PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

DATA QUALITY PARA LA MIGRACIÓN DE UNA BASE DE DATOS ORIENTADA A COLUMNAS

JUAN PABLO HERRERA GUZMÁN

JUAN CARLOS SALINAS ORELLANA

INFORME FINAL DEL PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

JULIO 2014

Pontificia Universidad Católica de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

DATA QUALITY PARA LA MIGRACIÓN DE UNA BASE DE DATOS ORIENTADA A COLUMNAS

JUAN PABLO HERRERA GUZMÁN

JUAN CARLOS SALINAS ORELLANA

Profesor Guía: **José Rubio León**

Profesor Correferente: **Broderick Crawford labrín**

Carrera: INGENIERÍA CIVIL INFORMÁTICA

JULIO 2014

Dedicatoria

Dedicamos este trabajo de memoria de título a toda nuestra familia, a aquellos que nos han apoyado incondicionalmente no sólo ahora, sino toda nuestra vida, a nuestros padres y abuelos, a los que están y a los que ya han partido y a todos aquellos que han llegado a formar parte de nuestra vida y que con su preocupación nos han motivado a concluir esta etapa del camino. A todos ellos, muchas gracias.

Índice

Índice de tablas

Índice de Ilustraciones

Resumen

Este informe de proyecto describe la implementación de un marco metodológico para el desarrollo de proyectos de calidad de datos. Si bien un proyecto de calidad de datos es un proceso lleno de casuísticas y particularidades que dificultan en gran medida concebir algún tipo de generalización, hemos querido plasmar en el desarrollo de este trabajo un esbozo de lo que pudiera perfilarse como una metodología a la hora de abordar problemas de esta naturaleza. Aspectos como la correcta tipificación de los datos y sus dominios, contar con un conjunto de reglas y técnicas que permitan la corrección, homologación y eliminación de valores erróneos, así como criterios de evaluación acerca de la bondad de las mismas son parte de lo que queremos proponer como marco de trabajo. Los datos a analizar son el resultado de un proceso llevado a cabo en el contexto de un proyecto de migración de base de datos de una empresa de servicios eléctricos. Se definen e implementan las mediciones de calidad a realizar sobre la base de datos, y se establece la forma de registrar los resultados de las mismas. Luego se definen e implementan procesos de limpieza (automáticos y semiautomáticos) con el fin de corregir los errores detectados. Para las implementaciones se utilizan sentencias SQL, obteniendo como resultado un Script que automatiza el ciclo completo de la calidad de datos: medición, registro y limpieza necesaria para llevar a cabo la migración de la base de datos.

Abstract

This project report describes the implementation of a methodological framework for the development of data quality projects. While a quality project data is full of casuistry and special process that make it very difficult to conceive of any kind of generalization, we wanted to capture in the development of this paper an outline of what might emerge as a methodology when addressing problems of this nature. Aspects such as the correct classification of the data and their domains, have a set of rules and techniques to enable correction, approval and removal of erroneous values and evaluation criteria of the goodness of these are part of what we want propose as a framework. The data analyzed are the result of a process carried out in the context of a migration project database of electric utility. They are defined and implemented quality measurements to be performed on the database, and how to record the results of the same set. Then they define and implement cleaning processes (automatic and semiautomatic) in order to correct the errors found. For implementations SQL statements are used, resulting in a script that automates the full cycle of data quality: measurement, registration and cleaning required to perform the migration of the database.

1. Introducción

Los sistemas de bases de datos se han convertido en elementos imprescindibles en la vida cotidiana de la sociedad moderna. Cada día, la mayoría de nosotros nos encontramos con actividades que requieren algún tipo de interacción con una base de datos. Por eso, es muy importante para todas las empresas poseer un sistema de bases de datos de calidad, en tanto la calidad de datos es uno de los fundamentos del éxito de las organizaciones. El tener acceso a información exacta y completa es fundamental para la toma de decisiones estratégicas y de misión crítica.

La calidad de la información en todo sistema informático resulta cada día más importante, fundamental, ya que es un factor de gran peso para cualquier actividad que se realice en base a dicha información. En la actualidad en todas las organizaciones, principalmente en las de servicios, se genera gran cantidad de datos, en base a los cuales se obtienen conclusiones acerca de cómo se están llevando a cabo sus procesos. Los datos con calidad apoyan y fortalecen virtualmente todas las funciones de los negocios y son especialmente importantes para las iniciativas que involucran servicios al cliente, comunicaciones y manejo de relaciones.

En este trabajo se presenta un caso de estudio en Calidad de Datos para el proceso de migración de una base de datos de una empresa de servicios eléctricos. El caso de estudio consiste en el análisis de errores a partir de reglas de aceptación de los datos entregados por parte de la organización. Los datos sobre los cuales se realizó el proceso de Calidad de Datos son parte de una base de datos de test en la que se encuentra replicada la base de datos productiva de la compañía, los defectos encontrados se guardan en una base de datos orientada a columnas (Sybase). Los datos resultantes del experimento deben ser analizados estadísticamente para poder concluir el nivel de errores que presentan los datos previos a la migración de los mismos con el fin de obtener resultados que reflejen más fielmente la realidad.

2. Motivación del Proyecto

La abstracción con la cual se define el concepto de calidad, puede hacer difícil la valoración de los resultados y beneficios que se obtienen a partir de la aplicación de distintas técnicas y/o actividades de la calidad. Tal como establece Robert Pirsig (filósofo, define la "metafísica de la calidad"), "Even though quality cannot be defined, you know what it is". A pesar de que no se puede definir la calidad, sabemos lo que es. Sabemos su importancia, sabemos lo que significa.

En los últimos años ha ido adquiriendo mayor relevancia, convirtiéndose en un aspecto fundamental que es necesario considerar para todo sistema de información. Dentro de las disciplinas que forman parte de la Ingeniería de Software, la Verificación ha ido tomando un papel de relevancia cada vez mayor.

El área de Calidad de Datos también ha tenido esta tendencia, debido a la creciente cantidad de información que se genera y almacena, incrementando también su valor e importancia para las organizaciones. La mala calidad de los datos influye de manera muy significante y profunda en la efectividad y eficiencia de cualquier organización, llevando en algunos casos a pérdidas multimillonarias. Cada día se hace más notoria la importancia y necesidad en distintos contextos de un nivel de calidad adecuado para los datos.

3. Descripción del Proyecto

3.1. Definición de Objetivos

3.1.1. Objetivo General

El objetivo general de este proyecto es asegurar la calidad de los datos en el proceso de migración desde un sistema a otro mediante la implementación de un proceso automático de análisis sobre el modelo de datos.

3.1.2. Objetivos Específicos

- Generar a partir de los resultados obtenidos y el estudio realizado una metodología para la implementación de un proceso de calidad de datos.

- Generar a partir de los resultados obtenidos y los estudios realizados una base de conocimiento de técnicas y criterios de calidad ampliable y aplicable a distintos escenarios.

  Desarrollar un conjunto de scripts y patrones de diseño asociados al proceso de calidad de datos parametrizables a distintos escenarios y requerimientos de análisis de los datos.

- Realizar la implementación de un proceso de Calidad de Datos para la migración de una base datos en una empresa de distribución de energía eléctrica.

4. Marco Teórico

Se denomina “error” a los errores que son encontrados en los datos con respecto a la evaluación de su calidad a partir de métricas dadas por el estudio, como cantidad de blancos, nulos, tipos de datos.

Un error es la instanciación de un tipo de error sobre un atributo y/o tabla específica. El término “tipo de error” se utiliza para definir conceptualmente un error genérico para determinado factor y dimensión de la calidad de los datos.

Se denomina “regla correcta” cuando al realizar la query SQL, esta entrega los datos que cumplen con la regla de negocio entregada por la empresa, dentro de esta “regla correcta” se maneja otro concepto que es el de “número de registros de regla correcta” que contempla el conteo de registros que cumple con la regla de negocio

Se denomina “regla incorrecta” cuando al realizar la query SQL, esta entrega los datos que no cumplen con la regla de negocio entregada por la empresa, dentro de esta “regla correcta” se maneja otro concepto que es el de “número de registros de regla incorrecta” que contempla el conteo de registros que no cumple con la regla de negocio.

4.1. Características de una base de Datos orientada a columnas

Las Bases de Datos Orientadas a Columnas son sistemas de bases de datos que tienen la característica de almacenar los datos en forma de columna. La ventaja principal de este tipo de sistema es que permite el acceso a grandes volúmenes de datos de forma rápida porque se puede acceder como una unidad a los datos de un atributo particular en una tabla. Un SMBD orientada a columnas es un sistema de gestión de bases de datos que almacena su contenido por columnas (atributos) y no por filas (registros) como lo hacen los SMBD relacionales.

Cada columna es almacenada contiguamente en un lugar separado en disco, usando generalmente unidades de lectura grandes para facilitar el trabajo al buscar varias columnas en disco. Para mejorar la eficiencia de lectura, los valores se empaquetan de forma densa usando esquemas de compresión ligera cuando es posible. Los operadores de lectura de columnas se diferencian de los comunes (de filas) en que son responsables de traducir las posiciones de los valores en locaciones de disco y de combinar y reconstruir, si es necesario, tuplas de diferentes columnas.[2]

Con este cambio ganamos mucha velocidad en lecturas, ya que si se requiere consultar un número reducido de columnas, es muy rápido hacerlo pero no es eficiente para realizar

escrituras. Por ello este tipo de soluciones es usado en aplicaciones con un índice bajo de escrituras pero muchas lecturas. Típicamente en data warehouses y sistemas de inteligencia de negocios, donde además resultan ideales para calcular datos agregados. Cabe resaltar que parte del auge actual que está provocando NoSQL se debe a la adopción de Cassandra (originalmente desarrollada por y para Facebook, luego donada a la fundación Apache) por parte de Twitter y Digg. Apache Cassandra es la base de datos orientada a columnas más conocida y utilizada actualmente.

4.1.1. Tiempo de carga

¿Cuánto tiempo se necesita para convertir datos de origen en el formato de Columna? Esta es la pregunta más básica de todas. Tiempos de carga son a menudo medidos en gigabytes por hora, que puede ser extremadamente lento, cuando de decenas o cientos de gigabytes de datos se trata. La cuestión a menudo carece de una respuesta sencilla, porque la velocidad de carga puede variar en función de la naturaleza de los datos y las elecciones realizadas por el usuario. Por ejemplo, algunos sistemas pueden almacenar varias versiones de los mismos datos, ordenados en diferentes secuencias o en los diferentes niveles de agregación. Los usuarios pueden construir un menor número de versiones a cambio de una carga rápida, pero puede pagar un precio más adelante con consultas más lentas. Pruebas realistas basadas en sus propios datos son el mejor camino para una respuesta clara.[1]

4.1.2. Carga Incremental

Una vez que un conjunto de datos se ha cargado, todo debe ser recargado cada vez que hay una actualización. Muchos sistemas columnares permiten carga incremental, teniendo solo los registros nuevos o modificados y la fusión de los datos anteriores. Pero la atención al detalle es fundamental, ya que las funciones de carga incremental varían ampliamente. Algunas cargas incrementales tardan hasta una completa reconstrucción y algunos resultados son el rendimiento más lento, algunos pueden agregar registros, pero no cambiar o suprimirlos. Las cargas incrementales a menudo deben completarse periódicamente con una reconstrucción completa.

4.1.3. Compresión de Datos

Algunos sistemas columnares pueden comprimir mucho la fuente de datos y archivos resultantes a fin de tomar una fracción de espacio en el disco original. Puede ocasionar en estos casos un impacto negativo en el rendimiento por la descompresión de datos a realizar la lectura. Otros sistemas utilizan menos compresión o almacenan varias versiones de los datos comprimidos, teniendo más espacio en disco, pero cobrando otros beneficios a cambio. El enfoque más adecuado dependerá de sus circunstancias. Tenga en cuenta que la diferencia de los requisitos de hardware pueden ser sustanciales.

4.1.4. Limitaciones Estructurales

Las bases de datos columnares utilizan diferentes técnicas para imitar una estructura relacional. Algunos requieren la misma clave principal en todas las tablas, es decir, la jerarquía de la base de datos está limitada a dos niveles. Los límites impuestos por un sistema en particular no parecen tener importancia, pero recuerde que sus necesidades pueden cambiar mañana. Limitaciones que parece aceptable ahora podrá evitar que la ampliación del sistema en el futuro.

4.1.5. Técnicas de Acceso

Algunas bases de datos de columnares solo se pueden acceder utilizando su propio proveedor de lenguaje de consultas y herramientas. Estos pueden ser muy poderosos, incluyendo capacidades que son difíciles o imposibles usando el estándar SQL. Pero a veces faltan funciones especiales, tales como las consultas que comparan valores con o en los registros. Si necesita acceder al sistema con herramientas basadas en SQL, determine exactamente que funciones SQL y dialectos son compatibles. Es casi siempre un subconjunto completo de SQL y, en particular, rara vez se dispone de las actualizaciones. También asegúrese de encontrar si el rendimiento de las consultas SQL es comparable a los resultados con el sistema de la propia herramienta de consulta. A veces, el ejecutar consultas SQL mucho más lento.

4.1.6. Rendimiento

Los sistemas columnares por lo general superan a los sistemas de relaciones en casi todas las circunstancias, pero el margen puede variar ampliamente. Las consultas que incluyen cálculos o acceso individual a los registros puede ser tan lento o más que un sistema relacional adecuadamente indexado.

4.1.7. Escalabilidad

El punto de las bases de datos columnares es obtener buenos resultados en grandes bases de datos. Pero no puede asumir todos los sistemas pueden escalar a decenas o centenares de terabytes. Por ejemplo, el rendimiento puede depender de determinados índices de carga en la memoria, de modo que su equipo debe tener memoria suficiente para hacer esto. Como siempre, en primer lugar preguntar si el vendedor tiene en ejecución los sistemas existentes a una escala similar a la suya y hablar con las referencias para obtener los detalles. Si el suyo sería más grande que cualquiera de las instalaciones existentes.

4.2. Base de Datos Orientadas a Columnas vs Orientadas a Filas

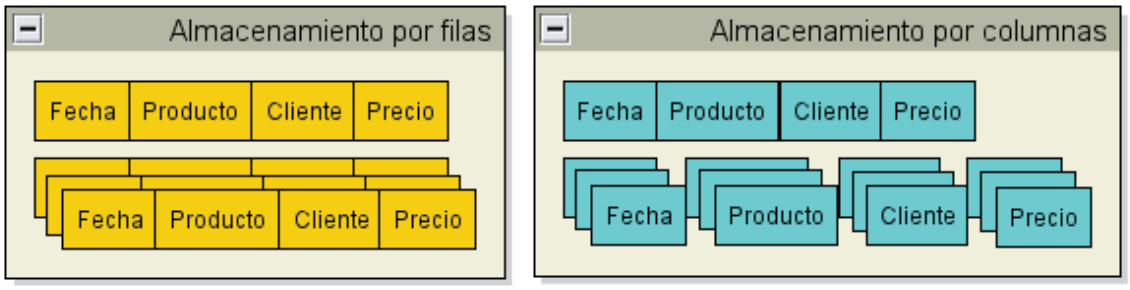

Ilustración 1 Almacenamiento Filas vs Columnas

La base de datos orientada a filas debe leer toda la fila con el fin de acceder a los atributos necesarios. Como resultado, las consultas analíticas y de inteligencia de negocios terminan leyendo más datos de lo necesario para satisfacer su consulta. Además este tipo de bases de datos habiendo sida diseñada para actividades transaccionales, es a menudo construida para la recuperación óptima y unión de conjunto de datos pequeños en lugar de

grandes, cargando así los subsistemas de entrada y salida que soportan el almacenamiento analítico. En respuesta, los administradores de base de datos tratan de ajustar el entorno de las diferentes consultas mediante la construcción de índices adicionales así como la creación de vistas especiales. Esto requiere mayor tiempo de procesamiento y consumo adicional de almacenamiento de datos.

Debido a que cada columna puede ser almacenada por separada, para cualquier consulta, el sistema puede evaluar las columnas que se están accediendo y recuperar solo los valores solicitados en las columnas específicas. En lugar de exigir los índices separados para las consultas de forma óptima los datos se valoran dentro de cada forma de columna del índice, reduciendo los sistemas de entrada y salida lo que permite un acceso rápido a los datos mejorando el tiempo y el rendimiento de las consultas.[3]

4.2.1. Ventajas

- La principal ventaja de este tipo de sistemas es el rápido acceso a los datos: esto ya lo hemos demostrado con el modelo DSM el cual nos permite consultar rápidamente los datos columna a columna, al guardarse físicamente de manera contigua.

- Un BBMS en una base de datos orientada a columnas, lee solo los valores de columnas necesarios para el procesamiento de una consulta determinada por lo cual las bases de datos orientadas a columnas tienen una mayor eficiencia en entornos de almacenes, donde las consultas, típicas incluyen los agregados realizados por un gran número de elementos de datos.

- Se comprime la información asignable de cada columna con el fin de mejorar el procesamiento desde el ancho de banda del acceso a disco.

- Cambios en el esquema tiene menor impacto y por lo tanto el coste de realizarlos es menor.

4.2.2. Desventajas

- No orientado a transacciones: este es el factor más débil de esta tecnología. El hecho de tener los datos guardados columna a columna nos permite retornarnos las filas más rápidamente, pero al insertar, actualizar o borrar un registro, se deberá hacer en más de una ubicación (al tener que actualizar todos los pares clave-valor asociados a una relación). Por esta razón, este tipo de bases de datos no se recomienda para sistemas de tipo OLTP orientados a transacciones y alta concurrencia.

- Reportes operacionales: también llamados reportes de seguimiento en los que se desea ver toda la información de una relación que puede contener muchas tuplas. En algunos casos esto puede resultar ineficiente comparado con los Row-Stores.

- No existe un modelo de datos que soporte teóricamente este modelo de base de datos.

- No existe un estándar que unifique los criterios de implementación de este modelo de base de datos.

5. Estado del Arte

5.1. Base de Datos Columnares y Calidad de Datos

5.1.1. Sybase

Sybase es una base de datos relacional basada en columnas que es intrínsecamente más apropiado para el adecuado procesamiento de consultas que un enfoque basado en filas. Debido a que está basado en columnas, Sybase IQ aprovecha las características de cada columna en la tabla, en un número de diferentes caminos.

Sybase soporta los esquemas relacionales tradicionales, incluyendo la normalización de esquemas usados para procesos de transacción. Como se puede ver Sybase incluye una API SQL que permite el acceso a SQL, también incluye ODBC, JDBC y XML, provee java para que pueda ser usado para escribir procedimientos almacenados y funciones de usuario.

Ofrece una serie de índices especializados para el adecuado rendimiento de las consultas.

Una consecuencia de utilizar el almacenamiento columnar en conjunción con la indexación de Sybase IQ Bit Wise, es que las agrupaciones pueden hacerse bajo la marcha. Dado que una parte significativa de extraer, trasformar y cargar es la anterior agrupación de transacciones. Compresión de datos es mucho más fácil de implementar en un enfoque basado en columnas que cuando se utilizan los métodos convencionales. Es significativamente más eficiente. En la práctica Sybase IQ ha demostrado una compresión de datos de un 50% a un 70% del conjunto de datos original. Es fácil agregar y cargar una columna de datos a una tabla como seria agregar una fila a una base de datos relacional convencional.

Un enfoque basado en columnas es mucho más fácil de mantener y requiere menos sintonización que un DWH convencional.

A parte de las características ya mencionadas, también apoya RCube, estructura plana que puede proveer importantes beneficios en comparación con los esquemas convencionales. En particular RCube puede acelerar significativamente la implementación, así como el rendimiento en tiempo de ejecución y proporcionar una mayor flexibilidad. Sybase ha sido creado para soportar el mayor número de consultas posible corriendo en paralelo en lugar de concentrarse en el uso del paralelismo para optimizar el rendimiento de una consulta en particular.

