Webinar on Building a Central Statistical Data Warehouse
Doha, 17 May 2022

حلقة عمل عن بعد بشأن بناء قاعدة بيانات إحصائية مركزية
الدوحة، 17 مايو 2022

PRIME MINISTER'S DEPARTMENT
DEPARTMENT OF STATISTICS MALAYSIA

StatsDW®
Statistics Data Warehouse
Brings Data To Life

MODERNIZING DATA GOVERNING: StatsDW AS NATIONAL AGENDA

17th MAY 2022 | 2.00 p.m
QATAR

Data Management & Integration Division
Department of Statistics Malaysia

StatsMalaysia
www.DOSM.gov.my

#KELUARGA MALAYSIA

20 OCT

2016 - 2030

CONTENTS

01 Introduction

02 StatsDW Phase I

03 StatsDW Phase II

04 Challenges

05 Innovation Impact

06 Benefits & Way Forward of StatsDW

StatsDW®
Statistics Data Warehouse
Brings Data To Life

INTRODUCTION

StatsMalaysia
www.DOSM.gov.my

#KELUARGAMALAYSIA

STRENGTHENING OF NATIONAL STATISTICAL SYSTEM

Efforts to **continuous** modernisation and innovation in order to provide effective service delivery.

Overflow of Data from Various Sources

1. Data is a highly valued commodity that is widely available across agencies and platforms.
2. Requires efficient and integrated data resource management.
3. A system framework is essential for holistic statistical production.

Technology Revolution

The need for complex and granular information down to the smallest geographical level and based on various specific demographic characteristics.

Data requests become more complex

1. Supports the transition to a more efficient and effective high-volume data integration and coordination platform.
2. In line with the digital government and public data era, for better and more effective and cost-effective data processing, structuring, analysis, and handling.

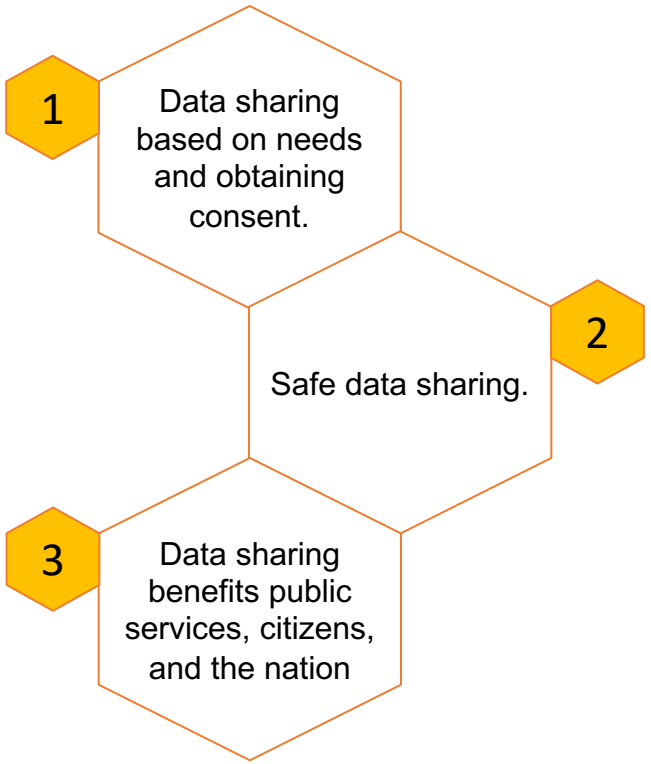
Environmental change

1. The global situation's uncertainties, and the need for readiness in the preparation of the latest inputs.
2. Rapid economic evolution and business models.
3. Social and environmental management are becoming more complex.

PUBLIC SECTOR DATA SHARING POLICY

The Public Sector Data Sharing Policy (DPDSA) was clarified in the Public Administration Development Circular number 2 year 2021 Public Sector Data Sharing Policy, which became effective on 27th August 2021.

Principles of Public Sector Data Sharing



Compliance in Data Sharing

INTRODUCTION

Statistics Data Warehouse (StatsDW) is an Enterprise Data Warehouse (EDW) containing **historical data** from various censuses and surveys conducted by the Department of Statistics Malaysia (DOSM) as well as compiled data for economic indicators.

