

**UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS**

*Universidad del Perú,* ***DECANA DE AMÉRICA***

**FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA**

E.A.P. Ingeniería de Software

**CURSO:**

* Datawarehouse

**DOCENTE:**

* Bartra More, Arturo Alejandro

**INTEGRANTES:**

* Armas Yalico, Camilo Armas
* Chavez Arzapalo, Alejandro
* Arotuma Martinez, Victor Miguel
* De la Cruz Padilla, Jorge Alberto
* Guerra Cruzado, Diego Alonso
* Huaypar Sotelo, Shany Sonaly

**2016**

**“Año de la consolidación del Mar de Grau”**

**ÍNDICE**

[INTRODUCCIÓN](#h.ahzr7dwgdtaz)

[ANTECEDENTES](#h.7mt0w8fpg3pw)

[DATA STEP](#h.4job9yoiouyz)

[Diagrama de flujo básico de un típico DATA STEP](#h.twvzcy5kyr0q)

[VENTAJAS DE USAR SAS EN GENERAL](#h.96x5jq58uhwi)

[CONTROL DE VERSIONES, DE CAMBIOS, PROMOCIÓN, BACKUP Y RECUPERACIÓN](#h.2fx1n85op192)

[PROGRAMACIÓN, DEPENDENCIA Y GESTIÓN DE RESTABLECIMIENTO (INCLUYENDO PARALELIZACIÓN)](#h.r201x9c4x57b)

[GESTIÓN DE METADATOS, SEGURIDAD Y ANÁLISIS DE IMPACTO](#h.xrjauvo2rc67)

[VENTAJAS DE USAR SAS CON RELACIÓN A SUS COMPETIDORES](#h.mhs44rrkbu2p)

[CARACTERÍSTICAS PRINCIPALES](#h.gc3zzu5tidkn)

[Diseño y elaboración de perfiles de datos](#h.88suepo2p9hr)

[Fuente de extracción de datos](#h.883jxlcuep5n)

[Transformación y carga](#h.mncwtnk7tq4z)

[Captura de datos modificados](#h.3xv8tae7npk5)

[Estudio de la calidad, auditoría y manejo de excepciones/gestión](#h.ukjjvqrnzf94)

[La integración con el entorno de producción y componentes de procesos de negocio](#h.4k2k6o9mzwnx)

[FUNCIONALIDADES](#h.rh0xlyyhnuq0)

[Metadata Import and Export in SAS Data Integration Studio](#h.pz9uq7r1z7tk)

[SAS Metadata Bridges](#h.h3kqp1rpvokk)

[JOBS](#h.n90a1eafep3i)

[Versions](#h.srb81jtjl2i)

[Impact Analysis and Data Lineage](#h.bdz5pqxles67)

[SAS Lineage](#h.55r54rl2k7ch)

[Working with SPD Server Cluster Tables](#h.bm1ahziepbiz)

[Working with Hadoop](#h.g87og4ufoqxr)

[UTILIZACIÓN](#h.8ul6i28m8owk)

[CASO PRÁCTICO](#h.svtjg8u8uvmv)

[LIMITACIONES Y/O PROBLEMAS](#h.srg5jevik9c)

[CONCLUSIONES](#h.15gu1vi559rr)

[BIBLIOGRAFÍA](#h.qjw6rhgs5qr)

[ENLACES EXTERNOS](#h.8bxcwjmj8abv)

# **INTRODUCCIÓN**

Hasta hace una década, el manejo de información de gran volúmen en la computadora era tedioso por razones técnicas, sea por la lentitud del procesador, la poca capacidad de memoria, el bajo procesamiento de los datos a nivel software, etc.; sin embargo, el pasar del tiempo y, en paralelo, el avance de la tecnología de la información ha fomentado una cantidad de software de manejo de información, es así que comenzamos a ver software que nos permiten tener un mejor manejo de información empresarial para futuras tomas de decisiones, es así que observamos, por ejemplo, el caso de SAS Institute, una empresa americana que apuesta por el business intelligence (inteligencia de negocios).

Se imaginan tener a la mano un software que nos muestre, mediante gráficos, estadísticas, animaciones, data, un conjunto de información acorde a lo solicitado por nosotros (usuarios). Pues esto es real, ya que el SAS Data Integration Studio nos proveerá de una serie de herramientas de diseño visual para la construcción, implementación y gestión de los procesos de integración de datos, todo esto independientemente de las fuentes de datos, aplicaciones o plataformas.

Este maneja múltiples usuarios y es fácil de manejar, permitiendo la colaboración de grandes proyectos empresariales con procesos repetibles que son compartidos fácilmente. La creación y gestión de datos y metadatos se mejoran con un amplio análisis de impacto de los posibles cambios realizados en todos los procesos de integración de datos.

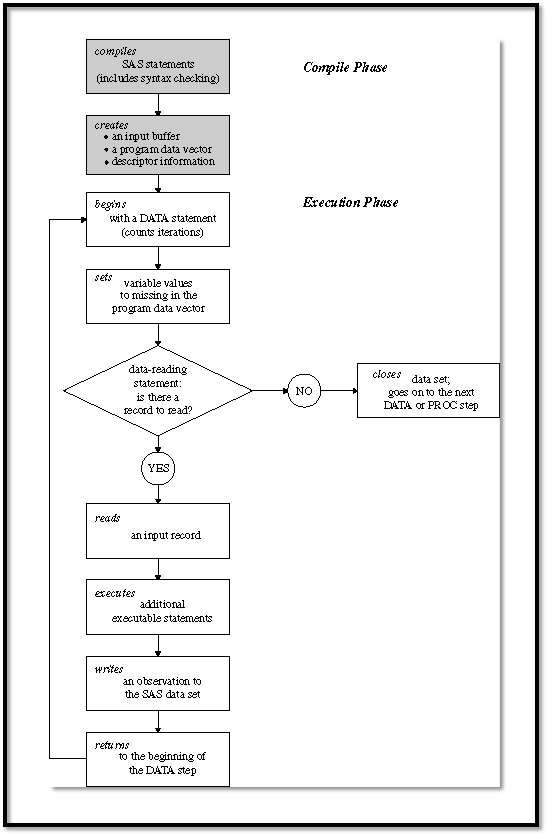
# **ANTECEDENTES**

Primero hablemos acerca de SAS como un conjunto de herramientas. SAS ha existido durante más de 30 años - se inician en la investigación académica como un conjunto de herramientas fundamentales para el análisis estadístico. Uno de los primeros inventos en SAS fue el DATA STEP.

## **DATA STEP**

la respuesta del software de SAS a cómo procesamos los registros de datos, sigue siendo un componente central y base del sistema y productos SAS.

El paso de datos consiste en un grupo de instrucciones SAS que comienza con una instrucción de datos. La instrucción de datos se inicia el proceso de construcción de un conjunto de datos SAS y nombres del conjunto de datos. Los estados que componen el paso de datos se compilan y se comprueba la sintaxis. Si la sintaxis es correcta, entonces se ejecutan las instrucciones. En su forma más simple, el DATA STEP es un bucle con una acción de salida y retorno automático. La siguiente figura ilustra el flujo de la acción en una etapa típica DATOS.



## **Diagrama de flujo básico de un típico DATA STEP**

Con el tiempo, SAS reconocido que con el fin de saciar la necesidad de obtener resultados estadísticamente derivados, se dio cuenta no solo debía tratar o transformar los datos, sino también tenía que presentar los datos de maneras significativas y relevantes.

Como SAS evolucionó a partir de un conjunto de procs (procedimientos) y el paso de datos a una plataforma completa para el soporte de decisiones también las herramientas evolucionaron a partir de un 4GL (lenguaje de 4ª generación) a algo más parecido a lo que los usuarios necesitan para construir almacenes de datos complejos, la inteligencia empresarial y aplicaciones analíticas anticipadas. Así que cuando pensamos en una sola pieza (ETL) dentro de toda esa arquitectura de la información, no olvidemos que **una clara ventaja de SAS es que lo hace responder a la llamada de una plataforma integrada**.

## **VENTAJAS DE USAR SAS EN GENERAL**

### **CONTROL DE VERSIONES, DE CAMBIOS, PROMOCIÓN, BACKUP Y RECUPERACIÓN**

En el caso de Base SAS todo lo que realmente tenemos es un entorno de desarrollo, DMS, Enterprise Guide, etc., sin la esperanza de llevar a cabo cualquiera de las otras tareas puramente en el SAS solo.

Por Data Integration Studio, hemos visto un enorme salto de hacerlo todo usted mismo, a un modesto intento de crear un proceso simplificado para la promoción al pasar de prueba para las pruebas de integración.

La migración de metadatos de plataforma en plataforma es todavía un trabajo en progreso Data Integration Studio pretende ser una herramienta de desarrollo de muchas maneras. Una característica relacionada con el proceso es su apoyo a la gestión del cambio. Esta característica permite que varios desarrolladores trabajen en el mismo proceso de ETL dentro de los confines del contexto de metadatos. Proporciona soporte básico para el comportamiento típico de gestión del cambio: el registro de salida módulos, así como el seguimiento de cuándo, por quién, los comentarios, etc. Estos datos se almacenan como metadatos.

En concreto, no se puede ver o revertir a una versión anterior de un trabajo o una mesa. Por otra parte, cualquier código escrito por el usuario, ya que no es metadatos, no se administra dentro de Data Integration Studio.

