Using alternative data sources to produce consumer price indices

Liam and Lefteris

Office for National Statistics

02 December 2019

Overview of the Alternative Data Sources Project

Liam Greenhough

Consumer Prices Methods Transformation

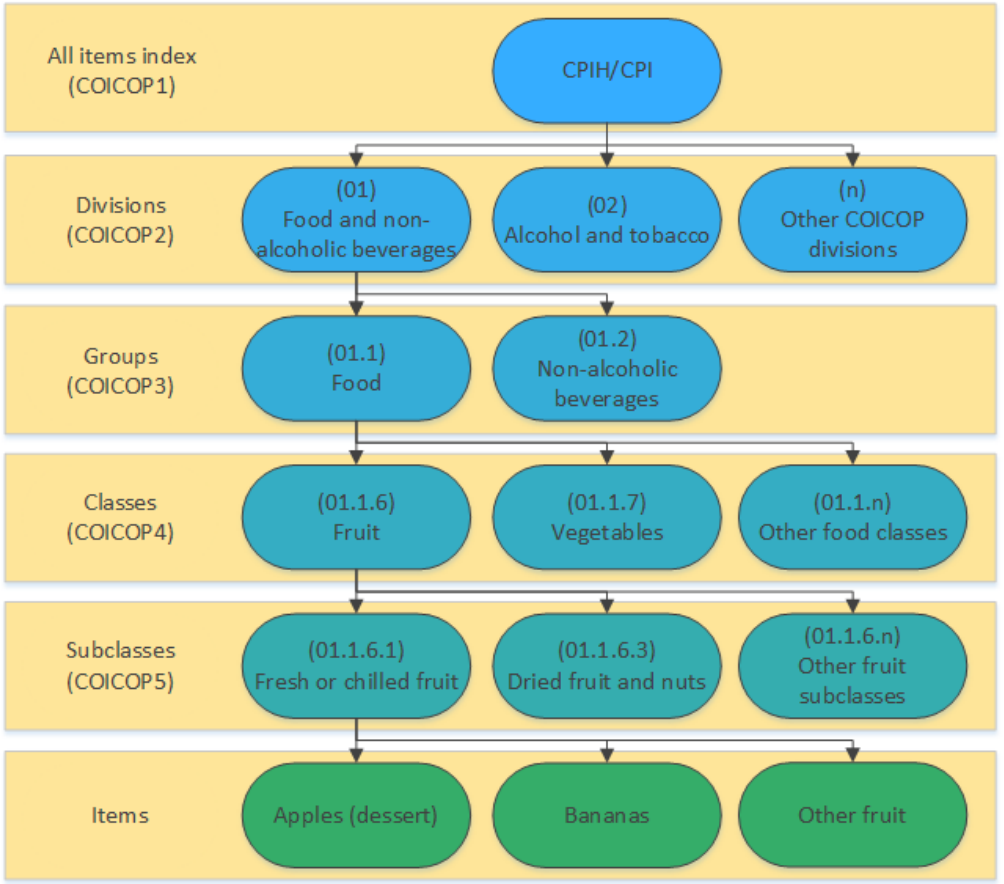
How price statistics are measured

In January, select 700 “items” to track over year. Known as the fixed basket. Each year the basket is “refreshed” to account for changing consumer behaviours.

How price statistics are measured

For each item, select a group of products to track over the year.

Each item is an aggregate – but is also a “subset” of higher aggregates.



How price statistics are measured

Collect prices of products each month. These are collected:

