Database opereert in niche van tijdsafhankelijke data

RRDtool en de kunst van het vergeten

Rick van Rein

In een tijd waarin datasets genadeloos groeien is de stijl van RRDtool even verfrissend als nuttig. RRDtool slaat alle data op in een vaste hoeveelheid ruimte en gooit verouderd spul op een slimme manier weg.

RRDtool is een database voor een niche. Daar waar de meetgegevens je om de oren vliegen en je overzichten wilt hebben zoals maandelijkse gemiddelden of weekpatronen, is deze tool de perfecte manier om heel veel metingen in heel overzichtelijke diagrammen om te zetten.
RRDtool is ontstaan uit software die bijhoudt hoeveel netwerkverkeer door een router vliegt. Maar het is veralgemeniseerd en leent zich nu voor een heel scala aan toepassingen met reguliere metingen. Het lopende voorbeeld in dit artikel is de opbrengst van een paneel zonnecellen.

Periodiek meten

De basis-aanname onder RRDtool is dat je met een vast interval een meetwaarde oplevert. Voor elke meetperiode alloceert RRDtool een slot waarin de waarde wordt opgeslagen. Als je dus elke vijf minuten even kijkt hoeveel je zonnecellen hebben opgewekt, dan vul je elke vijf minuten zo'n slot met een meetwaarde. Dat bepaalt dan meteen in hoeveel detail je op een meting kunt inzoomen.

> Het leuke is dat RRDtool bij uitstek geschikt is om met falende systemen om te gaan

Al die meetgegevens stapelen zich op, en het is niet afdoende om dat te doen op de manier van een relationele database. Immers, als je elke vijf minuten een INSERT doet, dan heb je na 35 jaar (dat is een realistische levensduur voor zonnecellen) krap vier miljoen meetpunten. De gedachtesprong die RRDtool maakt is dat het detail van een meting per vijf minuten leuk is om terug te kijken over een beperkte periode, maar dat je gegevens van 25 jaar terug niet met datzelfde detail zult willen bekijken.
Van zulke oude gegevens ben je alleen in overzichtsgegevens geïnteresseerd. De uitwerking van deze gedachte is een cyclische structuur, waarin RRDtool per meetperiode een meetwaarde plaatst, daarbij een verouderde meting weggooiend. Bij de constructie van de database geef je aan wanneer je een meetwaarde verouderd vindt, dus je hebt controle op de tijd dat je terug kunt bladeren op elke losse meetwaarde.
Deze opmerkelijke constructie heeft een aangenaam gevolg, namelijk dat de database een vaste grootte heeft. Dit maakt hem perfect geschikt voor toepassing in gesloten omgevingen, eventueel zelfs zonder harde schijf. En doordat alle meetgegevens zijn geïndexeerd op tijd is het zelfs mogelijk om onmiddellijk naar het juiste meetgegeven te springen. Dit alles maakt RRDtool een aangenaam lichtgewicht stuk gereedschap.

Hoe te meten

Er zijn verschillende mogelijkheden om metingen te verrichten voor RRDtool. Het is bijvoorbeeld mogelijk om eens in de vijf minuten een momentopname op te vragen van de hoeveelheid vermogen die de zonnecellen op dat moment opwekken. Het kan zijn dat er net een wolk voor de zon zit of juist even niet, maar dat zou bij voldoende middelen wegvallen als een detailfout. Statistisch is dit verantwoord, maar het kan beter.
Waar dingen geteld worden, zoals het aantal auto's dat door een stoplicht rijdt, kan de teller telkens gereset worden bij het uitlezen, zodat steeds de laatste data worden opgeslagen. Het verdient dan wel aanbeveling om de reset op te nemen in dezelfde transactie die ook de gegevens invoert in RRDtool, en dus ontstaan kansen op locking-problemen. Het werkt, maar is technisch nog altijd niet het eenvoudigst haalbare.

De simpelste oplossing is uit te gaan van monotoon stijgende (dus nooit dalende) waarden. In het voorbeeld van de zonnecellen is dat de totaal opgewekte energie die sinds de installatie werd opgeleverd. Of de totale tijd dat er energie is opgewekt. Doordat zulke waarden niet door de meting worden beïnvloed kan er gewoon worden gesampled, waarna die samples zo goed mogelijk worden verwerkt. RRDtool kan zo worden geconfigureerd dat de verschillen tussen opeenvolgende meetwaarden worden opgeslagen, en niet de daadwerkelijk gemeten waarden. Deze laatste oplossing is uiteraard de beste. Waar een momentopname van het vermogen kan worden beïnvloed door een wolk voor de zon, is de meting van de toename in opgewekte energie over de laatste vijf minuten een goede maat voor het gemiddelde vermogen over die hele periode.

