# Situering

Catchem wil analyses trekken rond de logs die geregistreerd werden voor de Treasures**.** Er werd besloten een DWH op te zetten en als pilootproject een ster rond de **subject area TreasureFound** te bouwen. In onderstaande diagram zie je wat het analistenteam geschetst heeft. Bouw op basis van dit beperkte model de ster en vul deze met data uit de correcte brontabellen met behulp van de ETL-tool. De data in de ster beperkt zich tot data van ‘Found’ logs. Hou rekening met de regels van de kunst en zorg ervoor dat op basis van de inhoud van de tabellen propere en leesbare rapporten gemaakt kunnen worden.

Lees eerst de opgave in zijn geheel door zodat je een zicht hebt op wat er allemaal van je verwacht wordt.

Nadat je het datawarehouse correct hebt gevuld moet je analyse query’s opstellen voor een aantal analysevragen. Naast de query’s voor de hieronder reeds opgesomde analysevragen voorzie je zelf 4 bijkomende analysevragen met bijhorende query.

**Studenten die individueel werken: Analysevragen in verband met het weer moeten niet uitgewerkt worden.**

* Welke rol hebben datumparameters (dagen, weken, maanden, seizoen) op het aantal caches?
* Wat is de invloed van het type user op de duur van de treasurehunt? Doet een beginner er langer over?
* Vinden users de cache gemiddeld sneller in de regen?
* Zoeken beginnende users gemiddeld naar grotere caches?
* Worden er gemiddeld minder caches gezocht op moeilijker terrein als het regent?
* Worden er in weekends meer moeilijkere caches gedaan?

# 

# Detail van het feit en de dimensies

## De datum dimensie

GeoCache wil statistieken kunnen trekken op basis van seizoenen. Voorzie in de datumdimensie (hiervoor heb je al een basis) één veld dat aangeeft welk seizoen het is. We gaan er voor de eenvoud vanuit dat er enkel op het noordelijk halfrond aan GeoCaching gedaan wordt. Je kan zelf kiezen hoe je dit aanpakt (Spoon steps gebruiken om de seizoenen te berekenen, extra tabel aanmaken…)

## De user dimensie

Geocache wil kunnen rapporteren op een aantal belangrijke zaken die gekoppeld zijn aan de gebruikers. In de dimensie moet naast de Voornaam, Achternaam, Woonplaats en Land ook bijgehouden worden wat de ‘ExperienceLevel’ is. Volgende levels komen voor:

* Newbie: Nog geen ‘Found’ logs geplaatst
* Amature: Minder dan drie ‘Found’ logs geplaatst
* Regular: 3-10 ‘Found’ logs geplaatst
* Expert: Meer dan 10 found logs geplaatst

Daarnaast wordt ook bijgehouden of hij al dan niet een ‘Dedicator’ is. Dit is het geval als hij zelf beheerder is van minimum één Cache

Voor woonplaats, land, experiencelevel en dedicator wil Geocache bij analyses steeds de situatie op het moment van de logdatum kunnen raadplegen. Als een gebruiker op dag d experiencelevel:Regular had en op dag d+1 experiencelevel:Expert dan moet een CacheFoundEvent van d+1 gekoppeld worden aan de dimensieversie van de gebruiker met experiencelevel:Expert. Voor een CacheFoundEvent van dag d moet dat Regular zijn.

