Inlämningsuppgift: Skapa SSIS paket från ax till limpa

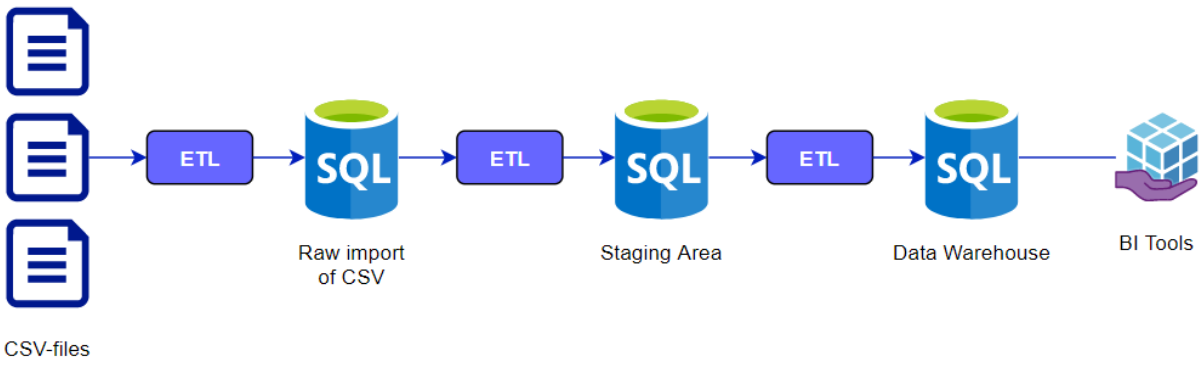
Bakgrund

U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics kontrollerar alla stora flygbolag i USA för att se om de är i tid eller inte. Ni har genom det fått tillgång till summerad data som innehåller kommersiella inrikesflygningar i USA under år 2015; totalt cirka 6 miljoner rader flight-data (eller strax under 1 GB data).

Hämta filen 2015FlightDelayAndCancellations.zip, och packa upp den. När den är uppackad ska det finnas tre filer: airlines.csv, airports.zip, samt flights.zip. Tänk på att ni alltså behöver utrymme att lagra csv-filerna både när de hämtas och packas upp, men då de även ska tankas in i databasen så måste det även finnas utrymme för databas och dess mdf- och ldf-filer.

Du har fått i uppdrag att utföra migrationsprocessen från rådata till strukturerad information för ett företag som vill etablera sig på marknaden för kommersiellt flyg i USA. Det som i slutändan är mest intressant är att se förseningar i olika former.

Uppdragsgivaren har gett er följande schema som illustration vad som ska ingå:



CSV-files och Raw import of CSV: Data som innehåller kommersiella inrikesflygningar i USA under 2015, kopierade från CSV in i SQL.

Staging Area: Innehåller den tvättade data

Data Warehouse: Innehåller datat uppdelat i ett StarSchema

BI Tools: OLAP Cubes i form av en Tabular Models (SSAS) för snabb förfrågning från analysverktyg såsom PowerBi, Excel, eller liknande.

Specifikation CSV-filer + Raw import of CSV

Du bör lära känna ditt data innan du börjar skapa SSIS-paket.

Skapa en databas som heter ”SourceSystems” eller liknande. Ladda in data från CSV och inspektera i SQL Server Management Studio för att få en uppfattning om vilka kolumner som är mest besvärliga, d v s har många NULLs, blanks, eller andra värden som inte passar in. Detta utgör en bra start innan du börjar med ETL processen i SSIS. Man vill alltid få en första anblick på data innan man börjar arbeta med det så att man inte arbetar i blindo. Om du inte vill ladda in filerna manuellt via SSMS kan du alltid skapa ett SSIS paket som laddar in filer med exempelvis Bulk Insert Task i Control Flow.

När du väl lärt känna datat som ska hanteras, ska ett SSIS-paket skapas. I paketet ska allt laddas in från CSV till SQL automatiskt i ett enda flöde.

Specifikation Staging Area

Skapa en databas som heter ”StagingArea” eller liknande. Här ska samtliga data från Source Systems laddas efter en enklare tvättning av data. Även transformationer sker här såsom datatypskonverteringar.

Även denna del ska skötas med automatik via SSIS-paket. Skicka in det tvättade datat till Staging Area.

Airlines:

Ha i bakhuvudet att detta ska vara grunden för en dimension senare.

Se till att du hanterar eventuella tomma fält (dvs ””) och NULL-värden på ett korrekt sätt.

Airports:

Ha i bakhuvudet att detta ska vara grunden för en dimension senare.

Se till att du hanterar eventuella tomma fält (dvs ””) och NULL-värden på ett korrekt sätt.

Flights:

Ha i bakhuvudet att detta ska vara grunden till en faktatabell senare.

Se till att du hanterar eventuella tomma fält (dvs ””) och NULL-värden på ett korrekt sätt.

Vissa klockslag är troligen av INT-datatyp men bör egentligen vara av annan datatyp (beror på hur du importerat datat). INT lagrar heltal och alla nollor som bör finnas för exempelvis 0005 (00:05) kommer att visas som 5. Gör en transformation till string och lägg på rätt antal 0-or.

Sätta ihop Year, Month, och Day till formatet YYYYMMDD och lämplig datatyp (så du senare kan hantera datum-dimensionen).

Routes:

Denna finns inte med i csv-filer, utan denna ska vi skapa själva. Ha i bakhuvudet att detta ska vara grund för en dimension senare.

