Big Data for Official Statistics: Strategic Considerations and Recommendations on Data Infrastructure and Governance

Authors:

Eric Deeben
Richard Tonkin

Statistics Working Papers

SD/WP/18/July 2025

Please cite this paper as:

Deeben, Eric, and Tonkin, Richard (2025). *Big Data for Official Statistics: Strategic Considerations and Recommendations on Data Infrastructure and Governance.* Statistics Working Paper Series (SD/WP/18/July 2025). Statistics Division, ESCAP, United Nations: Bangkok.

About the authors:

This working paper was drafted by Eric Deeben, Consultant and Richard Tonkin, Statistician, Economic and Environment Section, Statistics Division, ESCAP, with inputs and guidance from Rikke Munk Hansen (Statistics Division, ESCAP), Neil Jackson (independent consultant) and Samrat Maskey (Statistics Division, DESA).

For further information on this paper, please contact stat.unescap@un.org.

Available at: https://www.unescap.org/knowledge-products-series/statistics-working-papers

Tracking number: SD/WP/18/July 2025

Photo credit: shutterstock_623711519.jpg

Contents

Figures and Boxes

Abstract

This working paper provides guidance for Senior Leaders of National Statistical Offices (NSOs) seeking to modernise their data infrastructure in order to harness big data and other non-traditional data sources for official statistics. Drawing on country experiences from Asia and the Pacific, the paper outlines key infrastructure components, highlights best practices, and presents recommendations tailored to different levels of institutional readiness.

The development of this paper was supported by ESCAP's project on Big Data for Official Statistics, funded by the 2030 Agenda Sub-Fund of the UN Peace and Development Trust Fund.

Executive summary

Across Asia and the Pacific, National Statistical Offices (NSOs) face increasing demands from policymakers and other stakeholders for more timely and granular statistics. At the same time, the vast and expanding array of data sources, including satellite imagery, mobile phone data, and administrative records, offers significant opportunities for enhancing official statistics to meet such increased demands. To effectively utilise existing and emerging data sources, NSOs need modern, secure, and scalable data infrastructure supported by strong governance and skilled personnel.

This working paper, drawing on experiences from ESCAP's project on Big Data for Official Statistics and complementary global and regional initiatives, outlines practical considerations for modernising data infrastructure. Aimed primarily at senior NSO leaders, including Chief Statisticians, Deputy Heads, and Directors of IT or Data Systems, it provides clarity on core concepts, shares real-world examples from the region, and presents actionable recommendations.

Case studies from across the Asia-Pacific region (Annex 1) illustrate diverse pathways towards data infrastructure modernisation, including:

- **Indonesia:** Implementing end-to-end automation of satellite imagery processing for agricultural statistics.
- **Uzbekistan:** Developing a structured and scalable system for Consumer Price Index (CPI) production, integrating scanner and web-scraped data.
- **Maldives:** Creating secure, automated workflows for administrative data sharing through open-source storage solutions.
- **Viet Nam:** Establishing a modular data ecosystem to streamline the production and dissemination of national household survey data.

These examples demonstrate that infrastructure modernisation is achievable within varying resource constraints. They underscore the importance of strategic vision, modular and interoperable designs, automation, data governance, and targeted capacity development.

Key strategic recommendations for NSO leadership include:

- **Recognising infrastructure as a strategic asset:** Infrastructure investments underpin NSOs' ability to remain responsive and relevant amid rapid technological and data changes.
- **Prioritising scalability, security, and interoperability:** Adopting modular, cloud-based solutions and common standards ensures infrastructure can adapt and scale effectively.
- **Strengthening data governance and institutional capacity:** Clear governance frameworks, supported by ongoing skills development, ensure data are managed responsibly and strategically.
- **Taking context-driven, incremental actions:** NSOs should adopt modernisation strategies suited to their institutional readiness, resources, and national priorities.

All depending on the starting point, initial steps may include baseline assessments and pilot projects using open-source tools. NSOs with more developed capacities could prioritise scalable infrastructure and robust governance arrangements. Those with an even higher degree of institutional readiness could focus on expanding automation, adopting advanced analytics, and leading regional knowledge-sharing efforts.

Modernising data infrastructure is not a one-off effort but a continuous process. By making strategic, incremental investments aligned with national development priorities, NSOs can build the capabilities needed to meet future data demands and support evidence-based policymaking.

Introduction

Across the Asia-Pacific region, National Statistical Offices (NSOs) find themselves at the intersection of increasing demands and unprecedented opportunities. Policymakers and other stakeholders require timely, granular, and integrated statistical information to address urgent issues such as climate resilience, economic transformation, population change and progress toward the Sustainable Development Goals (SDGs). At the same time, new non-traditional data sources—such as earth observation, mobile phone data, and web-scraped and commercial datasets—promise richer insights and improved responsiveness. However, harnessing the potential of these sources presents new challenges for NSOs in terms of data governance, infrastructure, methods and capacity.

Recognising these challenges, the ESCAP Committee on Statistics decided at its seventh session to feature big data for official statistics as a priority area of future work. This commitment has focused on sharing country research, experiences, and good practices, and on facilitating capacity development across the region. As part of this programme of work, ESCAP launched a project on Big Data for Official Statistics, funded by the 2030 Agenda Sub-Fund of the UN Peace and Development Trust Fund. The project has aimed to support NSOs in moving beyond experimentation with big data and data science, and toward embedding these sources and methods within regular statistical production.

ESCAP's support under this project has highlighted significant variability in the data infrastructure capabilities and governance readiness of NSOs across the region, revealing both the critical gaps and the strategic opportunities that exist. Expert meetings and workshops organised under the project have revealed a common set of challenges: around the ability to ingest and integrate diverse data sources, maintain data security and privacy, ensure system interoperability, and generate regular, high-quality outputs. Underlying these gaps is a broader need for strategic planning and investment, particularly in foundational systems and frameworks that support secure, efficient, and scalable data use.

The table below highlights some of the challenge areas that have been identified through this work:

Challenge Area	Key Issues
Data Access and Privacy	Accessing privately controlled data; need for legislative reforms; importance of privacy protections and compliance with existing data protection regulations.
Technological Readiness	Lack of high-performance computing, scalable storage, and secure cloud in low-resource NSOs
Data Quality and Integration	Difficulties integrating unstructured or non-official data; high demand on tools for validation, interoperability, and metadata
Capacity Building	Shortage of skilled professionals; inadequate and under-funded training programs
Sustainability	Dependence on external partners or proprietary tools; risk of vendor lock-in; need for sustainable, open-source-friendly infrastructure
Legal and Ethical Concerns	Challenges around data ownership, algorithmic bias, transparency, and Intellectual Property (IP) compliance; need for ethical safeguards in system design

Responding to these challenges requires a new approach to statistical operations—one that is dynamic, cross-sectoral, and technology-supported. It also requires greater attention to data governance: the policies, responsibilities, standards, and practices that ensure data is managed ethically, legally, and effectively across its lifecycle. ESCAP has developed a five-pillar framework to guide NSOs in strengthening governance for both traditional and new data sources. This framework, further discussed later in this paper, complements efforts to modernise infrastructure, enhance institutional capacity, and promote innovation within official statistics.

Together, these challenges underscore the need for NSOs to prioritise strategic investments in modern, adaptable, and secure data infrastructure, underpinned by robust governance and enabled through regional and international cooperation. Yet too often, data infrastructure discussions remain narrowly focused on technology and take place among technical staff, rather than by senior management as part of broader efforts to strengthen organisational performance and relevance.

This working paper is intended specifically for senior leaders within NSOs: Chief Statisticians, their deputies, and directors responsible for IT and data management systems. Its purpose is not to prescribe a rigid technical model or provide exhaustive detail on technical implementation. Instead, it seeks to provide clarity on key concepts and offer practical guidance that can inform strategic planning and decision-making.

Structure of this paper

The paper begins by defining and distinguishing among core concepts such as data infrastructure, data architecture, and IT infrastructure, highlighting why these distinctions matter in practice, as well as an exploration of ESCAP's five-pillar framework for describing data governance.

The central section of the paper introduces a practical framework for understanding the components of a modern NSO data infrastructure, drawing on regional good practices and established models. This is followed by a discussion of strategic considerations for infrastructure modernisation

The paper concludes with actionable recommendations designed specifically for senior NSO leadership. These recommendations outline steps that can be taken at strategic and operational levels to begin or accelerate modernisation efforts, regardless of an NSO's current infrastructure.

Country examples drawn from the Asia-Pacific region are shared in Annex 1 to illustrate how NSOs are navigating these issues in diverse contexts, with varying resources and priorities. These case studies demonstrate both practical solutions and persistent obstacles, offering insight into the real-world pathways for improving infrastructure and governance.

Ultimately, infrastructure modernisation is less about technology for its own sake and more about positioning NSOs to effectively meet evolving national and global needs for statistics. This paper, produced with reference to regional and global frameworks and initiatives, aims to support NSOs in this journey by helping them build infrastructures capable of reliably supporting new data sources, new methods, and, ultimately, better policy outcomes.

