Breaking the wall – building an infrastructure to enable multi-disciplinary analyses for social sciences and the Internet of Things

Darren Bell

Repository Architect – UK Data Service

IASSIST 2018: Once Upon A Data Point:
Sustaining Our Data Storytellers

30 May 2018

UK Data Service

UK • DATA ARCHIVE

University of Essex

30,000 foot view

- More data was created in 2017 than the previous 5,000 years of humanity.
- Only 0.5% is actually being analysed operationally
- Biotech, Energy, IoT, Healthcare, Automotive, Space, Deep sea explorations, Cybersecurity, Social media, Telecom, Consumer electronics, Manufacturing, Gaming and Entertainment are just some
- It will be critical for organizations to deploy or employ platforms that have the capability to consume huge amounts of data and present that data in a way that helps them make the right decisions.
- This is leading to frenetic competition among enterprises and start-ups. If data is the new oil, who gets to process and refine it?

Repository Infrastructures

- A “repository” is a collection of lifecycles, functions and processes
- There will always be new data, new file formats, new objects and new tech – this is business as usual
- BUT “Big Data”/NNFD is different. The architecture remains the same but demands a different parallel infrastructure.
- This new infrastructure enables new research methods and hopefully opens up new research funding opportunities
- We do not expect the repository “architecture” to change significantly

A future and a USP

- “Big Data” tech gives us opportunities at a smaller scale for re-evaluating how we process and re-use social science data

- Keynote at “Data for Policy” Conference in Sept 2017 London:
Policy value comes from crossing domains – this is “collective intelligence”

- RDA 11th Plenary Berlin:
“*in the modern world, data is no longer composed of static files*”

A secure, trusted platform for cross-disciplinary linkage

1. Secure machine-assisted linkage with privacy guarantees
2. Dynamic creation and re-use of derived information products
3. Cast-iron provenance chains
4. Domain-agnostic research

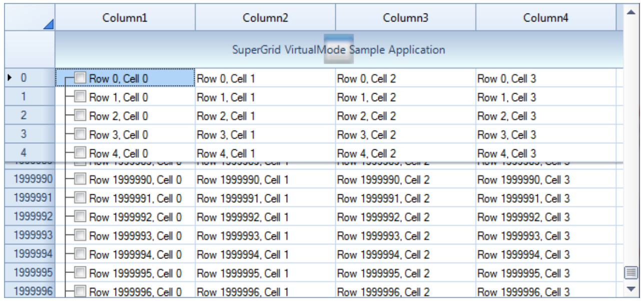
- **PAST** – Relational Databases and files – small and tightly structured
- **PRESENT** - Big Data – lots of it but chaotic
- **FUTURE** “Intelligent Enterprise” – when all this data is tagged, processed and joined up

Hadoop in one slide

- Hadoop started out from a 2003 paper: "The Google File System" https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
- Hadoop is the name for a bunch of different pieces of software that allows you to store and process data across a network (more commonly called a "**cluster**") of computers.
- You can use some or all of these pieces of software. We use some.
- This **cluster** of computers can consist of two or ten thousand computers (or "nodes").
- This **cluster** effectively functions as a single supercomputer
- In a nutshell, it's affordable supercomputing for the masses

What problem does Hadoop solve in practice?

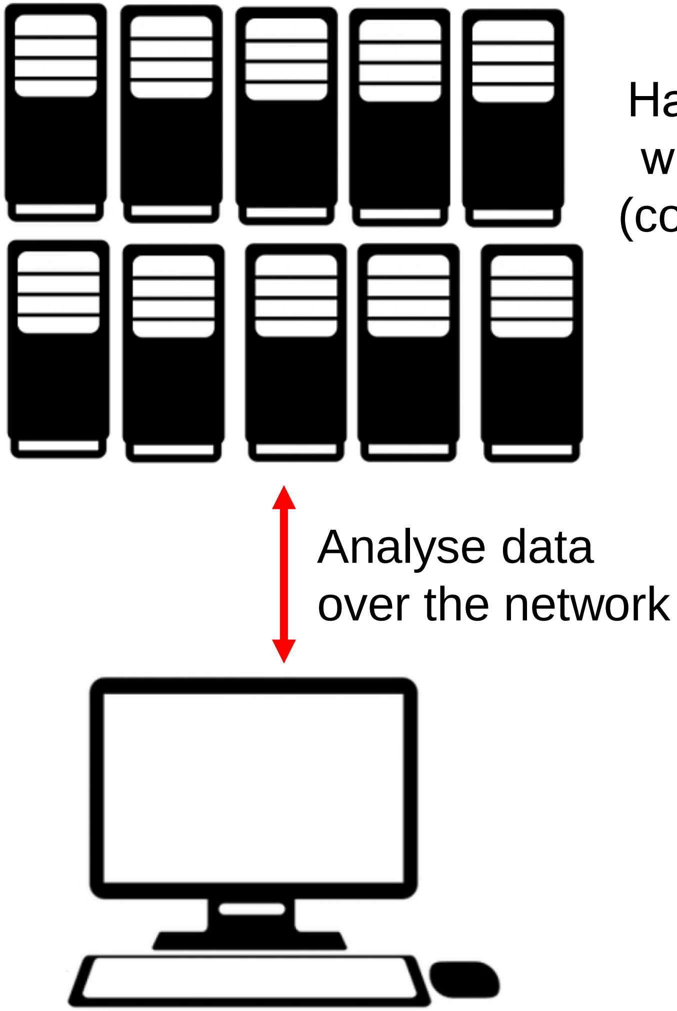
- I have a 2 Terabyte dataset I want to analyse