StatsDW **enable fast and easy access to data** and facilitate wider access for users according to user requirements and level of access permitted.

"StatsDW is the 1st initiative by the DOSM to consolidate & integrate all the Statistical Data in one **Enterprise Data Warehouse (EDW)**"

EDW consolidates data from multiple sources, giving the right people access to the right information so that they can take necessary action. In turn, this helps in reporting, data analysis and acts as a core component of business intelligence/analytics environment.

OBJECTIVES

01

To consolidate all historical micro and aggregate data in an Enterprise Data Warehouse (EDW)

02

To enhance the quality and accuracy of data

03

To enable fast and easy access to data repository

04

To disseminate data to wider users in timely manner

MOTIVATION

1

DATA MANAGEMENT

- Various sources and form of data
- Management and compilation of time series data
- Data Migration

2

LIMITED DATA WAREHOUSE APPLIANCES

- Business Intelligence (BI)
- Extract Transfer and Loading (ETL)
- BI Integration vs Geographical Information
- Data Warehouse

3

SYSTEM CAPABILITY

- System integration not available
- Data access and preparation
- Limited data storage

4

SECURITY AND USER PROFILE

- Non-centered data storage location
- Difficulties in controlling data access

Data is stored in 66 silo /stand alone databases

Approximately 7,700 softcopies / hardcopies of publication are stored at various location

Approximately 5,000 micro data are stored separately at various Subject Matter Divisions

BENCHMARK STUDY

STATISTICS CANADA

CANSIM
-Data warehouse
-Data management
(http://www5.statcan.gc.ca)

STATISTICS NETHERLANDS

StatLine
-Micro data management
-Aggregate data management
-eDataBank
-Visualisation
(http://statline.cbs.nl)

EUROSTATS

-Data warehouse
-Micro data management
-Aggregate data management
(http://ec.europa.eu/eurostat/web/europeanstatistical-system)

STATISTICS KOREA

Korean Information Statistical Service (KOSIS)
-Micro data management
-eDataBank
(https://kosis.kr/eng/)

Guiding Principles

https://sqa.mampu.gov.my/index.php/ms/garis-panduan/garis-panduan-pengurusan-projek-ict-prisa

AUSTRALIAN BUREAU OF STATISTICS

ABS.Stat
-Data warehouse
-MyLab
(http://stat.data.abs.gov.au/)

STATISTICS NEW ZEALAND

NZ.Stat
-Data warehouse
-MyLab
- Visualisation
(http://nzdotstat.stats.govt.nz)

StatsDW JOURNEY

Enterprise data warehouse study

Report of study
- Technical needs
- Business needs
- Financials

Development of StatsDW
- Go Live 2016
- 7 Modules

Study for enhancement
- POC
- POT
- Paper work

Development of StatsDW 2.0
- Expansion of subjects
-System Integration within DOSM & OGAs

StatsDW PHASE I

StatsMalaysia
www.DOSM.gov.my

#KELUARGAMALAYSIA

ARCHITECTURE

Data Source

Structured Data

- **Microdata**
- **Aggregated Data**
- ***Historical Data***

Unstructured Data

Data Governance Data Integration (ETL)

Data Profiling & Data Quality

Data Migration, Data Integration (ETL)

Extract, transform & load (ETL)

Enterprise Data Warehouse (EDW)

