TALK DATA TO ME:
SPARKING INSIGHTS AT ELSEVIER

Emlyn Whittick
Principal Software Engineer, Elsevier

SPARK SUMMIT
EUROPE 2016

Elsevier

- Founded in 1880 (136 years old)
- Started life as a traditional publisher
- Specialised in scientific and medical publishing
- Now an information solutions provider

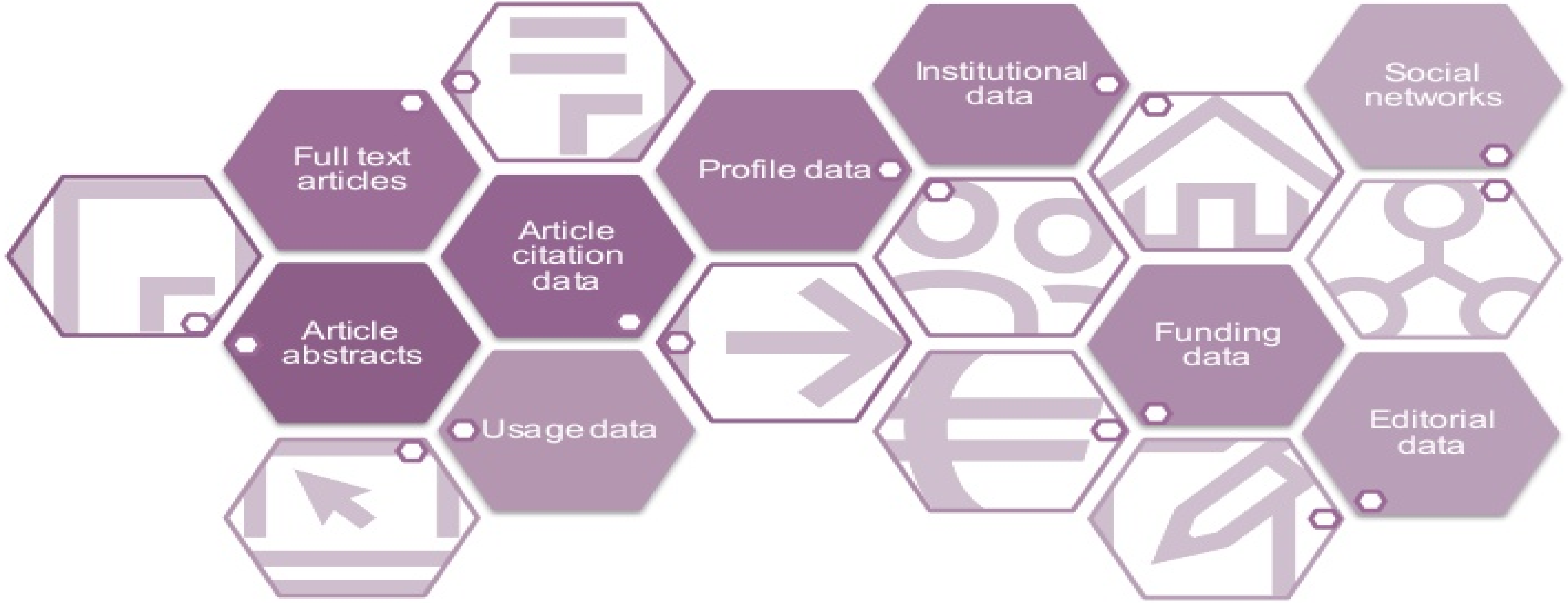
NON SOLUS

ELSEVIER

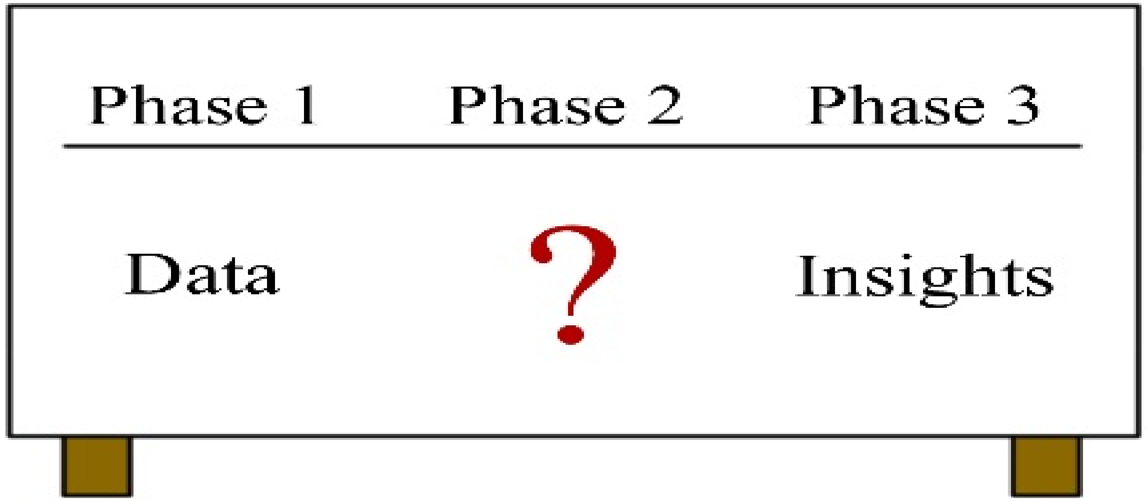
LEAD THE WAY IN ADVANCING SCIENCE, TECHNOLOGY AND HEALTH

Elsevier’s Mission