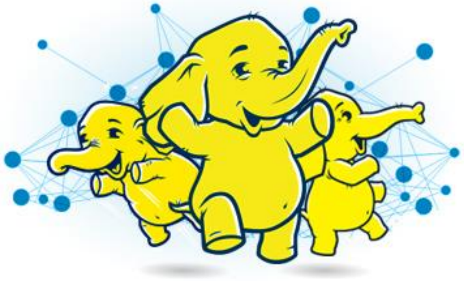
	Column1	Column2	Column3	Column4
	SuperGrid VirtualMode Sample Application			
0	□ Row 0, Cell 0	Row 0, Cell 1	Row 0, Cell 2	Row 0, Cell 3
1	□ Row 1, Cell 0	Row 1, Cell 1	Row 1, Cell 2	Row 1, Cell 3
2	□ Row 2, Cell 0	Row 2, Cell 1	Row 2, Cell 2	Row 2, Cell 3
3	□ Row 3, Cell 0	Row 3, Cell 1	Row 3, Cell 2	Row 3, Cell 3
4	□ Row 4, Cell 0	Row 4, Cell 1	Row 4, Cell 2	Row 4, Cell 3
1999990	□ Row 1999990, Cell 0	Row 1999990, Cell 1	Row 1999990, Cell 2	Row 1999990, Cell 3
1999991	□ Row 1999991, Cell 0	Row 1999991, Cell 1	Row 1999991, Cell 2	Row 1999991, Cell 3
1999992	□ Row 1999992, Cell 0	Row 1999992, Cell 1	Row 1999992, Cell 2	Row 1999992, Cell 3
1999993	□ Row 1999993, Cell 0	Row 1999993, Cell 1	Row 1999993, Cell 2	Row 1999993, Cell 3
1999994	□ Row 1999994, Cell 0	Row 1999994, Cell 1	Row 1999994, Cell 2	Row 1999994, Cell 3
1999995	□ Row 1999995, Cell 0	Row 1999995, Cell 1	Row 1999995, Cell 2	Row 1999995, Cell 3
1999996	□ Row 1999996, Cell 0	Row 1999996, Cell 1	Row 1999996, Cell 2	Row 1999996, Cell 3

- I cannot load it into Excel, SPSS etc. on my PC

Answer: split the file across many PCs



Hadoop "cluster" with 10 "nodes" (could be 10,000)

