Enhancing the Canadian Consumer Price Index

Challenges and lessons learned developing production processes for alternative data sources

Serge Goussev
serge.goussev@statcan.gc.ca

Presented at the Meeting of the Group of Experts on Consumer Price Indices, 2023-06-08

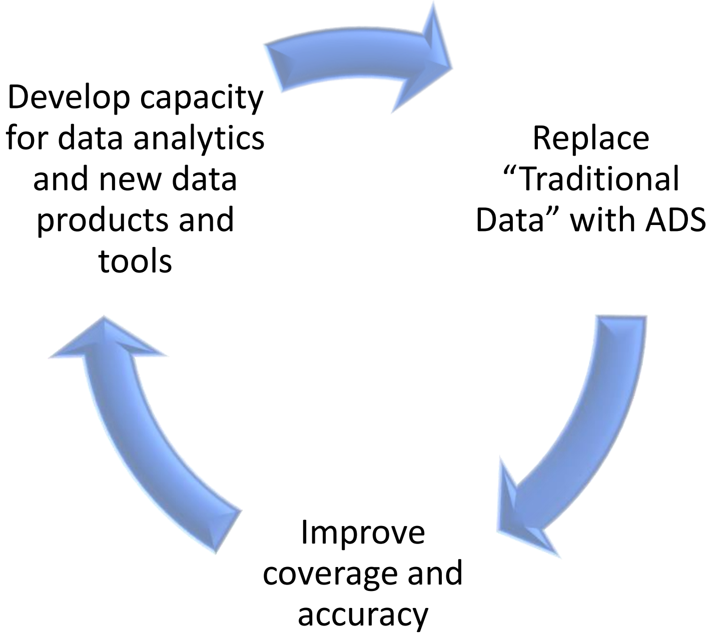
Delivering insight through data for a better Canada

Objective of the presentation

- ✓ Provide an update on Statistics Canada’s enhancements of the CPI
- ✓ Share some challenges faced, and lessons learned from Statistics Canada’s journey of adopting alternative data for the CPI
- ✓ Through the lens of business capabilities, discuss the data and application architecture we are in the process of developing to create a modern, modular, and scalable production processing platform

Business context: Enhancement vision

- Increasingly adopting alternative data sources (ADS) for a more accurate and relevant Canadian Consumer Price Index.
- Utilize Machine Learning (ML) and advanced price index methods to process near universe set of products consumed in Canada.
- Develop dynamic processing systems to be more adaptable, scalable and easier to use.
- Produce experimental series and alternative data products to support insight on price trends in Canada.

