

Efficient tabular data ingestion and manipulation with MonetDBLite

Hannes Mühleisen*

Thanks previous Speakers

- Javier for DBI Intro
- Ben for proposing persistent SQL DBs for medium data

Running a DB is hard

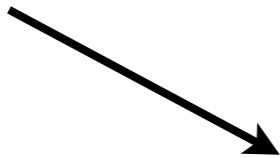
- Installation / Automatic startup difficult
- Configuration for workload often crucial
 - `checkpoint_completion_target`?
 - `effective_cache_size`?
- Maintenance, updates, ... workload increase
 - Most people don't bother if not forced

What about SQLite ?

- In-process SQL database, data either in memory or in a file, rock-solid, used on every smartphone, browser, OS,
- People also use it for large-ish dataset analysis
- Bad idea, SQLite was never built for this
 - e.g. row-based storage model

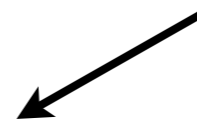
Postgres, MySQL, etc.:

Conceptual



class	speed	flux
NX	1	3
Constitution	1	8
Galaxy	1	3
Defiant	1	6
Intrepid	1	1

Physical (on Disk)



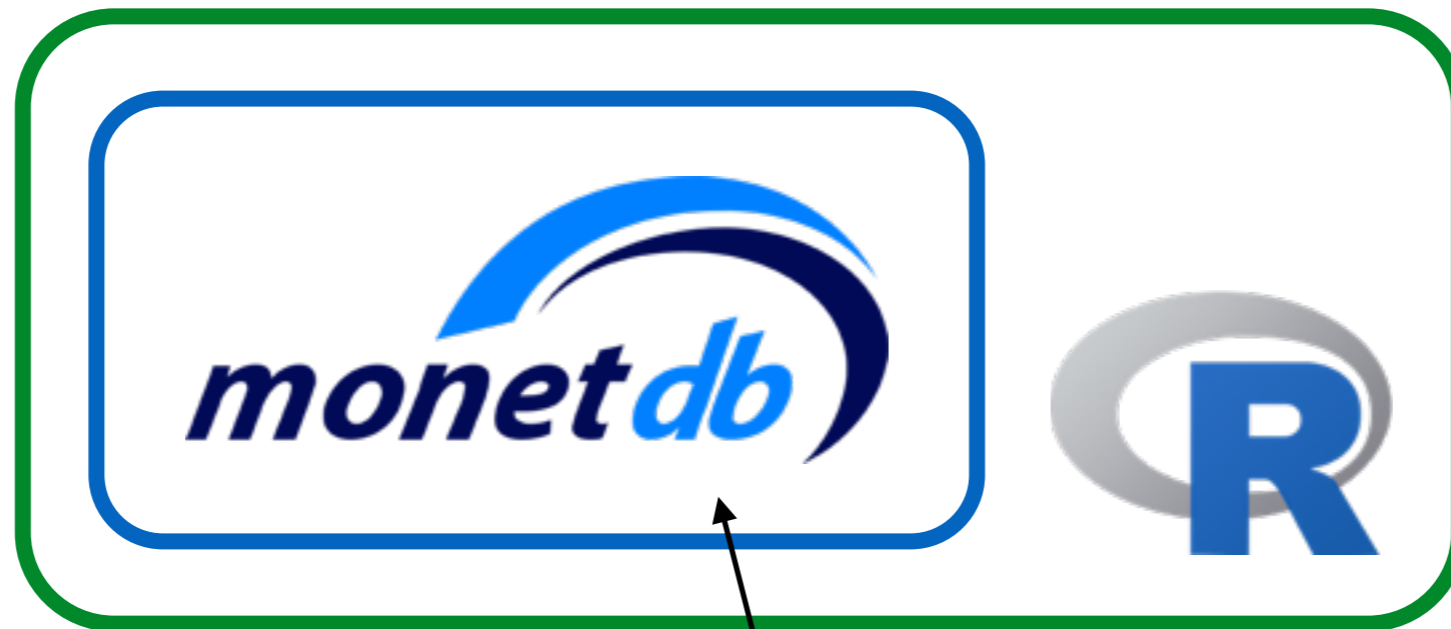
NX		1	3	Constitution		1	8	Galaxy	
1	3	Defiant		1	6	Intrepid		1	1

