



**UNIVERSIDAD NACIONAL DE ENTRE RÍOS**

**FACULTAD DE INGENIERÍA**

**Tecnicatura Universitaria en Procesamiento y Explotación  
de Datos**

**MATERIA: Espacio Integrador 2**

**Nombre de la Actividad: Informe final**

**Título: Desarrollo de un sistema automatizado para el  
procesamiento de ofertas laborales**

**Fecha de Entrega: 14-11-2023**

**Profesores: Lic. Walter R. Elias**

**Alumno: Ruiz Diaz, Enzo**

<b>Introducción.....</b>	<b>3</b>
Contexto del Informe.....	3
<b>Situación Problemática.....</b>	<b>4</b>
Hipótesis.....	4
<b>Objetivos generales.....</b>	<b>5</b>
<b>Marco Teórico.....</b>	<b>6</b>
Base de datos multidimensional.....	6
Data Warehouse.....	6
Metodología de diseño Kimball.....	6
ROLAP.....	7
<b>Desarrollo.....</b>	<b>8</b>
Diseño.....	10
Modelado dimensional.....	13
<b>Conclusión.....</b>	<b>16</b>
<b>Bibliografía.....</b>	<b>17</b>

# Introducción

## Contexto del Informe

En el marco de una beca de formación y práctica educativa dentro de la asignatura "Espacio Integrador 2" de la carrera Tecnicatura universitaria en procesamiento y explotación de datos dictada en la Facultad de Ingeniería de la Universidad Nacional de Entre Ríos, se realizó este informe a partir de la necesidad de resolver un problema planteado por la Secretaría de Extensión de dicho establecimiento, donde el objetivo principal es el procesamiento de ofertas laborales publicadas en formato HTML durante el periodo comprendido entre 2012 y 2022.

El contexto de nuestra problemática surge de un período en el que el mundo laboral experimentó notables transformaciones en cuanto a la demanda de habilidades y perfiles profesionales. Las empresas e instituciones se han adaptado a un entorno crecientemente digitalizado y globalizado, lo que ha impulsado una evolución en los requisitos y competencias que buscan en sus futuros colaboradores.

La problemática se ve agravada por la ausencia de una base de datos centralizada que almacene y organice las ofertas laborales en formato HTML. Esta carencia impide un proceso fluido de recopilación y análisis de datos estadísticos relevantes sobre el perfil y las habilidades requeridas por las empresas e instituciones para cubrir puestos relacionados con los graduados de la FIUNER.

# Situación Problemática

El problema que enfrentamos puede definirse de la siguiente manera:

Nos encontramos ante el desafío de procesar y analizar ofertas laborales en formato HTML correspondientes al periodo 2012-2022.

Esta tarea se dificulta significativamente debido a la carencia de una base de datos centralizada que limita la eficacia en la recopilación y el análisis de información crucial para la formación de sus estudiantes y la adaptación de sus programas académicos al entorno laboral actual.

Al tener un gran volumen de fuentes de datos, que deben ser revisados periódicamente y deben ser filtrados de alguna manera. Es necesario buscar una herramienta que ayude a cumplir los objetivos. Con esto surge, **¿Qué herramienta/s se puede/n diseñar para unificar de forma automatizada las fuentes de datos ?**

## Hipótesis

Dentro del contexto descrito previamente, es posible plantear una hipótesis que oriente nuestra investigación y brinde una base sólida para abordar la problemática. Dada la naturaleza del problema, podemos formular la siguiente hipótesis:

### **Hipótesis:**

Si logramos desarrollar un sistema automatizado, utilizando una base de datos multidimensional con ETL, donde sea efectivo el procesamiento de ofertas laborales en formato HTML, lo cual incluye la extracción de datos relevantes, su organización y definición de métricas. Para obtener una comprensión más precisa de las demandas laborales específicas de las empresas e instituciones para cubrir puestos relacionados con graduados de las carreras de la FIUNER. Esto permitirá que sea utilizable a largo plazo, de forma automatizada y accesible para los trabajadores de las secretaría de extensión.