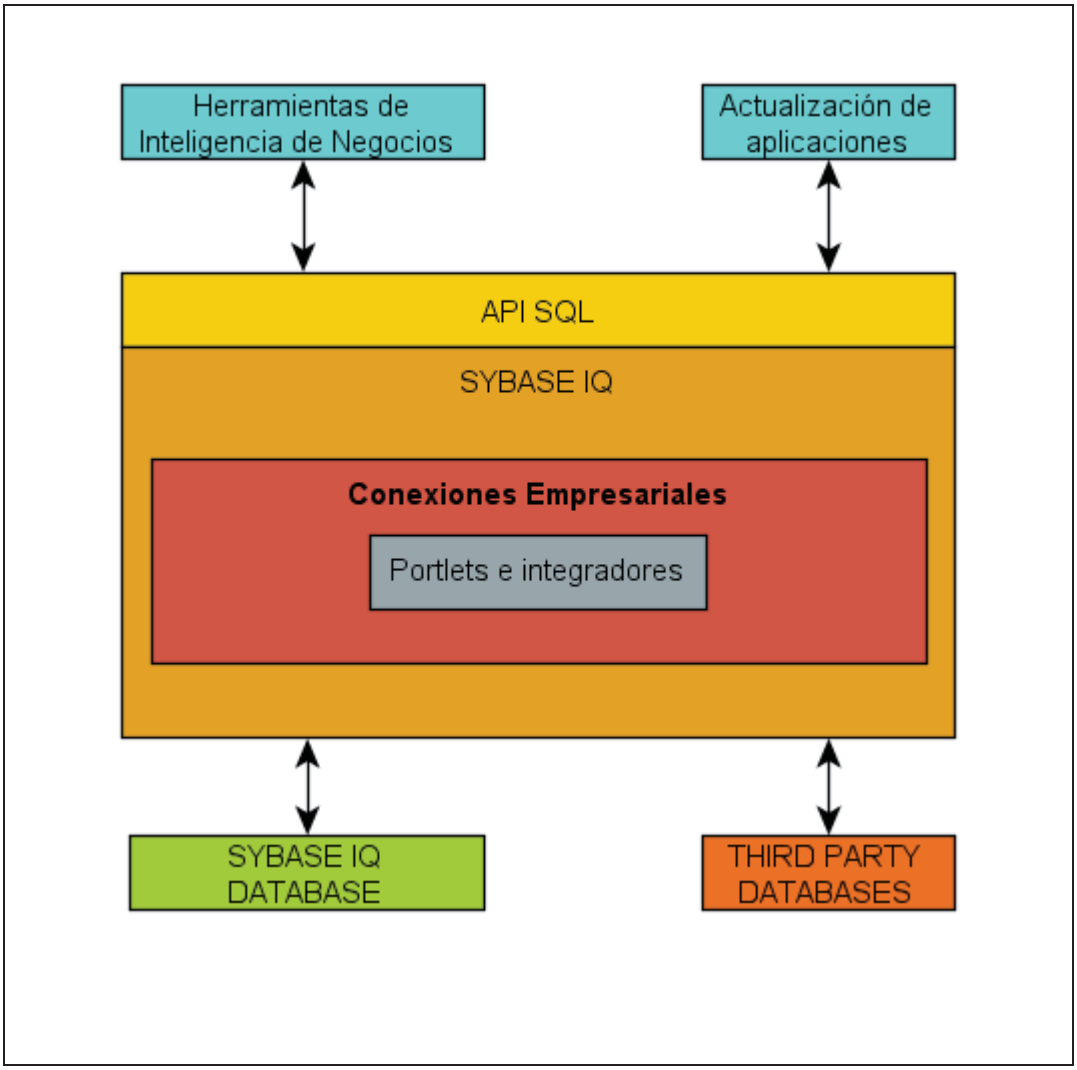

Ilustración 2 Componentes Sybase IQ

5.2. Inteligencia de negocios

La inteligencia de negocios se define como la habilidad corporativa para tomar decisiones. Esto se logra mediante el uso de metodologías, aplicaciones y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de conocimiento, los datos pueden ser estructurados para que indiquen las características de un área de interés, generando el conocimiento sobre los problemas y oportunidades del negocio para que pueden ser corregidos y aprovechados respectivamente.

Implementar herramientas de BI dentro de la organización permite soportar las decisiones que se toman; al nivel interno ayuda en la gestión del personal y del lado externo produce ventajas sobre sus competidores. Existen ocasiones en las cuales no se pueden lograr todos los beneficios que tiene BI; debido al proceso que lleva consigo implementar un proyecto de estas características, se puede cometer errores en la definición del planteamiento de las necesidades de conocimiento de la empresa; el no determinar la magnitud de los problemas de información a solucionar generalmente repercute en el fracaso del proyecto.

En la actualidad se está planteando un concepto nuevo llamado Agile BI Governance, el cual propone, arquitecturas, métodos y herramientas necesarios para implantar una infraestructura para BI. Esta definición, combina conceptos de IT Governance, Manifiesto Ágil y Data Governance, para lograr un alcance que contemple las diferentes unidades de negocio, y soporte el proceso estratégico de obtención de valor del Business Intelligence en la empresa. Permite conocer cómo controlar un sistema de estas características, qué políticas debo aplicar, qué métodos de control tengo que poner en marcha y cómo debo gobernar los sistemas de BI.

Agile BI Governance establece 4 valores básicos, pero dependiendo de cada organización puede incluir los que vayan en relación con su propia estrategia.

- Adaptabilidad Continúa. La incertidumbre y el cambio continuo son el estado natural de los sistemas de toma de decisiones, pero parece ser que muchas organizaciones aún no son conscientes de ellos. En este tipo de proyectos siempre se está cambiando el punto de vista analítico.

- Trabajo Conjunto. El usuario operativo del software ha de ser parte activa dentro de los grupos de IT que desarrollan los sistemas de BI.

- Jerarquías Flexibles. Los grupos de trabajo dentro del Agile BI Governance deberán estar estructurados con jerarquías flexibles que fomenten el intercambio de información.

- Personas Antes que Procesos. Priorizar la entrega de la información a las personas que controlan los procesos y no tanto en definir los procesos que han de controlar las personas.

5.2.1. Data WareHouse

Es el proceso de extraer datos de distintas aplicaciones (internas y externas), para que una vez depurados y especialmente estructurados sean almacenados en un depósito de datos consolidado para el análisis del negocio. Requiere una combinación de metodologías, técnicas, hardware y los componentes de software que proporcionan en conjunto la infraestructura para soportar el proceso de información. La estructura que se defina debe reflejar las necesidades y características del negocio, sus departamentos, equipos de trabajo y directivos, esto permitirá responder a interrogantes generados al tratar de tomar las decisiones y con el tiempo se va convirtiendo en la memoria corporativa, describiendo el pasado y el presente de la empresa. Data Warehouse desglosa, resume, ordena y compara, pero no descubre, ni predice.

Para la construcción de un Data Warehouse se establecen tres etapas; la primera está dedicada a examinar el esquema Entidad Relación de la base de datos operacional, generando los esquemas multidimensionales candidatos.

La segunda etapa, consiste en recoger los requisitos de usuario por medio de entrevistas, para obtener información acerca de las necesidades de análisis de estos, y la tercera etapa, contrasta la información obtenida en la segunda etapa, con los esquemas multidimensional candidatos formados en la primera etapa generando así, una solución que refleja los requisitos de usuario.[5]

Por otra parte implementar una solución de este tipo, ocasiona un costo que no todas las organizaciones están dispuestas a pagar (debido a sus capacidades de inversión), es por eso que los promotores del proyecto dentro de la empresa deben persuadir a los directivos y compañeros de trabajo, una buena alternativa de hacerlo es mediante el uso de técnicas administrativas, que permitan conocer a los directivos como se puede establecer el retorno de la inversión del proyecto equiparando inversión contra beneficios.

Al ser un depósito de datos consolidado para el análisis del negocio necesita tomar datos de distintas fuentes, Internas y Externas y como las características de las empresas son diferentes la cantidad de registros almacenados en algunas de ellas puede llegar a ser de proporciones exponenciales; es por esta razón que se necesita de procesos que optimicen los tiempos de extracción, transformación y transferencia de los datos del sistemas de información a la fuente de datos esto se logra implementando técnicas incrementales que mediante el uso de Snapshots y Triggers, se encarguen de sacar, transformar y transferir los registros que existen en el sistema de información a la fuente de datos.

El uso de Data Warehouse es tan amplio que llega a diferentes tipos de organizaciones y distintos temas de interés, puede ser implementado con conceptos Administrativos, en la administración; ayuda en la identificación de elementos de cambio que definan una nueva manera de hacer negocios, en donde la competencia debe estar orientada a trabajar no sólo de forma aislada, sino en colaboración con los diversos grupos de interés o actores de la industria, buscando referencias diferenciadoras para alcanzar el éxito, en empresas petroquímicas; incrementa la exactitud y precisión en la toma de decisiones con un 93.9% en la rentabilidad, en la Web optimiza búsqueda Web de metadatos con características semi-inteligentes y también suministra el soporte necesario para crear comunidades de colaboración científica, en transformadores de potencia; almacenando, la monitorización del estado del flujo de energía.

5.2.2. OLAP

El procesamiento analítico en línea permite obtener acceso a datos organizados y agregados de orígenes de datos empresariales, organiza subconjuntos de datos con una estructura multidimensional de manera que represente un significado especial o responda a una pregunta en particular. Estas herramientas soportan el análisis interactivo de la información de resumen, soportando muchas tareas de agrupación de datos que no pueden realizarse empleando las facilidades básicas de agregación y agrupamiento.[3]

5.2.3. Tipos de Sistemas OLAP

Tradicionalmente, este sistema se clasifica según las siguientes categorías:

- ROLAP. Implementación que almacena los datos en un motor relacional. Típicamente, los datos son detallados, evitando las agregaciones y las tablas se encuentran normalizadas.

- MOLAP. Esta implementación almacena los datos en una base de datos multidimensional. Para optimizar los tiempos de respuesta, el resumen de la información es usualmente calculado por adelantado.

- HOLAP (Hybrid OLAP). Almacena algunos datos en un motor relacional y otros en una base de datos multidimensional.[6]

Al igual que Data Warehouse, OLAP también es aplicable a un amplio rango de temas diferentes, uno de ellos es en Bases de Datos espaciales proporcionando características necesarias para los sistemas de tipo geográfico; como hechos, dimensiones, miembros, niveles, jerarquías, operaciones de navegación, operaciones de consolidación y comportamiento del clima. También se utiliza el almacenamiento MOLAP y ROLAP, para generar índices que mejoran los tiempos de accesos a las consultas de manera que los tiempos de entrega de la información demore el menor tiempo posible. Otra de las aplicaciones es en la educación al ser aplicado en ambientes de aprendizaje proporcionando las dimensiones y los indicadores necesarios para hacer la definición de un modelo de evaluación académica.[7]

5.3. Calidad De Datos

El objetivo de esta sección es abordar la temática de la calidad en los datos, llegando a conocer sus conceptos y características fundamentales, y sobretodo comprender su relevancia para nuestro proyecto. En primera instancia se introducen sus principales conceptos, las dimensiones y factores de calidad. Luego se explican las técnicas y actividades que se llevan a cabo en el área de la calidad de datos, y en línea con este último punto se trata la limpieza de datos, cuyo objetivo final es la mejora en la calidad de los mismos.

Previo a cualquier análisis de datos, es importante conocer acerca de la relevancia de la calidad de datos. Es por esto que se menciona de manera breve de qué trata la calidad de datos y el motivo por el cual resulta importante (por no decir imprescindible) su estudio.

Finalmente se trata cuáles son las áreas de investigación que le competen.[8]

5.3.1. ¿Qué es la Calidad de Datos?

Los datos representan objetos del mundo real. Estas representaciones resultan ser aplicables en diferentes contextos y con variadas características. Por otro lado, los datos pueden ser almacenados o sometidos a algún proceso o transformación, siendo siempre de suma importancia para garantizar la sobrevivencia y éxito de las organizaciones.

El problema de la calidad de datos ha sido objeto de estudio desde varias perspectivas y por diferentes áreas a lo largo de los años, como la ha si do la Estadística, Gestión o Computación. A medida que su importancia se hace más evidente a los ojos de estas y otras áreas, se incrementan también las investigaciones e intenciones de mejora en este sentido. Es

indudable que el almacenamiento y/o procesamiento de datos es de vital importancia en la vida de todas las personas y organizaciones, en una gran variedad de actividades (más allá de la informática y los sistemas de información).

Existen varios ejemplos de situaciones de la vida cotidiana, donde se hace necesario almacenar, procesar, transmitir y utilizar datos. Uno de ellos, cuando elaboramos una lista para hacer las compras almacenamos datos correspondientes a qué productos comprar, en qué cantidad, de qué marca. En cuanto al concepto de calidad de datos, suele suceder que intuitivamente se piensa en ciertos aspectos de los datos.

Por lo general se tiende a pensar en que los datos sean exactos. Sin embargo, hace falta ahondar más en este concepto, para entender que hay varias "caras" o aspectos (las llamadas dimensiones), que hacen a la calidad de los datos. Más adelante en el documento se explican algunas dimensiones (exactitud, completitud, actualidad, entre otras) en detalle.

Como ejemplo trivial, se puede pensar en la situación de la elaboración de una lista para compras:

- Si se omite anotar un producto o la cantidad a comprar de cierto producto, se enfrenta el problema de completitud.

- Si ocurre una equivocación en la cantidad de cierto producto o se escribe mal su marca, se enfrenta el problema de exactitud.

- Si en lugar de llevar la lista de hoy se lleva la de ayer, se enfrenta el problema de actualidad.

Entonces, se puede decir que la definición de la calidad de los datos está relacionada estrechamente con la exactitud, completitud, consistencia y actualidad de los datos (entre otros). Es por esto que la calidad de datos es denominada un concepto "multifacético", ya que depende y es función de las dimensiones que la definen.

5.3.2. La Importancia de la Calidad de Datos

Son pocas las ocasiones en las cuales se es consciente de las consecuencias que la mala calidad de datos trae aparejada. Sin embargo, es de suma importancia lograr identificar sus causas para eliminar, o en su defecto mejorar, la problemática de raíz.

En el ejemplo anterior de elaboración de la lista de compras, la mala calidad de los datos puede acarrear consecuencias no deseadas (como omitir comprar un producto que se necesitaba, o una cantidad equivocada), ninguna de ellas de gravedad. Pero no es difícil pensar en otro tipo de situaciones (listas de productos para importación en cantidades masivas, nombres de clientes duplicados, errores en cobros, error en mediciones eléctricas,) donde una falta puede provocar problemas de gravedad.

La mala calidad de los datos influye de manera muy significante y profunda en la efectividad y eficiencia de las organizaciones así como en todo el negocio, llevando en algunos casos a pérdidas multimillonarias. Cada día se hace más notoria la importancia y necesidad en distintos contextos de un nivel de calidad adecuado para los datos.[8]

5.3.3. Áreas de Investigación en Calidad de Datos

Lograr calidad en los datos es una tarea compleja y multidisciplinaria, debido a su importancia, naturaleza, y la variedad de tipos de datos y sistemas de información que pueden estar implicados. La investigación dentro del área de calidad de datos incluye los siguientes puntos:

- **Dimensiones:** las mediciones sobre el nivel de calidad de los datos se aplican a las dimensiones de interés.
- **Metodologías:** proveen guías de acción.
- **Modelos:** representan las dimensiones y otros aspectos de la calidad de datos.
- **Técnicas:** proveen soluciones a problemas de calidad de datos.
- **Herramientas:** son necesarias para que las metodologías y técnicas puedan llevarse a cabo de manera efectiva.

5.3.4. Dimensiones de la Calidad de Datos

En la sección anterior, se introdujeron a modo de ejemplo conceptos como exactitud, completitud y actualidad. Todas estas características (y varias más) de los datos, se denominan dimensiones de la calidad de los datos.

Cada **dimensión** refleja un aspecto distinto de la calidad de los datos. Las mismas pueden estar referidas a la extensión de los datos (su valor), o a la intensión (su esquema). De esta

manera podemos distinguir entre calidad en los datos y calidad en los esquemas. El foco del presente proyecto es en la calidad inherente a los datos.

Se define **factor de calidad** como un aspecto particular de una dimensión. En este sentido, una dimensión puede ser vista como un agrupamiento de factores de calidad que tienen el mismo propósito.

Es claro que la mala calidad en los datos puede provocar varios problemas, así como también la mala calidad de un esquema (por ejemplo un esquema de una base de datos relacional sin normalizar) podría provocar problemas mayores, tales como redundancias.

Ambos tipos de dimensiones, tanto las referidas a los datos como a los esquemas, proveen una visión cualitativa de la calidad, mientras que las medidas cuantitativas se representan mediante las métricas.

Una **métrica** es un instrumento que define la forma de medir un factor de calidad, un mismo factor de calidad puede medirse con diferentes métricas. Por otro lado, definimos **método de medición** como un proceso que implementa una métrica. A su vez, una misma métrica puede ser medida por diferentes métodos.

Existen varias dimensiones que reflejan distintos aspectos de los datos. Esto no resulta ser una sorpresa al considerar que los datos pretenden representar todo tipo de características de la realidad, desde espaciales y temporales, hasta sociales. A continuación se describen algunas dimensiones de la calidad de datos.

5.3.5. Exactitud y Unicidad

La exactitud se puede definir como la cercanía que existe entre un valor v del mundo real, y su representación v'.

De acuerdo al enfoque teórico que se trata más adelante, la exactitud se define como una correcta y precisa asociación entre los estados del sistema de información y los objetos del mundo real.

Existen tres factores de exactitud:

- **Exactitud sintáctica** en la cual se refiere a la cercanía entre un valor v y los elementos de un dominio D. Esto es, si v corresponde a algún valor válido de D (sin importar si ese valor corresponde a uno del mundo real). Para poder medir la exactitud sintáctica se puede utilizar la comparación de funciones, métrica que mide la distancia entre un valor v y los valores en el dominio D. Otras alternativas posibles son la

utilización de diccionarios que representen fielmente el dominio, o el chequeo de los datos contras reglas sintácticas.[8]

- **Exactitud semántica** se refiere a la cercanía que existe entre un valor v y un valor real v'. Esta dimensión se mide fundamentalmente con valores booleanos (indicando si es un valor correcto o no), para lo cual es necesario conocer cuáles son los valores reales a considerar.

En este caso, interesa medir que tan bien se encuentran representados los estados del mundo real. Una de las métricas utilizadas es la comparación de los datos con referenciales considerados válidos.

- **Precisión** por otra parte, se refiere al nivel de detalle de los datos.

El enfoque hasta ahora ha sido en la exactitud a nivel de valores, o sea, del valor de una celda (o campo) de una tupla. Sin embargo, es posible pensar en la exactitud a nivel de tupla, o a nivel de 8 tablas, e incluso considerando la base entera. Es decir, se pueden considerar distintos niveles de granularidad a la hora de evaluar la calidad de los datos. Es por esto que se definen funciones de agregación, las cuales miden la exactitud de conjuntos de datos. Por ejemplo, obtener la medida de una tupla a partir de la medida de exactitud de cada una de sus celdas. El ratio es una función de agregación que consiste en identificar la cantidad de valores correctos sobre la cantidad de valores totales. Brinda un porcentaje de valores correctos. Otros ejemplos de funciones de agregación son los promedios y promedios ponderados.

Para aclarar los conceptos se plantea un ejemplo sencillo. Se posee una base de datos donde se almacena el nombre y la edad de determinadas personas. Para el dato "Edad" se especifica que su valor estará en el rango 0 a 120. Además, se sabe que existe una persona llamada Oscar Javier Morales, de 23 años de edad.

Se consideran entonces los siguientes casos:

- Si existe un registro para una persona donde el campo edad tiene el valor 234, entonces se trata de un error sintáctico (valor fuera del rango 0 a 120).

- Si existe un registro para Oscar donde el campo edad tiene el valor 19, entonces se trata de un error semántico, ya que es sabido que Oscar no tiene 19 años, sino que tiene 23 (en este caso no hay error sintáctico, pues 19 es un valor válido para la edad).

- Se enfrenta un problema de precisión si existe el interés de conocer la edad exacta de Oscar, ya que solo se conoce la cantidad de años, no los meses ni días de vida.

A pesar de que la exactitud semántica es generalmente más compleja de medir que la exactitud sintáctica (ya que se requieren conocer los valores del mundo real), cuando ocurren

errores de tipeo ambos tipos de exactitud coinciden. Al modificar su valor, se logrará exactitud sintáctica, ya que el valor escrito correctamente se corresponderá con alguno del dominio, y semántica, ya que existirá un valor real asociado al valor escrito correctamente.

Una forma de chequear la exactitud semántica es comparar diferentes fuentes de datos, y encontrar a partir de estas el valor correcto deseado. Esto también requiere de la resolución del problema de identificación de objetos, el cual consiste en identificar si dos tuplas representan el mismo objeto en el mundo real.

En el caso en que la exactitud sea considerada en un conjunto de valores, es necesario considerar también la duplicación. Dicha problemática ocurre cuando un objeto del mundo real se encuentra presente más de una vez (más de una tupla representa exactamente el mismo objeto).

Sin embargo, podrían existir también tuplas que representan el mismo objeto del mundo real pero con diferentes claves. Este aspecto es considerado por la dimensión de Unicidad. Es importante destacar aquí que existen diferentes situaciones que pueden llevar a la duplicación de datos:

- cuando la misma entidad se identifica de diferentes formas.
- cuando ocurren errores en la clave primaria de una entidad.
- cuando la misma entidad se repite con diferentes claves.

Distinguimos dos factores de la dimensión Unicidad:

Duplicación la misma entidad aparece repetida de manera exacta.

Contradicción la misma entidad aparece repetida con contradicciones.

5.3.6. Completitud

La completitud se puede definir como la medida en que los datos son de suficiente alcance y profundidad.

De acuerdo al enfoque teórico, esta dimensión se define como la capacidad del sistema de información de representar todos los estados significativos de una realidad dada.

Existen dos factores de la completitud

- **Cobertura** se refiere a la porción de datos de la realidad que se encuentran contenidos en el sistema de información. Al igual que para la exactitud semántica, la cobertura involucra una comparación del sistema de información con el mundo real. Una vez

más un referencial es requerido. Debido a que suele ser difícil obtenerlo, otra alternativa es estimar el tamaño de tal referencial.

- **Densidad** se refiere a la cantidad de información contenida, y la faltante acerca de las entidades del sistema de información.

Completitud de Datos Relacionales la completitud en un modelo relacional puede caracterizarse por los siguientes aspectos:

- **Valores nulos:** el significado de los valores nulos puede ser variado. Un valor nulo puede indicar que dicho valor no existe en el mundo real, que el valor existe en el mundo real pero no se conoce, o que no se sabe si el valor existe o no en el mundo real. Es importante conocer la causa de su presencia.

- **Suposiciones:**
    - CWA (Suposiciones del Mundo Cerrado, Closed World Assumption): todos los valores del mundo real se encuentran en el modelo relacional. En un modelo CWA con valores nulos, la completitud se define a partir de la granularidad de los elementos del modelo (completitud del valor, de la tupla, de un atributo, o de la relación).
    - OWA (Suposiciones del Mundo Abierto, Open Worl Assumption): no se puede asegurar que todos los valores del mundo real se encuentran en el modelo relacional. En un modelo OWA sin valores nulos, la completitud se mide como la cantidad de tuplas representadas en la relación sobre su tamaño total (la cantidad de objetos del mundo real que constituye la totalidad de la relación).