Column Store:

class	speed	flux
NX	1	3
Constitution	1	8
Galaxy	1	3
Defiant	1	6
Intrepid	1	1

NX	Constitution	Galaxy	Defiant	Intrepid
1	1	1	1	1
3	8	3	6	1

Enter MonetDBLite



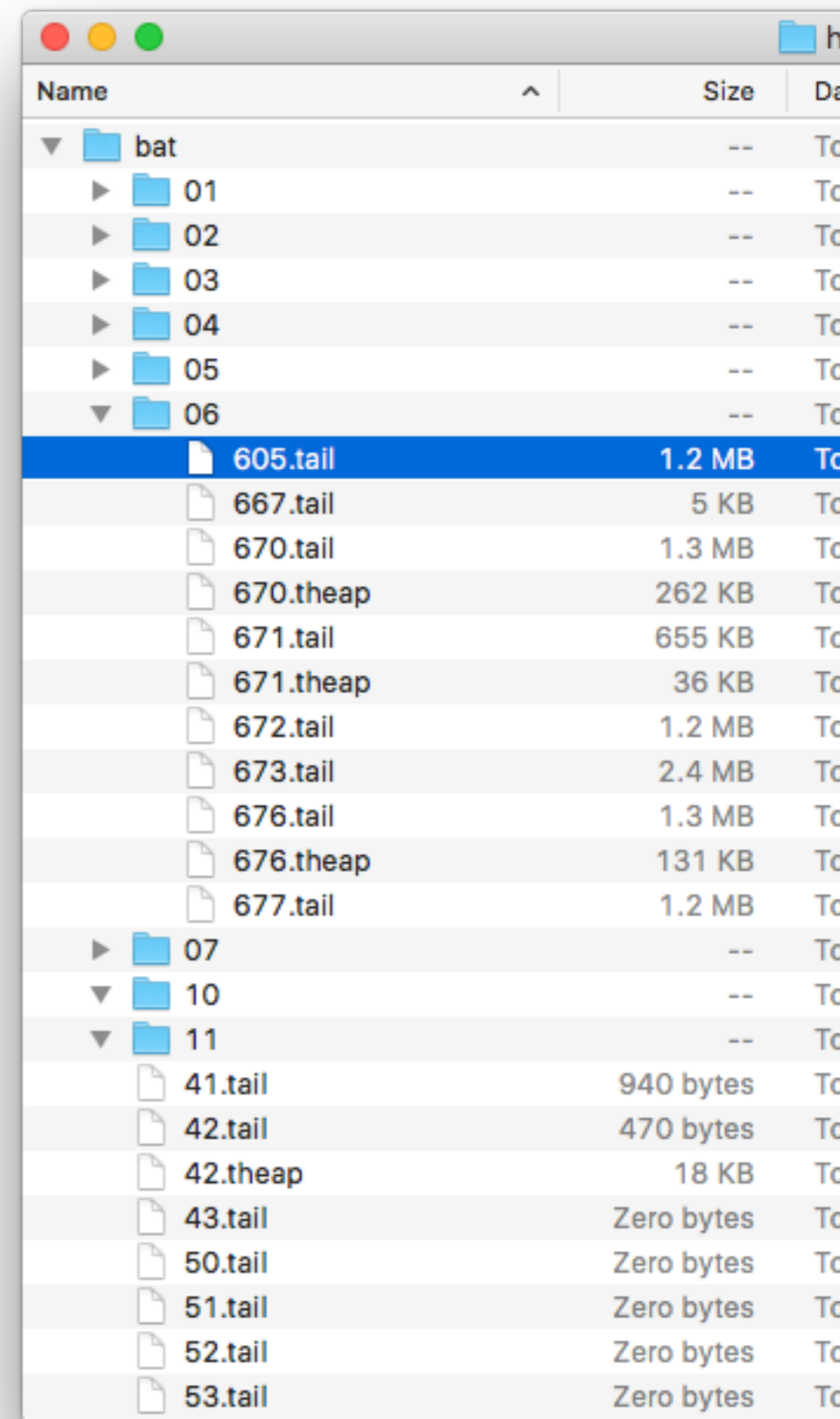
2016 SIGMOD systems award winner

What is MonetDBLite

- Embedded & streamlined MonetDB for R
- All of MonetDB: Transactions, complex SQL etc.
- In-Process operation
- Query results are data frames
- Fast data append from data frames
- DBI and dplyr backends

Persistence

- Table data is stored on disk in a native *columnar* format
- Persistent, super-fast access
- No more waiting at beginning of script to load data...