Esta hipótesis se basa en la premisa de que la recopilación sistemática y el análisis de datos provenientes de ofertas laborales permitirán identificar patrones y tendencias que reflejen las necesidades del mercado laboral. Además, supone que la disponibilidad de esta información ayudará a la FIUNER a adaptar sus programas académicos y actividades de formación para satisfacer mejor estas demandas, preparando así a sus estudiantes para una inserción laboral más exitosa.

Es importante destacar, que esta hipótesis es una suposición inicial que se someterá a pruebas y verificaciones a lo largo del proceso de investigación. También se aclara se que acotará al diseño de la herramienta y su implementación se deja para abordar más adelante en otra instancia

## Objetivos generales

El objetivo principal de este trabajo consiste en recopilar datos estadísticos detallados sobre los perfiles profesionales demandados en el mercado laboral actual, así como las habilidades específicas que los empleadores valoran en los candidatos. Este proyecto tiene como finalidad proporcionar una comprensión profunda de las necesidades del mercado laboral actual.

Para alcanzar este objetivo principal, se plantea como primer objetivo específico desarrollar una base de datos multidimensional robusta que sirva como infraestructura para llevar a cabo análisis estadísticos de manera más sencilla y accesible. Esta base de datos multidimensional facilitará la organización, gestión y análisis de la información recopilada, permitiendo así una interpretación más precisa y detallada de los datos relacionados con los perfiles profesionales y las habilidades demandadas por los empleadores. Al lograr este objetivo específico, se sentarán las bases para obtener una comprensión integral de las dinámicas del mercado laboral y cumplir con el objetivo general del proyecto

# Marco Teórico

## Base de datos multidimensional

### Data Warehouse

Un almacén de datos (data warehouse, DW) según Inmon (Inmon02, Imhoff & Galletto 03), es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se trata, sobre todo, de un historial completo de la organización, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y divulgación eficiente de datos (especialmente con herramientas OLAP, de procesamiento analítico en línea). Por otra parte Kimball (Kimball 98) la define como “una copia de los datos transaccionales estructurados específicamente para consultas y análisis”

### Metodología de diseño Kimball

En el contexto del diseño de almacenes de datos (Data Warehousing), la metodología propuesta por Kimball se basa en lo que él denomina el Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle), tal como se describe en las obras de Kimball et al. (1998, 2008) y Mundy y Thornthwaite (2006). Este enfoque estructura el ciclo de vida del proyecto de Data Warehousing en torno a cuatro principios fundamentales. En primer lugar, destaca la importancia de centrarse en el negocio, enfocándose en la identificación precisa de los requerimientos del negocio y su valor asociado. Esta fase implica el desarrollo de relaciones sólidas con el negocio, así como el perfeccionamiento del análisis y la competencia consultiva de los implementadores. El segundo principio aboga por la construcción de una infraestructura de información adecuada, caracterizada por ser única, integrada, fácil de usar y de alto rendimiento, capaz de reflejar la amplia gama de requerimientos identificados en la empresa. Además, se propone realizar entregas en incrementos significativos, estructurando la creación del almacén de datos en plazos de 6 a 12 meses, utilizando el valor de negocio de cada elemento identificado para determinar la

secuencia de aplicación de los incrementos. Este enfoque se asemeja a las metodologías ágiles utilizadas en la construcción de software. Por último, el cuarto principio aboga por ofrecer la solución completa, proporcionando todos los elementos necesarios para entregar valor a los usuarios de negocios. Esto incluye un almacén de datos sólido, bien diseñado y de calidad probada, así como herramientas de consulta ad hoc, aplicaciones para informes y análisis avanzado, capacitación, soporte, sitio web y documentación (Rivadera, 2010)