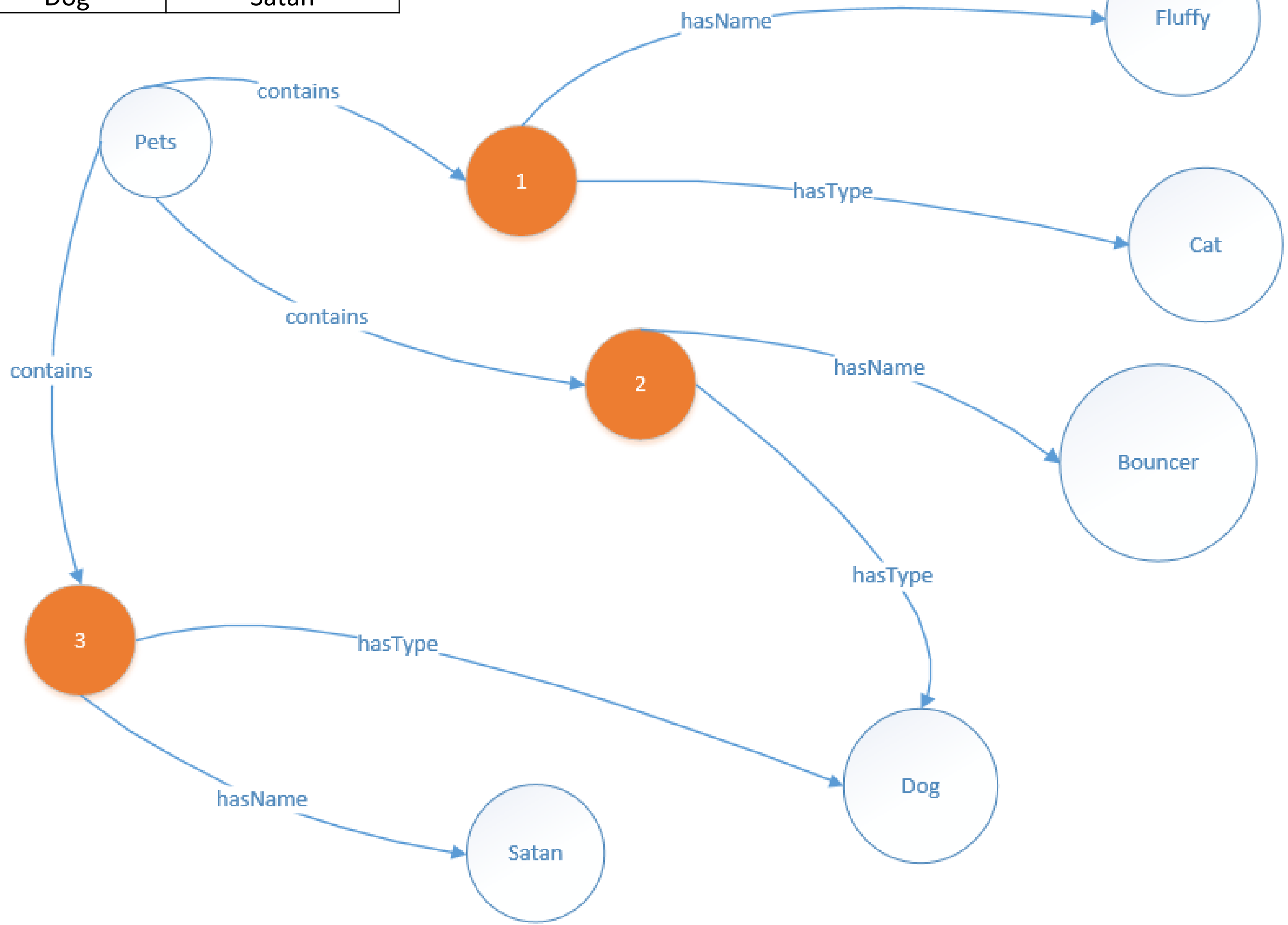


Analyse data over the network

LINKED DATA – from a grid to a “graph”

