Universidad Nacional Autónoma de México

Facultad de Ciencias, 2024 - 1

Almacenes y Minería de Datos

Tarea 02. Procesos de Extracción, Transformación y Carga

PostgresandoesoSQLazos

Integrantes

Adrian Aguilera Moreno	421005200
Marco Antonio Rivera Silva	318183583
Sebastían Alejandro Gutiérrez Medina	318287021
Israel Hernández Dorantes	318206604
Alejandra Ortega García	420002495

1. Lectura del Articulo

1.1. Resumen

Este artículo discute el concepto de limpieza de datos, también conocido como limpieza de datos o depuración, que implica identificar y corregir errores e inconsistencias en conjuntos de datos para mejorar la calidad de los datos. Los problemas de calidad de datos a menudo surgen debido a errores de ortografía, información faltante u otros datos no válidos en colecciones de datos individuales como archivos y bases de datos. Cuando se integran múltiples fuentes de datos, como en almacenes de datos o sistemas de bases de datos federadas, la necesidad de limpieza de datos se vuelve aún más significativa debido a la redundancia de datos y representaciones variadas.

Los almacenes de datos, en particular, requieren procesos sólidos de limpieza de datos, ya que manejan grandes volúmenes de datos de diversas fuentes y la precisión de los datos es crucial para la toma de decisiones informadas. La limpieza de datos generalmente se realiza como parte del proceso ETL (extracción, transformación, carga) antes de cargar los datos en el almacén. Aunque existen varias herramientas para respaldar estas tareas, a menudo se requiere intervención manual y programación de bajo nivel.

También se destaca la importancia de la limpieza de datos en sistemas de bases de datos federadas y sistemas de información basados en la web, que enfrentan desafíos similares de transformación de datos al extraer, transformar e integrar datos de múltiples fuentes. La limpieza de datos es esencial para lograr resultados precisos y eficientes en consultas en tales sistemas.

Algunos de los requisitos clave para un enfoque de limpieza de datos incluyen la detección y eliminación de errores e inconsistencias importantes, el soporte de herramientas para reducir el esfuerzo manual, la capacidad de ampliación para cubrir fuentes de datos adicionales y la integración con transformaciones de datos relacionadas con el esquema. Es fundamental la especificación declarativa de funciones de mapeo y una infraestructura de flujo de trabajo para la limpieza de datos en almacenes de datos.

A pesar de la importancia de la limpieza de datos, se le ha prestado menos atención en comparación con la traducción e integración de esquemas. Sin embargo, algunos esfuerzos de investigación se han centrado en la identificación y eliminación de duplicados, enfoques de minería de datos, transformaciones de datos basadas en la coincidencia de esquemas y soluciones integrales de limpieza de datos que abarcan múltiples fases de transformación y operadores.

Asimismo, se da una descripción general de los problemas de limpieza de datos, categorizándolos en problemas de fuente única y fuente múltiple, así como problemas relacionados con el esquema y la instancia y resalta en la importancia del diseño del esquema y las restricciones de integridad para prevenir datos incorrectos en fuentes individuales y discute cómo estos problemas pueden volverse más complejos al integrar múltiples fuentes con representaciones de datos variadas.

1.2. Preguntas

- **Sebastián Alejandro Gutiérrez Medina**

1. ¿Qué es la limpieza de datos? ¿Qué importancia tiene la limpieza en el proceso ETL y para la implementación de un DWH?

   La limpieza de datos es un proceso que busca mejorar la calidad de los datos y se utiliza para corregir o eliminar registros inexactos en una base de datos o conjunto de datos, esto significa identificar y sustituir los datos o registros incompletos, inexactos, corruptos o irrelevantes, después de una limpieza de datos correctamente realizada, todos los conjuntos de datos deben ser coherentes y estar libres de errores.

   La limpieza de datos es muy importante en el proceso ETL, ya que asegura la calidad de los datos que se van a procesar, evita la información no veraz o errónea, ahorra costes de espacio en disco al eliminarse la información duplicada y agiliza las consultas por la ausencia de datos repetidos o inservibles.

   En cuanto a la implementación de un DWH, la limpieza de datos garantiza que los datos almacenados sean precisos y confiables, un DWH alimentado con datos limpios permite tomar decisiones estratégicas correctas basadas en información precisa y útil. sin una etapa previa de limpieza de datos, no es posible disponer de una base de datos de calidad que permita la toma de decisiones acertadas a nivel estratégico o ejecutivo.

2. ¿Cuál es el objetivo de la limpieza de datos?

   Su objetivo es corregir o eliminar registros inexactos en una base de datos o conjunto de datos, esto significa identificar y sustituir los datos o registros incompletos, inexactos, corruptos o irrelevantes

3. ¿Qué significa Calidad de Datos?

   La calidad de datos se refiere a la capacidad de los datos para cumplir con ciertos criterios específicos, incluyendo precisión, integridad, consistencia, relevancia, actualidad y validez, mide hasta qué punto cumplen los conjuntos de datos con los criterios de exactitud, completitud, validez, coherencia, unicidad, oportunidad y adecuación a un propósito.

4. ¿Qué significa Gobierno de Datos?

   El Gobierno de Datos es una estructura organizativa que apoya la gestión de los datos empresariales, consiste en un conjunto de normas, políticas y procesos que permiten asegurar que los datos de la organización sean correctos, fiables, seguros y útiles.

   Estos procesos determinan los propietarios ,medidas de seguridad y los usos previstos para los datos, el objetivo del gobierno de datos es mantener datos de alta calidad que sean seguros y

fácilmente accesibles para extraer información de negocio más detallada

5. ¿Cuáles son los problemas que enfrenta hoy en día la Limpieza de Datos?

   Los problemas que enfrenta son principalmente la calidad de los datos, que se pueden clasificar como de una sola fuente de datos o de multiples fuentes, a su vez, cada una se puede subdividir en problemas a nivel de esquema y a nivel de instancia.

   Por la parte de una fuente individual de datos, los problemas a nivel de esquema se refieren a la falta de restricciones de integridad y/o mal diseño del esquema, por otro lado, los problemas a nivel de instancia se refieren a los problemas en la entrada de datos, como fallas ortográficas o valores contradictorios.

   En cuanto a los problemas de múltiples fuentes, los de esquema se refieren a modelos heterogéneos de datos y esquemas, como problemas entre varios nombres para los mismos datos o problemas estructurales, asimismo, los problemas a nivel de instancia se refieren a datos contradictorios, inconsistentes o que se encimen unos a otros.

6. ¿Qué enfoques aborda para solventar dichos problemas?

   Para una sola fuente de datos la prevención de producir datos sucios es fundamental, diseñando apropiadamente los esquemas de la BD y sus restricciones de integridad para la entrada de datos.

   Para múltiples fuentes de datos la principal manera de prevenir datos sucios es una correcta integración de esquemas y procesos de limpieza de datos, en particular la detección de de duplicados basados en un representación uniforme de nombres de atributos.

7. ¿Qué es el análisis de datos y cómo se puede utilizar para apoyar las tareas de limpieza de datos?

   El análisis de datos es un proceso que implica inspeccionar, limpiar, transformar y modelar datos con el objetivo de descubrir información útil, informar conclusiones y apoyar la toma de decisiones.

   Puede ser utilizado en la limpieza de datos de las siguientes maneras :

   - Identificación de errores: Las herramientas de análisis pueden ser utilizadas para inspeccionar sistemáticamente los datos, utilizando diferentes reglas y algoritmos para identificar valores no válidos. Esto puede incluir la detección de errores de sintaxis, valores atípicos y puntos de datos corruptos.
   - Corrección de errores: Una vez que se han identificado los errores, el análisis de datos puede ayudar a corregirlos. Esto puede implicar la eliminación o corrección de entradas problemáticas.
   - Eliminación de duplicados: El análisis de datos también puede ser utilizado para identificar y eliminar registros duplicados4. Esto es especialmente útil cuando se trabaja con grandes conjuntos de datos.

- Transformación de datos: El análisis de datos puede ayudar a transformar los datos en un formato adecuado para el análisis2. Esto puede implicar la conversión de los datos a un formato estándar o la normalización de los valores

8. ¿De qué forma los procesos ETL ayudan a efectuar la Limpieza de Datos?

   Los procesos ETL ayudan a asegurar la calidad de los datos que se van a procesar, evitan la información no veraz o errónea, ahorran costes de espacio en disco al eliminar la información duplicada y agilizan las consultas por la ausencia de datos repetidos o inservibles.

9. ¿Por qué crees que se considera que la etapa ETL consume entre el 60% y el 80% de los recursos en un proyecto de implementación de DWH? ¿Cuál sería el proceso ETL que te sería más complicado llevar a cabo sí lo tuvieras que realizar?

