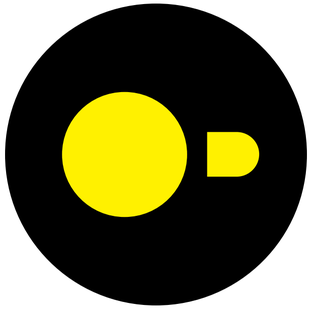


Developing Systems in Academia: The Good, the Bad, and the Not-so-Ugly Duckling

Hannes Mühleisen

DuckDB

Backstory


LAMP
Linux
Apache
MySQL
PHP


D
S QL 1337
BADEN-WUERTTEMBERG

Accelerating Queries with Group-By and Join by Groupjoin

Guido Moerkotte
Universität Mannheim
Mannheim, Germany
moerkotte@informatik.uni-mannheim.de

Thomas Neumann
Technische Universität München
Munich, Germany
neumann@in.tum.de

ABSTRACT

Most aggregation queries contain both group-by and join operators, and spend a significant amount of time evaluating these two expensive operators. Merging them into one operator (the *groupjoin*) significantly speeds up query execution.

We introduce two main equivalences to allow for the merging and prove their correctness. Furthermore, we show experimentally that these equivalences can significantly speed up TPC-H.

The groupjoin is more than 20 years old. To the best of our knowledge, von Bültzingsloewen invented the groupjoin [19]. He named it *outer aggregation*. Thereafter, several new names were invented. For example, Nakano used the name *general aggregate formation* [16]. The reason might be that grouping is called *aggregate formation* by Klug [12]. Steenhagen, Apers, and Blanken introduced the name *nest-join* [17]. Cluet and Moerkotte called it *binary grouping* [8]. Lately, Chatziantoniou, Akinde, Johnson, and Kim introduced the name *MD-Join* [5].

The groupjoin is quite versatile, and we strongly believe that no DBMS can do without it. For example, it has