De meest ruwe vorm van dergelijke gegevens bestaat uit twee componenten:

1. De hoeveelheid opgewekte energie tot nu toe;
2. De totale tijd dat er tot nu toe is opgewekt.

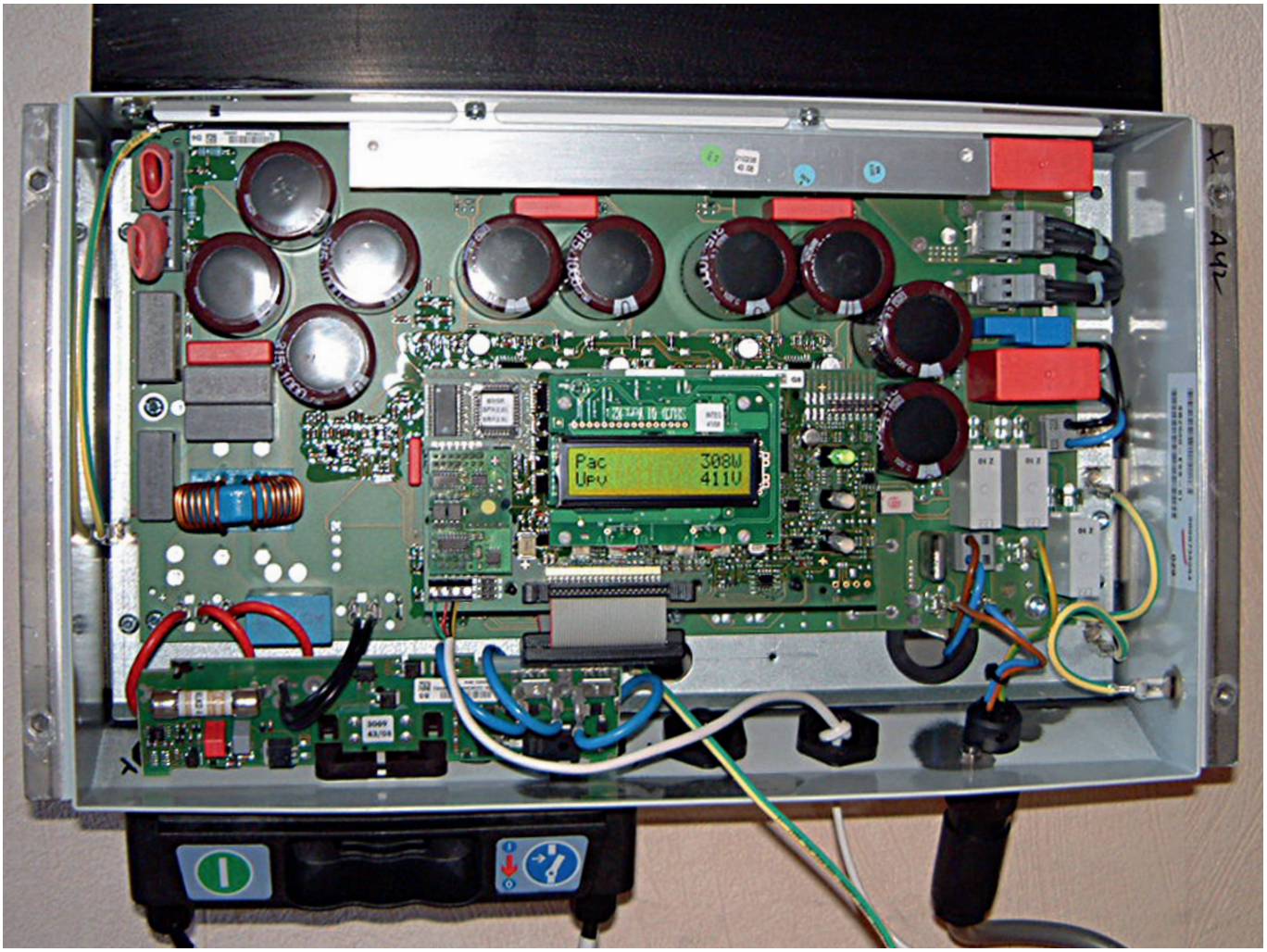
Op basis van deze twee gegevens kan van alles worden herleid over de prestatie van de zonnecellen.

