UTEC
Universidad Tecnológica

Licenciatura en Tecnologías de la Información

Base de Datos

Proyecto Final Tecnicatura

Integrantes:

Sebastián Barboza - Genaro Delgado

Victoria Rey- Milagros Villanueva- Matias Fernandez

Generación 2024

Tutor:

Sergio Baudine

Año de entrega 2025

Índice

Base de datos Operativa

Introducción

Para el área de **Base de Datos** del Proyecto Final de Tecnicatura, partimos de la estructura diseñada y utilizada en el **Proyecto de Desarrollo del segundo semestre**. A partir de un modelo inicial, se realizaron ajustes y mejoras con el fin de adaptarlo mejor a los requerimientos.

En el proceso de diseño y creación de la base de datos se contemplaron todos los módulos definidos en el sistema, garantizando la correcta representación de entidades, relaciones e integridad de los datos. Los módulos que debían estar reflejados en la base son los siguientes:

1. Módulo de Gestión de Usuarios
2. Módulo de Gestión de Perfiles
3. Módulo de Gestión de Funcionalidades
4. Módulo de Auditoría
5. Módulo de Gestión de Actividades
6. Módulo de Gestión de Espacios
7. Módulo de Gestión de Tipos de Actividades
8. Módulo de Gestión de Registro de Pagos de Usuarios

Herramientas

- Se utilizó el Sistema de Gestión de Base de Datos **PostgreSQL**, y con ello la herramienta PgAdmin para la creación de la base de datos del proyecto.
- Se usó la herramienta **Microsoft Excel** para la creación del Modelo Relacional, y Modelo Físico
- Se utilizó **Draw.io** para la creación del Modelo Entidad-Relación..

Modelo Entidad- Relación

- Link Modelo Entidad-Relación
  https://drive.google.com/file/d/1vG-Dhkc2r0n2ZYgn1RT2YpdTnGV4kXog/view?usp=sharing

Modelo Lógico

- Link a modelo logico
  https://docs.google.com/spreadsheets/d/12-hpcRYOIQdgZp2D_gNrCGOcE_vEax-WaZT3XrljKrs/edit?usp=sharing

Scripts de base de datos

- Link a scripts

  https://drive.google.com/file/d/1zjnrHRNfLTJp8OWk9f7CZE8DQ0o0A7xo/view?usp=sharing

Documentación base de datos - Proyecto de Desarrollo

- Link a documento
  https://drive.google.com/file/d/1-f21pmw2Kfso4y41OTs4MUsFr3lI6NpB/view?usp=sharing

Data Warehousing

Objetivo

El objetivo de esta sección del proyecto, es construir una capa de Data Warehousing a partir de la base de datos operativa de ASUR. Buscamos unificar y limpiar la información para poder analizarla de forma más simple, y obtener indicadores que ayuden a la directiva a tomar decisiones.

Con esto se podrán ver tendencias, comparar períodos y tener tableros con métricas claras sobre las actividades, socios y la gestión general de la institución.

Definición de preguntas de negocio

Se seleccionaron tres preguntas basadas en las necesidades operativas y de gestión de ASUR. El objetivo es identificar patrones y métricas que apoyen la toma de decisiones respecto a la participación en actividades, la utilización de espacios y la segmentación entre socios y no socios.

1. **¿Qué tipos de actividades y qué meses presentan mayor cantidad de inscripciones y menor tasa de cancelación?**
   Esta pregunta permite analizar la popularidad de las actividades y su nivel de efectividad, identificando tendencias estacionales o categorias más atractivas para los participantes.

2. **¿Qué espacios presentan mayor nivel de uso y tasa de cancelación, y cómo varía según el mes?**
   Busca evaluar la utilización de los espacios físicos de ASUR, determinando cuáles tienen mayor demanda y en qué períodos se registran más cancelaciones, con el fin de optimizar la planificación y gestión de recursos.

3. **¿Cuál es la participación de socios y no socios en las actividades, y cómo varía por tipo de actividad y mes?**
   Permite medir la proporción de inscripciones según el tipo de usuario, con el propósito de comprender el grado de involucramiento de los socios frente a otros participantes y orientar estrategias de promoción o fidelización.

Estas preguntas posibilitan obtener indicadores relevantes para la **planificación institucional**, **mejora de la oferta de actividades**, **optimización del uso de espacios** y **fortalecimiento del vínculo con los socios**.