## ROLAP

Dentro del panorama de almacenes de datos, el enfoque de almacén de datos Relational OLAP (ROLAP), como lo define Kimball y Ross (fuente), se distingue como un tipo específico de almacén de datos multidimensional. En este contexto, un almacén de datos ROLAP se caracteriza por almacenar los datos en una base de datos relacional. En este modelo, los datos se organizan en tablas, donde cada tabla representa una dimensión o medida del almacén de datos. La peculiaridad distintiva de los almacenes de datos ROLAP radica en que las consultas a estos almacenes se realizan mediante el lenguaje de consulta estructurado (SQL), permitiendo a los usuarios acceder y analizar eficientemente los datos almacenados en el sistema. Este enfoque fusiona las ventajas de las estructuras relacionales con las capacidades analíticas multidimensionales, proporcionando así una plataforma flexible para la exploración y análisis de datos en el contexto de un almacén de datos (Kimball & Ross, "Data Warehousing: Concepts, Techniques, and Applications")

La elección de ROLAP para nuestro proyecto se fundamenta en su capacidad para gestionar volúmenes extensos de datos y su escalabilidad, aspectos que se alinean de manera óptima con los requisitos específicos de nuestra iniciativa. Este enfoque no solo posibilita el manejo eficiente de grandes cantidades de información, sino que también permite un análisis detallado y una exploración ad hoc de los datos. La característica distintiva de ROLAP en la capacidad de realizar análisis detallados facilita la investigación de patrones, tendencias o relaciones en los datos, incluso aquellos que no han sido previamente planificados. La flexibilidad inherente a ROLAP facilita la exploración de datos de manera creativa, utilizando herramientas analíticas para descubrir información relevante y obtener insights inmediatos. Esta versatilidad hace que ROLAP sea una elección apropiada para nuestro proyecto, ofreciendo la capacidad de adaptarse a las demandas cambiantes y proporcionando una plataforma robusta para la toma de decisiones informada

# Desarrollo

En el proceso de análisis detallado de las fuentes de información de la Secretaría de Extensión, se identificaron desafíos significativos relacionados con la programación no estructurada en HTML de los boletines y la variabilidad en la redacción. Esta complejidad subraya la necesidad de una solución eficiente y específica para el procesamiento de datos.

En respuesta a esta problemática, la Secretaría de Extensión planteó inicialmente la implementación de un software en Python, respaldado por un modelo de machine learning. Este software tiene como objetivo extraer y procesar el texto del código HTML, generando un archivo CSV con la información relevante. Además, se propuso el desarrollo de una interfaz de usuario para automatizar el proceso y facilitar la creación de gráficos y estadísticas basadas en el conjunto de datos procesado.

Dada la complejidad y amplitud de esta propuesta, se optó por una contra propuesta más enfocada y alineada con los conocimientos adquiridos durante el cursado de la materia de bases de datos multidimensionales. Se propuso diseñar un Data Warehouse con un proceso ETL automatizado y su integración con Power BI. Este enfoque no sólo centraliza los datos, sino que también permite la obtención de métricas, cumpliendo con los objetivos de la Secretaría. Además, se prevé la instalación de los programas necesarios y la capacitación del usuario final para garantizar la autonomía en el manejo del sistema.

Con la aceptación de esta propuesta, se decidió, en una primera fase, llevar a cabo el diseño del Data Warehouse bajo la metodología ROLAP, dejando aspectos más específicos, como la implementación del ETL, para una segunda fase del proyecto. Este enfoque estratégico permitirá abordar de manera eficiente y escalonada los desafíos planteados, asegurando una implementación exitosa del sistema propuesto.



## Imagen 1

Ejemplo de boletín, en este caso tiene tres ofertas laborales



Facultad de  
UNER Ingeniería



### Búsquedas Laborales

8 de febrero de 2021

#### Nuevas ofertas

---

**Posiciones a cubrir en empresa multinacional de Equipos Médicos.**

**- Gerente de Ventas.**  
Tendrá a su cargo 15 Ejecutivos de Ventas, 1 Analista de Negocios y 1 Especialista de producto. Asimismo llevará adelante proyectos de mejora de procesos interdepartamentales y el desarrollo de competencias de su equipo.