För att skapa data för rutterna så ska du utgå från Flights som vi städat lite.

Ladda bara in ORIGIN_AIRPORT och DESTINATION_AIRPORT kolumnerna in i en Routes-tabell, och ta bara med unika rader, d v s använd den här SQL satsen i Source:

```
SELECT DISTINCT ORIGIN_AIRPORT, DESTINATION_AIRPORT FROM [Flights]
```

Skapa därefter en unik kolumn och helst även ett index för Routes tabellen.

```
ALTER TABLE Routes ADD RouteID INT identity(1,1)
```

Det kan finnas en del flighter, flygplatser etc som saknar data... Om det saknas något, exempelvis om en flygplats finns angiven i Flights men denna finns inte med i Airports, så ska vi göra en ”John Doe” av det hela; vi måste alltså tar reda på vilka flygplatser som saknas i Airports och lägger in en dafault-hänvisning så vi får vår struktur/vårt star schema/vår join att fungera.

Specifikation Data Warehouse

Vi ska i denna del se till att data kopieras in till ett Star Schema. Döp gärna denna Data Warehouse-databas till DW.

- Faktatabellen i Star Schemat ska vara baserad på Flights.
- Det ska finnas en dimension för Datum
- Det ska finnas en dimension för Routes
- Det ska finnas en dimension för Airports

Flights + Routes:

Men innan vi kopierar allt, så måste vi kontrollera/uppdatera så Flights och Routes stämmer överens mot varandra.

Använd en Lookup transformation för att koppla kolumnerna ORIGIN_AIRPORT och DESTINATION_AIRPORT i Flightdata-tabellen Destination med motsvarande i Routes-tabellen. Hämta RouteID från Routes-tabellen.

Ge Flights ett lämpligt namn i DW databasen och ladda in dess data dit.

Date:

Skapa en tabell för DimDate / datum-dimension, och fyll den med data för aktuella årtal. Se till att ha kolumner för svenska namn för veckodagar och månader.

Routes (och Airports):

Slå ihop tabellerna Routes och Airports till en platt struktur, så vi får en lämplig dimension i ett Star Schema i samband med att data kopieras till DW-databasen.

Airlines:

Skapa en dimension för Airlines, och kopiera över dess data till DW

För VG ska ni i ovanstående steg förklara och kortfattat skriva en kort förklaring av stegen du tog för att slutföra uppgiften, inklusive eventuella utmaningar du stötte på och hur du överkom dem. Jag vill gärna att ni argumenterar för era val, såsom era val av datatyper; synpunkter och åtgärder kring prestanda etc. Jag är INTE på jakt efter en lång roman, utan korta kommentarer som typ annotations duger utmärkt.

Specifikation SSAS

Vi ska skapa en kub i SSAS som vi ska använda våra tabeller till.

När det gäller Measures så ska ni åtminstone kunna ta fram största förseningen och försening som medelvärde.

ETL-flödet behöver inte vara inkrementellt, utan det räcker med att allt tankas in en gång och att kuben genereras.

Beskrivning av data och dess olika fält

airlines.csv

Column	Explanation
IATA_CODE	Airline Identifier
AIRLINE	Airport's Name

airports.csv

Column	Explanation
IATA_CODE	Location Identifier
AIRPORT	Airport's Name
CITY	Atlanta
STATE	Georgia
COUNTRY	Country Name of the Airport
LATITUDE	Latitude of the Airport
LONGITUDE	Longitude of the Airport

flights.csv

Column	Explanation
YEAR	Year of the Flight Trip
MONTH	Month of the Flight Trip
DAY	Day of the Flight Trip
DAY_OF_WEEK	Day of week of the Flight Trip
AIRLINE	Airline Identifier
FLIGHT_NUMBER	Flight Identifier
TAIL_NUMBER	Aircraft Identifier
ORIGIN_AIRPORT	Starting Airport
DESTINATION_AIRPORT	Destination Airport
SCHEDULED_DEPARTURE	Planned Departure Time
DEPARTURE_TIME	WHEEL_OFF - TAXI_OUT
DEPARTURE_DELAY	Total Delay on Departure
TAXI_OUT	The time duration elapsed between departure from the origin airport gate and wheels off
WHEELS_OFF	The time point that the aircraft's wheels leave the ground
SCHEDULED_TIME	Planned time amount needed for the flight trip
ELAPSED_TIME	AIR_TIME+TAXI_IN+TAXI_OUT
AIR_TIME	The time duration between wheels_off and wheels_on time
DISTANCE	Distance between two airports
WHEELS_ON	The time point that the aircraft's wheels touch on the ground
TAXI_IN	The time duration elapsed between wheels-on and gate arrival at the destination airport
SCHEDULED_ARRIVAL	Planned arrival time
ARRIVAL_TIME	WHEELS_ON+TAXI_IN
ARRIVAL_DELAY	ARRIVAL_TIME-SCHEDULED_ARRIVAL
DIVERTED	Aircraft landed on airport that out of schedule
CANCELLED	Flight Cancelled (1 = cancelled)
CANCELLATION_REASON	Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
AIR_SYSTEM_DELAY	Delay caused by air system
SECURITY_DELAY	Delay caused by security
AIRLINE_DELAY	Delay caused by the airline
LATE_AIRCRAFT_DELAY	Delay caused by aircraft
WEATHER_DELAY	Delay caused by weather