Building a cross-disciplinary platform for smart meter data

Darren Bell

Director of Technical Services
UK Data Service

CLOSER Webinar
24 Feb 2021

UK Data Service

UKDA “Traditional” data infrastructure

- Approx. 8000 survey datasets outputted in SPSS/STATA/TAB with PDF documentation
- Data held in files on Windows folders (with different levels of security) as binary files
- Metadata held in MS-SQL databases
- Discovery (search and download) website in C# asp.net, based on Umbraco CMS
- Other data exploration tools like Nesstar (based on MySQL, JSP)
- Very much a classical architecture disseminating “pre-designed” datasets

DDI Codebook metadata

- Flat and simple. Easy to understand. Limited in sophistication
  The “Study” is the core unit of management
- Supports discovery operations (this is the primary use case) but less expressive around variable relationships or rights management.

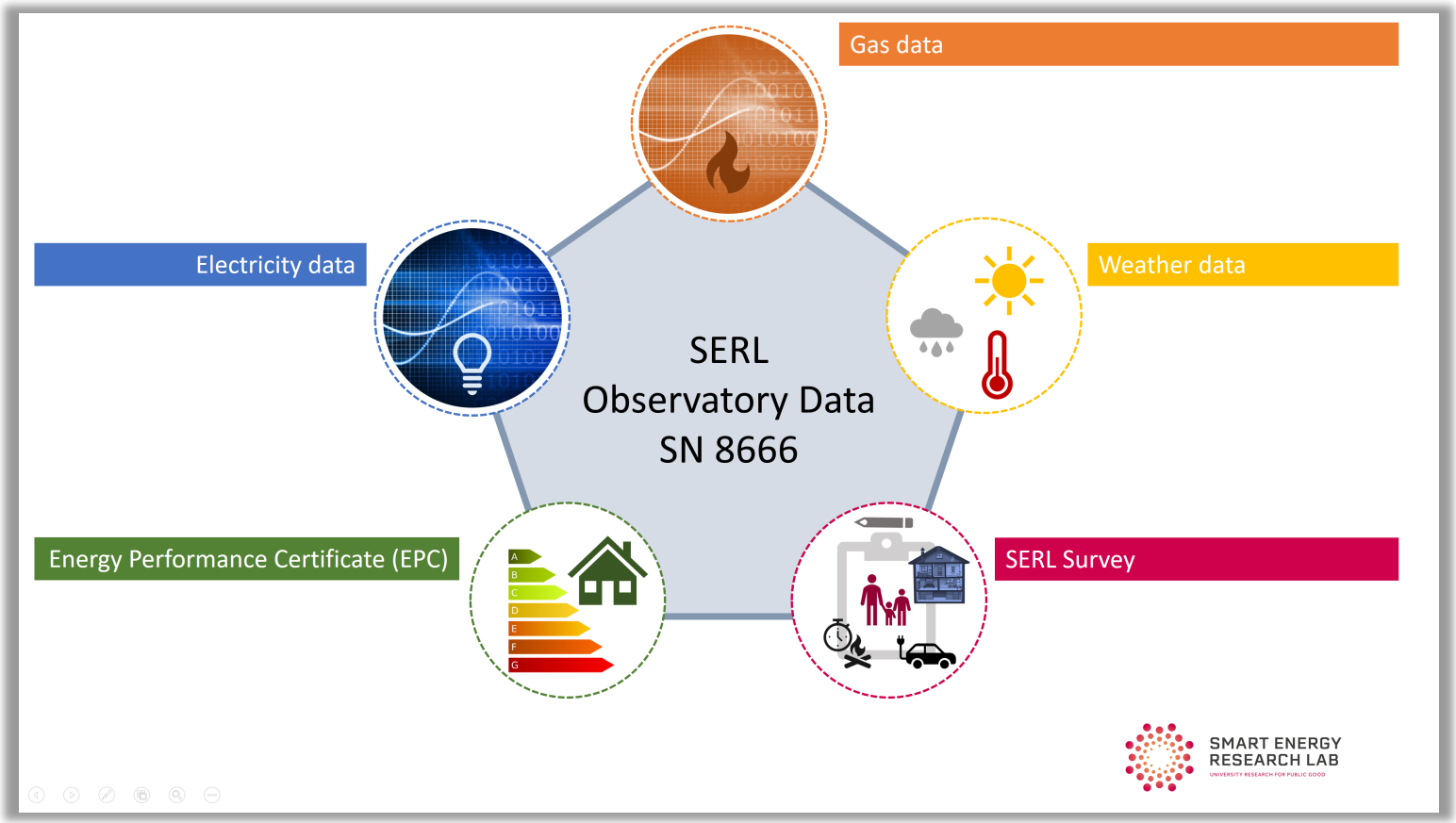

```xml
<codeBook xsi:schemaLocation="http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd"
    <docDscr>
        <citation>
            <titlStmt>
                <titl xml:lang="en" >DDI2.5 XML CODEBOOK RECORD FOR STUDY NUMBER 7481</titl>
            </titlStmt>
            <prodStmt>
                <prodDate>2014-04-04T14:42:58Z</prodDate>
            </prodStmt>
        </citation>
    </docDscr>
    <stdyDscr>
        <citation>
            <titlStmt>
                <titl xml:lang="en" >Integrated Census Microdata (I-CeM), 1851-1911</titl>
                <subTitl></subTitl>
                <altTitl>I-CeM</altTitl>
                <IDNo agency="UKDA" xml:lang="en" >7481</IDNo>
                <IDNo agency="datacite" xml:lang="en" >10.5255/UKDA-SN-7481-2</IDNo>
            </titlStmt>
            <rspStmt>
                <AuthEnty xml:lang="en" >Schurer, K., University of Essex, Department of History</AuthEnty>
                <AuthEnty xml:lang="en" >Higgs, E., University of Essex, Department of History</AuthEnty>
            </rspStmt>
            <prodStmt>
                <copyright>Copyright K. Schürer, E.J. Higgs and BrightSolid</copyright>
                <fundAg>Economic and Social Research Council</fundAg>
                <grantNo>RES-062-23-1629</grantNo>
            </prodStmt>
            <distStmt>
                <distrbtr xml:lang="en" >UK Data Service</distrbtr>
                <depositr>Schurer, K., University of Essex, UK Data Archive</depositr>
                <depDate date="2018-09-06T16:25:04Z" >06 September 2018</depDate>
                <distDate date="2014-04-04T14:42:58Z" >04 April 2014</distDate>
```


