**TRABAJO PRÁCTICO**

**Business Intelligence & Datawarehousing**

**Cuatrimestre: 02 – Año: 2021**



**UNIVERSIDAD ARGENTINA DE LA EMPRESA**

**FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS**

**Integrantes del Grupo:**

Abelini, Fernando Nicolas 1106578

Bollella, Lucas Emiliano 1114426

Bustamante Zuluaga, July Marcela 1114012

Mirabile Viola, Fernando Pablo 1110820

Shevchuk Calo, Miguel Omar 1116936

Strficek, Ivo Andres 1118312

Vazquez Crossetto, Juan Manuel 1106850

**Profesora:**

Combet, Mercedes Cecilia

**Tabla de Contenidos:**

1. [**Situación inicial, problemas detectados y mejoras propuestas**](#_heading=h.lcjvl0616yd6) **2**

[**Situación inicial**](#_heading=h.qqb7f7boths5) **2**

[**Problemas detectados**](#_heading=h.9gdkk4yw2922) **5**

[**Mejoras propuestas**](#_heading=h.schbuk3hwiah) **5**

1. [**Diseño del modelo final – (Diagrama lógico documentado)**](#_heading=h.7ggld5tonbs2) **6**
2. [**Informe resumen de los problemas de calidad de datos en los sistemas fuente.**](#_heading=h.ne3egrdu7hd7) **7**

# Situación inicial, problemas detectados y mejoras propuestas

## Situación inicial

Desde el negocio se provee un set de datos de origen en diferentes formatos como .xls, csv y accdb

Los orígenes de datos mencionados, serán utilizados para poder dar cumplimiento a los requerimientos del negocio que llevan las siguientes premisas:

1. El área de inscripciones requiere poder consultar:

* *Evolución de las inscripciones en cada temporada, según curso, tipo de sala y profesor.*
* *Ranking de cursos más pedidos por temporada, horario e instructor.*
* *Cumplimiento de capacidad real versus esperada.*
* *Características de la dotación de instructores para asignación de clases según modalidad.*

2. El área de Pedagogía requiere poder consultar:

* *Análisis de demanda por duración de los cursos.*
* *Índice de asistencia a clases por modalidad.*
* *Disponibilidad de instructores por temporada y tasa de asistencia a clases.*

Se encontraron algunas inconsistencias y faltantes que son importantes para poder conocer a más detalle el funcionamiento de los procesos involucrados en los requerimientos, por lo cual se elaboró un cuestionario que sirvió como referencia para hacer una entrevista de relevamiento y así poder corroborar las necesidades del negocio.

**Personal entrevistado**: Mercedes Cecilia Combet

**Rol**: Jefe operaciones.

Cuestionario:

1- ¿ A qué se refiere con temporada?

**Respuesta:** Son las estaciones del año, como primavera, verano, etc.

2- ¿ Que serían grupos dimensionados?

**Respuesta:** Son tamaños pequeños y medianos.

4- ¿ Cuáles son las características de la dotación de instructores para asignación de clases según modalidad?

**Respuesta:** todo lo que tengamos de instructores , la edad, la antigüedad, y la familia. Cuanto más se pueda sacar mejor. Por ejemplo los que tienen hijos se puede suponer que se puede asignar cursos remotos.

5- ¿ Cómo sacamos la asistencia?

**Respuesta:** No tenemos ese dato.

7- • Construcción de los procesos ETL, utilizando la herramienta adquirida por el área de tecnología: Pentaho.

**Respuesta:** Descargar Spunk

8- ¿ Qué es el “comissionRate” y la comisión y si es por curso?.

comissionRate: Porsentaje que tiene el profesor de lo que se paga

comision: lo que se paga. En función de todos los cursos se puede sumar las comisiones y saber cuanto se le pagó al profesor.

9- En la disponibilidad del profesor es necesario tener en cuenta la locación o con las horas es suficiente? Porque no encontramos la dirección de la sala.

**Respuesta:** Si es necesario tener en cuenta la locación.

para tener en cuenta, la ubicación de la sala y la ubicación del docente, aunque no son posibles cruzar la ubicación de la sala con la del docente, para el caso de los docentes que tengan familia se le puede ofrecer cursos remotos.

