Juan Luis Acebal Rico

PR 1 BDA

- *Introducción*
  - *Este documento es para desarrollar y diseñar un almacén de datos, para el soporte al análisis de los resultados de las elecciones presidenciales en Estados Unidos.*
  - *Según el contexto del caso y las fuentes de datos proporcionadas, hay que desarrollar concretamente en este caso práctico el análisis de requisitos, de todas las fuentes de datos, análisis funcional y diseño del modelo conceptual, lógico y físico del almacén de datos.*

1. Análisis de requisitos.

En base a toda la información dada, a la guía de muestra, y el conjunto de todo lo que tenemos, creo que:

Estamos ante el análisis de los resultados de las elecciones de Estados Unidos, estos resultados conllevan posibles tomas de decisiones, según quien gane.

- *De los requisitos identificados están:*
  - *Por año*
  - *Por estados*
  - *Por tipo*
- *De otros requisitos, quizás hay que valorar otras perspectivas de análisis, lo menciono por si tuviera cabida; tales como afecta la demografía, la participación electoral, o si tuviéramos la utilización concreta del almacén de datos, se podía valorar incluir aspectos socioeconómicos de los estados, por ejemplo, para además de saber los cambios posibles según el signo político de quien gane, en qué zonas geográficas o poblaciones será más pronunciado el cambio. Se podrían llegar a responder preguntas de hasta qué punto hay una correlación entre el PIB per cápita y las tendencias políticas, o el crecimiento económico de cada estado o de EEUU según el partido político en el poder. Aquí se podrían hacer muchas métricas, aunque es una idea, lo pongo como complemento del ejercicio ya que no sé si excede lo que se demanda o como no se han dado esos datos, no puedo utilizarlos.*

- *Dicho esto, las preguntas clave que debemos responder:*
  - *¿Cuál ha sido la evolución por partido a lo largo del tiempo?*
  - *¿Cuál ha sido la evolución de los índices bursátiles en años electorales en EEUU?*
  - *¿Cuál ha sido la evolución de los índices bursátiles después de los años electorales?*
  - *¿Cómo han variado los resultados en cada estado?*
  - *¿Cómo cambian los resultados según cada estado?*
  - *¿Cuál es la participación electoral en las elecciones por estado?*
  - *¿Cómo ha variado la participación electoral?*
  - *¿Qué partido ha tenido más éxito histórico electoral? (Desde 1976)*

2. Análisis de todas las fuentes de datos

Archivo 1976-2020-president.tab:

Aquí están las presidenciales en Estados Unidos desde 1976 hasta 2020. Incluye información sobre los candidatos, los partidos a los que pertenecen y los votos obtenidos.

Estructura del archivo:

Nombre del campo	**Descripción**	**Tipo**	**Ejemplo**
year	Año de la elección	Entero	1976
state	Nombre del estado	Texto	Arizona
state_po	Código de dos letras del estado	Texto	AL
state_fips	Código FIPS del estado	Entero	1
state_cen	Código CEN del estado	Entero	63
state_ic	Código IC del estado	Entero	41
office	Cargo en elección	Texto	US PRESIDENT
candidate	Nombre del candidato	Texto	CARTER, JIMMY
party_detailed	Partido del candidato	Texto	COMMUNIST PARTY USE
writein	Está ya en las papeletas de elección o hace falta escribirlo.	Boleano	FALSE
candidatevotes	Votos obtenidos por el candidato	Entero	9198
totalvotes	Total de votos del estado	Entero	1182850
version	Version de los datos	Entero	20210113
notes	Notas adicionales	Texto	null
party_simplified	Partido político del candidato abreviado	Texto	OTHER

Observaciones: state, state_po , state_fips, state_cen y state_ic ofrecen la misma información en todas las tablas, están relacionadas entre si. Notes siempre está vacío.

Total de registros: 4287.

2) Archivo "1976-2020-senate.tab":

Éste archivo contiene las elecciones al Senado de Estados Unidos de 1976 a 2020. Igual que el anterior nos da los candidatos, los partidos políticos y los votos obtenidos.