Por ejemplo, si se requiere tener registrados en una base de datos los datos (nombre, edad y sexo) de todas las personas que habitan en el planeta Tierra, entonces cada persona no registrada en la base degradará la completitud de los datos (esto sería completitud a nivel de la relación). También se verá disminuida la completitud si no se cuenta con la edad de ciertas personas, o con su sexo (esto último se refiere a la completitud a nivel de tupla o registro).

5.3.7. Dimensiones Relacionadas con el Tiempo

Los cambios y actualizaciones de los datos son un aspecto importante de la calidad de datos a tener en cuenta. Es posible afirmar que en determinados contextos un dato no actualizado es de mala calidad y puede llegar a ocasionar problemas graves.

Como ejemplo, suponer que se planean unas vacaciones a una isla del Caribe. Además de los preparativos correspondientes, se verifica el pronóstico del clima para asegurar que no ocurran huracanes en los días que se estará allí. Si la información climática no fue debidamente actualizada (por ejemplo si se consulta una página web que no posee mantenimiento), puede que se esté recibiendo el pronóstico equivocado, y por ende, que se estropeen las vacaciones. Por lo tanto, el pronóstico podría ser muy completo y exacto desde el punto de vista de la información climática que brinda, pero si es antiguo de nada serviría.[8]

Se describen las siguientes dimensiones relacionadas con el tiempo:

- **Actualidad:** trata sobre la actualización de los datos y su vigencia. Esta dimensión puede ser medida de acuerdo a la información de "última actualización".

- **Volatilidad:** se refiere a la frecuencia con que los datos cambian en el tiempo. Una medida para esta dimensión es la cantidad de tiempo que los datos permanecen siendo válidos.

- **Edad:** especifica qué tan actuales/viejos son los datos para la tarea/evento en cuestión. Para medir esta dimensión es necesario considerar una métrica de actualidad, y verificar que los datos se encuentren dentro del límite establecido por la tarea/evento en cuestión.

5.3.8. Consistencia

Esta dimensión hace referencia al cumplimiento de las reglas semánticas que son definidas sobre los datos.

De acuerdo al enfoque teórico, la inconsistencia de los datos se hace presente cuando existe más de un estado del sistema de información asociado al mismo objeto de la realidad. Una situación que podría ocasionar inconsistencias en los datos es la incorporación de datos externos o con otros formatos.

Un ejemplo sencillo: si en una tabla se almacenan datos de personas, tales como fecha de nacimiento y edad, entonces si en un registro se tiene como fecha de nacimiento el 01/01/2005 y como edad 42 años, existe una inconsistencia (como se explica a continuación, se estaría violando una regla intra-relacional).[9]

- **Restricciones de integridad**

Las restricciones de integridad definen propiedades que deben ser cumplidas por todas las instancias de un esquema relacional.

Se distinguen tres tipos de restricciones de integridad:

- **Restricciones de dominio:** se refiere a la satisfacción de reglas sobre el contenido de los atributos de una relación.

- **Restricciones intra-relacionales:** se refiere a la satisfacción de reglas sobre uno o varios atributos de una relación.

- **Restricciones inter-relacionales:** se refiere a la satisfacción de reglas sobre atributos de distintas relaciones.

Existen además diferentes tipos de dependencias:

- **Dependencias de clave:** no existen dos instancias de una relación r con la misma clave k.

- **Dependencias de inclusión (restricciones referenciales):** algunas instancias de la relación r están contenidas en instancias de otra relación s. Un ejemplo de esta dependencia son las restricciones de clave foránea.

- **Dependencias funcionales:** una relación r satisface la dependencia funcional $X \rightarrow Y$ si para todo par de tuplas t1 y t2 se cumple que: Si : $t_1.x = t_2.x \rightarrow t_1.y = t_2.y$

5.3.9. Relaciones entre las Dimensiones

Es claro que las dimensiones no son independientes entre sí, sino que se interrelacionan de manera estrecha. Es necesario ser cuidadoso a la hora de invertir esfuerzo en mejorar un aspecto (dimensión) de la calidad de datos, ya que podría estar afectando negativamente otro aspecto de estos.

En línea con lo mencionado anteriormente, dependiendo del contexto particular en el cual nos situemos elegiremos mejorar aquellas dimensiones que consideramos de mayor valor para la calidad de nuestros datos, e ignorar las que no la perjudican o afectan de manera significativa.

A modo de ejemplo, se mencionan algunas de las relaciones negativas más comunes entre diferentes dimensiones de la calidad de datos:

- Datos exactos, completos o consistentes podría implicar su desactualización debido al tiempo que es necesario invertir en actividades de chequeo y corrección.
- La completitud (muchos datos) tiene mayores probabilidades de acarrear errores de inconsistencia en los datos.

Sin embargo, también existen correlaciones positivas, esto es, que mejoran más de un factor. Es importante identificar en primera instancia cuáles son los factores o dimensiones que se requiere mejorar de acuerdo al contexto de aplicación, para luego evaluar si es posible realizarlo de forma conjunta.[9]

A modo de ejemplo, mencionamos algunas de las correlaciones positivas más comunes entre diferentes factores de la calidad de datos:

- La corrección de errores de tipeo mejora tanto la exactitud semántica como sintáctica.
- Si se logran obtener datos más actualizados, se podría mejorar la exactitud semántica (más datos corresponderían a la realidad).
- Si se completan los valores nulos (densidad) también se podría mejorar la exactitud semántica.

5.3.10. Enfoque en las Dimensiones de la Calidad de Datos

A continuación se definen tres enfoques distintos que es posible adoptar con respecto a las definiciones de las dimensiones en Calidad de Datos.

Enfoque Teórico

Este enfoque considera la correcta representación de la realidad en un sistema de información. En este aspecto, interesa conocer las deficiencias que se generan cuando ocurren desviaciones en dicha representación. Dentro de las deficiencias relativas al diseño del sistema de información, se destacan las siguientes:

- **Representación incompleta:** cuando un objeto del mundo real no se asocia con ningún estado del sistema de información.

- **Representación ambigua:** cuando varios objetos del mundo real se asocian con el mismo estado del sistema de información.

- **Representación sin significado:** cuando existen estados del sistema de información que no se encuentran asociados con ningún objeto del mundo real.

En lo que respecta a las deficiencias operacionales destacamos los errores (garbling), que se refieren a una incorrecta asociación entre los objetos de la realidad y los estados del sistema de información.

Enfoque Empírico

En este caso la información es obtenida a partir de entrevistas, cuestionarios y experimentos.

Se destacan cuatro categorías:

- **Calidad de Datos intrínseca:** calidad que los datos deben tener por sí sola (ejemplo: exactitud).

- **Calidad de Datos contextual:** toma en cuenta el contexto en que los datos son utilizados (ejemplo: completitud).

- **Calidad de Datos representacional:** referente a la calidad de la representación de los datos (ejemplo: interpretación).

- **Calidad de Datos:** para la accesibilidad de los mismos.

Enfoque intuitivo

Las dimensiones son definidas de acuerdo al sentido común y la experiencia práctica.

Se destacan tres categorías:

- Esquema conceptual
- Valor de los datos
- Formato de los datos.

5.4. Técnicas y Actividades de Calidad de Datos

En esta sección se explican algunas actividades y técnicas desarrolladas para mejorar la calidad de los datos.

Las actividades relativas a la calidad de datos se refieren a cualquier proceso (o transformación) que se aplica a los datos con el objetivo de mejorar su calidad. Para llevar a cabo dichas actividades, se hace uso de distintas técnicas.[9]

A continuación se describen algunas actividades relativas a la calidad de los datos:

- **Obtención de nueva información**: es el proceso de refrescar la información almacenada en la base con datos de mayor calidad (por ejemplo ingresar datos más precisos, de mayor actualidad).

- **Estandarización:** es el proceso de "normalizar" los datos almacenados, de manera que queden almacenados respetando cierto formato (por ejemplo todos los números de teléfono deben incluir el código de región).

- **Identificación de Objetos:** es el proceso por el cual se identifican registros (dentro de una misma tabla, o entre tablas) que hacen referencia al mismo objeto de la realidad.

- **Integración de datos:** hace referencia a la actividad de unificar datos provenientes de distintas fuentes, resolviendo los problemas que esto trae aparejados (redundancias, problemas de consistencia, duplicación).

- **Confiabilidad de las fuentes:** implica "calificar" a las distintas fuentes de información de acuerdo a la calidad de los datos que proveen (esto tiene más sentido considerando un sistema P2P por ejemplo).

- **Composición de calidad:** hace referencia a la definición de un álgebra para calcular la composición (o agregación) de las medidas de las dimensiones de calidad de datos. Por ejemplo, calcular la completitud de una unión de relaciones, a partir de la completitud de cada relación.

- **Detección de errores:** dadas una o más tablas, y ciertas reglas que los registros de dichas tablas deben cumplir, este es el proceso de detectar qué registros no cumplen con dichas reglas.

- **Corrección de errores:** luego de la detección, esta actividad se encarga de corregir los registros con errores, de manera que se respeten todas las reglas correspondientes.

- **Optimización de costos:** implica obtener la mejor relación costo-beneficio al aplicar procesos de mejora de la calidad de los datos.

6. Caso de Estudio

Durante la última década hemos asistido al surgimiento de grandes temas de investigación en el área de la gestión de la información, el análisis de los datos, la gestión del conocimiento y los sistemas decisionales. Si bien el foco se encuentra hoy en aprovechar la gran masa de datos que las plataformas de información generan en cada una de las áreas productivas de las empresas, el constante avance de la tecnología y los esfuerzos que hacen hoy las compañías por mejorar sus procesos internos y mantener su plataforma tecnológica al día dan lugar a continuos proyectos de renovación de hardware y software.

En la mayoría de los casos, uno de los requisitos fundamentales de estos procesos de modernización de la plataforma tecnológica es la migración de los datos históricos de la operación de la compañía. Y cuando hablamos de migración de los datos históricos no nos referimos a un mero traspaso desde las estructuras antiguas a las nuevas. Muchas veces la migración de los datos implica un proceso de revisión ya que se aprovecha de auditar la historia, tipos de datos (dado que muchas veces las nuevas tablas y estructuras tienen tipos de datos distintos), categorizaciones y conversiones puesto que no pocas veces junto con el cambio de plataforma tecnológica viene todo un cambio a nivel de procesos y conceptualizaciones en las unidades de negocio. En este contexto es el que se sitúa nuestro trabajo.

Una importante empresa de energía de nuestra región se encuentra en proceso de actualización y cambio de su plataforma tecnológica. Ha sido un proyecto de largo aliento (casi 1 año y medio) y a meses de su salida a producción han comenzado los trabajos de calidad de los datos. Los trabajos de evaluación de la calidad de los datos tienen un doble objetivo:

- En primer lugar se trata de auditar que los instructivos de la rigurosa normativa de regulación de nuestro país referente a los procesos de tarificación y facturación de los servicios eléctricos se hayan implementado de manera adecuada en la nueva plataforma tecnológica, produciendo los datos esperados (tipos y dominios).

- Dependiendo de la tarifa a la que está asociado un cliente determinado, su consumo histórico (promedio del último período especificado por la normativa) es una variable en el cálculo de su facturación presente. Es por este motivo que, en segundo lugar, se debe garantizar que los datos históricos han sido correctamente almacenados en la nueva estructura para que puedan seguir siendo utilizados en el cálculo actual de las tarifas de los clientes de la empresa.

El sistema desde el cual se migra es un desarrollo hecho “en casa” y que soporto por muchos años la operación de la compañía, tanto en su casa matriz como en sus 4 filiales. El sistema se desarrolló sobre una base de datos transaccional SYBASE ASE.

La plataforma que llega a suplir al antiguo sistema es Open Smartflex. Se trata de un sistema que integra ventas, gestión de pedidos, quejas y reclamaciones, facturación, cobranzas, inventario geográfico y operaciones de campo, sobre una plataforma de software unificada y flexible.

Las estructuras de datos que caen dentro del alcance de este proyecto son descritas en detalle en la sección “6.6.1.2. Entidades a Analizar” así como las principales reglas de validación entregadas por la compañía como índices de bondad de la data.

7. Desarrollo de la solución

7.1. Análisis previo a la Implementación

A continuación de describe el proceso previo a la implementación de modelo de Calidad de Datos.

Se comenzará con la definición tanto de los objetivos como de los límites del alcance del proyecto, lo que se llevará en conjunto con el equipo de la empresa donde se llevó a cabo la implementación.

Se realizará un análisis de las herramientas a utilizar como los clientes del motor de base de datos y la forma conexión a la red corporativa de la empresa.

También se recopilara la información (reglas de validación), que se utilizarán en la verificación de los datos para llevar a cabo el proceso de calidad de los datos.

Por último, se especificará una metodología y se realizará una evaluación de los datos para precisar las métricas que se utilizarán en el análisis de los resultados

7.2. Objetivos de la implementación

El objetivo principal de un proceso de Calidad de Datos es la evaluación y control de los datos para mejorar la rentabilidad, seguridad y eficacia del sistema, así como también la comprobación de la bondad de los mismos para análisis como de desempeño del motor de base de datos y generación de reportes para la toma de decisiones.

El proyecto tiene como elementos de estudio la verificación de los datos para así describir debilidades y disfunciones de las bases de datos de la empresa.

Posteriormente se presentan algunas sugerencias y planes de acción para eliminar las disfunciones y debilidades encontradas anteriormente.

7.2.1. Objetivos Generales de la implementación

El proyecto de Calidad de Datos tiene como principal objetivo medir la calidad de los datos que se encuentran en el universo a migrar desde una Base de Datos Sybase (antigua) a una Oracle (nueva), según las especificaciones definidas y entregadas por los administradores del nuevo motor de base de datos. La bondad de estos datos se determina mediante la evaluación de métricas y reglas especificadas tanto por la empresa como por el equipo que participo en la implementación del proyecto.

7.2.2. Objetivos Específicos de la implementación

Para alcanzar el objetivo principal del proyecto mencionado anteriormente, es necesario que se cumplan los siguientes objetivos específicos:

- Mediante la lectura y estudio de documentos de metodologías, preparar y desarrollar una metodología a utilizar en implementación del proceso evaluación de calidad de datos.

- Proponer métricas informáticas para realizar la evaluación de las distintas tablas involucradas en el proceso.

- Estudiar y analizar las reglas de validación de datos consideradas hasta la fecha por la compañía.

- Desarrollar una estructura de análisis que sea capaz de almacenar los datos necesarios y ejecute las métricas propuestas para la evaluación del sistema Sybase de la compañía.

- Analizar los resultados obtenidos tras la evaluación de los datos, para la generación de un reporte final, entrega de recomendaciones y conclusiones.

7.3. Alcance de la implementación

Este proyecto pretende medir y cuantificar el nivel de error en la calidad de los datos contenidos en la base de datos Sybase de la compañía, para poder realizar la migración de éstos a su nuevo motor de base de datos Oracle.

Se medirá este nivel de error a través de métricas definidas por el grupo de trabajo a cargo del proceso de calidad de datos, y mediante una serie de reglas de negocio definidos por la empresa que se deben cumplir, para lograr una migración exitosa a la nueva plataforma.

Este examen a los datos almacenados en la base Sybase, tiene una duración aproximadamente de dos meses y será realizada por el grupo de implementación del proceso.

Este análisis de calidad de datos, se realiza mediante la evaluación de las reglas entregada por parte de la compañia, a través de *querys* SQL. Estas *querys* cumplen un rol binario, es decir, los datos a analizar cumplen o no cumplen con las reglas, será esto lo que se cuantificará.

7.4. Metodología

En el contexto empresarial actual resulta ser común encontrarse con grandes depósitos de datos heterogéneos originados por los múltiples sistemas operacionales presentes en cada una de las áreas productivas de la organización. Es frecuente también observar diversos intentos por organizar esta enorme cantidad de datos bajo el alero de proyectos de inteligencia de negocios, data warehouse corporativos, data mining, big data, data discovery, data science, gestión del conocimiento y un largo etc. de términos que no dejan de acuñarse, conceptos que sin duda tienen un propósito bien definido en su concepción original, propósito que lamentablemente muchas veces no alcanza a concretarse en los tiempos estimados ni en la forma deseada debido principalmente a la carencia de entendimiento de las reales implicancias que este tipo de proyectos trae asociadas consigo, elementos que van desde cambios en la cultura organizacional hasta reingeniería de procesos. Una de las actividades más olvidadas y dejadas de lado a la hora de la evaluación e implementación de proyectos de esta naturaleza es el análisis de la calidad de los datos y las posteriores acciones que deben tomarse para su corrección.

Ya en el 2003, Dasu, et al. [10] afirmaban que "es común que las bases de datos tengan entre un 60% y 90% de problemas en los datos". En la misma línea, Gartner señalaba en el 2007 que más del 25% de los datos críticos de las compañías presentan algún tipo de error.

Si atendemos al principal objetivo de la implementación de este tipo de proyectos en las organizaciones, el cual es por supuesto, mejorar el proceso de toma de decisiones ya no solo estratégicas, sino también tácticas y operacionales, podemos concluir sin mucho esfuerzo que un inadecuado manejo de la calidad de los datos llevará ineludiblemente a un deterioro en la calidad de las decisiones tomadas en la empresa en toda la cadena de valor.

Existen diversas categorías de errores asociadas a la calidad de los datos y en coherencia con esto han surgido múltiples enfoques y técnicas para detectarlos y corregirlos. Conceptos como calidad de los datos (data quality), heterogeneidad de los datos (data heterogeneity), limpieza de los datos (data cleaning) o reconciliación de los datos (data reconciliation) son los que se pueden encontrar más frecuentemente en la literatura sobre el tema.

Si bien no existe una taxonomía aceptada por todos, el trabajo realizado por Oliveira et. al. [11] ha sido referencia para nuestro trabajo ya que no sólo realizan una taxonomía con treinta y cinco problemas de calidad de los datos, sino que plantean métodos semiautomáticos para detectarlos, los cuales representan mediante árboles binarios.

Las técnicas desarrolladas por los investigadores hasta el momento, son variadas y casi siempre aplican a un tipo de problema en particular. Es así como existen técnicas para tratar el problema de la detección de duplicados, para detección y corrección de valores atípicos, para tratar con los valores faltantes y para cada posible problema que puedan presentar los datos, hecho del cual, podemos desprender una primera conclusión, evidente pero muy importante a la hora de plantearnos a definir una metodología para nuestro trabajo: la calidad de la limpieza lograda sobre los datos depende de la técnica aplicada y la elección de la técnica está íntimamente ligada con la naturaleza de los datos específicos sobre los que se está trabajando.

Ahora bien ¿Cómo determinar las técnicas que deben ser empleadas para realizar procesos de depuración a los datos en un caso particular? Ya que sabemos que ninguna de las técnicas cubre todos los posibles aspectos de los diversos tipos de datos, en consecuencia estableceremos los pasos a seguir para determinar cuál de todas es la que mejor se adapta a la problemática que nos toca enfrentar:

- **Entendimiento de la naturaleza de los datos a evaluar**. Se trata en este punto de tomar nota de todos los aspectos relacionados con los datos que vamos a analizar, como su tipo, dominio, granularidad, cardinalidad, dispersión, etc.

- **Identificación de las técnicas a aplicar**. De acuerdo a la naturaleza de los datos se debe escoger la técnica a aplicar. Es evidente que esto conlleva un amplio

conocimiento de las técnicas desarrolladas hasta el momento y de su aplicabilidad a cada escenario.

- **Formulación de criterios para la evaluación de las técnicas seleccionadas**. Se trata de definir criterios que permitan medir la eficacia de una técnica a la hora de aplicarla a los datos.

- **Definición de índices de eficacia de la evaluación.** Se trata de definir los grados de bondad que se le otorgarán a cada técnica aplicada según los criterios establecidos. (Alta, media, baja).

- **Definición de condiciones y consideraciones generales a la hora de aplicar la técnica a los datos.** Existen una serie de consideraciones y restricciones que deben quedar explícitamente descritas en forma de manual o guía de forma tal que la aplicación de la técnica no dé lugar a ambigüedades de ningún tipo. Es altamente recomendable establecer estas condiciones en forma de árbol decisional con la finalidad de facilitar la elección de la técnica más adecuada.

A modo de ejemplo abordaremos la problemática planteada por la detección de duplicados en los valores de un atributo individual.

El campo en cuestión es del tipo string, se trata de *nombres de clientes del tipo empresa*. Se espera una cardinalidad baja menor a mil.

Se aplicará la técnica de la distancia de edición estándar o distancia de Levehnstein [12], calcula la distancia existente entre dos textos como el número de operaciones de edición (inserciones, borrados y reemplazos) necesarias para transformar un texto en el otro.

Se definen como índices de evaluación los siguientes:

- Alta: Cuando la similitud entre los dos textos se acrecienta hasta un 80% o más al aplicarse el criterio
- Media: Cuando la similitud entre dos textos al aplicarse el criterio fluctúa entre un 60% y un 79%.
- Baja: cuando la similitud entre los textos decae bajo el 59% al aplicarse el criterio.

Algunas de las condicionantes que podemos establecer a modo de ejemplo para este caso son las siguientes:

¿Cuál es el tipo de datos? Si el campo es de texto, aplican técnicas de detección de duplicados como la distancia de edición, pero si el campo es numérico los problemas a buscar pueden ser de datos atípicos (outliers) y se debe establecer la distribución de los datos antes de aplicar alguna fórmula.

