



Universiteit  
Leiden  
The Netherlands

## **Integrating analytics with relational databases**

Raasveldt, M.

### **Citation**

Raasveldt, M. (2020, June 9). *Integrating analytics with relational databases*. *SIKS Dissertation Series*. Retrieved from <https://hdl.handle.net/1887/97593>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/97593>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/97593> holds various files of this Leiden University dissertation.

**Author:** Raasveldt, M.

**Title:** Integrating analytics with relational databases

**Issue Date:** 2020-06-09

---

## Samenvatting

---

Database onderzoekers hebben enorme voortgang geboekt in het ontwikkelen van krachtige database systemen die efficient analytische queries kunnen beantwoorden. Deze krachtige systemen worden echter zelden gebruikt door analytici. Dat komt voornamelijk omdat het gebruik van huidige relationele database systemen in combinatie met de programma's die zij gebruiken traag en onhandig is. In plaats van deze database systemen te gebruiken, zijn analytici database systemen opnieuw aan het uitvinden. Ze schrijven hun eigen programma's die vergelijkbare functionaliteit hebben, maar de innovaties van het database veld van de afgelopen decennia negeren.

In dit proefschrift proberen we dit probleem op te lossen. We doen dit door te onderzoeken hoe we de integratie van database systemen met deze analytische programma's efficiënter en gebruiksvriendelijker kunnen maken. Ons onderzoek is gefocussed op de drie primaire methodes van database-client integratie: client-server verbindingen, in-database analyses en gintegreerde database systemen.

Voor elk van deze methoden onderzoeken wij de implementaties in bestaande database systemen, en evalueren wij hoe efficient deze zijn voor standaard analytische gebruik. We kijken naar de voor en nadelen van elk van deze technieken, zowel in termen van efficiëncie als in gebruiksvriendelijkheid.

We introduceren meerdere nieuwe technieken die verbeteren op de huidige state-of-

the-art. We demonstreren een nieuw client-server protocol dat wij hebben ontwikkeld dat geoptimaliseerd is voor bulkoverdracht van grote data sets. We laten onze MonetDB/Python user-defined functions zien, die efficiënte grootschalige in-database analyses versnellen door gebruik te maken van vectorisatie. Uiteindelijk beschrijven wij MonetDBLite, een versie van het MonetDB database systeem die wij hebben geventueerd in R en Python. Al onze technieken zijn getest in de context van echte systemen, wat laat zien dat onze oplossingen niet alleen theoretisch maar ook praktisch toepasbaar zijn.

In het laatste hoofdstuk introduceren wij DuckDB, een nieuw data management systeem dat wij hebben gebouwd met als specifiek doel om deze analytici te ondersteunen. Bij het bouwen van DuckDB hebben wij alle lessen die we hebben geleerd over de integratie van database systemen met analytische applicaties toegepast.

In conclusie, de algoritmes die wij hebben ontwikkeld maken het mogelijk om database systemen veel efficiënter te integreren met analytische applicaties. Desalniettemin is er nog meer om te onderzoeken in dit gebied. We sluiten dit proefschrift af met suggesties voor toekomstig onderzoek, samen met ideeën voor eventuele oplossingen.