UNIVERSIDAD CENFOTEC

Maestría en Tecnologías de Bases de Datos

Proyecto de Investigación Aplicada

Construir un modelo de negocios que consolide las diferentes fuentes de datos, con el fin de mejorar la calidad de la información actualmente utilizada por la organización WW TS Support Solutions Pricing en Hewlett Packard Enterprise para la toma de decisiones.

Elaborado por:

Araya Elizondo, Ian

Muñoz Alcázar, Jorge

Julio 2016, Costa Rica

Lista de Figuras & Tablas

Tabla de contenidos

Capítulo 1.

Introducción

1.1 Generalidades

Hewlett Packard (HP), como parte de los servicios que brinda a los clientes, tiene una plataforma dedicada a la venta de garantías, que cubren no solo el hardware sino también el software. Como parte de un servicio de alta calidad y nivel de personalización detallado, se le permite al cliente crear su propio paquete de cobertura para los productos adquiridos.

Existe a nivel mundial una gran cartera de productos que van desde impresoras, hasta servidores de alta gama, agrupados en diferentes organizaciones, cada una con su propio conjunto de herramientas.

1.2 Antecedentes del problema

La organización de Pricing está a cargo de monitorear las diferentes herramientas para conseguir estandarización de los procesos y homologación de precios de los paquetes de garantía.

Actualmente se cuenta con una plataforma obsoleta conformada por un grupo de herramientas, que entró en funcionamiento en el año 2007 y no se le han realizado actualizaciones desde entonces. Estas herramientas están trabajando en hojas de Excel con macros, usando Visual Basic 6, que también está fuera de soporte por parte de Microsoft.

Las herramientas con las que se cuenta son: INDICO y DC CONFIG TOOL.

El seguimiento de los contratos de garantía se ve afectado por este tipo de solución, los mismos se almacenan en listas de Sharepoint, lo que complica el reporteo y análisis del comportamiento de la información.

La calidad de los datos se ha visto severamente afectada por este sistema, y se dan innumerables inconsistencias en la información, debido al nivel de complejidad en la creación de reportes y no se cuenta con una herramienta transaccional única, sino con múltiples hojas de Excel creadas por una gran cantidad de vendedores según su conveniencia.

En el 2014 se contrataron los servicios internos para el desarrollo de software, con esto se pretendía relevar una de las herramientas obsoletas, DC CONFIG TOOL, por un nuevo sistema basado en Java con SQL Server como base de datos, que se llamó CASPER.

Con la llegada de CASPER, se incrementó la necesidad de crear un sistema que unifique la información de todas las herramientas utilizadas por la organización para estandarización de datos y la generación de reportes para la toma de decisiones que hasta este punto no ha sido posible, generando una sesgada toma de decisiones.

Al ser HP una organización multinacional y de grandes dimensiones, cuenta con una división de TI que se encarga de ofrecer soluciones para solventar las necesidades de las diferentes organizaciones. Dichas necesidades se priorizan y se desarrollan según las necesidades de la compañía, esto genera un retraso importante en el desarrollo de herramientas para la toma de decisiones y la prioridad son las herramientas transaccionales.

1.3 Definición y descripción del problema

Se han identificado problemas muy específicos que son:

- Al ser una organización estrictamente de carácter financiero, no cuenta con el recurso humano necesario para el desarrollo de soluciones de tipo tecnológico, lo que influye en el grado de obsolescencia en las herramientas usadas.
- Se tienen múltiples fuentes de datos y todas las herramientas son manejadas por diferentes grupos y almacenadas de manera independiente. Las fuentes son:
  - Hojas de Excel
  - SharePoint (Formularios y listados de archivos)
  - SQL Server
- La calidad de los datos es muy deficiente debido a desarrollos independientes y no calculados, cada grupo realizó su herramienta basado en su propio criterio y sus estándares regionales.

- Carencia de un depósito de datos centralizado, que reúna la información de las diferentes herramientas.
- Al no existir un repositorio centralizado, no hay una plataforma de reportes que permita la toma de decisiones basadas en la información.

1.4 Justificación

Basado en los problemas identificados, se nota de forma urgente la necesidad de cambios en la organización de manera que tenga un giro en la forma en la que toman decisiones debido a que con la solución de los problemas identificados, la organización de Pricing, va a ser capaz de:

- Obtener una solución tecnológica moderna, que le permita un manejo de la información de forma ágil y amigable con los usuarios.
- Centralizar las diferentes fuentes de datos en un repositorio único.
- Tener confiablidad e integridad de la información.
- Contar con un sistema de reportes personalizado para cada sección de la organización.
- Tomar decisiones a partir de la información que se recopiló en reportes antes mencionados.

La finalidad de esta implementación es que la organización, a mediano plazo, cuente con herramientas de inteligencia de negocios, que le permitan de forma moderna y tecnológica aprovechar la información que hasta este momento no ha sido explotada de ninguna forma.

1.5 Viabilidad

1.5.1 Punto de vista técnico

Se cuenta con todos los recursos necesarios para la realización de la solución, entre los que están:

- Hardware
    - Repositorio de datos (servidor)

    - Laptops para desarrollo
    - Máquinas de prueba
        - Físicas
        - Virtuales
- Software
    - Licencia de Servidor (Windows Server 2012)
    - Licencia de SQL Server (2014)
    - Licencia de Visual Studio 2013
    - Licencias de Microsoft Office 2013
    - Licencia de VMWare 11.1

1.5.2 Punto de vista operativo

Se tiene el aval y apoyo de la dirección de la organización de Pricing para la realización de este proyecto.

Se cuentan con los siguientes recursos humanos:

- 1 Auditor del proyecto
- 1 Estadista
- 1 Desarrollador de las aplicaciones
- 2 Desarrolladores del proyecto

Se logró el aval de la organización para el uso de los recursos y de realización de este proyecto en tiempo laboral.

1.5.3 Punto de vista económico

Se cuenta con un presupuesto tangible, todos los involucrados en este proyecto son colaboradores de la misma compañía y todos los recursos técnicos están a disposición de la organización sin costo adicional.

1.6 Objetivos

Utilizar la Taxonomía de Bloom, debido a que es la más confiable y utilizada para este tipo de entregables académicos.

1.6.1 Objetivo general

Construir un modelo de negocios que consolide las diferentes fuentes de datos, con el fin de mejorar la calidad de la información, actualmente utilizada por la organización WW TS Support Solutions Pricing en Hewlett Packard Enterprise para la toma de decisiones.

1.6.2 Objetivos específicos

1. Identificar las diferentes fuentes de datos utilizadas por WW TS Support Solutions Pricing en Hewlett Packard Enterprise como insumo para el desarrollo de las actividades del día a día en la organización.
2. Producir un modelo de manejo de la calidad de los datos que garantice la integridad y la confiabilidad de la información que va a ser utilizada por la organización.
3. Construir un sistema de almacén de datos que consolide la información arrojada por las diferentes fuentes identificadas en el objetivo específico #1.

1.7 Alcances y limitaciones

1.7.1 Alcances

1. Modelo de manejo de la calidad de los datos.
2. Sistema de almacén de datos.
3. Conjunto de reportes.

1.7.2 Limitaciones

1. Por políticas de la empresa, utilizar la plataforma móvil es prohibido.

2. El funcionamiento actual de las fuentes de información no puede ser modificado.

3. El desarrollo del proyecto se hará en un servidor de pruebas, si la organización lo aprueba, será migrado a producción.

4. No toda la información, por ser de carácter confidencial, se va a compartir con la universidad.

5. Dado que el idioma oficial de la universidad es el español y el de la organización es el inglés, se entregará a la organización únicamente un documento con especificaciones técnicas y no el trabajo de tesis completo.

1.8 Marco de referencia organizacional y socio-económico

1.8.1 Historia[1]

En 1938, los ingenieros de la Universidad de Stanford William Hewlett y David Packard, compañeros de carrera y grandes amigos, decidieron montar en el pequeño garaje de Packard su primera empresa de electrónica. El garaje en cuestión, ubicado en el número 367 de Addison Avenue, en Palo Alto (California) incita a pensar que, en aquel entonces, en Palo Alto había más garajes que coches, porque futuros grandes empresarios comenzaron sus meteóricas carreras en un lugar similar.

Bill Gates y Paul Allen desarrollaron la primera versión de su BASIC para un Altair 8800 en un garaje parecido y Steve Jobs y Steve Wozniak fabricaban a mano sus primeros Ordenadores Apple en otro garaje. También Chad Hurley y Steve Chen, fundadores de YouTube, comenzaron su proyecto en un garaje y Larry Page y Sergey Brin diseñaron su primera versión de Google en un garaje alquilado.

Hewlett y Packard desarrollaron un oscilador de audio de precisión, el conocido como Modelo 200A. Utilizaron una bombilla como resistencia, para

[1] Tomado de http://www8.hp.com/us/en/hp-information/about-hp/history/hp-garage/hp-garage.html

estabilizar la temperatura del circuito, lo que les permitió simplificar el dispositivo y reducir el precio de venta a $54,40, en lugar de los $200 que valían otros modelos menos estables del mercado.

El 1 de enero de 1939, los dos ingenieros fundaron la empresa Hewlett-Packard y consiguieron sacar al mercado el modelo 200B de su oscilador. Este aparato tuvo como primer cliente a los estudios Walt Disney Pictures, que compraron ocho para sincronizar los efectos de sonido a la película Fantasía. HP ha llegado a convertirse hoy en una de las compañías de tecnologías de la información más importantes del mundo.

En 1987 este garaje fue declarado lugar de nacimiento de lo que, con los años, sería Silicon Valley, el territorio técnicamente más avanzado del mundo. Los empresarios fundadores no pararon hasta intentar recuperar la pequeña construcción de madera y en el año 2000, HP logró hacerse con el garaje y la vivienda anexa. En el año 2005 terminaron su restauración por completo con el fin de preservar este legado.

En el año 2007, el garaje de HP fue declarado Lugar Histórico de Estados Unidos, una conmemoración realmente importante y que otorga a la construcción carácter de lugar histórico como lo pueden ser la isla de Alcatraz, las cataratas del Niágara o el Gran Cañón.

David Packard falleció en 1996, a la edad de 84 años; William Hewlett murió en el año 2001, contando con 88 años de edad. Seguramente ambos se fueron felices al ver sus sueños cumplidos y su garaje en manos de la empresa que surgió del interior de sus cuatro paredes.

1.8.2 Tipo de negocio y mercado meta[2]

HP Financial Services

HPFS apoya y mejora las soluciones de productos y servicios globales de HP, proporcionando una amplia gama de servicios de gestión del ciclo de vida económica con valor agregado. HPFS permite a los clientes en todo el mundo adquirir soluciones de TI completas, incluyendo hardware, software y servicios. El grupo ofrece arrendamiento, financiamiento, programas de servicios públicos y servicios de recuperación de activos, así como servicios de gestión de activos financieros para grandes clientes globales y de la empresa. HPFS también ofrece una gama de servicios financieros especializados a pymes y entidades educativas y gubernamentales. HPFS ofrece alternativas innovadoras, personalizadas y flexibles para equilibrar el flujo de caja única del cliente, la obsolescencia tecnológica y las necesidades de capacidad.

1.8.3 Misión, Visión y Valores[3]

Misión

"Proveer soluciones de calidad, a través de la iniciativa y respuesta de sus integrantes, ofreciendo tecnologías de vanguardia y servicios de valor agregado para asegurar la satisfacción de nuestros clientes."

Visión

"Ofrecer la mejor experiencia digital del mercado, capaz de motivar la interacción con los clientes y garantizar que HP sea la marca elegida en todo el mundo posicionándonos como líderes del mercado. "

[2] Tomado de http://www.sec.gov/Archives/edgar/data/47217/000104746910010444/a2201180z10-k.htm
[3] Tomado de http://hewlett-packard-unaq.blogspot.com/p/analisis-del-entorno.html

Objetivos Corporativos

"Es necesario que las personas trabajen juntas al unísono en pos de objetivos comunes y eviten en todos los niveles trabajar en metas contrarias si desean obtener lo mejor en eficacia y logros." ***Dave Packard***

1.8.4 Políticas institucionales[4]

Debido a la naturaleza internacional del negocio, los cambios políticos o económicos u otros factores podrían perjudicar futuros ingresos, costos y gastos y la situación financiera.

Como aproximadamente el 65% de las ventas son de países fuera de los Estados Unidos, otras monedas, especialmente el euro, la libra esterlina, renminbi yuan chino y el yen japonés, puede tener un impacto en los resultados de HP (expresado en dólares estadounidenses). Variaciones de divisas también contribuyen a las variaciones en las ventas de productos y servicios de las jurisdicciones afectadas.

En consecuencia, las fluctuaciones en las tasas de cambio extranjeras, especialmente el fortalecimiento del dólar frente al euro, podrían afectar adversamente el crecimiento de ingresos en períodos futuros. Además, las variaciones de cambio pueden afectar negativamente los márgenes de las ventas de los productos en los países fuera de los Estados Unidos y los márgenes de las ventas de productos que incluyen componentes obtenidos de proveedores ubicados fuera de los Estados Unidos.

Utilizamos una combinación de contratos de futuros y opciones designados como coberturas de flujo de efectivo para proteger contra riesgos de tipo de cambio de moneda extranjera. La eficacia de nuestras coberturas depende de nuestra capacidad para predecir con precisión los flujos de efectivo futuros, lo que es particularmente difícil durante los períodos de demanda incierta para nuestros productos y servicios y tipos de cambio altamente volátiles.

[4] Tomado de http://www.sec.gov/Archives/edgar/data/47217/000104746910010444/a2201180z10-k.htm

Como resultado, se podría incurrir en pérdidas significativas de nuestras actividades de cobertura si nuestras previsiones son incorrectas. Además, nuestras actividades de cobertura pueden ser ineficaces o no pueden compensar cualquiera o más de una parte del impacto financiero adverso como consecuencia de cambios de divisas. Las ganancias o pérdidas asociadas a las actividades de cobertura también pueden afectar nuestros ingresos y en menor medida el costo de ventas y situación financiera.

En muchos países, particularmente en aquellos con economías en desarrollo, es común participar en las prácticas de negocios que están prohibidas por las leyes y reglamentos aplicables a nosotros, como la Ley de Prácticas Corruptas en el Extranjero. Por ejemplo, como se explica en la Nota 18 a los Estados Financieros Consolidados, la Oficina del Fiscal General alemán, el Departamento de Justicia de Estados Unidos y la SEC ha denunciado que investigan que ciertos empleados y ex empleados de HP participan en el soborno, la malversación y evasión de impuestos o eran involucrados en sobornos u otros pagos indebidos.

“Aunque ponemos en práctica políticas y procedimientos diseñados para facilitar el cumplimiento de estas leyes, nuestros empleados, contratistas y agentes, así como de aquellas empresas a las que externalizar algunas de nuestras operaciones de negocios, podrá adoptar acciones en violación de nuestras políticas. Cualquier violación, aunque prohibido por nuestras políticas, podría tener un efecto adverso en nuestro negocio y reputación.”

1.9 Estado de la cuestión

Para la implementación de este apartado, se va a usar de referencia el informe técnico UCLM-TSI-003. (Carlos Blanco, 2008)

1.9.1 Introducción

En los estudios sobre calidad de datos que se hacen en la actualidad, se destaca constantemente la importancia de disponer de una versión acertada de la realidad para tomar decisiones acertadas. Según estos mismos estudios, éste es un asunto que se encuentra entre los temas que más preocupan a los ejecutivos de TI de las organizaciones.

Sin embargo, la gestión de la información acaba siendo uno de los primeros proyectos que cae de la cartera cuando hay una reducción de presupuesto.

La organización de este informe es la siguiente: en primer lugar, se comienza señalando en la Sección 1 a modo de introducción la importancia de la calidad de los datos, para terminar introduciendo la técnica de revisión sistemática utilizada.

El resto de secciones corresponden a las etapas de la revisión sistemática de modo que en la Sección 2 se planifica la revisión, en la Sección 3 se ejecuta sobre las fuentes seleccionadas y en la Sección 4 se analizan los resultados obtenidos realizando la comparación formal de las principales propuestas. Finalmente, la Sección 5 muestra las conclusiones de este trabajo.

1.9.1.1 Calidad de los datos

Se refiere a los procesos, técnicas, algoritmos y operaciones encaminados a mejorar la calidad de los datos existentes en empresas y organismos.

Sin embargo, la calidad de datos generalmente se refiere al mejoramiento de la calidad de los datos de personas físicas y jurídicas, por ser éstos los que más tienden a degradarse y cuya falta de calidad impacta en la productividad de las organizaciones.

Los principales beneficios de la calidad de datos son:

- Ahorrar costos directos: evitando tener información duplicada y por lo tanto evitar el envío replicado de cartas a un mismo cliente.
- Potenciar las acciones de mercadeo y la gestión: la normalización de archivos mejora el análisis de datos y permite segmentaciones precisas

para que sus acciones de mercadeo y su gestión ganen en precisión y eficacia.

- Optimizar la captación y la fidelización de clientes: con los datos correctos, se mejoran las tasas de respuestas y el cliente se siente plenamente identificado con la empresa.
- Mejorar la imagen corporativa: el cliente sólo recibe el envío que le corresponde, una sola vez y con sus datos correctos.
- Mejorar el servicio: identificación más rápidamente del cliente que llama a un Centro de llamadas, reduciendo los tiempos de espera y, dejando tiempo al operador para centrarse en el mensaje de negocio.

1.9.1.2 Revisión sistemática

Realizar una revisión de la literatura existente mediante el enfoque de una revisión sistemática permite identificar, evaluar e interpretar todos los estudios importantes o significativos (llamados estudios primarios) para una pregunta de investigación en particular. De este modo hay varias razones que remarcan la utilidad de usar este enfoque:

- Resume la evidencia existente concerniente a un tratamiento o tecnología.
- Identifica brechas en la investigación actual para sugerir áreas de ulterior investigación.
- Proporciona un marco para posicionar apropiadamente nuevas actividades de investigación.

La técnica de revisión sistemática comienza definiendo un protocolo de revisión que especifica la cuestión investigada y los métodos a utilizar, documentando su estrategia de búsqueda para que los lectores puedan conocer su rigor y compleción. Se basa en una estrategia de investigación definida que pretende detectar toda la literatura relevante posible, requiriendo de criterios explícitos de inclusión y exclusión para evaluar cada estudio primario potencial y especificando la información de cada uno de estos estudios primarios incluyendo criterios de calidad.

Barbara Kitchenham propone un método para la realización de revisiones sistemáticas en el contexto de la ingeniería del software, que será utilizado en la presente revisión sistemática.

1.9.2 Planificación de la revisión

En esta etapa se identificó la necesidad de la revisión indicando cuáles son sus objetivos, qué fuentes se utilizarán para identificar los estudios primarios, si hubo algunas restricciones, cuáles son los criterios de inclusión y exclusión, qué criterios se utilizarán para evaluar la calidad de los estudios primarios y cómo se extraerán y sintetizarán los datos de los estudios.

1.9.2.1 Formulación de la pregunta

Se comenzó formulando la pregunta de investigación de forma que se focalice el área de interés del trabajo y queden definidos tanto el problema a tratar como sus principales características.

1.9.2.2 Foco de la pregunta

En esta revisión sistemática se pretende localizar trabajos centrados en estándares en el manejo de la calidad de los datos y realicen aportaciones importantes en el área.

1.9.2.3 Amplitud y calidad de la pregunta

Para definir la amplitud y calidad de la pregunta, se propone la pregunta de investigación y el conjunto de palabras clave identificadas, así como los resultados que se esperó obtener y cómo serán analizados.

