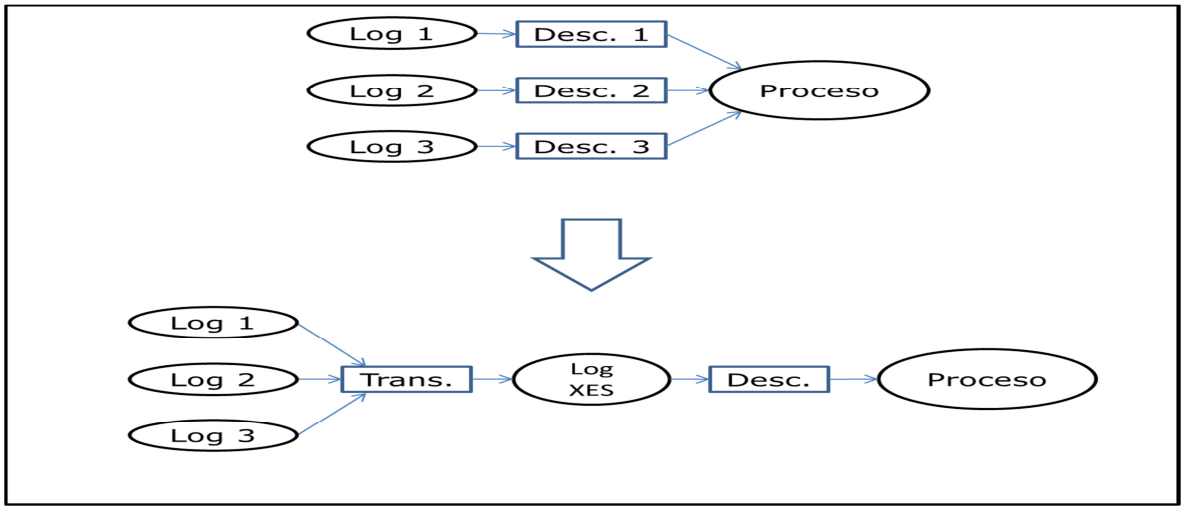


Figura 1: Aplicación en contexto

En la parte superior de esta figura vemos la situación actual, donde por cada tipo de formato de logs de eventos, se debe construir un programa que realice el descubrimiento de procesos y se adecúe al formato del log. Mientras que en la parte inferior se encuentra la solución a desarrollar, la cual independiza a la aplicación que realiza el descubrimiento de procesos de las especificaciones de los formatos de origen, concentrándose en un solo tipo de log XES y transfiriéndole la responsabilidad de entender los formatos de origen al transformador.

1.2 Objetivos

1.2.1 Objetivo General

Desarrollar una herramienta computacional para transformar logs de procesos desde distintos formatos definidos por el usuario al formato XES para permitir descubrimiento de procesos.

1.2.2 Objetivos Específicos

- Implementar descripción de formatos de origen.

- Desarrollar un software de extracción de datos de logs.
- Implementar reglas de filtros y operaciones sobre los datos por el usuario.
- Transformar datos desde un formato de origen al formato de destino XES.
- Implementar interfaz gráfica para uso de la aplicación.

1.3 Alternativas

Dentro de las distintas alternativas analizadas que nos permiten transformar formatos, nos encontramos con algunas aplicaciones que abarcan en parte este problema.

1.3.1 Disco

Este software desarrollado por Fluxicon [6] tiene la posibilidad de importar logs de eventos en formato "comma-separated values" (CSV), Excel y XES para luego poder exportar al estándar requerido. Tiene un gran potencial ya que posee una interfaz gráfica que permite visualmente elegir los datos a procesar y distribuirlos como se ve en la figura 2. También construye una red de Petri que representa al proceso descubierto, como se ve en la figura 3, dándonos valiosa información acerca del proceso. A esto se suma una gran cantidad de datos estadísticos respecto a la muestra tales como frecuencia de los eventos que componen el proceso, por ejemplo. Una de las desventajas es que este producto es pagado y su versión DEMO sólo nos permite trabajar hasta 100 registros. Aunque este programa permite exportar datos al estándar XES, el log exportado tiene un formato específico para ser utilizado por Disco y no puede ser configurado para otros programas.

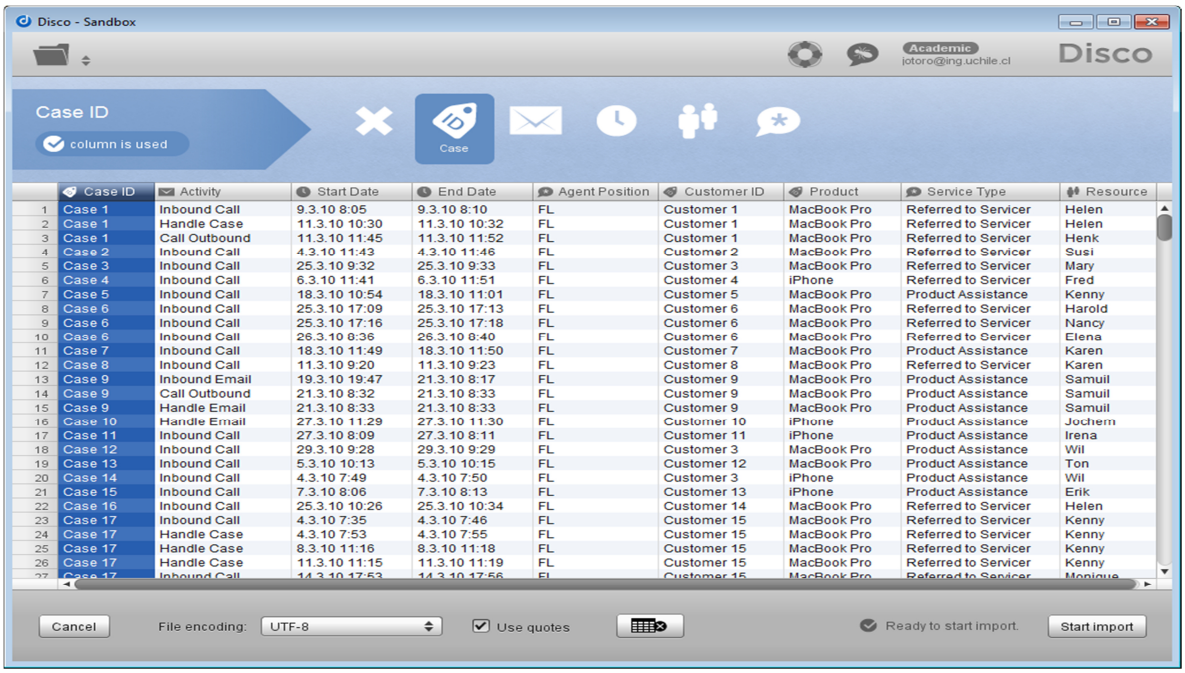


Figura 2: Disco, selección de datos

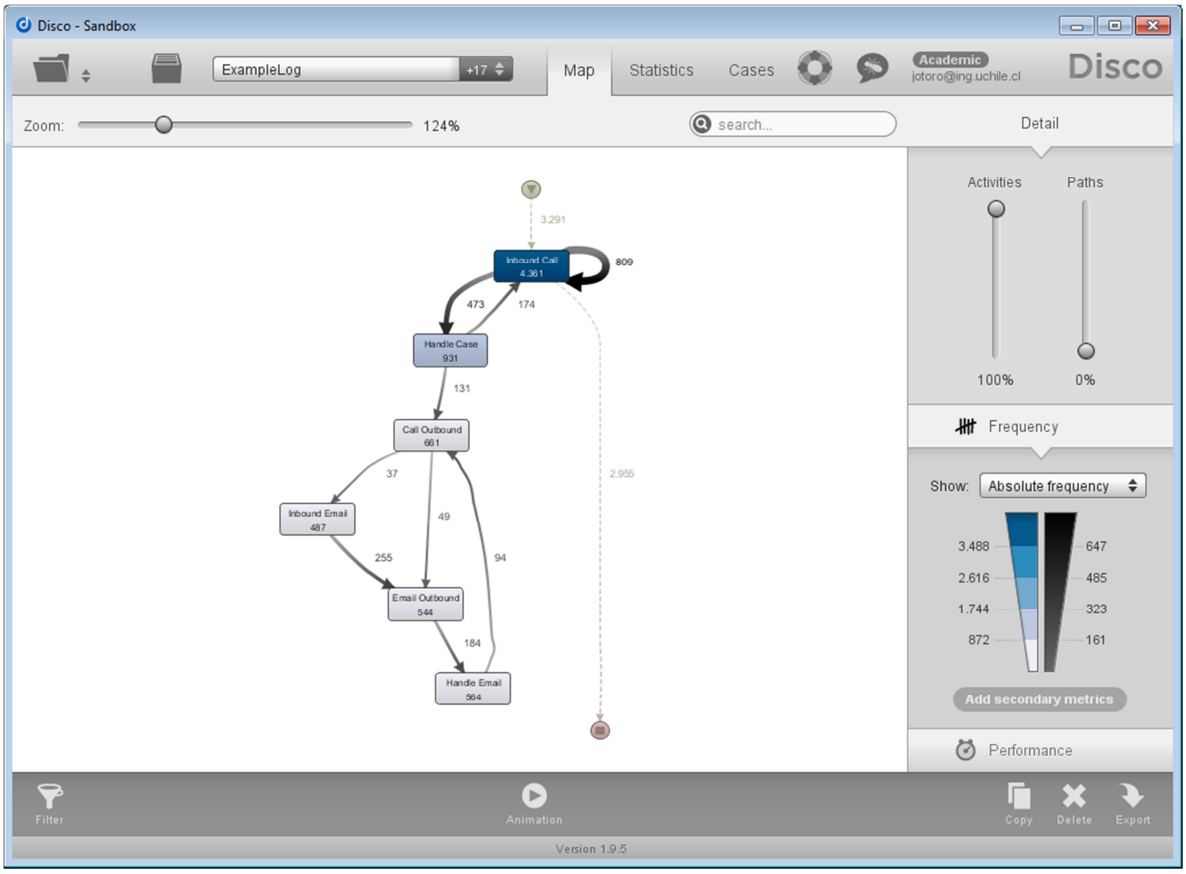


Figura 3: Disco, modelamiento del proceso

1.3.2 Prom y XESame

Ambos programas fueron creados por Technische Universiteit Eindhoven [7] [13]. El software XESame inserto en Prom (figura 4) es el encargado de realizar la transformación de formato deseada. Es una solución que, al igual que el software anterior, ocupa un solo tipo de entrada, pero en este caso se conecta a una base de datos relacional y de forma genérica. Esto nos permitiría utilizar distintos tipos de log almacenados en distintos tipos de bases de datos, como MySQL u otros. Este aspecto nos aumenta a un mayor número de tipos de logs de eventos. Una de las principales desventajas es la dificultad a la hora de usar este software ya que, aunque posee una interfaz gráfica, esta no es intuitiva y no permite de manera fácil poder llegar a obtener la conversión requerida. Además, para poder utilizar las distintas bases de datos, el usuario tiene que proveer todas las herramientas requeridas tales como las librerías necesarias para poder conectarse y las clases a usar. Esto agrega un problema extra al existente, ya que es necesario poseer esas herramientas.

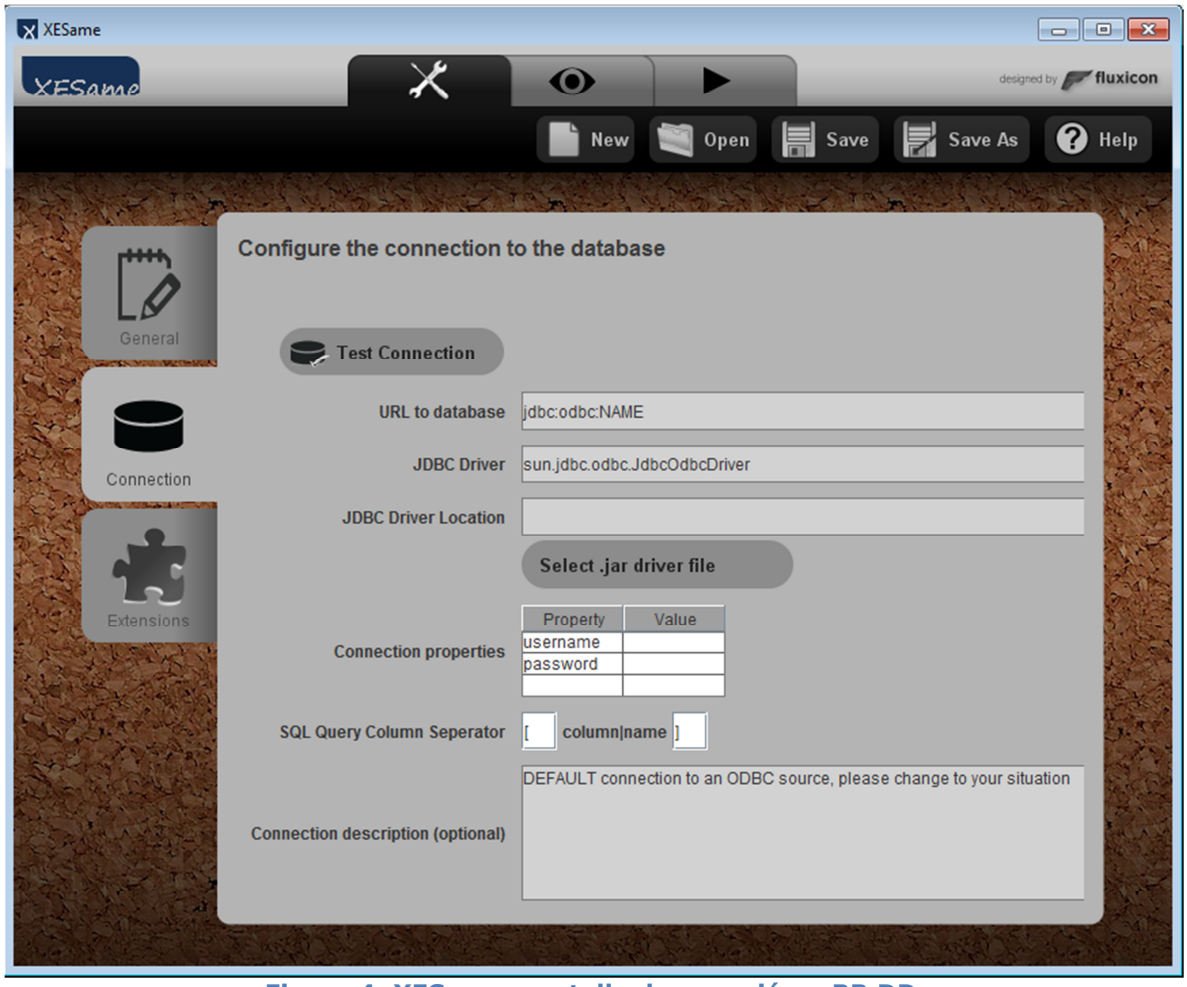


Figura 4: XESame, pantalla de conexión a BB.DD.