Herramientas externas complementarias se deben utilizar para administrar el código, como el de código abierto 'cvs' y el 'ClearCase' comercial. Estas herramientas son sistemas de gestión de control de versiones, desde donde se puede recuperar versiones históricas de archivos.

### **PROGRAMACIÓN, DEPENDENCIA Y GESTIÓN DE RESTABLECIMIENTO (INCLUYENDO PARALELIZACIÓN)**

En el ámbito de la supervisión de los progresos y presentación de informes, a través de un tablero de instrumentos o de otro modo, Data Integration Studio no es tan completo. El producto es capaz de enviar alertas cuando las cosas simples, tales como correos electrónicos simples. Nos gustaría ver una pila-trace incluida en esa función de alertas para que el administrador tenga la oportunidad de saber dónde empezar a buscar - en especial para> 1000 líneas de registro que tienen que analizar.

Relacionada con la programación es la idea de la capacidad de reinicio. Ser capaz de reanudar un trabajo que se ha detenido, o de retirarse de todo un trabajo y reiniciarlo, es clave para el almacenamiento de datos en el mundo real. El sistema de sucesos también permite esta característica que no está disponible en Data Integration Studio por sí mismo.

### **GESTIÓN DE METADATOS, SEGURIDAD Y ANÁLISIS DE IMPACTO**

Con respecto a ser capaz de gestionar la información de los metadatos, SAS ha hecho un gran trabajo en la plataforma SAS®9 para ayudar a localizar el centro de información sobre los datos, la empresa usa de esos datos y los elementos relacionados con el proceso. El servidor de metadatos en SAS®9 es verdaderamente un sistema integral para la captura y el mantenimiento de todos los metadatos ETL, incluyendo toda la lógica de transformación e incluye metadatos proceso y metadatos técnicos. Sin embargo, no se incluye una gran cantidad de metadatos de negocio y nivel de aplicación o porque para éstos, que todavía necesitan tablas de referencia.

Una clara ventaja de Data Integration Studio y el servidor de metadatos es que tenemos una enorme cantidad de metadatos a nuestra disposición acerca de dónde vienen los datos, cómo se transforma, y ​​los procesos que sufrió y cuando todo esto sucedió. Kimball habló de esto cuando él identificó algunas de las capacidades para capturar información linaje y lo que sucedería si tuviéramos que cambiar algo (análisis de impacto). Base SAS no tiene ninguna de estas características y que tienen que ser construido a través de la codificación personalizada.

Por último, en lo que respecta a la seguridad, Base SAS precisa el sistema operativo host para reforzar la seguridad. Data Integration Studio se basa en los metadatos a través de la consola de administración de SAS para hacer cumplir las reglas de autorización a través de usuarios, grupos y roles. Esto permite el control del diseñador de quién tiene acceso y a qué, en toda la cadena de información. La única área débil en todo esto es la característica de auditoría de la SAS® servidor de metadatos. Si bien es posible crear registros de auditoría, no es ni útil para el consumo humano ni está exento de problemas de escalabilidad.

## **VENTAJAS DE USAR SAS CON RELACIÓN A SUS COMPETIDORES**

Un área en la que sobresale SAS está en la programación. En lugar de escribir su propio programador, incluye un planificador de terceros llamado LSF Programador. Este es un sistema para programar y poner en marcha todos los trabajos de ETL y puede manejar algunas dependencias bastante complejas entre estos.

Otra área en la que SAS se ha destacado es en el aprovechamiento de múltiples procesadores y entornos, incluso a base de rejilla. Su estrategia se basa en la plataforma SAS®9 que introdujo la computación grid.

Con MP CONNECT, un proceso ETL se puede ejecutar para que el desarrollador no tiene que solicitar los datos intermedios establece que se escriben en el disco. También no tiene que esperar hasta que otro proceso ha terminado antes de pasar a la siguiente etapa. Con la tecnología de red, estos procesos pueden ocurrir al tomar ventaja de una gran cantidad de recursos informáticos en toda la organización - no sólo los grandes servidores dedicados a estas tareas.

# **CARACTERÍSTICAS PRINCIPALES**

Describiremos las principales características que se pueden realizar en el SAS Data Integration Studio, que incluye: acceso a los datos; integración de datos; gestión de metadatos; limpieza de datos y enriquecimiento; extraer, transformar y cargar (ETL); extraer, cargar y transformar (ELT); y la arquitectura orientada al servicio (SOA) y la integración de cola de mensajes.

Sabemos que normalmente pasamos alrededor del 70% de nuestro tiempo en el almacenamiento de datos al "hacer" ETL. ¿Qué soluciones nos ofrece esta potente herramienta?

## **Diseño y elaboración de perfiles de datos**

En SAS DIS el primer paso crítico es buscar en los datos – generar listados, frecuencias y estadísticas de resumen de los campos en los sistemas de datos de origen. Esta herramienta automatiza el descubrimiento de anomalías en los datos.

## **Fuente de extracción de datos**

Tiene la capacidad de llegar a casi cualquier fuente de datos en el mundo. Desde los sistemas de bases de datos relacionales, aplicaciones propietarias, etc; ya sea que SAS DIS puede acceder a ellos de forma nativa (motores de acceso nativo) o por medio de cualquier método que incluye ODBC / JDBC o estándares abiertos compatibles con SAS Integration Technologies. Como fundación, éstas se apoyan tanto en la Base SAS, así como herramientas como Data Integration Studio.

## **Transformación y carga**

Los que han estado involucrados en ETL durante un tiempo se darán cuenta rápidamente de que la transformación y la carga son verdaderamente lo primordial del almacén de datos. Aquí es donde hacemos la magia que une los sistemas de origen y de destino para crear sistemas de datos útiles para el consumo.

Esta metodología prevé la creación de documentos de diseño en torno a cómo debe producirse esta asignación. De hecho, llamamos a un documento SMT que es parte integral del diseño de un almacén de datos. SMT significa fuente-movimiento-objetivo, y es un documento que describe lo que se va a acceder; incluyendo formatos y valores esperados, como debe ser movido o se transforma, y ​​cuáles son los valores con objetivo similar (de nuevo, los valores esperados, las dimensiones conformadas, etc.). Este documento de diseño es una herramienta esencial para asegurar que todos los involucrados está hablando el mismo idioma.

Esto es especialmente útil en la herramienta usada, ya que, al crear un trabajo en el Process Designer, se trabaja hacia atrás. Saber lo que la tabla de destino debe ser similar es prácticamente un requisito.

## **Captura de datos modificados**

Una de las decisiones fundamentales de diseño al respecto es cómo hacer frente a cambios en el tiempo. Kimball dio referencia a tres de sus subsistemas que pertenecían a cómo debemos gestionar el cambio en nuestra estrategia de ETL y que SAS DIS ofrece:

* Cambio en el sistema de captura de datos. Archivos de registro, la fecha y la fuente, filtros de número de secuencia, y la comparación de registro basados ​​en la CRC en el sistema ETL
* Lógica de transformación para el manejo de tres tipos de variación de tiempo posible para un atributo de dimensión: Tipo 1 (sobrescribir), Tipo 2 (crear nuevo registro), y el Tipo 3 (crear nuevo campo).
* La inserción y actualización de la lógica cambios de dimensión que se han retrasado en llegar al almacén de datos.

## **Estudio de la calidad, auditoría y manejo de excepciones/gestión**

En términos de desarrollo de software, este ítem podría ser llamado simplemente como “programación defensiva” y es lo que le da la característica al proceso de ETL como autorregulado:

* Sistema de limpieza de datos: normalmente, un diccionario del sistema impulsado por el análisis completo de los nombres y direcciones de personas y organizaciones, posiblemente también productos o ubicaciones.

"De-duplicación": incluyendo la identificación y eliminación general de los individuos y organizaciones, posiblemente, productos o lugares. A menudo utiliza la lógica difusa. "Sobrevivir": usando la lógica de combinación de datos especializada que conserva campos especificados de ciertas fuentes que sean las versiones finales guardados. Mantiene la espalda referencias (como las llaves naturales) a todas las fuentes originales participantes.

* Ensamblador en el contexto de metadatos: que rodea a cada carga de tabla de hechos de tal manera que en el contexto los metadatos se puede unir a la tabla de hechos como una dimensión normal.
* Controlador de pantalla de calidad: en línea de pruebas ETL aplican sistemáticamente a todos flujos de datos con la comprobación de los problemas de calidad de datos.
* Controlador de eventos de error: Sistema integral de información y responder a todos los eventos de error de ETL. Incluye una lógica ramificada que manejar varias clases de errores, e incluye la monitorización en tiempo real de la calidad de los datos ETL

## **La integración con el entorno de producción y componentes de procesos de negocio**

Como todos hemos aprendido hace mucho tiempo, que no es todo acerca de la tecnología. Ahí es donde el proceso es tan importante. La integración de productos como Base SAS Data Integration Studio y en la forma de pensar acerca de su estrategia de ETL es crítica.

