جهـــاز التخطيـــط والإحصـــاء
Planning and Statistics Authority
State of Qatar • دولـــة قطـــر

FINAL REPORT OF THE WEBINAR ON BUILDING A CENTRAL STATISTICAL DATA WAREHOUSE

FINAL REPORT OF THE WEBINAR ON BUILDING A CENTRAL STATISTICAL DATA WAREHOUSE

TABLE OF CONTENTS

FORWARD

I am delighted to share with dear readers the final report of the Webinar on "Building A Central Statistical Data Warehouse," which was held in Doha on 17th May 2022. The webinar shed light on the essential role of the timely, relevant, and robust official statistics for policy making in economic, social, human and environmental and the national development strategies.

The contemporary data revolution that resulted in advanced technology, digital transformation, and artificial intelligence call for the modernization of the official statistics systems for improved methods in using data to facilitate the needed transformation. This data revolution was not only about large amounts of data produced, but also about the increasing demand for it. This has led to increased pressure on national statistical organizations and their budgets and resources. This necessitates the adoption of efficient methods that improves users' access to what they need; including the standard data and indicators used to measure progress in achieving SDGs of the 2030 Agenda for Sustainable Development.

In this regard, a data warehouse will ensure consistency and a uniform format to standardize all collected data, facilitate data access for decision-makers and further help them analyze, visualize, and share insights. The data warehouse system decreases the costs and the risk of interpretation error by advancing the overall data accuracy, allowing decision-makers to access historical data where they can monitor and adjust their strategy.

A crucial aspect of data warehousing system is the maintenance of personal data confidentiality, protecting privacy, and building partnerships with data providers, whether they are individuals or business institutions. According to the statistical legislations issued by the UN Statistical Commission (UNSC) and the UN Economic and Social Council (ECOSOC), the main objective of official statistics systems is to provide statistics, data, statistical services and analyses to the public users, supported by metadata, visualization and spatial data

We are aware in the Planning and Statistics Authority of the ever-increasing demand for detailed and disaggregated data and statistics by national and international users and the new challenges for data in light of the digital technological advances, that generate large amounts of data which can be used and employed in making sound decisions, and in preparing studies.

Taking these factors into account, PSA has organized this forum in response to the rapid developments, to discus and exchange knowledge and methods of best practice with national and international partners, The proceedings of the forum will shed light on the importance of this work for all data users in the country It will emphasize the importance of national and international partnerships necessary to build the Data Warehouse, identify the statistical, technical, and the necessary structural processes required to build and operate an agile and responsive official national data and statistics system.

In conclusion, the workshop, was informative due to its valuable presentations. Moreover, it witnessed a remarkable virtual presence of about 120 participants from various ministries and government agencies: and an effective partnership with the Department of Statistics Malaysia (DOSM), The Italian National Statistical Institute (ISTAT), the Ministry of Interior (MOI), and Microsoft.

The workshop came out with an important report which set up the fundamental knowledge for building a central statistical data warehouse in the State of Qatar. The final report of the Webinar is available on PSA's website (visit here).

Dr. Saleh bin Mohammed Al Nabit

President of the Planning and Statistics Authority

OUR VISION

Leading national efforts to realize aspirations of Qatar National Vision 2030 for sustainable development and a decent life for its citizens, supported by an effective development planning system that prepares coherent development strategies, and by a pioneering national statistical system that produces high-quality statistics.

ACKNOWLEDGMENT

This is to acknowledge all those, without whom this Webinar would not have been a reality.

Firstly, a special thanks goes to HE Dr. Saleh bin Mohammad Al Nabit, President of Planning and Statistics Authority (PSA) and Mr. Mohammed Abdulaziz Mohammed Al-Nuaimi, Assistant to the President, for imparting thier vision, guidance, support and valuable contribution to this webinar.

Secondly, PSA would wish to thank the presenters Ms. Jamaliah Jaafar, Executive Director at Malaysian Bureau of Labour Statistics (MBLS) of Department of Statistics, Malaysia (DOSM); Mr. Herzie Mohamed Nordin, Director at Integration and Data Management Division, Department of Statistics (DOSM); Mr. Francesco Rizzo, Senior Executive Officer in The Italian National Statistical Institute (ISTAT); Ms. Latifa Jaber Al-marri, Database and Business Intelligence Expert at the Technical Affairs Department, Ministry of Interior,State of Qatar; and Ms. Mirna Malaeb, Data & AI Solutions Specialist – Middle East Cluster at Microsoft. The success of this webinar wouldn't have become possible without their technical inputs. Furthermore, we would like to thank all the PSA departments for their seamless support and help provided during the organization of Webinar, the attendees and participants from various Ministers, Institutions, Universities, Research Centers and Non-Governmental Organizations (NGOs).

A special acknowledgment is for the PSA staff for their professional dedication and expertise and contributions at the various stages of the webinar including the production of this report: Dr. Ahmad Hussein who organized and moderated the webinar, Ms. Sana Elashie for her outstanding support during the preparation, implementation, and reporting on the webinar, Dr. Mostafa Kharoufi and Dr. Madhavi Lalitha for their help in reviewing the webinar's final report, and Mr. Hilal Yazbek for his technical assistance during the webinar.

Disclaimer:

The views and opinions expressed in the regional workshop do not necessarily reflect the views of PSA.

Istat

Microsoft

وزارة الداخليـــــــــة
Ministry of Interior
State of Qatar • دولـــة قطـــر

INTRODUCTION

This webinar which was held on 17 May 2022 received the attention of data users and decision-makers from State of Qatar as well as from other nations, as it dealt with emerging concepts like advanced technology, digital transformation, and artificial intelligence. The present scenario of data revolution in the world showcased different aspects data. On one hand, there is an enormous data being produced across the globe and on the other, there is an increasing requirement for quality data for drawing better insights. Likewise, the national statistics and official statistics producers are looking for better means to manage the enormous data that is being piled up, day by day. At this juncture, developing a central statistical data warehouse that can address all the requirements of the stakeholders in developing a range indicator, particularly for SDGs along with the indicators related to Qatar National Vision 2030 is the need of the hour.

OBJECTIVES

In this context, the webinar aimed to:

A. Shed light on the importance of developing a Data Warehouse for all data users in the state of Qatar and the role of national and international partnerships required, to build the Data Warehouse.

B. Identify the statistical, technical, and structural processes required for its operation and update.

C . Learn about national and international experiences on the advanced practices and approaches in this field.

INAUGURATION

Mr. Mohammed A-Aziz M. Alnaimi, Assistant of the PSA President initiated the webinar with a brief inaugural address, in which he affirmed the need for such webinars that can not only provide room for widening the knowledge frontiers regarding the emerging topics related to statistics and data but also calls for the necessity to be on par with the technological ascents that are happening around the world. He also opined that knowledge imparting and awareness development programs, like the webinar, can help the nations to help each other in attaining the SDGs in an optimized manner, by fostering universal brotherhood (See Annex 1).

Dr. Ahmad Hussein, Official statistics expert moderated the webinar and provided a comprehensive background on the importance of building a central statistical data warehouse as a statistical dissemination tool and its relationship with the forthcoming 3rd National Development Strategy. He elaborated the statement of First UN Fundamental Principles of Official Statistics (UNFPOS), 'official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information' . These principles emphasize that official statistics is a public good, and access to statistical data should be provided to all users, and that the means of access dissemination of statistical data are to be considered for the convenience of users and their needs.

The webinar was attended by nearly 120 people from various Ministries, government institutions, universities, research centers, and non-government organizations. It was conducted fully virtual with simultaneous interpretation provided in Arabic and English languages.

PART I: PRESENTATIONS

Session One: Department of Statistics Malaysia **(DOSM)**

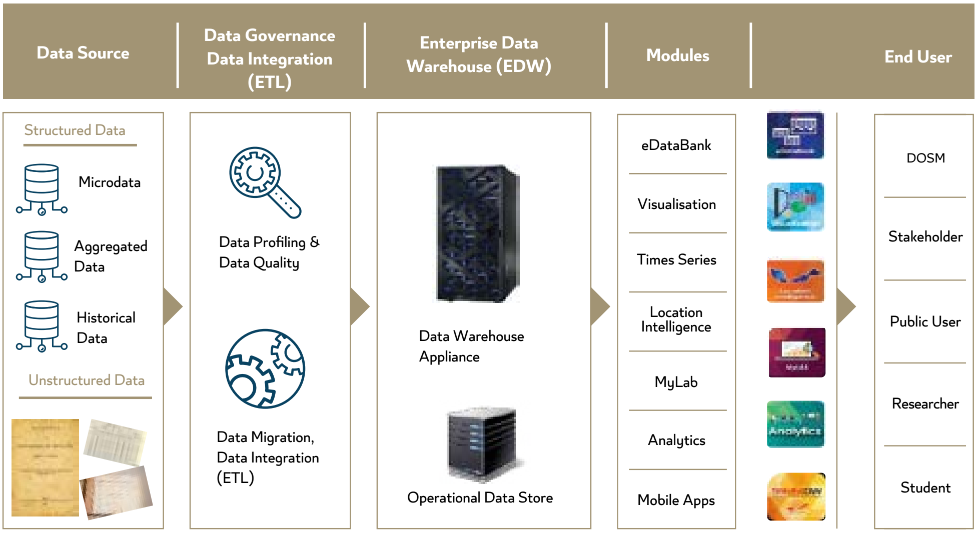
Modernizing Data Governing: StatsDW as national agenda (Presentation Link)

• Ms. Jamaliah Jaafar; Executive Director at Malaysian Bureau of Labour Statistics (MBLS)

• Mr. Herzie Mohamed Nordin; Director at Integration and Data Management Division

Abstract of the Presentation:

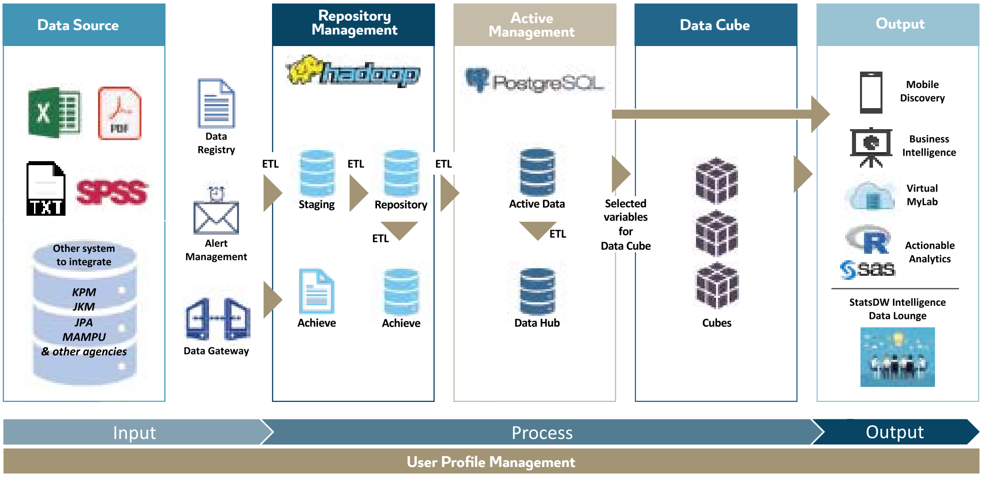
The presentation about StatsDW (Statistics Data Warehouse) has enthralled the gathering by covering a range of emerging topics from basic to advanced level that have helped to achieve the national agenda by modernizing the data governance. Discussing the strategy that StatsDW, Malaysia has adopted with well-defined objectives, motivation and benchmarking, the phase-wise journey of StatsDW has been articulated by the team in an effective manner. Also, the intricacies involved in modernizing data governance and management of statistics data warehouse with modules of StatsDW, all-inclusive architectures, Smart Data Partnership, Smart Data Lake Management, extract, transform and load (ETL) process, Business Intelligence, BI-eDATABANK, Data Visualization, and concerned analytics are discussed with enough priority.



Extract, transform & load (ETL)

Figure 1: Architecture of Phase-I provided by StatDW

Notwithstanding this, the presentation dealt with the benefits of deploying latest data-related technologies, methodology to overcome the challenges as well as many novel technologies which are vital for nurturing a data-driven society. An illustration of the Architecture provided by StatsDW is given in figure-1 & 2 for a bird's eye view.