Se orienta a graduados de **Bioingeniería**, Ing. Biomédica, Farmacia, Bioquímica o Lic. en Comercialización con 5 años de experiencia en empresas de dispositivos médicos, inglés intermedio y disponibilidad para viajar.

Se ofrece auto compañía entre otros beneficios y posibilidades crecimiento en un entorno de primer nivel.

Los interesados pueden enviar CV a [selector@intersearch.com.ar](mailto:selector@intersearch.com.ar) con Asunto "Gte. Ventas 04"

**- Analista de Asuntos Regulatorios Sr**  
Para llevar adelante un proyecto de cambio de registros ante las autoridades sanitarias.

Serán sus funciones:

- Preparar presentaciones para asegurar aprobaciones para el lanzamiento de productos
- Mantener actualizada la lista de requisitos para el armado del expediente de registro de producto y comunicar los requisitos al equipo de asuntos regulatorios de la unidad de negocios.
- Verificar y revisar la documentación necesaria para cumplir con ANMAT, con la Autoridad Sanitaria de Chile y/o regulaciones corporativas
- Identificar los cambios de producto que representen un impacto en la región y distinguir cuándo un cambio requiere aprobación antes de la comercialización.
- Revisar etiquetado, los materiales promocionales, los cambios del producto y la documentación de los cambios que requieren la aprobación de la agencia regulatoria.

Se orienta a graduados de Bioquímica, Farmacia, **Bioingeniería** o carreras afines, con 3 años de experiencia en dispositivos médicos (pref. reactivos de diagnóstico) e inglés avanzado.

Los interesados pueden enviar CV a [selector@intersearch.com.ar](mailto:selector@intersearch.com.ar) con Asunto "Reg. 03"

**- Ingeniero de Servicio Técnico Jr.**  
Liderará la instalación y puesta en funcionamiento de los equipos. El puesto incluye la capacitación y atención de los usuarios, por lo que se requieren excelentes habilidades de comunicación y orientación al detalle. Asimismo será responsable de mantener un nivel adecuado de inventario de repuestos.

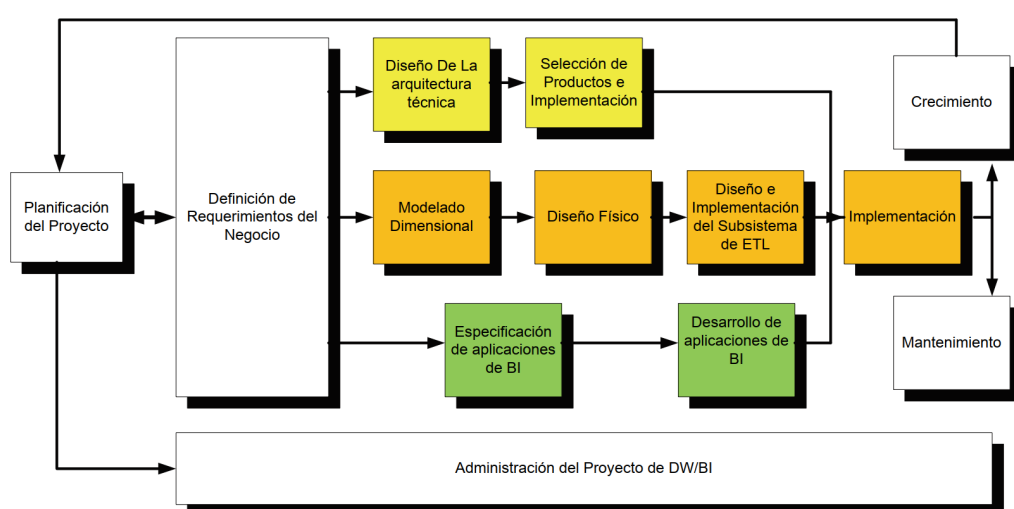
Se requiere graduado de Ing. en Sistemas, Mecánica, Electrónica, **Bioingeniería** o afines, con 2-3 años de experiencia en la función, inglés intermedio y disponibilidad para viajes al Interior.