Perfilado y limpieza

Se realizó una evaluación de la base de datos operativa con el fin de comprender el estado actual de los datos y su calidad. Esta revisión incluyó:

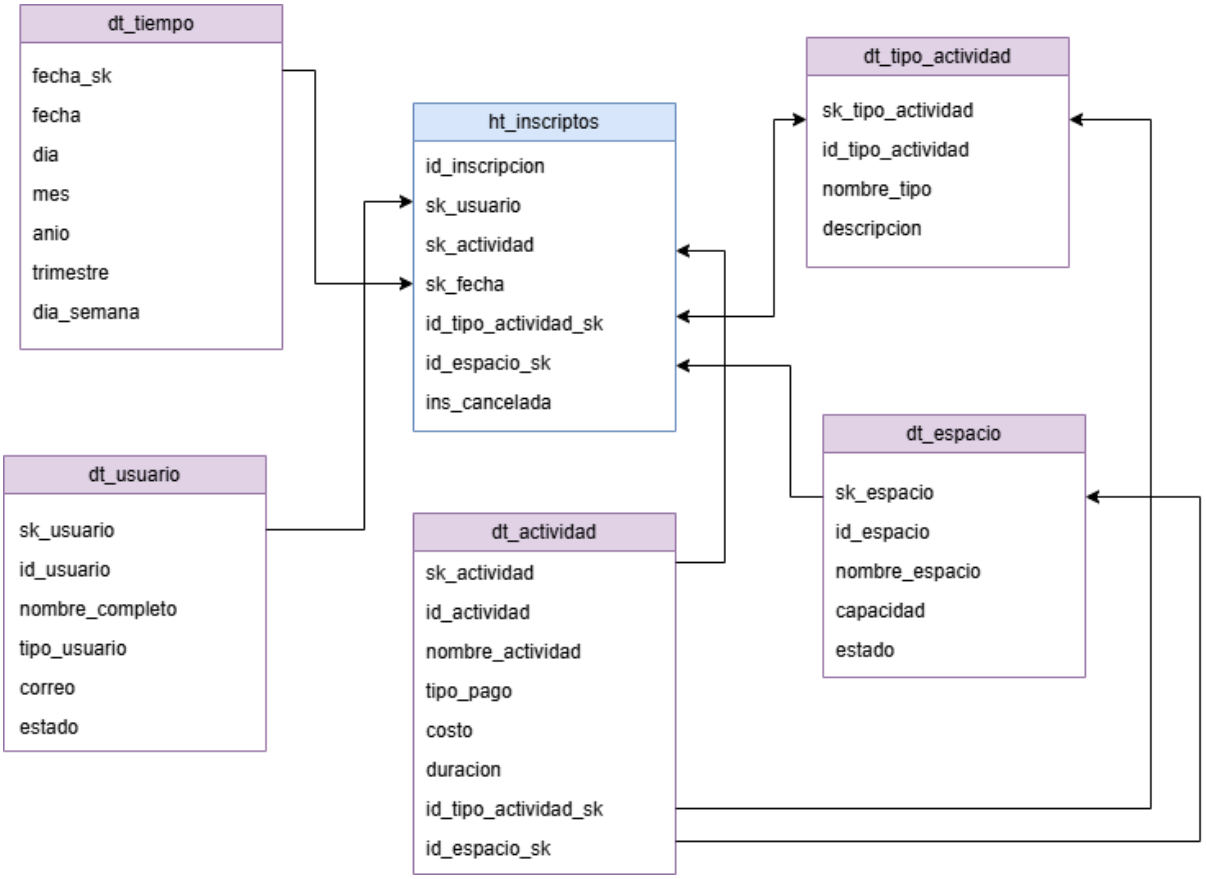
- Detección de valores nulos y formatos inconsistentes.
- Identificación de posibles duplicados.
- Validación de dominios y reglas de negocio.
- Revisión de coherencia entre registros relacionados.

Este proceso permitió identificar ajustes necesarios para garantizar que los datos ingresados al Data Warehouse sean consistentes y adecuados para análisis.

A partir de este análisis se definieron las siguientes reglas de limpieza:

- Normalización de textos para evitar variaciones por espacios o diferencias en mayúsculas/minúsculas.
- Conversión de todos los campos de fecha a un formato único y validación de fechas fuera de rango (por ejemplo, no admitir fechas futuras en pagos o inscripciones).
- Eliminación de registros duplicados, conservando el más reciente.
- Validación de valores como categorías de socio, formas de pagos, tipos de documento, entre otros.
- Validación de valores numéricos, por ejemplo no se permiten valores negativos en pagos.
- Validación de capacidades de espacios, deben ser mayores a 0 y numéricas.
- Ocultación de datos sensibles que no son necesarios para análisis, como emails o teléfonos, para proteger la información personal.

Modelo dimensional



Grano del hecho

Cada fila en ht_inscriptos representa una inscripción individual de un usuario a una actividad en una fecha determinada.

