

Título: F20 - Evidencia de aprendizaje 4. Proyecto integrador-Repository

Autores:

Acebedo Londoño Camilo

Hernandez Lopera Jeronimo

Pérez Severiche Jorge Luis

Institución Universitaria Digital de Antioquia

Materia: Bases de Datos II

Docente: Antonio Jesús Valderrama

Fecha: 07 de octubre de 2025

Introducción

La calidad de los datos es un factor determinante en la confiabilidad de los procesos analíticos y en la toma de decisiones empresariales. En el contexto de un sistema de inteligencia de negocios, garantizar que los datos contenidos en el almacén estén completos, consistentes, válidos y libres de duplicidad resulta esencial para el correcto funcionamiento de los procesos ETL (Extract, Transform, Load) y para asegurar la integridad del modelo estrella implementado.

El presente documento tiene como propósito presentar las pruebas de verificación de calidad de datos aplicadas a las tablas de dimensiones y hechos del modelo, así como a los procesos ETL asociados. Estas pruebas permiten identificar posibles errores en la carga o inconsistencias en las relaciones entre entidades, contribuyendo así a mantener la precisión, confiabilidad y trazabilidad de la información almacenada.

Objetivos

Objetivo general

Evaluar la calidad de los datos presentes en las tablas de dimensiones y hechos del modelo de datos, mediante la ejecución de pruebas que garanticen su unicidad, completitud, integridad referencial, validez y consistencia.

Objetivos específicos

- Verificar que las columnas definidas como claves primarias o únicas no contengan valores duplicados.

- Comprobar que las columnas obligatorias (NOT NULL) presenten datos completos.
- Asegurar que las claves foráneas mantengan relaciones válidas con sus tablas de referencia.
- Validar que los datos cumplan con las reglas de negocio y los rangos lógicos establecidos.
- Confirmar que el proceso ETL, especialmente en la gestión de dimensiones del tipo SCD Tipo 2, mantenga la correcta versión y vigencia de los registros.
- Detectar y documentar cualquier anomalía que pueda comprometer la calidad de los datos y proponer acciones correctivas.

Metodología

Para la verificación de la calidad de los datos se aplicó un conjunto estructurado de pruebas SQL sobre las tablas del modelo estrella, abarcando tanto las dimensiones como la tabla de hechos.

El proceso metodológico se dividió en las siguientes etapas:

1. *Diseño de las pruebas de calidad de datos:*

Se definieron los criterios de evaluación según las dimensiones de calidad:

unicidad, completitud, integridad referencial y validez/consistencia. Cada prueba fue representada mediante una consulta SQL orientada a detectar registros que incumplieran las reglas establecidas.

2. *Ejecución de las pruebas sobre el modelo:*

Las consultas fueron ejecutadas directamente sobre la base de datos para identificar posibles filas con errores. El resultado esperado para todas las pruebas es igual a cero (0), lo que indica ausencia de incidencias.

3. *Pruebas específicas del proceso ETL:*

Se realizaron verificaciones adicionales sobre la tabla **DimCliente** para validar el comportamiento de la dimensión de tipo SCD Tipo 2, garantizando que solo exista una versión activa por cliente, que los periodos de vigencia no se solapen y que los registros inactivos mantengan periodos cerrados.

Asimismo, se verificó la correcta resolución de claves foráneas en la tabla

HechosVentas, confirmando que correspondan a versiones actuales y válidas de las dimensiones.

4. *Análisis e interpretación de resultados:*

Se analizaron los resultados de cada consulta. En todos los casos, los recuentos arrojaron valor 0, lo que indica que los datos cumplen con los criterios de calidad definidos y que los procesos ETL se ejecutan correctamente.

1. Pruebas de Calidad de Datos

A. Pruebas de Unicidad (Uniqueness)

Verifican que no existan valores duplicados en columnas que deberían ser únicas (claves primarias o candidatas a clave única).