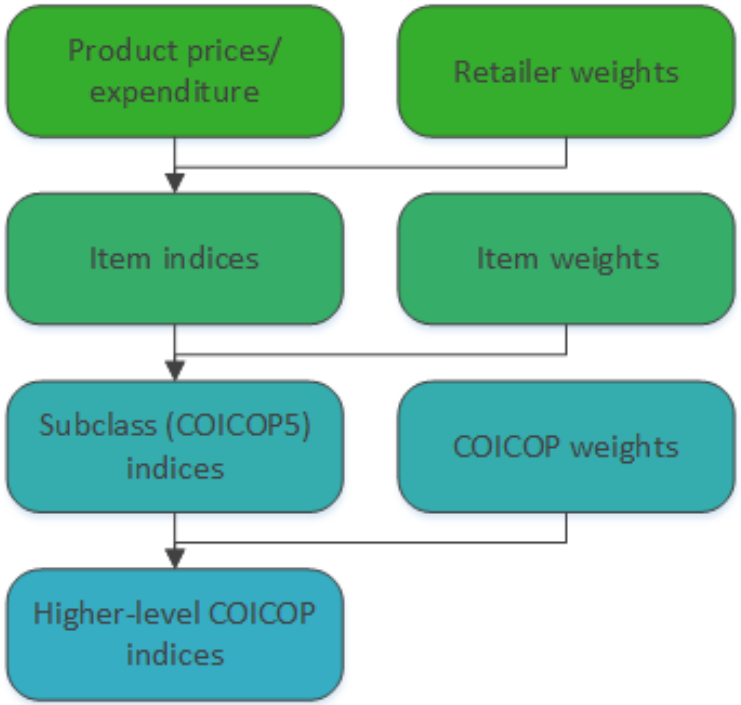
- Locally, and
- Centrally

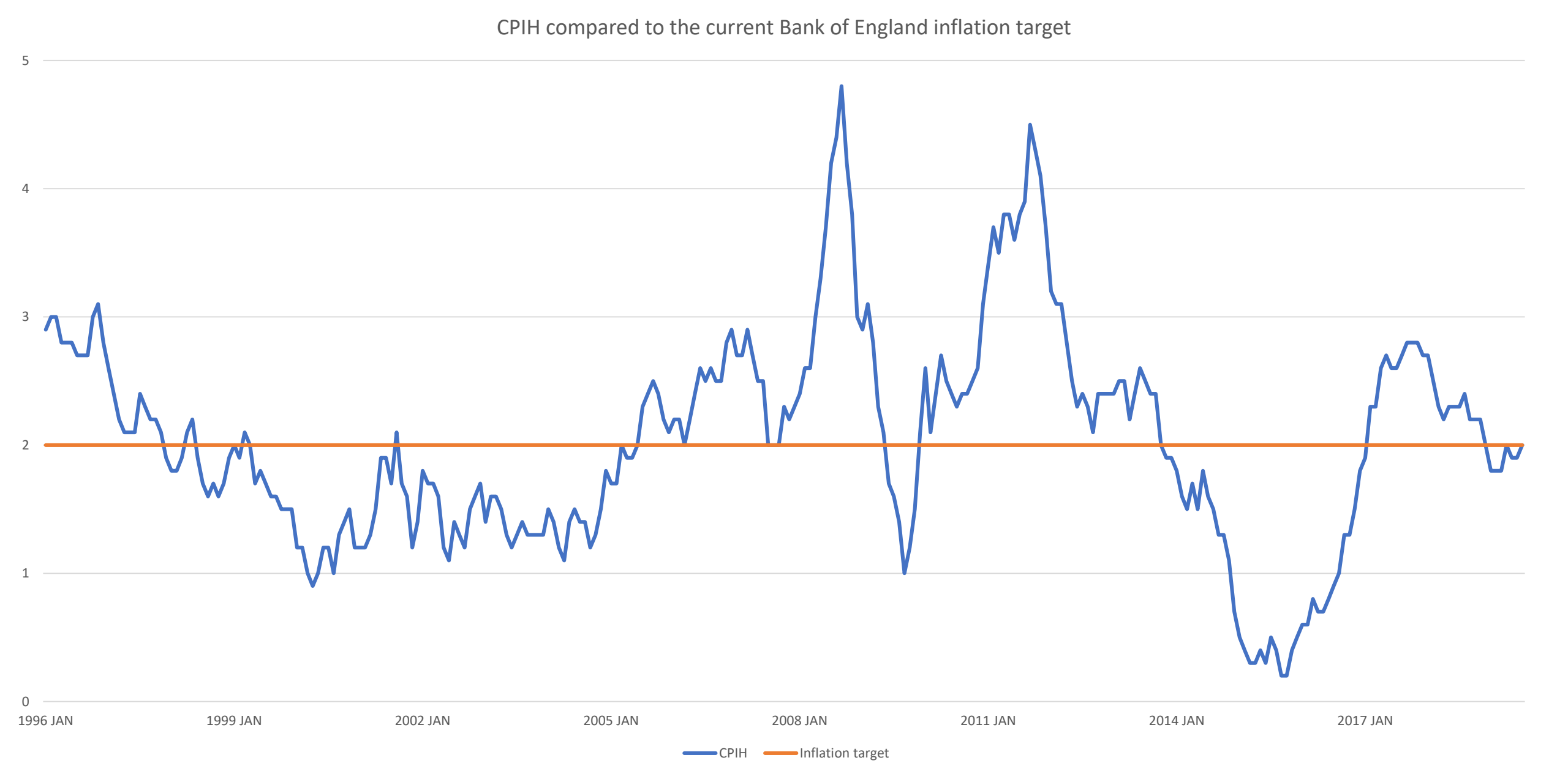
Approximately 180,000 price quotes are collected per month.

How price statistics are measured

Use index formulae to compare prices of products across months. Most common index is the Jevons.

Use weights to aggregate upwards to higher-level indices.




CPIH compared to the current Bank of England inflation target
5
4
3
2
1
0
1996 JAN
1999 JAN
2002 JAN
2005 JAN
2008 JAN
2011 JAN
2014 JAN
2017 JAN
CPIH
Inflation target

Consumer Price Statistics: Alternative Data Sources

Office for National Statistics

4 December 2019

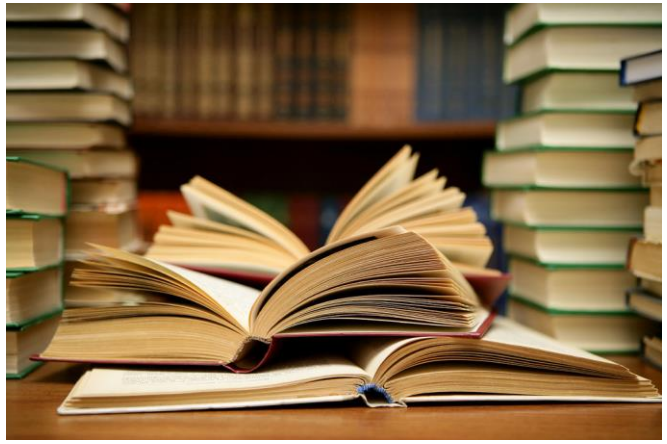
Alternative Data

Looking to implement two new data sources:

- Scanner data – transactional data from large retailers
- Web scraped data – data scraped from online retailers

Aim to use in conjunction with traditional!

