ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN Y VINCULACIÓN CON LA COLECTIVIDAD

MAESTRÍA EN GERENCIA DE SISTEMAS
IX PROMOCIÓN

TESIS DE GRADO MAESTRÍA EN GERENCIA DE SISTEMAS

TEMA: "DESARROLLO DE UN CASO DE ESTUDIO APLICANDO EL MODELO DE GESTIÓN DE CALIDAD DE DATOS BASADO EN LA METODOLOGÍA IBM DATA QUALITY PARA EL PORTAFOLIO DE BUSINESS INTELLIGENCE DE LA EMPRESA DWCONSULWARE"

AUTORES:
ANDRADE TIRIRA, CHRISTIAN ANDRÉS
MADRID RUIZ, DAVID ALEJANDRO

DIRECTOR:
ING. FERNANDO GALARRAGA, Msc.

SANGOLQUÍ, ENERO DE 2015

UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE

MAESTRÍA EN GERENCIA DE SISTEMAS

CERTIFICACIÓN

ING. FERNANDO GALÁRRAGA MSc.

C E R T I F I C A

El proyecto de grado titulado "DESARROLLO DE UN MODELO DE GESTIÓN DE CALIDAD DE DATOS BASADO EN LA METODOLOGÍA IBM DATA QUALITY PARA EL PORTAFOLIO DE BUSINESS INTELLIGENCE DE LA EMPRESA DWCONSULWARE", realizado por los señores Christian Andrés Andrade Tirira y David Alejandro Madrid Ruiz, ha sido guiado y revisado periódicamente y cumple normas estatuarias establecidas por la Universidad de las Fuerzas Armadas ESPE, por tanto, se autoriza su presentación para los fines legales pertinentes.

Sangolquí, Enero de 2015

Ing. Fernando Galárraga MSc.

Director

UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE

MAESTRÍA EN GERENCIA DE SISTEMAS

DECLARACIÓN DE RESPONSABILIDAD

CHRISTIAN ANDRADE Y DAVID MADRID

DECLARAMOS QUE:

El proyecto de grado denominado "DESARROLLO DE UN CASO DE ESTUDIO APLICANDO EL MODELO DE GESTIÓN DE CALIDAD DE DATOS BASADO EN LA METODOLOGÍA IBM DATA QUALITY PARA EL PORTAFOLIO DE BUSINESS INTELLIGENCE DE LA EMPRESA DWCONSULWARE", ha sido desarrollado en base a una investigación exhaustiva, respetando los derechos intelectuales de terceros, conforme las citas que constan en cada texto correspondiente, cuyas fuentes se incorporan en la bibliografía.

Consecuentemente este trabajo es nuestra autoría.

En virtud de esta declaración, nos responsabilizamos del contenido, veracidad y alcance científico del proyecto de grado en mención.

Sangolquí, Enero de 2015

Christian Andrade Tirira

David Madrid Ruiz

UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE

MAESTRÍA EN GERENCIA DE SISTEMAS

AUTORIZACIÓN

Nosotros, Christian Andrade y David Madrid

Autorizamos a la Universidad de las Fuerzas Armadas – ESPE la publicación, en la biblioteca virtual de la Institución del trabajo “DESARROLLO DE UN CASO DE ESTUDIO APLICANDO EL MODELO DE GESTIÓN DE CALIDAD DE DATOS BASADO EN LA METODOLOGÍA IBM DATA QUALITY PARA EL PORTAFOLIO DE BUSINESS INTELLIGENCE DE LA EMPRESA DWCONSULWARE”, cuyo contenido, ideas y criterios son de nuestra exclusiva responsabilidad y autoría.

Sangolquí, Enero de 2015

Christian Andrade Tirira

David Madrid Ruiz

DEDICATORIA

Este trabajo se lo dedicamos a Dios, quien supo guiarnos en todo momento y darnos fuerzas para continuar y no desmayar a pesar de todas las adversidades que se presentaron en el camino. A nuestras familias por su permanente apoyo y solidaridad, aun cuando la meta parecía inalcanzable.

Christian - David

AGRADECIMIENTO

A nuestros compañeros, profesores y el coordinador de la Novena Promoción de la Maestría en Gerencia de Sistemas, por hacer de cada clase un lugar de aprendizaje y encuentro, donde fue posible compartir conocimientos, experiencias y opiniones, que nos han ayudado a convertirnos en mejores profesionales y personas.

Christian - David

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

RESUMEN

Los modelos de madurez permiten evaluar la percepción, aprovechamiento y soporte de la inteligencia de negocios (BI) dentro de una organización, por otro lado la Gestión de Calidad de Datos se enfoca en corregir y disminuir los defectos en los datos utilizados para procesos de BI, la combinación de ambos conceptos nos permite efectivamente evaluar las ventajas de la aplicación de la Gestión de Calidad de Datos. El objeto de este trabajo es determinar el impacto que tiene aplicar un Modelo de Gestión de Calidad de Datos en el nivel de madurez de BI. Partiendo desde una evaluación inicial del nivel de madurez en un proyecto de BI, hemos utilizado una herramienta de Gestión de Calidad de Datos y finalmente se ha realizado una segunda evaluación. Al final de este trabajo, se encuentra que si bien la aplicación del Modelo de Gestión de Calidad de Datos mejora el puntaje de un área específica del modelo de madurez, no es suficiente por sí mismo para elevar el nivel de madurez de BI de la organización en su conjunto.

PALABRAS CLAVES:
CALIDAD DE DATOS
MODELOS DE MADUREZ DE BI
PROYECTO DE BI
TDWI

ABSTRACT

Maturity models allow measuring the perception, use and support of Business Intelligence (BI) within an organization, Data Quality Management focuses on correct and reduce defects in the data used for BI processes, the combination of both concepts effectively allows us to evaluate the benefits of implementing Data Quality Management processes. The purpose of this study is to determine the impact of implementing a Data Quality Management Model on the BI maturity level. Starting from an initial assessment of the level of maturity in a BI project, we used a Data Quality Management suite and finally made a second assessment. At the end of this paper, we found that while the implementation of the Management Model of Data Quality improves the score in a specific area of the maturity model is not sufficient by itself to raise the level of BI maturity within the complete organization.

KEY WORDS:
CALIDAD DE DATOS
MODELOS DE MADUREZ DE BI
PROYECTO DE BI
TDWI

CAPÍTULO I

1 ANTECEDENTES

La Gestión de Calidad de Datos es el proceso que comprueba que los datos tengan los valores correctos y tipos de datos válidos, comprende una serie de procedimientos y técnicas aplicadas con el fin de asegurar que los datos disponibles cumplan con una serie de requisitos y características que los hagan “apropiados para los usos previstos en las operaciones, toma de decisiones y planificación” (Menéndez, 2013).

Este proyecto de tesis se enfocará en desarrollar un modelo de Gestión de la Calidad de Datos que permita estandarizar los proyectos de desarrollo de Business Intelligence basado en la metodología IBM Data Quality la cual básicamente define a la Calidad de Datos como información completa, válida, coherente y precisa, lo que hace a los datos apropiados para un uso específico.

El Modelo de Gestión desarrollado será aplicado en el portafolio de Business Intelligence de la empresa DWConsulware, la cual es líder en soluciones de administración de desempeño corporativo (CPM). Los productos y servicios de esta empresa permiten que las organizaciones alcancen un alto desempeño planificando, presupuestando, pronosticando, monitoreando con alertas o tableros de control y analizando con BI.

DWConsulware es una empresa con más de 15 años en el mercado, presta sus servicios a las más importantes empresas del país en diferentes industrias. Su alto nivel de servicio le ha permitido destacarse en Latinoamérica recibiendo premios importantes de su asociado principal IBM.

Para la elaboración de este proyecto se analizarán en forma general los Modelos de Gestión de Calidad de Datos y de forma particular el Modelo de

Gestión de IBM, igualmente se analizarán los diferentes Modelos de Madurez de Inteligencia de Negocios y específicamente el que plantea The Data Warehousing Institute (TDWI).

En cuanto a la Metodología y Técnicas de Investigación:

- Es una *investigación Aplicada* ya que a través del trabajo se obtendrá un modelo de gestión de calidad de datos para DWConsulware.

- La investigación se abordará de forma *Cualitativa* esto debido a que el trabajo se basa en estándares definidos por IBM.

- Los objetivos son propios de una *investigación Explicativa* porque el trabajo está dirigido a establecer un procedimiento que permita administrar la calidad de datos en las empresas.

- Para esta investigación se utilizará el procedimiento técnico del Estudio de Caso, el cual se basa en el *método científico inductivo* ya que se va a desarrollar un caso en base al modelo de gestión de calidad de datos.

1.1 Justificación e Importancia

Fundada en 1998, DWConsulware es líder en el país en cuanto la implementación de proyectos de inteligencia de negocios. Su principal base de operación se encuentra en Quito, Ecuador, sin embargo tiene instalaciones en Ecuador y Perú, su portafolio de clientes se extiende a varios países de América Latina. Para seguir innovando y continuar con su crecimiento la empresa está incursionando en nuevas líneas de negocio como la Calidad de Datos. Esta línea requiere de un modelo definido para poder llevar a cabo los proyectos de Gestión para sus clientes. Este proyecto de tesis se enfocará en aplicar el modelo de Gestión de la Calidad de Datos basado en la

metodología IBM Data Quality de la empresa DWConsulware en un caso de estudio concreto.

1.2 Planteamiento del Problema

DWConsulware tiene definida una metodología de proyectos de Inteligencia de Negocios cuya implementación realiza para sus clientes, sin embargo, dentro del portafolio de negocios de Calidad de Datos, al ser una nueva línea de negocio no se dispone de un caso de estudio referencial que utilice el modelo de gestión de Calidad de Datos para la implementación de este concepto, lo cual es vital para asegurar el éxito de este tipo de proyectos.

Las preguntas a ser resueltas en este proyecto son:

- ¿Se ha desarrollado en DWConsulware un caso de estudio aplicando el Modelo de Gestión de Calidad de Datos?

- En el marco de un proyecto específico de Calidad de Datos: ¿Cómo identificar y aplicar las etapas de desarrollo para este tipo de proyectos según el modelo de gestión IBM Data Quality?

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar un caso de estudio aplicando el modelo de Gestión de Calidad de Datos basado en la metodología IBM Data Quality en un proyecto de Business Intelligence de la empresa DWConsulware.

1.3.2 Objetivos Específicos

- Implementar los procesos del Modelo de Gestión de Calidad de Datos en la herramienta IBM Infosphere Information Server para el caso de estudio.

- Determinar el nivel de madurez del caso de estudio mediante la guía de evaluación del Modelo de Madurez de Business Intelligence de TDWI. (TDWI, Maturity Models and Assessments, 2014)

- Aplicar las directrices de implementación de los procesos de mejora y Gestión de Calidad de datos para el caso de estudio.

1.4 Alcance

Este proyecto culmina con la presentación de un caso de estudio aplicando el Modelo de Gestión de Calidad de Datos desarrollado en el proyecto previamente presentado como primera parte de esta tesis.

En este caso de estudio se incluye el desarrollo de los procesos de Gestión de Calidad de Datos mediante el uso de la herramienta IBM Infosphere Information Server, estos son:

- Perfilamiento
- Investigación
- Estandarización
- Coincidencia
- Supervivencia
- Monitoreo

En este trabajo no se incluye ningún otro proceso de Inteligencia de Negocios, tales como: carga de tabla de hecho y dimensiones, verificación de integridad, etc.

CAPÍTULO II

2 INFOSPHERE INFORMATION SERVER

2.1 Introducción a Infosphere Information Server

IBM Infosphere Information Server para la Calidad de los Datos es una suite de aplicaciones que permiten mejorar y monitorear la calidad de los datos. Proporciona altas capacidades de procesamiento que permiten ejecutar las tareas antes mencionadas, convirtiendo datos en información de calidad. A través del análisis, la limpieza, el control y la gestión de calidad de datos, se puede tomar mejores decisiones y optimizar la ejecución de procesos de negocio. Esta herramienta de gestión tiene las siguientes características:

- Automatiza la investigación de origen de datos y estandariza la información.

- Permite aprovechar las capacidades de limpieza de datos tanto en lotes o procesos batch como en tiempo real.

- Enriquece los datos y se asegura de que los mejores datos sobreviven a través de las fuentes.

- Eliminación de duplicados, householding (Consiste en agrupar datos similares de diferentes fuentes, es el acto de identificar un grupo de registros de datos de una fuente, por ejemplo por el orden de ingreso, y cómo estos se relacionan con otros datos ya presentes en el sistema. (Mancuso, 2004), entre otras operaciones.

- Monitorea y mantiene la calidad de los datos.

- Establece indicadores de calidad de datos alineados con los objetivos de negocio que le permiten descubrir rápidamente los problemas de calidad de datos y establecer un plan de remediación.

La **Figura 1** muestra las funciones clave de Infosphere Information Server que se pueden utilizar para implementar una estrategia completa de integración de datos. El núcleo de estas funciones es un repositorio común de metadatos que almacena los metadatos importados, configuraciones, informes y resultados del proyecto para todos los componentes de Infosphere Information Server. Cuando se comparten datos en el repositorio de metadatos, otros usuarios de la organización pueden interactuar con los activos importados y utilizarlos en otros componentes de Infosphere Information Server.

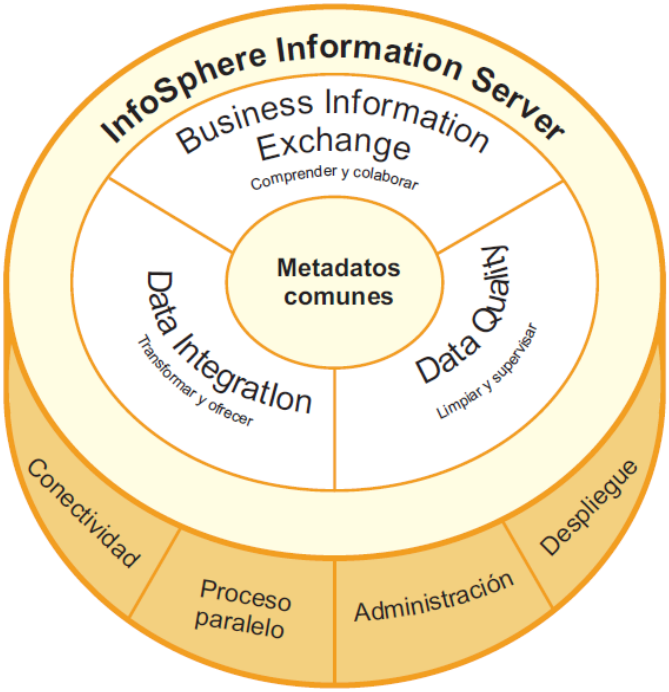


Figura 1: Funciones de Integración de Infosphere Information Server
Fuente: (InfoSphere Information Server Intro, 2012)

La suite de Infosphere Information Server consta de varios componentes, cada uno proporciona distintas funciones para la integración de datos. Juntos, estos componentes forman los bloques de construcción necesarios para

ofrecer información fiable en toda la empresa, independientemente de la complejidad.

Las soluciones de Infosphere Information Server incluyen diferentes componentes que cumplen sus necesidades. Cada solución incluye Infosphere Blueprint Director, Infosphere Discovery e Infosphere Metadata Workbench como componentes base. Los componentes adicionales en cada solución proporcionan diferentes posibilidades que se centran en la calidad de los datos, la integración de datos y la conexión de usuarios funcionales y usuarios de TI.

1. **Infosphere Information Server Business Information Exchange** proporciona funciones que le ayudan a establecer y mantener conocimientos de la empresa a medida que cambia el proyecto.

2. **Infosphere Information Server for Data Integration** proporciona funciones que permiten la transformación y entrega de datos a diferentes aplicaciones.

3. **Infosphere Information Server for Data Quality** proporciona amplias funciones que permiten gestionar la calidad de los datos.

En el **Cuadro 1** se listan los componentes de cada solución de Infosphere Information Server.

Cuadro 1

Componentes de cada solución de Infosphere Information Server

COMPONENTE	BUSINESS INFORMATION EXCHANGE	DATA INTEGRATION	DATA QUALITY
Infosphere Blueprint Director	X	X	X
Infosphere Discovery	X	X	
Infosphere Metadata Workbench	X		
Infosphere Data Architect	X		
Infosphere Business Glossary		X	X
Infosphere Business Glossary Anywhere		X	X
Infosphere Information Analyzer		X	X
Infosphere QualityStage		X	X
Infosphere Information Services Director		X	
Infosphere DataStage & Quality Stage Designer		X	
Infosphere FastTrack		X	

Fuente: (InfoSphere Information Server Intro, 2012)

Cada componente tiene una funcionalidad específica que permite a las empresas obtener soluciones de integración de datos. A continuación, se detallan cada uno de los componentes que forman parte de cada solución:

1. **Infosphere Blueprint Director:** Se utiliza para definir y gestionar planes de un proyecto de integración de datos, desde bosquejos iniciales hasta la entrega.

   Infosphere Blueprint Director proporciona coherencia y comunicación en un proyecto de integración de datos enlazando la visión general de la solución y documentos de diseño detallados. Los miembros del equipo pueden comprender el proyecto a medida que este evoluciona. La creación de un plan completo y bien documentado de la visión de negocio y la visión técnica ayuda al departamento de TI a alinear los requisitos de negocio con la arquitectura de referencia de la empresa.

2. **Infosphere Discovery:** Proporciona técnicas de exploración y análisis de datos innovadores para descubrir automáticamente relaciones y correlaciones entre datos estructurados en la empresa. El análisis se basa en valores reales en los datos, en lugar de sólo metadatos.

   Este análisis de control de valor significa que Infosphere Discovery puede detectar relaciones entre tablas y columnas cuyos nombres o metadatos de por sí no sugieren ninguna conexión. Infosphere Discovery puede identificar y generar transformaciones altamente complejas que se pueden utilizar para describir las ubicaciones y los formatos de datos confidenciales, describir las relaciones de elementos de datos en las aplicaciones o generar salida como código SQL o bien extraer, transformar y cargar código ETL (Extract, Transform, Load) para utilizarlo en trabajos de transformación de datos.

3. **Infosphere Metadata Workbench:** Proporciona informes completos de flujo de datos y análisis de impacto de los activos de la información que utilizan los componentes de Infosphere Information Server. Analistas y especialistas utilizan Infosphere Metadata Workbench para explorar y gestionar los activos de información mediante informes de flujos de datos y análisis de impacto completos que facilitan el descubrimiento y análisis de las relaciones entre activos de información en el repositorio de metadatos de Infosphere Information Server.

4. **Infosphere Data Architect:** Mediante el uso de Infosphere Data Architect, la organización puede diseñar activos de datos, comprenderlos así como sus relaciones y agilizar el proyecto de integración de datos.

Infosphere Data Architect proporciona herramientas para descubrir, modelar, visualizar y relacionar activos de datos heterogéneos. Los usuarios pueden diseñar y desarrollar modelos lógicos, físicos y dimensionales. Las funciones de modelado de datos tradicionales se combinan con funciones de correlación y análisis de modelos exclusivos, todos organizados en una aplicación modular basada en un proyecto.

Infosphere Data Architect descubre la estructura de los orígenes de datos heterogéneos examinando y analizando los metadatos subyacentes. Puede examinar la jerarquía de elementos de datos para comprender las propiedades detalladas y visualizar tablas, vistas y relaciones en un diagrama contextual.

5. **Infosphere Business Glossary:** Es una herramienta interactiva basada en web que permite a los usuarios crear, gestionar y compartir un vocabulario empresarial y un sistema de clasificación. Infosphere Business Glossary Anywhere, su producto complementario, mejora la productividad, incrementando la colaboración y asignando propiedad de datos empresariales a representantes de datos.

Infosphere Business Glossary proporciona un entorno de creación colaborativo que ayuda a los miembros de una empresa a crear una colección central de terminología específica de la empresa, incluyendo las relaciones con activos de información técnica. Dicha colección, denominada glosario empresarial, está diseñada para ayudar a los usuarios a comprender el lenguaje empresarial y el significado empresarial de activos de información como las bases de datos, los trabajos, las tablas y columnas de base de datos y los informes de inteligencia empresarial.

Desde Infosphere Business Glossary, los usuarios designados pueden definir términos, categorías, políticas de control de información y reglas de control de información.

Mediante el uso de Infosphere Business Glossary y Infosphere Business Glossary Anywhere, los usuarios pueden adquirir información de la terminología empresarial común, tener descripciones de datos, propiedad de términos y metadatos y saber cómo se relacionan los términos con activos de información.

6. **Infosphere Information Analyzer:** Proporciona funciones para crear perfiles de datos y analizarlos para ofrecer información fiable a la organización.

Los especialistas en calidad de datos utilizan Infosphere Information Analyzer para explorar ejemplos y volúmenes completos de datos a fin de determinar su calidad y estructura. Este análisis ayuda a descubrir las entradas del proyecto de integración de datos, que van de campos individuales a entidades de datos de alto nivel. El análisis de la información permite a la organización corregir problemas de estructura o validez antes de que afecten al proyecto de integración de datos.

Tras analizar los datos, los especialistas en calidad de datos crean reglas de calidad de datos para evaluar y supervisar orígenes de datos heterogéneos para tendencias, patrones y condiciones de excepción. Estas reglas ayudan a descubrir problemas de calidad de datos y contribuyen a que la organización alinee medidas de calidad de datos en todo el ciclo vital del proyecto. Los analistas empresariales pueden utilizar estas medidas para crear informes de calidad que realicen un seguimiento de los datos a lo largo del tiempo y supervisen su calidad. Los analistas empresariales pueden utilizar IBM Infosphere Data Quality Console para rastrear y examinar excepciones que genera Infosphere Information Analyzer.

7. **Infosphere QualityStage:** Proporciona funciones para crear y mantener una vista precisa de las entidades de datos tales como cliente, ubicación, proveedores y productores en toda la empresa.

   Infosphere QualityStage utiliza reglas tanto predefinidas como personalizables para preparar información compleja sobre las entidades empresariales para aplicaciones transaccionales, operativas y analíticas en proceso por lotes, en tiempo real o como un servicio web. La información se extrae del sistema de origen, se mide, limpia, enriquece, consolida y carga en el sistema de destino.

   Al completar el análisis en nivel de carácter o de palabra, Infosphere QualityStage ayuda a descubrir anomalías de datos e incoherencias antes de que se produzca el proceso de transformación. Los datos provenientes de orígenes diversos se estandarizan automáticamente en campos fijos, tales como un determinado nombre, la fecha de nacimiento, el género y el número de teléfono. Utiliza las reglas de calidad de datos y, a continuación, asigna el significado semántico correcto a los datos de entrada para facilitar la comparación.

   Al garantizar la calidad de los datos, Infosphere QualityStage reduce el tiempo y el coste necesarios para implementar la gestión de datos maestros (MDM), la inteligencia empresarial, la planificación de recursos empresariales (ERP) y otras iniciativas de TI relacionadas con el cliente.

8. **Infosphere Information Services Director:** Proporciona un entorno integrado que permite a los usuarios desplegar rápidamente la lógica de Infosphere Information Server como servicios.

   La infraestructura de la arquitectura orientada a servicios (SOA) de Infosphere Information Services Director asegura que la lógica de

integración de datos desarrollada en Infosphere Information Server se pueda utilizar en cualquier proceso empresarial. Los datos más adecuados están disponibles todo el tiempo, para todas las personas y para todos los procesos.

Infosphere Information Services Director define servicios como objetos empresariales verdaderos, que se despliegan localmente en el servidor de aplicaciones. Esta integración oculta por completo la complejidad de la implementación del servicio que consume lógica de integración de datos publicados.

9. **Infosphere DataStage:** Es una herramienta de integración de datos que permite a los usuarios mover y transformar datos entre sistemas de transacciones y analíticos. La transformación y el movimiento de datos es el proceso mediante el cual se seleccionan, convierten y correlacionan datos de origen en el formato que requieren los sistemas de destino. El proceso manipula datos para que sean conformes con las reglas de negocio, de dominio y de integridad y con otros datos en el entorno de destino.

Infosphere DataStage proporciona conectividad directa a aplicaciones empresariales como orígenes o destinos, garantizando que los datos más relevantes, completos y precisos se integren en el proyecto de integración de datos.

10. **Infosphere FastTrack:** Proporciona funciones para automatizar el flujo de trabajo del proyecto de integración de datos. Los usuarios pueden hacer un seguimiento de las tareas de integración de datos múltiples y automatizarlas entre desarrollar requisitos empresariales e implementar una solución.

Los analistas empresariales utilizan Infosphere FastTrack para convertir requisitos empresariales en un conjunto de especificaciones,

que los especialistas en integración de datos utilizan a continuación para producir una aplicación de integración de datos que incorpora los requisitos empresariales.

Con Infosphere FastTrack los analistas de datos crean especificaciones de correlación que convierten los requisitos de la empresa en aplicaciones. Los especialistas en integración de datos utilizan Infosphere DataStage and QualityStage para desarrollar procesos que extraen, transforman y comprueban la calidad de los datos basados en las especificaciones generadas por los analistas. Adicionalmente se puede utilizar Infosphere Information Services Director para desplegar procesos como servicios de información coherentes y reutilizables (Servicios Web), esta tarea debe ser realizada por un arquitecto de SOA (Service-Oriented Arquitecure)

2.2 Infosphere Information Server for Data Quality

Infosphere Information Server for Data Quality es la suite que va a ser utilizada en el proyecto de Caso de Estudio, la suite proporciona funcionalidades que permiten limpiar los datos y controlar la calidad de datos, convirtiendo los datos en información confiable. El **Cuadro 2** lista los productos que contiene la solución:

Cuadro 2

Componentes de Infosphere Information Server for Data Quality

COMPONENTE	INFOSPHERE INFORMATION SERVER FOR DATA QUALITY
Infosphere Blueprint Director	**X**
Infosphere Business Glossary	**X**
Infosphere Business Glossary Anywhere	**X**
Infosphere Information Analyzer	**X**
Infosphere QualityStage	**X**

Fuente: (InfoSphere Information Server Intro, 2012)

La Figura 2 indica cómo los componentes de Infosphere Information Server funcionan para crear una solución de integración de datos unificada.

Un producto base de metadatos común permite que distintos tipos de usuarios creen y gestionen metadatos utilizando herramientas que están optimizadas para sus roles, lo cual facilita la colaboración entre estos.

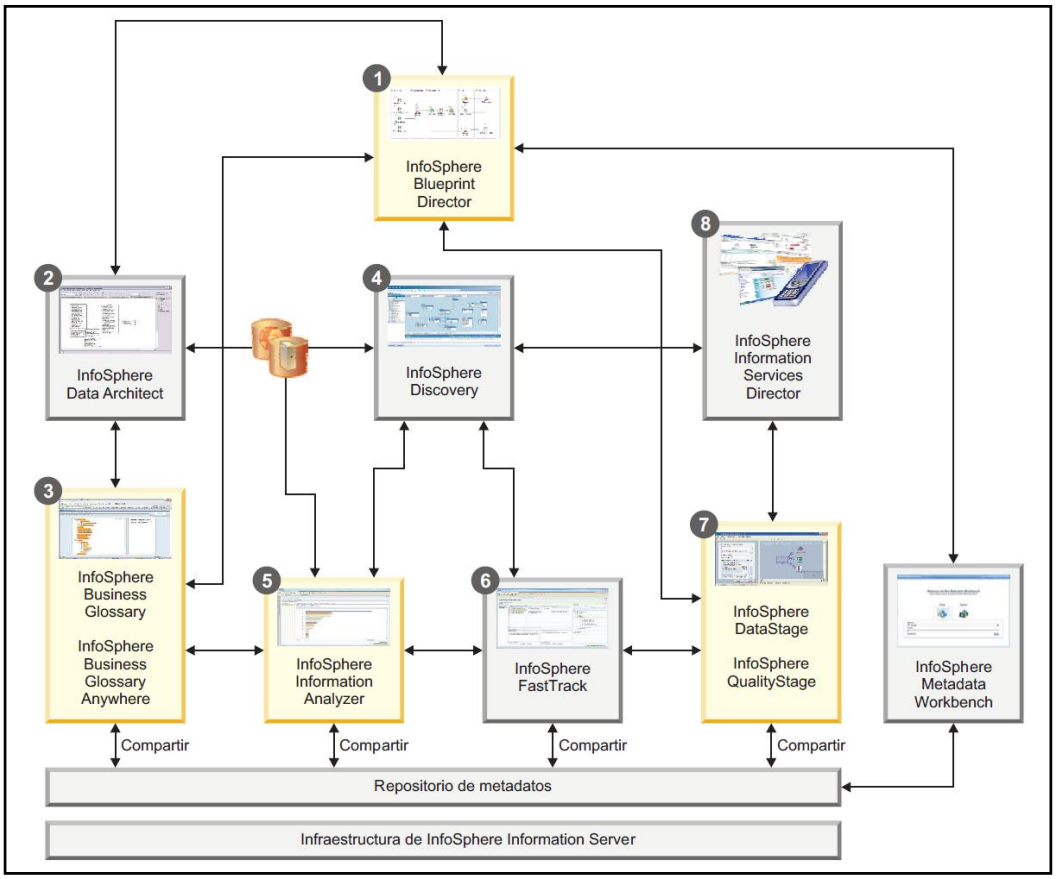


Figura 2: Componentes de Infosphere Information Server
Fuente: (InfoSphere Information Server Intro, 2012)

El primer producto, Infosphere Blueprint Director (1), es la herramienta que los arquitectos de información utilizan para **planificar** y gestionar los proyectos de integración de datos, mediante planes el equipo de trabajo puede conectar la visión empresarial con el proyecto. Una vez generado el plan de proyecto se utiliza Infosphere Data Architect (2) para **diseñar** la estructura de los datos, establecer relaciones y crear modelos físicos y lógicos de los activos de datos de la organización. Los analistas de datos definen y establecen un criterio común de los conceptos de negocio tomando el plan de proyecto y de los modelos de datos del negocio mediante Infosphere Business Glossary (3) con el objetivo de **mantener** un mismo lenguaje para todos los participantes del proyecto, es decir generar un glosario empresarial para

desarrollar y compartir el mismo vocabulario común entre los usuarios de la empresa y de TI.

A través del repositorio de metadatos compartido Infosphere Information Server permite a los analistas de datos **descubrir** las estructuras de los datos y analizar el significado de sus relaciones con Infosphere Discovery (4). La información obtenida sirve de insumo para Infosphere Information Analyzer (5) e Infosphere FastTrack (6).

Con Infosphere Information Analyzer (5) los especialistas de calidad de datos diseñan, **desarrollan** y gestionan las reglas de calidad de datos de la organización, al ser un proceso cíclico las reglas pueden cambiar porque los datos evolucionan constantemente, para mantener actualizado el proyecto esta información es utilizada por Infosphere Business Glossary (3), Infosphere FastTrack (6), Infosphere DataStage and QualityStage (7) y otros componentes de Infosphere Information Server.

Para importar información técnica al repositorio de metadatos como, por ejemplo, informes BI, modelos lógicos, esquemas físicos y procesos de Infosphere DataStage and QualityStage (7) se utiliza la herramienta Infosphere Metadata Asset Manager.

2.3 Arquitectura de la Solución

IBM Infosphere Information Server proporciona una arquitectura unificada que funciona con todos los tipos de integración de información. Los servicios compartidos, los procesos paralelos unificados y los metadatos unificados son la base de la arquitectura del servidor, la cual está orientada a los servicios, lo que permite utilizar IBM Infosphere Information Server en las arquitecturas en evolución como las orientadas a servicios.

La Figura 3 muestra la arquitectura de la solución que se implementará en el proyecto:

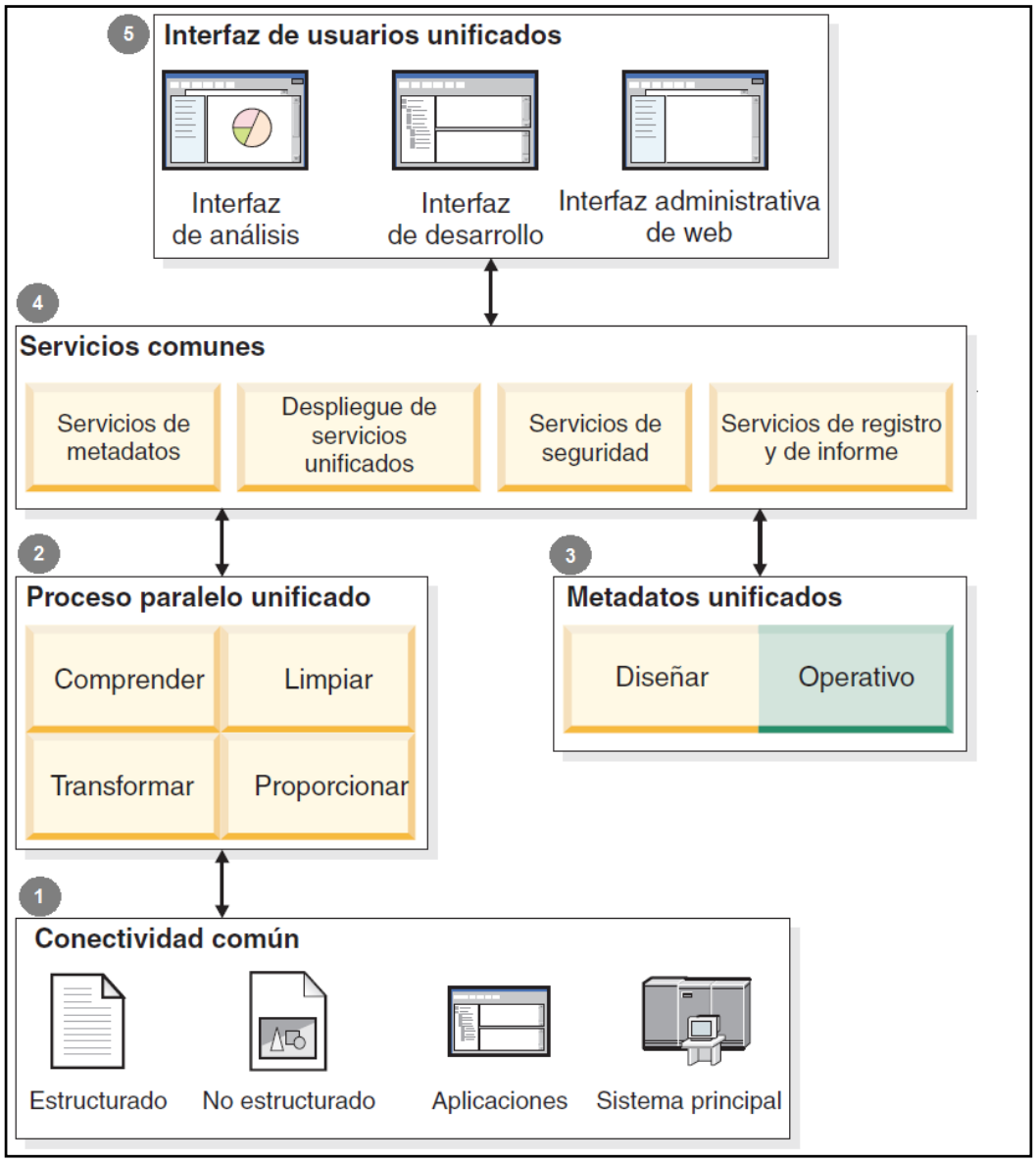


Figura 3: Arquitectura de la solución
Fuente: (InfoSphere Information Server Intro, 2012)

1) **Conectividad común:** Infosphere Information Server puede conectarse con fuentes de información estructurada, no estructurada o del sistema principal, o con aplicaciones. La conectividad es controlada por metadatos que se comparten entre los componentes de la suite y los objetos de conexión pueden volver a utilizare en las distintas funciones.
2) **Motor de procesos paralelos unificados:** La mayor parte del trabajo que realiza Infosphere Information Server tiene lugar en el motor de

procesos paralelos. El motor realiza tareas como el análisis de bases de datos de gran tamaño con Infosphere Information Analyzer, la limpieza de datos con Infosphere QualityStage y transformaciones complejas con Infosphere DataStage. El motor está diseñado para brindar ventajas como:

- **Paralelismo e interconexión de datos.** Permite disminuir el tiempo de transferencia de grandes volúmenes de datos.

- **Escalabilidad**. Permite adición de hardware sin realizar cambios en el diseño de la solución.

- **Procesos optimizados de bases de datos, archivos y colas**. Permite gestionar archivos de gran tamaño que no caben en la memoria todos al mismo tiempo o gestionar un elevado número de archivos pequeños.

3) **Metadatos unificados**: La base de Infosphere Information Server es una infraestructura de metadatos unificados que permite el uso compartido de dominios empresariales y dominios técnicos. Esta infraestructura reduce el tiempo de desarrollo y proporciona un registro permanente que puede mejorar la confianza en la información. Todas las herramientas de Infosphere Information Server comparten el mismo metamodelo, lo que facilita la colaboración de distintos roles y funciones. El repositorio contiene dos tipos de metadatos:

- **Dinámicos**. Los metadatos dinámicos incluyen información de diseño.

- **Operativos**. Los metadatos operativos incluyen datos de supervisión del rendimiento, auditoría y registro, y datos de muestro de creación de perfiles de datos.

4) **Servicios comunes:** Infosphere Information Server está basado en un conjunto de servicios compartidos que centralizan tareas como la seguridad, la administración de usuarios, el registro cronológico y la generación de informes. Estas tareas se administran desde un solo lugar, independientemente del componente de la suite que se utilice. Los servicios compartidos también incluyen servicios de metadatos, que proporcionan acceso estándar orientado a servicios y análisis de metadatos en toda la plataforma.

5) **Interfaz de usuario unificada:** Infosphere Information Server se presenta como una infraestructura de herramientas y una interfaz gráfica común. Las interfaces compartidas ofrecen homogeneidad, controles visuales y una experiencia del usuario similar en los distintos productos.

CAPÍTULO III

3 IMPLEMENTACIÓN DEL MODELO DE GESTIÓN

3.1 Introducción

La empresa que hemos seleccionado para aplicar el modelo de gestión se dedica a la producción de alimentos de consumo masivo y se va a denominar Empresa CS. Los directivos de la misma requieren de información confiable para tomar decisiones que permitan mejorar el desempeño de los procesos de negocio, es por esto que se justifica un proyecto de Gestión de Calidad de Datos, cuyo objetivo es ayudar a la organización a mantenerse siempre un paso delante de sus competidores.

3.1.1 Situación Actual

La empresa actualmente posee bases de datos centralizadas con la información de las diversas fuentes de datos, entre ellas están un Sistema de Maestros, un Sistema de Transacciones y un Data Warehouse. El proceso de ETL se lo realiza mediante la herramienta de Infosphere Datastage, este proceso consiste en tomar los datos de los sistemas fuente y llevarlos a los repositorios centralizados de la empresa aplicando una lógica de negocio. Esta información luego es consumida a través de herramientas para la toma de decisiones. La Figura 4 describe la arquitectura actual de la organización.

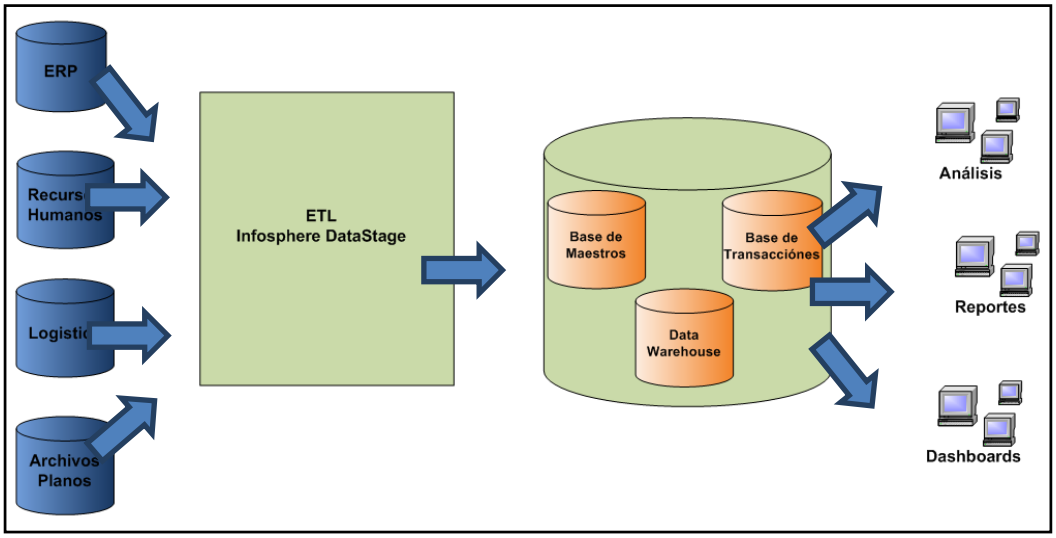


Figura 4: Arquitectura del Caso de estudio

Para este caso de estudio nos vamos a enfocar en el Sistema Central de Maestros el cual es un conjunto de procesos que cargan información desde los sistemas fuente de la empresa inicialmente hacia un área de paso temporal *(staging)* en donde se realizan transformaciones simples, luego se aplica una lógica de negocio para finalmente guardar la información en una base de datos centralizada.

Debido a que la información debe ser libre de defectos, es necesario incluir procesos de calidad de datos que permitan obtener datos confiables y precisos para las aplicaciones de la empresa.

3.1.2 Alcance del Proyecto

En función del escenario planteado se propone un proyecto de Calidad de Datos con la herramienta Infosphere information Server for Data Quality, el cual va a permitir identificar el estado actual de los datos y realizar mejoras en los casos que sean necesarios.

El proyecto se aplicará para el Sistema Central de Maestros específicamente en el catálogo de Clientes, el cual es ampliamente utilizado por la mayoría de aplicaciones de los distintos modelos de negocio de la organización.

La Figura 5 describe la arquitectura del caso de estudio, modificada para incluir el proyecto de calidad de datos.

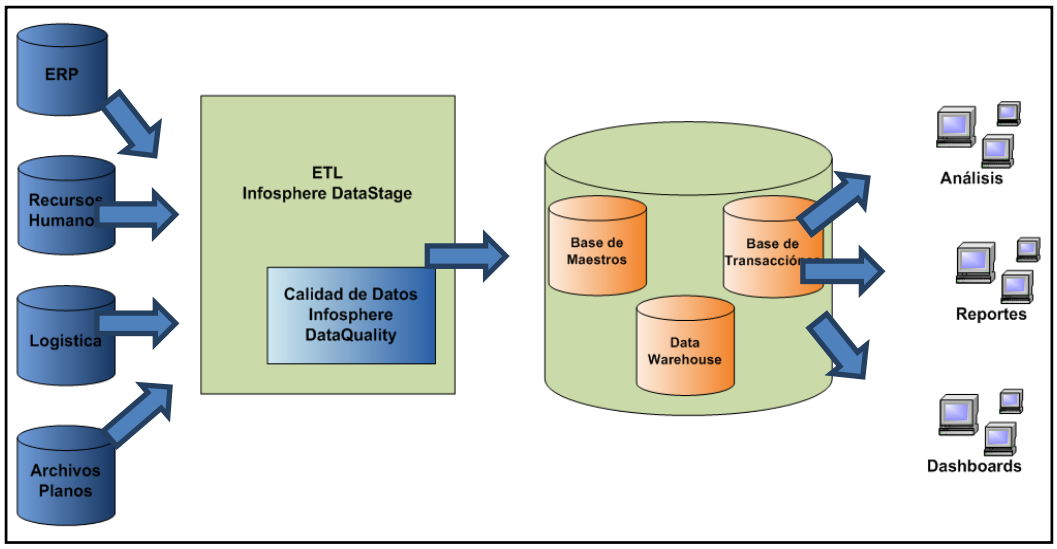


Figura 5: Arquitectura General del Caso de Estudio + Calidad de Datos

Para el caso de estudio tomaremos como fuente de datos principal el ERP como punto de partida del análisis, esta fuente se caracteriza por que maneja un concepto de multiempresa lo cual se debe tomar en cuenta en el momento de realizar el estudio de los datos.

Uno de los principales catálogos de la organización es el Maestro de Clientes, ya que este es consumido por la mayoría de aplicaciones, como se indicó anteriormente es el catálogo que será analizado en el caso de estudio. La Figura 5 describe la arquitectura señalando específicamente el Maestro de Clientes.

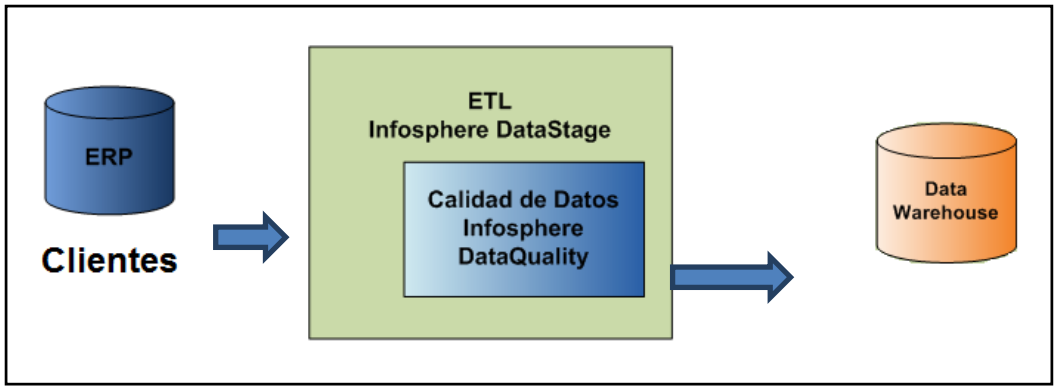


Figura 6: Arquitectura de Clientes del Caso de Estudio + Calidad de Datos

La información de clientes está separada en función de los datos del cliente y de las direcciones en dos tablas del Sistema ERP, para poder agrupar está información se debe combinar las tablas de la base de datos. Las tablas a analizar son las siguientes:

- TTCCOM100 – Clientes
  - Código de Cliente
  - Cédula
  - Nombre o Razón Social
  - Nombre Corto

- TTCCOM130 – Direcciones
  - Código de Cliente
  - Teléfono
  - Calle Principal
  - Código Postal
  - Provincia
  - Fax

3.2 Desarrollo del Proyecto

Una vez que se ha definido la situación actual de la empresa y se ha determinado tanto la arquitectura actual de BI de la organización como la arquitectura combinada con la Gestión de Calidad de Datos es posible iniciar el desarrollo del proyecto, este se compone dos evaluaciones del nivel de Madurez de BI utilizando el Modelo de Madurez TDWI y la etapa de Gestión de Calidad de Datos propiamente, las evaluaciones se realizan antes y después de aplicar el Modelo de Gestión de Calidad de Datos y permiten determinar el impacto de la Gestión de Calidad de Datos en la organización.

3.2.1 Evaluación Inicial - Modelo de Madurez TDWI

Para evaluar el Modelo de Madurez de acuerdo a los parámetros del TDWI se utiliza la encuesta proporcionada por el mismo instituto, el cuestionario completo fue detallado en la primera parte de este trabajo y en esta etapa se han contestado las preguntas de acuerdo a las características de la empresa donde se está implementado el caso de estudio, los resultados de la encuesta se presentan a continuación en la **Figura 7**.