Note :

KPM – Ministry of Education

JKM – Welfare Department

JPA – The Public Service Department

MAMPU – Malaysia Administrative Modernisation and Management Planning Unit

Figure 2: Architecture of Phase-II provided by StatDW

Session Two: Italian National Institute of Statistics (ISTAT)

The Statistical Dissemination: Data Warehouse in ISTAT (Presentation Link)

- Mr. Francesco Rizzo; Senior Executive Officer in The Italian National Statistical Institute

Abstract of the Presentation:

Statistical Dissemination: Data Warehouse in ISTAT from Italy being the second session in the webinar, has orchestrated much more details about the need for modernizing the statistical processes in general along with the rationale for its adoption at ISTAT, in particular. ISTAT has been running a modernization program based on the guidelines of the UNECE "High-level Group for the Modernization of Statistical Production and Services". With the equation, Modernization = Standardization + Industrialization, the steps involved in implementation of DWH (Data Ware House), architecture followed coupled with the role of Distributed Dissemination of DWH in the National Statistical System is brought forth in an effective manner.

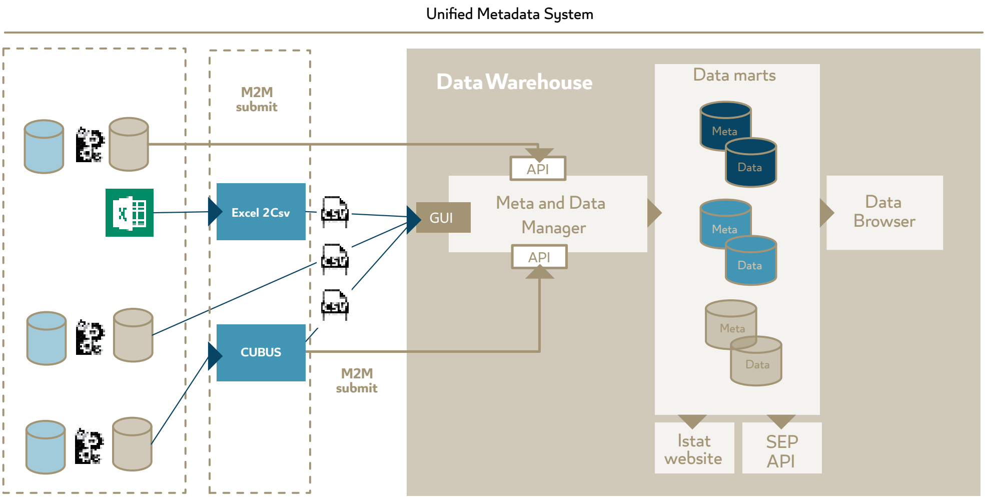


Figure 3: Dissemination Data Warehouse architecture in ISTAT

In addition to the information about the data flows involved and the synchronization needed between various components and technological tools (like Excel2Csv, Meta and Data Manager, Data Browser, Cubus, and others), the presentation also detailed the modus operandi of realizing the tasks with detailed pictures and content. The state-of-the-art framework of architecture, as well as the process flow of tasks that is carried out effectively with the support of technological tools, are provided in Figures 3 & 4 respectively.

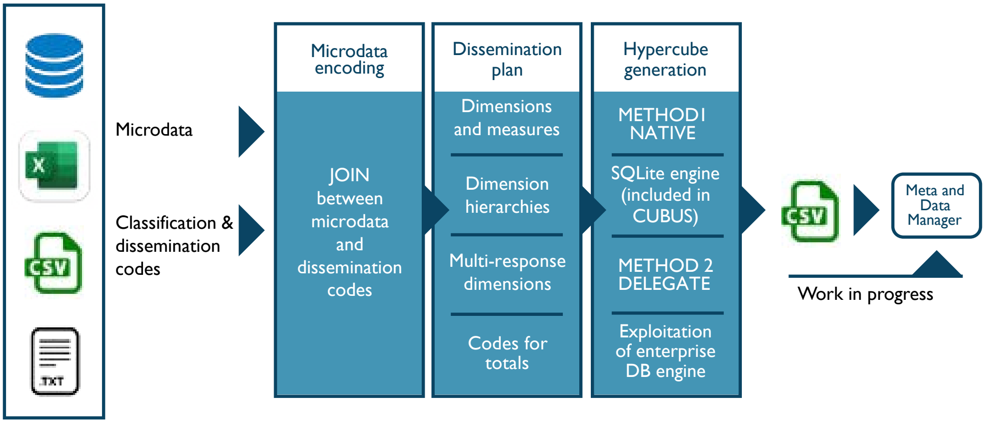


Figure 4: Data Processing with the support of Tools at ISTAT

Session Three: Ministry of Interior **(MOI)**

National Experience on Building Decision Making Support Systems Based on Data (Presentation Link)

- Ms. Latifa Jaber Al-marri; Database and business intelligence expert at the Technical Affairs Department at the Ministry of Interior

Abstract of the Presentation:

Under the title of Decision-Making Systems, Ms. Latifa Jaber Al-Marri gave a presentation in which she addressed various applications of decision-making system. The activities carried out by the Ministry of Interior, such as the EHTERAZ application that was developed to identify the situation of people in relation to the Covid-19 pandemic, and other applications related to the Center for Security Systems and decision support systems, including preparation for the country's to be hosted Football World Cup, electronic portals, facial recognition system and others.

Ms. Al-Marri illustrated the decision-making support mechanisms at the Ministry of Interior as follows (Figure 5)

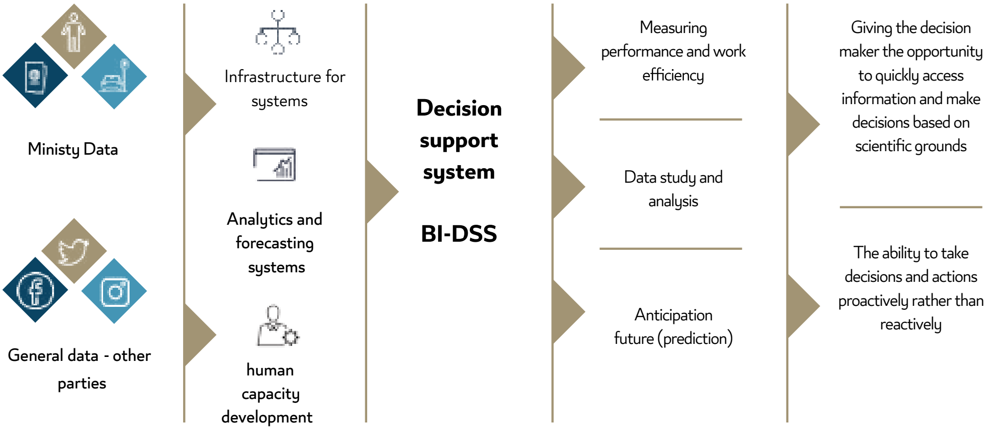

Ministy Data
General data - other parties
Infrastructure for systems
Analytics and forecasting systems
human capacity development
Decision support system
BI-DSS
Measuring performance and work efficiency
Data study and analysis
Anticipation future (prediction)
Giving the decision maker the opportunity to quickly access information and make decisions based on scientific grounds
The ability to take decisions and actions proactively rather than reactively

Figure 6: Decision Support Mechanism Ministry of Interior MOI

Ms. Al marri also presented the challenges that are being encountered by decision-making processes as well as solutions as follows:

1. Challenges:

- Presence of data in more than one system, which hinders the integrating process
- Limited data mining in current systems.
- Technical limitations in the current systems that hinder the extraction of the required data.
- The time required to generate reports and statistics.
- Inability to follow up on work procedures and data quality.

2. Solutions:

- Presence of a unified system that includes all data of the Ministry of Interior.
- Make data available in one data set.
- A special environment separate from the operational environment of the Ministry's systems has been provided.
- Providing the ability to extract and build reports in real-time for the user.
- Providing mechanisms and developing specialized training programs for the project.
-

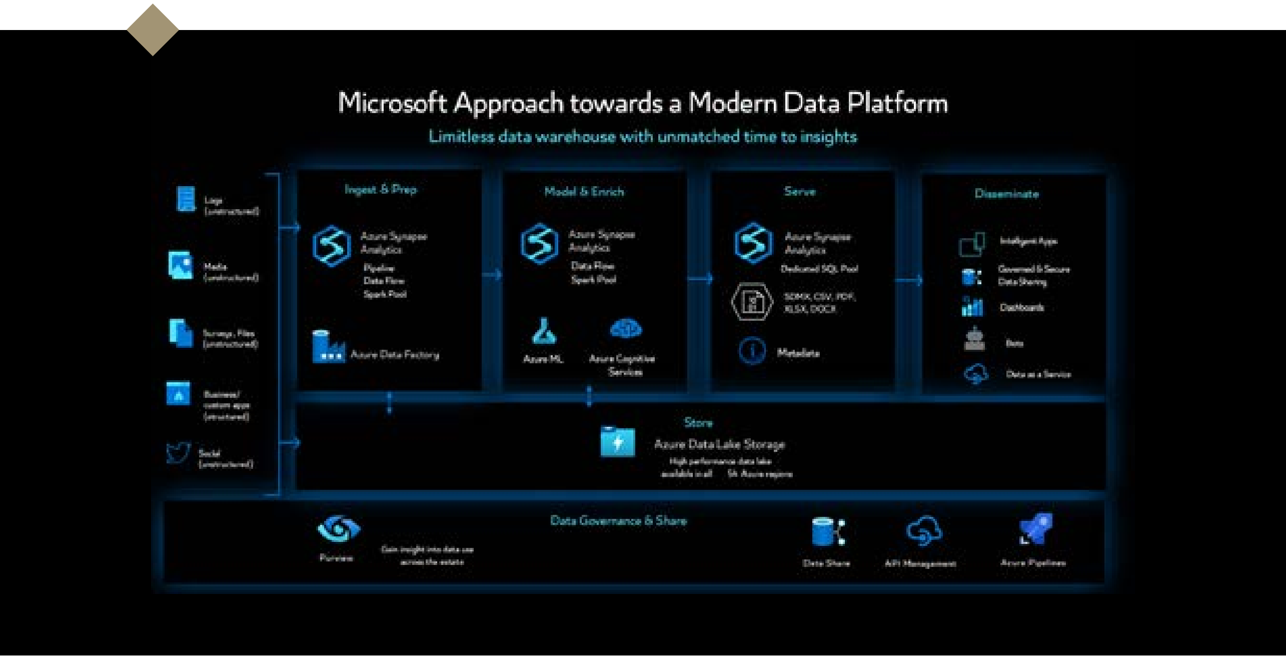
Ms. Al marri also identified the achievements made using decision-support system

1. The ability to view various types of information from a unified source in real-time.
2. Follow-up panels have been provided to decision-makers presenting the current situation.
3. Monitoring performance indicators in real-time and electronically.
4. The speed of extracting periodic reports (minutes instead of weeks).
5. Save time to conduct analytical studies instead of extracting data.
6. Improve data quality.
7. Training and accreditation of Qatari cadres in criminal security analysis.