Alternative Data – targeted items

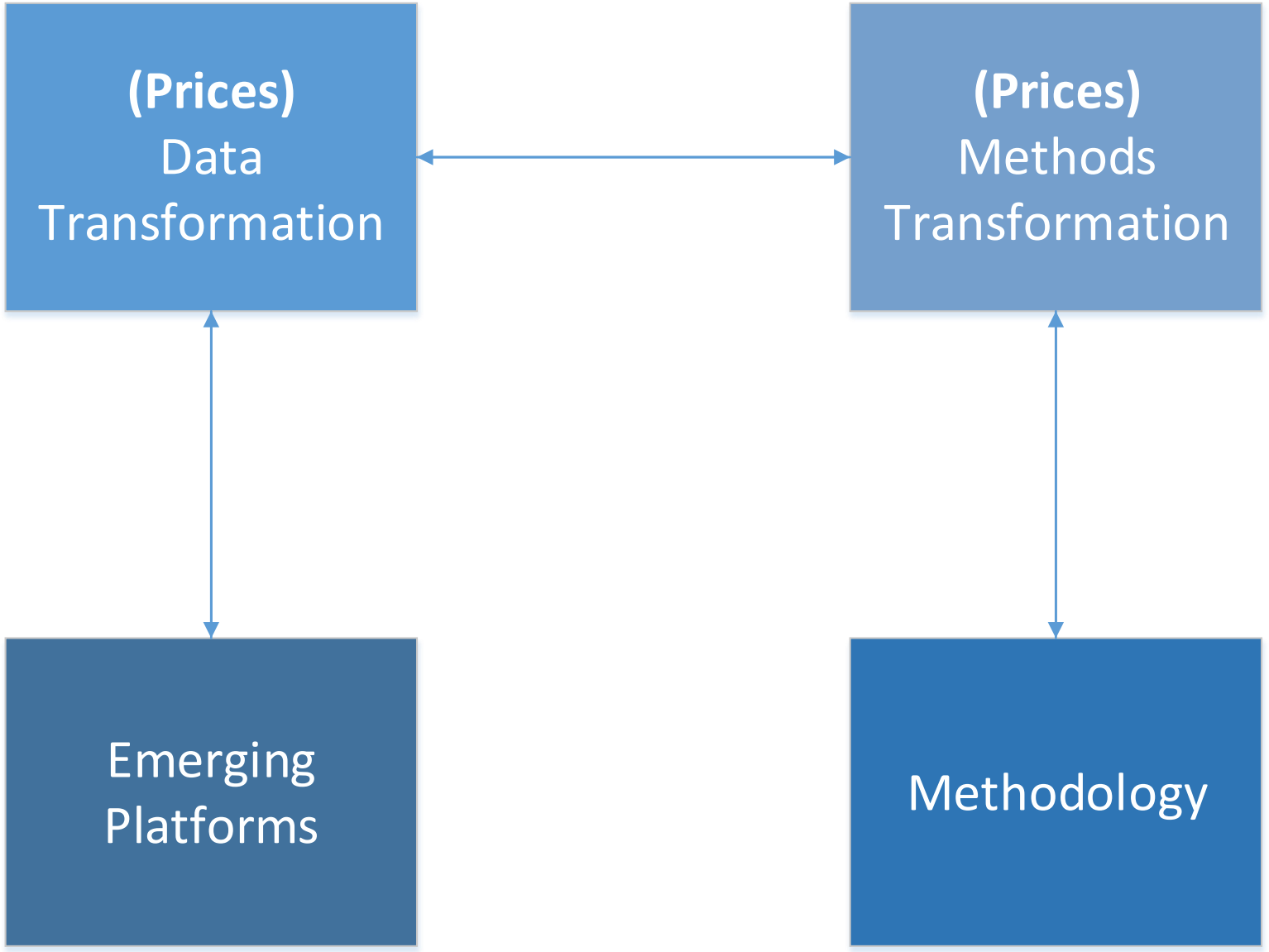


Data dimension	Traditional	Scanner data	Web scraping
Data acquisition	Manual	Automated	Automated
Completeness/scope	Sample from all retailers	All transactions (bulk) from medium to large retailers	Bulk or sample from online retailers
Metadata	Item description	Item description + limited attributes	Item description + attributes
Quantity data	N/A	Quantities sold	N/A
Timing	Single collection day	Daily	Daily

Big data

System needs to process big data.

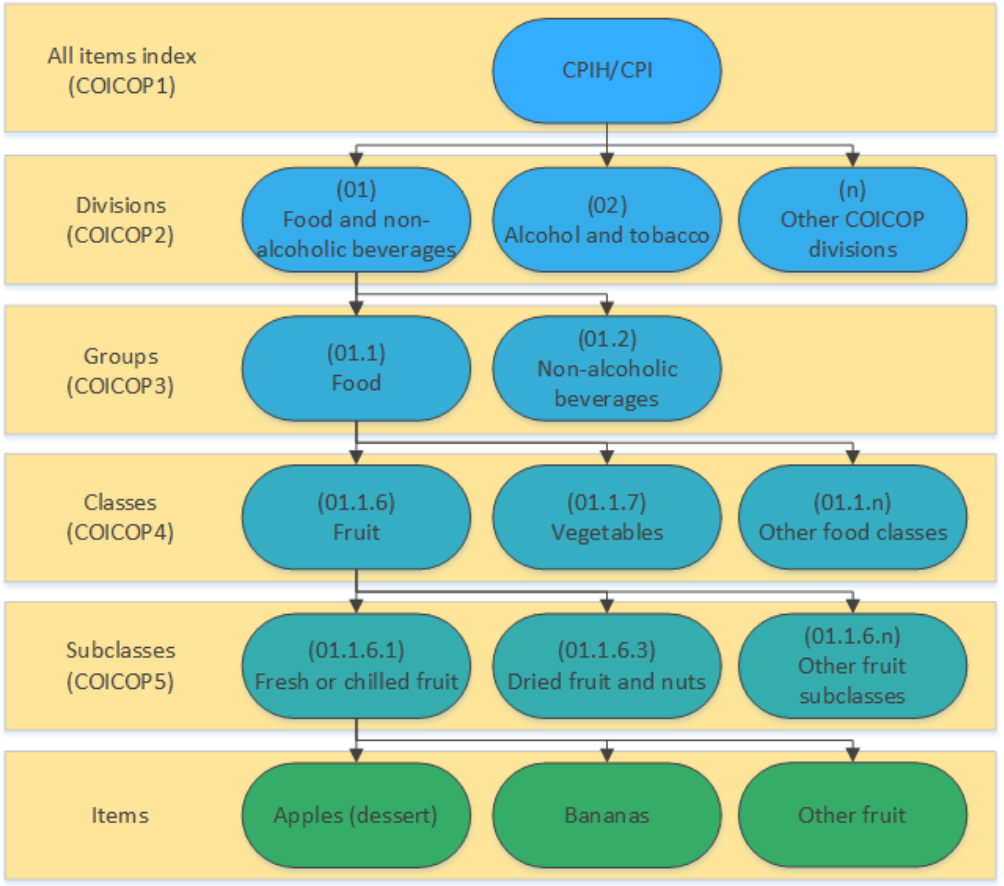
The Team


(Prices)
Data
Transformation
(Prices)
Methods
Transformation
Emerging
Platforms
Methodology