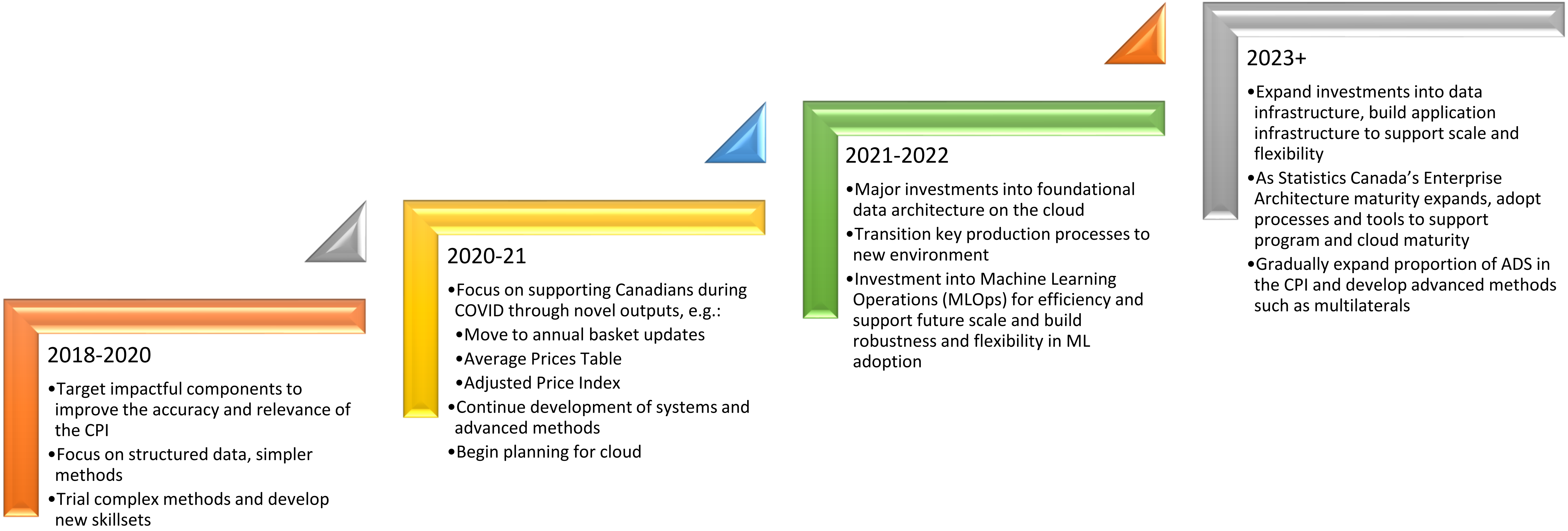


Develop capacity for data analytics and new data products and tools

Replace "Traditional Data" with ADS

Improve coverage and accuracy

Enhancement steps



2018-2020

- Target impactful components to improve the accuracy and relevance of the CPI
- Focus on structured data, simpler methods
- Trial complex methods and develop new skillsets

2020-21

- Focus on supporting Canadians during COVID through novel outputs, e.g.:
  - Move to annual basket updates
  - Average Prices Table
  - Adjusted Price Index
- Continue development of systems and advanced methods
- Begin planning for cloud

2021-2022

- Major investments into foundational data architecture on the cloud
- Transition key production processes to new environment
- Investment into Machine Learning Operations (MLOps) for efficiency and support future scale and build robustness and flexibility in ML adoption

2023+

- Expand investments into data infrastructure, build application infrastructure to support scale and flexibility
- As Statistics Canada's Enterprise Architecture maturity expands, adopt processes and tools to support program and cloud maturity
- Gradually expand proportion of ADS in the CPI and develop advanced methods such as multilaterals

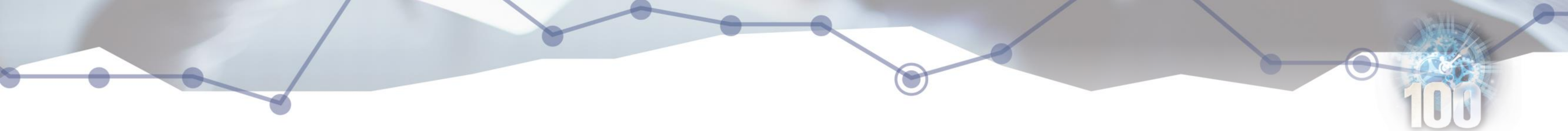
Limitations and challenges faced during initial phases

Technical

- Scale of the data considerable (billions of rows, millions of unique products, dozens of terabytes). Acquisition leads to exponential increase of data volume
  - High processing capacity key
- Infrastructure and powerful compute resources needed to support production
- Infrastructure and software for R&D even more pronounced
  - Maintain experiment and provenance
  - Horizontal access to data stores
  - Minimization of copies for iterative processes
- Access to software and hardware for robust Data Science & Machine Learning stack
- Automation and orchestration tools

Organizational

- Investing and upskilling staff, increasing technical skills
- Change management as data scale and approaches require adoption of new processes and tools
- Coordination within the program and agency for effective use of data
- Data governance framework to ensure accessibility control



Principles of planned CPI architecture & how to mitigate these issues

Transparency in production and R&D processes

Focus on enabling the development of reproducible pipelines for production or R&D

Ability to register models and datasets (including appropriate use of metadata for discovery and interoperability)