   Debido a la complejidad y tiempo que puede tomar, principalmente debido a la transformación de datos pues dependiendo de cuantos problemas tengan, mas tiempo sera necesario para poder limpiarlos, siendo esto así, considero que la transformación de los datos es el proceso ETL mas complicado.

10. ¿Qué mecanismos propondrías para eliminar o minimizar el impacto de la mala calidad de datos?

    Revisiones periódicas a la calidad de los datos, así como de tratar de reducir las posibles entradas erróneas de datos y al mismo tiempo tener mecanismos ya establecidos para la resolución de datos sucios y no solo intentar desarrollarlos cuando los errores sean detectados.

- **Adrian Aguilera Moreno**

1. ¿Qué es la limpieza de datos? ¿Qué importancia tiene la limpieza en el proceso ETL y para la implementación de un DWH?

   La limpieza de datos busca filtrar la información de nuestra base de datos. Entendemos como información sucia a la información que contiene datos no correspondientes a su asignación o datos que desprecian valores y por tanto no se puede generar información a partir de estos. Limpiar los datos es el proceso en el que eliminamos y/o filtramos esta los datos que no nos permiten generar información concreta.

   En un ETL es necesario realizar una limpieza de nuestros datos para poder generar información concreta y correcta. En un DWH al igual que en un ETL necesitamos una limpieza de los datos para poder generar un conocimiento correcto.

2. ¿Cuál es el objetivo de la limpieza de datos?

   Corregir. eliminar, y suplir datos erróneos en nuestros procesos ETL y/o DWH. La información que generan los procesos ETL o el conocimiento que se crea a partir de un DWH depende en gran parte de tener datos libres de errores y en la medida de lo posible este proceso reducirá el sesgo al preprocesar nuestros datos.

3. ¿Qué significa Calidad de Datos?

   La calidad de datos es la manera en la que clasificamos el nivel de integridad, consistencia y veracidad de los datos. Datos "libres de errores" suelen tener un buen porcentaje de calidad de datos. Sin embargo, puede que después de nuestro proceso de limpieza de datos nuestros datos no logren una calidad máxima. esto no quiere decir que no lo hayamos hecho bien. Un ejemplo sería realizar la limpieza de nuestros datos y que aún así haya elementos que no precisan y en el caso de los procesos ETL podrían quedar desactualizados. En el caso de los DWH podríamos tener valores aún inexactos pero que cumplen con el formato correspondiente para su celda precisa.

4. ¿Qué significa Gobierno de Datos?

   Cómo se mencionó en clase, el gobierno de los datos hace referencia a las políticas de la organización con la que se este trabajando. Estas políticas están inmersas en los siguientes puntos:

   - Calidad de los datos.
   - Seguridad de los datos.
   - Privacidad de los datos.
   - Responsabilidad y gobierno de los datos (cumplimiento de las normas estipuladas por la organización).

   Y aunque cada organización decide cuáles serán sus políticas a trabajar, es cierto que los puntos anteriores son un esqueleto común entre las organizaciones que manejan sus datos por medio de políticas (el efecto de estas es el gobierno de los datos).

5. ¿Cuáles son los problemas que enfrenta hoy en día la Limpieza de Datos?

   Existen diversos problemas que afectan el proceso de limpieza de datos, y lo vuelven en muchas ocasiones un proceso más tedioso. Estos problemas están relacionados a la *veracidad de los datos* (complicado observar estos errores si no tenemos un referéndum) y la *fragmentación* de los datos.

   Con el primer problema hay poco que hacer, pues asumimos que nuestras fuentes de datos son veraces. Mientras que con la fragmentación tenemos soluciones que suelen no ser tan triviales en ocasiones y conllevan actividades repetitivas. La fragmentación de los datos es uno de los problemas más comunes al momento de crear un DWH y posiblemente es el que conlleva más tiempo resolver, pues hay que buscar soluciones equivalentes para cada una de las versiones.

   El principal problema de las limpieza de datos son los errores o imprecisiones que puedan tener nuestras fuentes de datos y es, de hecho, por esto que solemos tomarnos un tiempo para este proceso tan importante que es la limpieza de datos.

6. ¿Qué enfoques aborda para solventar dichos problemas?

   Dependiendo el proceso que estemos trabajando podemos dar una respuesta u otra, por ejemplo:

   - Para procesos OLTP solemos normalizar y filtrar valores erróneos.
   - Para nuestros DWH no nos intersa preservar la normalización de sus fuentes y de hecho deseamos denormalizar. Aquí buscamos integrar las fuentes de la mejor manera posible y eliminar los errores más notorios en nuestras dimensiones (dividimos en partes atómicas respecto a lo que queramos realizar y unimos por medio de hechos y dimensiones).

7. ¿Qué es el análisis de datos y cómo se puede utilizar para apoyar las tareas de limpieza de datos?

   El análisis de datos es el proceso que filtra, elimina errores, y modela los datos de alguna manera. En este proceso va inmerso cómo proceso previo la limpieza de los datos, si al momento de realizar nuestro análisis llegamos a visualizar fallas, podemos regresar al preproceso de limpieza y, sencillamente, modificar los criterios que estén generando las fallas.
8. ¿De qué forma los procesos ETL ayudan a efectuar la Limpieza de Datos?

9. ¿Por qué crees que se considera que la etapa ETL consume entre el 60% y el 80% de los recursos en un proyecto de implementación de DWH? ¿Cuál sería el proceso ETL que te sería más complicado llevar a cabo sí lo tuvieras que realizar?

   Los procesos de extracción, transformación y carga añaden la información a nuestros DWH y son los encargados de filtrar, eliminar, modificar, añadir los datos que vayan pasando el proceso de limpieza, digamos, los datos que cumplen tener una gran calidad inmersa en ellos.
10. ¿Qué mecanismos propondrías para eliminar o minimizar el impacto de la mala calidad de datos?

    - Diccionarios de tuplas a tuplas. Cómo hemos observado, en ocasiones queremos transformar varios datos a diferentes dominios posibles, es aquí cuando podríamos tener mapeos que nos relacionen datos.
    - Especialización de reglas de limpieza de datos. Cómo hemos notado, nuestros podemos agrupar procesos en, por ejemplo, metanodos. Sin embargo, para reglas parecidas que se deban modificar en lo mínimo tenemos que reescribirlas, esto va encontra de los principios de reutilización de código.
    - Una buena sugerencia es que nuestro proceso de limpieza sea cíclico y poder retornar a el cuando sea necesario.

- **Alejandra Ortega García**

1. ¿Qué es la limpieza de datos? ¿Qué importancia tiene la limpieza en el proceso ETL y para la implementación de un DWH?

   La **limpieza de datos**, consiste en detectar y eliminar errores e inconsistencias de los datos para mejorar su calidad. En el **proceso ETL**, donde se extraen datos de múltiples fuentes, la limpieza de datos es esencial para asegurar que los datos sean compatibles y puedan integrar de manera efectiva.
   Para la **implementación de un DWH** la limpieza de datos es fundamental, pues estos sistemas se utilizan para la toma de decisiones, por lo que la exactitud de sus datos es vital para evitar conclusiones erróneas.

2. ¿Cuál es el objetivo de la limpieza de datos?

   El **objetivo** principal es **mejorar la calidad de los datos**, al detectar y eliminar todos los errores e inconsistencias importantes tanto en fuentes de datos individuales como al integrar múltiples fuentes, podemos garantizar que los datos sean confiables y útiles para su uso en análisis y toma de decisiones.

3. ¿Qué significa Calidad de Datos?

   La **Calidad de los Datos** mide hasta qué punto los datos son exactos, consistentes, completos, confiables y relevantes para satisfacer las necesidades y objetivos.

4. ¿Qué significa Gobierno de Datos?

   El **Gobierno de Datos** se refiere a un conjunto de políticas, procesos y estándares; fomenta la disponibilidad, la calidad y la seguridad de los datos de una organización.

5. ¿Cuáles son los problemas que enfrenta hoy en día la Limpieza de Datos?

   Algunos de los problemas a los que se enfrenta están entre problemas de que provienen de una sola fuente y problemas que provienen múltiples fuentes, así como entre problemas relacionados con esquemas e instancias.

   Los problemas a nivel de esquema pueden tratarse a nivel de esquema mediante un diseño de esquema mejorado, traducción de esquema e integración de esquema. Mientras que los problemas a nivel de instancia se refieren a errores e inconsistencias en el contenido real de los datos que no son visibles a nivel de esquema.