Session Four: Microsoft

Modern Data Platform (Presentation link)

- Ms. Mirna Malaeb; Data & AI Solutions Specialist – Middle East Cluster at Microsoft



Microsoft Approach towards a Modern Data Platform

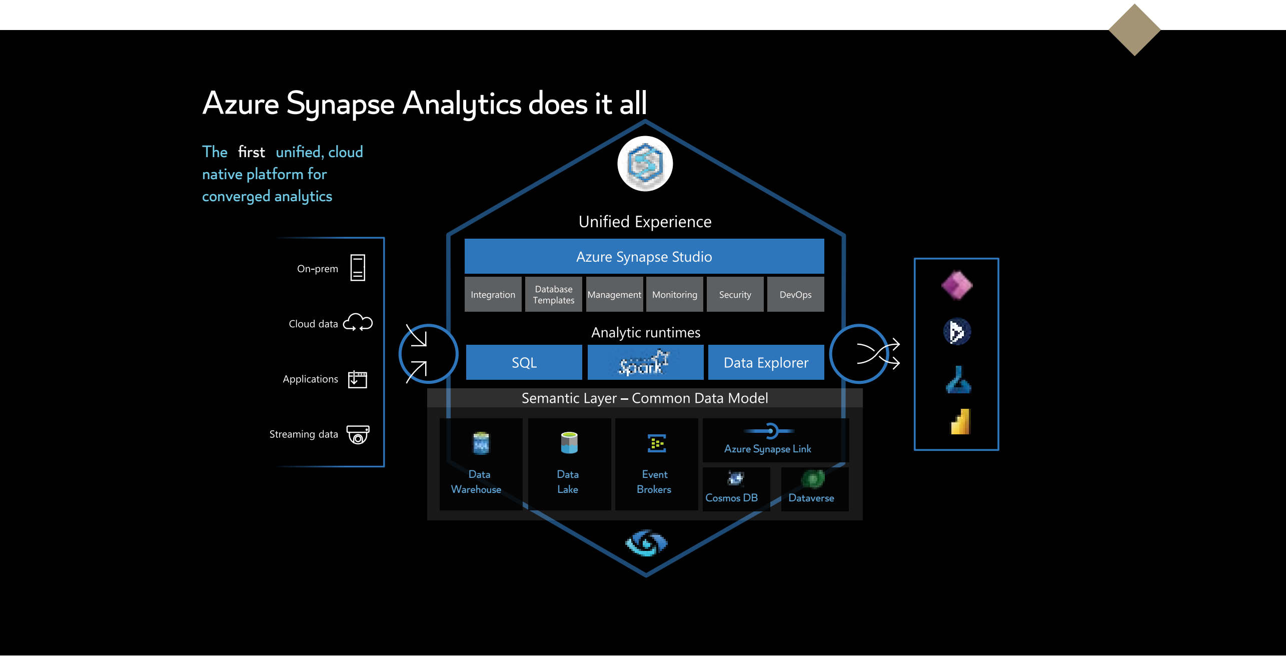
Limitless data warehouse with unmatched time to insights

Figure 7: Microsoft Approach towards a Modern Data Platform

Abstract of the Presentation:

As the final session, the fourth presentation by Microsoft Data Team dealt with all the technicalities involved in Building and Managing a Central Statistical Data Warehouse. Apart from providing the technology-oriented conceptual base for the unification of analytics for all types of data, the session could establish the means as to how the modern data platform can aid in achieving the data related to the four major pillars (Human + Social + Economic + Environment) of Qatar National Vision 2030. The presenter has also thrown light on the set of challenges that are generally encountered in managing the data like lack of uniformity in data, inconsistency of data, costs in involved in managing different kinds of data and others.

Also, the system that is practiced in hedging against different types of hurdles at Microsoft with the support of numerous frameworks is elaborated to widen the analytical outlook of the audience. The approach of Microsoft and the process of analytics performed with the help of Azure cloud services are depicted in figure 7 & 8 respectively.



Azure Synapse Analytics does it all

The first unified, cloud native platform for converged analytics

Figure 8:Azure Synapse Analytics operations

Further to the interesting technical tasks that are carried as the backbone to build a Centralized Statistical Data Warehouse, information about unique features of Azure Synapse like Limitless scale, Powerful Insights, Unified experience, Converged analytics and Unmatched security are detailed to emphasize the need for strong technical support to manage Central Statistical Data Warehouse. Highlighting the necessity to think out of the box, the session narrated how the Data Warehouse can help in Data Monetization, which further can ensure the self - sustainability of the nations. Finally, a roadmap is provided to achieve the overarching strategies as well as the goals of Qatar National Vision 2030.

PART II: SUMMARY AND KEY INSIGHTS

In this part of the report, the main findings of the webinar on building a Central Statistical Data Warehouse are reviewed and summarized, highlighting the path to develop a data lake, data warehouse and data mart .This part was drafted by Dr. Ahmad Hussein.

1. Data Warehouse

1.1. What Is a Statistical Data Warehouse?

A data warehouse centralizes and consolidates large amounts of structured data and metadata created by integrating data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making. It builds a historical record over the time that can be invaluable to data scientists and business analysts. Because of these capabilities, a data warehouse can be considered an organization's "single source of truth". A Statistical Data Warehouse can also be defined as a corporate Data Warehouse fully based on metadata, developed to support the production of multi-purpose statistical information. With an S-DWH different aggregate data on different topics should not be produced independently from each other but as integrated parts of a comprehensive information system where statistical concepts, microdata, macro data, and infrastructures are shared. Structured data such as Population data, Labor data, Education data, GDP, Foreign trade, price index, etc.

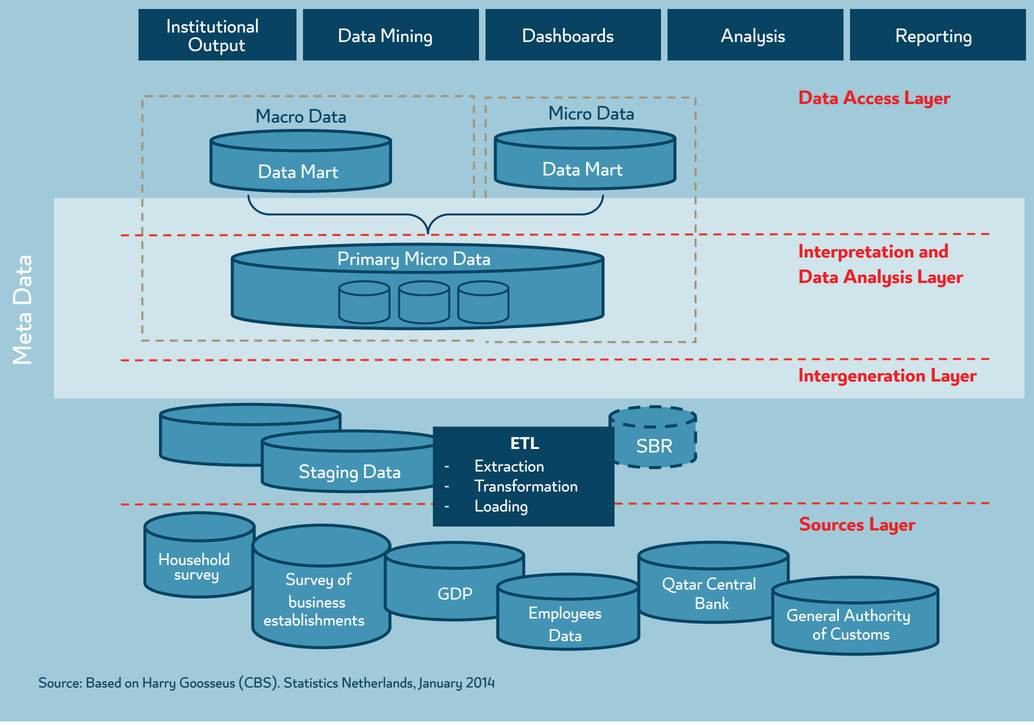
Data Integration Definition:

Statistical data integration involves combining data from different administrative and/or survey sources, at the unit level (i.e. for an individual person or organization) or micro level (e.g. information for a small geographic area), to produce new datasets for statistical and research purposes. This approach leverages more information from the combined datasets than is available from the individual datasets.

The data models underlying the data warehouse are not only oriented to produce specific statistical output or online analytical processing, as is the case of in many National Statistics Offices (NSOs) but also to sustain the production of statistical information in the various phases of the statistical life-cycle production process. A S-DWH model, instead of focusing on a process-oriented design, is based on data inter-relationships that are fundamental for different processes of different statistical domains.

The S-DWH data model must sustain the ability to realize data integration at micro and macro data granularity levels: microdata integration is based on the combination of different data sources with a common unit of analysis, one or a system of statistical registers, while macro data integration is based on the integration of different types of aggregate or dis-aggregate information in a common estimation domain .

S-DWH Information System Architecture



Source: Based on Harry Goosseus (CBS). Statistics Netherlands, January 2014

Figure 9 S-DWH Information System Architecture (Layers)

Data warehouses in the cloud offer the same characteristics and benefits of on-premises data warehouses but with the added benefits of cloud computin — such as flexibility, scalability, agility, security, and reduced costs. Cloud data warehouses allow enterprises to focus solely on extracting value from their data rather than having to build and manage the hardware and software infrastructure to support the data warehouse

Defining the S-DWH

European Statistics Systems network defined a Data warehouse as 'A central statistical data store for managing all available data of interest'. PSA adopted this definition to create new data / outputs, to produce the necessary information and to perform analysis for reporting, regardless of the data source. However, in statistical practice, the statistical data warehouse is the central data hub, which enables the connection and integration of all kinds of (new) data sources with statistical output.

Therefore, the S-DWH must not only support statistical production processes but also data collection processes by providing

1. A detailed and correct overview/insights of already available data sources

2. A framework for adequate data governance, including metadata management

3. Access to registries, sampling frames (like Business registries etcetera...)

4. Flexible data storage and data exchange between the processes

1.2. Why is a Data Warehouse important?

The process of modernizing the statistical system in the State of Qatar requires building a data warehouse, which includes economic, social, and environmental data, that makes different types of data available to users of different specializations. Data producers from various Ministers and institutions are required to take a role in building the said Data warehouse:

1. The data warehouse can contribute to the production of more indicators, to complete the quantitative aspects of the national development process, in accordance with the statistical activities specified by the United Nations Statistical Commission and its related working groups.

2. It can contribute to improve the quality and comprehensiveness of data, geographic and spatial, and making international comparisons.

3. The data warehouse can meet the needs of Ministries and government institutions, along with the private sector, universities, and research centers.

4. Information technology and advanced applications can contribute to build a data warehouse easily and enables the users to meet to their data needs easily.

5. The data warehouse can contribute to the process of data integration from different sources.

6. The data warehouse can contribute to to identify gaps in the available data that is provided by Ministries, government institutions and the private sector.

7. The process of building a data warehouse can contribute to the process of data integration from multiple sources.

8. The data warehouse can help to prepare national reviews and briefs the measures taken and the progress attained in achieving the SDG as well as for the sectors strategies of the 3rd National Development Strategy of Qatar.

1.3. The input datasets

The aim of developing a DWH is to create a set of fully integrated statistical data. Input for the data can be from different sources like surveys, administrative data, accounting data and census. Different data sources cover different populations. Some of the data sources like census cover all the population (units), some other cover all units with certain characteristics or influential units or other subpopulations. Other sources include less influential units but provide information only about few of them. To link these input data sources and to ensure that the data is linked to the specific unit and is compared with a specific target population is the crucial task. To do so, the source layer for the said data ware house ought to use the Classification of Statistical Activities (CSA) 2.0 to be endorsed by the statistical commission of the United nations on March 2022 (Annex 2)

1.4. Main Data Source

1.1.1. Surveys (censuses, household and business sample surveys)

Surveys based on the statistical data collection like labor force surveys, income-expenditure surveys, economic and business surveys, research and Development surveys, etc.

1.1.2. Administrative data

Administrative data is the set of units and data derived from an administrative source. A traditional definition of administrative sources is that they are files of data collected by government bodies. Editing data from different sources is required for different purposes: maintaining the register and its quality; for a specific output and its integrated sources; and to improve the statistical system.

1.1.3. Combined Data (Survey and Administration data)

A combination of both sources (Survey and Administration data) enhances the potential for research. Record linkage has advantages from a survey perspective and administrative data is used to update the frame of active units and to cover and estimate non-surveyed or non-responding units. The success of the actual linkage depends on the available information to identify a respondent in administrative records and on the quality of these identifiers.

1.1.4. Big Data (Based on availability)

Big Data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage and process data within a tolerable elapsed time.

Big data is often largely unstructured, UNECE Big Data classification:

- Social Networks (human-sourced information): Record of human experiences, previously recorded in books and works of art and later in photographs, audio and video. Data are loosely structured and often ungoverned.

- Traditional Business systems (process-mediated data): These processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc.

- Internet of Things (machine-generated data): Derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world .