Version control of code and orchestration pipelines

Horizontal access to the data for research, development, and analytics

‘Break down the data silos’

Provide analytical insight from all data sources

High processing capacity

Ability to process large data at scale, and scale down upon run completion

Adoption of appropriate tools, open standards, solutions

Access modern tools for R&D or production, especially critical for Data Science work

Security

Maintain access control for datasets throughout all environments and their entire lifecycle, not just at source

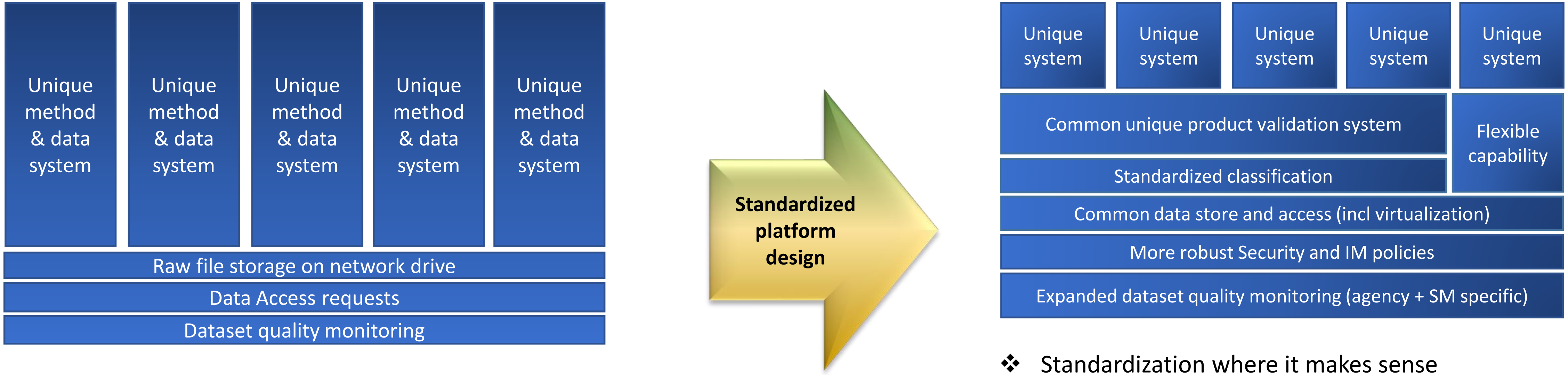
Auditability of access

Cost-effectiveness

Elastic processing capacity

Cost transparency

Paradigm shift: from on-premises to cloud processing



Unique method & data system

Unique method & data system

Unique method & data system

Unique method & data system

Unique method & data system

Raw file storage on network drive

Data Access requests

Dataset quality monitoring

Standardized platform design

Unique system

Unique system

Unique system

Unique system

Unique system

Common unique product validation system

Flexible capability

Standardized classification

Common data store and access (incl virtualization)

More robust Security and IM policies

Expanded dataset quality monitoring (agency + SM specific)

- Siloed monolithic systems (data + method specific)
- Duplicated processes
- Added complication slows down adoption and lowers transparency
- Need to reproduce what others have done

