Open Data Science
Big Data and Infographics

http://training.theodi.org/BigInfographics

Dr. Dave Tarrant
@davetaz – davetaz@theodi.org

Big Data

Outcomes

Explain the current trends in big data

Identify a number of “big” datasets

Perform a number of short investigations with “big data”

Exercise

What is Big Data to you?

Big Data

Dataset that are too large and complex to manipulate with standard methods or tools.

Excel

Workbook WAS limited to 65,536 rows (2^{16} aka 16-Bit)

64-Bit operating system addressing limit is 2^{64}

18,446,744,073,709,551,615
q q t b m t h

What is big data?

Volume

Velocity

Variety

Veracity

What is big data?

Volume

Velocity

Variety

Veracity

We create around 4 zettabytes of data day.

That’s 1 sextillion bytes per day (128-Bit OS required)

Exercise

6,000,000 rows of data.

Visualise it in 10 minutes…

Significance

Data stays on the web

We “download” the computer.

The computer is a cluster…

Amazon public datasets

Explore the connection here to commodity computing and the volume problem.

Come back to the visualisation later

What is big data?

Volume

Velocity

Variety

Veracity

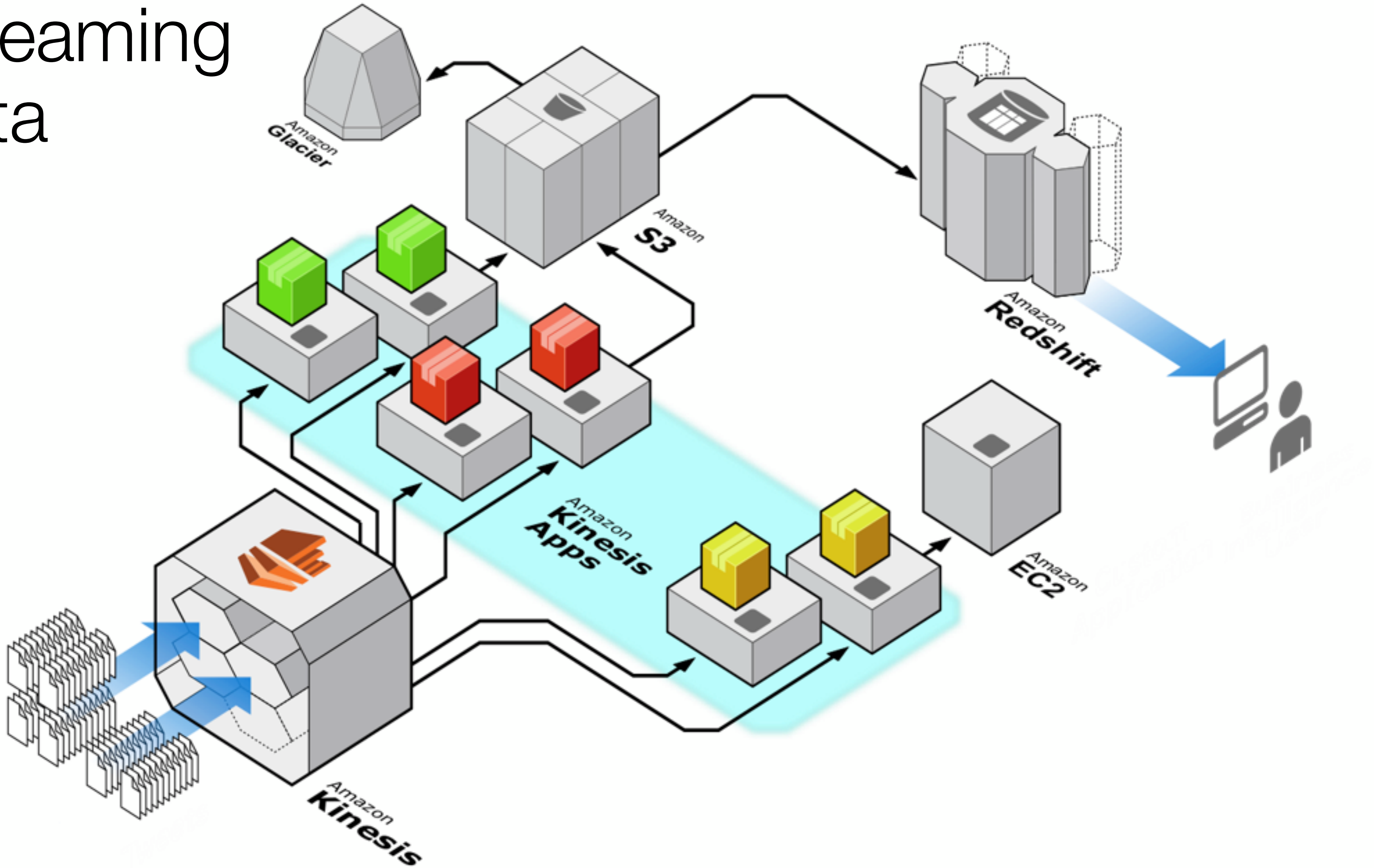
The data is created quicker than we can process it.

Create a stream processor

You are running a chain of Sushi restairants. You want to get a live view on what people are eating, so you fix an RFID tag to the bottom of each bowl on the conveyor in every shop.

How do you get a live summary view of everythng being consumed? Where is the processing power required, and where isn’t it?

Streaming data



What is big data?

Volume

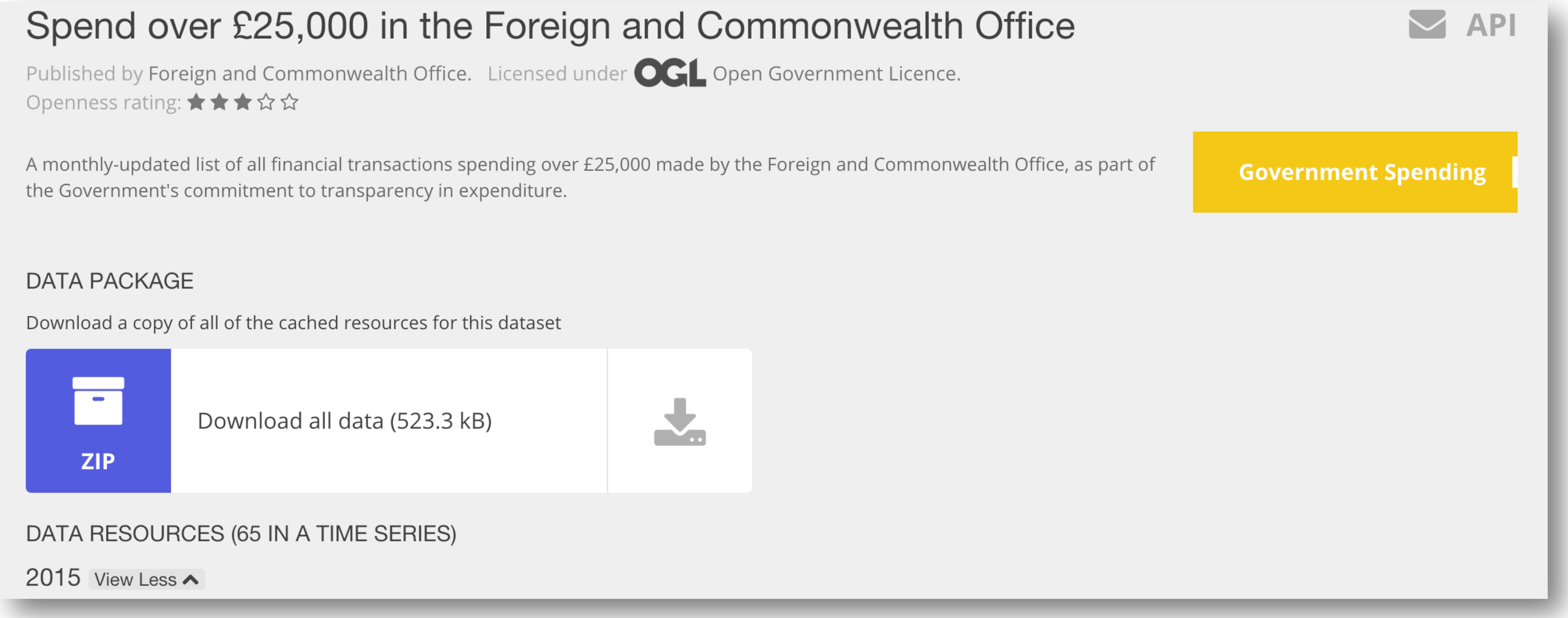
Velocity

Variety

The data is continuously changing in structure, format and detail.