Data Warehouse Appliance

Operational Data Store

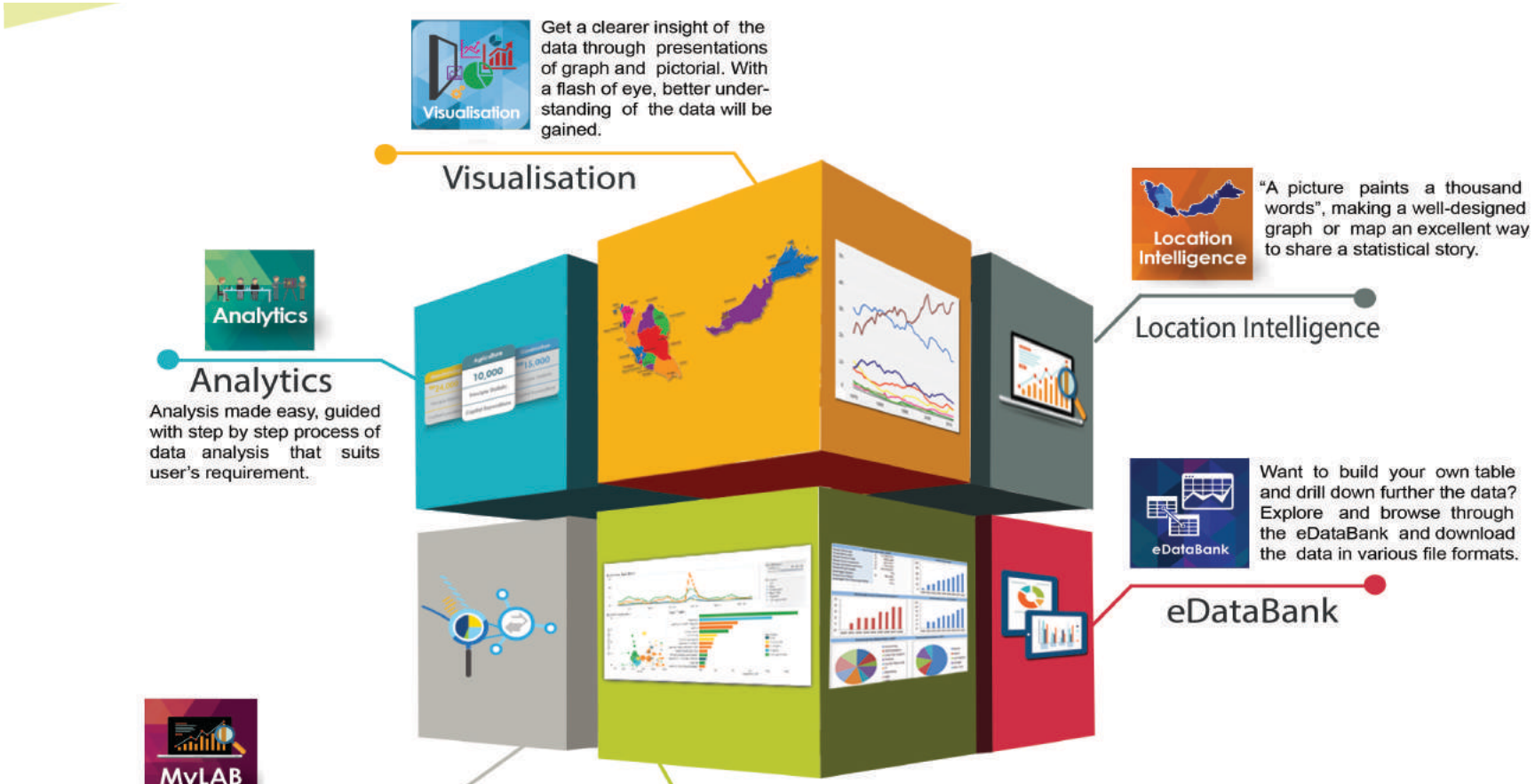
Modules

- eDataBank
- Visualisation
- Times Series
- Location Intelligence
- MyLab
- Analytics
- Mobile Apps