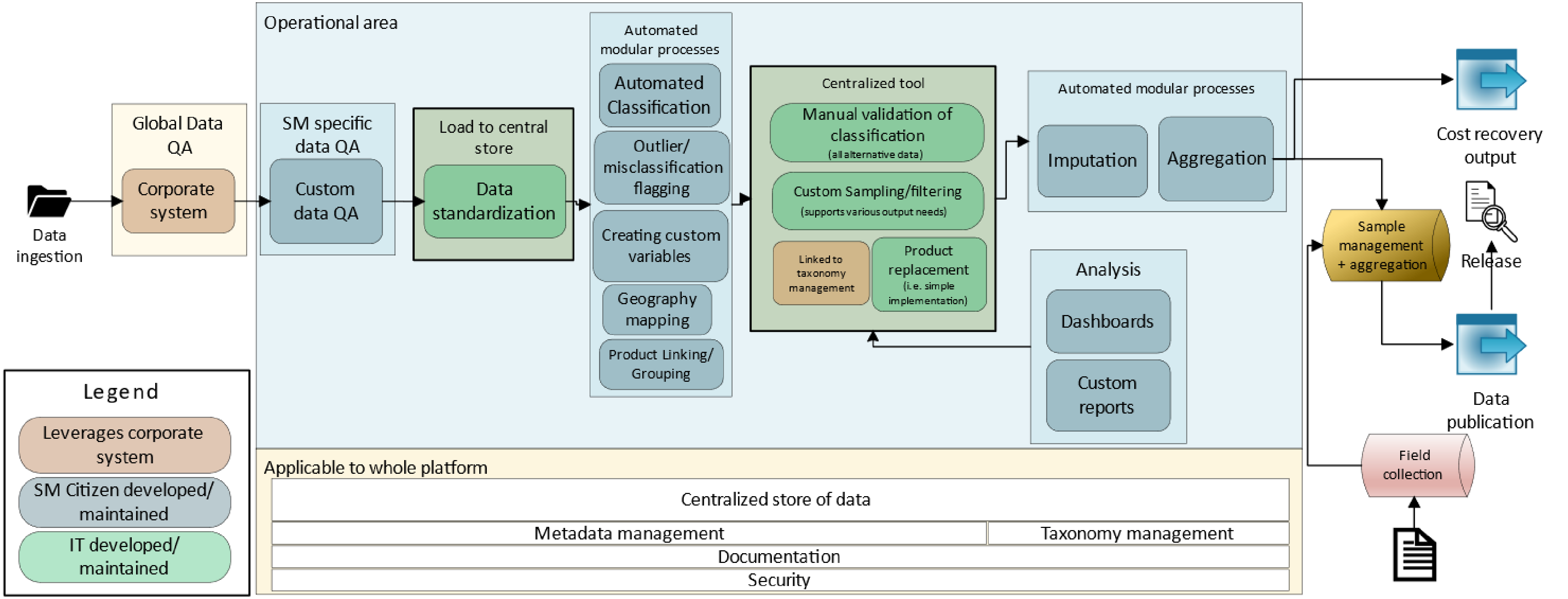
- Standardization where it makes sense
  - Standardization of common tasks for greater efficiency
- Flexible capabilities following a standard framework to support multiple outputs
- Enable adaptability to change through modularity
- Enable lighter and faster R&D

Standardized capabilities for price indices

Capability	Short summary
Dataset Quality Assurance prior to use	Each dataset received must be validated to make sure it can be used in production
Data standardization	Standardize diverse product information into standard prices and weights
Product Linking/Grouping	Related products need to be linked to align with product relaunch, or grouped when the granularity of data is lower than the homogeneity of a product
Creating new variables	To support imputation and quality adjustment methods, detect product attributes/variables
Imputation	Impute missing data, or perform more complex quality adjustment
Geography mapping	Location of each banner needs to be mapped to geography class
Automated Classification (and MLOps processes)	Assignment of category code that is utilized within the aggregation taxonomy of the CPI. For Machine Learning (ML) methods, MLOps processes need to be developed to support robust ML use in production
Outlier/misclassification flagging	Flag impactful and outlier records for manual validation (subsequent step)
Manual validation of classification	Validation that impactful records for quality control of price indices
Custom sampling/filtering	During production, only strata necessary for aggregation needs to be selected
Aggregation	Aggregate and calculate price statistics as per the taxonomy structure of the NSO
Analysis	Conduct analysis to understand and explain reason for price movements seen

Designing agility through the lens of standard capabilities

Capabilities for CPI production, focus on alternative data



Data ingestion

Global Data QA
Corporate system

Operational area

SM specific data QA
Custom data QA

Load to central store
Data standardization

Automated modular processes
Automated Classification
Outlier/ misclassification flagging
Creating custom variables
Geography mapping
Product Linking/ Grouping