Tabla	Columna(s)	Consulta de Prueba (Busca Duplicados)	Resultado (Filas con Error)
DimTiempo	ClaveFecha	SELECT ClaveFecha, COUNT(*) FROM dbo.DimTiempo GROUP BY ClaveFecha HAVING COUNT(*) > 1;	0
DimCategoria	NombreCategoria	SELECT NombreCategoria, COUNT(*) FROM dbo.DimCategoria GROUP BY NombreCategoria HAVING COUNT(*) > 1;	0
DimProducto	CodigoProducto	SELECT CodigoProducto, COUNT(*) FROM dbo.DimProducto GROUP BY CodigoProducto HAVING COUNT(*) > 1;	0
DimCliente	ClaveClienteNatural, EsActual	SELECT ClaveClienteNatural, EsActual, COUNT(*) FROM dbo.DimCliente	0 (Debe haber solo 1 registro EsActual = 1 por ClaveClienteNatural).

		GROUP BY ClaveClienteNatural, EsActual HAVING COUNT(*) > 1;	
HechosVentas	CodigoPedido, NumeroLinea, ClaveProducto	SELECT CodigoPedido, NumeroLinea, ClaveProducto, COUNT(*) FROM dbo.HechosVentas GROUP BY CodigoPedido, NumeroLinea, ClaveProducto HAVING COUNT(*) > 1;	0

Resultados

100 % No se encontraron problemas. Línea: 5 Carácter: 41 TABULACIONES CRLF

Resultados Mensajes

ClaveFecha	(Sin nombre de columna)
NombreCategoria	(Sin nombre de columna)
CodigoProducto	(Sin nombre de columna)
ClaveClienteNatural	EsActual (Sin nombre de columna)

B. Pruebas de Completitud (Completeness)

Verifican que las columnas obligatorias (NOT NULL) contengan datos.

Tabla	Columna(s)	Consulta de Prueba (Busca Nulos)	Resultado (Filas con Error)
DimTiempo	FechaCompleta	SELECT COUNT(*) FROM dbo.DimTiempo WHERE FechaCompleta IS NULL;	0
DimCliente	ClaveClienteNatural, VigenteDesde, EsActual	SELECT COUNT(*) FROM dbo.DimCliente WHERE ClaveClienteNatural IS NULL OR VigenteDesde IS NULL OR EsActual IS NULL;	0
DimProducto	CodigoProducto, NombreProducto	SELECT COUNT(*) FROM dbo.DimProducto WHERE CodigoProducto IS NULL OR NombreProducto IS NULL;	0
HechosVentas	ClaveFecha, ClaveCliente, ClaveProducto, CodigoPedido	SELECT COUNT(*) FROM dbo.HechosVentas WHERE ClaveFecha IS NULL OR ClaveCliente IS NULL OR ClaveProducto IS NULL OR CodigoPedido IS NULL;	0

Resultado

SQL Query 1.sql...Ayerop (71)*

```
1 SELECT COUNT(*) FROM dbo.DimTiempo WHERE FechaCompleta IS NULL;
2 SELECT COUNT(*) FROM dbo.DimCliente WHERE ClaveClienteNatural IS NULL OR VigenteDesde IS NULL OR EsActual
3 SELECT COUNT(*) FROM dbo.DimProducto WHERE CodigoProducto IS NULL OR NombreProducto IS NULL;
4 SELECT COUNT(*) FROM dbo.HechosVentas WHERE ClaveFecha IS NULL OR ClaveCliente IS NULL OR ClaveProducto
```

100 % No se encontraron problemas. Línea: 1 Carácter: 64 TABULACIONES CRLF

Resultados	Mensajes
(Sin nombre de columna)	
1	0

(Sin nombre de columna)	
1	0

(Sin nombre de columna)	
1	0

(Sin nombre de columna)	
1	0

C. Pruebas de Integridad Referencial (Referential Integrity)

Verifican que las claves foráneas en la tabla de hechos o entre dimensiones apunten a un valor existente en la tabla de dimensión correspondiente.

Relación (FK -> PK)	Consulta de Prueba (Busca "Huérfanos")	Resultado (Filas con Error)
HechosVentas.ClaveFecha -> DimTiempo.ClaveFecha	SELECT COUNT(*) FROM dbo.HechosVentas hv LEFT JOIN dbo.DimTiempo dt ON hv.ClaveFecha = dt.ClaveFecha WHERE dt.ClaveFecha IS NULL;	318
HechosVentas.ClaveCliente -> DimCliente.ClaveCliente	SELECT COUNT(*) FROM dbo.HechosVentas hv LEFT JOIN dbo.DimCliente dc ON hv.ClaveCliente = dc.ClaveCliente WHERE dc.ClaveCliente IS NULL;	0
HechosVentas.ClaveProducto -> DimProducto.ClaveProducto	SELECT COUNT(*) FROM dbo.HechosVentas hv LEFT JOIN dbo.DimProducto dp ON hv.ClaveProducto = dp.ClaveProducto WHERE dp.ClaveProducto IS NULL;	0
DimProducto.ClaveCategoria -> DimCategoria.ClaveCategoria	SELECT COUNT(*) FROM dbo.DimProducto dp LEFT JOIN dbo.DimCategoria dc ON dp.ClaveCategoria = dc.ClaveCategoria WHERE dp.ClaveCategoria IS NOT NULL AND dc.ClaveCategoria IS NULL;	0