Los interesados pueden enviar CV a [selector@intersearch.com.ar](mailto:selector@intersearch.com.ar) con Asunto "Ing. 02"

## Diseño

En el marco del diseño del Data Warehouse, se optó por implementar la metodología Kimball, como se ilustra en la Figura 1. Esta metodología desempeña un papel central al establecer los requisitos iniciales, sirviendo como la base fundamental para las fases subsiguientes del proyecto. La Figura 1 destaca tres rutas fundamentales: tecnología en el camino superior, datos en el camino del medio, y aplicaciones de inteligencia de negocios en el camino inferior

**Figura 1**



**Fig. 1:** Tareas de la metodología de Kimball, denominada *Business Dimensional Lifecycle* (Kimball et al 98, 08, Mundy & Thornthwaite 06)

Como se puede observar, esta metodología guía el proceso desde la etapa de planificación hasta el análisis de requerimientos. En esta última fase, se llevaron a cabo entrevistas con los usuarios finales en la Secretaría de Extensión. Estas interacciones permitieron identificar temas analíticos específicos que se detallan en la Tabla 1. Al avanzar desde la planificación hasta la identificación de los requisitos, se sientan las bases esenciales para la fase siguiente del proceso, consolidando así un enfoque estructurado y orientado a resultados.

**Tabla 1***Temas analíticos*

Tema analítico	Requerimiento inferido	Comentarios
Ofertas laborales	Análisis histórico de ofertas laborales	Por tipo de empresa, por provincia, por habilidad técnica y no técnica

Durante la fase de análisis de requerimientos bajo la metodología Kimball, se identificaron dimensiones clave que estructuran el Data Warehouse, brindando un marco integral para el procesamiento de las ofertas laborales estas fueron obtenidas de un csv modelo de resultados esperados proporcionado por la secretaria. Las dimensiones identificadas son:

- Empresa
- Especialidad
- Puesto
- habilidad blanda
- habilidad técnica
- región
- tiempo

Estas dimensiones ofrecen una visión completa y detallada de los aspectos fundamentales de las ofertas laborales, permitiendo una organización eficiente de los datos.

Asimismo, se estableció como métrica básica el número de ofertas laborales en un día. Esta métrica proporciona una visión general de la demanda laboral en el conjunto de datos, ofreciendo un indicador cuantitativo para evaluar la actividad laboral en un período específico.

En la etapa final del análisis de requerimientos, se definió un proceso de mantenimiento que se llevará a cabo cada 6 meses. Esta decisión se tomó para aliviar la carga de los datos, ya que se almacenarán en la computadora de la oficina y operará fuera de línea. Esta elección estratégica se basa en la consideración de que el Data Warehouse no requiere estar disponible las 24 horas.

## Modelado dimensional

En la siguiente etapa, el proceso de creación del modelo dimensional se llevó a cabo de manera iterativa, aproximándose al diagrama de flujo representado en la Figura 2. La identificación previa de dimensiones, nivel de granularidad y métricas fue esencial para guiar esta fase crucial del proyecto. El modelo adoptado para representar la estructura del Data Warehouse fue el diagrama de estrella, Figura 3