Key concepts and definitions

To enable meaningful strategic discussions about modernising data infrastructure within National Statistical Offices (NSOs), it is essential to clarify foundational concepts and terminology. Terms such as data infrastructure, data architecture, and IT infrastructure are frequently used interchangeably, leading to confusion. Clear definitions and a coherent understanding of their interrelationships are crucial for senior NSO leaders responsible for strategic planning and investment decisions.

Data Infrastructure

In the context of an NSO, *data infrastructure* refers to the integrated system of technologies, people, processes, policies, and governance frameworks that support the entire data lifecycle, from collection and storage to analysis and dissemination. It encompasses physical components (such as servers, storage systems, network equipment, and cloud platforms), digital tools (including databases, data pipelines, and analytics software), and governance mechanisms designed to ensure data quality, privacy, and compliance. A summary of the typical components of a modern NSO's data infrastructure is provided in Box 1

Box 1. Components of Modern NSO Data Infrastructure

A well-functioning data infrastructure in a National Statistical Office typically includes:

- **Technical capacity:** Digital environments capable of managing large volumes and diverse types of data from traditional and emerging sources.
- **Automated systems and workflows:** Processes that streamline data ingestion, transformation, validation, and analysis, reducing reliance on manual tasks.
- **Secure and scalable data storage:** Systems such as cloud-based data lakes or data warehouses that provide robust, secure, and efficient access to data.
- **Data governance frameworks:** Mechanisms ensuring data privacy, compliance with national regulations, and consistency and quality throughout the data lifecycle.
- **Interoperability and standardisation:** Clearly defined standards and APIs enabling seamless data exchange internally and externally.
- **Analytical and dissemination tools:** Software and platforms that support statistical analysis, visualisation, and public dissemination of results.
- **Human capacity and skills:** Staff—including data engineers, statisticians, IT specialists, and analysts—trained and equipped to develop, operate, and maintain infrastructure systems.
- **Strategic policies and frameworks:** Institutional policies that guide infrastructure development, aligned with national priorities and international standards.

The data infrastructure underpins all statistical production processes, supporting traditional survey and census activities, as well as the integration of emerging data sources like administrative records, mobile phone data, satellite imagery, and real-time sensor feeds. While the scale and sophistication of data infrastructure varies widely among NSOs, all offices have some form of data infrastructure—whether advanced and automated, or fragmented and largely manual. Modernisation efforts typically aim to strategically evolve and strengthen existing infrastructure rather than replace it wholesale.

Distinguishing Data Infrastructure, Data Architecture, and IT Infrastructure

Although closely related, the concepts of *data infrastructure, data architecture*, and *IT infrastructure* have distinct meanings and practical implications. Clarifying these distinctions can help NSO leaders better understand where to direct strategic efforts and resources.

- **Data Infrastructure:** As described above, data infrastructure refers to the overall environment—including hardware, software, processes, policies, and capacities—that enables data collection, management, analysis, and dissemination.
- **Data Architecture:** Data architecture refers more specifically to the high-level design or blueprint that guides how data moves within the organisation. It specifies data standards, structures, metadata management, and rules for interoperability across systems. It shapes how data should be stored, integrated, and accessed, ensuring that different parts of the infrastructure work coherently together. Good data architecture facilitates efficient data use and reuse, ensures consistency, and supports compliance with standards and governance requirements. Box 2 sets out the technical components of a modern data architecture.
- **IT Infrastructure:** The physical and virtual hardware and software systems required to support data operations. While vital, IT infrastructure alone does not ensure robust data management without a clearly defined data architecture and effective governance mechanisms. Conversely, IT infrastructure may also support digital operations beyond data management itself.

Understanding these distinctions helps NSOs to strategically prioritise investment decisions. For example, upgrading servers or migrating to cloud storage (IT infrastructure) may significantly enhance processing capacity, but without a coherent data architecture and robust data governance mechanisms, NSOs might still struggle with interoperability, data quality, or effective integration of big data sources.

Box 2. Technical Components of Modern Data Architecture

The figure below shows several key components commonly found in modern data architectures. While the relevance of each will depend on the NSO's context, size, and goals, understanding the role of these components can support better decision-making about infrastructure investments and coordination with technical teams.

Component	What it does	Why it matters for NSOs
Data Pipelines	Automate the flow of data from source to storage, transformation, and analysis.	Reduces manual effort, improves reliability, and enables faster data processing.
Cloud Storage	Provides flexible and scalable storage in a hosted environment.	Supports cost-effective handling of large datasets, especially from non-traditional sources.
APIs	Enable secure, structured exchange of data between systems.	Facilitates integration with partners (e.g. ministries) and improves internal interoperability.
AI/ML Models	Support predictive analysis, classification, and pattern detection.	Enhances the NSO’s ability to use complex data sources such as satellite or mobile data.
Data Streaming	Processes data in real time as it is generated.	Useful for time-sensitive applications like price monitoring, mobility tracking, or disaster response.
Kubernetes	Manages deployment and scaling of applications in containers.	Helps ensure reliable operation of analytics and data services, especially in cloud environments.
Cloud Computing	Provides on-demand access to processing power and services.	Supports scalable, distributed analysis without requiring heavy local infrastructure.
Real-time Analytics	Delivers insights as data becomes available, often integrated with dashboards or alerting systems.	Enables decision-makers to respond rapidly to changing conditions, increasing the relevance of NSO outputs.

These components support different stages of the data lifecycle and align with the broader goals of building flexible, interoperable, and sustainable NSO infrastructure. Not all components are required in every context, but collectively they illustrate the kinds of technologies shaping modern statistical systems.

Data governance

Data governance is a critical cross-cutting element underpinning effective data infrastructure, including when handling non-traditional or big data sources. Data governance refers to the structured set of policies, processes, responsibilities, standards, and practices that ensure data is managed ethically, legally, securely, and efficiently throughout its lifecycle.

ESCAP has developed a data governance framework structured around five pillars, which provides a basis for describing national data governance arrangements in different countries. These pillars comprise:

1. **Vision and policy intent**: Sets the high-level ambitions and strategic direction for data, aligning with national priorities and public expectations.

2. **Data management, policies, rules and institutions**: Establishes the formal rules, standards, and institutional structures - legislation, codes of conduct, or guidelines - that govern how data is handled.

3. **Role of data custodians and data stewards:** Clarifies who is responsible for day-to-day data management (custodians) and who manages data strategy, user engagement, and accountability (stewards).

4. **Data sharing, accessibility and integration:** Focuses on the mechanisms - technical, legal and operational - that allow data to be combined and accessed responsibly.

5. **Data sharing risks and mitigation**: Addresses data use risks (e.g., privacy, misuse) through frameworks such as the 'Five Safes', ensuring secure and trustworthy data sharing.

Security and privacy protections are essential elements of data governance, especially when working with sensitive or personal data. NSOs should implement technical and procedural safeguards such as data anonymisation, encryption, firewall protection, and role-based access controls. These safeguards must be complemented by internal policies and compliance with applicable national, regional and international regulations.

Auditability is also critical. Systems should include functionality for tracking data lineage, allowing users to understand how data has been transformed over time, as well as maintaining logs of access and modification activities. These practices build institutional trust and support external accountability, particularly when working with non-traditional or third-party data sources.

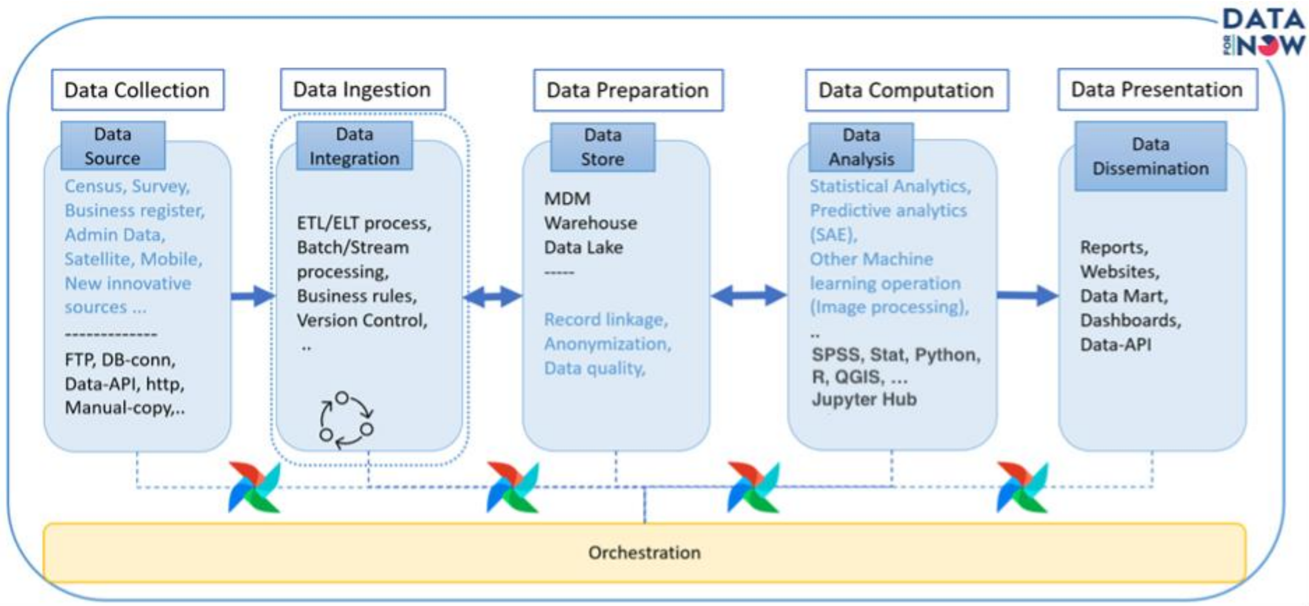
As NSOs increasingly exchange data with other government agencies and international partners, the use of common data standards becomes vital. Frameworks such as SDMX (Statistical Data and Metadata eXchange), RDF (Resource Description Framework), and DDI (Data Documentation Initiative) facilitate interoperability, support efficient data exchange, and reduce duplication. Adopting these standards improves long-term system compatibility and data reuse.