Veracity

Variety in simple data?



Spend over £25,000 in the Foreign and Commonwealth Office

Published by Foreign and Commonwealth Office. Licensed under OGL Open Government Licence.

Openness rating: ★★★☆☆

A monthly-updated list of all financial transactions spending over £25,000 made by the Foreign and Commonwealth Office, as part of the Government's commitment to transparency in expenditure.

DATA PACKAGE

Download a copy of all of the cached resources for this dataset

Download all data (523.3 kB)

DATA RESOURCES (65 IN A TIME SERIES)

2015 View Less

What is big data?

Volume

Velocity

Variety

Veracity

The data quality is highly variable and affected by changing perception of truth and fact.

Big Data

Taken collectively. All digital data is big data. Looking at a facet might reveal that you are looking at a dataset that only conforms to one or two of the **V**s.

A few more V’s

Value and **V**iability

More data does not mean better results.

In fact often entirely the opposite is true.

Sample selection is critical to all good statistic studies.

Not being able to control selection may lead to an incorrect conclusion.

Conclusion

The majority of datasets are large.

Lots of rows with lots of joins that can be processed. If you know how to exploit computing power available.

Scaling

Computing clusters

Injestors
Translators
Indexors
Caches
Validators
Data stores
Visualisers
General purpose units

British Library (in 2008)

BRITISH LIBRARY

80 terabytes of digitised newspapers,

60 terabytes of web-harvested information, e-journals and books,

25 million pages of digitised C19 literature, broadcast television, digital video and digital maps.

GOAL: All 80Tb of newspaper images migrated from TIFF to JP2000

http://www.planets-project.eu/docs/newsletters/Planetarium7_July09.pdf

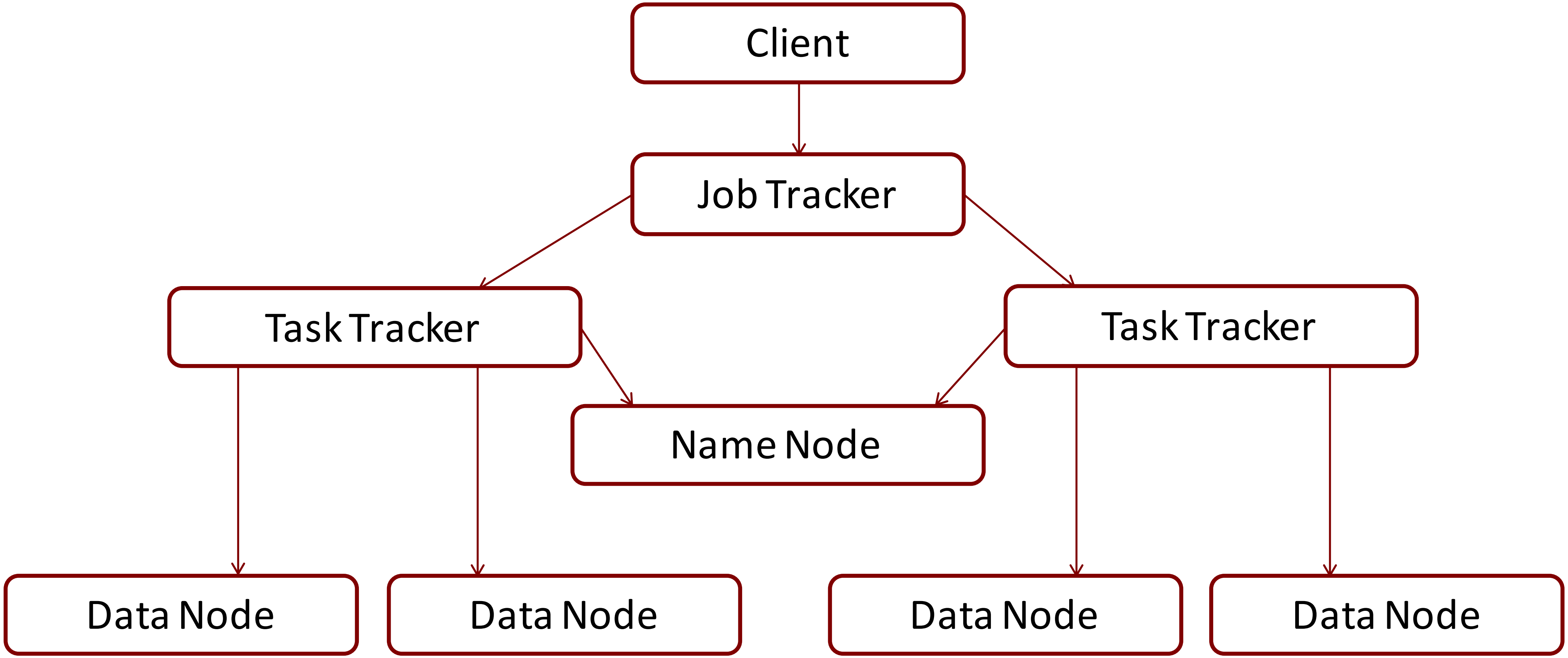
Exercise

Design a system to help the BL with their goal?

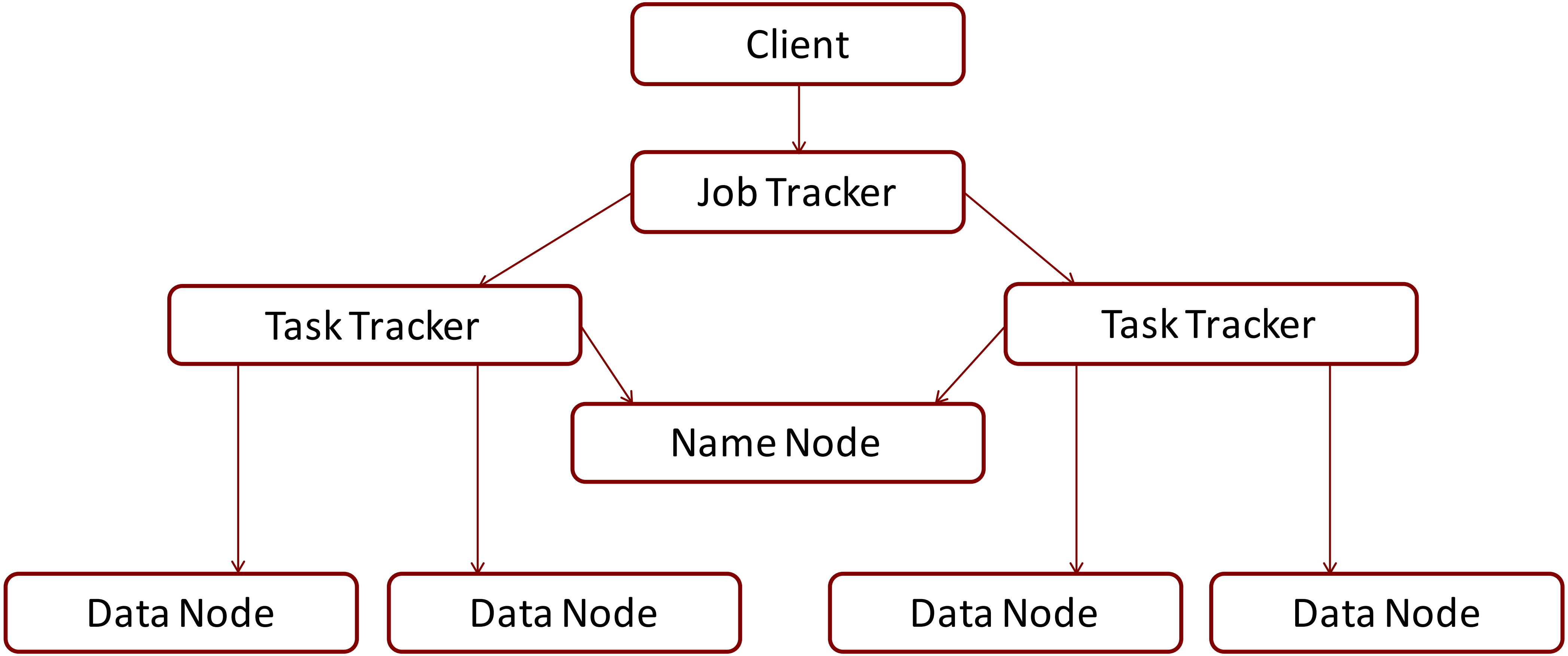
How long will it take?