UK Data Service

Smart Energy Research Lab (SERL) Project

- Smart Energy Research Lab (https://www.serl.ac.uk) is a data resource for UK research community
- High-quality smart meter and linked contextual data for innovative research
- Still very difficult for researchers to access high quality energy data in the UK
- However substantial barriers to accessing smart meter data
  - Technical
  - Legal
  - Financial

Cross-disciplinary by design



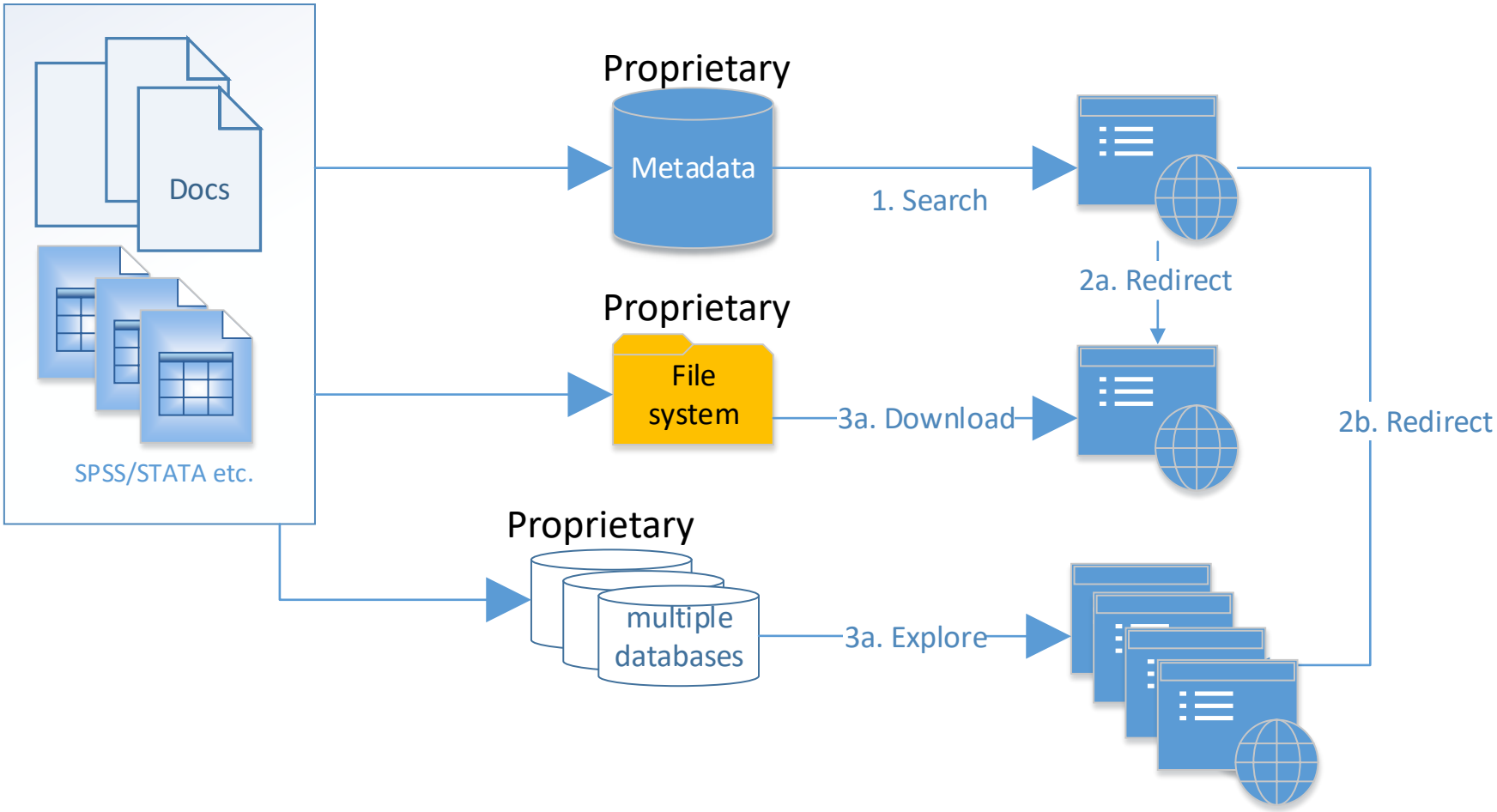
We need to move from this…