Llaves foráneas

Llave foránea	Dimensión relacionada
sk_usuario	dt_usuario
sk_actividad	dt_actividad
sk_tiempo	dt_tiempo

Diccionario de Datos

Tabla: ht_inscriptos

Campo	Tipo de dato	Descripción
id_inscripcion	SERIAL	Identificador único de inscripción
sk_usuario	BIGINT	FK → Usuario inscrito
sk_actividad	BIGINT	FK → Actividad inscrita
sk_tiempo	BIGINT	FK → Fecha de inscripción
ins_cancelada	BOOLEAN	Indica si la inscripción fue cancelada

Tabla: dt_usuario

Campo	Tipo de dato	Descripción
sk_usuario	BIGINT	Clave sustituta del usuario
nombre_completo	VARCHAR(150)	Nombre y apellido del usuario
tipo_usuario	VARCHAR(50)	Socio / No socio
ciudad	VARCHAR(100)	Ciudad de residencia

Tabla: dt_actividad

Campo	Tipo de dato	Descripción
sk_actividad	BIGINT	Clave sustituta de la actividad
nombre_actividad	VARCHAR(100)	Nombre descriptivo de la actividad
costo	NUMERIC	Costo de la actividad
duracion	INT	Duración en minutos
sk_tipo_actividad	INT	FK → Tipo de actividad
sk_espacio	INT	FK → Espacio donde se realiza

Tabla: dt_espacio

Campo	Tipo de dato	Descripción
sk_espacio	INT	Identificador del espacio
nombre_espacio	VARCHAR(100)	Nombre del espacio físico
capacidad	INT	Capacidad máxima del espacio

Tabla: dt_tipo_actividad

Campo	Tipo de dato	Descripción
sk_tipo_actividad	INT	Identificador del tipo de actividad
nombre_tipo	VARCHAR(50)	Nombre del tipo (taller, charla, evento, etc.)
descripcion	VARCHAR(255)	Breve descripción

Tabla: dt_tiempo

Campo	Tipo de dato	Descripción
sk_tiempo	INT	Identificador de tiempo
fecha	DATE	Fecha completa
mes	INT	Mes numérico
anio	INT	Año correspondiente

SQL/ETL

Extracto representativo:

```
INSERT INTO staging.d_usuario (
  id_usuario, nombre_completo, tipo_usuario, correo, telefono, id_ciudad, categoria_socio, dif_auditiva, len_sena)
SELECT *
FROM staging.dblink(
  'dbname=ProyectoS4 user=postgres password=123456'::text,
  $$
  SELECT
    u.id_usuario,
    CONCAT(u.pri_nombre, ' ', u.pri_apellido, ' ', u.seg_apellido) AS nombre_completo,
    u.tipo_usuario,
```

```
        u.correo,
        t.nro_telefono AS telefono,
        u.id_ciudad,
        s.cat_socio AS categoria_socio,
        s.dif_auditiva,
        s.len_sena
    FROM usuarios u
    LEFT JOIN socios s ON u.id_usuario = s.id_usuario
    LEFT JOIN telefonos t ON u.id_usuario = t.id_usuario
    $$
) AS t(
    id_usuario BIGINT,
    nombre_completo VARCHAR(150),
    tipo_usuario VARCHAR(50),
    correo VARCHAR(100),
    telefono VARCHAR(50),
    id_ciudad INT,
    categoria_socio VARCHAR(50),
    dif_auditiva BOOLEAN,
    len_sena BOOLEAN
);
```

Enlace a script completo

PII y seguridad

El modelo dimensional del Data Warehouse de ASUR incluye ciertos datos personales identificables (PII), como nombre completo, correo electrónico, C.I y teléfono, provenientes de la base operativa.

Durante el proceso de integración se aplicaron medidas concretas de protección de PII:

- El campo cédula de identidad (CI), por considerarse altamente sensible, no fue cargado en ninguna etapa del proceso.
- Los campos teléfono y correo electrónico se cargaron solo en el área *staging* para validaciones, pero fueron eliminados en las etapas posteriores (ODS y DW).

De este modo, el modelo analítico final no contiene información personal directa o identificable.

Control de acceso y exportación de datos

Actualmente, no se establece conexión directa entre Looker Studio y la base PostgreSQL.
La información analítica se exporta desde el DW a archivos CSV controlados manualmente, que luego se cargan en Looker.
Esto garantiza que solo se compartan vistas agregadas y anonimizadas, previamente revisadas.

Enmascaramiento y anonimización (plan futuro)

Se prevé implementar en el futuro:

- Anonimización de nombres mediante códigos o identificadores únicos en las vistas analíticas.

Dashboard en Looker Studio

Enlace al dashboard en looker studio

Limitaciones y oportunidades de mejora

Durante el desarrollo Data Warehouse, se identificaron ciertas limitaciones propias del alcance académico del proyecto, así como posibles mejoras para futuras entregas.

Limitaciones actuales

- **Conectividad con herramientas externas:**
  Looker Studio no permitió conectarse directamente con la base de datos PostgreSQL, por lo que fue necesario exportar los resultados desde el DW a archivos CSV y luego cargarlos manualmente.
  Aunque funcional, este método implica mayor trabajo manual y no permite actualización automática de los dashboards.

