Developing reproducible analytical pipelines for the transformation of consumer price statistics: rail fares

Matthew Price
Technical Lead

8 June 2023

Office for National Statistics

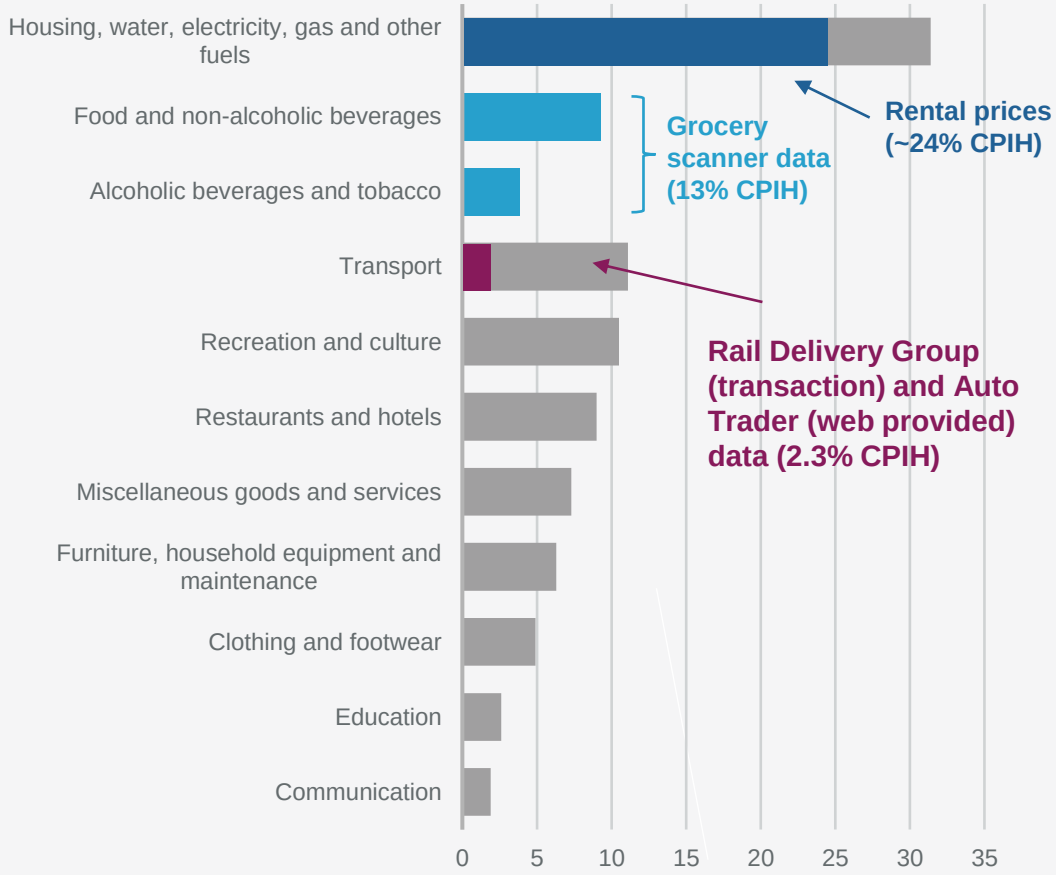
Transforming UK consumer price statistics

Continuous programme of improvements for consumer price statistics over several years beginning with rail fares

Aims:

- Obtaining **robust sources of alternative data** (scanner/web-scraped data)
- **Researching methodologies** to most effectively incorporate the data
- Developing **statistical systems** for existing and new data and methods
- Embedding new **systems** and **processes**

Primarily, new data will help us to **inform the narrative** around what is driving inflation for our users



ONS and UK Government platform strategy

- UK Government Cloud Strategy is “Cloud First” and “Cloud Agnostic”
- Aim to use “Infrastructure as a Service” and “Platform as a Service”
- ONS utilise a range of in-house and cloud platforms