Mientras pensamos en las implicaciones de cada uno de estos, hay algunas curiosidades de relevancia para nuestras implementaciones:

* El control de versiones, control de cambios, promoción, copia de seguridad y recuperación
* Programación, gestión de la dependencia y de la capacidad de reinicio, incluyendo la paralelización
* Gestión de metadatos y análisis de impacto
* Gestión de proyectos y de ampliación del problema

Finalmente, si bien todas estas características son parte integral de cómo las herramientas como SAS Data Integration Studio opera en los entornos de TI de las organizaciones, es interesante observar cuáles de ellas son puramente proceso y que puede ser ayudada por la tecnología.

# **FUNCIONALIDADES**

## **Metadata Import and Export in SAS Data Integration Studio**

SAS Data Integration Studio le permite importar y exportar metadatos para el individuo objetos o grupos relacionados. Se puede trabajar con dos tipos de metadatos:

* Metadata SAS en forma de paquete SAS
* Metadatos relacionales (metadatos para bibliotecas, tablas, columnas, índices y claves) en formatos que se puede acceder con un SAS Metadata Bridge.

Mediante la importación y exportación de metadatos del paquete SAS, puede mover los metadatos de los Jobs SAS Data Integration Studio y objetos relacionados entre servidores de metadatos de SAS. Por ejemplo, puede crear un job en un entorno de prueba, exportarlo como un paquete SAS, y la importación en otra instancia de SAS Data Integration Studio en un entorno de producción.

Mediante la importación y exportación de metadatos relacional en formatos externos, puede volver a utilizar los metadatos de las aplicaciones de terceros, y puede volver a utilizar SAS metadatos en estas aplicaciones también. Por ejemplo, puede utilizar el software de modelado de datos de terceros para especificar un esquema de estrella de un conjunto de tablas. El modelo se puede exportar en formato común Meta modelo de depósito (CWM). A continuación, puede utilizar un puente SAS metadatos para importar ese modelo en SAS Data Integration Studio.

## **SAS Metadata Bridges**

SAS Data Integration Studio puede importar y exportar metadatos relacionales en cualquier formato que sea compatible con un puente SAS metadatos. Mediante la importación y exportación de metadatos relacional en formatos externos, puede volver a utilizar los metadatos de las aplicaciones de terceros, y puede volver a utilizar SAS metadatos en estas aplicaciones también. Por ejemplo, puede utilizar el software de modelado de datos de terceros para especificar un esquema de estrella de un conjunto de tablas. El modelo se puede exportar en formato común Metamodelo de depósito (CWM). A continuación, puede utilizar un puente de metadatos de SAS para importar ese modelo en SAS Data Integration Studio.

El Asistente para exportar metadatos permite exportar metadatos relacionales de SAS Data Integration Studio en un archivo, en cualquier formato que sea compatible con un puente SAS metadatos.

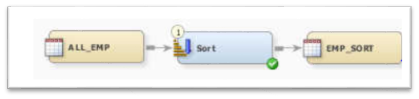
El Asistente de importación de metadatos que permite llevar a cabo las siguientes tareas:

* Importación de metadatos relacionales en un archivo, en cualquier formato que se puede acceder con una SAS
* Puente de metadatos
* Comparación de metadatos importados con los metadatos existentes
* Ver todos los cambios en la ventana Diferencias
* Ejecutar el análisis de impacto o análisis de impacto inverso en las tablas y columnas en la ventana Diferencias, para ayudarle a entender el impacto de un cambio dado en el entorno de destino
* Elegir qué cambia para aplicar al entorno de destino.

## **JOBS**

Un trabajo es un conjunto de tareas que crean SAS de salida. SAS Data Integration Studio utiliza los metadatos para cada puesto de trabajo para generar el código SAS que lee y crea fuentes de objetivos en el almacenamiento físico. Si desea SAS Data Integration Studio para generar código para una tarea, se debe definir un diagrama de flujo de proceso que especifica la secuencia de cada fuente, destino, y el proceso de trabajo en una profesión. En el diagrama, cada fuente, destino y proceso tiene su propio objeto de metadatos. Por ejemplo, el siguiente diagrama de flujo del proceso muestra un trabajo que lee los datos de una tabla de origen, ordena los datos, y luego escribe los datos ordenados en una tabla de destino.

Ejemplo: Diagrama de proceso de flujo de trabajo que ordena los datos



Los componentes de este flujo de proceso realizan las siguientes funciones:

* ALL\_EMP especifica los metadatos de la tabla de origen
* Sort especifica los metadatos para el proceso de clasificación
* EMP\_SORT especifica los metadatos de la tabla de destino

SAS Data Integration Studio utiliza estos metadatos para generar código que lee SAS ALL\_EMP, ordena esta información, y luego lo escribe en la tabla EMP\_SORT. También puede incluir tablas de salida y transformaciones temporales Tabla Loader, flujos de proceso.

Cada proceso en un diagrama de flujo del proceso es especificado por un objeto de metadatos llamado una transformación. En el ejemplo, SAS Sort es una transformación. Una transformación específica cómo extraer los datos, transformar los datos, o datos de carga en los almacenes de datos. Cada transformación que se especifica en un diagrama de flujo de proceso genera o recupera código SAS. Se puede especificar el código escrito por el usuario para cualquier transformación en un diagrama de flujo del proceso.

Para todos los Jobs, excepto los Jobs de sólo lectura que crean cubos, puede especificar el código escrito por el usuario para todo el trabajo o para cualquier transformación en el trabajo.

Hay cuatro (4) formas de ejecutar un job:

1. Enviar la tarea de ejecución inmediata.
2. Implementar el trabajo para la programación.
3. Implementar el trabajo como un proceso almacenado SAS.
4. Implementar un proceso almacenado como un web service.

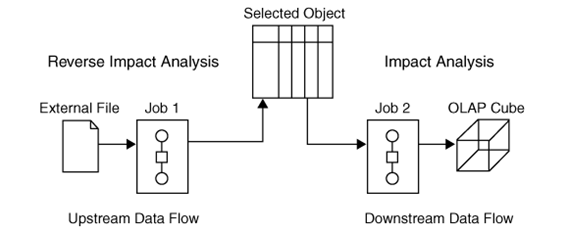
## **Versions**

El control de versiones le permite realizar un seguimiento de los cambios que se están produciendo con el tiempo a los objetos SAS Data Integration Studio. Control de versiones funciona debido al movimiento de contenido tales como jobs y otros objetos dentro de un archivo y archivando el archivo en un sistemas de versiones. SAS Data Integration Studio crea el archivo como un paquete SAS y lo escribe en el sistema de administración de origen. Para llevar el contenido al repositorio, SAS Data Integration Studio recupera el contenido almacenado en el sistema de administración de origen y lo coloca de nuevo en el repositorio de metadatos de SAS. De esta manera, se pueden crear diferentes versiones del contenido y restaurar versiones anteriores de contenido cuando sea necesario. Los objetos pueden ser versionadas de forma independiente o con otros objetos para hacer un paquete de contenido relacionado. Esta capacidad le permite archivar conjuntos de objetos que están relacionados lógicamente, como todo el contenido en un proyecto. También puede optar por generar código fuente 253 para un trabajo y lo almacena junto con el trabajo como el contenido del texto. Esta función hace que sea fácil de ver el código fuente asociado con una versión específica de un trabajo. Puede ver los resultados archivados de cualquier objeto para ver última vez que se versionado. Esta función le permite identificar la versión anterior de los objetos que puede que desee restaurar y mantener un historial de cambios. Después de haber creado versiones de un objeto seleccionado, puede acceder a los objetos versionados en la ventana Paquetes archivado SAS. La ventana muestra una lista de todas las versiones de todos los objetos archivados de manera que se puede acceder y mantener las versiones. Se puede seleccionar un objeto y ver las diferencias entre las versiones del objeto seleccionado o entre una versión archivada y la versión actual de ese objeto.

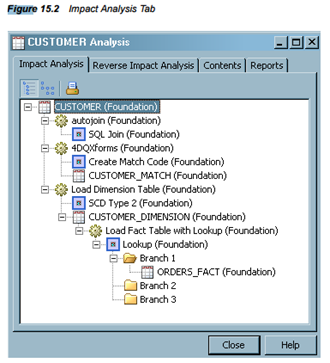
## **Impact Analysis and Data Lineage**

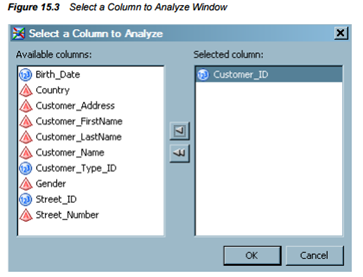
Análisis de impacto identifica las posibles consecuencias de un cambio, cómo agregar una nueva columna a una tabla que se utiliza en un puesto de trabajo. SAS Data Integration Studio le permite realizar análisis de impacto sobre columnas, tablas, archivos externos, mapas de información, informes, procedimientos almacenados, proyectos SAS Enterprise Guide y objetos asociados, y los niveles e indicadores en cubos OLAP. También se pueden generar los análisis de impacto de las transformaciones generadas. El análisis identifica el linaje de datos donde los datos se originan y cómo se utiliza. Se puede utilizar un cliente web, SAS Linaje, para mostrar la información de Lineage de tablas y archivos externos que se utilizan en trabajos Integración de datos SAS Studio.