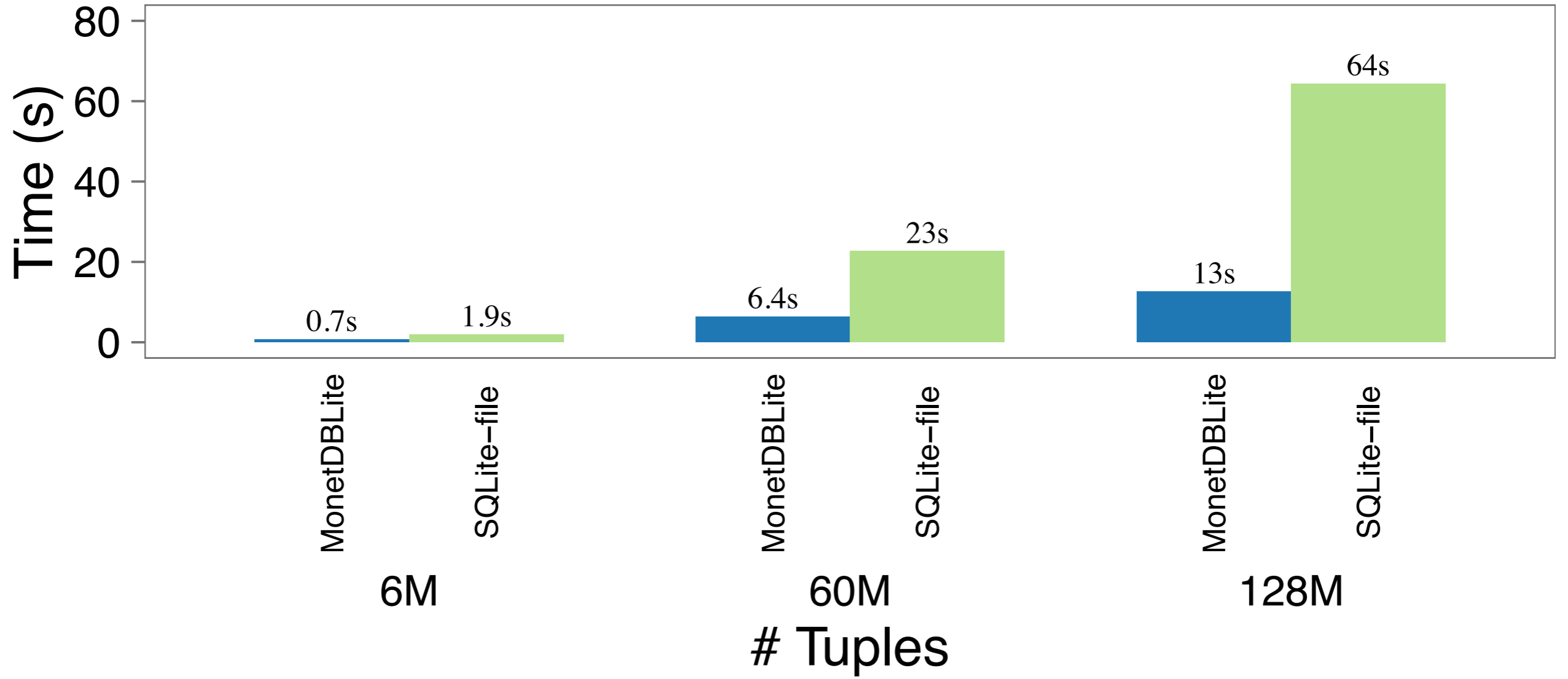
A screenshot of a file explorer window showing a directory structure. The window has a title bar with red, yellow, and green window control buttons. The main area displays a list of files and folders. The columns are labeled 'Name', 'Size', and 'Date'. The 'Name' column shows a hierarchy starting with a folder named 'bat', which contains subfolders '01' through '06'. Folder '06' is expanded, showing files '605.tail' (1.2 MB), '667.tail' (5 KB), '670.tail' (1.3 MB), '670.theap' (262 KB), '671.tail' (655 KB), '671.theap' (36 KB), '672.tail' (1.2 MB), '673.tail' (2.4 MB), '676.tail' (1.3 MB), '676.theap' (131 KB), and '677.tail' (1.2 MB). Below folder '06' are folders '07', '10', and '11'. Folder '11' is expanded, showing files '41.tail' (940 bytes), '42.tail' (470 bytes), '42.theap' (18 KB), '43.tail' (Zero bytes), '50.tail' (Zero bytes), '51.tail' (Zero bytes), '52.tail' (Zero bytes), and '53.tail' (Zero bytes).

Name	Size	Date
bat	--	To
01	--	To
02	--	To
03	--	To
04	--	To
05	--	To
06	--	To
605.tail	1.2 MB	To
667.tail	5 KB	To
670.tail	1.3 MB	To
670.theap	262 KB	To
671.tail	655 KB	To
671.theap	36 KB	To
672.tail	1.2 MB	To
673.tail	2.4 MB	To
676.tail	1.3 MB	To
676.theap	131 KB	To
677.tail	1.2 MB	To
07	--	To
10	--	To
11	--	To
41.tail	940 bytes	To
42.tail	470 bytes	To
42.theap	18 KB	To
43.tail	Zero bytes	To
50.tail	Zero bytes	To
51.tail	Zero bytes	To
52.tail	Zero bytes	To
53.tail	Zero bytes	To