Data Infrastructure components for an NSO

This section provides an overview of the main components of a modern NSO data infrastructure, using the data lifecycle structure developed by the Data for Now initiative (Figure 1). This lifecycle reflects how data typically flows through an NSO's systems, from collection to final dissemination, and highlights the infrastructure and process elements needed at each stage.

The data lifecycle structure provides a common language and framework that can help senior NSO leaders identify where current strengths lie, where gaps may exist, and how different components connect.

Figure 1. Data Lifecycle within a Modern NSO Infrastructure



Note: All key terms/acronyms included in Figure 1 are explained in the glossary in Annex 2.

Data Collection: The data lifecycle begins with the collection of data from a variety of sources. This can include traditional sources such as censuses and surveys, but increasingly also administrative data (e.g., civil registration systems, tax records), geospatial and satellite data, mobile phone records, and other forms of digital exhaust. These sources may be accessed in a variety of ways, including through Application Programming Interfaces (APIs), direct database (DB) connections, manual transfer, or file transfer protocols (FTP).

Infrastructure needs at this stage include secure access mechanisms, reliable systems for importing data in various formats and frequencies, and governance arrangements that ensure appropriate legal agreements and privacy safeguards are in place when using sensitive or third-party data.

Data Ingestion: Once data is obtained, it must be brought into the NSO's systems. This is the ingestion phase, where processes like Extract-Transform-Load (ETL), Extract-Load-Transform (ELT), batch and streaming ingestion, and version control are applied. This stage also typically includes initial business rules or validation checks to ensure data is formatted, structured, and ready for internal use.

For traditional sources such as surveys, censuses, and administrative data, ingestion may involve uploading data files received from field teams or partner agencies, validating their structure, and transforming formats (e.g. from Excel or CSV to database-friendly formats) to ensure compatibility with internal systems.

Well-functioning data ingestion infrastructure allows NSOs to work with large and complex data sources without extensive manual handling, minimising errors and supporting reproducibility.

Data Preparation: In this stage, ingested data is prepared for analysis. It is cleaned, integrated, harmonized, and stored in structured environments such as data warehouses, lakes, or master data management (MDM) systems. Processes like anonymisation, deduplication, and metadata tagging are also carried out here.

Data preparation is a critical step, especially when combining multiple sources or working with sensitive data. It requires secure and scalable infrastructure, as well as processes that support traceability, reproducibility, and compliance with data governance policies.

Data Computation: This is the analytical core of the infrastructure, where statistical and computational methods are applied. It includes both traditional techniques (such as descriptive and inferential statistics) and newer approaches like machine learning, image processing and other forms of data science.

Tools commonly used include R, Python, QGIS, SPSS and Jupyter notebooks. Computing infrastructure may include cloud-based platforms or on-premise systems, each with its own advantages in terms of scalability, cost, and compliance. Frameworks such as Hadoop or Spark may be used to support distributed or parallel processing, especially when dealing with large volumes of data.

Decision-making at this stage should consider long-term sustainability, cost-efficiency, and alignment with national security or data protection requirements. Governance mechanisms—such as access control, audit logging, and model validation procedures—should be in place to ensure outputs are reliable, reproducible, and ethically produced.

Data Presentation: The final stage is the dissemination of statistical outputs. This may take the form of publications, dashboards, open data portals, APIs, and downloadable datasets. Infrastructure at this stage needs to support timely, secure, and user-friendly access to statistical information for a variety of audiences.

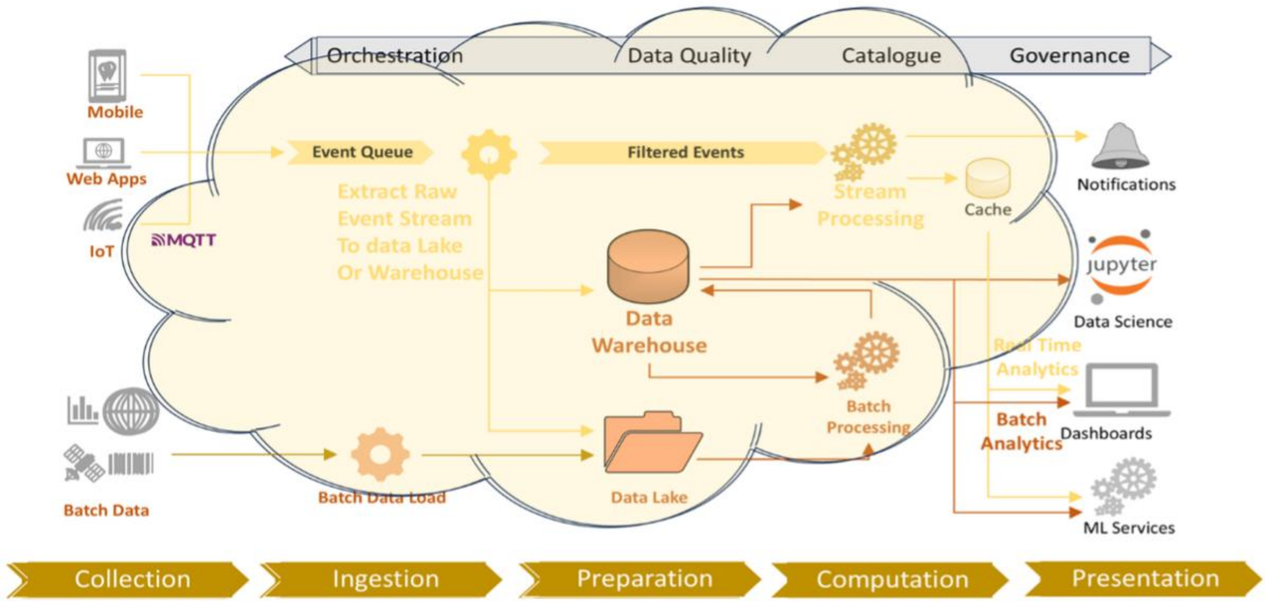
Dissemination infrastructure may also incorporate mechanisms for feedback, allowing users to raise issues or suggest improvements, supporting the continuous development of statistical products.

Orchestration: A key cross-cutting element is orchestration: the systems and processes that coordinate workflows across the data lifecycle. This includes task scheduling, resource management, monitoring, automation, and error handling. Effective orchestration ensures that infrastructure components operate together coherently, enabling end-to-end processes to be executed consistently and scaled as needed.

Figure 2 provides an example of how a Big Data infrastructure blueprint might appear.

The picture illustrates the five steps of the data lifecycle structure, from collection to presentation, and the various open-source technologies an NSO could adopt at each stage.