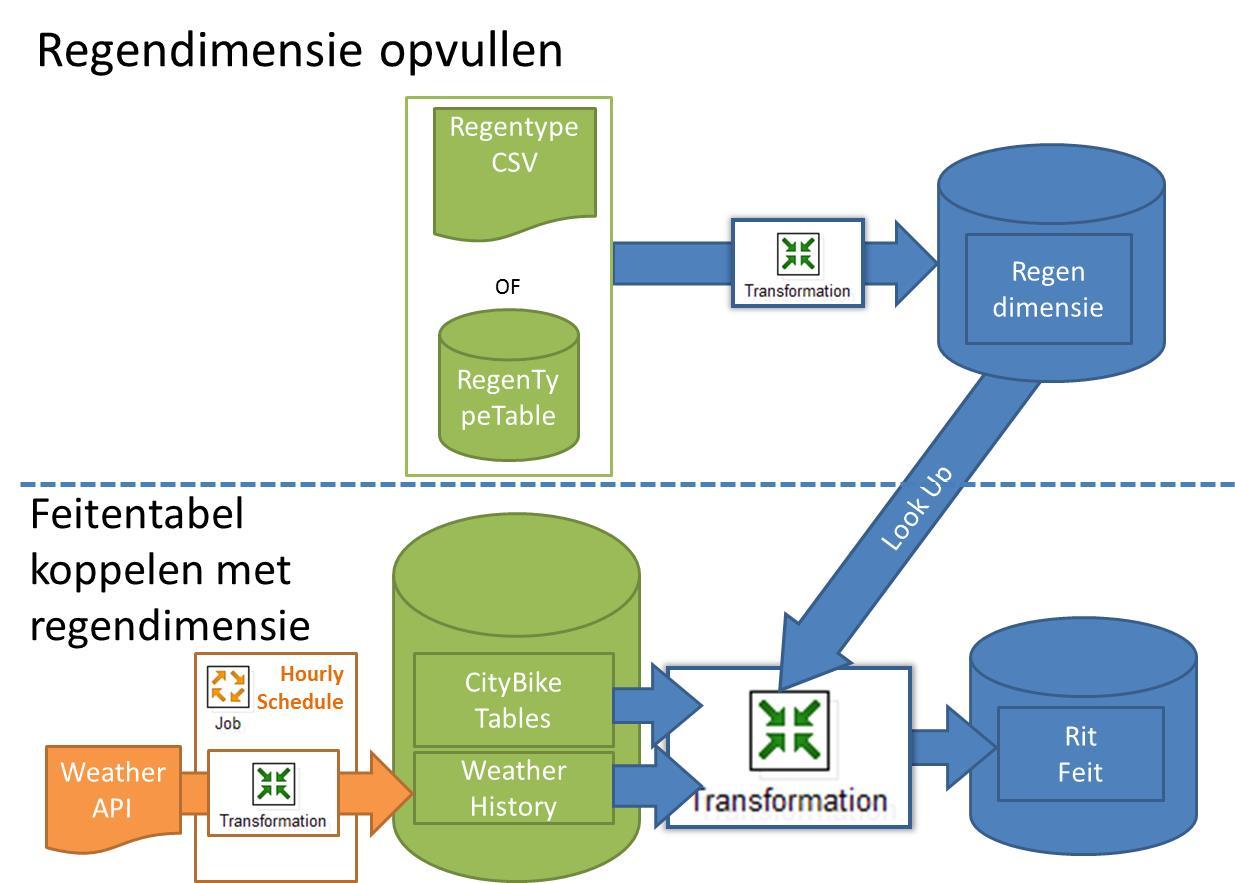
## De regendimensie

**Studenten die individueel werken: De weerdimensie en de historiektabel die hieraan gekoppeld is moet niet uitgewerkt worden.**

Ook het weer kan een belangrijke parameter zijn voor de analyse van gevonden treasures. Voorzie een regendimensie die de verschillende regentypes bevat. Je moet deze koppelen aan het feit op basis van de regensituatie op het moment van de logdatum. Maak om dit te verwezenlijken gebruik van een weer-api.

Let op: Het aantal rijen in de regendimensie is beperkt tot weertypes met regen in de weer-api (zie <http://openweathermap.org/weather-conditions>) + 1 rij voor alle weertypes zonder regen + 1 rij voor "Regensituatie onbekend". De bron van de regendimensie kan je zelf maken. Dit kan bijvoorbeeld via CSV of een tabel.

Informatie over hoe je de regensituatie op moet halen kan je vinden bij "Het feit TreasureFound".



## De dimensie TreasureType

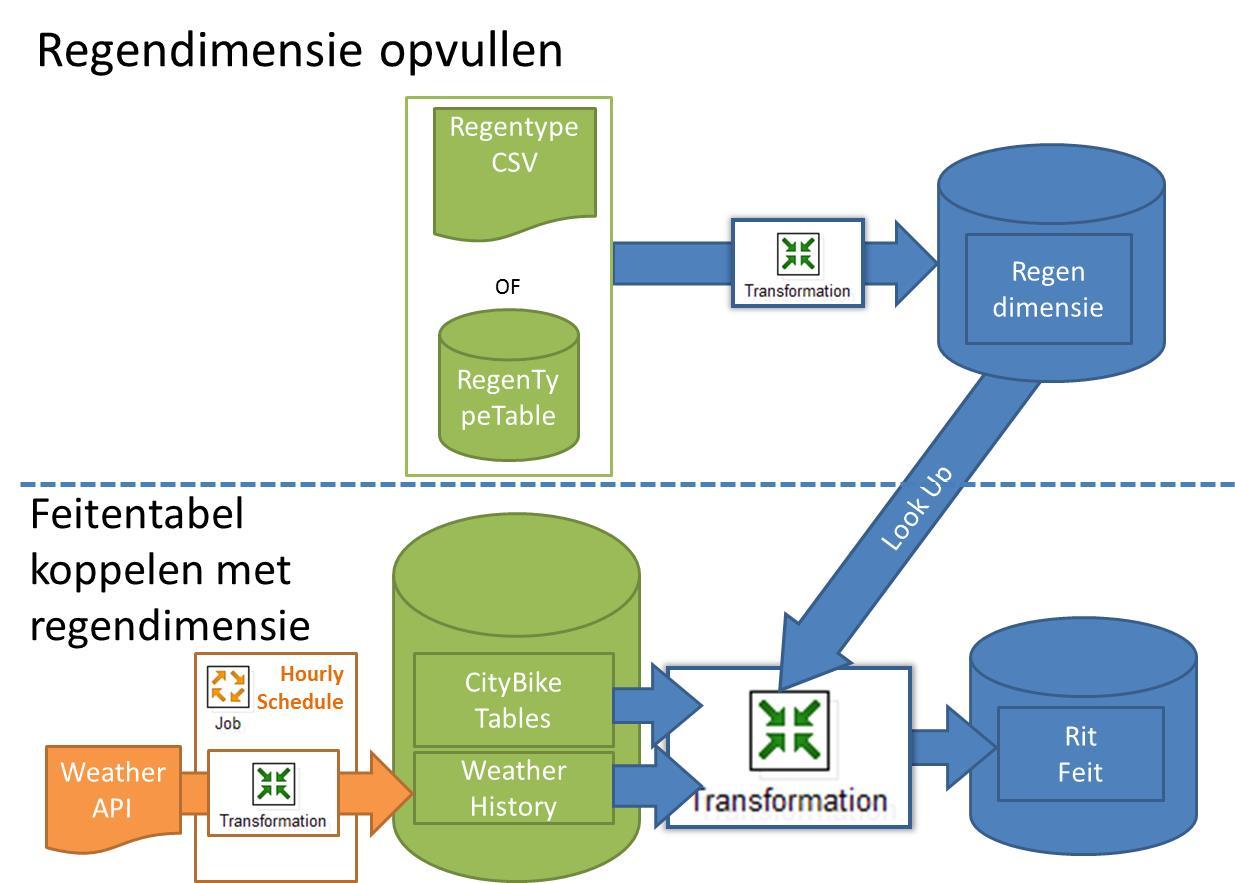
In de dimensie treasureType worden gegevens bijgehouden over de cache die gevonden is. Van een cache wordt volgende informatie in de dimensie opgeslagen: Difficulty, Terrain, Size (aantal stages binnen een treasure)

