Universidad de El Salvador
Facultad de Ingeniería y Arquitectura
Escuela de Ingeniería de Sistemas Informáticos
Curso de Especialización Ingeniería de Datos
Ciclo I-2024

EXAMEN PARCIAL 3

Catedrático: René Fabricio Quintanilla Gómez

Estudiante: Luis Javier Díaz Menéndez

Carnet: DM15019

Ciudad Universitaria, 25 de octubre del 2024

Enunciado del Problema

La empresa aeronautica “Para Volar” tiene en operación el sistema de vuelos “Gamatron”, el cual se encarga de operativizar los vuelos comerciales realizados por la empresa hacia diferentes aeropuertos en el mundo a través de reservaciones de vuelos y de abordajes por vuelos realizados.

Después de varios años operando, la compañía ha decidido contratarlo para diseñar su estrategia de Big Data, para lo cual le comparte la información que se descarga de sus sistemas donde se muestran datos sobre estados de vuelos y se busca que esta sea cargada en un repositorio donde se pueda exponer dicha información de forma ágil.

Esta información se puede descargar de este enlace: https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022

Para comenzar a implementar su estrategia de Big Data, el negocio requiere hacer una proyección de vuelos cancelados, retrasados y diferidos, así como las causas del posible incidente (clima, llantas, seguridad del viaje, llegadas tardías, etc).

Para solventar este requerimiento, la empresa le solicita lo siguiente:

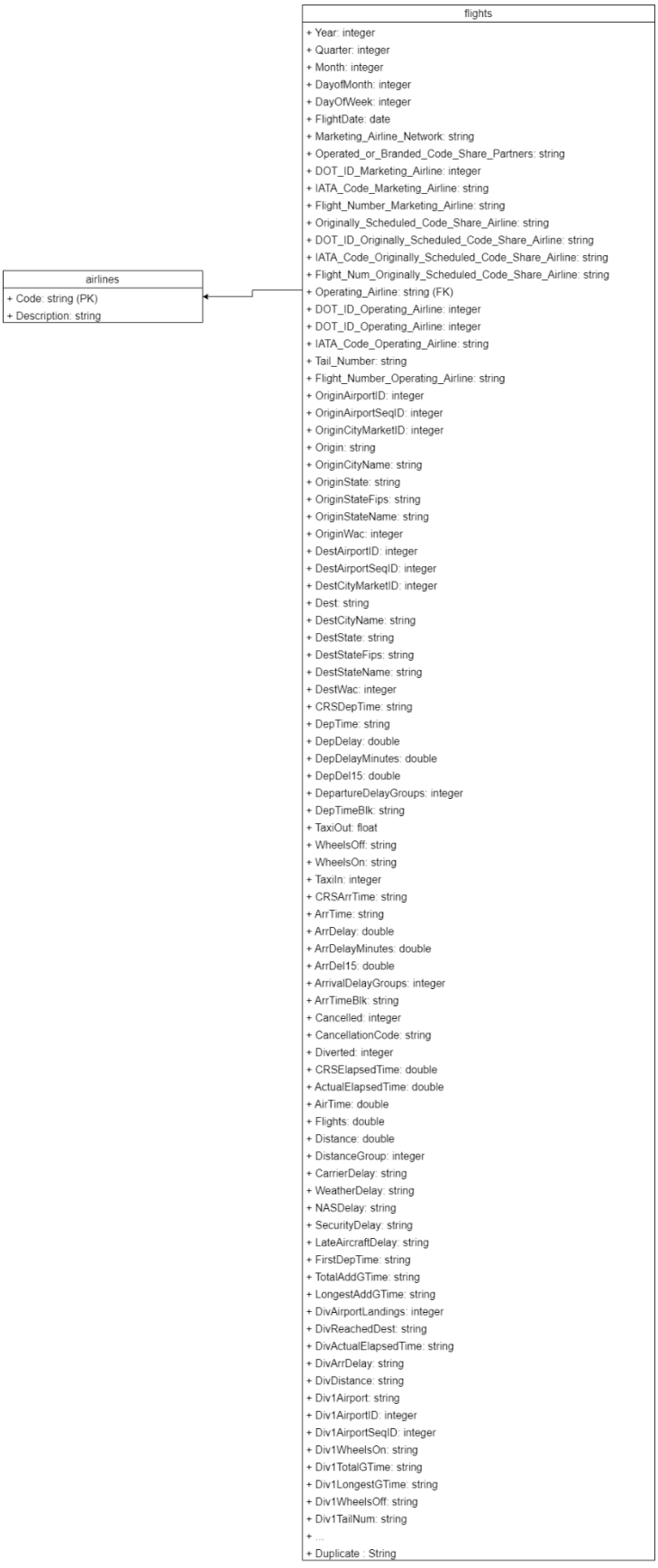
1. (25%) Diseñar una base de datos para almacenar el datawarehouse con el que se analizará esta información

2. (20%) Cree una estructura de carpetas para procesar la información

3. (30%) Cree una base de datos en SQL Server siguiendo el diseño de la base de datos elaborada en el punto 1.