1.5. Layers of S-DWH

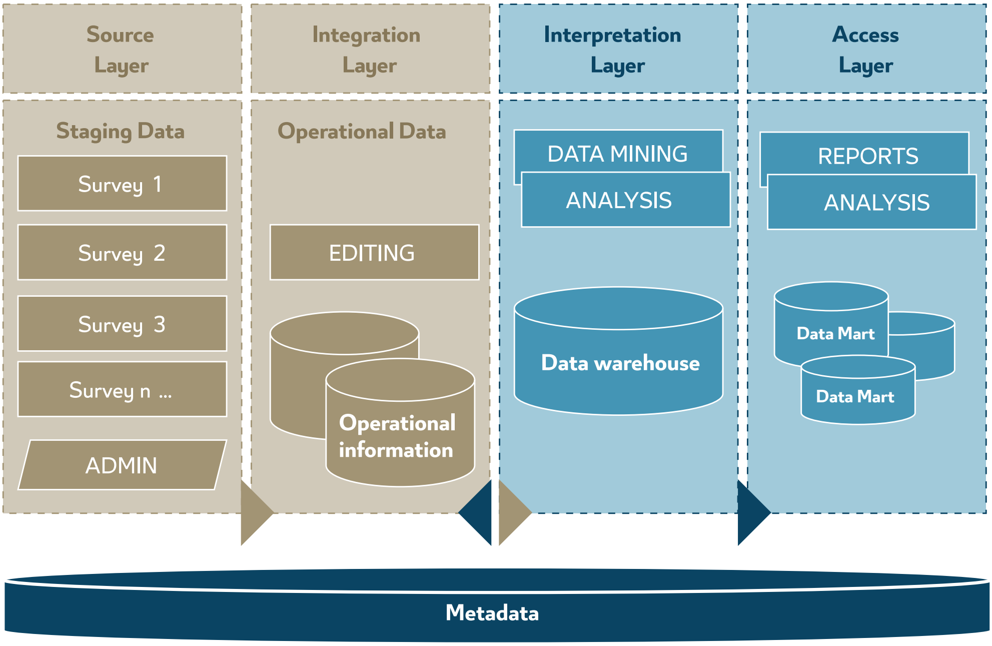
We can identify four conceptual layers for the S-DWH, starting from the bottom up to the top of the architectural pile, they are defined as:

A . Source layer is the level in which we locate all the activities related to storing and managing external data sources and is realized the reconciliation, the mapping, of statistical definitions from external to internal DW environment.

B . Integration layer is where all operational activities needed for any statistical production process are carried out; in this layer, data are manly transformed from raw to cleaned data.

C. Interpretation and data analysis layer enables data analysis or data mining function to support statistical design. Functionality and data are then optimized for internal users, specifically for statistician methodologists or statistician experts on specific domains.

D. Access layer for the access to the data is selected operational views, final presentation, dissemination and delivery of the information sought specialized for external, relatively to NSI or Eurostat, users .



Source: Harry Goosseus (CBS). Statistics Netherlands, January 2014

Figure 10 Information System Architecture

Figure 10 shows the grouped layers in two sub-groups:

- The first two layers for statistical operational activities, i.e. where the data are acquired, stored, coded, checked, imputed, edited and validated.

- The last two layers are for the effective data warehouse, i.e. levels in which data are accessible for analysis, design data re-use and for reporting (Figure 10).

1.6. Data Warehouse and Metadata management

Metadata is the DNA of the statistical data warehouse, defining its elements and how they work together. Thus, metadata plays a vital role in the S-DWH, satisfying two essential needs:

A. To guide statisticians in processing and controlling the statistical production;

B. To inform end-users by giving them insight in the exact meaning of statistical data and in order to meet these two essential functions, the statistical metadata must be:

i. Correct and reliable - the metadata must give a correct picture of the statistical data

ii. Consistent and coherent - the metadata driving the statistical processes and the reporting metadata presented to the end-users must be compatible with each other,

iii. Standardized and coordinated - the data of different statistics are described and documented in the same standardized way.

Meta Data and SDMX

The Statistical Data and Metadata Exchange (SDMX) is an initiative from several international organizations, which started in 2001 and aims to set technical standards and statistical guidelines in order to facilitate the exchange of statistical data and metadata using modern information technology.

The term metadata is very broad, and a distinction is made between “structural” metadata which defines the structure of statistical data sets and metadata sets, and “reference” metadata describing actual metadata contents. For instance, concepts and methodologies used, the unit of measure, the data quality (e.g. accuracy and timeliness) and the production and dissemination process (e.g. contact points, release policy, dissemination formats). Reference metadata may refer to specific statistical data, to entire data collections, or even to the institution that provides the data .

GOOD PRACTICE IN BUILDING A S-DWH: CASE FROM REPUBLIC OF KOREA (KOSIS)

As a gateway for Korea's official statistics, KOSIS offers a convenient one-stop service to full range of major domestic, international and Korean statistics. Currently, official statistics produced by over 120 statistical agencies covering more than 500 subject matters as well as the latest data on international finance and economy from international organizations (i.e. IMF, World Bank, OECD)

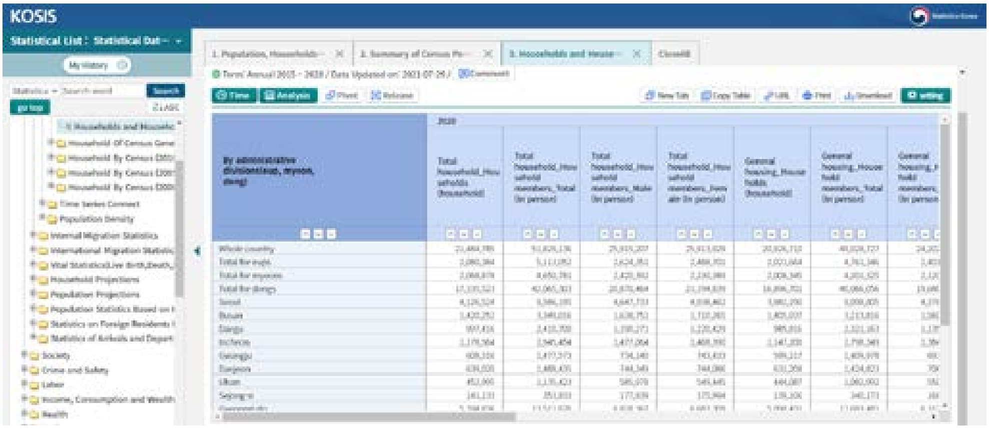


Figure 11: Screenshot of the KOSIS Data Warehouse

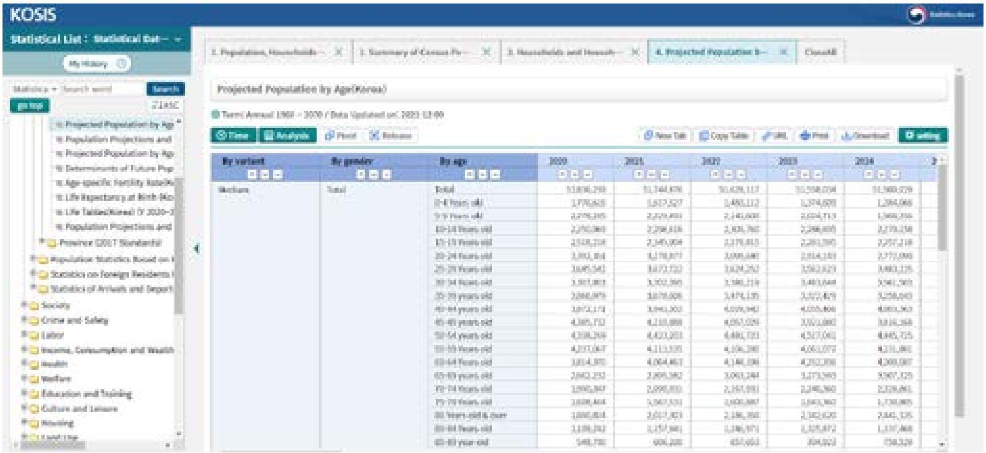


Figure 12: Screenshot of the KOSIS Data Warehouse 2

2. Data Lake

2.1. What is Data Lake?

The data lake is a storage repository that holds a large amount of data in its native, raw format. Data lake stores are optimized for scaling big amounts of data. The data typically comes from multiple heterogeneous sources and may be structured, semi-structured, or unstructured. The idea of a data lake is to store everything in its original, untransformed state. This approach differs from a traditional data warehouse, which transforms and processes the data at the time of ingestion .

2.2. Advantages of a Data Lake:

- Data is never discarded because it is stored in its raw format. This is especially useful in a big data environment, where you can't know what insights are available from the data, in advance.
- Users can explore the data and create their own queries.
- More flexible than a data warehouse because it can store unstructured and semi-structured data.

A complete data lake solution consists of both storage and processing. Data lake storage is designed for fault tolerance, infinite scalability and high-throughput ingestion of data with varying shapes and sizes. Data lake processing involves one or more processing engines built with these goals in mind and can operate on data stored in a data lake at scale.

2.3. Characteristics of Data Lake:

- Collect everything: A Data Lake contains all data; raw sources over extended periods of time as well as any processed data.

- Dive in anywhere: A Data Lake enables users across multiple business units to refine, explore and enrich data on their terms.

- Flexible access: A Data Lake enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines.

2.4. When to use a Data Lake?

Typical uses for a data lake include data exploration, data analytics and machine learning. A data lake can also act as the data source for a data warehouse. With this approach, the raw data is ingested into the data lake and then transformed into a structured query able format. Typically, this transformation uses an ELT (extract-load-transform) pipeline, where the data is ingested and transformed in place. Source data that is already relational may go directly into the data warehouse, using an ETL process, skipping the data lake.

2.5. Data Lake Stores

Often used in event streaming or Internet of Things (IoT) scenarios because they can persist large amounts of relational and nonrelational data without transformation or schema definition. They are built to handle high volumes of small writes at low latency and are optimized for massive throughput. Since It can be hard to guarantee the quality of the data going into the data lake.

Data Catalog – Data Lake will have a data catalog that provides metadata about data in the data store, which is used by the landing zone to provide data discovery and access capabilities for users. It provides data discovery and access capabilities to data scientists, business analysts, data engineers, and other users. It is integrated with data governance tools such as master data management (MDM) systems which store authoritative metadata about the organization's critical entities including customers, products, and suppliers. Data catalogs can be used in conjunction with master data management (MDM) systems to track data lineage and own data quality.

- **Data ETL pipelines:** Data Lake architecture uses data ETL pipelines to load data from existing data stores into the data store. This allows users to explore and access data from a variety of sources without moving it around or copying it, which can be very time-consuming and expensive.

- **Data ingestion pipelines** involve numerous steps in order to get data ready for analysis, including transformation, validation, and enrichment.

- **Data landing layer** – Data Lake architecture has a data landing zone, which is a layer that provides data management capabilities to prepare data for use by users. The goal of the data landing zone is to bring all data together, data from data warehouses as well as data from other sources.

- **Data access layer:** Data access layer provides a single point of entry into the lake which is used to discover, search and acquire data from data services that are exposed through the data access layer.

- **Data processing environment** – The data processing environment provides an execution context for interactive and batch analysis. It also manages workflows, which are used to orchestrate activities across data services. A data lake requires a data processing environment/context, which provides a data science workbench with the ability to process data from data services and applications where data is stored in a variety of formats.

- **Data governance** – Data lake's data needs to be governed by a data steward, who is responsible for ensuring data quality and protecting it from unauthorized access or changes. Data stewards can define security policies that govern how data can be accessed in the data store, catalog and landing zone by users with different roles. The architecture should also include data governance tools that can be used to review data lineage and data usage in real-time

- **Data security –** Data lakes require Authority security features such as authentication, authorization, auditing and encryption of data at rest and in data motion. Data security has become one of the most important (and often overlooked) data governance requirements for data lake adoption.

- **Data lineage:** Data Lake architecture should include lineage tools that can be used to monitor data movement within the architecture, which is critical for governance purposes. Data lineage provides insight into how data is accessed or moved throughout different parts of the architecture, so it can be reviewed in case there are any issues with data quality.