Pick pre-built datasets from the catalogue

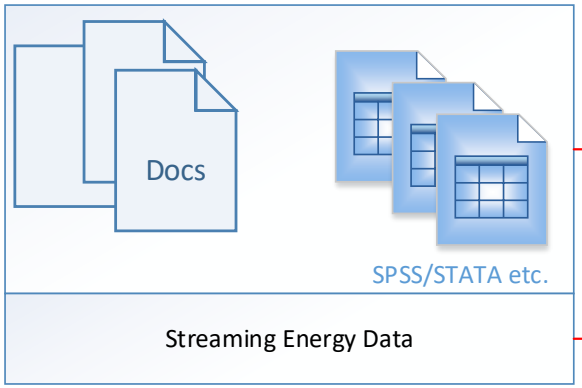
To this…

Build your own

In more detail – from this…

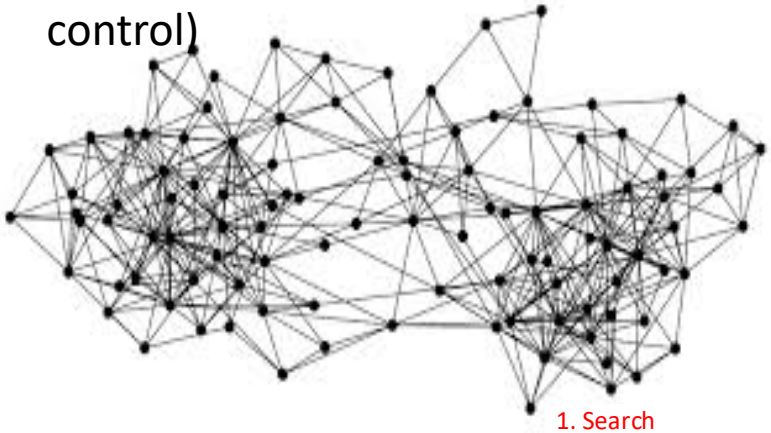


In more detail – to this…

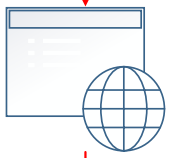


Other data sources e.g. Bio?

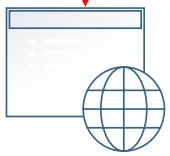
DDI-3/4 (data/metadata/access-control)



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



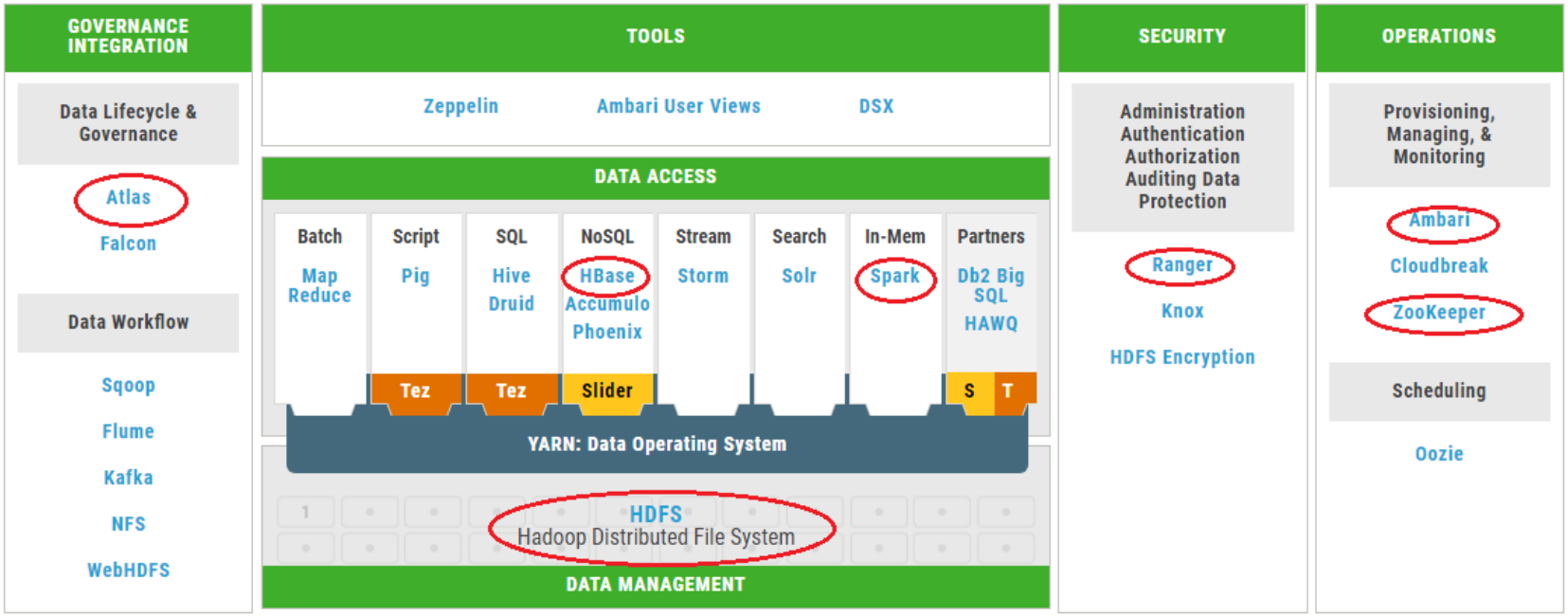
Drill down to virtual lab
Researchers tools of choice