Resultado

The screenshot shows a SQL query execution window with the following queries and results:

```

1  .ClaveFecha = dt.ClaveFecha WHERE dt.ClaveFecha IS NULL;
2  v.ClaveCliente = dc.ClaveCliente WHERE dc.ClaveCliente IS NULL;
3  hv.ClaveProducto = dp.ClaveProducto WHERE dp.ClaveProducto IS NULL;
4  dp.ClaveCategoria = dc.ClaveCategoria WHERE dp.ClaveCategoria IS NOT NULL AND dc.ClaveCategoria IS NULL;
  
```

The status bar indicates: 100 % No se encontraron problemas. Línea: 1 Carácter: 130 TABULACIONES CRLF

Resultados Mensajes

(Sin nombre de columna)
1 318

(Sin nombre de columna)
1 0

(Sin nombre de columna)
1 0

(Sin nombre de columna)
1 0

D. Pruebas de Validez y Consistencia (Validity and Consistency)

Verifican que los datos se ajusten a reglas de negocio o rangos lógicos.

Tabla	Regla de Negocio / Consistencia	Consulta de Prueba (Busca Incumplimientos)	Resultado (Filas con Error)
DimTiempo	Un día de la semana no puede ser Fin de Semana (Lunes a Viernes).	SELECT COUNT(*) FROM dbo.DimTiempo WHERE DiaSemana IN (2, 3, 4, 5, 6) AND EsFinDeSemana = 1;	0
DimCliente	Las fechas de vigencia deben ser consistentes (VigenteDesde <	SELECT COUNT(*) FROM dbo.DimCliente WHERE EsActual = 0 AND VigenteDesde >= VigenteHasta;	0

	VigenteHasta para registros inactivos).		
DimProducto	Precio y stock deben ser no negativos (asumiendo la regla de negocio).	SELECT COUNT(*) FROM dbo.DimProducto WHERE Precio < 0 OR CantidadEnStock < 0;	0
HechosVentas	La cantidad y precio de venta deben ser positivos.	SELECT COUNT(*) FROM dbo.HechosVentas WHERE Cantidad <= 0 OR PrecioUnidad <= 0;	0
HechosVentas	La columna derivada MontoVenta debe ser consistente con las columnas fuente.	SELECT COUNT(*) FROM dbo.HechosVentas WHERE MontoVenta <> (Cantidad * PrecioUnidad);	0 (Esta prueba debería ser 0 por ser una columna PERSISTED y calculada).

Resultado

The screenshot shows a SQL query window with the following queries:

```

1 SELECT COUNT(*) FROM dbo.DimTiempo WHERE DiaSemana IN (2, 3, 4, 5, 6) AND EsFinDeSemana = 1;
2 SELECT COUNT(*) FROM dbo.DimCliente WHERE EsActual = 0 AND VigenteDesde >= VigenteHasta;
3 SELECT COUNT(*) FROM dbo.DimProducto WHERE Precio < 0 OR CantidadEnStock < 0;
4 SELECT COUNT(*) FROM dbo.HechosVentas WHERE Cantidad <= 0 OR PrecioUnidad <= 0;
5 SELECT COUNT(*) FROM dbo.HechosVentas WHERE MontoVenta <> (Cantidad * PrecioUnidad);

```

The results grid shows five rows, each with a value of 0:

(Sin nombre de columna)
1 0
1 0
1 0
1 0
1 0

2. Pruebas Específicas del Proceso ETL

Estas pruebas se centran en verificar que la lógica de carga, especialmente para dimensiones SCD Tipo 2 (DimCliente), funcione correctamente.