End User

- DOSM
- Stakeholder
- Public User
- Researcher
- Student

MODULES StatsDW PHASE I

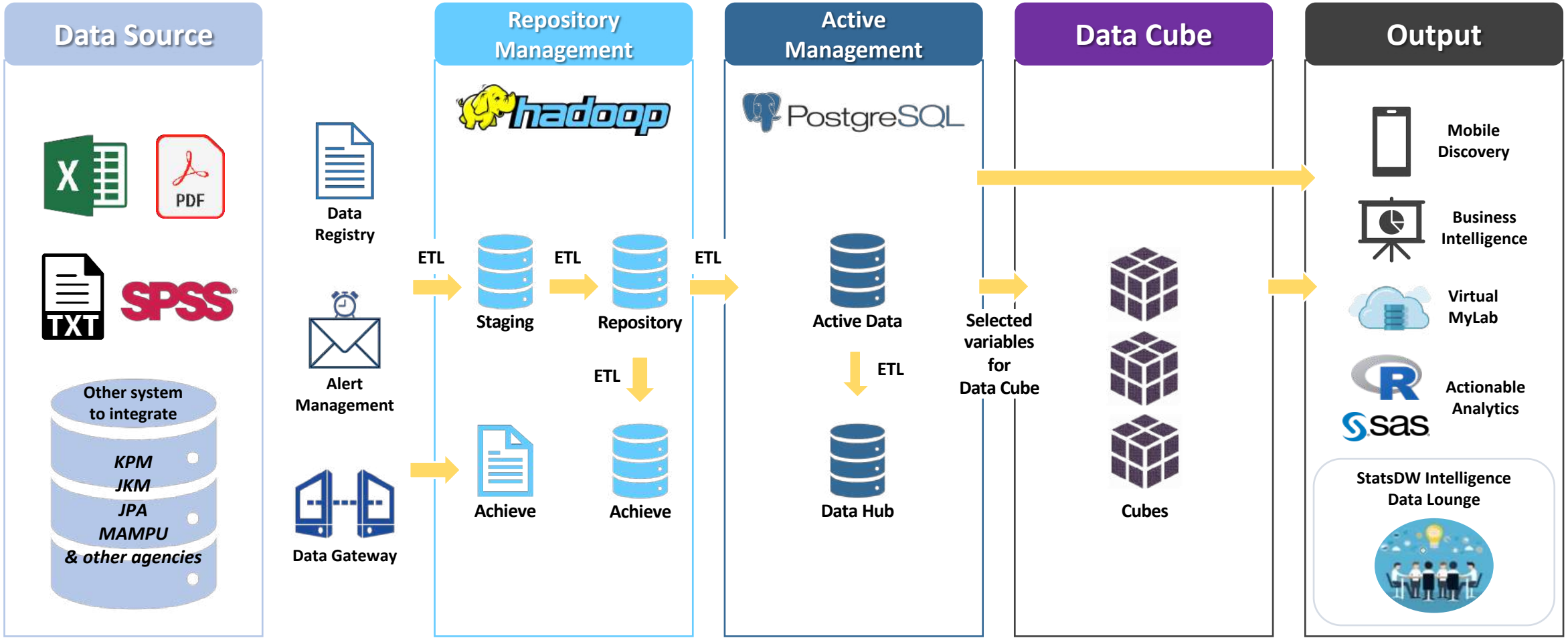


StatsDW Phase II

StatsMalaysia
www.DOSM.gov.my

#KELUARGAMALAYSIA

StatsDW PHASE II ARCHITECTURE