UK Data Service

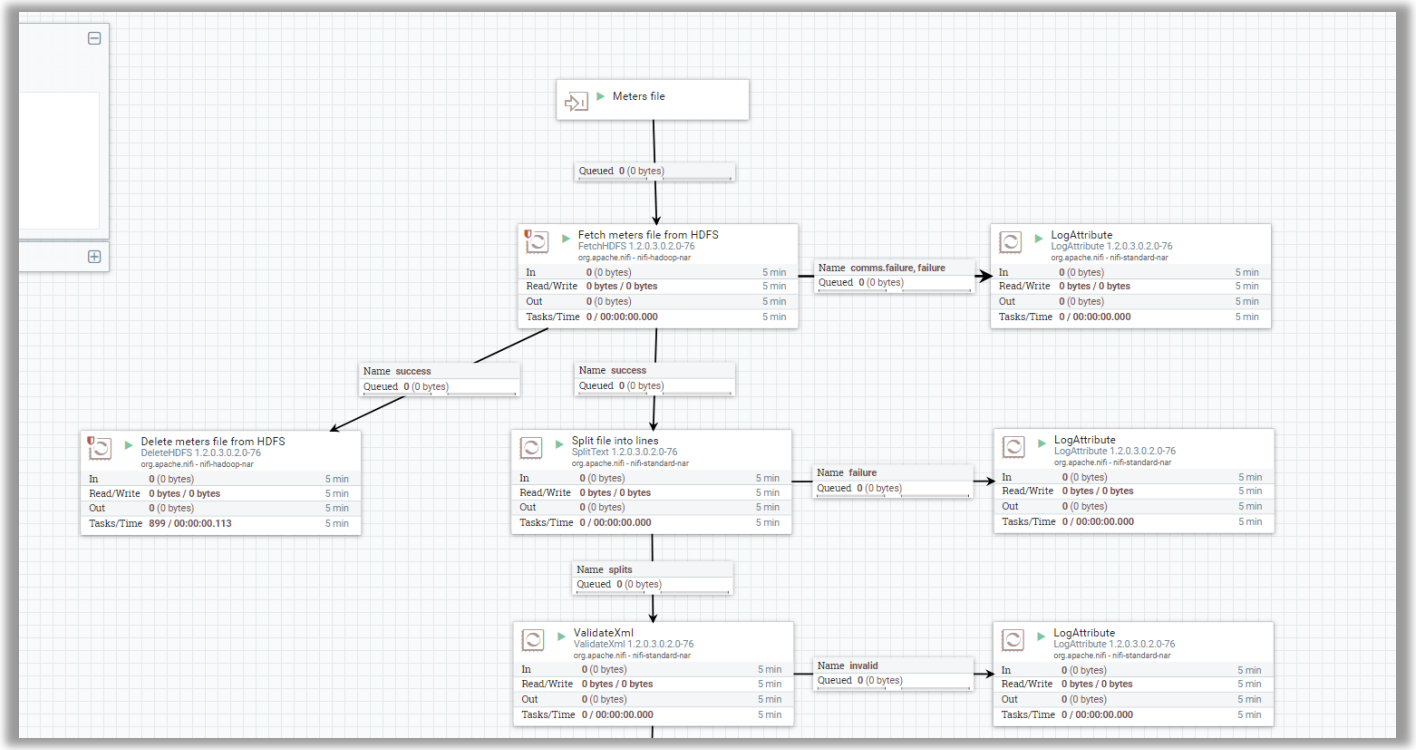
The infrastructure bit - Big data and Hadoop

- Hadoop is a *suite* of different products (like Office is a suite of Excel, Access, Word, Powerpoint, Publisher etc.)