1.3.3 OpenXES

A diferencia de los casos anteriores esto es una librería hecha en Java para poder utilizar el formato XES. Esto quiere decir que no es un programa que nos permita obtener otros formatos y convertirlos. La principal ventaja que nos propone es que está directamente relacionada con el estándar y a la vez que este va evolucionando también aparecen versiones nuevas de esta librería. A partir de esta ventaja es que se ha decidido usarla como parte del software propuesto.

1.4 Solución Propuesta

Viendo las características del problema y algunas cualidades comunes entre las alternativas estudiadas, es que se propone desarrollar una aplicación donde, además de poder importar y exportar datos, tenga como principal función la transformación de los formatos de estos. Este tipo de aplicaciones se conocen como ETLs (extract, transform, load).

Considerando el uso de la librería OpenXES para poder leer y escribir los logs en formato XES como parte de la solución, es que se ha decidido desarrollar el programa en Java. Este lenguaje también nos permite abarcar varios contextos de uso como, por ejemplo distintos sistemas operativos.

En las primeras 2 alternativas presentadas en la sección anterior, podemos apreciar la importancia de una interfaz gráfica que permita al usuario poder realizar la conversión de forma sencilla y evitar la necesidad de tener el conocimiento abstracto que se requiere en otros casos.

Con la información obtenida de las alternativas analizadas y el contexto en que se encuentran es que se propone una solución que debe contar con una interfaz gráfica que permita al usuario poder manipular de forma simple las características de estas transformaciones. Esta solución no sólo debe ser capaz de adaptarse a los distintos formatos según tipo de origen (Excel, MySQL, etc), sino que también a la disposición de los datos (nombre de columnas, datos agregados, etc).

1.5 Resultados Obtenidos

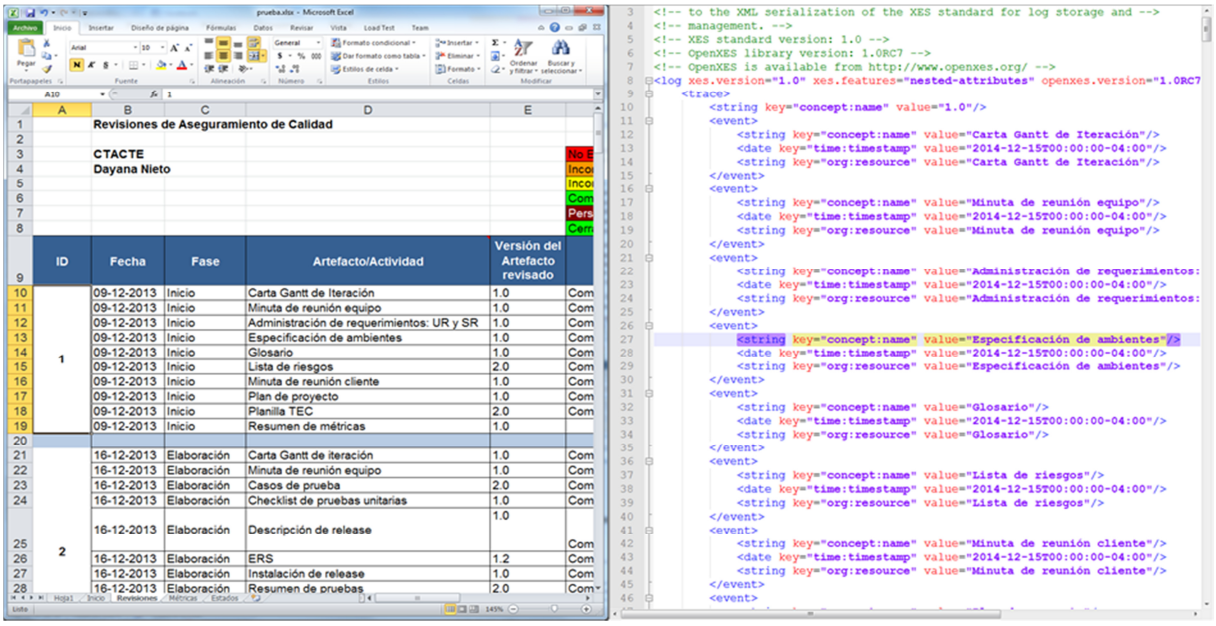
A partir de los requerimientos encontrados se diseñó e implementó una aplicación con una interfaz gráfica que nos permite elegir cómo realizar la transformación de los datos.

Debido a que los formatos de entrada de los logs de proceso pueden ser variados se decidió generalizar la solución dando la libertad a los usuarios para configurar desde las columnas a transformar hasta aplicar unas

funciones básicas sobre los datos. Esta libertad lleva por contraparte mayor trabajo por parte de usuario ya que se evitaron automatizaciones e inteligencia asociados a ciertos formatos específicos.

Se logró hacer distintas transformaciones desde bases de datos y planillas Excel, incluyendo la capacidad de guardar los perfiles de transformación, los cuales facilitaron la tarea al realizarse sobre distintos archivos e inclusive entre un archivo Excel y una consulta en base de datos donde se comparte la misma estructura.

A continuación, en la figura 5, se puede ver un ejemplo de datos transformados a partir del software desarrollado. A la izquierda una planilla Excel con los datos originales y a la derecha el archivo XES transformado.



Revisiones de Aseguramiento de Calidad

CTACTE
Dayana Nieto

ID	Fecha	Fase	Artefacto/Actividad	Versión del Artefacto revisado	
1	09-12-2013	Inicio	Carta Gantt de Iteración	1.0	Com
	09-12-2013	Inicio	Minuta de reunión equipo	1.0	Com
	09-12-2013	Inicio	Administración de requerimientos: UR y SR	1.0	Com
	09-12-2013	Inicio	Especificación de ambientes	1.0	Com
	09-12-2013	Inicio	Glosario	1.0	Com
	09-12-2013	Inicio	Lista de riesgos	2.0	Com
	09-12-2013	Inicio	Minuta de reunión cliente	1.0	Com
	09-12-2013	Inicio	Plan de proyecto	1.0	Com
	09-12-2013	Inicio	Planilla TEC	2.0	Com
	09-12-2013	Inicio	Resumen de métricas	1.0	
2	16-12-2013	Elaboración	Carta Gantt de iteración	1.0	Com
	16-12-2013	Elaboración	Minuta de reunión equipo	1.0	Com
	16-12-2013	Elaboración	Casos de prueba	2.0	Com
	16-12-2013	Elaboración	Checklist de pruebas unitarias	1.0	Com
	16-12-2013	Elaboración	Descripción de release	1.0	Com
	16-12-2013	Elaboración	ERS	1.2	Com
	16-12-2013	Elaboración	Instalación de release	1.0	Com
	16-12-2013	Elaboración	Resumen de pruebas	2.0	Com

```
<!-- to the XML serialization of the XES standard for log storage and -->
<!-- management. -->
<!-- XES standard version: 1.0 -->
<!-- OpenXES library version: 1.0RC7 -->
<!-- OpenXES is available from http://www.openxes.org/ -->
<log xes.version="1.0" xes.features="nested-attributes" openxes.version="1.0RC7
    <trace>
        <string key="concept:name" value="1.0"/>
        <event>
            <string key="concept:name" value="Carta Gantt de Iteración"/>
            <date key="time:timestamp" value="2014-12-15T00:00:00-04:00"/>
            <string key="org:resource" value="Carta Gantt de Iteración"/>
        </event>
        <event>
            <string key="concept:name" value="Minuta de reunión equipo"/>
            <date key="time:timestamp" value="2014-12-15T00:00:00-04:00"/>
            <string key="org:resource" value="Minuta de reunión equipo"/>
        </event>
        <event>
            <string key="concept:name" value="Administración de requerimientos:
            <date key="time:timestamp" value="2014-12-15T00:00:00-04:00"/>
            <string key="org:resource" value="Administración de requerimientos:
        </event>
        <event>
            <string key="concept:name" value="Especificación de ambientes"/>
            <date key="time:timestamp" value="2014-12-15T00:00:00-04:00"/>
            <string key="org:resource" value="Especificación de ambientes"/>
        </event>
        <event>
            <string key="concept:name" value="Glosario"/>
            <date key="time:timestamp" value="2014-12-15T00:00:00-04:00"/>
            <string key="org:resource" value="Glosario"/>
        </event>
        <event>
            <string key="concept:name" value="Lista de riesgos"/>
            <date key="time:timestamp" value="2014-12-15T00:00:00-04:00"/>
            <string key="org:resource" value="Lista de riesgos"/>
        </event>
        <event>
            <string key="concept:name" value="Minuta de reunión cliente"/>
            <date key="time:timestamp" value="2014-12-15T00:00:00-04:00"/>
            <string key="org:resource" value="Minuta de reunión cliente"/>
        </event>
        <event>
```

Figura 5: Datos Transformados

Este ejemplo, al igual que otras pruebas, se realizó con datos reales que validan la funcionalidad de la aplicación. Este programa nos da un punto de partida para realizar otras transformaciones abarcando una cantidad considerable de formatos.

Existe trabajo futuro de esto, tal como, optimizar tiempos de ejecución, detección de errores a la hora de configurar de mala manera las transformaciones, datos estadísticos y análisis básicos sobre los datos convertidos (número de trazas, número de eventos, etc.), pero se puede verificar que abarca el desafío propuesto de buena forma.

2. MARCO TEÓRICO

2.1 Logs de procesos

Un log es un registro de eventos y/o acciones realizadas. En el área de computación, la aplicación más habitual es registrar los diversos eventos de un sistema, que pueden ser realizadas por el sistema o por un usuario. Los logs pueden ser un archivo, un conjunto de archivos, una base de datos, o algún otro formato según la herramienta de apoyo que se use. Tanto la existencia de este log en un software o no, al igual que el nivel de detalle que posee, son definidas por el equipo que desarrolla el sistema.

Un log de software puede guardar acciones fundamentales para el sistema a la hora de ser ejecutado, así como mostrarnos comportamientos de nuestros procesos de desarrollo. Para los desarrolladores y personas encargadas del software puede entregar información de alta relevancia. Por ejemplo, en un software que posee un error, pero que a pesar de hacer distintas pruebas y testeo, este no ha podido ser detectado, el log lo captura y esto hace posible obtener información del por qué y cuándo ocurre, lo que puede ayudar a resolver estos problemas. Aunque dentro de estos logs de software podemos encontrar tareas fundamentales del proceso llevado a cabo utilizando una aplicación, estos logs poseen información no relevante y principalmente tienen otro enfoque a diferencia de un log de proceso.

De forma específica, los logs de procesos guardan datos relevantes sobre el proceso realizado a la hora de diseñar, desarrollar y evaluar una solución. Este puede estar compuesto por tareas y/o eventos que van caracterizando al proceso en su totalidad.

Análogamente a un log de software, un log de proceso nos permite encontrar errores y características fundamentales de éste. Incluso podemos descubrir el proceso ejecutado en la práctica lo que nos da la posibilidad de contrastarlo con la definición formal de éste si es que existe (figura 6).

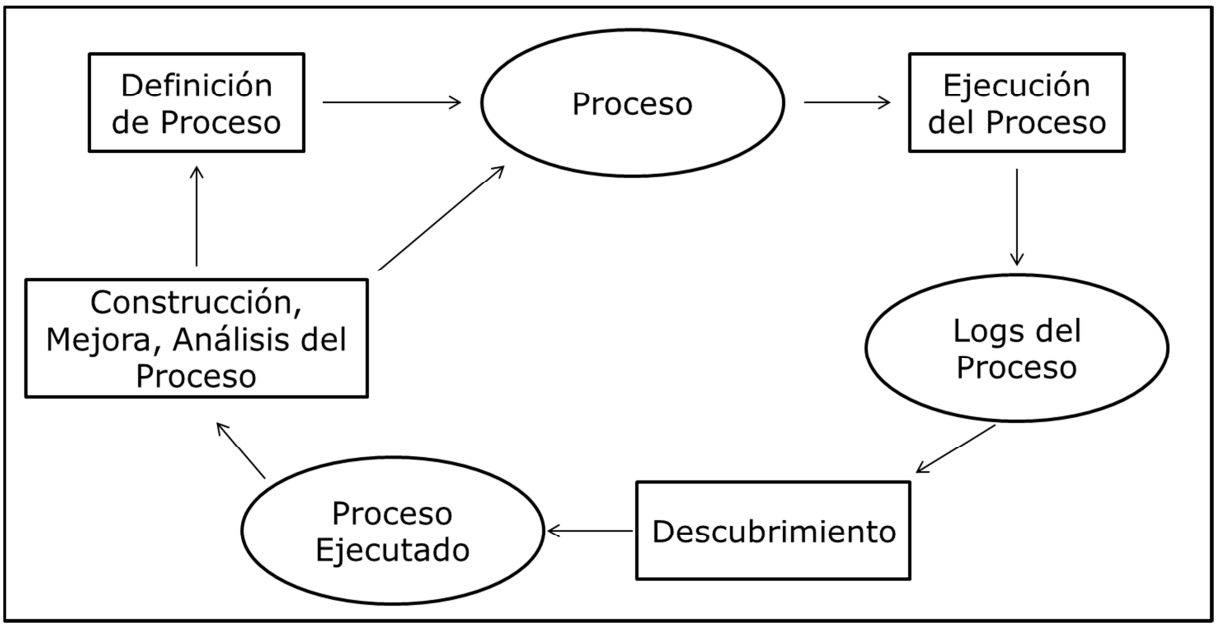


Figura 6: Ciclo de vida del proceso

Wolf y Rosenblum definen tres caracteristicas fundamentales que deben tener los logs de procesos para poder realizar un análisis de calidad: datos concisos, precisos y significativos [11].

Este tipo de logs son fundamentales a la hora de hacer análisis retrospectivos de procesos, es decir, aquellos que se hacen a partir de datos reales de los procesos ejecutados. Un tipo de logs se conoce como modelo basado en eventos. Estos logs consisten en una serie de eventos que contienen como principales atributos un identificador de la acción realizada y el tiempo donde se ejecutó. Cada evento representa una actividad instantánea. Un intervalo se define como el tiempo transcurrido entre dos eventos.

Al utilizarse un log basado en eventos para describir un proceso se asumen las siguientes caracteristicas [2]:

- Cada evento representa una tarea.
- Cada evento ocurre en un caso. Un caso puede ser una instancia del proceso (traza), o las tareas ejecutadas por una persona, entre otros ejemplos.
- Los eventos están ordenados. En la mayoría de los casos esto se logra a través del tiempo de ejecución.