A Wealth of Data

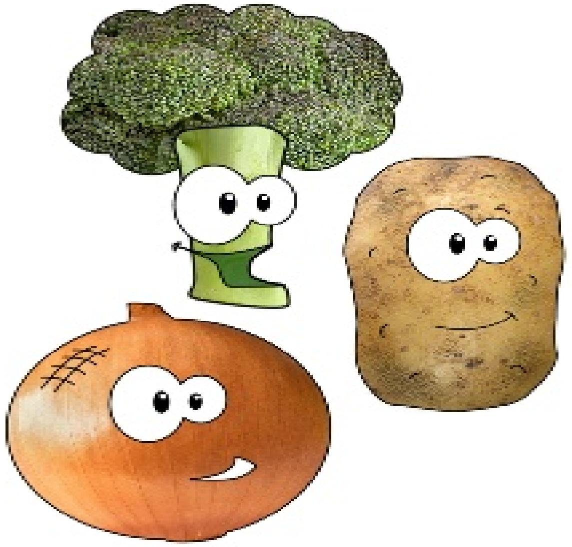


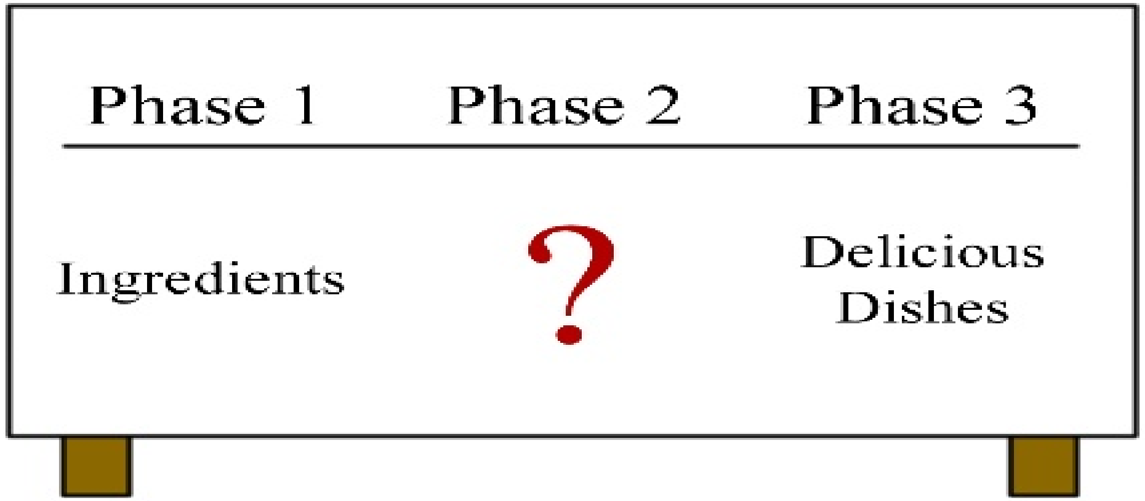
The Challenge



Phase 1	Phase 2	Phase 3
Data	?	Insights

The Challenge





Phase 1	Phase 2	Phase 3
Ingredients	?	Delicious Dishes

Part One

COLLECTING THE INGREDIENTS

Let’s Go Shopping

Supermarket

Robin’s Farm

YAMSDIRECT

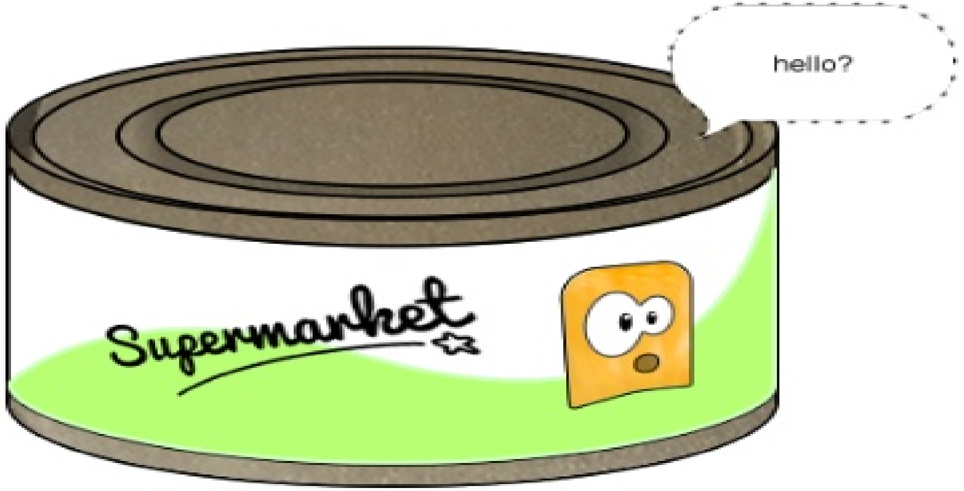
Collecting Ingredients

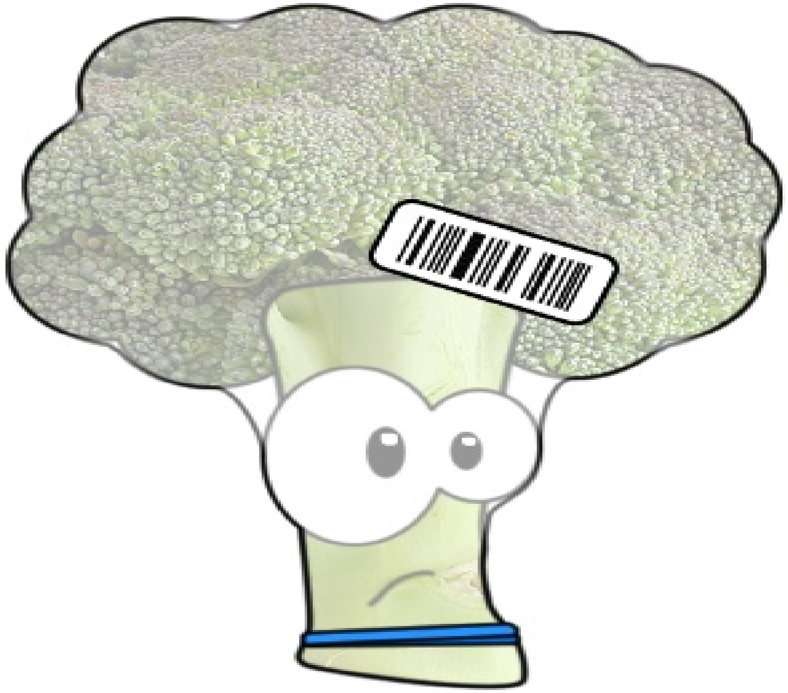
Standardised, automated collection and delivery

Proprietary, mostly automated mechanisms

Manual, ad-hoc collection and delivery

Getting Access





Part Two

ALONG CAME A SPARK

An Initial Spark

- **Use Case:** Application of NLP on Elsevier's entire body of published content
- Databricks selected for:
  - Mounting of data
  - Minimal operational overhead
  - Presentation of results

2014

An Initial Spark

- Databricks used by team of 15 people for content analytics
- Spark adopted as the main processing engine for Elsevier's big data platform