Researcher

Co-Founder & CEO

Professor

New

CWI

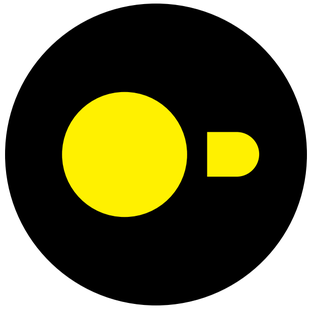
Database Architectures

DuckDB Labs

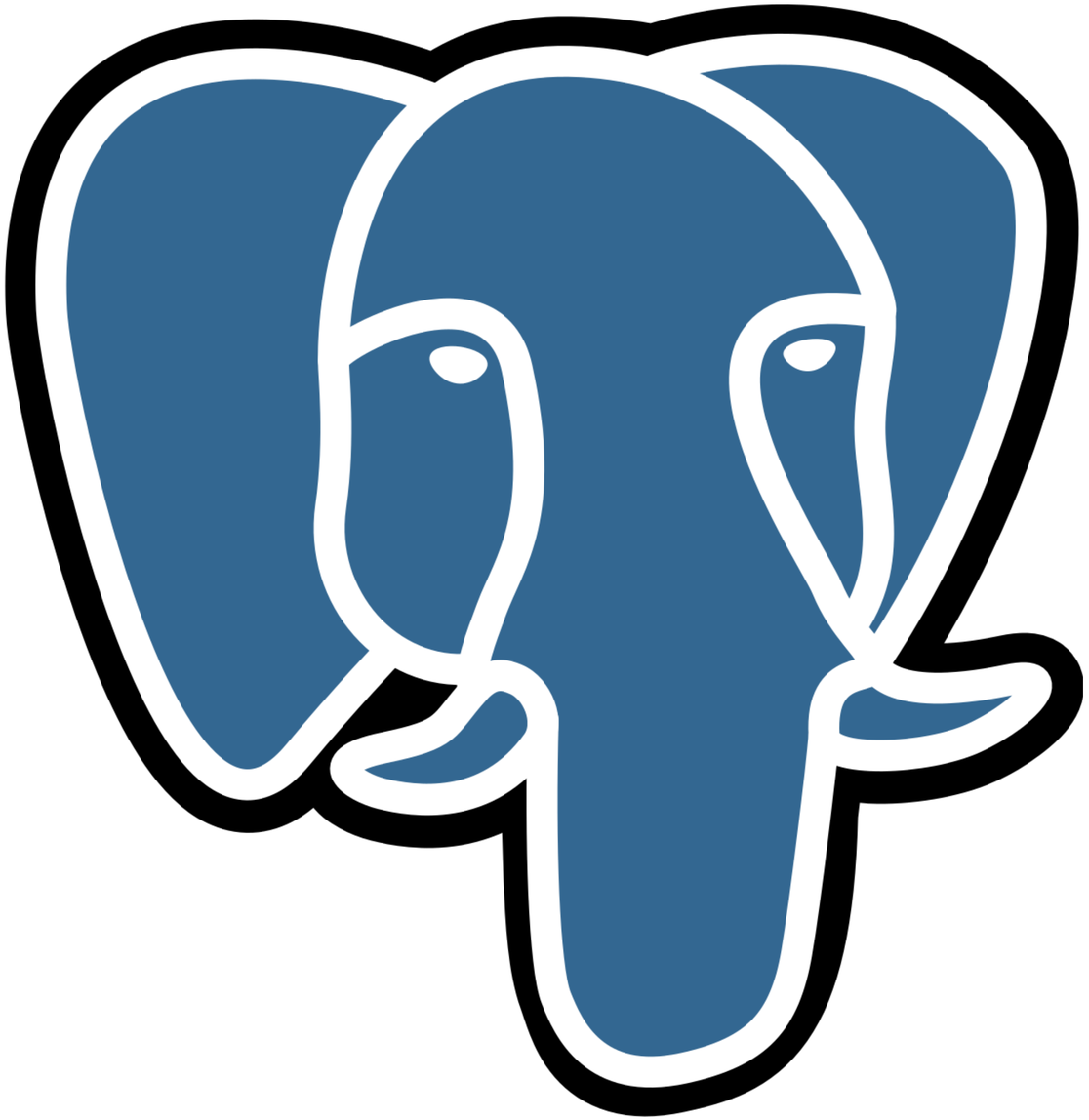
Radboud Universiteit Nijmegen

DuckDB

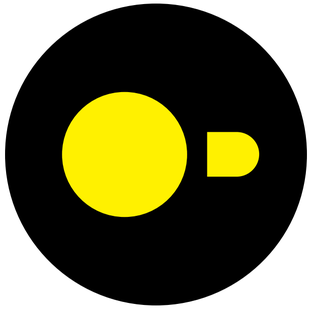
duckdb.org



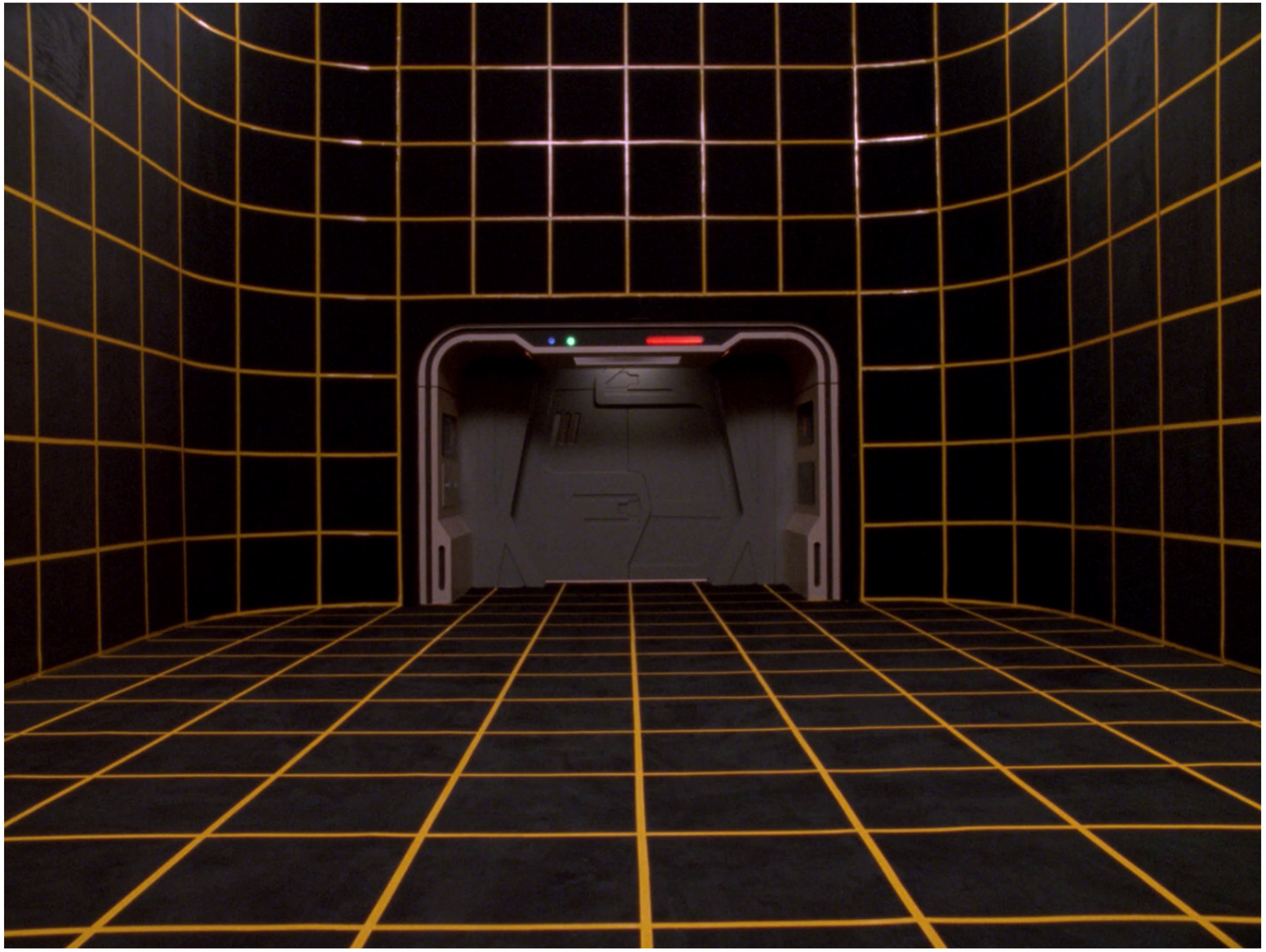
The **Good**



Impact



The **Bad**



Losing Game

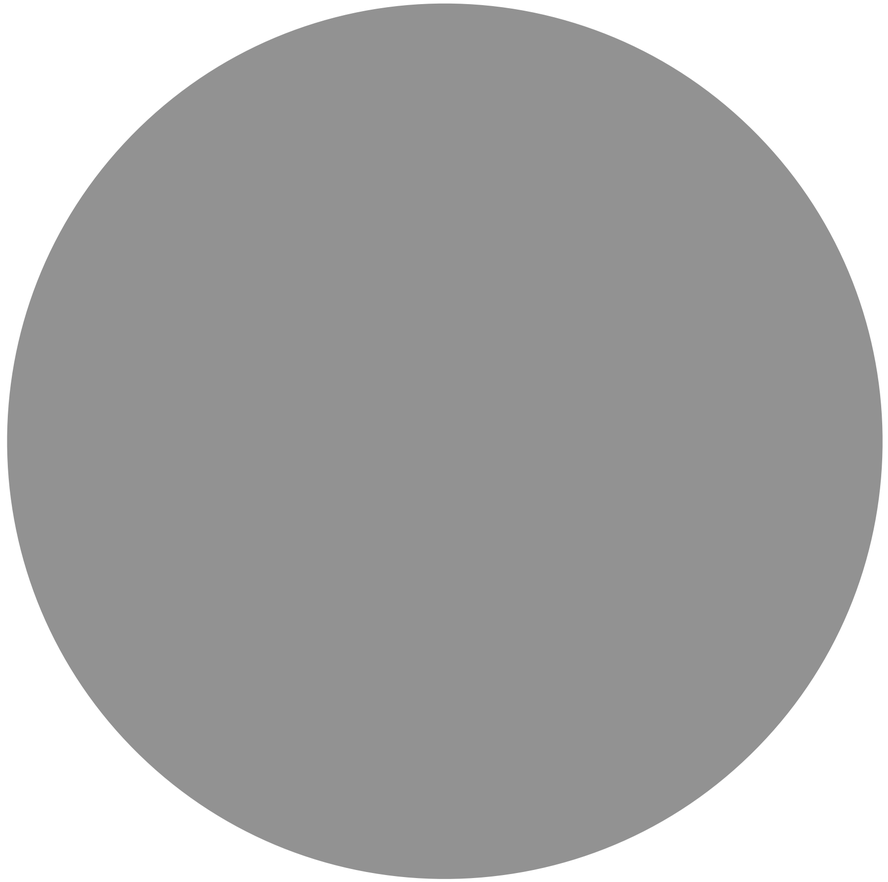
Resistance

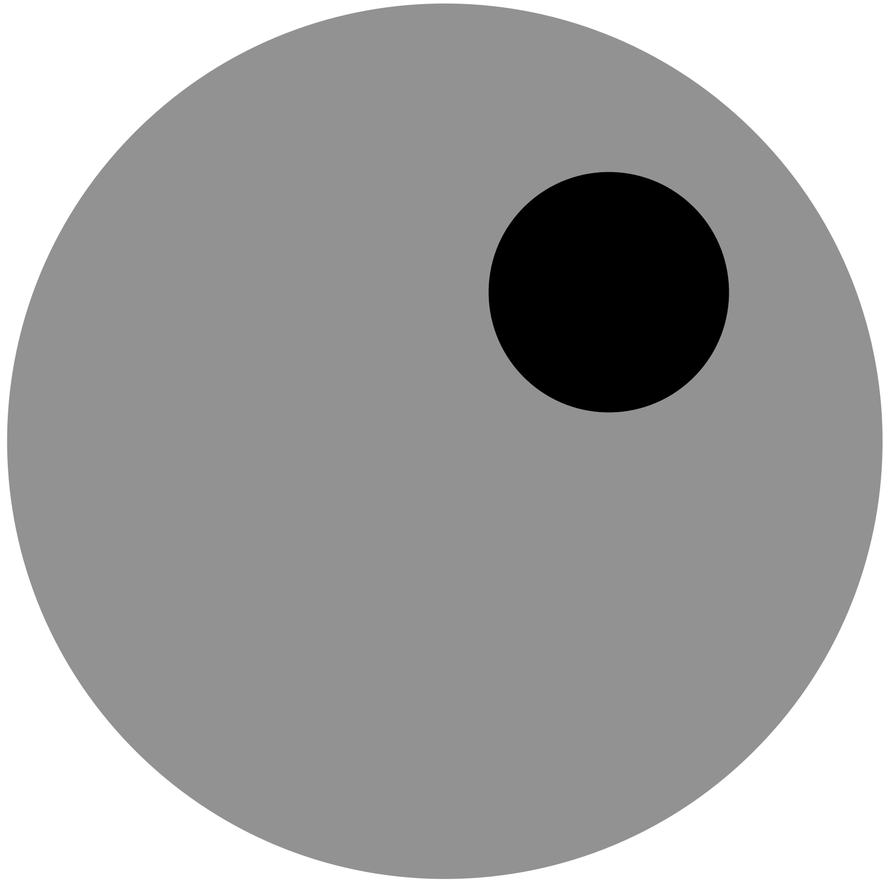
Career **Gamble**

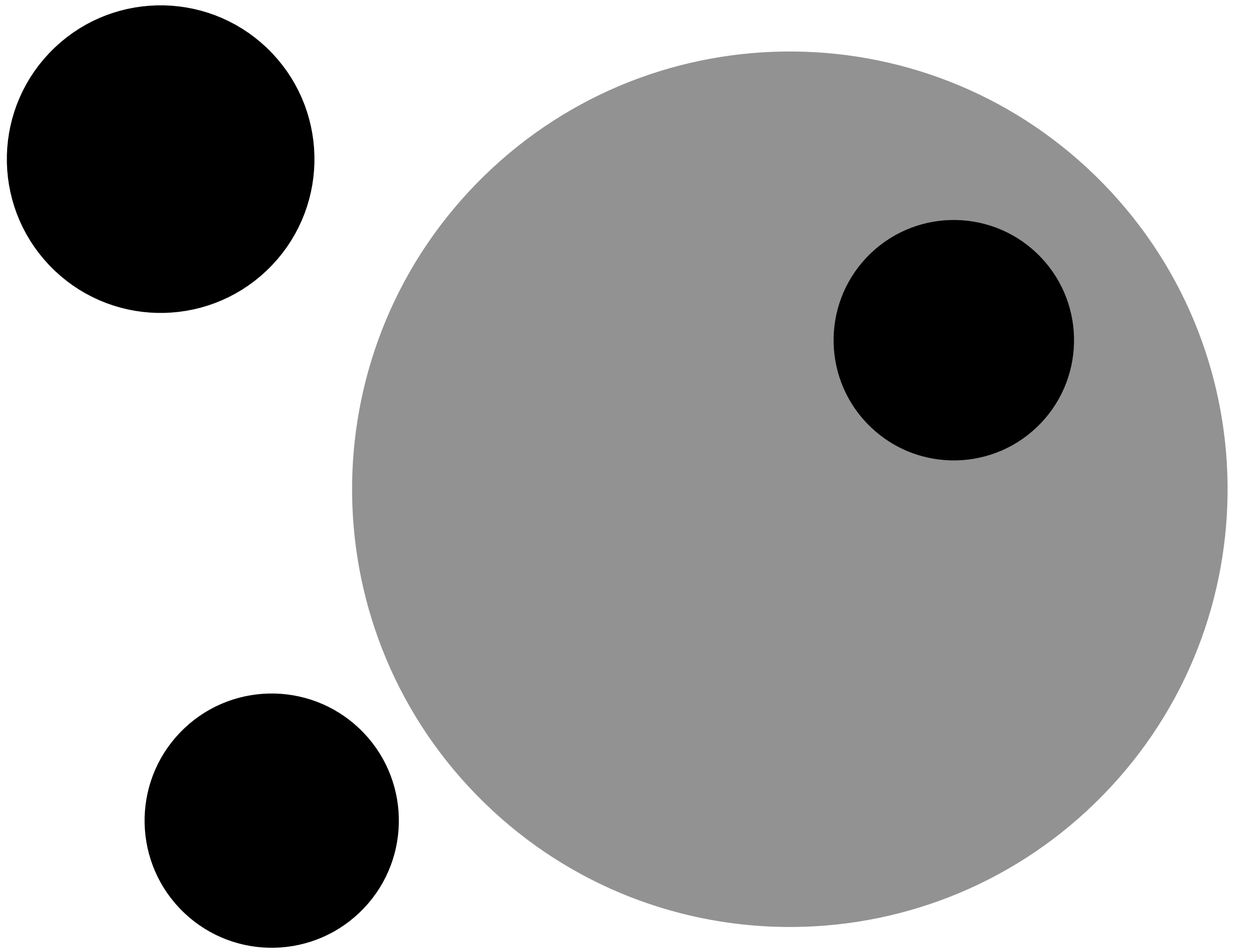
Funding

Recipe?

Weird Topics







Fewer **Papers**

Open Source

Information on Mission

Scope

The Journal of Statistical Software (JSS) publishes open-source software and corresponding reproducible scholarly articles discussing all aspects of the design, implementation, documentation, application, evaluation, comparison, maintenance and distribution of software dedicated to improvement of state-of-the-art in statistical computing in all areas of empirical research. Open-source code and articles are jointly reviewed and published in this journal and should be accessible to a broad community of practitioners, teachers, and researchers in the field of statistics.

Types of Papers

JSS aims at publishing articles describing comprehensive open-source implementations of broad classes of statistical models and procedures or computational infrastructure upon which such implementations can be built. JSS evaluates and publishes software and articles as a unity. Full reproducibility is mandatory for publication and the source code is published along with the article. Articles describing a specific software implementation must present the scope of the software and the underlying models or concepts such that statistically literate readers can understand what the software does. A careful comparison with other open-source implementations of similar models or procedures should highlight the capabilities of all implementations and the corresponding advantages or disadvantages. A description of the design principles and the actual implementation is at the heart of an JSS article. The general analysis workflow must be illustrated by an enlighting non-trivial case study.

JSS publishes special issues on topics in statistical computing and statistical software. Software described by articles

You can help us to improve our editorial management system by reporting possible issues [read more].