**Figura 2**

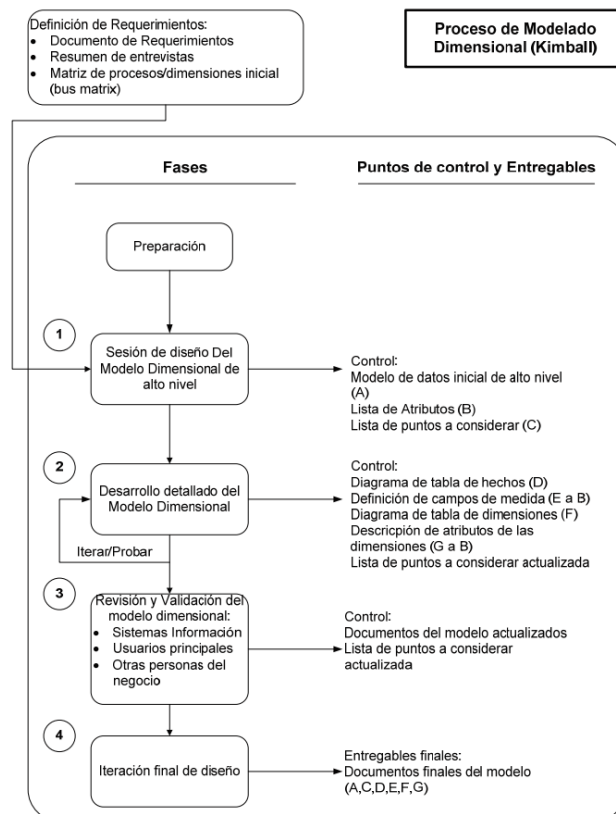
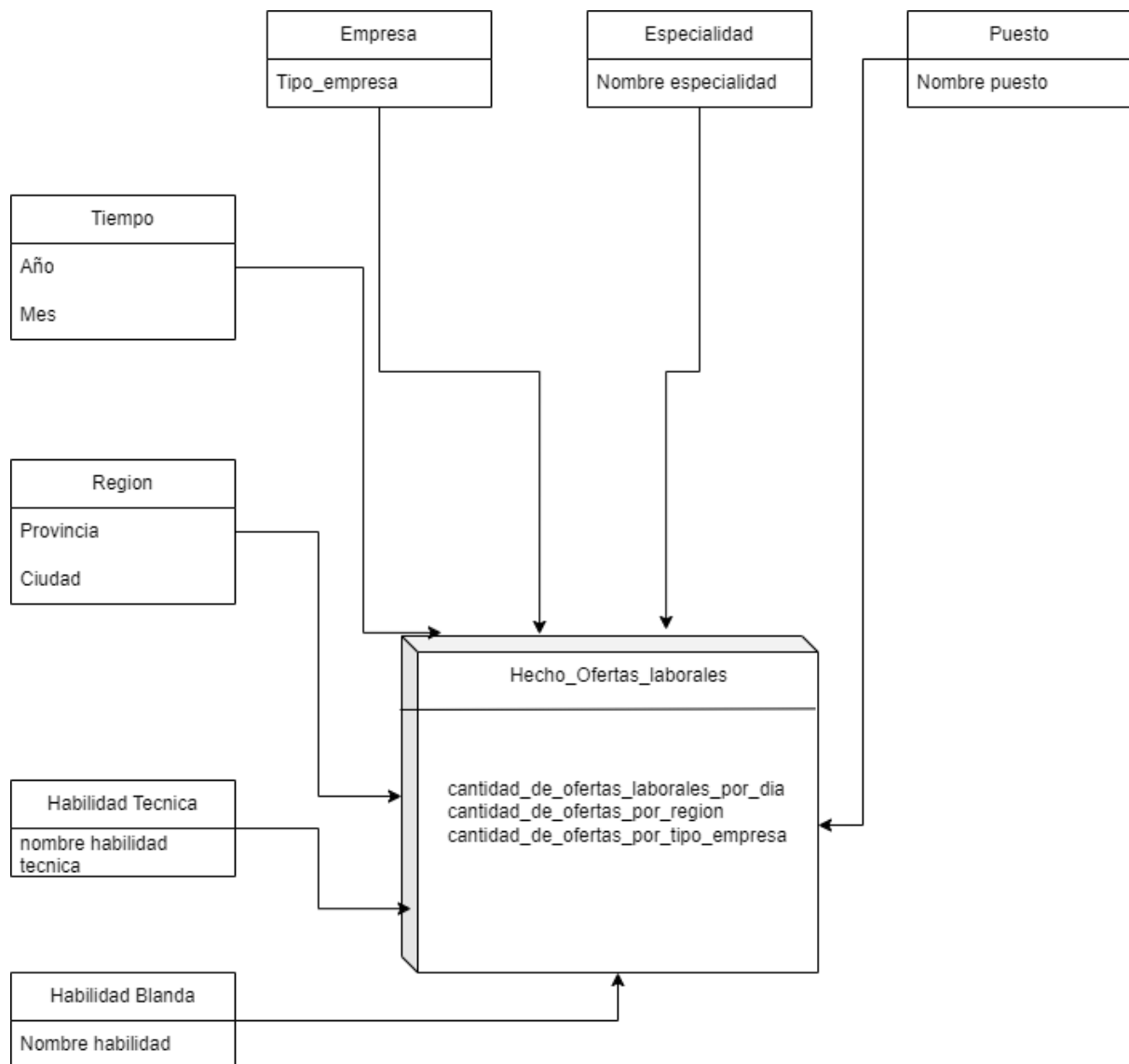