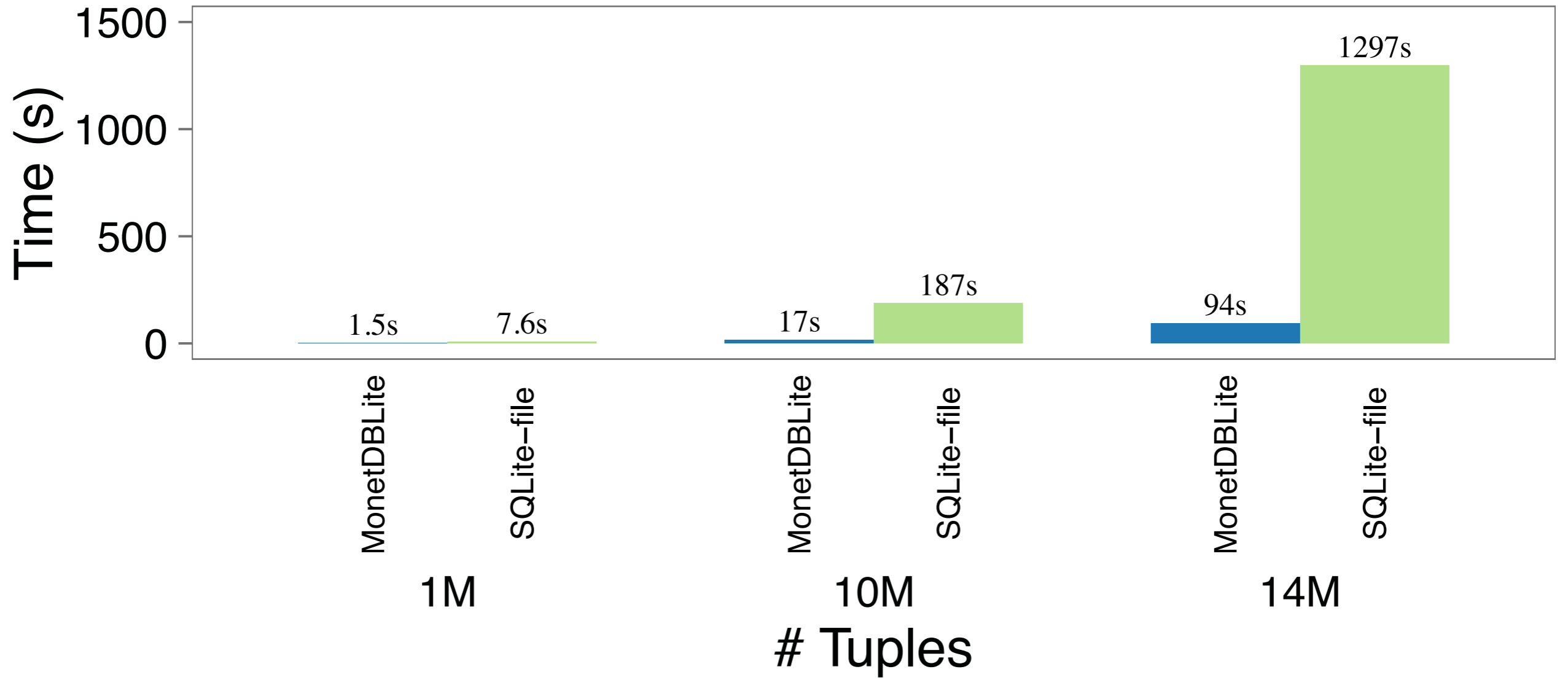
Performance

- Home Mortgage Disclosure Act dataset & queries
 - 128M records, 71 fields each, 56 GB CSV
 - Sampled down to 6M and 60M for testing
- Contenders: MonetDBLite & SQLite with a file backend
- Experiments:
 - HMDA queries
 - Table transfer

Run HMDA queries








Convert table to data.frame



End-to-end-testing

- “Sisyphus”
- <http://www.asdfree.com> , survey data with R
- Making sure analyses still runs with MonetDBLite

	acs	acs2	brfss	bsapuf	censo	ces	cps	hmda	limit1	misc	ncvs	nhts	nibrs	nppes	nvss	pisa	pnad	pums	sbo	seer	sipp	svyby	svymean	svytable	swmap_acs	swmap_cps	swmap_pnad	timss	
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Live Demo

Script at <https://gist.github.com/hannesmuehleisen/f6d590b4efcda539d0e8a27c420764dc>

MonetDBLite

Fast SQL, fast startup

<https://github.com/hannesmuehleisen/MonetDBLite>

Lazy typers: <http://ml.h4.ms>



@hfmuehleisen