Rekenen leert meer dan meten

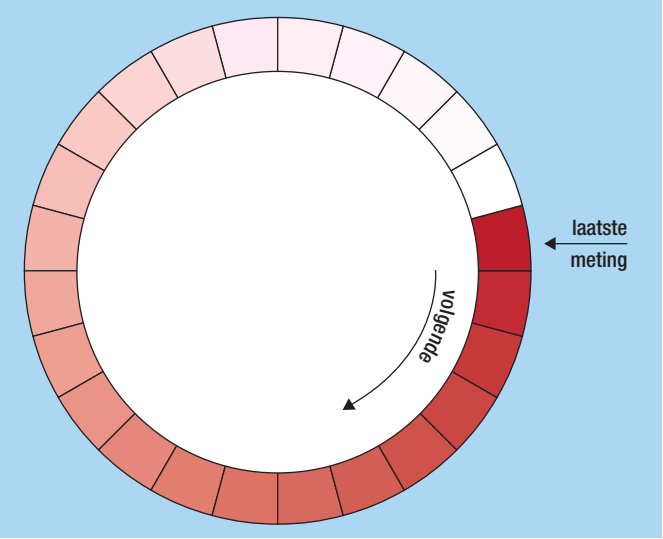
RRDtool is in staat om de meetgegevens op allerlei manieren te verwerken. Zo kunnen bijvoorbeeld de 288 meetwaarden in een etmaal worden opgeteld, gemiddeld, of het minimum en maximum kunnen worden bepaald. En daar kun je desgewenst weer maandgrafieken uit afleiden. Ook complexere rekentoeren met het beschikbare cijfermateriaal zijn mogelijk.

Per uitgerekende waarde geef je ook aan hoeveel meetwaarden er moeten worden opgeslagen, zodat ook die weer cyclisch kunnen worden opgeslagen. In de metingen per vijf minuten wil je misschien een week terug kunnen bladeren, in de dagelijkse gemiddelden misschien een hele maand. En als er maandgegevens worden bijgehouden dan wil je die misschien wel tien jaar terug kunnen zien. Zoveel maandgegevens zijn er niet, dus dat kost weinig ruimte.

Een klassieke aanpak met een relationele database zou zijn geweest om de daggemiddelden uit de meetgegevens per vijf minuten af te leiden, en voor dat doel zouden de metingen per vijf minuten jarenlang moeten worden onthouden. Terwijl al dat detail uit het verleden je eigenlijk niet meer interesseert.



Afbeelding 1: Dit is de binnenzijde van een wisselrichter, die de gelijkstroom van zonnecellen omzet in wisselende netstroom. Middenvoor de controllerprint, met daarheen voerend de RS-232 kabel die de link legt naar een computer met daarop RRDtool.



Afbeelding 2: De opslag van meetgegevens in RRDtool gebeurt cyclisch. Als een nieuwe meting binnenkomt dan schuift de pointer met 'nu' op om een verouderd meetgegeven te consumeren, en te vervangen door het nieuwe meetgegeven.

De extra kennis dat het verleden wat waziger mag worden wordt in RRDtool automatisch meegenomen, terwijl het in een relationele database expliciet zou moeten worden uitgeprogrammeerd, en zou leiden tot een complexer datamodel.

Het aantal meetwaarden dat in de twee werkwijzen zou worden bijgehouden voor bovenstaand voorbeeld:

Periode	Bewaartijd	RRDtool	SQL
5 min	1 week	2016	1051920
1 dag	1 maand	31	0
1 maand	10 jaar	120	0
		2167	1051920

De besparing is bijna een factor 500, en dan hebben we het nog niet gehad over de rekentijd die wordt bespaard, doordat RRDtool de herleide waarden als een soort afgeleid meetpunt opslaat in plaats van telkens opnieuw de data te verzamelen en die door te rekenen. Dit maakt heel duidelijk dat er een niche is waarin RRDtool veel beter functioneert dan het relationele model.

Bijzondere ontwerpkeuzes

Enkele aspecten van RRDtool zijn een beetje vreemd. Die komen voort uit de strakke structuren, maar ze zijn niet zo erg dat er niet mee te leven valt. Een maand duurt bijvoorbeeld zo'n 30,5 dagen. Met die insteek krijg je gemiddelden en andere overzichtswaarden die zich goed laten vergelijken; dat wordt anders bereikt door statistische correcties op de meetgegevens toe te passen. Bij RRDtool ligt het meer voor de hand elke periode een vast aantal meetgegevens te laten samenvatten. Blijkbaar is het daarvoor nodig om geen dagoverzichten te maken maar overzichten per halve dag.