Journal Information

Publisher: Foundation for Open Access Statistics
Editors-in-chief: Bettina Grün, Torsten Hothorn, Rebecca Killick, Achim Zeileis
Board: Editorial Team
ISSN: 1548-7660
CODEN: JSSOBK

Sponsored by

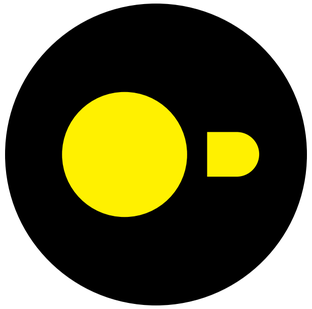
universität innsbruck Universität Zürich UCLA

Support

As a matter of principle, **JSS charges no author fees or subscription fees**. Our editors, reviewers, and programmers are volunteers. Universität Innsbruck, Universität Zürich, and UCLA Statistics provide support staff, website maintenance, website hosting, and some

De Vonk

‘Ik vind het niet ethisch als vooral investeerders profiteren van onze groei’



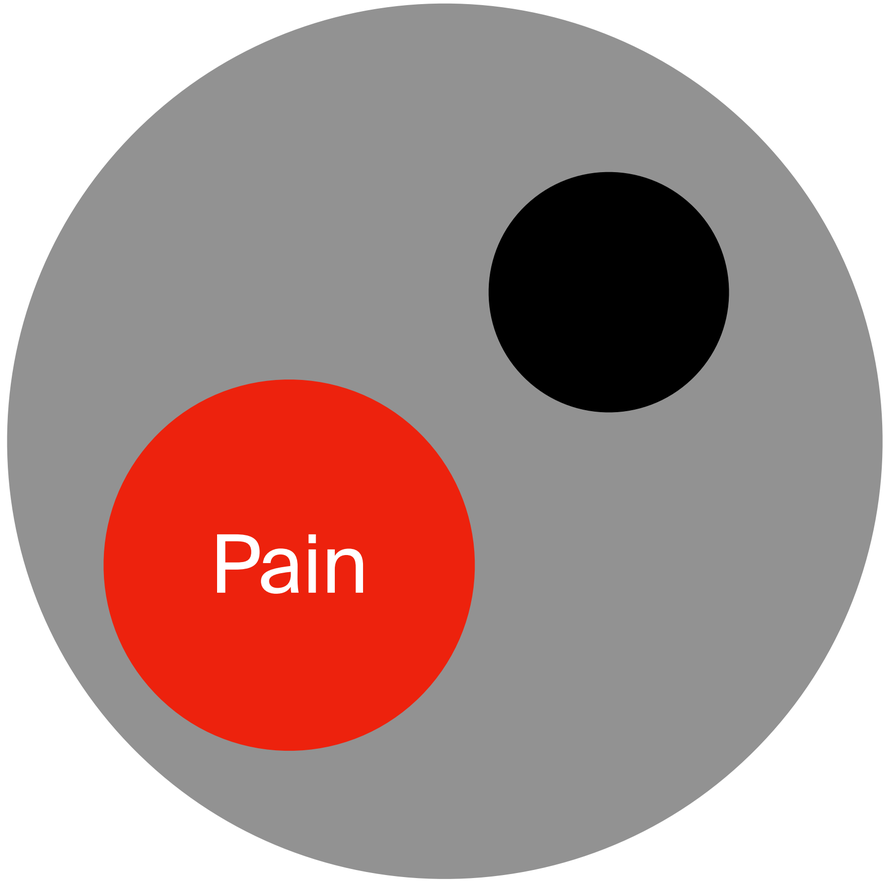
The
Ugly

The

not-so-ugly

Duckling

Seeing Pain

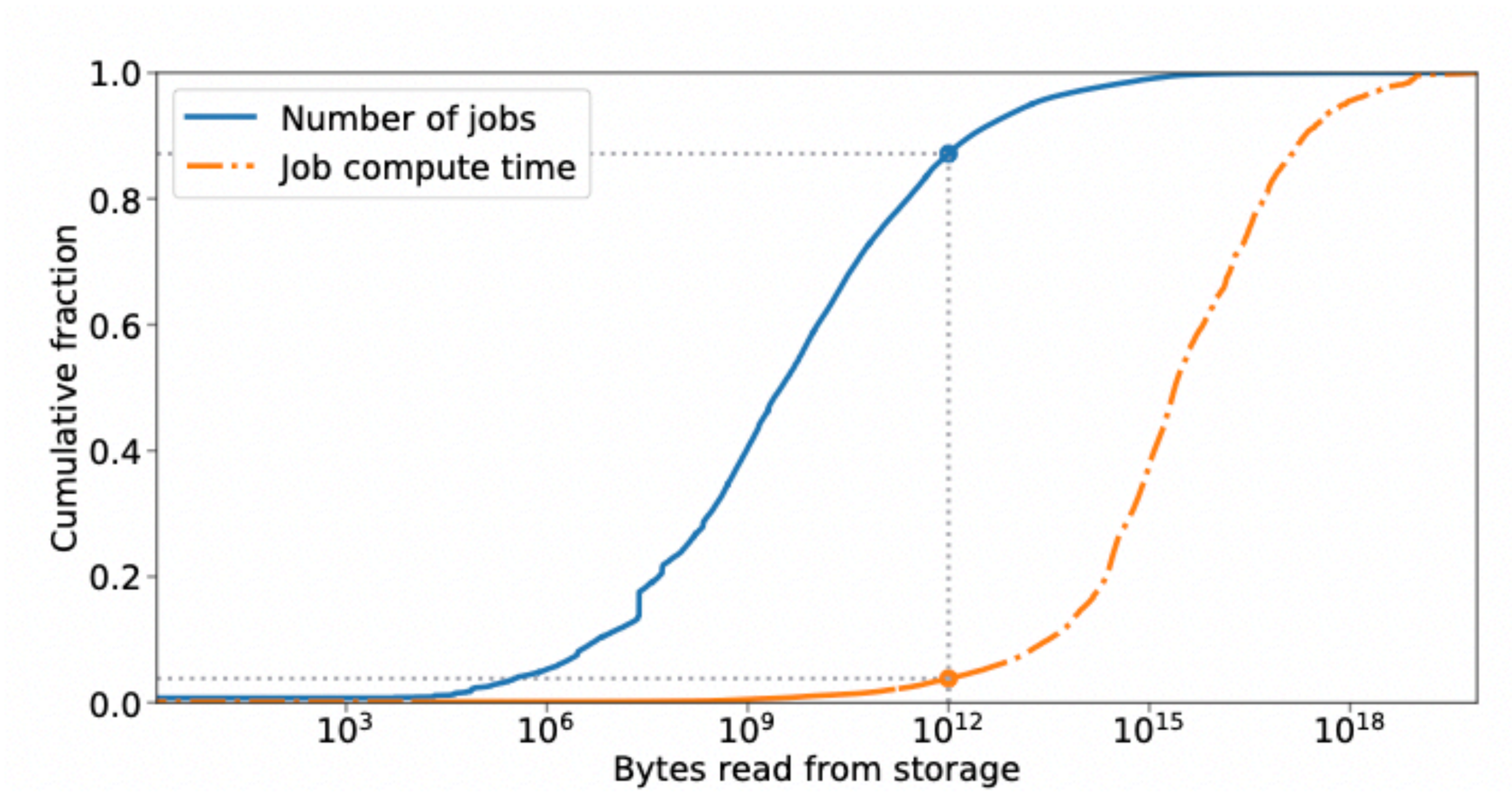

Pain

Treating Pain

Anomaly

Long-Term

Single Node



Source: tf.data: A Machine Learning Data Processing Framework, VLDB 2021


NASDAQ CONGRATULATES
MOTHERDUCK
ON ITS $47.5 M FUNDING
BY A16Z AND REDPOINT
MotherDuck
Nasdaq
Visit HARFORD MARYLAND
VisitHarford.com
MarylandOPEN
Absen
A NOAH BAUMBACH PICTURE
ACADEMY AWARD® NOMINEE GRETA GERWIG
ACADEMY AWARD® NOMINEE DON CHEADLE
H&M
WELCOME TO
Members get early access to

Investing in MotherDuck

by Martin Casado and Jennifer Li

enterprise & SaaS • data infrastructure

Big data's roots are in the hyperscalers. As a result, most tooling we use for analytics today has been built around a scale-out approach where performance is sacrificed for the ability to handle enormous amounts of data. So although as an industry we're very good at building heavy machinery to clean, format, and process petabytes of data, for modestly smaller workloads the tool chain is tremendously complicated, slow, and unfit for purpose. (When we say *modestly smaller*, we're still referring to many terabytes of data, which is adequate for a large swath of analytics use cases.)