## Het feit TreasureFound

Centraal in de ster staat de feitentabel **TreasureFound**. Voor elke gevonden treasure wordt hier een record in opgeslagen. Je kan een treasure als gevonden beschouwen als er een log van het type ‘Found’ (= 2) voor werd geregistreerd. Zorg voor de koppelingen met de dimensies die hierboven beschreven staan, een standaard meetwaarde (met telkens een 1 als waarde), een meetwaarde voor de duur van de quest en zorg ook voor een veld met de creatiedatum van het record. Zorg voor de juiste koppeling met de type 2 dimensie ‘User’ en de andere dimensies

**Studenten die individueel werken: De weerdimensie en de historiektabel die hieraan gekoppeld is moet niet uitgewerkt worden.**

Let op: In het bronsysteem is geen weersinformatie aanwezig. Je hebt hiervoor dus een bijkomende bron nodig.   
Je zou historische weersinformatie kunnen opvragen via een weer-api, maar dat kost geld.   
Bouw daarom op basis van de huidige weersituatie zelf historiek op. Maak hiervoor een transformatie in DI die als hij wordt uitgevoerd per stad een rij toevoegt aan een historiektabel met daarin het uur de dag en de stad van de waarneming en de waarneming zelf. Je beperkt je tot 10 steden uit de database (de api's hebben nogal strikte limieten). De weergegevens kan je bijvoorbeeld van <http://openweathermap.org/current> halen.



Als je voor een bepaalde stad voor dat uur geen regeninformatie kan vinden in de historiektabel dan zet je voor die record het regentype op "Regensituatie onbekend".

Maak ook een job die de feitentabel leegmaakt en die vervolgens terug helemaal opvult.

*Tip: Als het uitvoeren van de job te lang duurt voeg dan een logische filtering toe in je query bij het inlezen uit het bronsysteem. Je kan die nadat alle transformaties getest zijn terug weghalen om een volledige run te doen. Lang wachten tijdens testruns is immers verloren tijd.*

## Optimalisatiedossier (Niet voor teams met 2 leden of individuele studenten)

Vul het optimalisatiedossier uit opdracht deel 2 aan met een hoofdstuk rond datawareoptimalisatie. Ga hiervoor even grondig te werk maar zoek naar optimalisaties die voor de structuur van de datawarehouse databank extra goed geschikt zijn.

## Extra dimensie (Niet voor teams met 2 leden)

Voorzie een extra dimensie in het datawarehouse. Zoek zelf naar een nuttige dimensie. Kies het gepaste SCD type en werk de dimensie mee uit in Pentaho DI. Zorg ook voor 3 bijkomende analysevragen die gebruik maken van deze dimensie.

# Evaluatiecriteria

* De werking van de ETL-transformaties en jobs in DI.
* De maten waarin aan de analysebehoeften uit de opdracht is voldaan.
* De volledigheid van de query’s die een antwoord bieden op de analysevragen.
* De maten waarin de richtlijnen uit de theorie gevolgd werden bij de uitwerking van de ster.