- **Data masking:** Data masking can be used in data lake architecture to ensure sensitive information. It remains private when copies of data are shared with data scientists or data analysts. Data masking can be achieved through data masking tools, which are used to generate automatically masks for data elements based on column values.

- **Meta-data registry:** Meta-data registry stores metadata about data services, catalog data models, and data definitions. The metadata registry helps ensure the consistency of metadata by providing a single integration point for all data service layers within the architecture. Data lake requires a meta-data registry to help data stewards define data services and metadata, which will be used by data scientists in the data science layer .

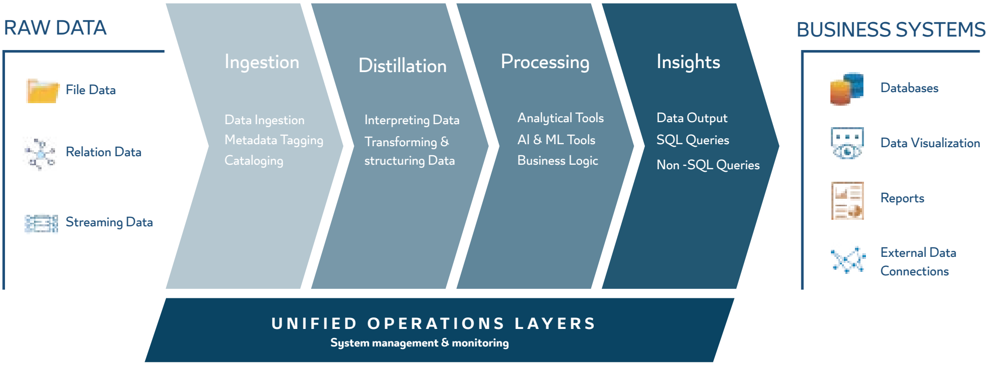
2.6. Data Lake project consist of three pillars:

1. Metadata repository (technical & conceptual).

2. Data Virtualization as technology to provide single data platform.

3. Security and Authorization to prevent data sets from unauthorized use.

2.7. Data Lake Architecture



Virtasant - Data Lake Architecture: A comprehensive Guide. Menur Hajdarbegovic . (2022)

Figure 13: Data Lake Architecture Layers

- **The Ingestion Layer** is tasked with ingesting raw data into the Data Lake. Modification of raw data is prohibited. Raw data can be ingested in batches or in real-time and is organized in a logical folder structure. The Ingestion layer can accommodate data from different external sources, such as social networks, IoT devices, wearable devices, data streaming devices.

- **The Distillation Layer** converts the data stored by the Ingestion Layer to structured data for further analysis. In this layer, raw data is interpreted and transformed into structured data sets and subsequently stored as files or tables. The data is cleansed, deformalized, and derived at this stage, and then becomes uniform in terms of encoding, format, and data type.

- **The Processing Layer** runs user queries and advanced analytical tools on structured data. Processes can be run in real-time, as a batch, or interactively. Business logic is applied in this layer and data is consumed by analytical applications. This layer is also known as trusted, gold, or production-ready.

- **The Insights Layer** is the output interface, or the query interface, of the Data lake. It uses SQL or non-SQL queries to request and output data in reports or dashboards.

- **The Unified Operations Layer** performs system monitoring and manages the system using workflow management, auditing, and proficiency management

3. Data Mart

3.1. What is a Data Mart?

A data mart is a subject-oriented data repository, similar in structure to the statistical data warehouse, but holding the data required for the decision support and BI needs of a specific group within PSA.

3.2. Why do we need a Data Mart?

The well-known IT company. Oracle confirmed that a data mart is a good solution for a data-driven organization that has a large central data warehouse. with a data mart, different departments use data and resources more efficiently because the only access to the data relates to them

Data Mart could be constructed solely for the analytical purposes of the specific group, or it could be derived from an existing data warehouse. Data marts are built using a dimensional data model as in (Figure 13)

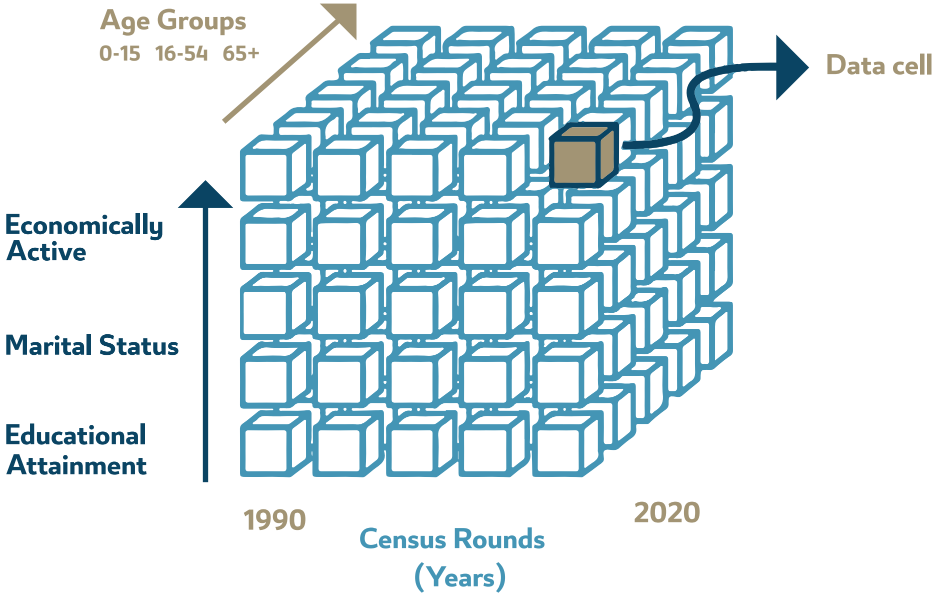


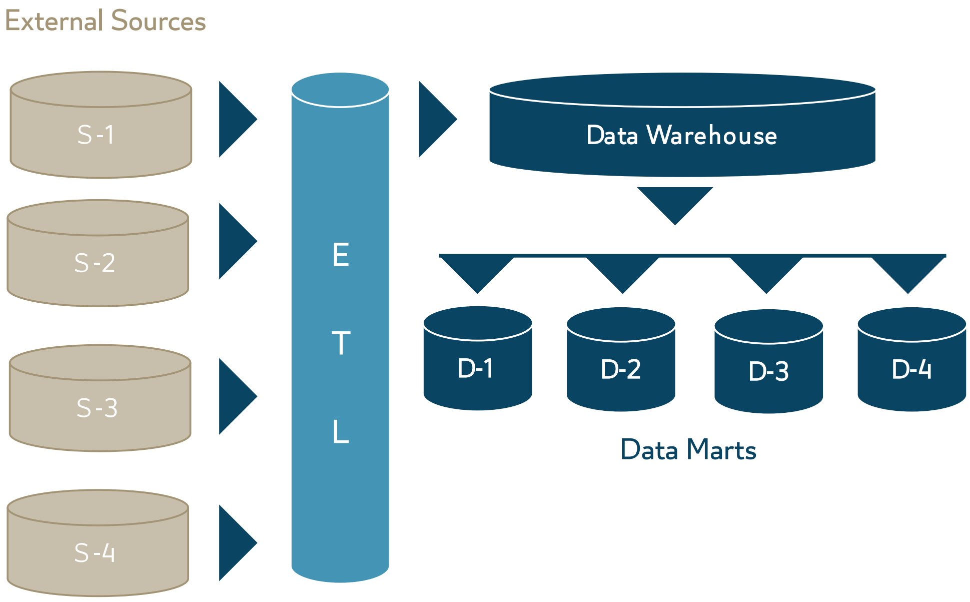
Figure 14 Data Mart (Data Cube)

It is usually designed for a smaller number of users. Data marts provide fast, specialized access and applications. There are differences between a data mart and a data warehouse, mostly because of the different natures of the desired results.

A data mart is likely to be configured for more generalized reporting for the specific business users within the department. Standard reports are more likely to be generated off the data mart, which will be much smaller than the data warehouse and will provide better performance. It is a normal practice for data marts to contain what are called "key performance indicators" (KPIs). KPIs often track the progress in implementing the projects related to the welfare of the Qatar population for example. .

A Data Mart begins with user-defined data analysis and emphasizes meeting the specific demands of the user in terms of content, presentation and ease of use. Data may be presented in data cubes with different formats, specialized to support different tools and software

3.3. Data Mart Architecture



Netsuite . Data Mart Defined: What It Is, Types & How to Implement. Andy Morris. (2021)

Figure 15: Data Marts Architecture

3.4. How is a Data Mart different than a Data Warehouse?

The key difference between a data lake and a data warehouse is that the data lake tends to ingest data very quickly and prepare it later the fly as people access it. With a data warehouse, on the other hand, you prepare the data very carefully upfront before you ever let it in the data warehouse.

Users tend to want to ingest data into the data lake as quickly as possible, so that companies with operational use cases, especially around operational reporting, analytics and business monitoring, have the newest data. This enables them to have access to the latest data and see the most updated information.

With the data lake, users often ingest data in the original form without altering it. This can be for speed reasons but can also be for other reasons including the desire to perform advanced analytics which can depend upon detailed source data. This would be analytics based on any kind of mining, whether it's (Text mining, Data mining, Statistical analysis, Anything involving clusters, Graph analytics)

PART III: CONCLUSION AND HOW TO GET STARTED?

Overall, the webinar session ran smoothly without any technical difficulties, The session was excellent, and all the resource persons presented their topics by imparting wealth of knowledge. The rich experience and exposure that the speakers shared in each session allowed for an in-depth and timely discussion that equipped participants with concrete ways to stay actively involved in efforts to achieve the goals of the webinar.

This part suggests some strategic insights on building a data warehouse **(Figure 14)**

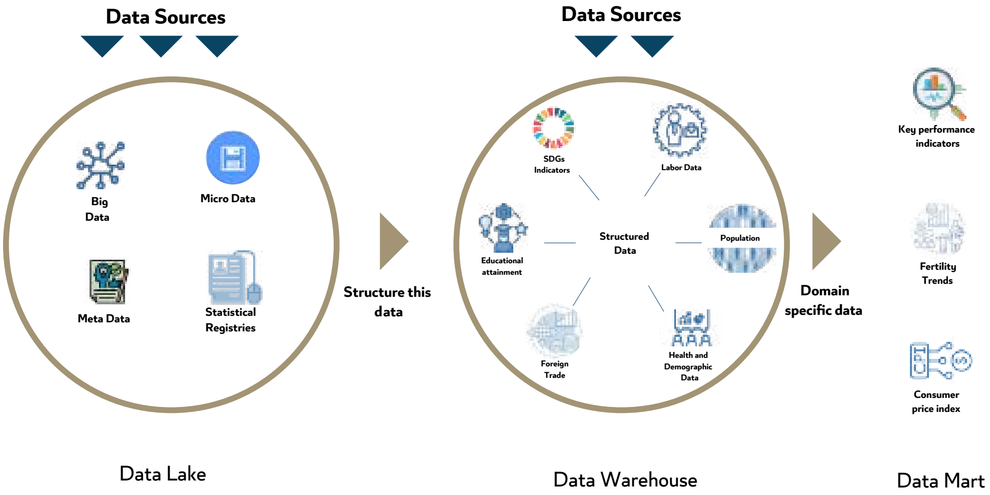
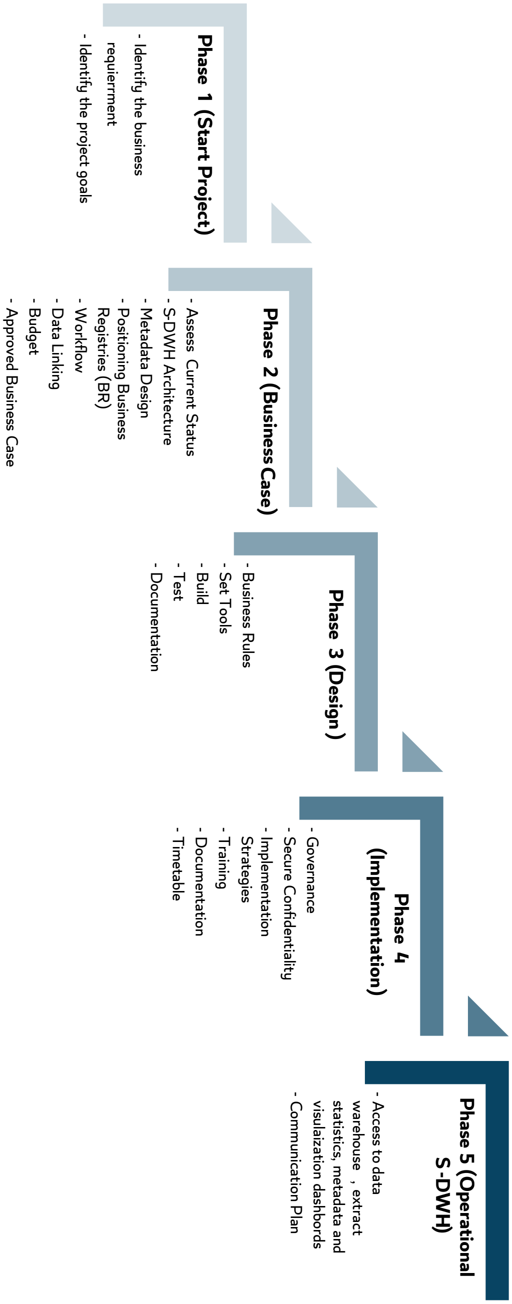


Figure 16 The linkage between data lake, data warehouse and data mart

HOW TO GET STARTED?