2.2 Formato XES

XES (extensible event stream), es un estándar para logs de eventos basado en XML. Este estándar fue diseñado en base a 4 principios [8]:

- Simplicidad a la hora de representar la información.
- Flexibilidad para capturar eventos desde cualquier tipo de proceso.
- Extensibilidad para poder fácilmente agregar definiciones en el futuro o modificarlo personalmente.
- Expresividad para poder mostrar la mayor cantidad de información necesaria y que pueda ser fácilmente interpretado por las personas.

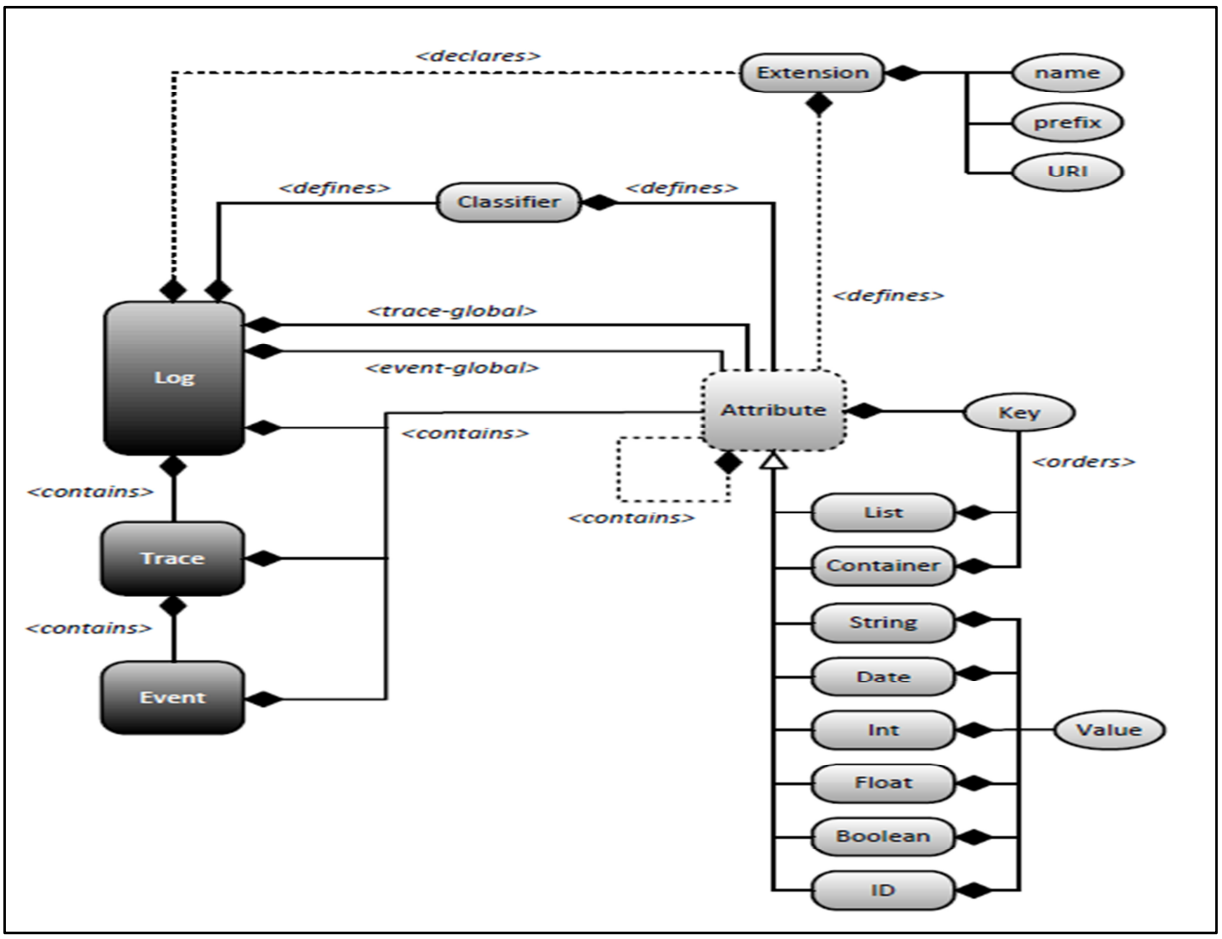


Figura 7: Meta-Modelo del estándar XES [8]

Es importante conocer las características del formato XES y cómo éstas estarán relacionadas a las descripciones de proceso a usar.

- Evento: Son las actividades y/o tareas que componen al proceso y se registran en el log.

- Traza: Es la secuencia de eventos que contiene los datos de una ejecución del proceso.
- Log: Es el elemento que contiene todas las trazas.

La figura 7 nos muestra el meta-modelo que describe a este estándar y cómo se relacionan las entidades nombradas.

2.3 Herramientas ETL

Esta sigla es la abreviación de programas con 3 funciones claras: extraer ("extract"), transformar ("transform") y cargar datos ("load") [9].

- La extracción de datos desde una o varias fuentes de origen también contempla la verificación de datos y filtrado de información relevante.
- La transformación de datos requiere relacionar los formatos de origen y destino, parte fundamental del problema que estamos considerando, y aplicar cualquier función que sea necesaria desde los datos de origen para corresponder a los datos de destino.
- La carga de datos considera el proceso de agregar los datos al nuevo sistema de destino y contemplar las distintas alternativas y/o estrategias para realizarlo.

Transversalmente a estas tres funciones se debe considerar la limpieza y aislamiento de datos, de forma que no se traspasen datos innecesarios y/o corruptos.

Acompañado de estos pasos suele encontrarse un plan de cómo y cuándo debe ejecutarse este tipo de procesos.

2.4 Descubrimiento de procesos

El descubrimiento de procesos es un tipo específico de análisis sobre procesos que tiene como objetivo inferir, a partir de datos recolectados por la ejecución de un proceso, el modelo formal que represente a éste, cómo se muestra en la figura 8.

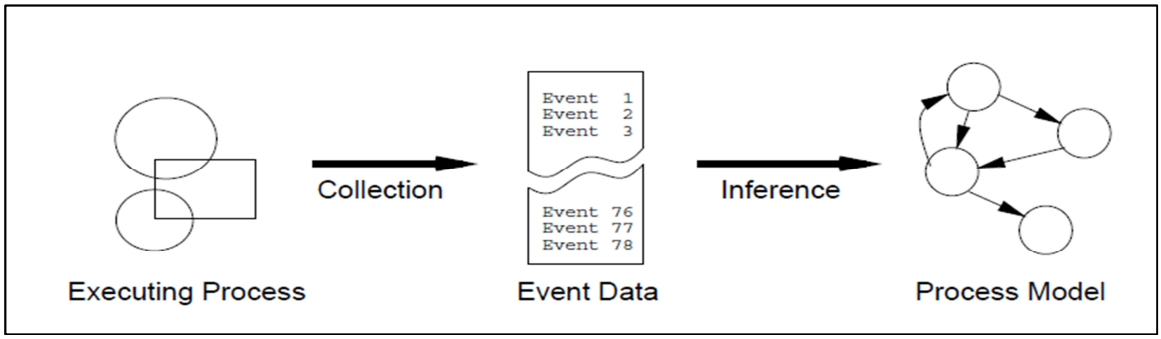


Figura 8: Desarrollo de Descubrimiento de Procesos [1]

También conocido como minería de procesos, este tipo de análisis puede ser diferenciado según la perspectiva que tenga. W.M.P. Van der Aalst [10] considera las siguientes perspectivas:

- Flujo de control: Esta perspectiva está centrada en el orden de las actividades o eventos que componen los procesos, así como los distintos caminos posibles. Esta perspectiva es considerada por varios autores y es también llamada perspectiva de procesos.
- Organizacional: En este caso el foco de estudio está en conocer los diferentes actores que participan en el proceso y cómo se relacionan.
- Casos: Similar a la perspectiva de flujo de control, esta perspectiva también se enfoca en los caminos generados por los eventos, pero se diferencia en que se enfoca en descubrir qué genera un caso y qué se necesita para llevarlo a cabo. En este sentido los casos corresponden a subprocesos.
- Temporal: Esta perspectiva se centra en el tiempo ocupado en los eventos que componen al proceso, así como la frecuencia de éstos.

Una de las clasificaciones más usada en la minería de procesos es la perspectiva de proceso y suele ser expresada en términos de algún modelo de procesos, por ejemplo, una red de Petri.

Hoy en día, el descubrimiento de procesos es considerado toda un área de estudio, en las cuales existen distintos estudios de cómo realizar su objetivo y cómo lidiar con distorsiones de los datos [1], [2], [3], [10].

3. ESPECIFICACIÓN DEL PROBLEMA

3.1 Problema a abordar y relevancia

Al desarrollar un software se deben considerar varios factores: tecnologías disponibles, tiempos de desarrollo, conocimientos del equipo, etc. Dentro de estos también encontramos las dependencias de la aplicación con los datos a utilizar. Específicamente, en el caso de software de descubrimiento de procesos, es necesario contar con datos sobre procesos ejecutados. Estos datos componen uno o varios logs de procesos, y deben contener información necesaria para poder hacer el análisis correspondiente y así, de forma sistemática, ir evolucionando nuestros procesos.

El solo recolectar la información ya requiere de un trabajo importante. Es posible para esto automatizarlo utilizando distintas herramientas computacionales ligadas a las empleadas para realizar las múltiples tareas realizadas al llevar a cabo los distintos procesos. Esto ya plantea un gran desafío y a su vez surgen varias preguntas: ¿En qué formato se registran los datos? ¿Cuántos tipos de formatos existirán? ¿Qué define un formato? Muchas veces los desarrolladores y analistas que utilizarán estos datos para realizar un software de descubrimiento de procesos no son los que definen esto, en la mayoría de los casos simplemente les darán los datos como la empresa ya los tenga. Esto condiciona y limita el desarrollo, pero además impacta de mayor forma si a lo largo de proyectos existentes el formato varía a lo largo del tiempo.

Por otra parte si una empresa adquiere un programa que realice el descubrimiento de proceso, se tiene que adecuar al formato establecido por el software y modificar cada uno de sus logs para poder adecuarse. Imaginemos que además adquieren otros programas para hacer otros tipos de análisis y cada uno necesita un formato específico. Simplemente el tiempo y trabajo solo para poder procesar la información y entregarla en los formatos necesarios es considerable.

Por ejemplo pongamos el siguiente escenario: se quiere poder conocer los distintos comportamientos de los usuarios en un sitio web para así poder optimizar ciertas secciones y asignar de mejor forma las prioridades a la hora de desarrollar nuevas funcionalidades y mejorar las ya existentes. Esto puede ser modelado a través de un proceso ya sea partiendo por loguearse en el sitio, ir a la página de inicio, etc. Para poder obtener los datos necesarios la plataforma ha sido modificada para guardar en su base de datos ciertas acciones de los usuarios. Esto correspondería a un primer log de procesos en un formato sobre bases de datos. Además se han organizado encuestas y entrevistas personalmente a la gente y se han ido recopilando la

información en distintas planillas Excel. Este sería un segundo formato, esta vez sobre archivos Excel. Los encargados del estudio quieren utilizar un software para poder hacer descubrimiento de procesos de las dos fuentes independientemente y reuniendo la información y comparar los distintos procesos descubiertos, como se muestra en la figura 9. Realizar esta tarea, asumiendo que el programa que realiza el descubrimiento de proceso acepte al menos uno de los formatos, ya es considerable y más aún si no acepta ninguno de ellos.

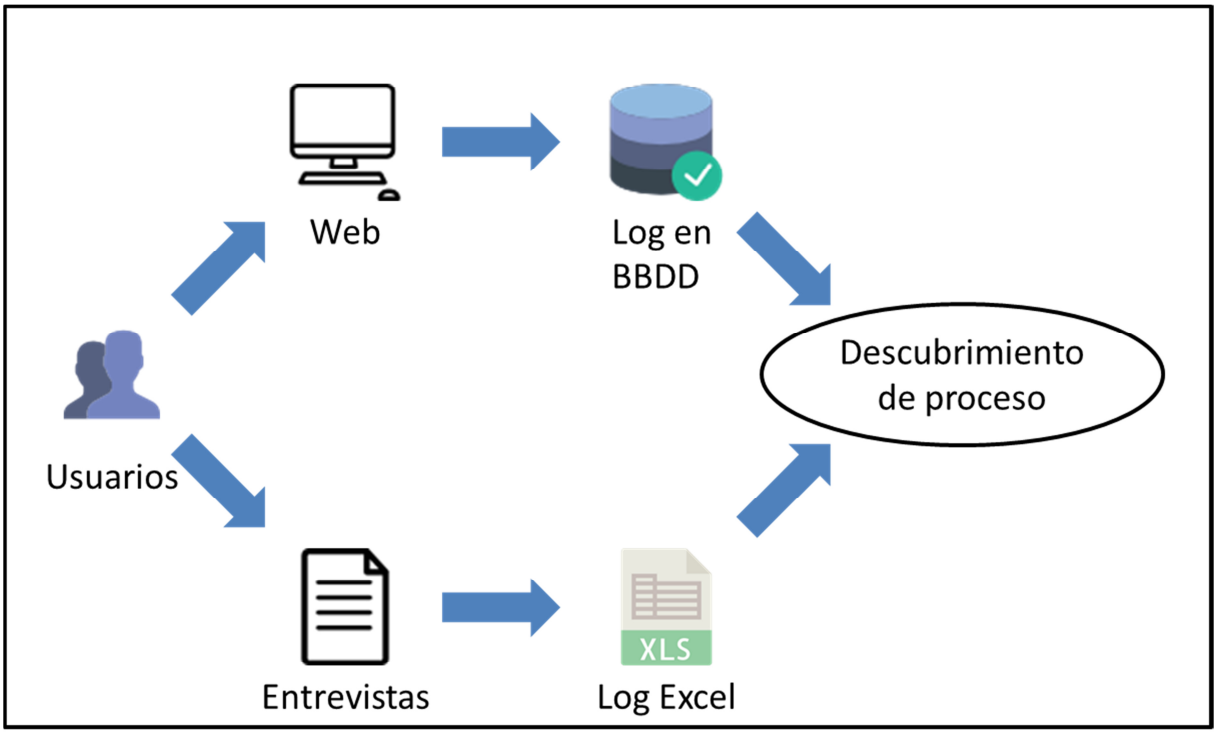


Figura 9: Caso ejemplo

Sabemos que hoy en día muchos software utilizan como formato de entrada, el estándar XES para poder realizar los distintos análisis sobre este tipo de logs. Dentro de estos programas encontramos, además de los ya analizados en la sección 1.3, BPMN Miner, ProConformance, ProDelta.