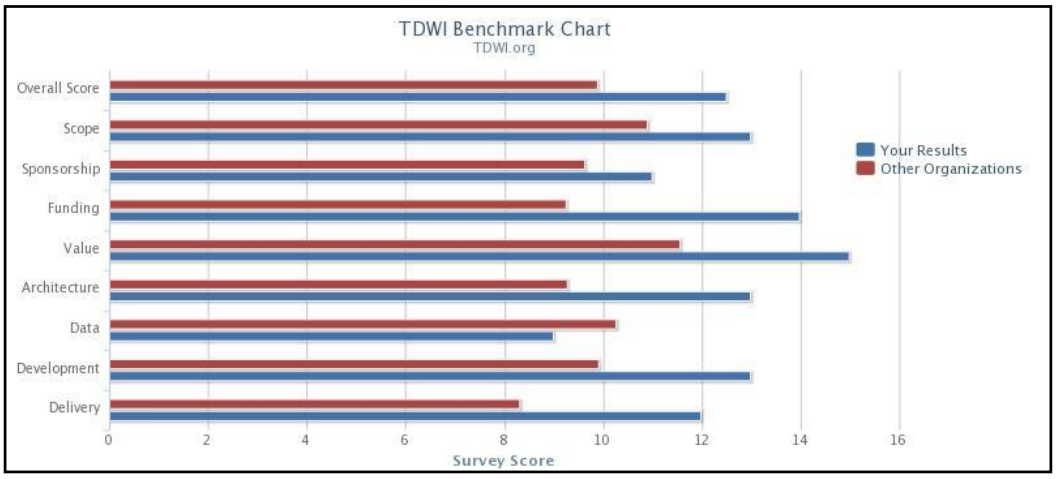


Figura 7: Resultados Iniciales de la Evaluación TDWI
Fuente: (TDWI, Evaluación del Modelo de Madurez de BI de TDWI, 2012)

Para cada una de las categorías obtenemos los siguientes resultados:

- Ámbito (scope): 13
- Patrocinio (sponsorship): 11
- Presupuesto (funding): 14
- Valor (value): 15
- Aquitectura (architecture): 13
- Datos (data): 9
- Desarrollo (development): 13
- Entrega (delivery): 13

3.2.1.1 Análisis de los Resultados

De acuerdo a las directrices del TDWI (TDWI, Evaluación del Modelo de Madurez de BI de TDWI, 2012) el puntaje obtenido en cada categoría se ubica en un rango, el cual a su vez nos indica la fase dentro del modelo de madurez TDWI donde se encuentra la organización, con estos parámetros podemos generar la siguiente tabla de resumen:

Tabla 1

Resultados de la Evaluación Inicial TDWI

CATEGORÍA	PUNTAJE OBTENIDO	NIVEL DE MADUREZ / PUNTO CRÍTICO
Ámbito	13	Repetible
Patrocinio	11	Preliminar
Presupuesto	14	Repetible
Valor	15	Repetible / Abismo
Arquitectura	13	Repetible
Datos	9	Preliminar / Golfo
Desarrollo	13	Repetible
Entrega	13	Repetible

La mayoría de las categorías nos ubican en el nivel Repetible, o “Adolescencia”, del modelo TDWI:

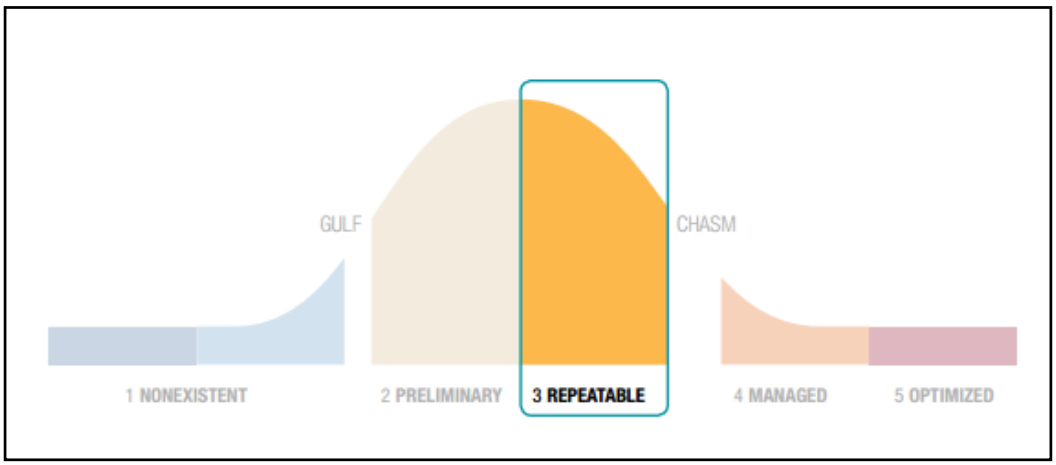


Figura 8: Nivel de Madurez de la Evaluación TDWI
Fuente: (TDWI, Evaluación del Modelo de Madurez de BI de TDWI, 2012)

Además, la categoría Valor se encuentra en el punto crítico del Abismo y la de Datos en el punto crítico del Golfo.

De acuerdo a los resultados entregados por la evaluación, podemos determinar los siguientes aspectos:

1. Las unidades de negocio de la organización reconocen el valor de integrar sus fuentes de datos en único repositorio (data warehouse).

2. Los proyectos de BI están pasando de ser proyectos ad hoc a convertirse en parte de un solo gran proyecto de BI, alineado con los objetivos estratégicos de la organización.

3. Las categorías que se encuentran en los puntos críticos requerirán un mayor trabajo para pasar al siguiente nivel, estas son:

   - Valor
   - Datos.

3.2.2 Modelo de Gestión de Calidad de Datos

Para el caso de estudio se van a aplicar las fases de la gestión de calidad de datos[1] con la ayuda de la suite de herramientas Infosphere Information Server for Data Quality.

El primer paso del proyecto es planificar el flujo de información y los activos de información que se van a requerir, para ejecutar este paso es necesario crear un plan *(blueprint)* del proyecto usando IBM Infosphere Blueprint Director.

[1] Las Fases de la Gestión de Calidad de datos se detallan en el primer proyecto de este trabajo.

Para entender la estructura de los sistemas fuente e identificar cualquier problema de calidad de datos se realizará el perfilamiento de la información a través de IBM Infosphere Information Analyzer.

Para investigar las fuentes de datos, conocer el grado de procesamiento que se necesita para obtener datos precisos, identificar errores en los datos y validar el contenido de cada campo se realizaran procesos en la herramienta IBM Infosphere DataStage and Quality Stage Designer.

Con la evaluación inicial obtenida se aplicaran técnicas de estandarización para asegurar la consistencia de los datos. La estandarización va a permitir usar técnicas de coincidencia o *matching* para agrupar entidades similares y eliminar los duplicados. Estos procesos también se desarrollarán con IBM Infosphere DataStage and Quality Stage Designer.

Se aplicarán Reglas de supervivencia en la información agrupada de las fuentes para consolidar en un solo registro la mejor información de las fuentes, todo esto a través de IBM Infosphere DataStage an Quality Stage Designer.

Con los datos limpios se requiere realizar el monitoreo de los datos a través de IBM Information Analyzer.

Finalmente se aplicarán reglas de validación contra los nuevos datos para verificar qué registros no cumplen con los parámetros de calidad usando IBM QualityStage Data Rules Stage.

3.2.2.1 Planificación del Proyecto

Utilizando IBM Infosphere Blueprint Director se obtiene la documentación técnica del proyecto, esto consiste en generar un diseño de la arquitectura del proyecto para ser usada como guía para el desarrollo de las etapas posteriores.

El diseño de la arquitectura facilita la comunicación entre los miembros del proyecto la siguiente figura indica la arquitectura de alto nivel para el proyecto de caso de estudio.

A continuación se presenta el detalle del modelo realizado para el caso de estudio, en la **Figura 9** se presenta el esquema general del todo el proyecto de inteligencia de negocios.

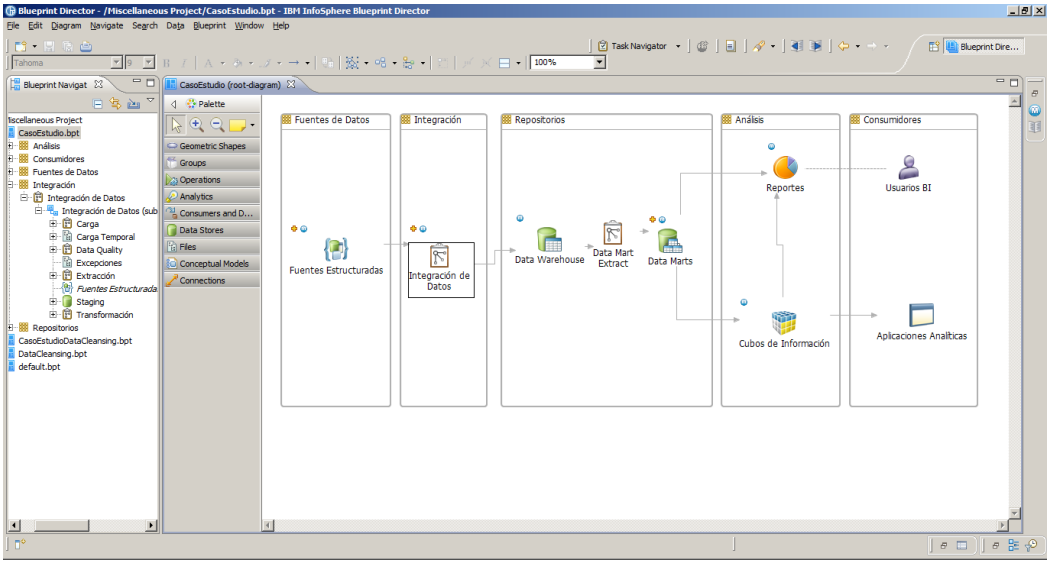


Figura 9: Modelo de Caso de Estudio

Una parte del proyecto y que concierne al caso de estudio es la Integración de Datos que tiene como característica principal el tema de Calidad de Datos. La **Figura 10** indica el proceso de la Integración de Datos que consiste en carga los datos desde la fuente hasta un repositorio temporal llamado Staging.

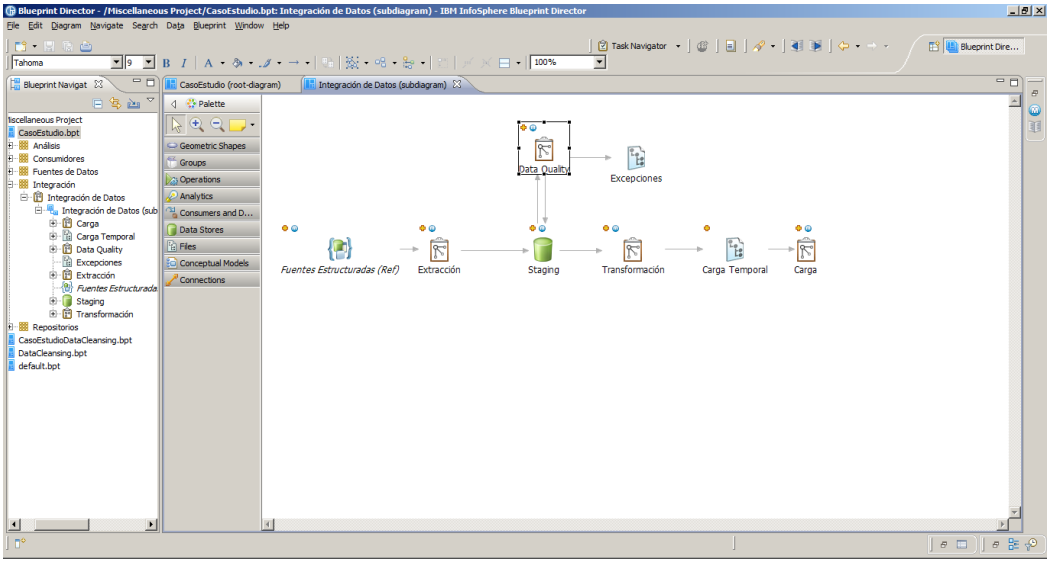


Figura 10: Modelo Caso de Estudio – Carga a Stage

La calidad de datos es un proceso que toma datos de Staging y los guarda en el mismo repositorio pero depurados.

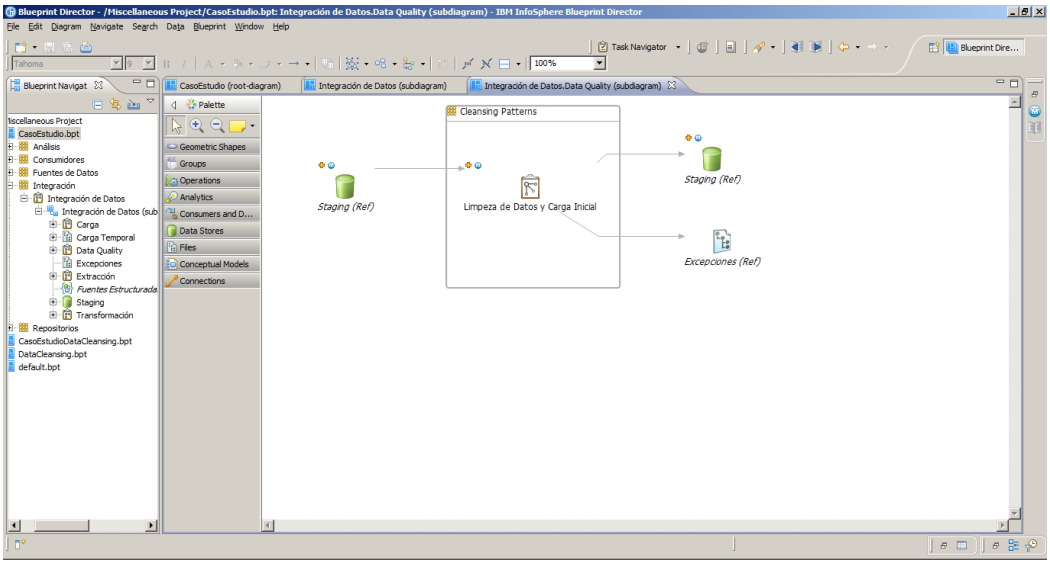


Figura 11: Modelo Caso de Estudio – Transformación

El proceso de calidad de datos se describe en la **Figura 12** con todas sus etapas las cuales van a ser desarrolladas en este proyecto de tesis.

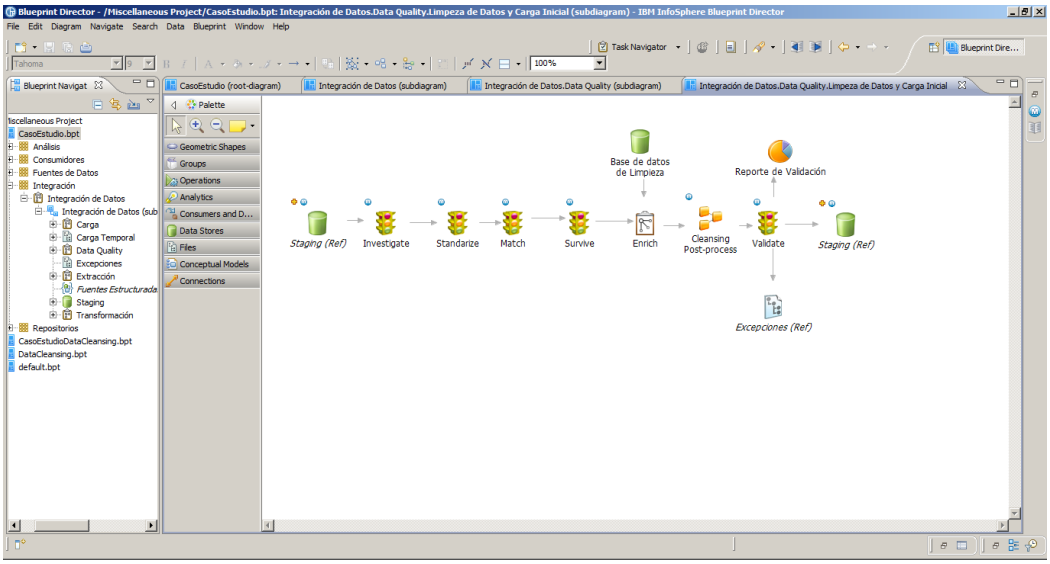


Figura 12: Modelo Caso de Estudio – Calidad de Datos

El modelo creado en la herramienta puede ser publicado para que sea compartido por los miembros del equipo. La **Figura 13** y la **Figura 14** detallan la configuración que se debe realizar para la publicación.

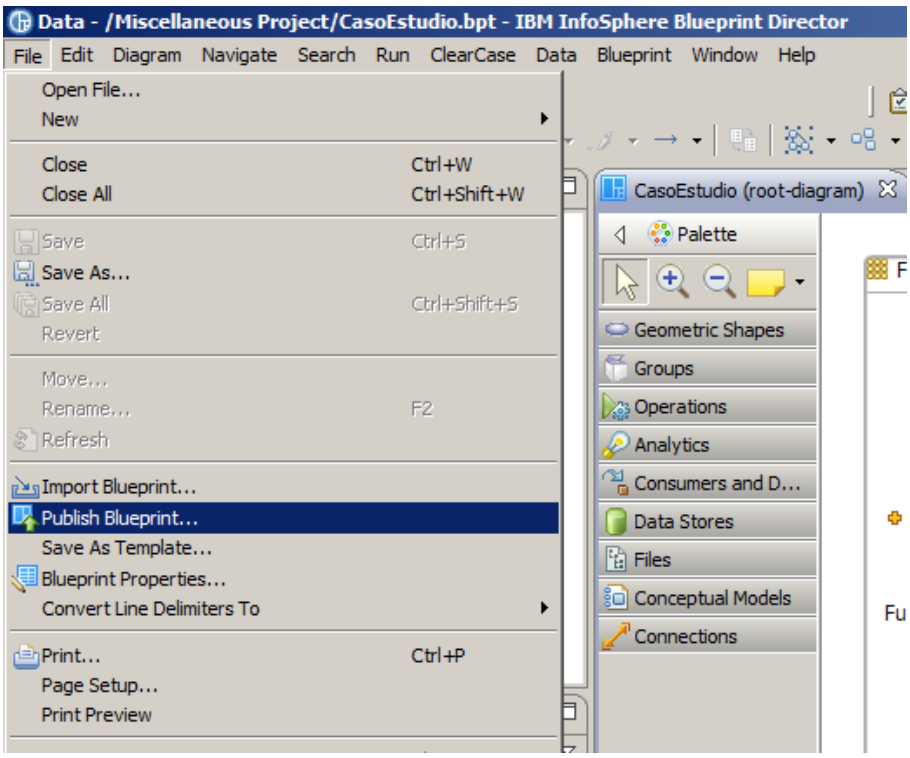


Figura 13: Publicación de Modelo Caso de Estudio – 1

Establezca el nombre y seleccione la conexión.

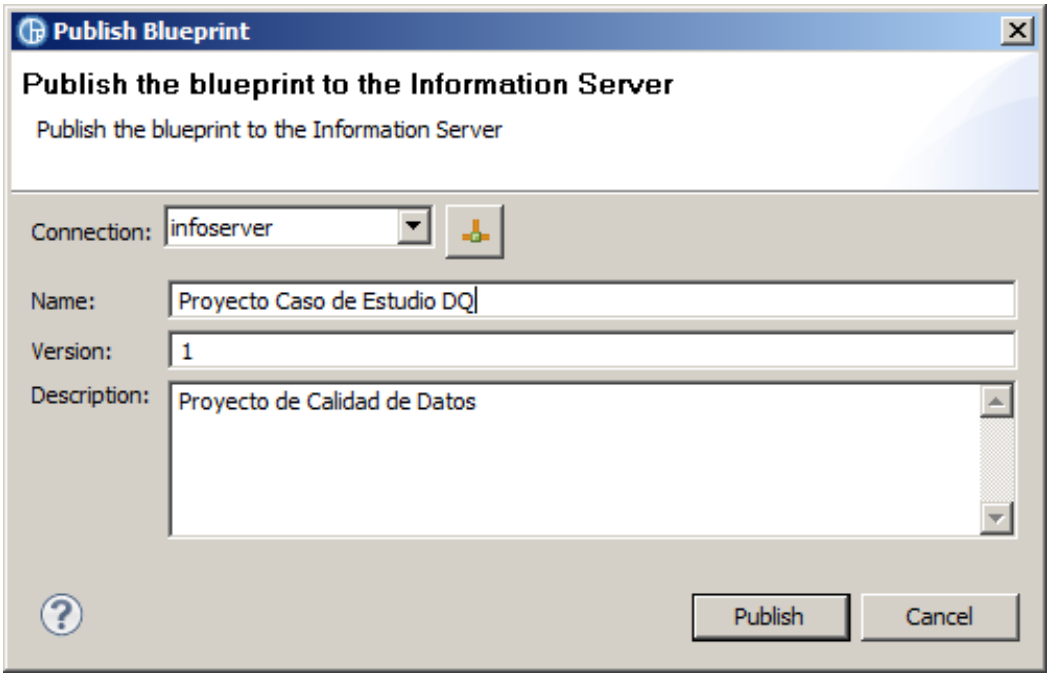


Figura 14: Publicación de Modelo Caso de Estudio – 2

La publicación puede ser revisada en la herramienta IBM Business Glossary como se describe a continuación.

3.2.2.2 Glosario de Términos

Dentro de una organización existe información y terminología que cambia constantemente, para llevar un control y difusión a los miembros del equipo existe la solución IBM Business Glossary, la cual permite tener actualizadas las últimas definiciones del negocio.

Para hacer búsquedas rápidas y navegar por el glosario de términos existe la opción con la herramienta IBM Business Glossary Anywhere. Por ejemplo realizamos la búsqueda como indica la **Figura 15**.

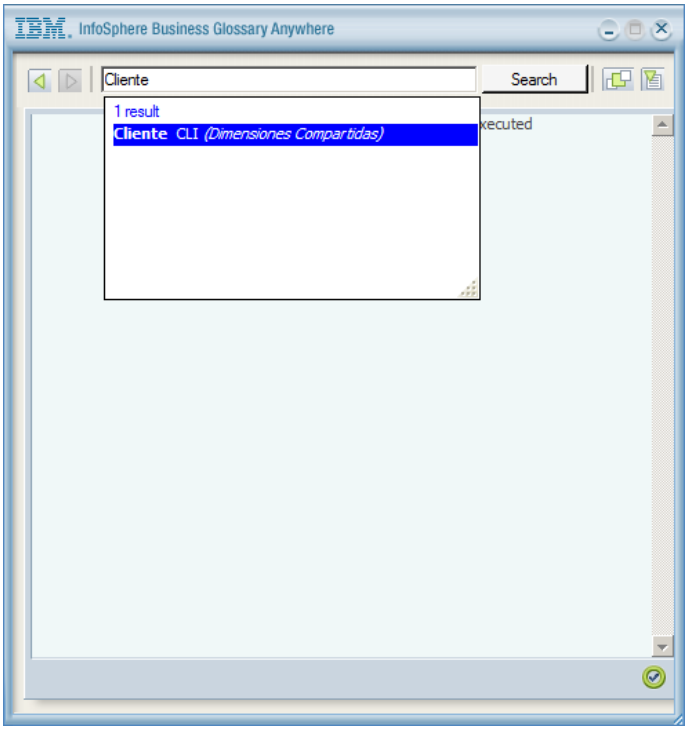


Figura 15: IBM Business Glossary Anywhere – 1

El resultado de la búsqueda se presenta en la **Figura 16**, donde se puede observar el detalle de las propiedades de la entidad Cliente.

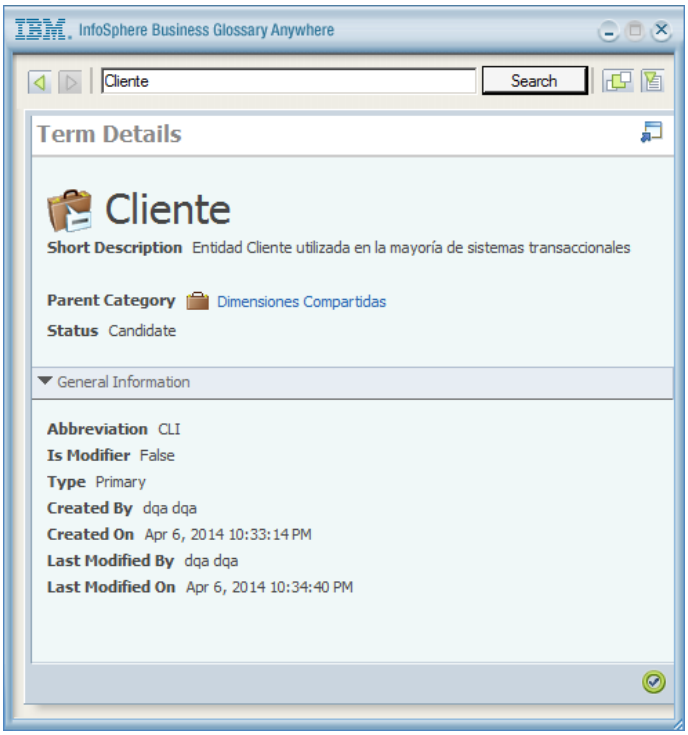


Figura 16: IBM Business Glossary Anywhere – 2

La herramienta permite establecer una estructura jerárquica que se define en términos y categorías siendo términos el nivel más bajo, si navegamos la padre del cliente son las dimensiones compartidas del negocio.

Figura 17: IBM Business Glossary Anywhere – 3

Para incluir o editar términos y categorías accedemos al portal de la herramienta el cual presenta la información de manera similar a IBM Business Glossary Anywhere.

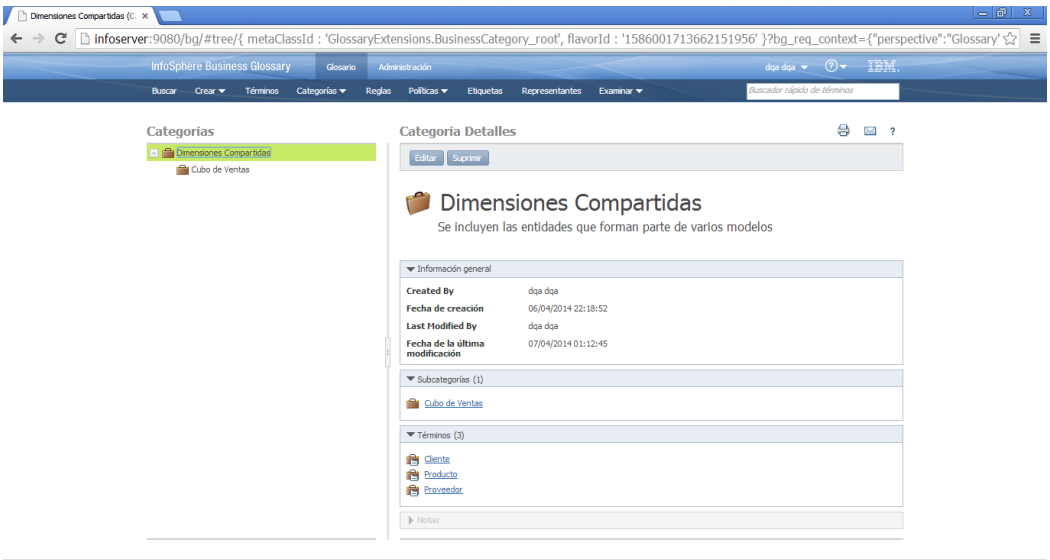


Figura 18: IBM Business Glossary Anywhere – 4

El modelo realizado en IBM Infosphere Blueprint Director fue publicado para ser compartido en la red de la organización. A continuación se describe cómo utilizar la herramienta para navegar por el modelo. La **Figura 19** presenta el modelo publicado.

Figura 19: IBM Business Glossary Anywhere – 5

Accedemos al modelo y se presenta la siguiente interfaz de validación, la cual presenta la descripción y las propiedades del modelo.

Figura 20: IBM Business Glossary Anywhere – 6

En la **Figura 21** se presenta el modelo con la funcionalidad de navegación.

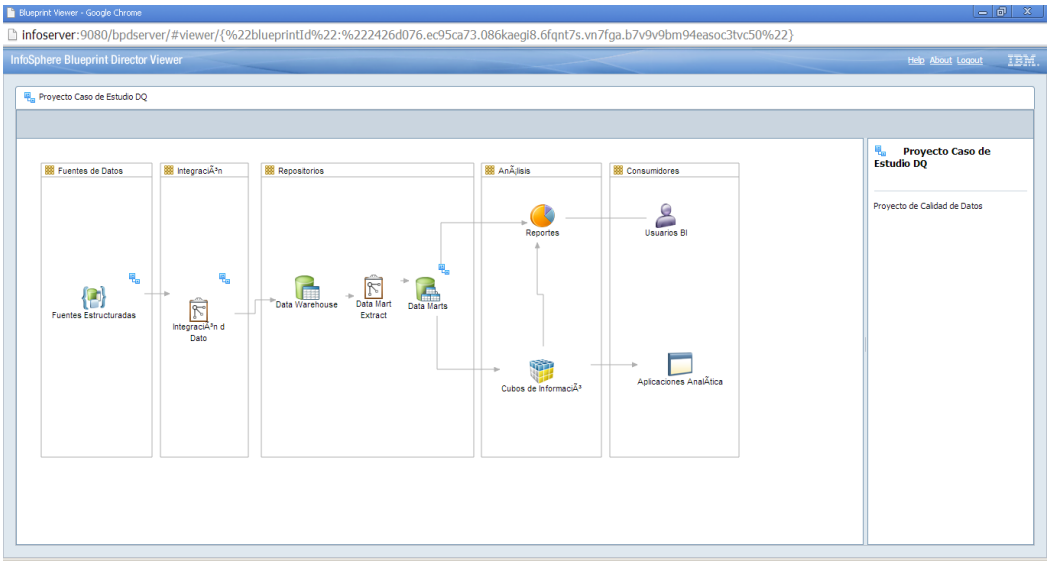


Figura 21: IBM Business Glossary Anywhere – 7

3.2.2.3 Perfilamiento de Datos

En la etapa de Investigación del caso de estudio se requiere entender la condición actual de los datos así como el contenido y la estructura de los mismos, para esto se debe evaluar los siguientes aspectos:

- Validez de los tipos de datos.

- Consistencia de formatos.
- Estructura y contenido

La herramienta IBM Information Analyzer permite realizar este análisis inicial, para lo cual se ejecutan las siguientes tareas:

1. **Importar metadata a Information Server Repository e Information Analyzer para poder analizar los datos**

   Utilizando la herramienta Metadata Asset Manager e ingresando con un usuario con permisos para importar metadatos crear una nueva área de importación, esta operación se detalla en el ANEXO A.

2. **Ejecutar el análisis de columnas para identificar posibles errores con los datos**

   Ingresar con un usuario del tipo “Analista” a Information Analyzer y crear un nuevo proyecto:

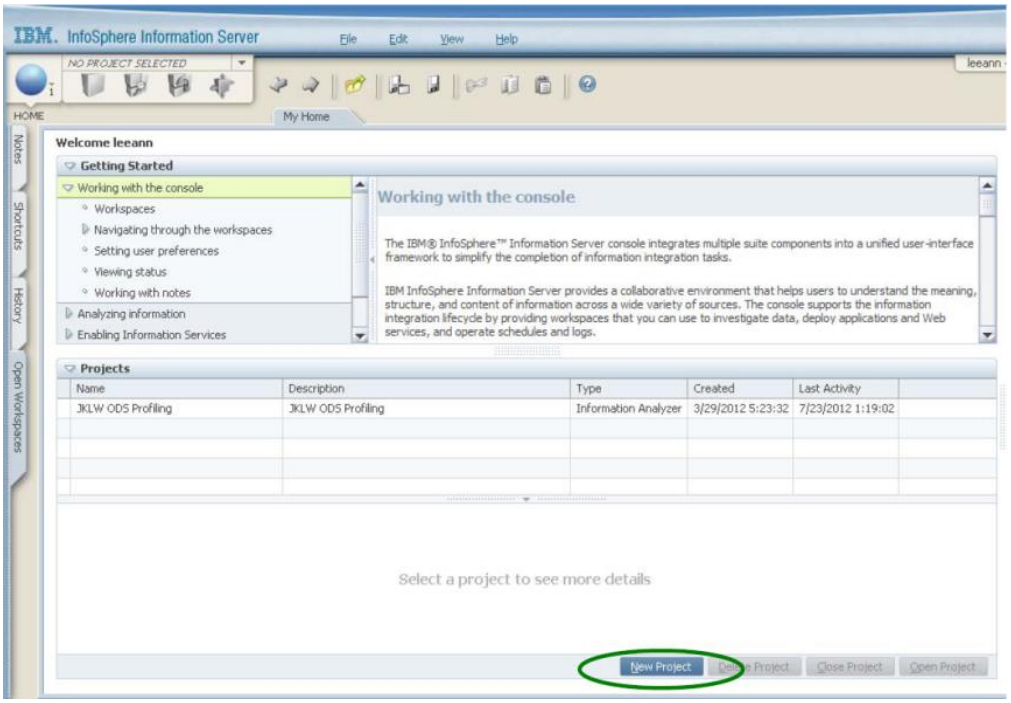


Figura 22: IBM Infosphere Information Analyzer – 1

El proyecto se llama DQ Caso de Estudio como indica la **Figura 23**:

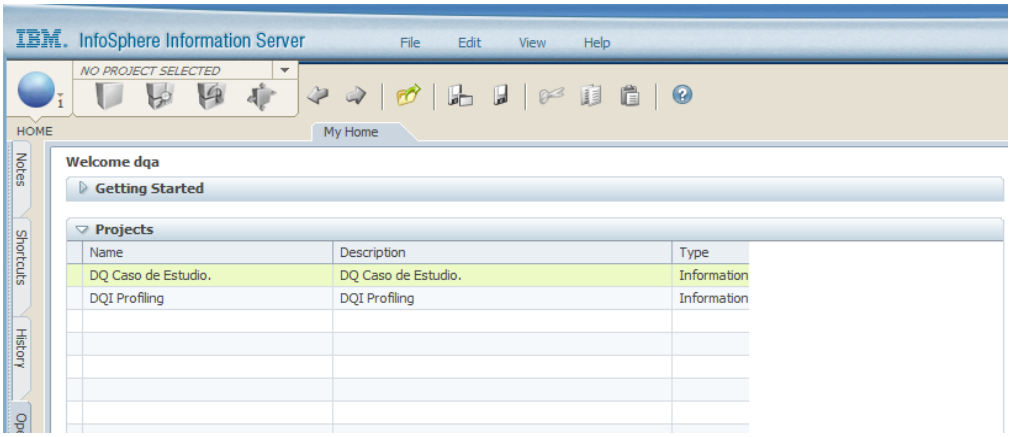


Figura 23: IBM Infosphere Information Analyzer – 2

Seleccionar las fuentes de datos que se van a analizar:

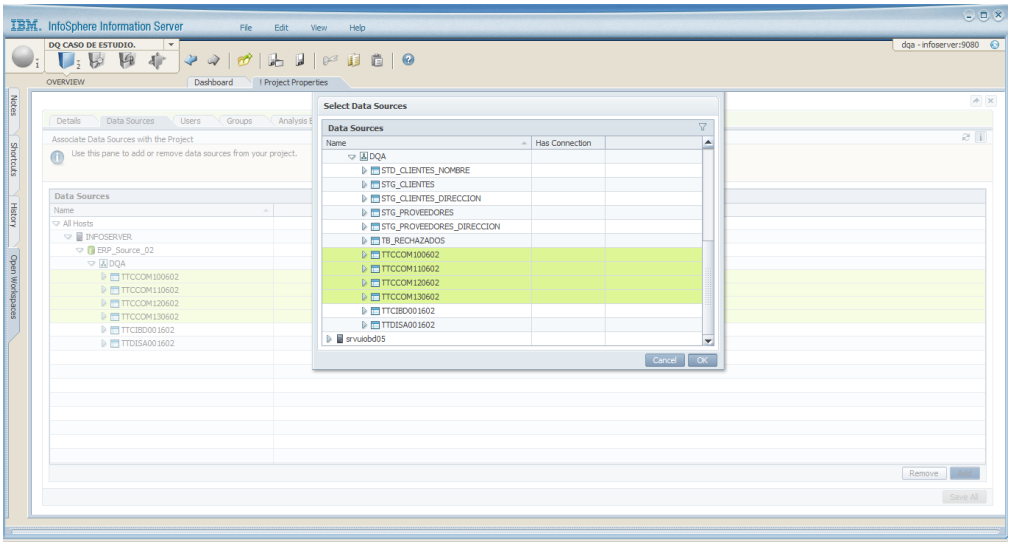


Figura 24: IBM Infosphere Information Analyzer – 3

Seleccionar la opción Column Analysis para empezar con el análisis de los datos, seleccionar la tabla que se desea analizar y presionar Run Column Analysis.

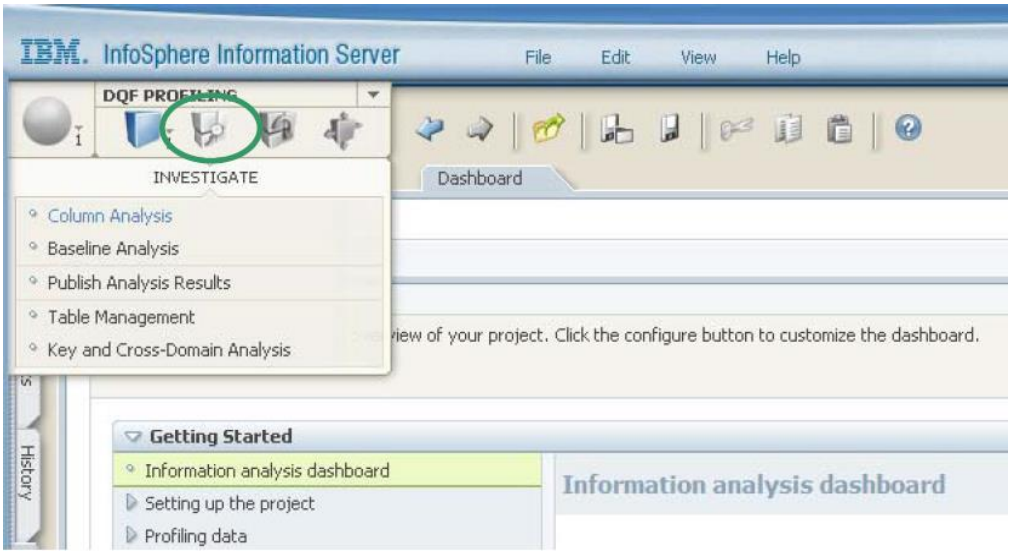


Figura 25: IBM Infosphere Information Analyzer – 4

Una vez terminado el análisis se presenta la siguiente pantalla donde se puede revisar el detalle mediante la opción Open Column Analysis:

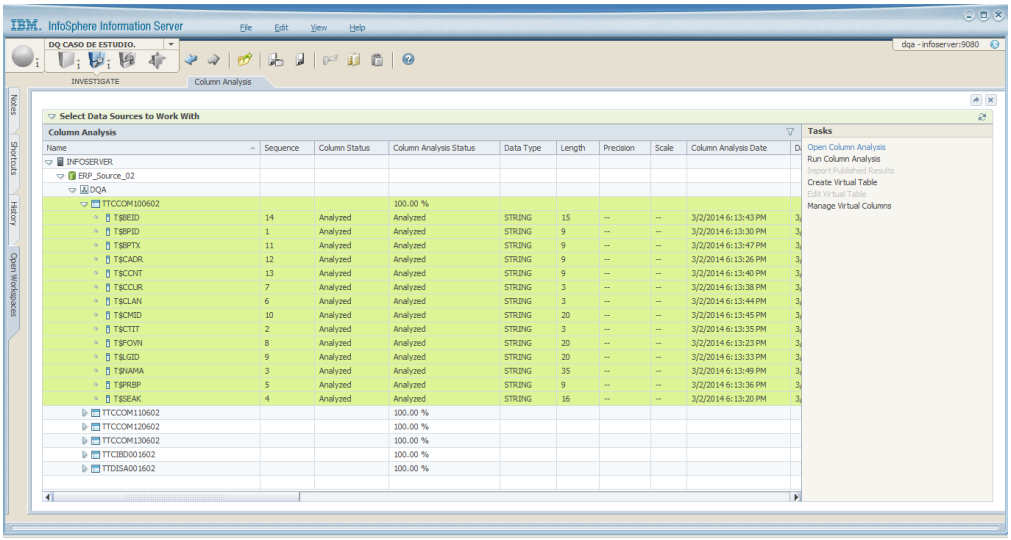


Figura 26: IBM Infosphere Information Analyzer – 5

Los resultados del análisis se pueden visualizar directamente en la herramienta Information Analyzer y se pueden publicar a través de Information Server Web Console en distintos formatos como PDF, Excel o HTML.

La **Figura 27** indica los resultados que se pueden verificar en Information Analyzer. Seleccionamos el campo T$NAMA que representa el nombre o razón social para ver mayor detalle:

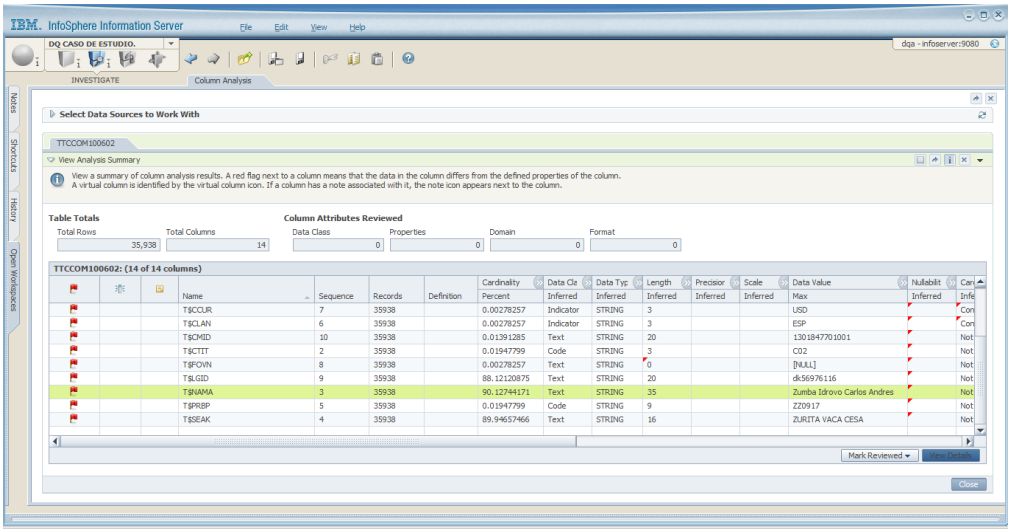


Figura 27: IBM Infosphere Information Analyzer – 6

Se puede observar distintos tipos de análisis por ejemplo Distribución de frecuencias o Formato.

El análisis de frecuencias nos indica que existen nombres repetidos y resaltado en color verde se marca un caso interesante “CONSUMIDOR FINAL” el cual se repite 44 veces, lo cual puede indicar un error en el ingreso de datos.

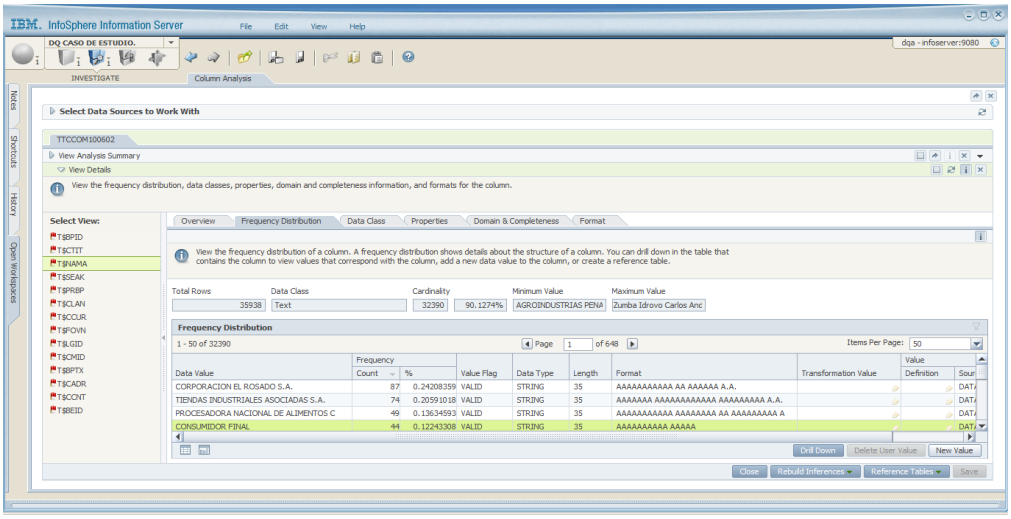


Figura 28: IBM Infosphere Information Analyzer – 7

Al visualizar los resultados a través de un gráfico de barras, se puede observar que existen varios nombres repetidos:

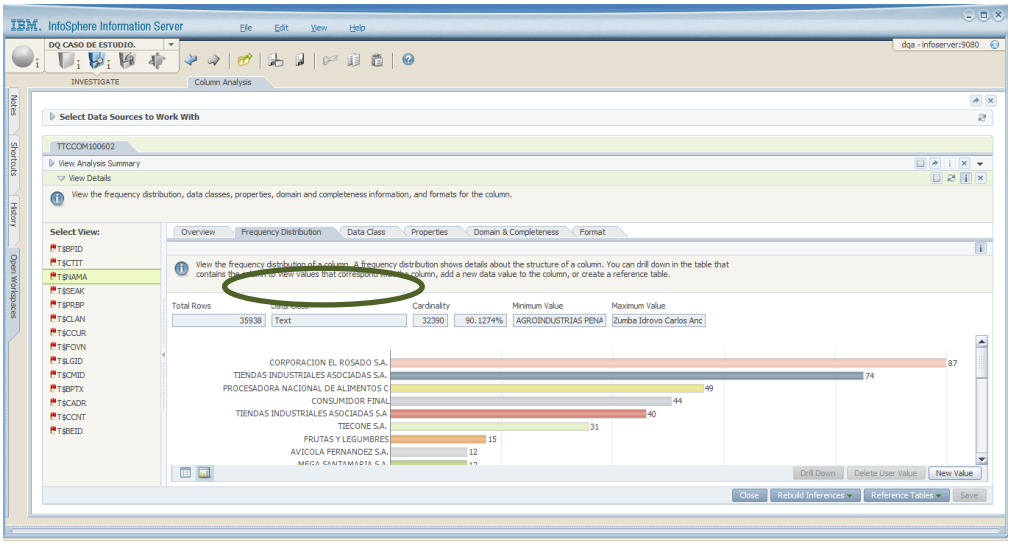


Figura 29: IBM Infosphere Information Analyzer – 8

El análisis de formatos nos indica que el formato “AAAAAAA AAAAAAA” es el más representativo con 234 coincidencias. El formato puede indicar “Nombre ApellidoPaterno ApellidoMaterno”.