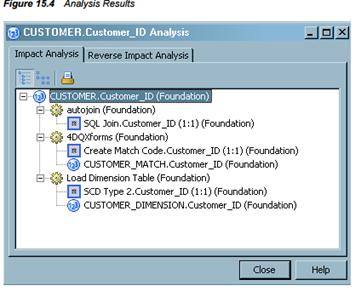
En SAS Data Integration Studio, las funciones de análisis de impacto identifican las tablas, columnas, trabajos y transformaciones que se ven afectados por un cambio a una tabla o columna seleccionada. Las funciones de análisis de impacto inverso identifican las tablas, columnas, trabajos y transformaciones que contribuyen al contenido de una tabla o columna seleccionada. Utilizar el análisis de impacto antes de cambiar o borrar un objeto de metadatos, para ver cómo ese cambio puede afectar a otros objetos. Utilice el análisis de impacto inverso de rastrear los datos fuente que contribuye al contenido de una tabla o columna seleccionada. La siguiente figura muestra la diferencia entre el análisis de impacto y análisis de impacto inverso de un objeto seleccionado.

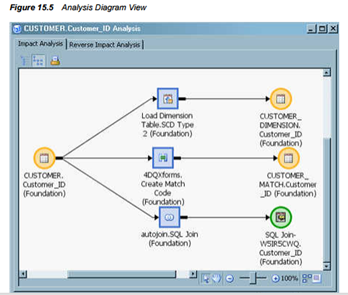


Como se muestra en la figura, el análisis del impacto traza el impacto del objeto seleccionado en objetos posteriores en el flujo de datos. Análisis de impacto inverso traza el impacto que los objetos anteriores en el flujo de datos han tenido en el objeto seleccionado. El análisis se realiza en todos los repositorios de metadatos en el servidor de metadatos actual. El análisis se extiende en cubos. Puede generar impacto y análisis de impacto para la mayoría de tipos de objetos de datos, incluidas las columnas, tablas, archivos externos, mapas de información, informes, procedimientos almacenados, proyectos Guía de Empresa y objetos asociados, así como los niveles y medidas en cubos OLAP inversa. También se pueden generar los análisis de impacto de las transformaciones generadas. Para llevar a cabo un análisis, haga clic en un objeto en el árbol de inventario, árbol de encargo, o Editor de trabajos y seleccione Analizar. Esta acción abre una nueva ventana que contiene hasta cuatro pestañas, que incluyen el análisis del impacto, el análisis del impacto de reversa, el material, e informes. Los resultados analíticos aparecen en las fichas de análisis de impacto o de análisis de impacto inverso. En esas pestañas, puede hacer clic derecho sobre la tabla y seleccione Analizar Columnas para determinar cómo esa mesa de trabajo o impactos o se ve afectada por el objeto seleccionado. Dentro de estas pestañas, también puede mostrar las propiedades o seleccione Abrir para ver los datos en una tabla. También puede seleccionar uno de los iconos en la parte superior de la pestaña para ver el objeto en una vista de árbol o diagrama o imprimir el contenido. Si ejecuta un análisis y los resultados no incluyen objetos que se sabe que existen en el sistema, pida al administrador que compruebe que tiene los privilegios apropiados para ver estos objetos.









## **SAS Lineage**

SAS Lineage es un cliente web que le permite ver el linaje de fuentes y destinos en un trabajo. Si el Servicio de Relación SAS contenido ha sido habilitado, los metadatos de SAS Data Integration Studio será recuperado por el servicio, en los que se puede acceder por SAS Linaje. Por ejemplo, podría utilizar SAS Lineage para mostrar el linaje de una tabla o archivo externo que se utiliza en un trabajo de SAS Data Integration Studio. Como alternativa, puede haga clic en un archivo de tabla o externa en el SAS Data Integration Studio y seleccione la opción Analizar Visualizador Web en. Esta opción le permite iniciar sesión en SAS información de Lineage y vista al linaje para esa tabla o archivo.

## **Working with SPD Server Cluster Tables**

El SAS Scalable datos de rendimiento (SPD) Server le permite crear tablas dinámicas de racimo. Una tabla de clúster dinámico es de dos o más tablas del servidor SPD que están encadenados virtualmente en una sola entidad, el uso de metadatos que es administrado por el SPD Server.

Tablas dinámicas de racimo se pueden utilizar como las entradas o salidas en los empleos de SAS Data Integration Studio. Antes de crear una tabla de clúster, los siguientes requisitos deben cumplirse:

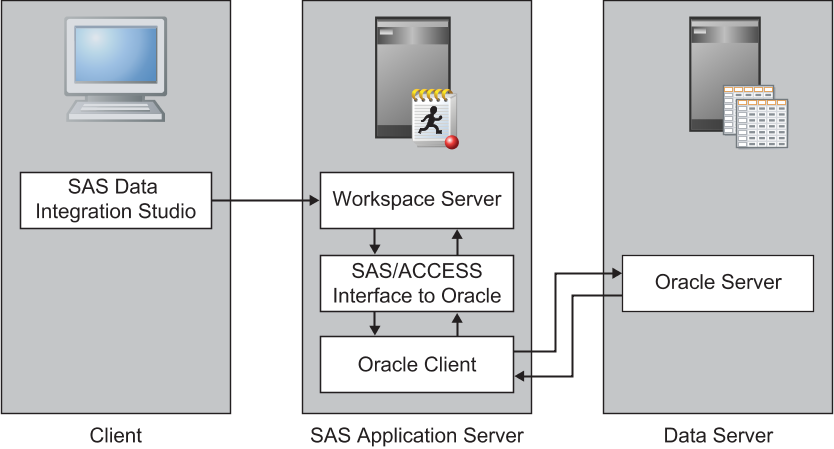
* Los administradores deben haber instalado, comenzaron, y registraron un servidor SPD. El servidor de aplicaciones que ejecuta el trabajo tabla de clúster debe ser capaz de acceder a la SPD Server. Para obtener más información acerca de los servidores SPD, ver los capítulos sobre las fuentes de datos comunes en la Plataforma de Inteligencia de SAS: Guía de administración de datos.
* Una biblioteca de servidor SPD debe estar disponible. Para obtener más información acerca de las bibliotecas de servidor SPD, ver los capítulos sobre las fuentes de datos comunes en la Plataforma de Inteligencia de SAS: Guía de administración de datos.
* Todas las tablas de origen que se van a añadir a una tabla de clúster se han registrado en la biblioteca SPD Server. Todas las tablas de origen deben tener la misma estructura de la columna.
* Una tabla de clúster se ha registrado en la biblioteca SPD Server. La tabla de clúster y todas sus tablas de origen deben tener la misma estructura de la columna. Una forma de asegurar que todas estas tablas tienen las mismas columnas es utilizar el asistente Nueva tabla para copiar los metadatos de una tabla de origen y guardarlo como los metadatos de la tabla de clúster.

## **Working with Hadoop**

El árbol Transformaciones en el SAS Data Integration Studio incluye una carpeta de Hadoop. Las transformaciones en esta carpeta le permiten realizar las siguientes operaciones dentro de un trabajo:

* Leer archivos de un clúster Hadoop y escribir archivos en el clúster
* Transferir archivos desde y hacia un cluster Hadoop
* Enviar su propio Pig latin, HiveQL, o código de MapReduce
* Utilizar una transformación para realizar una serie de pasos en una conexión con el clúster Hadoop, tales como traslados desde y hacia Hadoop, MapReduce procesamiento y procesamiento de Pig latin.

Hadoop es una tecnología de código abierto para el gran almacenamiento y procesamiento volumen de datos. Hadoop ofrece escalabilidad a través de la unión del Sistema Hadoop Distributed File (HDFS), su gran ancho de banda y el sistema de almacenamiento en clúster, y MapReduce, su algoritmo de procesamiento distribuido tolerante a fallos. Apache Pig es una plataforma de alto nivel para la creación de MapReduce programas que se utilizan con Hadoop. El lenguaje de esta plataforma se llama latín de cerdo. Apache Colmena es una infraestructura de almacenamiento de datos construida encima de Hadoop para consultas de datos, análisis y resumen. Proporciona un lenguaje SQL, llamado HiveQL.