Little **Things**

Dependencies

Single File

Parquet

Small* Footprint

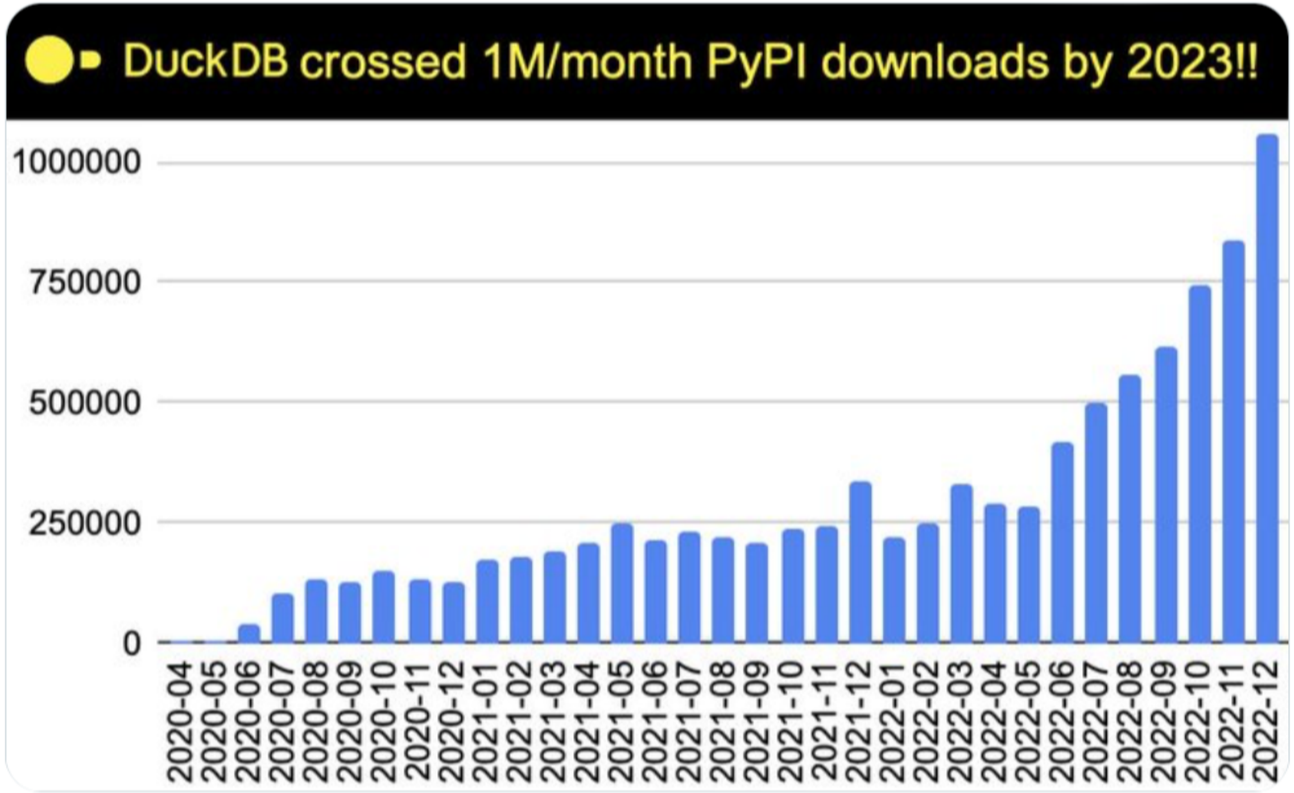
Promotion

DuckDB Labs

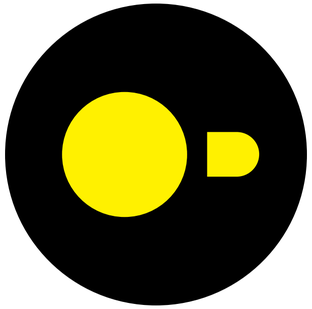
Outcome

Peter Boncz
@peterabcz

Happy 2023, @duckdb team! Congrats on many things, including on exceeding adoption expectations yet again.



6:06 PM · Jan 4, 2023



- Pedro Holanda
- Laurens Kuiper
- Richard Wesley
- Sam Ansmink
- Tania Bogatsch
- Thijs Bruineman
- ...