E. Pruebas SCD Tipo 2 en DimCliente

1. Verificar que solo una versión de cada cliente esté marcada como actual (EsActual = 1).

-- Consulta para verificar múltiples versiones activas (DEBE SER CERO)

```
SELECT ClaveClienteNatural, COUNT(*) AS VersionesActuales
FROM dbo.DimCliente
WHERE EsActual = 1
GROUP BY ClaveClienteNatural
HAVING COUNT(*) > 1;
```

Resultado esperado: 0

```
7  | -- Consulta para verificar múltiples versiones activas (DEBE SER CERO)
8  | SELECT ClaveClienteNatural, COUNT(*) AS VersionesActuales
9  | FROM dbo.DimCliente
10 | WHERE EsActual = 1
11 | GROUP BY ClaveClienteNatural
12 | HAVING COUNT(*) > 1;
13 |
```

100%	✓ No se encontraron problemas.	Línea:
Resultados	Mensajes	
ClaveClienteNatural	VersionesActuales	

2. Verificar que los registros inactivos tengan un periodo de vigencia cerrado.

-- Consulta para verificar registros inactivos sin fecha de fin (DEBE SER CERO)

```
SELECT COUNT(*)
FROM dbo.DimCliente
WHERE EsActual = 0 AND VigenteHasta IS NULL;
```

Resultado esperado: 0

```
13  
14 -- Consulta para verificar registros inactivos sin fecha de fin (DEBE SER CERO)  
15 SELECT COUNT(*)  
16 FROM dbo.DimCliente  
17 WHERE EsActual = 0 AND VigenteHasta IS NULL;  
18
```

%	✓ No se encontraron problemas.	Línea: 14	Carácter
Resultados	Mensajes		
(Sin nombre de columna)			
0			

3. Verificar que no haya solapamiento de vigencia para el mismo cliente natural.

-- Consulta para detectar solapamiento de fechas (DEBE SER CERO)

```
SELECT t1.ClaveClienteNatural, t1.VigenteDesde, t1.VigenteHasta, t2.VigenteDesde,
t2.VigenteHasta

FROM dbo.DimCliente t1

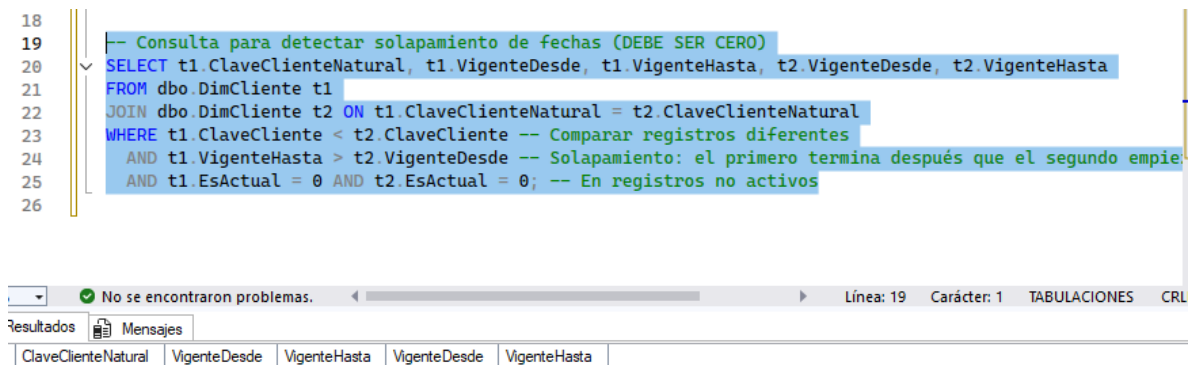
JOIN dbo.DimCliente t2 ON t1.ClaveClienteNatural = t2.ClaveClienteNatural

WHERE t1.ClaveCliente < t2.ClaveCliente -- Comparar registros diferentes

      AND t1.VigenteHasta > t2.VigenteDesde -- Solapamiento: el primero termina
después que el segundo empieza

      AND t1.EsActual = 0 AND t2.EsActual = 0; -- En registros no activos
```