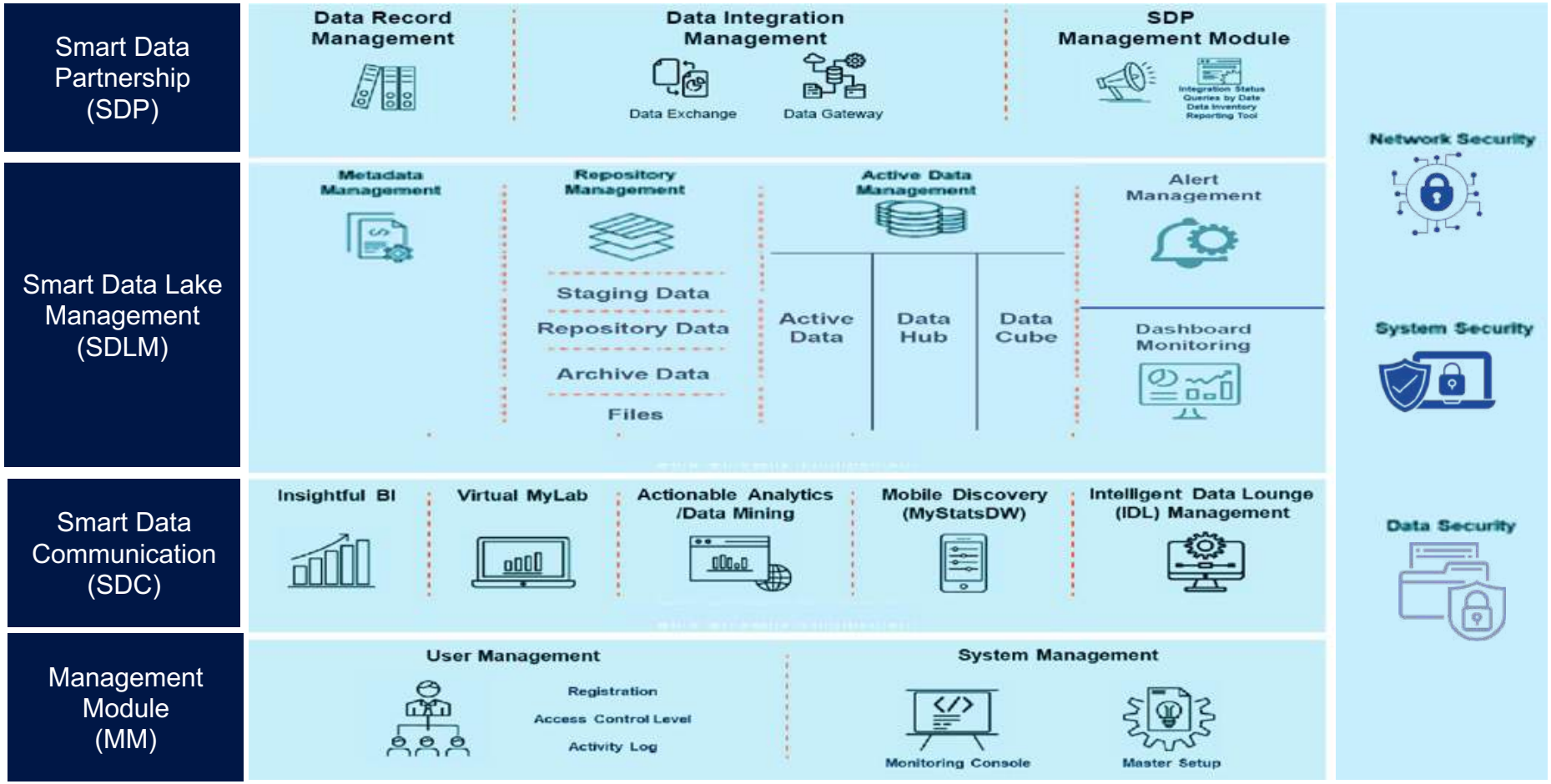


User Profile Management

Note :
KPM – Ministry of Education
JKM – Welfare Department
JPA – The Public Service Department
MAMPU – Malaysia Administrative Modernisation and Management Planning Unit