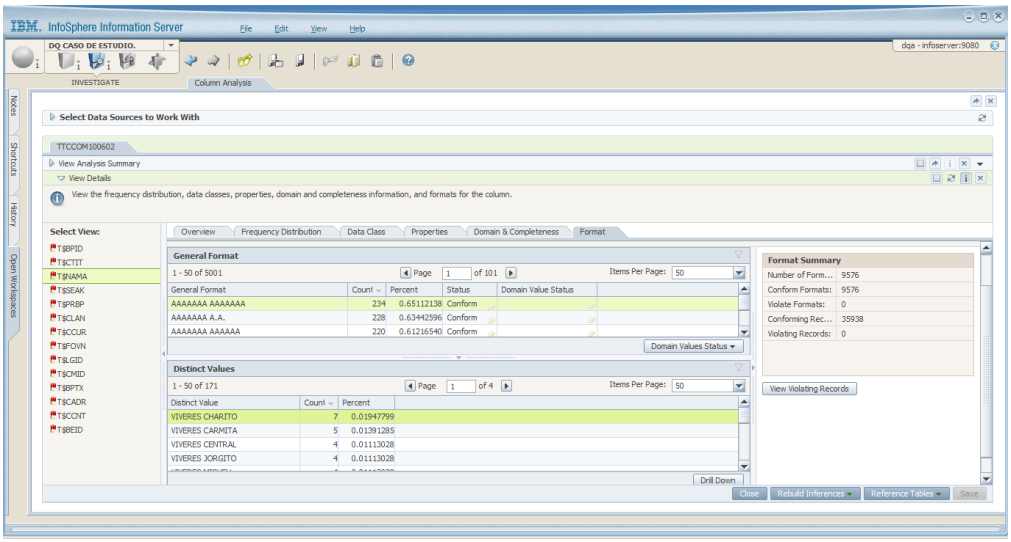


Figura 30: IBM Infosphere Information Analyzer – 9

3. Crear un reporte de análisis

Los reportes se generan en la herramienta Information Server Web Console. El reporte para el análisis de los formatos más frecuentes se obtiene con las opciones que se indican en la **Figura 31**.

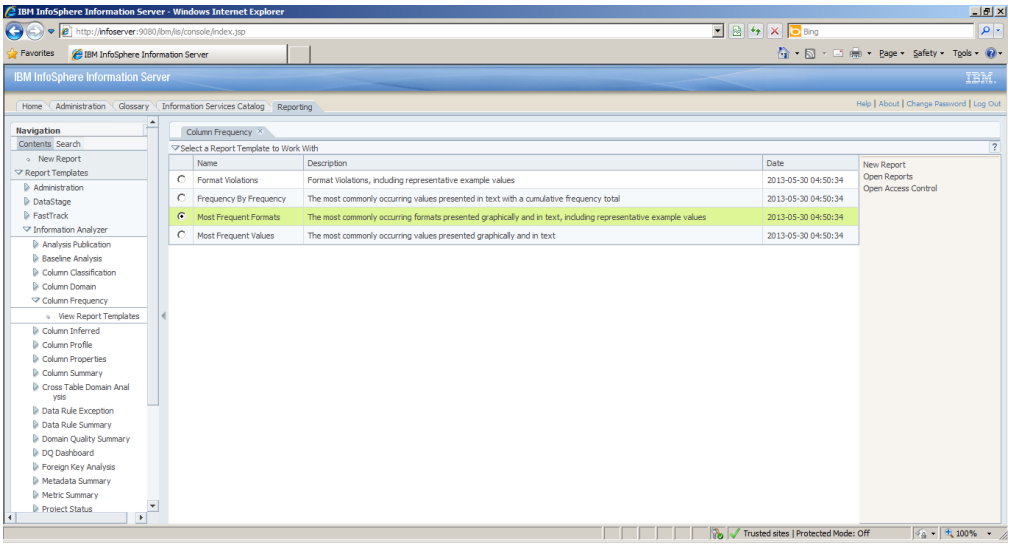


Figura 31: IBM Infosphere Information - Reporte 1

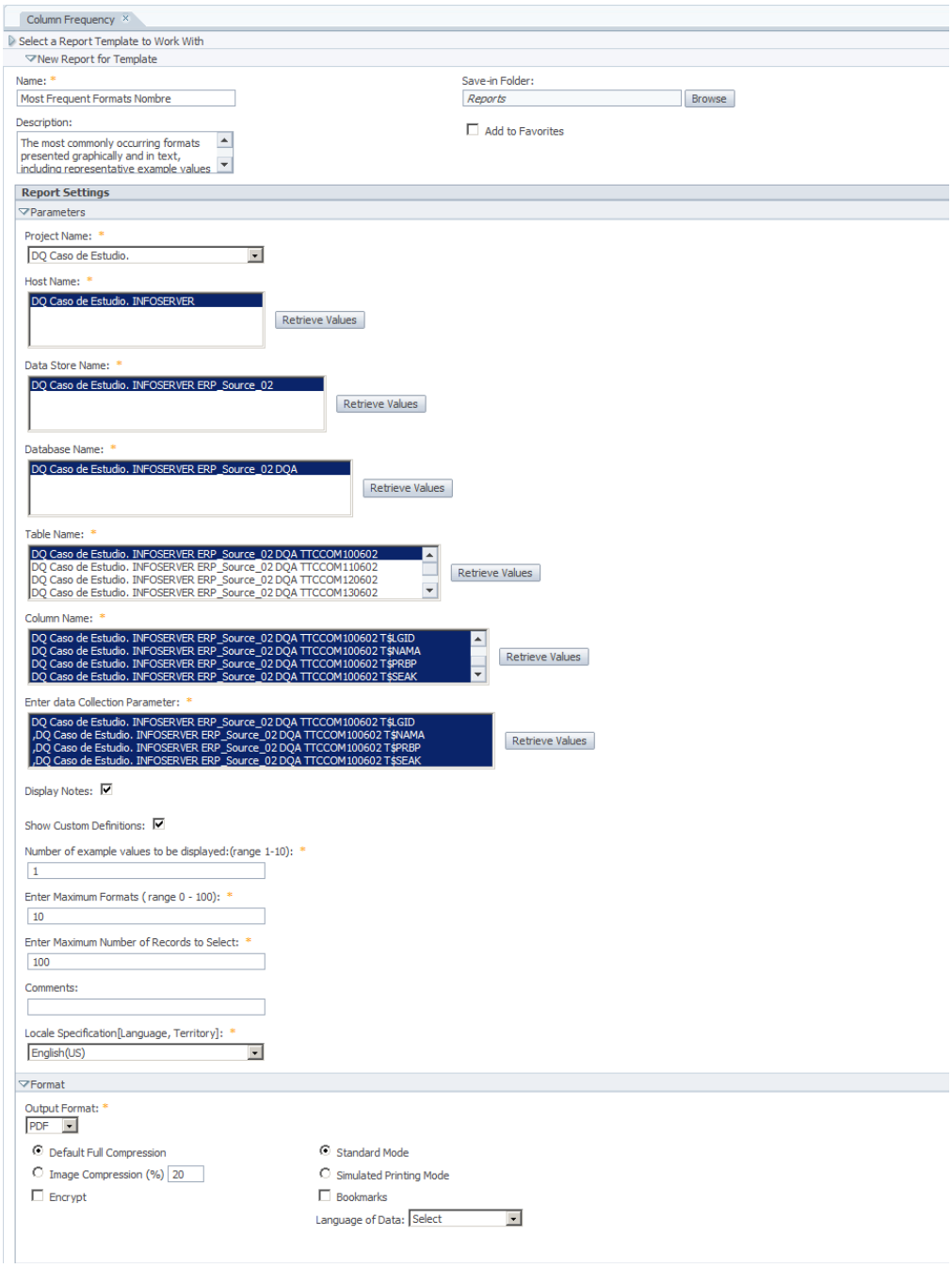


Figura 32: IBM Infosphere Information - Reporte 2

El resultado del reporte se indica en la **Figura 33**. En el gráfico de barras se puede observar que existen varios formatos para el nombre y no solo de personas sino de negocios también.

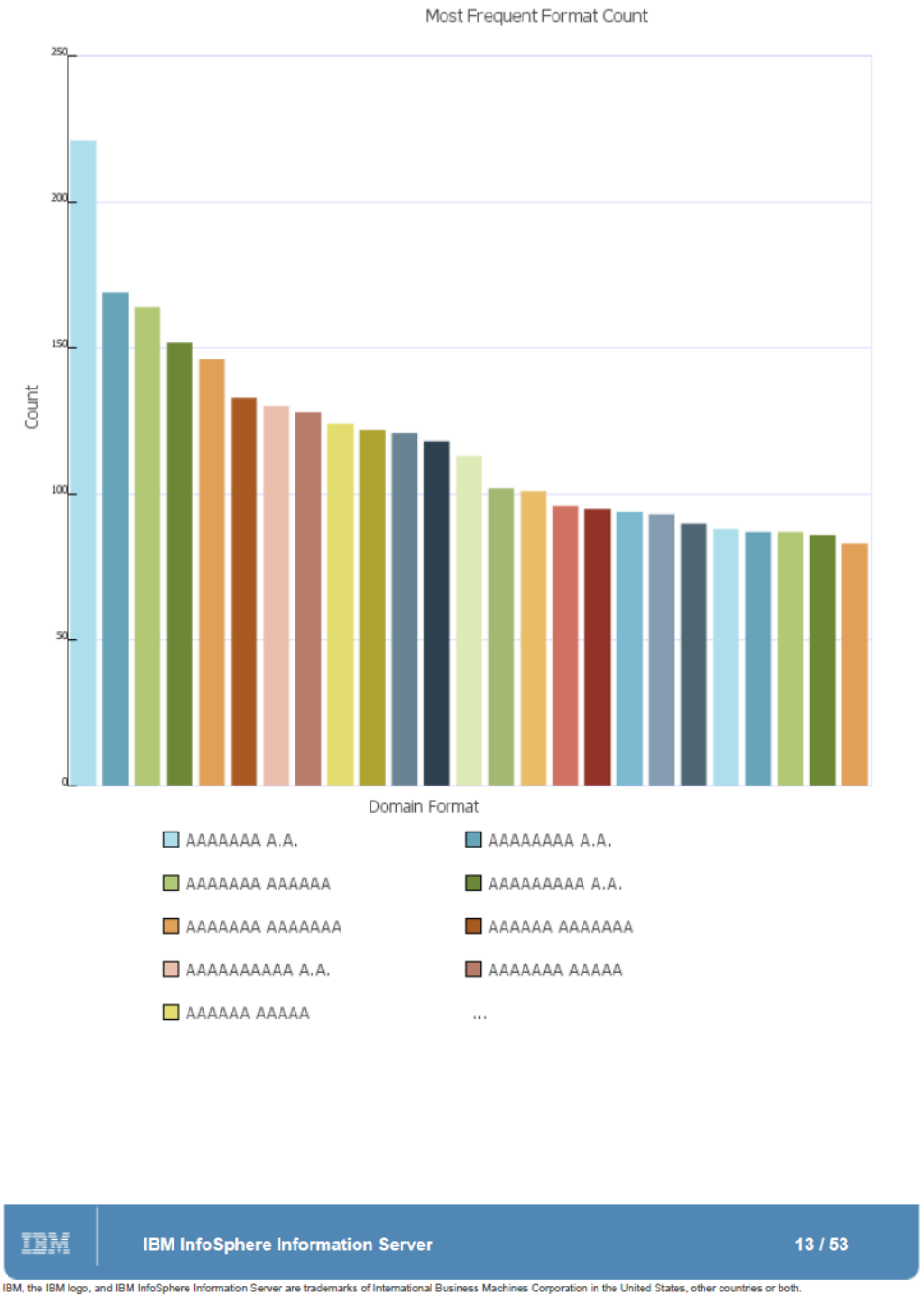


Figura 33: IBM Infosphere Information Analyzer – Resultado 1

El reporte también presenta los datos de manera tabular, donde indica uno de los formatos más utilizados con 221 repeticiones que indica el nombre de un negocio.

Column Level Details

Format	Count	Total Rows %	Total Rows Cumulative %	Example Values
AAAAAAA A.A.	221	0.66245017	0.66245017	TIECONE S.A. ATIMASA S.A. DULCAFE S.A. DOELDOS S.A. AQUAMAR S.A. NIKAMAR S.A. MISAGRO S.A. MARRIOT S.A. LADYBUS S.A. JERILEX S.A.
AAAAAAAA A.A.	169	0.50657954	1.16902971	ENMARDOS S.A. JUDISPRO S.A. BECROMAL S.A. INVERNEG S.A. CASABACA S.A. BORDAINS S.A. QUIFATEX S.A. NAPORTEC S.A. INTERDIN S.A. CONCLINA C.A.
AAAAAAA AAAAAA	164	0.49159198	1.66062169	VIVERES ROSITA VIVERES YOLITA SALINAS LIBANO TERCENA CANELA BALSECA MANUEL AVICOLA PIANGO

Figura 34: IBM Infosphere Information Analyzer – Resultado 2

El reporte con los valores más frecuentes se obtiene con las opciones que se indican a continuación:

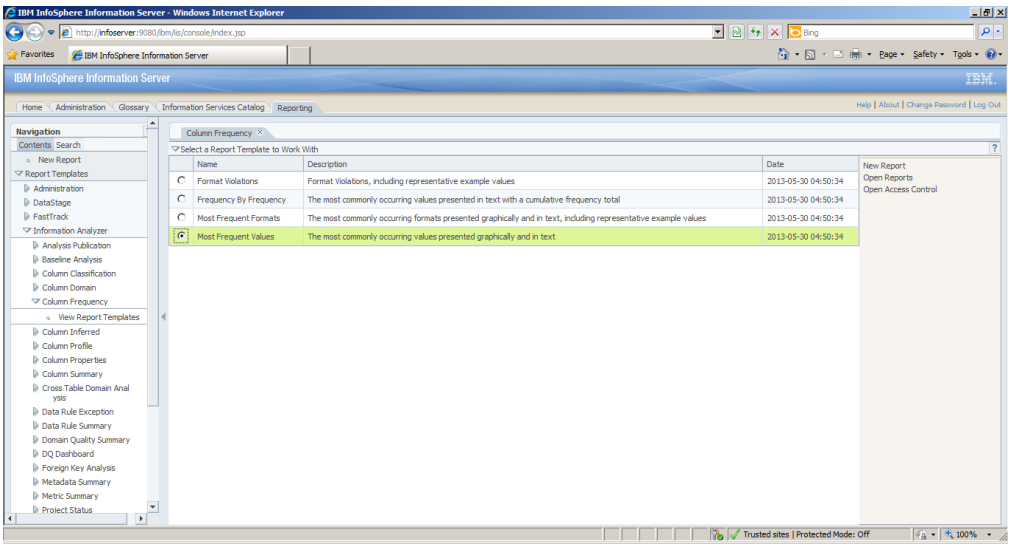


Figura 35: IBM Infosphere Information Analyzer – Reporte 3

Figura 36: IBM Infosphere Information Analyzer – Reporte 4

El resultado del reporte se indica en la **Figura 37**. En el gráfico de barras de puede observar que existen algunos valores repetidos que debemos depurar en las siguientes etapas de la calidad de datos.

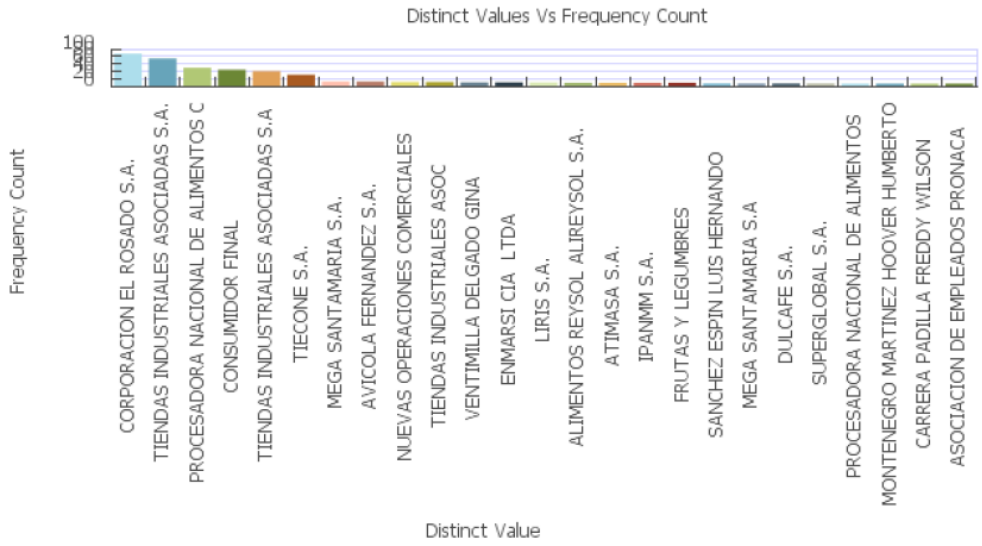


Figura 37: IBM Infosphere Information Analyzer – Resultado 4

El reporte también indica los clientes que fueron reconocidos como repetidos.

Frequency Distribution Data

Distinct Value	Frequency Count	Frequency %	Cumulative %
CORPORACION EL ROSADO S.A.	87	0.26078355	0.26078355
TIENDAS INDUSTRIALES ASOCIADAS S.A.	74	0.22181589	0.48259944
PROCESADORA NACIONAL DE ALIMENTOS C	49	0.14687809	0.62947753
CONSUMIDOR FINAL	44	0.13189053	0.76136806
TIENDAS INDUSTRIALES ASOCIADAS S.A	40	0.11990048	0.88126854
TIECONE S.A.	30	0.08992536	0.97119390
MEGA SANTAMARIA S.A.	12	0.03597014	1.00716404
AVICOLA FERNANDEZ S.A.	12	0.03597014	1.04313418
NUEVAS OPERACIONES COMERCIALES	11	0.03297263	1.07610681
TIENDAS INDUSTRIALES ASOC	11	0.03297263	1.10907944
VENTIMILLA DELGADO GINA	10	0.02997512	1.13905456
ENMARSI CIA LTDA	10	0.02997512	1.16902968

Figura 38: IBM Infosphere Information Analyzer – Resultado 5

4. Asignar permisos de visualización de reportes

Para poder publicar el reporte primero es necesario asignar seguridades, esto se realiza a través de Information Server Web Console. La configuración del control de acceso se resalta en la **Figura 39**:

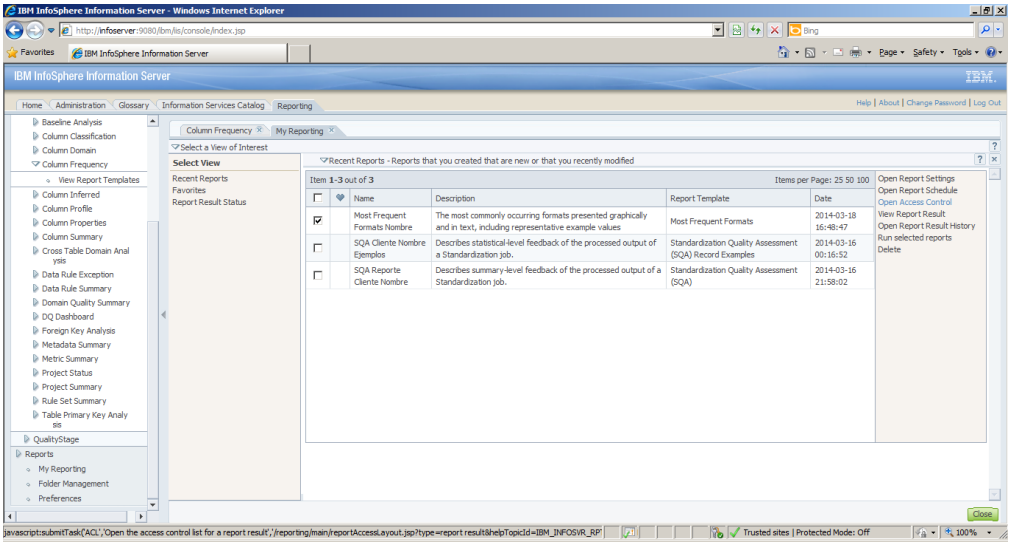


Figura 39: IBM Infosphere Information Analyzer – Permisos 1

Seleccione el usuario al que se van a asignar los permisos de acceso.

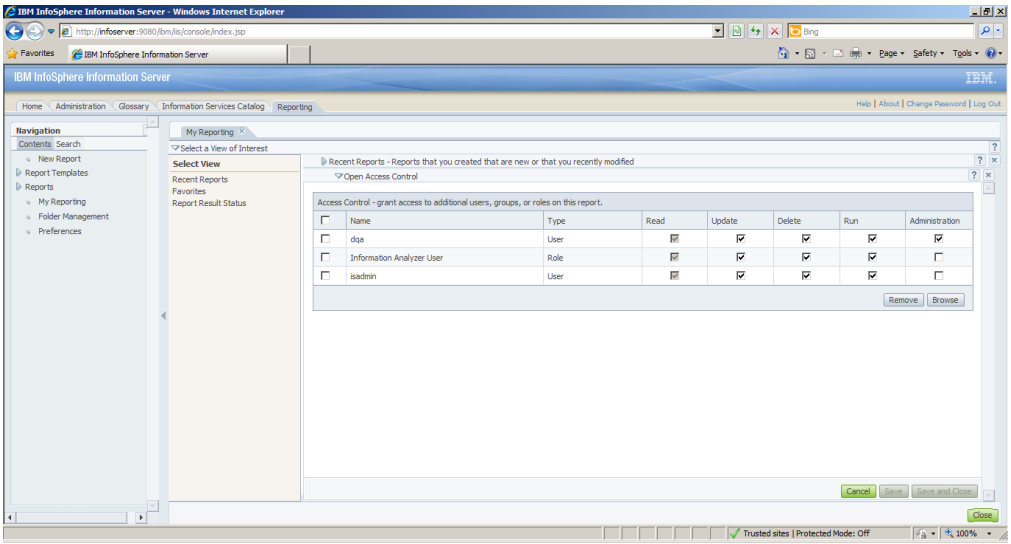


Figura 40: IBM Infosphere Information Analyzer – Permisos 1

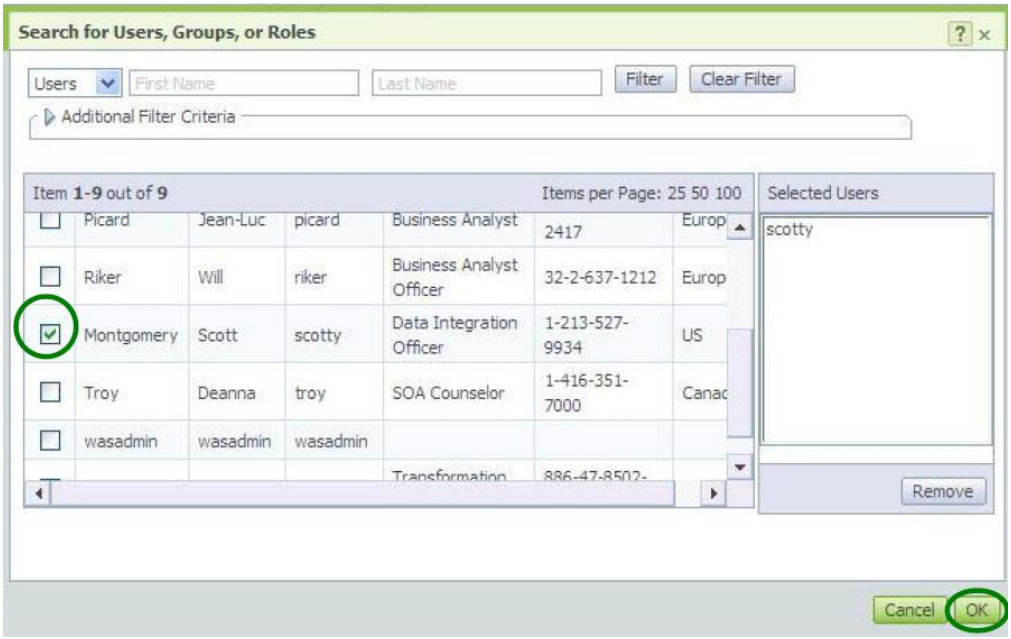


Figura 41: IBM Infosphere Information Analyzer – Permisos 2

Asignar los permisos de Actualización, Eliminación y Ejecución.

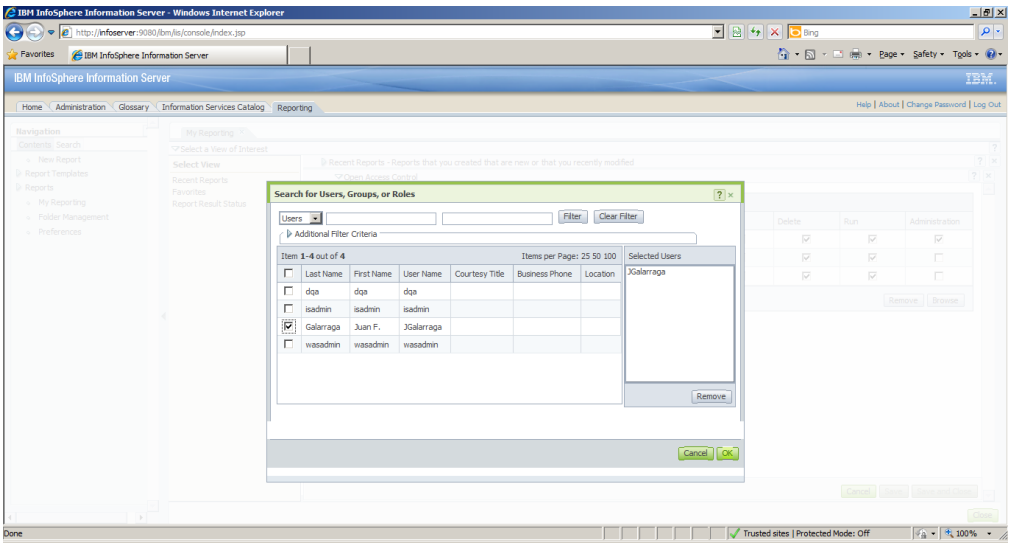


Figura 42: IBM Infosphere Information Analyzer – Permisos 3

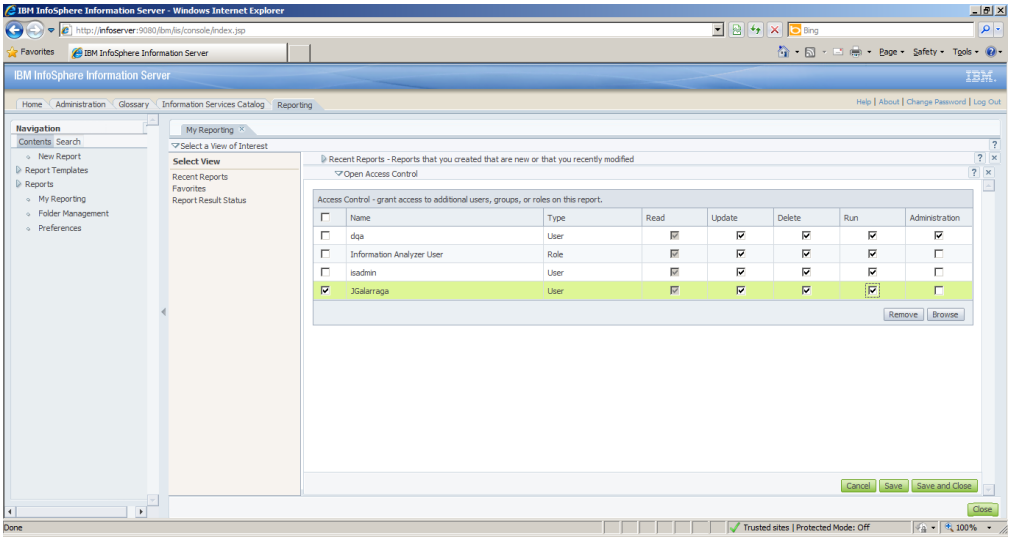


Figura 43: IBM Infosphere Information Analyzer – Permisos 4

5. Publicar los resultados del análisis

Para visualizar los reportes publicados se utiliza la herramienta IBM Infosphere Metadata Workbech.

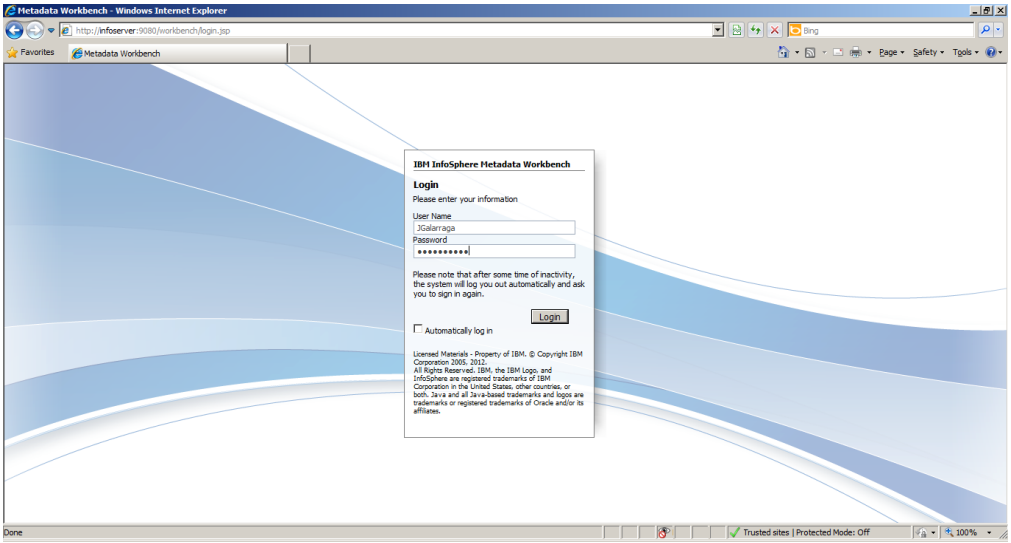


Figura 44: IBM Infosphere Information Analyzer – Publicación 1

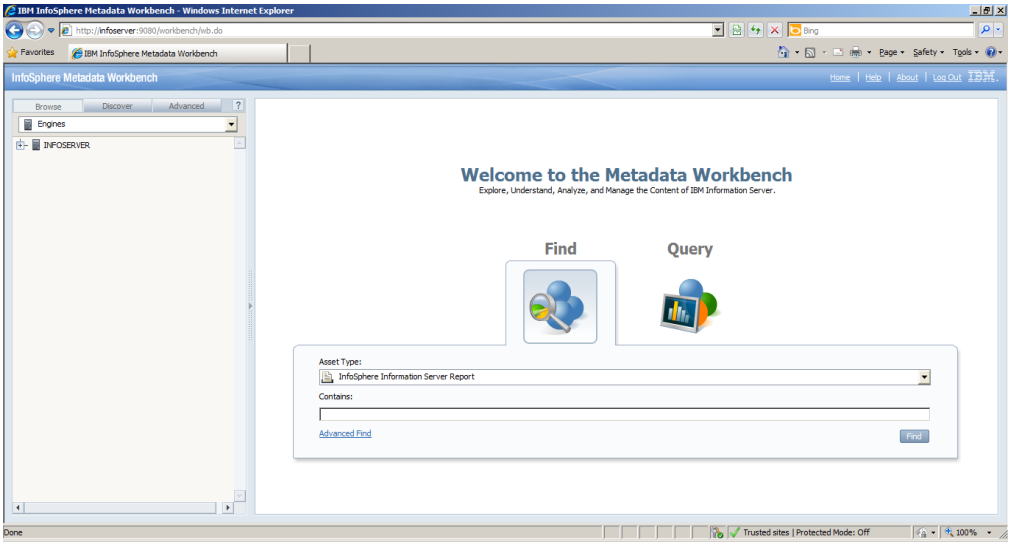


Figura 45: IBM Infosphere Information Analyzer – Publicación 2

Buscar el reporte que deseamos evaluar.

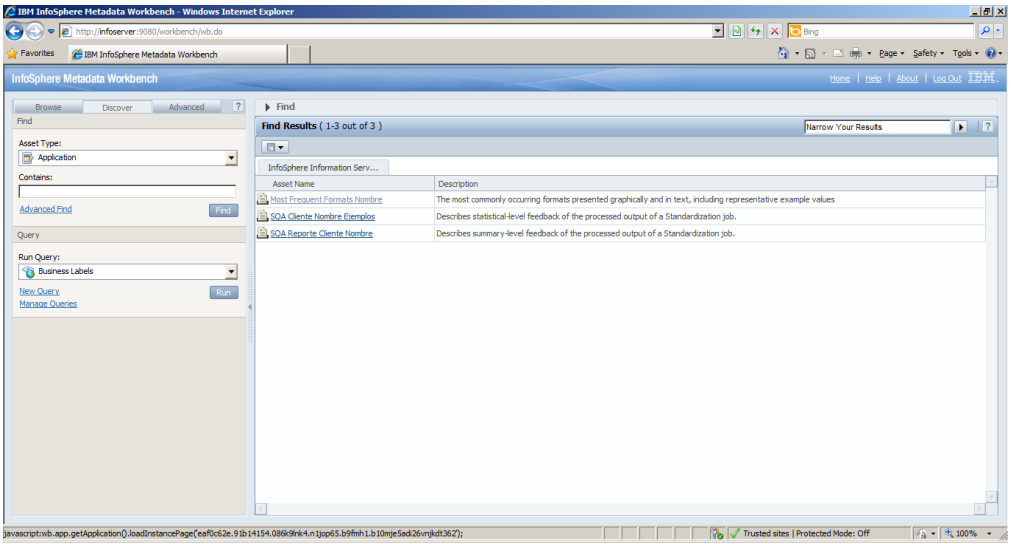


Figura 46: IBM Infosphere Information Analyzer – Publicación 3

A través del link se puede acceder al reporte en formato PDF.

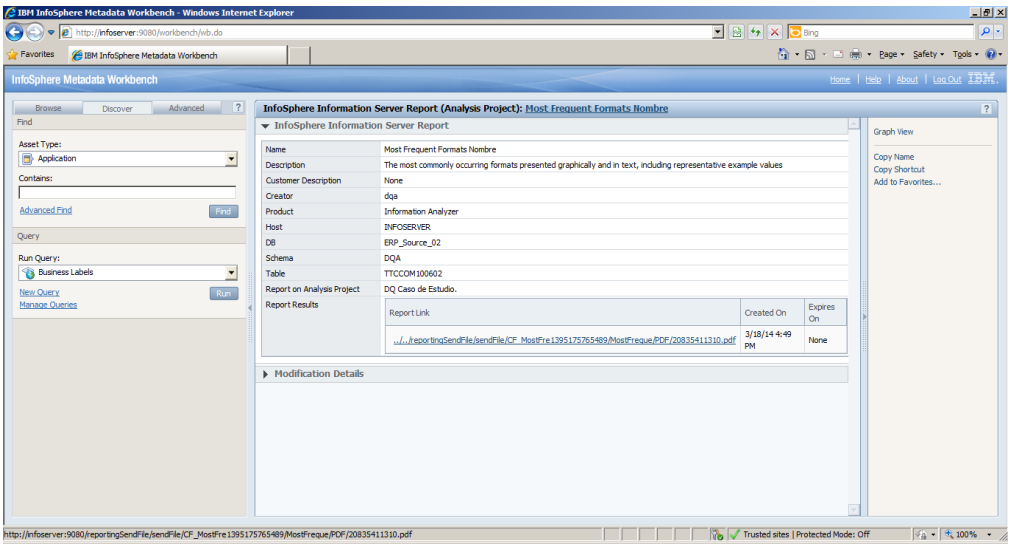


Figura 47: IBM Infosphere Information Analyzer – Publicación 4

Mediante estos reportes y el análisis en IBM Information Analyzer se puede establecer un estado inicial de los datos el cual sirve para las siguientes etapas del proceso.

3.2.2.4 Investigación

El objetivo principal de la etapa de investigación es identificar los problemas de datos en las fuentes de información para detectar datos que no cumplan las características de calidad y corregirlos antes de que estos sean consumidos por los sistemas de la organización.

Esta etapa nos va a permitir revelar reglas de negocio no documentadas, identificar formatos, valores por defecto y valores vacíos a través del análisis para cada columna o combinación de columnas.

Existen 3 métodos para realizar la investigación de los datos:

1. **Character Discrete:** Analiza valores y formatos, este método genera un reporte de frecuencias.

2. **Character Concatenate:** Analiza valores y formatos de campos relacionados.

3. **Word Investigation:** Identifica tokens que van a ser analizados para encontrar patrones de los datos, los cuales son obtenidos a través de la utilización de reglas definidas. Con este método obtenemos reportes de patrones y tokens de las columnas que definimos analizar.

Para el caso de estudio se realizara el análisis a través del método Word Investigation de la tabla de Clientes que se encuentra almacenada en una base de datos IBM DB2, este método se aplica sobre los datos utilizando la herramienta IBM Infosphere DataStage and QualityStage Designer.

Inicialmente se debe combinar la información de las dos tablas que tienen información del cliente, es decir TTCCOM100 y TTCCOM130, en una sola tabla para ser analizada mediante el siguiente proceso desarrollado en IBM Infosphere Datastage Designer.

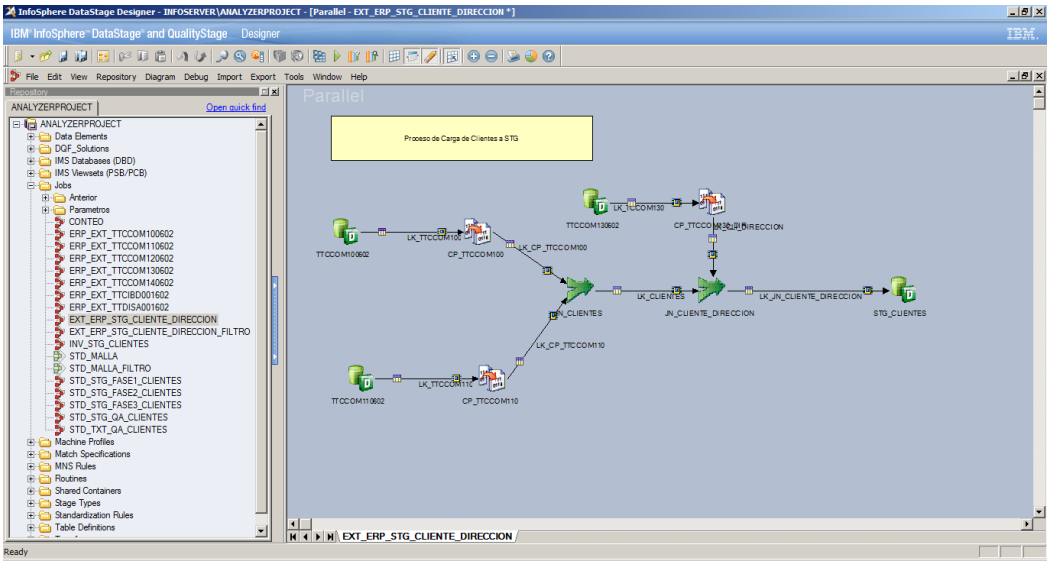


Figura 48: Investigación Paso – 1

Luego de obtener la información consolidada se establece el siguiente proceso para realizar la investigación por el método Word Investigation.

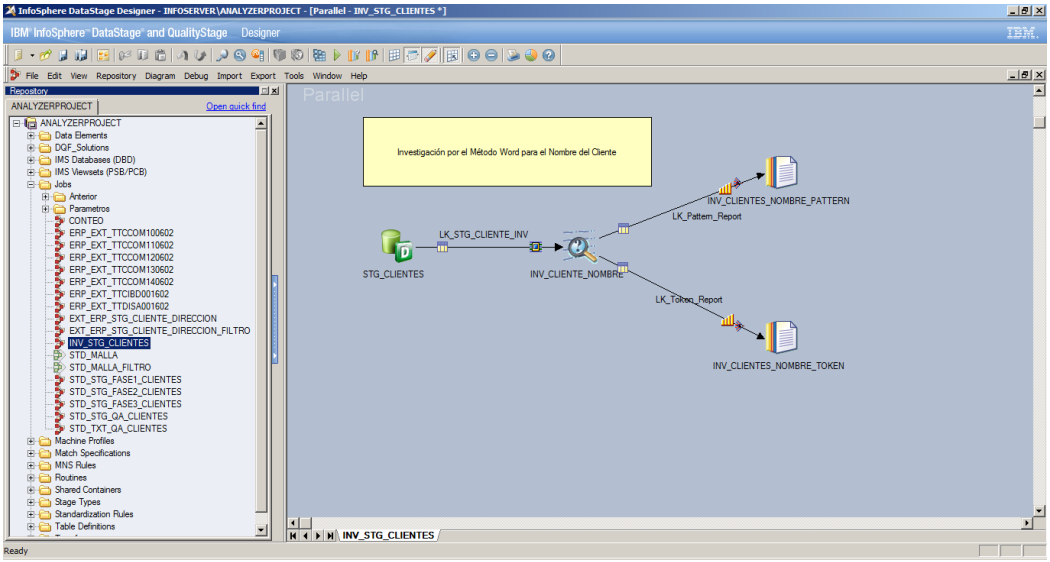
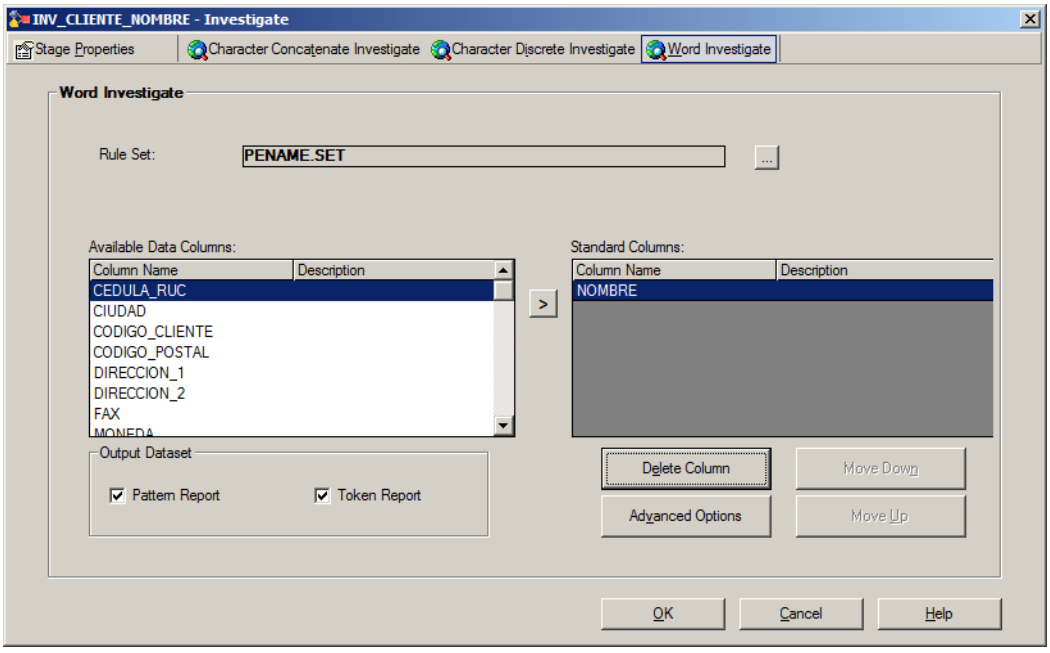


Figura 49: Investigación Paso – 2

Para realizar la investigación se requiere de una regla, en este caso utilizamos la regla PENAME incluida por defecto en la herramienta. Esta regla identifica información del nombre de las personas como nombres, apellidos e iniciales. Adicionalmente configuramos las salidas para que muestre el reporte de tokens[2] y de patrones.[3]



[2] Token: Un elemento sintáctico, tales como una frase, una palabra, o un conjunto de uno o más caracteres, que se utiliza para el análisis y el procesamiento de texto.

[3] Patrón: Una secuencia de caracteres utilizado para la notación de expresiones regulares, como un medio para seleccionar varios caracteres o cadenas de nombres, respectivamente.

Figura 50: Investigación Paso – 3

En la opción de Stage Properties se mapean todas las columnas a los destinos, luego a cada destino se asigna un archivo de salida. Al ejecutar el proceso se visualizan los resultados.