Estructura del archivo:

Nombre del campo	**Descripción**	**Tipo**	**Ejemplo**
year	Año de la elección	Entero	2020
state	Nombre del estado	Texto	Arizona
state_po	Código de dos letras del estado	Texto	AL
state_fips	Código FIPS del estado	Entero	1
state_cen	Código CEN del estado	Entero	63
state_ic	Código IC del estado	Entero	41

office	*Cargo en elección*	*Texto*	*US SENATE*
district	*Distrito electoral*	*Texto*	*statewide*
stage	*Etapa/vuelta de la elección*	*Texto*	*Runoff*
special	*Es una elección especial*	*Boleano*	*False*
candidate	*Nombre del candidato*	*Texto*	*JOE MANCHIN*
party_detailed	*Partido del candidato*	*Texto*	*DEMOCRAT*
writein	*Está ya en las papeletas de elección o hace falta escribirlo.*	*Boleano*	*TRUE*
mode	*Modo de votación*	*Texto*	*Total*
candidatevotes	*Votos obtenidos por el candidato*	*Entero*	*288808*
totalvotes	*Total de votos del estado*	*Entero*	*582911*
unofficial	*Es un resultado oficial o no.*	*Boleano*	*True*
version	*Version de los datos*	*Entero*	*20210114*
Party_simplified	*Partido político del candidato abreviado*	*Texto*	*DEMOCRAT*

Observaciones: state, state_po , state_fips, state_cen y state_ic ofrecen la misma información en todas las tablas, están relacionadas entre si. Office siempre es US SENATE en este archivo. Mode es siempre total.

Total de registros: 3629

3) Archivo "1976-2022-house.tab":

Este archivo es al congreso de EEUU (Cámara de Representantes) de 1976 hasta 2022. Igual que los anteriores, aquí se encuentran los candidatos, las informaciones respecto a ellos y resultados.

Estructura del archivo:

Nombre del campo	***Descripción***	***Tipo***	***Ejemplo***
year	*Año de la elección*	*Entero*	*1976*
state	*Nombre del estado*	*Texto*	*Arizona*
state_po	*Código de dos letras del estado*	*Texto*	*AL*
state_fips	*Código FIPS del estado*	*Entero*	*1*
state_cen	*Código CEN del estado*	*Entero*	*63*
state_ic	*Código IC del estado*	*Entero*	*41*
office	*Cargo en elección*	*Texto*	*US HOUSE*
district	*Distrito electoral*	*Texto*	*001*
stage	*Etapa elección*	*Texto*	*GEN*
runoff	*Segunda vuelta*	*Boleano*	*FALSE*
special	*Se trata de una elección extraordinaria o no*	*Boleano*	*FALSE*
candidate	*Nombre del candidato*	*Texto*	*BILL DAVENPORT*
party	*Partido del candidato*	*Texto*	*DEMOCRAT*
writein	*Está ya en las papeletas de elección o hace falta escribirlo.*	*Boleano*	*FALSE*

mode	*Modo de votación*	*Texto*	*TOTAL*
candidatevotes	*Votos obtenidos por el candidato*	*Entero*	*58906*
totalvotes	*Total de votos del estado*	*Entero*	*157170*
unofficial	*Son los resultados oficiales*	*Boleano*	*False*
version	*Version de los datos*	*Entero*	*20230706*
fusion_ticket	*Representa a varios partidos políticos*	*Boleano*	*False*

Registros: 32452

Observaciones: state, state_po , state_fips, state_cen y state_ic ofrecen la misma información en todas las tablas, están relacionadas entre si. Office siempre es US HOUSE en este archivo. Mode es siempre total.

4) Archivo "state_offices.txt":

Es un archivo de texto plano, de tipo clave-valor, para relacionar a los estados con su abreviación.