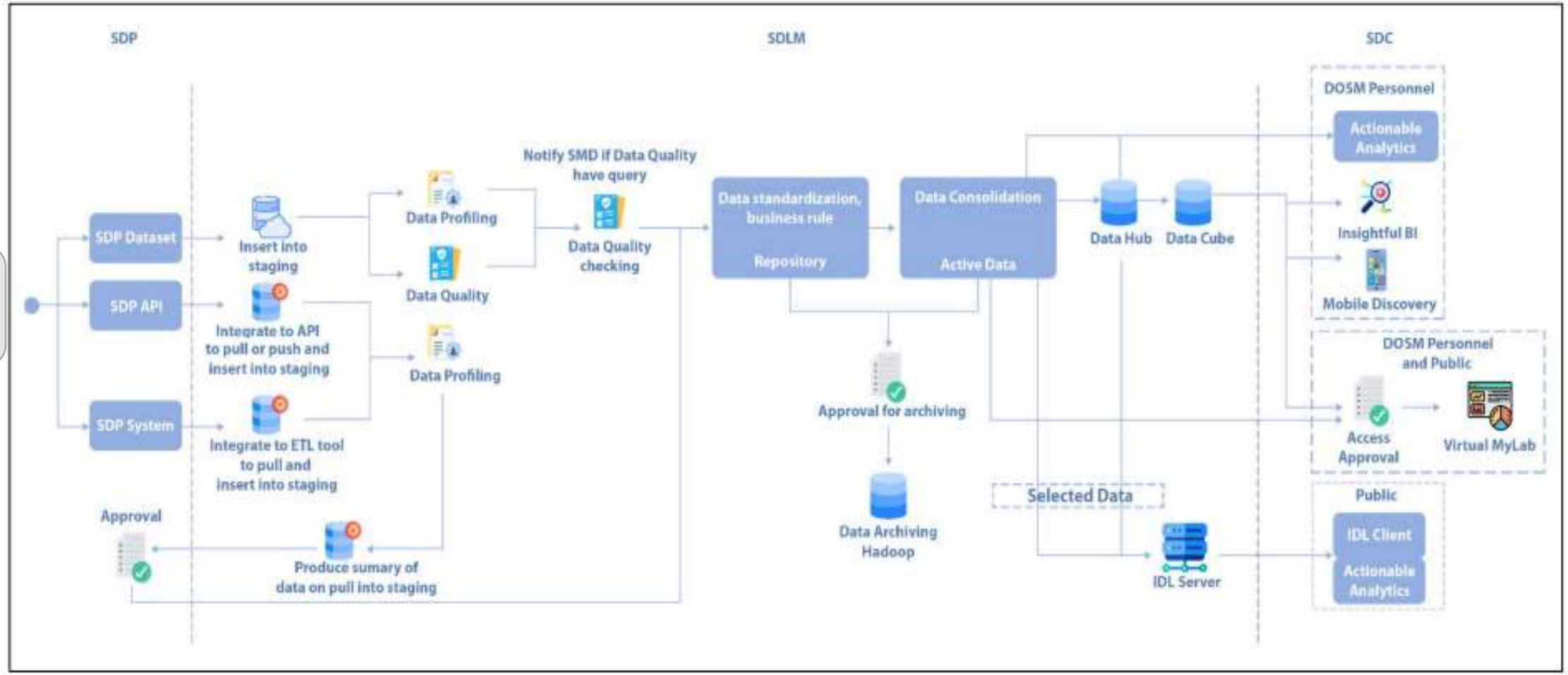
StatsDW PHASE II MODULES

StatsDW Phase II will facilitate the storing, retrieving, processing activities of data for more efficient organising, administration, governing and securing of datasets



DATA FLOW DIAGRAM

Direct API
1. BNM
2. MET M'sia
3. JKM



SMART DATA PARTNERSHIP (SDP)

SDP

consists of 3 sub-modules which are:

Data Record Management

Users able to register, submit & edit for acquisition by StatsDW 2.0 environment

Data Integration Management

Integration of data with internal DOSM systems & other agencies via API. Users able to download data from repository & staging environment

SDP Management Module

Dashboard for notification alerts and alerts & monitoring of data movement process from acquisition to storage

SMART DATA PARTNERSHIP (SDP)

User's Perspective

Data Record Management

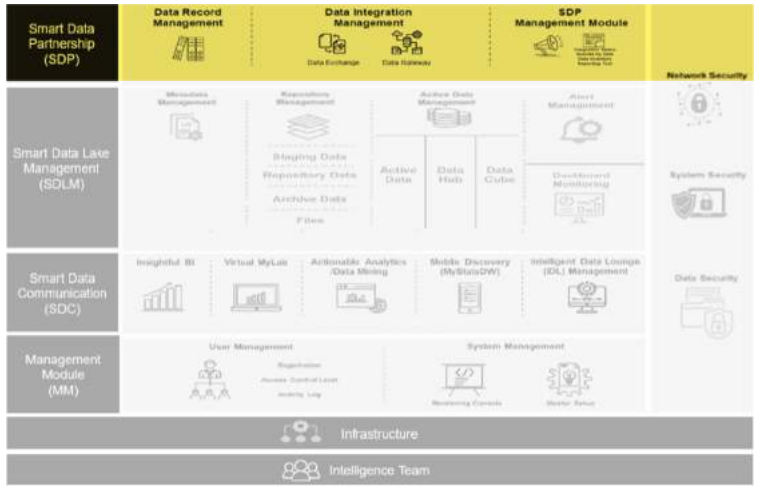
1. Dataset registration
   - Includes metadata, technical notes, questionnaire
2. Director's role to **approve dataset** to be registered
3. **Dissemination level** for the registered dataset – to be decided by the SMD
   - Whether the registered dataset to be stored in repository only or to be both, stored and visualized