2015

An Initial Spark

- Spark 2.0 and Spark 1.6
- RDD to DataFrame, and now to DataSet
- Production workflows in Scala
- Data Science and Analytics workflows may be in Python or R

2016

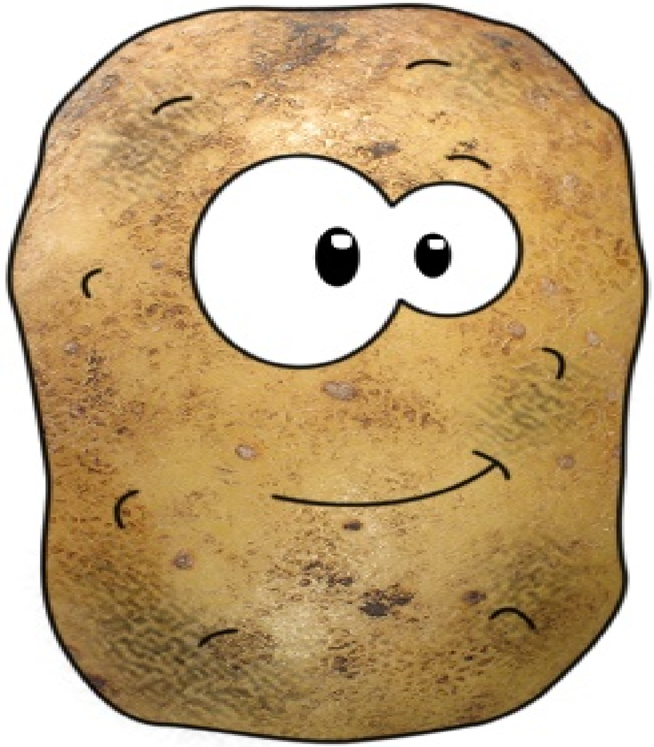
A Recipe for Insights

- Empower the master chefs
- Prepare the ingredients
- Share pre-prepared and cooked dishes
- Serve delicious dishes

Step Three

PREPARING OUR INGREDIENTS

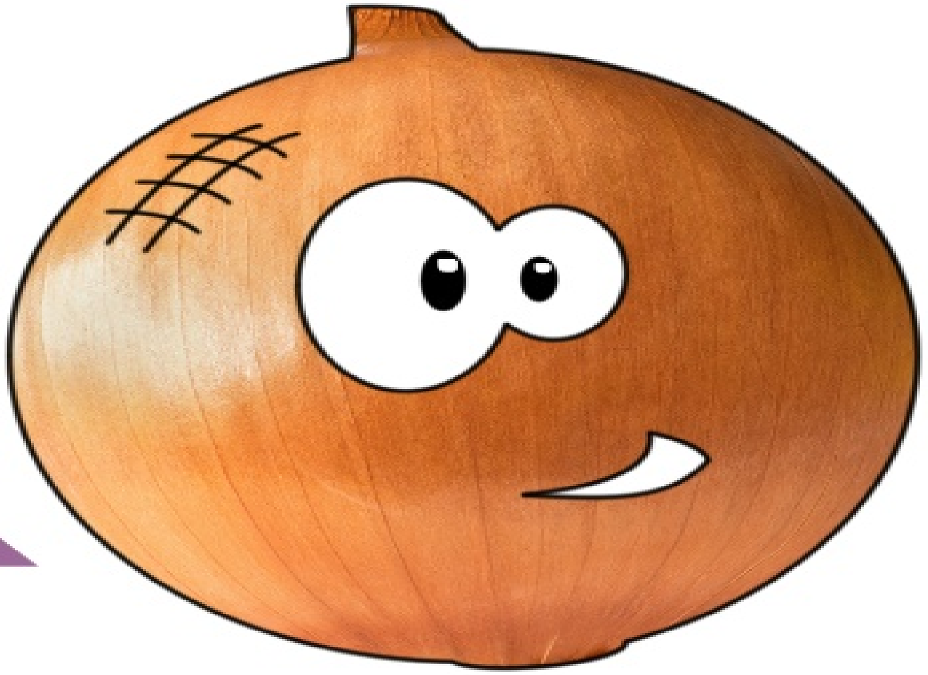
CSV Potato



- Old Classic
- All shapes and sizes
- Often dirty
- Sometimes it's a sweet potato