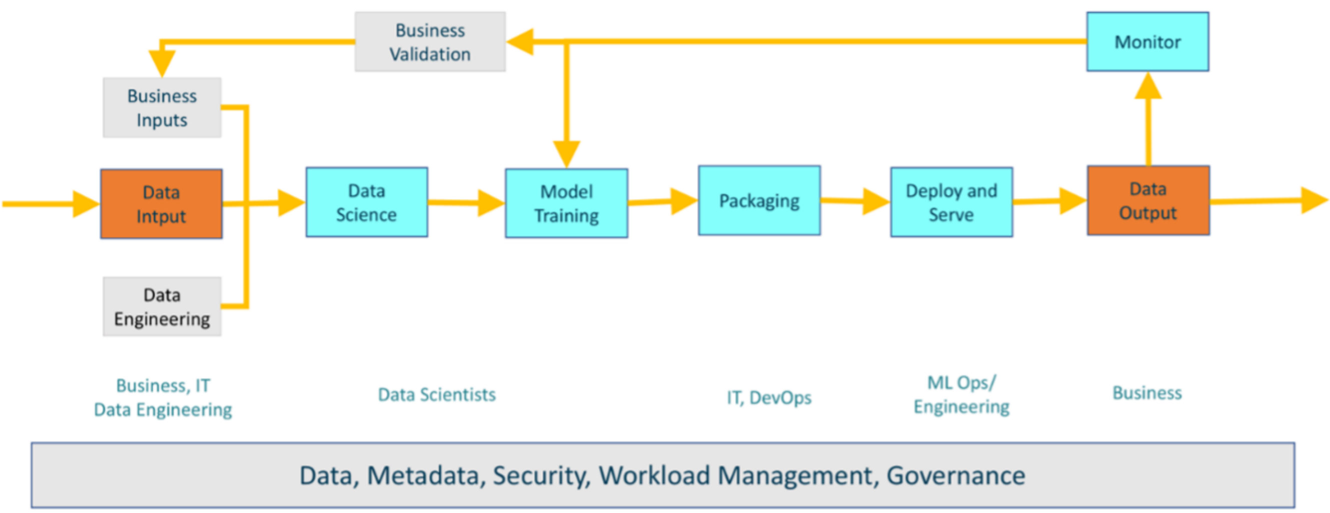
Figure 2: Example of Big Data Pipeline Infrastructure



An example: Earth observation data pipeline

Figure 3 provides an example of an end-to-end machine learning pipeline designed specifically for Earth Observation data, based on a well-designed Data Infrastructure.

Figure 3. Illustration: An End-to-End Pipeline for Earth Observation Data



While the terminology differs slightly, the stages in the pipeline broadly map onto the lifecycle described earlier:

- "Data Input" and "Data Engineering" correspond to **Data Collection** and **Ingestion**.
- "Model Training" and "Data Science" are forms of **Data Preparation** or **Computation**.
- "Packaging," "Deploy and Serve," and "Monitor" align with **Presentation** and post-dissemination monitoring.

This example illustrates how production machine learning (ML) is a collaborative and iterative process that involves multiple teams and disciplines working together. The pipeline begins with the ingestion of raw data, which is shaped by both business inputs and data engineering tasks to ensure it is usable for analysis. From there, data scientists take over, applying their expertise to explore and prepare the data for model training.

Once a machine learning model is trained, it is packaged for deployment, a process often managed by IT and DevOps teams. The model is then deployed into a production environment by ML operations and engineering teams, where it begins generating outputs based on real-world data. These outputs are delivered to the business, which validates their accuracy and relevance. Throughout this process, monitoring plays a crucial role in ensuring that the model continues to perform well, allowing for ongoing improvements and adjustments.

At the core of this pipeline are foundational elements such as data management, metadata tracking, security, workload management, and governance. These elements underpin the entire process, ensuring that data is handled responsibly, operations are efficient, and outputs are reliable and compliant with standards. Overall, the picture emphasises that machine learning for Earth Observation is not a one-off task, but a continuous, team-driven cycle aimed at delivering actionable insights through coordinated efforts across business, technical, and analytical functions.

Strategic value of the lifecycle view

Understanding the full data lifecycle and its infrastructure requirements is essential for strategic planning. Rather than viewing infrastructure as a collection of disconnected tools or systems, the lifecycle perspective helps NSO leadership consider:

- Which parts of the lifecycle are currently well-supported;
- Where bottlenecks or gaps may exist;
- What kinds of skills and capabilities are needed at each stage;
- And how infrastructure planning aligns with national priorities and future data needs.

The following sections will turn to best practices that can help NSOs strengthen each of these components, drawing on lessons from across the region.

Data Infrastructure: Strategic considerations

Modernising data infrastructure is a complex but necessary process for NSOs seeking to produce more timely, granular, and relevant statistics, especially as they begin working with new data sources such as mobile phone records, satellite imagery, and administrative systems. While approaches will differ depending on institutional mandates, resources and capacity, common themes have emerged across the region.

This section presents selected good practices under five thematic areas. These are not intended as a rigid framework but as a reference to support strategic planning and decision-making. Annex 1 provides concrete examples of how countries from the Asia-Pacific region have adopted some of these practices.

Strategic Planning and Governance

- **Align infrastructure with national data strategies:** Investments in infrastructure should directly support the NSO's strategic objectives and national priorities. Clearly define goals related to data use, disaggregated reporting, and responsiveness to emerging policy issues, ensuring infrastructure plans directly address these needs.
- **Reference governance arrangements clearly:** Infrastructure planning should be informed by the NSO's existing governance landscape, including legal mandates, stewardship roles, and data-sharing rules. ESCAP's five-pillar framework may be used to describe or assess current arrangements and identify areas for potential strengthening or coordination.
- **Coordinate cross-departmental planning:** Form integrated teams involving statisticians, IT specialists, and senior managers to ensure infrastructure developments are informed by real statistical needs, promoting alignment and shared understanding across the organisation.

Modular and Scalable Systems

- **Adopt a modular approach:** Design infrastructure in separable modules—such as ingestion, storage, computation, and dissemination—to enable flexible upgrades or replacements. This modularity supports easier maintenance, reduces risks, and allows incremental improvements tailored to available resources. For NSOs seeking implementation-oriented resources, UNSD's *Technical Document to Modernize IT Architecture in NSO/NSS using Open-Source Technology Stack* provides more detailed guidance on setting up modular data infrastructure using open-source tools and cloud-ready components.
- **Prioritise interoperability and open standards:** Use widely adopted standards such as SDMX, RDF, and DDI to improve data interoperability both internally and externally. Clearly documented Application Programming Interfaces (APIs) facilitate secure data exchange with government departments and international partners, reducing integration complexity.
- Evaluate cloud versus on-premise solutions: Carefully assess cloud computing options for flexibility, scalability, and cost-effectiveness, while considering on-premise solutions for scenarios with data sovereignty, legal, or security constraints. Tools like Hadoop or Spark may support efficient processing of large datasets, especially in distributed or hybrid computing environments.

Automation and Workflow Efficiency

- **Automate data processing pipelines:** Implement automated workflows for data ingestion, transformation, and quality assurance using orchestration tools such as Apache Airflow or NiFi. Automation minimises manual errors, accelerates processing time, and ensures consistent application of validation and quality checks.
- **Integrate continuous quality assurance:** Quality controls should be embedded at every stage of the data lifecycle. Automated validation routines, anomaly detection, and metadata standards help maintain data reliability, reproducibility, and trustworthiness.
- **Track data lineage and ensure reproducibility:** Maintain robust metadata systems and audit trails to clearly document how data is transformed and analysed. This transparency facilitates auditability, supports compliance with governance requirements, and enables reproducibility of statistical processes.

Capacity Development and Cultural Change

- **Invest in skills development:** Modern infrastructure requires skilled staff, including data engineers, systems administrators, data scientists, and analysts. NSOs must prioritise capacity development through structured training, mentoring programmes, and partnerships with academic and training institutions.
- **Encourage experimentation and innovation:** Create a supportive environment that allows experimentation with new technologies and methods. Leaders should actively encourage pilot projects, celebrate successful innovation, and view failures as learning opportunities rather than setbacks.
- **Enhance senior leadership understanding:** Senior NSO leaders require clear, concise briefings about data infrastructure and governance challenges to make informed strategic decisions. Regular dialogues with technical teams help leaders understand capabilities, constraints, and opportunities for improvement.

Leveraging Partnerships and Shared Tools

- **Utilise existing open-source solutions:** Avoid unnecessary duplication by reusing proven tools, methods, and platforms developed by other NSOs, regional bodies, or global statistical communities. Shared solutions significantly reduce costs and shorten development timelines.
- **Participate in regional and global networks and communities:** ESCAP, the United Nations Statistics Division (UNSD) and other partners offer platforms for peer exchange, joint projects, and shared services. Participating in these communities helps NSOs stay current with good practices and avoid duplication.
- **Collaborate strategically with external partners:** Partnerships with universities, research organisations, local technology companies, and international bodies can fill critical skill gaps, co-develop solutions, and strengthen institutional capacity. Clearly define roles, responsibilities, and expectations within such collaborations to maximise mutual benefit.

Infrastructure modernisation is an ongoing, iterative process requiring continuous leadership engagement, cross-sector collaboration, and careful strategic planning. By adopting these best practices tailored to specific contexts and capacities, NSOs can significantly enhance their capability to produce timely, reliable, and responsive statistics, effectively meeting evolving national and global data demands.

Recommendations

Modernising data infrastructure is no longer a peripheral task for national statistical systems. It is central to the ability of National Statistical Offices (NSOs) to produce timely, disaggregated, and policy-relevant statistics in a digital and data-rich era. Strategic decisions about infrastructure must be informed by a broad understanding of technological options, institutional needs, and the current stage of infrastructure readiness.

While there is no single pathway to follow, this report highlights a number of cross-cutting lessons and actions that can guide national efforts. These recommendations reflect experience from across Asia and the Pacific and aim to support senior NSO leaders in setting priorities, managing risks, and building capabilities over time.