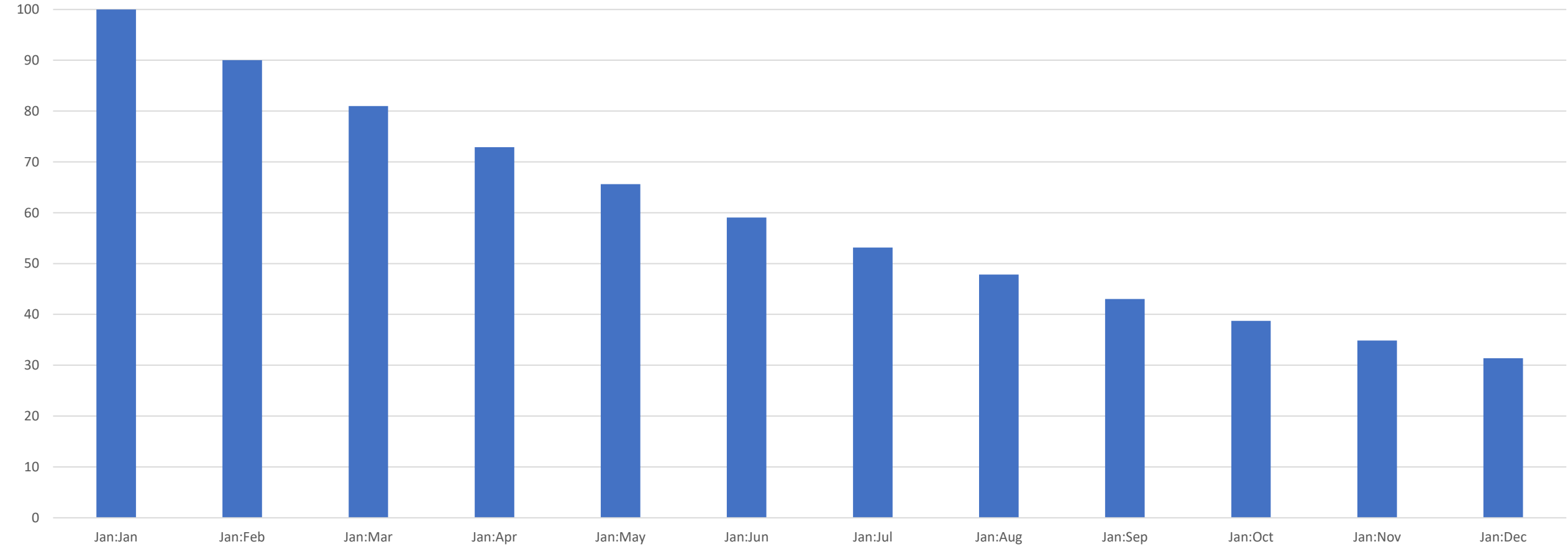
Some of the research

Scalability

Not possible to manually scrutinise big data, e.g. classification.



Product Churn – synthetic



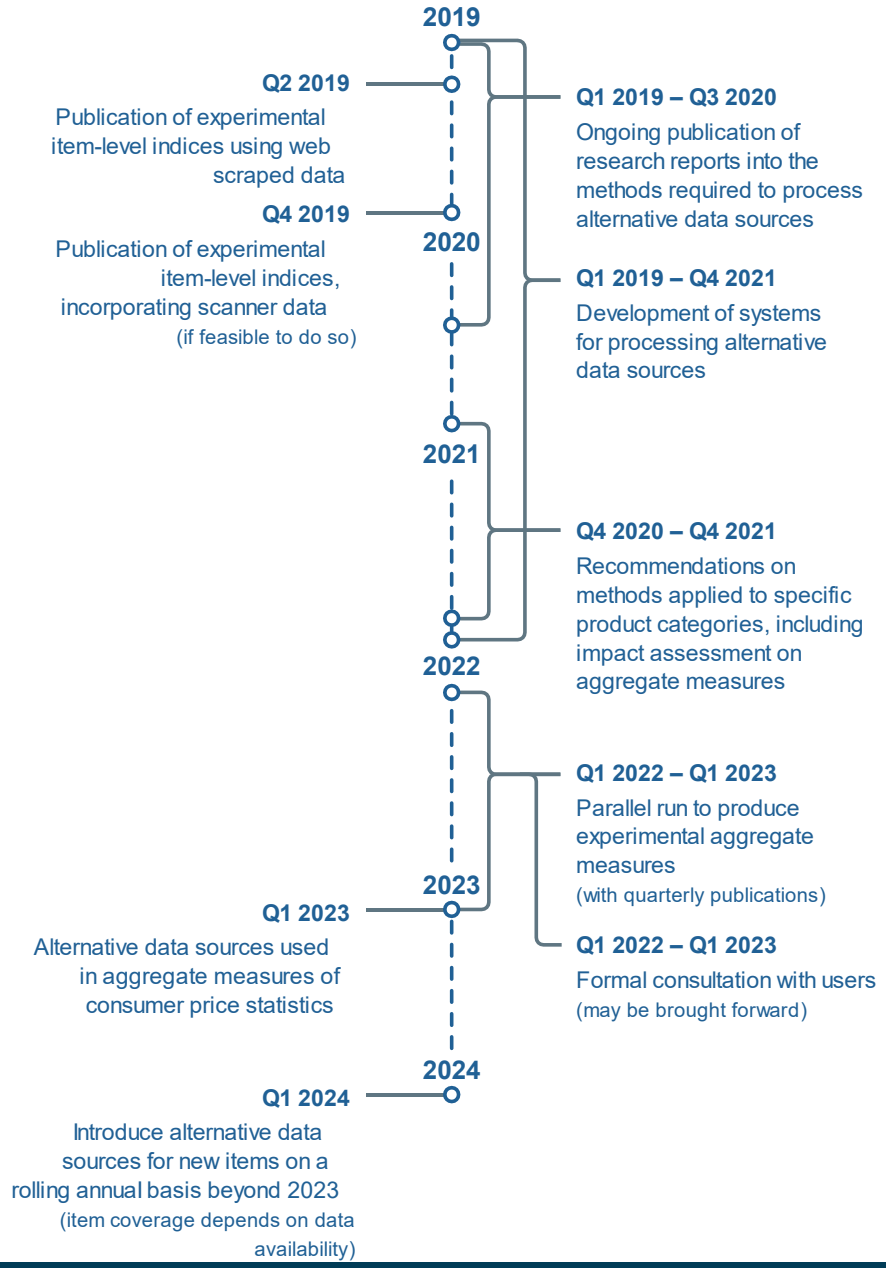
The combination problem

Lots of steps to calculate indices

Different methods at each step

Leads to many potential combinations!