Data Exchange Management

1. Dataset uploading
   - Includes metadata, technical notes, questionnaire (structured format)
2. If the dataset registered in the DR is not tally with the uploaded dataset, SMD will be notified via Alert Management

Data Gateway Management

1. **StatsDW®** integration with other systems – in DOSM and OA
2. Director's role to **approve the integration** that will be conducted

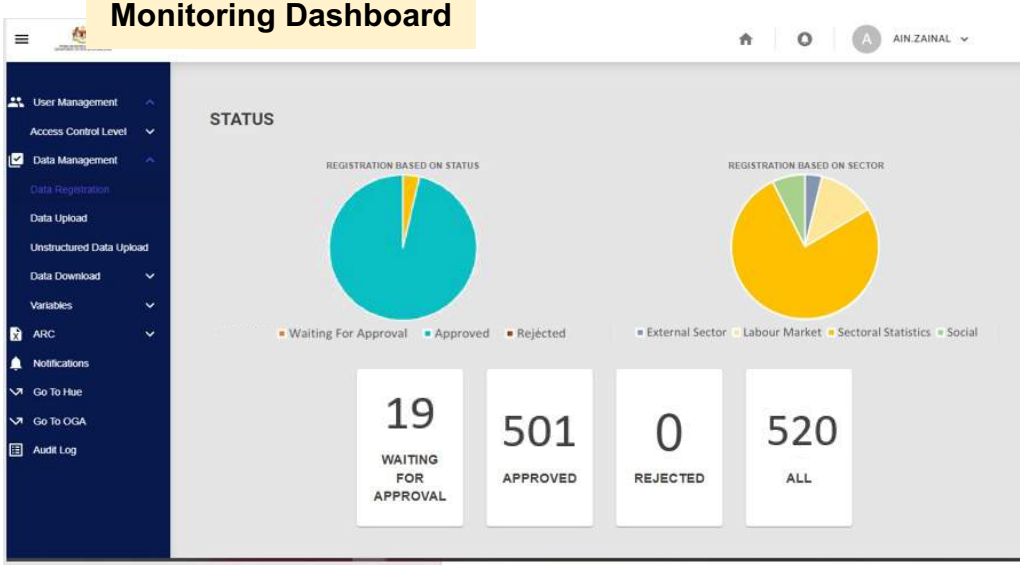


Key Points

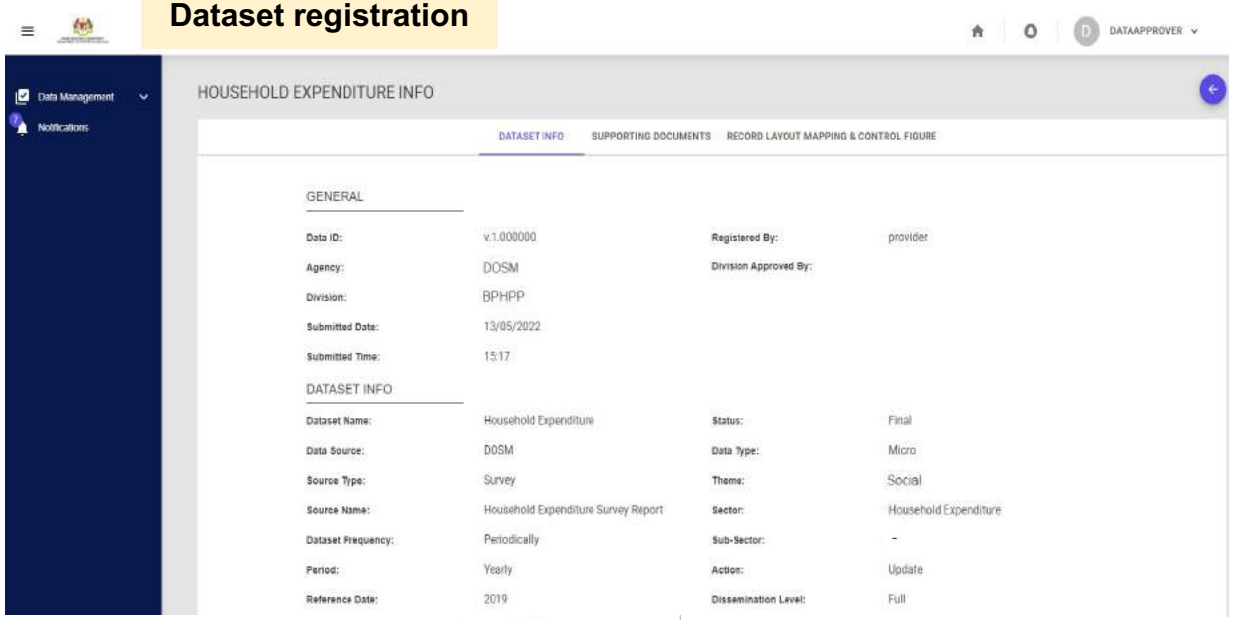
1. **Register**
2. **Upload**
3. **Integrate**