¿Es el atributo analizado un nombre de una empresa? Para este tipo de atributos, existen técnicas específicas como búsqueda en bases de datos del gobierno. Además, en estos casos, no será común que se varíe el orden de las palabras.

¿Es el atributo un nombre de persona? En nombres, dependiendo de la forma de captura, puede ser común, que se varíe el orden de las palabras o que existan palabras truncadas.

¿Es una dirección? Para atributos que almacenan direcciones, existen técnicas específicas relacionadas con georeferenciación.

¿Es un campo de contenido distinto a nombres o direcciones? (datos tipo texto). En un campo de contenido diferente a estos, como el título de un proyecto de investigación, no será común que se varíe el orden de las palabras.

La siguiente figura muestra un esquema de lo que podría ser un árbol decisional que plasme lo que hasta ahora hemos expuesto.

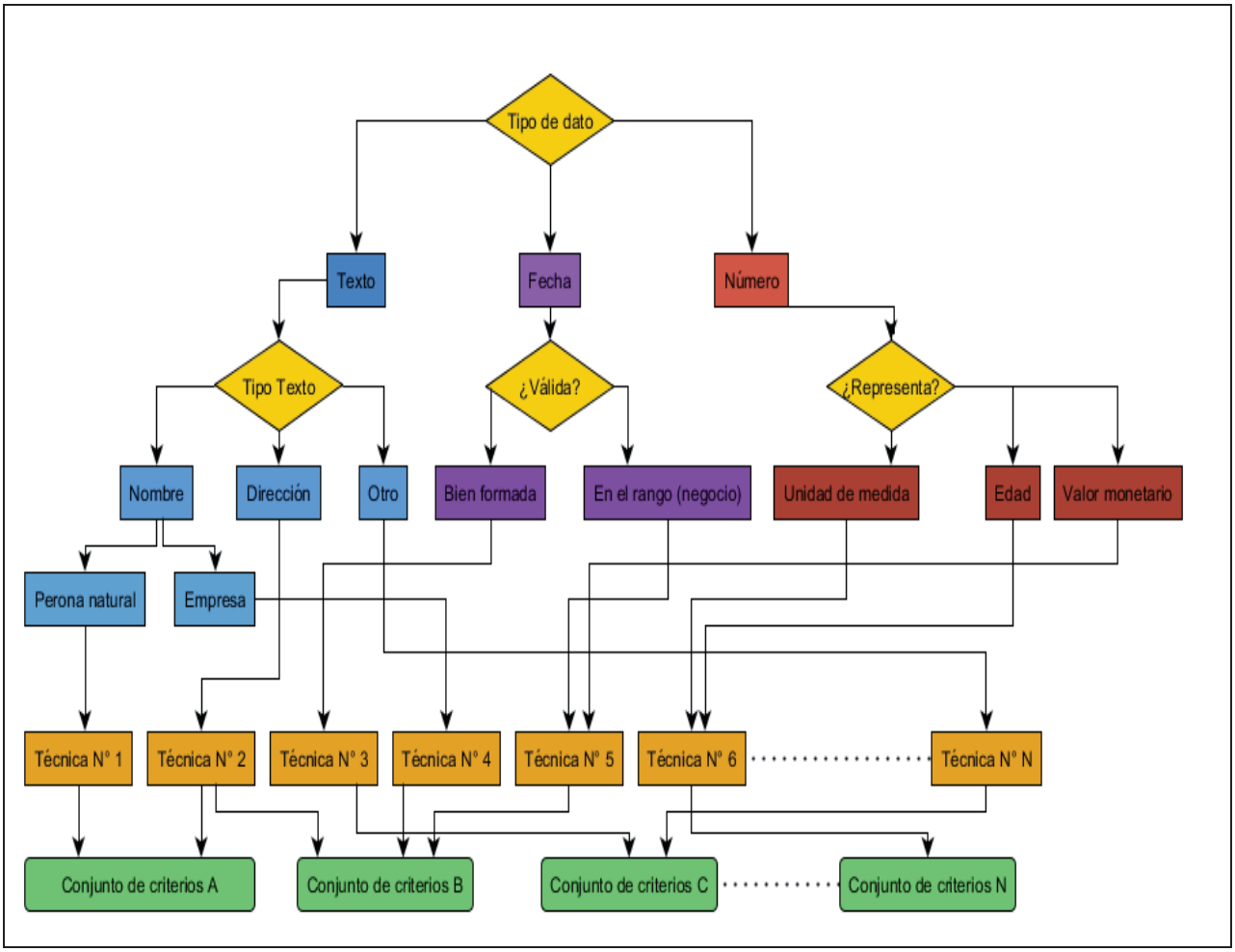

Ilustración 3 Esquema árbol decisional

La siguiente tabla resume los datos obtenidos tras aplicar la evaluación de la técnica seleccionada (distancia de edición estándar o distancia de Levehnstein) según los distintos criterios establecidos y de acuerdo a los rangos de los índices definidos.

Criterio	Definición	Eficacia
Palabras en orden	Textos equivalentes pero sus palabras están en distinto orden.	Baja
Mayúsculas Minúsculas	Textos equivalentes pero escritos con diferencias en términos de la utilización de mayúsculas y minúsculas.	Alta
Espacios en blanco	Textos equivalentes salvo en la utilización de espacios en blanco.	Baja
Palabras faltantes	Textos equivalentes salvo en la omisión de palabras no claves.	Media
Errores ortográficos	Textos equivalentes salvo en diferencias originadas en errores ortográficos en uno de ellos.	Baja
Errores tipográficos	Uno de los textos presenta caracteres faltantes, sobrantes o transpuestos respecto al otro.	Baja
Palabras truncadas	Uno de los textos presenta uso de palabras truncadas o abreviaciones respecto al otro.	Media
Prefijos y Sufijos	Textos equivalentes salvo en el uso de prefijos o sufijos.	Baja
Sinónimos	Uno de los textos es equivalente al otro pero utiliza una palabra sinónima	Baja

Tabla 1 Distancia de Levehnstein

De acuerdo a lo anteriormente analizado la metodología a utilizar será una mezcla de todos los conceptos, en la cual principalmente se utilizaran métricas para analizar los dados de acuerdo a los criterios de tipo de dato, análisis sintáctico, semántico y precisión, además de las reglas entregadas por la compañía generando una distribución de acuerdo a las distintas características encontradas.

7.5. Definición de tecnologías y herramientas

Un aspecto crítico en el proyecto fue definir el tipo de tecnología para las distintas actividades que abarcó su desarrollo. Durante el proceso de implementación del proyecto de calidad de datos se emplearon Sybase *ASE* versión 12.5, *Check Point Endponit Security*, Transact-SQL y Sybase *Power Designer*.

7.5.1. Sybase ASE

Sybase es la compañía de software empresarial más grande enfocada exclusivamente a la gestión y movilización de información, desde el centro de datos, hasta el punto de acción. Sus soluciones abiertas y multiplataforma entregan información de manera segura en cualquier momento y lugar. En mayo del 2010 fue comprada por SAP.

Sybase ASE 12.5 es la versión usada por el servidor de la Casa Central de la compañía, como motor de bases de datos, la cual es mucho más fácil de aprender, instalar y administrar. Los clientes pueden optimizar el rendimiento de muchas aplicaciones para hacer más fácil el manejo de documentos XML, el uso de encriptación avanzada para comunicaciones de red y la búsqueda en textos diferentes al inglés. Los nuevos límites definidos en esta versión significan que el servidor puede gestionar volúmenes de datos para una consulta más grande. Sybase es muy eficiente en términos de utilización de recursos del sistema operativo y uso del hardware [13].

7.5.2. Check Point Endpoint Security

Check Point Endpoint Security ofrece a los usuarios un acceso seguro y transparente a las redes y recursos corporativos cuando viajan o trabajan de forma remota. Integridad y privacidad de la información sensible está garantizada a través de la autenticación de múltiples factores. En esta ocasión se utilizó VPN E75.30 la cual nos permitió el acceso remoto a la red corporativa de la compañía para realizar las tareas necesarias para la implementación del proyecto.

7.5.3. T-SQL

En primer lugar, se define SQL (*Structured query language*), o en español, lenguaje de consulta estructurado como un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones. SQL es un lenguaje informático que se utiliza para interactuar con una base de datos relacional y nos permite realizar consultas a ésta, a través de clausuras, operadores y funciones de agregado. Estos elementos se combinan en instrucciones para crear, actualizar, modificar y manipular las bases de datos. A partir del estándar cada sistema ha desarrollado su propio SQL que puede variar de un sistema a otro, pero con cambios que no suponen ninguna complicación para alguien que conozca un SQL.[14]

T-SQL o Transact-SQL (*Transact Structured query language*) es una extensión al SQL de Microsoft y Sybase. Es un lenguaje muy potente que nos permite definir casi cualquier tarea a efectuar sobre la base de datos, ya que posee características que nos permiten definir la lógica necesaria para el tratamiento de la información, como tipos de datos, definición de variables, estructuras de control de flujo, gestión de excepciones, funciones predefinidas y bucles. Sin embargo, no permite crear interfaces de usuario, crear aplicaciones ejecutables, si no, elementos que en algún momento llegarán al servidor de datos y serán ejecutados. [14]

Esta tecnología es de vital importancia para la implementación del proyecto, ya que es el lenguaje central para desarrollar el Script SQL, el cual se encargará de realizar la evaluación de los datos del sistema.

7.5.4. Microsoft Excel

Microsoft Excel es un software desarrollado por Microsoft como herramienta de hojas de cálculo para el uso contable, financiero entre muchas otras utilidades. Permite crear tablas, calcular y analizar datos. Este tipo de herramienta se denomina software de hoja de cálculo. Excel permite crear tablas que calculan de forma automática el total de valores numéricos, imprimir tablas con diseños cuidados y crear gráficos simples. Excel forma parte de “Office”, un conjunto de productos que combina varios tipos de software para crear documentos, hojas de cálculo y presentaciones, y para administrar el correo electrónico.

Este software se utilizó para realizar los reportes dinámicos de la implementación y así, obtener una mejor visualización de los resultados obtenidos en la evaluación.

7.5.5. Power Designer

Power Designer de SAP es una herramienta empresarial colaborativa producida por Sybase, que provee un entorno simple de modelamiento que junta técnicas y notaciones de procesos de negocios con requerimientos de modelamiento, modelamiento de datos, modelamiento de arquitecturas y UML. Esta herramienta se ocupó en el proceso de diseño para generar de manera gráfica entidades.

7.6. Plan de Trabajo

La metodología es un conjunto de métodos o procedimientos racionales que parte de una base teórica que permite cumplir los objetivos planteados. Es la encargada de elaborar, determinar y sistematizar el conjunto de procedimientos a seguir en el desarrollo de esta auditoría, es decir, la organización de los pasos a través de los cuales se ejecutará el trabajo.

1. **Preparación de la implementación:** Corresponde a los preparativos para realizar la implementación. Se definen los objetivos, el alcance, la metodología, el equipo de trabajo y las herramientas a utilizar.

2. **Diseño preliminar:** Corresponde al trabajo realizado previo a la evaluación de los datos. Se realiza un análisis de los datos de origen, se define la estructura para el análisis, se identifican reglas particulares, se definen métricas de evaluación y criterios de aceptación.

3. **Evaluación:** Corresponde a la fase de evaluación en la cual los datos son analizados. Se obtienen resultados que se almacenan en la estructura definida en la fase de diseño. Se ejecuta el control de calidad y la validación de reglas.

4. **Análisis de Resultados:** Se analizan los resultados según las métricas definidas en la fase de diseño para describir fortalezas y debilidades. Luego se determina el modelo de madurez de la entidad evaluada.

5. **Conclusiones:** Es la fase final que presenta recomendaciones según los resultados evidenciados.

Cada etapa será realizada rigurosamente por el equipo de trabajo para asegurar el desarrollo efectivo y eficiente del proyecto. Estas etapas permitirán que la implementación realizada cumpla con el alcance propuesto dentro del proceso de calidad de datos.

A continuación las etapas realizadas se describen Ilustración 4:

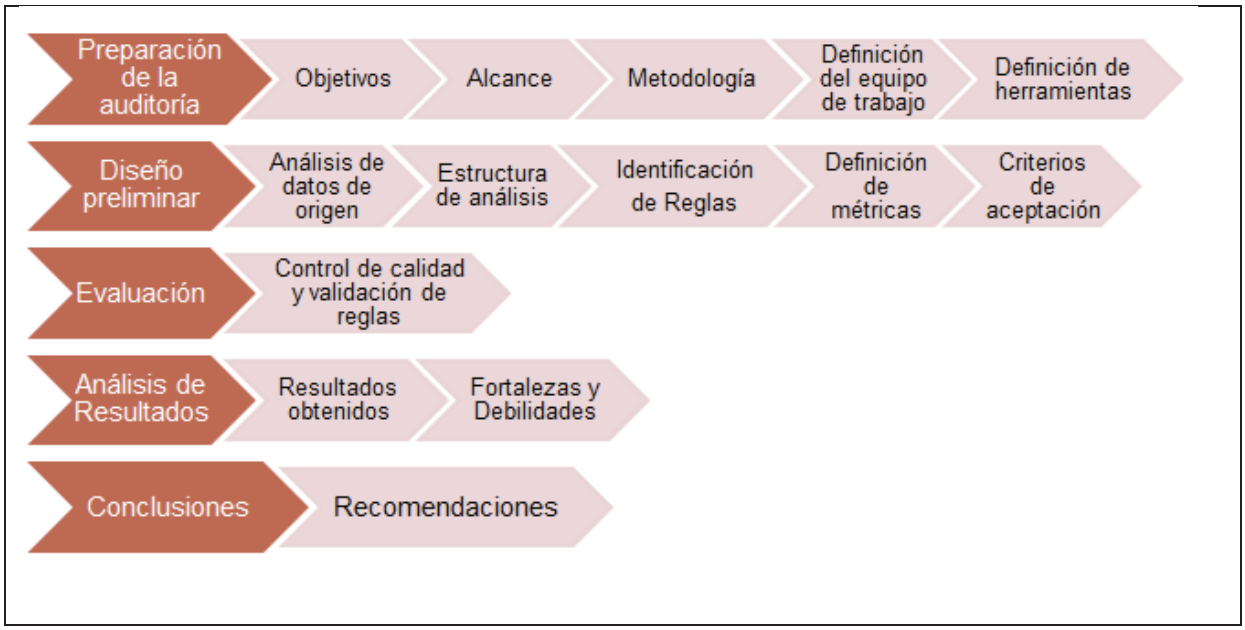

Ilustración 4: Plan de trabajo

7.6.1. Diseño Preliminar

En primera instancia se presentan los datos de origen que serán evaluados. Se evalúa la naturaleza y anomalía de los datos, con el objetivo de tener una idea general del proceso y los desafíos que se presentarán más adelante en la definición de métricas.

Se realiza el proceso de comprensión del modelo que soporta los datos a través de la documentación entregada por la compañía. Luego del análisis de los datos y la estructura que los soporta, se definen métricas de impacto comercial e informáticas con las cuales se cuantificará la calidad del universo de los datos evaluados.

Para finalizar, se definen los criterios de calidad de manera cuantitativa para medir el universo de datos evaluado.

7.6.1.1. Análisis de datos de origen

Para el análisis de los datos de origen, se estudia el universo de datos que se va a migrar desde el sistema Sybase al nuevo sistema Oracle. La base de datos inicial consta de dos servidores de *testing*, dentro de las cuales se encuentran diversas bases de datos que a su vez contienen las tablas específicas analizadas. El total de tablas a analizar es de 15 distribuidas de forma no equitativa en 4 bases de datos. Si bien estas 15 tablas tienen el foco del análisis, esto

no quiere decir que no se vayan a necesitar datos de otras tablas en caso de ser necesario. Uno de los servidores posee datos de la Casa Central de la compañía en tanto el otro contiene información de sus filiales.

7.6.1.2. Entidades a Analizar

A continuación se presenta el origen de las entidades dentro del sistema Sybase para ambos servidores descritos anteriormente. En la Tabla 2 se puede observar el nombre propio de cada entidad a evaluar en el servidor de la Casa Central de la compañía con su base de datos respectiva.

NOMBRE BASE DE DATOS	NOMBRE TABLA
CLIENTES	VAC_CLIENTES
CLIENTES	VAC_DADOR_RECEP
CLIENTES	VAC_DETALLE_FACT_UNICA
CLIENTES	VAC_DIR_POSTAL
CLIENTES	VAC_DTE_CONVENIO_AUTORIZADO
CLIENTES	VAC_EMPALMES
CLIENTES	VAC_EQUIPOS
CLIENTES	VAC_FACT_UNICA
CLIENTES	VAC_FONOS
CLIENTES	VAC_MEDIDOR
CLIENTES	VAC_ORDENSERVICIO
CLIENTES	VAC_SUMINISTROS
COBRANZAS	VCB_CREDITO
DIRECCIONES	VAD_DIRECCIONES
LECTURAS	VFL_LECTURA_CONSUMO_HIST

Tabla 2 Entidades de análisis servidor central

En el siguiente recuadro se muestran las tablas que pertenecen al servidor de filiales de la compañía.

NOMBRE BASE DE DATOS	NOMBRE TABLA
CLIENTES	VAC_CLIENTES
CLIENTES	VAC_DADOR_RECEP
CLIENTES	VAC_DETALLE_FACT_UNICA
CLIENTES	VAC_DIR_POSTAL
CLIENTES	VAC_DTE_CONVENIO_AUTORIZADO
CLIENTES	VAC_EMPALMES
CLIENTES	VAC_EQUIPOS
CLIENTES	VAC_FACT_UNICA
CLIENTES	VAC_FONOS
CLIENTES	VAC_MEDIDOR
CLIENTES	VAC_ORDENSERVICIO
CLIENTES	VAC_SUMINISTROS
COBRANZAS	VCB_CREDITO
DIRECCIONES	VAD_DIRECCION
LECTURAS	VFL_LECTURA_CONSUMO_HIST

Tabla 3 Entidades de análisis servidor filiales

Como se mencionó al comienzo de esta sección, a continuación se presentan en detalle las entidades que no entran en el proceso de la evaluación pero sí participaron durante la ejecución del análisis como tablas auxiliares, tablas enlaces y llaves compuestas.

En la Tabla 4 se presentan las entidades en servidor de la Casa Central de la compañia:

NOMBRE BASE DE DATOS	NOMBRE TABLA
CODIGOS	TIPO_REPARTO
CODIGOS	AREATIPICA_COMUNA
CODIGOS	TIPO_EMPALME
CODIGOS	INSTITU_PAC
CODIGOS	TIPO_TARIFA
CODIGOS	AREA_TELEFONICA
DIRECCIONES	VAD_COMUNAS
EQUIPOS	PROVEEDORES
DICCIONARIO	DD_TABLAS
DICCIONARIO	DD_COLUMNAS
COBRANZAS	VCB_SALDO

Tabla 4 Entidades complementarias servidor central

En la Tabla 5 se presentan las entidades en el servidor de Filiales:

NOMBRE BASE DE DATOS	NOMBRE TABLA
CODIGOS	TIPO_REPARTO
CODIGOS	AREATIPICA_COMUNA
CODIGOS	TIPO_EMPALME
CODIGOS	INSTITU_PAC
CODIGOS	TIPO_TARIFA
CODIGOS	AREA_TELEFONICA
DIRECCIONES	VAD_COMUNAS
EQUIPOS	PROVEEDORES
DICCIONARIO	DD_TABLAS
DICCIONARIO	DD_COLUMNAS
COBRANZAS	VCB_SALDO

Tabla 5 Entidades complementarias servidor filiales

La compañía aclara que la veracidad de los datos en sus servidores de *testing* es la misma que en servidores en producción.

Solamente habrá un periodo de tiempo de diferencia entre la actualización de los servidores de producción hacia los de *testing*. En ningún momento se realizarán análisis en servidores para desarrollo.

Cabe destacar que los datos, nombre de las tablas fueron modificados por nivel de confidencialidad a nivel contractual.

7.6.1.3. Universo a migrar

El análisis llevado a cabo por el equipo de trabajo se centró exclusivamente en los datos a migrar del servidor Sybase de la Casa Central de la compañía.

Dentro de los datos que se encuentran en las bases de datos, el universo de los datos a migrar es acotado y según la documentación entregada por la compañía se deben cumplir con los siguientes requisitos de migración.

Cabe recordar que los datos corresponden a una empresa de servicios eléctricos,

Para Suministros:

- Se consideran suministros activos.
- Se consideran suministros inactivos con deuda distinta de cero.
- Se consideran suministros castigados con deuda superior a $10.000 y que posea una deuda que acredite a ésta misma. Se excluyen suministros castigados con deuda menor a $10.001, castigados con un monto superior a $10.000 y que no tenga una factura que acredite la deuda. [10]

Para Órdenes de servicio:

- Se consideran las órdenes de servicio cuyo estado actual sea Ejecutada.
- Se consideran las órdenes de servicio cuya fecha de emisión sea mayor al 1 de enero del 2014.

Para Lecturas de consumo histórico:

- Se consideran las lecturas de consumo en un período de dos años. Este período se establece desde la fecha actual con una ventana de dos años atrás.

Para Dadores y receptores (Entidad DADOR_RECEP)

- Se consideran solo suministros activos.

7.6.2. Definición de Métricas

Una métrica es un instrumento que define la forma de medir un factor de calidad, asi mismo factor de calidad puede tener distintas formas de medición.