- **Cantidad reducida de datos históricos:**
  La base operativa (BDO) dispone de un número limitado de registros. Si bien se agregaron datos mockeados para enriquecer los análisis, aún se podría aumentar el volumen para obtener resultados más representativos.

- **Ausencia de anonimización completa:**
  Aunque los campos sensibles (como teléfono, correo y CI) se eliminaron o no se cargaron en las etapas analíticas, todavía no se implementó un proceso de anonimización o codificación de nombres de usuario.

Oportunidades de mejora

- **Ampliar el conjunto de datos simulados**, incorporando más registros que representen diferentes períodos y situaciones reales de la asociación, a fin de fortalecer los indicadores de análisis.

- **Agregar mecanismos de auditoría y control de calidad**, como una tabla de auditoría para registrar los procesos de carga y validaciones, planificado para la próxima entrega.

- **Aplicar anonimización o codificación de nombres** en las vistas del DW, para reforzar las medidas de seguridad y privacidad.

- **Optimizar los dashboards en Looker Studio**, incorporando filtros y comparativas que faciliten la interpretación de los resultados obtenidos.

Ajustes incorporados para la entrega final

Tras la devolución del docente se realizaron correcciones puntuales en la capa de Data Warehousing con el fin de alinear el proyecto a la metodología vista en clase. Los ajustes principales fueron:

1. **Revisión del ODS:**
   Se eliminaron los IDs y claves primarias innecesarias, manteniendo únicamente los identificadores provenientes de la base operativa para asegurar que el ODS sea un reflejo limpio del sistema transaccional.

Bitácora de decisiones

1. Selección del hecho principal

Se decidió establecer como hecho principal las inscripciones a actividades (`h_inscriptos`), ya que este proceso representa la interacción más relevante entre los usuarios y la asociación. A partir de él se pueden analizar variables esenciales como la participación, las cancelaciones y la evolución mensual de la asistencia. Esta elección permitió mantener un modelo claro y sencillo.

A partir de esta decisión se definieron los principales KPIs del proyecto, alineados con las preguntas de negocio:

- Cantidad de inscripciones por actividad y mes.

- Tasa de cancelación por actividad, espacio y mes.

- Nivel de uso de los espacios.

- Participación de socios y no socios por tipo de actividad.

Estos indicadores se consideraron suficientes para reflejar el comportamiento operativo de ASUR y responder a los objetivos planteados, sin agregar complejidad innecesaria al modelo.

2. Alcance de los datos personales (PII)

Durante la integración de los datos se optó por excluir completamente la cédula de identidad, al considerarse un dato altamente sensible, y eliminar los campos de teléfono y

correo electrónico en las etapas de ODS y DW. Esta medida tuvo como objetivo proteger la información personal de los usuarios y aplicar el principio de minimización de datos.

3. Conectividad con herramientas de análisis

Dado que Looker Studio no permitió la conexión directa con PostgreSQL, se resolvió exportar las vistas creadas en el Data Warehouse a archivos CSV y luego cargarlas manualmente en la herramienta. Esto permitió compartir únicamente datos ya transformados y agregados, evitando la exposición directa de las tablas internas. Aunque implica una carga manual, fue una solución práctica dentro del alcance académico del proyecto y garantizó un mayor control sobre la información publicada en los dashboards.

4. Carga de datos mockeados

Debido a la escasa cantidad de registros en la base operativa original, se generaron datos simulados (mockeados) para enriquecer los análisis. Esto permitió realizar pruebas más representativas y obtener resultados más completos al momento de construir las vistas analíticas. El agregado de estos datos permitió observar patrones más claros sin comprometer la estructura real de la base.

5. Estructura del modelo y nomenclatura

Se adoptó una nomenclatura estandarizada y una separación por esquemas para cada etapa del proceso. En la base de datos se crearon los esquemas staging, ods y dw, cada uno con sus propias tablas y vistas correspondientes. Además, se siguió la convención sugerida institucionalmente: prefijos d_ y h_ para staging, ods_ para la capa intermedia y dt_ y ht_ para el data warehouse. Esta organización permitió mantener una clara trazabilidad entre las capas, mejorar la comprensión del flujo de datos y facilitar el mantenimiento futuro del modelo.

6. Incidencias encontradas y ajustes

Durante la etapa de generación de datos mockeados se detectó una incidencia en la base operativa: los campos correspondientes al nombre y la cédula estaban intercambiados. Este error surgió al momento de simular los registros para ampliar el conjunto de datos. Se resolvió corrigiendo la transformación en la carga hacia staging, concatenando correctamente los nombres y apellidos y excluyendo la cédula del proceso. Esta corrección permitió mantener la coherencia de los datos y asegurar su integridad sin afectar el resultado final.