Hacer este tipo de análisis de forma manual es imposible a medida que el volumen de datos aumenta, lo que eventualmente se desea para así poder tener un estudio mucho más preciso. En este mismo sentido, realizar la transformación manual de formatos hacia XES también se vuelve inmanejable.

De escenarios como éste nace la necesidad de poder transformar formatos distintos al formato XES. Más aún, los formatos pueden variar en cómo presentan la información. Por ejemplo, para hacer el descubrimiento y optimización de procesos, la aplicación necesita saber el tiempo de inicio y de fin de una tarea. Es posible que, tomando el escenario planteado

anteriormente, la automatización del sitio guarde esos datos en 2 fechas pero las entrevistas guardadas en formato Excel quizás sólo tengan la hora de inicio y la duración de la actividad. Incluso pueden existir programas tan específicos que sólo acepten los datos con identificadores específicos que difieren a los usados (nombre de columnas, etc.).

Todas estas características hacen indispensable una herramienta capaz de resolver o simplificar estos problemas tan solo para recién poder hacer análisis más específicos de los procesos y lograr que las entidades que lo requieran puedan realizar mejoras a la hora de entregar sus soluciones, servicios, productos, etc.

De forma muy específica para empresas que tengan como servicio proveer de análisis tan complejos como el descubrimiento de procesos, esta herramienta se hace fundamental.

3.2 Detalle de requisitos

3.2.1 Formatos de origen

La característica principal de la herramienta a construir es poder obtener datos de distintas fuentes. Muchos logs de procesos están guardados en base de datos o archivos Excel por lo que el software debe ser capaz de aceptar estos formatos.

3.2.1.1 Bases de datos

De forma específica, las bases de datos pueden ser de distintos tipos aunque en su mayoría son del tipo relacional. Los datos pueden no estar sólo en una tabla sino que pueden estar distribuidos de distintas formas; por ejemplo podríamos considerar el siguiente diagrama como se muestra en la figura 10.

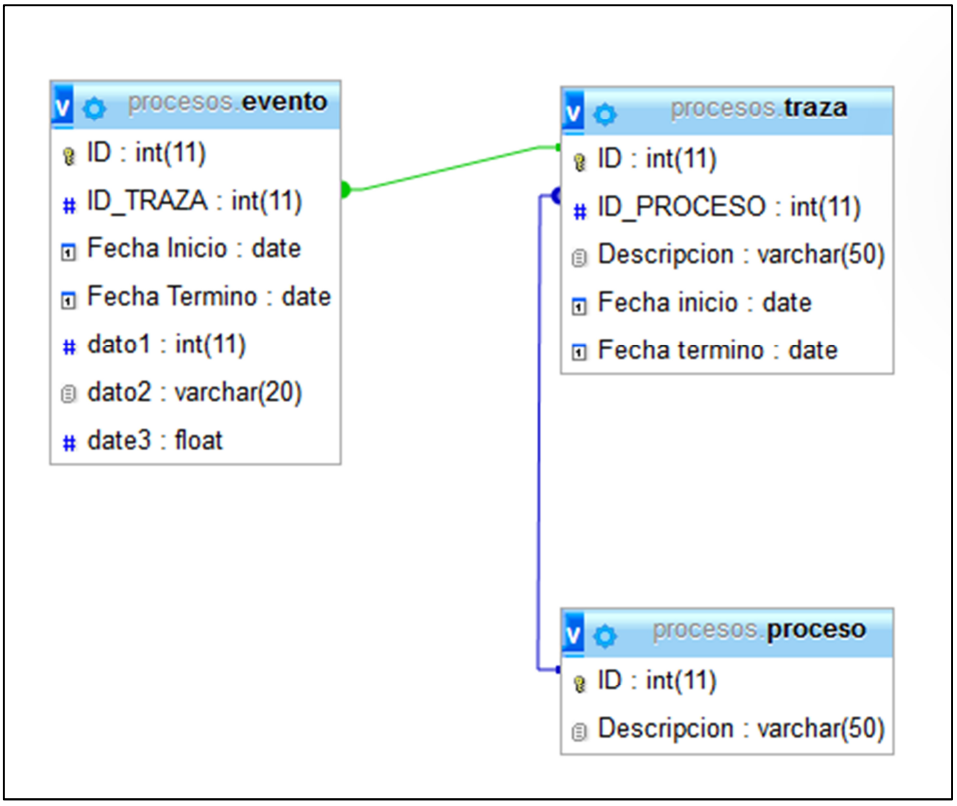


Figura 10: Diagrama BD ejemplo

En cuanto a la selección de los datos también se deberían poder elegir de forma libre. Siguiendo el mismo ejemplo es posible que se elijan sólo trazas de un proceso determinado o también sólo eventos de una traza específica.

3.2.1.2 Planillas Excel

En cuanto a los archivos Excel que deben ser trabajados, se pueden encontrar varias características importantes. La primera es que el software debe poder adaptarse a funcionalidades comunes de Excel, como lo son las funciones o el agrupamiento de datos. En cuanto a las funciones, el programa a desarrollar debe poder trabajar sin diferencia entre un valor estático o un valor calculado a través de alguna función. Si existe un valor agrupado el valor debe ser considerado para todos los eventos o trazas en los que esté presente como se ve en la figura 11, en la columna "ID".

9	ID	Fecha	Fase	Artefacto/Actividad	Versión del Artefacto revisado	Estado	Comentario de la Revisión
10		09-12-2013	Inicio	Carta Gantt de Iteración	1.0	Completo	sólo se encuentra la gantt de cta cte
11		09-12-2013	Inicio	Minuta de reunión equipo	1.0	Completo	Registro 02/12/2013
12		09-12-2013	Inicio	Administración de requerimientos: UR y SR	1.0	Completo	CTA CTE v1.24.0.0
13		09-12-2013	Inicio	Especificación de ambientes	1.0	Completo	CTA CTE: Act 19/12/2013
14		09-12-2013	Inicio	Glosario	1.0	Completo	Actualizado 27/12/2013
15	1	09-12-2013	Inicio	Lista de riesgos	2.0	Completo	
16		09-12-2013	Inicio	Minuta de reunión cliente	1.0	Completo	Actualizado 27/12/2013
17		09-12-2013	Inicio	Plan de proyecto	1.0	Completo	completo
18		09-12-2013	Inicio	Planilla TEC	2.0	Completo	CTA CTE v1.24.0.0
19		09-12-2013	Inicio	Resumen de métricas	1.0		

Figura 11: Datos agrupados Excel

Otras características de estos archivos es que los datos pueden estar en distintas páginas; por lo tanto, al igual que en Excel, debe poder elegirse la página en la cual trabajar.

Además de esto Excel nos permite ocultar ciertas columnas y filas. Estas, aunque no se ven, siguen estando presentes por lo que no pueden ser ignoradas.

3.2.2 Detalles de la transformación

Independientemente de los formatos con los que trabajar, existen ciertos requisitos generales que la aplicación debe considerar.

Para que este programa pueda ser usado de mejor forma, es necesario que posea una interfaz gráfica. Esta debe ser, dentro de lo posible, lo más usable posible, es decir, que alguien sin experiencia pueda ocuparla sin mayores dificultades, que sea intuitiva, que se entiendan los conceptos usados, etc.

Aunque esta herramienta simplifique las tareas de transformación de datos, es necesario configurar estas conversiones y no tener que hacerlo cada vez que se use la aplicación. Se espera que una misma configuración que represente al formato de entrada pueda ser utilizada varias veces y por lo tanto debe poder recuperarse esta y usarla de forma más rápida en ocasiones posteriores. Este caso se dará con frecuencia cuando existan empresas que recuperen y analicen un proceso a lo largo del tiempo como parte de un programa de mejora continua de procesos.

Ya que este software tiene como fin utilizar programas y herramientas ya existentes en más formatos, es consistente que la aplicación a desarrollar pueda ser usada en diversos escenarios. Considerando lo anteriormente expuesto, es que una de las condiciones de validación de este desarrollo es poder utilizar los logs convertidos desde los formatos de origen propuestos, en herramientas de análisis de procesos que utilicen el formato XES.

3.2.3 Resumen de requisitos

- Aceptar formatos desde base de datos.
- Poder trabajar sobre más de una tabla de la base de datos.
- Poder elegir los datos a usar desde la base de datos.

- Aceptar formatos desde planillas Excel.
- Poder convertir desde celdas con funciones.
- Poder utilizar celdas agrupadas.
- Considerar columnas y filas ocultas.
- Poder elegir página de trabajo.

- Contar con una interfaz gráfica.

- Reutilizar configuraciones de conversión.
- Trabajar en distintos escenarios

4. DESCRIPCIÓN DE LA SOLUCIÓN

A partir del problema planteado y los distintos requisitos detallados en la sección anterior es que se ha determinado desarrollar una aplicación ETL. Este transformador responder a escenarios similares a los planteados en la sección 3.1 cómo se ve en la figura 12.

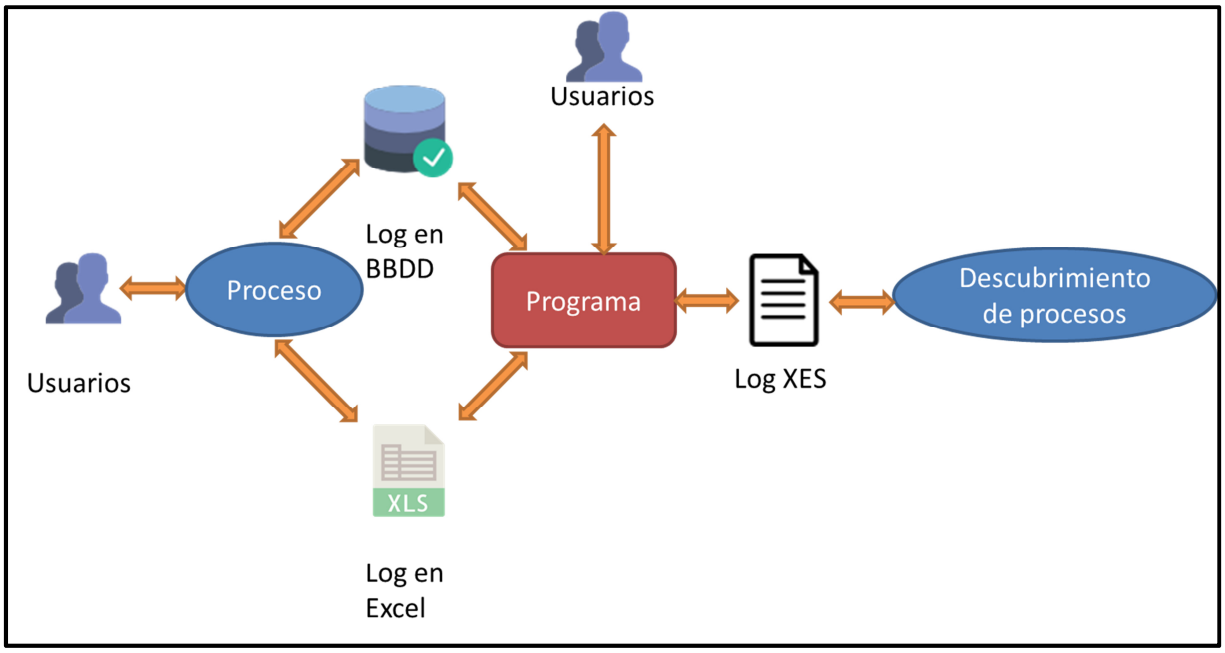


Figura 12: Programa aplicado en ejemplo

4.1 Alcance determinado

Se ha determinado distintos alcances que abarcará el proyecto según las distintas características analizadas. Aunque esto dejará ciertos escenarios fuera, se cree que el espectro el cual se abarcará lo compensa y quedan propuestos como trabajo a futuro varios de los demás casos.

4.1.1 Formatos de origen

Se ha considerado utilizar formatos de origen basados en planillas Excel y bases de datos

4.1.1.1 Bases de datos

Dentro de las bases de datos se consideraron tres tipos: MySQL, PostgreSQL y SQLServer.

Estas bases de datos relacionales tienen la característica de trabajar con datos distribuidos a través de distintas tablas y cada una de estas puede tener varias columnas. Estas características serán consideradas de forma de no restringir al usuario. Se dará la libertad al usuario para manejar la lógica relacionada a las tablas y columnas con las que quiera trabajar.

Se incluirá también la posibilidad de usar las distintas funciones que proveen este tipo de bases de datos; por ejemplo, sumas, máximo, etc. Esto para poder abarcar la mayor cantidad de escenarios posibles y llegar a definir una mayor cantidad de formatos sobre este tipo de origen.

Dentro de las características analizadas, se considera que un evento dentro de este tipo de log corresponde a una fila de datos de las consultas que se puedan realizar.

Para poder acceder a las bases de datos, la herramienta a desarrollar debe poder manejar el uso de cuentas basadas en usuario y contraseña. Dentro de las bases de datos SQLServer se ha decidido dejar de lado la posibilidad de conexión con credenciales de Windows, lo cual se realiza con más frecuencia en ambientes de este tipo debido a la integración de distintos programas.

4.1.1.2 Planillas Excel

Dentro de los archivos Excel sólo se consideró los archivos xlsx, los cuales corresponden a versiones de Excel 2007 en adelante. Además solo se considera planillas que contengan eventos en disposición horizontal, es decir, a lo largo del eje horizontal están los distintos atributos de un evento y a lo largo del eje vertical los distintos eventos. Además está pensado que cada evento esté caracterizado por solo una fila.

Tampoco se consideraron datos distribuidos entre distintas páginas del archivo, aunque se podrá elegir la página del archivo con la cual se desee trabajar.

Al considerar las planillas Excel, es probable que nos encontremos con una mayor cantidad de usuarios que no tengan conocimientos avanzados de computación, por lo que el software debe ser capaz de trabajar con datos en duro o funciones indiferentemente. De la misma manera se permitirá usar columnas ocultas y datos agrupados.

4.1.2 Características de la transformación

Una primera cualidad que tendrá el programa, es realizar las transformaciones de forma independiente del origen de los datos. Esto para poder reducir las dificultades en cuanto a aprendizaje de la herramienta y otras características de usabilidad.

Esta cualidad permite que futuras extensiones a otros orígenes de datos, tengan un pequeño impacto y simplifica las tareas a la hora de futuros desarrollos.

4.1.2.1 Trazas

Se dará la posibilidad de dividir los eventos en distintas trazas a través de un atributo identificador que el usuario podrá elegir. Si no se elige este atributo se considerará que los datos corresponden a una única traza.

4.1.2.2 Atributos

Dado que no se quiere restringir las características del log XES de destino, se ha determinado que los atributos que contengan los eventos serán elegidos por el usuario. Esto permite, por un lado, elegir un subconjunto de los atributos de origen. Por otro lado se podrá elegir más de una vez un mismo atributo de origen para tenerlo varias veces en el formato de salida, con distinto nombre, para así dar más libertades a los posibles programas en los que se utilizarán.

