MANCHESTER
1824

The University of Manchester

CATHIE MARSH INSTITUTE

Leading the way in
quantitative social science

Columnar Datastores on the Desktop

Questions to answer

- What is a columnar datastore?
- What are the advantages of using one?
- What do I have to do different to use them?
- Where are they available?

What is a columnar datastore?

- Starting with a ‘normal’ relational databases
  - Made up of tables
  - Imagined as being like a spreadsheet - Rows and Columns

	A	B	C	D	E	F	G
1	Date	year	month	day	min	max	hrs_sunshine
2	01/01/2016	2016	1	1	-2	1	2
3	02/01/2016	2016	1	2	3	14	7
4	03/01/2016	2016	1	3	3	9	0
5	04/01/2016	2016	1	4	2	4	10
6	05/01/2016	2016	1	5	-5	1	0
7	06/01/2016	2016	1	6	-2	5	2
8	07/01/2016	2016	1	7	6	14	10

What is a columnar datastore?

- When we add data we add it a Row at a time

Date	year	month	day	min	max	hrs_sunshine
01/01/2016	2016	1	1	-2	1	2
02/01/2016	2016	1	2	3	14	7
03/01/2016	2016	1	3	3	9	0
04/01/2016	2016	1	4	2	4	10
05/01/2016	2016	1	5	-5	1	0
06/01/2016	2016	1	6	-2	5	2
07/01/2016	2016	1	7	6	14	10
			New data goes here			

- The data is stored on disk by rows and read from disk in rows

What is a columnar datastore?

- This approach can be very wasteful of resources. A simple SQL Query of :

SELECT Date, max(Sunshine_hrs)
FROM Weather_date;

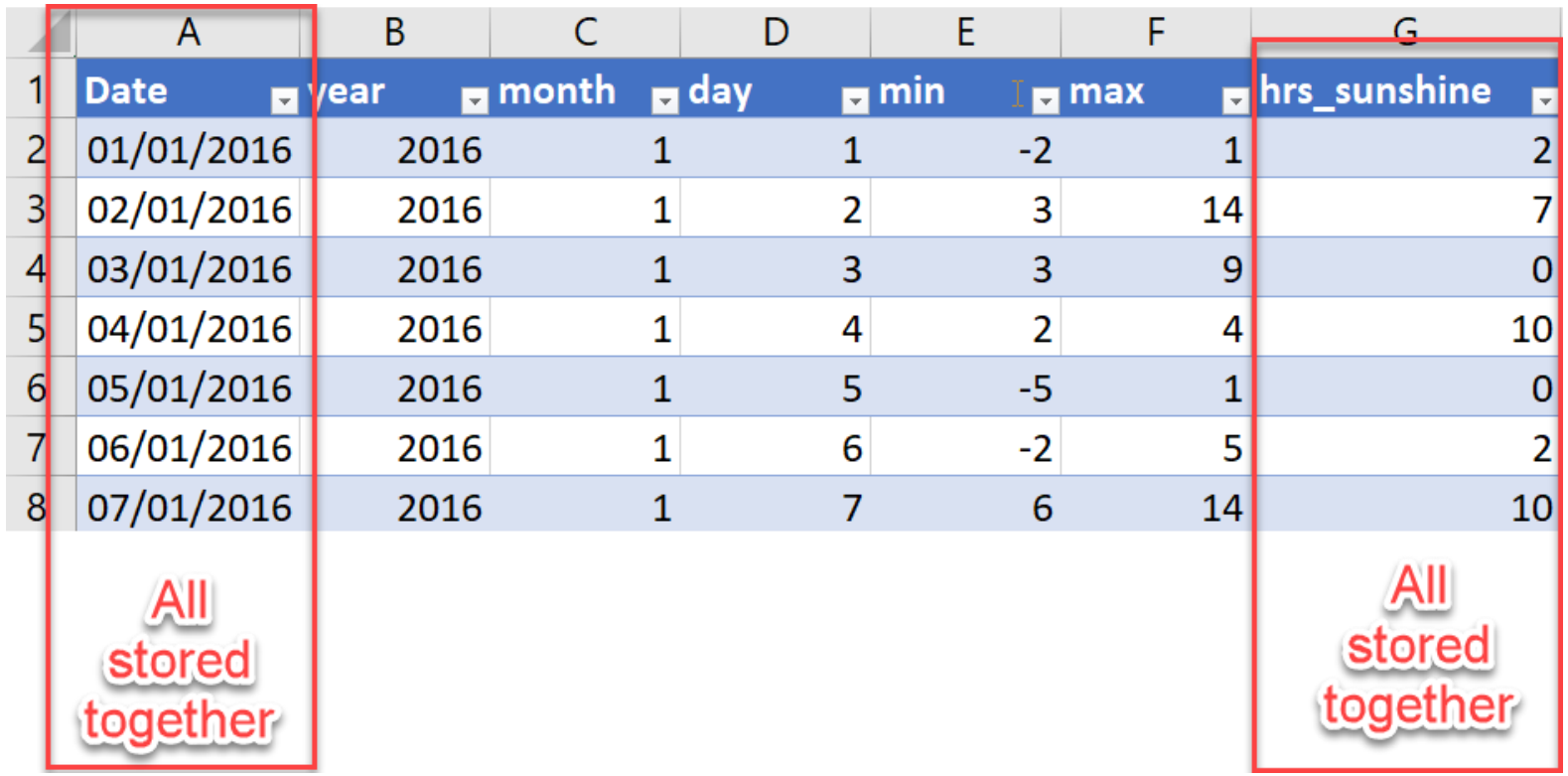
would have to read all of the data to extract the Date and Sunshine_hrs columns to find the Date(s) which have the maximum number of sunshine hours