GOAL: All 80Tb of newspaper images migrated from TIFF to JP2000

Cloud computing



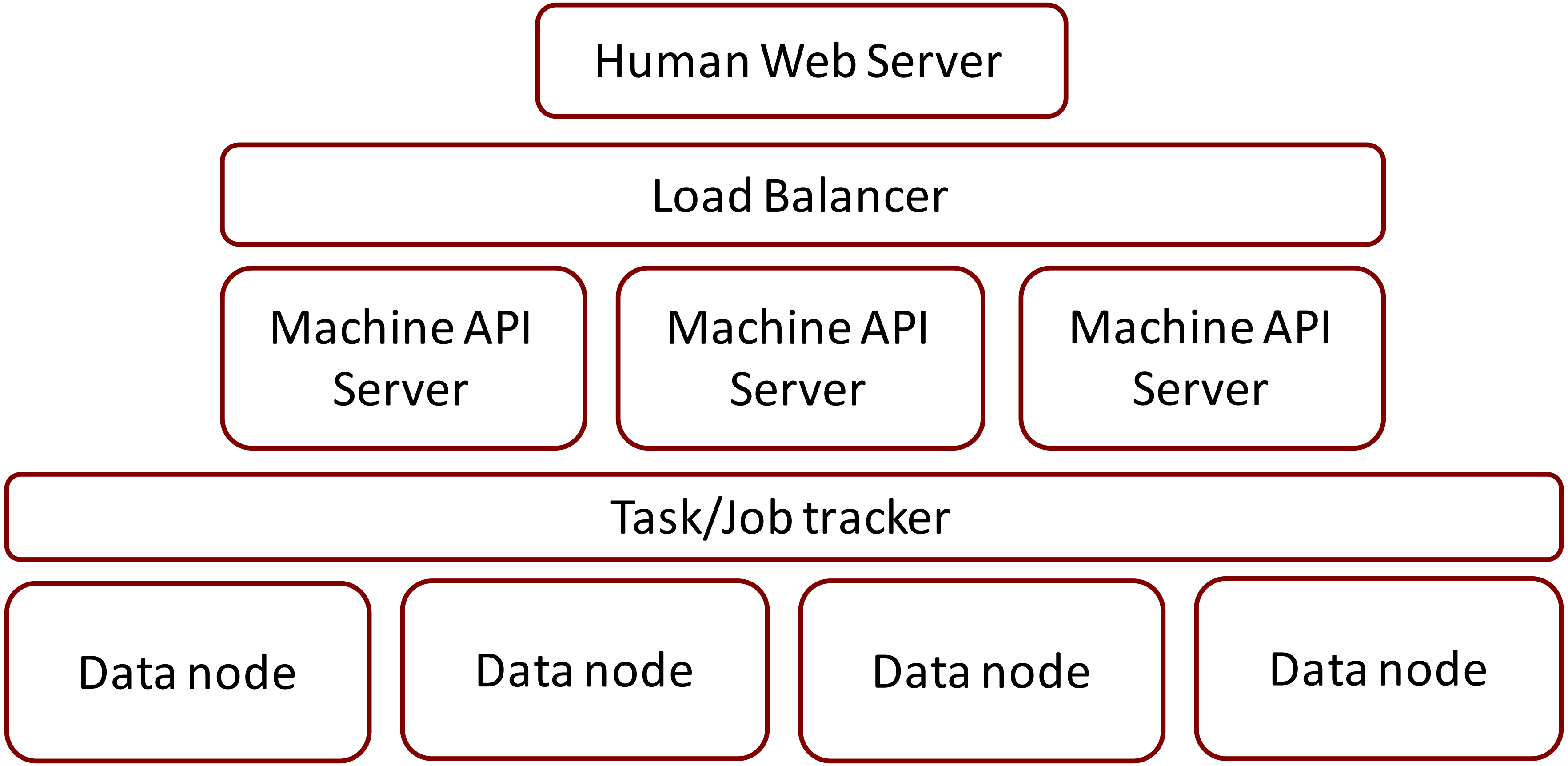
Cloud computing



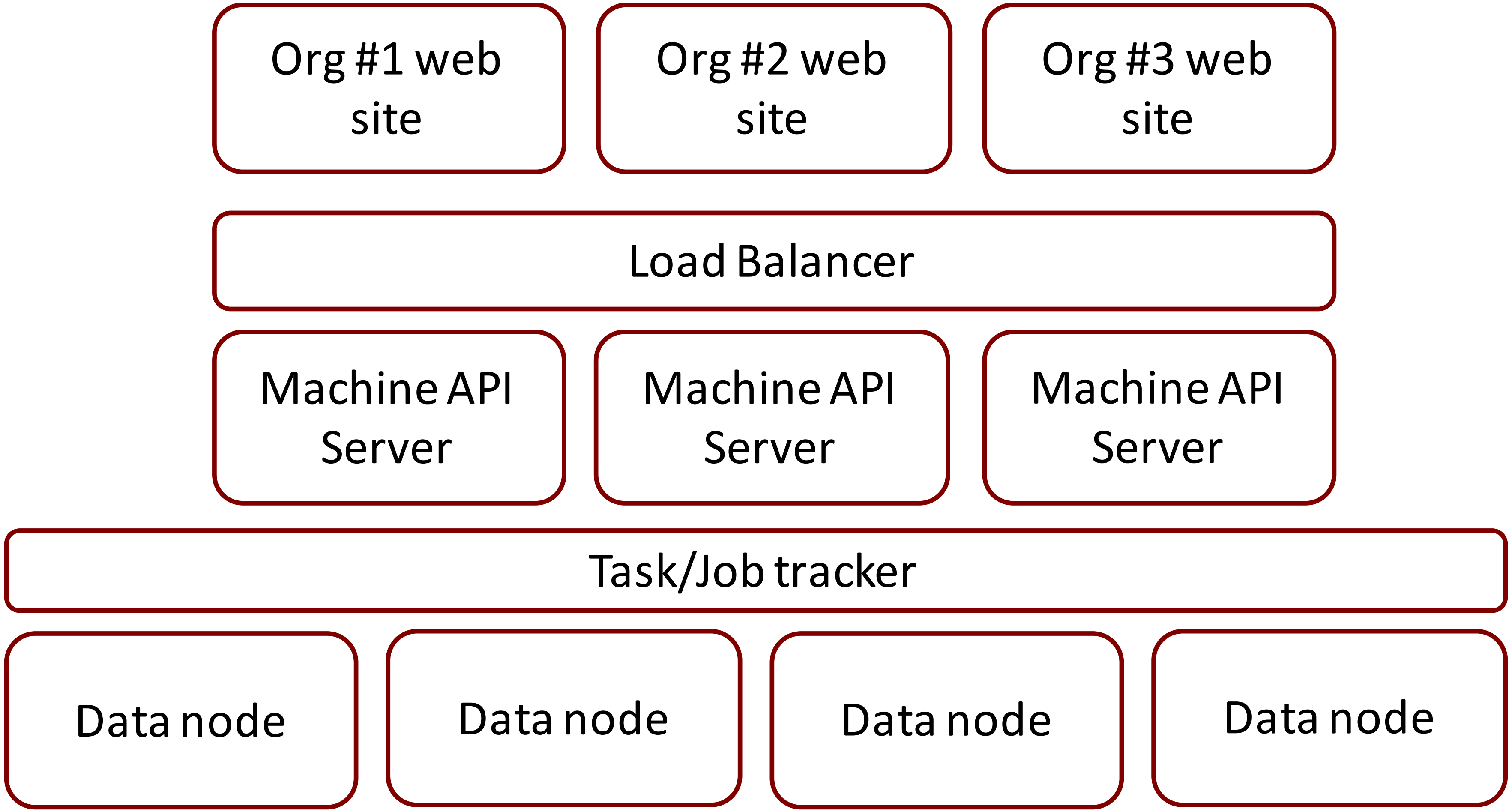
Separate data store from website

API vs Human readable

Open corporates



Socrata model



Databases

Data node

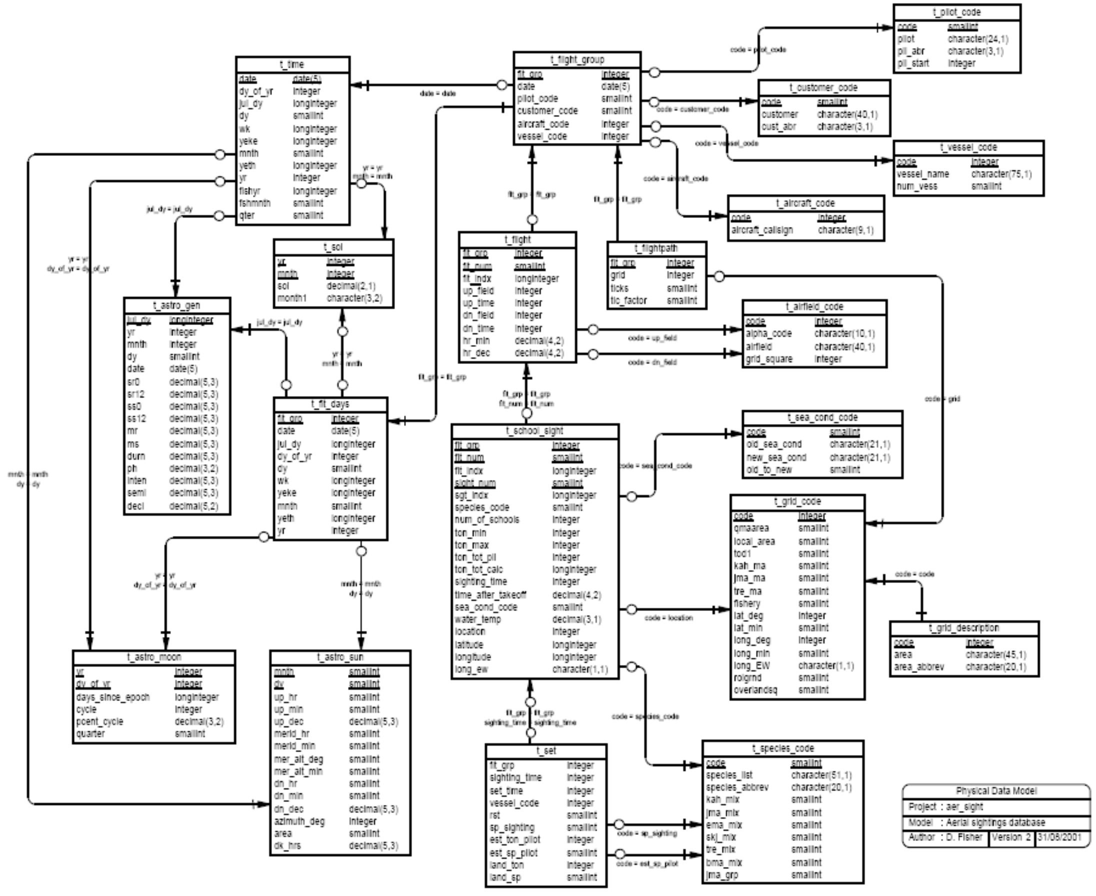


Figure 1: Entity Relationship Diagram (ERD) for the aer_sight database

Flat = fast

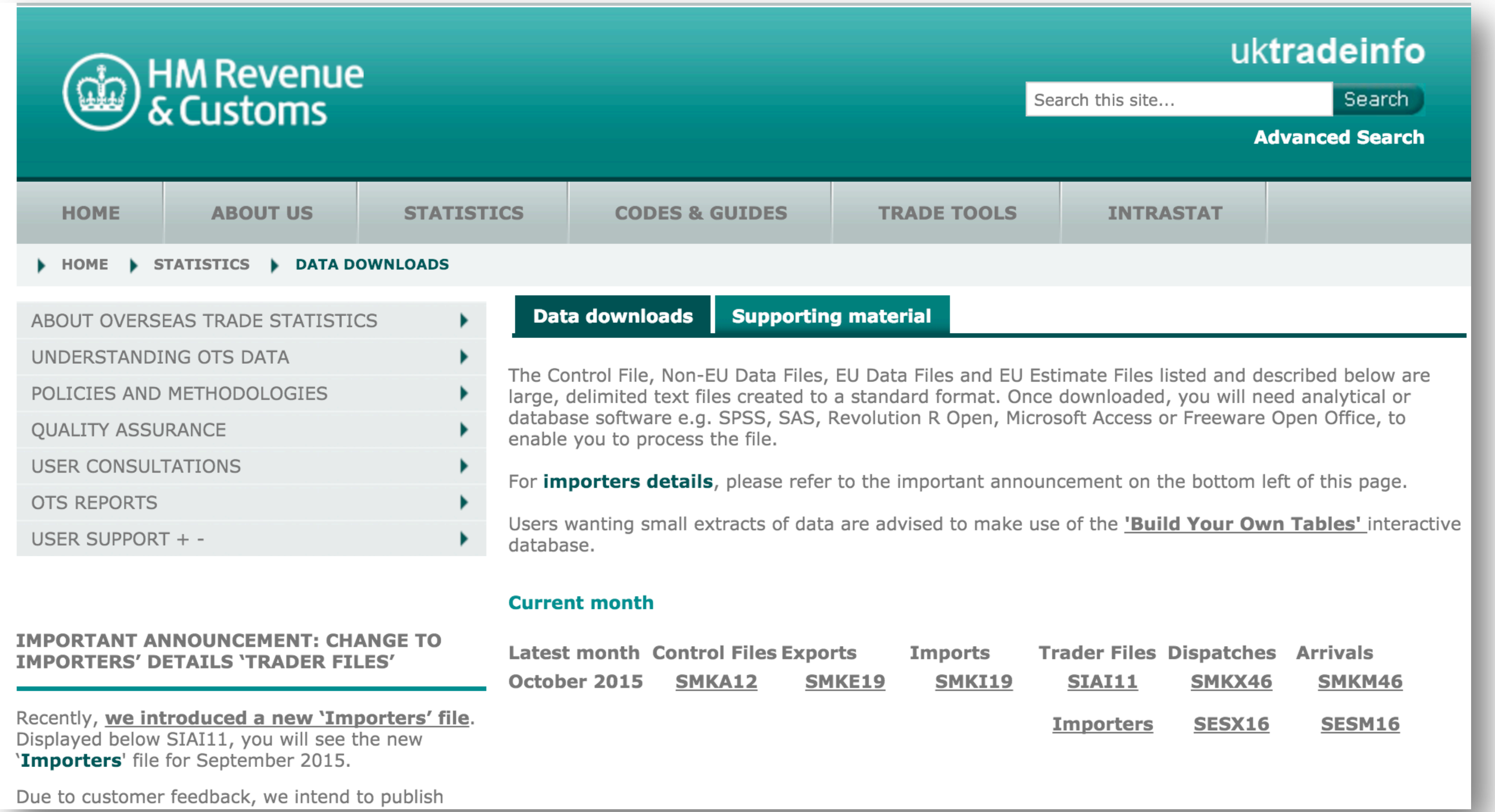
- Something about noSQL and differences

- Using big query…

- But first the dataset..

.

UK Trade data



The Control File, Non-EU Data Files, EU Data Files and EU Estimate Files listed and described below are large, delimited text files created to a standard format. Once downloaded, you will need analytical or database software e.g. SPSS, SAS, Revolution R Open, Microsoft Access or Freeware Open Office, to enable you to process the file.

For **importers details**, please refer to the important announcement on the bottom left of this page.

Users wanting small extracts of data are advised to make use of the **'Build Your Own Tables'** interactive database.

Current month

Latest month	Control Files	Exports	Imports	Trader Files	Dispatches	Arrivals
October 2015	SMKA12	SMKE19	SMKI19	SIAI11	SMKX46	SMKM46
				Importers	SESX16	SESM16

IMPORTANT ANNOUNCEMENT: CHANGE TO IMPORTERS' DETAILS 'TRADER FILES'

Recently, **we introduced a new 'Importers' file**. Displayed below SIAI11, you will see the new '**Importers**' file for September 2015.