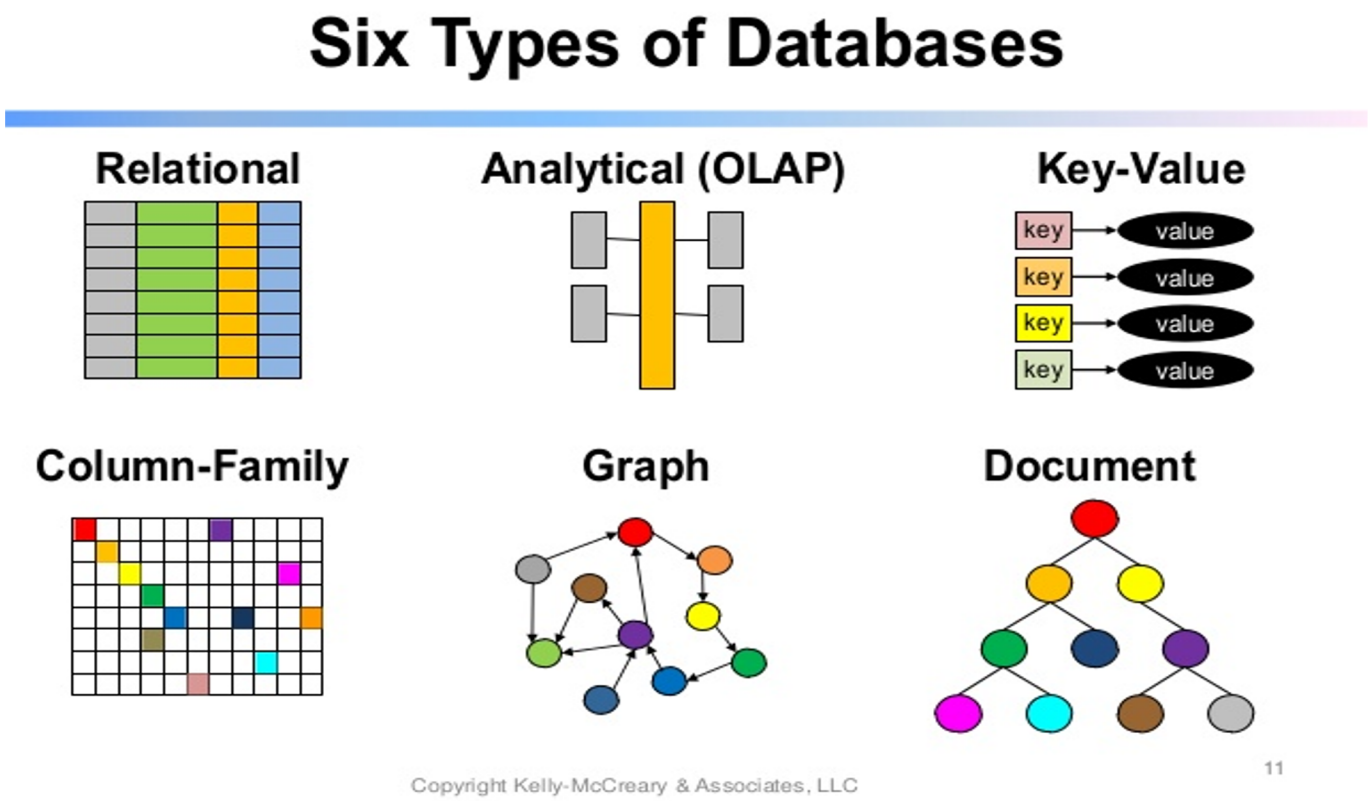


The ingest bit - Apache NiFi

- We use Hortonworks Data Platform (HDF) to move data around and implement QA / validation processes Captures logging and provenance for all operations



SQL - it was good while it lasted but it’s time to move on

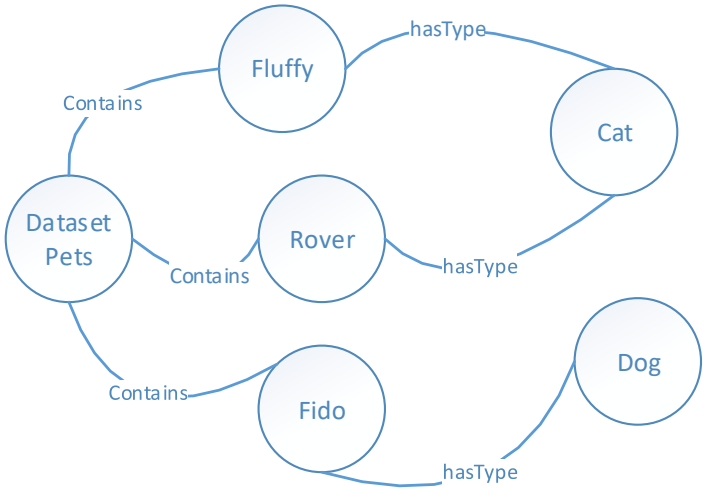


What and why of HBase/JanusGraph

- Traditional relational data stores will not scale and it's not always easy to alter the schema
- Instead of:

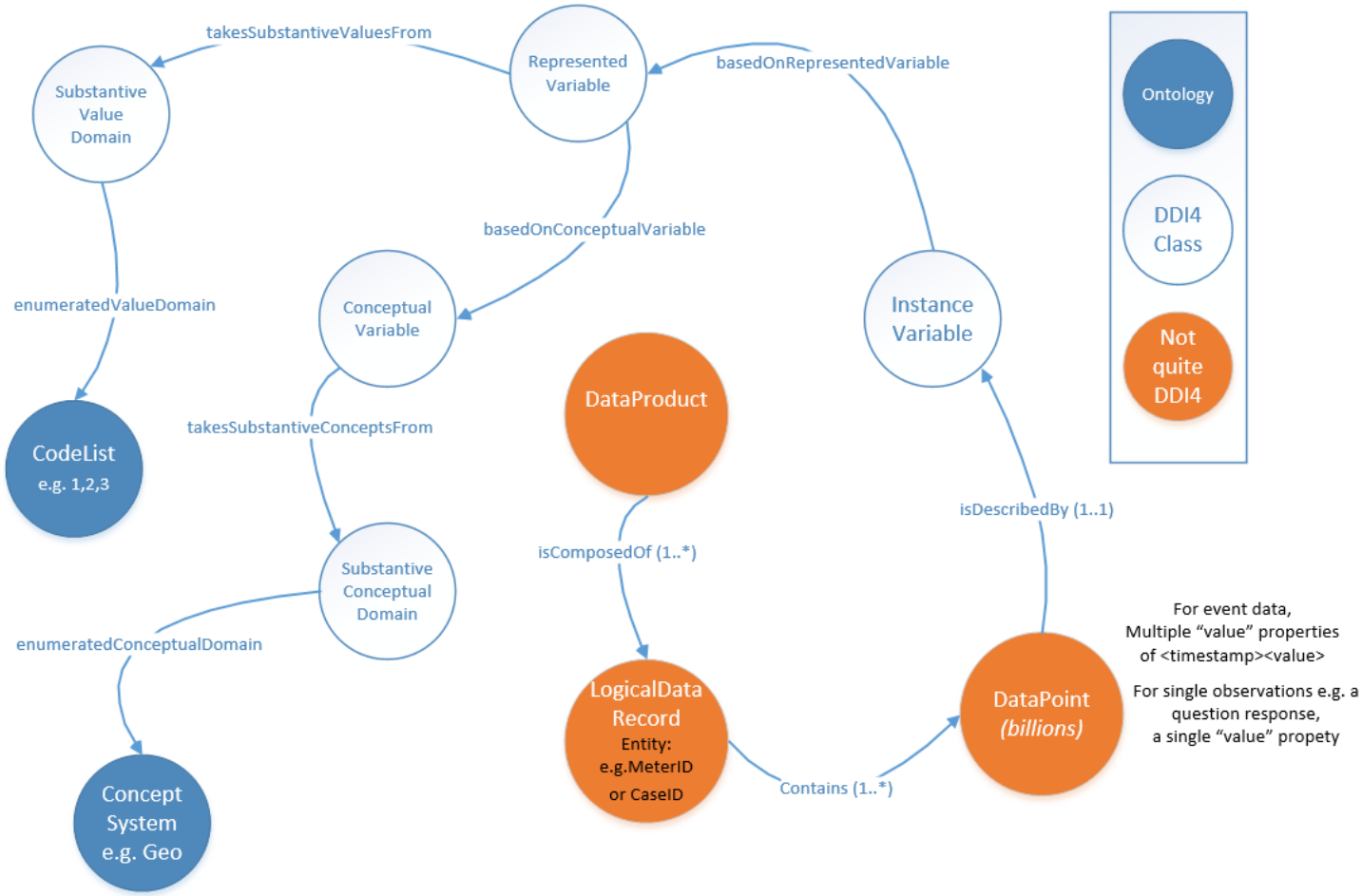
	Name	Type
Row1	Fluffy	Cat
Row2	Rover	Dog
Row3	Fido	Dog

- A property graph would be modelled like:



6 statements in the database		
PetsDataset	Contains	Fluffy
PetsDataset	Contains	Rover
PetsDataset	Contains	Fido
Fluffy	hasType	Cat
Rover	hasType	Dog
Fido	hasType	Dog

SERL Logical Data Model (DDI4)



UK Data Service

The access bit

To do FAIR properly, we <u>must</u> look to promote more machine-actionable access and rights models:

Unify:

- Consents
- Rights
- Licensing
- Access Mediation

in a single model

ODRL (open digital rights language) provides a ready-made machine-actionable "vocabulary" to describe these.

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

Simple 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

Front-end UX for SERL

- (1) Single entry point. We allow researcher to search for the variables, time-ranges and geography they are interested in (a “Universal Query”)

- (2) Next step is to identify possible linkages and access criteria while incrementally filtering and/or aggregating.