Key requirements for platform

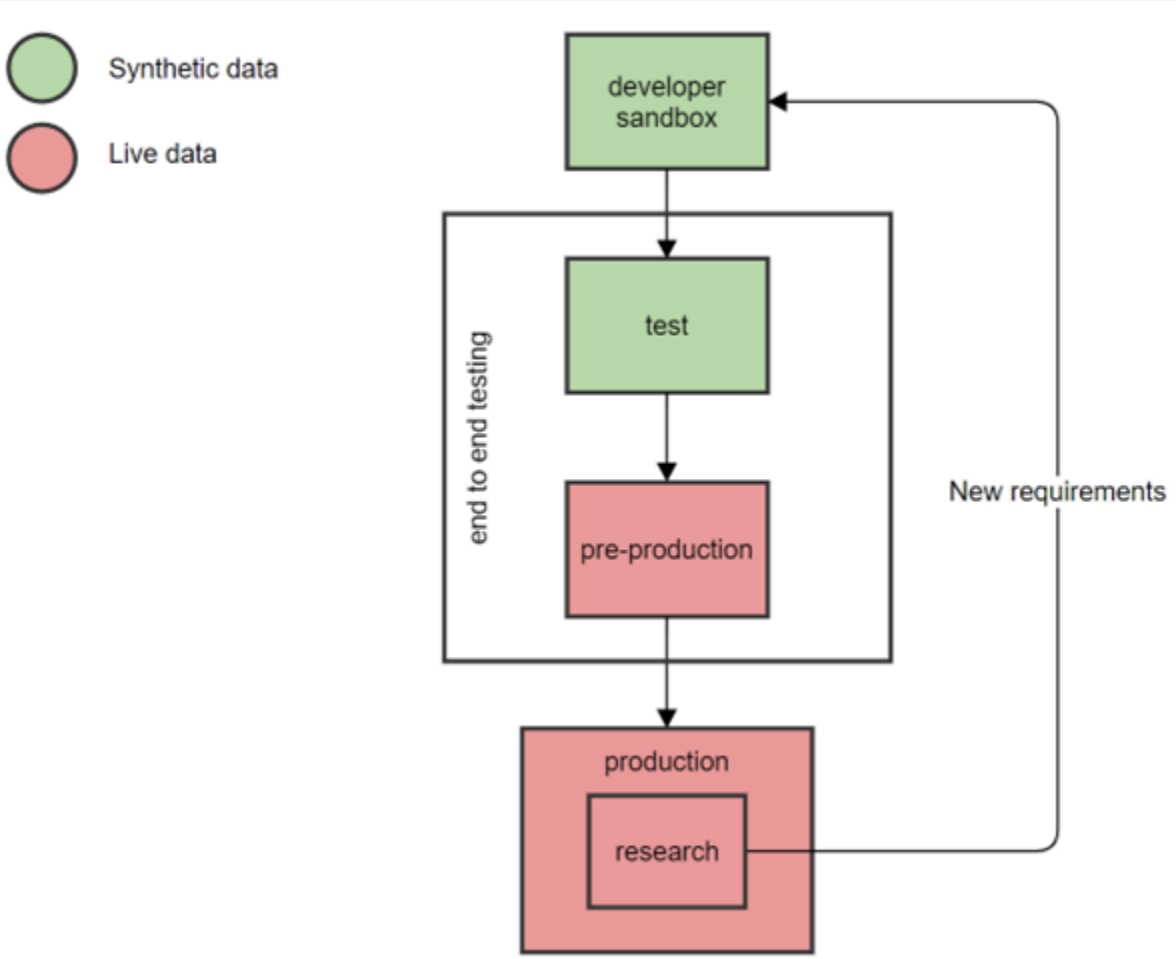
- Secure
- Scalable data storage
- Distributed, scalable compute system
- Dashboard capabilities
- Ability to host web applications
- Interactive research space