Our plans


2019
Q2 2019
Publication of experimental item-level indices using web scraped data
Q4 2019
Publication of experimental item-level indices, incorporating scanner data (if feasible to do so)
2020
2021
2022
2023
Q1 2023
Alternative data sources used in aggregate measures of consumer price statistics
2024
Q1 2024
Introduce alternative data sources for new items on a rolling annual basis beyond 2023 (item coverage depends on data availability)
Q1 2019 – Q3 2020
Ongoing publication of research reports into the methods required to process alternative data sources
Q1 2019 – Q4 2021
Development of systems for processing alternative data sources
Q4 2020 – Q4 2021
Recommendations on methods applied to specific product categories, including impact assessment on aggregate measures
Q1 2022 – Q1 2023
Parallel run to produce experimental aggregate measures (with quarterly publications)
Q1 2022 – Q1 2023
Formal consultation with users (may be brought forward)

Consumer Prices Data Transformation: Development

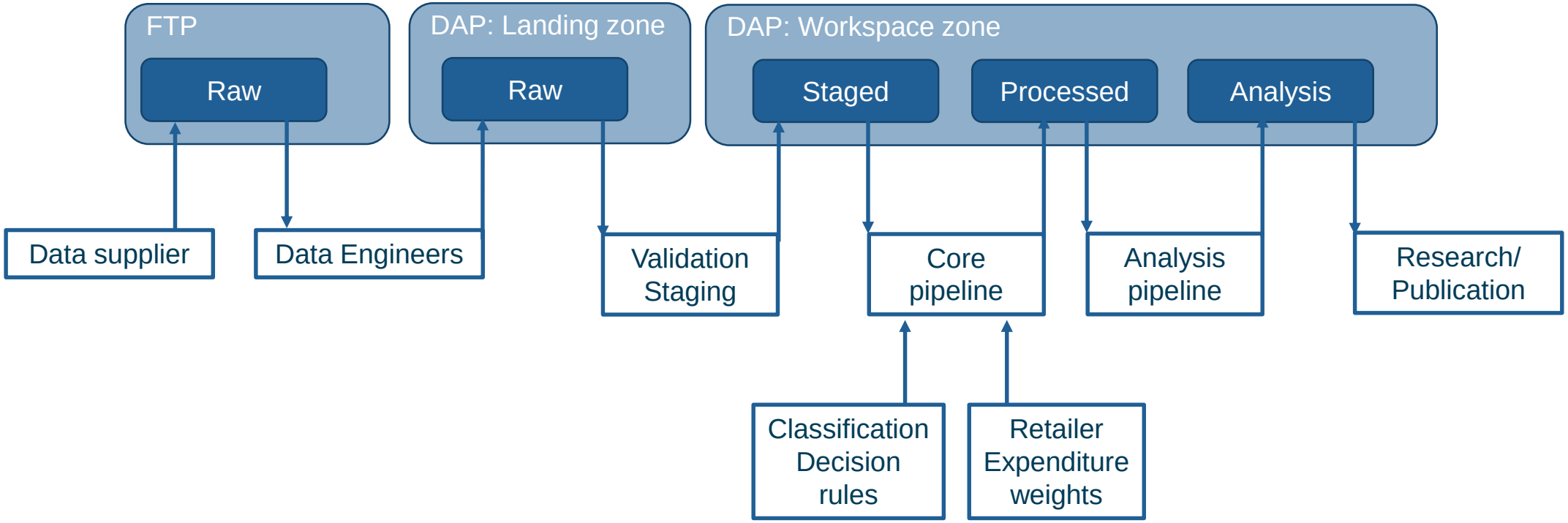
Lefteris Karachalias

Emerging Platforms Development and Support Team

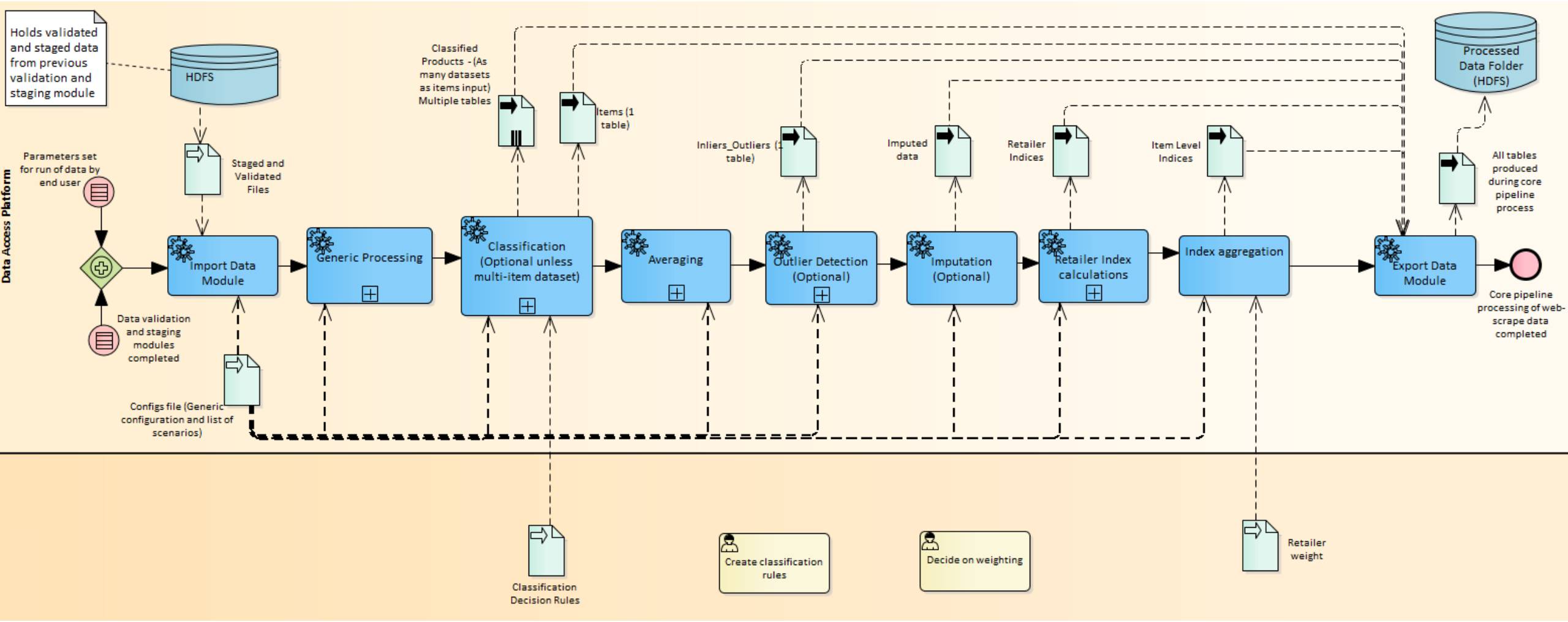
Overview

- The system
  - Overall system architecture
  - User interaction
- Development framework
  - Development project delivery team
  - Tools
  - Documentation
  - Dev&Test