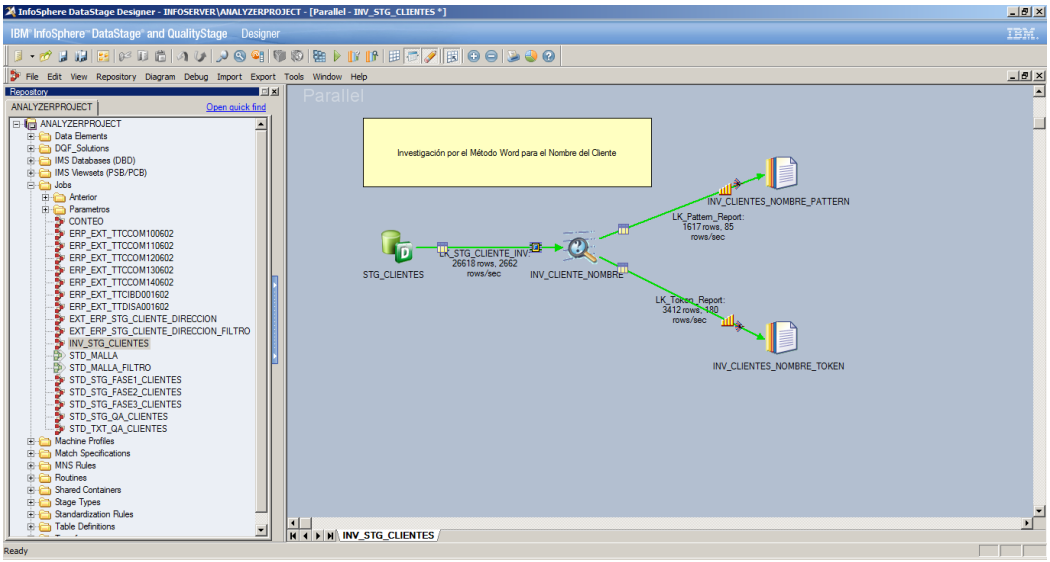


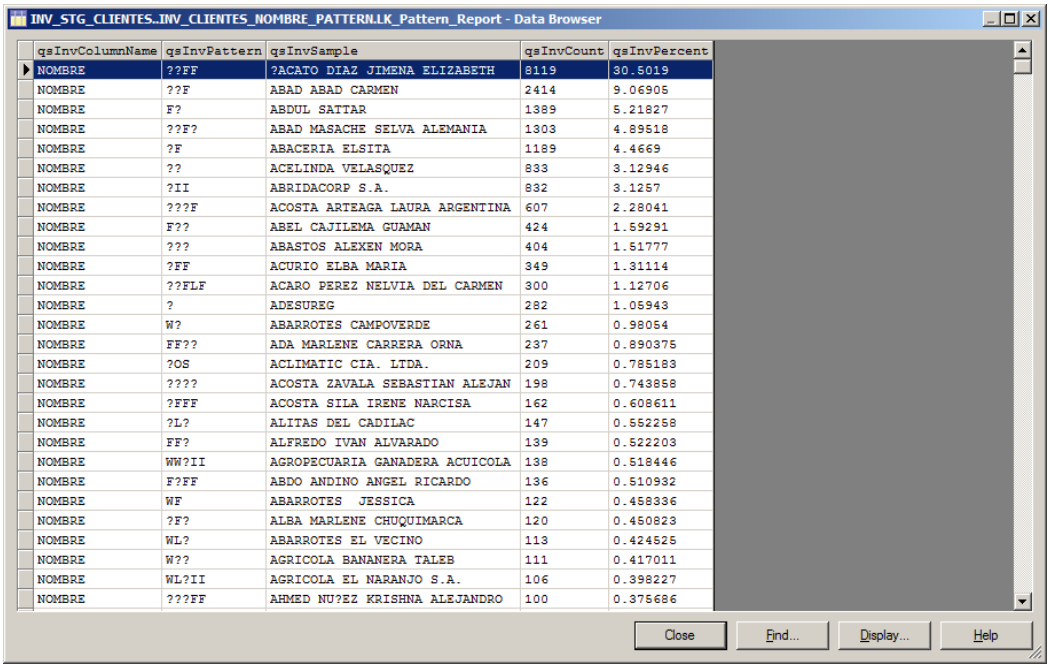
Figura 51: Investigación Paso – 4

Análisis de Patrones: Se puede apreciar en la **Figura 52** que existen 2 patrones que son los más representativos:

1) **??FF:** Donde **F =** Primer Nombre? **=** Palabra (Se puede inferir como Apellido)

2) **??F:** Donde **F =** Primer Nombre? **=** Palabra (Se puede inferir como Apellido)

Se puede decir que existen algunos patrones que son los mismos ya que al parecer hay algunos nombres que no se reconocen pero la cadena tiene misma estructura.



qsInvColumnName	qsInvPattern	qsInvSample	qsInvCount	qsInvPercent
NOMBRE	??FF	?ACATO DIAZ JIMENA ELIZABETH	8119	30.5019
NOMBRE	??F	ABAD ABAD CARMEN	2414	9.06905
NOMBRE	F?	ABDUL SATTAR	1389	5.21827
NOMBRE	??F?	ABAD MASACHE SELVA ALEMANIA	1303	4.89518
NOMBRE	?F	ABACERIA ELSITA	1189	4.4669
NOMBRE	??	ACELINDA VELASQUEZ	833	3.12946
NOMBRE	?II	ABRIDACORP S.A.	832	3.1257
NOMBRE	???F	ACOSTA ARTEAGA LAURA ARGENTINA	607	2.28041
NOMBRE	F??	ABEL CAJILEMA GUAMAN	424	1.59291
NOMBRE	???	ABASTOS ALEXEN MORA	404	1.51777
NOMBRE	?FF	ACURIO ELBA MARIA	349	1.31114
NOMBRE	??FLF	ACARO PEREZ NELVIA DEL CARMEN	300	1.12706
NOMBRE	?	ADESUREG	282	1.05943
NOMBRE	W?	ABARROTES CAMPOVERDE	261	0.98054
NOMBRE	FF??	ADA MARLENE CARRERA ORNA	237	0.890375
NOMBRE	?OS	ACLIMATIC CIA. LTDA.	209	0.785183
NOMBRE	????	ACOSTA ZAVALA SEBASTIAN ALEJAN	198	0.743858
NOMBRE	?FFF	ACOSTA SILA IRENE NARCISA	162	0.608611
NOMBRE	?L?	ALITAS DEL CADILAC	147	0.552258
NOMBRE	FF?	ALFREDO IVAN ALVARADO	139	0.522203
NOMBRE	WW?II	AGROPECUARIA GANADERA ACUICOLA	138	0.518446
NOMBRE	F?FF	ABDO ANDINO ANGEL RICARDO	136	0.510932
NOMBRE	WF	ABARROTES JESSICA	122	0.458336
NOMBRE	?F?	ALBA MARLENE CHUQUIMARCA	120	0.450823
NOMBRE	WL?	ABARROTES EL VECINO	113	0.424525
NOMBRE	W??	AGRICOLA BANANERA TALEB	111	0.417011
NOMBRE	WL?II	AGRICOLA EL NARANJO S.A.	106	0.398227
NOMBRE	???FF	AHMED NU?EZ KRISHNA ALEJANDRO	100	0.375686

Figura 52: Investigación Paso – 5

El listado del significado de cada letra se presenta en la **Figura 53**:

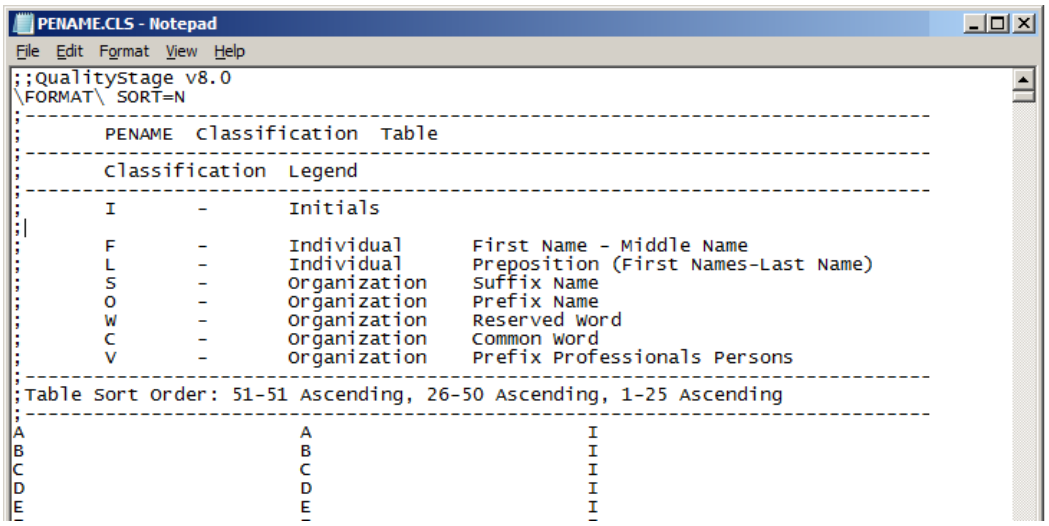


Figura 53: Investigación Paso – 6

Análisis de Tokens: El reporte de tokens nos indica cuales son las cadenas que fueron encontradas con mayor frecuencia. Por ejemplo la Inicial “A” se repite en 2030 registros y el primer nombre “MARIA” se repite en 1623 ocasiones.

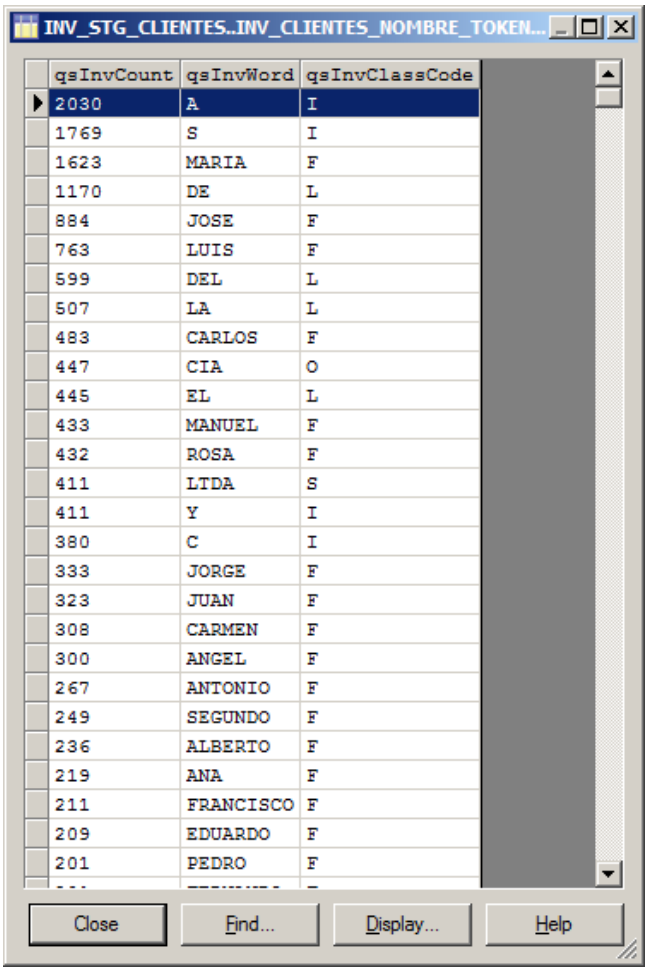


Figura 54: Investigación Paso – 7

3.2.2.5 Estandarización

Con la estandarización la compañía asegura que sus datos son consistentes, es decir datos con el mismo tipo de contenido y el mismo formato.

Basado en la salida que entrega la fase de Investigación se decide que reglas aplicar para dar formato a los datos de los diferentes sistemas. Estas reglas permiten incorporar estándares de la industria y del negocio. Cabe señalar que la herramienta con la cual se desarrolla este proceso es IBM DataStage and QualityStage Designer la cual tiene reglas predefinidas para los nombres, direcciones, números telefónicos, etc.

El proceso de estandarización consiste en aplicar al dato de entrada un patrón el cual permite clasificar cada token y estructurar la salida en función de la regla establecida, así lo indica la **Figura 55**:

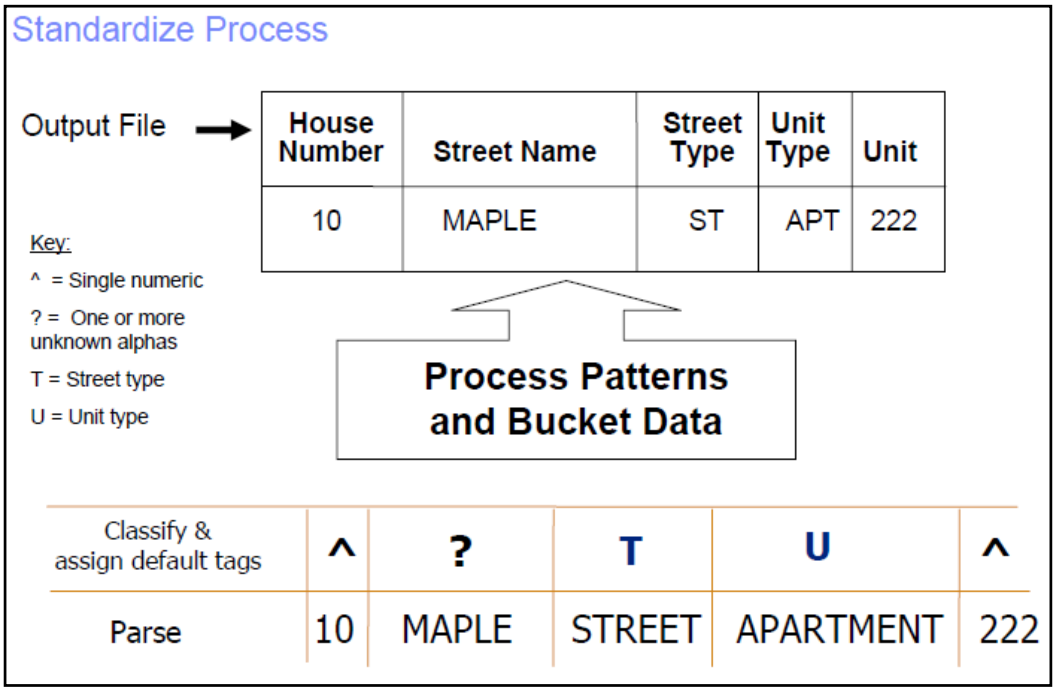


Figura 55: Proceso de Estandarización
Fuente: (IBM, 2012)

Las Reglas de dominio son usadas para estandarizar datos aplicando reglas acorde al criterio de un país. Para el caso de estudio es necesario identificar este tipo de reglas para clasificar los datos añadiendo el código ISO del país “EC”. Esto se aplica especialmente cuando la información viene de más de 2 países.

1. **Creación Proceso de Estandarización para añadir el código ISO del País.**

Crear un nuevo proceso paralelo en IBM Infosphere DataStage con la configuración que indica la figura siguiente:

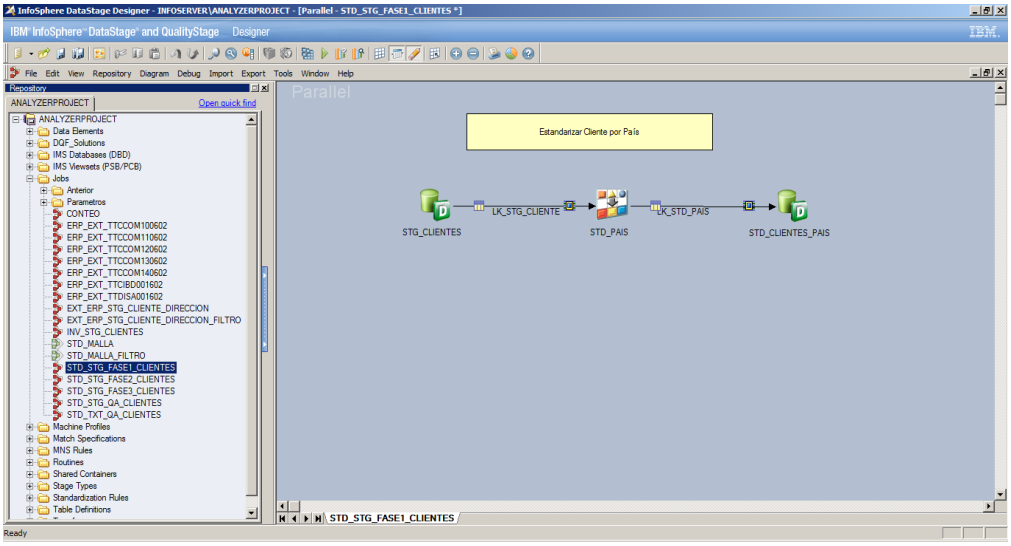


Figura 56: Estandarización Paso – 1

Configure una nueva regla en el objeto Standarize. La siguiente figura indica como las reglas están cargadas por defecto en la herramienta:

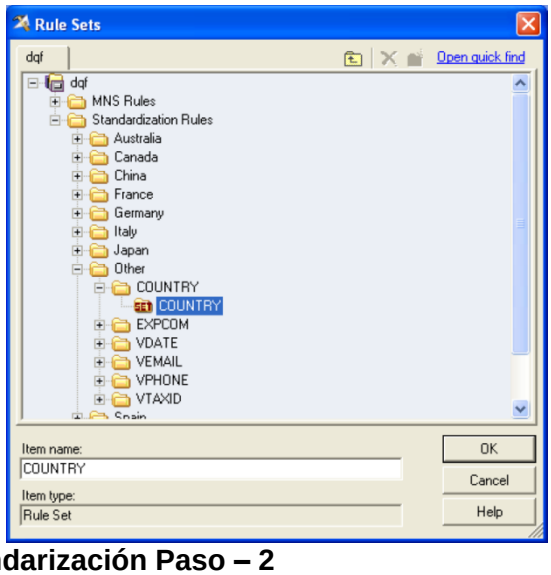


Figura 57: Estandarización Paso – 2

Seleccione los campos de ubicación como dirección, ciudad y provincia para poder aplicar la regla de dominio para el país.

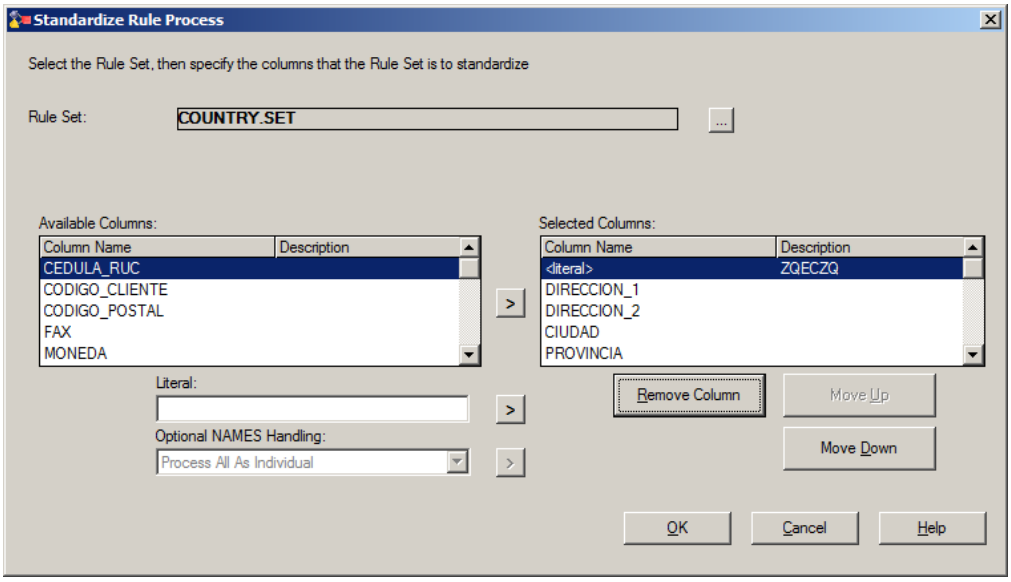


Figura 58: Estandarización Paso – 3

Se puede observar que para la ejecución de ECU(EC) algunas de las direcciones son clasificadas correctamente y se marcan con la letra "Y", otras direcciones solo se deducen y se marcan con "N"

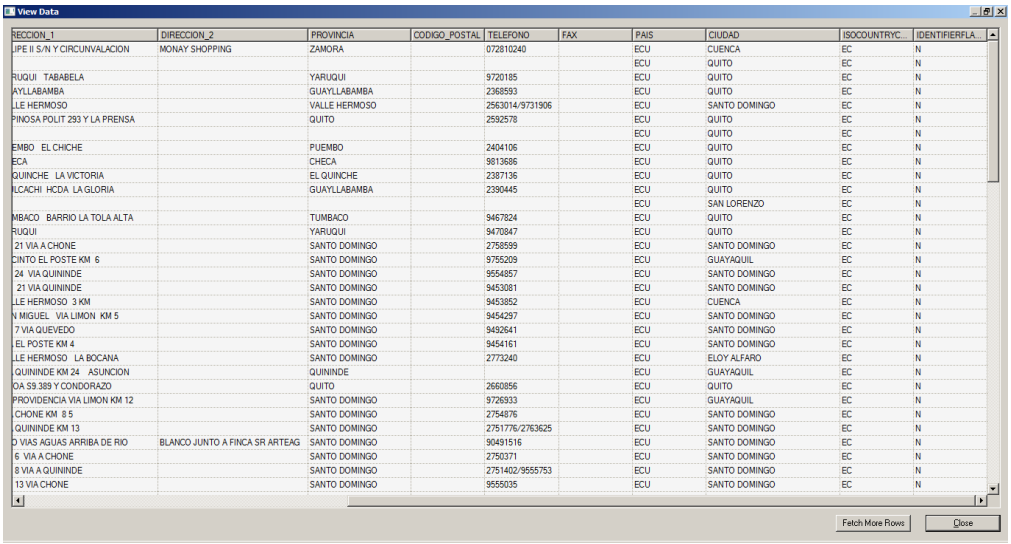


Figura 59: Estandarización Paso – 4

2. Preparar los datos para estandarización usando el pre-procesador

Crear un nuevo proceso paralelo en IBM Infosphere DataStage con la configuración que indica la siguiente figura.

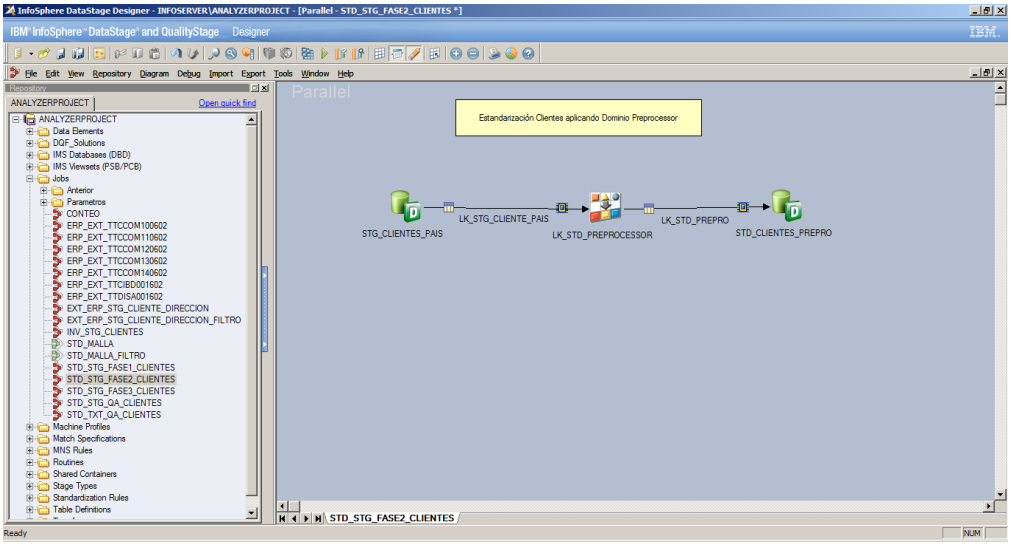


Figura 60: Estandarización Paso – 5

Después de configurar las entradas del proceso, configuramos la etapa de Estandarización en función de la regla PEPREP, configurando cada columna referente a la ubicación como indica la figura:

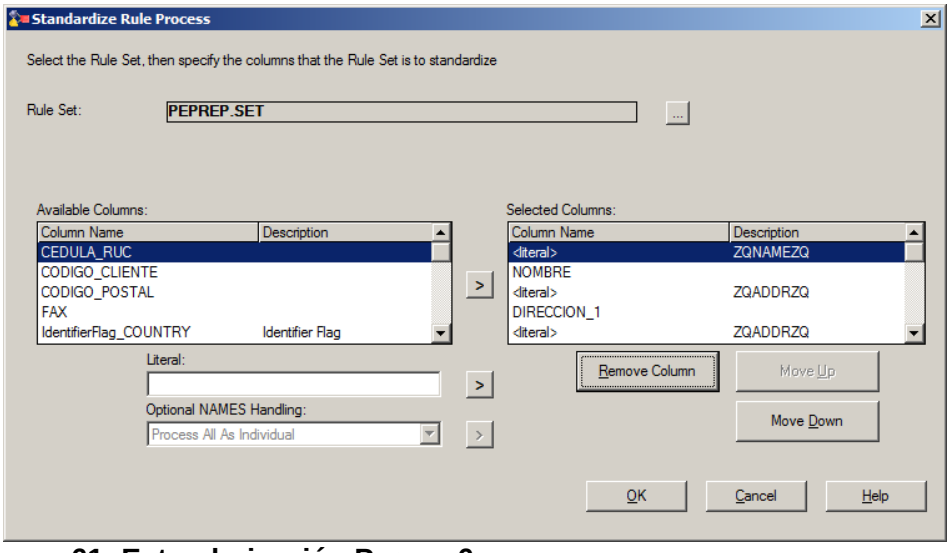


Figura 61: Estandarización Paso – 6

Literal: ZQNAMEZQ

Columna: NOMBRE

Literal: ZQADDRZQ

Columna: DIRECCION_1

Literal: ZQADDRZQ

Columna: DIRECCION_2

Luego de ejecutar el proceso se pueden verificar que las columnas DOMINIONOMBRE_PEPREP, DOMINIODIRECCION_PEPREP y DOMINIOAREA_PEPREP tienen el formato y la estructura necesaria para proceder con el siguiente paso de la estandarización. En las figuras se puede verificar que los campos mencionados están en letra mayúscula y juntan los campos que configuramos con la combinación de Literal y Columna.

Figura 62: Estandarización Paso – 7

3. **Crear un proceso para estandarizar los datos usando reglas específicas de dominio**

En este paso se deben crear procesos similares al que se indica en la **Figura 63** en donde se toma como fuente los datos resultantes del proceso anterior para aplicar reglas específicas de dominio sobre la información. Para la estandarización de los datos de Cliente se configura el siguiente proceso.

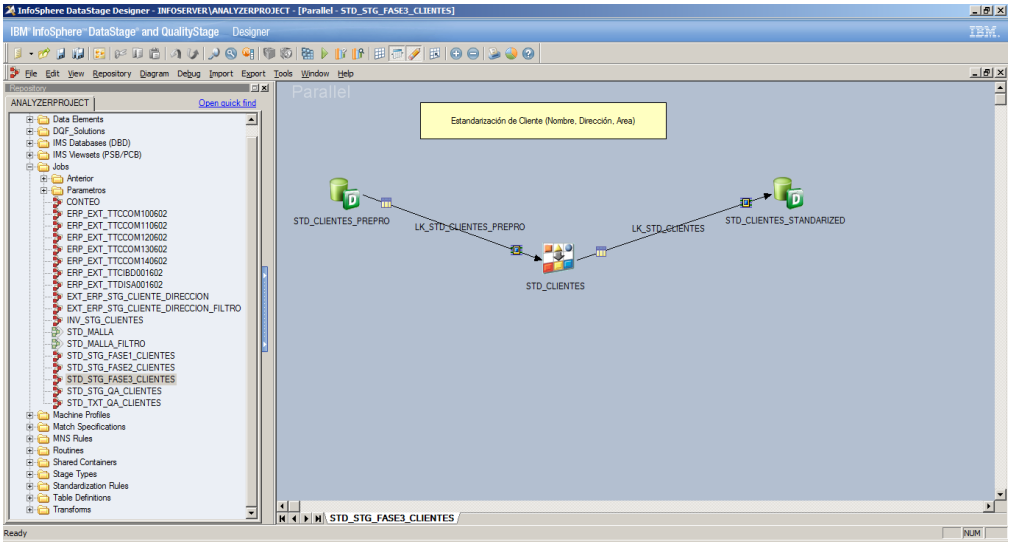


Figura 63: Estandarización Paso – 8

En la etapa de Estandarización se configuran tres procesos, cada uno debe tener una de las columnas que se generaron en el proceso anterior (DOMINIONOMBRE_PEPREP, DOMINIODIRECCION_PEPREP y DOMINIOAREA_PEPREP)

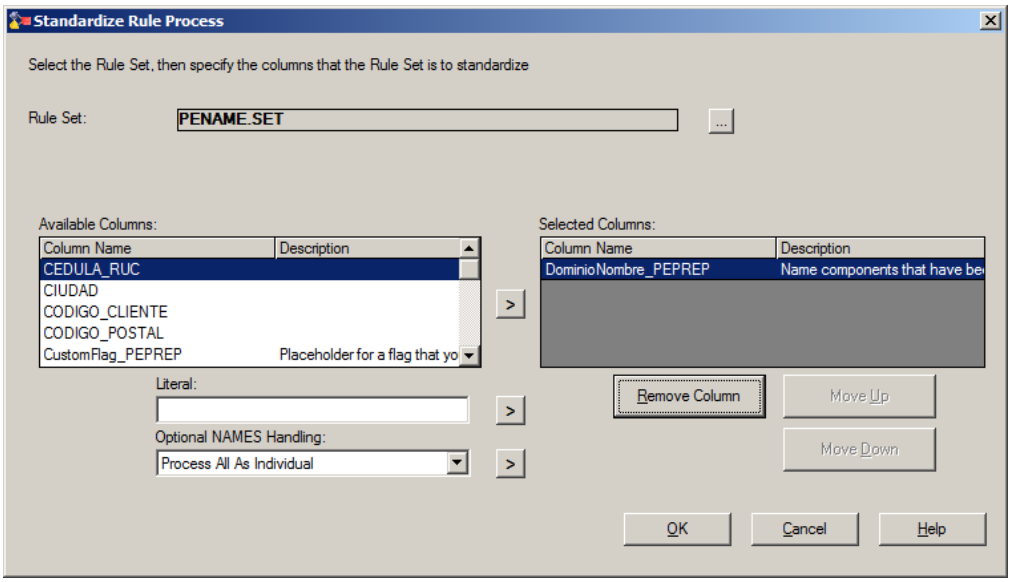


Figura 64: Estandarización Paso – 9

El proceso debe configurarse de la siguiente manera:

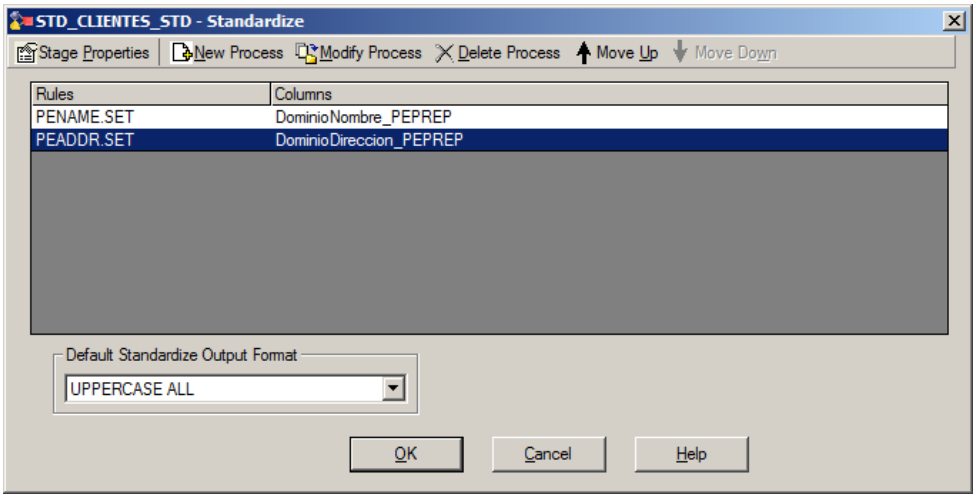


Figura 65: Estandarización Paso – 10

Al final de la configuración se configuran las salidas y se ejecuta el proceso. Se puede observar en las salidas que los datos de entrada han sido separados y organizados en diferentes campos dependiendo del dominio al cual representan por ejemplo si es nombre inicial, apellido, calle, número, etc.

En la figura se indica como el campo NOMBRE se ha dividido en diferentes campos: PRIMERNOMBRE_PENAME, SEGUNDONOMBRE_PENAME, APELLIDO_PENAME, etc.

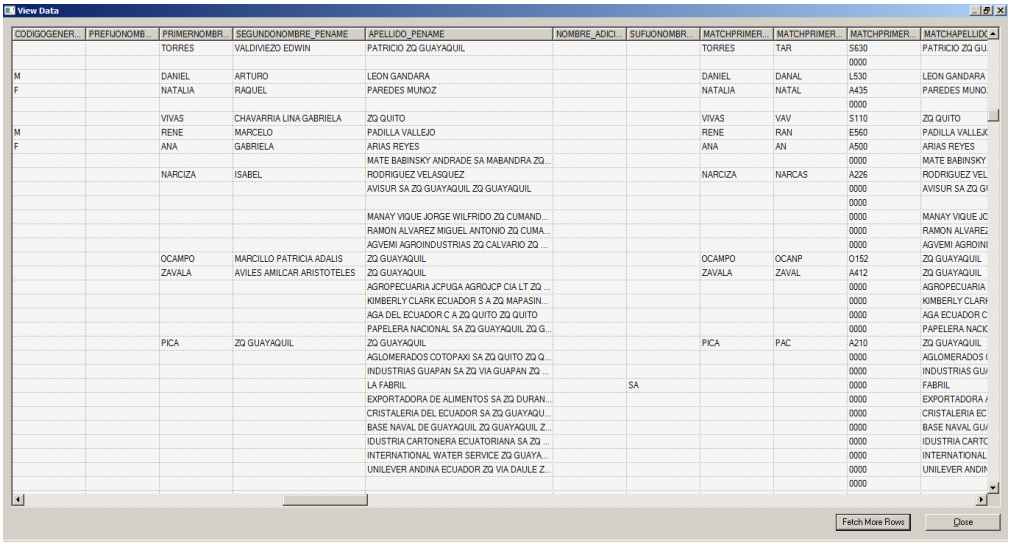


Figura 66: Estandarización Paso – 11

Para la dirección se dividió la información en columnas como: CALLE_PEADDR, NUMERO_PEADDR, etc.

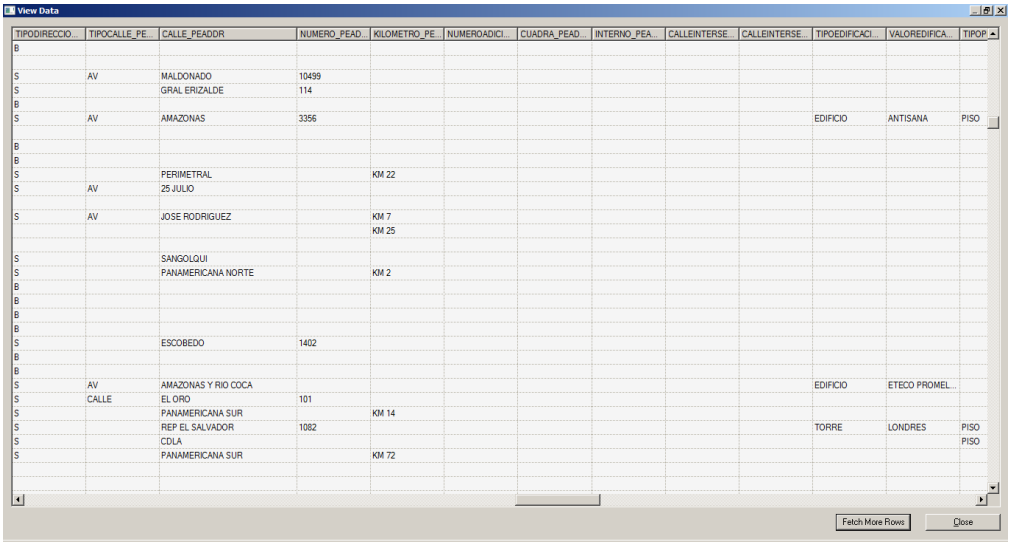


Figura 67: Estandarización Paso – 12

4. Crear un proceso para crear los reportes QA(Quality Assessment)

La estandarización, así como la calidad de datos, es un proceso cíclico en el cual se aplican reglas y se evalúan los resultados. Si estos resultados no son los esperados se requiere redefinir y mejorar las reglas. Para el caso de estudio se va a desarrollar un proceso de Quality Assessment(QA) para evaluar el cumplimiento de las reglas.

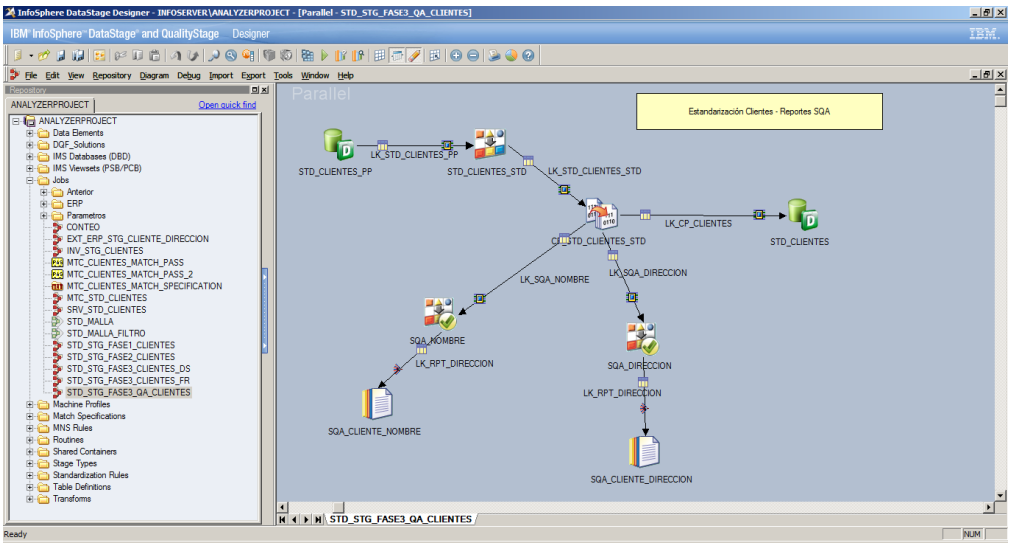


Figura 68: Estandarización Paso – 13

Establezca la configuración para obtener los reportes de QA.

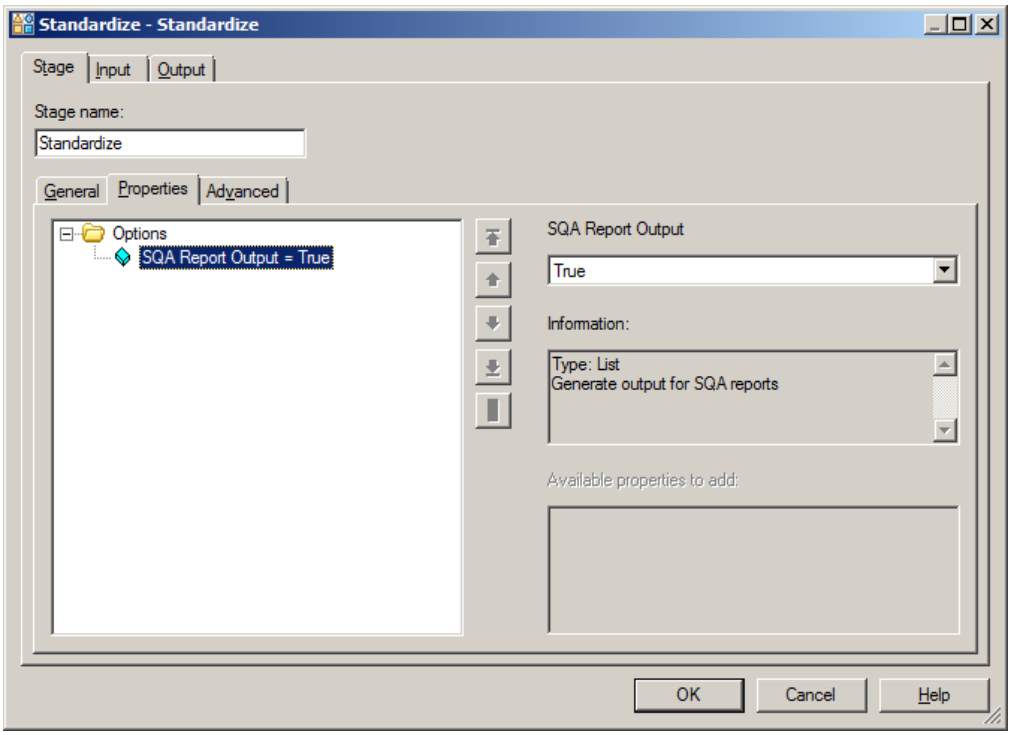


Figura 69: Estandarización Paso – 14

Configure las dos salidas de reporte para cada una de las columnas creadas en la estandarización.

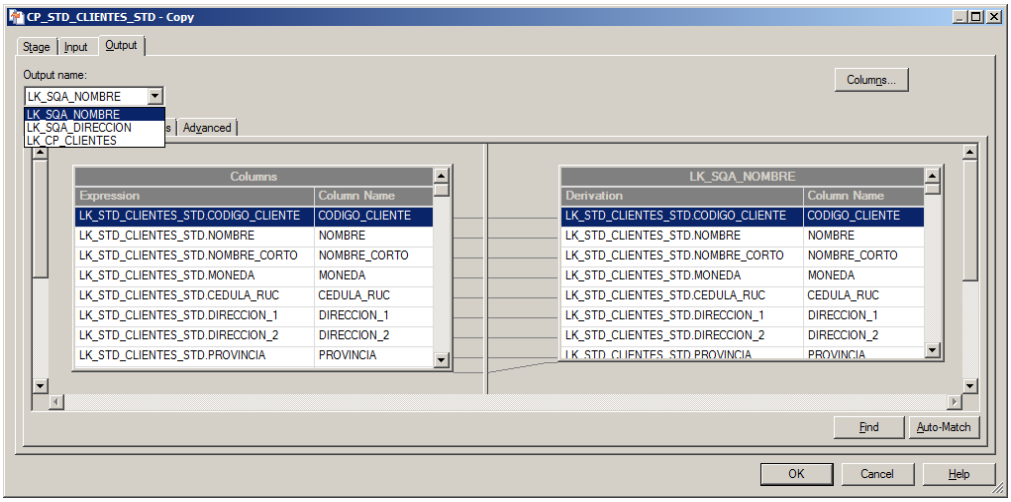


Figura 70: Estandarización Paso – 15

Configurar la opción para los datos que aplican a la regla para cada reporte.

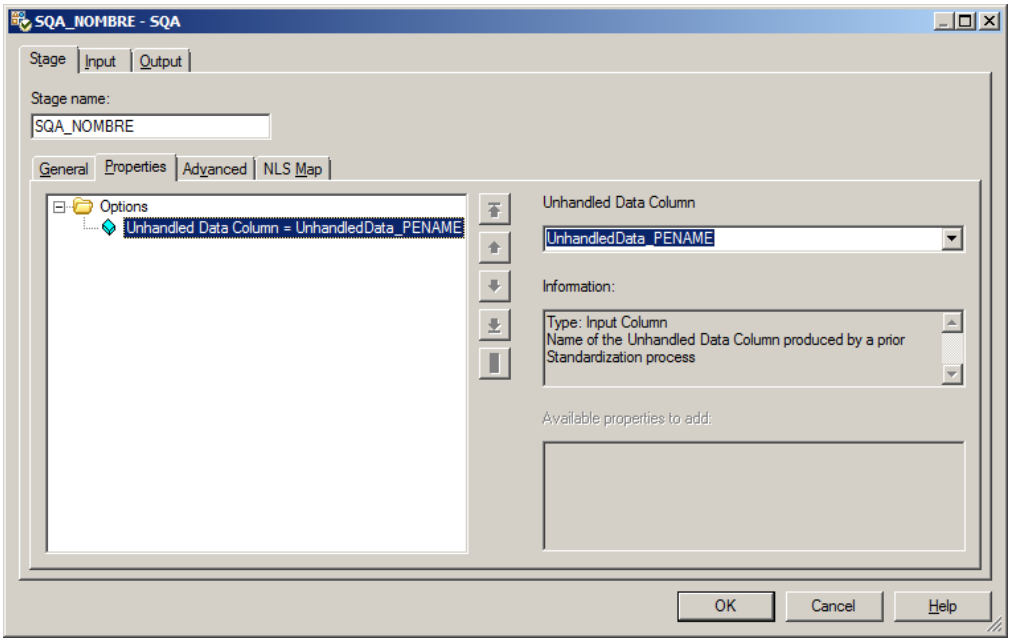


Figura 71: Estandarización Paso – 16

Mapear y asignar una ubicación al archivo de salida con la información de reporte de QA.

5. Crear los reportes QA(Quality Assessment)

Luego de ejecutar el proceso, abrir la consola Web de Infosphere y crear un nuevo reporte según se indica en la **Figura 72**.

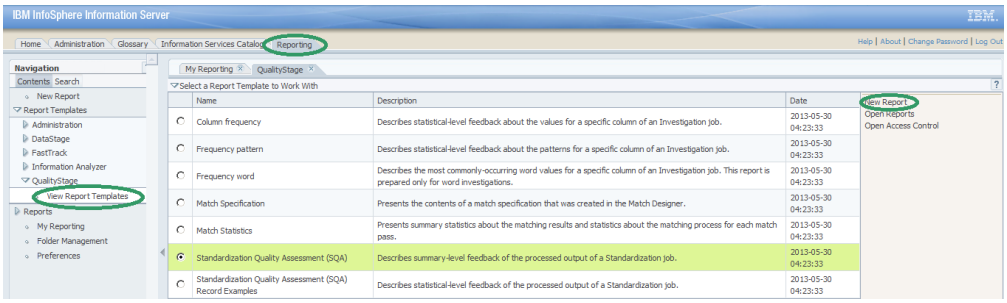


Figura 72: Estandarización Paso – 17

Establezca la siguiente configuración para crear el reporte del campo nombre en formato PDF:

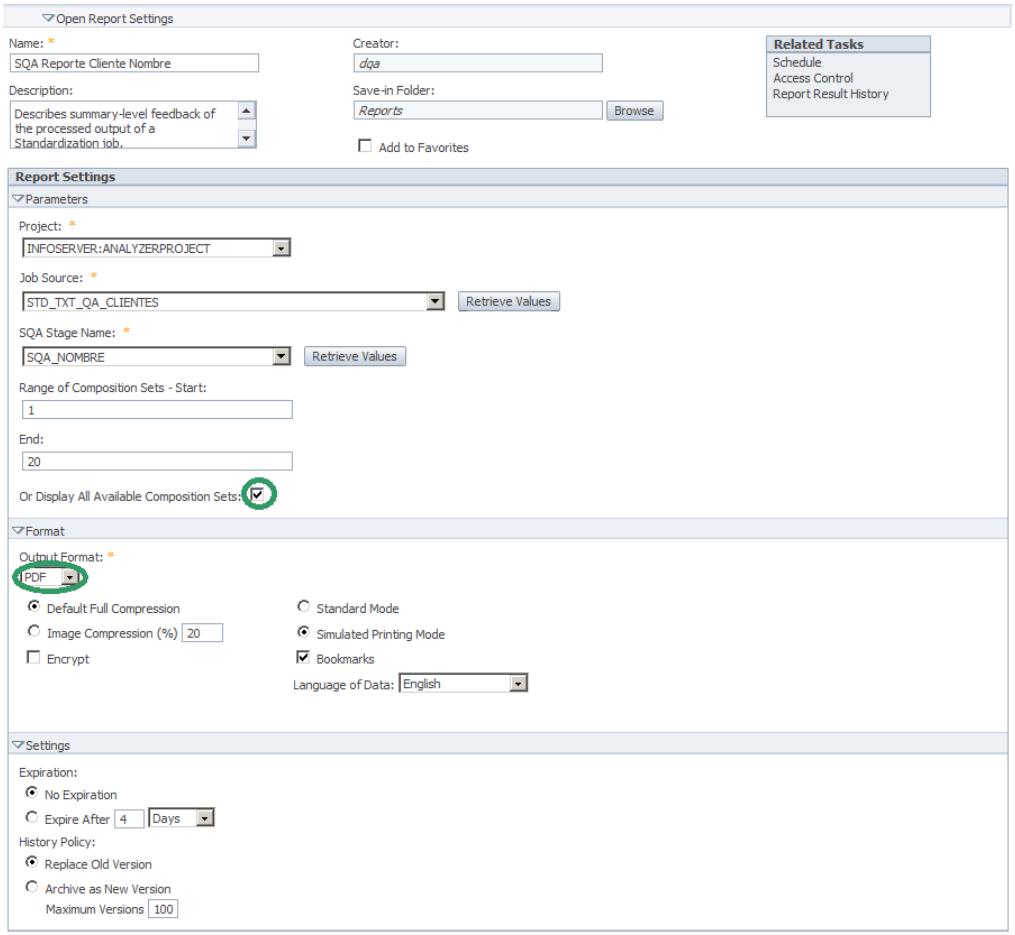


Figura 73: Estandarización Paso – 18

El reporte de estandarización del Nombre indica en primera instancia que existen datos que no están siendo estandarizados completamente, solo el 36.83% si lo están.

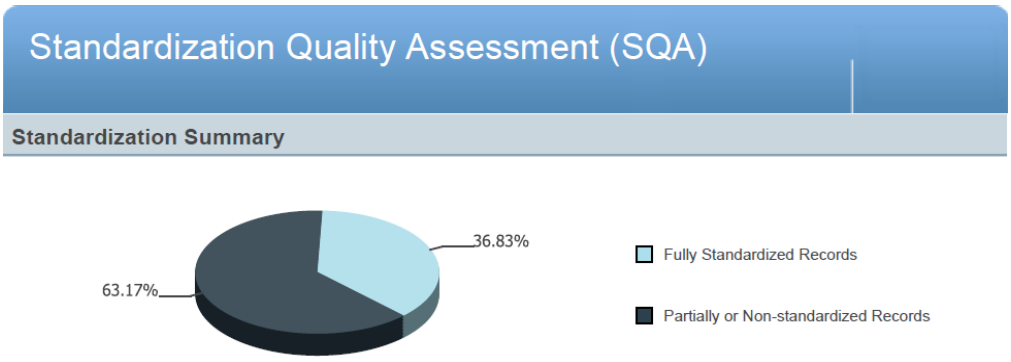


Figura 74: Estandarización Paso – 19

El reporte de estandarización de las direcciones indica un porcentaje similar de los datos completamente estandarizados: 35.9 %.