6. ¿Qué enfoques aborda para solventar dichos problemas?

   Esto requiere un diseño adecuado del esquema de la base de datos y de las restricciones de integridad, así como de las aplicaciones de entrada de datos. Además, el diseño del almacén permite descubrir reglas de limpieza de datos que pueden sugerir mejoras de las restricciones impuestas por los esquemas existentes.

   Hay que eliminar la información duplicada y consolidar y integrar la información complementaria para obtener una visión coherente de las entidades del mundo real.

7. ¿Qué es el análisis de datos y cómo se puede utilizar para apoyar las tareas de limpieza de datos?

   El **análisis de datos** es el proceso de examinar, limpiar, transformar e interpretar conjuntos de datos para obtener información valiosa, detectar patrones, tendencias o relaciones, y tomar decisiones.

   Llevar a cabo un análisis dentro de las tareas de **limpieza de datos**, ayuda a detectar los tipos de errores e inconsistencias hay que eliminar. Así, podemos obtener metadatos reales sobre las características de los datos o patrones de valores inusuales.

8. ¿De qué forma los procesos ETL ayudan a efectuar la Limpieza de Datos?

   Durante la etapa de **extracción**, se pueden identificar datos inconsistentes, faltan tes o duplicados, lo que proporciona la primera oportunidad para identificar problemas de calidad de datos.

   En el proceso de **transformación** de **ETL** es donde ocurre la mayor parte de la limpieza de datos, se identifican los errores y se eliminan o se corrigen.

   Para el proceso de **carga**, se pueden aplicar restricciones de integridad de datos y validaciones adicionales para garantizar que solo los datos limpios y válidos se almacenen en la base de datos.

9. ¿Por qué crees que se considera que la etapa ETL consume entre el 60% y el 80% de los recursos en un proyecto de implementación de DWH? ¿Cuál sería el proceso ETL que te sería más complicado llevar a cabo sí lo tuvieras que realizar?

   Considero que la etapa más complicada de llevar a cabo, es la de **transformación**, ya que se necesitan aplicar distintas modificaciones para mejorar la calidad de los datos. Asimismo, esta etapa consume más recursos, ya que implica detectar y corregir errores, eliminar duplicados, garantizar la consistencia y la integridad de los datos.

10. ¿Qué mecanismos propondrías para eliminar o minimizar el impacto de la mala calidad de datos?

    - Establecer políticas para mantener la calidad de los datos, que describan los estándares que deben cumplir los datos.
    - Normalizar y estandarizar los datos para que sigan un formato coherente.

- **Marco Antonio Rivera Silva**

1. ¿Qué es la limpieza de datos? ¿Qué importancia tiene la limpieza en el proceso ETL y para la implementación de un DWH?

   La **limpieza de datos** son una serie de varios procesos destinados a mejorar la calidad de los datos. Estos procesos se utilizan para corregir o eliminar registros inexactos en una base de datos o conjunto de datos. En general, esto significa identificar y sustituir los datos o registros incompletos, inexactos, corruptos/dañados o irrelevantes.
   La limpieza de datos tiene un papel muy importante en el proceso ETL, estas son algunas razones:

   - Asegura la calidad de los datos con los que trabajamos.
   - Previene problemas legales.
   - Facilita el análisis.

   En cuanto a la implementación de un DWH, el proceso ETL nos proporciona las mismas ventajas anteriormente mencionadas, además de que reduce los costos y ayuda a que se cumplan las normativas de un DWH.

2. ¿Cuál es el objetivo de la limpieza de datos?

   El objetivo de la limpieza de datos es mejorar la calidad y la integridad de los datos almacenados en nuestro sistema o base de datos. La limpieza de datos implica la identificación, corrección y eliminación de errores, inconsistencias y valores incorrectos o no válidos en los conjuntos de datos.

3. ¿Qué significa Calidad de Datos?

   La calidad de los datos mide hasta qué punto cumplen los conjuntos de datos con los criterios de exactitud, completitud, validez, coherencia, unicidad, oportunidad y adecuación a un propósito, y es fundamental para todas las iniciativas de gobierno de datos de una organización. Los estándares de calidad de datos garantizan que las empresas tomen las decisiones basadas en datos que les permitan cumplir sus objetivos de negocio.

4. ¿Qué significa Gobierno de Datos?

   El gobierno de datos es un enfoque riguroso para gestionar los datos durante su ciclo de vida, desde la adquisición hasta la eliminación, pasando por el uso. El gobierno de datos es imprescindible en todas las empresas de todos los sectores, ya que los datos se han ido convirtiendo en su recurso más valioso al avanzar en su transformación digital.

5. ¿Cuáles son los problemas que enfrenta hoy en día la Limpieza de Datos?

   Algunos de los problemas que enfrenta la limpieza de datos en la actualidad son:

   - El volumen de datos, pues con el crecimiento exponencial de la cantidad de datos generados y almacenados por las organizaciones, la limpieza de datos cada vez necesita de mas tiempo y control.
   - La gran cantidad de medios de donde viene la información, ya que ahora hay muchos medios como aplicaciones, APIs, redes sociales, etc.
   - Que los datos no se encuentren estructurados de forma correcta. Esto dificulta mucho la limpieza de los datos.

6. ¿Qué enfoques aborda para solventar dichos problemas?

   Puede haber muchos enfoques, pero los principales que podemos tomar seria una mejor estandarización de los procesos, tanto de limpieza de datos como de su estructuración, pues muchas organizaciones no siguen los mismos estándares. Otro posible enfoque seria una mejor automatización, pues también varios de los problemas que se tienen al limpiar datos son generados por el error humano.

7. ¿Qué es el análisis de datos y cómo se puede utilizar para apoyar las tareas de limpieza de datos?

   El análisis de datos se refiere al proceso de examinar, limpiar, transformar y modelar datos con el objetivo de descubrir patrones, tendencias, relaciones y conocimientos significativos. El análisis de datos se utiliza para obtener información valiosa a partir de conjuntos de datos y tomar decisiones informadas basadas en evidencia. A menudo, el análisis de datos implica la aplicación de técnicas estadísticas y herramientas de software para explorar, visualizar y entender los datos en profundidad.
   Se puede utilizar para apoyar a la limpieza de datos para identificar posibles datos que no se usen, mejorar los procesos de limpieza, al igual que reducir el tiempo de limpieza.

8. ¿De qué forma los procesos ETL ayudan a efectuar la Limpieza de Datos?

   ETL automatiza las tareas de procesamiento de datos repetibles para un análisis eficiente. Las herramientas ETL automatizan el proceso de migración de datos y pueden configurarlas para integrar cambios de datos periódicamente o incluso en tiempo de ejecución. Como resultado, los ingenieros de datos pueden dedicar más tiempo a innovar y menos tiempo a administrar tareas tediosas como mover, formatear y limpiar datos datos.

9. ¿Por qué crees que se considera que la etapa ETL consume entre el 60% y el 80% de los recursos en un proyecto de implementación de DWH? ¿Cuál sería el proceso ETL que te sería más complicado llevar a cabo sí lo tuvieras que realizar?

Debido a la gran cantidad de datos a procesar, pues el crecimiento exponencial de los datos han hecho que la complejidad de extraerlos, transformarlos, cargarlos e integrarlos incremente de igual forma exponencial.

10. ¿Qué mecanismos propondrías para eliminar o minimizar el impacto de la mala calidad de datos?

    Utilizar fuentes que vengan de fuentes confiables y que sigan un estándar óptimo para el manejo de los datos. También automatizar los procesos de manera mucho mas eficientes pero sin dejar de lado la labor humana para supervisar los casos en que la automatización pueda llegar a fallar.

- **Israel Hernández Dorantes**

1. ¿Qué es la limpieza de datos? ¿Qué importancia tiene la limpieza en el proceso ETL y para la implementación de un DWH?

   La **limpieza de datos** es el proceso de detectar y eliminar errores e inconsistencias en los datos para poder mejorar su calidad. La importancia en el proceso **ETL** es que la limpieza desempeña un paso fundamental en este proceso ya que el corregir errores en los datos como duplicados y mejorar su calidad ayuda a que se pueda hacer un mejor análisis de la información, y por consecuente evitar errores al **implementar un DWH**.

2. ¿Cuál es el objetivo de la limpieza de datos?

   Su objetivo es **mejorar la calidad de los datos** detectando y eliminando todos los errores e inconsistencias que se encuentren en una o varias fuentes de datos.

3. ¿Qué significa Calidad de Datos?

   Se refiere al grado en el que los datos son precisos, confiables, completos, consistentes y relevantes para garantizar que las decisiones con estos datos sean las correctas.

4. ¿Qué significa Gobierno de Datos?