# **UTILIZACIÓN**

Como se ha ido mencionado SAS ofrece una gran cantidad de productos y soluciones orientadas al análisis de datos, inteligencia de negocios y la gestión de datos, por lo que son muchas las áreas en la industria que lo prefieren con respecto a los diferentes productos y soluciones existentes en el mercado. Las áreas de la industria más destacadas y que gozan de un mayor soporte por parte de SAS son:

1. Automovilística

* Maruti Suzuki

1. Bancaria

* Scotiabank

1. Mercados de capitales
2. Casinos

* Foxwoods Resort Casino

1. Comunicaciones

* Telenor

1. Bienes de consumo envasados

* Nestle

1. Seguridad y defensa

* Delaware State Police

1. Soluciones gubernamentales

* HM Revenue & Customs

1. Salud

* Physicians Pharmacy ALLIANCE

1. Seguro médico

* Blue Cross Blue Shield of New Jersey

1. Fabricación de productos de alta tecnología

* Semiconductor Manufacturing International Corporation

1. Educación superior
2. Hoteles

* Foxwoods Resort Casino

1. Seguros

* CNA

1. P-12 Education

* Educator Quality Dashboard

1. Ciencias de la salud
2. Manufactura

* DOW

1. Media

* DirecTV

1. Energetica (Oil & Gas)

* Shell Americans

1. Retail

* DSW (Designer Shoe Warehouse)