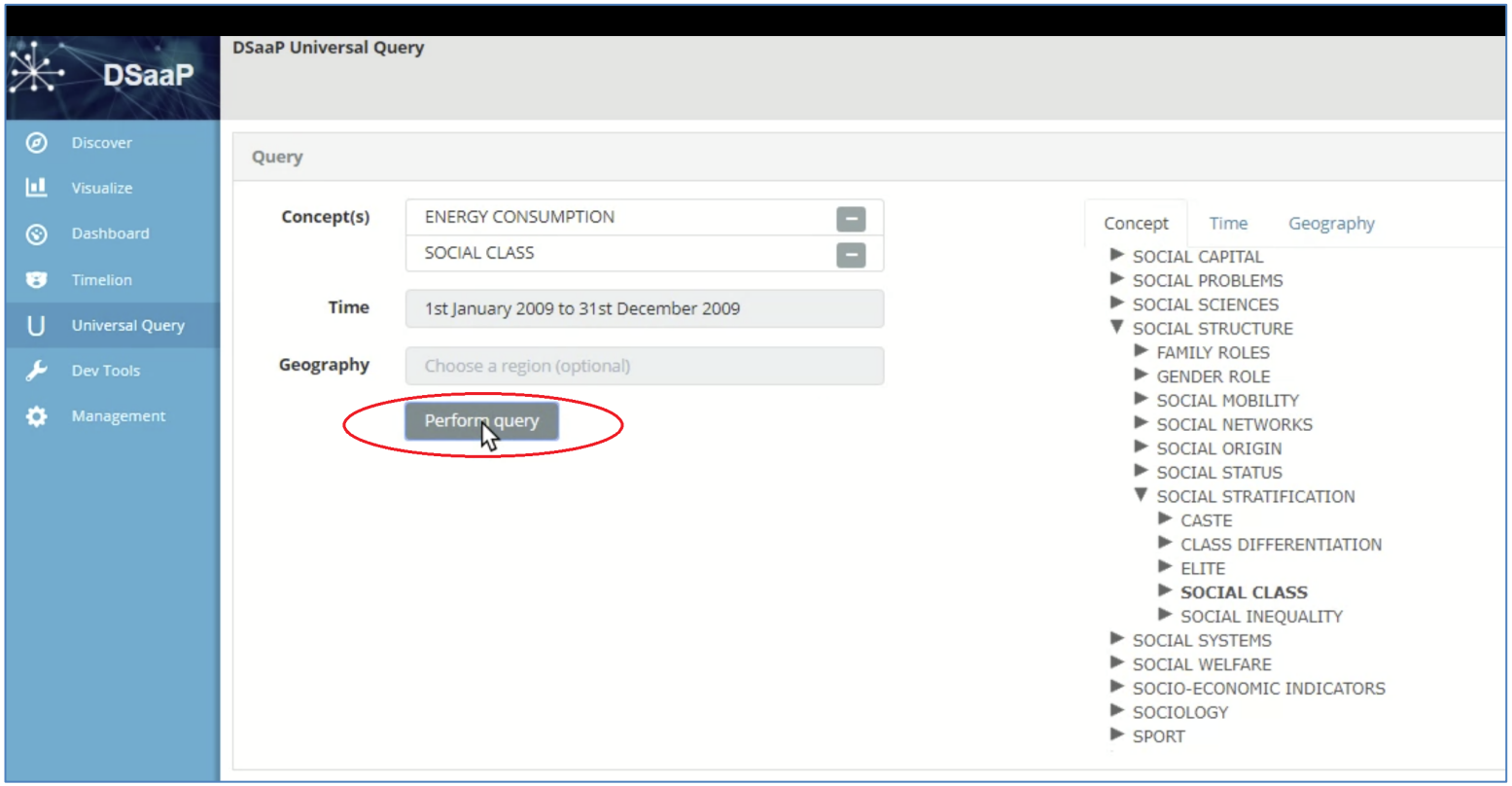
- (3) Once data product is defined, we then execute the access conditions and backhaul the data into a virtual environment, with the analytic tool they have chosen to use

Example use case on PoC platform

- Research question: Average energy consumption by social class
- Two demographics : affluent and poor
- Geospatial visualisation by local authority district
- Uses smart meter energy data and household survey data
- Uses Spark to quickly process millions of records

Select concepts and time

In this case: “ENERGY CONSUMPTION” and “SOCIAL CLASS” in 2009



Filter/aggregate by variables

DSaaP

- Discover
- Visualize
- Dashboard
- Timelion
- Universal Query
- Dev Tools
- Management

7591_e	edrp_elec		
	Variable	Question	Concept
	ELECKWH		ENERGY CONSUMPTION
7591_eg_y	edrp_annual_total_energy		
	Variable	Question	Concept
	TOTAL_y		ENERGY CONSUMPTION
7591_g	edrp_gas		
	Variable	Question	Concept
	GASKWH		ENERGY CONSUMPTION
7591_hh	edrp_household.sav		
	Variable	Question	Concept
	ACORN_Category		SOCIAL CLASS
	LAD		Geography (Local Authority)