Platform environment strategy

Environment	Description	Main users	Data used	Stability
Develop	Sandbox environments where teams have full access to explore, test, and do their work.	•Software engineers •Data engineers •Infrastructure engineers	Synthetic	Not stable
Test	Test environment where all systems can work together allowing testing of individual and multiple systems.	•Testers	Synthetic	Stable
Pre-production	Environment where testing on live data can occur to ensure changes will be stable before moving into production.	•Business change team	Production (data duplicated from prod environment)	Very stable
Production	Environment where production occurs via automated scheduling of pipelines. Also, where research on data and methods can occur.	•Production team •Research team	Production (where all new data is ingested, permissions set to prevent researchers seeing “current month” data for production datasets)	Most stable

Platform environment strategy

Also aids the development of new data sources, methods and pipelines



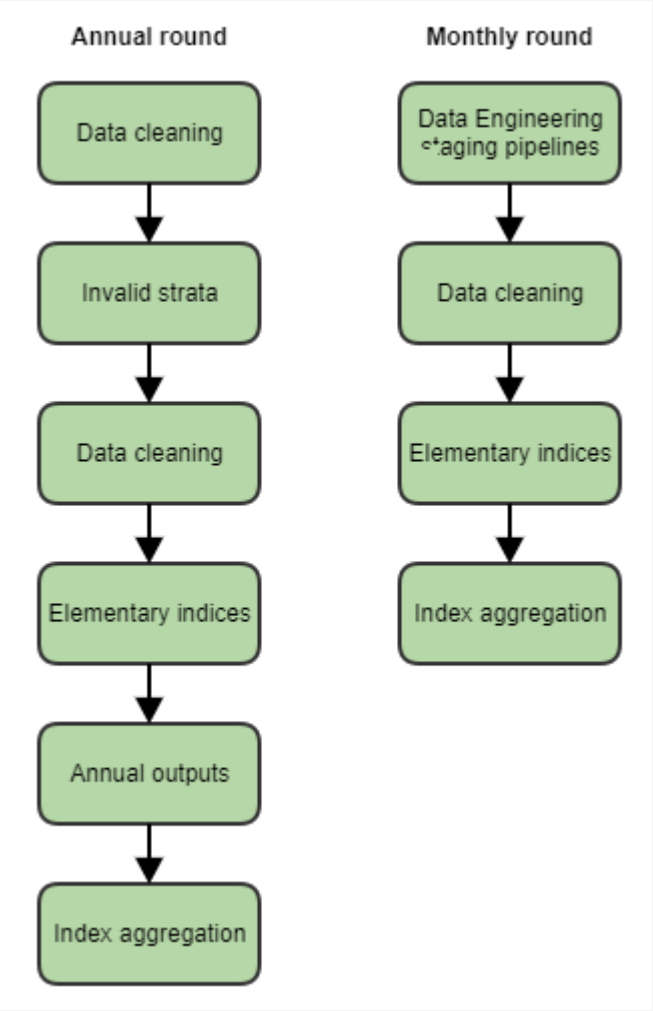
Data engineering

- Data sources are delivered as files in specific style for each supplier
- Data engineering stages data prior to processing by:
  - Virus scanning and ingesting data
  - Validating data against known metrics
  - Enrich data by applying appropriate mappers
  - Applying standardisation to each source

Data processing pipelines

- Several distinct pipelines make up the full system
  - E.g. railfares: data cleaning > elementary indices > aggregation
- Each pipeline follows the same code structure
- Controlled by a user and backend configuration file
- Each pipeline produces output for straightforward audit of data journey

Producing indices for production



- Production works with an “annual update”
- Annual round in February initialises data back series and calculates weights for next 12 months
- Monthly round (Feb – Jan) updates back series and produces the new indices

RAP – Reproducible analytical pipelines

- UK Government developed best practice principles for analytical systems
- Guidelines aim to:
  - improve the quality of the analysis
  - increase trust in the analysis by producers, their managers and users
  - create a more efficient process
  - improve business continuity and knowledge management