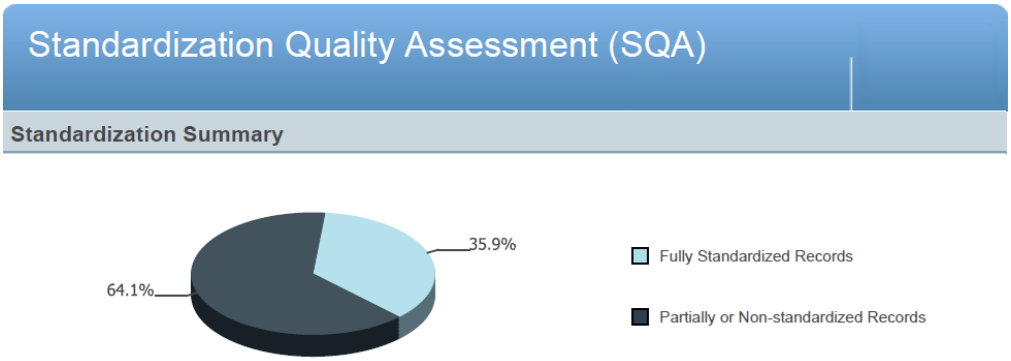


Figura 75: Estandarización Paso – 20

Los datos no estandarizados se pueden verificar con la opción UnhandledPattern que fue configurada en el proceso. El reporte generado nos da la posibilidad de rastrear los registros no estandarizados a través de la siguiente tabla indicada en la **Figura 76** donde observamos que en el conjunto 1 están los patrones no encontrados del dominio Nombre.

Standardization Quality Assessment (SQA)

Composition Sets — Displayed sets comprise 97.44% of the processed records

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
	63.17%	4.93%	4.39%	3.90%	3.35%	2.54%	2.27%	1.81%	1.75%	1.68%
TipoNombre		✓	✓	✓	✓	✓	✓	✓	✓	✓
CodigoGenero		✓							✓	
PrefijoNombre										
PrimerNombre		✓							✓	
SegundoNombre										
Apellido		✓	✓	✓	✓	✓	✓	✓	✓	✓
Nombre_Adicional										
SufijoNombre					✓		✓			✓
MatchPrimerNombre		✓							✓	
MatchPrimerNombreNYSIIS		✓							✓	
MatchPrimerNombreRVSNDX	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellido		✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellidoHashKey		✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellidoPackKey		✓	✓	✓	✓	✓	✓	✓	✓	✓
NumofMatchApellido	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellido1		✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellido2			✓	✓		✓	✓	✓	✓	✓
MatchApellido3			✓			✓		✓		✓
MatchApellido4						✓		✓		
MatchApellido5								✓		
MatchApellido1NYSIIS		✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellido1RVSNDX	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellido1SNDX	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellido2NYSIIS			✓	✓		✓	✓	✓	✓	✓
MatchApellido2RVSNDX	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MatchApellido2SNDX	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
UnhandledPattern	✓									
InputPattern	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ExceptionData										
UserOverrideFlag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Figura 76: Estandarización Paso – 21

Para conocer cuáles son los problemas con los conjuntos marcados, se genera el siguiente reporte:

QualityStage | My Reporting

Select a Report Template to Work With

	Name	Description	Date
○	Column frequency	Describes statistical-level feedback about the values for a specific column of an Investigation job.	2012-01-17 08:23:02
○	Frequency pattern	Describes statistical-level feedback about the patterns for a specific column of an Investigation job.	2012-01-17 08:23:02
○	Frequency word	Describes the most commonly-occurring word values for a specific column of an Investigation job. This report is prepared only for word investigations.	2012-01-17 08:23:02
○	Match Specification	Presents the contents of a match specification that was created in the Match Designer.	2012-01-17 08:23:02
○	Match Statistics	Presents summary statistics about the matching results and statistics about the matching process for each match pass.	2012-01-17 08:23:02
○	Standardization Quality Assessment (SQA)	Describes summary-level feedback of the processed output of a Standardization job.	2012-01-17 08:23:02
◉	Standardization Quality Assessment (SQA) Record Examples	Describes statistical-level feedback of the processed output of a Standardization job.	2012-01-17 08:23:02

New Report
Open Reports
Open Access Control

Figura 77: Estandarización Paso – 22

Configurar el reporte tal como indica la siguiente figura.

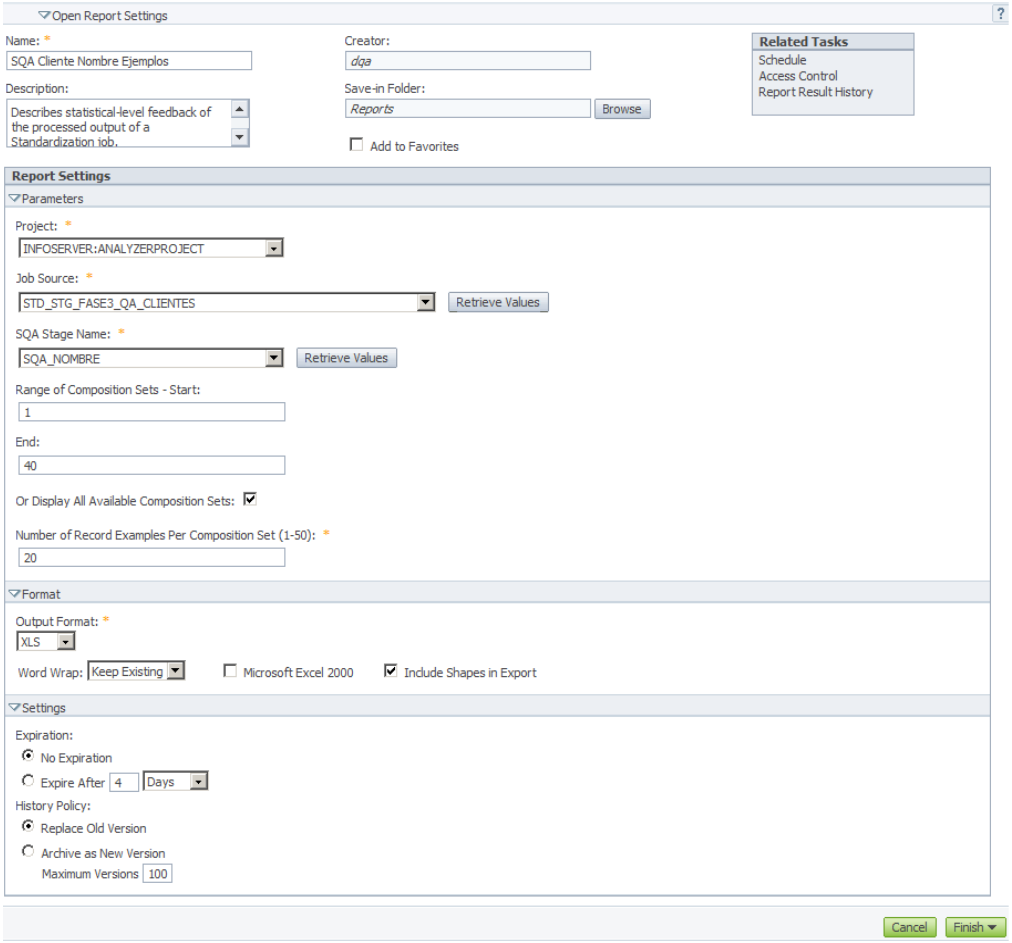
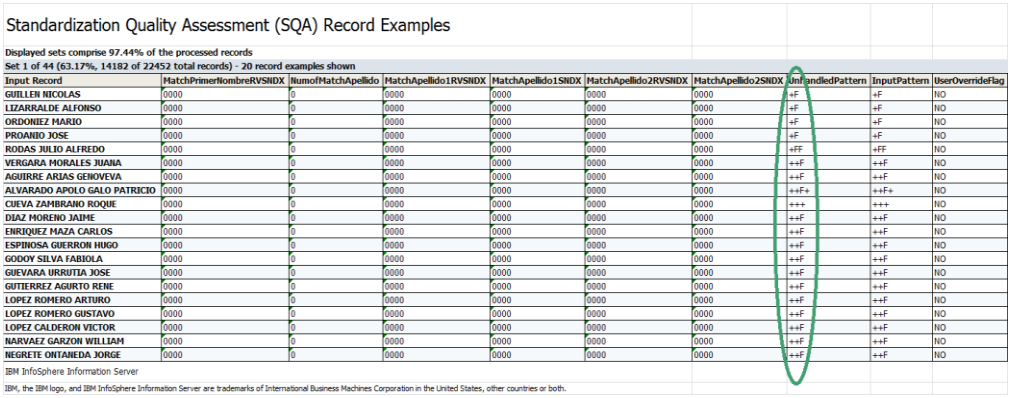


Figura 78: Estandarización Paso – 23

El reporte se genera en formato Excel. Para el conjunto 2 se puede observar que los datos no son estandarizados porque varios patrones no fueron reconocidos y no existen reglas que los soporten.



Standardization Quality Assessment (SQA) Record Examples

Displayed sets comprise 97.44% of the processed records

Set 1 of 44 (63.17%, 14182 of 22452 total records) - 20 record examples shown

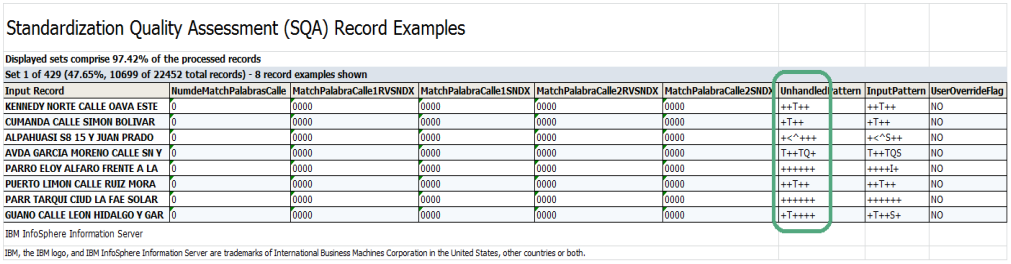
Input Record	MatchPrimerNombreRVSNDX	NumofMatchApellido	MatchApellido1RVSNDX	MatchApellido1SNDX	MatchApellido2RVSNDX	MatchApellido2SNDX	UnhandledPattern	InputPattern	UserOverrideFlag
GUILLEN NICOLAS	0000	0	0000	0000	0000	0000	+F	+F	NO
LIZARRALDE ALFONSO	0000	0	0000	0000	0000	0000	+F	+F	NO
ORDONIEZ MARIO	0000	0	0000	0000	0000	0000	+F	+F	NO
PROANIO JOSE	0000	0	0000	0000	0000	0000	+F	+F	NO
RODAS JULIO ALFREDO	0000	0	0000	0000	0000	0000	+FF	+FF	NO
VERGARA MORALES JUANA	0000	0	0000	0000	0000	0000	++F	++F	NO
AGUIRRE ARIAS GENOVEVA	0000	0	0000	0000	0000	0000	++F	++F	NO
ALVARADO APOLO GALO PATRICIO	0000	0	0000	0000	0000	0000	++F+	++F+	NO
CUEVA ZAMBRANO ROQUE	0000	0	0000	0000	0000	0000	+++	+++	NO
DIAZ MORENO JAIME	0000	0	0000	0000	0000	0000	++F	++F	NO
ENRIQUEZ MAZA CARLOS	0000	0	0000	0000	0000	0000	++F	++F	NO
ESPINOSA GUERRON HUGO	0000	0	0000	0000	0000	0000	++F	++F	NO
GODOY SILVA FABIOLA	0000	0	0000	0000	0000	0000	++F	++F	NO
GUEVARA URRUTIA JOSE	0000	0	0000	0000	0000	0000	++F	++F	NO
GUTIERREZ AGURTO RENE	0000	0	0000	0000	0000	0000	++F	++F	NO
LOPEZ ROMERO ARTURO	0000	0	0000	0000	0000	0000	++F	++F	NO
LOPEZ ROMERO GUSTAVO	0000	0	0000	0000	0000	0000	++F	++F	NO
LOPEZ CALDERON VICTOR	0000	0	0000	0000	0000	0000	++F	++F	NO
NARVAEZ GARZON WILLIAM	0000	0	0000	0000	0000	0000	++F	++F	NO
NEGRETE ONTANEDA JORGE	0000	0	0000	0000	0000	0000	++F	++F	NO

IBM InfoSphere Information Server

IBM, the IBM logo, and IBM InfoSphere Information Server are trademarks of International Business Machines Corporation in the United States, other countries or both.

Figura 79: Estandarización Paso – 24

Al realizar el mismo análisis para el dominio Direcciones, se encontraron que los siguientes patrones no están siendo reconocidos.



Standardization Quality Assessment (SQA) Record Examples

Displayed sets comprise 97.42% of the processed records

Set 1 of 429 (47.65%, 10699 of 22452 total records) - 8 record examples shown

Input Record	NumdeMatchPalabrasCalle	MatchPalabraCalle1RVSNDX	MatchPalabraCalle1SNDX	MatchPalabraCalle2RVSNDX	MatchPalabraCalle2SNDX	UnhandledPattern	InputPattern	UserOverrideFlag
KENNEDY NORTE CALLE OAVA ESTE	0	0000	0000	0000	0000	++T++	++T++	NO
CUMANDA CALLE SIMON BOLIVAR	0	0000	0000	0000	0000	+T++	+T++	NO
ALPAHUASI S8 15 Y JUAN PRADO	0	0000	0000	0000	0000	+<^+++	+<^S++	NO
AVDA GARCIA MORENO CALLE SN Y	0	0000	0000	0000	0000	T++TQ+	T++TQS	NO
PARRO ELOY ALFARO FRENTE A LA	0	0000	0000	0000	0000	++++++	+++I+	NO
PUERTO LIMON CALLE RUIZ MORA	0	0000	0000	0000	0000	++T++	++T++	NO
PARR TARQUI CIUD LA FAE SOLAR	0	0000	0000	0000	0000	++++++	++++++	NO
GUANO CALLE LEON HIDALGO Y GAR	0	0000	0000	0000	0000	+T++++	+T++S+	NO

IBM InfoSphere Information Server

Figura 80: Estandarización Paso – 25

6. Modificación de reglas de estandarización

Esta ha sido la primera ejecución de la estandarización de los datos, con estos resultados se deben modificar las reglas e incluir los patrones que no se consideraron al inicio. Primero se editará la regla con respecto al nombre, se van a incluir los patrones que se detectaron como faltantes, tanto de tipo Input como Unhandled. Seleccionar la regla desde el repositorio y abrimos sus propiedades.

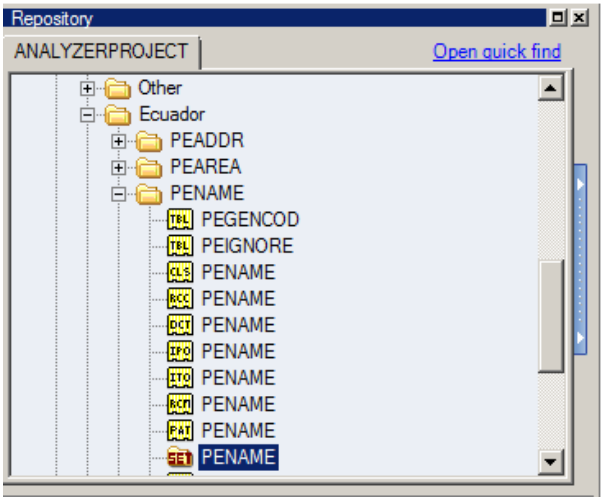


Figura 81: Estandarización Paso – 25

Seleccionar la opción Overrides para editar e incluir los nuevos patrones encontrados.

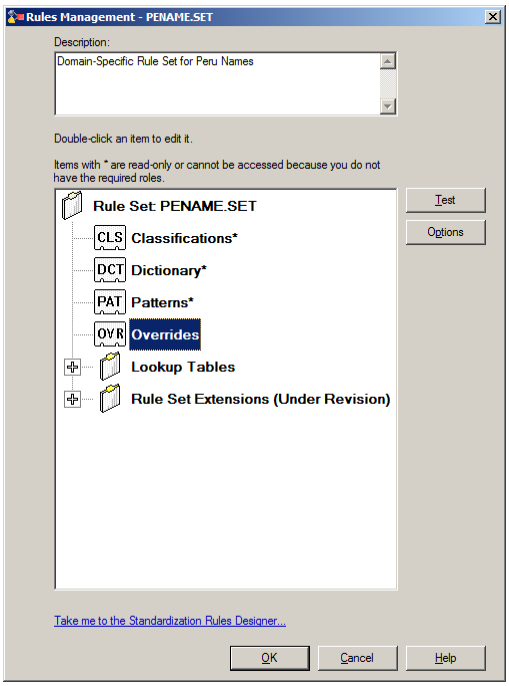


Figura 82: Estandarización Paso – 26

Dependiendo del tipo de patrón seleccionar la pestaña correspondiente. En este caso se incluye un patrón tipo Input.

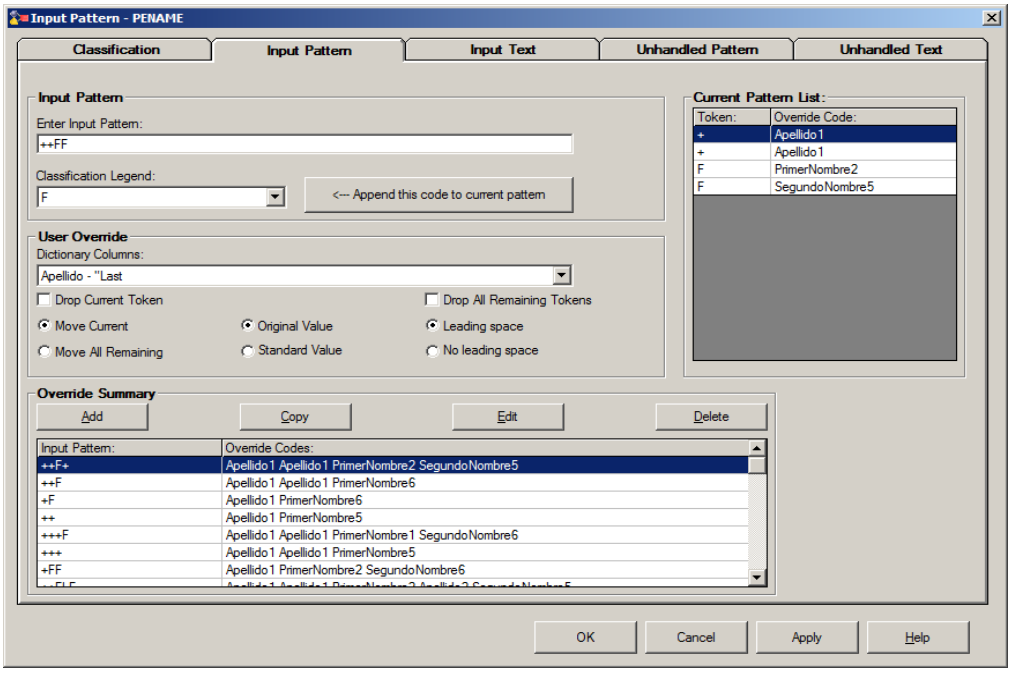


Figura 83: Estandarización Paso – 27

PrimerNombre contienen el número 2 que representa un valor estándar (1 representa valor Original), SegundoNombre contiene el 5 que significa mover el resto, es decir todo lo que se detecte después se va a ingresar como segundo nombre. Presionar añadir. Para el Nombre se deben incluir los siguientes Patrones.

Input Pattern:	Override Codes:	Other (
++FF	Apellido1 Apellido1 PrimerNombre2 SegundoNombre5	
++F+	Apellido1 Apellido1 PrimerNombre2 SegundoNombre5	
++F	Apellido1 Apellido1 PrimerNombre6	
+F	Apellido1 PrimerNombre6	
++	Apellido1 PrimerNombre5	
+++F	Apellido1 Apellido1 PrimerNombre1 SegundoNombre6	
+++	Apellido1 Apellido1 PrimerNombre5	
+FF	Apellido1 PrimerNombre2 SegundoNombre6	
++FLF	Apellido1 Apellido1 PrimerNombre2 Apellido2 SegundoNombre5	
+	Apellido5	
++++	Apellido1 Apellido1 PrimerNombre1 SegundoNombre1	
F+FF	Apellido1 Apellido1 PrimerNombre2 SegundoNombre6	
+++FF	Apellido1 Apellido1 Apellido1 Apellido2 Apellido6	

Figura 84: Estandarización Paso – 28

Para las Direcciones incluir los siguientes patrones en unhandled Pattern:

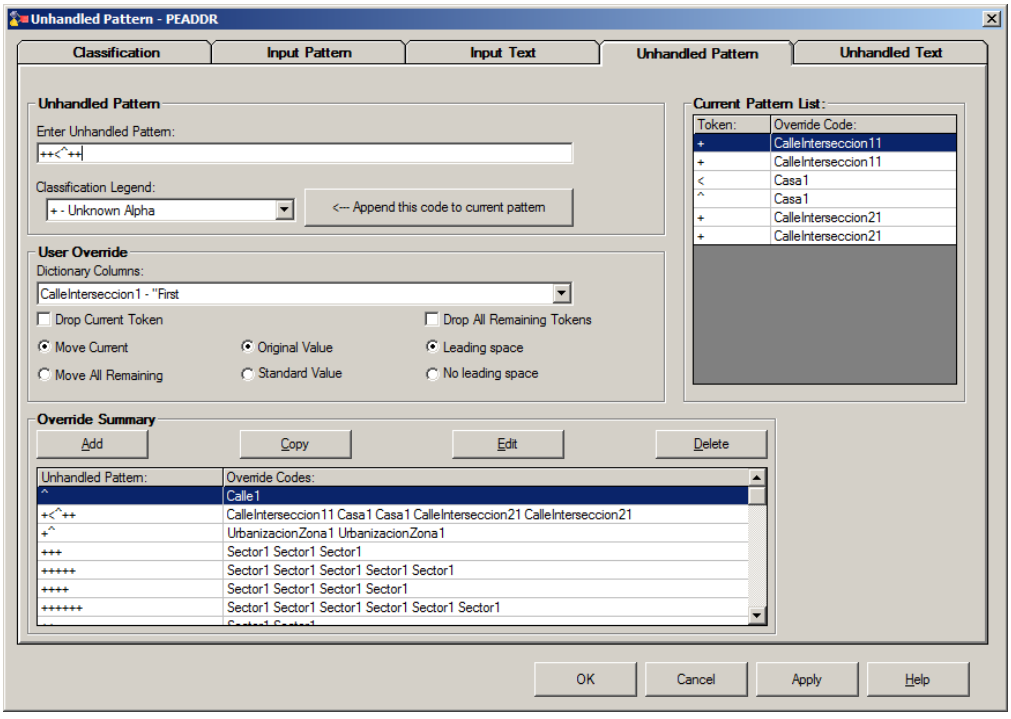


Figura 85: Estandarización Paso – 29

Este es el listado de patrones:

Unhandled Pattern:	Override Codes:
^	Calle1
++<^++	CalleInterseccion11 CalleInterseccion11 Casa1 Casa1 CalleInterseccion21 CalleInterseccion21
+<^++	CalleInterseccion11 Casa1 Casa1 CalleInterseccion21 CalleInterseccion21
+^	UrbanizacionZona1 UrbanizacionZona1
+++	Sector1 Sector1 Sector1
+++++	Sector1 Sector1 Sector1 Sector1 Sector1
++++	Sector1 Sector1 Sector1 Sector1
++++++	Sector1 Sector1 Sector1 Sector1 Sector1 Sector1
++	Sector1 Sector1
++<^+++	CalleInterseccion11 CalleInterseccion11 Casa1 Casa1 CalleInterseccion21 CalleInterseccion21
+	UrbanizacionZona1
++^	Calle1 Calle1 Casa1
+<^+++	Calle1 Casa1 Casa1 Calle1 Calle1 Calle1
<	Casa1
+++^	Calle1 Calle1 Calle1 Casa1
+++<^++	Calle1 Calle1 Calle1 Casa1 Casa1 Calle1 Calle1
T++++	TipoCalle1 Calle1 Calle1 Calle1 Calle1
T+++	TipoCalle1 Calle1 Calle1 Calle1
T++++++	TipoCalle1 Calle1 Calle1 Calle1 Calle1 Calle1
+^++	Calle1 Casa1 Calle1 Calle1
+T+	Calle1 TipoCalle1 Calle1
++<^	Calle1 Calle1 Casa1 Casa1
+^+	Calle1 Casa1 Calle1
++S	Calle1 Calle1 TipoCalle1
++<^+	Calle1 Calle1 Casa1 Casa1 Calle1

Figura 86: Estandarización Paso – 30

Al finalizar los cambios para cada regla ejecutar la opción Provision All, que aparece al dar clic derecho sobre la regla.

Una vez editadas las reglas procedemos a ejecutar nuevamente el proceso para los reportes QA. Estos reportes presentaron los siguientes resultados:

Se observa que para el nombre el resultado del cambio produjo el 93.15% de datos estandarizados, cuando al inicio fue de 36.83%.

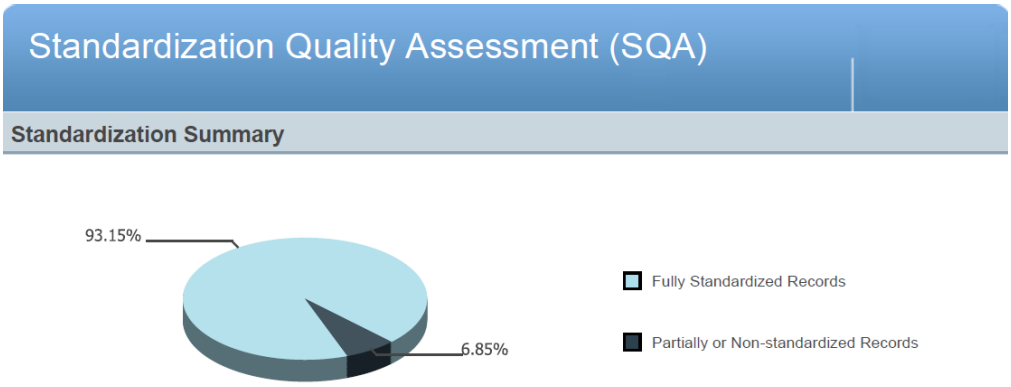


Figura 87: Estandarización Paso – 31

Para las direcciones el resultado del cambio produjo el 75.36% de datos estandarizados, el cual es un valor que se puede mejorar pero se tiene que hacer un trabajo desde la fuente de datos para alcanzar el resultado requerido, sin embargo se pudo mejorar el porcentaje inicial de 35.9% de estandarización que se presentó al inicio.

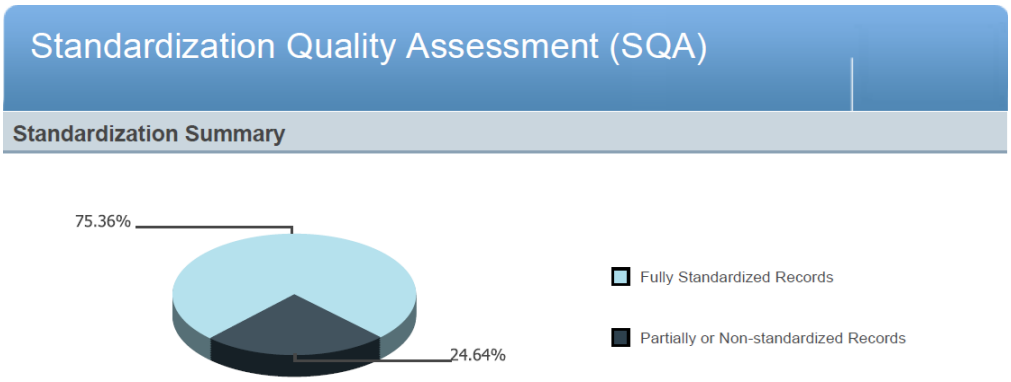


Figura 88: Estandarización Paso – 31

Con la segunda ejecución el resultado de la estandarización es más cercano a lo que desea lograr, esto permite seguir con el siguiente paso la eliminación de duplicados o etapa de Coincidencia *(Match)*.

3.2.2.6 Coincidencia

La etapa de coincidencia o matching permite identificar registros cuyos datos representan la misma información pero con códigos distintos, es decir encontrar los duplicados. En esta etapa se procede con el desarrollo de una especificación de coincidencias y la creación de un proceso que permita eliminar los duplicados utilizando la especificación.

1. **Crear un proceso para cargar los datos estandarizados a un DataSet.**

   Establecer el siguiente proceso para cargar los datos a un DataSet de DataStage,

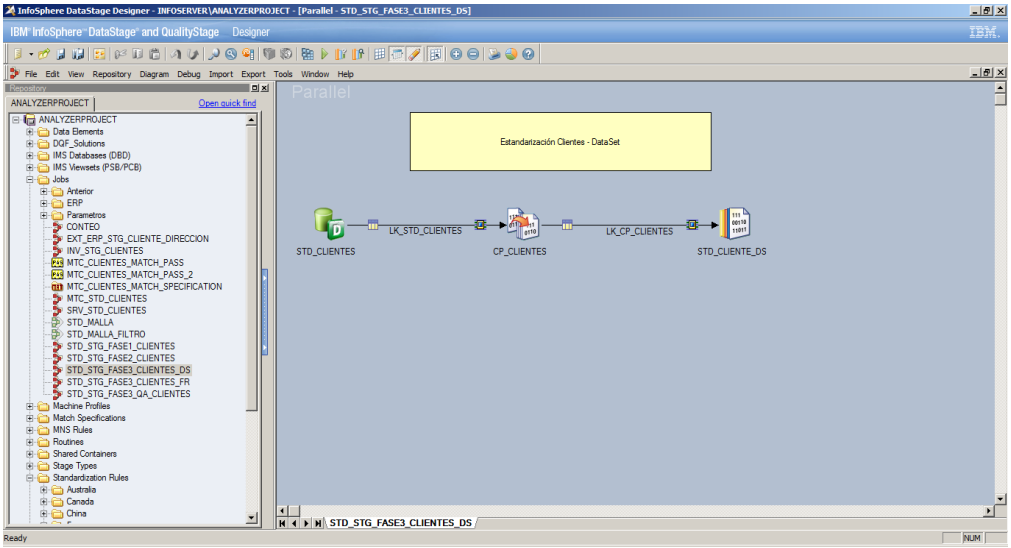


Figura 89: Coincidencia Paso – 1

2. **Crear un proceso para el análisis de frecuencias de los datos.**

   Establecer el siguiente proceso para generar las frecuencias de los datos, utilizando el DataSet creado en el paso anterior.

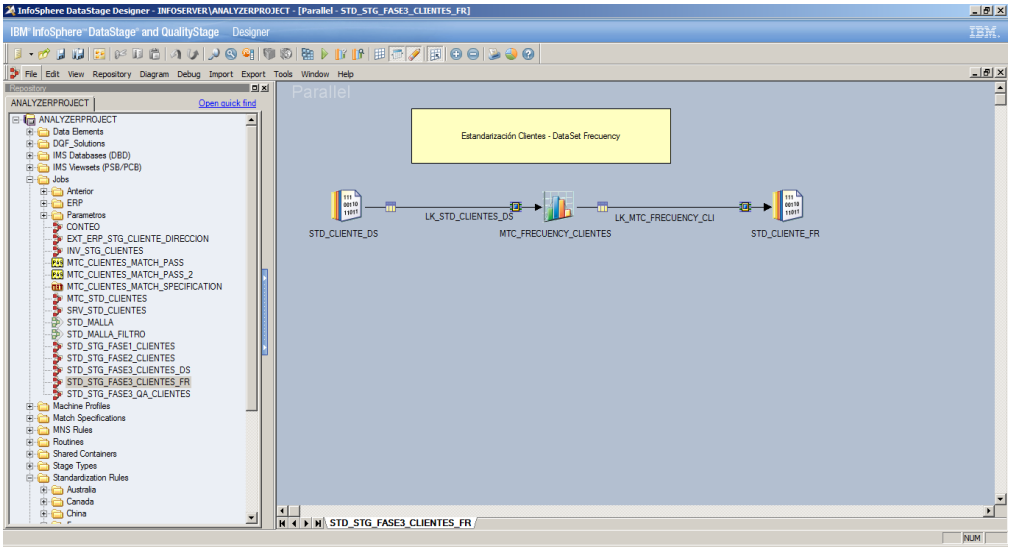


Figura 90: Coincidencia Paso – 2

Editar el stage Frecuency para configurar en función de la siguiente figura.

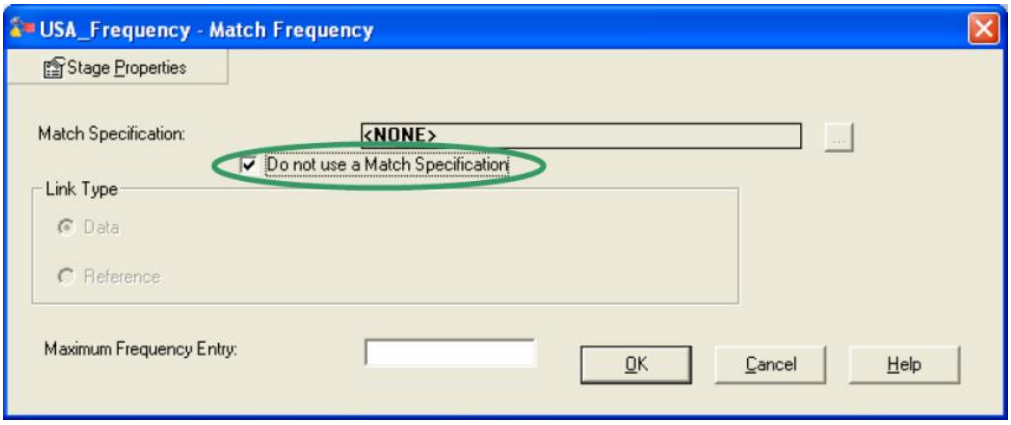


Figura 91: Coincidencia Paso – 3

Estos datos van a servir de insumo para el proceso de coincidencias que se explica a continuación.

3. Crear la especificación de coincidencias

Para crear la especificación de coincidencias se necesita realizar la siguiente configuración:

Crear una nueva especificación de coincidencias

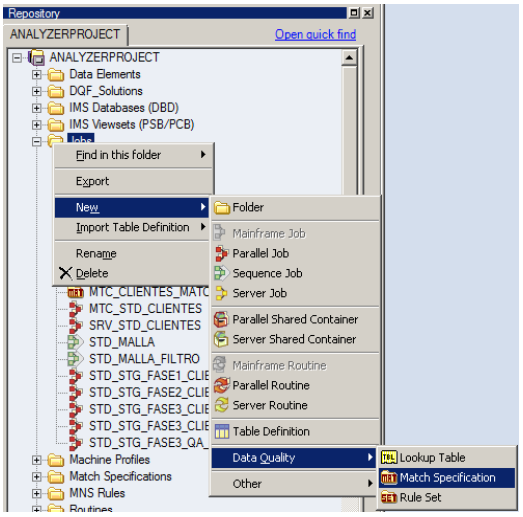


Figura 92: Coincidencia Paso – 4

Para facilitar se utiliza la herramienta Match Designer:

Figura 93: Coincidencia Paso – 5

Seleccione One-Source Dependant para agrupar duplicados:

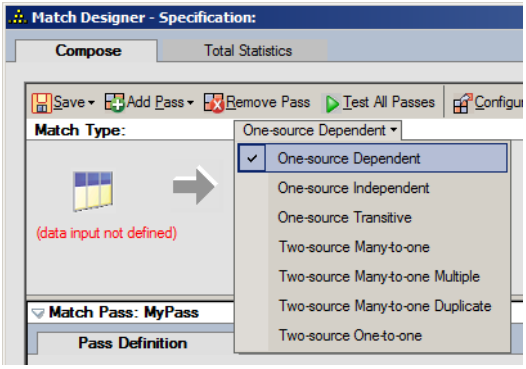


Figura 94: Coincidencia Paso – 6

En base a la estructura del DataSet creado cargue las columnas como entrada para la especificación.

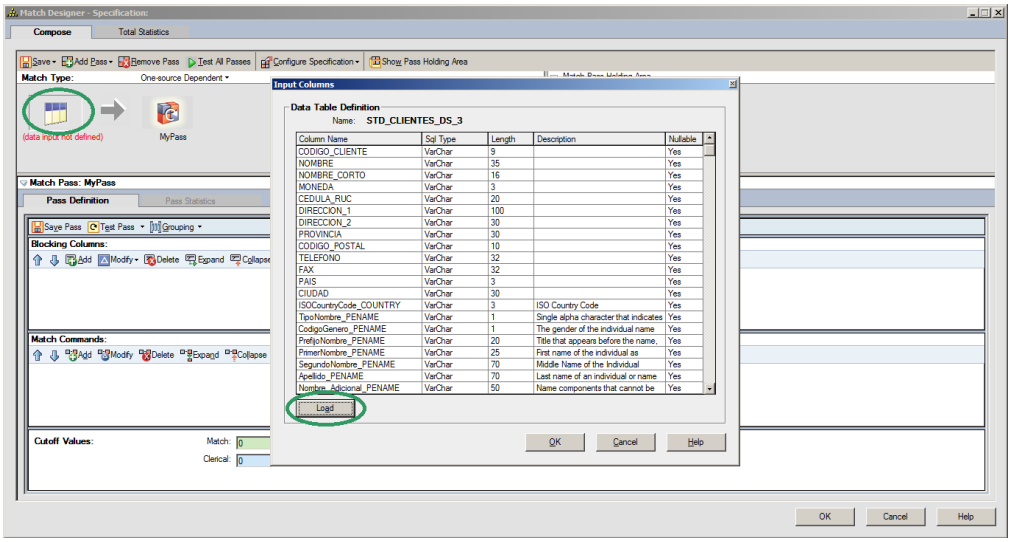


Figura 95: Coincidencia Paso – 7

Guardar la especificación y cambiar los nombre de la entrada y la primera pasada como indica la siguiente figura. Además seleccionar la opción Test Environment para configurar las fuentes de entrada de la especificación.

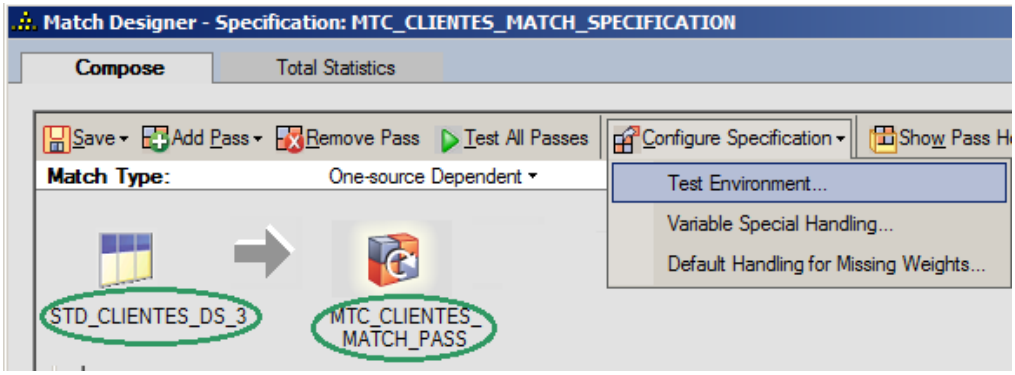


Figura 96: Coincidencia Paso – 8

La siguiente figura indica como configurar el ambiente:

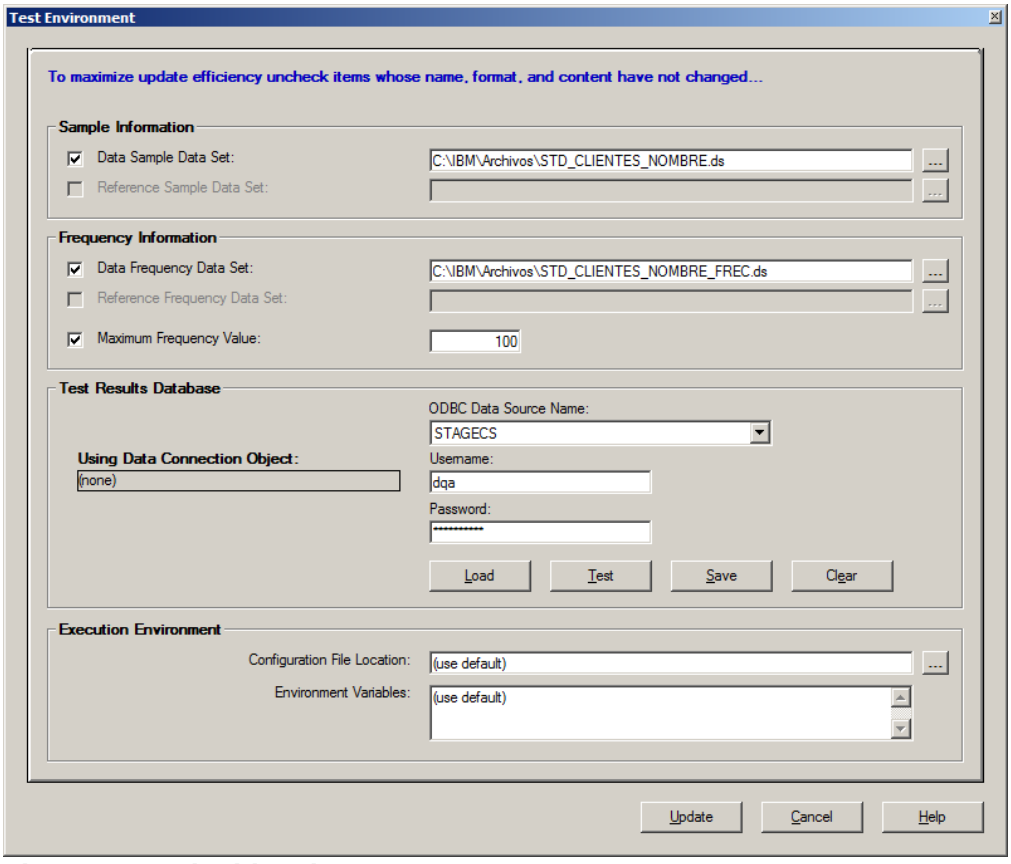


Figura 97: Coincidencia Paso – 9

Configurar la primera pasada, asignar los campos de dirección que se indican en la siguiente figura para agrupar la información en la sección Blocking Columns. Esto nos permite agrupar posibles coincidencias en función de la dirección en la entidad.

En la columna Record Type pueden existir 3 valores:

1. **MP**: Match - Identifica el registro base

2. **DA**: Duplicate – Señala los registros que tienen un peso que lo identifica como duplicado.

3. **CP**: Clerical – Señala los regristros que no tienen el peso necesario para ser duplicados pero necesitan revisión.

También se establece un rango en la sección Cutoff Values, para asegurar la identificación de los duplicados por dirección. Match: 8, Clerical: 7

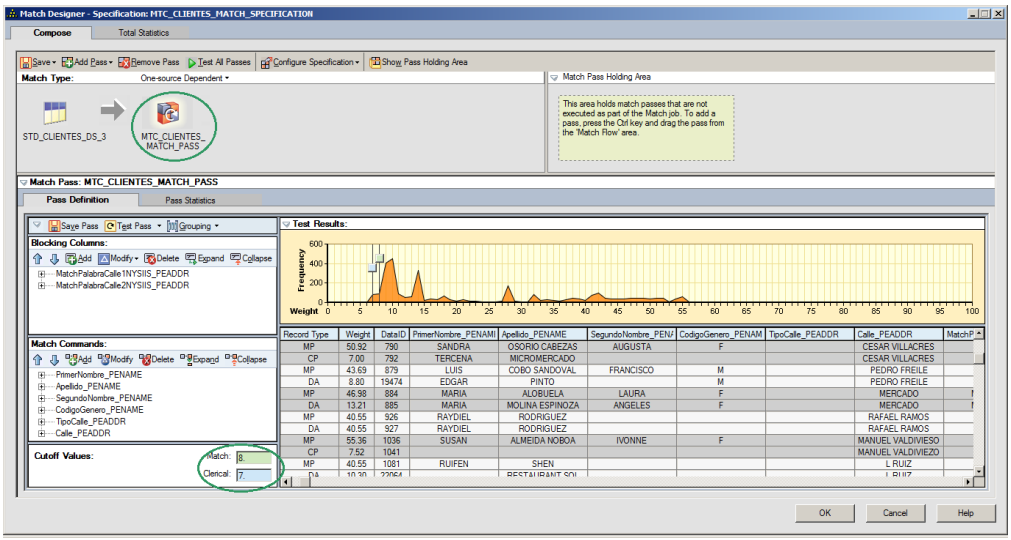


Figura 98: Coincidencia Paso – 10

Con la primera pasada la herramienta no arroja la siguiente estadística que nos indica que el 8% de registro puede ser que sea duplicado. Por lo que se configura una segunda pasada sub-agrupando los datos bajo otro criterio.

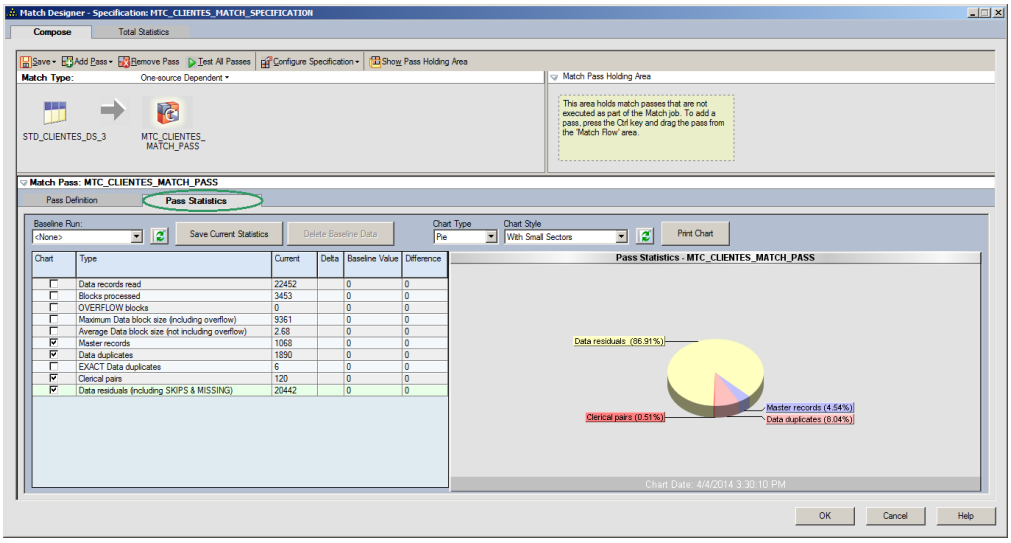


Figura 99: Coincidencia Paso – 11

Para configurar la segunda pasada se debe asignar el campo Apellido que se indican en la siguiente figura para agrupar la información en la sección Blocking Columns. Luego de agrupar por dirección se produce otra agrupación que es por apellido para poder conocer cuáles son los clientes duplicados.