   Es un conjunto de políticas y reglas establecidas que permiten gestionar los datos durante su ciclo de vida; regulando cómo se recopilan, almacenan, procesan y eliminan los datos.

5. ¿Cuáles son los problemas que enfrenta hoy en día la Limpieza de Datos?

   Los principales problemas o dificultades que puede enfrentar la limpieza de datos hoy en día son:

   - **Mala estructuración en los datos**, ya que ante ellos se necesitan técnicas específicas que dificultan la limpieza.
   - **Grandes volúmenes de datos**, ya que su limpieza a realizar procesos no muy eficientes.
   - **Gran complejidad en los datos**, ya que si se tienen varios datos de diferente tipo, estos requerirán de ”limpiezas” específicas, lo cual implicará mayor trabajo y menos eficiencia en el proceso de limpieza.
   - **Diferentes fuentes de datos**, ya que pueden existir inconsistencias como tener valores repetidos, valores fuera de los rangos, valores faltantes, valores contradictorios, etc.; los cuales hacen más difícil la tarea de limpiar datos.

6. ¿Qué enfoques aborda para solventar dichos problemas?

   Uno de los principales son tener una validación y verificación de los datos, para que de este modo se apliquen ciertas reglas que comprueben y validen la integridad de los datos y se resuelvan varios problemas que deriven de inconsistencias en los datos. Otro enfoque es el hacer uso de herramientas especializadas que automaticen el proceso de limpieza, siendo de esta manera más eficiente y costando menos.

7. ¿Qué es el análisis de datos y cómo se puede utilizar para apoyar las tareas de limpieza de datos?

   Es un proceso que explora, transforma y examina datos que permite encontrar información relevante para la toma de decisiones para poder garantizar resultados de manera inmediata.

   Y el análisis de datos puede ayudar a la limpieza de datos detectando patrones inusuales que implica que existan datos incorrectos; detectando valores faltantes o fuera del rango usando las estadísticas; validando el formato en el que vienen los datos; comparar entre varias fuentes de datos para que verificar que no existan inconsistencias entre cada una; entre otros más.

8. ¿De qué forma los procesos ETL ayudan a efectuar la Limpieza de Datos?

   Presentan una forma en la que se deben estructurar y automatizar el proceso y manipulación de los datos; lo cual permite que identificar y corregir errores o inconsistencias que puedan existir en los datos que se leen de cualquier fuente.

9. ¿Por qué crees que se considera que la etapa ETL consume entre el 60% y el 80% de los recursos en un proyecto de implementación de DWH? ¿Cuál sería el proceso ETL que te sería más complicado llevar a cabo sí lo tuvieras que realizar?

   Considero que es debido a la **transformación** de los datos, ya que implica una gran manipulación de éstos para que se puedan tener datos coherentes y consistentes para la carga de datos en la DWH que se está implementando. Y, de hecho, este proceso de **transformación** es el más complicado a realizar ya que implica hacer un gran número de operaciones con los datos que garantizan que éstos estén limpios y bien estructurados para ser cargados en el DWH.

10. ¿Qué mecanismos propondrías para eliminar o minimizar el impacto de la mala calidad de datos?

    Realizar evaluaciones y verificaciones periódicas de los datos para poder identificar qué aspectos mejorar; utilizar herramientas que ayuden a la limpieza y transformación de datos de una manera eficiente; y realizar pruebas que aseguren que los datos cumplan con ciertas normas de calidad definidas.

2. Detección de violaciones

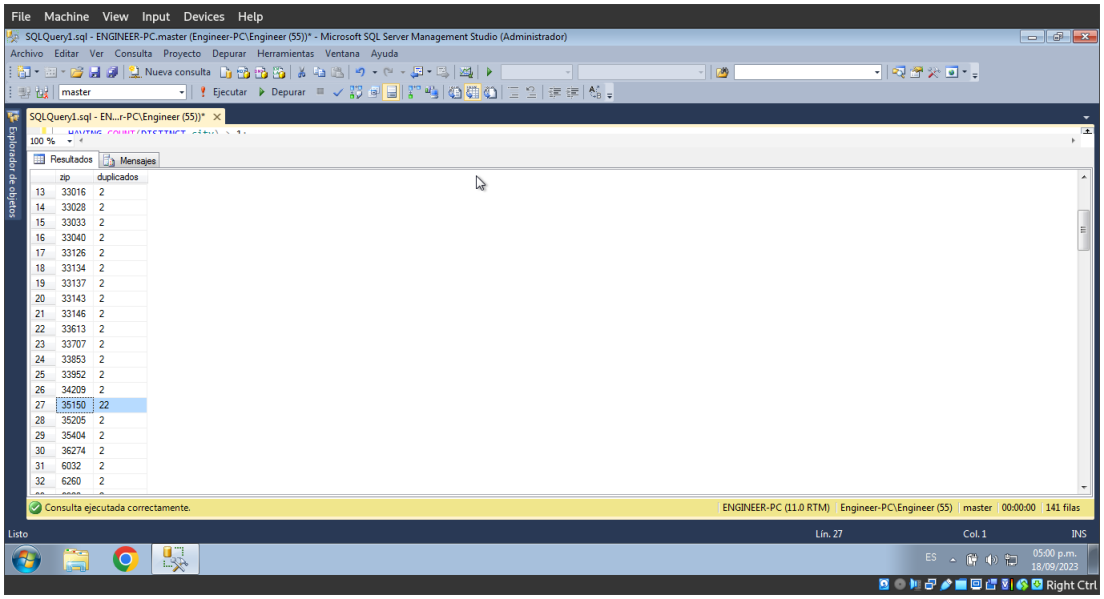
- Escribir consultas en SQL para detectar qué restricciones se violan.
- Ejecutar estas consultas y muestra a través de capturas de pantalla las violaciones que detectaste.

  Para ver que se violan las restricciones de una dependencia funcional, basta ver que para un valor del lado izquierdo (determinante), existen dos valores distintos del lado derecho (dependiente).

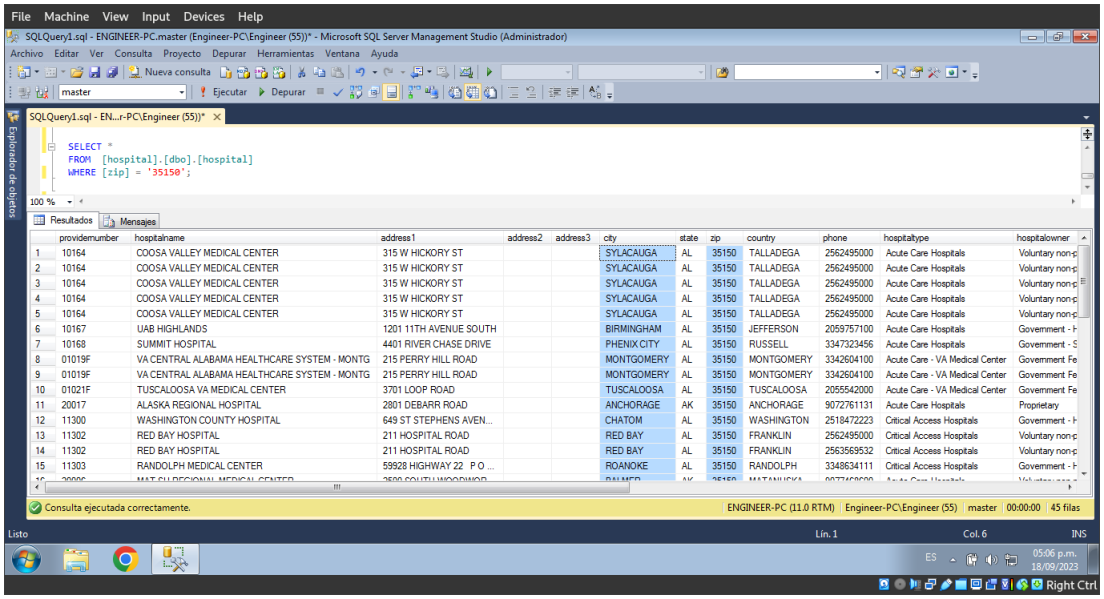
  1. $df0 : zip \longrightarrow city$

```sql
SELECT zip, COUNT(DISTINCT city) AS duplicados
FROM [hospital].[dbo].[hospital]
GROUP BY zip
HAVING COUNT(DISTINCT city) > 1;
```

  En este caso, para algún valor en zip, se tiene que hay valores distintos en la columna $city$. Por tanto, se violan las restricciones.



