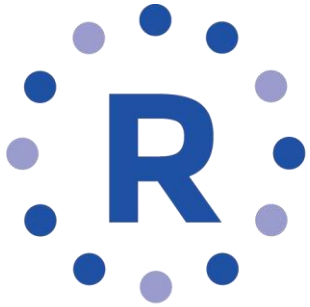


consortium

Part of the new
R/Insurance
Webinar Series

High performance programming in R

(arrow package and Parquet files)

31 January 2024

Welcome to the webinar!

R/insurance webinar series

1) From Excel to programming in R
2) From programming in R to putting R into production
3) R performance culture
4) **High performance programming in R**

Delivered on behalf of the R Consortium by Georgios Bakoloukas and Benedikt Schamberger, Actuarial Control, Group Risk Management, Swiss Re

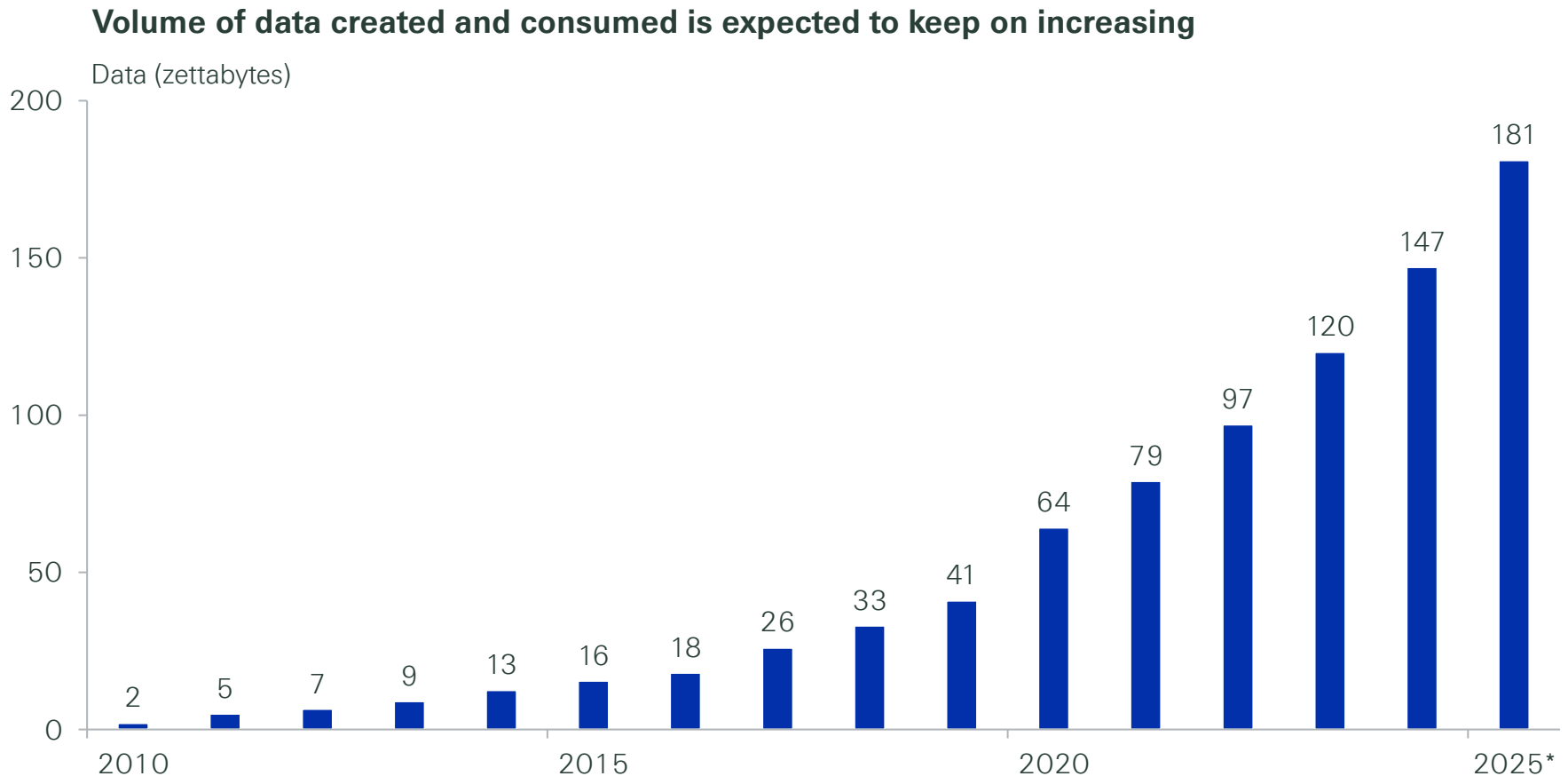
Background to Swiss Re's R community

Large actuarial R programming, Atelier, community

- Swiss Re internal R community sponsored by our Group Chief Actuary Philip Long (Atelier programme)
- 2000+ community with 500+ regular coders who also support each other
- The case we see today relates to code optimisations we did for experience study work, ie comparing how an insurance portfolio performed to initial expectations
- Views expressed belong solely to the speakers and not necessarily to the speaker's employer

Data sizes increasing

Sizes have been growing exponentially and are expected to keep this trend



Source: Statisa: Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025, September 2022

Parquet file format

Uses common compression methods like RLE, dictionaries and other optimisations

Parquet — **Fast and efficient columnar-storage**

Run length encoding (RLE)

Row	Name,...
1	Benedikt,...
2	Benedikt,...
...	...
1,000,000	Benedikt,...