1.9.2.4 Problema

Tal y como se pudo ver en la sección de introducción, quedó de manifiesto tanto la necesidad de considerar la calidad de los datos. De este modo, el problema se centra en el estudio de los trabajos realizados en el área de la calidad de los datos.

1.9.2.5 Pregunta de investigación

Una vez conocido el problema se puede definir la pregunta de investigación de este proyecto de la siguiente forma:

¿Qué trabajos se han realizado, dirigidos a estándares de manejo de la calidad de datos?

1.9.2.6 Palabras clave y sinónimos

Es necesario definir un conjunto de palabras clave presentes en los trabajos a localizar, que sirvan como base en la creación de las consultas que se aplicarán a las distintas fuentes para obtener los trabajos primarios. Para la selección de las palabras clave y sus conceptos relacionados, así como para conocer la traducción correcta al inglés de cada término, se revisaron varios artículos tanto de ingeniería ontológica como de seguridad de la información, que fueron proporcionados por los autores de este trabajo y por una búsqueda inicial.

A continuación, se puede ver un cuadro resumen en el que aparecen agrupadas por área las palabras clave y conceptos relacionados que se utilizarán en esta revisión.

Área	Palabra Clave
Calidad de datos	Data quality
	Management Information Systems (MIS)
Estándares	Data Quality Standard (DQS)
	ISO 8000

Cuadro 1. Muestra las palabras clave en la investigación. Fuente: autoría propia

1.9.2.7 Población

La población a analizar se compone de las publicaciones presentes en los repositorios de las fuentes de datos seleccionadas que estén relacionadas con el objetivo de esta revisión.

1.9.3 Selección de fuentes

En este apartado se analizaron principalmente las fuentes que se usaron para realizar la ejecución de la revisión. Posteriormente se utilizarán los elementos definidos en la planificación para aplicar el procedimiento de obtención de estudios primarios en cada una de las fuentes seleccionadas.

1.9.3.1 Definición del criterio de selección de fuentes

El criterio para la selección de las fuentes de búsqueda está basado en la opinión de los autores de este trabajo que se establece en su experiencia profesional recomendaron la lista de fuentes sobre las que realizar la revisión. Otros requisitos exigidos a las fuentes para su selección son su accesibilidad vía web y la inclusión de motores de búsqueda que permitan consultas avanzadas.

1.9.3.2 Lenguaje de estudio

El lenguaje de los estudios primarios será el inglés y serán extraídos mediante consultas en las que las palabras clave están en inglés. El informe de la revisión sistemática se realiza en español.

1.9.3.3 Identificación de fuentes

En este punto se identificaron las fuentes sobre las que se ejecutó la revisión sistemática planificada, definiendo el método de selección de fuentes y la lista de fuentes consideradas, así como estableciendo la cadena de búsqueda que se usará en la ejecución de la revisión.

1.9.3.4 Lista de fuentes

La lista de fuentes obtenida sobre la cual se ejecutó la revisión sistemática es la siguiente:

- IEEE Digital Library

1.9.3.5 Cadenas de búsqueda

Utilizando combinaciones AND sobre las palabras clave y conceptos relacionados que identificados anteriormente, se establece la cadena de búsqueda a utilizar en la presente revisión.

((Data quality) AND database quality)

1.9.3.6 Selección de fuentes después de la evaluación

Tras la ejecución de la revisión sobre las fuentes seleccionadas se realizó una fase de refinado en la que se identificaron e incluyeron los estudios primarios relevantes que a juicio de los expertos son necesarios para completar el estudio y no hayan podido ser recuperados en las búsquedas realizadas.

1.9.3.7 Comprobación de las fuentes

La principal fuente identificada fue IEEE como base para completar el estudio.

1.9.4 Selección de los estudios

Una vez que se han sido definidas las fuentes, es necesario describir el proceso y el criterio que se siguieron en la ejecución de la revisión para la selección y evaluación de los estudios primarios. Para ello se definió el proceso completo de selección, así como los criterios de inclusión y exclusión que se utilizaron.

1.9.4.1 Definición del criterio de inclusión y exclusión de estudios

El criterio de inclusión actúa sobre los resultados obtenidos al ejecutar la búsqueda sobre la fuente, que permitieron realizar una primera selección de documentos que serán considerados en el contexto de la revisión como candidatos a convertirse en estudios primarios.

Como criterio de inclusión se realizó principalmente un análisis sobre el título, las palabras claves y resumen de cada documento, de esta forma se pudo ver en una primera instancia cómo están relacionadas estas palabras y porqué ha sido seleccionado dicho documento. Con este criterio se localizó y eliminó la mayor parte de los resultados obtenidos que no realizaron aportes sobre seguridad dentro de la ingeniería ontológica. El criterio de exclusión actúa sobre el subconjunto de documentos obtenidos en la etapa anterior y permitió obtener el conjunto de estudios primarios.

Como criterio de exclusión se usó principalmente la lectura y análisis del resumen del documento y sus conclusiones, teniendo en algunos casos que profundizar en el mismo y realizar una lectura más detallada sobre otras partes del documento. Con este criterio se pudo ver en más detalle de qué trata cada documento, ver la relación real que presenta con los objetivos buscados y si es verdaderamente relevante para la revisión, seleccionándolo como estudio primario. El criterio de exclusión que también fue añadido fue el periodo de la búsqueda, el mismo se redujo a los últimos dos años.

1.9.4.2 Definición de tipos de estudio

Los tipos de estudios primarios que van a ser seleccionados durante la ejecución de la revisión sistemática son los artículos presentes en las fuentes seleccionadas, que cumplan con los criterios establecidos y hayan sido obtenidos como producto al aplicar el procedimiento de selección.

1.9.5 Ejecución de la revisión

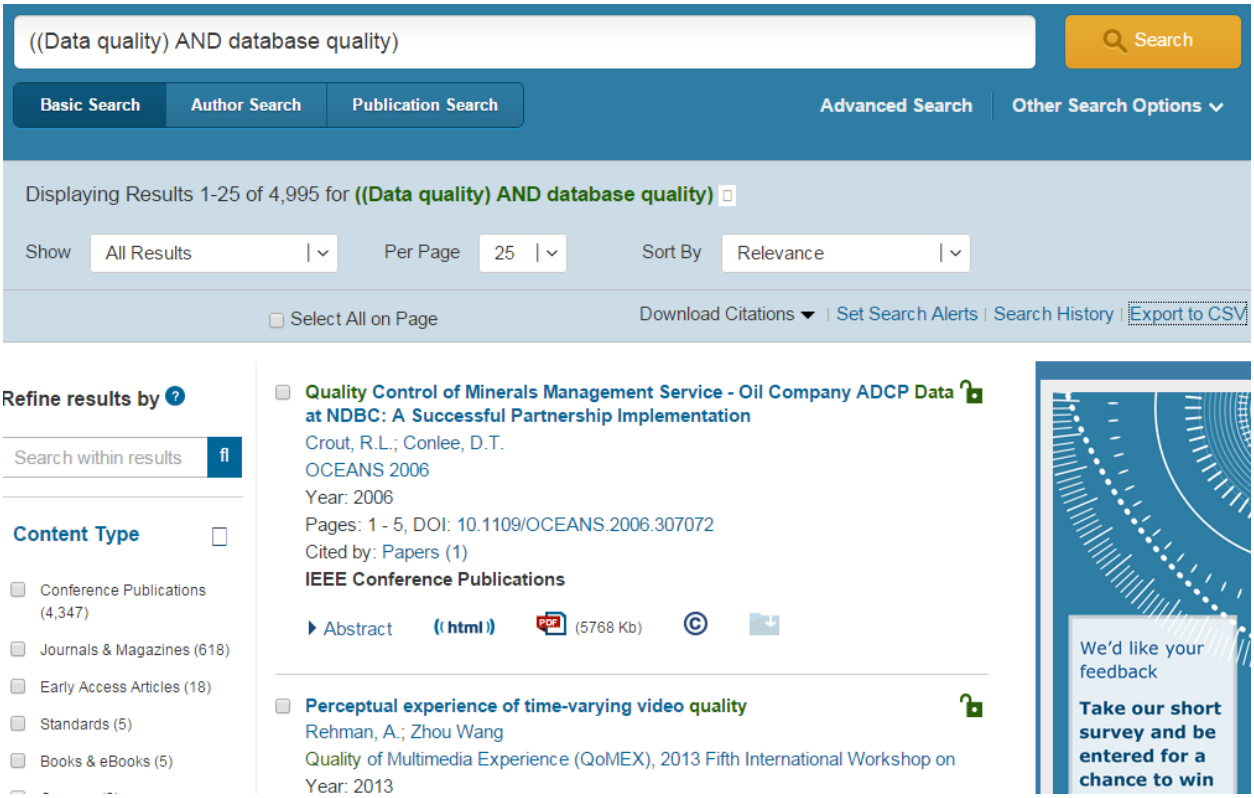
Una vez planificada la revisión, en esta sección se pasa a ejecutar la revisión sistemática en cada una de las fuentes seleccionadas aplicando todos los criterios y procedimientos especificados. A continuación, hay varias secciones en las que se detalla la ejecución en cada una de las fuentes para finalizar con una sección de refinamiento en la que se valida y añaden a juicio de los expertos los estudios importantes que hayan podido quedar atrás en estas búsquedas.

1.9.5.1 Ejecución de la selección en la fuente IEEE

En esta sección se aplican la revisión a IEEE, obteniendo nuevos estudios primarios que complementan la revisión.

1.9.5.2 Selección de estudios iniciales

Al realizar la consulta inicial adaptada al motor de búsqueda y sobre todo el texto, se obtuvo más de 4995 de resultados, los que nos llevó a limitar la misma de forma que la búsqueda final se realizó centrada en el título del artículo e indicando los términos:



Todos los documentos presentes en la fuente IEEE son de alta calidad, pasan por una serie de filtros y evaluaciones antes de ser debidamente publicados, además es conocido que IEEE es una de las fuentes más reconocidas en el campo de la ciencia e ingeniería.

1	ATLAS online data quality monitoring Real Time Conference (RT), 2010 17th IEEE-NPSS Almenar, C.C.; Corso-Radu, A.; Hadavand, H.; Ilchenko, Y.; Kolos, S.; Slagle, K.; Taffard, A.
2	Data quality requirements analysis and modeling Data Engineering, 1993. Proceedings. Ninth International Conference

	Wang, R.Y.; Kon, H.B.; Madnick, S.E.
3	Data quality: The other face of Big Data Data Engineering (ICDE), 2014 IEEE 30th International Conference Saha, B.; Srivastava, D.
4	A Process for Assessing Data Quality Software Testing, Verification and Validation Workshops (ICSTW), 2013 IEEE Sixth International Conference Sneed, H.M.; Demuth, B.; Freitag, B.
5	An extension of the relational model to intuitionistic fuzzy data quality attribute model Intelligent Systems, 2008. IS '08. 4th International IEEE Conference Boyadzhieva, D.; Kolev, B.
6	Data quality: a rising e-business concern IT Professional Paulson, Linda Dailey
7	Tools for data warehouse quality Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference Gebhardt, M.; Jarke, M.; Jeusfeld, M.A.; Quix, C.; Sklorz, S.
8	Ranking for data repairs Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference Yakout, M.; Elmagarmid, A.K.; Neville, J.

9	Detecting inconsistencies in distributed data Data Engineering (ICDE), 2010 IEEE 26th International Conference Wenfei Fan; Geerts, F.; Shuai Ma; Muller, H.
10	Fuzzy measures for data quality Fuzzy Information Processing Society, 1999. NAFIPS. 18th International Conference Janta-Polczynski, M.; Roventa, E.

Cuadro 2. Listado de estudios preliminares en la investigación. Fuente: autoría propia

1.9.5.3 Evaluación de la calidad de los estudios

Todos los documentos presentes en la fuente IEEE son de alta calidad, pasan por una serie de filtros y evaluaciones antes de ser debidamente publicados, además es conocido que IEEE es una la de las fuentes más reconocidas en el campo de la ciencia e ingeniería.

1.9.5.4 Revisión de la selección

La selección de los estudios primarios realizada ha sido validada por los expertos de forma que tengamos la seguridad de no haber dejado atrás ningún estudio relevante en esta fuente.

1.9.5.5 Extracción de resultados objetivos y subjetivos

Identificación 1
Título ATLAS online data quality monitoring
Publicación Real Time Conference (RT), 2010 17th IEEE-NPSS
Autores Almenar, C.C.; Corso-Radu, A.; Hadavand, H.; Ilchenko, Y.; Kolos, S.; Slagle, K.; Taffard, A.
Referencia <Ref>
Descripción

Área Calidad de Datos y Visualización
Palabras clave • data acquisition • data visualization • monitoring

Identificación 2
Título Data quality requirements analysis and modeling
Publicación Data Engineering, 1993. Proceedings. Ninth International Conference
Autores Wang, R.Y.; Kon, H.B.; Madnick, S.E.
Referencia
Descripción
Área Calidad de datos durante el diseño de bases de datos
• Palabras clave data structures • database management systems • query processing • software quality

Identificación 3
Título Data quality: The other face of Big Data
Publicación Data Engineering (ICDE), 2014 IEEE 30th International Conference
Autores Saha, B.; Srivastava, D.
Referencia
Descripción

Área Calidad datos en Big Data
• Palabras clave Big Data • Internet • decision making • quality management

Identificación 4
Título A Process for Assessing Data Quality
Publicación Software Testing, Verification and Validation Workshops (ICSTW), 2013 IEEE Sixth International Conference
Autores Sneed, H.M.; Demuth, B.; Freitag, B.
Referencia
Descripción Auditorias para control de la calidad de datos
Área
• Palabras clave Business Rules • Data Consistence • Data Models • Data Quality • Data Validation • Model-Driven Testing • Normalization • Schema Analysis

Identificación 5
Título An extension of the relational model to intuitionistic fuzzy data quality attribute model
Publicación Intelligent Systems, 2008. IS '08. 4th International IEEE Conference
Autores Boyadzhieva, D.; Kolev, B.
Referencia
Descripción Manejo de datos confusos en DB relacionales
Área Estrategias de medición de la calidad
• Palabras clave data quality • intuitionistic fuzzy • quality model • relational database

Identificación 6
Título Data quality: a rising e-business concern
Publicación IT Professional
Autores Paulson, Linda Dailey
Referencia
Descripción
Área Calidad de datos para DataWarehouse
• Palabras clave business data processing • data handling • data integrity • database management systems • information resources • professional aspects

• quality control

Identificación 7
Titulo Tools for data warehouse quality
Publicación Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference
Autores Gebhardt, M.; Jarke, M.; Jeusfeld, M.A.; Quix, C.; Sklorz, S.
Referencia
Descripción
Área Herramientas para calidad de datos
Palabras clave calidad,datos

Identificación 8
Titulo Ranking for data repairs
Publicación Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference
Autores Yakout, M.; Elmagarmid, A.K.; Neville, J.
Referencia
Descripción
Área Experimentos de calidad de datos en BD relacionales
• Palabras clave data handling • database management systems • software maintenance • user interfaces
Identificación 9
Título Detecting inconsistencies in distributed data

Publicación Data Engineering (ICDE), 2010 IEEE 26th International Conference
Autores Wenfei Fan; Geerts, F.; Shuai Ma; Muller, H.
Referencia
Descripción
Área Inconsistencias en las bases de datos
• Palabras clave SQL • computational complexity • distributed databases • optimization

Identificación 10
Título Fuzzy measures for data quality
Publicación Fuzzy Information Processing Society, 1999. NAFIPS. 18th International Conference
Autores Janta-Polczynski, M.; Roventa, E.
Descripción
Área Lógica manejo de datos confusos
• Palabras clave database management systems • fuzzy logic

Cuadro 3. Extracción de resultados objetivos y subjetivos. Fuente: autoría propia

Capítulo 2.

Marco teórico

2. Marco teórico

Este capítulo constituyó el fondo que apoyó la investigación y ofreció al lector una justificación para el estudio del problema anteriormente señalado.

La descripción de las variables de interés en el contexto de la revisión de la literatura ayudó a comprender las relaciones teóricas. Se iniciará con la descripción de lo que se conoce acerca de sus variables, lo que se conoce acerca de su relación y lo que puede ser explicado hasta el momento. En esencia, su objetivo es transmitir lo que se sabe de sus variables, las relaciones y la inclusión de la investigación y las teorías que apoyan la investigación.

Basado en la revisión sistemática realizada en el capítulo anterior, se logró identificar una serie de variables o términos que son de alta competencia para esta investigación.

Los datos son la estructura más atómica de la cadena en la formación de la información y se definen como los elementos que por sí solos no tienen un sentido o una interpretación específica.

Un número de teléfono o el nombre de alguna persona, no son datos que por sí solos tienen sentido, a menos que se usen juntos.

Los datos pueden ser una colección de hechos almacenados en algún lugar físico como un papel, un dispositivo electrónico (CD, DVD, disco duro...), o la mente de una persona. En este sentido las tecnologías de la información han aportado mucho a recopilación de datos.

Como cabe suponer, los datos pueden provenir de fuentes externas o internas a la organización, pudiendo ser de carácter objetivo o subjetivo, o de tipo cualitativo o cuantitativo.

La información se puede definir como un conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto), y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre.

Al trabajar de manera independiente cada una de las herramientas, los resultados no se pueden comparar o utilizar para tomar mejores decisiones a nivel gerencial, ya que no se ha podido entender la relación real entre ellas, que por su naturaleza financiera, por definición la tienen.

Una investigación realizada por Oracle en 2013, mostró en el “Oracle Technology Day”, un sinfín de ventajas que trae por definición la consolidación de bases de datos y de información per-se para las compañías, como lo es mejoras de desempeño, accesibilidad, seguridad y escalabilidad de las plataformas de información.

Hay muchas definiciones de calidad de los datos, pero los datos generalmente se consideran de alta calidad si "son aptos para los usos previstos en las operaciones, la toma de decisiones y la planificación." (J. M. Juran, sf). Alternativamente, los datos se consideran de alta calidad si representan correctamente la construcción del mundo real al que se refiere. Además, aparte de estas definiciones, a medida que aumenta el volumen de datos, la cuestión de la coherencia interna de datos llega a ser significativa, independientemente de la aptitud para el uso para cualquier propósito externo particular. Las opiniones de la gente sobre la calidad de los datos a menudo pueden estar en desacuerdo, incluso cuando se habla del mismo conjunto de datos que se utilizan para el mismo propósito.

Mientras la calidad de datos fue brevemente definida, también es importante considerar el control de calidad de datos, que es el proceso de controlar el uso de datos con las mediciones de calidad conocidos para una aplicación o un proceso. Este proceso generalmente se realiza después de un proceso de datos de Garantía de Calidad (QA), que consiste en el descubrimiento de la inconsistencia de datos y corrección.

En su libro de implantación de un sistema de calidad, López describe un sistema de gestión de calidad como: “una estructura organizativa, las responsabilidades, los procedimientos, los proceso y los recursos necesarios para llevar a cabo la gestión de la calidad.” (2005)

Cuando se trata de estándares, el Instituto de Manejo de Proyectos (PMI), define un estándar como: “Un documento establecido por consenso, aprobado por un cuerpo reconocido, y que ofrece reglas, guías o características para que se use repetidamente.” (2011)

Enfocando esta definición a la calidad, los estándares no son más que los niveles mínimo y máximo deseados o aceptables de calidad que debe tener el resultado de una acción, una actividad, un programa, o un servicio. En otras palabras, el estándar es la norma técnica que se utilizará como parámetro de evaluación de la calidad.