Key Messages

1. **Data infrastructure is a strategic asset:** Rather than a back-office concern, data infrastructure underpins the delivery of official statistics and enables innovation in the use of new and alternative data sources. It plays a central role in enabling automation, integration, and the delivery of more timely and granular data products.
2. **Modern systems must be scalable, secure and interoperable:** Investments should favour modular and standards-based designs that allow components to evolve over time. Cloud platforms, containerisation, APIs, and orchestration tools offer powerful ways to build flexibility and resilience into national statistical systems.
3. **Governance and human capacity are essential enablers:** Technical solutions cannot succeed in isolation. Clear data governance arrangements—including roles, responsibilities, access controls, and safeguards—are critical. So too is building a workforce that can develop, operate and improve these systems over time.
4. **Progress is incremental and context-dependent:** Each NSO has different starting points and constraints. Successful modernisation depends on identifying practical entry points, aligning efforts with broader national priorities, and learning from regional and global peers.

Suggested Priorities by Capacity Level

To support practical decision-making, the following table outlines recommended next steps tailored to different levels of infrastructural readiness. These are indicative only and should be adapted to each national context.

Current Capacity Level	Recommended Next Steps
Emerging (low capacity)	• Start with small-scale pilot projects using open-source tools. • Undertake a baseline assessment of infrastructure readiness and skills. • Map key governance and data architecture gaps. • Participate in peer-learning initiatives and regional communities of practice.
Developing (mid-level capacity)	• Adopt modular, scalable infrastructure components. • Strengthen institutional data governance, including formal roles and responsibilities. • Invest in training on data engineering, cloud services, and security. • Explore pay-as-you-go cloud models and hybrid architectures.
Established (high capacity)	• Expand the use of automation, orchestration, and real-time analytics. • Pilot advanced tools such as federated learning or privacy-preserving methods. • Contribute to regional or global initiatives on shared infrastructure. • Incorporate sustainability goals and green computing principles into infrastructure planning.

Annex 1 offers illustrative country experiences that reflect the recommendations outlined here, adapted to varied institutional and technical contexts. NSOs can also make use of more detailed technical guidance such as UNSD's Technical Document to Modernize IT architecture in NSO/NSS using open-source technology stack.

Modernising data infrastructure is an ongoing process that requires sustained commitment. The aim is not to build the most advanced system, but one that is appropriate, reliable, and responsive to national priorities. With the right strategies and partnerships in place, NSOs can strengthen their foundational capabilities and play a more central role in informing evidence-based policymaking and sustainable development.

Annex 1: Country experiences

Efforts to modernise data infrastructure across the Asia-Pacific region reflect the diversity of starting points, resources, and national priorities. While some National Statistical Offices (NSOs) are already deploying advanced systems and integrating new data sources, others are taking initial steps to build core capacity, strengthen governance, or consolidate legacy systems. Country experiences shared through ESCAP's project on Big Data for Official Statistics and related initiatives offer valuable insights into the challenges and opportunities faced across this spectrum.

The examples presented here are not intended as direct comparisons, nor as models to be replicated. Instead, they illustrate different pathways towards infrastructure modernisation and highlight how context-specific factors—such as institutional mandates, technical capacity, partnerships, and external shocks—shape national approaches.

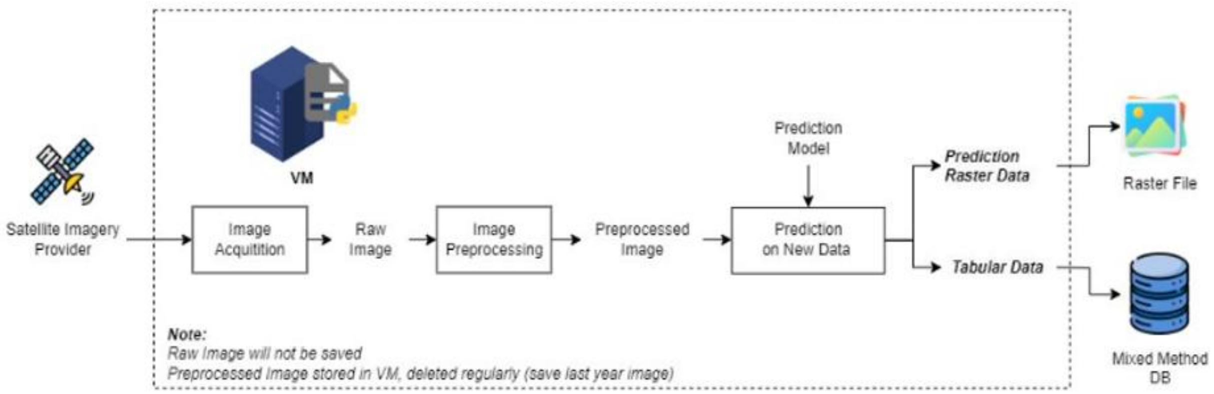
This section draws on experiences from four countries: two that participated in ESCAP's Big Data project—**Indonesia and Uzbekistan**—as well as two countries engaged in complementary work supported by UNSD under Data for Now initiative—**Viet Nam and the Maldives**. Together, they reflect the varied conditions and ambitions that characterise the region's modernisation landscape.

Indonesia

Indonesia's experience illustrates how modern infrastructure and internal technical capacity can enable the integration of Earth observation data into official statistics. BPS - Statistics Indonesia, with support from ESCAP's Big Data for Official Statistics project, has been working through their One Rice initiative to integrate satellite imagery with ground truth data to produce estimates of rice production that are more timely, comprehensive and cost-effective than the current survey-based estimates. This video provides more insights into the data and methods used.

Supporting this mixed-methods approach is a data architecture comprising a data processing pipeline that prepares and analyses satellite imagery using predictive models, and a dissemination pipeline that delivers the resulting data in both tabular and map formats.

Figure 4. Processing and estimation pipeline for rice estimation in Indonesia



The upstream pipeline begins with image acquisition and preprocessing, conducted within a controlled virtual machine (VM) environment:

- **Image acquisition:** Satellite imagery is sourced from providers outside of BPS and transferred into the VM.
- **Preprocessing:** Raw images are cleaned and enhanced, including processes such as cloud masking and atmospheric correction. These pre-processed images are temporarily stored in the VM.
- **Prediction:** The cleaned images are fed into trained machine learning models (e.g., supervised classification algorithms) to generate spatial predictions of harvest area and crop conditions.

The model outputs include:

- **Raster data**, stored as geospatial layers for visualisation;
- **Tabular data**, summarising productivity and production, structured for statistical analysis.

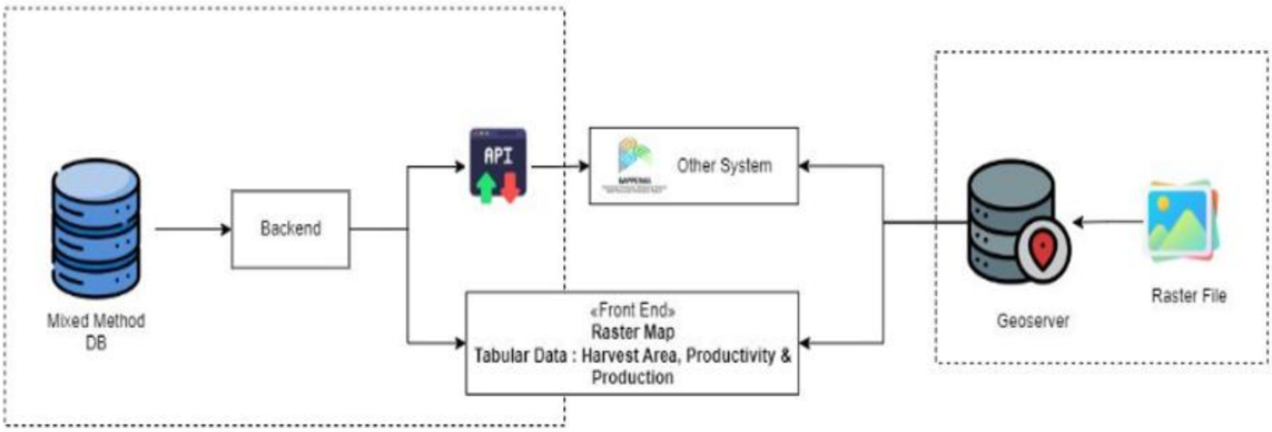
Given the volume and complexity of spatial data, significant computational resources are required to ensure efficient processing. A storage capacity of 30 TB is required due to Indonesia's vast territory and the size of raw satellite imagery files (e.g., Sentinel-1). This estimate is based on the number of Military Grid Reference System (MGRS) tiles covering Indonesia from January 2021 to December 2023.

To support this processing, a RAM capacity of 512 GB and a processor with 64 cores are used. Spatial operations on satellite imagery are memory-intensive, and these specifications allow for concurrent processing of up to three provinces, each of which can take one to two days to complete.

Raw satellite images are not stored permanently. Pre-processed images are deleted regularly, helping to manage limited storage capacity and reduce security risks.