Estructura del archivo:

Nombre del campo	***Descripción***	***Tipo***	***Ejemplo***
State	*Nombre del estado*	*Texto*	*Alabama*
State_po	*Abreviación de ese mismo estado*	*Texto*	*AL*

Número de registros: 49

5) Archivo "SP_500.csv":

Este archivo contiene datos del índice S&P 500 desde 1975 hasta 2022. Proporciona información en vela (intervalo) semanal sobre los precios, volumen y los cambio (%).

Estructura del archivo:

Nombre del campo	***Descripción***	***Tipo***	***Ejemplo***
Date	*Fecha del registro semanal*	*Texto/Fecha*	*12/25/2022*
Price	*Precio del índice (de cierre) de esa semana*	*Float*	*3,839,50*
Open	*Precio de apertura del índice de esa semana*	*Float*	*3,845,30*
High	*Precio mas alto habido durante esa semana*	*Float*	*3,780.20*
Low	*Valor más bajo del S&P 500 durante esa semana*	*Float*	*3,780.20*
Change %	*Cambio porcentual respecto al cierre de la semana anterior*	*Float*	*0,14*
Electoral_year	*Es año electoral en EEUU*	*Boleano (YES/NO)*	*NO*

Total de registros:2504

Estimación de volumetría:

Fichero	Registros	Valores	Datos
1976-2020-president.tab	*4287*	*15*	*64305*
1976-2020-senate.tab	*3629*	*19*	*68951*
1976-2022-house.tab	*32452*	*20*	*649040*
State_office.txt	*49*	*2*	*98*
SP_500.csv	*2504*	*7*	*17528*
Total			*799922*

3.Análisis funcional

Requisito	Prioridad	Exigible/Deseable
Se extraerá de forma adecuada la información de las fuentes de datos	*1*	*E*
Se creará un almacén de datos	*1*	*E*
Se realizará un análisis descriptivo de los datos para comprender su estructura y contenido.	*1*	*E*
Se identificarán y manejarán valores nulos, faltantes o incoherentes	*2*	*E*
Se visualizarán las tendencias temporales y patrones identificados en datos	*2*	*E*
Se calcularán estadísticas básicas	*2*	*E*
Se compararán resultados electorales entre diferentes años, lugares y tipos de elecciones.	*2*	*E*
Gráficos variados para mostrar los datos.	*2*	*E*
Se identificarán relaciones o correlaciones entre diferentes datos	*3*	*D*
Se integrará todo en un único almacén de datos	*3*	*D*
Análisis entre los resultados y movimientos del S&P 500	*3*	*D*

Arquitectura funcional propuesta: Arquitectura de Data Warehousing de tipo OLAP, la razón principal es su capacidad de tratar gran cantidad de datos de forma multidimensional.

Elementos de la arquitectura:

- ***Fuentes de datos: Los archivos 1976-2020-president.tab, 1976-2020-senate.tab, 1976-2022-house.tab state_office.txt y SP_500.csv serán las fuentes de datos principales.***
- ***ETL: Se realizará la extracción de datos de las fuentes, seguida de su transformación para limpiar y procesar los datos si fuera necesario, aunque una exploración preliminar de los datos no he visto muchos a hacer transformación y finalmente se cargarán en el almacén de datos.***

- ***Almacén de datos: Se creará un almacén de datos donde se consolidarán todos los datos relevantes para su análisis posterior. Este almacén contendrá tablas dimensionales y de hechos para facilitar consultas y análisis.***
- ***Herramientas de análisis y visualización de datos: Se utilizará la herramienta que designe el profesor en próximas etapas del proyecto.***
- ***Análisis de datos: Será importante que la herramienta para analizar los datos sea potente y con capacidad para analizar y visualizar los datos de diferentes puntos de vista.***

Otros requisitos funcionales:

Se garantizará la integridad y la seguridad de los datos en todo momento.

Se documentará adecuadamente el proceso de extracción, transformación y carga de datos.

Se asegurará la escalabilidad y la capacidad de expansión del sistema para futuras incorporaciones de datos o mejoras en el análisis junto con la adicción de otras fuentes de datos.

En nuestro caso, al no tratar datos personales privados, no se procesará la información en ese sentido para quitar datos no relevantes.