La creación e implantación de reglas personalizadas, permitirá que se guarden parámetros de manejo de los datos, pudiendo estos ser revisitados de ser necesario, lo que permite flexibilidad a la hora de manejar los cambios en las tendencias de las diferentes herramientas, ya que el cliente genera una regla con un par de clicks y éstas se guardan para su futuro uso y de ser necesario, se modifican.

Una vez programadas las actividades de solución al problema de gestión, los círculos de calidad deberán definir los estándares de calidad del resultado, o los resultados esperados.

Se debe cuidar que los estándares no sean influenciados por lo que actualmente hace el personal, quienes son los responsables de la gestión o ejecución de la actividad, componente o programa con un problema. Los estándares deben ser monitoreados y evaluados periódicamente, aplicando indicadores, para saber si se está asegurando la calidad.

La definición del tipo de organización, en este caso de tipo financiera, ha marcado una limitante en lo que a mejoras de procesos y herramientas respecta, porque se ha enfocado en crear lógicas de precios e índices de cálculo de costos, pero por definición no se permite realizar herramientas de calidad porque el fundamento de la lógica es lo que rige las prioridades de éstas, no su calidad de

ejecución y funcionamiento, dejando esa parte a un lado y generando herramientas con una excelente lógica pero muy básicas y muy homogéneas.

Esta solución pretende no solo resolver los problemas descritos en el apartado de definición del problema sino crear un cambio de mentalidad en lo que a creación de soluciones respecta y demostrar que se pueden realizar soluciones eficientes y rápidas, usando diferentes lenguajes de programación y usando bases de datos funcionales, sin necesidad de invertir recursos de más.

Capítulo 3.

Marco metodológico

4. Marco metodológico

En esta sección lo que se espera es definir de forma satisfactoria el tipo de investigación que se aplicó en este proyecto. La elección del tipo de investigación permitirá no solo eficiencia sino ayudara a que la investigación pueda ser satisfactoriamente replicada en un futuro.

3.1 Tipo de investigación

En este apartado se propone explicar o definir un tipo de investigación a realizar en este proyecto.

Se conocen tres tipos de investigación de tipo formal:

- Pura
- Aplicada
- Evaluativa

Basado en los tipos de investigación dados o descritos como formales que son: pura, la que busca encontrar respuestas o corroborar tesis en base al método científico, la aplicad que pretende resolver un problema específico a un “Cliente” o beneficiario en específico y la evaluativa que lo que hace es probar las capacidades o características de un objeto o método en especial.

Según lo que se busca en esta investigación: “Construcción de modelo para el manejo, calidad de la información y toma de decisiones en WW TS Support Solutions Pricing, organización de Hewlett Packard Enterprise” se puede intuir que por su naturaleza va a ser de tipo APLICADA, ya que lo que se espera es usar los conocimientos aprendidos durante el curso de la maestría de administración de bases de datos en el provecho de la compañía Hewlett Packard, específicamente en la organización de Pricing.

3.2 Alcance investigativo

Para entender mejor los tipos de investigación que se tienen y como se deben elegir, se va a tomar este cuadro comparativo de la publicación de Carballo para el Instituto Tecnológico de Sonora (2013)

	Exploratoria	Descriptiva	Correlacional	Explicativa
Propósito	Examinar un tema o problema de investigación poco estudiado, del cual se tienen muchas dudas o no se ha abordado antes	Describir un fenómeno: especificar propiedades, características y rasgos importantes	Identificar relación o grado de asociación que existe entre dos o más variables en un contexto	Explicar las casusas de relación entre variables (eventos, sucesos o fenómenos)
Utilidad	Familiarizarse sobre fenómenos nuevos o relativamente desconocidos Establecer prioridades para estudios futuros	· Mostar con precisión las dimensiones de un fenómeno.	Predecir el valor de una variable a partir del valor de otra relacionada. Explicar un fenómeno, aunque de manera parcial	Explicar por qué ocurre un fenómeno y en qué condiciones se manifiesta.
Método	Flexibles. Al final identifican conceptos o variables promisorias a estudiar en otra investigación	Identificar el fenómeno y los objetos/sujetos involucrados; definir las variables a medir; recolectar datos para medir las variables; concluir	Identificar variables; establecer hipótesis; medir cada variable; analizar la vinculación entre variables;	Describir y relacionar múltiples variables; explicar por qué se relacionan dichas variables

			probar o no las hipótesis	
Relación con otros estudios	Prepara el terreno para otros estudios (descriptivo, correlacional o explicativo)	Son la base para investigaciones correlacionales	Proporciona la base para llevar a cabo estudios explicativos	Genera un sentido de entendimiento sobre un fenómeno
Amplitud de investigación	Amplia y dispersa	Focalizada a las variables	Focalizada a las variables	Diversas variables (más estructurado)
Meta del investigador	Investigar un problema poco estudiado o desde una perspectiva innovadora	Describir fenómenos, situaciones, contextos y/o eventos	Asociar variables que permita predecir	Determinar la causa de los fenómenos
Riesgo implicado	Alto	Bajo	Obtener relaciones falsas	
Rasgos del investigador	Gran paciencia, serenidad y receptividad	Precisión, ser observador	Análisis	Análisis, ser crítico

Cuadro 4. Alcances de una investigación. Fuente: Carballo (2013)

Según las descripciones de los alcances y analizándoles desde el punto de vista de investigación se encuentra que la más conveniente para este propósito es la Exploratoria, porque luego de realizar investigaciones y lecturas de publicaciones se llega a la conclusión de que la calidad de los datos es un tema de suma importancia, pero no se ha presentado un modelo o estándar de manejo de datos formal para bases de datos, por lo que el concepto de investigación exploratoria propuesto por Carballo (2013) y que define el método investigativo como: “Examinar un tema o problema de investigación poco estudiado, del cual

se tienen muchas dudas o no se ha abordado antes ”, sería lo más cercano al objetivo de esta investigación.

3.3 Enfoque de la investigación

En esta investigación se utiliza un abordaje alternativo a la hora de definir el enfoque. Se parte de las ideas expresadas por autores como Chavarría (2011) y Padrón (1992), que insisten en la necesidad de reconocer que los enfoques cualitativos y cuantitativos no pueden existir de manera separada así que, teniendo la misma perspectiva y visión, se va a proceder de la misma forma.

Es por ello que en este proyecto se explican las dimensiones epistemológica, ontológica y axiológica de la investigación, sin hacer alusión explícita a un enfoque en particular.

En la dimensión epistemológica de la calidad de los datos, se van a considerar los diferentes modelos o marcos teóricos de gestión de la calidad de datos con el fin de entender el fenómeno como tal. Desde este punto se deduce que la calidad debe ser un ciclo sin fin dentro de todo modelo de datos, por lo que se deberán reconocer esas partes críticas considerando los diferentes marcos y metodologías actualmente identificadas como formales.

Para describir la dimensión ontológica, se parte del concepto expresado por Gruber (1993): "lo que existe es exactamente aquello que puede ser representado". La representación del conocimiento requiere de mecanismos formales, por eso se analizará la calidad desde muchos puntos de vista, representados en la figura 1.

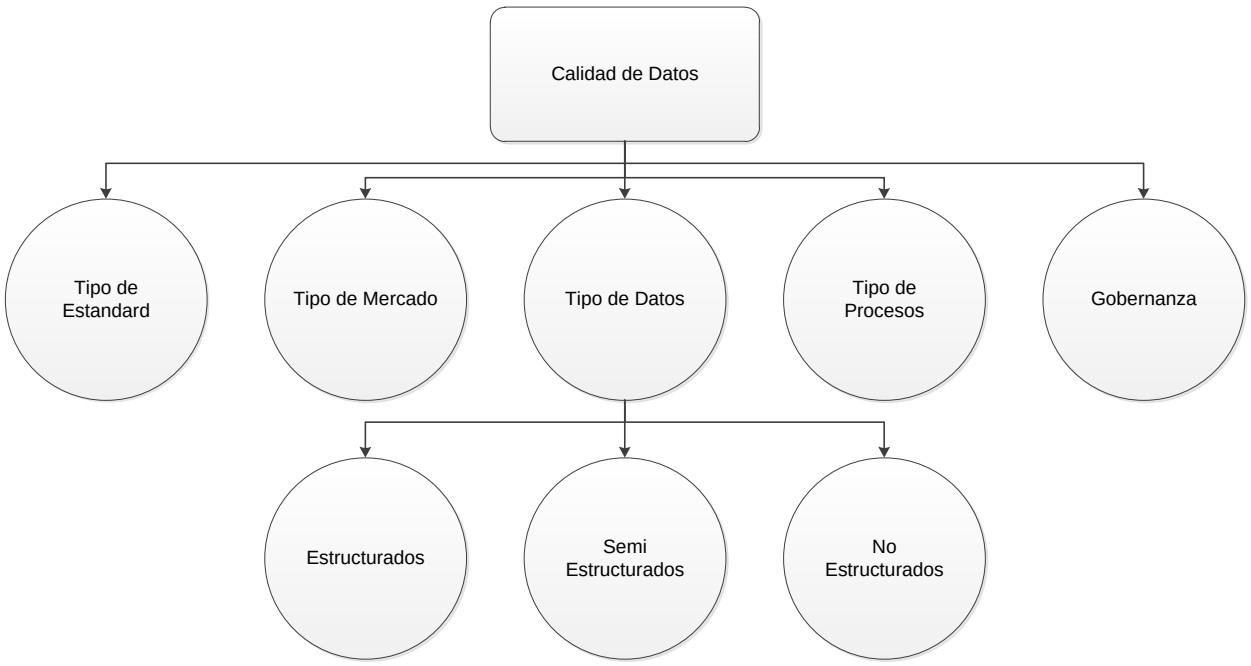


Figura 1. Calidad de datos desde dimensión ontológica. Fuente: autoria propia

La dimensión axiológica se enfocó en criterios considerados de alta importancia, como la eficacia, la eficiencia y la facilidad de implementación y mantenimiento del modelo a elegir.

Descripción	**Valor**	**Opción 1**	**Opción 2**	**Opción 3**
Granularidad: que tan granular es el modelo	45%	El modelo no describe de forma clara el marco de implementación ni especifica un manejo claro y detallado de los datos 15%	El modelo describe de forma clara el marco de implementación, pero no especifica un manejo claro y detallado de los datos 30%	El modelo describe de forma clara el marco de implementación, además especifica un manejo claro y detallado de los datos 45%

Documentación: que tanta riqueza y calidad de documentación se encuentra para facilitar la implementación y mantenimiento	30%	No se encuentra documento formal del modelo más que implementaciones realizados por terceros 10%	Se encuentra un documento formal de implementación pero no describe en su totalidad la forma de implementarse 20%	Se encuentra un documento formal de implementación, donde se describe en su totalidad la forma de implementarse 30%
Implementación: cuanta facilidad o elasticidad de implementación presenta ante diferentes ambientes, mercados o políticas	15%	Diseñado solo para un mercado o un tipo de cliente en específico 5%	Diseñado solo para dos tipos de mercado en específico 10%	Diseñado para cualquier mercado 15%
Ciclo de vida: que tan manejable y amigable es el ciclo de vida.	10%	Ciclo de vida no cubre en su totalidad la estructura de calidad requerida 5%	Ciclo de vida cubre en su totalidad la estructura de calidad requerida 10%	

Cuadro 5. Rúbrica de dimensión ontológica. Fuente: autoría propia

Entre mayor sea el porcentaje asignado, más fácil de implementar es el modelo de datos. Se considerará que un 80% o más significan que el modelo es fácil de implementar, de un 50% a un 80% significa que es medianamente complejo de implementar y menos de un 50% significa que es de implementación muy compleja.

3.4 Diseño

En términos generales, el diseño metodológico responde a la pregunta de cómo se debe proceder en la evaluación y planea lo que se debe hacer. Es una metodología puesto que establece un modelo, un camino y una manera de proceder.

El diseño incluye unas áreas técnicas que, según Stufflebeam y Shinkfiel (1995:25), deben conocer a fondo los evaluadores profesionales. Se encuentran las siguientes:

Las entrevistas, los informes preliminares, el análisis de contenidos, la observación, el análisis político, el análisis económico, el examen investigativo, los informes técnicos, el estudio de caso, la evaluación sin metas, la escucha de opiniones contrapuestas, las asociaciones de defensa, las listas de control, la elaboración de pruebas, el análisis estadístico, el diseño de la investigación, el análisis sistemático, la teorización y la administración del proyecto.

Tomando la aportación de Martínez Mediano (1996: 45): “El propósito de la investigación aplicada es contribuir al conocimiento que ayude a la gente a comprender la naturaleza de un problema de modo que los hombres puedan controlar de un modo eficaz su ambiente”.

Con base en estos supuestos y consideraciones, se realizarán evaluaciones por medio de:

- Análisis de informes y documentos.
- Análisis y colección de conclusiones funcionales por juicio de expertos usando herramientas como:
    - Diagramas de Ishikawa
    - Lluvia de ideas
    - Round Tables con expertos y usuarios
- Administración del proyecto

3.5 Población y muestreo

Según la página web statistics.com, que es el ente referente en lo que respecta a educación sobre estadística, la población, se define como: “un gran conjunto de objetos de la misma naturaleza, por ejemplo, los seres humanos, los hogares, las lecturas de un aparato de medición, que son de interés en su conjunto. Un concepto relacionado es una muestra, un subconjunto de objetos se extrae de una población.” (2013)

Como población en este caso específico, se considera la totalidad de las herramientas que están dentro del marco de la implementación:

- DC Configuration Tool
- Indico

De acuerdo a Hunt (2001) un muestreo: “Es una forma de representación estadística que muestra cómo se comporta una característica o variable en una población a través de hacer evidente el cambio de dicha variable en sub-poblaciones o estratos. Consiste en la división previa de la población de estudio en grupos o clases que se suponen homogéneos respecto a característica a estudiar y que no se solapen” (2001)

Como muestreo para propósitos de implementación y pruebas, se trabajará con la herramienta DC Config Tool, que se podría considerar de tipo estratificado no al azar porque se pueden clasificar las diferentes herramientas que conforman la población y se va a usar esa de forma adrede, por su complejidad y porque es de la que más detalles se conocen.

3.6 Instrumentos de recolección de datos

Debido a que la solución esperada luego de la investigación es de tipo aplicada, gran parte de la solución será propuesta por un conjunto de especialistas en diferentes áreas del conocimiento como lo son:

- Especialistas en bases de datos
- Especialistas en estándares de calidad
- Especialistas en administración de negocios
- Especialistas en sistemas de información

Éstos reunieron y analizaron los diferentes puntos de vista en una sesión de grupo focal, que generará importantes conclusiones de cómo diseñar el marco de la calidad que se espera en este caso específico y de ser necesario, generar un marco estándar de calidad personalizado.

Según el Departamento de Salud y Servicios Humanos del Gobierno de los Estados Unidos, un grupo focal se define como: “Una discusión moderada que implica típicamente 5 a 10 participantes. A través de un grupo de enfoque, usted

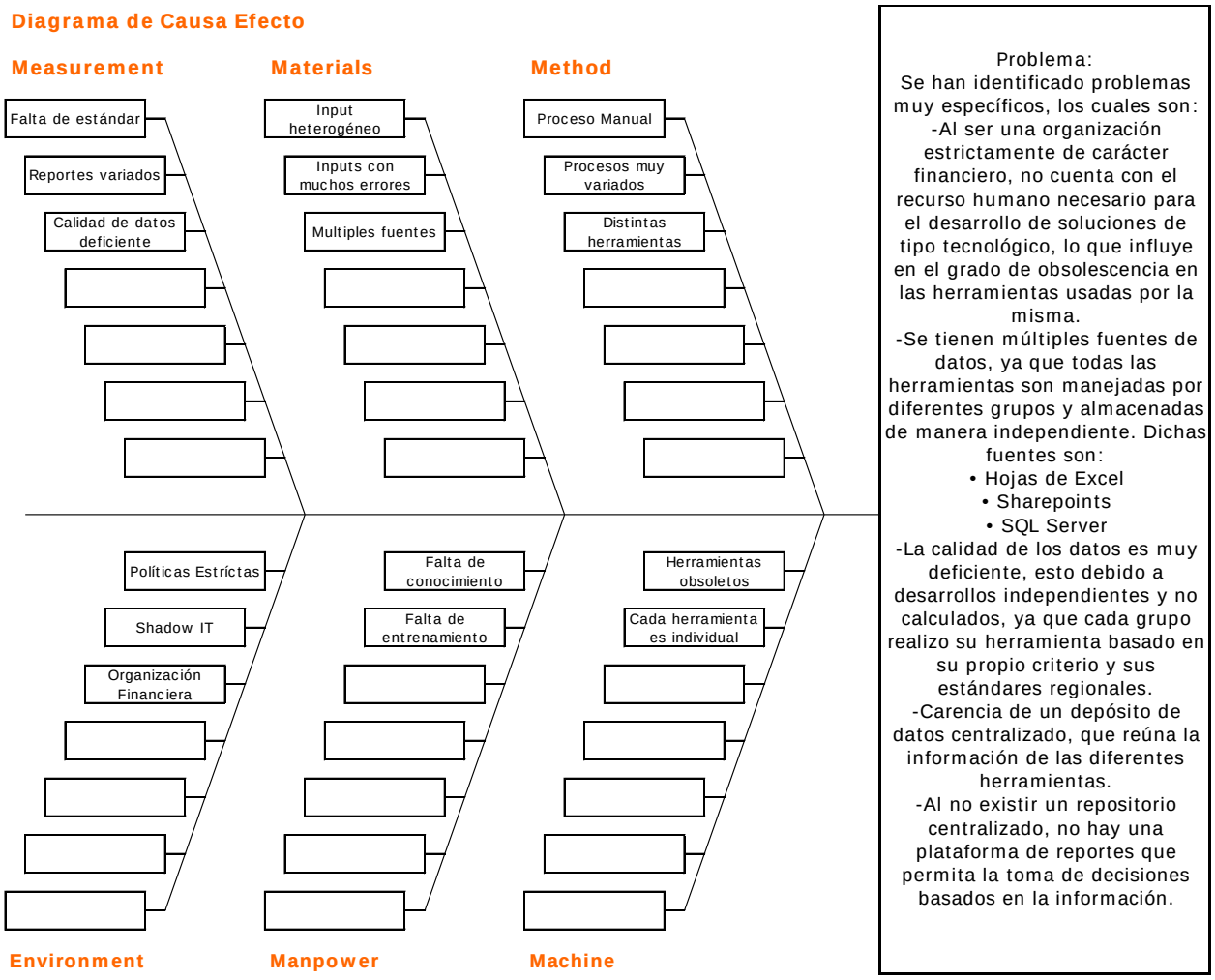
puede aprender acerca de los usuarios actitudes, creencias, deseos y reacciones a los conceptos.”