This processing pipeline demonstrates key features of scalable, modern infrastructure: automation, efficient storage strategies, and the operational use of predictive models.

Figure 5. Dissemination and visualisation data flow



Outputs from the prediction workflow are then channelled into BPS's dissemination system:

- Processed tabular data is stored in the **Mixed Method Database**, which feeds the backend of the dissemination system.
- A dedicated **API layer** enables structured and secure data exchange, including links with internal and external platforms.
- **GeoServer** is used to serve the raster files, allowing them to be visualised as

geospatial layers in interactive front-end applications.

- The final front-end displays, accessed by policymakers, researchers, and the public, combine **raster maps** with **summary statistics** on harvest area, productivity, and production.

Indonesia's implementation highlights several best practices discussed throughout this paper:

- **Modular infrastructure:** Each component—from data acquisition to dissemination—is functionally distinct but interoperable, enabling flexibility and scalability.
- **Automation and efficiency:** Machine learning is used to generate outputs without requiring manual interpretation of images, and automated scripts manage processing workflows.
- **Storage management and security:** Temporary file retention and isolated VM environments reduce infrastructure demands and mitigate risk.
- **Interoperability:** APIs and the use of open standards support integration with other national systems and future scalability.

This system represents a concrete example of how an NSO can move from experimentation with big data toward fully embedded operational use. The success of the initiative rests on long-term investment in internal capacity, alignment with strategic priorities, and the use of adaptable infrastructure that can support expansion over time, for example, to other crops, regions, or statistical domains

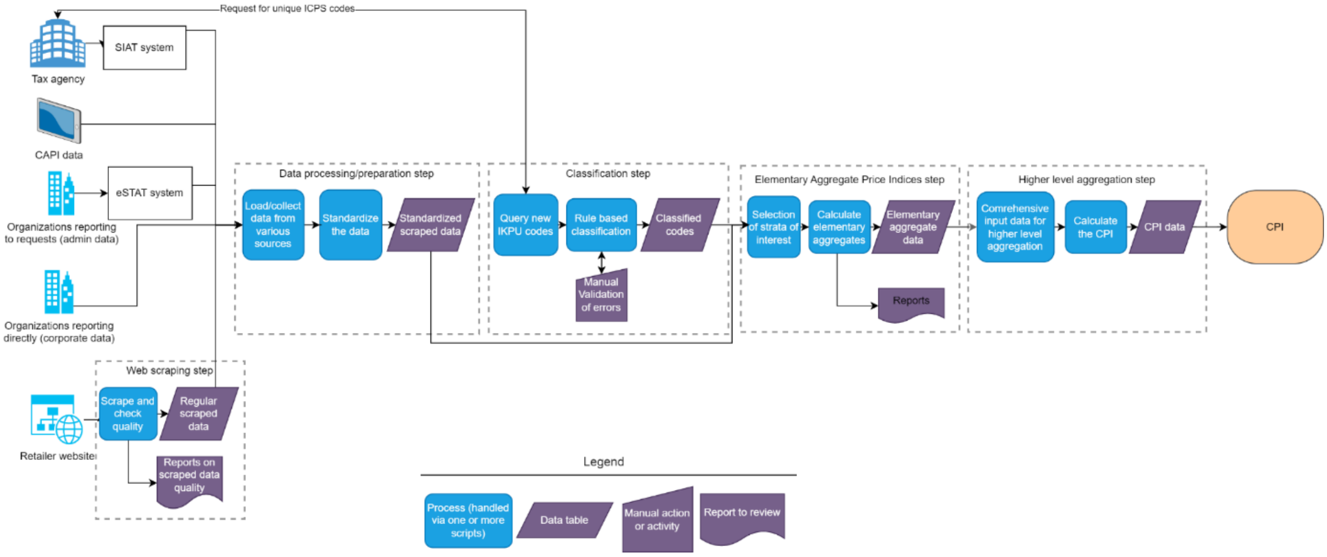
Uzbekistan

With support from ESCAP, the Statistics Agency under the President of the Republic of Uzbekistan (UZStat) is undertaking a major modernisation effort to transform its Consumer Price Index (CPI) production. This initiative is centred on the integration of new data sources—particularly scanner data and web-scraped price data—into regular statistical workflows. This Stats Brief provides a summary of the experiences, achievements and lessons learned by UzStat through this work, along with a set of recommendations for those looking to make similar use of alternative data sources for their price statistics.

To support this transformation, UZStat has made significant technical investments. The Agency began by shifting away from desktop tools such as Microsoft Access, moving instead to more scalable statistical processing in Stata. This transition is being complemented by the gradual adoption of open-source tools and best practices, including the use of GitLab for version control and the incorporation of Python-based processing, aligned with evolving team capacity.

At the core of this data infrastructure modernisation is a target data architecture developed by UZStat with support from ESCAP. The architecture reflects both international best practices and Uzbekistan's institutional and operational context. It is built around principles of modularity, automation, and traceability, using a staged processing workflow to ensure long-term scalability and resilience.

Figure 6. Uzbekistan's Target Architecture for CPI Modernisation.



Processing begins with the collection of data from 5 sources: scanner (or EPOS) data from the Tax Committee (shared with UZStat through the SIAT system), CAPI data, administrative data, corporate datasets, and web scraped data. All the data from all systems are loaded into a raw data layer (the main ingestion point) for downstream use in the form it is received. Failure at this level can be manually investigated by NSO staff.

In the data processing and preparation step, dedicated pipelines clean and standardise the different data sources. The use of separate processing pipelines for each data type allows for isolation of errors and easier adaptation to changes in data format or structure.

Several distinct elementary aggregation pipelines (as there may be many depending on the method needed) then ingest this data. Analysis can be done on both the underlying data and the aggregated data to understand why the elementary aggregate moved the way it did. A final pipeline, the main CPI aggregation system, then feeds the CPI release.

This approach separates changes to the data into distinct processing steps that communicate through the different layers of the data architecture. This allows changes to be made to one pipeline without impacting another.

UZStat is considering the approach as it creates a scalable and evolvable architecture, suitable due to the many changes in data sources and elementary aggregation methods that will need to occur in the future. It also allows the creation of smaller pipelines that can be developed with Reproducible Analytical Pipeline (RAP) principles in mind and supported between the IT and price statistics teams.

Many of the best practices discussed in this paper are illustrated in Uzbekistan's approach, including:

- **Modular Design:** A clearly structured architecture supports phased implementation and long-term flexibility.
- **Automated Workflows:** Even partial automation (e.g. validation and classification) yields significant gains in efficiency and data quality
- **Scalability and Open Source:** Transitioning from legacy tools like MS Access to scalable platforms is a recognised next step, in line with best

practices highlighted throughout this paper.

- **Strategic Partnerships:** Institutional agreements with the Tax Committee enabled scanner data integration. Further collaboration with retailers will be essential for sustainable web scraping.
- **Incremental Modernisation:** UZStat's progress shows that substantial gains are possible even before full systems replacement. The reform approach is staged and pragmatic

Maldives

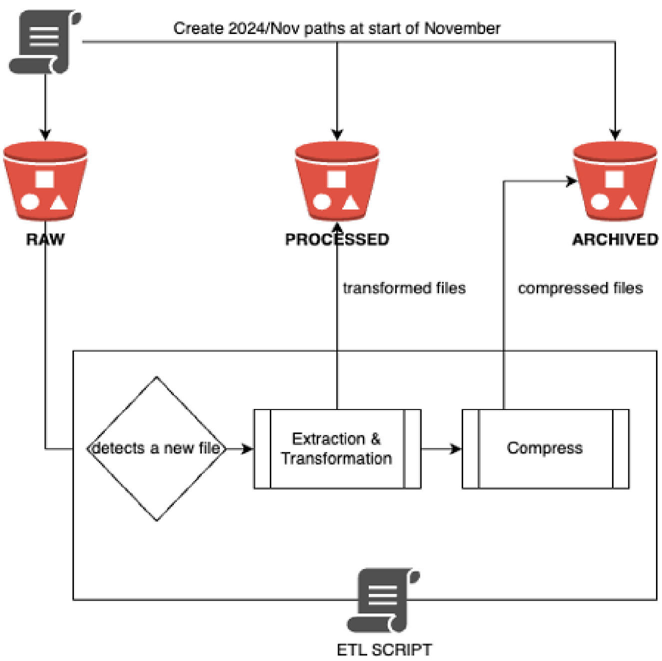
The Maldives Bureau of Statistics (MBS), with technical support from the United Nations Statistics Division (UNSD) under the 13th Tranche of the UN Development Account (DA-13), has implemented a secure and structured data management system designed to enhance the integration and use of administrative data. The work has focused on establishing foundational infrastructure for handling sensitive data securely and efficiently, while reducing reliance on manual processes.