Due to customer feedback, we intend to publish

UK Trade Data

Exports

Non-EU

150,000 to 200,000

per month

Imports

Non-EU

190,000 to 220,000

per month

Dispatches

EU

210,000 to 250,000

per month (+estimates)

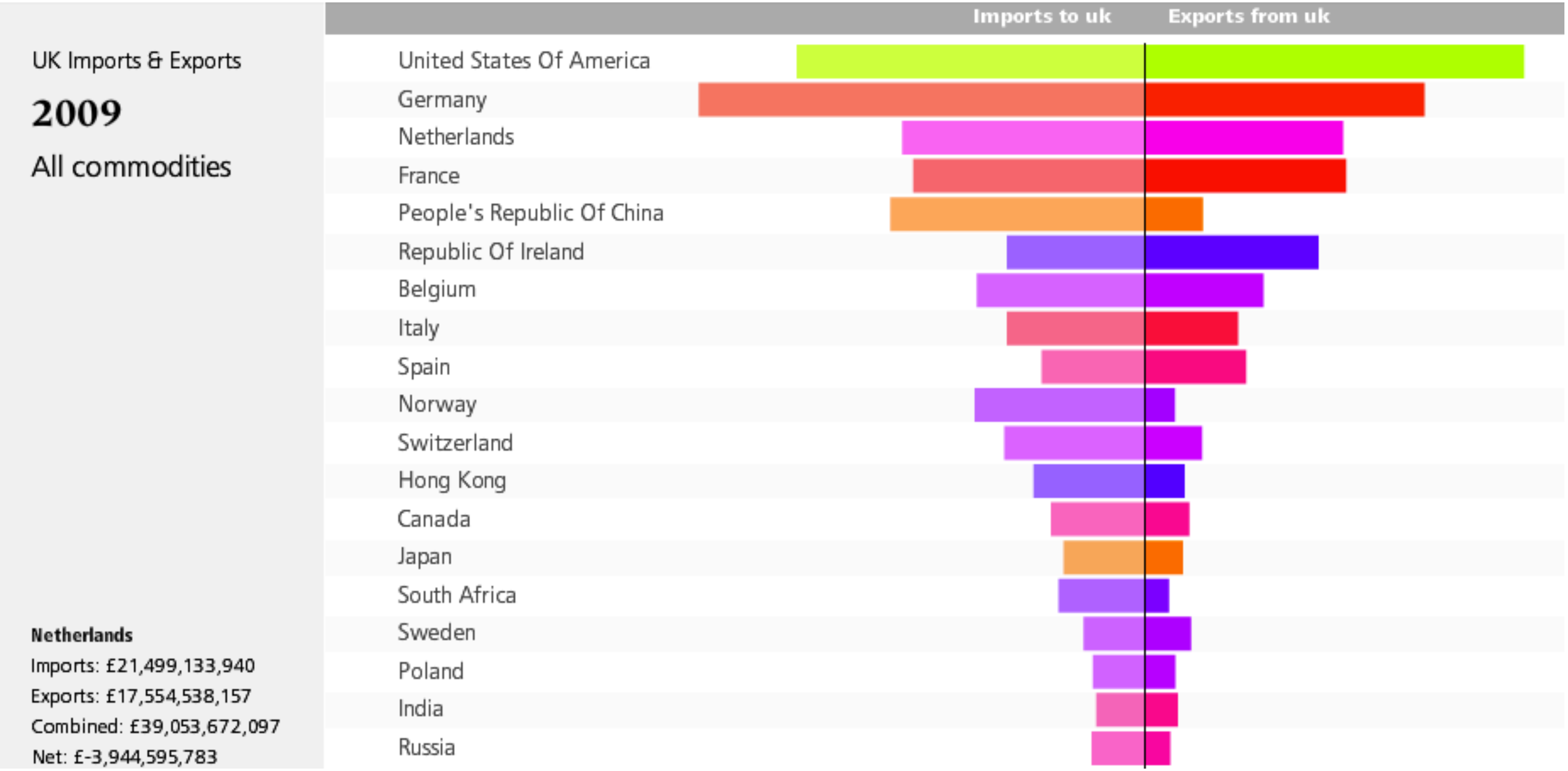
Arrivals

EU

125,000 to 135,000

per month (+estimates)

Distilled information



UK Imports & Exports

2009

All commodities

Netherlands

Imports: £21,499,133,940

Exports: £17,554,538,157

Combined: £39,053,672,097

Net: £-3,944,595,783

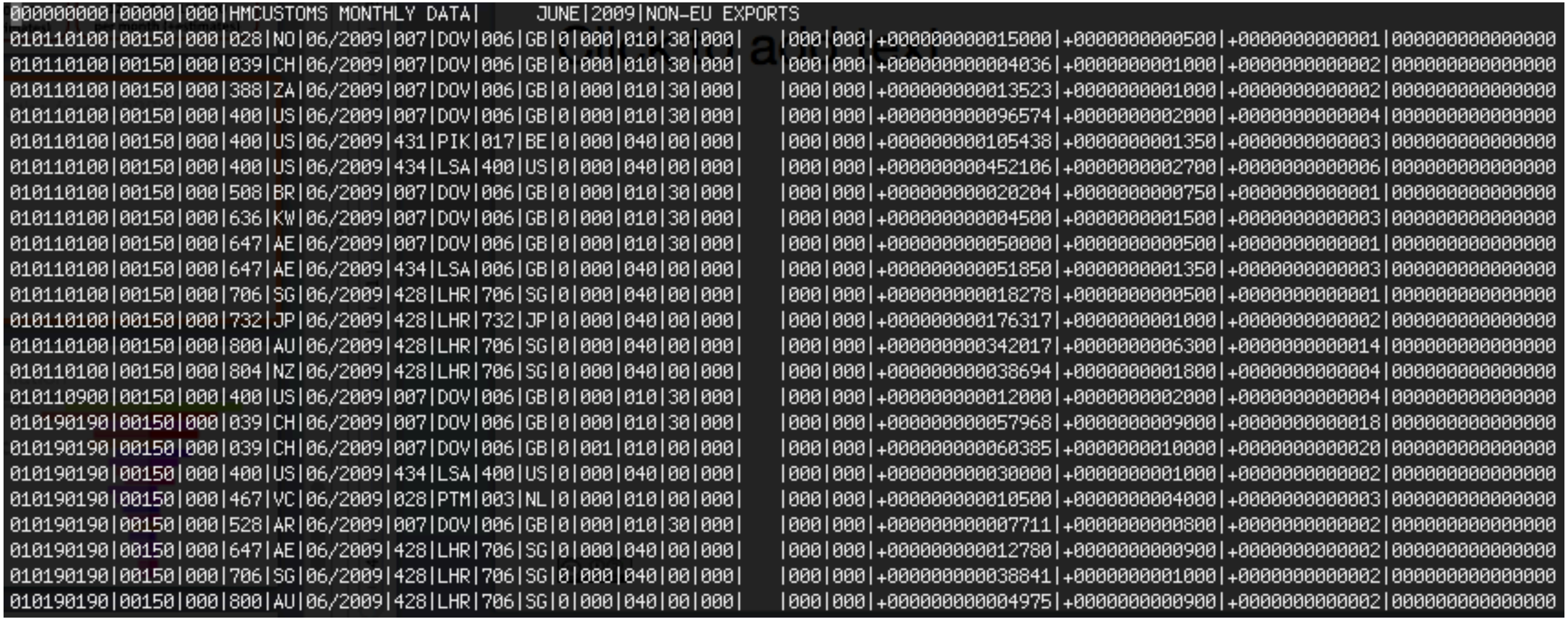
Exercise

Q: How have imports and exports on Jet Engines changed over the years?

Design a processing pipeline that can answer this question from the data.

bit.ly/uk_trade

Stage 1: What the format????



```
000000000|00000|000|HMCUSTOMS MONTHLY DATA|     JUNE|2009|NON-EU EXPORTS
010110100|00150|000|028|NO|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000015000|+0000000000500|+0000000000001|000000000000000
010110100|00150|000|039|CH|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000004036|+0000000001000|+0000000000002|000000000000000
010110100|00150|000|388|ZA|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000013523|+0000000001000|+0000000000002|000000000000000
010110100|00150|000|400|US|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000096574|+0000000002000|+0000000000004|000000000000000
010110100|00150|000|400|US|06/2009|431|PIK|017|BE|0|000|040|00|000|   |000|000|+000000000105438|+0000000001350|+0000000000003|000000000000000
010110100|00150|000|400|US|06/2009|434|LSA|400|US|0|000|040|00|000|   |000|000|+000000000452106|+0000000002700|+0000000000006|000000000000000
010110100|00150|000|508|BR|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000020204|+0000000000750|+0000000000001|000000000000000
010110100|00150|000|636|KW|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000004500|+0000000001500|+0000000000003|000000000000000
010110100|00150|000|647|AE|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000050000|+0000000000500|+0000000000001|000000000000000
010110100|00150|000|647|AE|06/2009|434|LSA|006|GB|0|000|040|00|000|   |000|000|+000000000051850|+0000000001350|+0000000000003|000000000000000
010110100|00150|000|706|SG|06/2009|428|LHR|706|SG|0|000|040|00|000|   |000|000|+000000000018278|+0000000000500|+0000000000001|000000000000000
010110100|00150|000|732|JP|06/2009|428|LHR|732|JP|0|000|040|00|000|   |000|000|+000000000176317|+0000000001000|+0000000000002|000000000000000
010110100|00150|000|800|AU|06/2009|428|LHR|706|SG|0|000|040|00|000|   |000|000|+000000000342017|+0000000006300|+0000000000014|000000000000000
010110100|00150|000|804|NZ|06/2009|428|LHR|706|SG|0|000|040|00|000|   |000|000|+000000000038694|+0000000001800|+0000000000004|000000000000000
010110900|00150|000|400|US|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000012000|+0000000002000|+0000000000004|000000000000000
010190190|00150|000|039|CH|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000057968|+0000000009000|+0000000000018|000000000000000
010190190|00150|000|039|CH|06/2009|007|DOV|006|GB|0|001|010|00|000|   |000|000|+000000000060385|+0000000010000|+0000000000020|000000000000000
010190190|00150|000|400|US|06/2009|434|LSA|400|US|0|000|040|00|000|   |000|000|+000000000030000|+0000000001000|+0000000000002|000000000000000
010190190|00150|000|467|VC|06/2009|028|PTM|003|NL|0|000|010|00|000|   |000|000|+000000000010500|+0000000004000|+0000000000003|000000000000000
010190190|00150|000|528|AR|06/2009|007|DOV|006|GB|0|000|010|30|000|   |000|000|+000000000007711|+0000000000800|+0000000000002|000000000000000
010190190|00150|000|647|AE|06/2009|428|LHR|706|SG|0|000|040|00|000|   |000|000|+000000000012780|+0000000000900|+0000000000002|000000000000000
010190190|00150|000|706|SG|06/2009|428|LHR|706|SG|0|000|040|00|000|   |000|000|+000000000038841|+0000000001000|+0000000000002|000000000000000
010190190|00150|000|800|AU|06/2009|428|LHR|706|SG|0|000|040|00|000|   |000|000|+000000000004975|+0000000000900|+0000000000002|000000000000000
```