3.7 Técnicas de análisis de la información

Las diferentes técnicas para analizar los datos o lo resultados de los análisis van a variar dependiendo del tipo de análisis que se va a realizar. Entre las herramientas seleccionadas para esta investigación, se usa el Diagrama de Ishikawa (cc: diagrama de espina de pescado, diagrama de causa-efecto, diagrama causal)

Según Simon (2000), este diagrama permite analizar las causas de un efecto esperado o conocido, por medio del análisis de sus diferentes factores, en inglés conocidos como las 6 Emes:

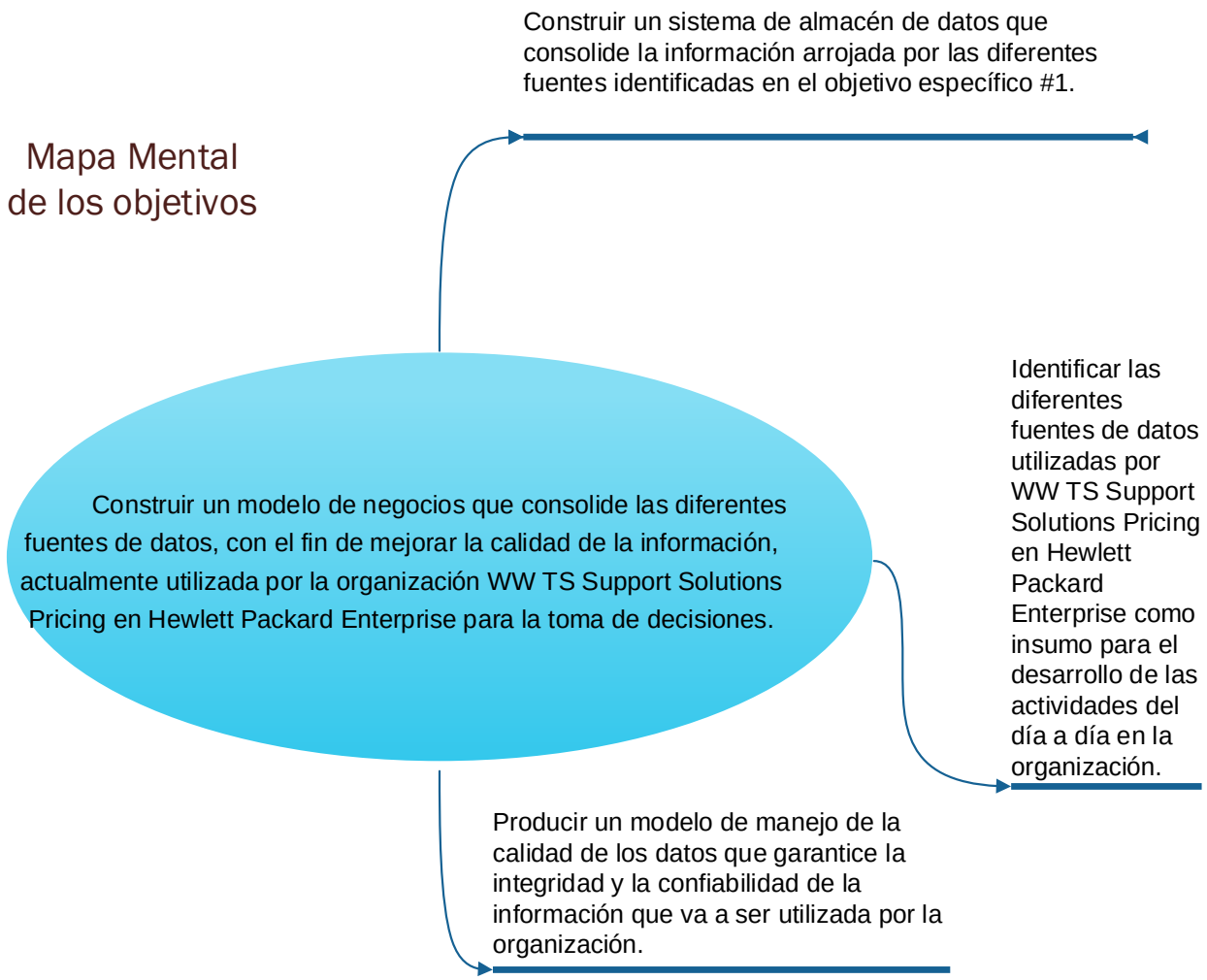
- Machine, Method, Material, Measurement, Mother Nature (Enviroment) y Manpower



De acuerdo a Buzan (2013): “Un mapa mental es un diagrama utilizado para organizar visualmente la información, a menudo se crea en torno a un solo concepto, elaborado como una imagen en el centro de una página horizontal en blanco, al que se añaden representaciones asociadas de ideas como imágenes, palabras y partes de palabras.

Las ideas principales están conectadas directamente con el concepto central, y otras ideas se ramifican de aquellos.

Los mapas mentales se pueden dibujar a mano, ya sea como "notas" durante una conferencia, reunión o sesión de planificación, por ejemplo, o como imágenes de mayor calidad cuando hay más tiempo disponible.”



Fuente: autoría propia.

3.8 Estrategia de desarrollo de la propuesta

De acuerdo a Mcclendon (2006): “La creación de prototipos de software es la actividad de creación de prototipos de aplicaciones de software, es decir, versiones incompletas del programa de software están desarrollando.”

Según el tipo de trabajo que se realizó, un prototipo fue la mejor opción en este caso en particular y se diseñó un modelo en DC Config tool también en Indico que puede ser replicable en todas las demás herramientas.

Capítulo 4.

Análisis del diagnostico

4. Análisis de diagnóstico

4.1 Análisis del negocio

Es esta etapa del proyecto, luego de analizar la situación actual de la operación dentro de la organización, es evidente la necesidad de realizar una solución que mitigue la carga laboral que genera esta tarea.

Se determinó que se requirieron 4 personas por 4 días laborales para generar reportes a partir de los datos crudos. Esto aunado al hecho de que no se realizaba ningún tipo de análisis de los datos una vez colectados fue una brecha de calidad de datos más concisa.

Estos procesos son clave para la organización, con esta información se crean los marcos de referencias para actualizar los precios de los servicios provistos, lo que levanta una bandera de alerta en estos procesos y la necesidad de que sean datos con altos niveles de calidad.

Se pudo determinar, con base en lo descrito anteriormente, que la necesidad de realizar el análisis de calidad es altamente requerida, pero había sido ignorado por la organización hasta la implementación de esta solución.

4.2 Análisis de las fuentes

Dado que los documentos que conforman los datos crudos son documentos de carácter legal, no deben ser corregidos o modificados ni parcial ni totalmente, por lo que el análisis de calidad se realizó en una copia de la base de datos basada en los datos crudos.

Capítulo 5.

Propuesta de solución

5.1 Selección de los modelos

Según el modelo propuesto en la sección 3 del capítulo 3, los modelos elegidos para ser analizados fueron:

Descripción	Valor	ISO 9001:2008 Sistemas de gestión de la calidad	Data quality concepts methodologies and techniques	Modelo de Calidad Aternativo
Granularidad: Que tan granular es el modelo	45%	El modelo no describe de forma clara el marco de implementación ni especifica un manejo claro y detallado de los datos 15%	El modelo no describe de forma clara el marco de implementación ni especifica un manejo claro y detallado de los datos 15%	El modelo describe de forma clara el marco de implementación, además especifica un manejo claro y detallado de los datos 45%
Documentación: Que tanta riqueza y calidad de documentación se encuentra para facilitar la implementación y mantenimiento	30%	No se encuentra documento formal del modelo más que implementaciones realizados por terceros 10%	Se encuentra un documento formal de implementación pero no describe en su totalidad la forma de implementarse 20%	Se encuentra un documento formal de implementación, donde se describe en su totalidad la forma de implementarse 30%
Implementación: Cuanta facilidad o elasticidad de implementación presenta ante diferentes ambientes, mercados o políticas	15%	Diseñado solo para un mercado o un tipo de cliente en específico 5%	Diseñado solo para dos tipos de mercado en específico 10%	Diseñado para cualquier mercado 15%
Ciclo de vida: Que tan manejable y amigable es el ciclo de vida.	10%	Ciclo de vida no cubre en su totalidad la estructura de calidad requerida 5%	Ciclo de vida no cubre en su totalidad la estructura de calidad requerida 5%	Ciclo de vida cubre en su totalidad la estructura de calidad requerida 10%
Calificación	100%	35%	50%	100%

Al analizar los modelos elegidos y siguiendo la rúbrica establecida en el modelo ontológico, se encontró más factible la realización de un modelo alternativo de calidad que fue la opción más viable basándose en la necesidad que se quiere sopesar.

Este modelo es una combinación de los dos primeros modelos, Sistemas de gestión de calidad propuesto por ISO 9001:2008 y el Modelo de Conceptos, Metodologías y Técnicas de Calidad de datos propuestas por Batini en su libro.

5.2. Análisis comparativo de los modelos elegidos

A continuación, se realizará una breve descripción de cada modelo.

5.2.1 ISO 9001:2008

El siguiente es un resumen del manual de calidad ISO 9001:2008 donde se compilan los puntos más importantes que aplican a un modelo de calidad basado en este estándar.

Enfoque

El enfoque principal es aumentar la satisfacción del cliente mediante el cumplimiento de los requisitos a la hora de desarrollar, implementar y mejorar la eficacia de un sistema, en este caso para la evaluación de la calidad de los datos que se generan de las herramientas a evaluar.

Requisitos de documentación

-Generalidades

- Declaraciones de una política de calidad y objetivos de calidad
- Manual de calidad
- Procedimientos documentados
- Los documentos, registros de operación y control de los procesos

-Manual de calidad

- Alcance de sistema de gestión
- Procedimientos documentales

Control de documentos

- Aprobación
- Revisión
- Control de cambios
- Manejo de versiones
- Accesibilidad
- Identificar los documentos externos
- Actualización y control de obsoletos

Planificación con la dirección

- Realizar homologación de la calidad con los objetivos gerenciales
- Planificación del sistema de gestión de calidad
- Control y planificación de la actualización

Revisión

- Realizar auditorias
- Retroalimentación del cliente
- Controles de desempeño de los procesos
- CAPA
- Control de seguimiento de CAPA
- Recomendaciones y posibles cambios
- Control de los resultados para medir la mejora de los procesos basados en los cambios que se han realizado
- El control de la mejora de los productos o servicios que se realizan

Infraestructura

- Se debe determinar, proporcionar y mantener la infraestructura necesaria para lograr la conformidad con los requisitos del producto

Realización del producto:

- Planificación
  - Objetivos y requisitos de calidad
  - Necesidad de establecer procesos y documentos con sus respectivos recursos
  - Todos los controles requeridos para realizar la revisión
  - Todos los registros que evidencian la satisfacción de la realización de los productos
- Cliente
  - Requisitos establecidos por el cliente
  - Especificaciones
  - Estándares de calidad
  - Plazos de entrega
  - Legales

- Diseño y desarrollo
    - La revisión, verificación y validación apropiadas para cada etapa del diseño y desarrollo
    - Definición de responsabilidades y autorización
    - Definición de elementos de entrada y salida cada etapa del desarrollo
    - Resultados de diseño y desarrollo
    - Definición de la revisión del diseño y desarrollo
        - Control de cambios del diseño y desarrollo
        - Validación del diseño y desarrollo

Medición, análisis y mejora

- Satisfacción del cliente
- Auditorías internas
    - Medición de procesos
    - Productos
    - Control de productos no conformes
- Análisis de datos
- Mejora continua
    - CAPA

Como se mostró en el anterior resumen, es un modelo muy bien estructurado, sin embargo, se notaron varias áreas de mejora, las que son:

- Granularidad: este modelo no tiene un nivel de granularidad requerido, ya que no explica el cómo usar el modelo para llegar directamente a los datos.
- Documentación: el modelo permite ser aplicado, pero no existe evidencia de cómo aplicarse en este caso específico, debido a lo complejo de los datos que se manejan en esta organización.
- Implementación: la finalidad de esta implementación es proveer un modelo poco flexible, lo que lo hace difícil de utilizar en casos muy específicos como lo es este.

- Ciclo de vida: según ISO, el modelo no propone el ciclo de vida especificado según las necesidades de la organización, el cual es el cambio continuo.

5.2.2 Modelo de conceptos, metodologías y técnicas de calidad de datos propuesto por Batini

El modelo de Carlo Batini y Monica Scannapieca, será analizado por medio de flujos de información.

Figura 2. Flujo general de datos:

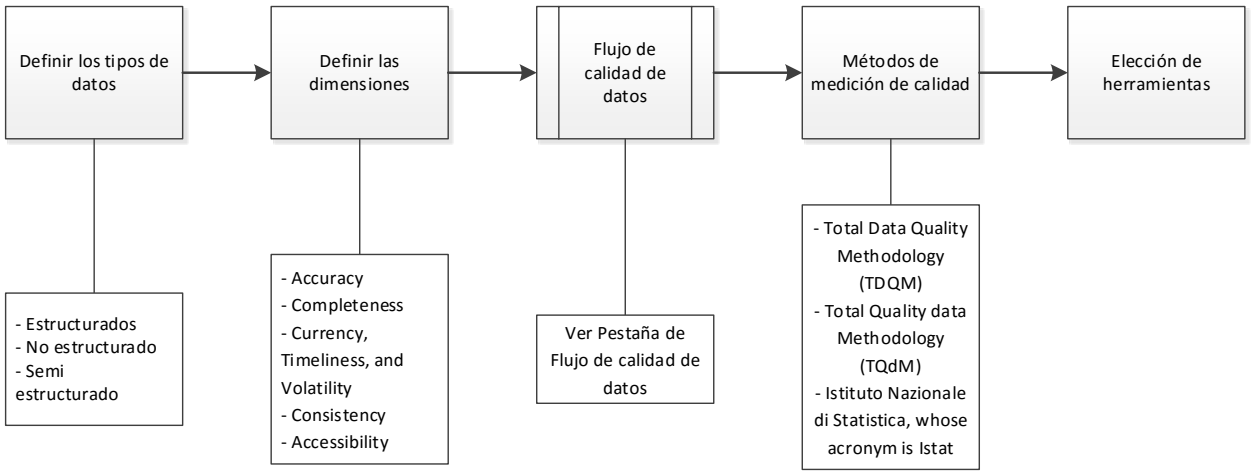


Figura 2. Flujo de datos. Fuente: autoría propia

Figura 3. Flujograma:

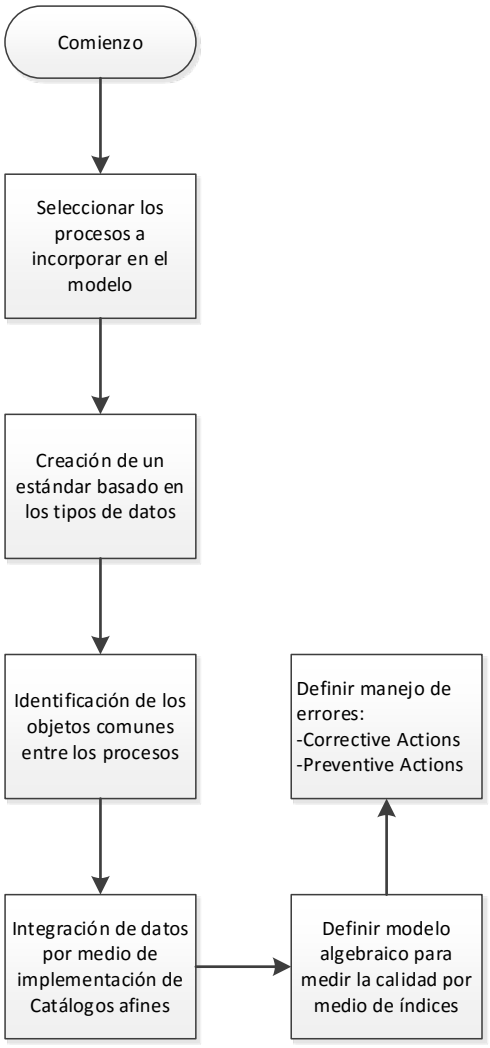


Figura 3. Flujograma lógico. Fuente: autoría propia

Las dimensiones de calidad propuestas por Batini:

Dimension Name	Type of dimension	Definition
Accuracy	data value	Distance between v and v', considered as correct
Completeness	data value	Degree to which values are present in a data collection
Currency	data value	Degree to which a datum is up-to-date
Consistency	data value	Coherence of the same datum, represented in multiple copies, or different data to respect integrity constraints and rules
Appropriateness	data format	One format is more appropriate than another if it is more suited to user needs
Interpretability	data format	Ability of the user to interpret correctly values from their format
Portability	data format	The format can be applied to as a wide set of situations as possible
Format precision	data format	Ability to distinguish between elements in the domain that must be distinguished by users
Format flexibility	data format	Changes in user needs and recording medium can be easily accommodated
Ability to represent null values	data format	Ability to distinguish neatly (without ambiguities) null and default values from applicable values of the domain
Efficient use of memory	data format	Efficiency in the physical representation. An icon is less efficient than a code
Representation consistency	data format	Coherence of physical instances of data with their formats

Cuadro 6. Dimensiones de calidad. Fuente: autoría propia

Como se aprecia, este es un modelo bastante completo y muy efectivo en el uso de calidad de datos, sin embargo, no se usó en este proyecto por las siguientes razones:

Implementación: el modelo está muy bien descrito en la teoría, pero falta una perspectiva más eficiente de implantación, lo cual redujo en gran medida la claridad de la aplicación en la práctica, dejando serias dudas al respecto.

Ciclo de vida: el ciclo de vida es una seria falla en este modelo, ya que no se aclara ni se explica nada concreto al respecto.

5.3 Flujo de la solución

El proceso como se aplicaba es sencillo pero muy manual y lleno de faltantes:

Figura 4. Proceso aplicado antes

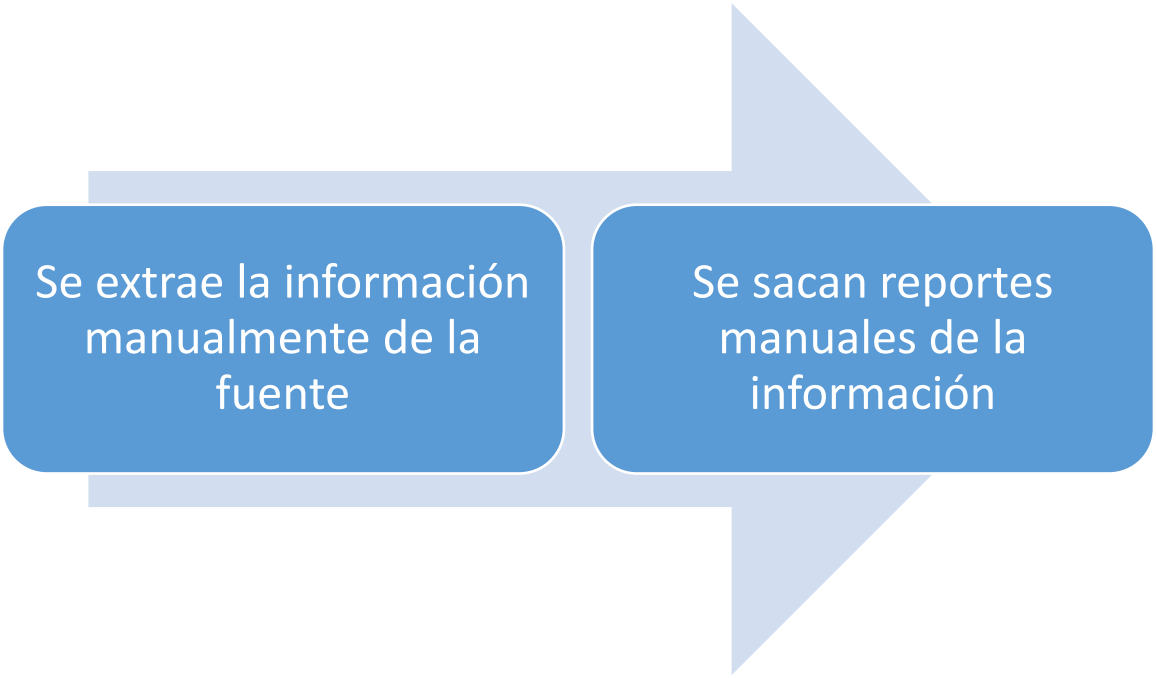


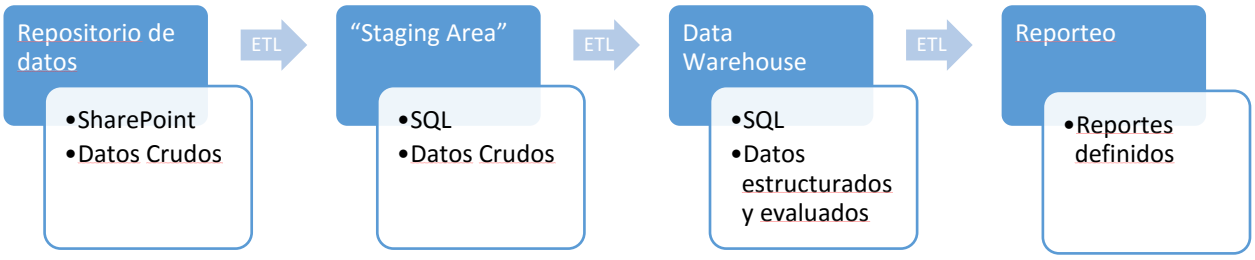
Figura 4. Proceso aplicado antes. Fuente: autoría propia

Se aclara que el proceso tiene deficiencias en lo que respecta a:

- Estandarización del proceso de extracción de datos, ya que por ser manual, no existe un método, lo cual hace que cada extracción sea "única" y personalizada basada las condiciones dadas.
- No existía un orden específico ni un control de calidad alguno en la extracción de datos, los datos extraídos y la creación de los reportes.
- No existía un análisis de calidad de los datos extraídos.