Size
~8,000,000 B

Item	Times
Benedikt	1,000,000

Size
~16 B

Dictionary encoding

Company name
Swiss Reinsurance Company, Ltd.

Size
~31 B

Company name
42

Size
~4 B

Performance optimisation

Row group	Age		Age < 20
1	30 31 27	min: 27 max: 31	min > 20 Skip
2	19 14 17	min: 14 max: 19	max < 20 Search
3	55 41 23	min: 23 max: 55	min > 20 Skip

Examples for illustration, exact sizes and optimisations will depend on your data and options used in creating the parquet file

Moving from CSV to Parquet with arrow

Drop-in replacement for CSV

CSV (with data.table)

Parquet (with arrow)

Reading data

```
my_data <- data.table::fread("my_data.csv")
```

```
my_data <- arrow::read_parquet("my_data.parquet")
```

Writing data

```
data.table::fwrite(my_data, file = "my_data.csv")
```

```
arrow::write_parquet(my_data, sink = "my_data.parquet")
```

Case study: Aggregating exposure information

Regular dplyr


```
df <- fread("my_data.csv") # or read_parquet("my_data.parquet)"

df |>
  filter(calendar_year == 2019) |>
  summarise(
    total_exposure = sum(exposure_amount),
    .by = c("benefit_id", "insured_age_group", "region")
  )
```

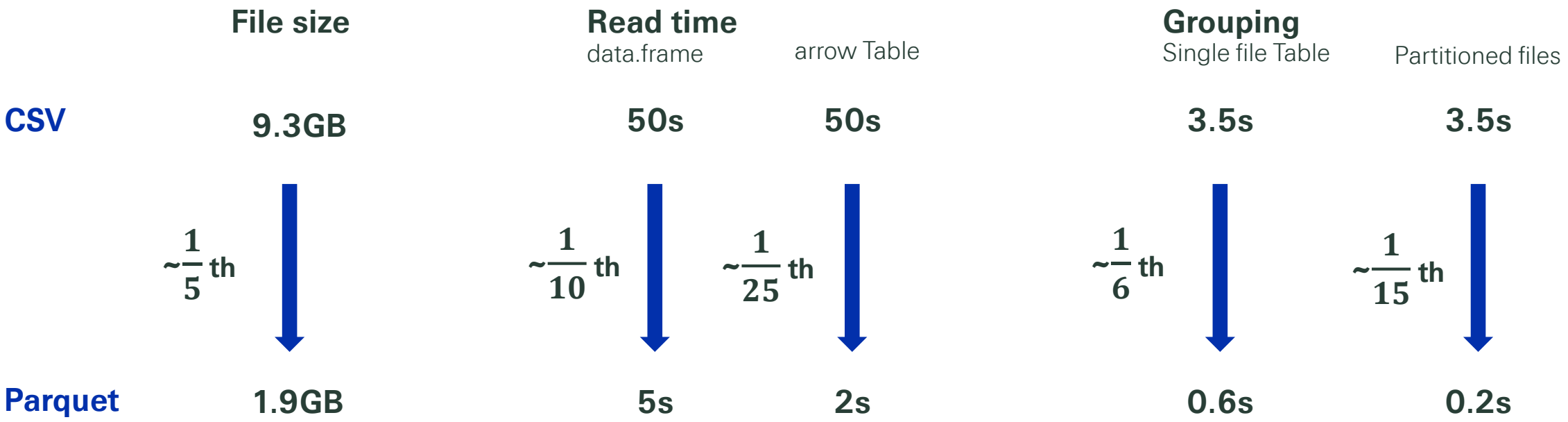
Partitioned parquet


```
df <- open_dataset("my_data") # partitioned parquet folder

df |>
  filter(calendar_year == 2019) |>
  summarise(
    total_exposure = sum(exposure_amount),
    .by = c("benefit_id", "insured_age_group", "region")
  ) |>
 collect()
```

Why switching to Parquet is worthwhile

Smaller files, faster reading and potentially much faster aggregations



Test file: L&H insurance treaty about 24m rows, with 51 columns, size 9.3GB (CSV)

Parquet with arrow is a valuable tool

Give it a try, you can switch back at any time

R Consortium Impact

- R Consortium Community **Grants** and Sponsorships Over **USD $1.4 Million**
- Organize large scale **collaborative projects**
  - R Validation Hub
  - R-Ladies
  - Diversity and Inclusion Working Group
- Co-host multidisciplinary **data science forums**
  - Stanford Data Institute
- Direct support for key **R events**
  - R/Medicine, R/Pharma, useR!, LatinR, more
- Direct support for **R User Groups**

Organizations Can Become a Member Today!

Email Joseph Rickert at **director@r-consortium.org** to set up first call