10- No vemos nada de alumno, lo dimensionamos?

**Respuesta:** No

11- De donde obtener la capacidad esperada?

**Respuesta:** Está en temporada.

12- El cumplimiento se saca de la capacidad real y esperada.

## Problemas detectados

* Poner en dimensión tiempo día y mes y darse cuenta después de que no se cuenta con esa información.
* Se pensó en dimensionar una tabla alumnos, para luego encontrar que no existía ningún dato como tal.
* En el relevamiento se encontró que no se posee información de la tasa de asistencia, así que no se puede responder a esa pregunta.
* Errores de cursos.
* Datos inconsistentes.
* Datos faltantes y campos null como por ejemplo encontrar dos docentes .
* Existen errores en los cursos.
* Existen registros duplicados
* Nombre de un mismo profesor de formas diferentes lo que dificulta el análisis de la información

## Mejoras propuestas

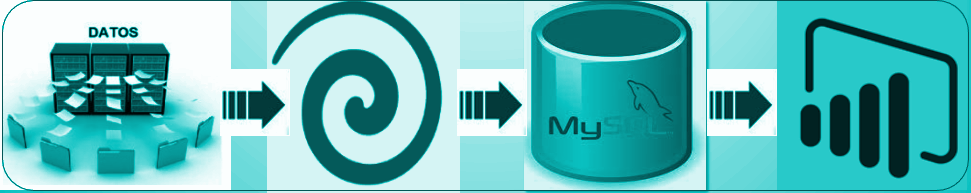
Dado que se posee múltiples datos de los cuales se puede obtener la información necesaria para poder comunicar las cuestiones planteadas, se propone crear un **datawarehouse** con el fin de contener los datos de los sistemas transaccionales del negocio.

Este datawarehouse contendrá datos organizados en áreas temáticas con versiones de los mismos registros según fueron variando en el tiempo.

La arquitectura de nuestro datawarehouse usará el modelo de capas ETL ósea **Extracción de datos, transformación y carga** como se explica mejor a continuación:

1. **Extracción:** En la capa de extracción los datos se extraen de los *orígenes* que es una variedad de fuentes internas de Commodore64.
2. **Transformación:** Debido a que los datos los tenemos en diferentes formatos necesitamos una capa de extracción para poder adaptar los datos a los requerimientos utilizando la herramienta **PENTAHO que nos permitira manipular los datos y manejar errores en caso de que ocurra**.
3. **Carga:** Estos datos van a estar almacenados en una base de datos **Mysql** para poder tener independencia de las fuentes de origen.

Finalmente se propone agregar una capa de presentación que es de vital importancia para que los usuarios del negocio puedan interactuar con los datos limpios y organizados, dicha herramienta será **PowerBI**.



**Figura 1: Arquitectura datawarehouse.**

Esta arquitectura va poder dotar al usuario de la capacidad de consultar y obtener información sobre la oferta de cursos disponible, los cursos realizados, la disponibilidad y todo lo que se desee analizar con respecto al negocio y así poder desarrollar informes automatizados que sirva a futuras tomas de decisiones para permitir que el negocio se siga expandiendo.