Centralized tool
Manual validation of classification (all alternative data)
Custom Sampling/filtering (supports various output needs)
Linked to taxonomy management
Product replacement (i.e. simple implementation)

Automated modular processes
Imputation
Aggregation

Analysis
Dashboards
Custom reports

Cost recovery output

Sample management + aggregation

Release

Data publication

Field collection

Applicable to whole platform
Centralized store of data
Metadata management
Taxonomy management
Documentation
Security

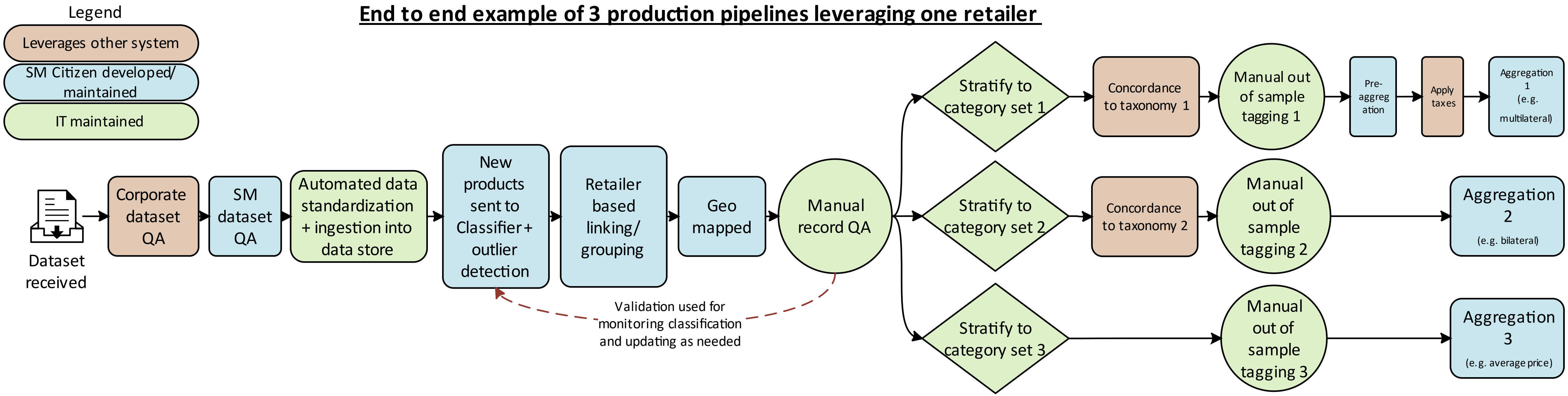
Legend
Leverages corporate system
SM Citizen developed/ maintained
IT developed/ maintained

Key takeaways:

- Alternative data requires a separate pipeline from traditional field collection
- A balance can be struck between IT and citizen development – the ‘rate of change' concept important in determining which capabilities can be designed by citizen developers (programs that change frequently, from monthly to every several years). Stable long-term systems and the foundational architecture are appropriate to be built by IT (due to low likelihood of changing the whole system and considerable complexity)

Example pipeline

End to end example of 3 production pipelines leveraging one retailer


Legend
Leverages other system
SM Citizen developed/ maintained
IT maintained
Dataset received
Corporate dataset QA
SM dataset QA
Automated data standardization + ingestion into data store
New products sent to Classifier + outlier detection
Retailer based linking/ grouping
Geo mapped
Manual record QA
Validation used for monitoring classification and updating as needed
Stratify to category set 1
Concordance to taxonomy 1
Manual out of sample tagging 1
Pre-aggregation
Apply taxes
Aggregation 1 (e.g. multilateral)
Stratify to category set 2
Concordance to taxonomy 2
Manual out of sample tagging 2
Aggregation 2 (e.g. bilateral)
Stratify to category set 3
Manual out of sample tagging 3
Aggregation 3 (e.g. average price)