4. (25%) Construya procesos ETL para la carga de información en la carpeta designada y su procesamiento para prepararlo para su carga en SQL Server.

Diagrama del DataSet


airlines
+ Code: string (PK)
+ Description: string
flights
+ Year: integer
+ Quarter: integer
+ Month: integer
+ DayofMonth: integer
+ DayOfWeek: integer
+ FlightDate: date
+ Marketing_Airline_Network: string
+ Operated_or_Branded_Code_Share_Partners: string
+ DOT_ID_Marketing_Airline: integer
+ IATA_Code_Marketing_Airline: string
+ Flight_Number_Marketing_Airline: string
+ Originally_Scheduled_Code_Share_Airline: string
+ DOT_ID_Originally_Scheduled_Code_Share_Airline: string
+ IATA_Code_Originally_Scheduled_Code_Share_Airline: string
+ Flight_Num_Originally_Scheduled_Code_Share_Airline: string
+ Operating_Airline: string (FK)
+ DOT_ID_Operating_Airline: integer
+ DOT_ID_Operating_Airline: integer
+ IATA_Code_Operating_Airline: string
+ Tail_Number: string
+ Flight_Number_Operating_Airline: string
+ OriginAirportID: integer
+ OriginAirportSeqID: integer
+ OriginCityMarketID: integer
+ Origin: string
+ OriginCityName: string
+ OriginState: string
+ OriginStateFips: string
+ OriginStateName: string
+ OriginWac: integer
+ DestAirportID: integer
+ DestAirportSeqID: integer
+ DestCityMarketID: integer
+ Dest: string
+ DestCityName: string
+ DestState: string
+ DestStateFips: string
+ DestStateName: string
+ DestWac: integer
+ CRSDepTime: string
+ DepTime: string
+ DepDelay: double
+ DepDelayMinutes: double
+ DepDel15: double
+ DepartureDelayGroups: integer
+ DepTimeBlk: string
+ TaxiOut: float
+ WheelsOff: string
+ WheelsOn: string
+ TaxiIn: integer
+ CRSArrTime: string
+ ArrTime: string
+ ArrDelay: double
+ ArrDelayMinutes: double
+ ArrDel15: double
+ ArrivalDelayGroups: integer
+ ArrTimeBlk: string
+ Cancelled: integer
+ CancellationCode: string
+ Diverted: integer
+ CRSElapsedTime: double
+ ActualElapsedTime: double
+ AirTime: double
+ Flights: double
+ Distance: double
+ DistanceGroup: integer
+ CarrierDelay: string
+ WeatherDelay: string
+ NASDelay: string
+ SecurityDelay: string
+ LateAircraftDelay: string
+ FirstDepTime: string
+ TotalAddGTime: string
+ LongestAddGTime: string
+ DivAirportLandings: integer
+ DivReachedDest: string
+ DivActualElapsedTime: string
+ DivArrDelay: string
+ DivDistance: string
+ Div1Airport: string
+ Div1AirportID: integer
+ Div1AirportSeqID: integer
+ Div1WheelsOn: string
+ Div1TotalGTime: string
+ Div1LongestGTime: string
+ Div1WheelsOff: string
+ Div1TailNum: string
+ ...
+ Duplicate : String

Modelo dimensional propuesto

Pasos para la creación del modelo dimensional

1. Proceso: Análisis de vuelos.
2. Granularidad: una fila de datos representa un vuelo la empresa, que representa una aerolínea o fecha.
3. Dimensiones:
   a) DimAirline
      - airlineKey
      - airlineCode
      - airlineDescription