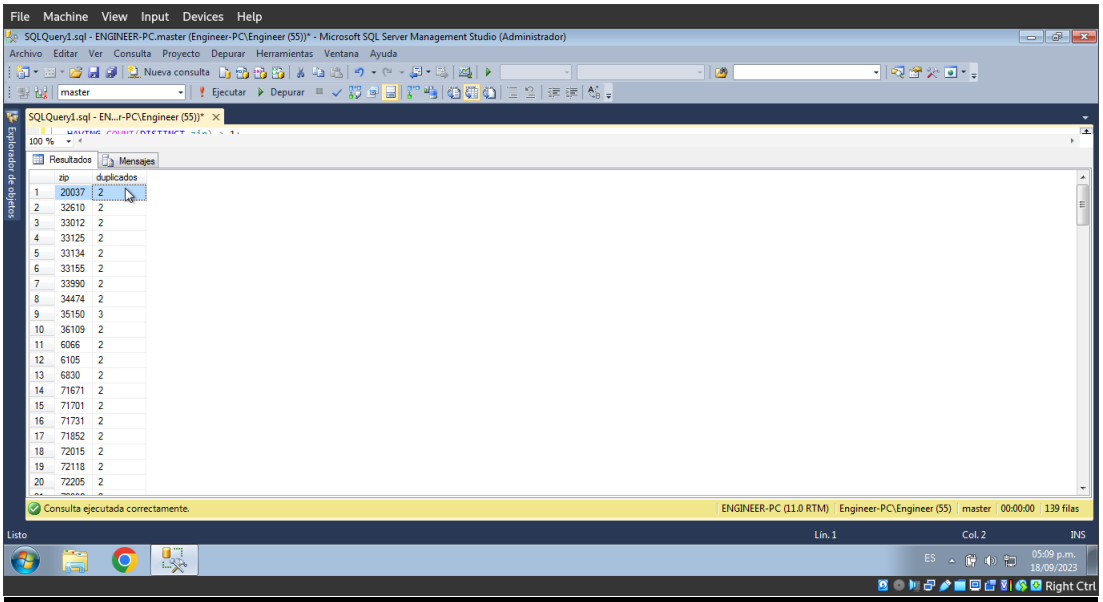
Al realizar la consulta anterior, obtenemos el siguiente resultado. Para visualizar y verificar que se violan las restricciones, seleccionamos el zip, 35150.

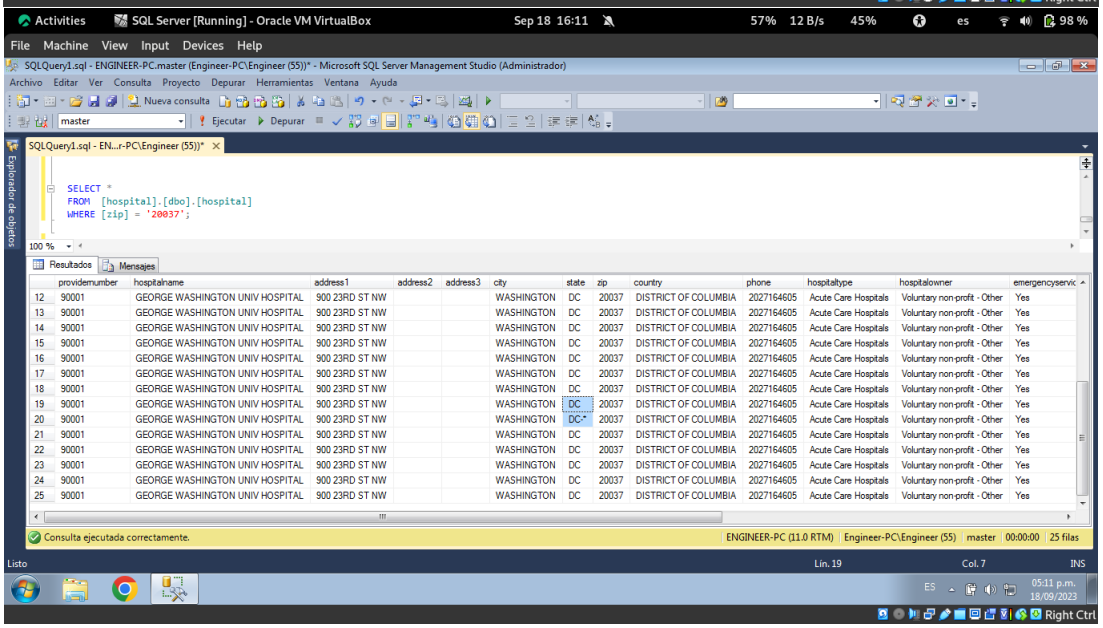


2. $df1 : zip \longrightarrow state$

   En este caso, se tiene algunos valores de la columna $state$ están sucios. Por ejemplo, DC, aparece como $DC - *$.

```
1  SELECT zip, COUNT(DISTINCT state) AS duplicados
2  FROM [hospital].[dbo].[hospital]
3  GROUP BY zip
4  HAVING COUNT(DISTINCT state) > 1;
```

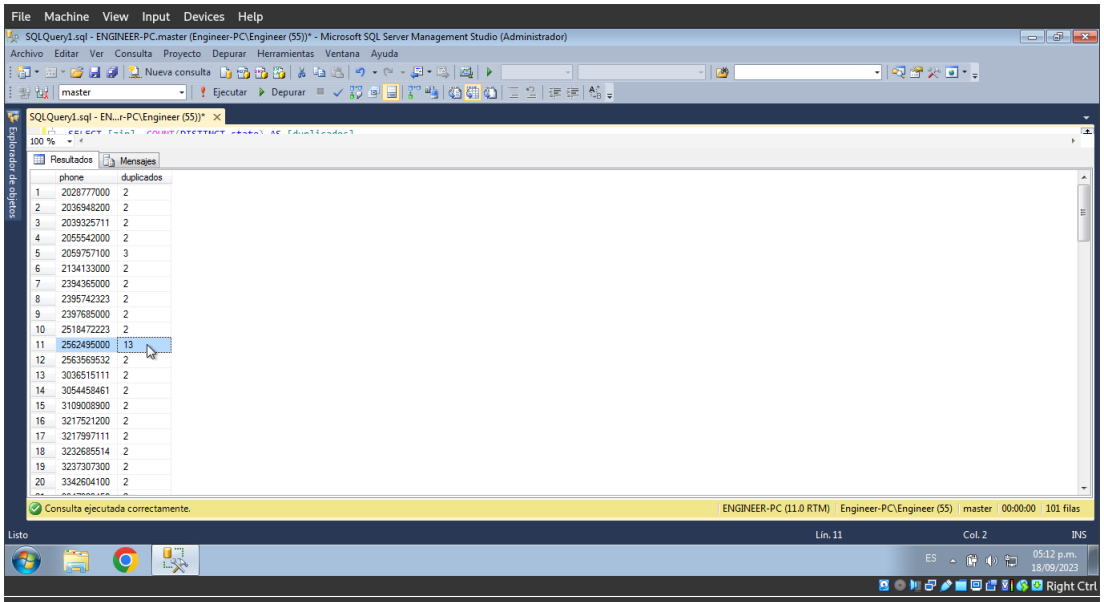


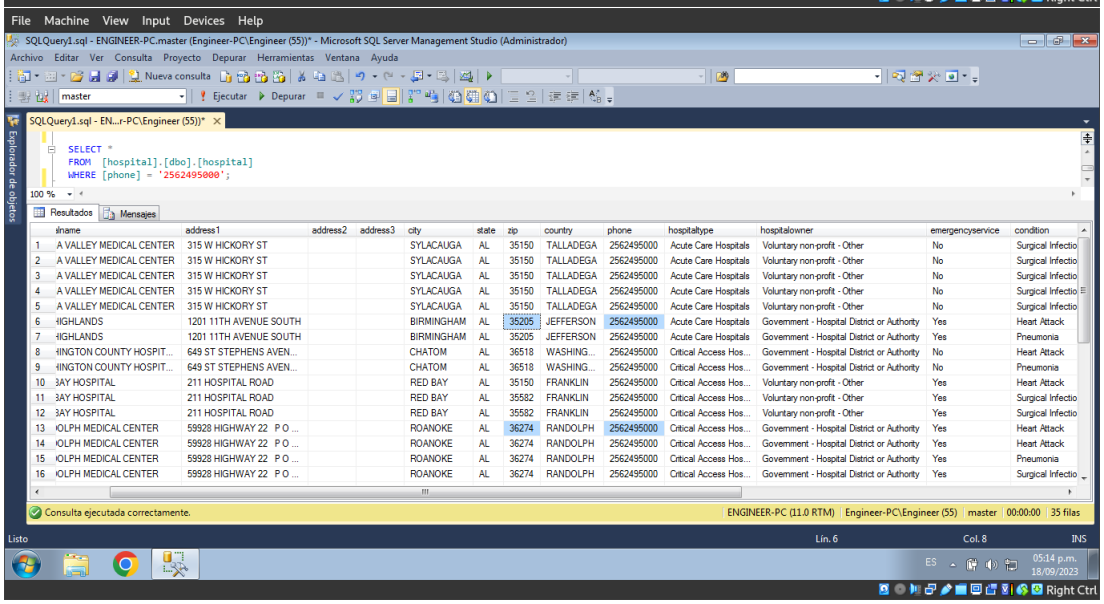


3. $df2 : phone \longrightarrow zip$

   En este caso se tiene que, para algún valor en *phone*, hay valores distintos en la columna *zip*. Por tanto, se violan las restricciones.

```sql
SELECT phone, COUNT(DISTINCT zip) AS duplicados
FROM [hospital].[dbo].[hospital]
GROUP BY phone
HAVING COUNT(DISTINCT zip) > 1;
```