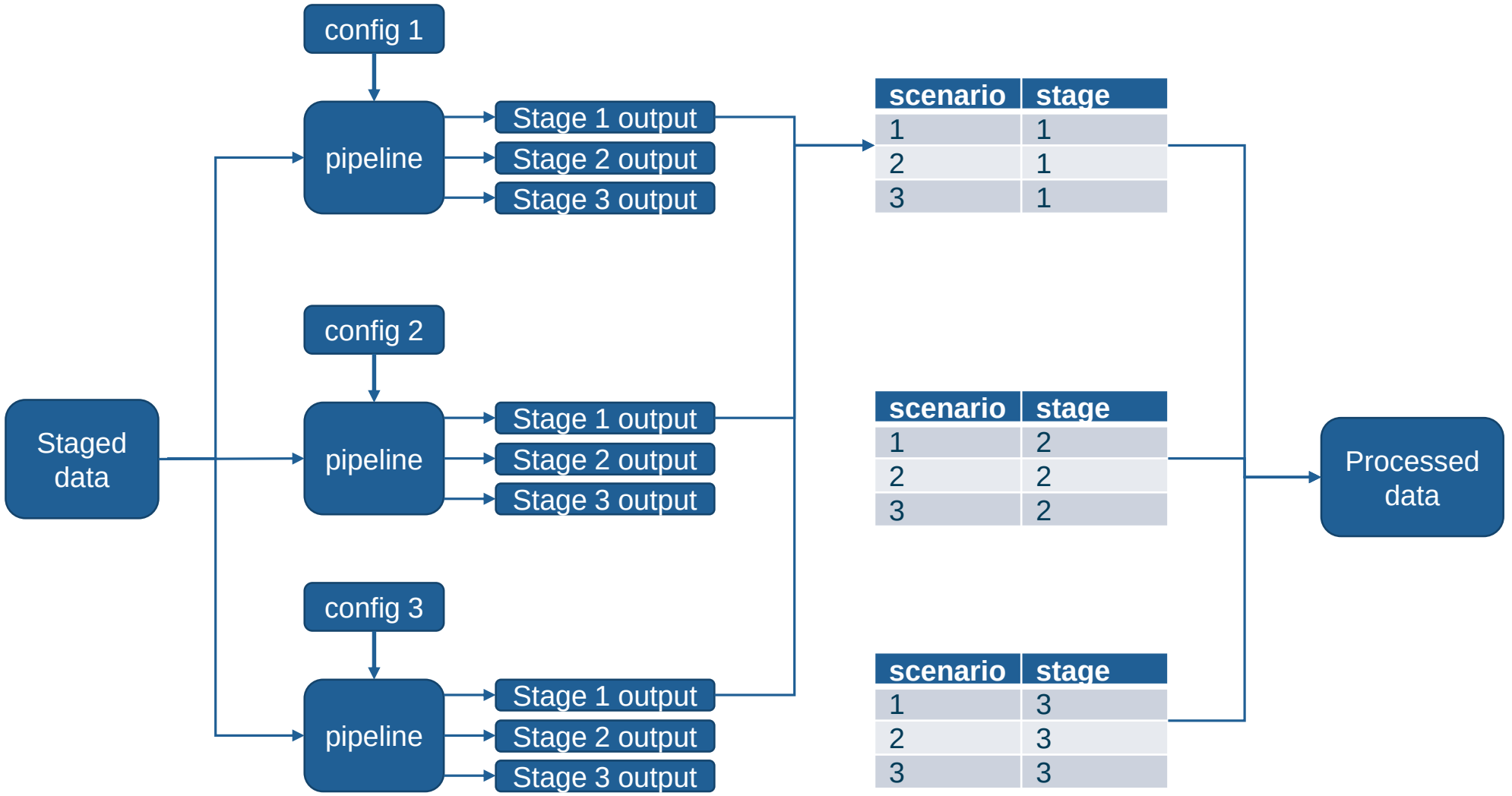
Overall system architecture



Core pipeline



Multiple configuration scenarios



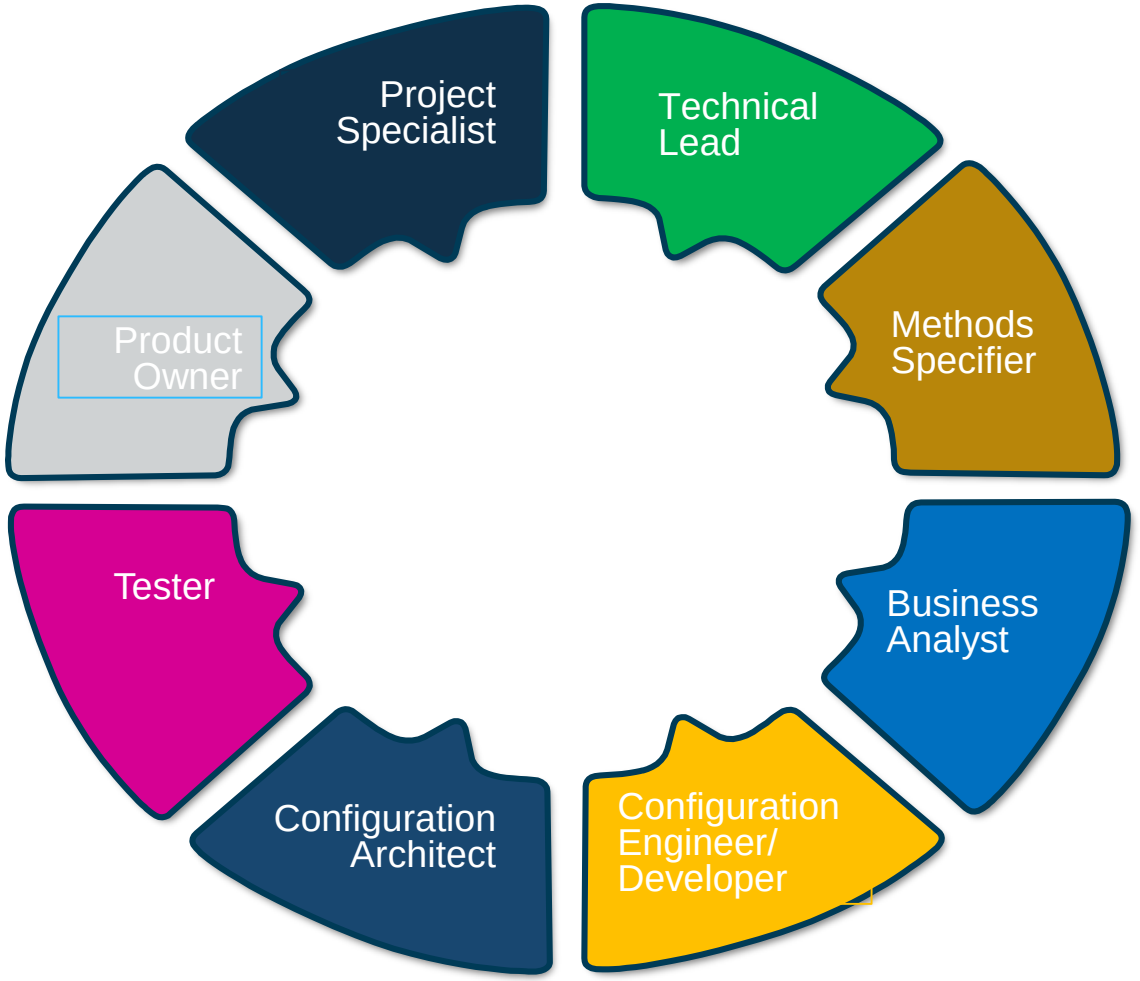
User interaction

- UI: CDSW / HUE
- Manual
- Configuration
- Mappers (BAU)
- Dashboard (cannot share VDI)
- Output tables

Development framework (1)

- Project delivery team
- Development phase: Between Discovery and Alpha phase