Dan is er voor zonnecellen nog een punt over het meten van het eind van een dag en het begin van de volgende dag. Als de installatie niets opwekt dan schakelt alles uit, en zijn er geen meetgegevens beschikbaar. Dat houdt in dat er een meetverzoek na vijf minuten is waarop niets meer wordt geantwoord. De vraag is dan of de meetgegevens die de volgende ochtend naar voren komen niet deels komen uit de vijf minuten aan het eind van de vorige dag.

Een wisselrichter kan daar soms uitkomst bieden, als deze behalve de totalen sinds het moment van installatie ook dagtotalen toont; hiermee is (op een laat moment) nog te bepalen wat de vorige dag had moeten worden gemeten aan energie en tijd, en vervolgens kunnen de dagtotalen van zo'n eerste meting worden gebruikt om die eerste periode van vijf minuten te vullen. Op die manier meet je exact. Overigens vraagt RRDtool zo weinig van de computer dat het prima mogelijk is om met kleine tussenpozen te meten. Daardoor vallen de eerste en laatste metingen vanzelf in het begin en eind van de dag, als de zon amper iets oplevert. Dat betekent dat de afwijkingen bijzonder klein zullen zijn. Maar RRDtool kan dus wel overweg met een exacte benadering van wanneer de zon onder is gegaan – gewoon door achteraf meetwaarden toe te voegen, met het tijdstip waarop ze plaatsgevonden hebben, in plaats van het normale patroon om een waarde voor 'nu' op te geven.

Het is mogelijk om grafieken uit RRDtool te trekken. Die gaan dan wel altijd over meetgegevens die aan boord zijn, in de aanwezige resolutie. Naast zo'n terugblik over de laatste twaalf maanden is het vaak ook leuk om maandoverzichten per jaar te hebben, en dat soort grafieken trek je het beste uit het systeem op 1 januari. Of een overzicht per maand, dat maak je het beste op de eerste van die maand. De grafieken sla je op en lepel je (bijvoorbeeld via webpagina's) op als erom gevraagd wordt.

Zonnestroom bijhouden

Zonnestroom is een middel om af te koppelen van schaarser wordende fossiele bronnen. Maar de zelfstandige rol als energieproducent maakt het ook aantrekkelijker om te meten hoe goed het loopt. De gelijkstroom die uit zonnecellen komt wordt door een zogenaamde wisselrichter omgezet in wisselspanning op een iets hoger voltage dan het stroomnet, zodat het terugstroomt. 's Avonds haal je het er dan weer uit, eventueel tegen een lager tarief. De wisselrichter bevat vaak een klein computertje dat de omzetting optimaal doet verlopen, en gegeven een seriële aansluiting kunnen zaken worden gemeten, zoals de totale en vandaag opgewekte energie, de totale en vandaag gedraaide opwektijd, en het huidige vermogen. Deze gegevens zijn in een servertje op te vangen en om te zetten in gegevens die in RRDtool kunnen worden ingevoerd.

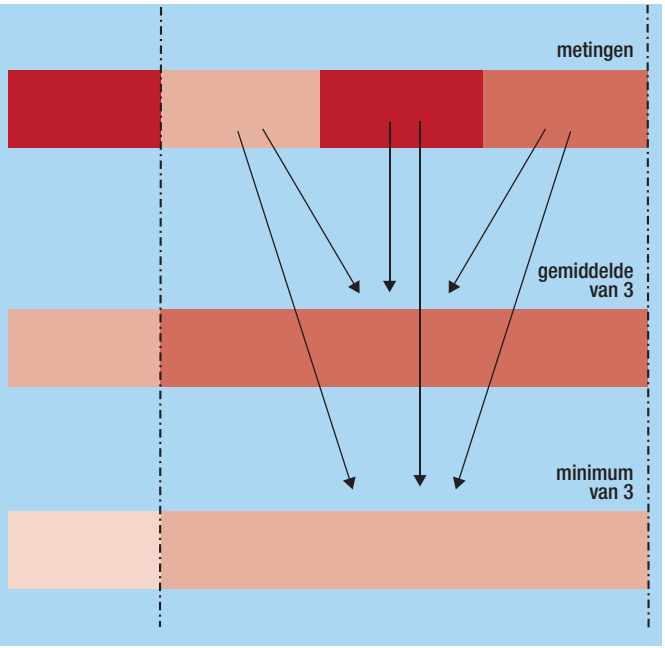
Het eerder vermelde rekenvoorbeeld is dus eigenlijk wat hypothetisch; het ligt meer voor de hand dat we jaaroverzichten uit RRDtool trekken en dat het er daarna niet meer toe doet wat er in zit. Om wat correctietijd te hebben bij fouten in de jaargegevens, kunnen maandgegevens dus bijvoorbeeld 14 maanden lang aanwezig zijn, maar tien jaar is niet eens nodig, tenzij er behoefte zou zijn aan een overzicht per decennium.