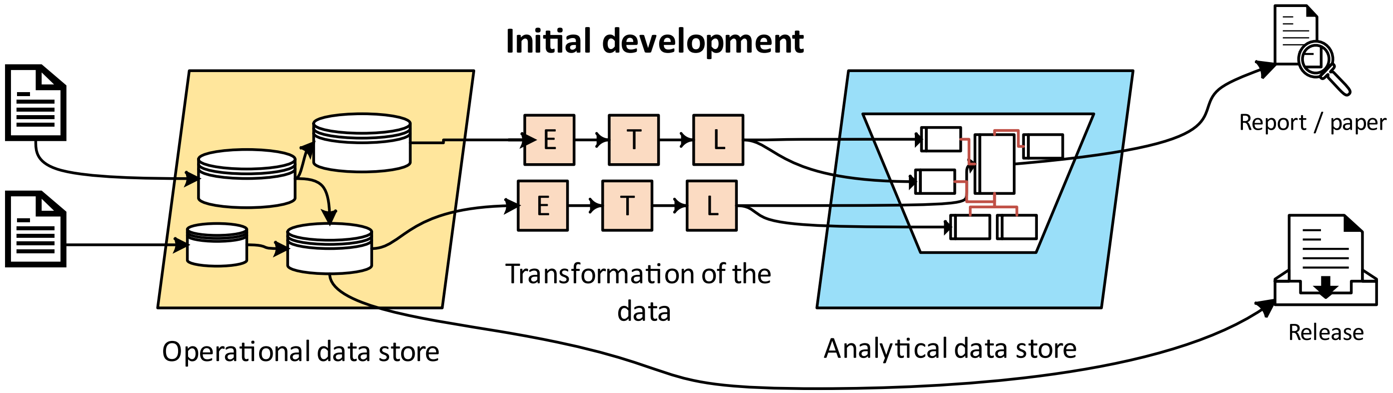
Key takeaways:

- Not all capabilities needed all at once – each retailer will need a different set of production pipelines to produce several outputs.
- Development of capabilities in a modular way allows interchangeability as methods need to evolve or to incorporate improvement in technology or tools
- Transparent development enables trust and partnership between statistical programs in the agency, allowing robust integration of one data source for multiple statistical outputs

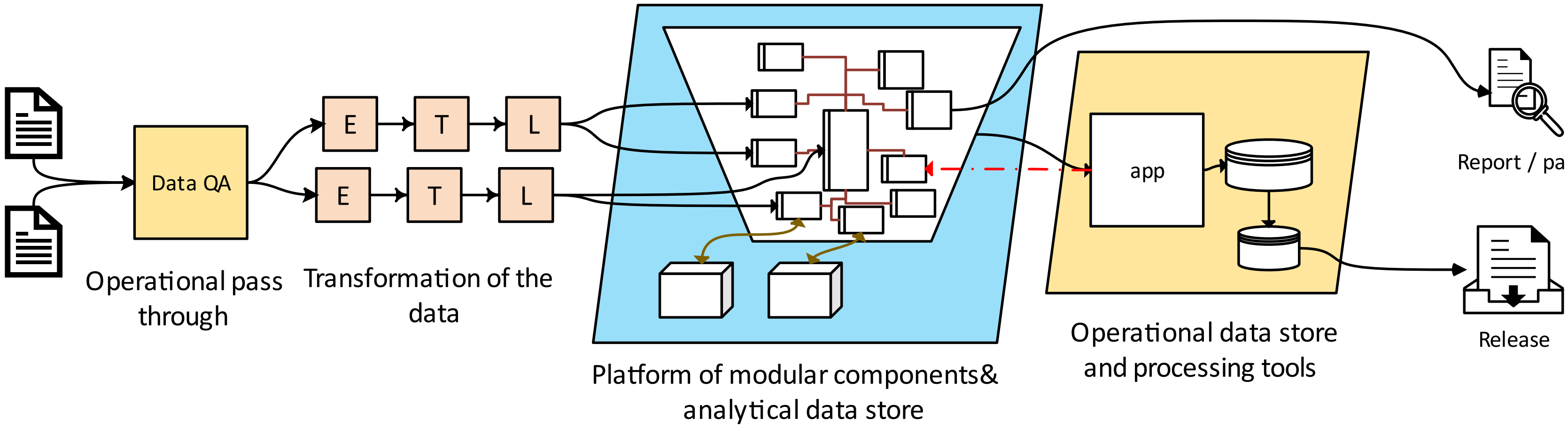
Connecting capabilities to data and application architecture