Las dos mayores características es la libertad del usuario a elegir el tipo de salida de cada atributo así como el nombre de éste.

4.1.3 Archivo de destino

En cuanto al log ya transformado, se podrá elegir la carpeta de destino del archivo así como su nombre.

4.2 Arquitectura del software

Analizaremos la arquitectura propuesta del programa desde el contexto general hasta el detalle interno. En la figura 13 vemos cómo se relaciona con otros sistemas al ser utilizado.

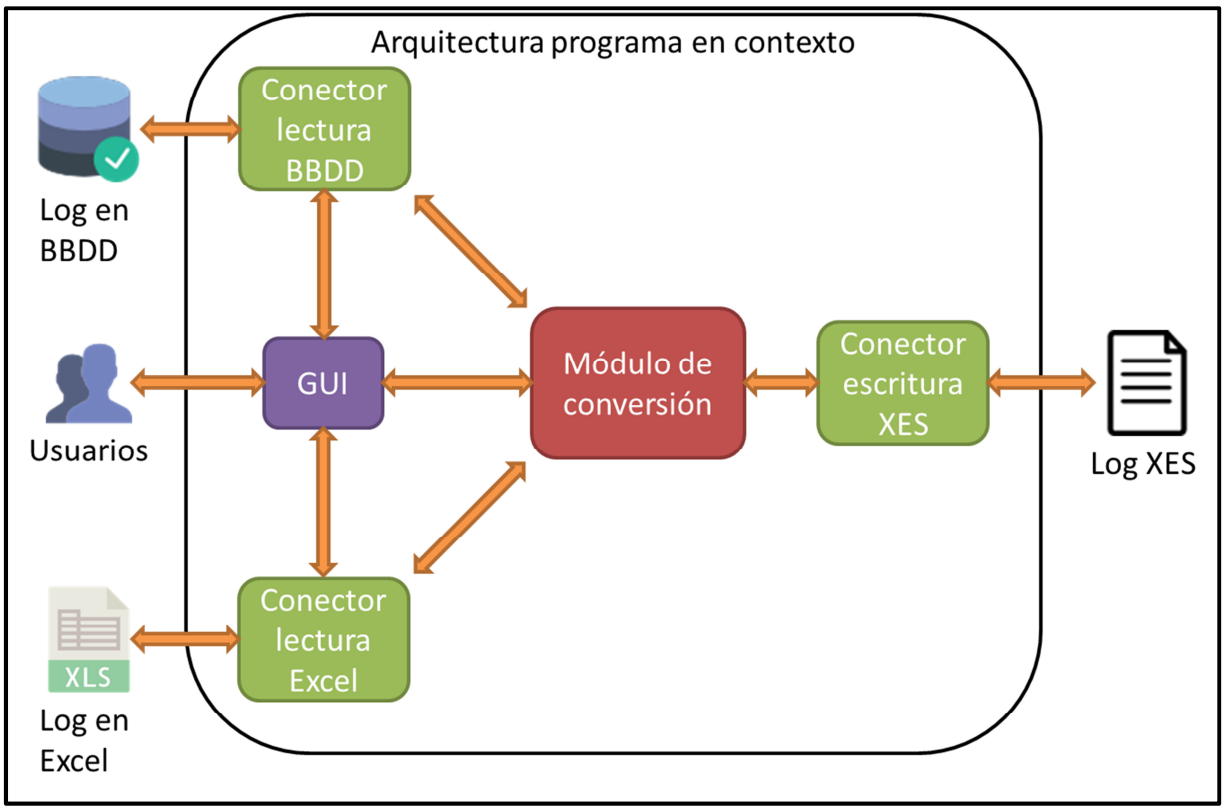


Figura 13: Arquitectura en contexto

El usuario se relaciona con el programa a través de una GUI en la que configura desde dónde quiere obtener los datos a transformar y dónde guardar el archivo XES convertido. En esta GUI también podrá configurar todos los aspectos relacionados con la transformación, eligiendo las columnas que quiere obtener y sus tipos.

Dependiendo del tipo de entrada configurada, es decir, si los datos vienen desde una base de datos o un archivo Excel, el módulo de conversión se encarga de realizar la transformación deseada y finalmente escribe el archivo XES que necesitamos.

Internamente el programa desarrollado define una serie de estructuras de datos con tal de poder describir y convertir los atributos de los eventos y trazas de nuestro proceso, como se pueden ver en color celeste en la figura 14.

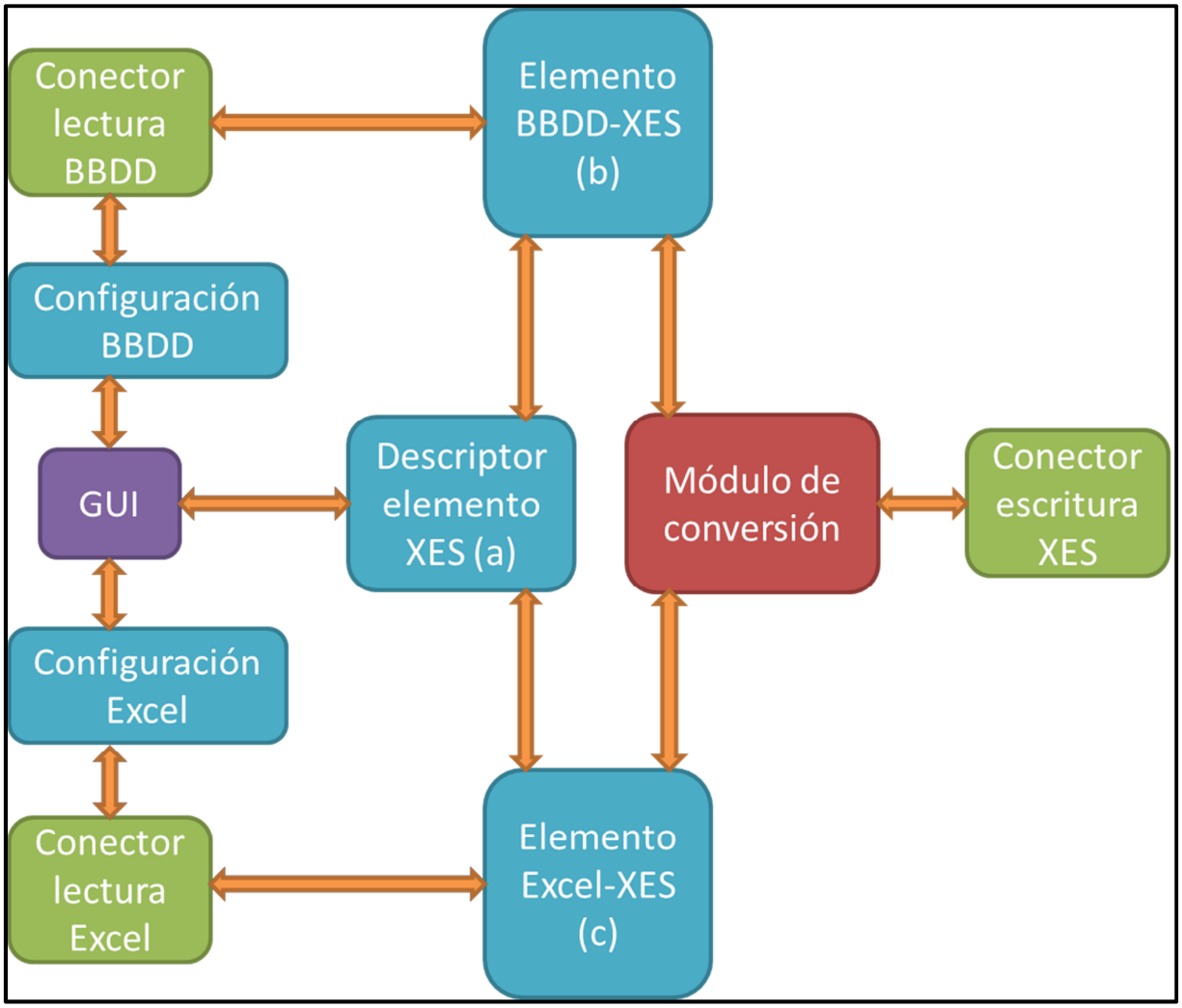


Figura 14: Arquitectura interna

4.3 Diseño de estructuras de datos

Para el desarrollo de la estructura de datos de la aplicación, se decidió trabajar con la unidad mínima de un log XES, esta es, un atributo dentro de un evento, traza o log. Esta estructura se encuentra transversalmente en la aplicación. En la figura 14 podemos encontrar esta estructura en "a", "b" y "c".

Esta estructura nos permite abarcar una gran cantidad de escenarios. Algunos de estos nos abren posibilidades para interesantes trabajos a futuro. En la figura 15 se puede ver la relación de las distintas clases que representan estos tipos de atributos.

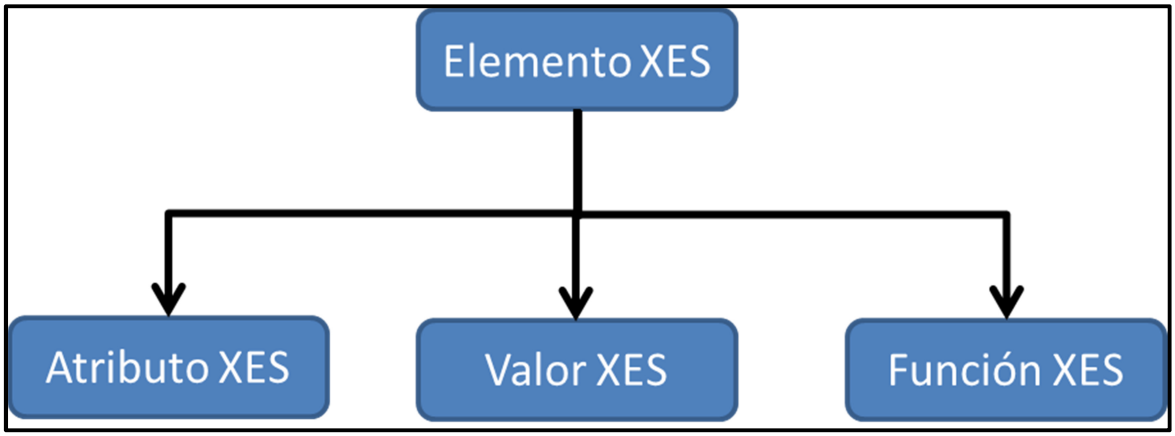


Figura 15: Base de la estructura

Elemento XES nos generaliza lo más básico de la unidad: nombre y tipo.

Las clases que representan a los atributos relacionan los datos desde el formato origen al formato XES. Estas clases son las esenciales para llevar a cabo la transformación.

Las clases que representan a los valores XES son valores libres que no dependen de los datos de origen. Esto fue agregado para poder ser usado por funciones y agregar ciertos valores extras que puedan necesitar los programas de análisis que usen el archivo XES generado por la aplicación. Además se usa en caso de que ciertos datos de origen no existan en algún evento y se quiera usar un valor por defecto.

Las clases que representan funciones son usadas para realizar ciertas conversiones internas entre los formatos. Las implementadas en esta versión sólo están destinadas a poder tener fechas de inicio y término de eventos, y/o duración de estos mismos.

Al implementar esta estructura de datos, lo único que depende directamente de los datos de origen es "Atributo XES" en los elementos "b" y "c" de la figura 14. Esto está pensado para que el desarrollo y extensión de la aplicación tenga un impacto mínimo y pueda hacerse de manera mucho más simple e independiente del resto de las estructuras.

4.4 Diseño Interfaz Gráfica

La interfaz gráfica que permite a los usuarios utilizar el programa fue pensada de forma que fuera simple, funcional y mostrara la esencia del software desarrollado.

La interfaz cuenta sólo con 2 pantallas que resumen los conceptos de "origen de datos" y "transformación de datos". La primera pantalla nos permite configurar desde dónde queremos obtener los datos, ya sea desde una base de datos o un archivo Excel. En la figura 16 se ve las distintas pestañas que nos permiten configurar nuestros datos de origen, eligiendo si queremos usar una planilla Excel o una base de datos.

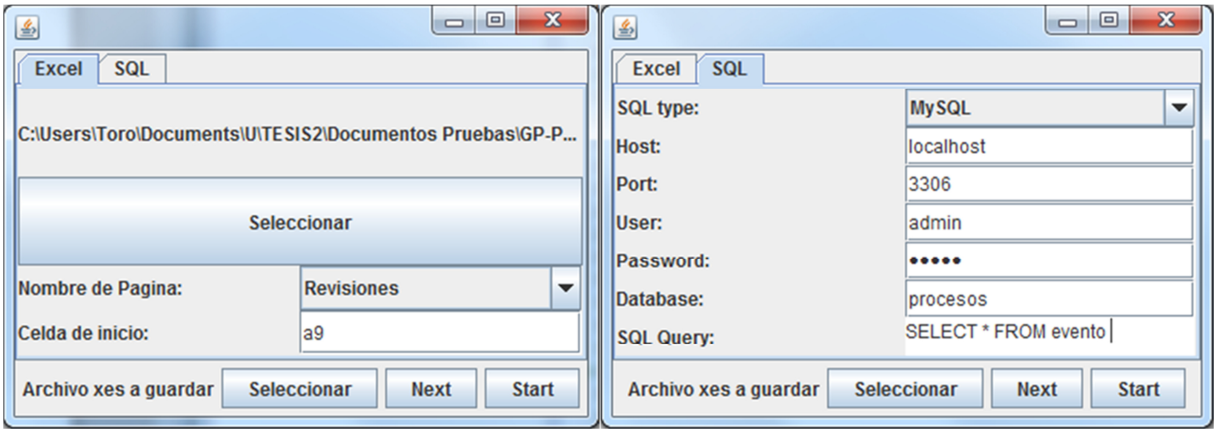


Figura 16: Origen de datos

Con esta pantalla podemos confirmar las libertades que se le otorgan al usuario a la hora de elegir sus datos de origen. Los formatos de Excel sólo requieren el archivo, la página y celda de inicio de los datos. La celda de inicio se considera donde empiezan la tabla con los datos incluyendo el encabezado, es decir, la primera fila solo se considera como los nombres de los atributos de los eventos a continuación.

En relación a la conexión a una base de datos, sólo es necesario, además de los datos de conexión, tener la consulta SQL que se desea realizar. Esto fue pensado para dar la mayor libertad posible a la hora de buscar los datos.

En la figura 17 podemos ver la pantalla de transformación de datos que se diseñó independientemente del origen de los datos.

