



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

# Integrating Analytics with Relational Databases

Mark Raasveldt

# Integrating Analytics with Relational Databases

Proefschrift

ter verkrijging van  
de graad van Doctor aan de Universiteit Leiden,  
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,  
volgens besluit van het College voor Promoties  
te verdedigen op dinsdag 9 juni 2020  
klokke 11:15 uur

door

Mark Raasveldt  
geboren te Leiderdorp  
in 1992

## Promotiecommissie

Promoter:	prof. dr. Stefan Manegold	(CWI & Leiden University)
Co-promoter:	dr. Hannes Mühleisen	(CWI)
Overige leden:	prof. dr. Aske Plaat	(Leiden University)
	prof. dr. Thomas Bäck	(Leiden University)
	dr. Mitra Baratchi	(Leiden University)
	prof. dr. Jens Dittrich	(Saarland University)
	prof. dr. Torsten Grust	(University of Tübingen)

The research reported in this thesis has been carried out at the CWI, the Dutch National Research Laboratory for Mathematics and Computer Science, within the Database Architectures group.

The research reported in this thesis has been carried out under SIKS, the Dutch Research School for Information and Knowledge Systems.

This research is financially supported by the Dutch funding agency NWO, under project number 650.002.001 (the PROMIMOOC project), in collaboration with Tata Steel IJmuiden, BMW Group Regensburg, Leiden University and Centrum voor Wiskunde en Informatica (CWI).



# Acknowledgments

When I was studying for my masters in university I always thought that I would never do a PhD. After all, you get paid less than working in industry and you need to work on theoretical research instead of solving practical problems. The reason that I opted to do a PhD anyway is because of Hannes, Stefan and the rest of the Database Architectures group. They showed me that not only can research be useful and practical, it can be amazingly fun and engaging as well. I have learned so much in my time here and am very thankful to each of the members of the DA group that have provided me with their knowledge and wisdom.

I am particularly indebted to Hannes, who took me in as a master student and has worked closely with me ever since. All of the days that we spent peer programming were extremely fun and informative. I would also like to give special thanks to my promotor Stefan. Firstly for giving me the opportunity of doing my PhD at the CWI, and secondly for being extremely kind and compassionate and creating such an accepting and amazing workplace. They say that how you experience your PhD depends entirely on your supervisor, and I had the best supervisors that I could wish for in Hannes and Stefan.

In my time at the CWI I have made many friends that have made my time there extremely pleasant. Pedro, Tim and Diego who I could always count on to have a good time and with whom I share many amazing memories. Thibault, who has taught me how to enjoy the pleasantries of life and how to write introductions. Abe, who has always impressed me with his math skills. Till, who was always ready to beat me in a game of table tennis. And finally, I would like to thank all the other people of the DA group for making my time there so special and amazing.

Finally, I would like to acknowledge my family and friends for their support. My mother and father - both also computer scientists - who have always supported me in doing whatever I wanted to do, even though I ended up following in their footsteps anyway. I would also like to thank my siblings, Maarten and Marieke. My close friends David and Dirk, and especially Ana who has always supported me.





---

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>11</b>
1	The Rise of Data Science . . . . .	11
2	Tools of the Trade . . . . .	12
3	Data Science & Data Management . . . . .	13
4	Our Contributions . . . . .	14
5	Structure and Covered Publications . . . . .	15
<b>2</b>	<b>Background</b>	<b>17</b>
1	Relational Database Management Systems . . . . .	18
2	RDBMS Design . . . . .	19
2.1	Workload Types . . . . .	19
2.2	System Types . . . . .	20
2.3	Physical Database Storage . . . . .	20
2.4	Database Processing Models . . . . .	21
3	Database Connectivity . . . . .	23
3.1	Client-Server Connection . . . . .	23
3.2	In-Database Processing . . . . .	24
3.3	Embedded Databases . . . . .	25

4	MonetDB . . . . .	25
5	Python . . . . .	28
<b>3</b>	<b>Database Client-Server Protocols</b>	<b>31</b>
1	Introduction . . . . .	31
	1.1 Contributions . . . . .	32
	1.2 Outline . . . . .	33
2	State of the Art . . . . .	33
	2.1 Overview . . . . .	34
	2.2 Network Impact . . . . .	37
	2.3 Result Set Serialization . . . . .	39
3	Protocol Design Space . . . . .	43
	3.1 Protocol Design Choices . . . . .	44
4	Implementation & Results . . . . .	54
	4.1 MonetDB Implementation . . . . .	54
	4.2 PostgreSQL Implementation . . . . .	55
	4.3 Evaluation . . . . .	57
5	Summary . . . . .	62
<b>4</b>	<b>Vectorized UDFs in Column-Stores</b>	<b>63</b>
1	Introduction . . . . .	63
	1.1 Contributions . . . . .	64
	1.2 Outline . . . . .	65
2	Types of User-Defined Functions . . . . .	65
3	MonetDB/Python . . . . .	66
	3.1 Usage . . . . .	66
	3.2 Processing Pipeline . . . . .	67
	3.3 Parallel Processing . . . . .	70
	3.4 Loopback Queries . . . . .	75
4	Evaluation . . . . .	75
	4.1 Systems Measured . . . . .	76

4.2	Modulo Benchmark . . . . .	77
5	Related Work . . . . .	79
5.1	Research . . . . .	79
5.2	Systems . . . . .	82
6	Applicability To Other Systems . . . . .	83
7	Development Workflow: devUDF . . . . .	85
7.1	The devUDF Plugin . . . . .	87
7.2	Usage . . . . .	87
7.3	Implementation . . . . .	89
8	Summary . . . . .	90
<b>5</b>	<b>In-Database Workflows</b>	<b>93</b>
1	Introduction . . . . .	93
1.1	Contributions . . . . .	94
1.2	Outline . . . . .	94
2	Related Work . . . . .	94
2.1	Machine Learning Integration . . . . .	94
2.2	Machine Learning Model Management . . . . .	95
3	Machine Learning integration . . . . .	96
3.1	Training . . . . .	97
3.2	Classification . . . . .	98
3.3	Ensemble Learning . . . . .	99
4	Experimental Analysis . . . . .	99
5	Summary . . . . .	102
<b>6</b>	<b>MonetDBLite</b>	<b>103</b>
1	Introduction . . . . .	103
1.1	Contributions . . . . .	104
1.2	Outline . . . . .	104
2	Design & Implementation . . . . .	104
2.1	Internal Design . . . . .	104

2.2	Embedding Interface . . . . .	105
2.3	Native Language Interface . . . . .	107
2.4	Technical Challenges . . . . .	111
3	Evaluation . . . . .	113
3.1	Setup . . . . .	113
3.2	TPC-H Benchmark . . . . .	114
3.3	ACS Benchmark . . . . .	121
4	Summary . . . . .	123
<b>7</b>	<b>DuckDB: an Embeddable Analytical Database</b>	<b>125</b>
1	Introduction . . . . .	125
1.1	Contributions . . . . .	127
2	Design and Implementation . . . . .	128
3	Summary . . . . .	130
<b>8</b>	<b>Conclusion</b>	<b>133</b>
1	Big Picture . . . . .	133
2	Future Research . . . . .	134
2.1	Client-Server Connections . . . . .	134
2.2	In-Database Processing . . . . .	135
2.3	Embedded Databases . . . . .	138
	<b>Bibliography</b>	<b>140</b>
	<b>Summary</b>	<b>151</b>
	<b>Samenvatting</b>	<b>153</b>
	<b>Publications</b>	<b>155</b>