In building our Data warehouse we have to consider the following phases:



Phase 1 (Start Project)

- Identify the business requierrment
- Identify the project goals

Phase 2 (Business Case)

- Assess Current Status
- S-DWH Architecture
- Metadata Design
- Positioning Business Registries (BR)
- Workflow
- Data Linking
- Budget
- Approved Business Case

Phase 3 (Design)

- Business Rules
- Set Tools
- Build
- Test
- Documentation

Phase 4 (Implementation)

- Governance
- Secure Confidentiality
- Implementation Strategies
- Training
- Documentation
- Timetable

Phase 5 (Operational S-DWH)

- Access to data warehouse , extract statistics, metadata and visulaization dashbords
- Communication Plan

Figure 17: Phases of building S-DWH

REFERENCES

1. Amazon. What is a data lake? https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

2. Antonio Castro, Jorge Machado, Matthias Roggendorf and Henning Soller. (2020) Mckinsey Technology. - How to build a data architecture to drive innovation – today and tomorrow

3. Classification of Statistical Activities (CSA). 2.0 up published

4. Digital/ McKinsey. A smarter way to jump into data lakes (August 2017)

5. ESSnet Data. - WareHousing overview and results. Harry Goossens . (2014)

6. ESSnet Data. - WareHousing Stocktaking Pieter Vlag, Viviana di Giorgi, Sonia Queresma https://slidetodoc.com/essnet-on-datawarehousing-the-business-register-pieter-vlag/

7. Irene Salemink. - Dutch Enterprise Data Lake Fishing in clear water. Nether land Bureau of Statistics

8. Netsuite. Data Mart Defined: What It Is, Types & How to Implement. Andy Morris. (2021)https://www.netsuite.com/portal/resource/articles/data-warehouse/data-mart.shtml#:~:text=There%20are%20three%20types%20of,dependent%20and%20independent%20data%20marts.

9. Oracle, what is a Data Lake? https://www.oracle.com/big-data/what-is-data-lake/#:~:text=Here's%20a%20simple%20definition%3A%20A,diverse%20data%20from%20diverse%20sources

10. Oracle, what is Data Mart? https://www.oracle.com/autonomous-database/what-is-data-mart/

11. Techtarget. What is data architecture? A data management blueprint. (Craig Stedman & Ben Lutkevich). https://www.techtarget.com/searchdatamanagement/definition/data-lake

12. Trends Technology.- Architecture, Chief Technology Officer Branch Version 0.1 Date 2019-7-12

13. UNECE. - Seminar on New Frontiers for Statistical Data Collection (Geneva, Switzerland, 31 October-2 November 2012)

14. Virtasant - Data Lake Architecture: A comprehensive Guide. Menur Hajdarbegovic. (2022)

15. W.H. Inmon, Daniel Linstedt, in Data Architecture: A Primer for the Data Scientist, (2015) Data Marts

16. What is KOSIS? https://kosis.kr/eng/aboutKosis/Introduction.do

ANNEX

Annex 1: Opening Session

Speech of His Excellency Mr. Muhammad Abdulaziz Al-Nuaimi - Assistant to the President of Planning and Statistics Authority

In Opening the Webinar on Building Central Statistical Data Warehouse

Distinguished Guests and Dear Colleagues

Peace be upon you,

At the outset, I would like to welcome all of you to this webinar, which sheds light on an important topic related to building a central statistical data warehouse in the State of Qatar, and studying the best ways to provide data to users at the national and international levels. Further, I would like to thank you very much for accepting our invitation to participate in this webinar that is of great importance to us. In fact, we count on the results of the webinar, which are expected to enrich our knowledge related to the process of disseminating data and statistics and delivering them to users as quickly as possible through a smart and innovative applications.

Dear Guests,

The "Fundamental Principles of Official Statistics" issued by the United Nations Statistical Commission (UNSC) and approved by the Economic and Social Council (ECOSOC), emphasized that official statistics are a "public good" access to them must be provided simultaneously to all users. Means of access to statistical data must consider the diverse needs of users. Such statistics should be sufficiently comprehensive, so that they are made available to the public in an easy and transparent manner without them having advanced or specialized technical knowledge. They also enshrine the confidentiality of personal data related to individuals and establishments, which the Planning and Statistics Authority attaches great importance to.

Nowadays, as you all know, we have seen a data revolution that resulted in advanced technology, digital transformation and artificial intelligence, etc. This revolution was not only about large amount of data produced, but also about the increasing demand for it. This, in turn, requires the improvement of users' access to what they need; including the standard data, indicators and evidence that are used to measure progress towards achieving the SDGs of the 2030 Agenda for Sustainable Development. It is anticipated that the results of this workshop will contribute to developing a clear vision of the statistical and technical aspects of the data warehouse, enabling users who are not specialized in statistics and data to reach what they need in their work, and at the forefront of these are decision makers, policy makers, researchers, academics and workers in civil society organizations. Moreover, the data warehouse will also be helpful to us in preparing the Third National Development Strategy (NDS-3).

Dear Guests,

I welcome you once again and I am looking forward to the results of the webinar, which will undoubtedly help us in drawing a roadmap for building the desired data warehouse. This can only be achieved if the ministries, government agencies, the private sector and households provide the Planning and Statistics Authority with the necessary data supported by the relevant definitions and methodology.

In conclusion, I welcome and thank the presenters for their presentations, which will undoubtedly help us in achieving our aspirations. Besides, let me warmly thank the interpreters, for helping us understand each other. In particular, I would like to thank Dr. Ahmad Hussein and my colleagues in the Planning and Statistics Authority for their efforts in preparing and organizing this webinar. May God bless you in your efforts.

Peace be upon you,

Annex 2: The Classification of Statistical Activities, version 2.0

The Classification of Statistical Activities, version 2.0

1. The aim of the classification is to classify information about statistical activities (such as data collection, processing, dissemination, capacity development, statistical events, working groups, etc.). It provides a top-level structure to make it easier to find information. The domains 1-5 (subject-matter domains) can also be used to classify statistical data and products.

2. It is an analytical classification, and its components are not fully mutually exclusive. In some cases, an item can be classified in several areas and users can decide where to place it according to their specific need.

A. Domain 1 – Demographic and social statistics

1.1 Population

1.2 Migration

1.3 Labour

1.4 Education

1.5 Health

1.6 Income and consumption

1.7 Social protection

1.8 Human settlements and housing

1.9 Culture

1.10 Time-Use

B. Domain 2 – Economic statistics

2.1 Macroeconomic accounts and statistics

- 2.1.1 System of National Accounts
- 2.1.2 Balance of payments and international investment position
- 2.1.3 Government finance statistics
- 2.1.4 Monetary and financial statistics
- 2.1.5 System of Environmental Economic Accounting
- 2.1.6 Other macroeconomic statistics

2.2 Business statistics

2.2.1 Short-term business statistics

2.2.2 Structural business statistics

2.2.3 Business demography and business dynamics

2.2.4 Entrepreneurship

2.2.5 Multinational enterprise statistics

2.2.6 Other business statistics

2.3 Agriculture, forestry, fisheries

2.4 Energy

2.5 Mining, manufacturing, construction

2.6 Transport

2.7 Tourism

2.8 Banking, insurance, financial statistics

2.9 Commerce and other services

2.10 International trade

2.11 Prices

2.12 Science, technology, and innovation

C. Domain 3 – Environment statistics

3.1 Environmental conditions and quality

3.2 Environmental resources and their use

3.3 Environmental residuals

3.4 Hazardous events and disasters

3.5 Human settlements and environmental health

3.6 Environmental protection, management and engagement

D. Domain 4 – Governance statistics

4.1 Non-discrimination and equality

4.2 Participation

4.3 Openness

4.4 Access to and quality of justice

4.5 Responsiveness

4.6 Absence of corruption

4.7 Trust

4.8 Safety and security

E. Domain 5 – Cross-cutting statistics

5.1 Sustainable development

5.2 Human rights

5.3 Gender and special population groups

5.4 Living conditions and poverty

5.5 Climate change

5.6 Information society and digitalization

5.7 Circular economy

5.8 Other cross-cutting statistics

F. Domain 6 – Statistical infrastructure and methodology

6.1 Metadata

6.2 Statistical infrastructure

- 6.2.1 Classifications

- 6.2.2 Statistical business registers

6.2.3 Registers of population

6.2.4 Registers of dwellings and buildings

6.2.5 Agricultural registers

6.3 Data sources

6.3.1 Population and housing censuses

6.3.2 Business and agricultural censuses

6.3.3 Household surveys

6.3.4 Business and agricultural surveys

6.3.5 Administrative sources

6.3.6 Data science

6.3.7 Geospatial data

6.4 Data exchange and data sharing

6.5 Data editing and data linkage

6.6 Data analysis

6.7 Data disaggregation

6.8 Statistical confidentiality and disclosure protection

6.9 Data dissemination and communication

G. Domain 7 – Strategic and managerial issues

7.1 Institutional frameworks and principles; role and organization of official statistics

7.2 National statistical coordination

7.3 Quality management

7.4 Management of human resources

7.5 Management of IT, information and knowledge

7.6 Management of other resources

7.7 International statistical coordination

7.8 Capacity development

II. Explanatory notes

A. Domain 1 – Demographic and social statistics

1.1 Population

Covers population and demographic statistics; topics like demography, vital statistics, population structures and growth, demographic projections, families and households (marriages, divorces, household size).

Excludes:

- Causes of death (1.5)

1.2 Migration

Covers statistics on the movement of people, refugees, asylum seekers, workers, the duration of migration stay or absence, immigration, emigration, migrant flows and stocks, etc.

Excludes:

- Human trafficking (4.8)

1.3 Labour

Covers statistics on labour force, labour market, employment and unemployment; the more detailed topics include economically active population, labour conditions, health and safety at work (accidents at work, occupational injuries and diseases, work-related health problems), working time and other working conditions, strikes and lockouts, job vacancies, job creation, wages and salaries and labour costs.

Excludes:

- Migrant workers (1.2)
- Unemployment insurance and unemployment benefits (1.7)
- Trade union membership (4.2)
- Labour statistics collected from businesses (2.2).

1.4 Education

Covers educational participation, illiteracy, educational institutions and systems, human and financial resources invested in education, lifelong learning, vocational training and adult learning, impact of

1.5 Health

Covers health and mortality related statistical activities, including topics like life expectancy, health status, health and safety, health determinants (including lifestyle, nutrition, smoking, alcohol abuse), health resources and expenditure, health care systems, morbidity and mortality (including infant and child mortality), hospital admission, causes of illness and death, specific diseases (e.g. AIDS), disabilities, pharmaceutical consumption and sales, health personnel, remuneration of health professions, environmental health status, health inequality, health accounts.