4.Diseño del modelo conceptual, lógico y fisico del almacén de datos

4.1. Diseño conceptual.

Lo primero de todo es saber los hechos que vamos a incluir, en nuestro caso el hecho es una elección, y luego las dimensiones serían los estados, el tipo de elecciones y el momento de la elección.

Tabla de Hechos: FACT_Elections_USA

Métricas: Votos obtenidos por candidato.

Dimensiones:

DIM_Time: Año de la elección.

DIM_State: Información del estado americano donde se produce la votación.

DIM_Election_Type: Tipo de elección (Presidencial, Senado, Cámara de representantes).

DIM_Candidate: Detalles de los candidatos a cada elección.

4.1 Diseño Conceptual Ampliado teniendo en cuenta la evolución del S&P500.

Dimensiones Adicionales:

DIM_Date: Fechas específicas que corresponden a los registros del S&P 500.

Tabla de Hechos Adicional:

FACT_SP500:

Métricas: Precio de cierre semanal, cambio porcentual semanal.

4.2 Diseño Lógico

En este modelo, se especifican las claves y las relaciones entre las tablas de hechos y dimensiones:

FACT_Elections_USA: Contiene claves foráneas para cada dimensión y la métrica de votos.

DIM_Time: time_id (PK), year.

DIM_State: state_id (PK), name, state_po.

DIM_Election_Type: type_id (PK), description.

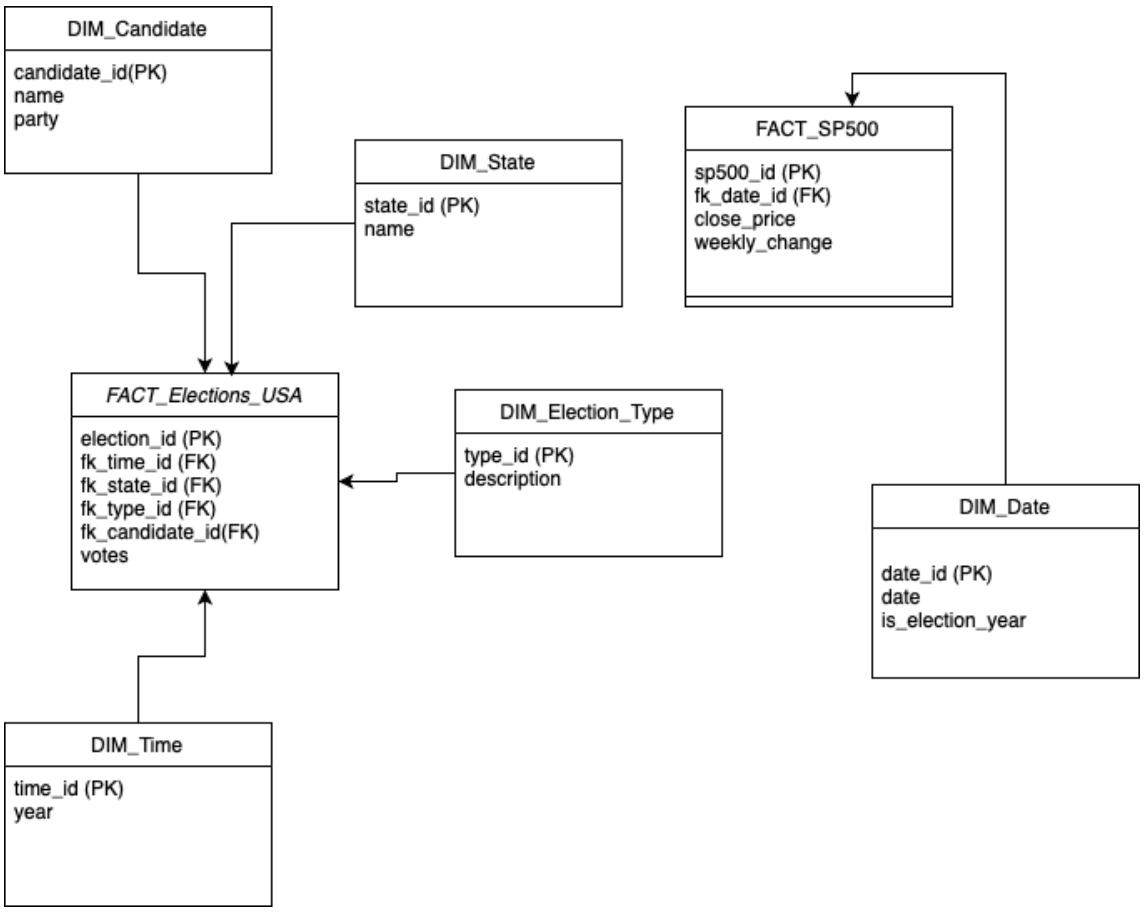
DIM_Candidate: candidate_id (PK), name, party.