- We could improve the efficiency of the query by using Indexes, but Indexes increase the data size – and they have to be created in advance. I.e you would have had to anticipated a query of this type.

What is a columnar datastore?

- In a columnar datastore, the data is stored in a different format
- You guessed it! - By columns

What is a columnar datastore?



What are the advantages of using one?

- If you only need to access certain columns, then only those columns need to be read.

```
SELECT Date, max(Sunshine_hrs)
FROM Weather_date;
```

- Because the data is stored by column, each column is effectively its own index, which can improve access times in many cases.

What are the advantages of using one?

	A	B	C	D	E	F	G
1	Date	year	month	day	min	max	hrs_sunshine
2	01/01/2016	2016	1	1	-2	1	2
3	02/01/2016	2016	1	2	3	14	7
4	03/01/2016	2016	1	3	3	9	0
5	04/01/2016	2016	1	4	2	4	10
6	05/01/2016	2016	1	5	-5	1	0
7	06/01/2016	2016	1	6	-2	5	2
8	07/01/2016	2016	1	7	6	14	10

- Notice that the values in the columns have limited and repeating values - something that doesn't typically occur with the values across a row

What are the advantages of using one?

- Because sets of values in columns are often less diverse than in rows, it is possible to compress the data from a column more effectively. This has the effect of being able to keep far more data in memory. The more data we keep in memory, the faster the access and query run times.

What are the advantages of using one?

- Overall Advantages
  - Only read the columns needed
  - Columns are compressed
  - Both of which reduces the amount to memory used – so more data in memory at a time

- Disadvantage
  - Writing new data to a table can be slower

  But we don’t tend to spend our time writing new records – so no need to worry about this.