Praktische details

De praktijk is weerbarstig, en de afhankelijkheid van RS-232 en de Nederlandse zonneschijn en een gecommercialiseerd elektriciteitsnet maakt dat er allemaal niet beter op. Het leuke is dat RRDtool bij uitstek geschikt is om met falende systemen om te gaan.

Als een meetwaarde niet beschikbaar is dan wordt er expliciet genoteerd dat die er niet is; een soort NULL dus eigenlijk. Bij het berekenen van gemiddelden, maxima en minima wordt daar op de juiste wijze mee omgegaan. Bij een monotoon stijgende waarde betekent het ontbreken van een meetwaarde zelfs dat de waarde ervoor en erna niet uitgerekend kan worden. Er kan bovendien worden geconfigureerd dat een afgeleide waarde bij ontbreken van te veel meetpunten zelf ook NULL oplevert. Dus iets als "een uurgemiddelde bepaal je ideaal uit twaalf, maar minimaal uit acht metingen per vijf minuten".

Vaker dan volledig ontbrekend, zal een meetgegeven te laat aankomen. In zo'n geval kan RRDtool zelfs interpoleren bij het toevoegen van een meetwaarde. Dat houdt in dat we ons kunnen beperken tot het toevoegen van de meetwaarde op moment 'nu' en de verdere zorgen aan RRDtool over laten. Natuurlijk blijft het wel zo dat er minder geïnterpoleerd hoeft te worden wanneer de meetwaarde dichter bij het ideale meettijdstip wordt ingevoerd. Tenslotte is het noemenswaardig dat RRDtool faciliteiten heeft voor trendanalyse. Deze trends kunnen bijvoorbeeld worden gebruikt om een boven- en ondergrens voor normaal gebruik af te bakenen. Wanneer een grens wordt doorbroken kan aan een alarmbel worden getrokken.



Afbeelding 3: Meetgegevens worden gecombineerd tot diverse afgeleide waarden. Hier worden drie metingen gecombineerd tot een gemiddelde en een minimum. Als de meetgegevens verouderd zijn verdwijnen ze uit het zicht, maar de samenvattende overzichtsgegevens zijn dan allang veilig gesteld.

Toepassingsmogelijkheden

RRDtool is een database die in een heel eigen niche opereert. Alles wat het bijhoudt moet een functie van de tijd zijn, en op zo'n manier worden bijgehouden dat historische gegevens samengevat kunnen worden. Er zijn flink wat toepassingen waarin dit bijzonder nuttig is. Een snelle brainstorm: beschikbaarheid van een server; gebruik van een service; mensenstromen (bus, warenhuis, concerten); productiviteit in een fabriek; ontwikkelingen in politieke opinie; de valutamarkt; allerhande CBP-metingen; urenregistraties per werknemer.

Grafische dondersteen

RRDtool staat bol van de grafische configureermogelijkheden. Een flink aantal voorbeelden is te vinden op de RRDtool homepage. Het is zeer de moeite waard om daar de Gallery te bekijken om een indruk te krijgen. Het is mogelijk om in de grafische vraagtaal met expressies aan te geven welke kleuren moeten worden gebruikt. Valt iets buiten de trend, dan kleur je het rood bijvoorbeeld.

De uitvoer van dit deel van RRDtool is bijvoorbeeld een PNG. Dat is een vreemd resultaat als je RRDtool ziet als een SQL-achtige database, wat het dus niet is. Je stopt er getallen in en trekt er plaatjes uit, wat kan er nog mooier zijn? Tja, het kan nog iets handiger. Er zijn lagen bovenop RRDtool gelegd die je wat low-level beheer van RRDtool kunnen besparen. Wat dat betreft wordt Munin bijvoorbeeld warm aangeraden. Het is nog steeds een commandline utility, maar kan gemakkelijker over netwerkverbindingen gebruikt worden.

Conclusies

RRDtool is een vreemde eend in de bijt, maar het is beslist een database die zijn eigen niche vult. Tijdsafhankelijke data waarvan vooraf bekend is welke afgeleide overzichtsdata eruit gehaald moeten worden, vormen een zeer geschikt toepassingsdomein. De specifieke trekjes die RRDtool voor dit domein heeft maken het regelmatig een veel geschikter tool dan een relationele database.

RRDtool is te vinden op http://oss.oetiker.ch/rrdtool/

Rick van Rein

Dr. ir. H. van Rein (rick@openfortress.nl) is ontwikkelaar en beheerder bij OpenFortress Digital signatures.