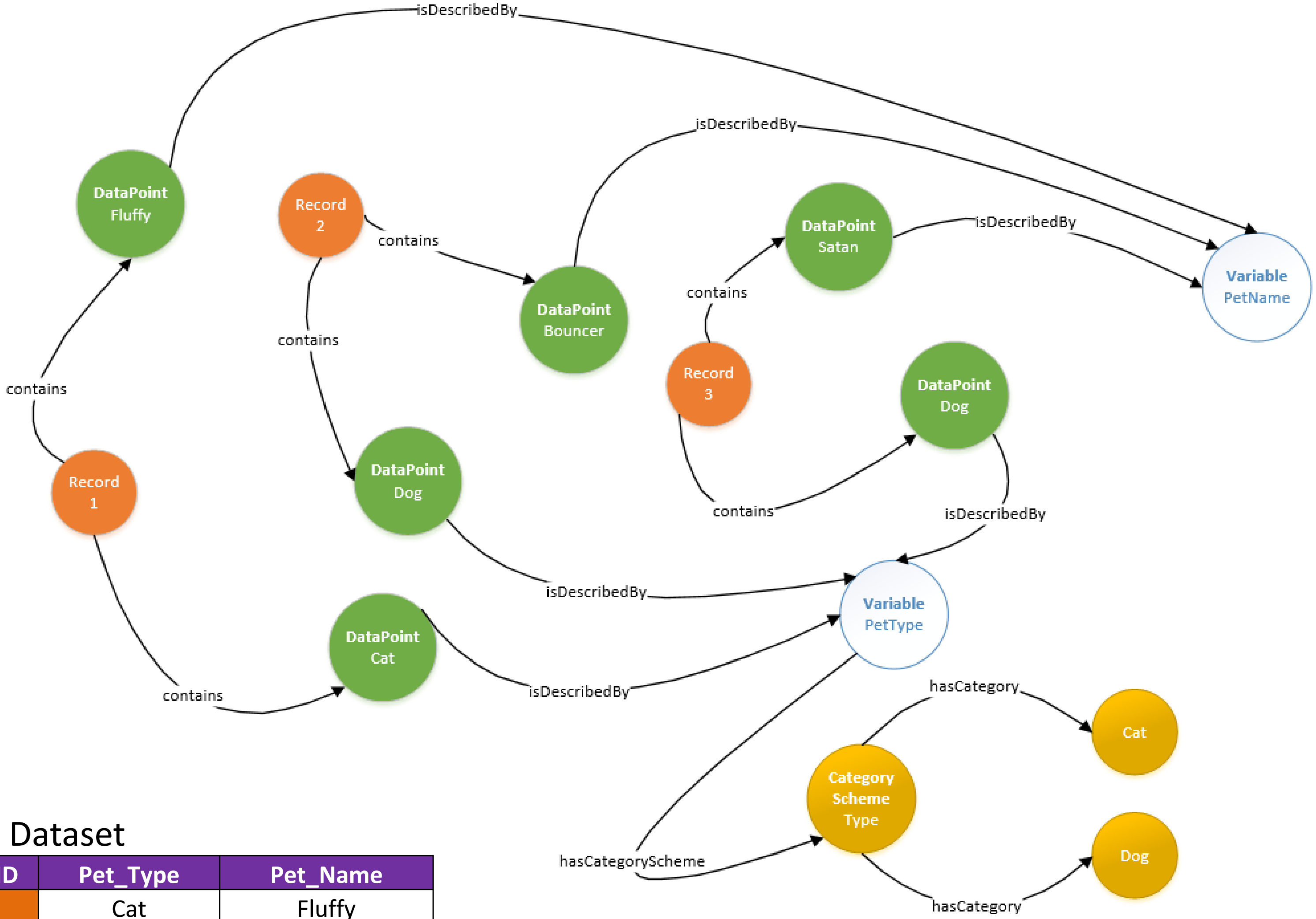
PETS Dataset

Pet_ID	Pet_Type	Pet_Name
1	Cat	Fluffy
2	Dog	Bouncer
3	Dog	Satan



UK Data Service

DDI4 allows us to do data as a graph – it can be pets, energy or social science



PETS Dataset

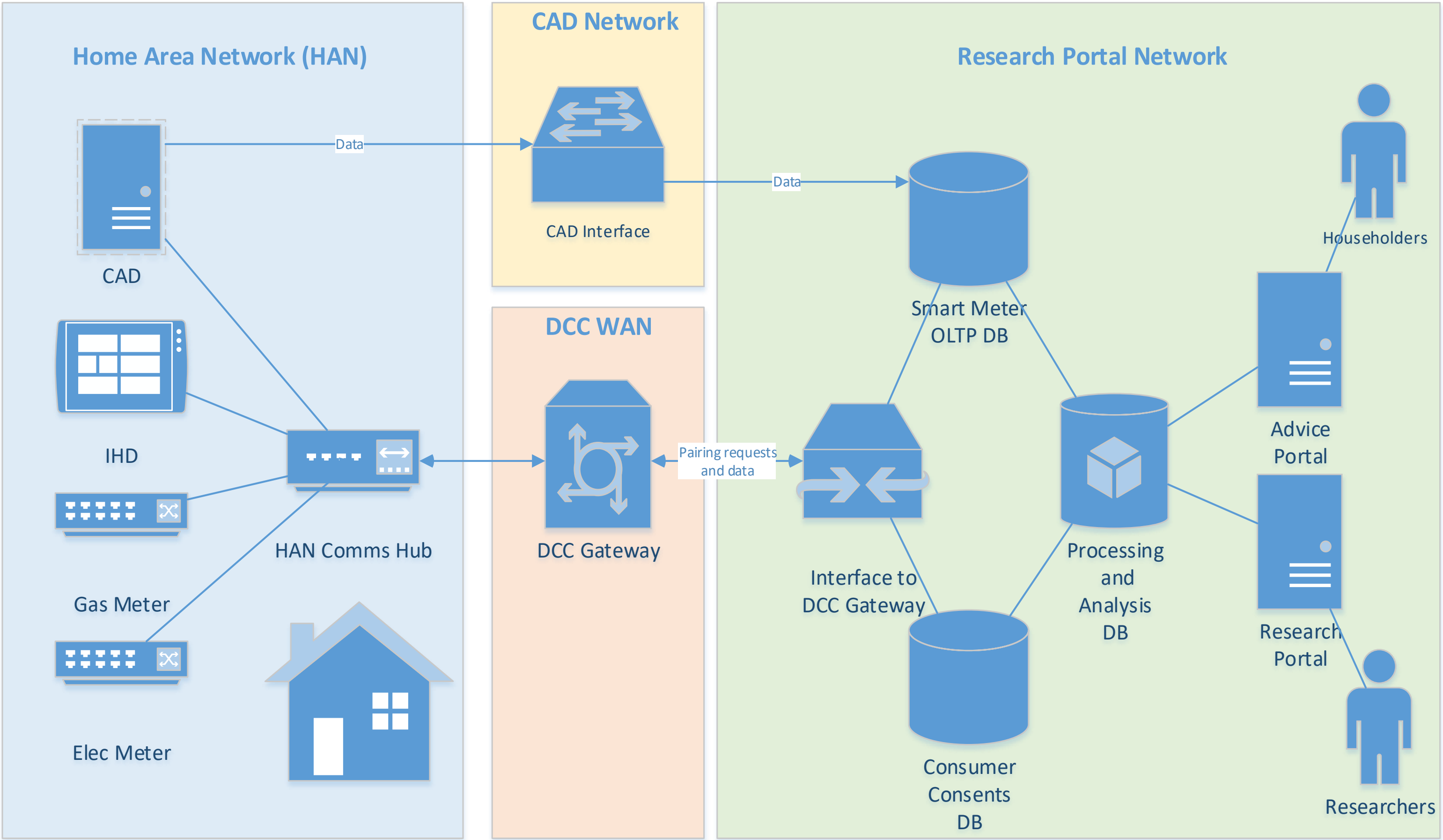
Pet_ID	Pet_Type	Pet_Name
1	Cat	Fluffy
2	Dog	Bouncer
3	Dog	Satan