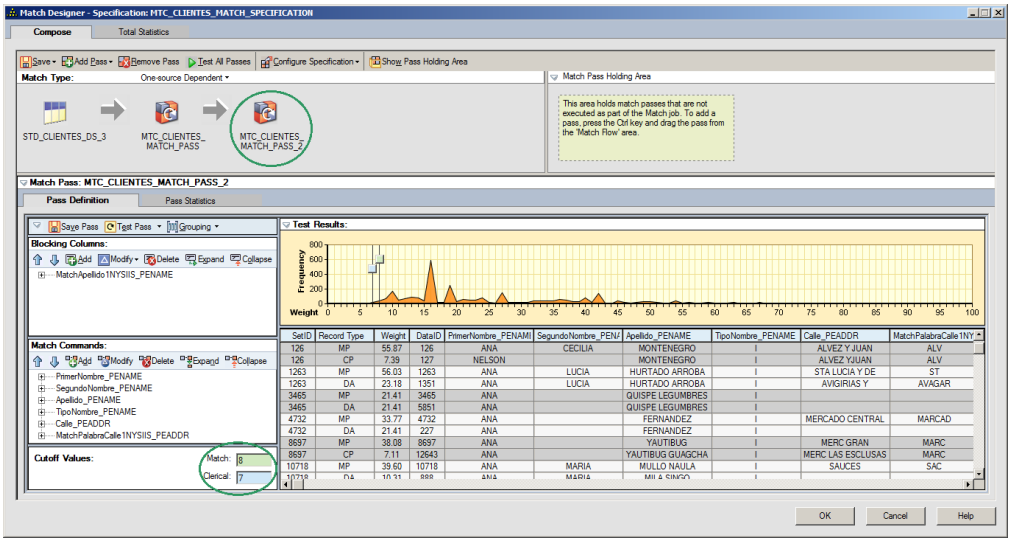


Figura 100: Coincidencia Paso – 12

Con la segunda pasada las estadísticas presentan que existe 7.8% de registros duplicados.

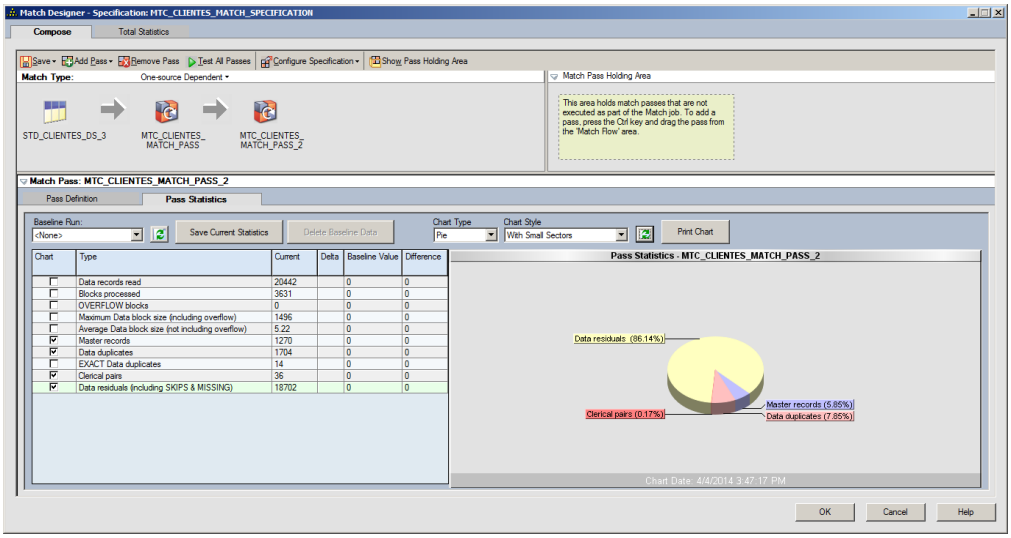


Figura 101: Coincidencia Paso – 13

A continuación se genera el proceso en IBM DataStage and QualityStage Designer para generar las coincidencias en una estructura que sirva para la Etapa de Supervivencia.

4. Crear proceso para eliminar duplicados

Con la ayuda de IBM DataStage and QualityStage Designer, realizar la configuración de la **Figura 102** para crear el proceso de Coincidencias, el cual no solo sirve de insumo para la etapa de supervivencia sino también para generar reportes de datos duplicados.

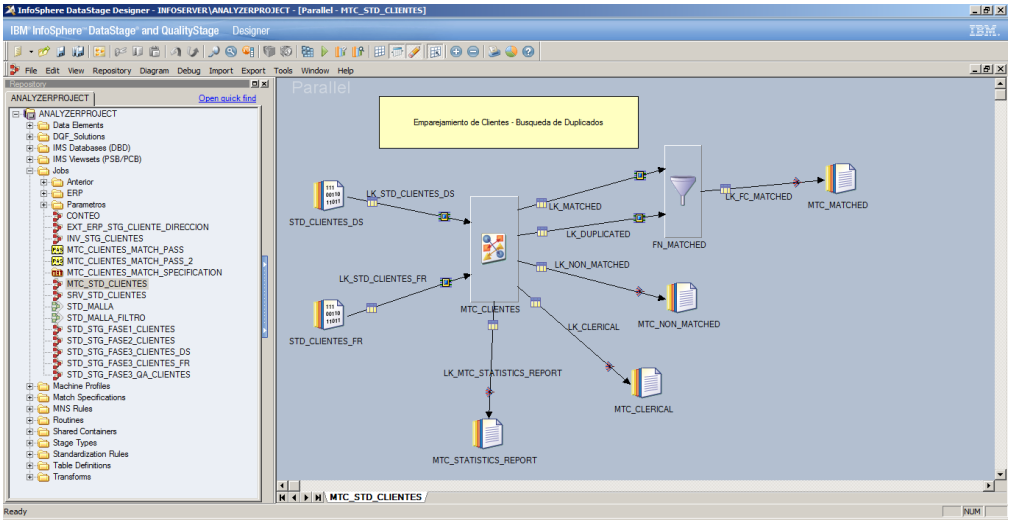


Figura 102: Coincidencia Paso – 14

Establecer la siguiente configuración en el stage de Matching, utilizando la especificación creada anteriormente

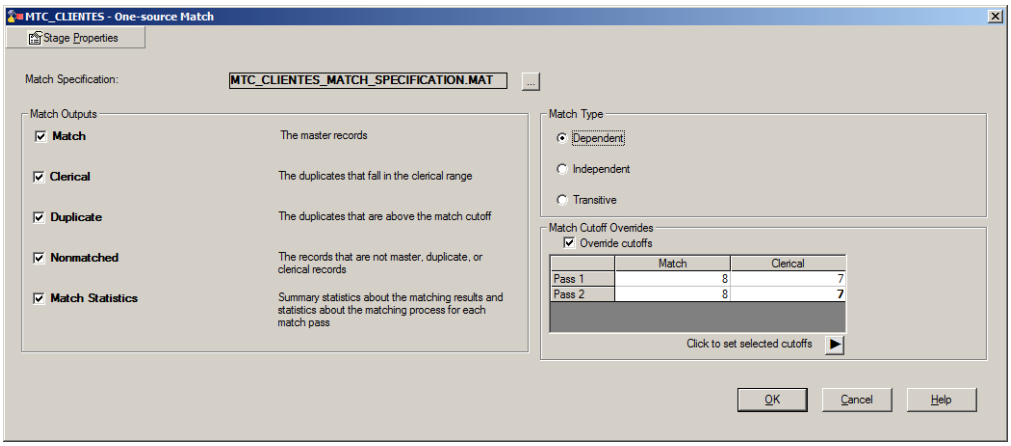


Figura 103: Coincidencia Paso – 15

Asegurarse de que las salidas están mapeadas adecuadamente.

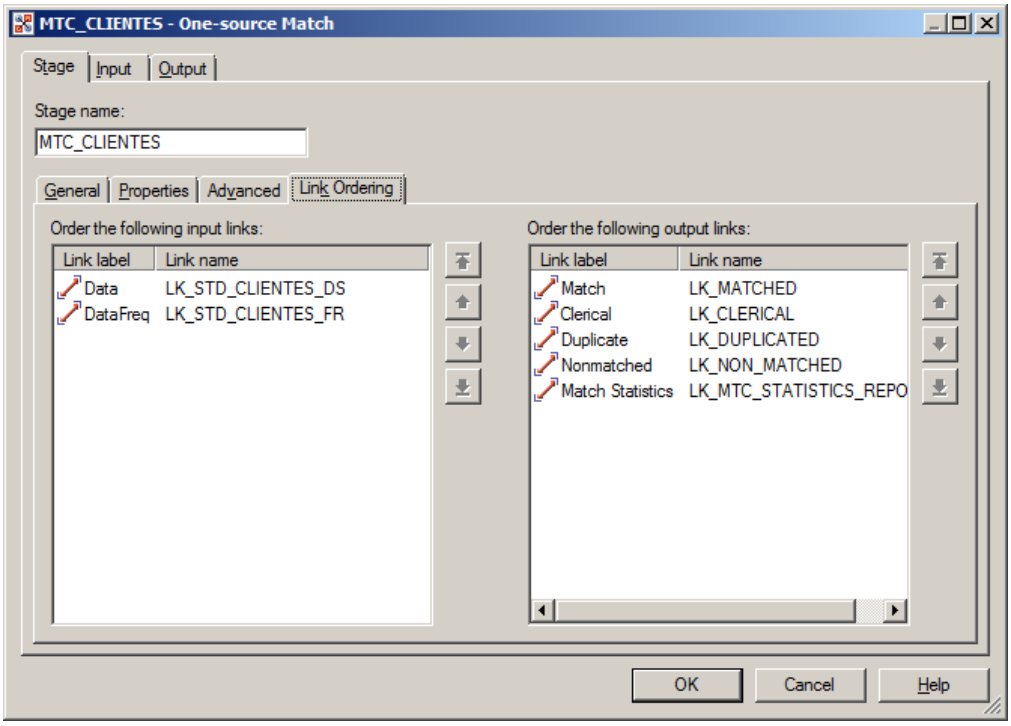


Figura 104: Coincidencia Paso – 16

Luego de establecer las entradas y las salidas de información ejecute el proceso para obtener los resultados.

5. Crear reportes de estadística de coincidencias

En la Consola Web genere un nuevo reporte en base al template Match Statistics con la siguiente configuración.

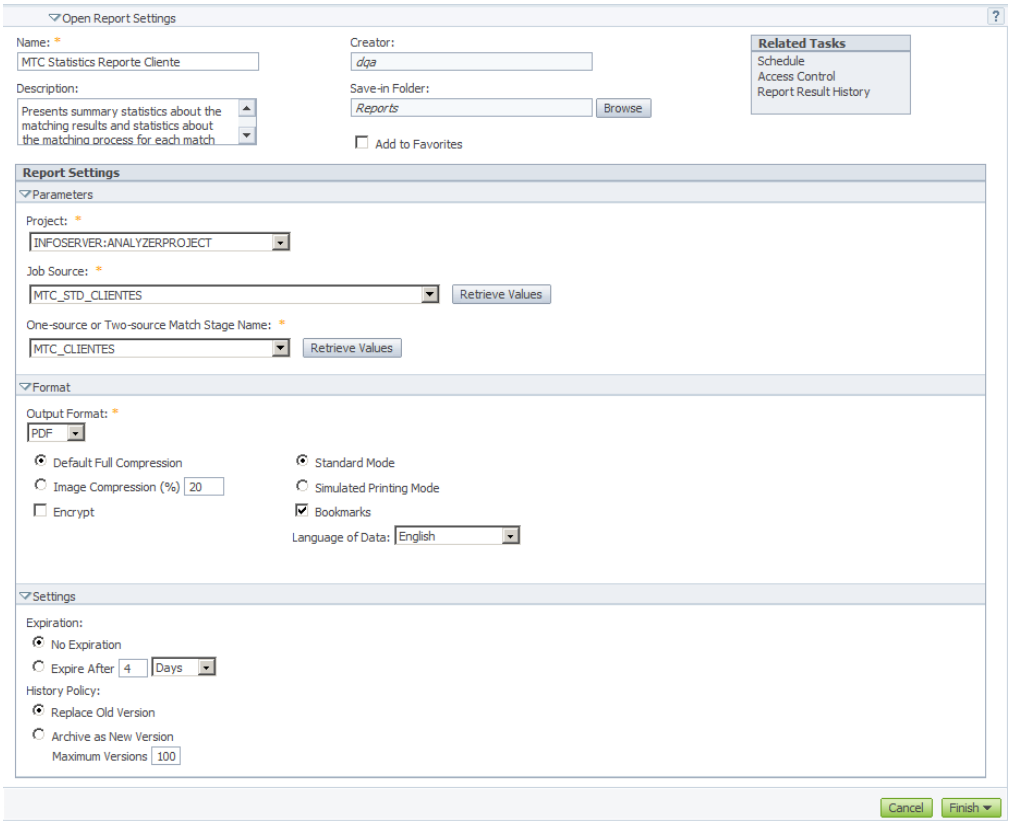
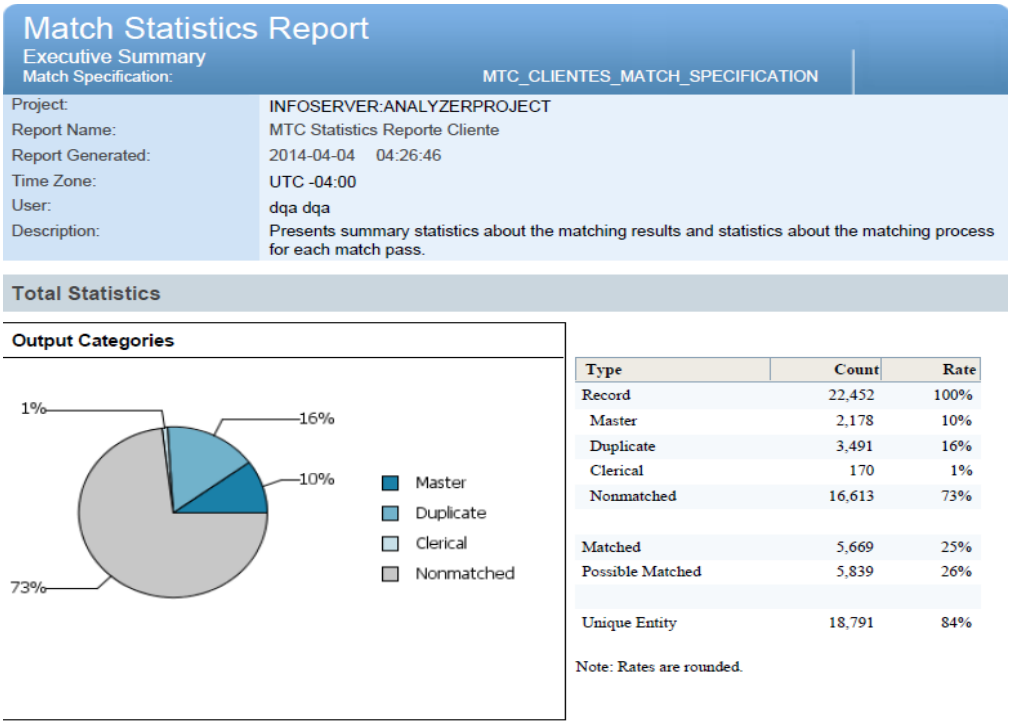


Figura 105: Coincidencia Paso – 17

Los resultados indican que se tiene 16% de duplicados en total con las 2 pasadas, la diferencia con los resultados anteriores es que antes aquí estamos viendo un resumen de todos los datos y antes eran solo una muestra de los datos.



Match Statistics Report

Executive Summary

Match Specification: MTC_CLIENTES_MATCH_SPECIFICATION

Project:	INFOSERVER:ANALYZERPROJECT
Report Name:	MTC Statistics Reporte Cliente
Report Generated:	2014-04-04 04:26:46
Time Zone:	UTC -04:00
User:	dqa dqa
Description:	Presents summary statistics about the matching results and statistics about the matching process for each match pass.

Total Statistics

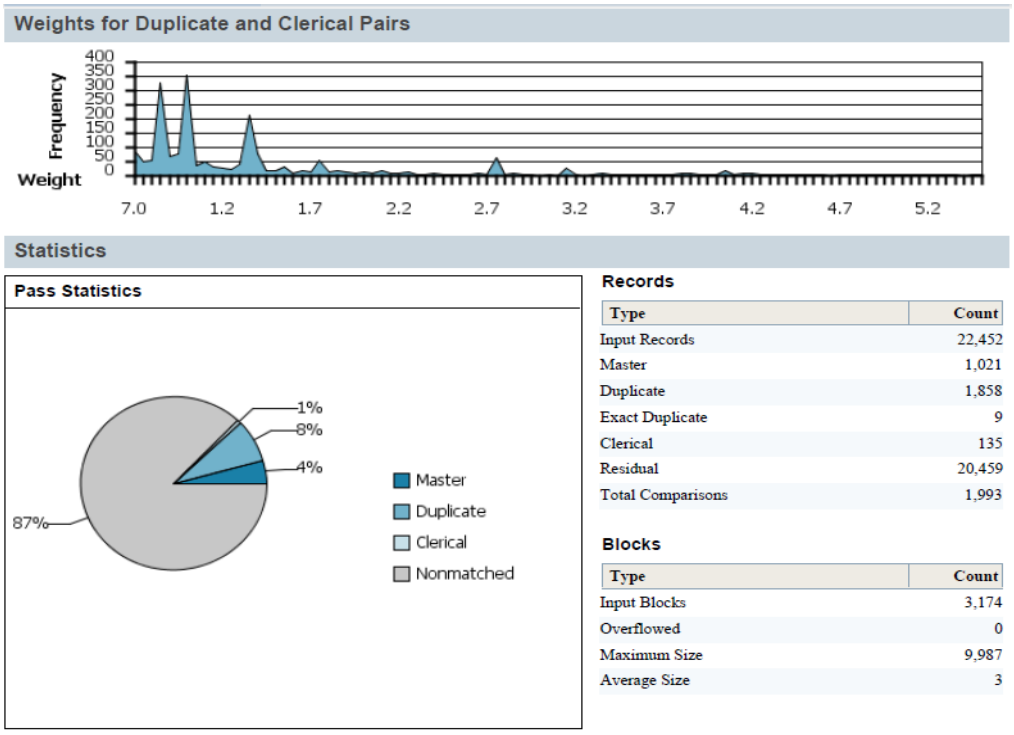
Output Categories

Type	Count	Rate
Record	22,452	100%
Master	2,178	10%
Duplicate	3,491	16%
Clerical	170	1%
Nonmatched	16,613	73%
Matched	5,669	25%
Possible Matched	5,839	26%
Unique Entity	18,791	84%

Note: Rates are rounded.

Figura 106: Coincidencia Paso – 18

Para la primera pasada se tiene el siguiente resultado que no difiere tanto del anterior, se tiene el 8% de duplicados, con 1858 registros del total.



Weights for Duplicate and Clerical Pairs

Statistics

Pass Statistics

Records

Type	Count
Input Records	22,452
Master	1,021
Duplicate	1,858
Exact Duplicate	9
Clerical	135
Residual	20,459
Total Comparisons	1,993

Blocks

Type	Count
Input Blocks	3,174
Overflowed	0
Maximum Size	9,987
Average Size	3

Figura 107: Coincidencia Paso – 19

Para la segunda pasada se tiene 8% de duplicados también pero con 1633 registros del total.

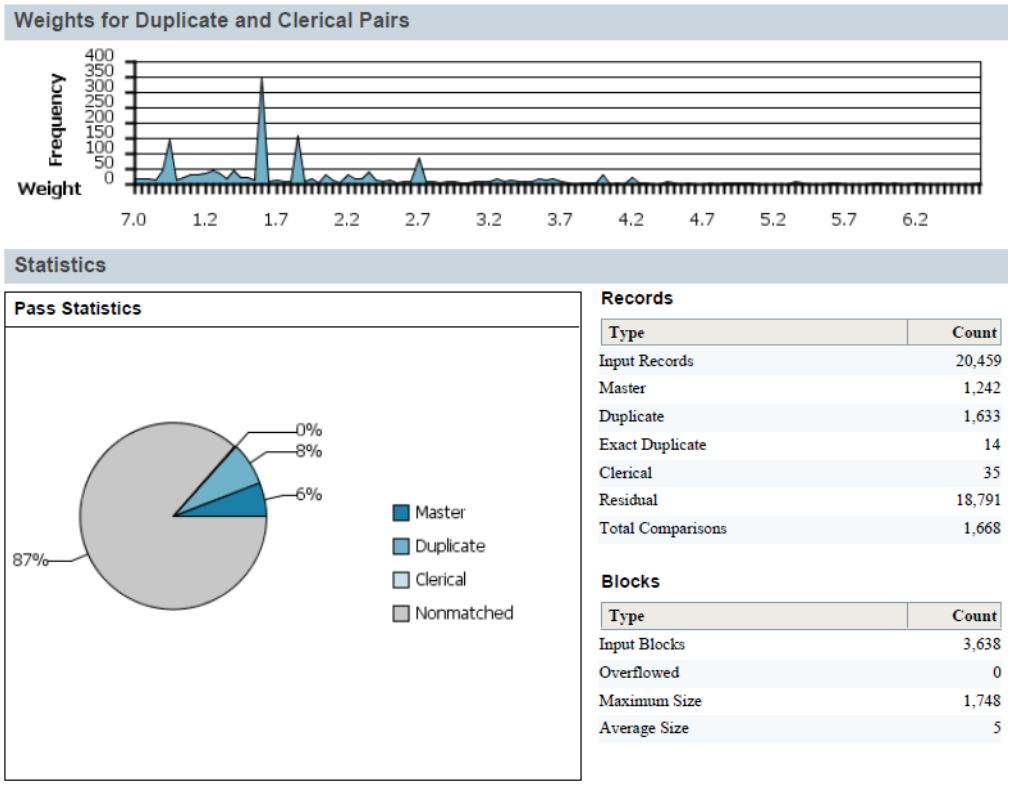


Figura 108: Coincidencia Paso – 20

Hasta este punto se ha identificado los posibles duplicados, a continuación se realiza el proceso que elimina los duplicados dejando los mejores datos en función del criterio que se defina.

3.2.2.7 Supervivencia

Después de agrupar los registros en la etapa de coincidencias, se pueden aplicar técnicas de supervivencias, como reglas para escoger al mejor candidato de varios registros y consolidarlo en uno solo. A continuación se crea un proceso para la supervivencia de los datos, el cual toma como insumo los datos agrupados generados en el proceso anterior.

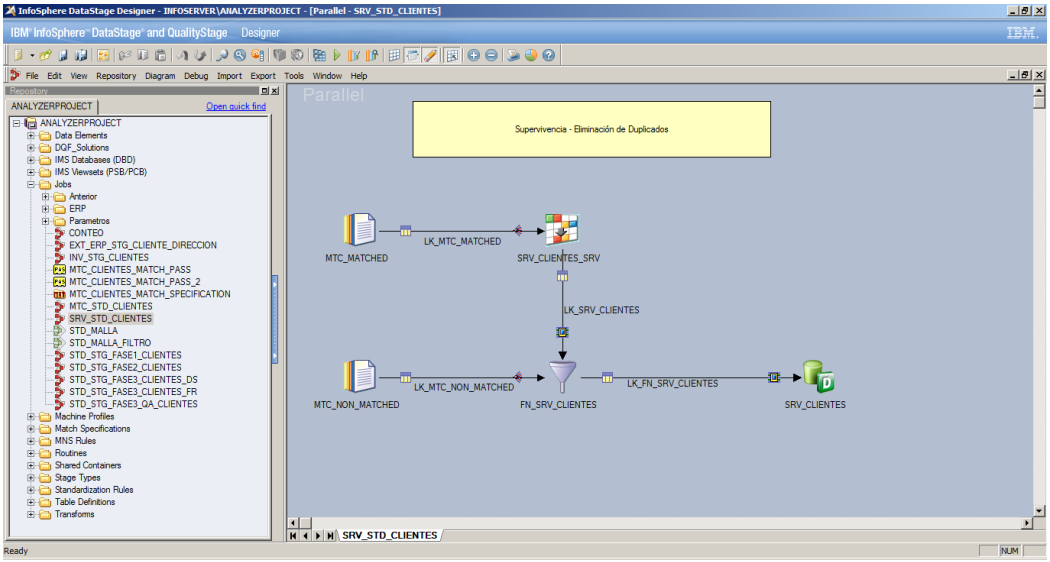


Figura 109: Supervivencia Paso – 1

Realizar la configuración del stage Survive en función de la **Figura 110**, incluir las columnas que van a pasar a la tabla final. No olvidar establecer el identificador qsMatchSetId.

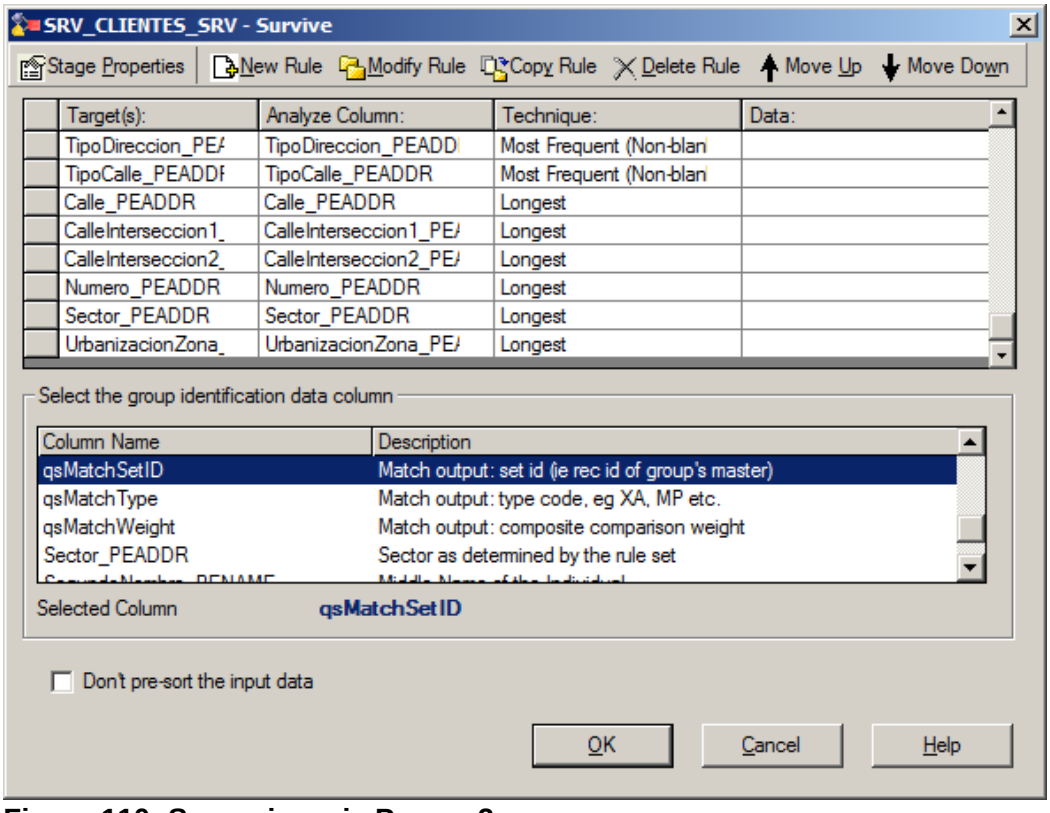


Figura 110: Supervivencia Paso – 2

Los resultados son los siguientes:

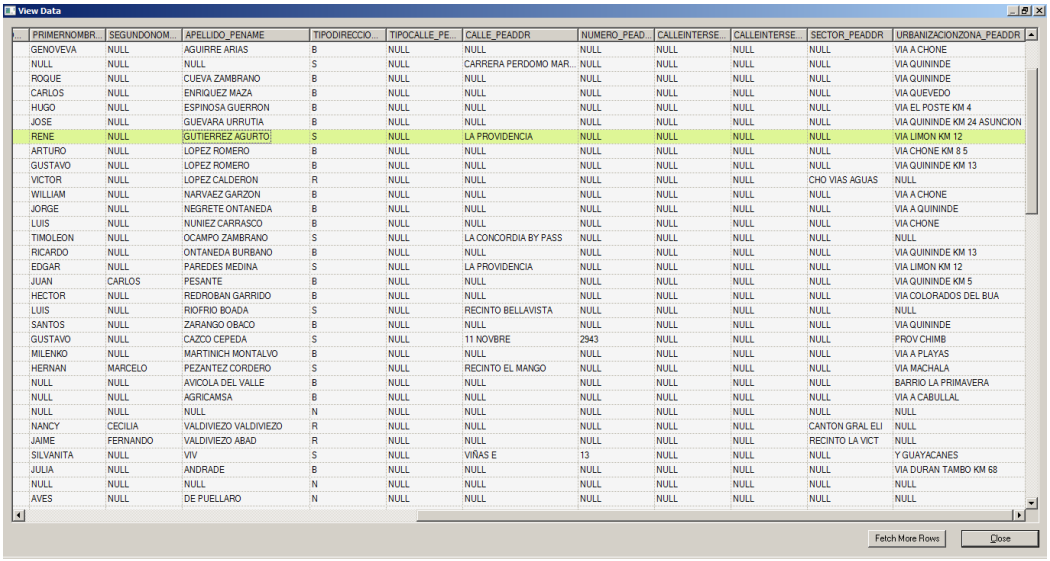


Figura 111: Supervivencia Paso – 3

3.2.2.8 Muestreo de Resultados

Existen varios casos en los que se presentan duplicados en la información, a continuación se describe un ejemplo de un cliente el que se identificó duplicados. Antes del proceso se tenía que el cliente CI/RUC 0990017514001 con duplicados en cuanto a dirección tal como indica la **Figura 112** marcado en verde y rojo.

CODIGO_CLIENTE	NOMBRE	CEDULA_RUC	DIRECCION_1
203690	TIENDAS INDUSTRIALES ASOCIADAS TIA	0990017514001	10 DE AGOSTO CALLE PASTAZA
203691	TIENDAS INDUSTRIALES ASOCIADAS TIA	0990017514001	10 DE AGOSTO CALLE PASTAZA
144974	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	10 DE AGOSTO Y BOTA
203694	TIENDAS INDUSTRIALES ASOCIADAS TIA	0990017514001	10 DE AGOSTO Y CIRCUNVALACION
105742	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CALLE TOACAZO Y MULALILLO ELOY
205318	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CALLE VELEZ Y LORENZO DE
286591	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CC GARZOCENTRO 2000
289878	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CDLA ACACIAS CALLE G. MORENO
131485	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CDLA IBARRA AV MARTHA BUCARAM
205313	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
205314	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
205315	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
205521	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
259023	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE
259024	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
259025	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
285005	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE
288654	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
288672	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE
401151	TIENDAS INDUSTRIALES ASOCIADAS S A	0990017514001	CHIMBORAZO 217 Y LUQUE
104745	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
104746	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
104935	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
176183	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
185605	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
205258	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
205526	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
205529	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
206105	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
273408	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
286580	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
288074	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
288083	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
288098	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
288099	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
289041	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
289075	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300766	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300767	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300776	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300777	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300778	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300782	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300829	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300884	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300887	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
300978	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
301137	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
301183	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
401102	TIENDAS INDUSTRIALES ASOCIADOS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
401113	TIENDAS INDUSTRIALES ASOCIADOS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
401824	TIENDAS INDUSTRIALES ASOCIADOS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ
401259	TIENDAS INDUSTRIALES ASOCIADOS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ.
401277	TIENDAS INDUSTRIALES ASOCIADOS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQ.
289865	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQUINA
300841	TIENDAS INDUSTRIALES ASOC	0990017514001	CHIMBORAZO 217 Y LUQUE ESQUINA
301101	TIENDAS INDUSTRIALES ASOCIADAS	0990017514001	CHIMBORAZO 217 Y LUQUE ESQUINA
301189	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUE ESQUINA
301197	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CHIMBORAZO 217 Y LUQUE ESQUINA
288675	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CHIMBORAZO 217 Y LUQUES ESQ
203672	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	CRISTOBAL Y CALLE ANDRES MZ 83
131295	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	CUSUBAMBA Y APUELA 2343
131381	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	ECUATORIANA E IGNACIO NOBOA
185603	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	ELIZARIO QUEVEDO 111 FELIX VAL
131323	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	EQUINOCCIAL Y 13 DE JUNIO
105829	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	GARCIA MORENO Y BUNINE
273603	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	GERRERO VALENZ Y LA A 1 G VALE
185607	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	GRAN COLOMBIA 127 J MARTIN
104925	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	GUAYAQUIL 958 Y ESPEJO
203080	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	JOSE RODRIGUEZ BONIN URB.RENAC
131347	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	JUAN MONTALVO Y FRAY GONZALO
203707	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	KM 11 1/2 VIA DAULE LOT.
203411	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	KM 11 1/2 VIA DAULE LOTIZ.
203709	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	KM 11 1/2 VIA DAULE LOTZ.
203710	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	KM 11 1/2 VIA DAULE LOTZ.
203345	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	KM.11 1/2 VIA DAULE LOTIZ.
203559	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	LA 17 Y PORTETE
203414	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	LA MANA
185801	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	LIZARDO RUIZ Y ALFAREROS 1 LIZ
505270	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	LOS ESTEROS AVENIDA 103
131374	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	MACHACHI AMAZONAS 3604 11NBRE
203695	TIENDAS INDUSTRIALES ASOCIADAS TIA	0990017514001	MACHALA 169 E/ROCAFUERTE 10
202790	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	MUCHO LOTE 1ERA.ETP. AV.ISIDRO
203332	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	MUCHO LOTE 5TA ETA.CALLJON 23
131453	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	PANAMERICANA NORTE ENTRADA A
105393	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	PEDRO V. MALDONADO Y LAS LAJAS
131311	TIENDAS INDUSTRIALES ASOCIODAS S.A.	0990017514001	PEDRO VICENTE MALDONADO Y GRAL
203231	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	PORTETE Y LA 9NA.
203348	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	REC.2 SALEM ENT.PRINC.1 ETP.
203347	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	RECREO 1 AV.PRINC.9 OCT.PEAT
131350	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	ROCAFUERTE E/BOLIVAR-LIBERTAD
203575	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	RODOLFO BAQUERIZO NAZUR Y CJON
203628	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	SAN FCO. DE MILAGRO Y ERNESTO
288676	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	SAUCES 5 CALLE RODRIGO ICAZA Y
288677	TIENDAS INDUSTRIALES ASOCIADAS S.A	0990017514001	SAUCES 6 CALLE GABRIEL ROLDOS
203282	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	TRINITARIA 1 COOP.POLO SUR
203283	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	TRINITARIA 2 COOP. JACOBO
202789	TIENDAS INDUSTRIALES ASOCIADAS S.A.	0990017514001	VIA 12 1/2 DAULE FTE.C.COMERC.

Figura 112: Estado Previo de los Datos

Una vez ejecutado el proceso se puede identificar que los clientes que estaban duplicados se simplificaron a uno solo como indica la **Figura 113** en los colores respectivos.

CODIGO_CLIENTI	CEDULA_RUC	APELLIDO_PENAME	TIPOCALLE	CALLE_PEADDR	NUMERO_F	KILOMETRO_I	SECTOR_PEADDR	CASA_P	URBANIZACIONZONA_
104925	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	GUAYAQUIL	958	NULL	NULL	NULL	Y ESPEJO
105742	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	CALLE	TOACAZO Y MULALILLO ELOY	NULL	NULL	NULL	NULL	NULL
105829	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	NULL	NULL	NULL	NULL	NULL	NULL
131295	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	CUSUBAMBA Y APUELA	2343	NULL	NULL	NULL	NULL
131401	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	AV	GUAYASAMIN UNA	NULL	NULL	NULL	NULL	NULL
170116	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	AV	MARISCAL SUCRE ENTRE OSORIO	NULL	NULL	NULL	NULL	NULL
185603	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	ELIZARIO QUEVEDO	111	NULL	NULL	NULL	FELIX VAL
185801	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	LIZARDO RUIZ Y ALFAREROS	1	NULL	NULL	NULL	LIZ
203080	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	JOSE FERNANDEZ DAVILA	NULL	NULL	NULL	NULL	URB RENAC
185613	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	AMAZONAS Y DE NOVIEMBRE	360	NULL	NULL	NULL	NOVIEMBRE
202789	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	NULL	NULL	NULL	NULL	NULL	NULL
202791	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	AV	ARQUITECTO MODESTO LUQUE	NULL	NULL	NULL	NULL	NULL
202874	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	AV	D COMIN FTE PARADA MET	NULL	NULL	NULL	NULL	VIA
203347	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	RECREO	1	NULL	NULL	NULL	NULL
203348	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	REC	2	NULL	NULL	NULL	NULL
203575	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	NULL	NULL	NULL	NULL	NULL	NULL
203631	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	CALLE	25 ENTRE LA L Y LA LL	NULL	NULL	NULL	NULL	NULL
203681	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	AV	MACHALILLA E ALEJO LASCAN	NULL	NULL	NULL	NULL	NULL
203695	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	MACHALA E ROCAFUERTE	169	NULL	NULL	10	NULL
205318	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	CALLE	VELEZ Y LORENZO DE	NULL	NULL	NULL	NULL	NULL
203332	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	MUCHO ETA CALLJON	NULL	NULL	NULL	NULL	NULL
203282	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	TRINITARIA	1	NULL	NULL	NULL	COOP POLO SUR
570091	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	NULL	NULL	KM 11	SAN FCO DE MILA	NULL	VIA DAULE LOTIZ
203692	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	BOLIVAR Y GONZALO CORDOV	NULL	NULL	SECTOR 6	NULL	DE NOVIEMBR
258636	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	LA	11	NULL	NULL	NULL	Y CAMILO DESTRUGE
288676	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	CALLE	SAUCES RODRIGO ICAZA Y		NULL	NULL	NULL	NULL
401824	0990017514001	TIENDAS INDUSTRIALES ASOCIADOS	ESQ	CHIMBORAZO Y LUQUES	217	NULL	NULL	NULL	Y LUQUE
273603	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	GERRERO VALENZ Y LA A	1	NULL	NULL	NULL	G VALE
286591	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	NULL	CC GARZOCENTRO	2000	NULL	NULL	NULL	NULL
500578	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	CALLE	ALAJUELA ENTRE PRIMERO	NULL	NULL	NULL	NULL	NULL
570386	0990017514001	TIENDAS INDUSTRIALES ASOCIADAS	CALLE	CHILE ENTRE	9	NULL	NULL	NULL	OCTUBRE

Figura 113: Resultados Calidad de Datos

3.2.2.9 Monitoreo de la Calidad de Datos

Hasta el momento se ha desarrollado todo el proceso de calidad de datos con sus fases, luego de esto se debe realizar el seguimiento para asegurar que los datos cumples con las reglas establecidas. A continuación se detallan los pasos a seguir para hacer el monitoreo de la información en la herramienta IBM Information Analyzer.

1) Crear un nuevo proyecto

Inicialmente se debe importar la metadata de la fuente que se va a hacer el seguimiento. Ingresar a la opción de importación como indica la **Figura 114**.

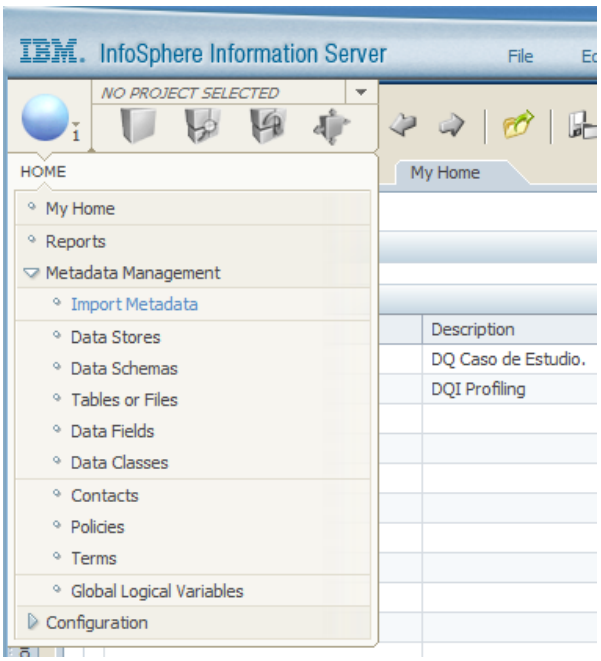


Figura 114: Monitoreo Paso – 1

Seleccione la fuente que se desea importar en la **Figura 115** se indica la base de datos en donde se guardan los datos del cliente.

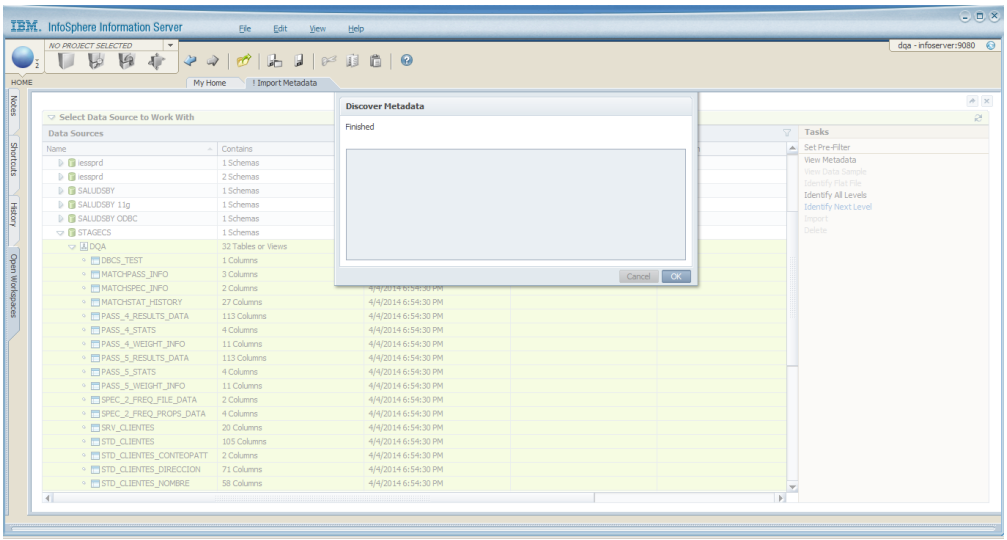


Figura 115: Monitoreo Paso – 2

Ubicar la tabla que servirá para la gestión. En la **Figura 116** se indica la tabla que guarda los datos depurados de clientes, SRV_CLIENTES

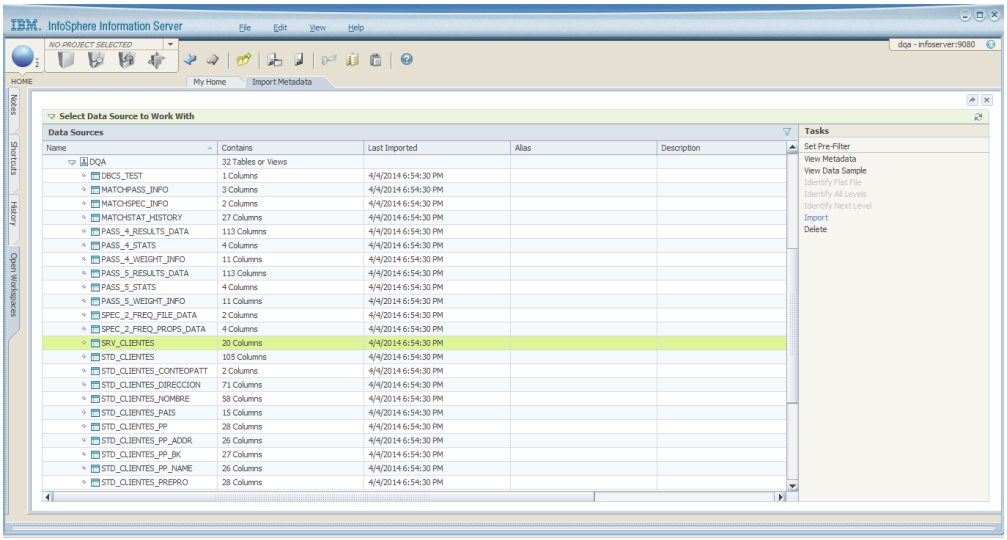


Figura 116: Monitoreo Paso – 3

Crear un nuevo proyecto como se indica en la **Figura 117**.

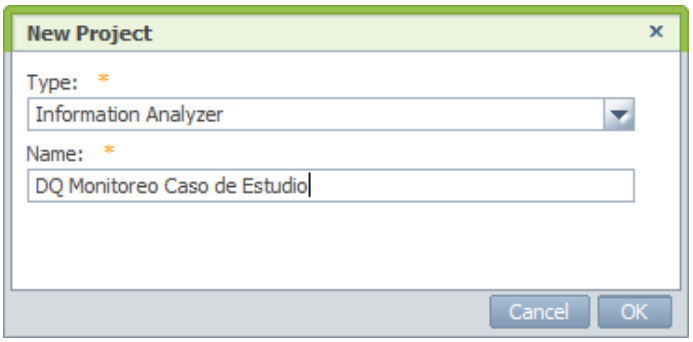


Figura 117: Monitoreo Paso – 4

Tomar como fuente la tabla importada y configurar las opciones de conexión para el proyecto.

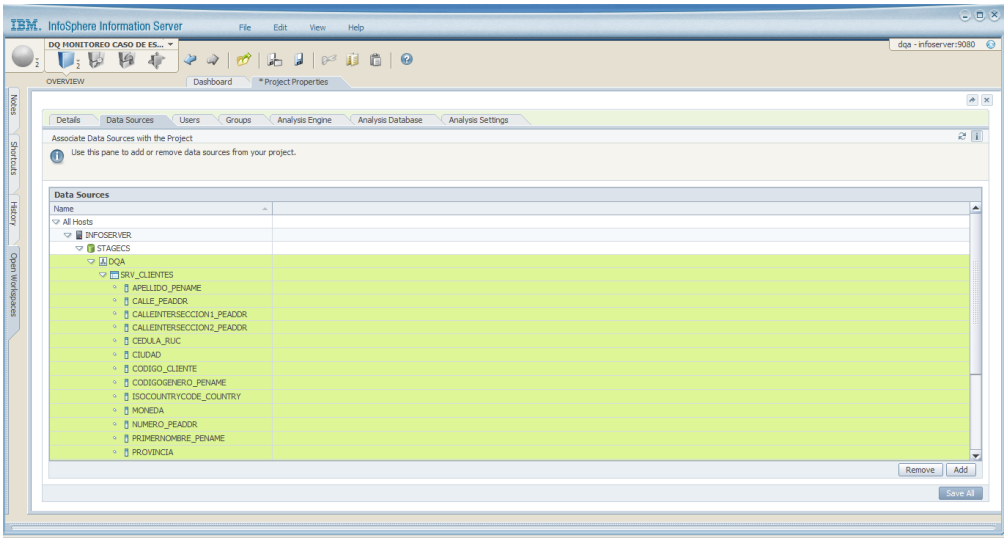


Figura 118: Monitoreo Paso – 5

2) Definir reglas de Monitoreo

Para el caso de estudio se definen 2 reglas de monitoreo, la primera va a ayudar a revisar que los tipos de dirección estén incluidos en una lista de predefinida. La segunda valida que un tipo de calle sea parte de un listado definido. Para los dos casos se define un umbral de 1% para alertar el cumplimiento de la regla, es decir si más del 1% de los datos dejan de cumplir con alguna de las reglas se marca como advertencia.