Por otro lado, definimos método de medición como un proceso que implementa una métrica. A su vez, una misma métrica puede ser medida por diferentes métodos.

La definición de métricas para este proyecto son de dos tipos:

- Métricas de impacto comercial
- Métricas informáticas.

7.6.2.1. Métricas de impacto comercial

La medición de calidad de los datos se realizará a través de una serie de reglas que se deberán cumplir para que la migración del sistema Sybase a Oracle sea exitosa.

Existe un numero de 49 reglas consideradas como las métricas que tendrán impacto comercial para la compañía (entregadas por la misma).

Según esto último, la métrica a nivel de columna es:

- Número de registros que no cumplen la descripción de una regla (incorrectos)

La medición entregará si una regla es cumplida o no. Técnicamente, las reglas se aplican tanto al servidor de la Casa Central como al servidor de Filiales de la compañía.

De la misma forma, se especifican reglas que debido a su descripción compleja han sido divididas para un mejor análisis a nivel técnico. La compañía ha especificado tipos de reglas "Bloqueantes" y "No Bloqueantes" para su proceso de migración, por lo tanto, se considerará como prioridad la ejecución de la reglas Bloqueantes.

Las 49 reglas se presentan en detalle en el anexo Reglas de Migración.

7.6.2.2. Métricas Informáticas

A través de las métricas desarrolladas por el equipo de trabajo se logrará cuantificar la calidad del universo de datos evaluados.

Las métricas informáticas entregarán una visión general del universo a migrar desde la base de datos Sybase. A continuación se definen las métricas a utilizar:

A nivel de tabla:

- Cantidad total de registros de una tabla.

A nivel de columna:

- Largo soportado de una columna.
- Largo del valor mínimo de una columna.
- Valor mínimo de una columna.
- Largo del valor máximo de una columna.
- Valor máximo de una columna.

- Cantidad de registros distintos de una columna.
- Número de registros nulos de una columna.
- Tasa de registros nulos (cantidad de registros nulos divididos por cantidad de total de registros de una tabla).

A nivel de celda:

- Frecuencia de ocurrencia de una celda específica

Esta última métrica se aplicará para las columnas cuya cantidad de registros distintos sea menor o igual a 30 (este número depende de cada implementación principalmente por la cardinalidad de los campos a analizar).

Las métricas descritas anteriormente se aplican a todos los registros que deben participar en la migración de datos desde Sybase a Oracle.

7.7. Definición de criterios de calidad

La definición de criterios de calidad permite al grupo de trabajo instaurar valores cuantitativos que serán base para las recomendaciones que se entregarán a compañía posterior al proceso de calidad de datos.

En primer lugar, la tasa de nulos se fijará a un 60% máximo [15]. Tasas por encima de ese valor se recomendará normalizar el campo respectivo.

Para el resto de las métricas informáticas, inconsistencias de largo soportado y cantidad de columnas, se deberán revisar la información en las tablas DD_COLUMNAS en la base de datos DICCIONARIO (catálogo de la base de datos) de cada servidor analizado en conjunto con los archivos anexos entregados a la compañía.

Para métricas de impacto comercial, la cantidad de registros incorrectos tendrá el siguiente criterio:

- Para regla Bloqueante: se esperan 0 registros incorrectos.

Tomando en cuenta esto, un 0% será aceptable para asegurar que ninguna regla Bloqueante presente registros que puedan bloquear el proceso de migración.

7.8. Estructura de Análisis

El sistema Sybase posee dos servidores de *testing* Filiales Y Casa Central, los cuales serán el foco del análisis. Para llevar esto a cabo, el equipo de trabajo desarrolló una estructura de análisis que permite la captura de resultados según las métricas propuestas en las secciones anteriores.

A continuación se describen las entidades que forman la estructura de análisis:

- **DQ_TABLA**: describe cada tabla que se va a evaluar en el análisis con su cantidad de registros involucrados. De esta forma, la entidad contendrá: identificador de tabla, nombre de su base de datos, nombre propio de tabla, número de columnas, número de registros y servidor al que pertenece la tabla.

- **DQ_COLUMNA**: describe columna o campo de cada tabla que se va a evaluar en el análisis. Contendrá las métricas informáticas asociadas a columnas. De esta forma, la entidad contendrá: identificador de tabla, identificador de columna, servidor asociado a la columna, tipo de dato, largo soportado, largo del valor mínimo y máximo de la columna, valor mínimo y máximo de la columna, número de registros distintos, número de registros nulos, tasa de nulos de la columna y la consulta necesaria en SQL para obtener los valores nulos de la columna.

- **DQ_COLUMNA_DISTRIBUCIÓN**: describe los valores de una columna específica en la cual sus valores distintos sean menos o igual a 100, capturando la frecuencia de ocurrencia de cada valor. De esta forma, la entidad contendrá: identificador de tabla, identificador de columna, nombre de tabla, nombre de columna, servidor asociado a la columna, valor específico de una columna y frecuencia de ocurrencia de dicho valor.

A continuación se presenta el Modelo Relacional de la estructura de análisis descrita:

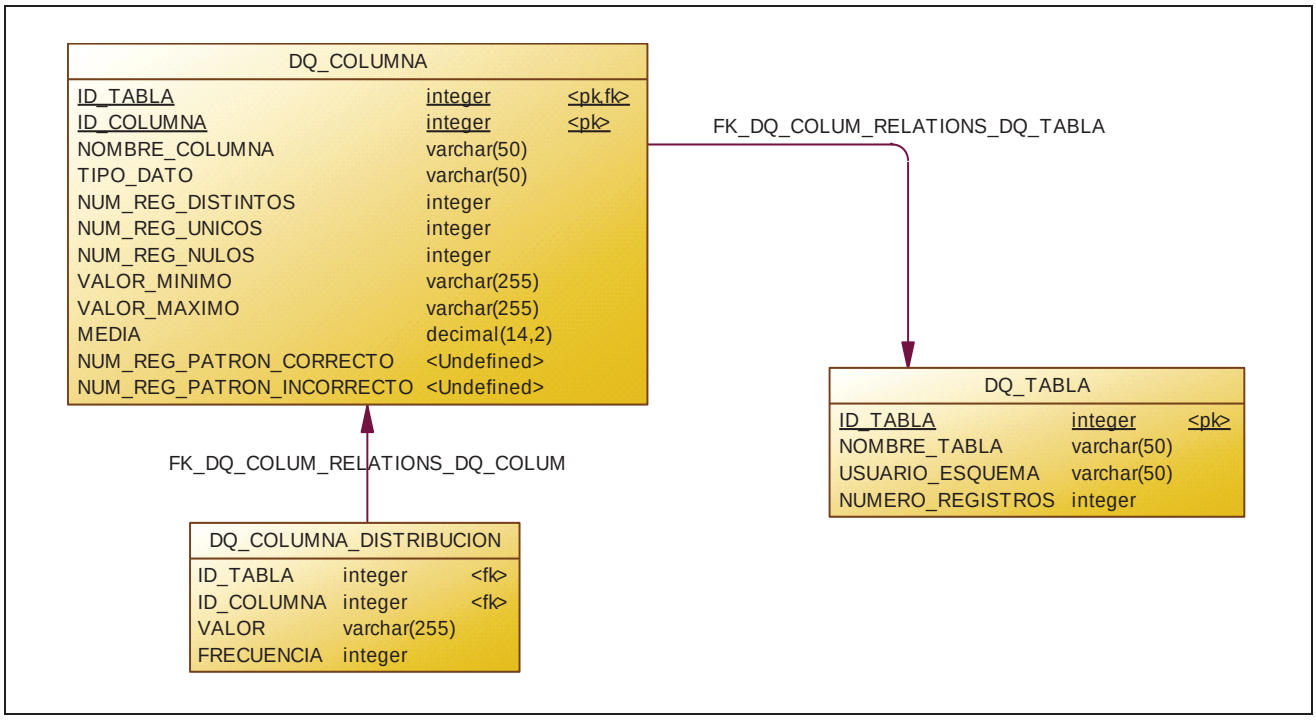

Ilustración 5: Modelo relacional DQ

La Ilustración 5 representa el Modelo Relacional en el cual se almacenarán los datos obtenidos de todas las tablas y así, obtener las métricas informáticas.

Para las métricas de impacto comercial se especifica una única entidad para el almacenamiento de datos. Dicha entidad se describe como:

- **DQ_REGLAS**: representa una entidad que almacenará los datos obtenidos de la validación de las reglas entregadas por la compañia. De esta forma, los datos que almacenará esta tabla corresponden a: identificador de regla, servidor asociado, nombre, tablas involucradas, columnas involucradas, tipo de regla, número de registros correctos, número de registros incorrectos, total de registros identificados por la regla, total de registros en la tabla asociada a la regla y las consultas SQL para la validación de la regla en forma correcta e incorrecta.

A continuación se presenta la entidad descrita en notación de Modelo Relacional:

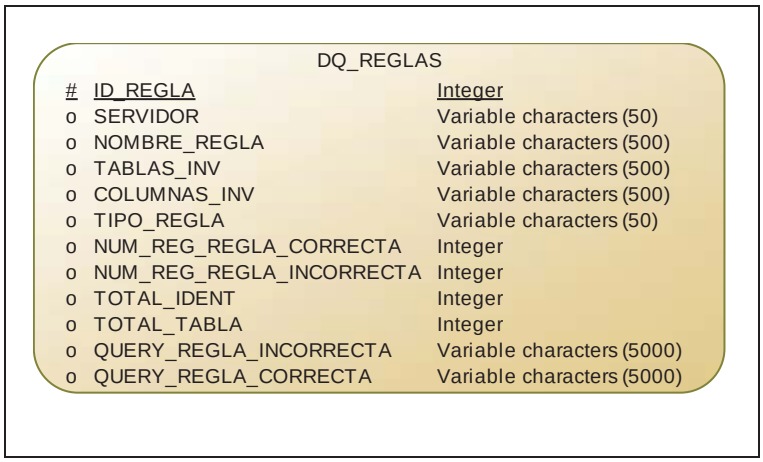

Ilustración 6 Modelo relacional DQ reglas

Las estructuras que almacenarán las métricas comerciales e informáticas físicamente estarán en el ambiente de compañía. Esta misma a ha propiciado credenciales para acceso remoto y una base de datos, para que luego de realizar el análisis, los datos queden en el mismo ambiente de compañía.

7.9. Evaluación

La etapa de evaluación está enfocada en la ejecución de las métricas a nivel de comercial e informático, definidas en la sección anterior. Esta sección presenta cómo se desarrolla la evaluación a nivel técnico, tanto los pasos a seguir como los formatos de salida que se obtienen de la aplicación de las métricas.

El proceso de evaluación se tradujo físicamente en un s*cript* SQL, el cual corresponde al archivo principal que contiene los procedimientos necesarios para recorrer tanto las 15 tablas como las 49 reglas. El lenguaje técnico utilizado por el *script* SQL corresponde a T-SQL definido en la sección "Definición de tecnologías y herramientas" en los anexos. El algoritmo de evaluación se representa a través del siguiente pseudocódigo:

Pseudocódigo *script* SQL de evaluación:

```
Inicio
        Creación de estructuras de análisis
                Captura de datos básicos de las entidades
                Repetir: por cada entidad
                        Seleccionar el filtro
                        Actualizar cantidad de registros
                        Capturar datos básicos de las columnas
                        Repetir: por cada columna
                                Actualizar métricas informáticas
                                Sí: Número de registros distintos <= 30
                                        Capturar frecuencias en la columna
                                Fin Sí
                        Fin Repetir
                Fin Repetir
        Insertar datos básicos de cada regla de validación
        Repetir: por cada regla
              Actualizar cantidad de registros incorrectos
                Actualizar cantidad de total de la tabla analizada
        Fin Repetir
Fin
```

Se debe destacar en el pseudocódigo anterior que al mencionar la acción de "Capturar" se hace referencia a la acción de insertar datos desde el sistema Sybase de la Casa Central de la compañía, en cambio "Insertar" será cuando el equipo de trabajo deba ingresar manualmente datos.

Si bien el proceso de evaluación en el pseudocódigo se presenta como uno solo, en las siguientes secciones se presenta el Control de calidad que se realizó a las 15 tablas capturando las métricas informáticas. La sección posterior presenta la Validación de reglas para las 49 reglas entregadas por la compañía capturando las métricas de impacto comercial.

7.9.1. Control de Calidad

El Control de calidad se enfoca en el análisis acotado de las 15 tablas a evaluar. El término acotado es debido a que al universo total de datos debió ser filtrado para solo analizar los registros que la compañía va a migrar.

Es por esto que en esta sección se tratarán los filtros ejecutados, consideraciones de tiempo en las ejecuciones y se presenta una muestra del formato de salida con una tabla de ejemplo. Cabe destacar que el análisis completo se presenta en los archivos anexos a este informe.

7.9.1.1. Filtros ejecutados

Según las definiciones del Universo de datos en el capítulo anterior, se entiende que dichas condiciones se deben aplicar al nivel técnico como restricciones SQL. En primera instancia, se identifica para suministros la siguiente consideración:

"Suministros activos, suministros inactivos que posean una deuda distinta de cero, suministros castigados con deuda superior a $10.000 y que posean una factura que acredite la deuda .Se excluyen los suministros castigados con deuda menor a $10.001 y suministros castigados con deuda superior a $10.000 y que no tengan una factura que acredite la deuda."

Por ende, se desarrolló la tabla **SUMINISTROS_CLIENTES_VALIDOS** que contiene datos de suministros y clientes que se migrarán de forma válida. La definición de dicha tabla para el servidor de la Casa Central de la compañía se presenta a continuación:

```
CREATE TABLE SUMINISTROS_CLIENTES_VALIDOS(
        nis                         INTEGER NULL,
        cd_dv_nis                   VARCHAR      NULL,
        cd_estado_suministro        INTEGER NULL,
        nr_rut_usuario              INTEGER      NULL,
        cd_dv_usuario               VARCHAR(1)   NULL,
        nr_rut_cliente              INTEGER      NULL,
        cd_dv_cliente               VARCHAR(1)   NULL
)
```

La definición de dicha tabla para el servidor de Filiales se presenta a continuación:

```
CREATE TABLE SUMINISTROS_CLIENTES_VALIDOS(
        nis                         INTEGER NULL,
        cd_empresa                  INTEGER NULL,
        cd_dv_nis                   VARCHAR      NULL,
        cd_estado_suministro        INTEGER NULL,
        nr_rut_usuario              INTEGER      NULL,
        cd_dv_usuario               VARCHAR(1)   NULL,
        nr_rut_cliente              INTEGER      NULL,
        cd_dv_cliente               VARCHAR(1)   NULL
)
```

Se aprecia similar a la tabla creada para servidor de la Casa Central solo que SUMINISTROS_CLIENTES_VALIDOS para servidor de Filiales contiene el atributo ***cd_empresa*** para diferenciar a las distintas filiales contenidas dentro del servidor.

La inserción de datos en servidor Central se realiza a través del siguiente código T-SQL:

```
INSERT INTO
        SUMINISTROS_CLIENTES_VALIDOS
SELECT
        nis,
        cd_dv_nis,
        cd_estado_suministro,
        nr_rut_usuario,
        cd_dv_usuario,
        nr_rut_cliente,
        cd_dv_cliente
FROM
        CLIENTES..VAC_SUMINISTROS
WHERE
        (cd_estado_suministro = 9 AND nis IN
                (SELECT nr_nis FROM COBRANZAS..VCB_SALDO WHERE vl_saldo > 10000  )
        )
        OR
        cd_estado_suministro = 2
         OR
        (cd_estado_suministro = 3 AND nis IN
                (SELECT  nr_nis FROM COBRANZAS..VCB_SALDO WHERE vl_saldo != 0  )
        )
```

La inserción de datos en el servidor de Filiales considera el campo cd_empresa como parte de las restricciones a cumplir en el siguiente código T-SQL:

```
INSERT INTO
        SUMINISTROS_CLIENTES_VALIDOS
SELECT
        nis,
        cd_empresa,
        cd_dv_nis,
        cd_estado_suministro,
        nr_rut_usuario,
        cd_dv_usuario,
        nr_rut_cliente,
        cd_dv_cliente
FROM
        CLIENTES..VAC_SUMINISTROS AL1
WHERE
(
        (AL1.cd_estado_suministro = 9 AND EXISTS
                (SELECT * FROM COBRANZAS..VCB_SALDO AL2 WHERE AL2.vl_saldo > 10000
        AND AL1.nis=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa  )
        )
        OR
        AL1.cd_estado_suministro = 2
        OR
        (AL1.cd_estado_suministro = 3 AND EXISTS
                (SELECT * FROM COBRANZAS..VCB_SALDO AL2 WHERE AL2.vl_saldo != 0  AND
        AL1.nis=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa)
        )
)
```

Se adicionará la restricción "AND cd_empresa = n" al momento de solo obtener resultados de una filial donde "n" es el código de ésta.

Luego de tener los suministros a migrar definidos, se procede a crear la tabla **CLIENTES_VALIDOS** que solo almacenará los clientes a migrar. Dicha tabla se define de la siguiente manera para servidor Casa Central:

```
CREATE TABLE CLIENTES_VALIDOS(
        nr_rut_cliente              INTEGER         NULL,
        cd_dv_cliente               VARCHAR(1)      NULL
)
```

Para servidor de Filiales su definición es la siguiente:

```
CREATE TABLE CLIENTES_VALIDOS(
        cd_empresa                  INTEGER NULL,
        nr_rut_cliente              INTEGER         NULL,
        cd_dv_cliente               VARCHAR(1)      NULL
)
```

La inserción en esta tabla de clientes válidos para la migración se realiza utilizando tablas temporales que almacenan los rut de clientes y usuarios de la tabla SUMINISTROS_CLIENTES_VALIDOS. Para servidor de la Casa Central la inserción tiene la siguiente codificación:

```
CREATE TABLE #CLIENTES1(
        nr_rut_cliente              INTEGER         NULL,
        cd_dv_cliente               VARCHAR(1)      NULL
)
CREATE TABLE #USUARIOS1(
        nr_rut_cliente              INTEGER         NULL,
        cd_dv_cliente               VARCHAR(1)      NULL
)
INSERT INTO #CLIENTES1
        SELECT nr_rut_cliente,cd_dv_cliente
        FROM SUMINISTROS_CLIENTES_VALIDOS

INSERT INTO #USUARIOS1
        SELECT nr_rut_usuario,cd_dv_usuario
        FROM SUMINISTROS_CLIENTES_VALIDOS

CREATE TABLE CLIENTES_VALIDOS(
        nr_rut_cliente              INTEGER    NULL,
        cd_dv_cliente               VARCHAR(1)  NULL
)

INSERT INTO CLIENTES_VALIDOS
        SELECT AL1.nr_rut_cliente,AL1.cd_dv_cliente FROM (
                SELECT nr_rut_cliente,cd_dv_cliente FROM #CLIENTES1
                UNION
                SELECT nr_rut_cliente,cd_dv_cliente FROM #USUARIOS1
        ) AL1
```

Las inserciones en CLIENTES_VALIDOS para el servidor de Filiales incluyen el campo cd_empresa. Por lo anterior, la codificación es la siguiente:

```
CREATE TABLE #CLIENTES1(
        cd_empresa                  INTEGER NULL,
        nr_rut_cliente              INTEGER      NULL,
        cd_dv_cliente               VARCHAR(1)   NULL
)

CREATE TABLE #USUARIOS1(
        cd_empresa                  INTEGER NULL,
        nr_rut_cliente              INTEGER      NULL,
        cd_dv_cliente               VARCHAR(1)   NULL
)

INSERT INTO #CLIENTES1
        SELECT cd_empresa,nr_rut_cliente,cd_dv_cliente
         FROM SUMINISTROS_CLIENTES_VALIDOS

INSERT INTO #USUARIOS1
        SELECT cd_empresa,nr_rut_usuario,cd_dv_usuario
        FROM SUMINISTROS_CLIENTES_VALIDOS

CREATE TABLE CLIENTES_VALIDOS(
        cd_empresa                  INTEGER NULL,
        nr_rut_cliente              INTEGER      NULL,
        cd_dv_cliente               VARCHAR(1)   NULL
)

INSERT INTO CLIENTES_VALIDOS
        SELECT AL1.cd_empresa,AL1.nr_rut_cliente,AL1.cd_dv_cliente FROM (
                SELECT cd_empresa,nr_rut_cliente,cd_dv_cliente FROM #CLIENTES1
                UNION
                SELECT cd_empresa,nr_rut_cliente,cd_dv_cliente FROM #USUARIOS1
        ) AL1
```

Luego de almacenar los suministros a migrar en **SUMINISTROS_CLIENTES_VALIDOS**, se desarrollan el resto de las condiciones específicas:

- Para lecturas de consumo histórico se aplicará la siguiente restricción para cada *query* T-SQL que involucre a la tabla **VFL_LECTURA_CONSUMO_HIST**:

```
(CONVERT(DATE,fc_lectura_act) > CONVERT(DATE,DATEADD(yy,-2,Getdate())))
```

El cual describe la sustracción de dos años desde la fecha actual.

- Para Órdenes de servicio se aplicará la siguiente restricción para cada query T-SQL que involucre a la tabla VAC_ORDENSERVICIO:

```
CONVERT(DATE,AL1.fc_emision) > '20140101'  AND st_actual='Ejecutada'
```

Lo cual hará referencia a solo las órdenes de servicio con estado actual Ejecutada.