RAP for code

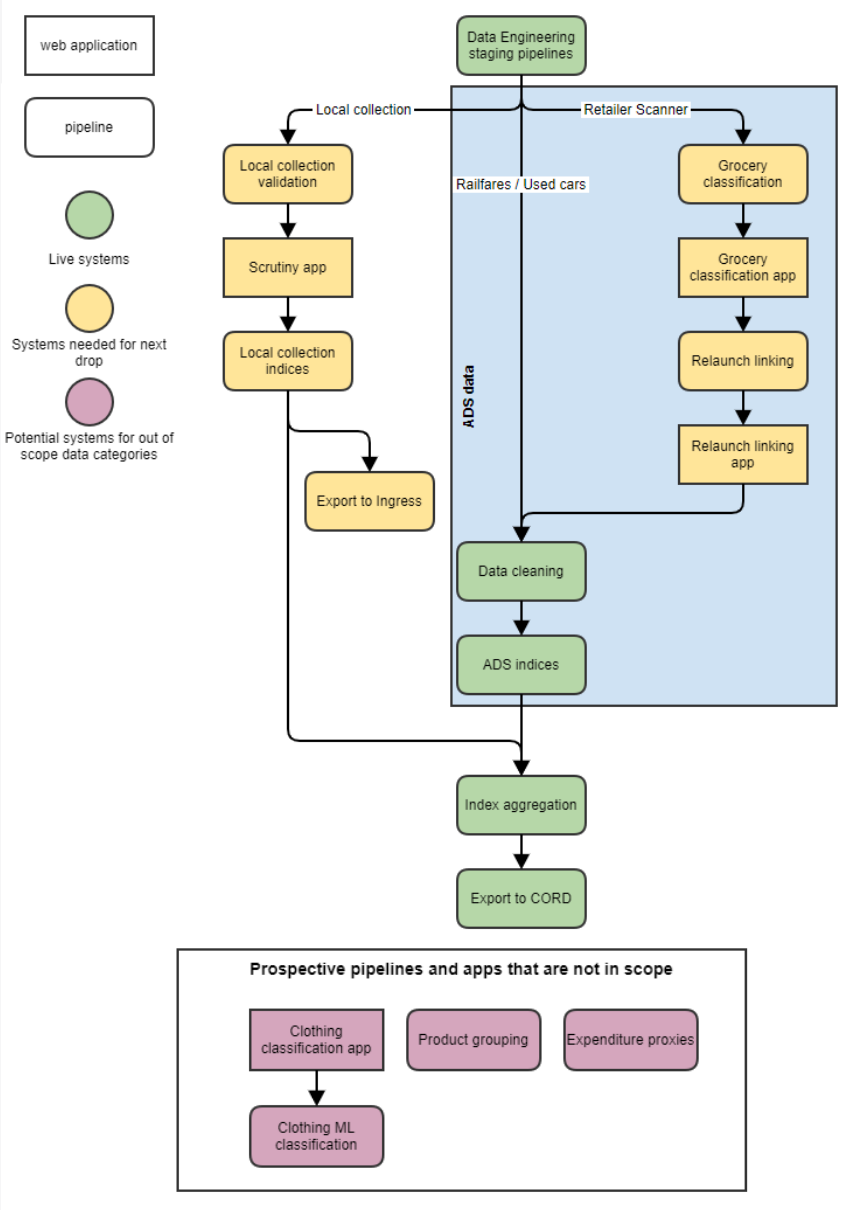
- Minimise manual steps
- Use open source software
- Peer reviewed
- Uses version control
- Open sourced code and data
- Follows department good practice
- Well documented
- Tested
- Uses CI/CD
- Appropriate logging

RAP for platform

- Automated pipelines
- Restricted access to production
- Reproducible infrastructure

Future developments timeline

- Multiyear transformation project
- Systems design template will allow scaling out of systems easily
- Rolling out new categories every year



Thank you

Office for National Statistics