- Agile, Jira
- DAP, PySpark, HDFS, HIVE (sensitivity)
- Git, GitLab

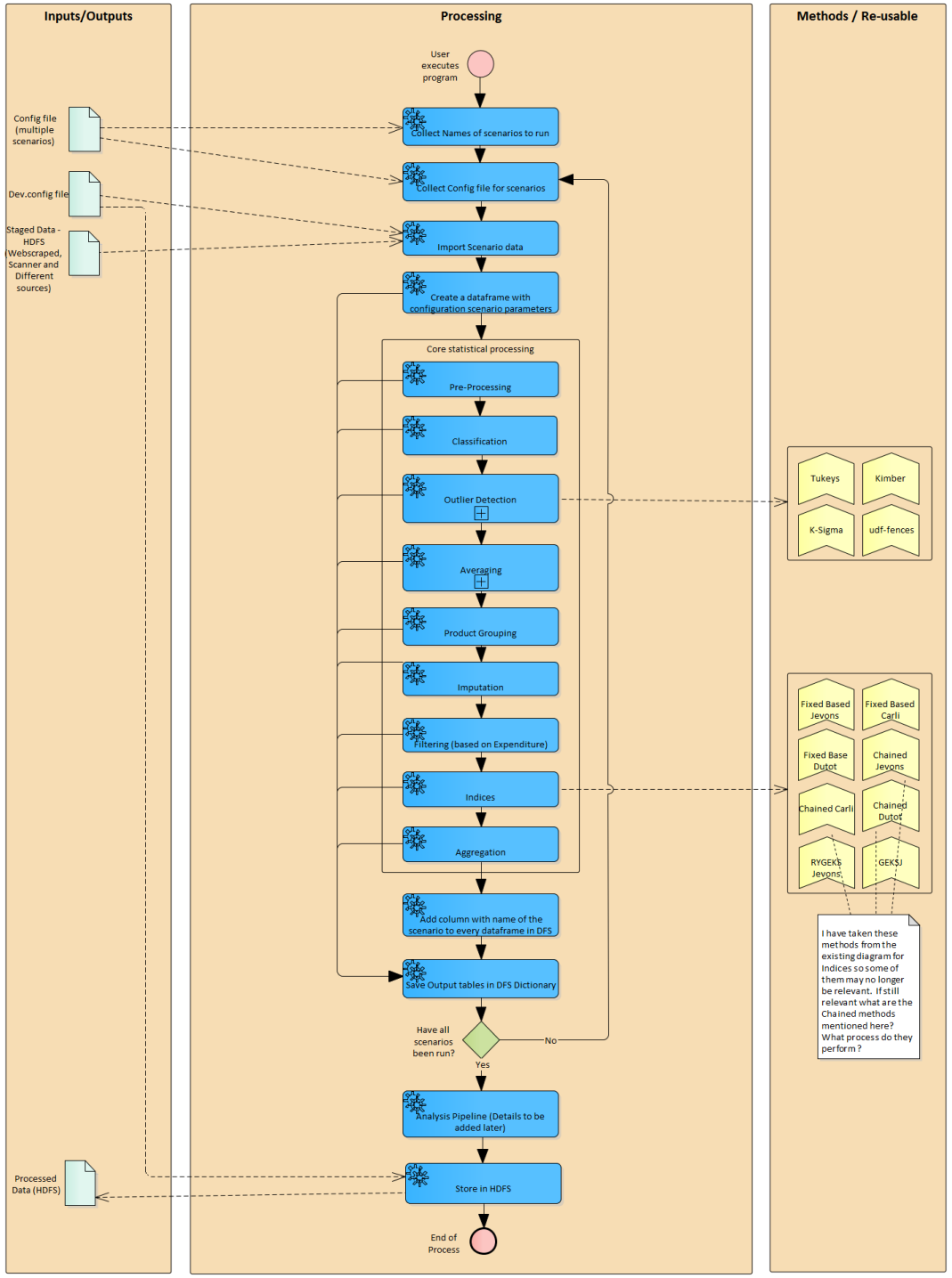
Project delivery team



Development framework (2)

- Unit testing, CI with Jenkins, UAC
- Documentation Sphinx, user manuals
- Business Analysis models, Sparx
- Business Architecture: pushing to the SML
- Synthetic data, Dev&Test, packaging