Visualise | Visualise | Visualise | Advanced

Advanced

Variables to visualise

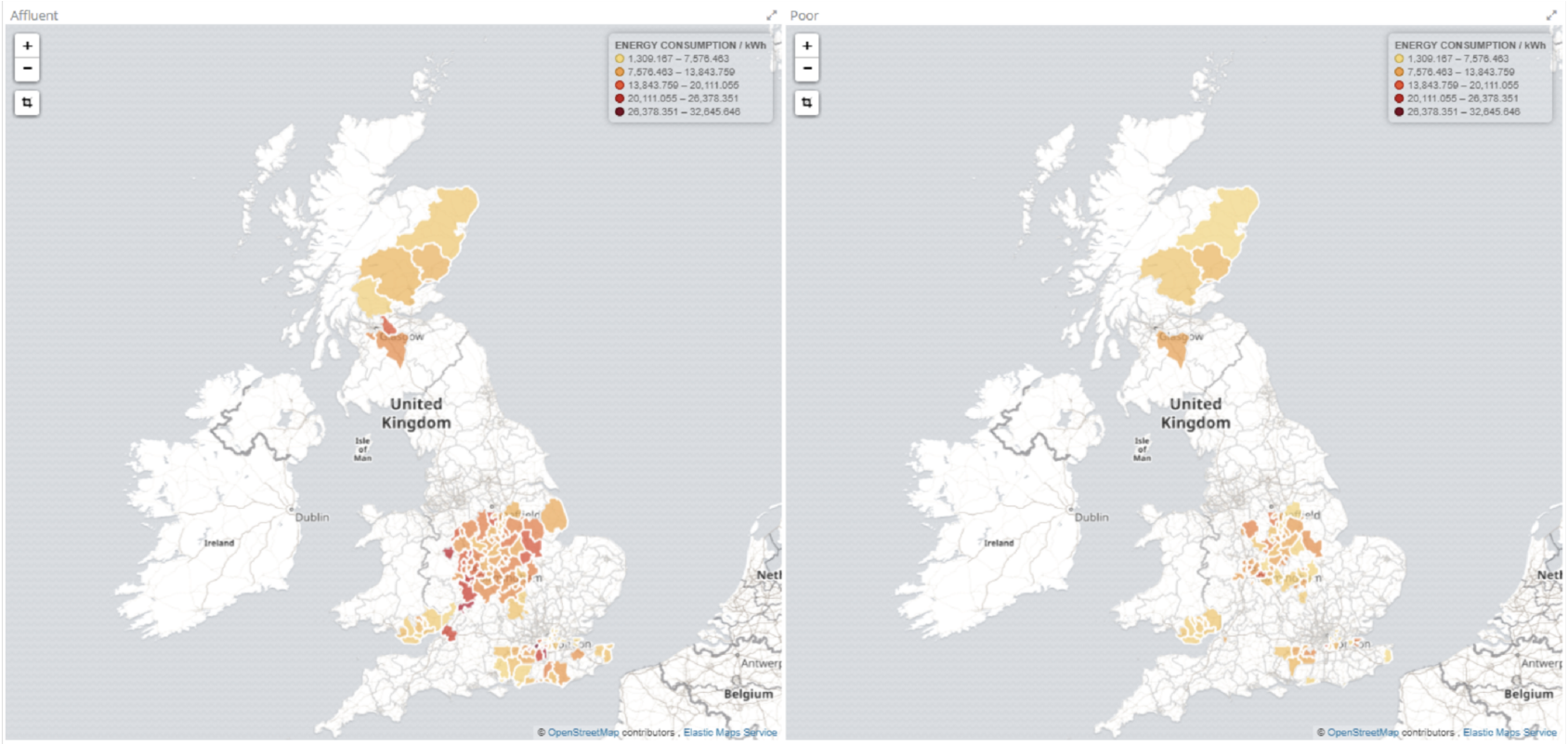
TOTAL_y — ENERGY CONSUMPTION

Concepts to aggregate on

SOCIAL CLASS

'Affluent Achievers' to 'Comfortable Communities'	Affluent
'Comfortable Communities' to 'Financially Stretched'	Mid range
'Financially Stretched' to 'Not Private Households'	Poor

Result


Affluent
ENERGY CONSUMPTION / kWh
1,309.167 – 7,576.463
7,576.463 – 13,843.759
13,843.759 – 20,111.055
20,111.055 – 26,378.351
26,378.351 – 32,645.646
United Kingdom
Isle of Man
Dublin
Ireland
Antwerp
Belgium
© OpenStreetMap contributors , Elastic Maps Service
Poor
ENERGY CONSUMPTION / kWh
1,309.167 – 7,576.463
7,576.463 – 13,843.759
13,843.759 – 20,111.055
20,111.055 – 26,378.351
26,378.351 – 32,645.646
United Kingdom
Isle of Man
Dublin
Ireland
Antwerp
Belgium
© OpenStreetMap contributors , Elastic Maps Service

Core messages

- Unification of metadata and data at lifecycle, function and process level is now possible and moreover, essential.
- Concept driven data discovery at the variable level and lower allows for powerful and flexible generation of bespoke data products.
- Standards based around semantic web and DDI(based on GSIM – Generic Statistic Information Model).
- Datum based approach = *domain-agnostic research*
- Unified access model
- Derived and reproducible *data products*