4.2 Diseño Lógico Ampliado teniendo en cuenta la evolución del S&P500.

Se introduce una nueva tabla de hechos junto con una dimensión de tiempo detallada que quizás pueda ser la misma que la usada con FACT Elections, ya que los campos son parecidos, pero yo finalmente me decido por ponerlo por separado.

DIM_Date: date_id (PK), date, is_election_year.

FACT_SP500: sp500_id (PK), fk_date_id (FK), close_price, weekly_change.



4.3 Diseño Físico.

Esta parte es la implementación de un SGBD, hacer las estructuras correctas y las relaciones adecuadas, es decir, muy importante que cada clave foránea este referenciada a cada clave primaria, y que los datos estén correctamente definidos.

Cada clave primaria de las dimensiones debe de ser única y bien referenciada a la clave que le corresponda dentro de la tabla de hechos.

Hechos FACT_Elections_USA:

Campo	Tipo de Dato	Descripción	Ejemplo
election_id	INTEGER	Clave primaria	2346
fk_time_id	INTEGER	Clave foránea que enlaza con DIM_Time	1
fk_state_id	INTEGER	Clave foránea que enlaza con DIM_State	10
fk_type_id	INTEGER	Clave foránea que enlaza con DIM_Election_Type	2
fk_candidate_id	INTEGER	Clave foránea que enlaza con DIM_Candidate	101
votes	BIGINT	Número total de votos	8767543

Dimensiones:

DIM_Time:

Campo	*Tipo de Dato*	*Descripción*	*Ejemplo*
time_id	*INTEGER*	*Clave primaria*	*1*
year	*INTEGER*	*Año de la elección*	*2020*

DIM_State:

Campo	Tipo de Dato	Descripción	Ejemplo
state_id	INTEGER	Clave primaria	10
name	VARCHAR(100)	Nombre completo del estado	Arizona

DIM_Election_Type:

Campo	Tipo de Dato	Descripción	Ejemplo
type_id	INTEGER	Clave primaria	2
description	VARCHAR(255)	Descripción del tipo de elección	Presidencial

DIM_Candidate:

Campo	Tipo de Dato	Descripción	Ejemplo
candidate_id	INTEGER	Clave primaria	101
name	VARCHAR(255)	Nombre del candidato	Joe Biden
party	VARCHAR(100)	Partido político del candidato	Democratic

4.3 Diseño Físico Ampliado teniendo en cuenta la evolución del S&P500.

Hechos FACT_SP500

Campo	Tipo de Dato	Descripción	Ejemplo
sp500_id	INTEGER	Clave primaria	9900
fk_date_id	INTEGER	Clave foránea que enlaza con DIM_Date	5001
close_price	DECIMAL(10,2)	Precio de cierre del índice S&P 500	3500.23
weekly_change	DECIMAL(5,2)	Cambio porcentual desde la semana anterior	-0.50

DIM_Date (para hechos FACT_SP500):

Campo	Tipo de Dato	Descripción	Ejemplo
date_id	INTEGER	Clave primaria	5001
date	DATE	Fecha específica	2020-11-03
is_election_year	BOOLEAN	Indica si es un año electoral	TRUE