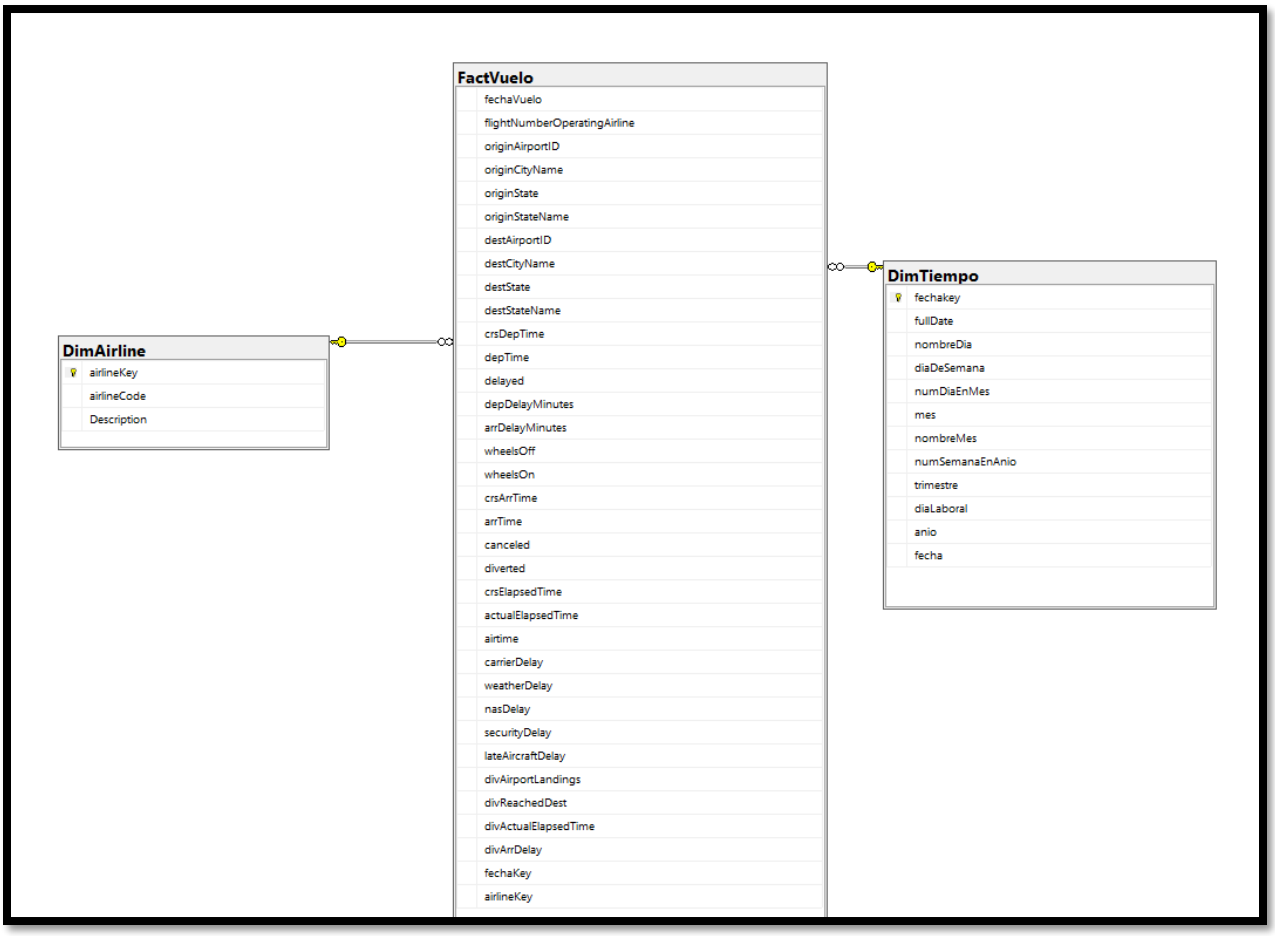
   b) DimFecha
      - fechaKey
      - fullDate
      - nombreDia
      - diaDeSemana
      - numDiaEnMes
      - numDiaAnio
      - Mes
      - nombreMes
      - numSemanaEnMes
      - numSemanaEnAnio
      - trimestre
      - nombreTrimestre
      - diaLaboral
      - año

   c) FactVuelo:
      - fechaKey
      - airlineKey
      - originAirportID
      - originCityName
      - originState
      - originStateName
      - destAirportID
      - destCityName
      - destState
      - destStateName

- flightNumberOperatingAirline
- crsDepTime
- depTime
- delayed
- depDelayMinutes
- arrDelayMinutes
- wheelsOff
- wheelsOn
- crsArrTime
- arrTime
- cancelled
- diverted
- crsElapsedTime
- actualElapsedTime
- airtime
- carrierDelay
- weatherDelay
- nasDelay
- securityDelay
- lateAircraftDelay
- divAirportLandings
- divReachedDest
- divActualElapsedTime
- divArrDelay

4. Métricas:
   - Vuelos exitosos: vuelos realizados sin incidencias.
   - Vuelos cancelados: vuelos que fueron cancelados.
   - Vuelos atrasados: vuelos que se atrasaron de su rutina programada
   - Vuelos diferidos: vuelos que tuvieron que ser desviados (diferidos) de su ruta establecida.

Diagrama Dimensional del DataWarehouse



Enlace

Repositorio de GitHub donde se almacenan los archivos de la solución:

https://github.com/Javier767/Examen_Parcial3_UES.git