Market Research - 'flights case'

De Case

Als free-lancer ben je benaderd door een luchtwaartmaatschapij die zich mogelijk wil gaan richten op de Amerikaanse markt. Jouw opdrachtgever heeft een aantal vragen waarbij hij jouw hulp vraagt.

> Voor het juist beantwoorden van de vragen heb je kennis nodig van: SELECT, WHERE, COUNT(), AVG(), MIN(), MAX(), SUM() DISTINCT, FROM, INNER JOIN, GROUP BY en HAVING.
>
> Kijk eventueel op w3schools hoe het ook weer precies zat.

Vragen

De vragen hebben betrekking op de Amerikaanse vliegmarkt:

1. Hoeveel vluchten vliegen er op het vliegveld *Philadelphia International Airport* *(tip: (1)* met *vliegen op* wordt de bestemming bedoeld; (2) de volledige naam van het vliegveld moet onderdeel van de query zijn).
2. Hoeveel vluchten zijn er in totaal gecancelled?
3. Hoevel vluchten zijn er in totaal omgeleid (diverted)?
4. Hoe groot was de grootste vertraging bij vertrek (departure delay)?
5. Soms wordt vertrekt een vlucht te vroeg en wordt er een negatieve departure delay vastgelegd. Voor het berekenen van de gemiddelde vertraging mogen deze vluchten niet meetellen. Wat is de gemiddelde vertraging (dus negatieve departure delay niet meegerekend).
6. Hoeveel vluchten zijn er die binnen de staat Calefornia vliegen, dus vertrekken en aankomen in Calefornia?
7. Hoeveel airtime (dus tijd in de lucht) heeft de kortste vlucht (in tijd)?
8. Wat is de kortste afstand van een vlucht en hoe lang duurde deze vlucht?
9. Wat is de airtime van de vlucht die de langste afstand heeft gevlogen?
10. Wat is de gemiddelde '*air time*' (dus tijd in de lucht) van alle vluchten die binnen de staat Calefornia vliegen?

11. Hoeveel vluchten zijn er die binnen een staat in Amerika blijven, dus vertrekken en aankomen indezelfde staat?
12. In welke staat is de gemidelde airtime van de vluchten die binnen een staat in Amerika blijven het laagst?
13. Op (naar) welke luchthavens wordt het meest gevlogen?
14. Vanaf welke luchthavens wordt het meest gevlogen?
15. Waarom kun je niet goed bepalen wat de drukste maand is? Maak de query en leg uit waarom je deze query niet goed kan testen.
16. Welk vliegveld ligt het meest zuidelijk?
    (tip: zoek eens op wat er in de kolom *latitude* staat, waar staat *latitide* voor?)
17. Welke luchtvaartmaatschappij legt de meeste vluchten af?
18. Welke luchtvaartmaatschappij legt legt de meeste airmiles af?
19. ~~Welke luchtvaartmaatschapij vliegt op de meeste bestemmingen?~~
    Ik weet niet of deze query mogelijk is zonder gebruik te maken van een temp table.
20. Maak een query die alle luchtvaarmaatschapijen laat zien die meer dan 350 bestemmingen hebben.
21. Tussen welke twee vliegvelden vindt de meeste vertraging bij vertrek plaats.?
22. Op welke vluchten (van-naar) is het vliegtuig gemiddeld het langs aan het taxiën?

Jouw opdrachtgever heeft alle vluchtinformatie van alle vluchten uit de USA uit 2015. Alle bovenstaande vragen kunnenn dus afgeleid worden uit de data van 2015.

Data

De data bestaat uit drie bestanden:

1. de vluchtgegevens
2. de luchthaven gegegevens
3. de airlines gegevens

Je krijgt een Excel sheet met drie tabjes waarin deze gegevens staan.

PoC

De vluchtgegevens zijn er dermate veel dat in de Excel sheet alleen de eerste 5000 vluchten zijn opgenomen. Het volledige bestand telt ruim 580 000 vluchten (en bijnan 600 MB). Als je PC krachtig genoeg is kun je het volledige bestand inlezen, anderes kun je volstaan met de eerste 5000 vluchten. Het gaat hierbij toch om een POC, Proof of Concept waarbij geldt dat als je queries

werken op een set van 5000 dan werken ze ook op een set van 580 000.

Het is in zijn algemeenheid aan te raden om eerst met een kleine set data te werken. Het importeren van data en het testen van queries gaat dan namelijk veel sneller. Als je eenmaal hebt bedacht hoe je de database kunt opbouwen en hoe de queries er uit moeten zien, dan kun je daarna opschalen naar de volledige data set.

Jij ontvangt een Excel sheet van de opdrachtgever met de drie bestanden in drie afzonderlijke tabjes.

Database

Eén en ander kun je wellicht in Excel uitvoeren, maar Excel zal geen 500 000+ regels kunnen verwerken, bovendien wil de klant mogelijk later nog meer vluchtgegvens (uit meer jaren) toevoegen en zal de data-set nog verder vergroten. We zullen de gegevens dus moeten importeren in een database.

De Uitdaging

Je hebt dus een aantal uitdagingen:

1. Hoe krijg ik de Excel data in een (mySQL) database?
2. Hoe controleer ik en weet ik zeker dat alle data goed in de database zit?
3. Hoe maak ik de queries die antwoord geven op de door de klant gestelde vragen?

Planning

Maar voordat je met de uitvoering begint, denk eerst eens na over *hoe* je dit gaat aanpakken en maak een urenschatting. In het (mbo) examen zul je een Programma van Eisen en projectplan moeten opstellen. Dat gat hier te ver, maar bedenk wel voor jezelf welke stapjes je gaat doen en hoeveel tijd die kosten. Schrijf dit of leg dit vast op de computer. Het opstellen van een plan mag best wat werk kosten, want als het goed is kun je hier later veel tijd mee besparen.

Succes!

Excel bestand staat in Teams.

Revision #28
Created 2019-09-13 20:15:10 UTC by Admin
Updated 2019-10-17 21:13:42 UTC by Max