Statistical process model

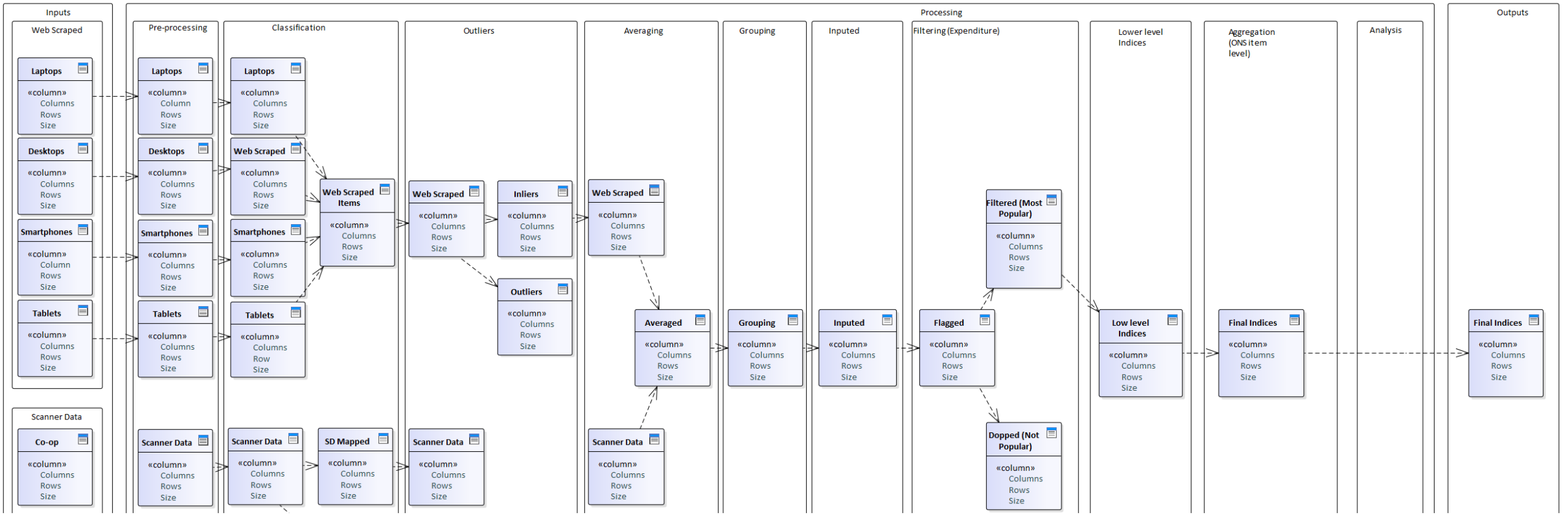


Data (journey) model

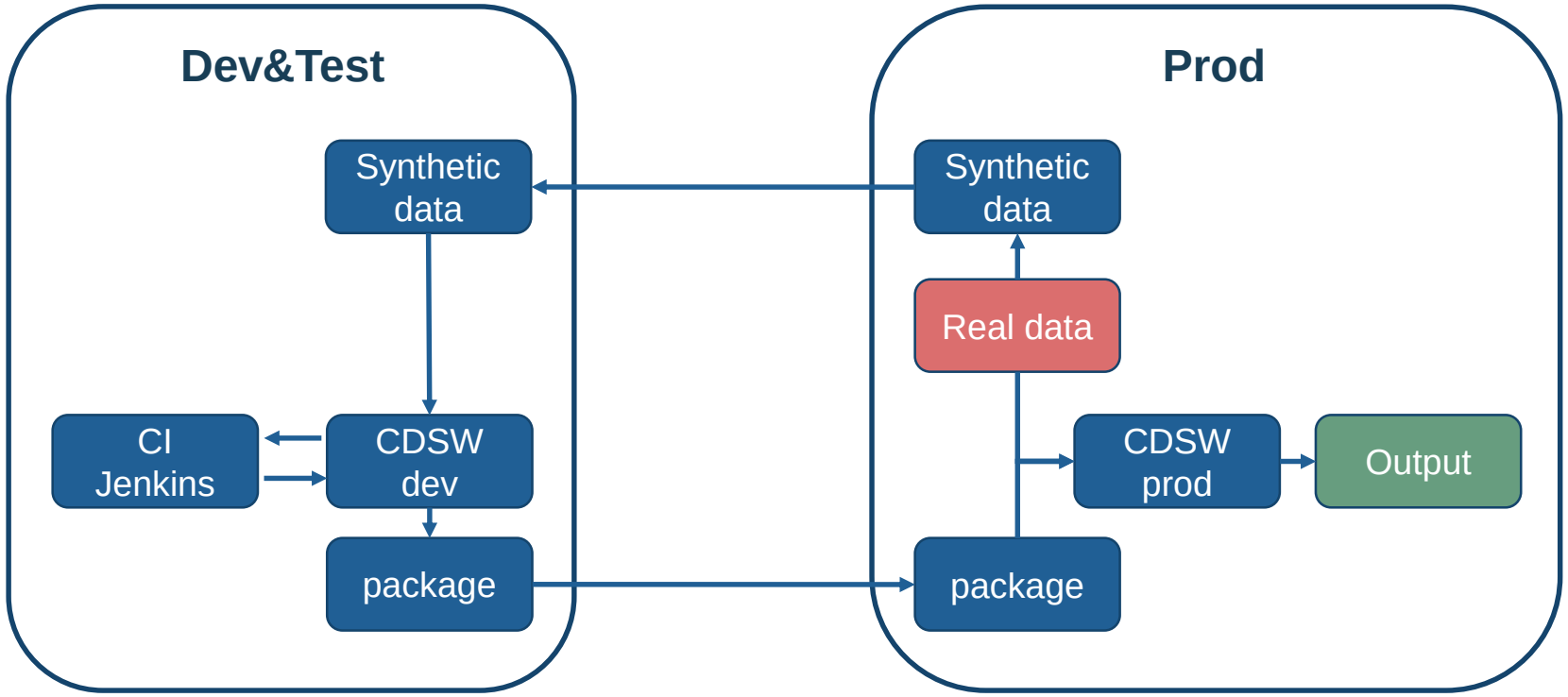
Input

Processing

Output



Dev and Test environment



Thank you!

Any questions?