Fig. 2: Diagrama de flujo del proceso dimensional de Kimball (Mundy & Thornthwaite 06)

**Figura 3**

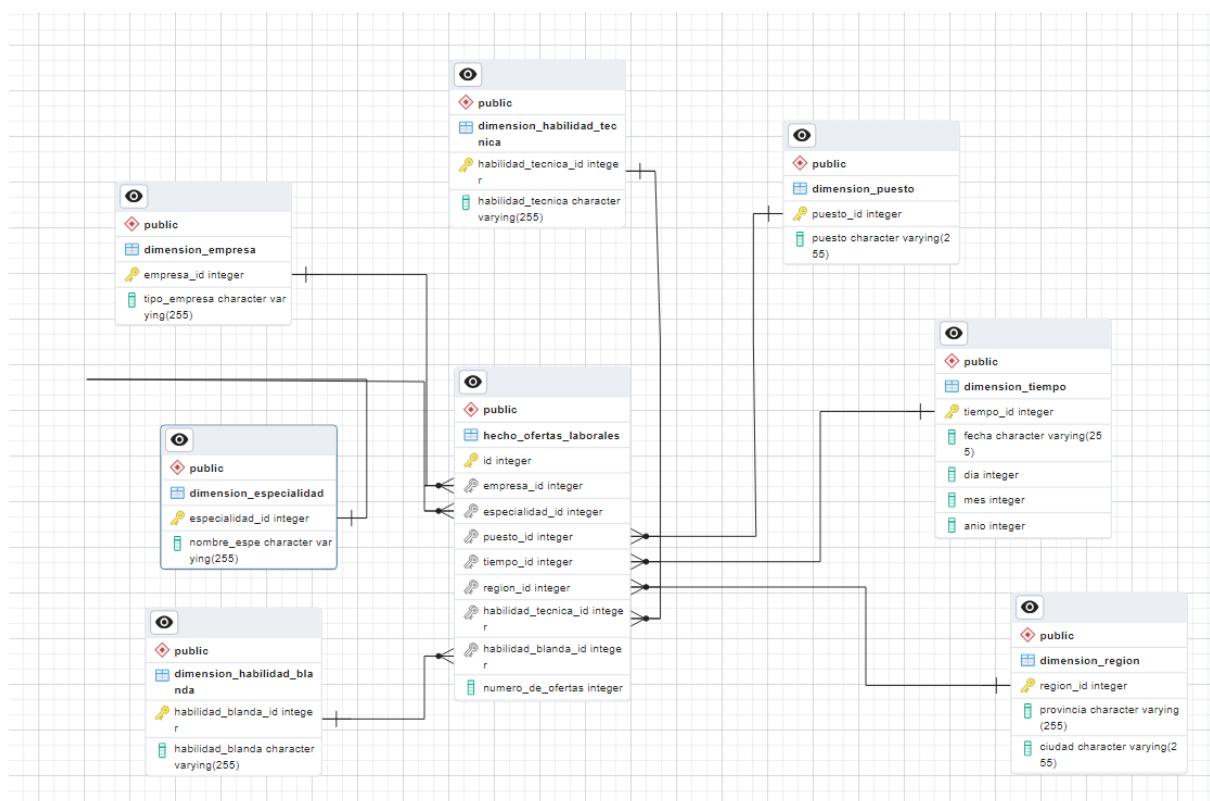
*Diagrama multidimensional*



Este enfoque de modelado dimensional proporciona una representación visual clara de la relación entre las dimensiones y la métrica seleccionada. La estructura de estrella permite un acceso eficiente a los datos y facilita consultas analíticas, al tiempo que mantiene una organización lógica y comprensible de la información. En la Figura 3, se presenta la tabla de hecho denominada "hecho ofertas laborales", que sirve como el núcleo central del modelo dimensional. Esta tabla está vinculada directamente a la métrica básica establecida, que es la "cantidad de ofertas laborales". Además, se muestran diversas métricas derivadas en la tabla de hecho, presentadas a modo de ejemplo. Estas métricas derivadas se calcularán mediante el uso de una herramienta de Business Intelligence (BI). La elección de esta herramienta permitirá la generación eficiente de análisis avanzados, informes personalizados y estadísticas detalladas.

En la Figura 4 se presenta el diagrama de tablas que refleja la estructura resultante de la implementación del [DDL](#). Este diagrama detalla las tablas, sus atributos y los tipos de datos asociados, ofreciendo una representación visual clara de la arquitectura de la base de datos diseñada.

**Figura 4**  
*Diagrama de tablas*



Con el diseño de la base de datos establecido, se abre el camino para comenzar a considerar el proceso de Extracción, Transformación y Carga (ETL). Sin embargo, se ha tomado la decisión estratégica de postergar la implementación de ETL para la segunda etapa del proyecto. Varios factores respaldan esta elección, incluyendo la necesidad de analizar detalladamente las fuentes de datos, la experimentación con la implementación de la propuesta inicial de la Secretaría de Extensión, la adquisición de conocimientos recientes sobre bases de datos multidimensionales y la complejidad inherente a la naturaleza de las fuentes de información.