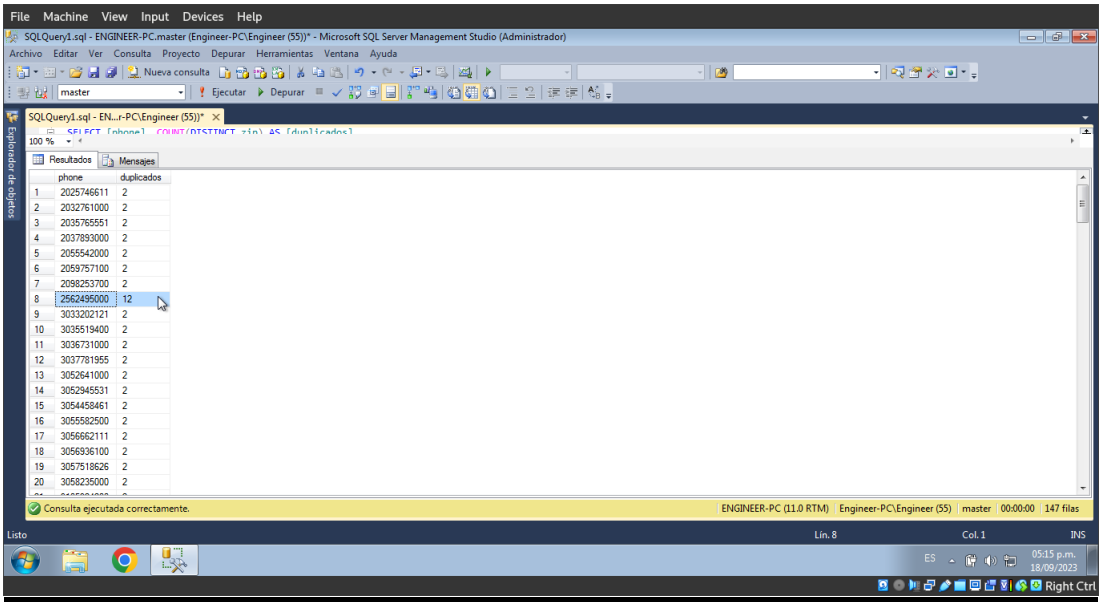


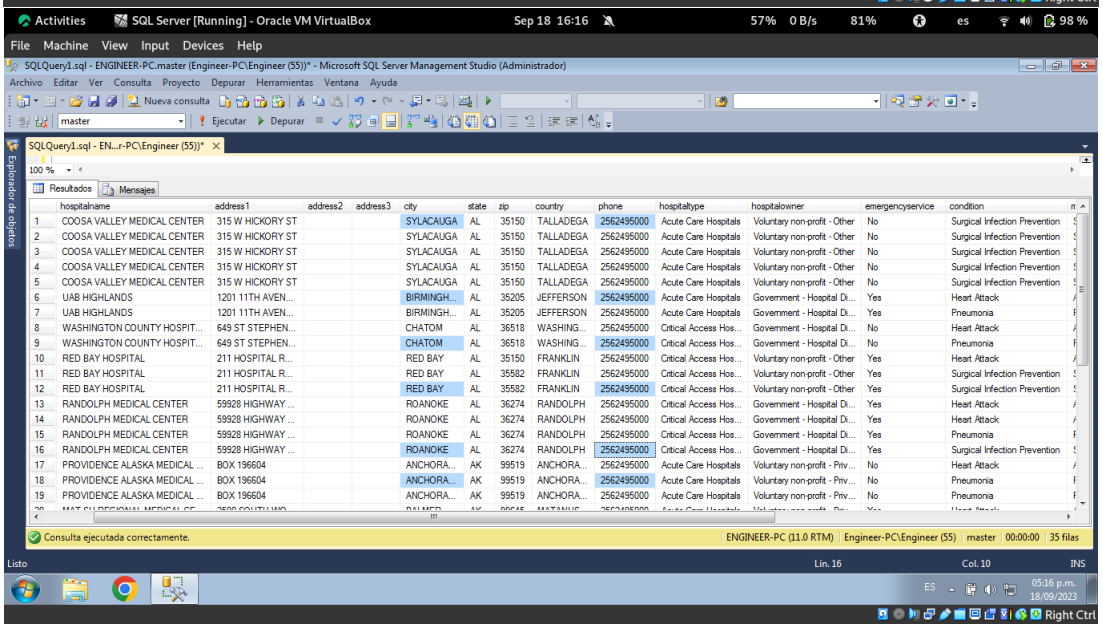


4. $df3 : phone \longrightarrow city$

   Al realizar la consulta notamos que, para algún valor en *phone*, se tiene que hay valores distintos en la columna $city$. Por tanto, se violan las restricciones.

```
1  SELECT phone, COUNT(DISTINCT city) AS duplicados
2  FROM [hospital].[dbo].[hospital]
3  GROUP BY phone
4  HAVING COUNT(DISTINCT city) > 1;
```