# 2. Diseño del modelo final – (Diagrama lógico documentado)

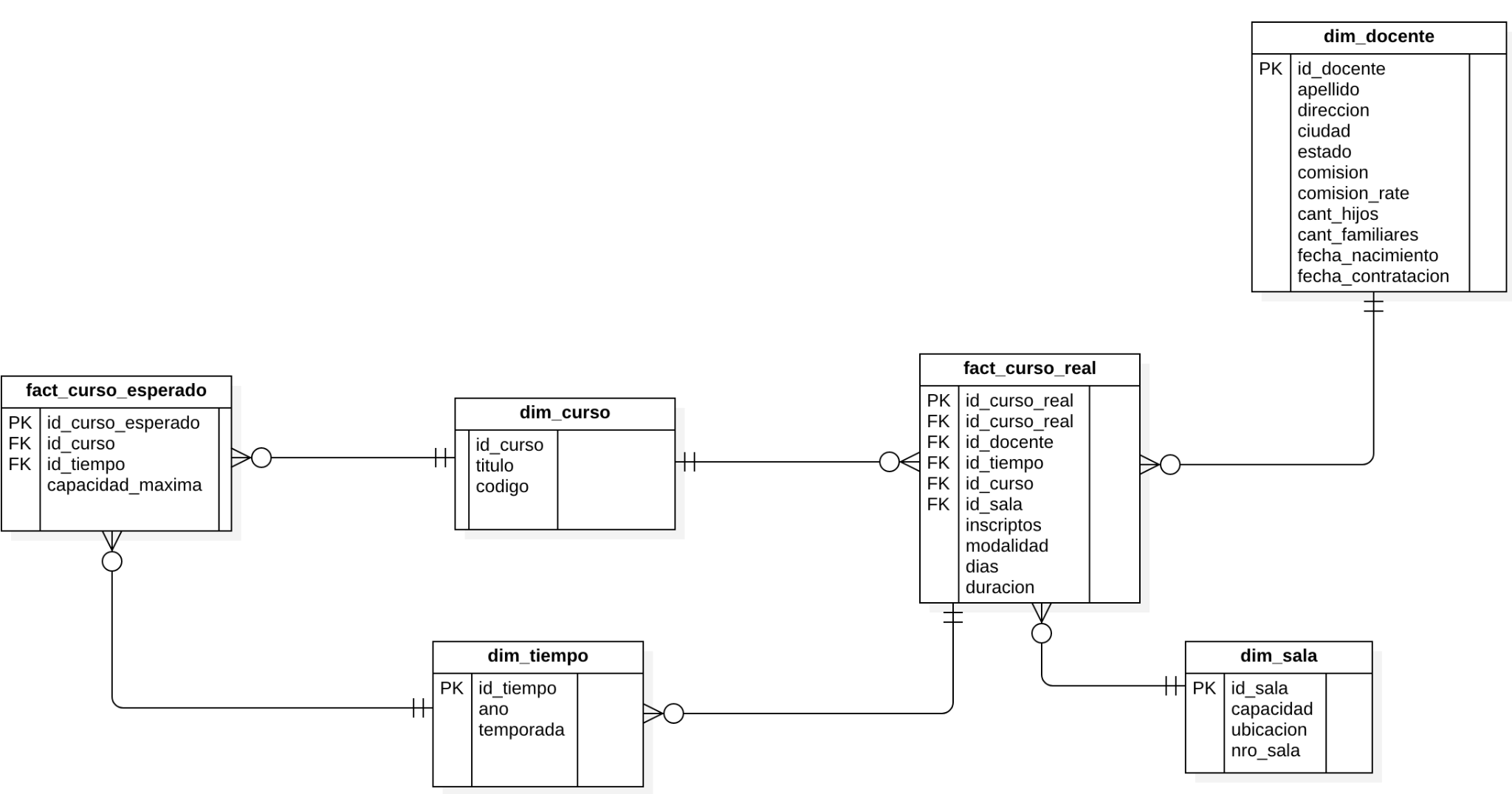
Dentro de los procesos de negocio que están relacionados con los requerimientos de negocio se encontró las siguientes *dimensiones* para nuestro modelo, las dimensiones propuestas son:

* Docente
* Sala
* Curso
* Tiempo

Para las tablas de hechos se prestó especial atención a los indicadores que se desea observar, para facilitar la disponibilidad de la información por eso se van a usar dos fact table uno para curso esperado y otro curso real, ya que el grano del negocio son los cursos, pero dichos cursos se deben dividir en antes y despuės de dictar los cursos para poder responder a las consultas.

El modelo lógico que se eligió fue en constelación que deriva del esquema estrella, ya que necesitamos dos tablas de hechos.

Las ventajas que encontramos en este diseño es que es fácil de entender y reduce el número de uniones, además de que genera metadatos más sencillos.