UK Data Service

Summary so far

- IoT data is about things and events. What we must be able to do is contextualise it.
- That could be people (SocScience), weather (Environment), places (Geospatial)
- **HADOOP** LETS US STORE ALL THIS DATA IN ONE PLACE
- A **GRAPH** LETS US ANALYSE THIS DATA IN A STRUCTURE THAT MORE NATURALLY REFLECTS THE CONNECTIONS BETWEEN THE DATA AND THE METADATA
- For us, Big Data is not just about the Big. It's where scale intersects new data paradigms like linked data and graphs.

Smart Meter Research Portal Sep 2019

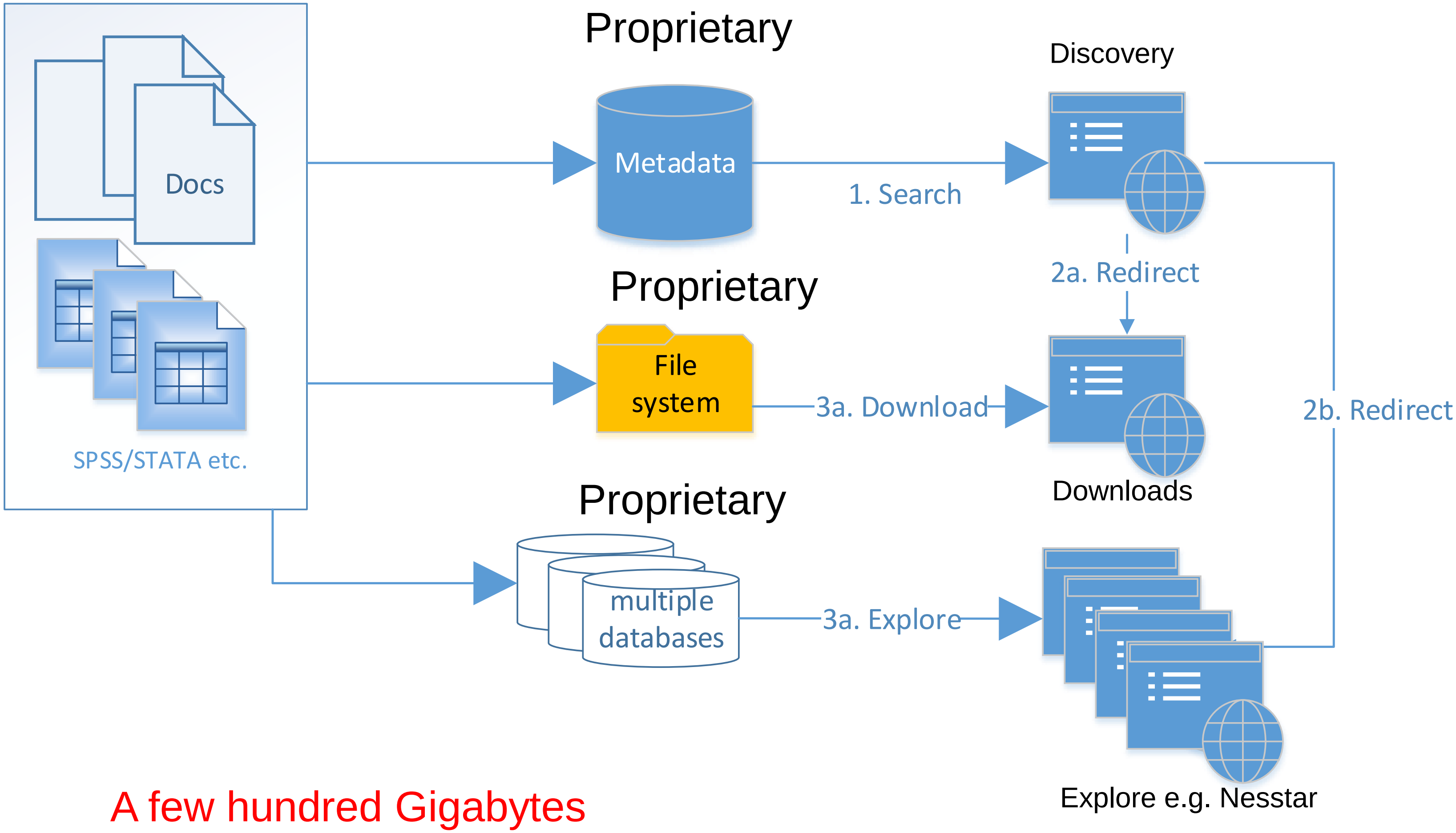


UK Data Service

Core Principles

- Open Source
- FAIR
- Scalable
- Standards-based
- TDR Compliant
- Domain-agnostic

Data Platform: The repository now

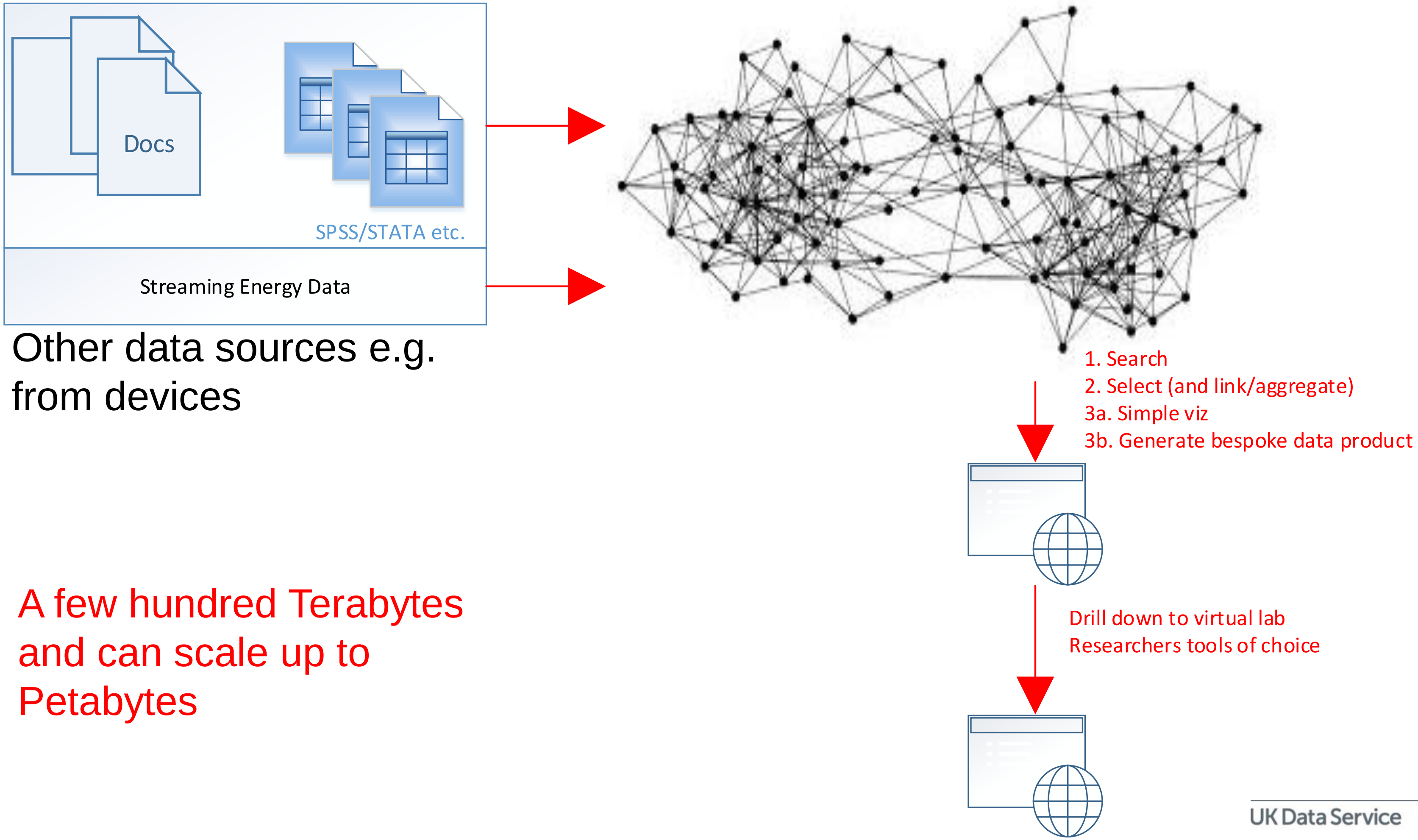


A few hundred Gigabytes

Data Platform:

Repository Target

DDI4: data/metadata
ODRL: access-control



Other data sources e.g. from devices

1. Search
2. Select (and link/aggregate)
3a. Simple viz
3b. Generate bespoke data product

Drill down to virtual lab
Researchers tools of choice

A few hundred Terabytes and can scale up to Petabytes

Or from this…

Pick pre-built datasets from the catalogue

UK Data Service

Plus this…

Build your own

Semantic Platform

- Unified approach to any re-usable components.
  CVs
  Code Lists
  Category Schemes
  Taxonomies
  Thesauri
  Ontologies – particularly GeoSpatial

- VocBench 3 management tool (http://vocbench.uniroma2.it/)

- This underpins the ability to perform machine-assisted harmonisation

Access Platform

Unify:

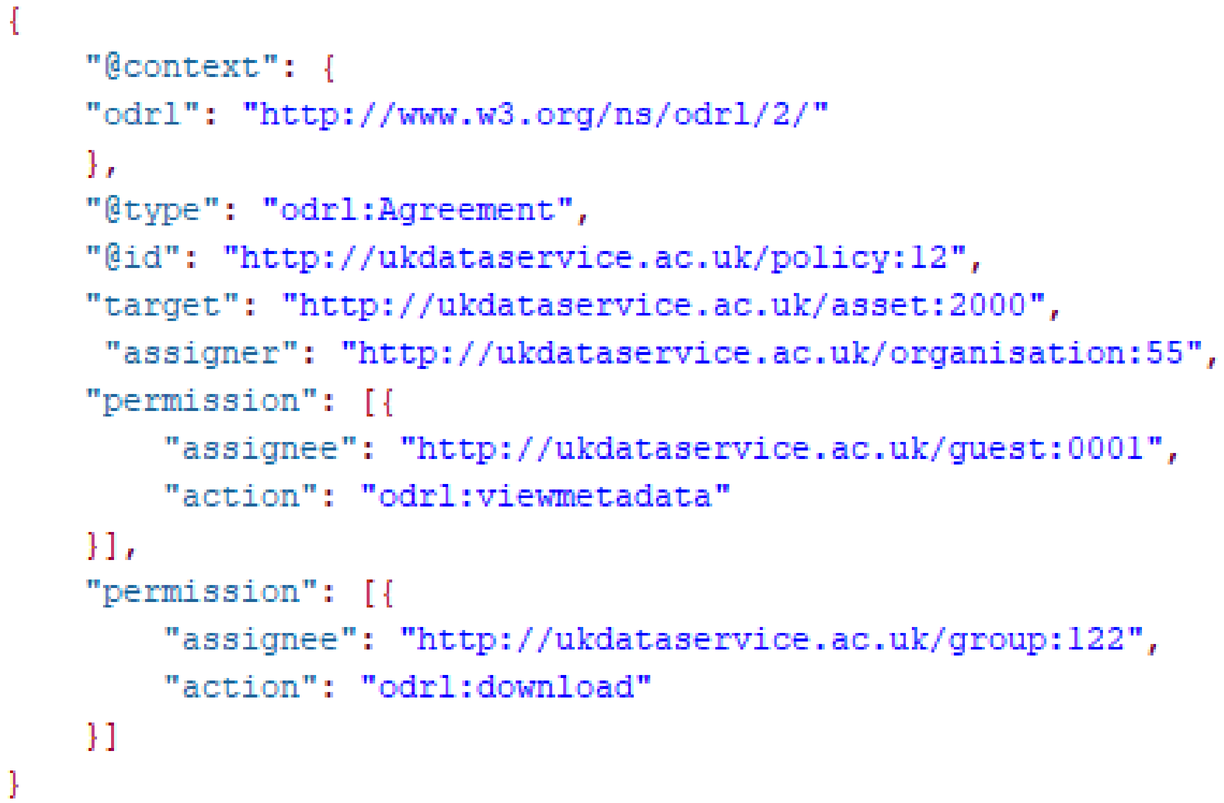
- Consents
- Rights
- Licensing
- Access Mediation

in a single infrastructure.

ODRL (open digital rights language) provides a machine-actionable “vocabulary” to formally describe these entities.