La regla tipo de calle se define a continuación, primero ingrese a la opción Data Quality del Menú, como indica la **Figura 119**:

Figura 119: Monitoreo Paso – 6

Seleccione crear una nueva Regla, tal como se indica en la **Figura 120**:

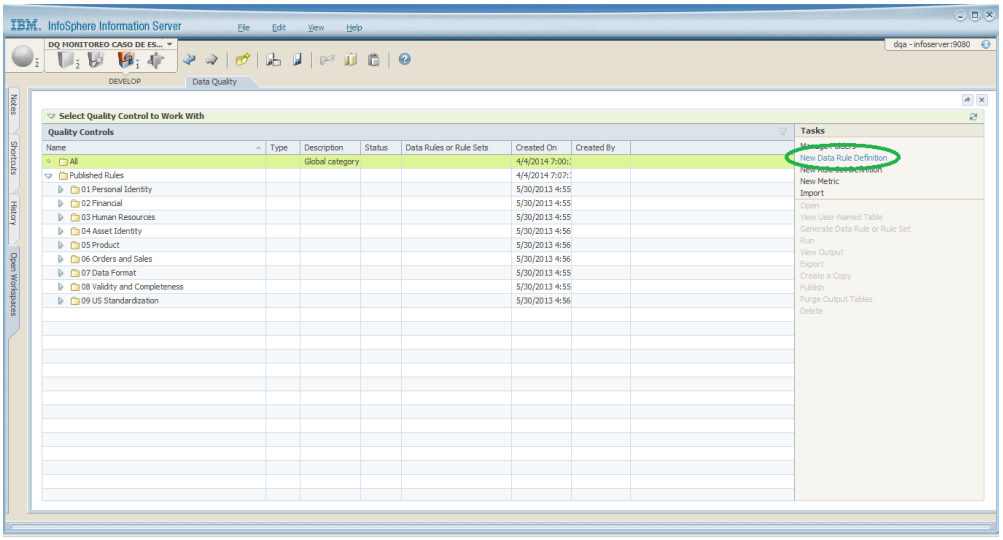


Figura 120: Monitoreo Paso – 7

Definir las reglas con los siguientes nombres PrefijoCalleValido y TipoCalleValido.

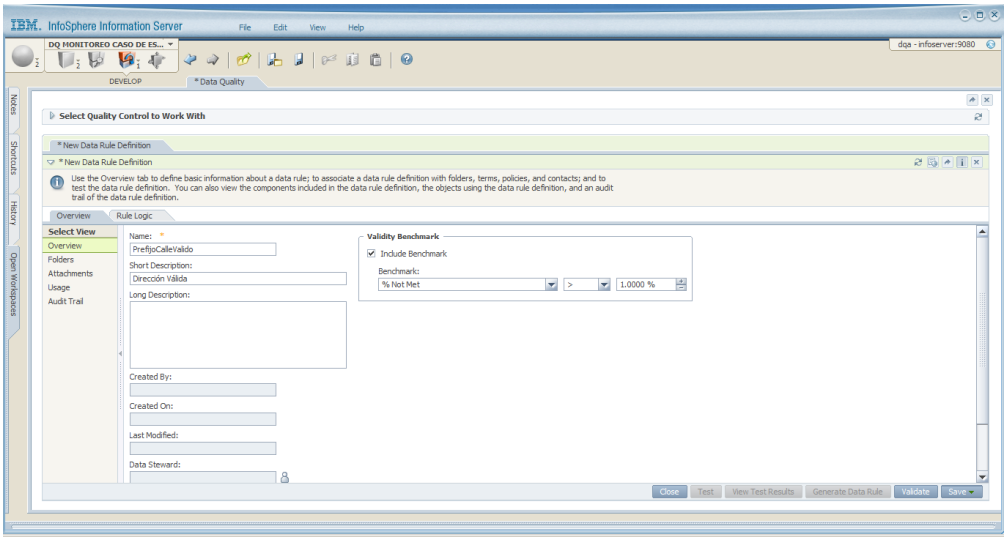


Figura 121: Monitoreo Paso – 8

Escriba la lógica de la regla con la siguiente definición:

- IF sourceData exists THEN sourceData in_reference_list {'B','L','M','P','R','S','U','Z'}
- IF sourceData exists THEN sourceData in_reference_list {'ACCESO', 'AU', 'AUTOPISTA', 'AV', 'AV CALLE', 'AV ESQ', 'AV

PASAJE', 'AVDA', 'AVE', 'AVENIDA', 'BOULEBARD', 'BOULEVAR', 'CALL', 'CALLE', 'CALLE AV', 'CALLE CALLE', 'CALLE ESQ', 'CARR', 'CL', 'DIAG', 'DIAGONAL', 'ESQ', 'ESQUI', 'ESQUINA', 'PASAJE', 'PEAT', 'PJE', 'PROLONG', 'PSJ', 'PSJE', 'Y', 'NULL'}

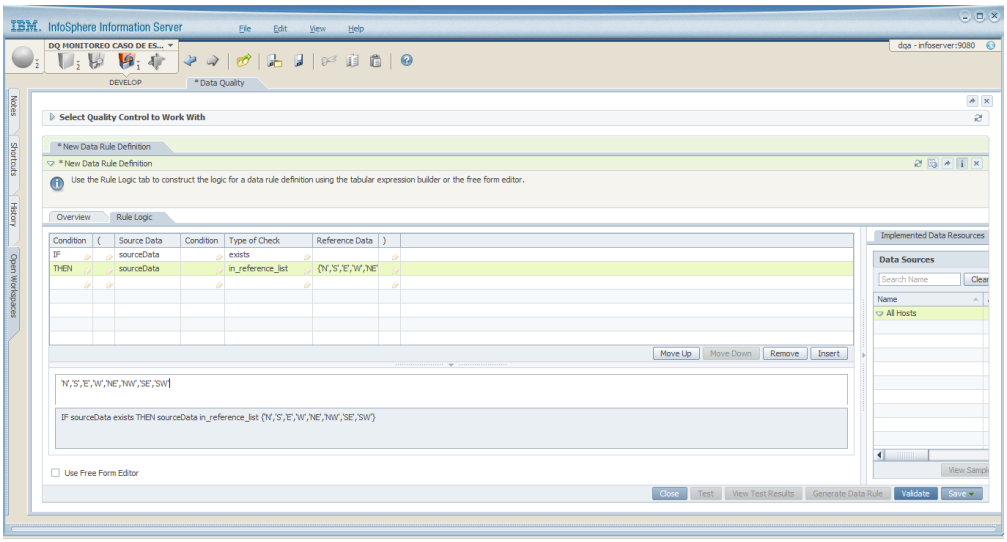


Figura 122: Monitoreo Paso – 9

Crear un conjunto de reglas, el cual sirve para monitorear que todas las reglas se cumplan.

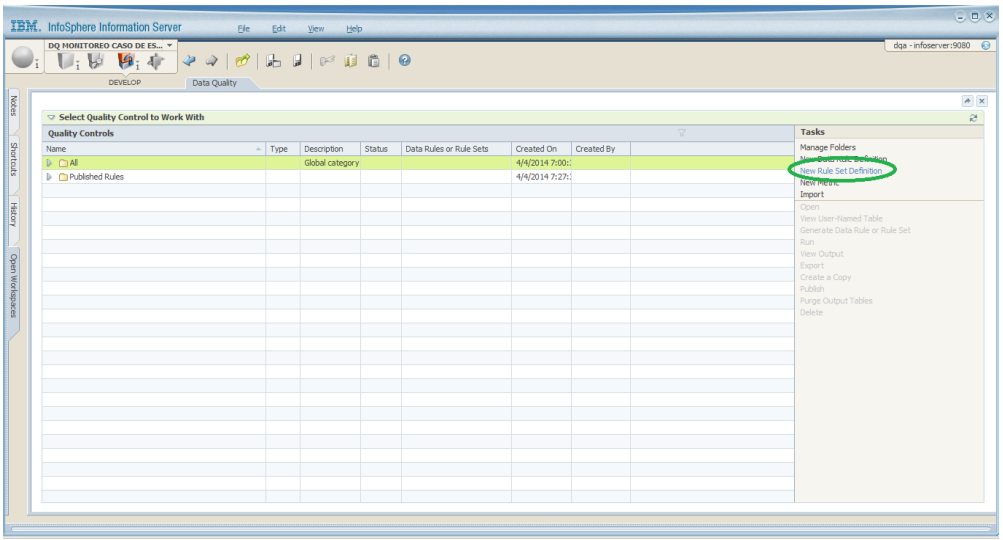


Figura 123: Monitoreo Paso – 10

Seleccionar las reglas creadas para que sean parte del conjunto.

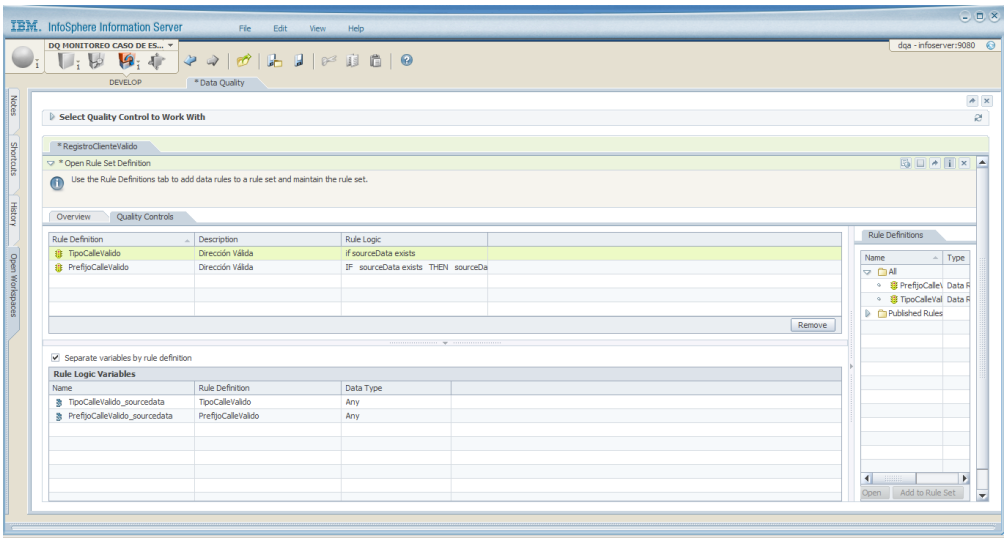


Figura 124: Monitoreo Paso – 11

Para terminar el proceso se genera el conjunto de reglas con la opción que indica la figura:

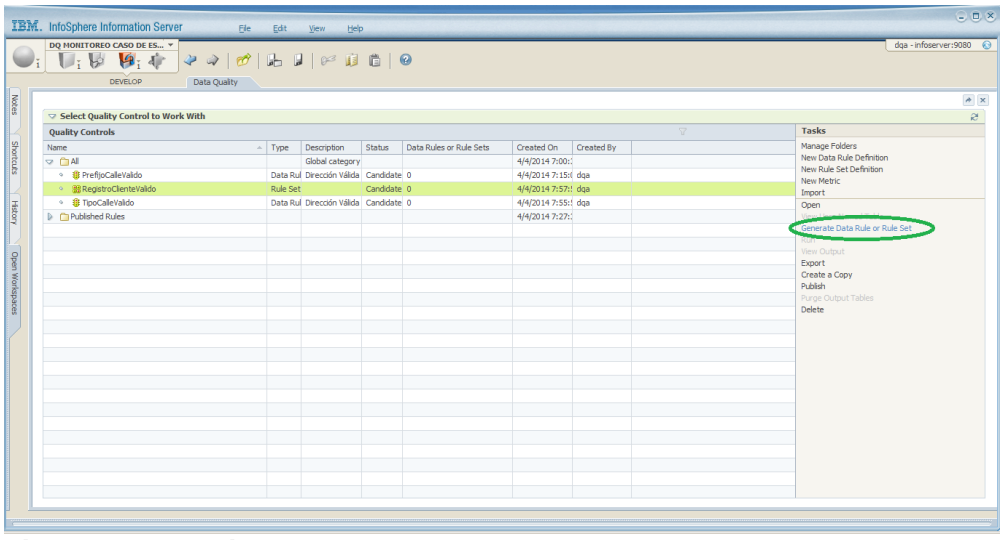


Figura 125: Monitoreo Paso – 12

Mapear la regla con la columna de la tabla que se va a evaluar:

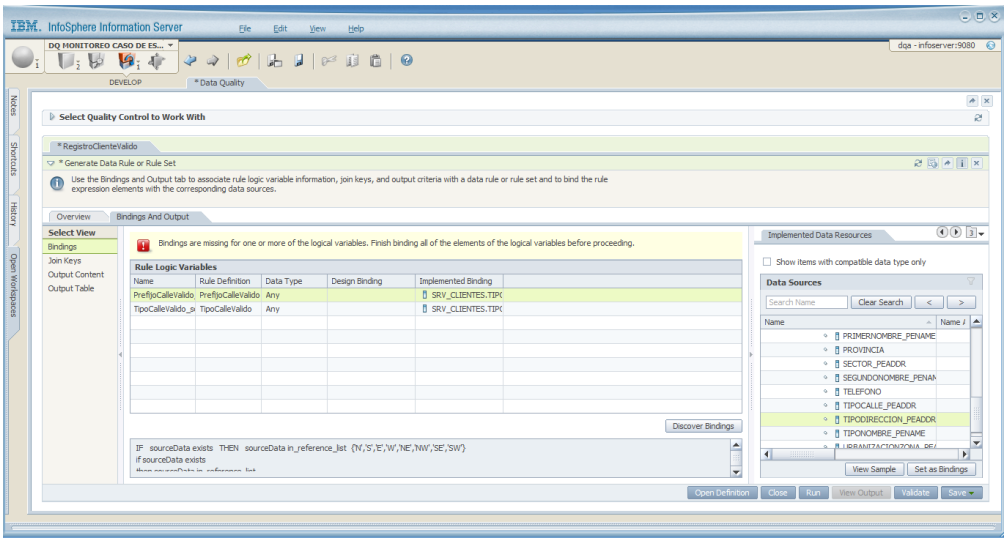


Figura 126: Monitoreo Paso – 13

3) Ejecutar el proceso de monitoreo

Ejecutar el conjunto de reglas para evaluar la calidad de los datos mediante la opción que se indica en la **Figura 127**:

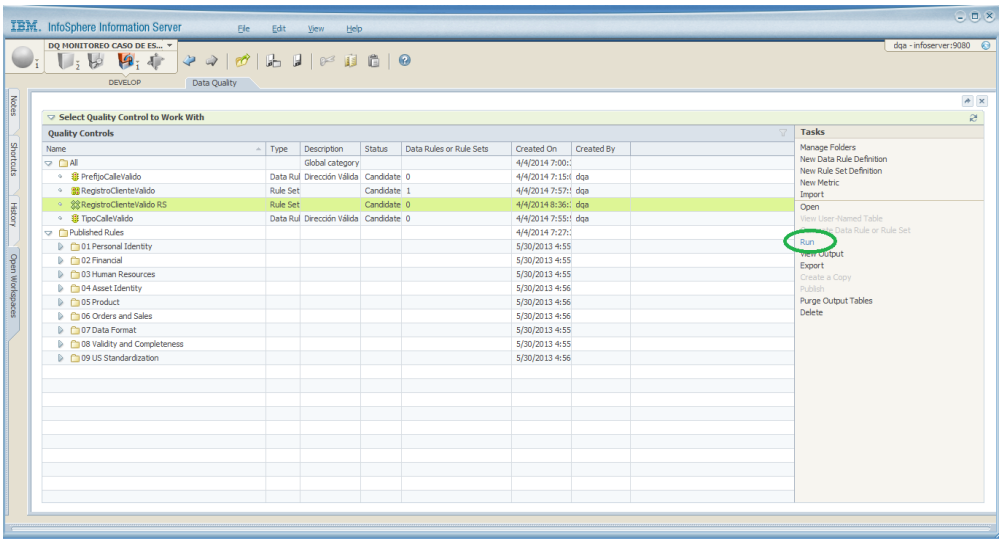


Figura 127: Monitoreo Paso – 14

Presionar Submit para iniciar la ejecución:

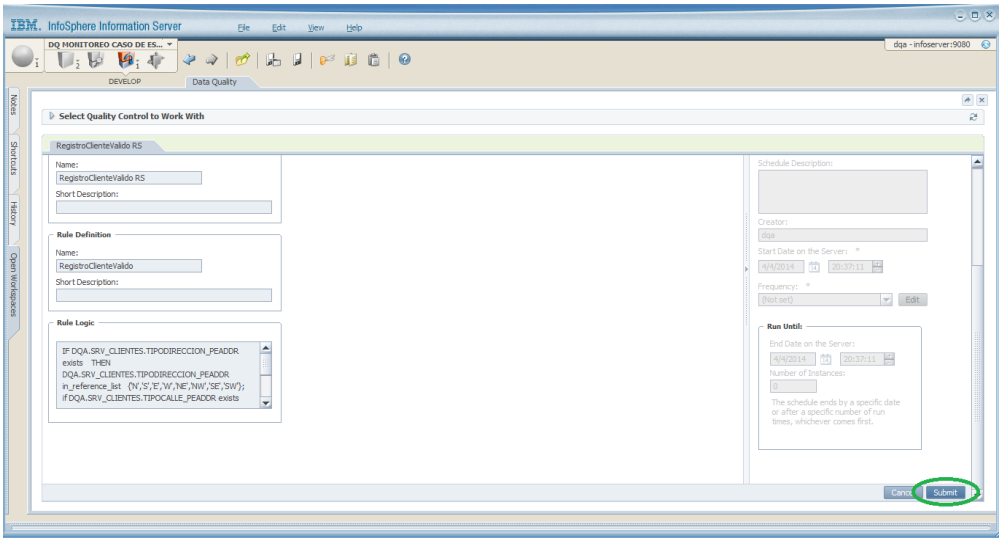


Figura 128: Monitoreo Paso – 15

Luego de ejecutar el proceso se presentan los resultados obtenidos

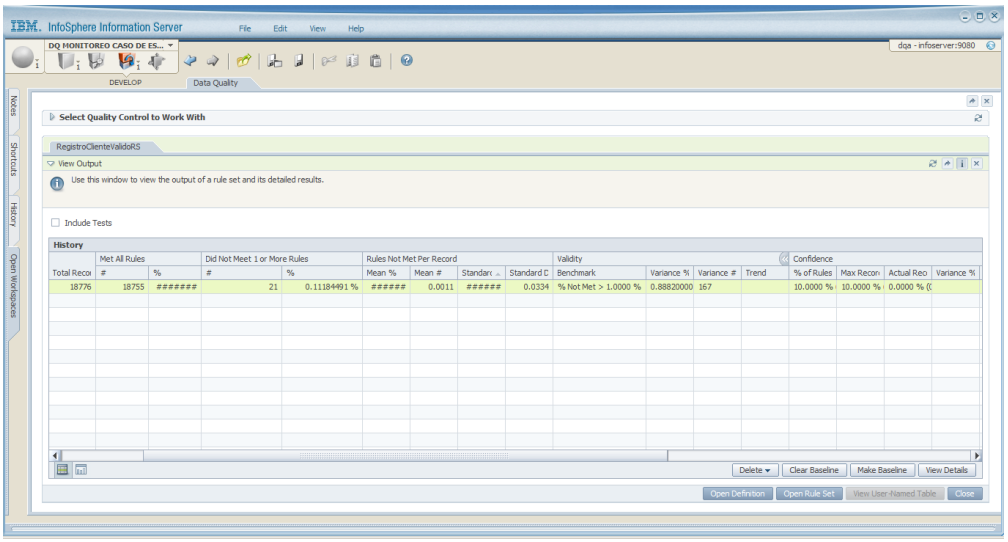


Figura 129: Monitoreo Paso – 16

El detalle de resultados nos indica que existen algunos datos que no cumplen con las reglas de monitoreo definidas.

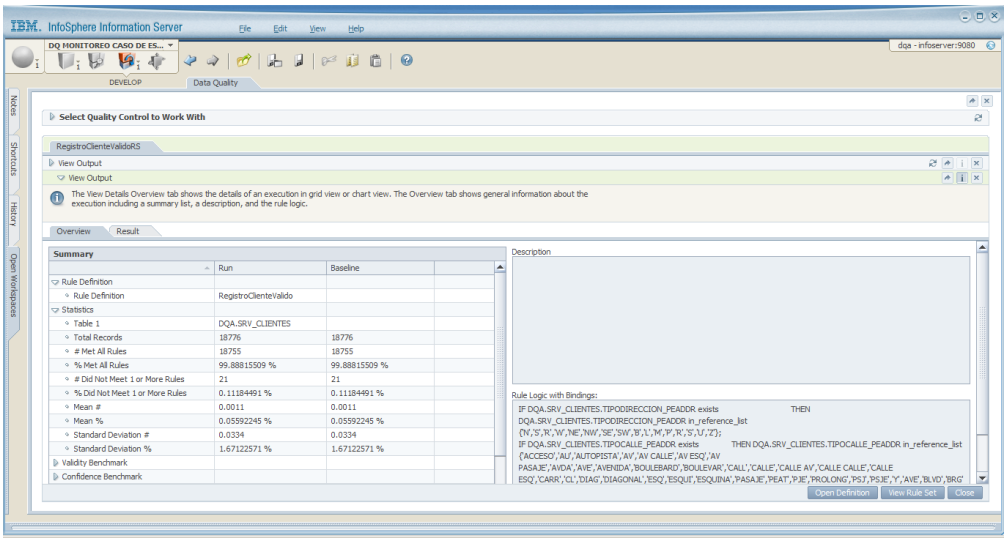


Figura 130: Monitoreo Paso – 17

Los resultados por distribución se presentan en la **Figura 131**. Se puede visualizar que existen datos que no cumplen con una de las reglas definidas.

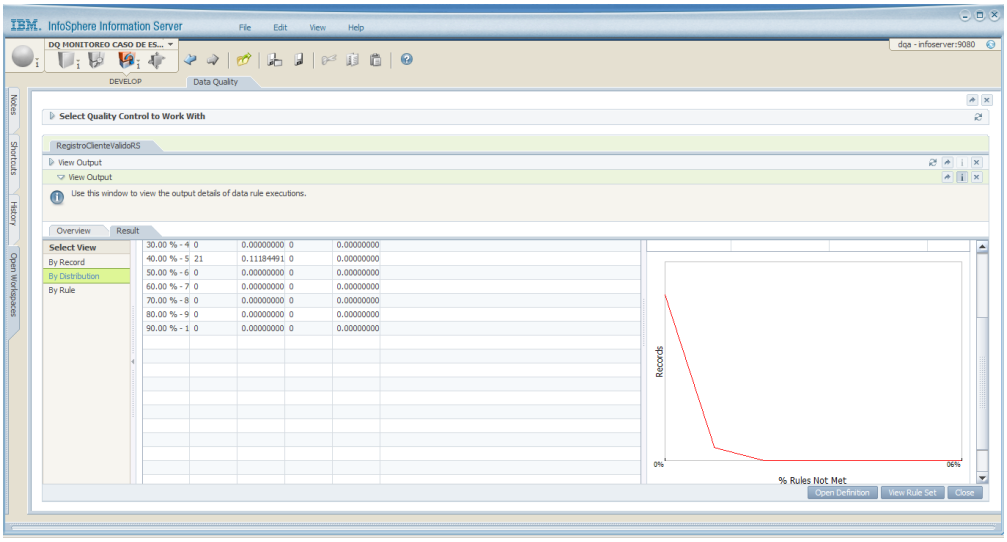


Figura 131: Monitoreo Paso – 18

Los resultados por registro se indican en la siguiente figura, para el tipo de dirección no se reconoce el dato F. Para que sea incluido debemos modificar la regla con la siguiente condición:

IF sourceData exists THEN sourceData in_reference_list {'B','L','M','P','R','S','U','Z',**'F'**}

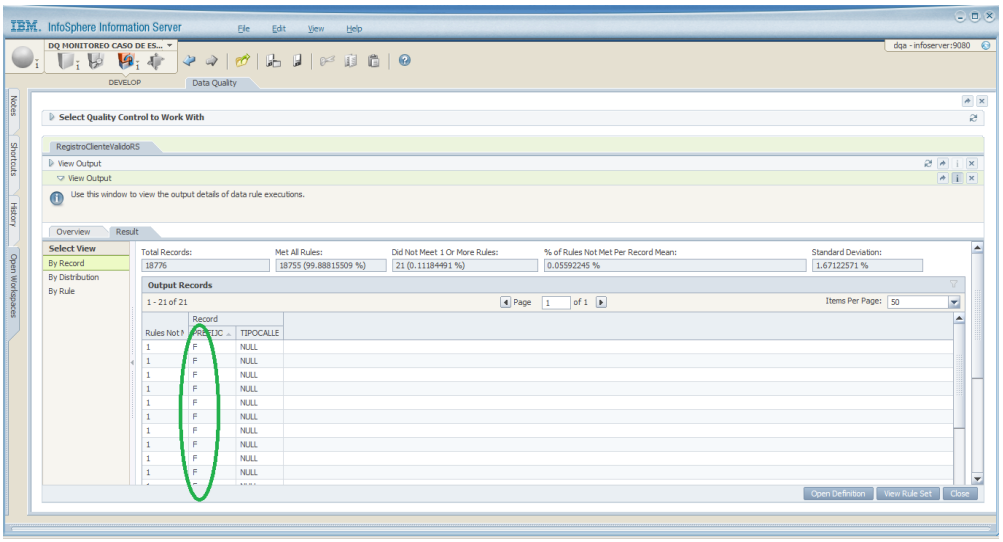


Figura 132: Monitoreo Paso – 19

Para poder realizar la evaluación debemos establecer esta corrida como línea base.

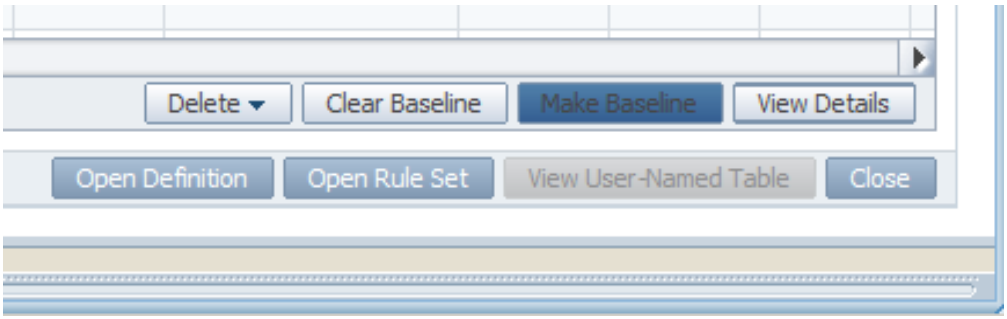


Figura 133: Monitoreo Paso – 20

Ejecutar nuevamente el proceso luego de incluir la nueva condición en la regla de tipo de dirección. El resultado indica que ha mejorado el porcentaje de cumplimiento de las reglas de monitoreo, así lo indica la **Figura 134**:

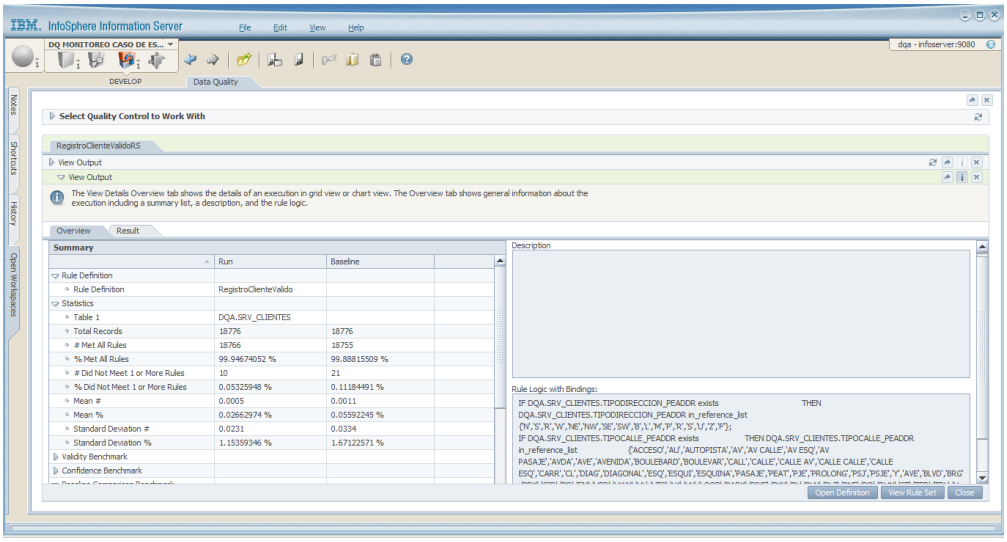


Figura 134: Monitoreo Paso – 21

La **Figura 135** indica la mejora que se presentan en los datos.

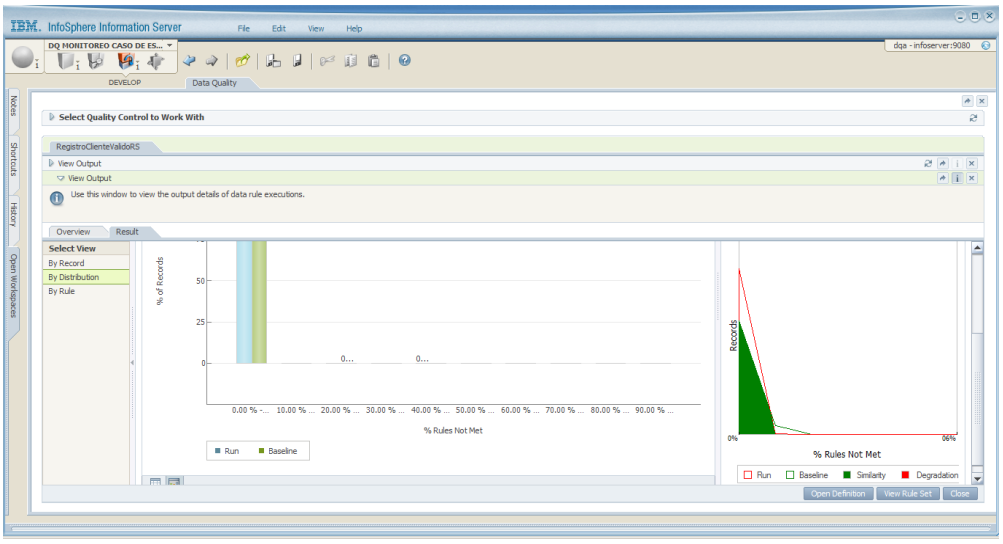


Figura 135: Monitoreo Paso – 22

El monitoreo sirve para detectar cambios en la calidad de los datos. Se pueden definir reglas para analizar la fuente de datos a diferentes niveles como columna, registro o regla. También se pueden definir alertas cuando la calidad de datos este degradando.

3.2.3 Análisis y Comparación de los Resultados Calidad de datos

Investigación

En la etapa de investigación se realizó el análisis de alto nivel mediante IBM Infosphere Information Analyzer, encontramos patrones que se repetían siendo ***AAAAAAA A.A.*** el más alto con **221** repeticiones. También encontramos valores repetidos como **Consumidor Final** que se repite **44** veces.

También se realizó la investigación mediante IBM Infosphere Data Quality Designer, mediante el método de investigación Word Investigation obtuvimos como resultado patrones que nos permitieron estandarizar la información por ejemplo estos 2 patrones:

1. ??FF: Donde F = Primer Nombre? = Palabra (Se puede inferir como Apellido)

2. ??F: Donde F = Primer Nombre? = Palabra (Se puede inferir como Apellido)

Estandarización

El proceso de estandarización se creó mediante un conjunto de reglas predefinidas en la herramienta para estandarizar el nombre y la dirección al inició se obtuvo para para el nombre **36.83%** de datos que cumplían las reglas y para la dirección **35.9%,** en base a estos resultados se sobre-escribieron las reglas predefinidas añadiendo patrones que permitan un mayor porcentaje de datos estandarizados. Los siguientes son los patrones añadidos para el nombre. Podemos notar que se incluyen los patrones que notamos en la etapa de investigación.

Input Pattern:	Override Codes:	Other (
++FF	Apellido1 Apellido1 PrimerNombre2 SegundoNombre5	
++F+	Apellido1 Apellido1 PrimerNombre2 SegundoNombre5	
++F	Apellido1 Apellido1 PrimerNombre6	
+F	Apellido1 PrimerNombre6	
++	Apellido1 PrimerNombre5	
+++F	Apellido1 Apellido1 PrimerNombre1 SegundoNombre6	
+++	Apellido1 Apellido1 PrimerNombre5	
+FF	Apellido1 PrimerNombre2 SegundoNombre6	
++FLF	Apellido1 Apellido1 PrimerNombre2 Apellido2 SegundoNombre5	
+	Apellido5	
++++	Apellido1 Apellido1 PrimerNombre1 SegundoNombre1	
F+FF	Apellido1 Apellido1 PrimerNombre2 SegundoNombre6	
+++FF	Apellido1 Apellido1 Apellido1 Apellido2 Apellido6	

Figura 136: Patrones de Estandarización Cliente

Los siguientes son los añadidos para la dirección:

Unhandled Pattern:	Override Codes:
^	Calle1
++<^++	CalleInterseccion11 CalleInterseccion11 Casa1 Casa1 CalleInterseccion21 CalleInterseccion21
+<^++	CalleInterseccion11 Casa1 Casa1 CalleInterseccion21 CalleInterseccion21
+^	UrbanizacionZona1 UrbanizacionZona1
+++	Sector1 Sector1 Sector1
+++++	Sector1 Sector1 Sector1 Sector1 Sector1
++++	Sector1 Sector1 Sector1 Sector1
++++++	Sector1 Sector1 Sector1 Sector1 Sector1 Sector1
++	Sector1 Sector1
++<^+++	CalleInterseccion11 CalleInterseccion11 Casa1 Casa1 CalleInterseccion21 CalleInterseccion21 (
+	UrbanizacionZona1
++^	Calle1 Calle1 Casa1
+<^+++	Calle1 Casa1 Casa1 Calle1 Calle1 Calle1
<	Casa1
+++^	Calle1 Calle1 Calle1 Casa1
+++<^++	Calle1 Calle1 Calle1 Casa1 Casa1 Calle1 Calle1
T++++	TipoCalle1 Calle1 Calle1 Calle1 Calle1
T+++	TipoCalle1 Calle1 Calle1 Calle1
T+++++	TipoCalle1 Calle1 Calle1 Calle1 Calle1 Calle1
+^++	Calle1 Casa1 Calle1 Calle1
+T+	Calle1 TipoCalle1 Calle1
++<^	Calle1 Calle1 Casa1 Casa1
+^+	Calle1 Casa1 Calle1
++S	Calle1 Calle1 TipoCalle1
++<^+	Calle1 Calle1 Casa1 Casa1 Calle1

Figura 137: Patrones de Estandarización Dirección

Se realizó nuevamente el análisis y se obtuvo para el nombre el **93.15%** de datos estandarizados, para la dirección se mejoró a **75.36%**.

Coincidencia

Mediante la creación de una especificación de coincidencias se realizaron 2 agrupaciones para identificar registros duplicados., las cuales se procesan

en secuencia. Para la primera pasada se tiene el **8% de duplicados**, con **1858** registros del total.

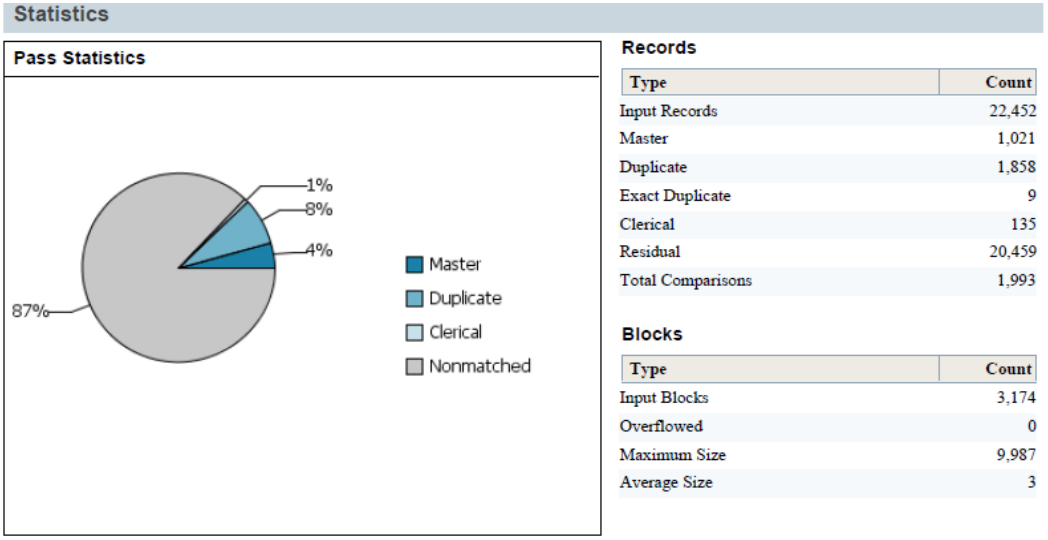


Figura 138: Proceso de Coincidencias Primera Corrida

Para la segunda pasada se tiene **8% de duplicados** también pero con **1633** registros del total.

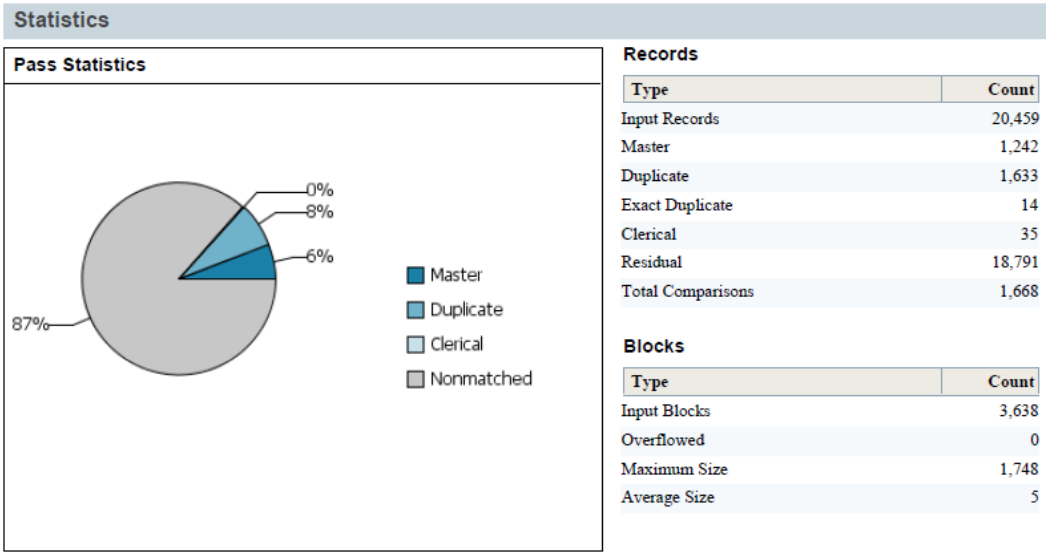
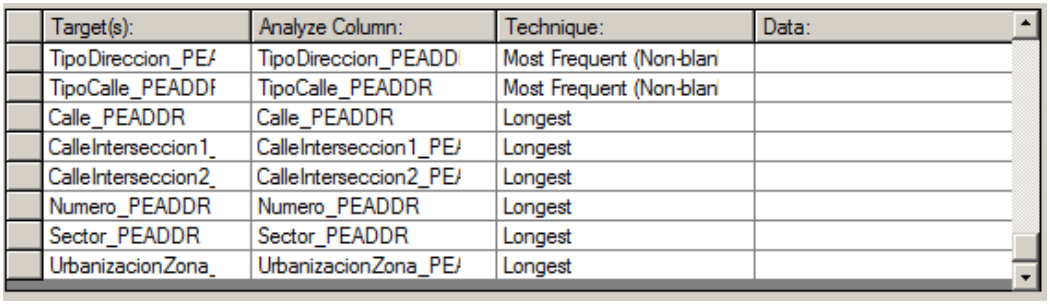


Figura 139: Proceso de Coincidencias Segunda Corrida

Supervivencia

Para la eliminación de los duplicados identificados en la etapa anterior se establece en función de 2 criterios Valores Repetidos o por longitud de la cadena. Con esto logramos asegurarnos que se cargue valores correctos a la tabla final.



Target(s):	Analyze Column:	Technique:	Data:
TipoDireccion_PE/	TipoDireccion_PEADD	Most Frequent (Non-blan	
TipoCalle_PEADDI	TipoCalle_PEADDR	Most Frequent (Non-blan	
Calle_PEADDR	Calle_PEADDR	Longest	
CalleInterseccion1_	CalleInterseccion1_PE/	Longest	
CalleInterseccion2_	CalleInterseccion2_PE/	Longest	
Numero_PEADDR	Numero_PEADDR	Longest	
Sector_PEADDR	Sector_PEADDR	Longest	
UrbanizacionZona_	UrbanizacionZona_PE/	Longest	

Figura 140: Configuración del Proceso de Supervivencia

3.2.4 Evaluación Final - Modelo de Madurez TDWI

3.2.4.1 De igual forma que en la sección 3.2.1 - Evaluación Inicial - Modelo de Madurez TDWI

Para evaluar el Modelo de Madurez de acuerdo a los parámetros del TDWI se utiliza la encuesta proporcionada por el mismo instituto, el cuestionario completo fue detallado en la primera parte de este trabajo y en esta etapa se han contestado las preguntas de acuerdo a las características de la empresa donde se está implementado el caso de estudio, los resultados de la encuesta se presentan a continuación en la **Figura 7**, se han contestado las preguntas del cuestionario TDWI después de aplicar el Modelo de Gestión de Calidad de Datos, en esta oportunidad los resultados se detallan en la **Figura 141**, donde es posible apreciar que ha existido un aumento en el parámetro Datos *(Data)*:

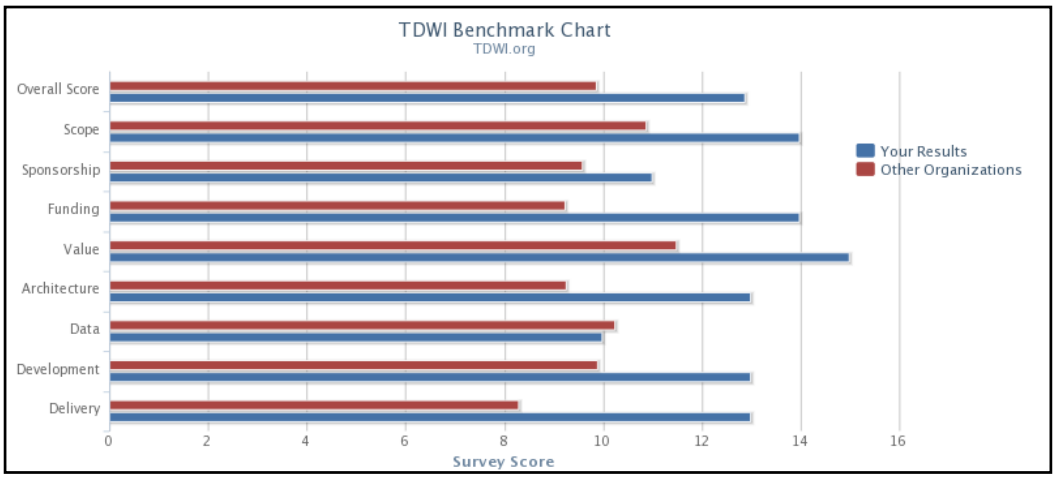


Figura 141: Resultados Evaluación Final TDWI
Fuente: (TDWI, Evaluación del Modelo de Madurez de BI de TDWI, 2012)

Presentando los resultados de forma tabular obtenemos lo siguiente:

Tabla 2

Resultados de la Evaluación Final TDWI

CATEGORÍA	PUNTAJE OBTENIDO	NIVEL DE MADUREZ / PUNTO CRÍTICO
Ámbito	13	Repetible
Patrocinio	11	Preliminar
Presupuesto	14	Repetible
Valor	15	Repetible / Abismo
Arquitectura	13	Repetible
Datos	**10**	**Preliminar**
Desarrollo	13	Repetible
Entrega	13	Repetible

3.2.4.2 Análisis y Comparación de los Resultados

Al comparar los puntajes y niveles de madurez/puntos críticos obtenidos anteriormente con los nuevos resultados, obtenemos la Tabla 3:

Tabla 3

Comparación de Resultados de la Evaluación TDWI

CATEGORÍA	PUNTAJE OBTENIDO		NIVEL DE MADUREZ / PUNTO CRÍTICO	
	Inicial	**Final**	**Inicial**	**Final**
Ámbito	13	13	Repetible	Repetible
Patrocinio	11	11	Preliminar	Preliminar
Presupuesto	14	14	Repetible	Repetible
Valor	15	15	Repetible / Abismo	Repetible / Abismo
Arquitectura	13	13	Repetible	Repetible
Datos	**9**	**10**	**Preliminar / Golfo**	**Preliminar**
Desarrollo	13	13	Repetible	Repetible
Entrega	13	13	Repetible	Repetible

Es importante notar que la categoría Datos mejora su calificación en un punto y además eso le permite sobrepasar el punto crítico del Golfo, sin embargo permanece en el nivel de madurez Preliminar. En las demás categorías no se evidencia un aumento o disminución de los puntajes.

El aumento de la calificación de la categoría Datos, así como el hecho de que las otras categorías hayan permanecido estables es el resultado esperado de la aplicación del Modelo de Gestión de Calidad de Datos, lo cual será analizado en mayor detalle en la respectiva sección del presente trabajo.

CAPÍTULO V

4 CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- El Modelo de Gestión de Calidad de Datos basado en la metodología Data Quality de IBM, propuesto en la primera parte de este trabajo, fue utilizado en el desarrollo de un caso de estudio práctico, basándose en una fuente de datos real obtenida del ambiente de producción de una reconocida empresa nacional. Esto comprueba que el Modelo de Gestión de Calidad de Datos propuesto es aplicable en los proyectos de Calidad de Datos que son una nueva línea de negocio de la empresa auspiciante DWConsulware.

- La Gestión de Calidad de Datos fue aplicada mediante el uso de la herramienta IBM Infosphere Information Server, el uso de la herramienta permitió mejorar la calidad de los datos iniciales desarrollando las etapas planteadas por la Metodología IBM DataQuality.

- La mejora en la calidad de datos proporciona a los usuarios finales una visión más clara de los mismos, efectivamente convirtiendo los datos en información, lo cual indica la conveniencia de generar reglas de negocio para la captura de datos en el campo de trabajo.

- El nivel de madurez de BI del caso de estudio analizado se incrementa únicamente en la categoría de Datos del Modelo de Madurez TDWI lo cual indica que:

- La Gestión de Calidad de Datos no es suficiente por sí misma para elevar el nivel de madurez de BI de la organización.

- Las categorías del Modelo de Madurez TDWI que se encuentran en los puntos críticos (el Golfo y el Abismo) requieren más iteraciones de un modelo de gestión, en este caso de estudio Gestión de Calidad de Datos, para avanzar hacia las siguientes etapas del modelo.

- La Gestión de Calidad de Datos es un proceso de ejecución a corto plazo, el cual influye en la categoría correspondiente en los Modelos de Madurez de BI como la categoría Datos en el Modelo TDWI, por lo tanto se requieren varias iteraciones de este proceso para influir de forma efectiva en otras áreas que por su naturaleza tienen ciclos de ejecución a mediano y largo plazo, como son las categorías de Presupuesto y Patrocinio. Esto se debe a que la Gestión de Calidad de Datos es un proceso enfocado casi exclusivamente hacia el área técnica de BI, mientras que las categorías de Presupuesto y Patrocinio se enfocan hacia el área de negocios y dirección de las organizaciones.

4.2 Recomendaciones

- Aplicar el Modelo de Gestión de Calidad de Datos propuesto en la primera parte de este trabajo y probado en un caso de estudio práctico en la segunda parte del mismo, en los proyectos de Gestión de Calidad de Datos de la empresa DWConsulware.