Stage 2: RTFM

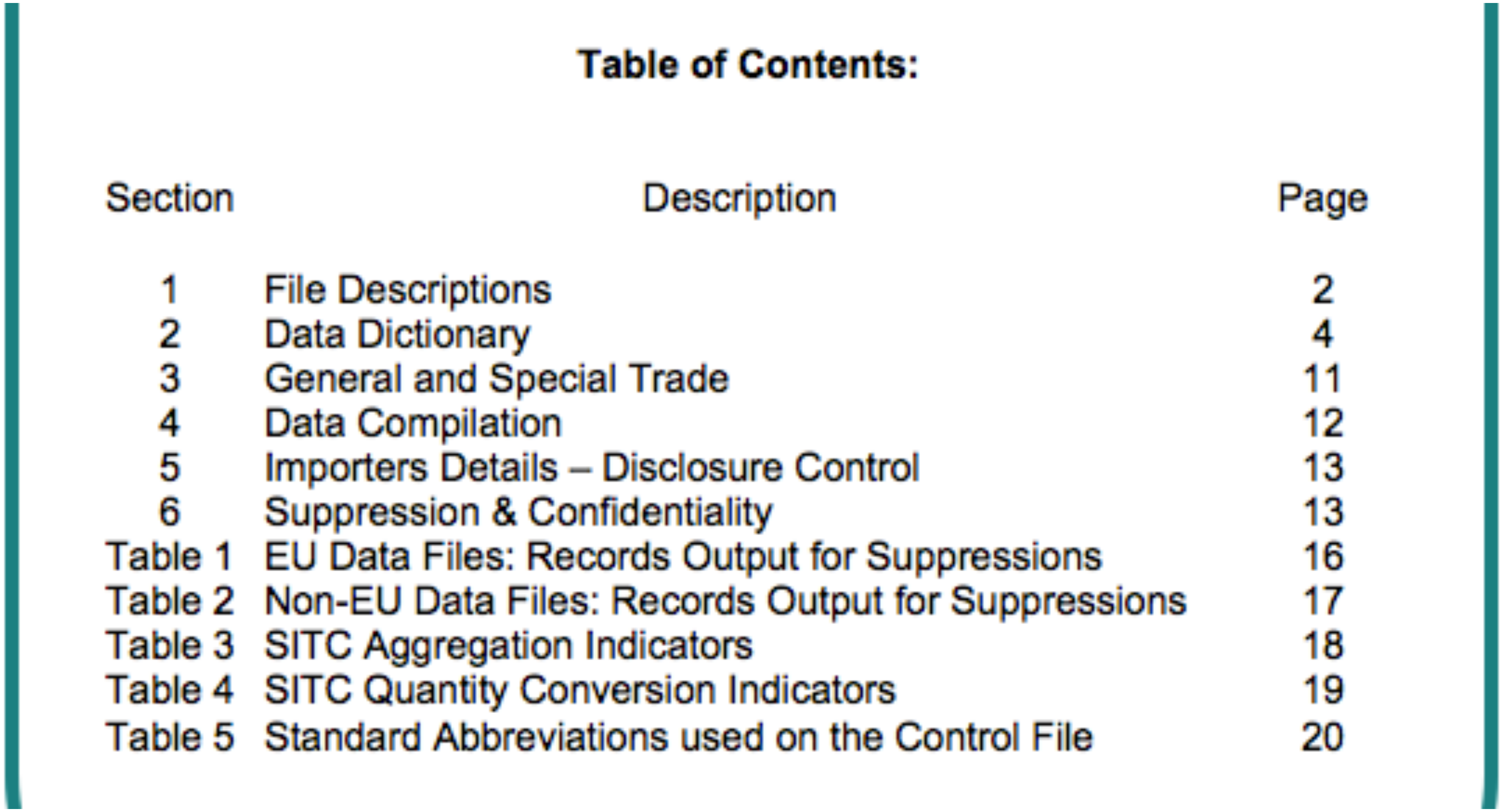
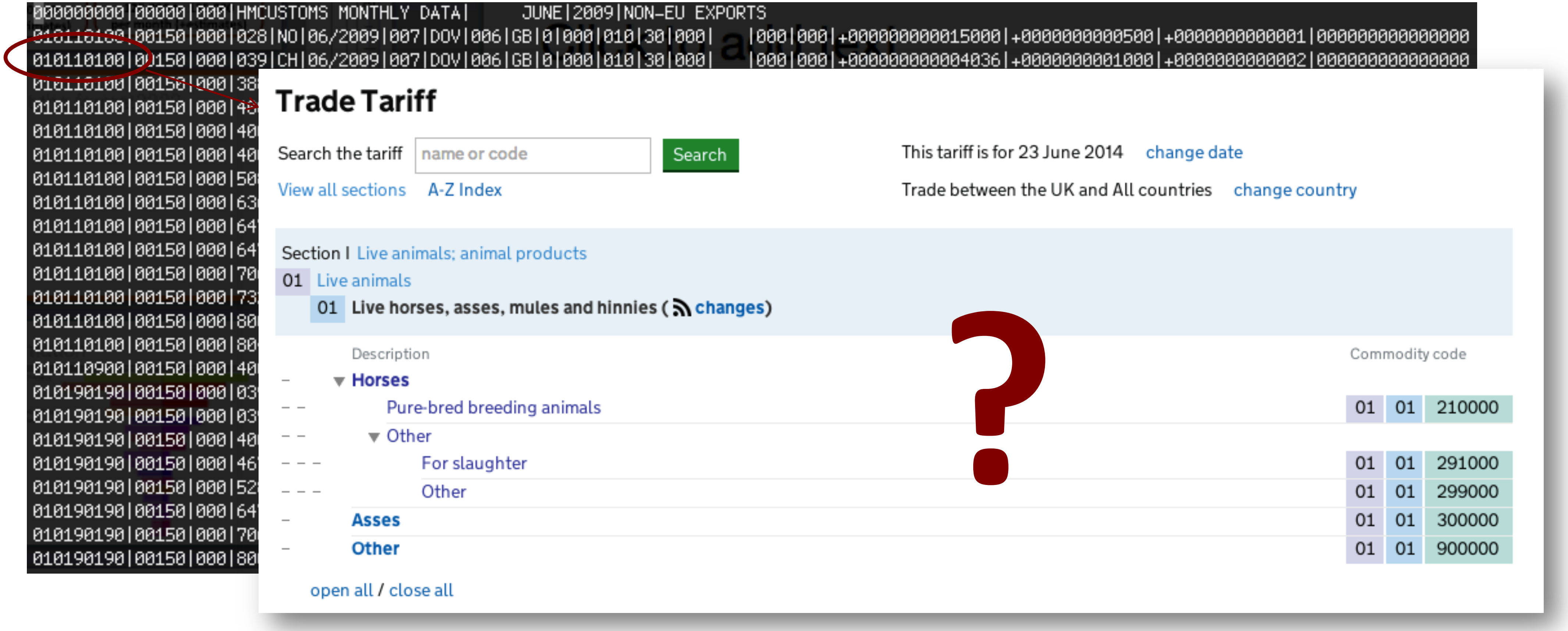