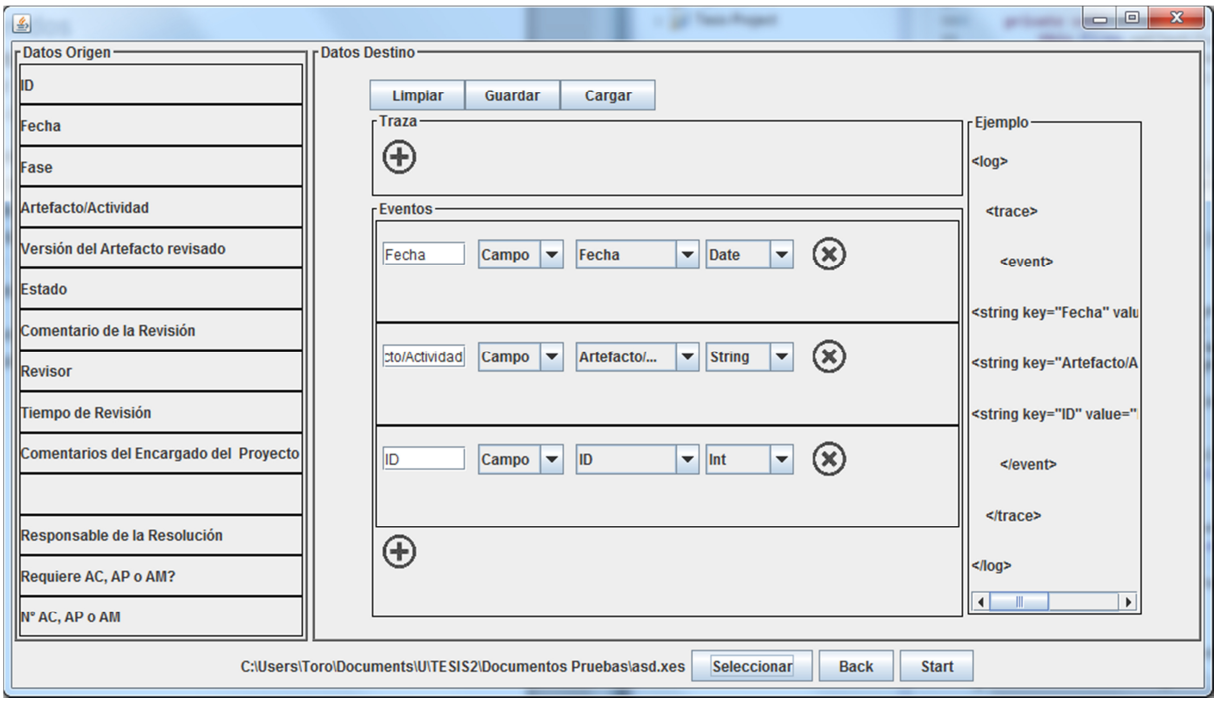


Figura 17: Transformación de datos

Esta pantalla está compuesta de 3 partes principales que nos muestran el flujo a ejecutar. A la izquierda una lista con los encabezados de los datos de origen según el orden en que se encuentran. En este caso podemos ver que el archivo utilizado tiene varias columnas con datos: ID, Fecha, Fase, etc. En el centro encontramos la configuración de la conversión. En esta sección podemos agregar distintos atributos a los eventos que queremos transformar. Además podemos definir si existe algún campo con el cual diferenciar las trazas que encontremos, de no existir este valor se considera una sola traza. En este caso sólo se utilizan 3 columnas de todas las disponibles. Además, no se eligió ninguna columna que identificara a qué traza corresponde por lo que el log de salida contendrá sólo una traza. Y a la derecha un ejemplo de cómo se verían los datos al ser convertidos al formato XES. En esta pantalla además encontramos 3 botones que nos permiten trabajar de forma simple la configuración de la conversión.

En ambas pantallas encontramos un footer donde se configura el archivo de salida, permite movernos entre pantallas y realizar la transformación a través de los botones “seleccionar”, “next/back” y “start” respectivamente.

4.5 Justificación del diseño

El diseño de la arquitectura, estructuras e interfaz gráfica fue desarrollado pensando en las características y requisitos planteados. La cualidad transversal en estos aspectos fue poder separar las distintas funciones de una herramienta ETL. A esto se le agrega el diseño enfocado a

la unidad mínima planteada para poder dar una flexibilidad a la herramienta, haciéndola capaz de adaptarse en distintos escenarios.

A nivel de extracción de datos podemos ver cómo se especifica según su fuente de origen y lo trabaja desde ese aspecto en concreto. De esta forma se logra desarrollar los conectores a las fuentes de origen de forma independiente unas de otras manteniendo la estructura planteada para la siguiente fase.

El nivel de transformación se logra en gran medida a través del enfoque de la unidad mínima haciendo que la conversión real sea simplemente de un dato a la vez.

Finalmente al llegar al nivel de escritura no se tiene referencia del origen de los datos logrando una conversión limpia.

El diseño modular realizado transversalmente a toda la aplicación, permite que esta herramienta sea fácil de extender a otros tipos de origen de datos así cómo agregar nuevas funciones en cuanto a la transformación de datos, con un impacto mínimo en lo ya desarrollado.

Considerando estas características y la ayuda de la interfaz gráfica, donde podemos destacar la libertad de los usuarios al crear sus propias configuraciones de conversión, se logra una herramienta que permite una amplia gama de los formatos de log más usados, logrando así el objetivo de ampliar aplicaciones de descubrimiento y otros análisis de procesos más allá del formato XES.

5. VALIDACIÓN DE LA SOLUCIÓN

Para poder evaluar esta aplicación se utilizaron una serie de datos, muchos de estos reales, de situaciones en las que se plantea su caso de uso. Asimismo se trabajó con un usuario de la aplicación de forma que pudiera la herramienta ser comprendida y usable por alguien ajeno al desarrollo de ésta.

Finalmente, y siendo el parámetro más relevante de validación, fue el uso de los programas que realizarán análisis de procesos a partir de los archivos XES generados.

Para estas pruebas fue utilizado el módulo de análisis incluido en Disco.

5.1 Desarrollo de prototipo

Se desarrolló un prototipo para poder realizar pruebas de conceptos y determinar los requerimientos fundamentales de la aplicación. Este prototipo sirvió para poder determinar el alcance del programa final.

El prototipo inicial ya contaba con algunas de las características más fundamentales del proyecto. Dentro de los formatos de origen encontramos las siguientes funcionalidades:

- Uso de planillas Excel
  - Elección de archivo
  - Elección de página. Se debía ingresar el nombre de ésta.
  - Elección de celda de inicio.
- Uso de base de datos MySQL
  - Uso de consulta sql.

A nivel de transformación, se contaba con las siguientes funcionalidades:

- Poder elegir los atributos de entrada a convertir.
- Poder elegir el tipo de atributo de destino.
- Poder identificar un atributo que represente a qué traza corresponde. Esto es opcional, de no haber elegido se crea una única traza.
- Una imagen de muestra de cómo quedarían los atributos en el archivo convertido.

Dentro de las pruebas realizadas con este prototipo por parte de un usuario final de la aplicación se obtuvieron las siguientes conclusiones:

- Poca claridad en la interfaz gráfica, especialmente a la hora de cambiar los atributos de los eventos.
- Poca funcionalidad real a la hora de usar el log XES de salida en otros programas. Esto se debió a que los formatos XES que se requerían necesitaban que los atributos tuvieran nombres específicos, el prototipo usaba el nombre de la columna elegida.
- Pobre manejo de errores. Existían varios errores en los datos con los que el prototipo fallaba o mostraba mensajes poco entendibles.

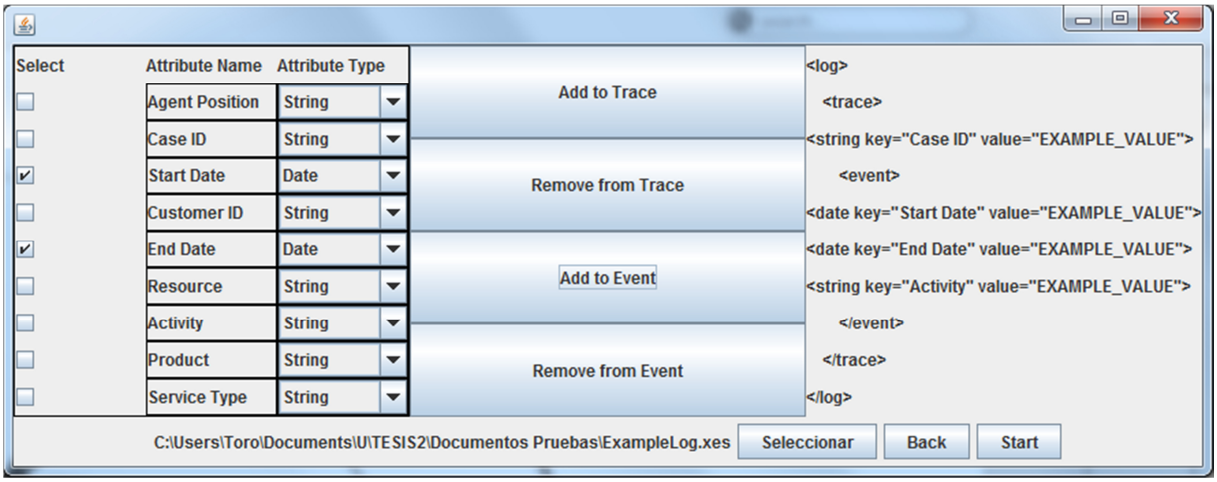


Figura 18: Pantalla de transformación en prototipo

Aunque a este nivel ya se encuentran muchas de las funcionalidades finales, muchas de estas no funcionaban 100% del modo deseado. Aunque sus problemas se fueron corrigiendo hasta la versión actual.

5.2 Validación con Disco

Disco es un software que permite el análisis de datos correspondiente al descubrimiento de procesos desde logs de eventos, como se explicó anteriormente, en la sección 1.3.1. Este programa permite cargar datos desde archivos Excel y también desde archivos XES. Además posee la posibilidad de exportar el logs trabajado en XES, este debe de tener un formato con ciertas especificaciones necesarias y otras opcionales.

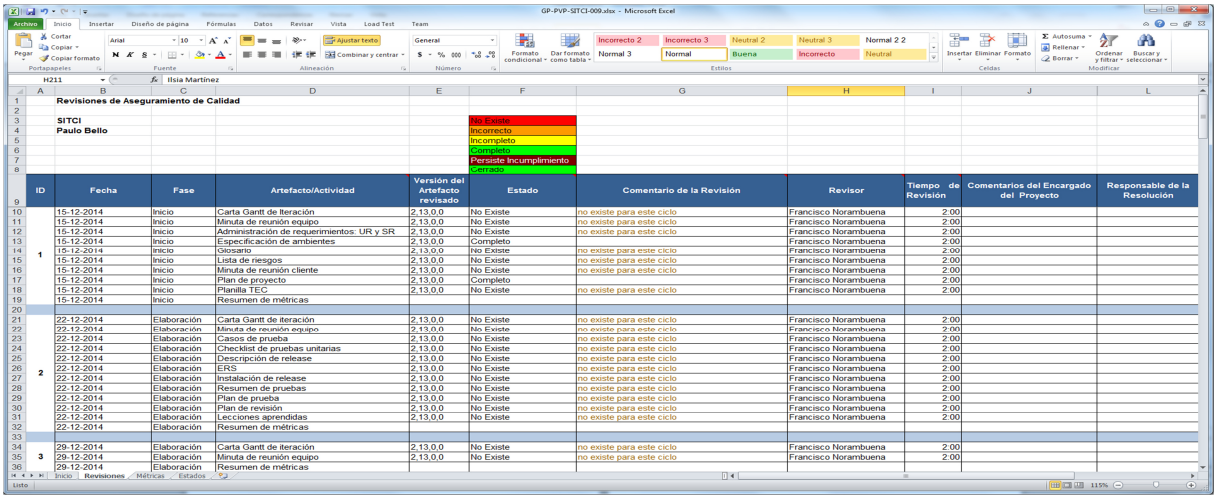
Aunque este software tiene la propiedad de transformar ciertos archivos Excel, tiene varias limitaciones. Entre estas, destaca el hecho que realiza la transformación sólo al formato XES que él utiliza. Estos archivos quedan, por consiguiente, con sólo algunas pocas columnas y con nombres específicos. Además los archivos Excel deben venir con ciertas características no transables: la tabla debe comenzar en la celda A1 de la primera página, no puede haber agrupaciones, etc.

5.2.1Excel

Se realizó el siguiente experimento:

1. Ejecutar el descubrimiento de procesos en Disco directamente desde un archivo Excel.
2. Usar la herramienta desarrollada para transformar el archivo Excel en un archivo XES.
3. A partir de este último, ejecutar el descubrimiento de procesos de Disco.
4. Comparar procesos obtenidos desde 1 y 3.

Para este ejemplo se usó el siguiente archivo que se muestra en la figura 19.



Revisiones de Aseguramiento de Calidad

SITCI

Paulo Bello

ID	Fecha	Fase	Artefacto/Actividad	Versión del Artefacto revisado	Estado	Comentario de la Revisión	Revisor	Tiempo de Revisión	Comentarios del Encargado del Proyecto	Responsable de la Resolución
1	15-12-2014	Inicio	Carta Gantt de Iteración	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	15-12-2014	Inicio	Minuta de reunión equipo	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	15-12-2014	Inicio	Administración de requerimientos: UR y SR	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	15-12-2014	Inicio	Especificación de ambientes	2,13,0,0	Completo		Francisco Norambuena	2:00		
	15-12-2014	Inicio	Glosario	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	15-12-2014	Inicio	Lista de riesgos	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	15-12-2014	Inicio	Minuta de reunión cliente	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	15-12-2014	Inicio	Plan de proyecto	2,13,0,0	Completo		Francisco Norambuena	2:00		
	15-12-2014	Inicio	Planilla TEC	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	15-12-2014	Inicio	Resumen de métricas							
2	22-12-2014	Elaboración	Carta Gantt de iteración	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Minuta de reunión equipo	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Casos de prueba	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Checklist de pruebas unitarias	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Descripción de release	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	ERS	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Instalación de release	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Resumen de pruebas	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Plan de prueba	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Plan de revisión	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Lecciones aprendidas	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	22-12-2014	Elaboración	Resumen de métricas							
3	29-12-2014	Elaboración	Carta Gantt de iteración	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	29-12-2014	Elaboración	Minuta de reunión equipo	2,13,0,0	No Existe	no existe para este ciclo	Francisco Norambuena	2:00		
	29-12-2014	Elaboración	Resumen de métricas							

Figura 19: Archivo para experimento

Para utilizar Disco, los archivos XES deben tener un formato determinado. Lo más importante es que distintos atributos deben tener nombres específicos. En la figura 20 se muestra la configuración para llevar a cabo la transformación.