**Figura 2: Diagrama lógico.**

# 3. Informe resumen de los problemas de calidad de datos en los sistemas fuente.

Archivo: **CourseCapacity.csv**

Casos:

- Existe el course Title “C Programming” y “C Programing”

- Existe el course Title “Database” y “Databases”

- Existe el course Title “JAVA” y “JAVA Programming”

- Existe el course Title “Prototyping & Rapid Appl Dev” y “Prototyping & Rapid Appl Dev.”

- Existe el course Title “Sem-Mgt Info Sys” y “Sem-Mgt Info Systems”

- Existe el course Title “Visual Basic Programming” y “Visual Basic”

Luego de evaluar la base de datos, se tomaron los siguientes supuestos y se aplicaron las siguientes correcciones.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Columna** | **Valor** | **Supuesto problema** | **Corrección** | **Problema** |
| Course Title | C Programing | Hubo un error de tipeo y faltó una m en el valor del campo | Se remplaza “C Programing” por “C Programming” | Precisión |
| Course Title | Database | Hubo un error de tipeo y falto una s en el valor del campo, la cual agrega la pluralidad. | Se remplaza “Database” por “Databases” | Precisión |
| Course Title | JAVA | Se encontró inconsistencia en un campo muy similar al otro pero sin uso en la base de datos. | Se remplaza “JAVA” por “JAVA Programming” | Precisión |
| Course Title | Prototyping & Rapid Appl Dev. | Hubo un error de tipeo y sobra un “.”en el valor del campo | Se remplaza “Prototyping & Rapid Appl Dev.” por “Prototyping & Rapid Appl Dev” | Precisión |
| Course Title | Sem-Mgt Info Systems | Se encontró inconsistencia en un campo muy similar al otro pero sin uso en la base de datos. | Se remplaza “Sem-Mgt Info Systems” por “Sem-Mgt Info Sys” | Precisión |
| Course Title | Visual Basic Programming | Se encontró inconsistencia en un campo muy similar al otro pero sin uso en la base de datos. | Se remplaza “Visual Basic Programming” por “Visual Basic” | Precisión |

Archivo: **Instructors.xlsx**

Casos:

- Existen dos registros sin valor en la columna Instructor\_Nu

- Existe el Instructor “Carte” y “Carter”

- Existe el Instructor “Mokrenko” y “Morenko”

- Existe un Instructor “Carter/Ferguson” tiene valor compuesto y con caracteres especiales.

Luego de evaluar la base de datos, se tomaron los siguientes supuestos y se aplicaron las siguientes correcciones.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Columna** | **Valor** | **Supuesto problema** | **Corrección** | **Problema** |
| Instructor | Morenko | Hubo un error de tipeo y sobra una “m” en el valor del campo | Se remplaza “Morenko” por “Mokrenko” | Precisión |
| Instructor\_Nu |  | Existen 2 registros sin valor en dicha columna, la cual consideramos obligatoria. | Se sigue con la correlación de dicho campo y se asignan los valores 56 y 57. | Validez |
| Instructor | Carter/Ferguson | Se ingresó un apellido con caracteres especiales para el cual suponemos que es un apellido doble. | Se mantiene el campo sin cambios, el cual es utilizado en la tabla Fall 2021 | Completitud |
| Instructor | Carter y Carte | Valores muy similares, pero como ambos tienen cursos asignados suponemos que son diferentes apellidos/instructores. | Se mantiene el campo sin cambios. | Unicidad |

Archivo: **Room Types.xlsx**

Casos:

- Un registro de la columna habitación está compuesto por números y letras (07H).

- Un registro de la columna type no tiene carácter especial (twa)

Luego de evaluar la base de datos, se tomaron los siguientes supuestos y se aplicaron las siguientes correcciones.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Columna** | **Valor** | **Supuesto problema** | **Corrección** | **Problema** |
| ROOM | 07h | Se encontró un valor compuesto por números y letras, pero al revisar la base de datos, suponemos que es correcto. | Se mantiene el campo sin cambios. | Validez |
| TYPE | twa | En principio se supone un error de tipeo al querer escribir ”w” en lugar de una “/”, pero por conclusión final llegamos a que se refiere a “table with arms” | Se mantiene el campo sin cambios. | Validez |

Archivo: **course offerings.accdb**

No se realizaron correcciones sobre el mismo.