A continuación se presenta en la Tabla 6 las entidades analizadas con su filtro respectivo para el servidor de la Casa Central:

BASE DE DATOS	NOMBRE TABLA	FILTRO
CLIENTES	VAC_SUMINISTROS, VAC_DIR_POSTAL, VAC_MEDIDOR, VAC_DTE_CONVENIO _AUTORIZADO, VAC_FONOS, VAC_EMPALMES, VAC_EQUIPOS	AL1.nis IN (SELECT nis FROM SUMINISTROS_CLIENTES_VALIDOS)
DIRECCIONES	VAD_DIRECCIONES	EXISTS(SELECT * FROM CLIENTES..VAC_SUMINISTROS AL2, SUMINISTROS_CLIENTES_VALIDOS AL3 WHERE AL1.cd_direccion=AL2.cd_direccion AND AL2.nis=AL3.nis)
CLIENTES	VAC_CLIENTES	EXISTS(SELECT * FROM CLIENTES_VALIDOS AL2 WHERE (AL1.nr_rut_cliente=AL2.nr_rut_cliente AND AL1.cd_dv_cliente=AL2.cd_dv_cliente))
LECTURAS	VFL_LECTURA_CONSUMO_HIST	AL1.nis IN (SELECT nis FROM SUMINISTROS_CLIENTES_VALIDOS) AND (CONVERT (DATE,AL1.fc_lectura_act) > CONVERT (DATE ,DATEADD(yy,-2,Getdate())))
CLIENTES	VAC_DADOR_RECEP	AL1.nis_dador IN (SELECT nis FROM CLIENTES..VAC_SUMINISTROS WHERE cd_estado_suministro=2)
CLIENTES	VAC_ORDENSERVICIO	AL1.nis IN (SELECT nis FROM SUMINISTROS_CLIENTES_VALIDOS) AND CONVERT(DATE,AL1.fc_emision) > '20140101' AND AL1.st_actual='Ejecutada'
CLIENTES	VAC_FACT_UNICA	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nr_rut_cliente=AL2.nr_rut_cliente AND AL1.nis_principal=AL2.nis)
CLIENTES	VAC_DETALLE_FAC_UNICA	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nr_rut_cliente=AL2.nr_rut_cliente AND AL1.nis_detalle=AL2.nis)
COBRANZAS	VCB_CREDITO	AL1.nr_nis IN (SELECT nis FROM SUMINISTROS_CLIENTES_VALIDOS)

Tabla 6 Filtros de entidades servidor Central

A continuación se presenta en la Tabla 7 las entidades analizadas con su filtro respectivo incluyendo el campo cd_empresa para el servidor de Filiales:

BASE DE DATOS	NOMBRE TABLA	FILTRO
CLIENTES	VAC_SUMINISTROS, VAC_DIR_POSTAL, VAC_MEDIDOR, VAC_DTE_CONVENIO _AUTORIZADO, VAC_FONOS, VAC_EMPALMES, VAC_EQUIPOS	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nis=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa)
DIRECCIONES	VAD_DIRECCION	EXISTS(SELECT * FROM CLIENTES..VAC_SUMINISTROS AL2, SUMINISTROS_CLIENTES_VALIDOS AL3 WHERE AL1.cd_direccion=AL2.cd_direccion AND AL2.nis=AL3.nis AND AL1.cd_empresa=AL3.cd_empresa)
CLIENTES	VAC_CLIENTES	EXISTS(SELECT * FROM CLIENTES_VALIDOS AL2 WHERE ((AL1.nr_rut_cliente=AL2.nr_rut_cliente AND AL1.cd_dv_cliente=AL2.cd_dv_cliente)) AND AL1.cd_empresa=AL2.cd_empresa)
LECTURAS	VFL_LECTURA_CONSUMO_HIST	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nis=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa) AND (CONVERT (DATE,AL1.fc_lectura_act) > CONVERT (DATE ,DATEADD(yy,-2,Getdate())))
CLIENTES	VAC_DADOR_RECEP	EXISTS(SELECT * FROM CLIENTES..VAC_SUMINISTROS AL2 WHERE (AL1.nis_dador=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa AND AL2.cd_estado_suministro=2))
CLIENTES	VAC_ORDENSERVICIO	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nis=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa) AND CONVERT(DATE,AL1.fc_emision) > '20140101' AND AL1.st_actual='Ejecutada'
CLIENTES	VAC_FACT_UNICA	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nr_rut_cliente=AL2.nr_rut_cliente AND AL1.nis_principal=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa)
CLIENTES	VAC_DETALLE_FAC_UNICA	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nr_rut_cliente=AL2.nr_rut_cliente AND AL1.cd_empresa=AL2.cd_empresa AND AL1.nis_detalle=AL2.nis)
COBRANZAS	VCB_CREDITO	EXISTS(SELECT * FROM SUMINISTROS_CLIENTES_VALIDOS AL2 WHERE AL1.nis=AL2.nis AND AL1.cd_empresa=AL2.cd_empresa)

Tabla 7 Fitrol entidades servidor Filiales

7.9.1.2. Ejecución del Script SQL

Se realizaron diversas ejecuciones sobre Script SQL durante su desarrollo antes de obtener los resultados finales. Las ejecuciones se llevaron a cabo de forma remota a través de VPN en las dependencias de la Escuela de Ingeniería Informática de la Pontificia Universidad Católica de Valparaíso.

7.9.1.3. Muestra de una salida de resultados

A continuación se presenta una muestra de los resultados finales de esta sección, teniendo como ejemplo la tabla VAC_FONOS desde la base de datos CLIENTES del servidor de *testing* de a Casa Central de la compañía:

Resultado en DQ_TABLA:

ID_TABLA	NOMBRE_BASEDATOS	NOMBRE_TABLA	NUM_COL	NUM_REG	SERVIDOR
9	CLIENTES	VAC_FONOS	7	1043553	Central

Tabla 8 Resultado DQ_Tabla

Resultado en DQ_COLUMNA considerando solo el nombre columna "cd_tipo_fono":

NOMBRE ATRIBUTO	VALOR
ID_TABLA	9
ID_COLUMNA	5
NOMBRE_TABLA	VAC_FONOS
NOMBRE_COLUMNA	cd_tipo_fono
SERVIDOR	Central
TIPO_DATO	tinyint
LARGO_SOPORTADO	1
LARGO_MIN_REAL	1
LARGO_MAX_REAL	1
NUM_REG_DISTINTOS	10
NUM_REG_NULOS	0
TASA_DE_NULOS	0
VALOR_MIN	1
VALOR_MAX	9
QUERY_NULOS	SELECT * FROM CLIENTES..VAC_FONOS AL1 WHERE ((AL1.cd_tipo_fono IS NULL) AND (AL1.nis IN (SELECT nis FROM SUMINISTROS_CLIENTES_VALIDOS)))

Tabla 9 Resultado DQ_Columna

Resultado en DQ_DISTRIBUCION (considerando solo la columna cd_tipo_fono y el primer valor analizado):

ID_TABLA	ID_COLUMNA	NOMBRE_TABLA	NOMBRE_COLUMNA	SERVIDOR	VALOR	FRECUENCIA
9	5	VAC_FONOS	cd_tipo_fono	Central	1	446710

Tabla 10 Resultado DQ_Distribucion

7.9.2. Validación de Reglas

En esta etapa de validación de reglas se desarrollaron las *querys* necesarias para validar cuantitativamente si una regla cumple o no en una o más columnas de una o más tablas.

Las 49 reglas a evaluar constan de una *querys* , una cadena de caracteres para guardar la *query* ejecutada y el total de la tabla involucrada.

Se considera que una regla puede o no ser cumplida, lo que se reflejará en cada *query* asociada a cada regla. El valor numérico obtenido si una regla es correcta o incorrecta corresponde a lo descrito en secciones anteriores como métricas de impacto comercial.

A continuación se presentan los resultados obtenidos en la ejecución de una regla de ejemplo, en este caso la regla número 16 en el servidor dela Casa Central de la Compañia. En primer lugar se define la inserción de los datos básicos de la regla.

Inserción datos básicos regla número 16:

```
INSERT INTO
        APUCV_REGLAS_CHILQUINTA
        (ID_REGLA,SERVIDOR,
        NOMBRE_REGLA,
        TABLAS_INV,
        COLUMNAS_INV,
        TIPO_REGLA,
        NUM_REG_REGLA_INCORRECTA,
        TOTAL_TABLA,
        QUERY_REGLA_INCORRECTA)
VALUES
        (16,
        'SINCOTEST',
        'Cd_tipo_fono_inexistente',
        'VAC_FONOS',
        'cd_tipo_fono',
        'Bloqueante',
        0,
        0,
         0,
        0,
        '',
         ''
        )
```

A continuación se presenta la codificación de la regla 16 al actualizar la cantidad de registros incorrectos:

```
UPDATE
        APUCV_REGLAS_CHILQUINTA
SET
        NUM_REG_REGLA_INCORRECTA = (SELECT COUNT(*) FROM CLIENTES..VAC_FONOS
                WHERE ((cd_tipo_fono NOT IN (1,2,3,4,5,6,7,8)) AND (nis IN (SELECT nis FROM
                SUMINISTROS_CLIENTES_VALIDOS) ))),
        TOTAL_TABLA = (SELECT COUNT(*) FROM CLIENTES..VAC_FONOS WHERE (nis IN (SELECT                nis
                FROM SUMINISTROS_CLIENTES_VALIDOS) )),
        QUERY_REGLA_INCORRECTA = "SELECT COUNT(*) FROM CLIENTES..VAC_FONOS WHERE
        ((cd_tipo_fono NOT IN (1,2,3,4,5,6,7,8)) AND (nis IN (SELECT nis FROM
        CLIENTES_VALIDOS) ))"
WHERE
        ID_REGLA = 16 AND SERVIDOR = 'SINCOTEST'
```

Se aprecia cómo el primer campo se actualiza ejecutando la *query* que comprueba la cantidad de registros incorrectos, luego se actualiza el total de la tabla evaluada y para finalizar se actualiza el campo **QUERY_REGLA_INCORRECTA** con la misma *query* que se ejecutó anteriormente pero como cadena de caracteres, de esta forma, la compañía sabrá cuál fue la consulta ejecutada al momento de revisar los resultados finales.

A continuación se presenta una muestra de resultados obtenidos de la regla 16 desde **DQ_REGLAS** en el servidor de la Casa Central de la empresa:

NOMBRE ATRIBUTO	VALOR
ID_REGLA	**16**
SERVIDOR	Central
NOMBRE_REGLA	Cd_tipo_fono_inexistente
TABLAS_INV	VAC_FONOS
COLUMNAS_INV	cd_tipo_fono
TIPO_REGLA	Bloqueante
NUM_REG_REGLA_INCORRECTA	109738
TOTAL_TABLA	1035137
QUERY_REGLA_INCORRECTA	SELECT COUNT(*) FROM CLIENTES..VAC_FONOS WHERE ((cd_tipo_fono NOT IN (1,2,3,4,5,6,7,8)) AND (nis IN (SELECT nis FROM SUMINISTROS_CLIENTES_VALIDOS)))

Tabla 11 Resultado DQ_Regla

7.10. Análisis de Resultados

Posterior a todo tipo de evaluación se da paso a informar los resultados obtenidos en la etapa anterior, a través de reportes, tablas, recomendaciones y conclusiones.

Es importante destacar que los datos analizados pertenecen a una ventana de tiempo que corresponde a la duración de la implementación. Existe una filial específica la cual presentará un análisis de resultados anexo cuyos datos corresponden a una ventana de tiempo distinta, por la urgencia en su migración, datos que fueron obtenidos desde el servidor FILIALES de producción.

Los resultados obtenidos en la etapa anterior se encuentran en forma cuantitativa los cuales serán analizados por métricas. Los detalles de los resultados obtenidos se encuentran en los archivos anexos a este documento. Las gráficas de cada métrica se encuentran en los anexos de este documento.

A continuación se presentan los resultados obtenidos y los elementos positivos detectados.

7.10.1.Resultados obtenidos

En esta sección se presentan los resultados obtenidos por las métricas definidas anteriormente según los registros de las tablas analizadas. A continuación se aprecian las cantidades totales de cada entidad analizada para servidor Central en las siguientes gráficas:

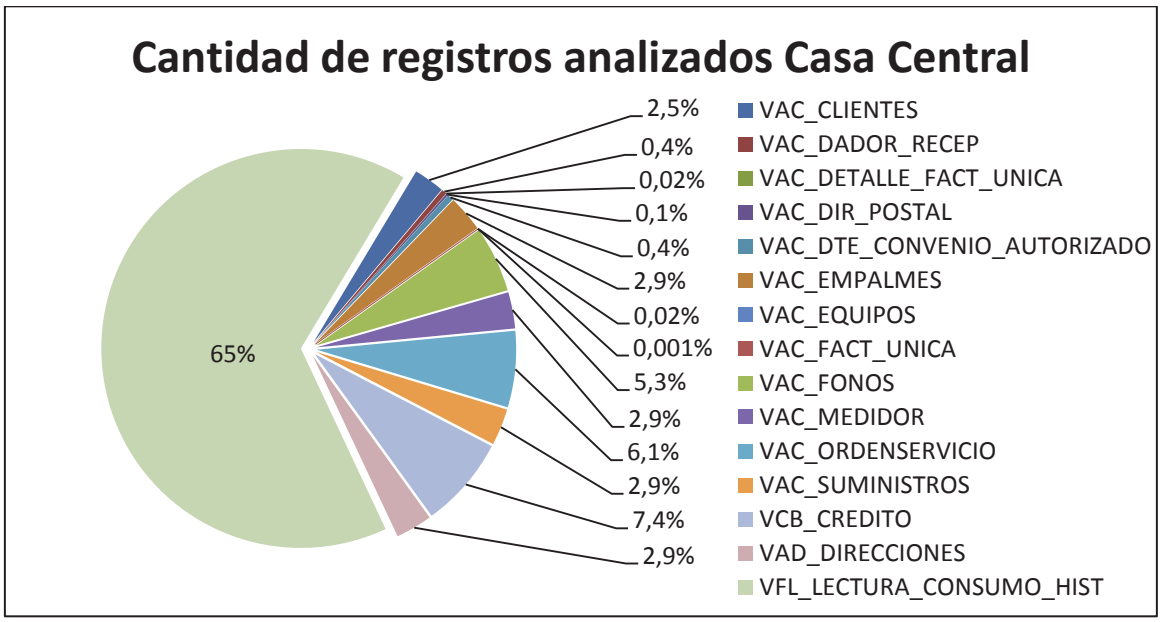

Ilustración 7 Grafica de total de datos analizados en el servidor central

Como se aprecia en la gráfica la tabla VFL_LECTURA_CONSUMO_HIST contiene un 66% del total de los registros analizados. El total de registros analizados en el servidor de la Casa Central corresponde a 19.506.375 registros distribuidos en 298 columnas de 15 tablas.

La cantidad de registros analizados para el servidor de Filiales se presenta en la siguiente gráfica:

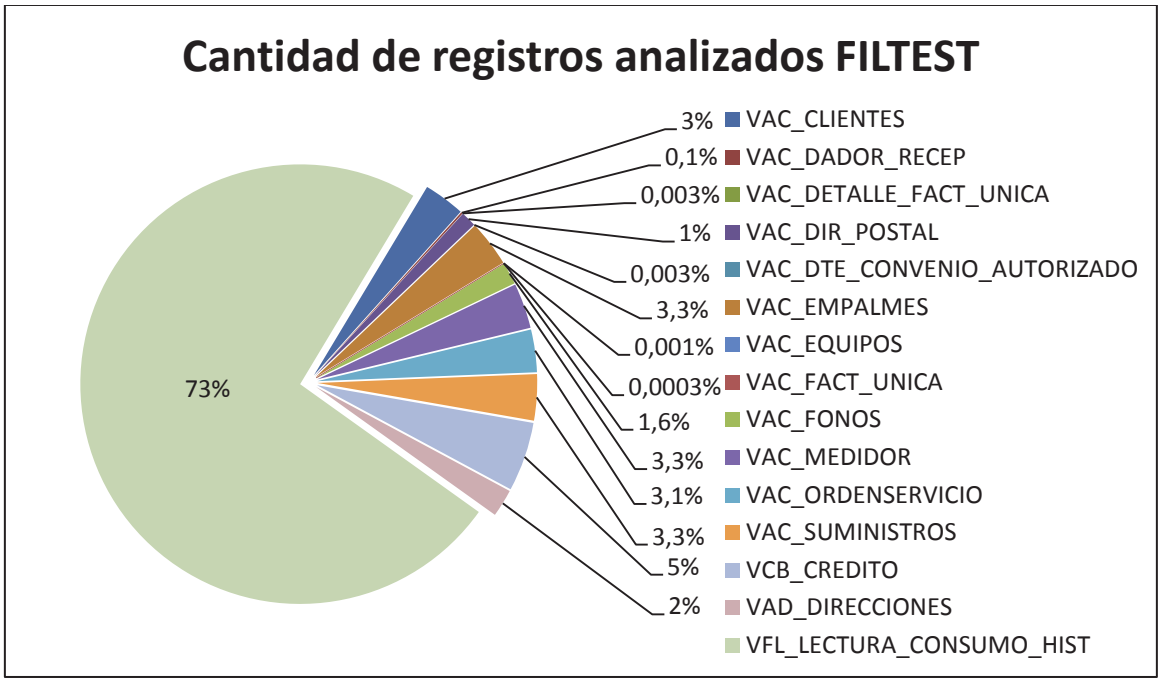

Ilustración 8 Grafica de datos totales analizados del servidor de filiales

Para servidor de Filiales la tabla VFL_LECTURA_CONSUMO_HIST se presenta como la entidad con mayores registros analizados. El total de registros analizados para FILTEST es de 3.365.542 registros distribuidos en 302 columnas de 15 tablas.

A continuación se presentan las distintas métricas ejecutadas por el Script SQL.

7.10.1.1. Tasa de Nulos

El control de los valores nulos (NULL) es un punto crítico para columnas, especialmente las que están sometidas a operaciones matemáticas. Se prefiere una tasa de nulidad baja para una columna.

En caso de no ser así se pondría en duda el propósito de dicha columna en su actual tabla recomendando algún tipo de normalización.

Recordar que un valor nulo es inválido, por ende es un tipo de error semántico que no tiene correspondencia en el mundo real.

La tasa de nulos entregará una medida de cómo la cantidad de registros nulos se relaciona con el total de registros de una tabla.

A continuación se presenta una tabla con las columnas cuya tasa de nulidad es mayor a cero en el servidor Casa Central de la compañía:

NOMBRE TABLA	NOMBRE COLUMNA	TIPO	REGISTROS NULOS	TASA DE NULOS
VAC_ORDENSERVICIO	nr_ficha_inspector	int	1206776	100%
VAC_ORDENSERVICIO	vl_valoriz_original	decimal	1194835	99%
VAC_ORDENSERVICIO	vl_valoriz_real	decimal	1194835	99%
VAC_ORDENSERVICIO	nr_factura	int	1194835	99%
VAC_ORDENSERVICIO	nr_os_anterior	int	1133582	94%
VAC_ORDENSERVICIO	tp_orden_servicio	tinyint	1133505	94%
VAC_ORDENSERVICIO	vl_ult_lectura	decimal	1108978	92%
VAC_ORDENSERVICIO	fc_asignacion	datetime	1066450	88%
VAC_ORDENSERVICIO	cd_ejecutor	int	1065223	88%
VAC_MEDIDOR	cd_formato_lect	int	484071	84%
VAC_ORDENSERVICIO	nr_ficha_ejecutor	int	937578	78%
VAC_ORDENSERVICIO	gl_visita	varchar	936582	78%
VAC_MEDIDOR	fc_fabricacion	datetime	333358	58%
VAC_DTE_CONVENIO_AUTORIZADO	nr_fono	varchar	49538	51%
VAC_EMPALMES	nr_rut_constructor	int	214036	37%
VAC_ORDENSERVICIO	gl_ejecutado	varchar	363732	30%
VAC_EMPALMES	nr_rut_instalador	int	128187	22%
VAC_FONOS	fc_ult_modif	datetime	167458	16%
VAC_SUMINISTROS	nr_folio_sec	char	91081	16%
VAC_FONOS	nr_anexo	smallint	66680	6%
VAC_ORDENSERVICIO	fc_recepcion	datetime	24648	2%
VAC_ORDENSERVICIO	gl_solicitud	varchar	16441	1%
VAC_DTE_CONVENIO_AUTORIZADO	gl_comentario	varchar	789	1%

Tabla 12 Tasa de nulidad servidor central

En la tabla anterior se presentan 12 campos que poseen una tasa de nulidad por encima del 60%. En su mayoría corresponden a la entidad VAC_ORDENSERVICIO.