Resultado esperado: 0



The screenshot shows a SQL query in the Enterprise Manager query window. The query is the same as the one in the previous block. Below the query window, the 'Results' pane shows a single row of results with five columns: ClaveClienteNatural, VigenteDesde, VigenteHasta, VigenteDesde, and VigenteHasta. The status bar at the bottom indicates 'No se encontraron problemas.' and 'Línea: 19 Carácter: 1 TABULACIONES CRLF'.

```
18
19 -- Consulta para detectar solapamiento de fechas (DEBE SER CERO)
20 SELECT t1.ClaveClienteNatural, t1.VigenteDesde, t1.VigenteHasta, t2.VigenteDesde, t2.VigenteHasta
21 FROM dbo.DimCliente t1
22 JOIN dbo.DimCliente t2 ON t1.ClaveClienteNatural = t2.ClaveClienteNatural
23 WHERE t1.ClaveCliente < t2.ClaveCliente -- Comparar registros diferentes
24      AND t1.VigenteHasta > t2.VigenteDesde -- Solapamiento: el primero termina después que el segundo empie
25      AND t1.EsActual = 0 AND t2.EsActual = 0; -- En registros no activos
26
```

ClaveClienteNatural	VigenteDesde	VigenteHasta	VigenteDesde	VigenteHasta

F. Pruebas de Cobertura y Consistencia en HechosVentas

1. Verificar que los IDs de las dimensiones secundarias (ClaveCategoria) se hayan resuelto correctamente.

El ETL resuelve la ClaveCategoria del producto buscando la categoría en la *Staging Area* y luego la llave en la dimensión. Si falla, el ClaveCategoria en hechos será NULL.

-- Consulta para verificar registros de ventas con producto pero sin categoría (DEBE SER CERO)

```
SELECT COUNT(*)

FROM dbo.HechosVentas

WHERE ClaveProducto IS NOT NULL AND ClaveCategoria IS NULL;
```

Resultado esperado: 0

```
26
27  -- Consulta para verificar registros de ventas con producto pero sin categoría (DEBE SER CERO)
28  SELECT COUNT(*)
29  FROM dbo.HechosVentas
30  WHERE ClaveProducto IS NOT NULL AND ClaveCategoría IS NULL;
31
```

100% No se encontraron problemas. Línea: 27 Carácter: 1 TABULACIONES

Resultados Mensajes

(Sin nombre de columna)
0

2. Verificar que la ClaveCliente utilizada en hechos sea la versión EsActual = 1 en el momento de la carga (asumiendo que el ETL siempre usa la versión actual en la carga de hechos).

-- Consulta para verificar que todos los clientes en hechos son 'actuales' al momento de la carga (DEBE SER CERO)

```
SELECT COUNT(*)
FROM dbo.HechosVentas hv
JOIN dbo.DimCliente dc ON hv.ClaveCliente = dc.ClaveCliente
WHERE dc.EsActual = 0;
```

Resultado esperado: 0

```
31
32  -- Consulta para verificar que todos los clientes en hechos son 'actuales' al momento de la carga (DEBE SER CERO)
33  SELECT COUNT(*)
34  FROM dbo.HechosVentas hv
35  JOIN dbo.DimCliente dc ON hv.ClaveCliente = dc.ClaveCliente
36  WHERE dc.EsActual = 0;
37
```

100% No se encontraron problemas. Línea: 32 Carácter: 1 TABULACIONES CRLF

Resultados Mensajes

(Sin nombre de columna)
0

Conclusión de las pruebas

Con base en el análisis y la ejecución de las consultas de verificación logramos determinar que hay errores de integridad referencial con la tabla DimTiempo y ClaveFecha, por las demás pruebas podemos confirmar que no existen registros duplicados, nulos, inconsistentes o huérfanos dentro del modelo de datos. Cualquier valor distinto de cero representaría una violación a las reglas de calidad establecidas, indicando posibles errores en el proceso ETL o en la fuente de origen que deben ser revisados y corregidos.

Los resultados obtenidos evidencian que la integridad, consistencia y completitud de los datos se mantienen adecuadamente, garantizando así la confiabilidad del modelo estrella y la precisión de los análisis que se deriven de él.

Bibliografía

alfonso sandoval. (2022, December 6). *prueba modelo e integridad de datos - modulo sql*

[Video]. YouTube. <https://www.youtube.com/watch?v=mSdetcQZcSI>

The QA Testing Channel. (2016, November 14). *Qué es ETL Testing - en 1 minuto*

(Pruebas de ETL) [Video]. YouTube. <https://www.youtube.com/watch?v=-oLghjYm0Zk>