# 4. Detalle de los pasos del proceso de ETL, procedimiento para ejecución en caso de cancelaciones.

TBD

# 

# 

# 

# 

# 5. Código fuente procesos ETL.

TBD

# 

# 6. Reportes construidos y Dashboard publicados.

# TBD(chicos)

# 

# 7. DDLs correspondientes a la versión final del modelo de datos.

create database if not exists commodore64;

CREATE TABLE `commodore64`.`dim\_docente` (

`id\_docente` INT NOT NULL,

`apellido` VARCHAR(255) NOT NULL,

`direccion` VARCHAR(255) NULL,

`ciudad` VARCHAR(255) NULL,

`estado` VARCHAR(255) NULL,

`fecha\_contratacion` VARCHAR(45) NULL,

`cant\_familiares` INT NULL,

`fecha\_nacimiento` VARCHAR(45) NULL,

`cant\_hijos` INT NULL,

`comision` DECIMAL NULL,

`comision\_rate` FLOAT NULL,

PRIMARY KEY (`id\_docente`));

CREATE TABLE `commodore64`.`dim\_salas` (

`id\_sala` INT NOT NULL,

`nro\_sala` VARCHAR(45) NULL,

`capacidad` INT NULL,

`ubicacion` VARCHAR(255) NULL,

`tipo` VARCHAR(255) NULL,

PRIMARY KEY (`id\_sala`));

CREATE TABLE `commodore64`.dim\_curso (

`id\_curso` INT NOT NULL,

`titulo` VARCHAR(255) NOT NULL,

`codigo\_curso` VARCHAR(255) NOT NULL,

PRIMARY KEY (`id\_curso`));

CREATE TABLE commodore64.dim\_tiempo (

id\_tiempo INT auto\_increment NOT NULL,

ano INT NOT NULL,

temporada VARCHAR(50) NOT NULL,

CONSTRAINT dim\_tiempo\_PK PRIMARY KEY (id\_tiempo)

);

CREATE TABLE `commodore64`.fact\_curso\_esperado (

`id\_curso\_esperado` INT UNIQUE,

`id\_curso` INT NOT NULL,

`capacidad\_maxima` INT NOT NULL,

`id\_tiempo` INT NOT NULL,

PRIMARY KEY (`id\_curso\_esperado`),

CONSTRAINT fact\_cursoEsperado\_FK FOREIGN KEY (id\_curso) REFERENCES commodore64.dim\_curso(id\_curso),

CONSTRAINT fact\_cursoEsperado\_FK\_2 FOREIGN KEY (id\_tiempo) REFERENCES commodore64.dim\_tiempo(id\_tiempo)

);

CREATE TABLE commodore64.fact\_curso\_real (

id\_curso\_real INT auto\_increment,

id\_curso INT NOT NULL,

id\_docente INT NOT NULL,

id\_tiempo INT NOT NULL,

id\_sala INT NOT NULL,

modalidad varchar(100) NOT NULL,

dias varchar(100) NOT NULL,

inscriptos INT NOT NULL,

duracion varchar(100) NOT NULL,

CONSTRAINT fact\_cursoReal\_PK PRIMARY KEY (id\_curso\_real),

CONSTRAINT fact\_cursoReal\_FK FOREIGN KEY (id\_curso) REFERENCES commodore64.dim\_curso(id\_curso),

CONSTRAINT fact\_cursoReal\_FK\_1 FOREIGN KEY (id\_docente) REFERENCES commodore64.dim\_docente(id\_docente),

CONSTRAINT fact\_cursoReal\_FK\_2 FOREIGN KEY (id\_sala) REFERENCES commodore64.dim\_salas(id\_sala),

CONSTRAINT fact\_cursoReal\_FK\_4 FOREIGN KEY (id\_tiempo) REFERENCES commodore64.dim\_tiempo(id\_tiempo)

);

# 8. Guía rápida de uso para los reportes y Dashboard para usuarios novatos.