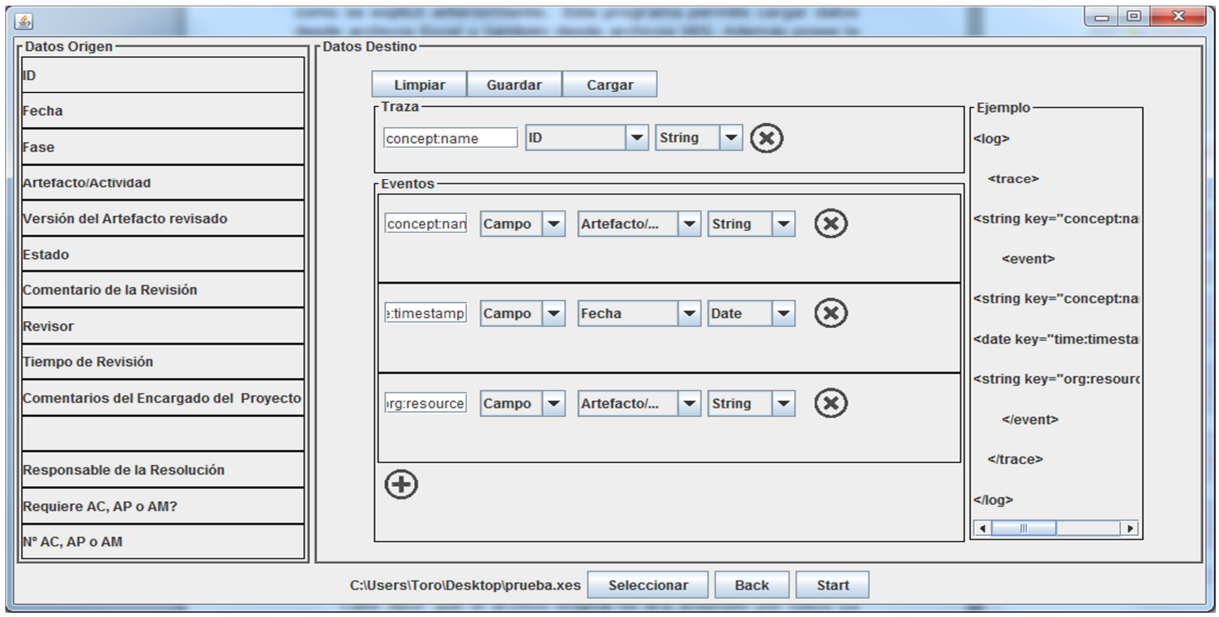


Figura 20: Conversión desde formato

Al comparar ambos archivos XES, lo primero que notamos es que hay varios atributos ausentes en el archivo generado por la aplicación desarrollada. Esto se discutirá más adelante.

Al utilizar ambos archivos, el Excel original y el XES generado por la herramienta hecha para descubrir el proceso en Disco, se obtuvo el mismo proceso que se muestra en la figura 21. Esto demuestra que el programa cumple su funcionalidad en el contexto esencial de uso.

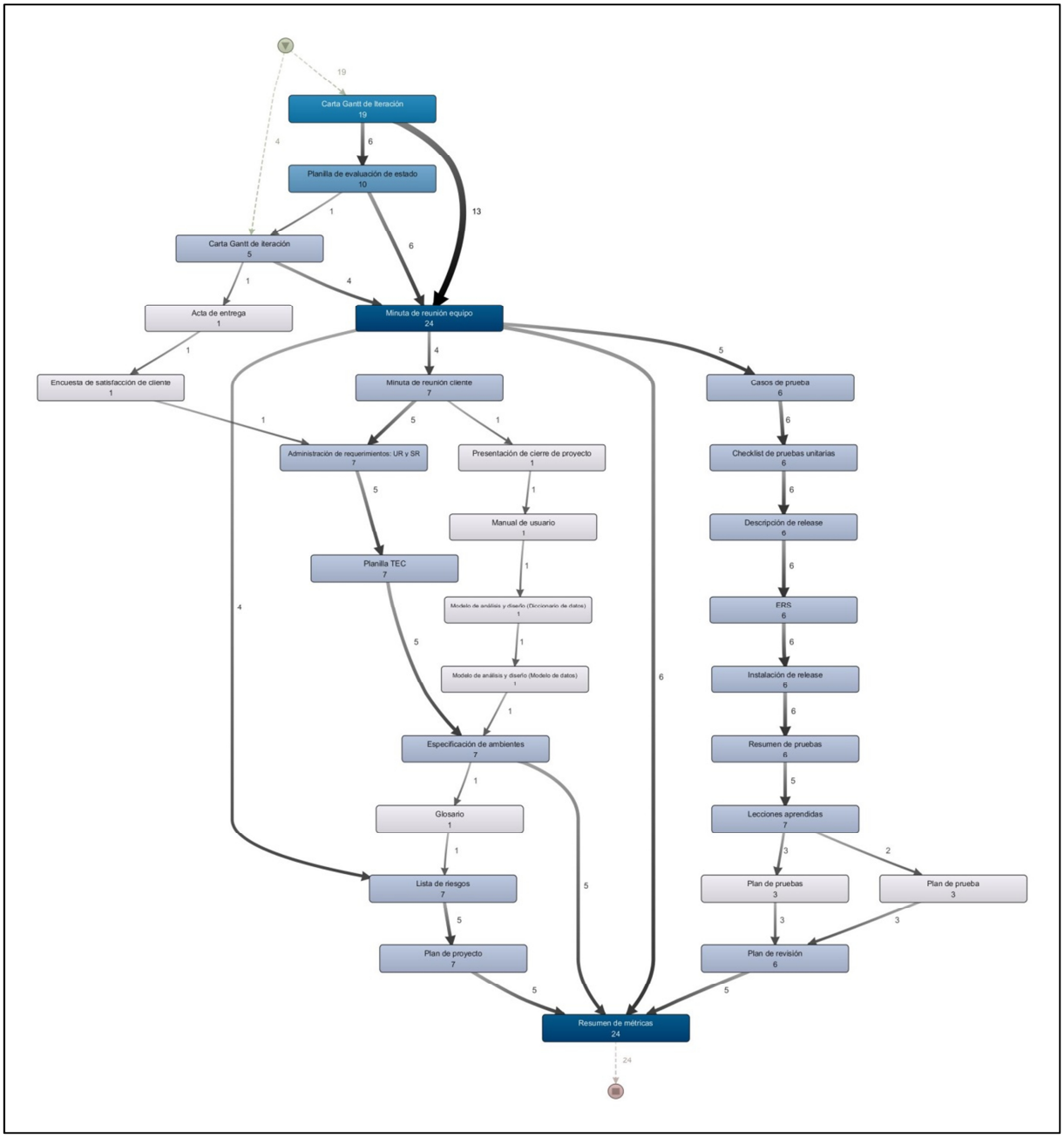


Figura 21: Proceso descubierto

Considerando este exitoso experimento además comparamos los pasos necesarios para lograr este resultado. En el caso del uso del archivo Excel los pasos fueron:

1. Copiar la tabla de datos a la posición “A1” en la primera página.
2. Desagrupar los datos agrupados y colocar los datos correspondientes.
3. Importar el archivo en Disco.
4. Configurar el descubrimiento de procesos.
5. Ejecutar descubrimiento de procesos.

En el caso usando la herramienta desarrollada, los pasos a seguir fueron los siguientes:

1. Configurar la conversión en la herramienta.
2. Realizar la conversión.
3. Importar el archivo generado en Disco.
4. Ejecutar el descubrimiento de procesos.

Al comparar podemos ver que al usar la herramienta desarrollada se ve una pequeña ventaja aunque no determinante, sobre todo si se considera que para poder realizar la transformación se necesitaba conocer el formato mientras que al usar la planilla Excel el programa lo adapto automáticamente. Esta ventaja recae en que el tiempo y el trabajo para ajustar la planilla Excel en el primer caso es mayor que configurar y realizar la transformación utilizando la aplicación desarrollada.

Esto plantea un nuevo escenario. ¿Qué pasaría si se quiere hacer este experimento para 100 archivos? Efectivamente hay varios pasos a repetir en cada una de las iteraciones pero debido a que nuestro conversor tiene la facilidad de guardar la configuración utilizada para la conversión, el paso más costoso en tiempo se reduce significativamente.

5.2.2 Base de datos

Conociendo el formato XES requerido por el programa Disco, se realizaron también pruebas con bases de datos. En este caso se utilizaron datos sintéticos en base a un proceso que se muestra en la figura 22.

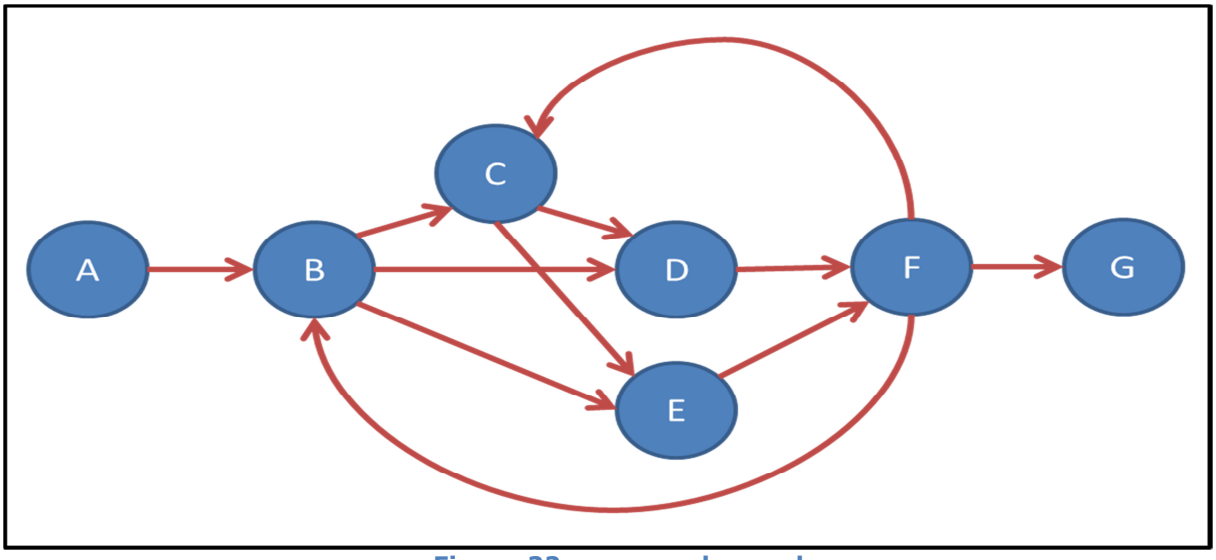


Figura 22: proceso de prueba

Se generaron 611 eventos repartidos entre 30 trazas. Cada evento tenía 3 atributos:

- Traza: un entero que representaba al número de la traza.
- Valor: letra dentro del proceso.
- Fecha: fecha en la que se desarrolló.

Con estos eventos y el formato requerido por Disco se creó la siguiente configuración que se muestra en la figura 23.

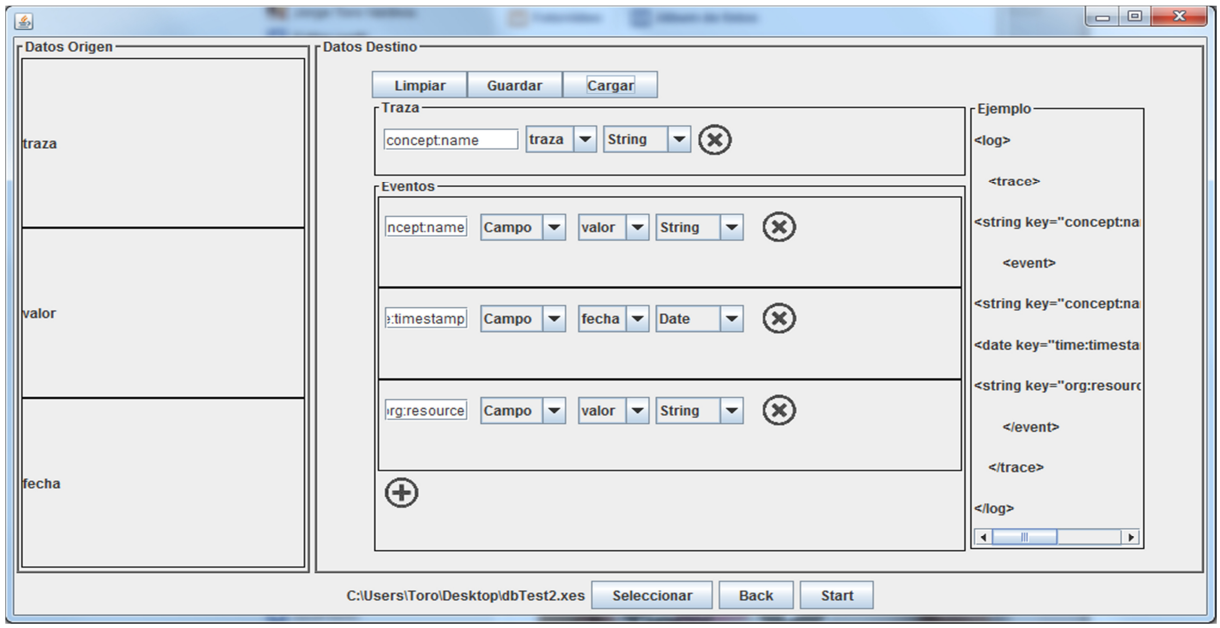


Figura 23: configuración prueba BBDD

Al momento de importar el archivo XES transformado a Disco se obtuvo el siguiente modelo de proceso que se muestra en la figura 24, el cual coincide con el proceso propuesto.

Con esto probamos que es posible también realizar transformaciones desde bases de datos y estas pueden ser utilizadas por programas que realicen descubrimiento de procesos.

Este ejemplo prueba el objetivo principal que fue planteado, ya que pudimos realizar este análisis en un contexto, que sin la herramienta desarrollada, no habría sido posible.