The MBS has adopted a three-tier data lake architecture using the open-source MinIO object storage system. This structure supports the full data pipeline from raw ingestion through to archiving:

- **Raw Bucket** – receives unmodified data from line ministries or other sources, ensuring traceability and data integrity;
- **Processed Bucket** – contains cleaned and transformed data, ready for analysis;
- **Archive Bucket** – stores compressed files for long-term retention and compliance.

These tiers provide a clear separation of responsibilities and allow the bureau to manage data across its lifecycle while supporting auditability and future reuse.

Figure 7. Three-Tier Data Storage Architecture – Maldives Bureau of Statistics



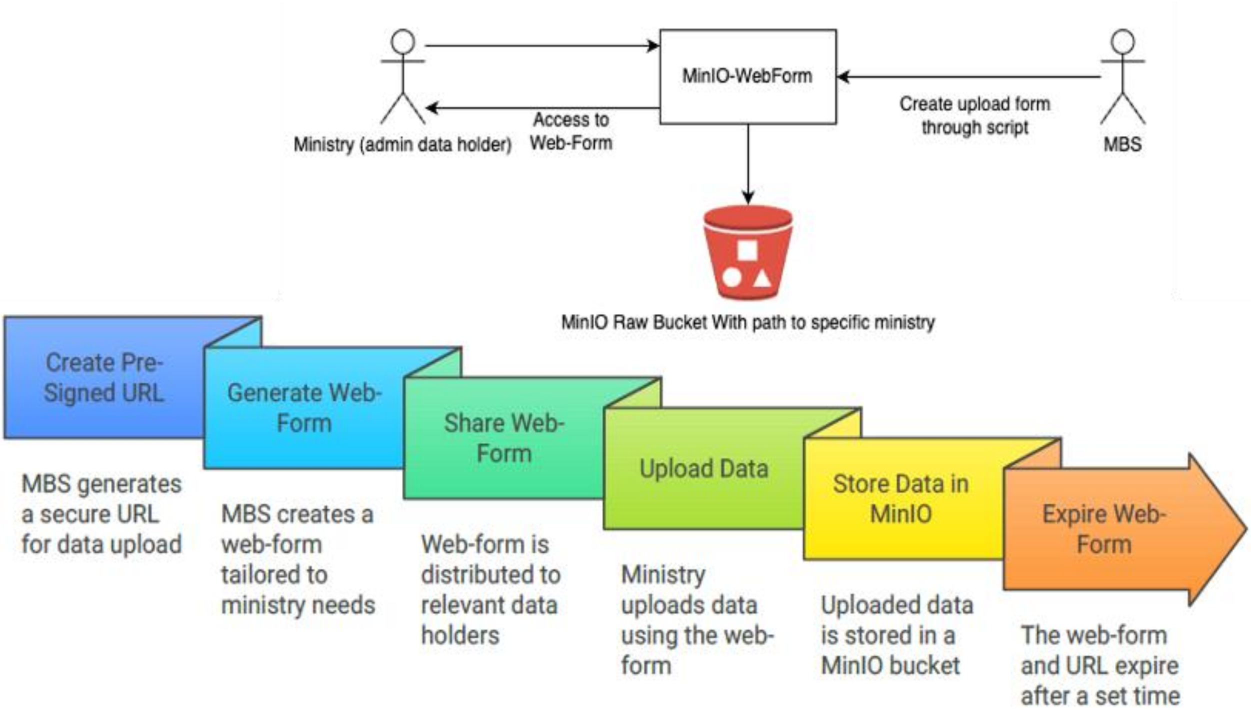
This process is supported by scheduled automation, using Python and Linux-based cron jobs. When new data is detected in the Raw Bucket, ETL scripts automatically extract and transform the data, move it to the Processed Bucket, and compress and archive the original files. This system minimises manual handling, ensures consistency, and supports reproducibility in data processing.

To improve administrative data acquisition from other government entities, MBS has developed two complementary methods for secure data exchange, both designed to work with the MinIO platform.

Web Form Upload Using Pre-Signed URLs:

MBS can generate temporary, secure upload links that are embedded into a web form, allowing ministries to easily submit files without needing direct access to the storage system.

Figure 8. Secure Data Upload via MinIO Web Form

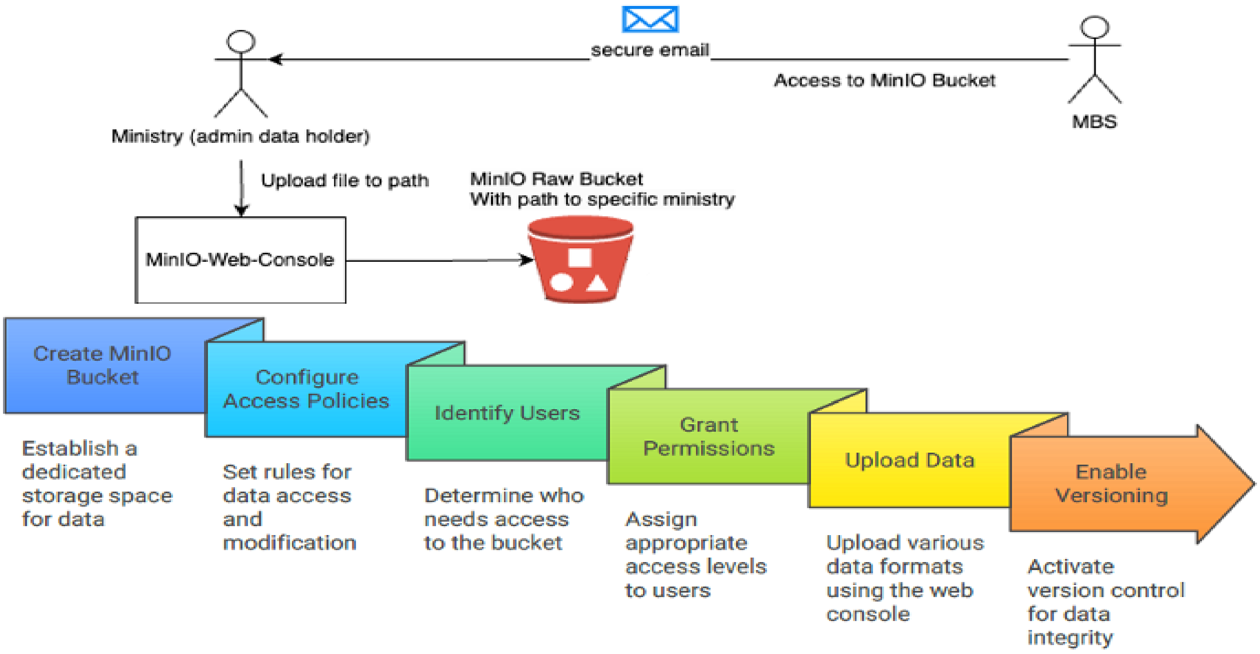


These web forms can be integrated into the MBS website or deployed as standalone tools. Once submitted, the data is placed directly into the appropriate Raw Bucket, ensuring confidentiality and preventing tampering.

Direct Access via the MinIO Web Console:

For ministries with more advanced needs or recurring submissions, MBS offers secure access to a dedicated MinIO bucket through its built-in web console. This allows administrators to upload files, configure access controls, and manage datasets directly.

Figure 9. Secure Upload via MinIO Web Console



Using this method, MBS can assign user-specific permissions, enforce version control, and track access through logs, supporting stronger accountability and auditability.

The Maldives' experience illustrates how even relatively lightweight and open-source technologies can support core elements of modern statistical infrastructure. Key strengths of the system include:

- **Security and governance:** All uploads are protected using secure URLs or user-level authentication, and version control ensures data integrity.
- **Automation and efficiency:** Scripts eliminate the need for manual handling, reducing the risk of errors and freeing up staff time.
- **Scalability:** The system can be expanded or replicated across ministries, supporting future ambitions for a more integrated national statistical system.

This approach aligns closely with the principles of secure, interoperable, and user-centred data infrastructure modernisation and highlights how targeted technical support can produce tangible operational improvements, even within constrained environments.

Viet Nam

The National Statistics Office (NSO) of Viet Nam is undergoing a transformation of its data infrastructure to support the modernisation of national survey operations. As part of the Data for Now initiative, co-led by the United Nations Statistics Division (UNSD), the World Bank, and other global partners, the NSO is focused on integrating open-source technologies into a modular data ecosystem that enhances automation, analytical capability, and data security.

This work is aligned with the NSO's strategic vision, which includes the establishment of a centralised statistical data system by 2030, and reflects broader efforts to improve the monitoring of Sustainable Development Goals

(SDGs), particularly where integration of administrative and geospatial data is needed.