A continuación se presentan las tasas de nulidad con los resultados del servidor de Filiales:

NOMBRE TABLA	NOMBRE COLUMNA	TIPO	REGISTROS NULOS	TASA DE NULOS
VAC_DIR_POSTAL	nm_destinatario	char	85	100%
VAC_SUMINISTROS	tp_localidad	tinyint	105121	100%
VCB_CREDITO	cd_empresa	tinyint	105121	100%
VCB_CREDITO	cd_servicio_producto	int	105493	100%
VCB_CREDITO	cd_saldo	tinyint	105493	100%
VFL_LECTURA_CONSUMO_HIST	bo_energia_reactiva	bit	105493	100%
VFL_LECTURA_CONSUMO_HIST	tp_propiedad	char	111005	99%
VCB_CREDITO	nis	int	97863	93%
VAC_SUMINISTROS	fc_ult_facturacion	datetime	71689	68%
VAC_SUMINISTROS	fc_ini_contrato	datetime	72113	68%
VAC_SUMINISTROS	nr_rut_cliente	int	71978	63%
VFL_LECTURA_CONSUMO_HIST	cl_lectura	smallint	66198	63%
VAC_FONOS	fc_ingreso	datetime	65655	59%
VAC_SUMINISTROS	fc_ult_lectura	datetime	55282	52%
VAC_DTE_CONVENIO_AUTORIZADO	fc_aceptacion	datetime	54768	48%
VAC_SUMINISTROS	fc_emision_factura	datetime	45787	43%
VAC_MEDIDOR	vl_cte_energia_reac	decimal	47244	41%
VAC_SUMINISTROS	nr_contrato	int	33483	32%
VFL_LECTURA_CONSUMO_HIST	vl_cte_demanda	decimal	23396	22%
VAC_DTE_CONVENIO_AUTORIZADO	gl_comentario	varchar	24168	21%
VAC_EQUIPOS	nr_serie_equipo	char	6128	11%
VAC_SUMINISTROS	gl_dir_destino_bolfac	varchar	1593	2%
VAC_CLIENTES	nm_apellido_paterno	char	802	1%
VAC_SUMINISTROS	cd_serv_comun	tinyint	1164	1%
VAD_DIRECCION	cd_comuna	char	761	1%
VFL_LECTURA_CONSUMO_HIST	vl_cte_energia_reac	decimal	1191	1%

Tabla 13 Tasa nulidad servidor fililales

Al igual que el servidor de la Casa Central, el servidor de Filiales presenta 12 campos con una tasa de nulidad por encima del 60%.

Existe un análisis a una filial en particular, que por su prioridad en el proceso de migración para la compañía, es por esto que se efectuó dicha filial como un servidor en particular.

A continuación solo se presentan los resultados solo para la filial específica:

NOMBRE TABLA	NOMBRE COLUMNA	TIPO	REGISTROS NULOS	TASA DE NULOS
VAC_ORDENSERVICIO	nr_ficha_inspector	int	11890	100%
VAC_ORDENSERVICIO	vl_valoriz_original	decimal	11890	100%
VAC_ORDENSERVICIO	vl_valoriz_real	decimal	11890	100%
VAC_ORDENSERVICIO	tp_orden_servicio	tinyint	11890	100%
VAC_ORDENSERVICIO	nr_os_anterior	int	11890	100%
VAC_MEDIDOR	cd_formato_lect	INT	5183	99%
VAC_ORDENSERVICIO	nr_factura	int	11273	95%
VAC_ORDENSERVICIO	cd_ejecutor	int	10833	91%
VAC_ORDENSERVICIO	vl_ult_lectura	decimal	10847	91%
VAC_ORDENSERVICIO	fc_asignacion	datetime	9524	80%
VAC_EMPALMES	nr_rut_constructor	int	3272	63%
VAC_SUMINISTROS	nr_folio_sec	char	3119	60%
VAC_SUMINISTROS	tp_localidad	tinyint	2756	53%
VAC_ORDENSERVICIO	nr_ficha_ejecutor	int	6003	50%
VAC_MEDIDOR	fc_fabricacion	datetime	2429	47%
VAC_ORDENSERVICIO	gl_visita	varchar	4640	39%
VAC_ORDENSERVICIO	gl_solicitud	varchar	2849	24%
VAC_FONOS	nr_anexo	smallint	275	12%
VAC_ORDENSERVICIO	gl_ejecutado	varchar	1417	12%
VAC_ORDENSERVICIO	fc_recepcion	datetime	1352	11%
VAC_EMPALMES	nr_rut_instalador	int	368	7%
VAC_CLIENTES	nm_fantasia	char	60	1%
VAC_CLIENTES	nm_razon_social	varchar	60	1%

Tabla 14 Tasa nulidad filial especifica

En el caso particular de la filial especifica solo se presentó 11 campos tienen una tasa de nulidad por encima del 60%.

7.10.1.2. Número de registros distintos

El análisis de registros distintos permite estudiar la cardinalidad de una columna específica. De esta forma se debe comprobar si la variabilidad de registros permitidos concuerda con la lógica de una columna en particular.

De esta forma se sabrá si la cantidad de registros distintos de una columna está dentro de los parámetros permitidos.

Dentro de los resultados expuestos en los archivos anexos, se evidencia que las claves primarias de cada tabla deben ser las columnas con mayor cantidad de registros distintos en

cada tabla. Se pueden encontrar excepciones si es que una tabla posee clave primaria compuesta.

También la entidad evaluada debe tener claridad solo los campos que poseen valores acotados correspondientes a aspectos del negocio.

Por ejemplo en la filial especifica, en la tabla VAC_SUMINISTROS el campo cd_tarifa presenta 20 números de registros distintos. Se debe verificar por la entidad evaluada que esos 20 valores sean válidos.

No es de utilidad realizar la revisión a campos como nr_rut_cliente donde la variabilidad es muy alta.

7.10.1.3. Inconsistencias en número de columnas

El Script SQL al momento de almacenar los datos básicos de las tablas a analizar realiza una consulta a la tabla DD_COLUMNAS solicitando la cantidad de columnas de una tabla específica. Luego, durante su ciclo iterativo el Script realiza la petición de todas las columnas de la tabla analizada para almacenarlas en la tabla DQ_COLUMNA en el campo LARGO_SOPORTADO.

El campo LARGO_MAX_REAL calcula el valor máximo de una columna específica. De esta forma se realiza la evaluación si el largo soportado por el campo corresponde al largo del valor máximo almacenado.

A continuación se presenta los casos que presentan inconsistencias para el servidor de la Casa Central de la compañía**:**

NOMBRE TABLA	COLUMNA	TIPO	LARGO SOPORTADO	LARGO MAX REAL
VAC_DADOR_RECEP	nis_dador	int	4	6
VAC_DADOR_RECEP	nis_receptor	int	4	6
VAC_DETALLE_FACT_UNICA	nr_rut_cliente	int	4	7
VAC_DETALLE_FACT_UNICA	nis_principal	int	4	6
VAC_DETALLE_FACT_UNICA	nis_detalle	int	4	6
VAC_DIR_POSTAL	nis	int	4	6
VAC_DTE_CONVENIO_AUTORIZADO	nis	int	4	6
VAC_DTE_CONVENIO_AUTORIZADO	fc_ingreso	datetime	8	19
VAC_DTE_CONVENIO_AUTORIZADO	fc_aceptacion	datetime	8	19
VAC_EMPALMES	nis	int	4	6
VAC_EMPALMES	fc_construccion	datetime	8	19
VAC_EQUIPOS	cd_equipo	smallint	2	3
VAC_EQUIPOS	nis	int	4	6

VAC_EQUIPOS	fc_instalacion	datetime	8	19
VAC_FACT_UNICA	nr_rut_cliente	int	4	7
VAC_FACT_UNICA	nis_principal	int	4	6
VAC_FONOS	nis	int	4	6
VAC_FONOS	fc_ult_modif	datetime	8	19
VAC_MEDIDOR	nis	int	4	6
VAC_MEDIDOR	fc_lect_ini	datetime	8	19
VAC_MEDIDOR	fc_instalacion	datetime	8	19
VAC_MEDIDOR	fc_fabricacion	datetime	8	19
VAC_ORDENSERVICIO	nr_orden_servicio	int	4	7
VAC_ORDENSERVICIO	nis	int	4	6
VAC_ORDENSERVICIO	nr_solicitud	int	4	6
VAC_ORDENSERVICIO	nr_rut_empresa	int	4	8
VAC_ORDENSERVICIO	fc_pactada_ejec	datetime	8	19
VAC_ORDENSERVICIO	fc_emision	datetime	8	19
VAC_ORDENSERVICIO	fc_asignacion	datetime	8	19
VAC_ORDENSERVICIO	fc_recepcion	datetime	8	19
VAC_SUMINISTROS	nis	int	4	6
VAC_SUMINISTROS	cd_oficina	tinyint	1	2
VAC_SUMINISTROS	fc_ult_facturacion	datetime	8	19
VAC_SUMINISTROS	fc_ult_lectura	datetime	8	19
VAC_SUMINISTROS	fc_emision_factura	datetime	8	19
VAC_SUMINISTROS	fc_ini_contrato	datetime	8	19
VAC_SUMINISTROS	fc_fin_contrato	datetime	8	19
VAC_SUMINISTROS	fc_ini_vigencia	datetime	8	19
VAC_SUMINISTROS	fc_fin_vigencia	datetime	8	19
VAC_SUMINISTROS	fc_ini_sec	datetime	8	19
VAC_SUMINISTROS	fc_fin_sec	datetime	8	19
VCB_CREDITO	nr_nis	int	4	6
VCB_CREDITO	cd_servicio_producto	int	4	6
VCB_CREDITO	fc_credito	datetime	8	19
VCB_CREDITO	fc_fin_vigencia	datetime	8	19
VCB_CREDITO	vl_tasa_cuota	decimal	4	7
VCB_CREDITO	fc_insert	datetime	8	19
VAD_DIRECCIONES	fc_revision_pt	datetime	8	19
VFL_LECTURA_CONSUMO_HIST	nis	int	4	6
VFL_LECTURA_CONSUMO_HIST	fc_lectura_act	datetime	8	19
VFL_LECTURA_CONSUMO_HIST	fc_lectura_ant	datetime	8	19
VFL_LECTURA_CONSUMO_HIST	fc_factur	datetime	8	19

Tabla 15 Inconsistencias de largo servidor central

Se presentan 52 casos con inconsistencias en el servidor de la Casa Central

A continuación se presentan las inconsistencias de largo para el servidor de Filiales

NOMBRE TABLA	NOMBRE COLUMNA	TIPO	LARGO SOPORTADO	LARGO MAX REAL
VAC_CLIENTES	nr_rut_cliente	int	4	8
VAC_CLIENTES	cd_giro_cliente	int	4	6
VAC_CLIENTES	cd_direccion	int	4	6
VAC_CLIENTES	cd_oficina	tinyint	1	2
VAC_CLIENTES	cn_suministros	smallint	2	4
VAC_DADOR_RECEP	nis_dador	int	4	5
VAC_DADOR_RECEP	nis_receptor	int	4	6
VAC_DADOR_RECEP	vl_pctje_pago	decimal	4	6
VAC_DETALLE_FACT_UNICA	nr_rut_cliente	int	4	8
VAC_DETALLE_FACT_UNICA	nis_principal	int	4	6
VAC_DETALLE_FACT_UNICA	nis_detalle	int	4	6
VAC_DIR_POSTAL	nr_rut	int	4	8
VAC_DIR_POSTAL	nis	int	4	6
VAC_DTE_CONVENIO_AUTORIZADO	nis	int	4	6
VAC_DTE_CONVENIO_AUTORIZADO	fc_ingreso	datetime	8	19
VAC_DTE_CONVENIO_AUTORIZADO	nr_solicitud	int	4	6
VAC_DTE_CONVENIO_AUTORIZADO	nr_rut_usuario	int	4	8
VAC_DTE_CONVENIO_AUTORIZADO	nr_orden_servicio	int	4	6
VAC_DTE_CONVENIO_AUTORIZADO	fc_aceptacion	datetime	8	19
VAC_EMPALMES	nis	int	4	6
VAC_EMPALMES	fc_construccion	datetime	8	19
VAC_EMPALMES	cn_proteccion	decimal	4	6
VAC_EMPALMES	cn_mt_edificio_fachada	decimal	4	6
VAC_EMPALMES	cn_mt_acometida	decimal	4	6
VAC_EMPALMES	cn_mt_bajada	decimal	4	6
VAC_EMPALMES	cn_mt_union_tablero	decimal	4	6
VAC_EMPALMES	nr_orden_servicio	int	4	6
VAC_EMPALMES	nr_rut_instalador	int	4	9
VAC_EQUIPOS	cd_equipo	smallint	2	3
VAC_EQUIPOS	nis	int	4	5
VAC_EQUIPOS	cd_tipo_equipo	tinyint	1	2
VAC_EQUIPOS	fc_instalacion	datetime	8	19
VAC_FACT_UNICA	nr_rut_cliente	int	4	8
VAC_FACT_UNICA	nis_principal	int	4	6
VAC_FONOS	nr_rut	int	4	9
VAC_FONOS	nr_fono	int	4	9
VAC_FONOS	nr_anexo	smallint	2	4
VAC_FONOS	nis	int	4	6
VAC_FONOS	fc_ingreso	datetime	8	19

VAC_MEDIDOR	cd_medidor	smallint	2	3
VAC_MEDIDOR	nis	int	4	6
VAC_MEDIDOR	vl_cte_energia_act	decimal	4	6
VAC_MEDIDOR	vl_cte_energia_reac	decimal	4	6
VAC_MEDIDOR	vl_cte_demanda	decimal	5	7
VAC_MEDIDOR	nr_lect_ini_act	int	4	7
VAC_MEDIDOR	nr_lect_ini_reac	int	4	6
VAC_MEDIDOR	fc_lect_ini	datetime	8	19
VAC_MEDIDOR	cn_lect_max_ener_act	int	4	6
VAC_MEDIDOR	cn_lect_max_ener_reac	int	4	6
VAC_MEDIDOR	fc_instalacion	datetime	8	19
VAC_MEDIDOR	nr_orden_servicio	int	4	6
VAC_MEDIDOR	fc_fabricacion	datetime	8	19
VAC_MEDIDOR	cd_tecnol_medidor	TINYINT	0	2
VAC_ORDENSERVICIO	nr_orden_servicio	int	4	6
VAC_ORDENSERVICIO	nis	int	4	6
VAC_ORDENSERVICIO	nr_rut_cliente	int	4	9
VAC_ORDENSERVICIO	nr_solicitud	int	4	6
VAC_ORDENSERVICIO	nr_ficha_emisor	int	4	8
VAC_ORDENSERVICIO	nr_ficha_ejecutor	int	4	6
VAC_ORDENSERVICIO	nr_rut_empresa	int	4	8
VAC_ORDENSERVICIO	cd_ejecutor	int	4	9
VAC_ORDENSERVICIO	fc_pactada_ejec	datetime	8	19
VAC_ORDENSERVICIO	fc_emision	datetime	8	19
VAC_ORDENSERVICIO	fc_asignacion	datetime	8	19
VAC_ORDENSERVICIO	fc_recepcion	datetime	8	19
VAC_ORDENSERVICIO	nr_visitas	tinyint	1	2
VAC_ORDENSERVICIO	nr_factura	int	4	7
VAC_ORDENSERVICIO	nr_ult_evento	tinyint	1	2
VAC_ORDENSERVICIO	vl_ult_lectura	decimal	6	11
VAC_SUMINISTROS	nis	int	4	6
VAC_SUMINISTROS	cd_oficina	tinyint	1	2
VAC_SUMINISTROS	cd_direccion	int	4	6
VAC_SUMINISTROS	nr_rut_usuario	int	4	8
VAC_SUMINISTROS	nr_rut_cliente	int	4	8
VAC_SUMINISTROS	cd_tarifa	smallint	2	3
VAC_SUMINISTROS	vl_lim_kwh_inv	smallint	2	5
VAC_SUMINISTROS	cn_distancia_ssee	decimal	3	4
VAC_SUMINISTROS	vl_factor_pot	decimal	4	5
VAC_SUMINISTROS	cn_dias_vencimiento	tinyint	1	2
VAC_SUMINISTROS	cn_consumo_pactado	decimal	6	7

VAC_SUMINISTROS	cn_potcontrat	decimal	6	7
VAC_SUMINISTROS	cn_ddamax_cont_fp	decimal	6	7
VAC_SUMINISTROS	cn_kwh_compra_anticip	decimal	6	7
VAC_SUMINISTROS	cn_pot_declarada	decimal	6	7
VAC_SUMINISTROS	cn_pot_solicitada	decimal	6	7
VAC_SUMINISTROS	fc_ult_facturacion	datetime	8	19
VAC_SUMINISTROS	fc_ult_lectura	datetime	8	19
VAC_SUMINISTROS	fc_emision_factura	datetime	8	19
VAC_SUMINISTROS	fc_ini_contrato	datetime	8	19
VAC_SUMINISTROS	fc_fin_contrato	datetime	8	19
VAC_SUMINISTROS	fc_ini_vigencia	datetime	8	19
VAC_SUMINISTROS	fc_fin_vigencia	datetime	8	19
VAC_SUMINISTROS	fc_ini_sec	datetime	8	19
VAC_SUMINISTROS	fc_fin_sec	datetime	8	19
VAC_SUMINISTROS	nr_ficha_rep_chilv	int	4	7
VAC_SUMINISTROS	nr_contrato	int	4	6
VCB_CREDITO	nis	int	4	6
VCB_CREDITO	nr_credito	tinyint	1	3
VCB_CREDITO	cd_servicio_producto	int	4	6
VCB_CREDITO	fc_credito	datetime	8	19
VCB_CREDITO	nr_docto_venta	int	4	8
VCB_CREDITO	nr_folio_int_bolfac	int	4	7
VCB_CREDITO	fc_fin_vigencia	datetime	8	19
VCB_CREDITO	nr_version	tinyint	1	2
VCB_CREDITO	nr_orden_servicio	int	4	6
VCB_CREDITO	nr_solicitud	int	4	6
VCB_CREDITO	ot_asociada	int	4	6
VCB_CREDITO	vl_tasa_cuota	decimal	4	7
VCB_CREDITO	vl_capital_credito	decimal	8	13
VCB_CREDITO	vl_intereses_credito	decimal	8	12
VCB_CREDITO	vl_saldo_capital_cred	decimal	8	13
VCB_CREDITO	vl_contado_inicial	decimal	8	12
VCB_CREDITO	vl_amortizacion	decimal	8	12
VCB_CREDITO	vl_interes_cuota	decimal	8	11
VCB_CREDITO	vl_cuota	decimal	8	15
VCB_CREDITO	cn_cuotas	tinyint	1	3
VCB_CREDITO	cn_meses_gracia	tinyint	1	2
VCB_CREDITO	cn_cuotas_facturadas	tinyint	1	2
VCB_CREDITO	cn_cuotas_canceladas	tinyint	1	2
VCB_CREDITO	dia_venc_cuota	tinyint	1	2
VCB_CREDITO	mes_cobra_cuota_uno	tinyint	1	2

VCB_CREDITO	id_usr_aprueba	int	4	6
VCB_CREDITO	id_usr_insert	int	4	6
VCB_CREDITO	fc_insert	datetime	8	19
VAD_DIRECCION	cd_direccion	int	4	6
VAD_DIRECCION	cd_ssee_poder	smallint	2	4
VAD_DIRECCION	cd_ssee_distribucion	int	4	6
VAD_DIRECCION	nis	int	4	6
VFL_LECTURA_CONSUMO_HIST	cd_medidor	smallint	2	3
VFL_LECTURA_CONSUMO_HIST	cd_oficina	tinyint	1	2
VFL_LECTURA_CONSUMO_HIST	nis	int	4	6
VFL_LECTURA_CONSUMO_HIST	vl_cte_energia_acti	decimal	4	6
VFL_LECTURA_CONSUMO_HIST	vl_cte_energia_reac	decimal	4	6
VFL_LECTURA_CONSUMO_HIST	vl_cte_demanda	decimal	5	7
VFL_LECTURA_CONSUMO_HIST	fc_lectura_act	datetime	8	19
VFL_LECTURA_CONSUMO_HIST	vl_lectura_activo_act	int	4	6
VFL_LECTURA_CONSUMO_HIST	vl_consumo_activo_act	int	4	7
VFL_LECTURA_CONSUMO_HIST	vl_lectura_reactivo_act	int	4	6
VFL_LECTURA_CONSUMO_HIST	vl_consumo_reactivo_act	int	4	6
VFL_LECTURA_CONSUMO_HIST	fc_lectura_ant	datetime	8	19
VFL_LECTURA_CONSUMO_HIST	vl_lectura_activo_ant	int	4	6
VFL_LECTURA_CONSUMO_HIST	vl_lectura_reactivo_ant	int	4	6
VFL_LECTURA_CONSUMO_HIST	vl_dem_leida_hpta	decimal	6	7
VFL_LECTURA_CONSUMO_HIST	vl_dem_leida_fpta	decimal	6	8
VFL_LECTURA_CONSUMO_HIST	cs_promedio_diario	decimal	6	11
VFL_LECTURA_CONSUMO_HIST	fc_factur	datetime	8	19

Tabla 16 Inconsistencia de largo servidor Filiales

Existen 146 casos de inconsistencia de largo en el servidor de Filiales.

7.10.1.4. Tasa de reglas incorrectas

Las reglas entregadas por la compañia fueron tomadas en cuenta como métricas de impacto comercial que afectarán de manera Bloqueante el proceso de migración en el sistema Sybase del servidor de la Casa Central de la empresa.

