

Data complete

Various authors

2023-08-22

Table of contents

Introduction	4
1 Data types	5
1.1 Basic data types	5
1.2 Matrices and tensors	5
1.3 Structured and unstructured data	5
1.4 Structure/record/user-defined types	5
1.5 List-like objects	5
1.6 Document-like objects	5
2 Data frames	6
2.1 Normalization/tidy data	6
2.2 Base R	6
2.3 Python	6
2.4 tidyverse	6
2.5 CSV and pitfalls	7
2.6 Parquet	7
3 Queries	8
3.1 Columnar and row-wise subsets.	8
3.2 Logical and ordinal slicing.	8
3.3 Random sampling, stratified sampling.	8
4 Combining data frames	9
4.1 Inner join	9
4.2 Outer joins	9
4.3 Anti-joins	9
4.4 Rolling joins	9
4.5 Union/binding rows	9
4.6 Inserting single rows	9
5 Transforms	10
5.1 Mutations	10
5.2 Aggregation/summarization	10
5.3 Window functions	10
5.4 Pivoting/spreading/unstacking	10

6	EDA	11
6.1	Tabular/query	11
6.2	Visual	11
7	Missing data	12
7.1	Types	12
7.2	How to address it	12
8	Relational database management systems	13
8.1	Pour one out for Edgar Codd	13
8.2	Normalization	13
8.3	ACID, CRUD and all that	13
8.4	Superficial notes about what a DBA does	13
9	NoSQL	14
9.1	Graph databases	14
9.2	Document data	14
10	Big data	15
10.1	Definition, examples	15
10.2	map/reduce	15
10.3	Spark	15
11	ETL, warehouses, etc.	16
12	Conclusion	17
	References	18

Introduction

This book will discuss electronic representations of data. Though this is intended as a practical guide for actuaries working with insurance and risk management data, the principles and techniques may be more general than that.

1 Data types

1.1 Basic data types

How computers represent floating point numbers.

Data type conversion.

1.2 Matrices and tensors

1.3 Structured and unstructured data

1.4 Structure/record/user-defined types

Not quite a thing anymore. More historical/background than practical.

1.5 List-like objects

Heterogeneous, recursive data, JSON/XML

1.6 Document-like objects

key-value pairs

2 Data frames

2.1 Normalization/tidy data

2.2 Base R

```
df_example <- data.frame(  
  col1 = 1:2,  
  col2 = 3:4  
)
```

2.3 Python

```
import pandas as pd  
  
df_example = pd.DataFrame(  
  {'col1': [1, 2], 'col2': [3, 4]}  
)
```

2.4 tidyverse

```
library(tibble)  
  
tbl_example <- tibble(  
  col1 = 1:2,  
  col2 = 3:4  
)
```

2.5 CSV and pitfalls

2.6 Parquet

3 Queries

3.1 Columnar and row-wise subsets.

3.2 Logical and ordinal slicing.

3.3 Random sampling, stratified sampling.

4 Combining data frames

4.1 Inner join

4.2 Outer joins

Also cover semi-joins

4.3 Anti-joins

4.4 Rolling joins

4.5 Union/binding rows

4.6 Inserting single rows

5 Transforms

5.1 Mutations

5.2 Aggregation/summarization

5.3 Window functions

5.4 Pivoting/spreading/unstacking

6 EDA

6.1 Tabular/query

6.2 Visual

7 Missing data

7.1 Types

Missing-at-random, missing-not-at-random, etc.

7.2 How to address it

Talk to data collectors, ignore it, impute it,

8 Relational database management systems

8.1 Pour one out for Edgar Codd

8.2 Normalization

8.3 ACID, CRUD and all that

8.4 Superficial notes about what a DBA does

9 NoSQL

See Luc Perkins and Wilson (2018).

9.1 Graph databases

Neo4j

9.2 Document data

MongoDB

10 Big data

10.1 Definition, examples

Four V's, etc.

10.2 map/reduce

10.3 Spark

11 ETL, warehouses, etc.

Need a succinct title for this chapter. Basic idea is: how do multiple data systems interact?
What is ETL, what the F\$#% is a data lake and so on.

12 Conclusion

Data is hard.

References

Luc Perkins, Eric Redmond, and Jim Wilson. 2018. *Seven Databases in Seven Weeks 2nd Edition a Guide to Modern Databases & the NoSQL Movement*. The Pragmatic Programmers.