Table of Contents:

Stage 3: Decode

010110100



```
000000000|00000|000|HMCUSTOMS MONTHLY DATA|     JUNE|2009|NON-EU EXPORTS
010110100|00150|000|028|NO|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+000000000015000|+0000000000500|+0000000000001|000000000000000
010110100|00150|000|039|CH|06/2009|007|DOV|006|GB|0|000|010|30|000|    |000|000|+000000000004036|+0000000001000|+0000000000002|000000000000000
010110100|00150|000|38
010110100|00150|000|4
010110100|00150|000|40
010110100|00150|000|40
010110100|00150|000|50
010110100|00150|000|63
010110100|00150|000|64
010110100|00150|000|64
010110100|00150|000|70
010110100|00150|000|73
010110100|00150|000|80
010110100|00150|000|80
010110900|00150|000|40
010190190|00150|000|03
010190190|00150|000|03
010190190|00150|000|40
010190190|00150|000|46
010190190|00150|000|52
010190190|00150|000|64
010190190|00150|000|70
010190190|00150|000|80
```

Trade Tariff

Search the tariff: name or code — Search

This tariff is for 23 June 2014 — change date

View all sections — A-Z Index

Trade between the UK and All countries — change country

Section I Live animals; animal products

01 Live animals

01 01 Live horses, asses, mules and hinnies (changes)

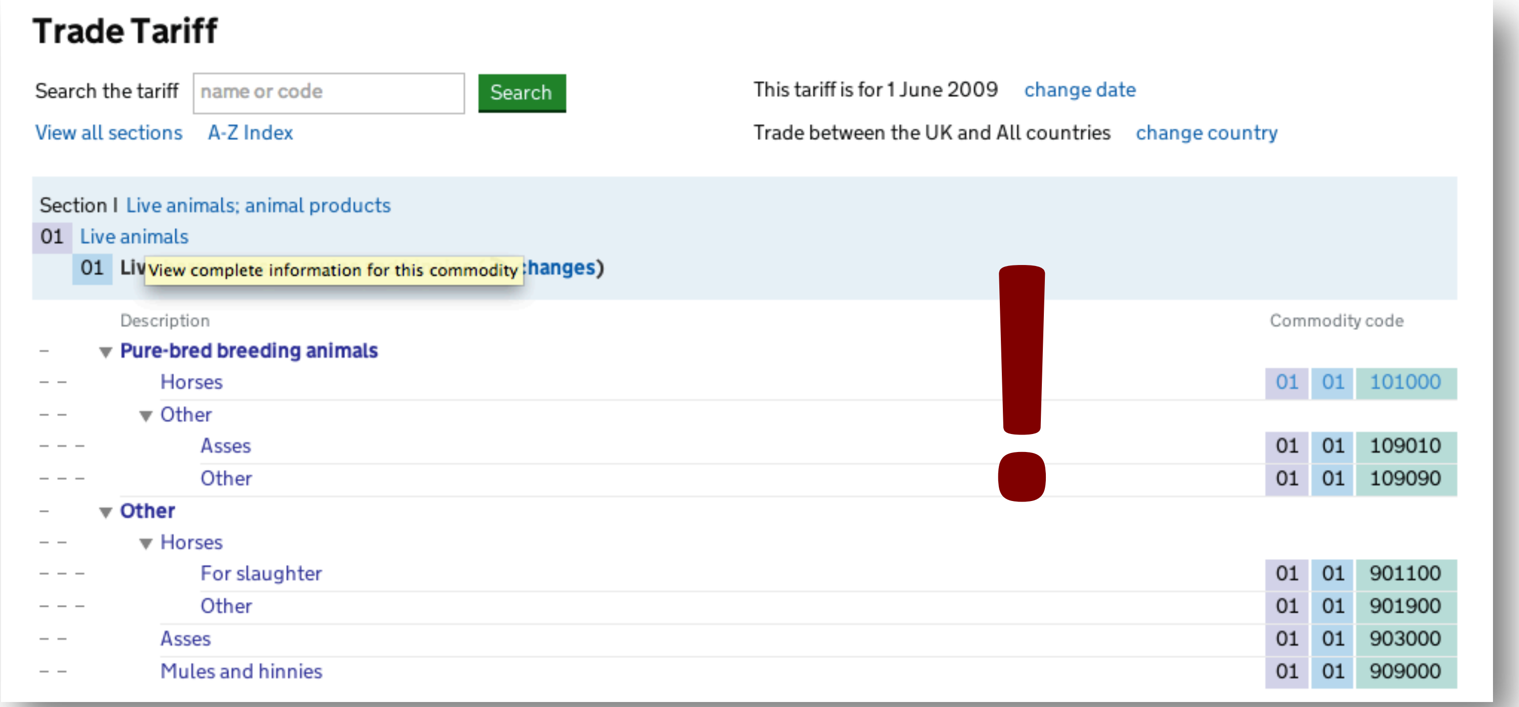
	Description	Commodity code
–	▼ Horses	
– –	Pure-bred breeding animals	01 01 210000
– –	▼ Other	
– – –	For slaughter	01 01 291000
– – –	Other	01 01 299000
–	Asses	01 01 300000
–	Other	01 01 900000

open all / close all

Stage 3b: API?

010110100

`https://www.gov.uk/trade-tariff/headings/0101?country=&day=1&month=6&year=2009`



Trade Tariff

Search the tariff [name or code] [Search]

View all sections A-Z Index

This tariff is for 1 June 2009 change date

Trade between the UK and All countries change country

Section I Live animals; animal products

01 Live animals

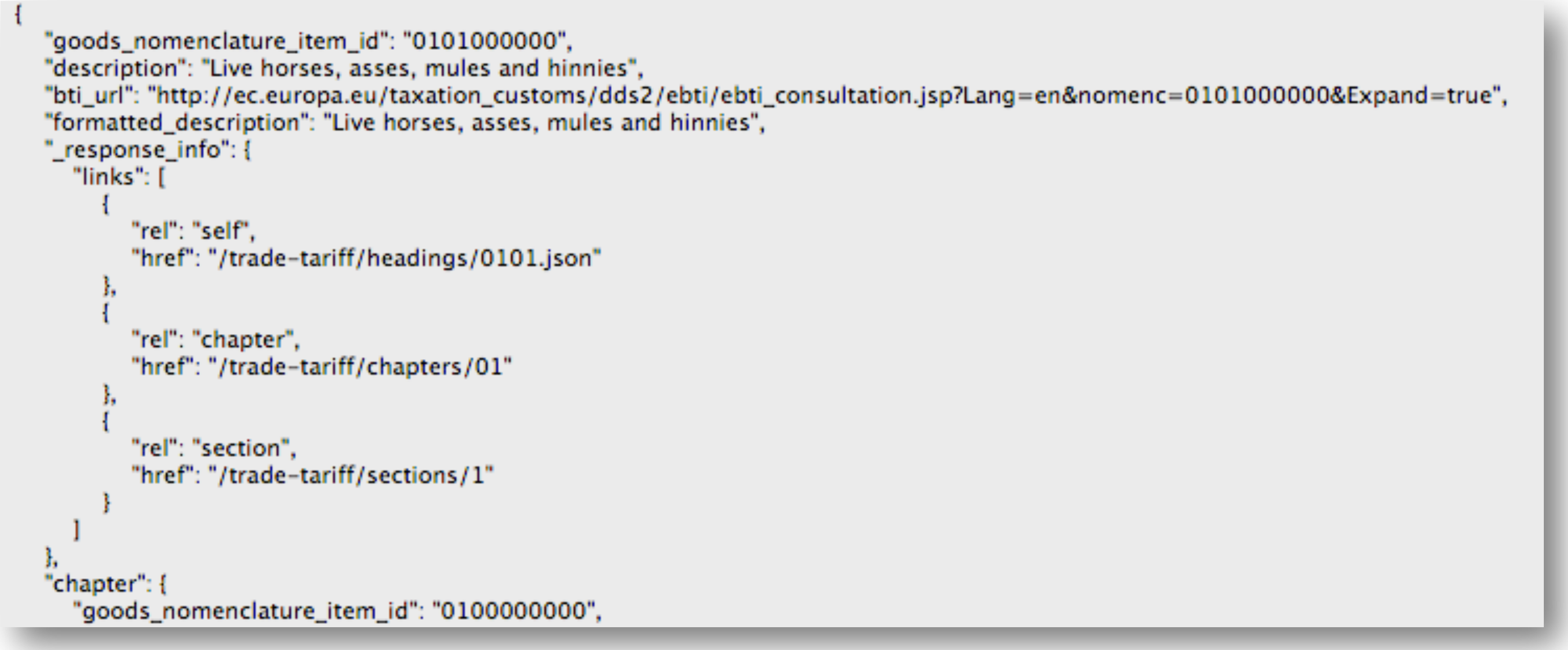
01 Liv View complete information for this commodity hanges)

	Description	Commodity code		
–	▼ **Pure-bred breeding animals**			
– –	Horses	01	01	101000
– –	▼ Other			
– – –	Asses	01	01	109010
– – –	Other	01	01	109090
–	▼ **Other**			
– –	▼ Horses			
– – –	For slaughter	01	01	901100
– – –	Other	01	01	901900
– –	Asses	01	01	903000
– –	Mules and hinnies	01	01	909000

The codes for the same things have changed. Meaning that we have to compare the text! Ahhh.