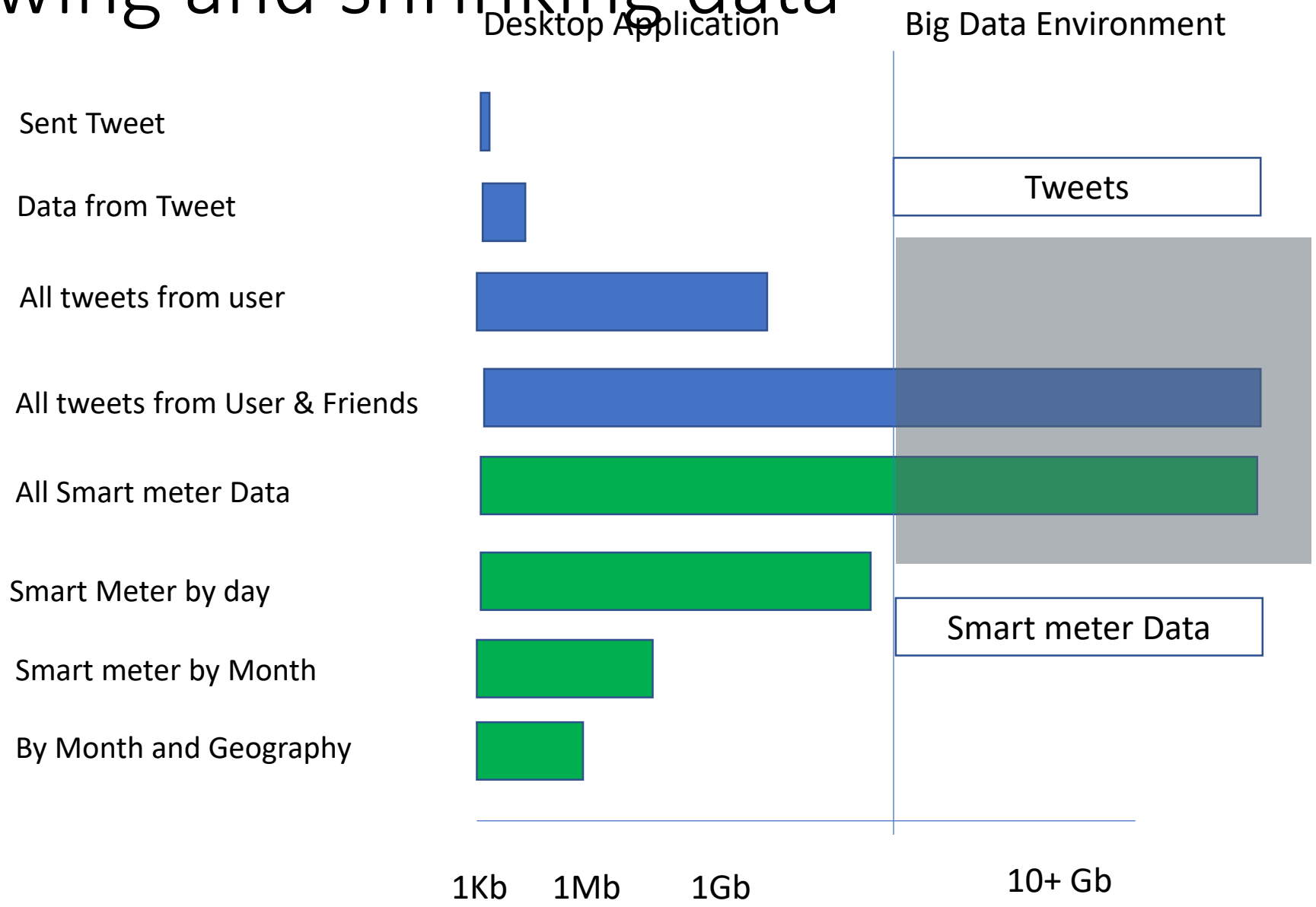
Digression into big data

Where does Big data come from?

- Grow it
- Inherit it (beg, borrow, steal)

- What is the difference between the two ?

Growing and shrinking data



Growing and shrinking data

- The line drawn on the previous slide is somewhat arbitrary.
- For better configured desktops, you would move it to the right.
- For a Desktop with 8Gb of RAM it would probably be at about the 5-6Gb mark.

What stops you processing it on the desktop?

- CPU power? Time? Ram?

- Lack of CPU power just means longer time
- In theory you can take all of the time in the world
- In technical terms it is really only the lack of memory.

- The ability of a columnar database to access the data by columns AND compress the column content means that your RAM can go a lot further.

What do I have to do different to use them?

- The differences between a columnar and non-columnar database system are mainly in the background.
- Accessing data stored in a columnar database is much the same as it is for any other Relational database system – SQL
- For many of the queries that you might write there will be no obvious difference in the code. However when you are accessing very large tables there may be a very discernible difference in the query run time.
- You won’t need indexes for single columns.

Where are they available?

- There are currently a variety of columnar database systems available.

https://www.predictiveanalyticstoday.com/top-wide-columnar-store-databases/ provides a list of the 'Top 9' (There opinion I think)

Top of the list is MariaDB.

- An offshoot of MySQL probably the most popular open source relational database, which means that If you have used MySQL you will feel at home with MariaDB
- Downside is that you need a special install to get the columnar version

Demonstration

Using PowerPivot in Excel 365

The dataset

- The dataset we will use is from the UKDS.
- You can find it as SN 7591 (Energy Demand Research Project: Early Smart Meter Trials, 2007-2010)

- We will use two of the files:
- edrp_gas (7.1 Gb, 246M records)
- edrp_geography (2Mb, 15K records)

edrp_gas

- Only four fields

ANON_ID	ADVANCEDATETIME	HH	GASKWH
12191	18FEB08:08:00:00	16	6.1

edrp_geography

- Has 17 fields but we are only interested in
  - anonid (to allow linking to the gas file)
  - NUTS1 A regional geography field

NUTS1 Regions

From Wikipedia

Hypothesis

- People who use gas, use less in the summer months, regardless of their geographical region

How to test

- Find average of gas usage by region (NUTS1) and month of year for 2009
- Graph the results

Using PowerPivot

- PowerPivot is included with some versions of Excel since Excel 2010
- It uses a columnar storage system allowing it to keep a lot more data in RAM
- PowerPivot tables are not restricted to the (very artificial) limit of 1M rows imposed by standard Excel.
- It provides tools for altering and adding columns to table.
- It can join tables and create Pivot tables based on all of the joined tables

Using PowerPivot

- Points to note
  - The ‘data model’ is stored within an Excel workbook. The max. size of which is 2Gb.
  - Our 7.1Gb file was stored in a file of approx. 575Mb
  - Not available in all versions of Excel
  - There is a free product Power BI Desktop which lets you do similar things.

Using PowerPivot

- Demonstration now!

The End

Any Questions?