El modelo propuesto se definió de la siguiente manera:

Figura 5. Modelo propuesto. Fuente: autoría propia



En este modelo, la información cruda es extraída a un "Staging Area" (donde esta va a permanecer cruda) por medio de herramientas ETL´s (Extraer-Transformar-Cargar) para la extracción de las diferentes fuentes.

Una vez en el "Staging Area" la información es llevada a un Data Warehouse para su posterior limpieza y análisis contra un catálogo de datos previamente creado y depurado para este fin.

Los análisis de calidad previos son tomados del "Staging Area" y comparados con el catálogo, para arrojar el porcentaje de CALIDAD INICIAL, el cual se comparará con los datos de CALIDAD FINAL, que son el resultado de la comparación de los datos corregidos en el DW con el catálogo.

Una vez los datos son corregidos y estructurados en el DW, estos son extraídos en forma de reportes estandarizados para su análisis.

5.3.1 Elección de las herramientas

Se intentó usar las herramientas de mercado, como por ejemplo SSIS para la creación de los ETL´s, pero el nivel de dificultad del manejo de los datos era tal, que fue

necesario utilizar herramientas ETL´s personalizados, programados en varios lenguajes como lo son TRANS-SQL y Scripts de Microsoft.

En el proceso de aplicación de la calidad, se probaron herramientas como Quality Services de Microsoft, pero una vez más, no fueron tan eficaces debido al detalle y uso de los datos que se buscaba, por lo que fue necesario realizar la aplicación de calidad en .NET, misma que será explicada posteriormente.

5.3.2 Desarrollo de la propuesta

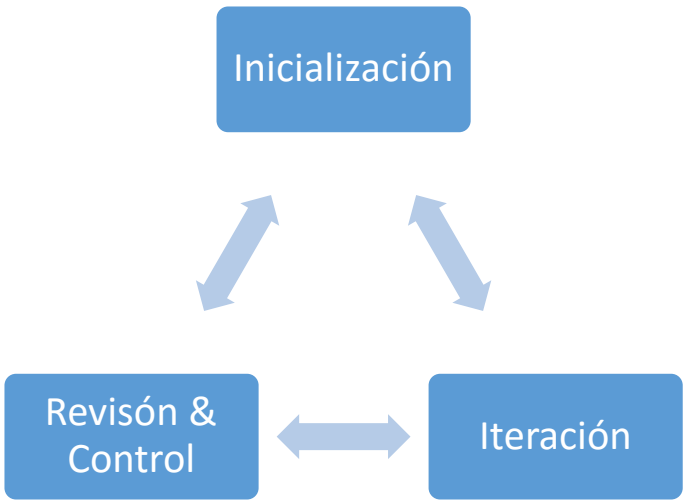
Ciclo de desarrollo

Inicialmente, se creó una propuesta de desarrollo en iteraciones, que agruparon tareas específicas en cada uno de los modelos entregados.

El modelo consta de tres etapas:

- Inicialización, en esta etapa se crea un listado de las metas a entregar basado en los problemas encontrados. Se realiza un plan de cómo desarrollar un prototipo.
- Iteración, es la etapa en la cual se implementan los cambios propuestos a la herramienta, listados en la etapa de iniciación.
- Revisión & Control, se refiere al control de mejoras esperado, revisado contra las mejoras implementadas, dejando para la siguiente iteración, los cambios que no se pudieron implementar en esta, además de listar los nuevos cambios que alimentaran la etapa de inicialización de la siguiente iteración.

Figura.6 Modelo en iteraciones



5.3.3 Modelo de iteración

5.3.4 Recolección de requerimientos

La recolección de los requerimientos fue realizada por medio de entrevistas con el usuario final, que arrojaron un listado importante de problemas, listados en la sección de definición del problema; los problemas fueron controlados y considerados a la hora de realizar el modelo final.

5.3.5 Desarrollo de aplicación final

En la versión final realizada se determinó la necesidad de un grupo de herramientas que cumplieran los requisitos necesarios para llevar a cabo las tareas. La lista de aplicaciones será descrita posteriormente.

Pruebas aplicación final: cada iteración que se realizó contó con un grupo de pruebas básico:

- Pruebas de funcionalidad, para asegurar que realice las tareas esperadas
- Pruebas de estrés, para probar como se desempeña en un ambiente atípico
- Pruebas de calidad, que confirman que la herramienta arroje los resultados consistentes

5.3.6 Análisis de resultados

Los resultados de cada iteración se convirtieron en el insumo de la siguiente iteración, resultando en un modelo progresivo y debidamente probado, buscando siempre solventar los problemas que se definieron en la sección de definición del problema.

5.3.7 Implementación de la aplicación

La versión final de la aplicación se realizó en un ambiente controlado, en una máquina virtual en AZURE, usando como Sistema Operativo Windows Server 2012 R2, que también fue probada en diversas máquinas físicas para descartar errores.

Sección de iteraciones de la etapa de desarrollo

Las iteraciones realizadas fueron las siguientes:

1. Implementación de herramienta de configuración preliminar de los insumos correspondientes
2. Implementación de ETL para la extracción de datos crudos, empleando filtros de selección
3. Implementación de ETL de traslado de datos al DW
4. Creación llenado y mantenimiento del catálogo
5. Aplicación de algoritmo de calidad de datos
6. Creación de mantenimiento de catálogos
7. Creación de reportes específicos
8. Creación de herramienta orquestadora de control de las herramientas requeridas

5.3.8 Aplicaciones

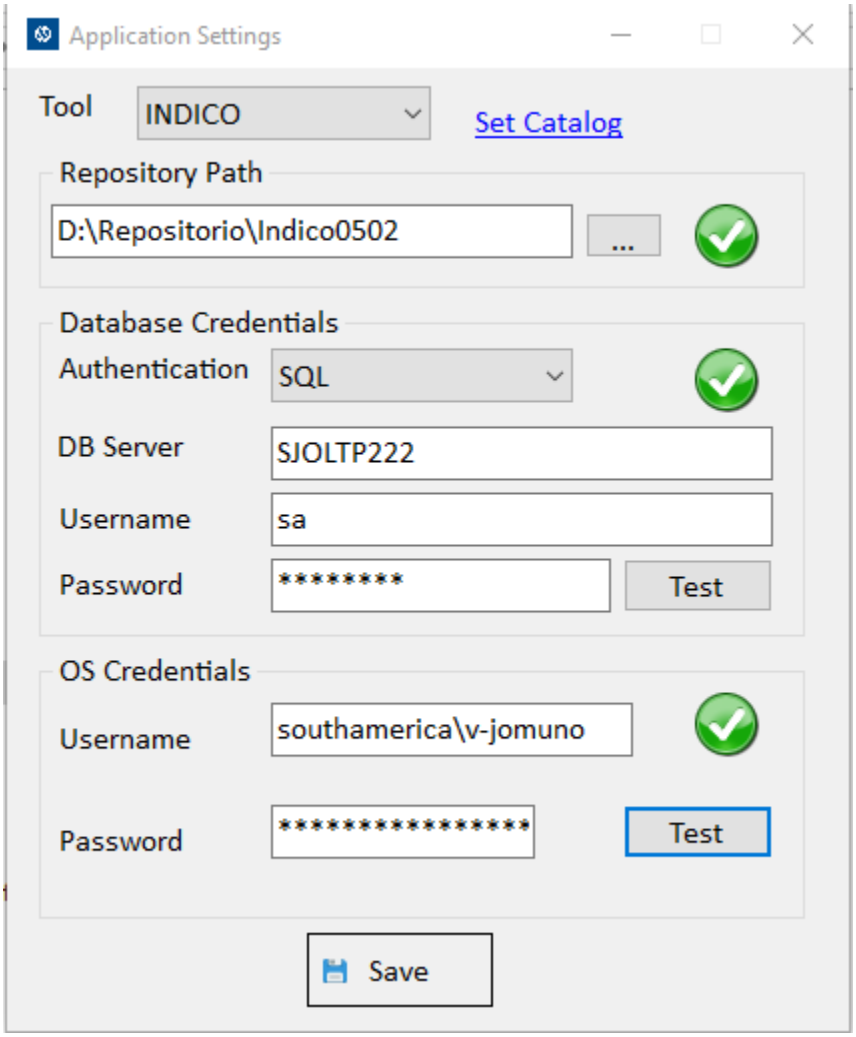
Descripción de cada aplicación:

1. Configuración del usuario y el acceso a las rutas

Esta herramienta busca colectar todos los detalles y confirmar accesos a las herramientas requeridas. Se utiliza para definir parámetros necesarios para ajustes de la aplicación. Fue desarrollada en .NET.

Nombre Aplicación: App_Settings.exe

Figura 7. App_Settings.exe



Sus opciones son:

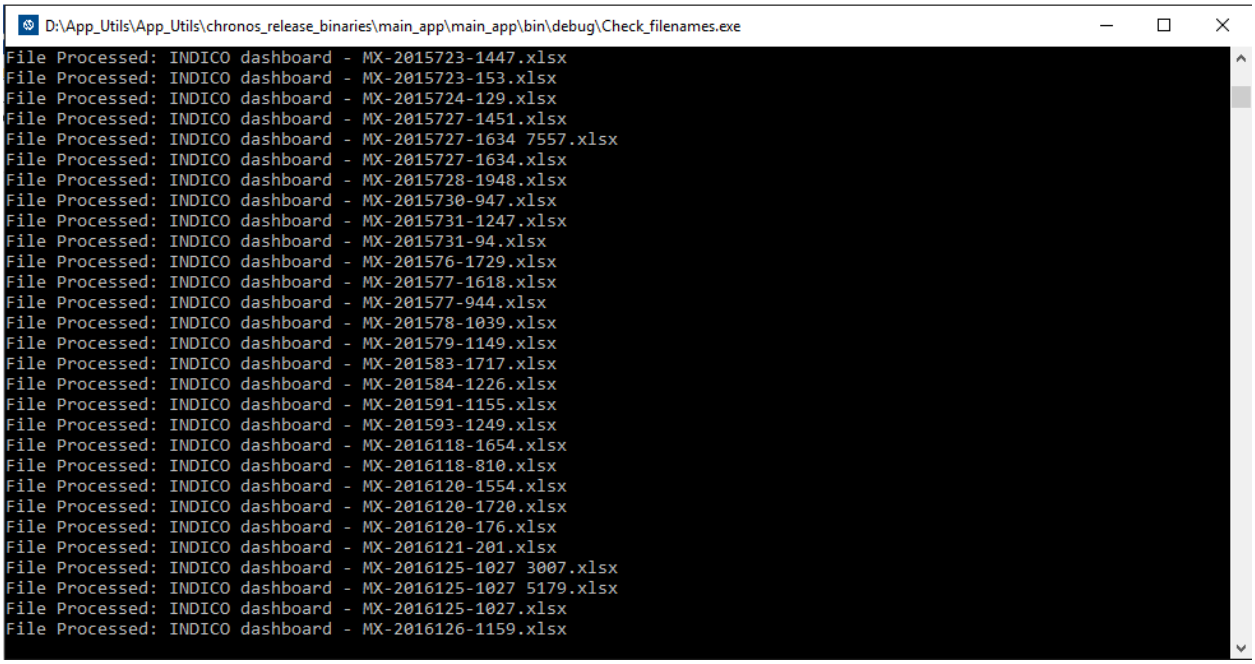
- Producto: se debe elegir entre Indico o DC_Config,los dos tipos de productos disponibles
- Repositorio: folder donde se encuentran los contratos que han de ser procesados por la herramienta. Debe estar de ser posible en una ruta local del servidor de bases datos.
- Catalogo: es un listado de los contratos a ser procesados, contiene información relevante tal como autor del archivo y fecha modificación. Únicamente disponible para Indico
- Servidor Bases Datos: parámetros de conexión para base datos
- Usuario Sistema Operativo: usuario y contraseña sistema operativo para ejecución de procesos

2. Línea de comando revisa nombres de archivos, elimina apostrofe

   Esta aplicación se encarga de realizar los filtros respectivos de información, previos a las corridas, basándose en los parámetros provistos por la organización, estos fueron realizados en .Net, que a su vez ejecutan Trans-SQL scripts para realización de tareas específicas en los datos.

   Nombre Aplicación: Check_filenames.exe

Figura 8 Check_filenames.exe

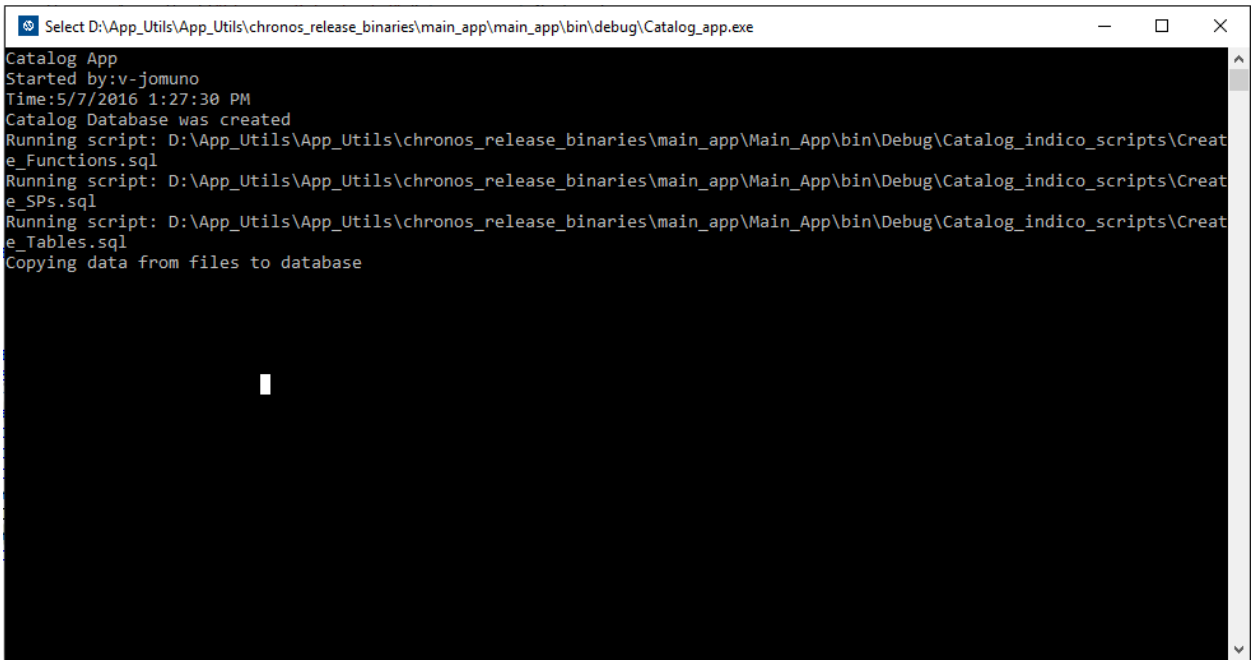


La función de este proceso es el de revisar la consistencia de los nombres de archivo de los contratos en Excel. Esto especialmente para detectar y corregir algunos nombres de archivo que contienen apostrofe y los cuales son rechazados por SQL Server.

3. Llenado de catálogo

Nombre aplicación: Catalog_app.exe

Figura 9. Catalog_app.exe

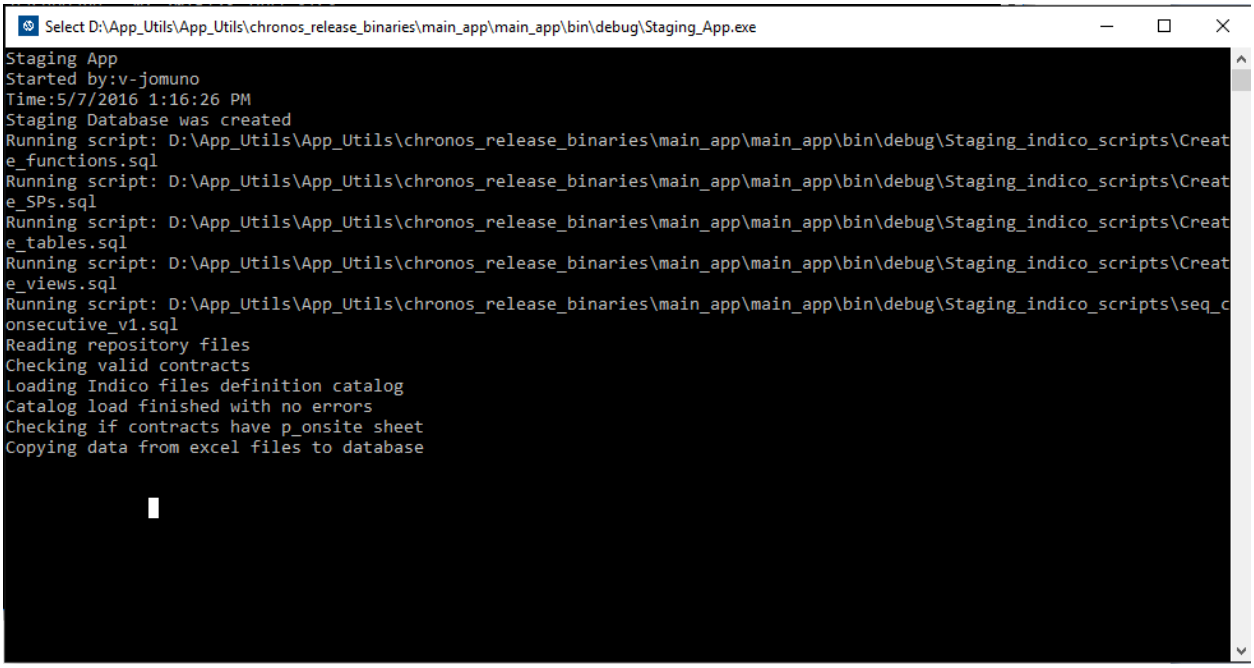


Aplicación que se ejecuta únicamente una vez, contiene una serie de valores prefijados por el cliente los cuales serán cotejados contra la información en el DW para evaluar la calidad de datos y de paso corregirlos para alcanzar una definición única