1. Pequeñas y medianas empresas

* Orlando Magic

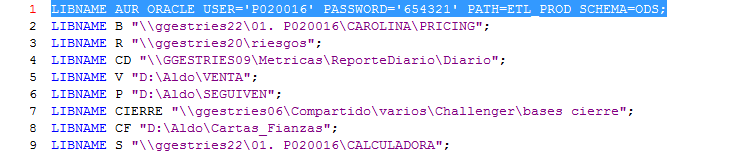
1. Deportes

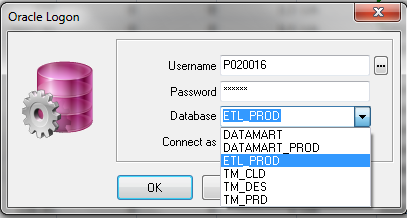
* New York Mets

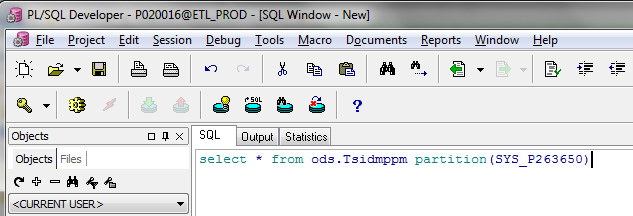
1. Viajes y transportes
2. Utilities

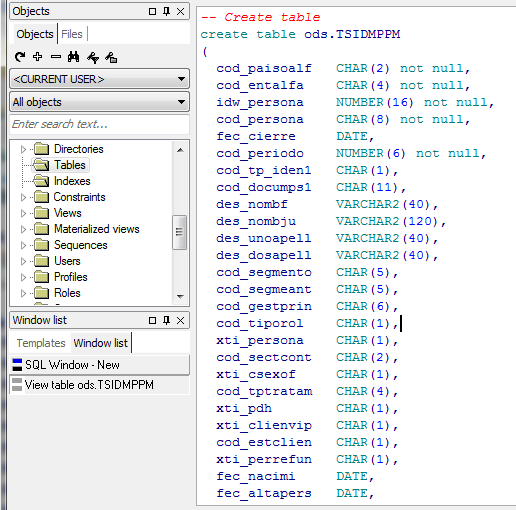
* MIT Sloan Management Review

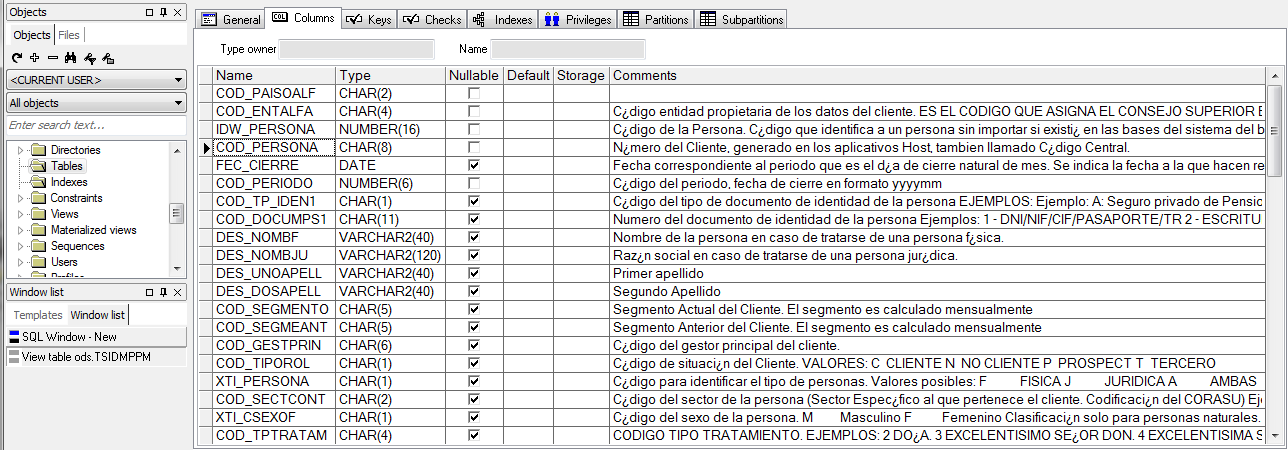
# **CASO PRÁCTICO**

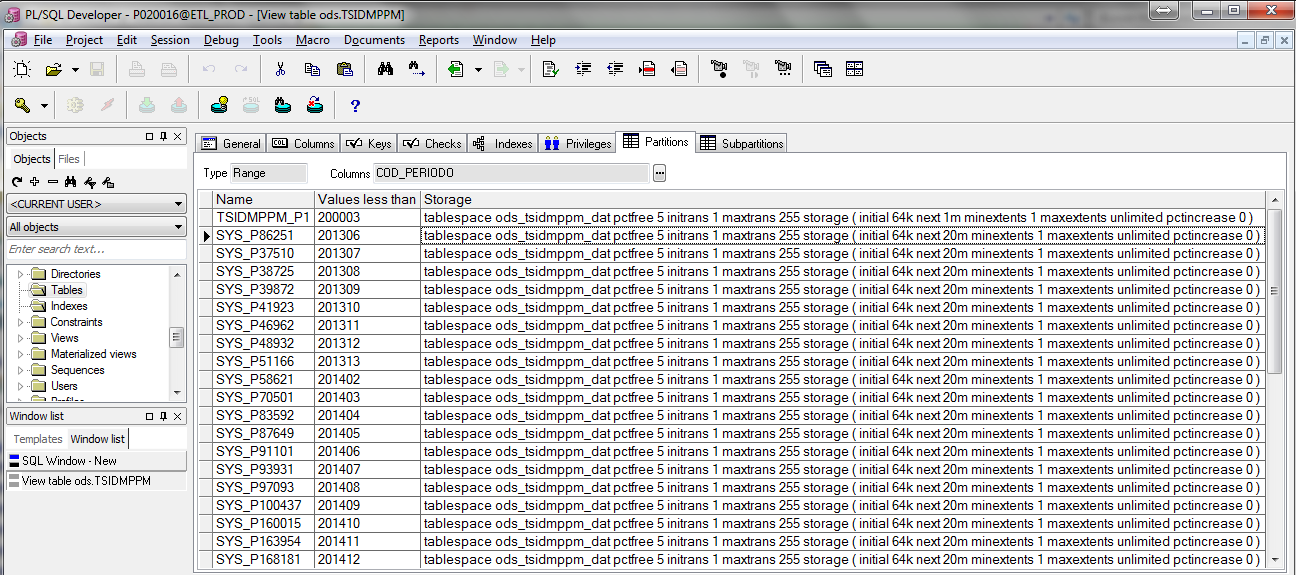


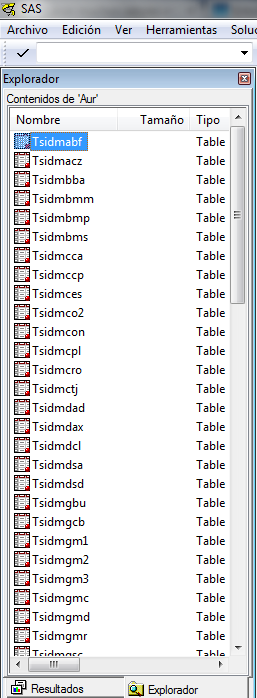


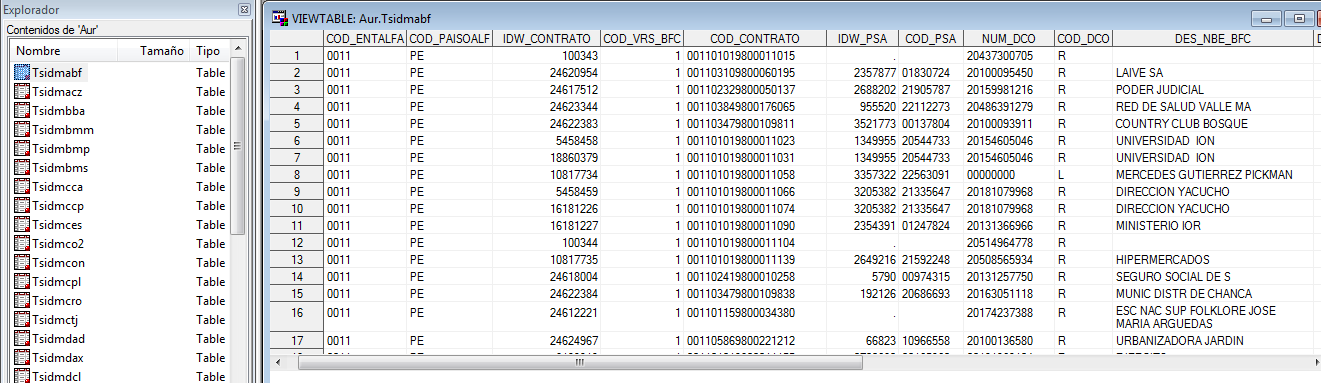


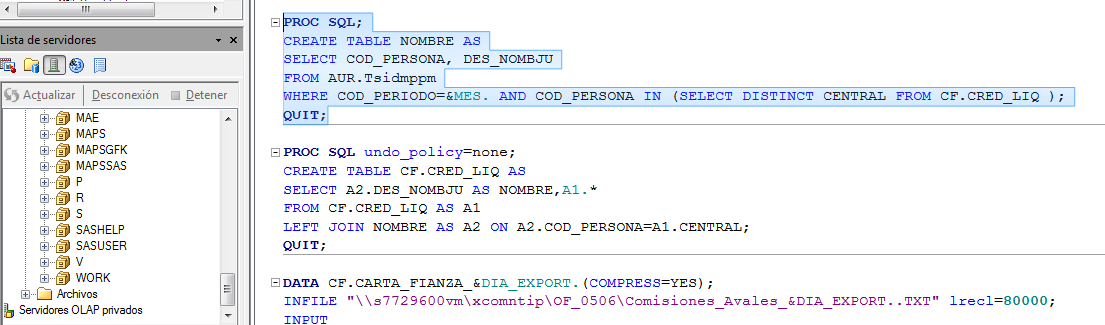


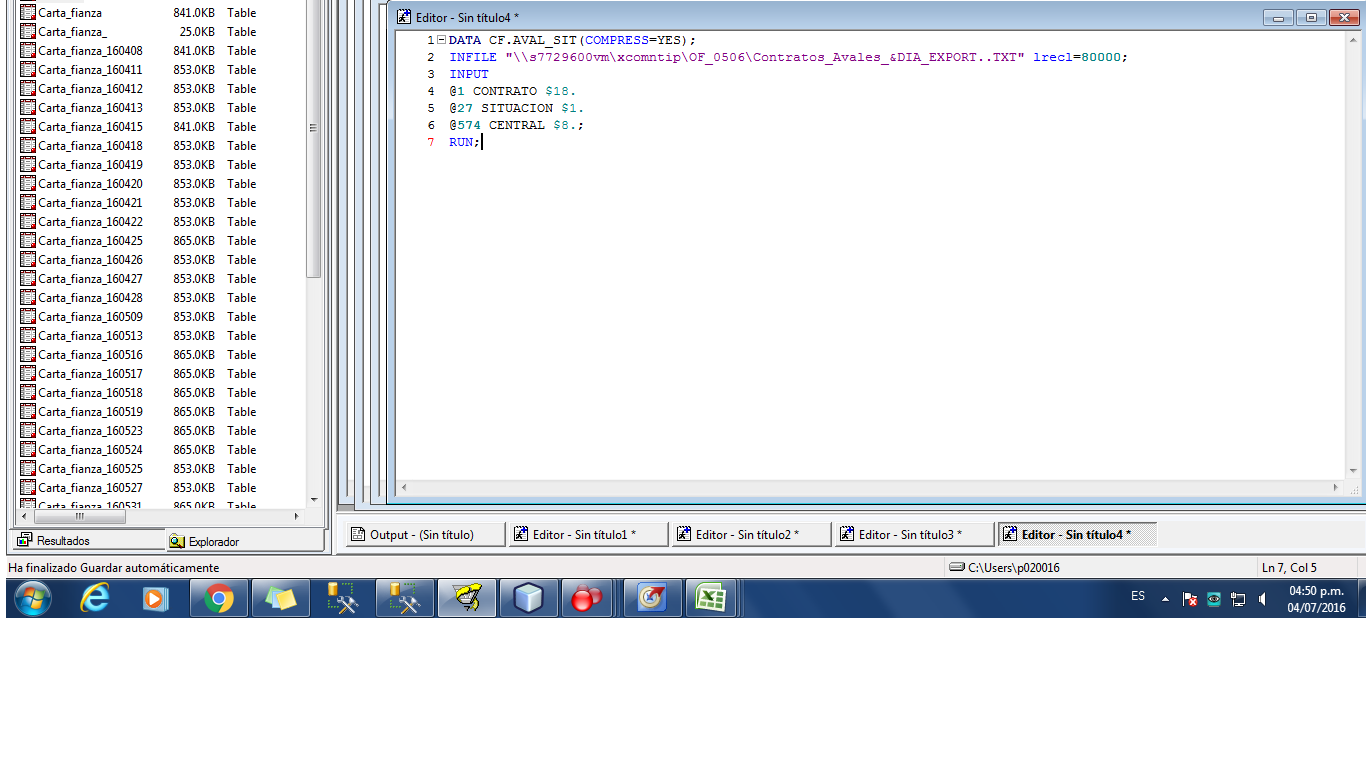


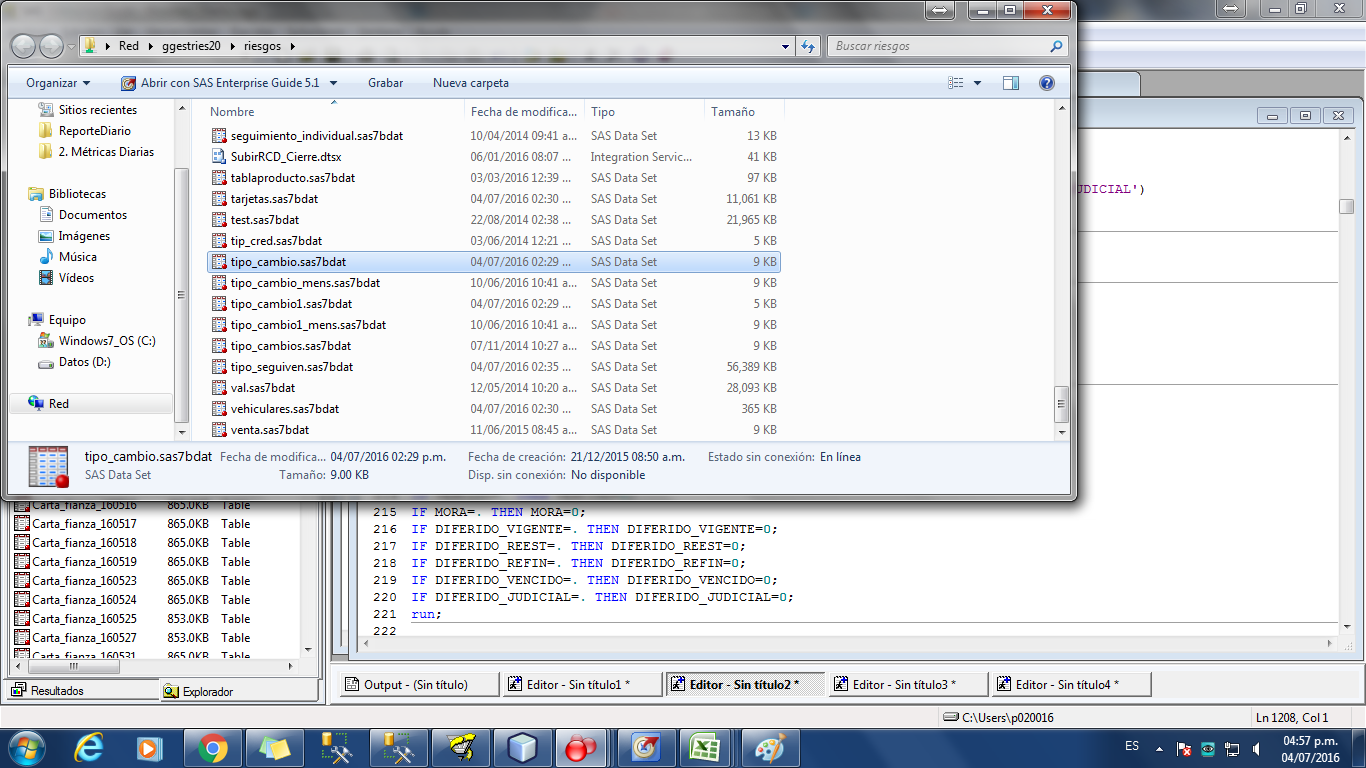


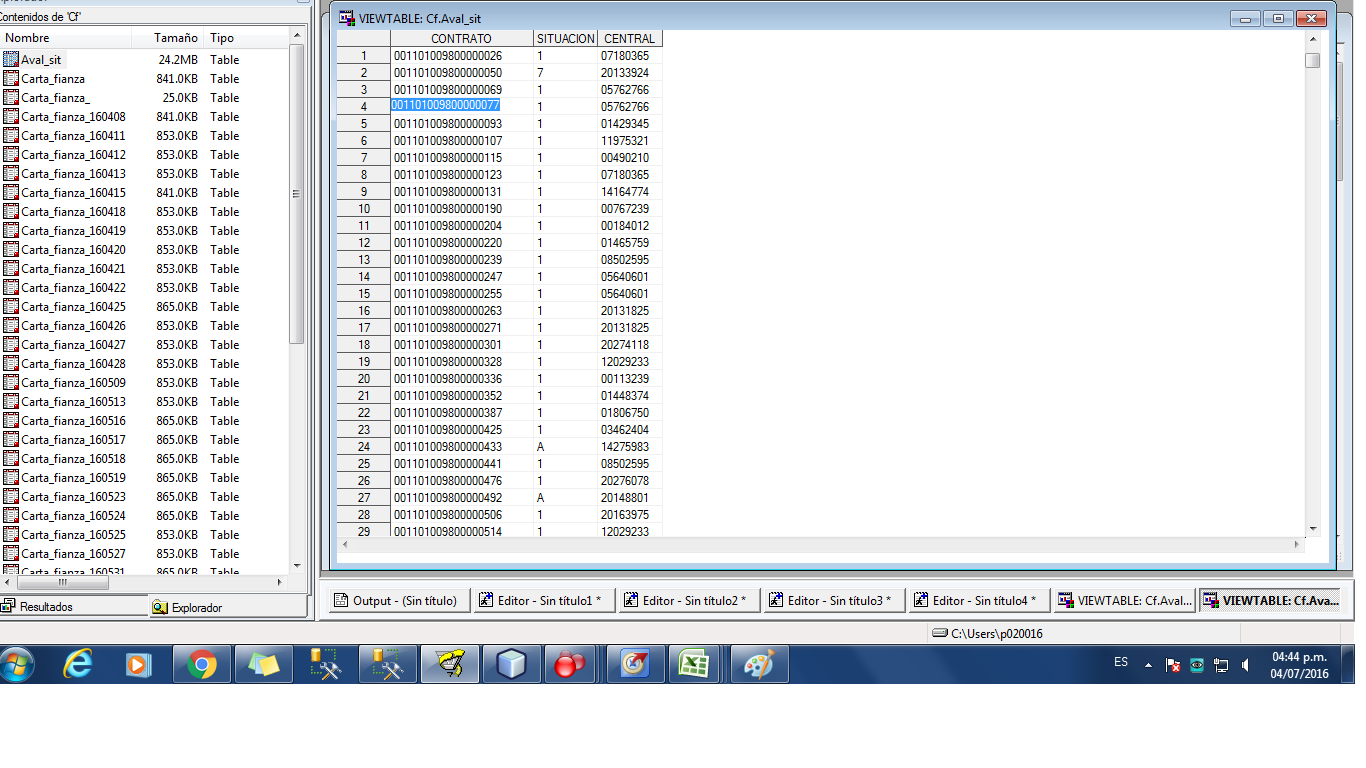


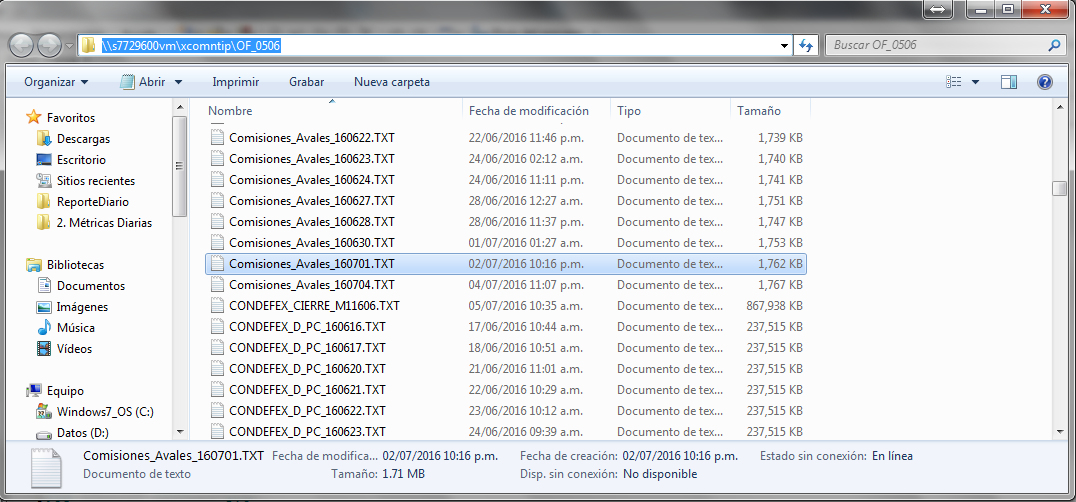


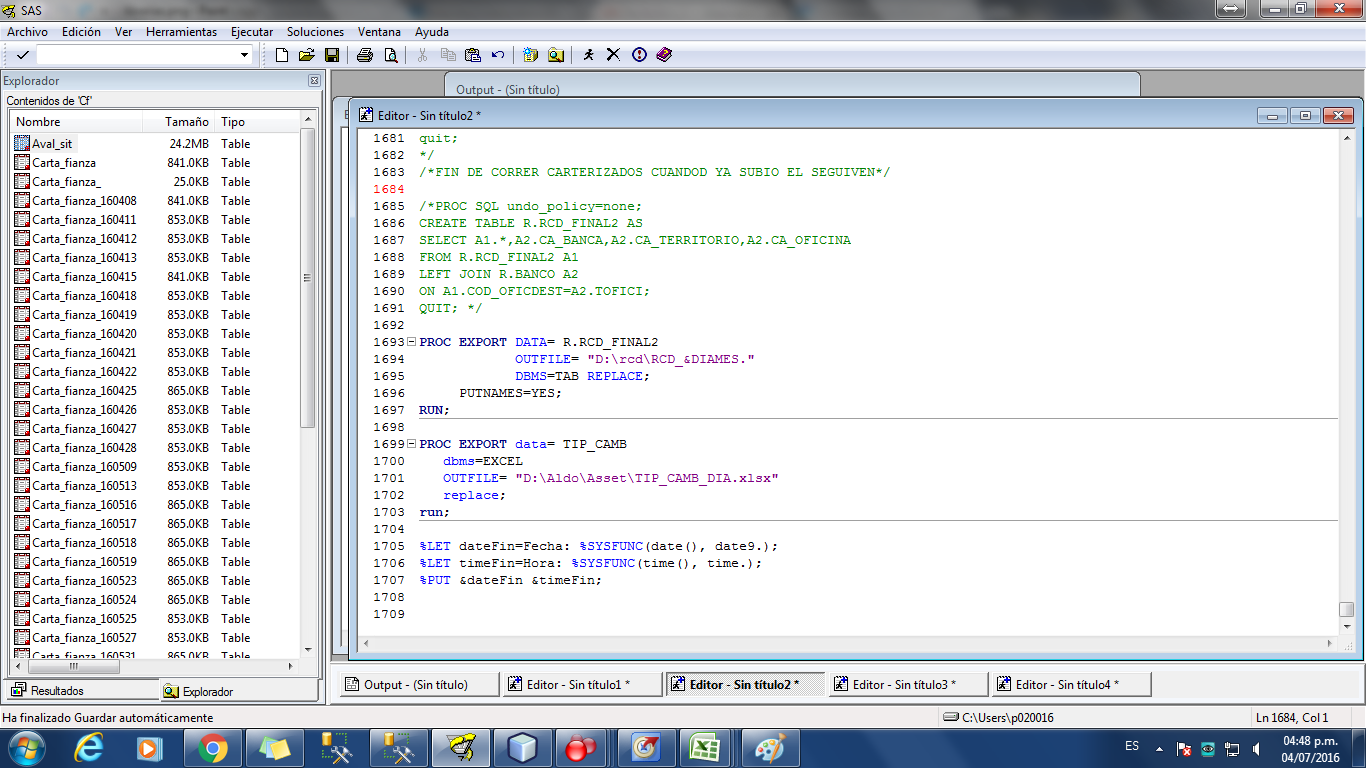


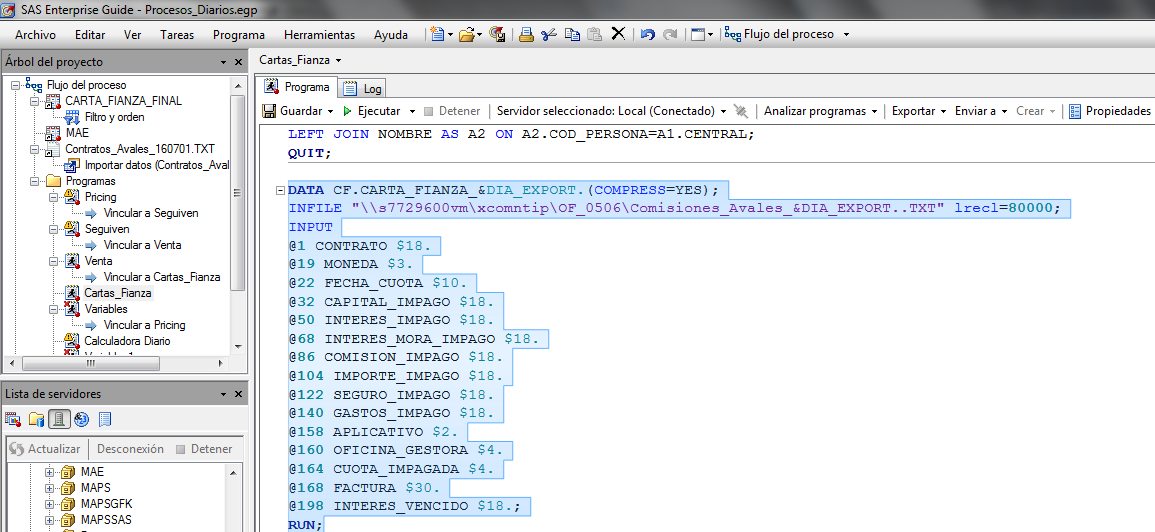


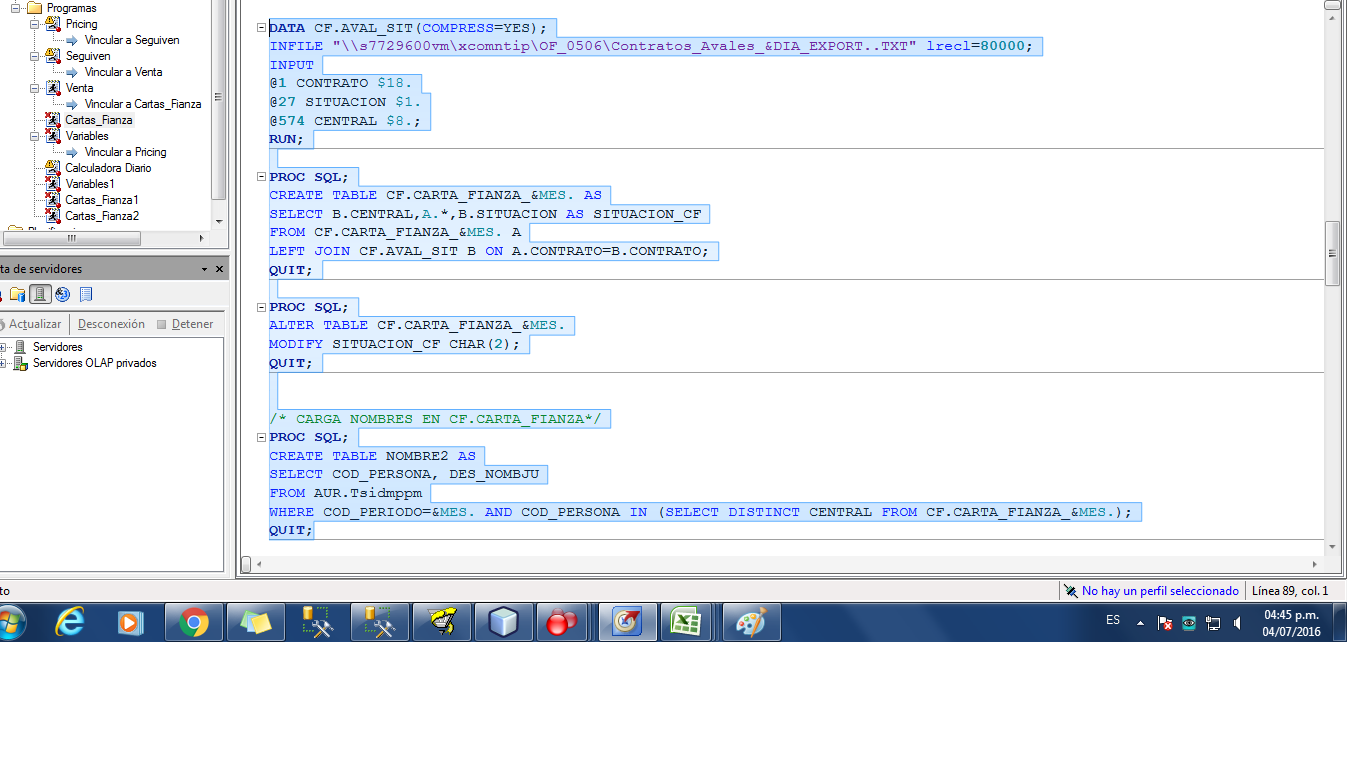


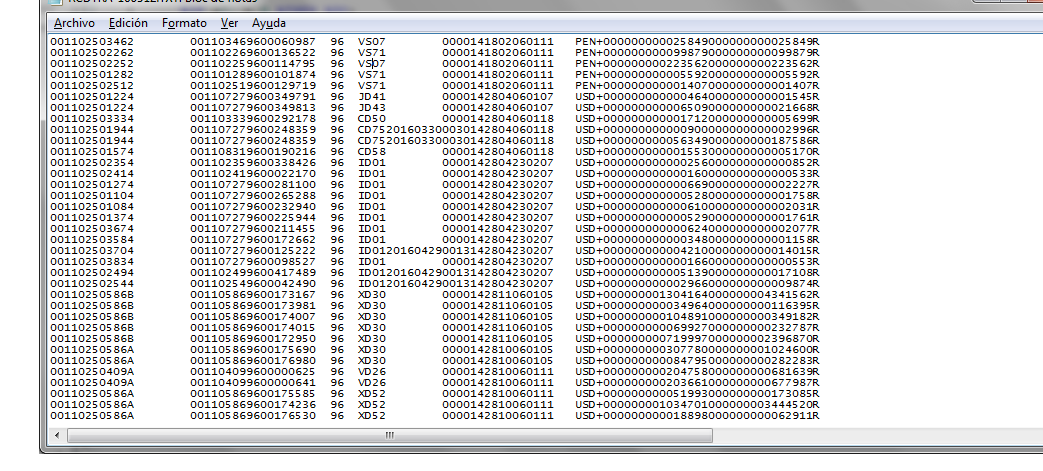


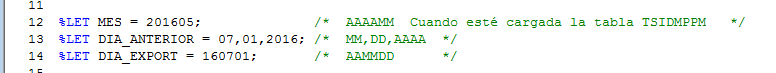












# **LIMITACIONES Y/O PROBLEMAS**

***Con referencia al SAS Data Integration Studio…***

En principio, la elección del SAS Data Integration Studio fue motivado por el interés en conocer sobre el software en cuestión, además, del poco conocimiento que se tiene sobre la infinidad de herramientas que nos provee la compañía americana SAS. SAS Institute es uno de los más importantes fabricantes de business intelligence software con una cifra anual de ventas de más de $1.680 millones.

Sin embargo, la realidad fue que dicha empresa cobra un precio elevado por sus productos y estos no están disponibles de manera gratuita, es decir, no es software libre en el mundo del internet contemporáneo. Es así que, a contracorriente, decidimos ir en busca de este software que cada vez parecía imposible de obtener.