- Integrar el Modelo de Gestión de Calidad de Datos con la metodología de proyectos de Inteligencia de Negocios para el portafolio de clientes de la empresa DWConsulware.

- Desarrollar reglas de estandarización de datos adaptadas a los formatos, leyes y reglamentos utilizados en el país para objetos comunes de negocio como: clientes, direcciones, personas naturales y jurídicas, que se puedan integrar con la plataforma IBM Infosphere Information Server.

- Considerar el desarrollo de Modelos de Gestión aplicables a otras áreas de los Modelos de Madurez de BI como posibles temas para la elaboración de proyectos de graduación del programa de Maestría en Gerencia de Sistemas.

- Considerar el desarrollo de un estudio comparativo entre los diferentes Modelos de Madurez de BI como posible tema para la elaboración de proyectos de graduación de pregrado.

REFERENCIAS

1. IBM. (2012). Data Quality Bootcamp.

2. InfoSphere Information Server Intro, I. (2012). IBM InfoSphere Information Server Introduction. Obtenido de http://publibfp.boulder.ibm.com/epubs/pdf/c1937950.pdf

3. Mancuso, G. (8 de July de 2004). The Common Problem – Working with Merge/Purge and Householding. Obtenido de Information Management: http://www.information-management.com/news/1006332-1.html

4. Menéndez, C. P. (13 de 08 de 2013). En los datos, la calidad importa. Obtenido de http://liberix.es/blog/en-los-datos-la-calidad-importa/

5. TDWI. (2012). Evaluación del Modelo de Madurez de BI de TDWI. Obtenido de http://tdwiorg0000.web711.discountasp.net/Content/TDWI_Benchmark_Final.pdf

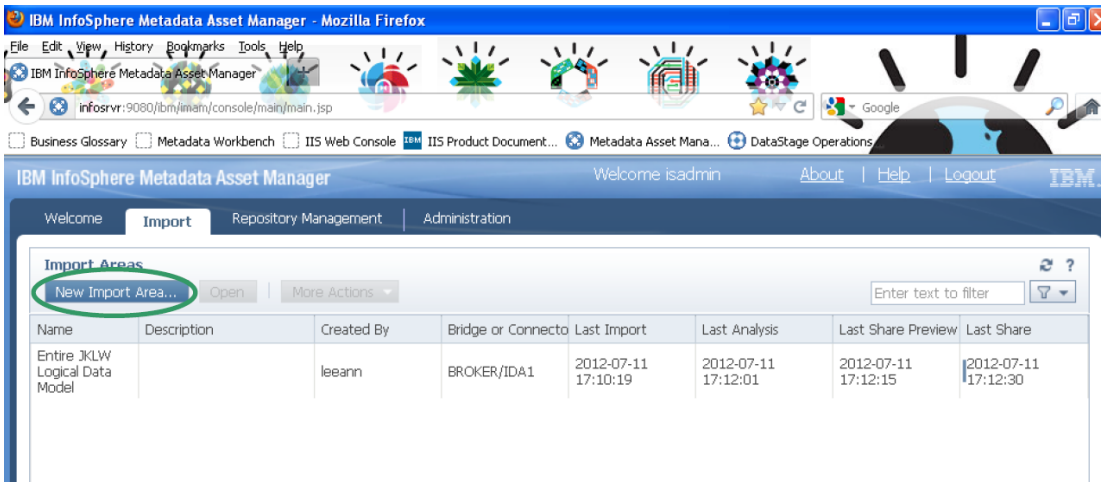
6. TDWI. (2014). Maturity Models and Assessments. Obtenido de http://tdwi.org/pages/maturity-model/maturity-model-home.aspx

ANEXOS

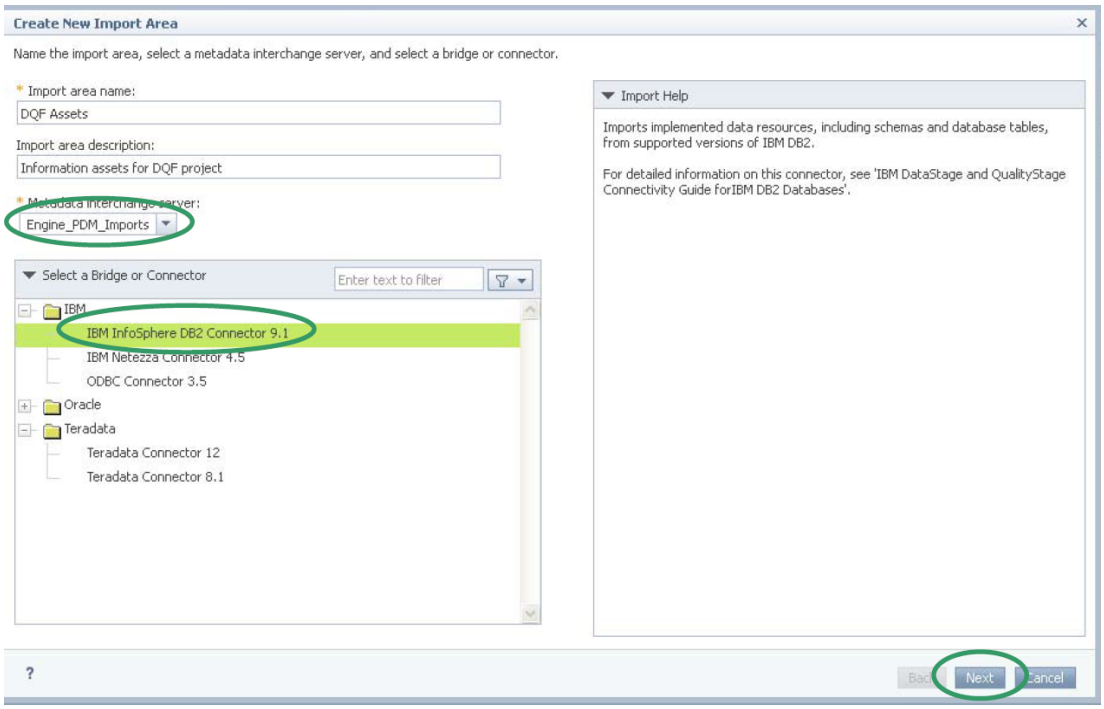
ANEXO A

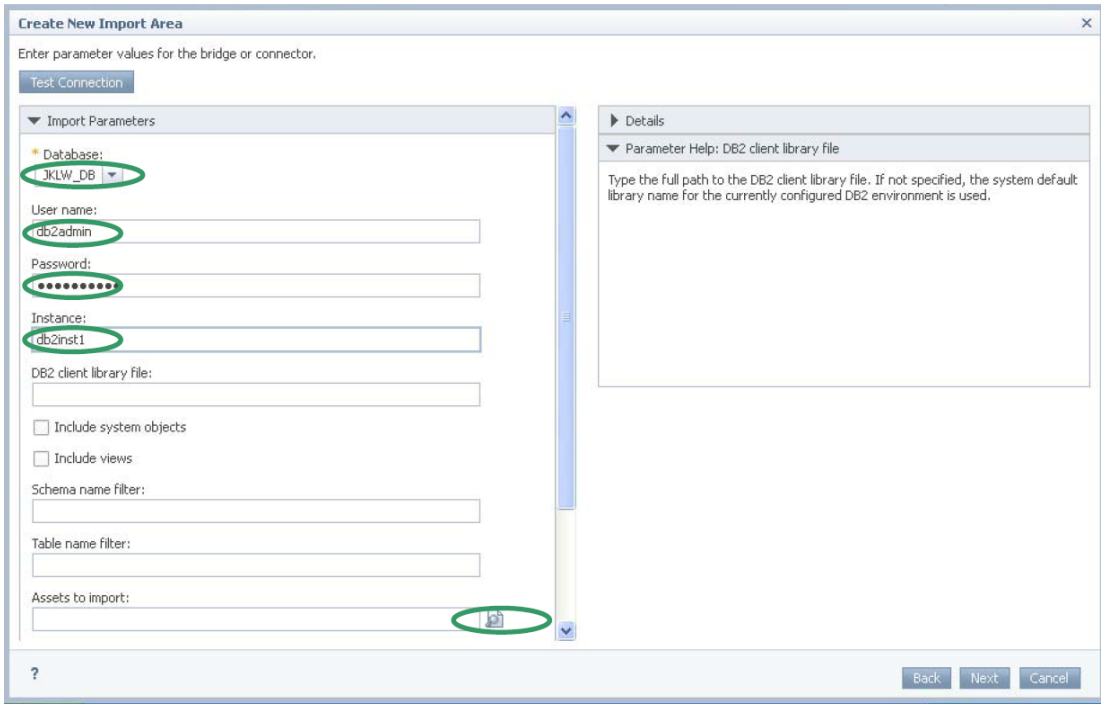
CREAR UNA NUEVA ÁREA DE IMPORTACIÓN EN IBM INFOSPHERE METADATA ASSET MANAGER

1) Crear una nueva área de importación

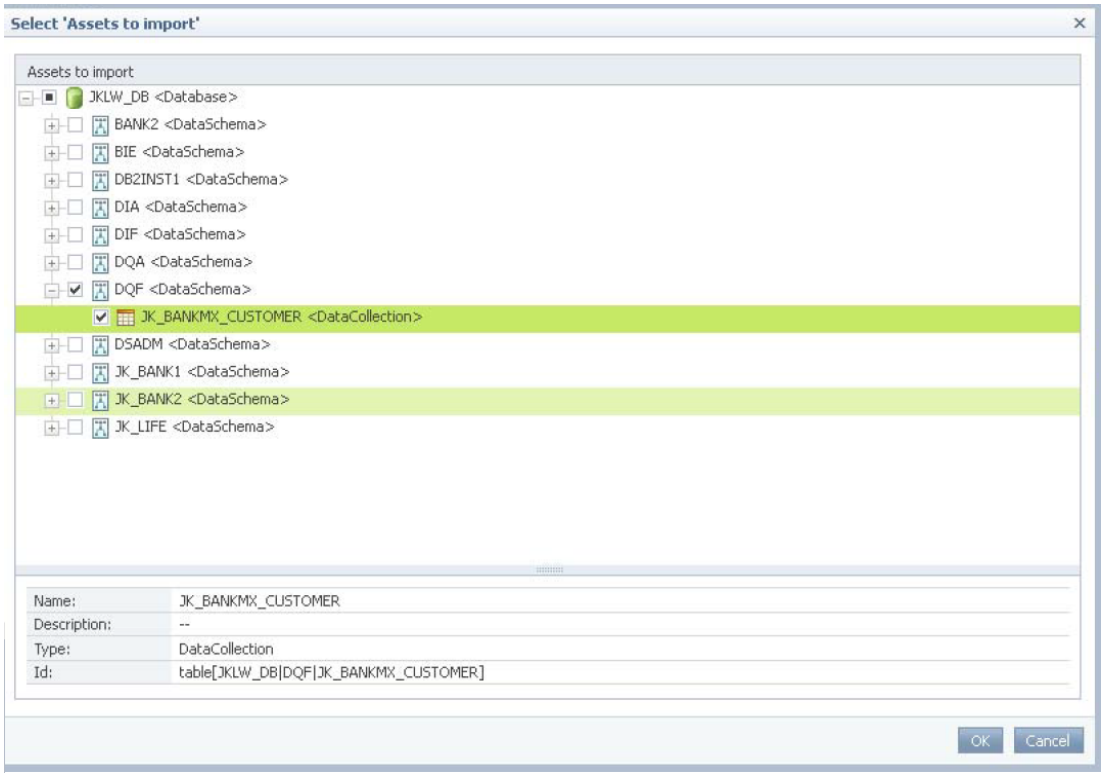


2) Realizar una conexión contra una fuente de la base de datos DB2.

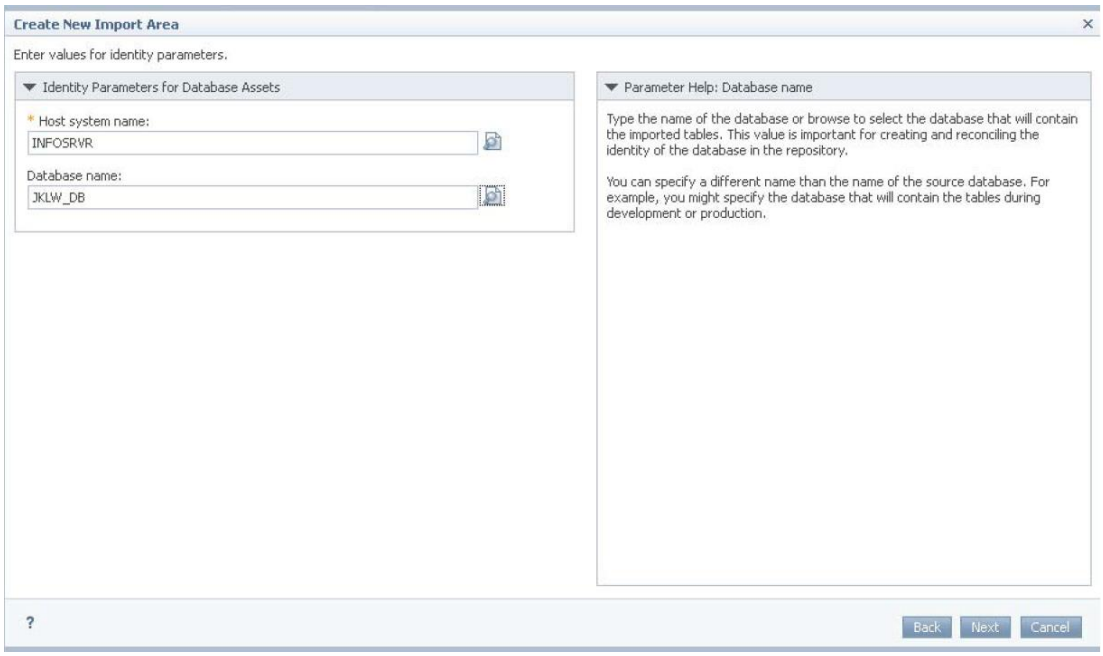




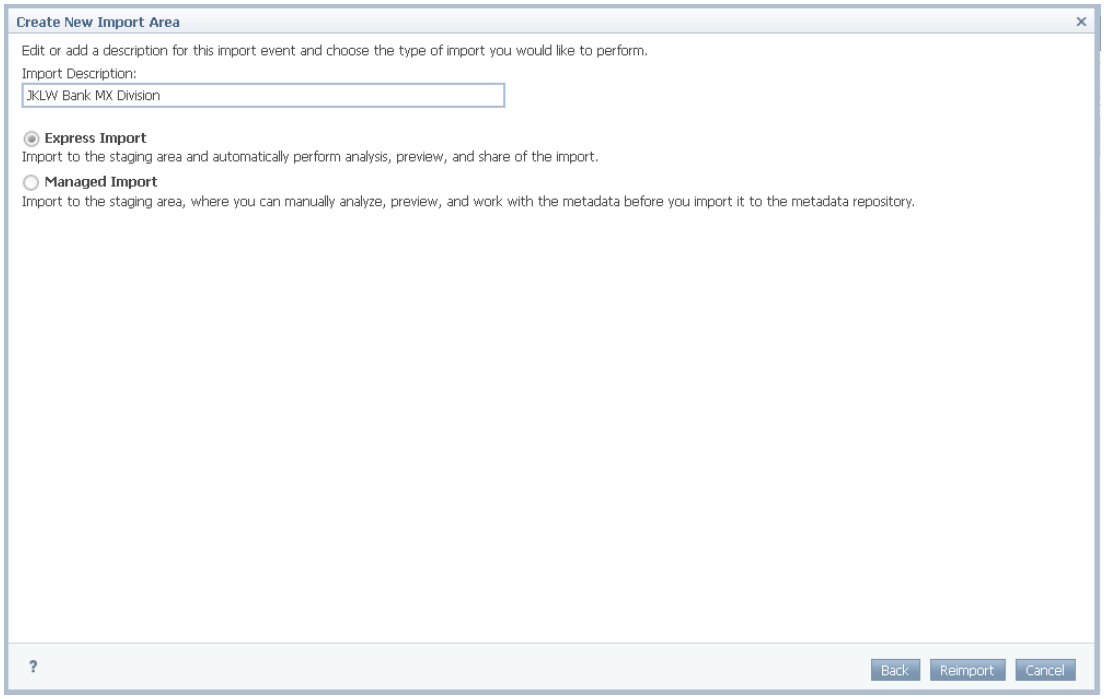
3) Seleccionar la tabla cuya definición se requiere importar.



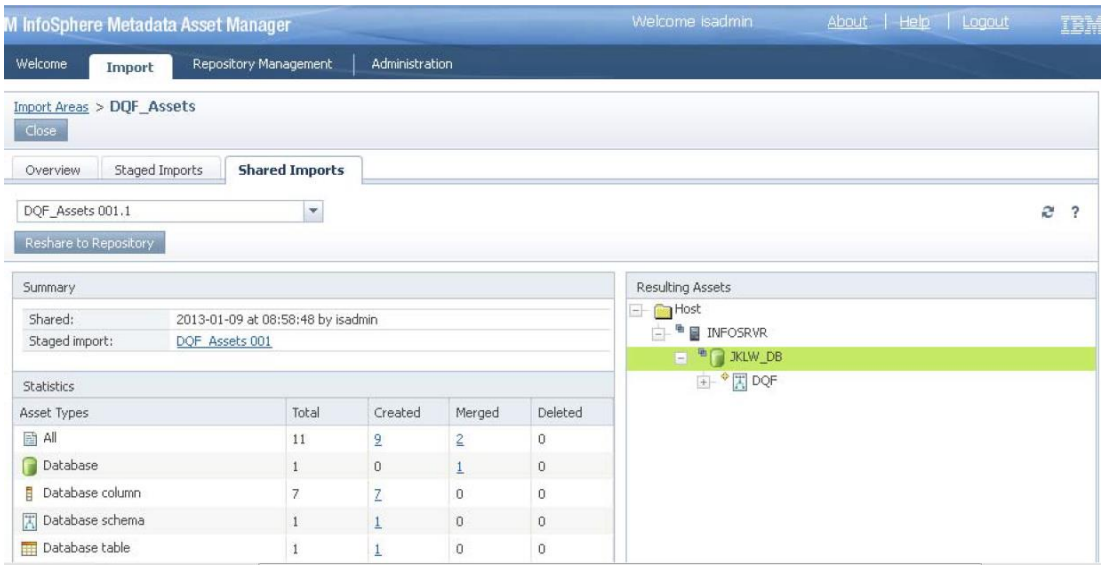
4) Escribir el nombre de la base de datos para guardar la importación.



5) Seleccionar Express Import y escribir un nombre para el evento. Presione Import.



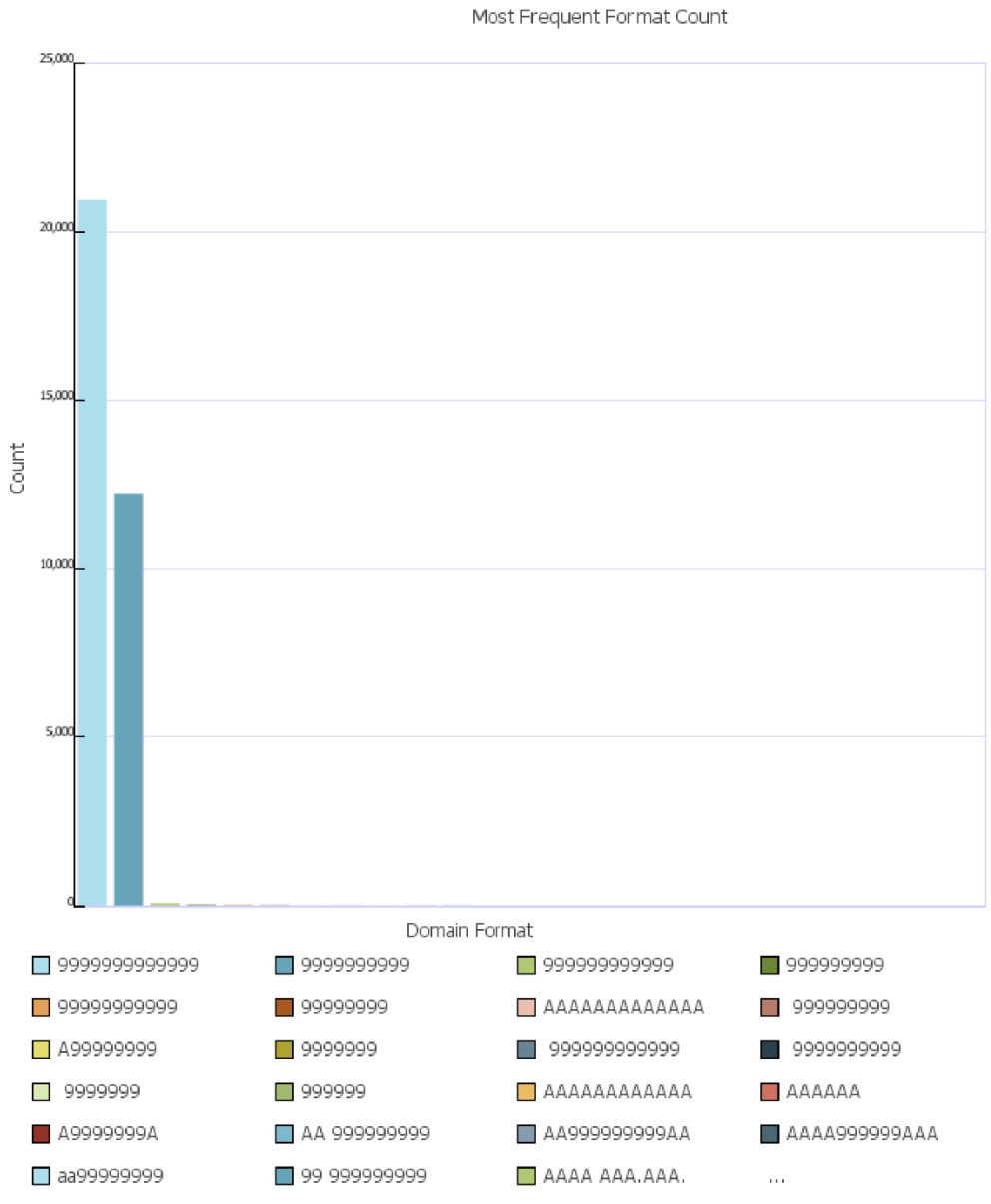
6) Se realizó la importación de los metadatos de la tabla a analizar, si se requiere mayor información visitar el sitio de IBM. (IBM Knowledge Center, 2012)



ANEXO B

REPORTES DE PROCESO INVESTIGATE

Reporte de Formatos más Frecuentes – Cédula Cliente



Column Level Details

Format	Count	Total Rows %	Total Rows Cumulative %	Example Values
9999999999999	20,940	62.76790264	62.76790264	0990017514001 0990004196001 1790319857001 1792239710001 1792060346001 1790710319001 0991306498001 1791997891001 0992142618001 1792129745001
9999999999	12,234	36.67156260	99.43946524	0501566129 1713684999 1309527420 0060284760 0100014513 0100025394 0100025915 0100044379 0100045004 0100149723
999999999999	74	0.22181589	99.66128113	010273794723 999999999999 673705100001 214423360011 204093978471 202131895365 091799205997 091733802296 091715730795 091661894594
999999999	28	0.08393034	99.74521147	033150039 999999999 911608052 895061219 812846057 770424202 710628109 410958766 271419456 205846120

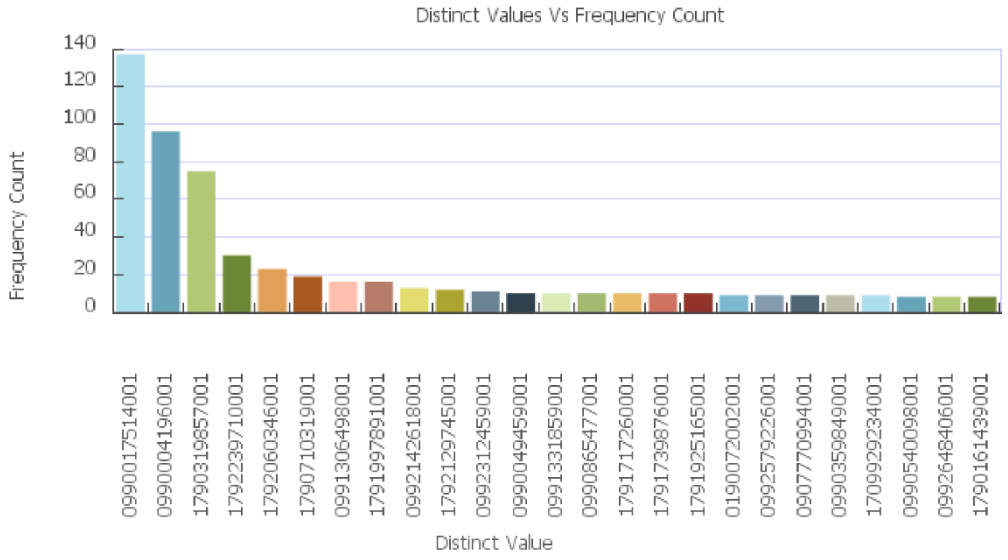
Reporte de Formatos más Frecuentes – Cédula Cliente

Host Name :	INFOSERVER
Data Store :	ERP_Clientes
Data Store Alias :	
Database Name :	BAAN
Database Alias :	
Table Name :	TTCCOM100602
Table Alias :	
Column Name :	T$LGID
Column Alias :	

Column Level Summary

Cardinality Count :	29,933
Null Count :	0
Actual Row Count :	33,361

Total Rows Covered :	2,980
%Rows Covered :	8.9326



Frequency Distribution Data

Distinct Value	Frequency Count	Frequency %	Cumulative %
0990017514001	137	0.41065915	0.41065915
0990004196001	96	0.28776116	0.69842031
1790319857001	75	0.22481340	0.92323371
1792239710001	30	0.08992536	1.01315907
1792060346001	23	0.06894278	1.08210185
1790710319001	19	0.05695273	1.13905458
0991306498001	16	0.04796019	1.18701477
1791997891001	16	0.04796019	1.23497496
0992142618001	13	0.03896766	1.27394262
1792129745001	12	0.03597014	1.30991276
0992312459001	11	0.03297263	1.34288539
0990049459001	10	0.02997512	1.37286051
0991331859001	10	0.02997512	1.40283563
0990865477001	10	0.02997512	1.43281075
1791717260001	10	0.02997512	1.46278587
1791739876001	10	0.02997512	1.49276099
1791925165001	10	0.02997512	1.52273611
0190072002001	9	0.02697761	1.54971372
0992579226001	9	0.02697761	1.57669133
0907770994001	9	0.02697761	1.60366894
0990359849001	9	0.02697761	1.63064655
1709929234001	9	0.02697761	1.65762416
0990540098001	8	0.02398010	1.68160426
0992648406001	8	0.02398010	1.70558436
1790161439001	8	0.02398010	1.72956446
0992106891001	7	0.02098258	1.75054704
1790049795001	7	0.02098258	1.77152962
1791744756001	7	0.02098258	1.79251220
1791242491001	7	0.02098258	1.81349478
1791309863001	7	0.02098258	1.83447736
1792207479001	7	0.02098258	1.85545994
1801909910001	7	0.02098258	1.87644252
0990967946001	6	0.01798507	1.89442759
1791305019001	6	0.01798507	1.91241266
1791334043001	6	0.01798507	1.93039773
1702550680001	6	0.01798507	1.94838280
1703779494001	6	0.01798507	1.96636787
1790007782001	6	0.01798507	1.98435294
1792073812001	6	0.01798507	2.00233801
1792265134001	6	0.01798507	2.02032308
0190311058001	5	0.01498756	2.03531064
0914598461001	5	0.01498756	2.05029820
0990006687001	5	0.01498756	2.06528576

ANEXO C

REPORTES DE PROCESO STANDARIZE

Standardization Quality Assessment (SQA)

Project:	INFOSERVER:ANALYZERPROJECT
Report Name:	SQA Reporte Cliente Nombre
Report Generated:	2014-04-04 11:23:47
Time Zone:	UTC -04:00
User:	dqa dqa
Job Name:	STD_STG_FASE3_QA_CLIENTES
Rule Set:	PENAME
Total Records Processed:	22452

Introduction

The Standardization Quality Assessment consists of two reports that are intended to be viewed as a pair. Together, these reports provide summary-level feedback of the processed output from a Standardization job. The Standardization Quality Assessment (SQA) report provides statistical-level feedback. The companion report, the Standardization Quality Assessment (SQA) Record Examples report, provides examples of processed records.

The following terms are referenced throughout this report and are defined as follows:

Composition Set
A set of records in which each record contains values for the same selected dictionary columns. See the SQA Record Examples report for more details about each composition set that is summarized in this report.

Fully Standardized Record
A record in which all of the selected dictionary columns are populated and the unhandled data column is empty.

Non-standardized Record
A record in which only the unhandled data column is populated and all of the selected dictionary columns are empty.

Partially Standardized Record
A record in which the unhandled data column is populated and all of the selected dictionary columns are populated.

Record Examples
The subset of processed records that are shown in the SQA Record Examples report. The SQA stage produces 50 examples. You can generate reports that contain all 50 of those examples, or you can limit the number of examples by changing the report settings in the Reporting tab of the IBM Information Server Web console.

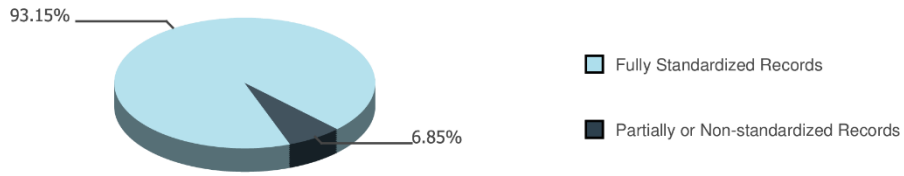
Selected Dictionary Columns
The dictionary columns that were selected in the Standardization stage for assessment in the SQA stage.

Total Records Processed
The total number of records that are processed by the Standardization job and summarized in this report.

Unhandled Data Column
The column that was selected in the SQA stage that reflects the presence of unprocessed data in the record.

Standardization Quality Assessment (SQA)

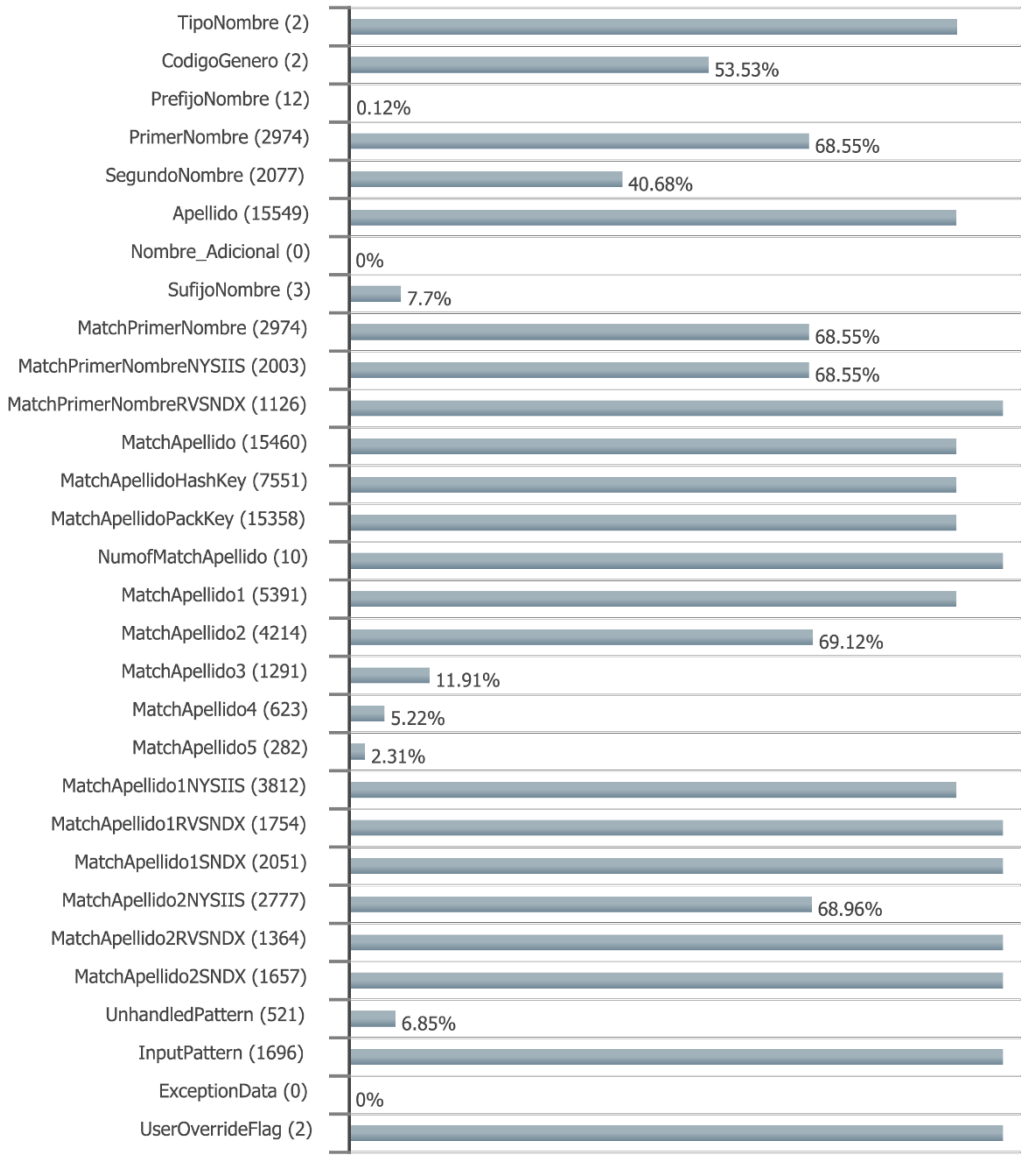
Standardization Summary



Standardization Quality Assessment (SQA)

Frequency of Records by Populated Dictionary Column

Dictionary Column (Unique Values)



Standardization Quality Assessment (SQA)

Composition Sets

Displayed sets comprise 97.44% of the processed records

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
	31.77%	10.61%	8.81%	6.85%	6.11%	5.35%	4.39%	3.90%	3.35%	2.54%
TipoNombre	✔	✔	✔		✔	✔	✔	✔	✔	✔
CodigoGenero	✔	✔	✔							
PrefijoNombre										
PrimerNombre	✔	✔	✔		✔	✔				
SegundoNombre	✔				✔					
Apellido	✔	✔	✔		✔	✔	✔	✔	✔	✔
Nombre_Adicional										
SufijoNombre									✔	
MatchPrimerNombre	✔	✔	✔		✔	✔				
MatchPrimerNombreNYSIIS	✔	✔	✔		✔	✔				
MatchPrimerNombreRVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido	✔	✔	✔		✔	✔	✔	✔	✔	✔
MatchApellidoHashKey	✔	✔	✔		✔	✔	✔	✔	✔	✔
MatchApellidoPackKey	✔	✔	✔		✔	✔	✔	✔	✔	✔
NumofMatchApellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1	✔	✔	✔		✔	✔	✔	✔	✔	✔
MatchApellido2	✔	✔			✔		✔	✔		✔
MatchApellido3							✔			✔
MatchApellido4										✔
MatchApellido5										
MatchApellido1NYSIIS	✔	✔	✔		✔	✔	✔	✔	✔	✔
MatchApellido1RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2NYSIIS	✔	✔			✔		✔	✔		✔
MatchApellido2RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
UnhandledPattern				✔						
InputPattern	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ExceptionData										
UserOverrideFlag	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Standardization Quality Assessment (SQA)

Composition Sets

Displayed sets comprise 97.44% of the processed records

	Set 11	Set 12	Set 13	Set 14	Set 15	Set 16	Set 17	Set 18	Set 19	Set 20
	2.49%	2.27%	2.18%	1.81%	1.68%	1.20%	0.85%	0.43%	0.36%	0.10%
TipoNombre	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CodigoGenero				✔				✔		
PrefijoNombre										
PrimerNombre	✔			✔			✔	✔		
SegundoNombre				✔			✔			
Apellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Nombre_Adicional										
SufijoNombre		✔			✔				✔	
MatchPrimerNombre	✔			✔			✔	✔		
MatchPrimerNombreNYSIIS	✔			✔			✔	✔		
MatchPrimerNombreRVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellidoHashKey	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellidoPackKey	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
NumofMatchApellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2	✔	✔	✔		✔			✔	✔	✔
MatchApellido3			✔		✔			✔	✔	✔
MatchApellido4			✔						✔	✔
MatchApellido5			✔							✔
MatchApellido1NYSIIS	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2NYSIIS	✔	✔	✔		✔			✔	✔	
MatchApellido2RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
UnhandledPattern										
InputPattern	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ExceptionData										
UserOverrideFlag	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Standardization Quality Assessment (SQA)

Composition Sets

Displayed sets comprise 97.44% of the processed records

	Set 21	Set 22	Set 23	Set 24	Set 25	Set 26	Set 27	Set 28	Set 29	Set 30
	0.06%	0.05%	0.04%	0.04%	0.03%	0.02%	0.01%	0.01%	0.01%	0.01%
TipoNombre	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CodigoGenero	✔									
PrefijoNombre				✔		✔	✔	✔		
PrimerNombre	✔	✔	✔	✔		✔	✔	✔	✔	✔
SegundoNombre	✔		✔				✔	✔		
Apellido	✔	✔	✔	✔	✔		✔	✔	✔	✔
Nombre_Adicional										
SufijoNombre					✔					
MatchPrimerNombre	✔	✔	✔	✔		✔	✔	✔	✔	✔
MatchPrimerNombreNYSIIS	✔	✔	✔	✔		✔	✔	✔	✔	✔
MatchPrimerNombreRVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido	✔	✔	✔	✔	✔		✔	✔	✔	✔
MatchApellidoHashKey	✔	✔	✔	✔	✔		✔	✔	✔	✔
MatchApellidoPackKey	✔	✔	✔	✔	✔		✔	✔	✔	✔
NumofMatchApellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1	✔	✔	✔	✔	✔		✔	✔	✔	✔
MatchApellido2	✔	✔	✔		✔		✔		✔	✔
MatchApellido3	✔	✔	✔		✔					
MatchApellido4					✔					
MatchApellido5					✔					
MatchApellido1NYSIIS	✔	✔	✔	✔	✔		✔	✔	✔	
MatchApellido1RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2NYSIIS	✔	✔	✔		✔		✔			✔
MatchApellido2RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
UnhandledPattern										
InputPattern	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ExceptionData										
UserOverrideFlag	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Standardization Quality Assessment (SQA)

Composition Sets

Displayed sets comprise 97.44% of the processed records

	Set 31	Set 32	Set 33	Set 34	Set 35	Set 36	Set 37	Set 38	Set 39	Set 40
	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%
TipoNombre	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CodigoGenero		✔	✔	✔						✔
PrefijoNombre		✔	✔		✔	✔				✔
PrimerNombre		✔	✔	✔		✔				✔
SegundoNombre			✔							
Apellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Nombre_Adicional										
SufijoNombre									✔	
MatchPrimerNombre		✔	✔	✔		✔				✔
MatchPrimerNombreNYSIIS		✔	✔	✔		✔				✔
MatchPrimerNombreRVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellidoHashKey	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellidoPackKey	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
NumofMatchApellido	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2	✔			✔		✔	✔	✔	✔	✔
MatchApellido3	✔			✔		✔	✔		✔	
MatchApellido4	✔								✔	
MatchApellido5										
MatchApellido1NYSIIS	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido1SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2NYSIIS										✔
MatchApellido2RVSNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MatchApellido2SNDX	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
UnhandledPattern										
InputPattern	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ExceptionData										
UserOverrideFlag	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Standardization Quality Assessment (SQA)

Composition Sets

Displayed sets comprise 97.44% of the processed records

	Set 41	Set 42	Set 43	Set 44						
	0.00%	0.00%	0.00%	0.00%						
TipoNombre	✔	✔	✔	✔						
CodigoGenero	✔	✔	✔							
PrefijoNombre	✔									
PrimerNombre	✔	✔	✔	✔						
SegundoNombre			✔	✔						
Apellido		✔	✔							
Nombre_Adicional										
SufijoNombre										
MatchPrimerNombre	✔	✔	✔	✔						
MatchPrimerNombreNYSIIS	✔	✔	✔	✔						
MatchPrimerNombreRVSNDX	✔	✔	✔	✔						
MatchApellido		✔	✔							
MatchApellidoHashKey		✔	✔							
MatchApellidoPackKey		✔	✔							
NumofMatchApellido	✔	✔	✔	✔						
MatchApellido1		✔	✔							
MatchApellido2		✔	✔							
MatchApellido3										
MatchApellido4										
MatchApellido5										
MatchApellido1NYSIIS		✔	✔							
MatchApellido1RVSNDX	✔	✔	✔	✔						
MatchApellido1SNDX	✔	✔	✔	✔						
MatchApellido2NYSIIS										
MatchApellido2RVSNDX	✔	✔	✔	✔						
MatchApellido2SNDX	✔	✔	✔	✔						
UnhandledPattern										
InputPattern	✔	✔	✔	✔						
ExceptionData										
UserOverrideFlag	✔	✔	✔	✔						

ANEXO D

REPORTES DE PROCESO MATCH

Home

Match Statistics Report

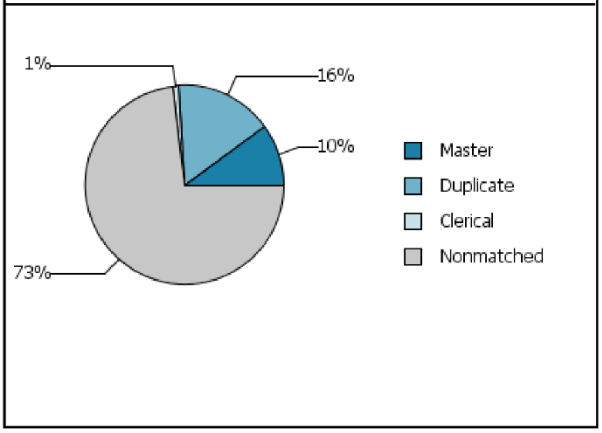
Executive Summary

Match Specification: MTC_CLIENTES_MATCH_SPECIFICATION

Project:	INFOSERVER:ANALYZERPROJECT
Report Name:	MTC Statistics Reporte Cliente
Report Generated:	2014-04-04 04:26:46
Time Zone:	UTC -04:00
User:	dqa dqa
Description:	Presents summary statistics about the matching results and statistics about the matching process for each match pass.

Total Statistics

Output Categories



Type	Count	Rate
Record	22,452	100%
Master	2,178	10%
Duplicate	3,491	16%
Clerical	170	1%
Nonmatched	16,613	73%
Matched	5,669	25%
Possible Matched	5,839	26%
Unique Entity	18,791	84%

Note: Rates are rounded.

Glossary

Master
A record that represents a group of matching records. Each group of two or more records has one master record.

Duplicate
A record that matches a master record. The duplicate record represents the same unique entity as the master record.

Clerical
A record that might be a duplicate. The matching process does not have enough information to be certain.

Nonmatched
A record that is not a master, duplicate, or clerical record.

Matched
A master or duplicate record.

Possible Matched
A master, duplicate, or clerical record.

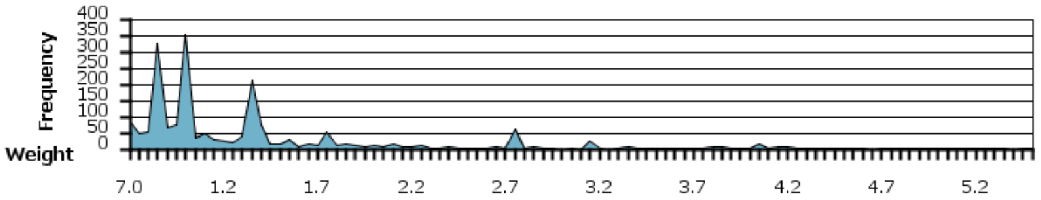
Unique Entity
A distinct real-world object that is represented by a group of one or more records.

Match Pass: 1 of 2

Match Specification: MTC_CLIENTES_MATCH_PASS

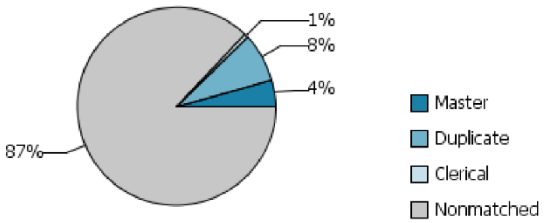
Project:	INFOSERVER:ANALYZERPROJECT
Report Name:	MTC Statistics Reporte Cliente
Report Generated:	2014-04-04 04:26:50
Time Zone:	UTC -04:00
User:	dqa dqa
Description:	Presents summary statistics about the matching results and statistics about the matching process for each match pass.

Weights for Duplicate and Clerical Pairs



Statistics

Pass Statistics



Records

Type	Count
Input Records	22,452
Master	1,021
Duplicate	1,858
Exact Duplicate	9
Clerical	135
Residual	20,459
Total Comparisons	1,993

Blocks

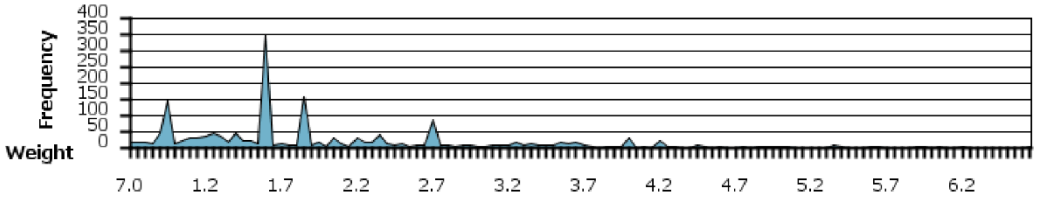
Type	Count
Input Blocks	3,174
Overflowed	0
Maximum Size	9,987
Average Size	3

Match Pass: 2 of 2

Match Specification: MTC_CLIENTES_MATCH_PASS_2

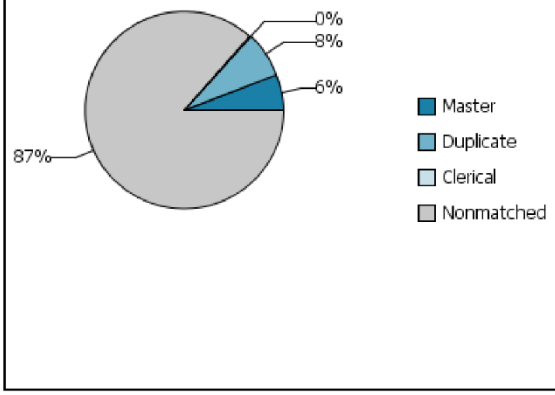
Project:	INFOSERVER:ANALYZERPROJECT
Report Name:	MTC Statistics Reporte Cliente
Report Generated:	2014-04-04 04:26:53
Time Zone:	UTC -04:00
User:	dqa dqa
Description:	Presents summary statistics about the matching results and statistics about the matching process for each match pass.

Weights for Duplicate and Clerical Pairs



Statistics

Pass Statistics



Records

Type	Count
Input Records	20,459
Master	1,242
Duplicate	1,633
Exact Duplicate	14
Clerical	35
Residual	18,791
Total Comparisons	1,668

Blocks

Type	Count
Input Blocks	3,638
Overflowed	0
Maximum Size	1,748
Average Size	5