4. ETL Trans-SQL

   Nombre aplicación: staging_app.exe

Figura 10. staging_app.exe

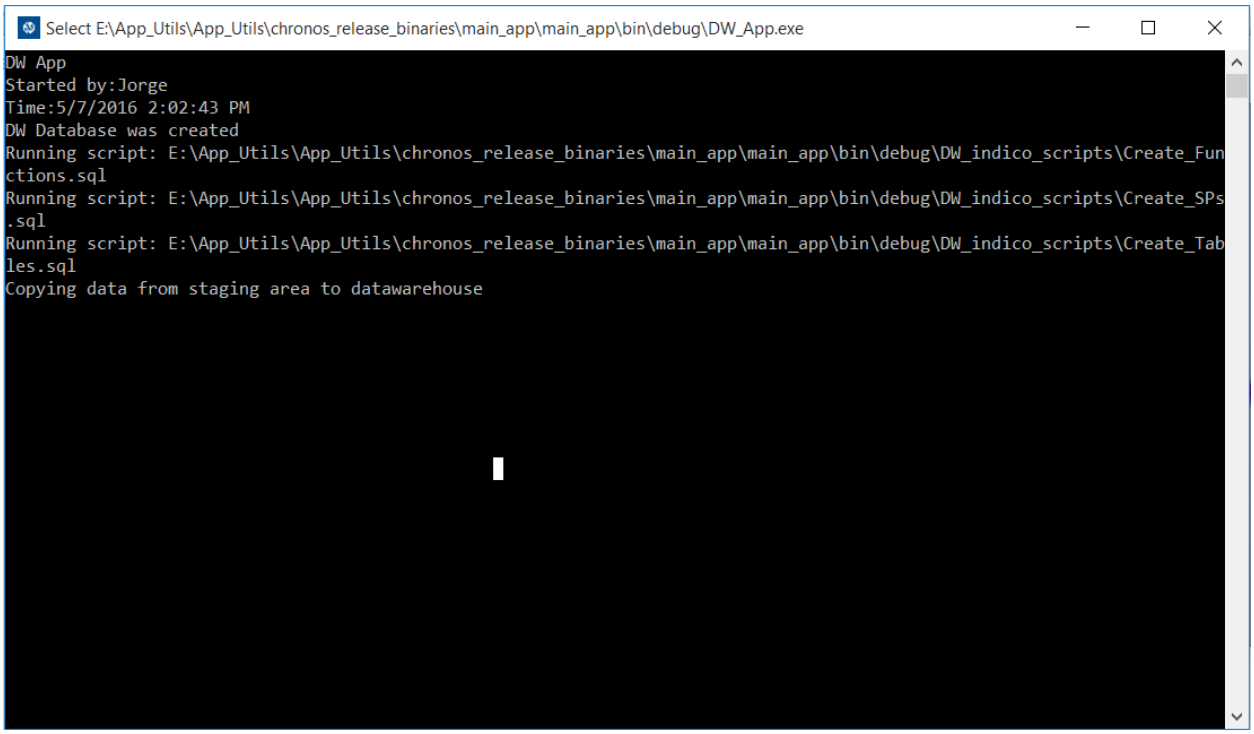


Este proceso se encarga de extraer la información de cada contrato (Excel) y transferirla hacia la base de datos. Dependiendo del producto algunos subprocesos pueden variar pero a nivel macro los más importantes son: detector fuentes de datos, esto es navegar dentro del repositorio y extraer los archivos en formato Excel y llenar una tabla especifica que la herramienta usara como referencia para acceder el archivo físico, revisar si el Excel es un contrato o simplemente es una hoja de Excel genérica, detectar si el contrato tiene o no las hojas correspondientes y entonces proceder con la trasferencia de datos.

5. Llenado de DW ETL

Nombre aplicación: DW_App.exe

Figura 11. DW_App.exe



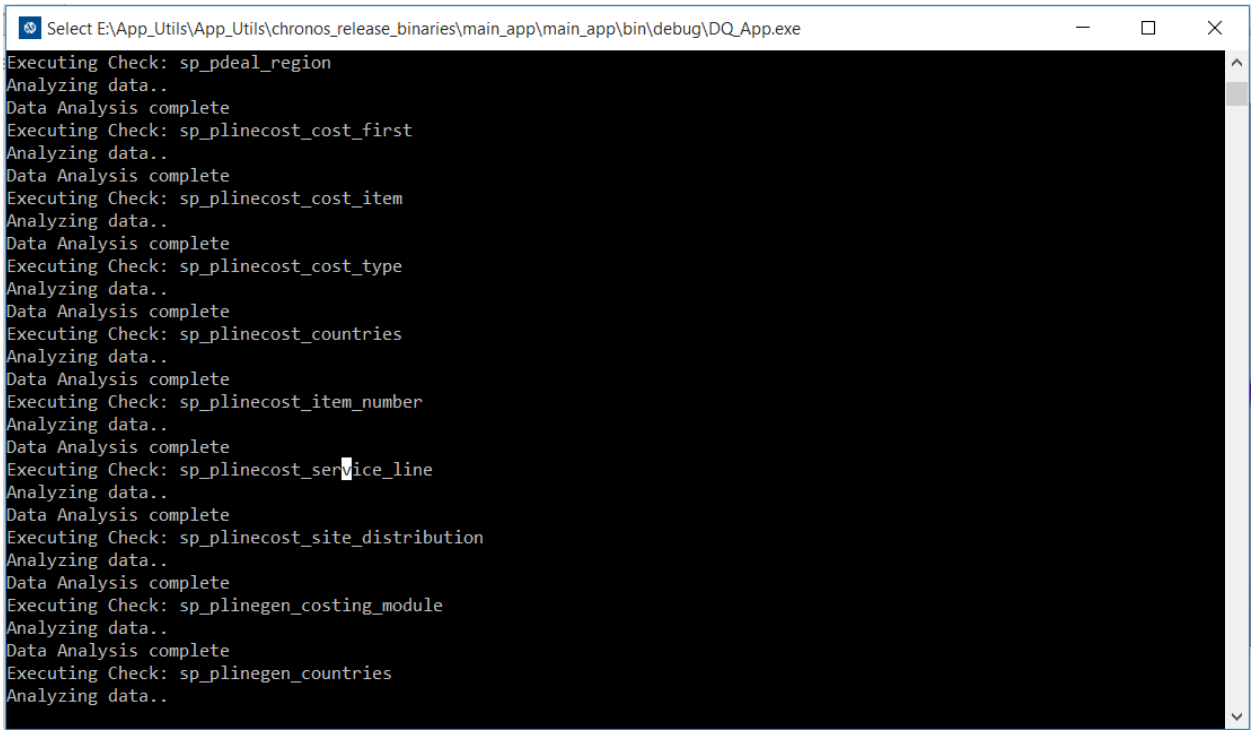
Su función es extraer de la base de datos del staging area únicamente aquella información que es relevante para el cliente. Esta base de datos contiene información ordenada y estructurada, desechando toda aquella información irrelevante de los archivos de Excel extraída en la fase anterior.

6. DQ valida DW vs catálogo

   Nombre aplicación: DQ_App

   El propósito es revisar la información contenida en el DW contra las tablas de catálogos. Si existen inconsistencias serán reportadas al usuario que deberá decidir qué hacer con cada una de ellas. Esta herramienta genera un reporte en completitud de contratos y en consistencia de datos.

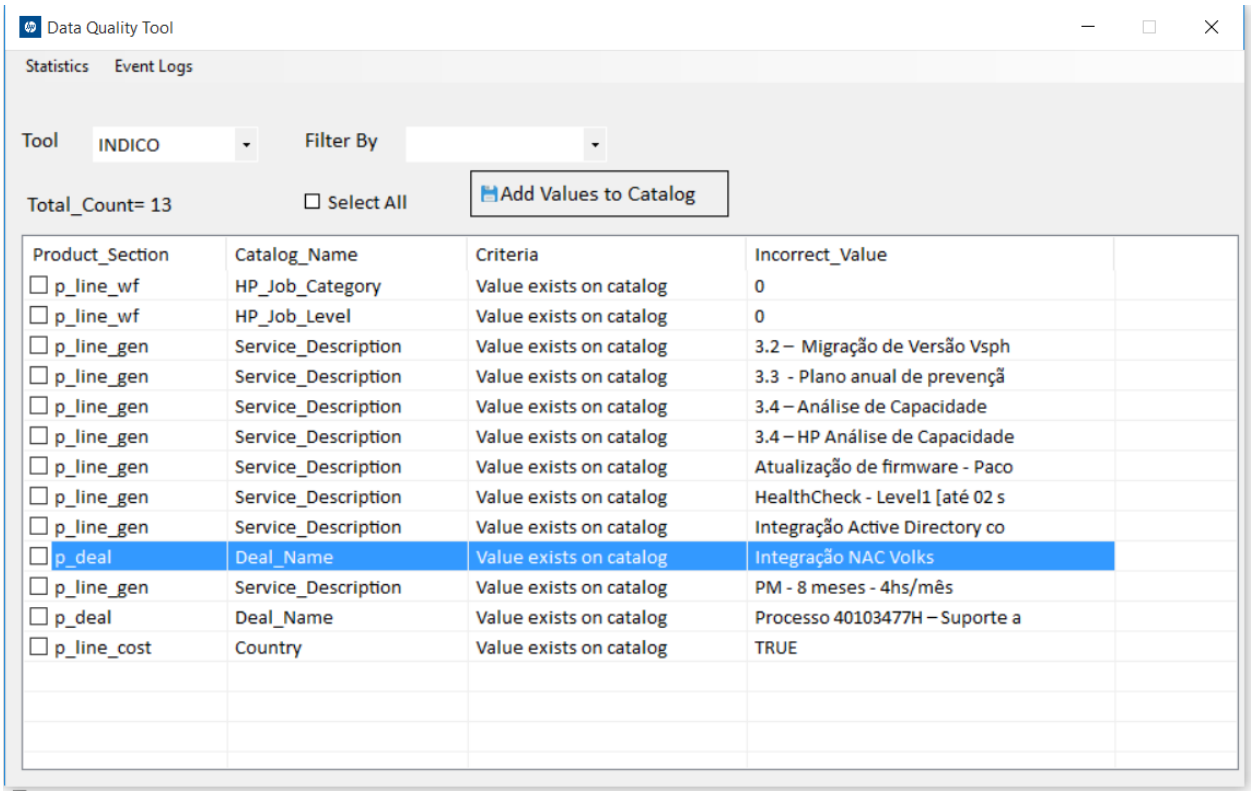
Figura 12. DQ_App



7. Mantenimiento del catálogo

La aplicación DataQuality_Maintenance es una aplicación con interfaz de usuario que permite reparar y visualizar las inconsistencias de datos encontradas en la fase anterior.

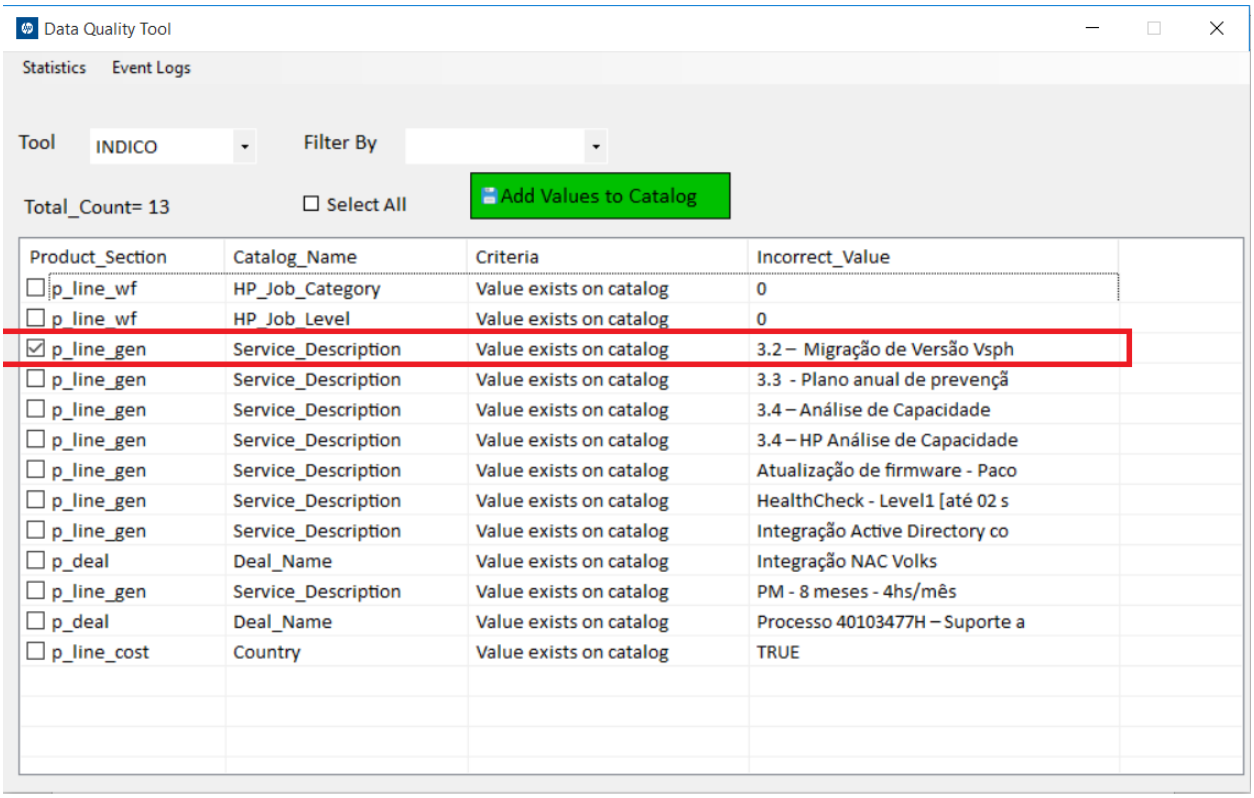
Figura 13. DataQuality_Maintenance



Los datos pueden ser reparados de varias maneras:

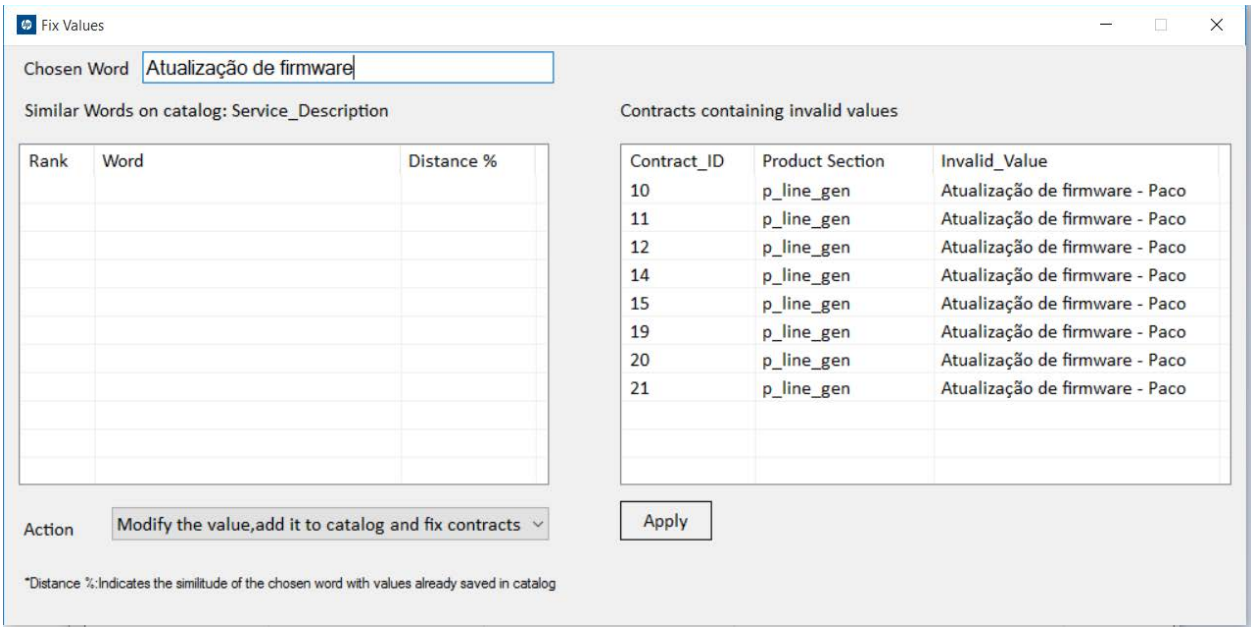
A- Agregar el valor al catálogo: se actualiza la base de datos del catálogo, de manera que se admiten nuevos valores como válidos.

Figura 14. Data Quality Tool



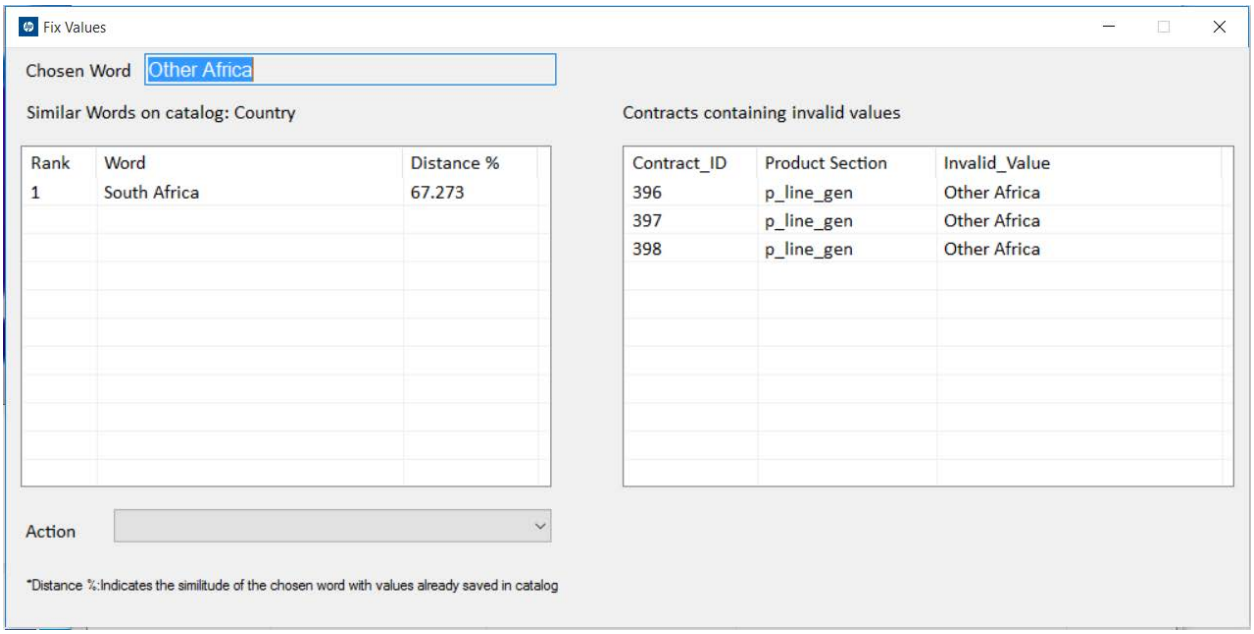
B- Modificar el valor y reparar tanto el catalogo como los datos en el DW: El valor existente se modifica y se agrega al catálogo como un Nuevo valor, mientras tanto en el DW se actualizan los valores existentes con el nuevo valor.

Figura 15. Fix values



C- Modificar los valores incongruentes del DW utilizando sugerencias emitidas por la aplicación: basado en la palabra a modificar la aplicación busca y califica las palabras que más se aproximan basado en criterio de similitud.