XML Onion

- Lots of nested layers
- Sometimes it makes you cry
- Pretty much everyone has some somewhere



Delivering the Groceries

- Large suppliers have well established delivery mechanisms
- The local store may be less reliable
- The local farm even less so

Sparking Data Preparation

- 200 million article abstracts
- Stored in Amazon S3
- XML files named by ID
- Data delivery notifications via SNS
- Variable file sizes (kB to MB)

Sparking Data Preparation

- Skewed key distribution made processing difficult
- Data pre-processing using Spark Streaming
- Data processing in batch using Spark and Parquet as an output format

Step Four

LET’S GET COOKING

Cooking our Ingredients

- Focus on sharing cooked ingredients
- Cooking in batches has its advantages
  - Verifying consistency
  - Testing and releasing batches
  - Recoverability

Cooking with Spark

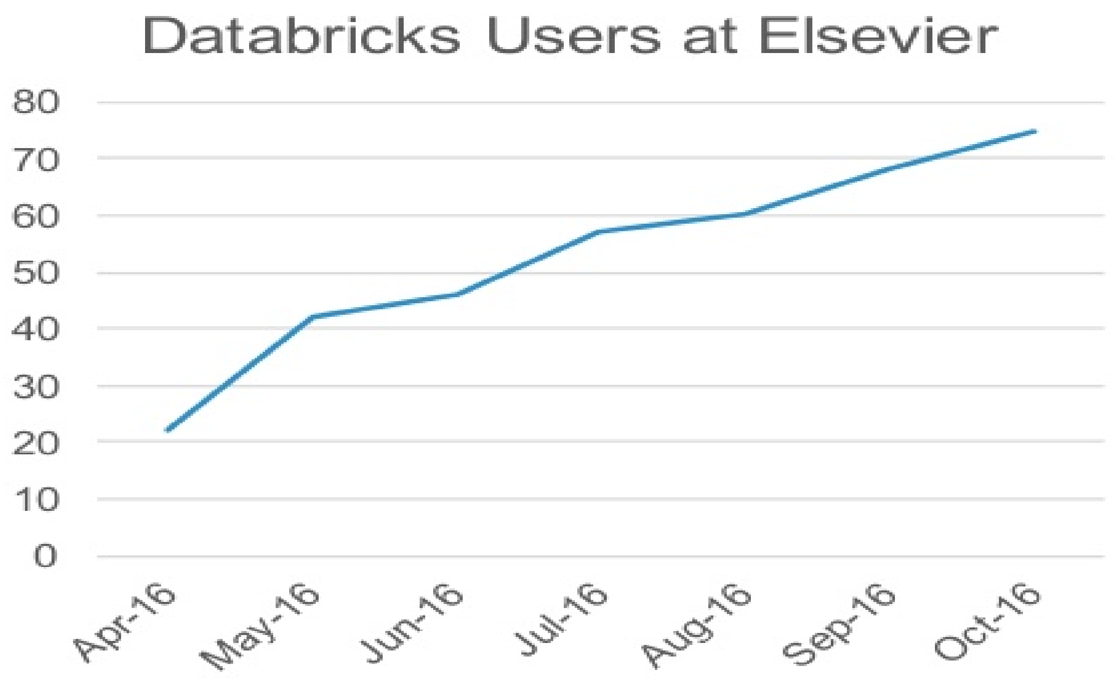
- Our data processing is exclusively done in Spark
- Cooking is mainly done in batch but we're looking at more and more streaming
- Most synthesized batch datasets stored as Parquet