***Con referencia al SAS Deployment Wizard…***

Al no encontrar referencia física sobre el SAS Data Integration Studio, optamos por adquirir el SAS Deployment Wizard 9.4. para comenzar a realizar el caso práctico, ya que nos encontrábamos a pocos días de la presentación final.

Este programa nos permite extraer data de diferentes formatos, además, podemos cruzar data, agregar estadísticas, imprimir análisis estadísticos de la data que se obtiene, importar data de una base de datos externa, manipular dicha BD desde SAS, convertir todo a formato DataSet de SAS, con lo cual se puede trabajar bajo el lenguaje de SAS y extraer estadísticos importantes para la toma de decisiones (no realmente), como media, varianza, coeficiente de correlación entre columnas, etc.. Además, nos permite exportar la data transformada a una base de datos relacional.

Sin embargo, no es una herramienta de BI propiamente dicha, puesto que no hace transformaciones a cubos OLAP, ni otra estructura de BI.

# **CONCLUSIONES**

**Con referencia al SAS Data Integration Studio…**

Es una herramienta intuitiva y, por su interfaz, será muy amigable al usuario. Una de sus principales características se rigen por el hecho que cualquier usuario podrá manejar dicho software sin temor a generar un problema técnico, ya que el SAS Data Integration Studio no está para la parte del desarrollador o analista de Base de Datos, sino por el lado del usuario. Esto, sin duda, es una ventaja que pocos software que manejan big data podrán sacar adelante.

Mediante la construcción de almacenes, procesamiento de datos y la visualización de cubos, que son las acciones que se tomarán en adelante, los esquemas y la interfaz, este software nos garantiza con su última versión y las plataformas con las que trabaja (sin obviar la ayuda a los desarrolladores) con referencia al almacenamiento de datos y la integración de datos, un procesamiento más eficiente, de mayor control y flexibilidad, sin olvidar que las principales áreas que toca se desarrollan en características de rendimiento en el trabajo y capacidad de gestión de metadatos.