Figura 16. Fix Values



5.4. Implementación de algoritmo predictivo para mantenimiento de catálogos

Algoritmo similitud palabras

- Longitud palabra X=palabra seleccionada.
  Cargar arreglo con el contenido palabras en la tabla correspondiente al catálogo al cual pertenece la palabra seleccionada
  Recorrer arreglo
  Longitud palabra Y=palabra arreglo
  Dividir en silabas o bigramas tanto la palabra X y Y.Estas silabas son formadas por cadenas de caracteres basados en el carácter anterior. Se limita a la longitud de palabra menor
  Se realiza una comparación de bigramas y se realiza un conteo basado primero en su posición.
  Se realiza comparación de bigramas totales, sin importar posición
  Calculo porcentaje longitud. Se asigna un peso de 10% que será otorgado según la similitud en cuanto a longitud entre ambas palabras siendo 10 el valor más alto y 0 el más bajo.
  Calculo porcentaje similitud:
  [(90/bigramas_totales)*bigramas_iguales]
  Calculo final:porcentaje longitud+porcentaje similitud
  Salvar arreglo de ranking
  Siguiente palabra
  Fin arreglo
  Impresión del top 10 valores ranking únicamente considerando los mayores a 40% similitud

Excepciones:

Se aplican ciertas excepciones para mantener confiabilidad del algoritmo. 1-Para considerar una palabra potencialmente parecida a otra, el conteo de bigramas consecutivos debe ser mayor a 0.Si el conteo es igual a 0 se desecha inmediatamente.

2-Dado que el algoritmo se basa en la longitud mínima de palabra para realizar la comparación, en caso de que una palabra obtenga el 90% de similitud(es decir todos sus bigramas sean iguales) pero el porcentaje de longitud sea=0 (lo cual significa que existen más de 10 caracteres de longitud entre ambas palabras) se castigara con rebaja de 5% del valor final.

Ejemplo

En el cuadro vemos la palabra seleccionada como “Other Africa”

Palabra X=Other Africa

Palabra Y=South Africa(tomada del catálogo existente de países)

Longitud palabra X=12 caracteres

Longitud palabra Y=12 caracteres

Construcción bigramas (|=espacio)

Palabra X

Ot th he er r| |A Af fr ri ic ca

Palabra Y

So ou ut th h| |A Af fr ri ic ca

Total bigramas=11

Bigramas iguales según posicion=6

Bigramas iguales sin importar posición=7 (el bigrama “th” se repite en ambas palabras)

Ya que los bigramas iguales sin importar posición es mayor y no sobrepasa el límite de total bigramas, se toma como valor valido para realizar calculo.

Calculo longitud=Dado que no hay diferencia, se asigna el 10% del total Calculo similitud=90/total bigramas*bigramas iguales=90/11*7

Total=67.27%

5.5 Reglas de calidad

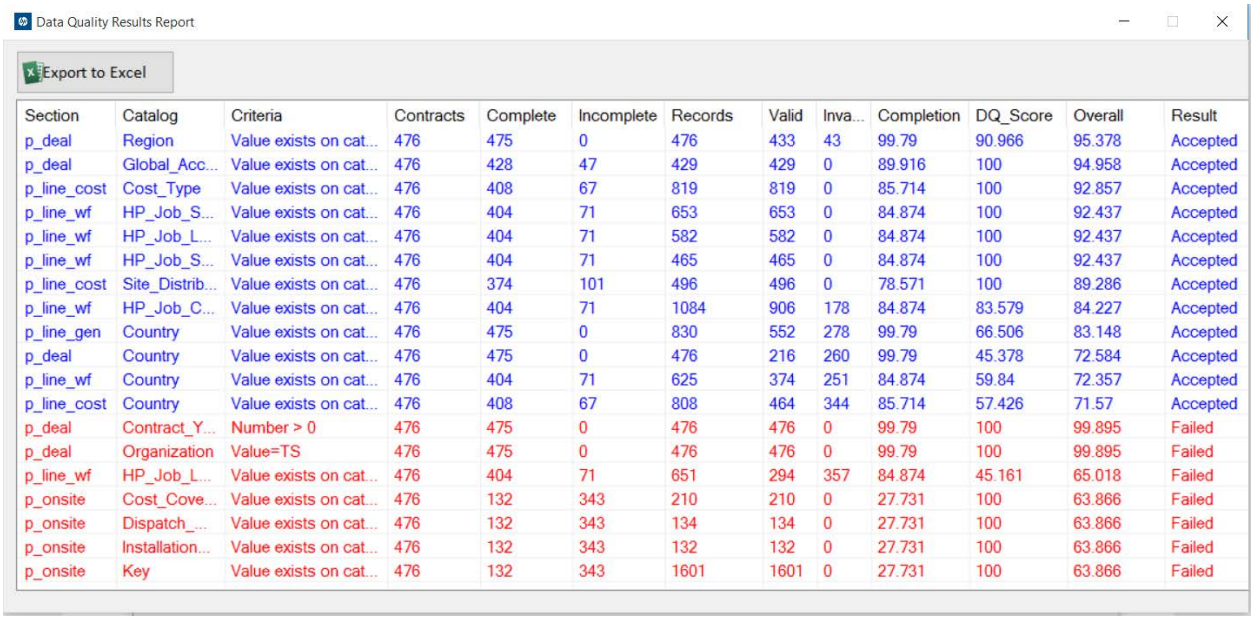
El sistema es capaz de “aprender” y replicar comportamientos del usuario esto con el fin de evitar que para cada corrida el usuario deba arreglar los mismos errores.

Tomando como ejemplo el caso anterior, si el usuario selecciona como valor valido “South África” para reparar “Other África”, se creará una regla de manera que el sistema automáticamente arreglará el valor durante la próxima ejecución.

5.6 Reportes calidad datos

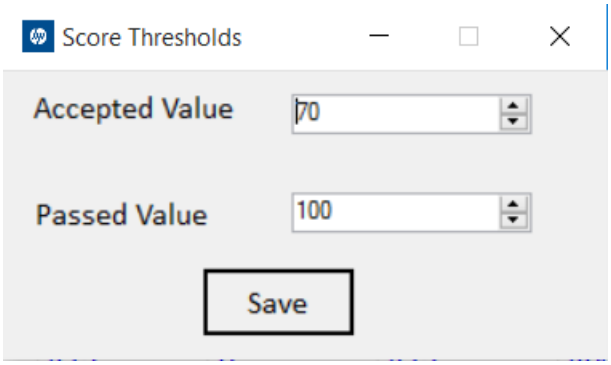
Brinda una perspectiva al usuario de que tan completos están los contratos y cuál es la consistencia de los datos.

Figura 17. Data Quality Report



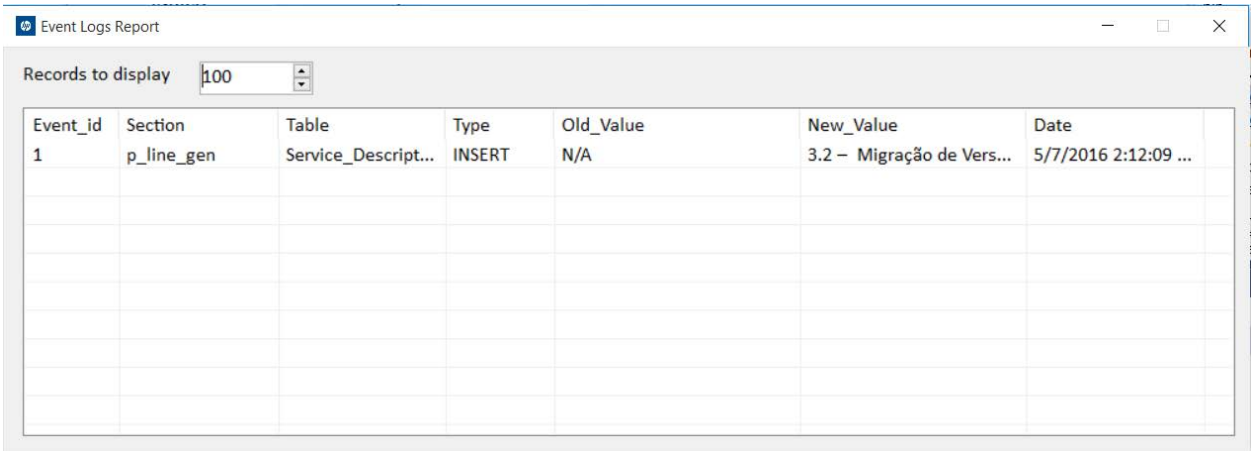
Section	Catalog	Criteria	Contracts	Complete	Incomplete	Records	Valid	Inva...	Completion	DQ_Score	Overall	Result
p_deal	Region	Value exists on cat...	476	475	0	476	433	43	99.79	90.966	95.378	Accepted
p_deal	Global_Acc...	Value exists on cat...	476	428	47	429	429	0	89.916	100	94.958	Accepted
p_line_cost	Cost_Type	Value exists on cat...	476	408	67	819	819	0	85.714	100	92.857	Accepted
p_line_wf	HP_Job_S...	Value exists on cat...	476	404	71	653	653	0	84.874	100	92.437	Accepted
p_line_wf	HP_Job_L...	Value exists on cat...	476	404	71	582	582	0	84.874	100	92.437	Accepted
p_line_wf	HP_Job_S...	Value exists on cat...	476	404	71	465	465	0	84.874	100	92.437	Accepted
p_line_cost	Site_Distrib...	Value exists on cat...	476	374	101	496	496	0	78.571	100	89.286	Accepted
p_line_wf	HP_Job_C...	Value exists on cat...	476	404	71	1084	906	178	84.874	83.579	84.227	Accepted
p_line_gen	Country	Value exists on cat...	476	475	0	830	552	278	99.79	66.506	83.148	Accepted
p_deal	Country	Value exists on cat...	476	475	0	476	216	260	99.79	45.378	72.584	Accepted
p_line_wf	Country	Value exists on cat...	476	404	71	625	374	251	84.874	59.84	72.357	Accepted
p_line_cost	Country	Value exists on cat...	476	408	67	808	464	344	85.714	57.426	71.57	Accepted
p_deal	Contract_Y...	Number > 0	476	475	0	476	476	0	99.79	100	99.895	Failed
p_deal	Organization	Value=TS	476	475	0	476	476	0	99.79	100	99.895	Failed
p_line_wf	HP_Job_L...	Value exists on cat...	476	404	71	651	294	357	84.874	45.161	65.018	Failed
p_onsite	Cost_Cove...	Value exists on cat...	476	132	343	210	210	0	27.731	100	63.866	Failed
p_onsite	Dispatch_...	Value exists on cat...	476	132	343	134	134	0	27.731	100	63.866	Failed
p_onsite	Installation...	Value exists on cat...	476	132	343	132	132	0	27.731	100	63.866	Failed
p_onsite	Key	Value exists on cat...	476	132	343	1601	1601	0	27.731	100	63.866	Failed

El resultado final está basado en límites fijados previamente por el usuario ya que en algunas ocasiones un 90% puede ser considerado bueno o se requiere que el 100% de los datos estén correctos.



5.7 Cambios en las tablas tanto de catálogos como DW

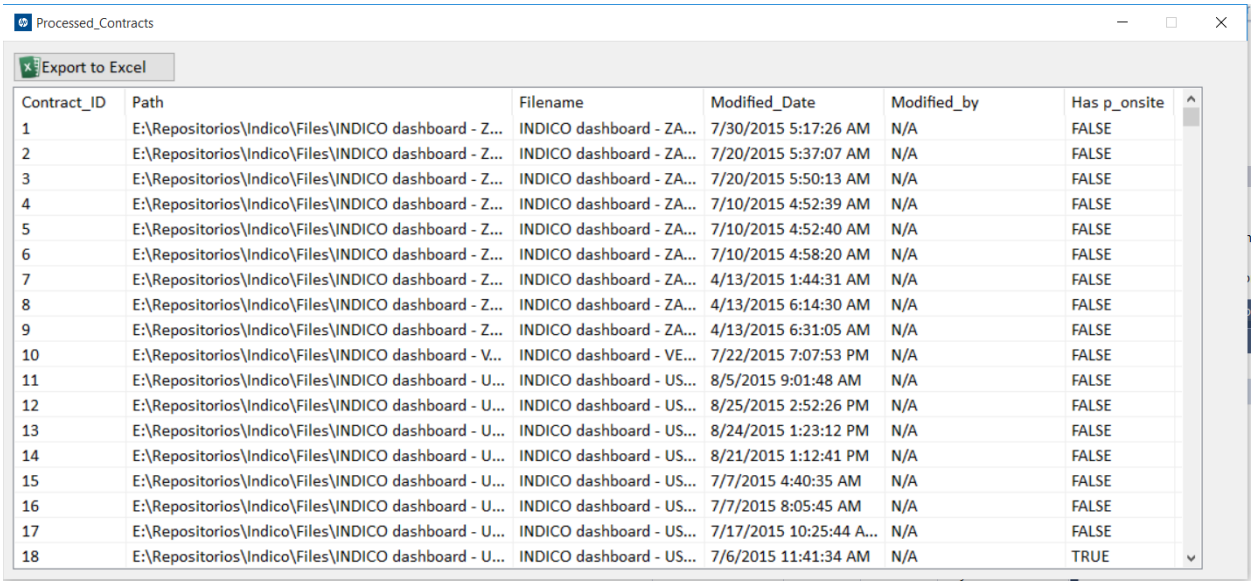
Permite la trazabilidad de los cambios realizados por el usuario tanto a nivel de catálogo como de DW



5.8 Reportes ejecución

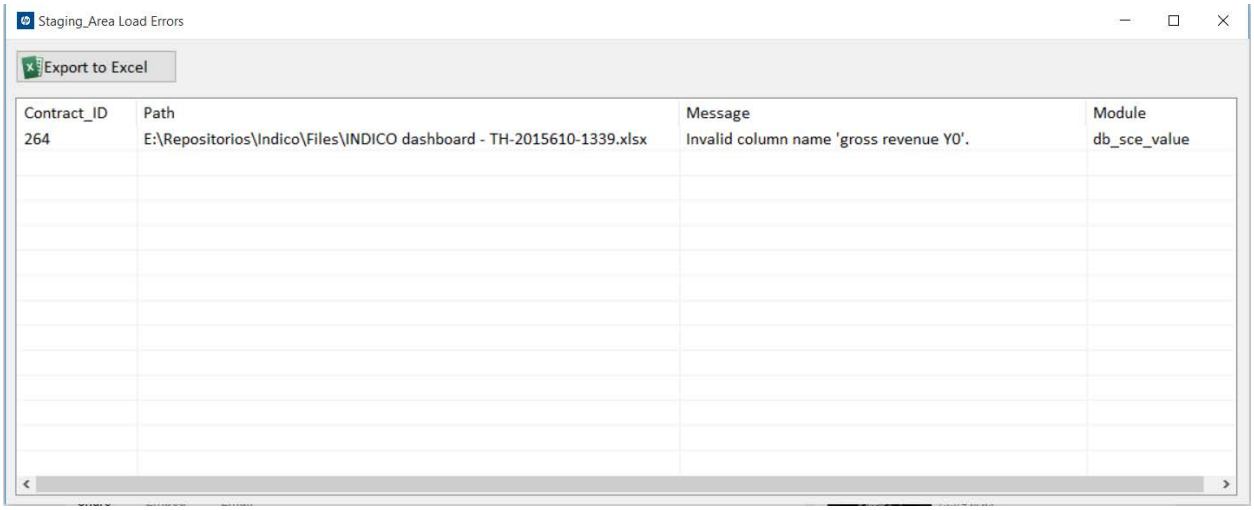
- **Contratos procesados**

Figura 18. Contratos procesados satisfactoriamente



- **Contratos con errores**

Figura 19. Contratos con inconsistencia



Capítulo 6.

Conclusiones & Recomendaciones

6.1. Conclusiones

En esta sección se van a resumir las conclusiones que se lograron recabar luego del proyecto aplicado, se pudo confirmar que el trabajo realizado cumplió efectivamente los objetivos y mejoró el desempeño de la organización en esta área.

Se logró identificar todas las fuentes con sus diferentes tipos de datos y formatos, lo que permitió crear una solución acorde a las necesidades de la organización.

Al desarrollar un algoritmo de comparación de datos para realizar comparaciones entre los datos crudos y el catálogo, se generó un porcentaje de proximidad con el fin de generar un listado de opciones a escoger por el usuario a la hora de corregir los datos en el Data Warehouse y esto incrementará la calidad de la información a partir de ese punto.

Al crear un sistema de almacenamiento de reglas o cambios en el Data Warehouse se preservan las preferencias de manejo de datos del usuario, lo que permitió mantener un control permanente de la información que contenga el catalogo.

El orquestador de aplicaciones que mantiene y centraliza el manejo de las diferentes aplicaciones dentro de la solución final permitió tener una administración clara de la solución desde el punto de vista del usuario, mejorando la experiencia del mismo en el uso de la solución.

El tiempo de procesamiento de esta tarea pasó de 30 horas a tan solo 45 minutos, dándole a la organización una mejora en tiempos de procesamiento de datos de aproximadamente 700% con respecto al proceso anterior.

El proyecto generó un ahorro de $4,200.00 por mes, que, al calcularlo por año, dará un ahorro de más $50,000.00, logrados del ahorro de la inversión de tiempo de la doctora en estadística.

El costo del proyecto se calculó tomando en consideración el salario de los 2 programadores por el tiempo de creación de la solución. Este costo es de $20,000.00, sin embargo, el retorno de la inversión se calcula en 4.8 meses luego de la implementación.

6.2. Recomendaciones

A la hora de usar herramientas de mercado, como Analysis Services e Integration Services, es necesario probar que cumplan todos los requerimientos listados.

Generar bitácoras de información para las acciones críticas del sistema, esto con el fin de que estos tengan históricos para control y para seguridad de los datos.

Trabajar con el usuario final en la prueba del proceso, para poder obtener retroalimentación en cada iteración que se le dé al desarrollo de la solución.

Es muy importante delimitar el alcance del proyecto y apegarse a este, ya que los cambios en los requerimientos son muy comunes y esto genera retrasos en las fechas de entrega del proyecto.

Buscar proyectos que son generados por la entrega de la solución, es una forma eficaz de poder entender que el proyecto fue efectivo y dejo secuelas positivas en la organización.

Medir los tiempos de procesamiento previos y posteriores a la implementación de la solución, permite generar un porcentaje de mejora del proceso intervenido.

Capítulo 7.

Reflexiones finales

Al cabo de la realización de este proyecto, muchas cosas son rescatables.

1. El uso de las herramientas de mercado.

Hay muchas herramientas de casas desarrolladoras muy reconocidas, sin embargo, estas herramientas no siempre logran el resultado esperado para una necesidad específica. Las soluciones desarrolladas en este proyecto fueron realizadas basadas en la necesidad de la organización y no pensadas desde el punto de vista de la capacidad de una aplicación, lo que muestra que estas herramientas tienen sesgos en su capacidad de ser implementadas en todos los ambientes para los que fueron diseñadas.

2. Trabajar con el cliente en cada etapa del desarrollo

Para el desarrollo de una herramienta, es crítico el contacto constante con el usuario final, esta premisa indica que si el usuario está en constante roce con el desarrollo de la solución va a ser un participe crucial en el éxito del resultado final. El usuario final marcó la pauta desde las primeras etapas del desarrollo y eso resultó en un proyecto exitoso, y garantizó que esta solución solucionó el problema y que esta herramienta va a ser usada de manera satisfactoria, siendo este último, un problema muy constante en el desarrollo de soluciones de automatización de tareas.

3. Mantener el enfoque inicial sin agregar nuevas funcionalidades

Al trabajar con el cliente en las etapas de desarrollo es muy natural que surjan cambios o mejoras en el camino, estas mejoras y cambios que son generalmente propuestos por el usuario, afectan el tiempo de desarrollo porque proponen nuevos retos y giros en los tiempos de realización de la solución. Es importante trabajar con el usuario en mantener el enfoque inicial para no afectar el tiempo de entrega original.