Excludes:

- Work related health and safety (1.3)
- Victimization from criminal behaviour (4.8)
- Traffic accidents and injuries (2.6).

1.6 Income and consumption

Covers statistics on household income and expenditures from the household or individual viewpoint (all types of income and expenditure), including topics like distribution of incomes, in-kind income, income transfers received and paid, income or expenditure-based measures of poverty, consumer protection, consumption patterns, consumer goods and durables, household wealth and debts.

Excludes:

- Social protection schemes against various risks (1.7)
- Tax schemes (2.1.3)
- Poverty in a multidimensional sense (5.4)
- Living conditions (5.4).

1.7 Social protection

Covers statistics on measures to protect people against the risks of inadequate incomes associated with unemployment, ill health, invalidity, old age, parental responsibilities, or following the loss of a spouse or parent, etc., includes statistics on pension beneficiaries, social security schemes, social protection expenditure, etc.

Excludes:

- Insurance companies as economic actors (2.8)
- Pension funds as actors in financial markets (2.8).

1.8 Human settlements and housing

Covers statistical activities on housing, dwellings and human settlements.

Excludes:

- Rents (2.11).

1.9 Culture

Covers statistics dealing with cultural activities in society, such as theatre, cinemas, museums, libraries, mass media, book production, sports and recreation, expenditure and financing of culture.

1.10 Time Use

Covers statistics on the use of time by individuals, often related to work-life balance (reconciling family responsibilities and paid work) and unpaid work.

Excludes:

- Working time (1.3).

B. Domain 2 – Economic statistics

2.1 Macroeconomic accounts and statistics

Covers the compilation of macroeconomic accounts and macroeconomic statistics. It includes institutional sectors and industries. This category includes the aggregation, consolidation and reconciliation of the different components of economic statistics.

2.1.1 System of National Accounts

Covers the compilation of National Accounts. It includes the sequence of accounts, institutional sector accounts, satellite accounts, and regional accounts. It can be further broken down into: GDP production and expenditure accounts, supply and use tables and input-output tables.

Excludes:

- Balance of payments and international investment position (2.1.2)

- Government finance statistics (2.1.3)
- Monetary and financial statistics (2.1.4)
- System of Environmental Economic Accounting (2.1.5)
- Composite indicators covering multiple areas of the economy (2.1.6).

2.1.2 Balance of payments and international investment position

Covers balance of payments and international investment positions.

Excludes:

- International trade on goods and services statistics (2.10).

2.1.3 Government finance statistics

Covers statistics on government finance.

2.1.4 Monetary and financial statistics

Covers monetary and financial statistics of the private sector. Includes Foreign Direct Investment (FDI) statistics.

Excludes:

- Government finance (2.1.3).

2.1.5 System of Environmental Economic Accounting

Covers the System of Environmental-economic Accounting (SEEA). Includes ecosystem accounts and thematic accounts such as water accounts, energy accounts, forest accounts, ocean accounts etc. Excludes traditional environment statistics which are not compiled in the format of accounts.

Excludes:

- Agriculture, forestry and fishery statistics (2.3)
- Energy statistics (2.4)
- Environment statistics (3)
- Environmental resources and their use (3.2).

2.1.6 Other macroeconomic statistics

Covers macroeconomic indicators that are nowhere else classified. It includes business cycle analysis, composite indicators, etc.

2.2 Business statistics

Covers economy-wide statistics on the activities of businesses across different sectors (as opposed to 2.3-2.9 that deal with specific individual sectors). Covers topics like statistics on economic activities of businesses, business demography, business investment, business services, demand for services, industrial performance, enterprises by size class, industrial production, commodities, structure of sales and services, outputs of the service industries, non-profit institutions, labour statistics collected from businesses.

Excludes:

- Statistics on specific industry groups (2.3-2.9)
- Statistical business registers (6.2.2).

2.2.1 Short-term business statistics

Covers the compilation of business statistics with higher than annual frequency (e.g., quarterly, monthly, etc.) across multiple industries. It includes business tendency surveys, and employment and wages surveys if they are conducted among businesses.

Excludes:

- Wages reported as a source of household income (1.6)
- Short-term statistics on a specific industry (2.3-2.9)
- Short-term indicators on international trade (2.10)
- PPI (2.11).

2.2.2 Structural business statistics

Covers the structure, economic activity, and performance of businesses across multiple industries on an annual basis.

Excludes:

- Business demography and business dynamics (2.2.3)
- Entrepreneurship (2.2.4)
- Statistics on the performance of a specific industry (2.3-2.9)
- TEC and STEC (2.10).

2.2.3 Business demography and business dynamics

Covers demographic events on businesses such as birth, death, survival and other demographic events in the entire economy or in specific industries.

Excludes:

- Structural business statistics (2.2.2)

- Entrepreneurship (2.2.4).

2.2.4 Entrepreneurship

Covers the measurement of the determinants, performance and impact of entrepreneurial activities of people and organizations. It also includes the studies of entrepreneurship from the perspective of gender and other population sub-groups.

2.2.5 Multinational enterprise statistics

Covers the study of multinational enterprises across different industries. It also includes foreign affiliate statistics (FATS) but excludes foreign direct investments of multinational enterprises.

Excludes:

- FDI (2.1.4).

2.2.6 Other business statistics

Covers other areas of business statistics that are not elsewhere classified. Includes, for example, statistics on the impact of businesses on well-being and sustainability, etc.

2.3 Agriculture, forestry, fisheries

Covers statistics related to agriculture, forestry and fishery. Includes agricultural monetary statistics (agricultural economic accounts), agricultural structures (farm structure), trade in agricultural products, crop and animal production, agricultural commodities, agro-industry statistics (including food production and safety), organic farming and organic food, products source and use tables, forest and forest product statistics, trade in forest products, fisheries.

Excludes:

- Agricultural labour input (1.3)

- Government expenditure for agriculture, forestry, and fisheries (2.1.3)

- Forest resource assessment (2.1.5)

- Environmental accounting for agriculture, forestry, fisheries (2.1.5)

- Forest fire (3.4).

2.4 Energy

Covers energy supply, energy use, energy balances, security of supply, energy markets, domestic trade in energy, energy efficiency, renewable energy sources.

Excludes:

- Expenditure of households on energy consumption (1.6)

- Government expenditure on energy (2.1.3)

- Energy accounts (2.1.5)

- International trade in energy products (2.10)

- Prices of energy products (2.11).

2.5 Mining, manufacturing, construction

Covers statistics on the specific industrial activities of mining, manufacturing and construction.

Excludes:

- Prices of manufactured products and PPI (2.11).

2.6 Transport

Covers statistics on all modes of domestic transport (air, rail, road, inland waterways, sea, pipelines); includes topics like transport infrastructure, equipment, traffic flows, personal mobility, safety, transport accidents, energy consumption, transport enterprises, passengers and freight transport, transport sector trends, road traffic accidents.

Excludes:

- Transport for tourism (2.7)

- International transport (2.10)

- Cost of transport services (2.11).

2.7 Tourism

Covers statistics regarding visitor activity (such as arrivals/departures, overnight stays, expenditures, purpose of the visit, etc.) associated to different forms of tourism (inbound, domestic and outbound), tourism industries activity and infrastructure, and employment as reported by tourism industries.

Excludes:

- Tourism satellite accounts (2.1.1)
- Prices for tourist services (2.11)
- Environmental impacts of tourism (3.5).

2.8 Banking, insurance, financial statistics

Covers monetary, banking and financial market statistics, including financial accounts, money supply, interest rates, exchange rates, stock market indicators, securities, bank profitability, private sector insurance and statistics on the management of pension funds, financial soundness indicators.

Excludes:

- Beneficiaries of private pension funds (1.7)
- Financing of state pension and of other state social security schemes (1.7).

2.9 Commerce and other services

Covers domestic commerce and services. It includes trade in goods and services, and service industries not elsewhere classified.

Excludes:

- Transport (2.6)
- Tourism (2.7)
- Financial services (2.8)
- International commerce (2.10).

2.10 International trade

Covers trans-border trade in goods and services. Includes statistics on trade in goods by enterprise characteristics (TEC) and services trade by enterprise characteristics (STEC), international transport, tariffs, market access, foreign aid, development assistance, resource flows to developing countries.

Excludes:

- Balance of payments (2.1.2)
- Foreign direct investment (FDI) (2.1.4)
- Foreign affiliates statistics (FATS) (2.2.5).

2.11 Prices

Covers any statistical activity dealing with prices, including Purchasing Power Parities (PPPs) and international comparisons of GDP; covers topics like Consumer Price Indices (CPI), inflation, Producer Price Indices (PPI), price indices for specific products and services (e.g., Information and Communication Technology products or rents).

Excludes:

- Wages (1.6 and 2.2.1)
- Interest rates (2.8)
- Exchange rates (2.8).

2.12 Science, technology, and innovation

Covers Research and Development (R&D), innovation, patents, human resources in science, technology, and innovation, high-tech industries and knowledge-based services, biotechnology, financing of R&D, and innovation.

Excludes:

- ICT (5.6).

C. Domain 3 – Environment statistics

The breakdown in this domain follows the structure of the Framework for the Development of Environment Statistics (FDES) approved by the UN Statistical Commission in 2013. More detailed breakdown of statistical areas can be found in the FDES.**Excludes:**

- System of Environmental-Economic Accounting (2.1.5).

3.1 Environmental conditions and quality

Covers statistics about the physical, biological and chemical characteristics of the environment. FDES further breaks it down to three subcomponents: (1) physical conditions of the atmosphere, hydrographical information, geological and topographic information, and soil characteristics; (2) land cover, ecosystems and biodiversity, protected areas and species; and (3) quality of air, water and soil.

3.2 Environmental resources and their use

Covers statistics on environmental resources (assets) and is closely related to the asset and physical flow accounts of the System of Environmental-economic Accounting Central Framework (SEEA-CF). The statistical area covers the stocks, changes, production, trade and use (consumption) of both renewable and non-renewable natural resources, including aquatic resources, crops and livestock.

3.3 Environmental residuals

Covers statistics on the amount and characteristics of residuals generated by human production and consumption processes, their management and final release to the environment. The main groups of residuals are emissions to air (including greenhouse gases), water or soil, wastewater and waste, and the release of chemical substances.

3.4 Hazardous events and disasters

Covers statistics on the occurrence of hazardous events and disasters and their impacts on human well-being and the infrastructure. It covers both natural and technological hazardous events and disasters. The area covers statistics on the occurrence, frequency and intensity of hazardous events and disasters, as well as their impact on human lives and habitats, and the environment as a whole.

3.5 Human settlements and environmental health

Covers statistics on the environment in which humans live and work, particularly with regard to living conditions and environmental health. FDES breaks it further down to two subcomponents: (1) human settlements, infrastructure and exposure to potentially harmful environmental conditions; (2) environmental health, i.e., the impacts of environmental conditions and pollution on human health.

3.6 Environmental protection, management and engagement

Covers the statistics on resources dedicated to environmental protection, management and engagement. It includes four subcomponents: (1) environmental protection and resource management expenditure, (2) environmental governance and regulation, (3) hazardous event preparedness and disaster management, and (4) environmental information and awareness.

D. Domain 4 – Governance statistics

The breakdown in this domain follows the structure of the Handbook of Governance Statistics (HGS) approved by the UN Statistical Commission in 2020. More detailed breakdown on statistical areas can be found in the HGS. (This Domain replaces and expands subdomain 1.8 'Justice and crime' of the earlier version of the classification (CSA 2009)).

4.1 Non-discrimination and equality

Covers statistics measuring any differential treatment based on age, sex, disability, race, ethnicity, origin, religion or economic or other status that has the intention or effect of impairing human rights and fundamental freedoms. It covers both direct and indirect discrimination.

4.2 Participation

Covers statistics on the ways in which individuals take part in political and public affairs, including by registering to vote, voting or standing as a candidate in elections; being members of legislative, executive and judicial bodies; accessing positions in the public service; being a member of a trade union; and engaging, individually or as members of political parties and other non-governmental organizations, in political activities.