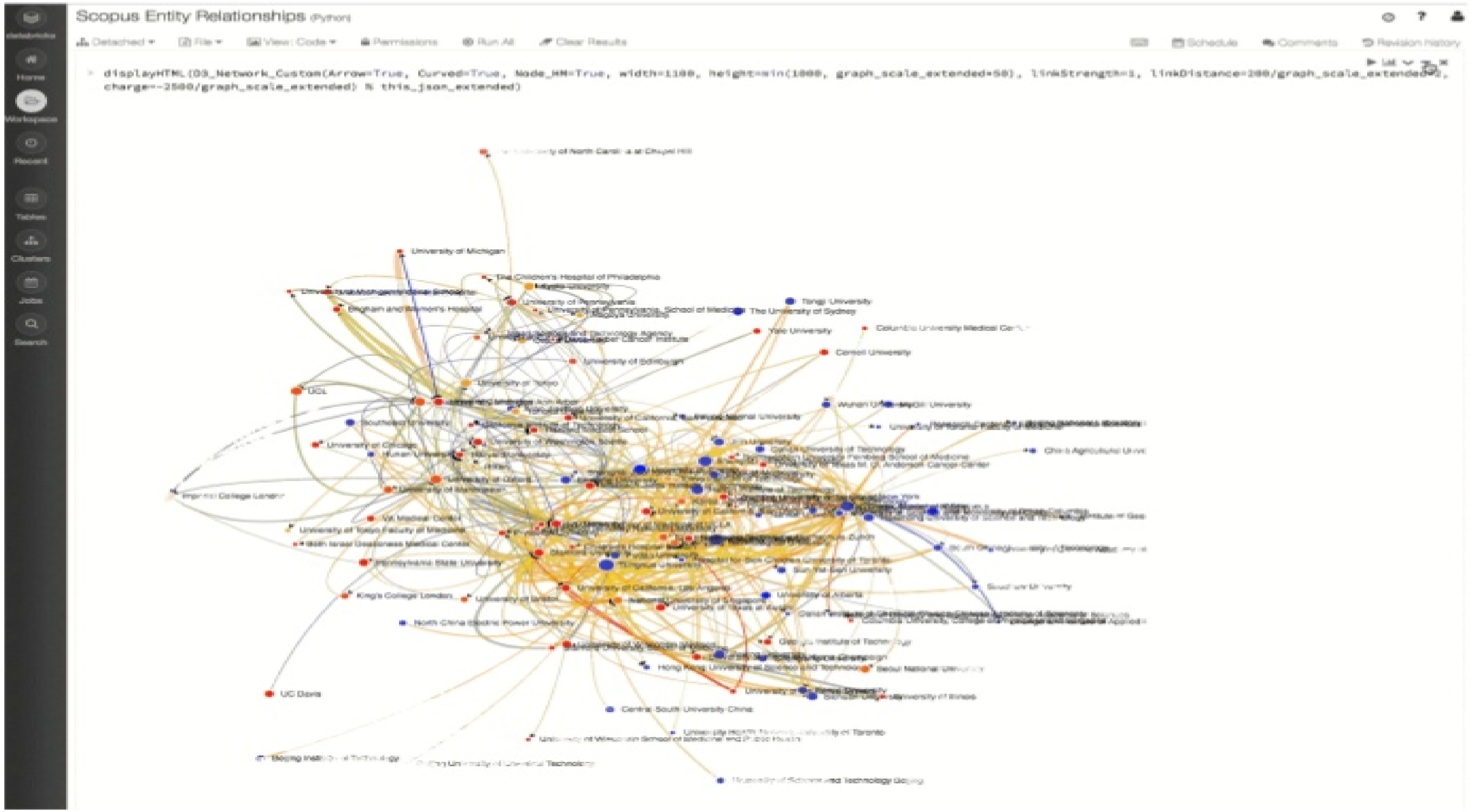
Cooking Citations

- 200 million article abstracts in XML format
- Calculation of article citations by article and author
- Transformed to key value JSON for REST API
- Mounted and shared in Databricks

Self Service with Databricks

- Data from the data lake mounted in DBFS
- Single shared cluster provided for general use
- Bespoke clusters for heavy workloads




databricks
Home
Workspace
Recent
Tables
Clusters
Jobs
Search
Detached
File
View: Code
Permissions
Run All
Clear Results
Schedule
Comments
Revision history
University of North Carolina at Chapel Hill
University of Michigan
The Children's Hospital of Philadelphia
Brigham and Women's Hospital
University of Pennsylvania
University of Pennsylvania, School of Medicine
Tongji University
The University of Sydney
Yale University
Columbia University Medical Center
Cornell University
University of Edinburgh
University of Tokyo
UCL
Ann Arbor
Southeast University
University of Chicago
Hunan University
University of Washington Seattle
Wuhan University
McGill University
China Agricultural University
Imperial College London
University of Oxford
University of Manchester
VA Medical Center
University of Tokyo Faculty of Medicine
Beth Israel Deaconess Medical Center
Pennsylvania State University
Tsinghua University
University of California, Los Angeles
King's College London
University of Bristol
North China Electric Power University
University of Alberta
University of Texas at Austin
Sun Yat-Sen University
Southwest University
University of Texas M. D. Anderson Cancer Center
Georgia Institute of Technology
Hong Kong University of Science and Technology
Seoul National University
University of Illinois
UC Davis
Central South University China
University of Wisconsin School of Medicine and Public Health
University Health Network University of Toronto
Beijing Institute of Technology
Beijing University of Chemical Technology
University of Science and Technology Beijing