**Con referencia al SAS Deployment Wizard…**

Esta es una herramienta bastante útil cuando se busca obtener estadísticas basadas en la data que fue ingresada, sea para nivel académico o nivel empresarial, para poder importarla en una base de datos externa. De esta forma, podemos tener mayor control de la información que será procesada hasta su resultado estadístico. Además, cuenta con un gran potencial si se pudiese añadir productos de SAS para complementar su funcionalidad, ya que el uso de la data que se tiene basada en la Inteligencia de Negocios y el aporte que nos brinda el SAS Data Integration Studio a nivel de visualización serían, en conjunto, un software apropiado para aquellos que manejen una big data.

Algo que se podría agregar con referencia al SAS Deployment Wizard, a diferencia del SAS Data Integration Studio, es que este programa base del SAS está más orientado a la parte de la programación, ya que converge lo que son códigos, base de datos y data estadística como fin último. Esto puede resultar no muy intuitivo para cualquier persona que desea utilizar dicha herramienta a nivel empresarial a nivel gerencial.

# **BIBLIOGRAFÍA**

* Reyes Villamizar, F. (2013). SAS: La Sociedad Por Acciones Simplificadas. Legis

# **ENLACES EXTERNOS**

* <http://www.mundoenlinea.cl/noticia.php?noticia_id=32420&categoria_id=54>
* <http://pyme.emol.com/el-big-data-las-pymes/>
* <http://abcblogs.abc.es/tareas-pendientes/2015/10/09/el-iot/>
* <http://worldofdata.es/entrevista-christian-gardiner/>
* <http://support.sas.com/resources/papers/proceedings11/133-2011.pdf>
* <http://www.sas.com/en_us/industry.html>