Assets *have*
Policies *consisting of*
Rules (Permissions, Obligations and Prohibitions)
which apply to Parties
and which determine Actions
which may have Constraints

Access Platform: ODRL example

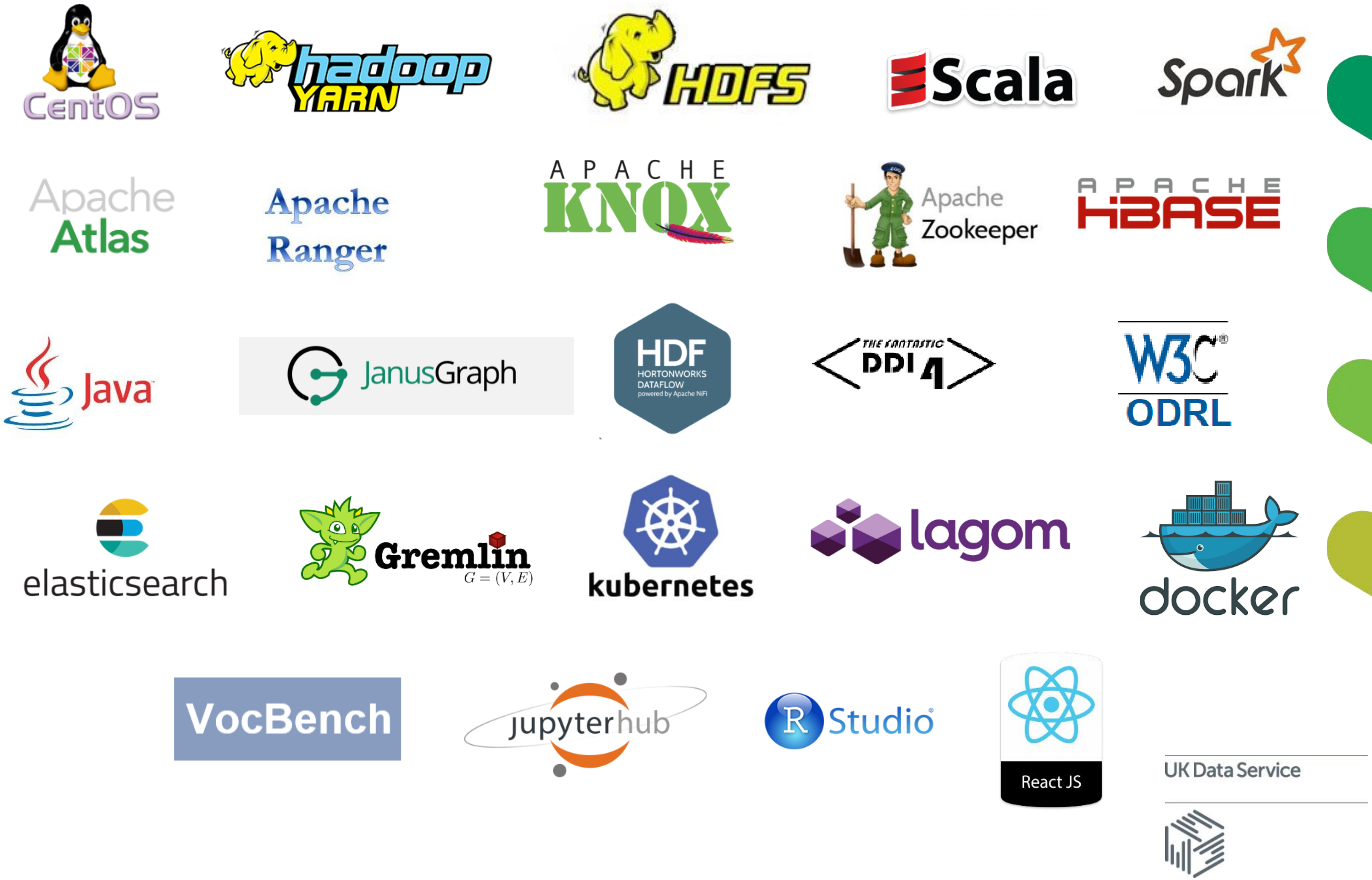


```
{
    "@context": {
    "odrl": "http://www.w3.org/ns/odrl/2/"
    },
    "@type": "odrl:Agreement",
    "@id": "http://ukdataservice.ac.uk/policy:12",
    "target": "http://ukdataservice.ac.uk/asset:2000",
     "assigner": "http://ukdataservice.ac.uk/organisation:55",
    "permission": [{
        "assignee": "http://ukdataservice.ac.uk/guest:0001",
        "action": "odrl:viewmetadata"
    }],
    "permission": [{
        "assignee": "http://ukdataservice.ac.uk/group:122",
        "action": "odrl:download"
    }]
}
```

=>
For Study 2000, ONS *(organisation #55)* have declared that guest users can view the metadata and UK users *(group #122)* can download the study

UK Data Service

The DSaaP ecosystem



Demo

Final messages

- The computational power of Hadoop enables management of complexity
- Unification of metadata and data at lifecycle, function and process level
- From dissemination of files (an archive)
  to enabling digital resources (a research data infrastructure)
- Concept driven data discovery at the variable level and lower
- Standards based around semantic web and DDI4
- Interoperability across domains
- Unified access model based on standard information model (ODRL)
- Derived and reproducible information products

Questions

Darren Bell

dbell@essex.ac.uk

UK Data Service