4.3 Openness

Covers statistics on the extent to which public institutions provide access to information and are transparent in their decision- and policy-making processes. More specifically, covers access to information, open government provisions, freedom of expression and media pluralism.

4.4 Access to and quality of justice

Covers statistics on the ability of people to defend and enforce their rights and obtain just resolution of justiciable problems, through impartial formal or informal institutions of justice and with appropriate legal support. Covers both criminal and civil justice, including the accessibility, effectiveness and quality of legal assistance, processes, decisions and outcomes.

4.5 Responsiveness

Covers statistics on whether people have a say in what government does and whether they are satisfied with the government's performance.

4.6 Absence of corruption

Covers statistics on: (1) the level of intolerance to corruption (i.e., ethical values, principles and norms that strengthen resistance to corruption practices); (2) the levels and patterns of corrupt practices; and (3) the response to corruption by the state. **4.7 Trust**

Covers statistics on people's trust in institutions as well as in other people, with a primary focus on the former, e.g., the parliament, the national government and the justice system.

4.8 Safety and security

Covers statistics on crime, victimization, violence, perceptions of safety, human trafficking, measurement of casualties directly provoked by armed operations, and the quality of law enforcement and criminal justice institutions.

E. Domain 5 – Cross-cutting statistics

Deals with conceptual or data work based on thematic approaches that require bringing together data across different domains to meet the data needs for policy agendas for development. This domain also covers statistical work that spans two or more of the previous subject-matter domains, or is not elsewhere classified.

5.1 Sustainable development

Covers work on indicators and frameworks to monitor sustainable development, well-being, etc., including indicators for assessing progress towards the Sustainable Development Goals.

Excludes:

- Environmental accounting (2.1.5)
- Environment statistics (3).

5.2 Human rights

Covers work on indicators and frameworks to monitor human rights (e.g., Human Rights-Based Approach to Data, Protocol of San Salvador indicators).

5.3 Gender and special population groups

Covers work on indicators and frameworks to monitor gender and special population groups, such as children, youth, older persons, persons with disabilities, minority groups, etc. (e.g., sets of gender indicators, indicators on ageing).

5.4 Living conditions and poverty

Covers work on indicators and frameworks to monitor living conditions and poverty, covering its different aspects: economic, social, etc. (e.g., Human Development Index, Multidimensional Poverty Index). Includes work on multidimensional methods to measure poverty, living conditions in the broad sense, social inclusion/exclusion, social indicators, and social situation.

5.5 Climate change

Covers work on climate change-related statistics, indicators and frameworks (e.g., sets of climate change-related indicators and statistics). This can concern the greenhouse gas emissions, climate change drivers, impacts, mitigation and adaptation.

Excludes:

- Environmental accounting (2.1.5)
- Environment statistics (3).

5.6 Information society and digitalization

Covers work on statistics, indicators and frameworks to monitor digital transformation and the use and impact of information and communication technologies (ICT) on society. Looks at these topics in a wider sense than economic aspects, including well-being in a digital society. Includes work on internet use, internet application, information and communications technology trade, computer use, broadband connectivity, e-commerce, etc.

Excludes:

- Business statistics (2.2).

5.7 Circular economy

Covers work on indicators to monitor the progress towards a circular economy (e.g., the European Commission's Circular Economy Indicators, the OECD Inventory of Circular Economy Indicators).

Excludes:

- Business statistics (2.2).

5.8 Other cross-cutting statistics, not elsewhere classified

Covers statistical subject-matter areas that are not classified in Domains 1-4 above or any of the previous cross-cutting statistics. Serves as a place-holder for any over-arching policy frameworks that may appear.

F. Domain 6 – Statistical infrastructure and methodology

6.1 Metadata

Covers developing, harmonizing and standardizing metadata models, structures and frameworks in the context of statistical information processing and dissemination, deals also with harmonizing statistical terminology and definitions.

6.2 Statistical infrastructure

6.2.1 Classifications

Activities related to developing, managing, maintaining and harmonizing classifications used in statistics, in subject matter areas (economic, social, environmental and governance statistics), as well as in methodology and management of official statistics.

6.2.2 Statistical business registers

Development and maintenance of statistical business registers.

Excludes:

- Agricultural registers (6.2.5)
- Business and agricultural censuses (6.3.2).

6.2.3 Registers of population

Development and maintenance of statistical population registers, covering the whole resident population. Includes civil and vital events registers.

Excludes:

- Register-based censuses (6.3.1)
- Registers of dwellings and buildings (6.2.4)
- Infrastructure (6.2.4).

6.2.4 Registers of dwellings and buildings

Development and maintenance of statistical registers of buildings, dwellings and infrastructure, covering all residential, non-residential buildings and dwellings, and elements of infrastructure.

6.2.5 Agricultural registers

Development and maintenance of agricultural registers.**Excludes:**

- Statistical business registers (6.2.2)
- Business and agricultural censuses (6.3.2).

6.3 Data sources

Deals with different methods of data collection from respondents and different forms of data sources at national level. Includes activities on electronic data reporting and Internet reporting which are not directly related to specific censuses or surveys. The two digit-level includes only activities that cannot be allocated to a three-digit item.

Excludes:

- Methods by which international organizations collect data from national producers (7.7)

6.3.1 Population and housing censuses

Covers methodology and organization of population and housing censuses, including register-based censuses.

Excludes:

- Development and maintenance of statistical registers of population (6.2.3)
- Collection and dissemination of national statistical results from population censuses by international organizations (1.1. or other relevant area of Domain 1)
- Civil and vital events registers (6.2.3)
- Administrative sources on persons generated by the social security system or kept for special population groups in their use for other statistical activities than population and housing censuses (6.3.5).

6.3.2 Business and agricultural censuses

Covers methodology and organization of economic and agricultural censuses.**Excludes:**

- Development and maintenance of statistical business registers (6.2.2)
- Development and maintenance of agricultural registers (6.2.5)
- Administrative sources on subsets of agricultural holdings or businesses and their activities in their use for other statistical activities than business and agricultural censuses (6.3.5)

6.3.3 Household and individual surveys

Covers methodology and organization of household sample surveys, and sample surveys of individuals, including sample designs; international surveys with direct data collection from households such as Living Standard Measurement Survey or World Health Survey.

6.3.4 Business and agricultural surveys

Covers methodology and organization of business and agricultural surveys, including sampling, and international surveys with direct data collection from businesses.

6.3.5 Administrative sources

Addresses the suitability of administrative sources for official statistics, the legal, organizational and conceptual problems of accessing administrative sources, the use of registers and other administrative sources in other contexts than censuses.

6.3.6 Data science

Addresses the practical use of combining multiple data sources (including big data) with the purpose of producing experimental and official statistics: methodological issues, covering quality concerns and fitness for purpose; legal and other issues in respect of access to data sources.

6.3.7 Geospatial data

Data and information having an explicit association with a location relative to Earth, such as topographic data, remote sensing, geodesy, satellite imagery and Earth observation data.

6.4 Data exchange and data sharing

Covers issues related to data sharing and data exchange at the collection, analysis, and dissemination phases, including data access, confidentiality and privacy issues. Covers data sharing and exchange both at national and international level.**Excludes:**

- Confidentiality and disclosure protection in the dissemination phase (6.8).

6.5 Data editing

Covers methodological, organizational and legal issues related to data editing at the collection phase, including data quality control, data imputation and use of geo-referenced data. Includes alignment to classifications and other statistical standards that facilitate data integration.

6.6 Data analysis

Covers methods of data analysis in official statistics for other purposes than editing/quality management, e.g., seasonal adjustment, methods for constructing composite indicators, identification of causal factors, extrapolation, scenario and model building, etc.

Excludes:

- Data editing and imputation for data quality control at the collection phase (6.5)

- Methods for data disaggregation (6.7).

6.7 Data disaggregation

Covers conceptual and methodological aspects to consider for data disaggregation. Includes small area estimation and use of novel sources to disaggregate traditional data.

6.8 Statistical confidentiality and disclosure protection

Covers legal, organizational and technical measures to safeguard confidentiality of statistical data; methods of releasing microdata while protecting against disclosure of individual data.

6.9 Data dissemination and communication

Policies, strategies, methods and techniques of data dissemination, design and organization of output databases such as data warehouses, data lakes and virtualized data. Includes feedback from users, data and metadata presentation, electronic dissemination (Internet), statistical portals and open data. Includes best practices for communicating with the media and work of NSO press offices.

G. Domain 7 – Strategic and managerial issues

The statistical areas in the domain are linked to the Generic Activity Model for Statistical Organizations (GAMSO). The updated structure of this domain is in line with the previous version of the CSA and includes also international activities, therefore the structure is somewhat different from GAMSO.

7.1 Institutional frameworks and principles; role and organization of official statistics

Covers activities dealing with developing, harmonizing and revising the institutional framework and principles of official statistics at national and international level, such as fundamental principles of official statistics, codes of practice, organizational and legal aspects of national statistical systems, functioning of the statistical systems, organization of statistical offices, ethics, value and promotion of official statistics, and the increasing role of national statistical offices in a wider data ecosystem including data stewardship.

Corresponds to GAMSO categories 1 Strategy and leadership (including 1.1 Define vision and 1.2 Govern and lead) and part of 3.1 related to managing legislation.

7.2 National statistical coordination

Covers the coordination within national statistical systems, as well as strategic partnerships with other data producers within a country. Covers the processes for setting up national statistical programs, including relationship with users and respondents etc.

Corresponds to GAMSO category 1.3 Manage strategic collaboration and cooperation

Excludes:

- Coordination between international statistical agencies (7.7).

7.3 Quality management

Covers quality frameworks and measurement of performance of statistical systems and offices: developing and administering a quality framework and tools to assure quality, i.e. compliance with the quality framework that should cover quality linked to the organizational framework, processes and products. Comprises work on quality indicators, user surveys, self-assessments, quality reviews or audits, certification and labelling of statistics. Quality documentation here refers to the organizational level and covers quality declarations, policies and relevant guidelines such as guidelines on handling of errors and revisions.

Corresponds to GAMSO category 3.10 Manage quality.

7.4 Management of human resources

Covers managing employee performance, recruitment, skills development, talent management and succession planning.

Corresponds to GAMSO category 3.8 Manage human resources.

7.5 Management of IT, information and knowledge

Covers coordination and management of information technology and information and knowledge. Management of IT includes management of the physical security of data and IT, IT assets and services, and managing technological change. Information and knowledge management includes the ownership or custody of records, documents, information and other intellectual assets held by the organization and the governance of information collection, arrangement, storage, maintenance, retrieval, dissemination and destruction. It also includes maintaining the policies, guidelines and standards regarding information management and governance.

Corresponds to GAMSO categories 3.4 Manage IT and 3.6 Manage information and knowledge.

Excludes:

- Data exchange and data sharing (6.4)
- Metadata (6.1)
- Data warehousing (6.9).

7.6 Management of other resources

Covers management of finances, buildings and physical space, and of any other resources not covered elsewhere in the classification. Managing finances covers the use of financial and accounting information to measure, operate and predict the efficiency and effectiveness of its activities, including procurement and contracts. Managing buildings and physical space covers maintenance of the building and allocation of physical space the organization occupies, including office space.

Corresponds to GAMSO categories 3.2 Manage finances and 3.9 Manage buildings and physical space.

7.7 International statistical coordination

Covers coordination of statistical activities across international and supranational statistical organizations, such as work of the UN Statistical Commission, Conference of European Statisticians and the Coordinating Committee of Statistical Activities.

7.8 Capacity development

Covers general bilateral and multilateral capacity development and technical cooperation activities, including coordination of capacity development.

Relates to GAMSO category 2 Capability management but is wider. GAMSO is focused on a view within a statistical organization while this area covers mainly international capacity development assistance offered by different kinds of donors (funds, countries, international organizations, etc.).

Excludes:

- Capacity development in specific subject areas (given under the relevant areas in Domains 1-5).

جهـــاز التخطيـــط والإحصـــاء

Planning and Statistics Authority

State of Qatar • دولـــة قطـــر

جهـــاز التخطيـــط والإحصـــاء
Planning and Statistics Authority
State of Qatar • دولـــة قطـــر

www.psa.gov.qa psaqa psa_qatar