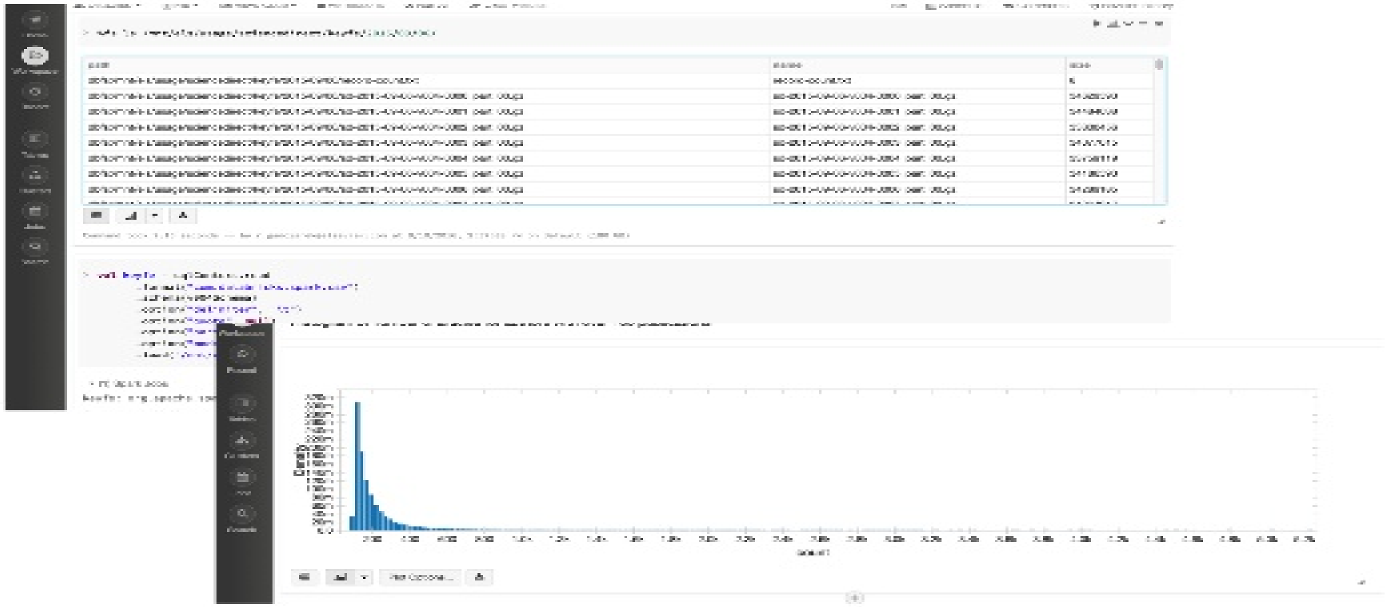
Scopus Entity Relationships (Python)

```
displayHTML(D3_Network_Custom(Arrow=True, Curved=True, Node_HM=True, width=1100, height=min(1000, graph_scale_extended*50), linkStrength=1, linkDistance=200/graph_scale_extended*2,
charge=-2500/graph_scale_extended) % this_json_extended)
```

Databricks Use Cases

- Author relationships and graphs
- Author disambiguation research
- Article recommendations
- Data profiling and exploration
- Learning

Databricks Use Cases



Next Up

HAUTE CUISINE

Our Road Ahead

- Fine-grained data access, privacy and security
- Data discovery and provenance
- Data cleansing and classification
- Enhanced operational support

THANK YOU.

e.whittick@elsevier.com

HPCC

- Developed for processing public records
- Limited flexibility
- Limited support and community
- High barrier to entry
- Difficult to get back to the original data
- Not tolerant to faults