En cuanto a los reportes, se optó por la implementación de Power BI en la segunda etapa del proyecto. Esta elección se basa en la eficacia y eficiencia que ofrece Power BI en comparación con la primera solución propuesta de codificar una interfaz de usuario desde cero.

## Conclusión

En respuesta a la pregunta inicial planteada al inicio del trabajo, se concluye que, sin lugar a dudas, el Data Warehouse emerge como la herramienta idónea para llevar a cabo un análisis estadístico y la generación de gráficos. La combinación de eficiencia en el diseño y desarrollo, junto con la agilidad inherente a la implementación de un Data Warehouse, supera significativamente la opción de desarrollar un software desde cero. La capacidad de integración que ofrece esta tecnología con herramientas de Business Intelligence (BI) como Power BI, entre otras, amplifica su utilidad en el análisis de datos. Las ventajas de estas herramientas facilitan de manera ágil e intuitiva las tareas analíticas, otorgando una mayor capacidad de análisis.

Además, es relevante destacar mi afinidad personal con la metodología de diseño propuesta por Kimball. Esta metodología no solo proporciona un marco robusto para el diseño de almacenes de datos, sino que también integra aspectos de gestión de proyectos. Esta conexión entre la planificación del proyecto y la metodología de diseño encaja de manera interesante, despertando un interés personal en profundizar aún más en este enfoque metodológico.



# Bibliografía

- Imhoff & Gallemmo, Mastering Data Warehouse Design: Relational and Dimensional Techniques, Wiley Publishing, 2003
- Inmon, Building the Data Warehouse, (Third Edition). John Wiley & Sons, 2002
- Kimball & Caserta, The Data Warehouse ETL Toolkit, Indianapolis, Wiley, 2004
- Rivadera, G. R. (2010). La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).