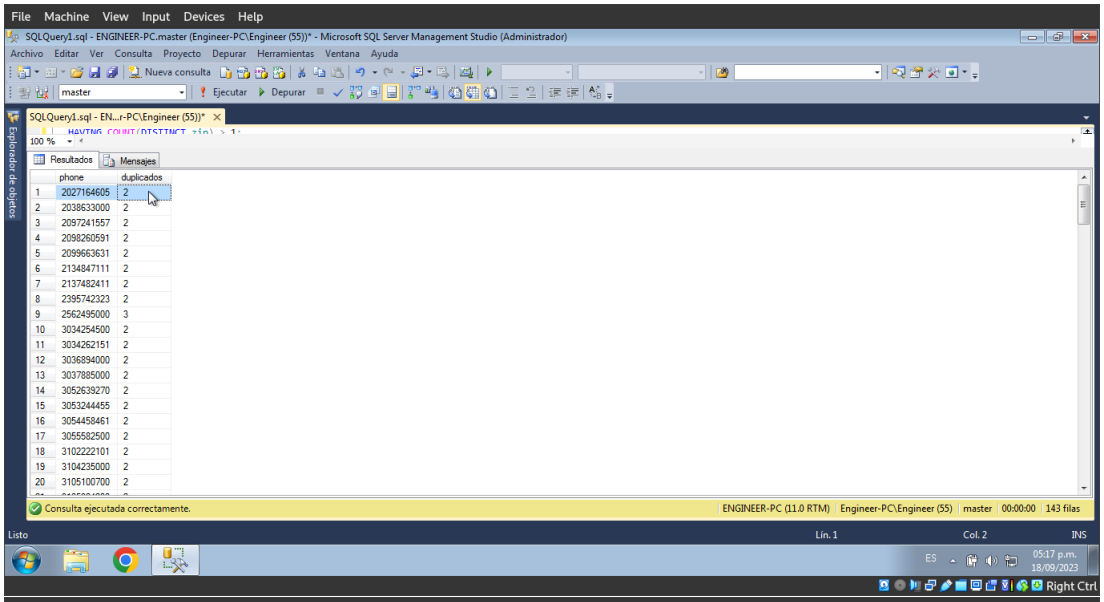


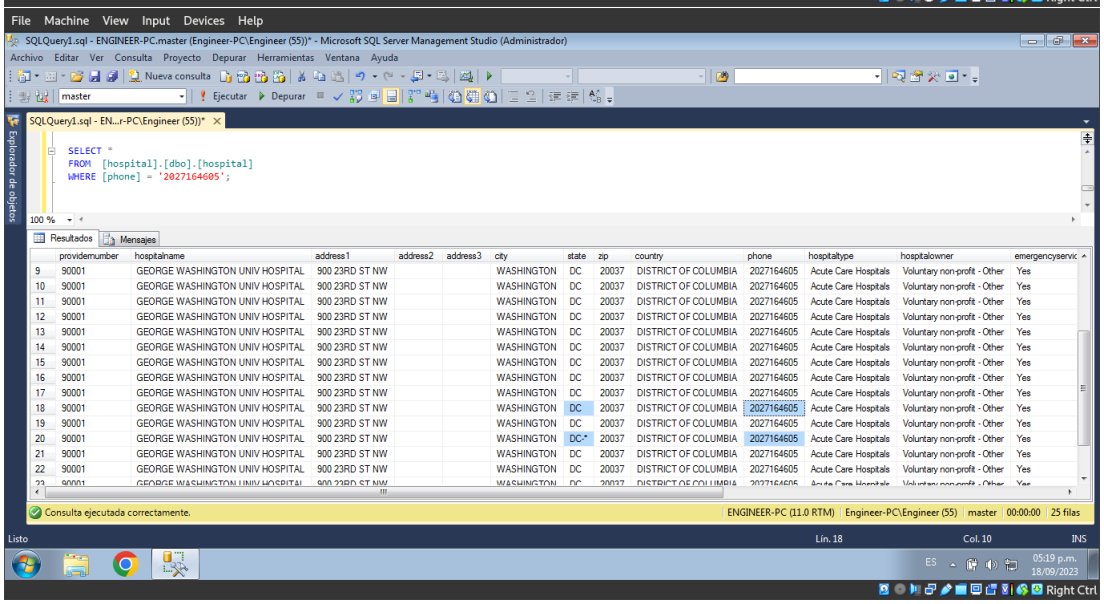


5. $df4 : phone \longrightarrow state$

   En este caso, se tiene algunos valores de la columna *state* están sucios. Por ejemplo, DC, aparece como $DC - *$.

```
1    SELECT phone, COUNT(DISTINCT state) AS duplicados
2    FROM [hospital].[dbo].[hospital]
3    GROUP BY phone
4    HAVING COUNT(DISTINCT state) > 1;
```

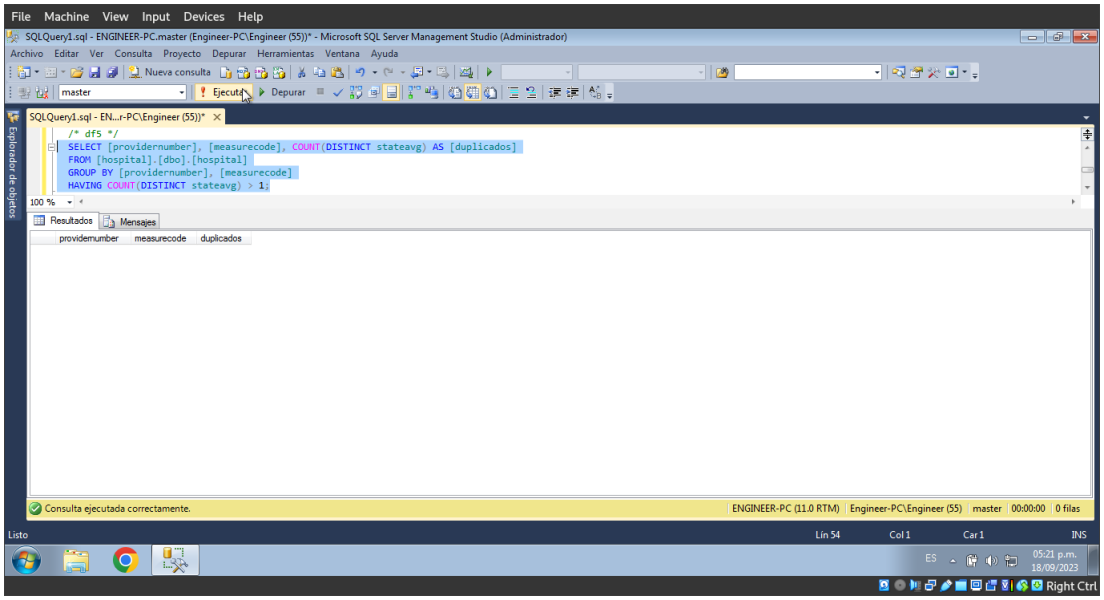




6. $df5 : providernumber, measurecode \longrightarrow stateavg$

   No se presentaron violaciones o anomalías para esta dependencia funcional.

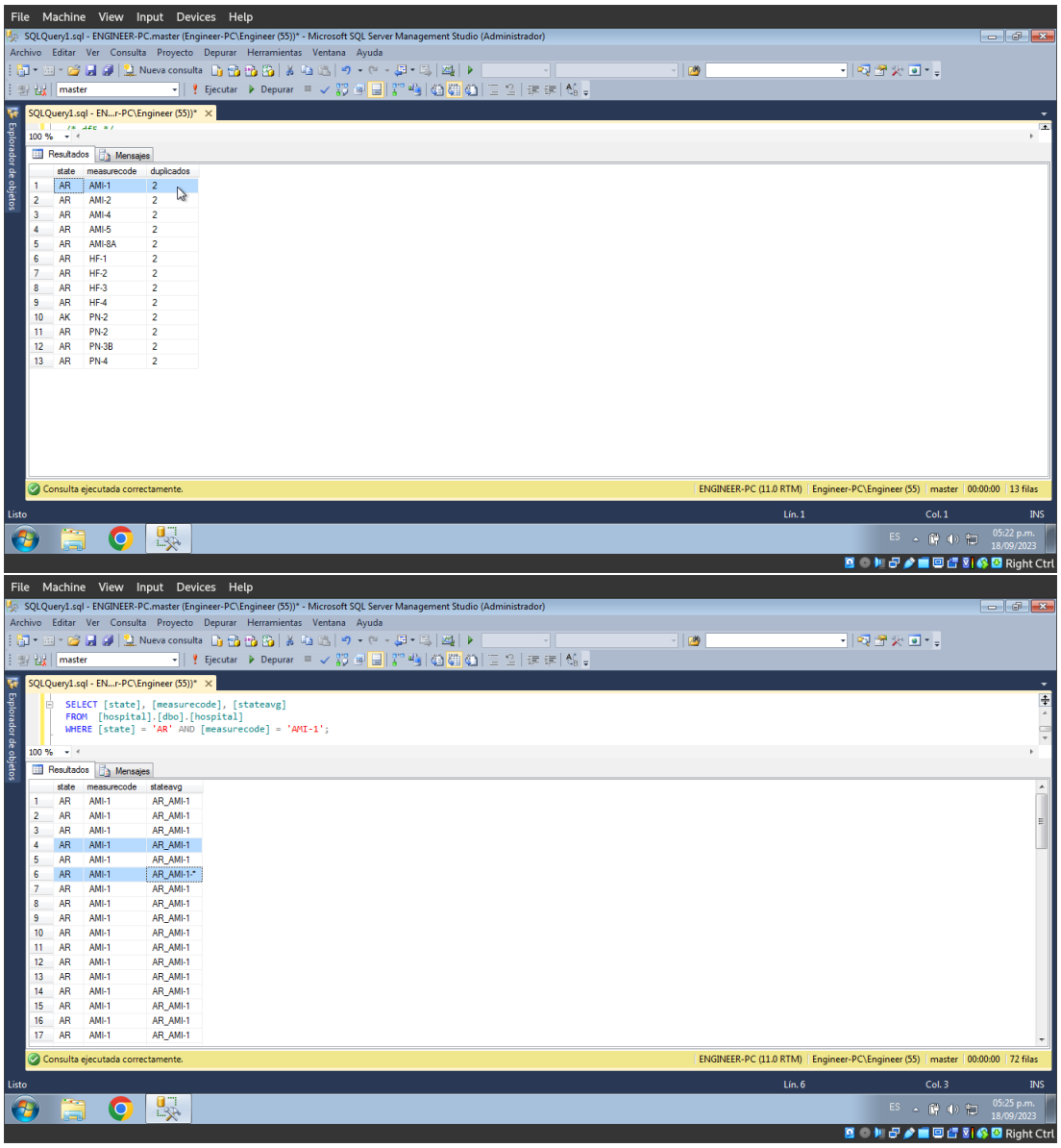
```sql
SELECT providernumber, measurecode,
COUNT(DISTINCT stateavg) AS duplicados
FROM [hospital].[dbo].[hospital]
GROUP BY providernumber, measurecode
HAVING COUNT(DISTINCT stateavg) > 1;
```



7. $df6 : state, measurecode \longrightarrow stateavg$

   En este caso, se tiene algunos valores de la columna $stateavg$ están sucios. Por ejemplo, AR_AMI-1, aparece como AR_AMI-*.

```
1  SELECT state, measurecode,
2  COUNT(DISTINCT stateavg) AS duplicados
3  FROM [hospital].[dbo].[hospital]
4  GROUP BY state, measurecode
5  HAVING COUNT(DISTINCT stateavg) > 1;
```



- Elije tres dependencias funcionales en las que hayas detectado violaciones e indica, apoyándote en Knime, para crear un flujo simple de limpieza para indicar las operaciones necesarias para corregir estas violaciones. Es posible que detectes algunos problemas adicionales como valores perdidos o problemas de calidad en algunas cadenas, también deberá agregar los nodos correspondientes para limpiarlos.