El Script SQL se encargó de almacenar si las reglas se cumplen de manera correcta o incorrecta. A continuación en la siguiente tabla se presentan las reglas su cantidad de registros incorrectos encontrados en servidor de la Casa Central:

NRO.	NOMBRE DE LA REGLA	TABLA	TIPO	INCORRECTAS
16	Cd_tipo_fono_inexistente	VAC_FONOS	Bloqueante	109738
36	Solicitudes y OS ejecutadas sin fechas	VAC_ORDENSERVICIO	Bloqueante	24648
60	Verifica si datos del suministro genera "clase empalme" en OSF	VAC_EMPALMES	Bloqueante	13076
19	Nr_fono inválido	VAC_FONOS	No Bloqueante	7868
38	Caracteres no permitidos por OSF en direcciones	VAD_DIRECCIONES	Bloqueante	7287
15	Email incosistente	VAC_DTE_CONVENIO_AUTORIZADO	Bloqueante	4941
2	Categoría vs Tipo	VAC_SUMINISTROS	Bloqueante	2848
32	Suministro con consumo pactado 0 kwh	VAC_SUMINISTROS	Bloqueante	2656
17	Cd_area_telefonica inválido	VAC_FONOS	Bloqueante	2194
25	Glosas de dirección repetidas	VAD_DIRECCIONES	Bloqueante	1282
41	Suministros-clientes con direcciones sin comuna	VAC_CLIENTES	Bloqueante	855
44	Validar fecha fin vigencia del suministro	VAC_SUMINISTROS	Bloqueante	814
44	Validar fecha fin vigencia del suministro provisorio	VAC_SUMINISTROS	Bloqueante	641
61	Verifica si datos del medidor genera "clase medidor"	VAC_MEDIDOR	Bloqueante	616
57	Valida existencia de transf, cuando existe recargo	VAC_SUMINISTROS	Bloqueante	372
24	Relación fc_lectura - fc_instalacion_medidor	VAC_MEDIDOR	Bloqueante	208
44	Validar fechas del suministro	VAC_SUMINISTROS	Bloqueante	183
65	Caracteres no permitidos por OSF en dirección postal	VAC_DIR_POSTAL	Bloqueante	178
27	Prorrateos que no suman 100%	VAC_DADOR_RECEP	Bloqueante	161
28	Inconsistencias entre protección y potencia	VAC_EMPALMES	No Bloqueante	87
53	Valida tecnología medidor	VAC_MEDIDOR	Bloqueante	77
3	Provisorios con tarifa no provisoria	VAC_SUMINISTROS	Bloqueante	68
47	Información en longitud de empalme	VAC_EMPALMES	Bloqueante	45

4	Ciudad despacho postal	VAC_DIR_POSTAL	Bloqueante	44
56	Valida consistencia en el sector para Fact única	VAC_SUMINISTROS	Bloqueante	39
62	Verifica si datos del suministro genera "clase transformador"	VAC_EQUIPOS	Bloqueante	35
35	Clientes y Suministros con cd_direccion = 0	VAC_CLIENTES	Bloqueante	31
6	Medidor incorrecto tarifa 3	VAC_MEDIDOR	Bloqueante	23
42	Área típica válida	VAC_SUMINISTROS	Bloqueante	23
26	Suministro sin ruta	VAC_SUMINISTROS	Bloqueante	20
59	si tiene reparto email, debe tener correo válido	VAC_SUMINISTROS	Bloqueante	18
31	Corrección cd_empalme inválido	VAC_EMPALMES	Bloqueante	16
39	Suministro con consumo pactado y medidor	VAC_SUMINISTROS	Bloqueante	8
8	Medidor incorrecto tarifa 4.3	VAC_MEDIDOR	Bloqueante	6
33	Identificar rut mayores a 100.000.000	VAC_CLIENTES	No Bloqueante	3
35	Clientes y Suministros con cd_direccion = 0	VAC_SUMINISTROS	Bloqueante	2
41	Suministros-clientes con direcciones sin comuna	VAC_SUMINISTROS	Bloqueante	2
54	Valida tipo equipo para medidor	VAC_MEDIDOR	Bloqueante	2
10	Suministros no provisorios con tarifa provisoria	VAC_SUMINISTROS	Bloqueante	1
29	Corrige rut del cliente según rut del usuario	VAC_SUMINISTROS	No Bloqueante	1

Tabla 17 Registros incorrectos servidor Central

Se detectaron 40 reglas que tienen registros incorrectos. Solamente 4 reglas tienen calidad de No Bloqueantes para el proceso de migración.

A continuación se presentan las reglas con valores incorrectos para el servidor de Filiales:

NRO	NOMBRE DE LA REGLA	TABLA	TIPO	INCORRECTAS
28	Inconsistencias entre protección y potencia	VAC_EMPALMES	No Bloqueante	46124
44	Validar fecha fin vigencia del suministro	VAC_SUMINISTROS	Bloqueante	13068
64	Tipo reparto custodia oficina y oficina santiago	VAC_SUMINISTROS	Bloqueante	3220
25	Glosas de dirección repetidas	VAD_DIRECCIONES	Bloqueante	2227
42	Área típica válida	VAC_SUMINISTROS	Bloqueante	1663
36	Solicitudes y OS ejecutadas sin fechas	VAC_ORDENSERVICIO	Bloqueante	1593
41	Suministros-clientes con direcciones sin comuna	VAC_CLIENTES	Bloqueante	1262
39	Suministro con consumo pactado y medidor	VAC_SUMINISTROS	Bloqueante	1117
61	Verifica si datos del medidor genera "clase medidor"	VAC_MEDIDOR	Bloqueante	1037
60	Verifica si datos del suministro genera "clase empalme" en OSF	VAC_EMPALMES	Bloqueante	812
30	Corrige rut del usuario según rut del cliente	VAC_SUMINISTROS	No Bloqueante	530
3	Provisorios con tarifa no provisoria	VAC_SUMINISTROS	Bloqueante	528
44	Validar fechas del suministro	VAC_SUMINISTROS	Bloqueante	410
57	Valida existencia de transf, cuando existe recargo	VAC_SUMINISTROS	Bloqueante	398
32	Suministro con consumo pactado 0 kwh	VAC_SUMINISTROS	Bloqueante	227
65	Caracteres no permitidos por OSF en dirección postal	VAC_DIR_POSTAL	Bloqueante	191
53	Valida tecnología medidor	VAC_MEDIDOR	Bloqueante	159
24	Relación fc_lectura - fc_instalacion_medidor	VAC_MEDIDOR	Bloqueante	118
38	Caracteres no permitidos por OSF en direcciones	VAD_DIRECCIONES	Bloqueante	113
4	Ciudad despacho postal	VAC_DIR_POSTAL	Bloqueante	38
26	Suministro sin ruta	VAC_SUMINISTROS	Bloqueante	27
13	cd_reparto_docto_cobro_inexistentes	VAC_SUMINISTROS	Bloqueante	14
16	Cd_tipo_fono_inexistente	VAC_FONOS	Bloqueante	11
2	Categoría vs Tipo	VAC_SUMINISTROS	Bloqueante	6
54	Valida tipo equipo para medidor	VAC_MEDIDOR	Bloqueante	5
29	Corrige rut del cliente según rut del usuario	VAC_SUMINISTROS	No Bloqueante	4
6	Medidor incorrecto tarifa 3	VAC_MEDIDOR	Bloqueante	3
15	Email incosistente	VAC_DTE_CONVENIO_AUTORIZADO	Bloqueante	3
31	Corrección cd_empalme inválido	VAC_EMPALMES	Bloqueante	3
8	Medidor incorrecto tarifa 4.3	VAC_MEDIDOR	Bloqueante	2
7	Medidor incorrecto tarifa 4.2	VAC_MEDIDOR	Bloqueante	1
33	Identificar rut mayores a 100.000.000	VAC_CLIENTES	No Bloqueante	1
55	Valida código de medidor	VAC_MEDIDOR	Bloqueante	1

Tabla 18 Registros incorrectos servidor Filiales

En el servidor de Filiales existen solo 33 reglas que poseen registros incorrectos. De las 33 reglas, solo 4 tienen calidad de No Bloqueante destacando la regla 28 como primera en el listado.

A continuación se presentan las reglas de validación con la cantidad de registros incorrectos presentados para la filial específica en estudio:

NRO	NOMBRE DE LA REGLA	TABLA	TIPO	INCORRECTAS
28	Inconsistencias entre protección y potencia	VAC_EMPALMES	No Bloqueante	1529
36	Solicitudes y OS ejecutadas sin fechas	VAC_ORDENSERVICIO	Bloqueante	1352
61	Verifica si datos del medidor genera "clase medidor"	VAC_MEDIDOR	Bloqueante	1037
41	Suministros-clientes con direcciones sin comuna	VAC_CLIENTES	Bloqueante	420
44	Validar fecha fin vigencia del suministro	VAC_SUMINISTROS	Bloqueante	367
44	Validar fechas del suministro	VAC_SUMINISTROS	Bloqueante	337
57	Valida existencia de transf, cuando existe recargo	VAC_SUMINISTROS	Bloqueante	81
60	Verifica si datos del suministro genera "clase empalme" en OSF	VAC_EMPALMES	Bloqueante	41
53	Valida tecnología medidor	VAC_MEDIDOR	Bloqueante	37
3	Provisorios con tarifa no provisoria	VAC_SUMINISTROS	Bloqueante	36
38	Caracteres no permitidos por OSF en direcciones	VAD_DIRECCIONES	Bloqueante	20
39	Suministro con consumo pactado y medidor	VAC_SUMINISTROS	Bloqueante	5
54	Valida tipo equipo para medidor	VAC_MEDIDOR	Bloqueante	5
65	Caracteres no permitidos por OSF en dirección postal	VAC_DIR_POSTAL	Bloqueante	4
25	Glosas de dirección repetidas	VAD_DIRECCIONES	Bloqueante	3
33	Identificar rut mayores a 100.000.000	VAC_CLIENTES	No Bloqueante	1

Tabla 19 Registros incorrectos filial específica

Según la tabla presentada, solo 16 reglas poseen registros incorrectos en la cual 2 de estas reglas tienen calidad de No Bloqueante.

7.10.1.5. Análisis de distribución de los datos

La distribución de los datos por cada tabla es la forma de saber qué valores tienen una elevada frecuencia de ocurrencia. Se consideró ejecutar la distribución para columnas con número de registros distintos menores o iguales a 30.

Se recomienda la revisión de la tabla DQ_DISTRIBUCION de todos los archivos anexos correspondientes para observar qué valores el sistema está permitiendo ingresar y no tienen sentido lógico con el propósito de una columna específica.

A través de esto se pueden generar más reglas de validación por parte de la empresa para corregir valores anómalos. Este análisis ayudará en la revisión de valores distintos con un rango acotado.

Por ejemplo en Casablanca el campo cd_zona de la tabla VAC_CLIENTES reporta en DQ_COLUMNA como números distintos existentes. En la tabla de distribución DQ_DISTRIBUCION efectivamente el cd_zona posee dos valores: "1" y " ". Existen 3750 registros con valor "1" y 456 registros con valor " ".

Sumando estos valores se comprueba que hay 4206 clientes como indica el campo cd_empresa de la misma tabla.

7.11. Validación de reglas para la filial específica en estudio

A continuación se presentan los resultados obtenidos al ejecutar las reglas de validación para la filial específica en estudio. El ambiente para la ejecución de las reglas es el servidor de producción FILIALES en el cual están almacenados los datos actuales de la filial.

Se ha especificado un conjunto de reglas que presentaban diferencias entre los resultados obtenidos por el equipo de trabajo del proyecto y el equipo de la compañía.

Posteriormente se realizó un trabajo en conjunto con un equipo de expertos de la compañía llegando a los siguientes resultados:

ID_REGLA	TIPO_REGLA	Incorrectos	Observación
3	Bloqueante	NO APLICA	
27	Bloqueante	0	
28	No Bloqueante	12	
36	Bloqueante	0	Corrige compañía en migración
38	Bloqueante	0	
39	Bloqueante	5	
41.2	Bloqueante	0	Corrige compañía en migración
42	Bloqueante	8	
44.3	Bloqueante	0	
44.2	Bloqueante	8	
44.1	Bloqueante	3	
57	Bloqueante	81	Corrige compañía en migración
60	Bloqueante	17	
61	Bloqueante	15	
65	Bloqueante	0	

Tabla 20 Aplicacion de reglas filial Casablanca

Con respecto a las reglas 27, 36, 38, 41.2, 44.3 y 65 de la filial específica se llevó un proceso de corrección obteniendo cero datos incorrectos. Se especifica que la regla 3 no se aplica para esta filial. Las reglas 36, 41.2 (segunda división de la regla) y 57 serán corregidas durante el proceso de migración por parte de la compañía. La regla 42 presenta **una excepción** con los 8 registros encontrados, esto ha sido informado al equipo de calidad de datos de la compañía.

El resto de reglas que contienen todavía datos incorrectos deberán ser corregidas por parte del equipo de migración de la compañía.

8. Conclusiones

Antes de presentar las conclusiones específicas de la implementación llevada a cabo en el contexto del proyecto, quisiéramos establecer algunas conclusiones de carácter general.

En primer lugar quisiéramos consignar que el árbol decisional propuesto en el apartado metodológico es una herramienta invaluable a la hora de establecer un punto de partida en el complejo escenario que estos proyectos plantean, sin embargo es necesario consignar que su aplicación tiene algunos supuestos base que es requisito conocer. El primer elemento a tener en presente es que se debe contar con una base de conocimiento de técnicas a aplicar para la corrección y limpieza de los datos. Estas técnicas se encuentran fuertemente vinculadas al tipo de datos con el que se está trabajando y el dominio semántico de los mismos. Si bien algunas técnicas rozan lo que podríamos llamar el sentido común, otras en cambio requieren un entendimiento mayor o al menos la conciencia de su existencia.

Junto con la base de técnicas asociadas a la limpieza de los datos y una vez que se ha establecido el conjunto de las mismas que tienen relevancia para un campo determinado es necesario contar con un set de criterios de evaluación para discriminar cuál de las técnicas en primera instancia seleccionadas es la que mejor cumple con los requerimientos del proceso de limpieza.

Existen numerosas categorizaciones de técnicas y criterios asociados, lo que este trabajo ha demostrado es que sin importar la elección de una u otra taxonomía lo importante es, una vez establecido el conjunto base con de técnicas y criterios con el cual trabajar, la correcta identificación de los tipos de datos y sus dominios.

Si bien la automatización de las tareas de análisis y corrección de los datos es una tarea específica y será prácticamente única para cada proyecto, es posible establecer ciertos patrones de diseño y modelos comunes que permitan estandarizar un conjunto de librerías que admitiendo adaptaciones, mantengan una estructura reconocible y constante.

Contar con estos elementos es una gran ayuda en esta clase de proyectos que si bien no pueden nunca ser considerados como un producto final, creemos firmemente ameritan tomarse el tiempo para estandarizar y protocolizar todo lo que se pueda el proceso de análisis, diseño y desarrollo de los mismos.

8.1. Aspectos positivos

Como elementos positivos el equipo de trabajo destaca:

- De acuerdo a la prioridad del proceso de migración de la filial específica, la validación de reglas presentó un bajo porcentaje de datos incorrectos en los servidores de producción.
- El equipo de calidad de datos de la compañía más allá de la implementación del proyecto se encuentra en el desarrollo de nuevas reglas de validación para asegurar un proceso de migración exitoso.
- La gestión de detección y corrección de errores por parte del equipo de calidad de datos de la compañía tiene un óptimo funcionamiento ya que se encuentra en constante generación de sentencias SQL con el fin de llevar la corrección de forma inmediata.
- A nivel general, el equipo de trabajo se ha percatado que la empresa en la cual se llevó a cabo implementación del proyecto ha realizado un trabajo extenso de mejoras sobre la calidad de sus datos para asegurar un proceso de migración exitoso.

8.2. Recomendaciones

A continuación se presentan una serie de recomendaciones elaboradas por el equipo de trabajo con el fin de tener un proceso de migración con la mínima cantidad de errores:

- Se recomienda disminuir la tasa de nulidad de algunas columnas. Las columnas con tasa de nulidad por encima del 60% deberían ser normalizadas para eliminar redundancias innecesarias en una tabla. Los datos nulos no entregan información relevante para el core de la empresa y a la vez no entregan información relevante para realizar análisis.

- Se recomienda realizar una revisión de números de registros distintos para cada columna que posea una cantidad de valores acotados menores a 30 necesarios para el negocio. Por ejemplo una columna de códigos de teléfonos debería solo tener valores correspondientes a los códigos de teléfonos permitidos por el sistema.

- En tanto para las reglas de validación, se recomienda al equipo de calidad de datos de a compañía revisar las *querys* desarrolladas por el equipo de trabajo en los casos donde existan diferencias de resultados. La revisión de las *querys* de control deben realizarse

de manera rápida, sobre todo la de la filial específica en estudio por su nivel de prioridad.

- Se recomienda realizar una revisión de errores semánticos en los datos. Por ejemplo una columna de códigos de empresa solamente puede poseer valores de empresas reales que están asociadas a la compañía. Revisar si los datos almacenados tienen una correspondencia en el mundo real.

- Se recomienda realizar una revisión de errores sintácticos. La compañía deberá definir formatos válidos para todas sus columnas y luego verificar si se cumplen. La definición de formatos deberá ser desarrollada pensando en la migración hacia la plataforma Oracle.

- Se recomienda realizar una revisión sobre la precisión de datos (nivel de detalle sobre la representación de los datos) para columnas cuya presentación de los datos no afecte a la entidad evaluada. Por ejemplo en los campos que impliquen una fecha, definir si los valores corresponden a día-mes, día-mes-año o
  dia-mes-año:tiempo. Esta recomendación no debería tener prioridad para el proceso de migración.

- Se recomienda actualizar el Diccionario de Datos para corregir inconsistencias entre el largo especificado como máximo soportado por cada columna y el definido realmente en la Base de Datos.

8.3. Conclusiones de la implementación

De acuerdo al análisis realizado se encontraron los siguientes problemas relacionados con la validación de reglas para calidad de datos:

- Para el servidor de la Casa Central de la compañía, con sus datos *actualizados hasta el término de este proyecto*, se concluye que 40 reglas tienen registros incorrectos, dentro de las cuales 4 reglas tienen calidad de No Bloqueantes para el proceso de migración.

- Para el servidor de Filiales en su totalidad con sus datos *actualizados hasta el término del proyecto*, se concluye que 33 reglas tienen registros incorrectos, dentro de las cuales 4 reglas tienen calidad de No Bloqueantes para el proceso de migración.

- Para la filial *especifica en estudio* con sus datos *actualizados hasta el término del proyecto*, se concluye que 7 reglas tienen registros incorrectos. Las reglas 36, 41.2 y 57

serán corregidas durante el proceso de migración. La regla 42 posee 8 registros incorrectos que son válidos, esto fue informado por el equipo de calidad de datos de la empresa como una excepción. Se destaca que la compañía se encuentra en proceso de corrección de estas reglas, se ejecutarán *queries* de corrección para tener cero reglas con registros incorrectos.

Como equipo de trabajo en la implementación del proyecto de calidad de datos, se concluye luego de los distintos análisis realizados a las bases de datos de una compañía de distribución servicios eléctricos, que tanto en su servidor de la Casa Central y Filiales presentan inconsistencias a nivel de calidad de datos las cuales se han evidenciado en el presente documento y sus archivos anexos.

Sin embargo, el equipo de migración de la compañía tiene detectadas dichas inconsistencias y a la fecha está trabajando en solucionarlas. La filial de específica en estudio presenta errores mínimos que a la fecha de este informe están siendo corregidos por el equipo de migración de la compañía.

Por consiguiente, con la calidad de datos actual, no debieran existir dificultades para ejecutar el paso a producción de la filial específica en el nuevo motor de base de datos Oracle. Es claro que posteriormente los esfuerzos de corrección deben abarcar a las demás filiales y la Casa Central de la misma.

9. Referencias

[1] DATE, C: J. Introducción a los Sistemas de Bases de Datos. 7 ed. Pearson Education, 2007.

[2] VALERO, Nelica. Bases de Datos Columnares. 2009.

[3] Sybase IQ Columnar Database, Column-Based & Oriented DBMS – Analytics Server Data Warehouse – Sybase Inc. Octubre 2011,

[4] Abril D., Pérez J. 2007. Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a bases de datos espaciales, Abril 2007.

[5] Arturo L., Carmona C. 2001. Guía para obtener el retorno a la inversion en proyectos de Data Warehouse.

[6] Fernández J., Mayol E. y Pastor J. 2008. Agile Business Intelligence Governance: Su justificación y presentación.

[7] Tamayo M., Moreno F.2006. Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP, Diciembre 2006.

[8] Data quality : concepts, methodologies and techniques / Carlo Batini, Monica Scannapieca. Publication *Info*. Berlin ; New York : Springer, 2006.

[9] Data Quality Maintenance in Data Integration Systems,Phd thesis, Marotta A. 2009

[10] Dasu, T., Vesonder, G. T., y Wright, J. R. 2003. Data quality through knowledge engineering. En: Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining ACM SIGKDD 2003 (Washington, D.C., Agosto 24 - 27, 2003). KDD '03. ACM, Nueva York, NY, 705-710

[11] Oliveira, P., Rodrigues, F., Henriques, P., y Galhardas, H. 2005. A Taxonomy of Data Quality Problems. En: Second International Workshop on Data and Information Quality IQIS 2005 (Porto, Portugal, Junio 13-17, 2005).

[12] Ristad, E., y Yianilos, P. 1998. Learning string edit distance. 1998. IEEE Trans. Pattern Analysis and Machine Intelligence, 20(5), 522-532, Mayo, 1998.

[13] Sybase Inc., "Lo nuevo en Sybase Adaptative ServerEnterpise" , Adaptative Server Enterprise 12.5, ID:32961-01-1250-01, última revisión junio 2001. Disponible vía web en : http://www.academia.edu/9877811/Enterprise_Adaptive_Server_Enterprise. Última vez visitado 07-04-2015.

[14] Technet, "Referencia de Transact-SQL (Transanct-SQL)", Referencia de lenguajes de SQL Sever, LIbrary, The TechNet Library of Microsoft. Disponible vía web en:https://technet.microsoft.com/es-es/library/ms189826(v=sql.90).aspx. Última vez revisado 07-04-2015.

[15] Peralta V., " Data Quality Evaluation in Data Integration Systems", Human Computer Interaction, Universit´e de Versailles-Saint Quentin en Yvelines, Universit´e de la R´epublique d'Uruguay, 2006.