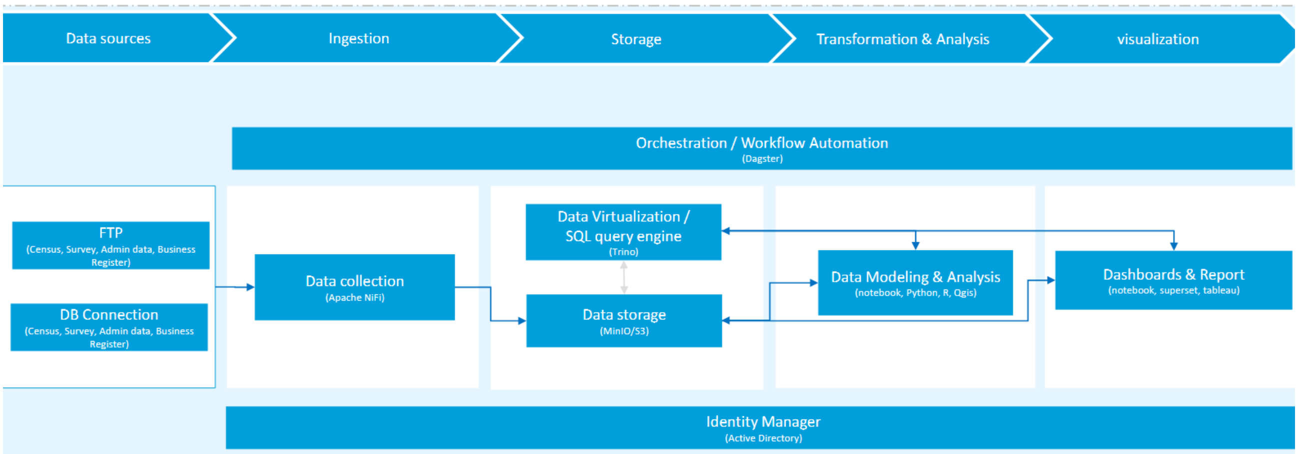
A suite of tools has been implemented to support the end-to-end data lifecycle, including:

- **Apache NiFi:** Enables secure, automated data ingestion and transformation. This tool supports real-time routing of diverse data formats (such as CSV, JSON, and XML) and helps ensure that incoming data is enriched, validated, and integrated with minimal manual intervention.
- **JupyterHub:** Offers GSO analysts a collaborative, browser-based platform for working with Python, R, and Stata. It supports interactive analysis, code sharing, and reproducible workflows, all of which are critical for complex statistical processing.
- **MinIO:** Provides high-performance object storage, structured into distinct "buckets" to support data lifecycle management:
    - *Raw Bucket:* Retains original data in its natural format.
    - *Anonymised Bucket:* Stores data processed to meet privacy requirements.
    - *Staging Bucket:* Hosts cleaned and transformed datasets prepared for analysis.
    - *Aggregate (Gold) Bucket:* Contains final datasets ready for dissemination and reporting.

  MinIO's support for policy-based access control ensures secure and auditable access to sensitive data.
- **Kubernetes:** Facilitates orchestration and management of containerised applications, supporting scalability, system reliability, and streamlined deployment of analytical services.

These tools are linked to form an integrated ecosystem that supports both flexibility and security, while allowing the NSO to scale its infrastructure as demands evolve.

Figure 10. National Statistics Office of Viet Nam's IT Infrastructure.



One of the major applications of this infrastructure is the transformation of Viet Nam's Labour Force Survey (LFS) and Viet Nam Household Living Standards Survey (VHLSS). These flagship national surveys have traditionally involved extensive manual data handling, resulting in long processing times and limited capacity for rapid analysis.

Through automation enabled by Apache NiFi and the flexible analysis environments of JupyterHub, data from these surveys can now be ingested, cleaned, and analysed far more efficiently. Storage is centralised using MinIO, enabling secure access by authorised staff. Kubernetes supports continuous integration and delivery, allowing NSO to manage infrastructure resources dynamically and ensure high system uptime.

Impact and Strategic Outlook

This modernised ecosystem has already demonstrated benefits in terms of speed, security, and analytical depth. Processing times for household surveys have been significantly reduced, and the system enables faster turnaround of key indicators for policy use. Enhanced governance features such as policy-based access control also help ensure compliance with national and international data protection standards.

The experience of NSO Viet Nam highlights how targeted investment in data infrastructure can strengthen the capacity of NSOs to integrate data sources, support high-frequency outputs, and ensure secure, governed data use. The architecture in place is scalable and standards-based, offering a model that may be adapted by other countries at different stages of digital development.

Looking ahead, the NSO's strategic vision includes building a national statistical data lake that will integrate administrative, survey, and geospatial data sources. These efforts aim to support a growing set of nationalised SDG indicators, including those requiring small-area estimation and geospatial visualisation.

Annex 2: Glossary of terms and acronyms

Term	Definition
Alternative Data Sources (ADS)	Non-traditional data sources such as mobile phone records, satellite imagery, scanner data, and web-scraped content used to complement or replace survey-based data collection.
Apache Airflow	An open-source platform used to programmatically author, schedule, and monitor workflows. Often used for orchestrating complex data pipelines.
Apache NiFi	An open-source tool for automating the flow of data between systems, supporting real-time and batch processing with built-in security and data provenance features.
Application Programming Interface (API)	A set of protocols and tools that allow different software applications to communicate with each other.
Automated Data Ingestion	The process of automatically collecting and importing data from various sources into a data system.
Cloud Computing	The delivery of computing services over the internet, allowing access to storage, databases, and applications without on-site infrastructure.
Computer-Assisted Personal Interviewing (CAPI)	A method of data collection where interviewers use electronic devices to input responses during interviews.
Consumer Price Index (CPI)	An index measuring the average change in prices over time that consumers pay for a basket of goods and services.
Containerisation	A lightweight form of virtualization that packages software and its dependencies into a "container" to ensure it runs consistently across different environments.
Cron Job	A time-based scheduler in Unix-like operating systems used to automate server-side tasks at specified intervals.
Data Architecture	The design framework that dictates how data is collected, stored, integrated, and used in an organisation.
Data Governance	The set of policies, procedures, and standards used to manage data integrity, security, and quality within an organisation.
Data Documentation Initiative (DDI)	An international metadata standard for describing data from the social, behavioural, and economic sciences.
Data Infrastructure	The composite hardware, software, and services that enable data collection, processing, storage, and dissemination.

Term	Definition
Data Lake	A centralised repository that allows storage of structured and unstructured data at any scale.
Data Lifecycle	The sequence of stages data goes through, from initial collection to eventual archival or deletion.
Data Warehouse	A centralized repository designed to store large volumes of structured data from multiple sources, optimized for query and analysis. Unlike data lakes, data warehouses typically store curated, cleaned, and structured data for reporting and business intelligence.
Database Connection (DB-Conn)	A method for accessing and interacting with a database, typically through structured queries and protocols to read, write, or update data.
Earth Observation (EO)	The collection of data about the Earth's physical, chemical, and biological systems via remote sensing technologies, including satellites.
Extract, Transform, Load (ETL)	A data integration process involving the extraction of data from sources, transformation into usable formats, and loading into a database.
File Transfer Protocol	A standard network protocol used to transfer files from one host to another over a TCP-based network, such as the internet.
Five Safes	A framework used to assess and manage risks in data access: Safe projects, people, settings, data, and outputs.
GeoServer	An open-source server for sharing geospatial data.
General Data Protection Regulation (GDPR)	A European Union regulation on data protection and privacy.
Jupyter Hub	A multi-user platform for interactive computing environments like Jupyter notebooks.
Kubernetes	A platform for managing containerised workloads and services, enabling automatic deployment, scaling, and operations.
Machine Learning (ML)	A subset of artificial intelligence where algorithms learn patterns from data and improve over time without being explicitly programmed.
Manual Copy	The manual transfer of data files between systems or locations, typically using USB drives, file explorers, or direct system access. Often used when automated pipelines are not in place.
Master Data Management (MDM)	A set of tools, processes, and policies used to ensure the uniformity, accuracy, stewardship, and accountability of shared master data assets—such as reference data for people, locations, or products—across an organisation.
MinIO	An open-source object storage server compatible with Amazon S3 APIs.

Term	Definition
Orchestration	Coordinating automated tasks and workflows across different components of a data system to ensure reliable and timely execution.
Policy-Based Access Control (PBAC)	A security model that uses policies to determine user access to resources.
RAP (Reproducible Analytical Pipeline)	A set of principles and practices that support the automation and reproducibility of data workflows, particularly in the context of official statistics.
Raster Data	Pixel-based data used to represent images or spatial information.
Resource Description Framework (RDF)	A framework developed by the World Wide Web Consortium (W3C) for representing and linking structured data on the web.
Scanner Data	Electronic transaction records from electronic point-of-sale (EPOS) systems, commonly used in compiling price statistics like the CPI.
Sentinel-1	A satellite mission from the European Space Agency providing radar imagery for Earth observation.
Statistical Data and Metadata eXchange (SDMX)	An international standard designed to support the exchange and sharing of statistical data and metadata among organisations.
Virtual Machine (VM)	An emulation of a computer system that provides the functionality of a physical computer.

Statistics Division

ESCAP
Economic and Social Commission
for Asia and the Pacific

Follow us:

unescap

unitednationsescap

united-nations-escap

unescap.org