4. Mejoramiento de la calidad de datos usando algoritmos de calidad

El uso de algoritmos de proximidad para el mantenimiento de los catálogos significó un gran reto, pero también fue uno de los aportes de innovación más claros de este proyecto, dando un giro de innovación y tecnología aplicada.

5. ROI, el valor del trabajo y el retorno de la inversión

El cálculo del Retorno de la inversión representa una ventaja a la hora de hacer un proyecto y permite evaluar el costo del proyecto y además definir cuando este va a comenzar a dejar ganancias en la organización. Este retorno de la inversión fue bastante positivo y mostró el gran ahorro que se va a generar con la implementación de este proyecto.

6. Hallazgos de brechas de calidad y reporteo en herramientas en desarrollo y en producción.

La implementación de este proyecto trajo a la luz un sinnúmero de “Áreas de mejora” que no se hubieran podido detectar fácilmente en las circunstancias en las que se llevaban a cabo los diferentes procesos, como, por ejemplo:

- Casper: no tiene una conexión o prevista para reportes, lo que dificulta la realización de controles y reportes basados en la configuración actual.
- Flexible Capacity: además de no contar con la prevista para reportes, tiene serios problemas de manejo de datos que se evidenciaron con la implementación de este proyecto.

Capítulo 8.

Trabajos a futuro

Como bien se dice en proyectos: “el mejor indicador de un proyecto exitoso es una lista de nuevos proyectos como consecuencia de este”. Esta frase ayuda a reconocer que este proyecto fue exitoso y generó una buena lista de proyectos con su llegada:

- Incluir a Flexible Capacity en el marco del proyecto de extracción de datos
- Crear una plataforma de reporteo y extracción de datos para Casper
- Incluir a Casper en el marco del proyecto de extracción de datos
- Crear una herramienta de precios que incluya un marco de reporteo y lógica de negocios para MRU que es una plataforma dentro de la organización que se dedica a la venta de repuestos en el departamento de garantías

Glosario

CASPER: Una herramienta que maneja un tipo de productos específicos para HP

Data Warehouse: Base de datos que integra información de varias fuentes, que permite el análisis de los datos de una forma consolidada y completa.

DC CONFIG: Una herramienta que maneja un tipo de productos específicos para HP

ETL: Proceso de Extracción, Transformación y Carga de datos

HPE: Hewlett Packard Enterprise

HPFS: Hewlett Packard Finantial Services

INDICO: Una herramienta que maneja un tipo de productos específicos para HP

.Net: de la familia Microsoft, es una solución para el desarrollo de aplicaciones.

Staging Area: concepto de administración de Bases de datos que define una tabla intermedia en un proceso de Extracción Transformación y Carfa de datos (ETL)

WW TS Support Solutions Pricing: Organización de HPE que se encarga de fijar los precios de los productos

Bibliografía

- (n.d.). Retrieved November 15, 2015, from http://www.sec.gov/Archives/edgar/data/47217/000104746910010444/a2201180z10-k.htmHewlett-Packard. (n.d.). Retrieved November 27, 2015, from http://hewlett-packard-unaq.blogspot.com/p/analisis-del-entorno.html
- (n.d.). Retrieved November 06, 2015, from http://www.sec.gov/Archives/edgar/data/47217/000104746910010444/a2201180z10-k.htm
- Baškarada, S, Koronios, A (2014) "A Critical Success Factors Framework for Information Quality Management" Information Systems Management, vol. 31, no. 4, pp. 1-20.
- Ivanov, K. (1972) "Quality-control of information: On the concept of accuracy of information in data banks and in management information systems". The University of Stockholm and The Royal Institute of Technology. Doctoral dissertation.
- Implantación de un sistema de calidad: los diferentes sistemas de calidad existentes en la organización:
- http://books.google.co.cr/books?hl=es&lr=&id=qdv2lr9yr3wC&oi=fnd&pg=PA1&dq=sistema+de+calidad&ots=uGcZ0XWtcR&sig=NQYVZUGGzrYRTPy03pSsixxLtCU
- Datos, información, conocimiento. (n.d.). Retrieved July 1, 2015, from http://www.sinnexus.com/business_intelligence/piramide_negocio.aspx
- ¿Qué es un estándar? (n.d.). Retrieved July 1, 2015, from https://americalatina.pmi.org/latam/PMBOKGuideAndStandards/WhatIsAStandar.aspx
- Conozca 3 tipos de investigación: Descriptiva, Exploratoria y Explicativa. (2013). Retrieved July 8, 2015, from http://www.creadess.org/index.php/informate/de-interes/temas-de-interes/17300-conozca-3-tipos-de-investigacion-descriptiva-exploratoria-y-explicativa
- MARTÍNEZ MEDIANO, C. (1996). Evaluación de programas educativos. Investigación evaluativa. Modelo de evaluación de programas. Madrid: UNED.
- HERNÁNDEZ, F., IGLESIAS, E., FUENTES, P. y SERRANO, F. J. (1994) Introducción al proceso de investigación en educación. Murcia. [DL: MU-2158-1992].
- Stufflebeam, Daniel; Shinkfiel, Anthony (1995) Evaluación Sistemática - Guía Teórica y práctica. España: Centro de Publicaciones del Ministerio de Educación y Ciencia, Ediciones Paidós Ibérica.
- Glossary of statistical terms. (n.d.). Retrieved August 12, 2015, from http://www.statistics.com/glossary&term_id=812

- Hunt, Neville; Tyrrell, Sidney (2001) Stratified Sampling. Páginaweb de Coventry University (visto julio 2015), from http://nestor.coventry.ac.uk/~nhunt/meths/strati.html
- Focus Groups. (2015). Retrieved August 12, 2015, from http://www.usability.gov/how-to-and-tools/methods/focus-groups.html
- Simon, K. (2000). The Cause and Effect (a.k.a. Fishbone) Diagram. Retrieved August 12, 2015, from http://www.isixsigma.com/tools-templates/cause-effect/cause-and-effect-aka-fishbone-diagram/
- Buzan, T. (2013), "Cómo crear mapas mentales". Barcelona.URANO, Enero 2013
- C. Melissa Mcclendon, Larry Regot, Gerri Akers: The Analysis and Prototyping of Effective Graphical User Interfaces. October 1996
- Beneficios de consolidar bases de datos en nubes privadas. (2013). Retrieved July 11, 2016, from http://itusersmagazine.com/2013/02/28/beneficios-de-consolidar-bases-de-datos-en-nubes-privadas/

Apéndices

Apéndice 1.

ROI

Improvement Projects FTE / ROI Calculation Template

Project name:	Please fill in the Project Name
Project Goal Statement:	Please fill in the projects Goal Statement
Activity(s) Involved:	Please fill in the name of the Activity(s) Involved

Data	Standard Time per Activity (Min)	Monthly Recurrence	Total Processing Time (minutes per month)	Minutes per month (1 FTE)	Head Count
Actual Process	7680	1	7680	9600	0.800
Improved Process	180	1	180	9600	0.019
Improvement Percentage	4267%		98%		98%

Project Duration (Months)	4.00
Project FTE Investment (per month)	2.000
FTE Cost (Monthly Salary)	$5,571.25
FTE Cost (Yearly Salary)	$36,000.00
TOTAL Project FTE Investment	8.000
FTE REDUCTION	0.781
Total Project Cost / Investment	$24,000.00
Months required to Repay the Investment	10

Employee Cost Per Minute	$0.58		
Activity Cost Per Month			
Current	$4,457.00	New	$104.46
Monthly Savings	$4,352.54		
Activity Cost Per Year			
Current	$53,484.00	New	$1,253.53
Annual Savings	$52,230.47		
Investment Monthly Return Percentage	18.14%		
Years required to Repay the Investment	0.46		
Yearly Cost Balance During Project Development	$28,230.47		

ROI Ratio
2.176269531

**Note: Please fill in Underlined values only

Apéndice 2.

Algoritmo de calidad

'El **propósito** de este **módulo** es encontrar el top 10 de palabras similares a la seleccionada por el usuario de manera que este pueda seleccionarla como valor valido en el **catálogo**.

```
Module Data_Qualityv2

'Inicio del modulo

    Public Sub Full_proc()

        'Limpieza de los arreglos e inicialización de variables

      For j As Integer = 0 To 999

         var.words_rank(j, 0) = Nothing

         var.words_rank(j, 1) = Nothing

         var.words_rank(j, 2) = Nothing

      Next

      For y As Integer = 0 To 4999

         var.catalog_word_data(y, 0) = Nothing

         var.catalog_word_data(y, 1) = Nothing

      Next

      var.index_rank = -1

      var.match_counter = 0

      If check_zero(var.word_x) = True Then

         var.words_rank(0, 0) = "N/A"
```

```
    var.words_rank(0, 1) = "0"
    Exit Sub
End If
load_data_array()
            'Flujo principal de la operación
For x As Integer = 0 To 1000
    If String.IsNullOrEmpty(var.catalog_word_data(x, 1)) = False Then
        flush_variables()
        var.word_y = var.catalog_word_data(x, 1)
        load_bigram_x()
        load_bigram_y()
        get_lowest_number(var.total_bigrams_x, var.total_bigrams_y)
        get_consecutives_bigrams()
        get_matching_bigrams()
        get_lenght_percentage(var.word_x_lenght, var.word_y_lenght)
        get_total_match(var.consecutive_match, var.nonconsecutive_match)
        get_word_distance(var.total_bigrams, var.total_match, var.lenght_percentage)
        distance_rules()
        load_similar_values()
    End If
Next
ordering_similar_values()
print_similar_values()
```

```
    End Sub
```

'Este procedimiento se encarga de procesar la palabra o cadena seleccionada por el **usuario. Su** principal **función** es primero tomar la longitud

'de la palabra y guardar los bigramas en un **arreglo. Un** bigrama es la **unión** de dos letras de manera consecutiva siempre tomando como base la **última** letra.

'Ejemplo la palabra puede ser **Aruba. Los** bigramas **serían** AR RU UB BA

```
    Public Sub load_bigram_x()
        logger.Debug("Word X:" & var.word_x)
        var.word_x_lenght = var.word_x.Length


        For i As Integer = 0 To var.word_x_lenght - 2
            var.bigram_x_array(i) = String.Concat(var.word_x.Substring(i, 2))
        Next
        For x As Integer = 0 To 100
            If String.IsNullOrEmpty(var.bigram_x_array(x)) = False Then
                var.total_bigrams_x = var.total_bigrams_x + 1


            End If
        Next
        logger.Debug("Counter:" & var.total_bigrams_x.ToString)
    End Sub
```

'Este procedimiento es exactamente igual al anterior a diferencia que carga en un arreglo todos los valores pertenecientes al **catálogo** relacionado al **módulo** seleccionado

```
        'y los convierte en bigramas
    Public Sub load_bigram_y()


        logger.Debug("Word y:" & var.word_y)
        var.word_y_lenght = var.word_y.Length


        For i As Integer = 0 To var.word_y_lenght - 2
            var.bigram_y_array(i) = String.Concat(var.word_y.Substring(i, 2))
        Next
        For x As Integer = 0 To 100
            If String.IsNullOrEmpty(var.bigram_y_array(x)) = False Then
                var.total_bigrams_y = var.total_bigrams_y + 1
            End If
        Next
        logger.Debug("Counter:" & var.total_bigrams_y.ToString)
    End Sub
```

'La **función** de este procedimiento es contar cuantos bigramas son iguales si se compara x vs y contando **únicamente** aquellos que sean iguales en el mismo **índice**.

```
    Public Sub get_consecutives_bigrams()
        For i As Integer = 0 To var.total_bigrams - 1
            If String.Compare(var.bigram_x_array(i), var.bigram_y_array(i), True) = 0 Then
                var.consecutive_match = var.consecutive_match + 1
                If i = 0 Then
```

```
            var.consecutive_flag = True
        End If
        logger.Debug("x:" & var.bigram_x_array(i))
        logger.Debug("y:" & var.bigram_y_array(i))
    End If

  Next
  logger.Debug("There are " & var.consecutive_match.ToString & " out of " & var.total_bigrams & " bigrams in a consecutive")
 End Sub
```

'La **función** de este procedimiento es contar cuantos bigramas son iguales si se compara x vs y contando todos aquellos que sean iguales sin importar el **índice**

```
 Public Sub get_matching_bigrams()
  For i As Integer = 0 To var.total_bigrams - 1
    For x As Integer = 0 To var.total_bigrams - 1
      If String.Compare(var.bigram_x_array(i), var.bigram_y_array(x), True) = 0 Then
        var.nonconsecutive_match = var.nonconsecutive_match + 1
        logger.Debug("x:" & var.bigram_x_array(i))
        logger.Debug("y:" & var.bigram_y_array(x))
      End If
    Next
  Next
```

```
    If var.nonconsecutive_match > var.total_bigrams Then
        var.nonconsecutive_match = var.total_bigrams
    End If
    logger.Debug("There are " & var.nonconsecutive_match.ToString & " out of " & var.total_bigrams & " bigrams in a non consecutive")
  End Sub
```

'Este procedimiento calcula la **penalización** de la palabra en caso de que no tengan la misma longitud.

'El peso asignado a la igualdad de longitud es de 10% y se calcula mediante la regla 10- el valor absoluto de la diferencia entre ambas palabras.

```
  Public Function get_lenght_percentage(ByVal x As Integer, ByVal y As Integer) As Double
    If x = y Then
        var.lenght_percentage = 10
        logger.Debug("Lenght %: " & var.lenght_percentage.ToString)
        Return var.lenght_percentage
        Exit Function
    End If
    If Math.Abs(x - y) < 10 Then
        var.lenght_percentage = 10 - (Math.Abs(x - y))
        logger.Debug("Lenght %: " & var.lenght_percentage.ToString)
        Return var.lenght_percentage
        Exit Function
    End If
```

```
        If Math.Abs(x - y) > 10 Then
            var.lenght_percentage = 0
            logger.Debug("Lenght %: " & var.lenght_percentage.ToString)
            Return var.lenght_percentage
            Exit Function
        End If
        Return 0
    End Function
        'Calcula el total de igualdades basado en bigramas consecutivos y no consecutivos
    Public Function get_total_match(ByVal consecutive As Integer, ByVal nonconsecutive As Integer) As Integer
        If consecutive > nonconsecutive Then
            var.total_match = consecutive
        End If
        If consecutive < nonconsecutive Then
            var.total_match = nonconsecutive
        End If
        If consecutive = nonconsecutive Then
            var.total_match = consecutive
        End If
        Return var.total_match
    End Function
```

'Calcula la distancia entre ambas palabras(x vs y).Se calcula divi**di**endo 100 entre el total de bigramas (el total de bigramas es la cantidad de bigramas de la palabra **más pequeña** en longitud)

'El valor resultante se multiplica por la cantidad de bigramas iguales y se multiplica por 90%.A este resultado se le suma el valor obtenido de la **penalización** por longitud que ocupa el restante 10%

```
Public Function get_word_distance(ByVal total_bigrams As Integer, ByVal total_match As Integer, ByVal lenght_percentage As Double) As Double
    var.word_distance = ((100 / total_bigrams) * total_match)
    var.word_distance = (var.word_distance * 0.9) + lenght_percentage
    logger.Debug("Word distance before rules: " & var.word_distance.ToString)
    Return var.word_distance
End Function
```

'Conjunto de reglas o excepciones para ciertas condiciones en la **comparación**

```
Public Sub distance_rules()
```

'Si las palabras son exactamente **iguales, el** valor da 100%.Esta **condición** nunca debe ocurrir ya que entonces el valor no **debería** aparecer como invalido para el usuario

```
    If String.Compare(var.word_x, var.word_y, True) = 0 Then
        var.word_distance = 100
        logger.Debug("Word distance: " & var.word_distance.ToString)
        Exit Sub
    End If
```

'Esta invalida una palabra si contiene bigramas iguales no consecutivos pero cero bigramas iguales **consecutivos. Esto evitaría palabras en la clasificación.**

'con una similitud considerable pero que no tengan una igualdad real

```
    If var.consecutive_match = 0 Then
        var.word_distance = 0
        logger.Debug("Word distance: " & var.word_distance.ToString)
        Exit Sub
    End If
    'Penaliza con 5% si la longitud entre palabras es mayor a 10 caracteres.
    If var.consecutive_match > 0 And var.lenght_percentage = 0 Then
        var.word_distance = var.word_distance - 5
        logger.Debug("Word distance: " & var.word_distance.ToString)
        Exit Sub
    End If
    'Anula igualdades si la regla de consecutivo es 0
    If var.consecutive_match = 1 And var.consecutive_flag = False Then
        var.word_distance = 0
    End If

End Sub
      'Limpieza de variables
Public Sub flush_variables()

    var.word_y = ""
    var.word_x_lenght = 0
    var.word_y_lenght = 0
```

```
        var.total_bigrams = 0
        var.total_bigrams_x = 0
        var.total_bigrams_y = 0
        var.consecutive_match = 0
        var.nonconsecutive_match = 0
        var.word_distance = 0
        var.lenght_percentage = 0
        var.consecutive_flag = False
        For x As Integer = 0 To 99
            var.bigram_x_array(x) = Nothing
            var.bigram_y_array(x) = Nothing
        Next
    End Sub
```

'Carga en un arreglo todas aquellas palabras que obtuvieron un **puntuación** mayor o igual a 35%

```
    Public Sub load_similar_values()
        If var.word_distance > 35 Then
            var.index_rank = var.index_rank + 1
            var.words_rank(var.index_rank, 0) = var.word_y
            var.words_rank(var.index_rank, 1) = var.word_distance.ToString
            var.match_counter = var.match_counter + 1
            logger.Debug("New word rank:" & var.words_rank(var.index_rank, 0) & " --> " & var.words_rank(var.index_rank, 1))
```

```
        End If
    End Sub
        'Impresion del arreglo anterior
    Public Sub print_similar_values()
        For x As Integer = 0 To 15
            If String.IsNullOrEmpty(var.words_rank(x, 0)) = False Then
                logger.Debug(var.words_rank(x, 0) & " --> " & var.words_rank(x, 1))
            End If
        Next

    End Sub
        'Ordenamiento del arreglo para que las palabras sean impresas de mayor a menor segun el porcentaje de similitud
    Public Sub ordering_similar_values()
        Dim aux_x_0 As String
        Dim aux_x_1 As String
        Dim aux_y_0 As String
        Dim aux_y_1 As String

        For x As Integer = 0 To var.match_counter - 1
            For y As Integer = 0 To var.match_counter - 1
                If System.Convert.ToDouble(var.words_rank(x, 1)) > System.Convert.ToDouble(var.words_rank(y, 1)) Then
```

```
                aux_x_0 = var.words_rank(x, 0).ToString
                aux_x_1 = var.words_rank(x, 1).ToString
                aux_y_0 = var.words_rank(y, 0).ToString
                aux_y_1 = var.words_rank(y, 1).ToString
                var.words_rank(x, 0) = aux_y_0
                var.words_rank(x, 1) = aux_y_1
                var.words_rank(y, 0) = aux_x_0
                var.words_rank(y, 1) = aux_x_1
            End If
        Next
    Next
End Sub
End Module
```

Apéndice 3.

Detalle de bases de datos

Nombre Base Datos	Funcionalidad
db_admin_<product>	Almacena reglas creadas por el usuario
Db_catalog_<product>	Catálogo de palabras validas definidas por el negocio
Db_dw_<product>	Almacena datos de forma ordenada mediante consultas al staging area
Db_sa_<product>	Almacena de manera cruda la información de cada contrato en excel
Db_dq_results_<product>	Almacena los resultados de las pruebas de calidad de datos.

Apéndice 4.

Flujograma de solución final