- ❖ Isolated analytical platform proved cumbersome and inefficient
- ✓ Planning foundational steps: build a platform with robust data layer, and enable modular development of citizen development. Platform should:
  - ✓ Enable reproducible production pipelines
  - ✓ Support iterative R&D through cloud SaaS tools (dataset registry, version control, orchestration, metadata standards, etc)
  - ✓ Enable development that follows blueprints or best practices (composed within the NSO or program) to enable robust development and maintain maturity and discipline

Initial development



Current plans: Flexible platform to enable modular development



Data QA
E
T
L
E
T
L
app
Report / paper
Release
Operational pass through
Transformation of the data
Platform of modular components& analytical data store
Operational data store and processing tools

Key part of the Agency

Statistics Canada is continuing to rapidly develop the agency’s IT maturity, with major investments aligned with a Target Enterprise Architecture and building on Data Mesh principles.

The CPI program, as a pivotal domain with the multi-domain agency ecosystem, both supporting other statistical programs and benefiting as others mature.

- Get support to develop modular platform and data architecture to achieve program needs
- Contribute back to make StatCan data discoverable
- Partner to build enhanced security processes: greater collaboration while maintaining privacy
- Partner to develop virtualization to minimize data copies
- Trial and bring in various analytical tools to support staff to do their work effectively
- Build monitoring and risk management solutions
- Maintain computing capacity without sacrificing cost

Lessons learned and continued development

Benefits from a paradigm of "a platform of modular components"

- Data architecture the foundation of a robust platform for multiple use
- Modularity of components enables reuse and interchangeability
- Facilitates coordination within the program and other programs (in alignment with agency architecture)

Standardize components with a holistic picture in mind

- Standardization and balance between low rate of change capabilities with IT, citizen development for high rate of change capabilities
- Compliance with a robust architecture direction builds maturity and accelerates enhancements to the program. Lower levels of coordination builds technical debt and decelerates innovation
- Centralization of common resources, such as for data validation or annotation, accelerates R&D
- Classification to low level, concordance to higher level taxonomies allows us to standardize better and enhance analytic capacity
- Management of the whole data lifecycle facilitated with a standardized approach

Enabling citizen development for R&D and production provenance

- Adoption of open standards and solutions builds robust and reproducible production processes and build provenance in the whole platform
- Code version control, experiment tracking, orchestration pipelines, deployment processes

Analytics approached holistically provides data driven insights to explain movements

Change management and production processes necessary alongside infrastructure development

Discoverability of datasets and methods

- Minimize copies and modification
- Adoption of proper metadata and dataset registry

Recent enhancements

- Statistics Canada has developed a rich set of tools and resources to help Canadians learn more about inflation and the CPI
  - CPI Portal, data visualization tools, personal inflation calculator
- Statistics Canada continues to enhance the CPI program through the ongoing introduction of alternative data sources which better reflect the prices Canadians are paying for the goods and services they consume and in systems, processes and tools to continually enhance the datasets used to calculate the CPI
  - Roughly 50% of the prices used to calculate the CPI come from alternative data sources, including retailer transaction data from the point of sale or scanner data, web scraping, application programming interface, and administrative data. The remaining 50% of prices are primarily collected online.
- Other recent CPI enhancements:
  - Annual CPI Basket Updates
  - Adjusted Price index to account for shifting consumer spending patterns
  - Expansion of the monthly average retailer prices (18-10-0245-01)
  - Measuring price change for used vehicles in the Canadian Consumer Price Index
  - Measuring the price of digital computing equipment and devices in the Consumer Price Index
  - And many more!

Thank You!

Questions, feedback, ideas?

serge.goussev@statcan.gc.ca