Stage 4: API for data?

`https://www.gov.uk/trade-tariff/headings/0101.json?country=&day=1&month=6&year=2009`



```
{
    "goods_nomenclature_item_id": "0101000000",
    "description": "Live horses, asses, mules and hinnies",
    "bti_url": "http://ec.europa.eu/taxation_customs/dds2/ebti/ebti_consultation.jsp?Lang=en&nomenc=0101000000&Expand=true",
    "formatted_description": "Live horses, asses, mules and hinnies",
    "_response_info": {
        "links": [
            {
                "rel": "self",
                "href": "/trade-tariff/headings/0101.json"
            },
            {
                "rel": "chapter",
                "href": "/trade-tariff/chapters/01"
            },
            {
                "rel": "section",
                "href": "/trade-tariff/sections/1"
            }
        ]
    },
    "chapter": {
        "goods_nomenclature_item_id": "0100000000",
```

Stage 5: Predict scale

(12 * 4) files per year

12 Comcode tables

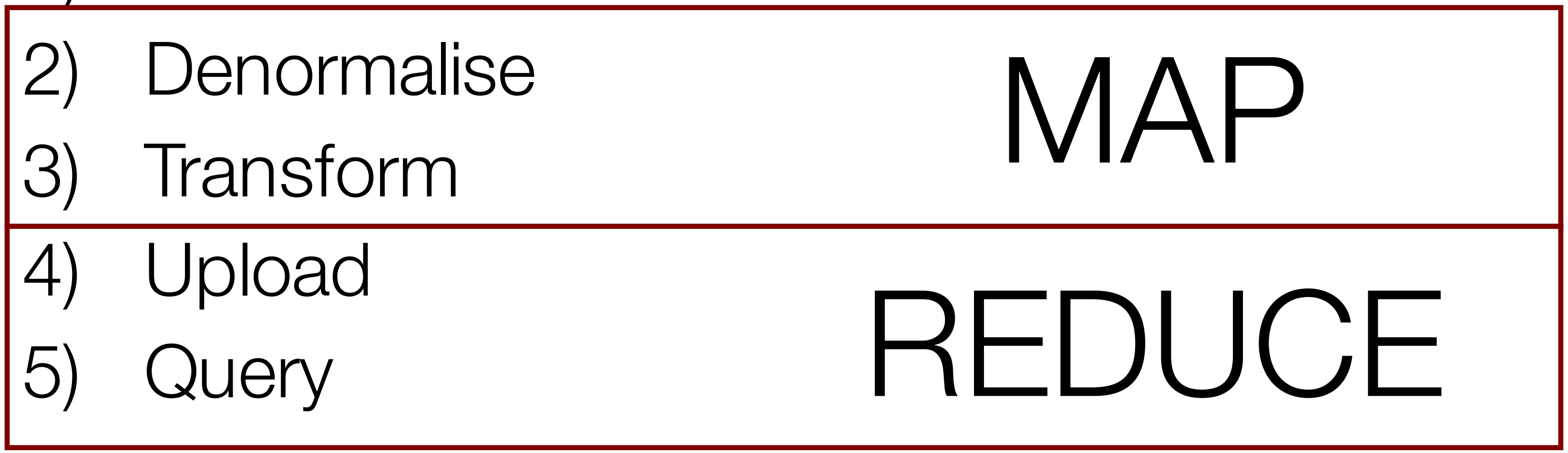
12 Portcode tables

To answer one query you may have to join 48 tables to 24 others to answer it.

This is not how map reduce and big data work.

Large databases

1) Extract data



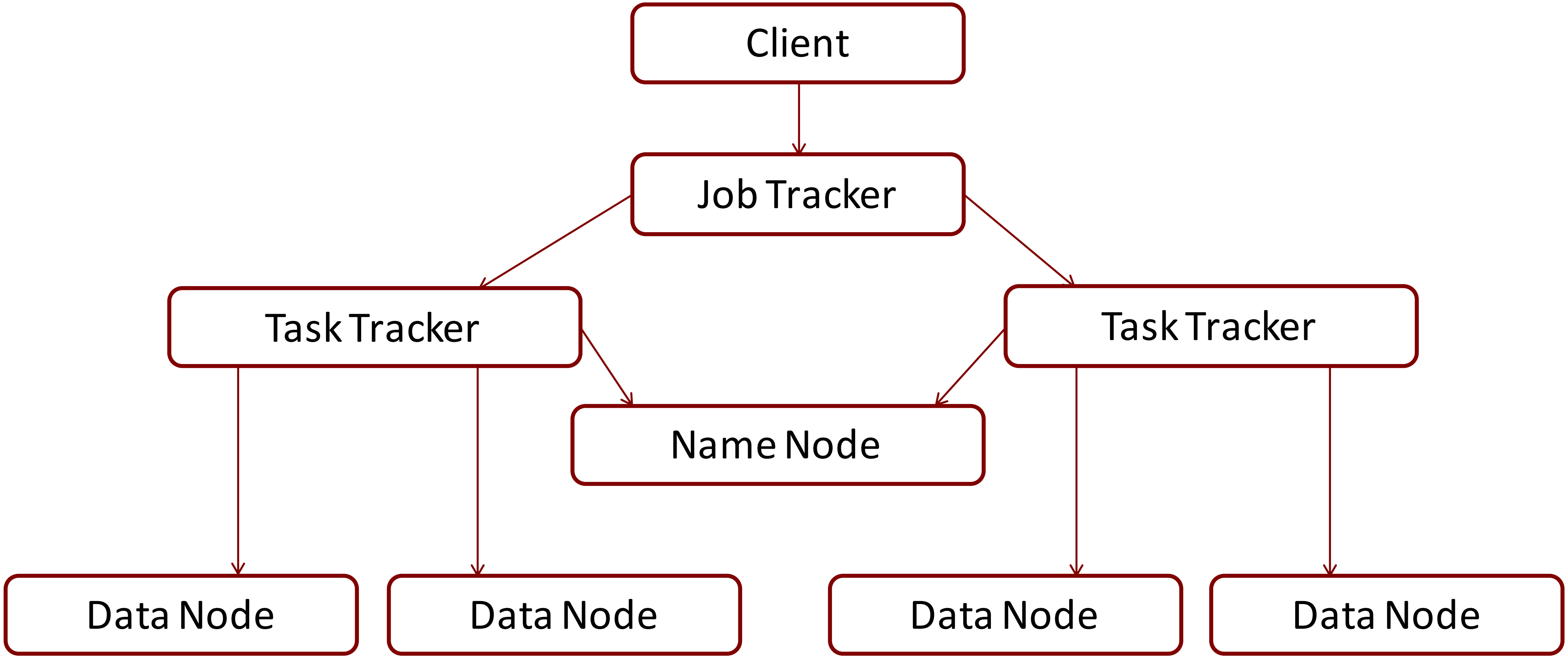
2) Denormalise
3) Transform

MAP

4) Upload
5) Query

REDUCE

Cloud computing



Client

Job Tracker

Task Tracker

Task Tracker

Name Node

Data Node

Data Node

Data Node

Data Node

Process pipeline

1) Translate to CSV (exports_makecsv)
2) Filter out supressed data (exports_process_supression)
3) Get ComCode data for that month (get_comcodes)
4) De-Normalise CSV with ComCodes and translate dates to timestamps (expand_csv)
5) Import into Big Query

DEMO & EXERCISE

Data in Socrata: bit.ly/uk_trade_socrata

Quesitons

Is the UKTrade data big data?

What are the biggest problems with the data?

How would you change your data to use cloud compute platforms?

Where are we now?