  Se eligen las dependencias funcionales, $df0$, $df1$ y $df2$. Dentro de knime, se llevo a cabo los siguientes pasos para la limpieza de datos:

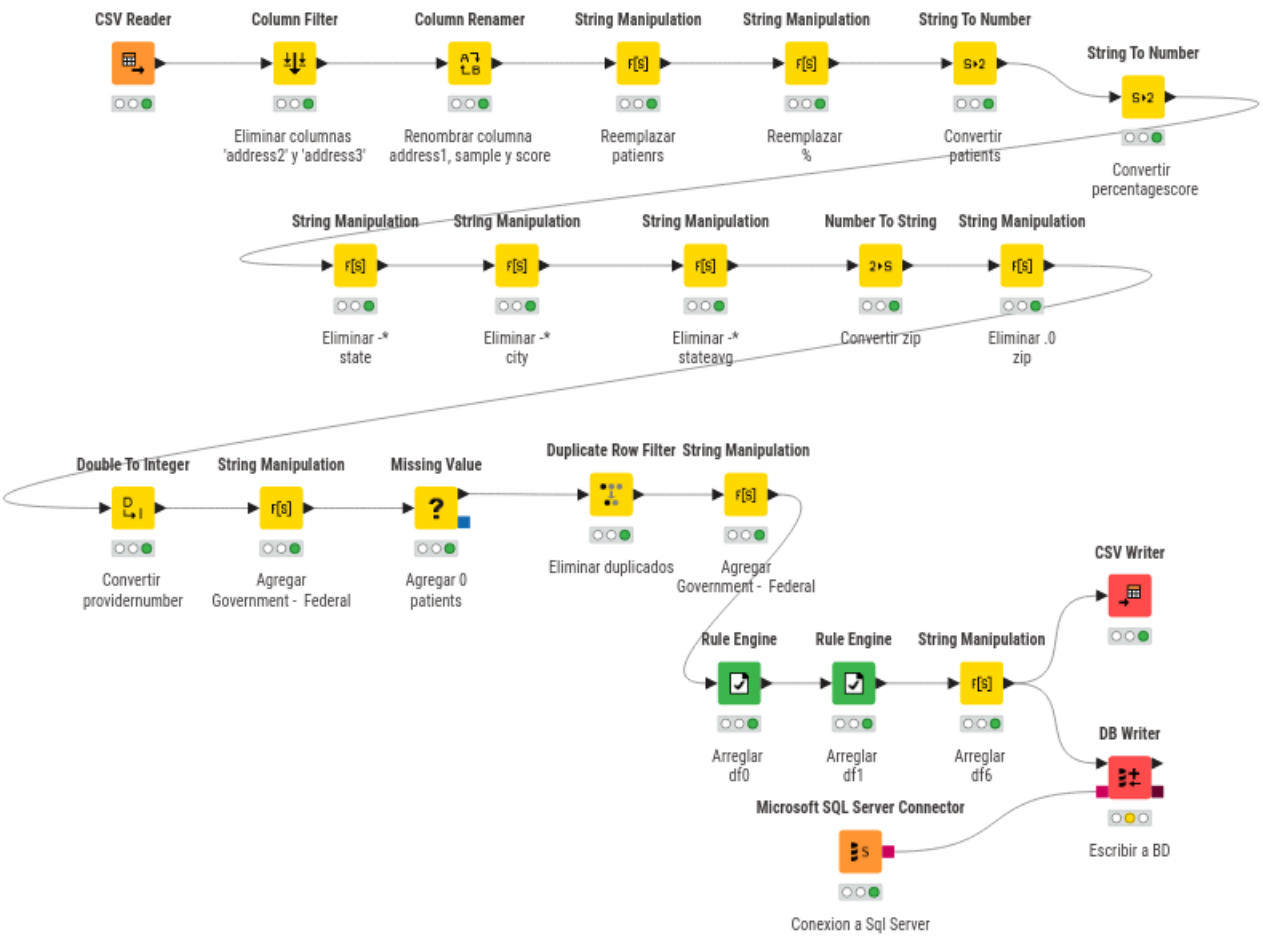
  a) Eliminación de las columnas $address2$ y $address3$.

  b) Corrección de algunos valores para las columnas $state$, $stateavg$ y zip.

  c) Renombrar las columnas $sample$ y $score$, y conversión de valores a $number$.

  d) Para $df0$ y $df1$, la mayoría de los restricciones que no se cumplían se arreglaron con b). Sin embargo, para el 35150, se opto por reemplazar los valores por el estado y ciudad con más ocurrencias.

  e) Para $df6$ se modifico el valor de $stateavg$ de acuerdo al estado en que se encontraba el hospital.



El archivo se encuentra en el directorio de `Programas` y ha sido nombrado `Hospital.knwf`, mientras que el archivo resultante de la limpieza se encuenytra en el directorio `Datasets` con el nombre de `Dataset1.csv`

3. Limpieza de datos

Realiza una tabla donde indiques todos los cambios que tuviste que realizar (de acuerdo con el ejercicio) y por qué se hicieron. Deberás entregar el conjunto de datos ya sin errores en formato Excel. ¿Qué utilidad consideras que tendría esta herramienta en tu ámbito profesional?

El ejercicio consistía en limpiar los siguientes datos: http://api.us.socrata.com/api/catalog/v1/domains
Y los cambios que se hicieron en ese conjunto de datos, usando OpenRefine fueron:

Cambios realizados	**Explicación**
Eliminación de espacios en blanco al inicio y al final de los datos	Se realizó esto para que todos los datos tengan un mejor formato y puedan ser manejados sin problemas en futuras implementaciones o usos.
Eliminación de columnas innecesarias	Se realizó la eliminación de las columna ”resultSetSize” ”_ - timings - serviceMillis” y ”_ - timings - searchMillis - searchMillis” debido a que no se le estaban dando un uso relacionado a los datos, y por lo tanto no eran necesarias.
Cambio de nombres de las columnas	Se realizó el cambio de nombre de las columnas ”results_domain” y ”“result_count” para colocar nombres más descriptivos a los datos que se presentaban en las celdas de dichas columnas, y así evitar ambigüedad.
Chequeo de datos duplicados	Se verificó si existían datos repetidos en la columna ”data_portals” para así evitar **redundancia** en los datos y guardar datos innecesarios.
Creación de nueva columna a partir de los valores de la columna ”data_portals”	Se realizó la creación de una nueva columna llamada ”region” a partir de los datos de la columna ”data_portals”, obteniendo sólo el valor que se encuentra después del primer punto (’.’), esto para así almacenar las organizaciones y gobiernos en una sola columna, y poder tener separados estos datos con un fácil acceso.
Transformar a mayúsculas la primera letra de los valores de la columna ”region”	Se cambió a mayúsculas la primera letra de los valores de la columna antes creada, ”region”, esto para darles un **formato** a estos datos y poderlos distinguirlos mejor.
Transformar a mayúsculas los datos de la columna ”region” que sólo tengan 2 letras	Se cambió a mayúsculas los datos que sólo tuvieran 2 letras, esto para poder dar un **formato** a las abreviaciones sobre los estados que se están guardando en esta columna.
Verificación de celdas vacías	Se verificó si existían celdas vacías en la columna ”region” ya que pudieran haber existido casos debido a la transformación de los datos de la columna ”data_portals” que se hizo, lo cual generaría inconsistencia entre los datos, pero en este caso no hubo.

La utilidad que tendría con esta herramienta en el ámbito profesional sería el poder tener la capacidad de ordenar y organizar un conjunto de datos que se me haya proporcionado, como una parte de una base de datos, la cual podría analizar de una manera más fácil y eficiente sin tener que modificar los datos reales.

El archivo resultante con el conjunto de datos sin errores en formato Excel se nombró `Dataset2.xlsx` ubicado en la carpeta `Datasets`.