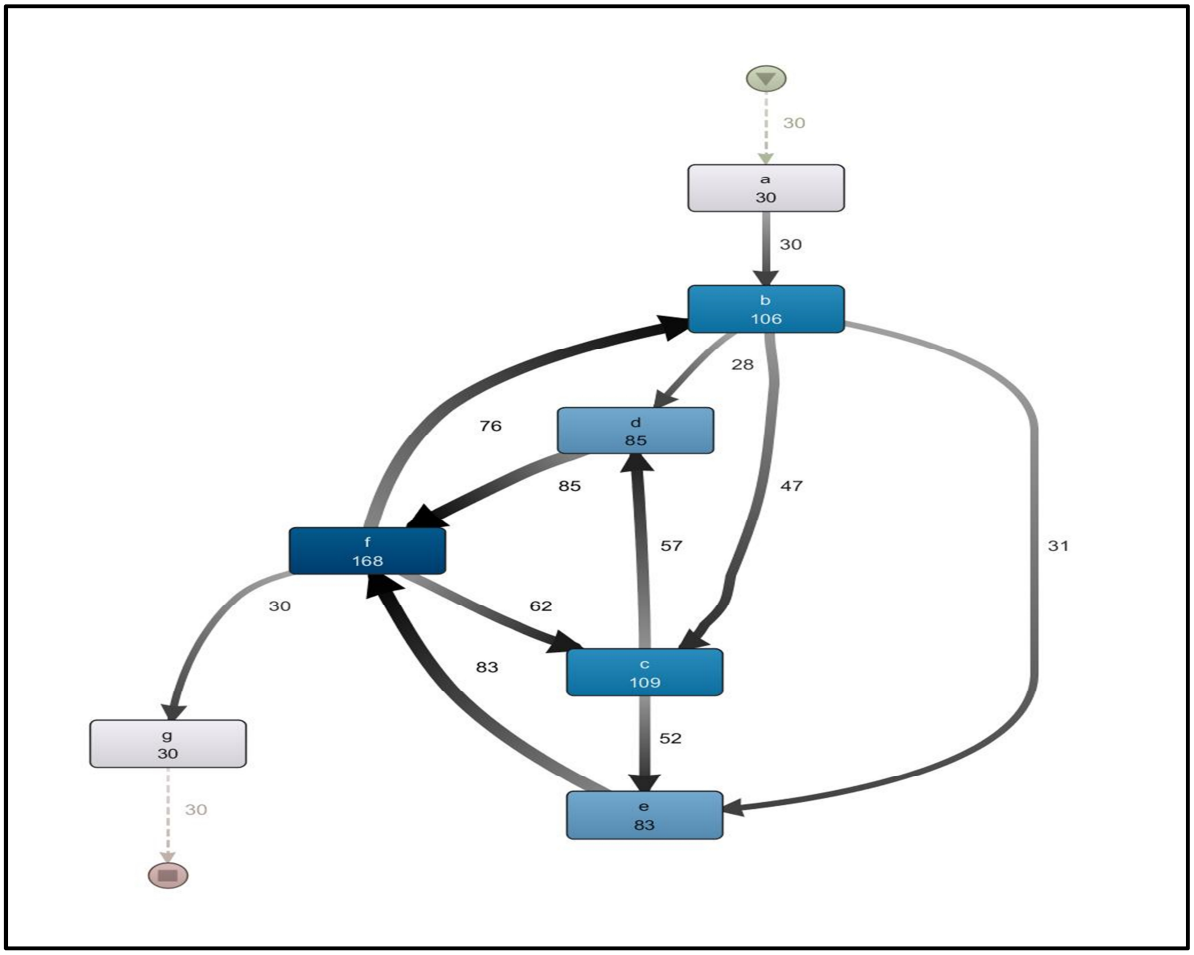


Figura 24: Proceso de prueba recuperado

6. CONCLUSIONES

Este proyecto se llevó a cabo desde la motivación de ser una herramienta de apoyo a la hora de realizar descubrimiento de procesos. Para poder comprender de mejor forma el problema, se trabajó y desarrolló el programa basado en archivos con datos reales y algunos ejemplos planteados de los posibles escenarios a encontrar. El mayor desafío durante el trabajo fue poder caracterizar de la forma más genérica posible lo necesario para definir un formato. A partir de ese trabajo se definieron tres grandes características: tipo de entrada, datos o columnas que contenían los datos, y tipo de dato en esas columnas. El poder entender esto fue fundamental a la hora de determinar cómo obtener los datos, lo que llevó también a ciertas consideraciones, en especial las distintas cualidades de los archivos Excel. Aunque funcionalmente se avanzaba, era necesario crear la interfaz gráfica para hacer que el uso de esta herramienta fuera posible y lo más fácil posible. La interfaz en sí sufrió varios cambios durante el proceso y aún así puede ser mucho mejor. Al hacer las distintas pruebas se logró comprobar que la herramienta cumple su función, contando varias pruebas con distintos niveles de error durante el camino hasta llegar a la última versión de este software.

6.1 Objetivos alcanzados

Recapitulando los objetivos específicos planteados al comienzo de este proyecto, estos fueron cumplidos así pudiendo lograr el objetivo principal planteado que era: "Desarrollar una herramienta computacional para transformar logs de procesos desde distintos formatos definidos por el usuario al formato XES para permitir descubrimiento de procesos". Cada uno de los objetivos específicos se considerara individualmente para poder construir un análisis desde lo particular a lo general.

6.1.1 "Implementar descripción de formatos de origen"

Al determinar el alcance de esta herramienta los formatos de origen fueron restringidos a aquellos que provenían desde ciertas bases de datos o archivos Excel.

Tomando en cuenta las consideraciones a la hora de abarcarlos, se logró tener acceso a todos los formatos que se encontraron a partir de datos reales, por lo que se confirma el cumplimiento de este objetivo.

Este objetivo se logró cumplir, en principal medida, debido a la decisión de abarcar cada atributo por separado. Esto permitió trabajar ciertos atributos más específicos de manera concreta y así en conjunto caracterizar un formato en su totalidad.

6.1.2 "Desarrollar un software de extracción de datos de logs"

Considerando los datos provenientes desde archivos Excel, estos abarcaron varias formas de uso simplificando de gran forma la extracción a nivel de usuario. Esto permite eliminar una barrera técnica importante.

A diferencia de lo comentado anteriormente, los datos provenientes desde bases de datos presentan en un mayor desafío técnico a considerar. Es posible que ciertos usuarios se vean complicados en su uso. A pesar de eso, se considera que la libertad otorgada compensa de buena forma esta debilidad.

6.1.3 "Implementar reglas de filtros y operaciones sobre los datos por el usuario"

La mayoría de los filtros y operaciones fueron abarcados en la ventana que nos permite configurar la transformación de los datos. Esto permite simplificar los archivos de salida y no tener que preocuparse de todos los datos existentes, concentrando la transformación en lo necesario para poder ejecutar un análisis posterior. De todas formas los filtros sólo son a nivel de atributos y sólo en algunos casos menores nos permiten discriminar ciertos eventos.

Las operaciones quedaron reducidas al uso de fechas.

Considerando esto, y si separamos el objetivo entre filtros y operaciones, se podría evaluar que el primer objetivo se cumple aunque en el segundo caso se cubre parcialmente.

6.1.4 "Transformar datos desde un formato de origen al formato de destino: XES"

Aunque el formato interno de los archivos XES generados es bastante básico, las transformaciones y pruebas llevadas a cabo nos confirman que este objetivo específico, que es el centro de este proyecto, se logró de buena manera permitiendo que una gran variedad de formatos puedan ser convertidos a XES. Este objetivo fue el que principalmente nos permite el uso de aplicaciones para el análisis de procesos, en especial el descubrimiento de estos. El hecho que se haya cumplido de esta manera es ratificado con las pruebas realizadas.

6.1.5 "Implementar interfaz gráfica para uso de la aplicación"

Este objetivo fue cubierto y resultó fundamental en el uso de la herramienta, de hecho abarca todas las funcionalidades. A pesar que no era una condición inicialmente, finalmente se convirtió en la única forma de uso. Esto permite que la aplicación pueda ser usada por distintos tipos de usuarios.

Considerando todo lo anteriormente expuesto es que el objetivo general planteado se evalúa como cumplido.

6.2 Impacto realizado

Aunque la motivación fue el descubrimiento de procesos, en el camino se vio otros usos a esta aplicación dando a conocer que esta herramienta tiene una relevancia mayor que la pensada. Dentro de los otros usos encontrados, más allá de otros tipos de análisis de procesos, se destacan el potencial uso para llevar a cabo integraciones entre sistemas y migraciones de datos.

Utilizando esta herramienta sería posible limitar las integraciones, especialmente de bases de datos, a un formato XES específico y centrar el trabajo sólo en las variaciones de datos de origen. Este uso permitiría reducir grandes programas que lleven esto a cabo, así como reemplazar bases de datos intermedias, permitiendo de mejor manera la distribución de trabajo al realizar dichas integraciones.

Asimismo sería posible poder unificar datos de distintas personas o grupos de trabajo y exportar datos desde planillas Excel a bases de datos de sistemas más complejos haciendo que la información se encuentre en un solo lugar y así no tener que imponer modelos que se adapten a la diversidad de formas de trabajo que puedan tener en relación a la información.

6.3 Aprendizaje

Dentro de las lecciones aprendidas, la primera a considerar es la importancia de, en especial, a la hora de comenzar un desarrollo, contar con datos reales y usuarios finales de la aplicación. Este punto obtiene principal énfasis a la hora de evaluar su posible uso, convirtiendo al proyecto en un impacto real y aplicado. Aunque contar con ciertos ejemplos y usuarios concretos puede sesgar el alcance del desarrollo, el tradeoff versus la posibilidad de realizar testing en tempranas etapas del desarrollo, el feedback entregado desde opiniones y casos de prueba específicos, y validación constante durante todo el proyecto, logra concretar un programa de mejor calidad que responde a una problemática concreta.

El segundo aprendizaje es cómo lograr caracterizar conceptos en forma genérica y con la menor pérdida de información. En el caso de este proyecto, definir cómo considerar los formatos fue un desafío más allá de un nivel informático. Buscar el equilibrio entre generalidad y completitud de

información es considerable. En este caso quedaron ciertos tipos de formatos fuera debido al alcance del proyecto. Definir el alcance sólo fue un primer paso, lograr definir los formatos desde unidades mínimas logró una independencia importante en cuanto a los orígenes de los datos. Además, debido a esta toma de decisiones es que se logró un trabajo con distintos módulos muy concretos, repartiendo las responsabilidades de buena forma.

El tercer aprendizaje va relacionado con el diseño de interfaces gráficas. Este punto es bastante complejo de lograr de perfecta forma para la totalidad de los usuarios. Aunque es importante, el desarrollo no se puede concentrar sólo en la usabilidad o estética del programa, al hacer esto se puede dejar de lado la funcionalidad del mismo. Es necesario que una interfaz dé acceso a toda la funcionalidad desarrollada o sino estas pasan a ser trabajo innecesario y con valor mínimo.

6.4 Trabajos futuros

El trabajo ha abierto varios trabajos a futuros que sería deseable poder seguir desarrollando dando a esta herramienta una utilidad mucho mayor logrando que sea un programa mucho más completo y que abarque más situaciones y escenarios.

Los trabajos que se proponen a continuación son los que han sido considerados más relevantes y/o de alcance cercano.

6.4.1 Origen XES

Al avanzar con el trabajo se determinó que puede existir más de un formato por tipo de origen, por ejemplo dos esquemas distintos de bases de datos. En este sentido, ya que se buscaba exportar datos al formato XES, puede ser posible que distintos programas de análisis de logs de procesos utilicen distintos formatos, aunque ambos sobre archivos XES. Un ejemplo se puede ver en el caso de prueba realizado, que existen más datos agregados al log que simplemente los atributos de los eventos.

Ya que se trabajó con el formato XES para realizar la exportación, es consistente agregar este tipo de archivos a los orígenes de los datos.

6.4.2 Agrupación de archivos

Debido a que se logró capturar la configuración de una conversión, debido a que muchos archivos usan el mismo formato, es posible transformar múltiples archivos a la vez logrando pasar de transformaciones 1:1 a transformaciones n:n y también n:1, si es que se considera que más de una entrada sólo son distintas trazas o alguna otra regla por el estilo. Inclusive si se busca el mismo formato de destino, puede ser que se usen distintas configuraciones e incluso tipos de orígenes distintos, por ejemplo,

construir un log donde las primeras trazas vengan desde un archivo Excel y las restantes desde una base de datos.

6.4.3 Extensión del formato XES actual

Un primer punto de este trabajo a futuro es poder agregar distintos datos a los logs XES para así abarcar más formatos y necesidades. Además de lo que ya se maneja, es posible que programas utilicen otros tipos de atributos que no han sido integrados, como las listas. Además XES fue pensado para poder extenderlo fácilmente y agregar más tipos de atributos de los conocidos y que ya están presentes en el estándar.

6.4.4 Automatización de conversiones

En este momento la única forma de realizar las conversiones es a través de su interfaz gráfica. Esto reduce su utilización en casos masivos y espaciados en el tiempo ya que, aunque el tiempo de conversión ya teniendo la configuración requerida es bajo, no permite que se ejecuten las conversiones si no existe un usuario. El poder usar esta aplicación como un servicio automático o como un plugin dentro de otras aplicaciones le daría un impulso determinante.

6.4.5 Interfaz gráfica

Este ámbito nunca tiene un fin real. Las interfaces gráficas suelen ir cambiando para ser más usables o agradables para los usuarios que las ocupan día a día. Esto requiere distintos estudios en casos de usos mucho más concretos y con distintos tipos de usuarios.

6.4.6 Inteligencia y análisis de datos

Es posible desarrollar cierta inteligencia en la aplicación que permita, entre otras cosas, saber el tipo de dato con el que se va a trabajar. Muchas veces no es necesario realizar la conversión completa como para saber que, debido a un dato ingresado por el usuario, la aplicación falle. Por ejemplo un usuario determinó una columna como un número entero mientras que en realidad es una cadena de texto. Asimismo poder sacar ciertos datos antes, como por ejemplo el número de eventos y/o trazas del archivo. Además la posibilidad de tener un editor de fórmulas con tal de no tener que editar la aplicación entera para agregar una nueva, en casos menores.

7. BIBLIOGRAFÍA

[1] J. Cook, A. Wolf: Automating process discovery through event-data analysis

[2] L. Maruster, A. Weijters, W. Van Der Aalst, A. Van Der Bosch: A Rule-Based Approach for Process Discovery: Dealing with Noise and Imbalance in Process Logs

[3] L. Maruster, A. Weijters, W. Van Der Aalst, A. Van Der Bosch: Process mining: Discovering direct successors in process logs

[4] J. Bézivin: In Search of a Basic Principle for Model Driven Engineering

[5] J. Favre: Towards a Basic Theory to Model Model Driven Engineering

[6] http://fluxicon.com/disco/

[7] http://www.promtools.org/doku.php?id=prom651

[8] C. Günther, E. Verbeek: XES Standard Definition 2.0.

[9] P. Vassiliadis: A survey of extract-transform-load technology.

[10] W.M.P. Van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011.

[11] A. L. Wolf, D. S. Rosenblum : A Study in Software Process Data Capture and Analysis,

[12] J. Gosling, H. McGilton: The Java Language Enviroment: A White Paper.

[13] J.C.A.M. Buijs. Mapping Data Sources to XES in a Generic Way. Master's thesis,Eindhoven University of Technology, 2010