DATA RECORD MANAGEMENT

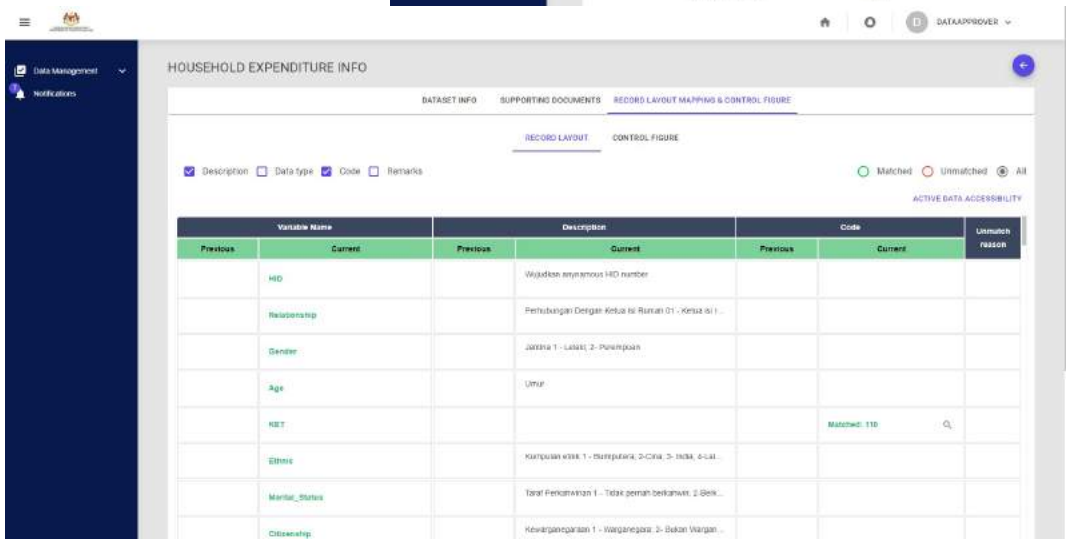
Monitoring Dashboard



Dataset registration



Record Layout Mapping



SMART DATA LAKE MANAGEMENT (SDLM)

SDLM Module Functions

- Able to store structured and unstructured data, as well as files with Hadoop equipped - an open source large storage system, and complemented with high-storage capacity.
- Stores and enables user control of data throughout its lifecycle in the system.
- Facilitate to store, retrieve, process activities of data by using Extract Transform & Load (ETL) tools.
- Deliver better solutions in preparing data, providing alerts and monitoring data activities.
- Monitoring dashboard of all data processes and every data element is given a unique identifier and tagged with metadata.

SMART DATA LAKE MANAGEMENT (SDLM)

User's Perspective

Metadata Management

1. Metadata input to be key in by SMD
   - If query – SMD to amend
2. Only query free metadata will be uploaded

Repository Management

1. SMD has to amend the data structure of the dataset if there is query from the RM
3. All datasets that is uploaded by SMD will be stored in RM

Active Data Management

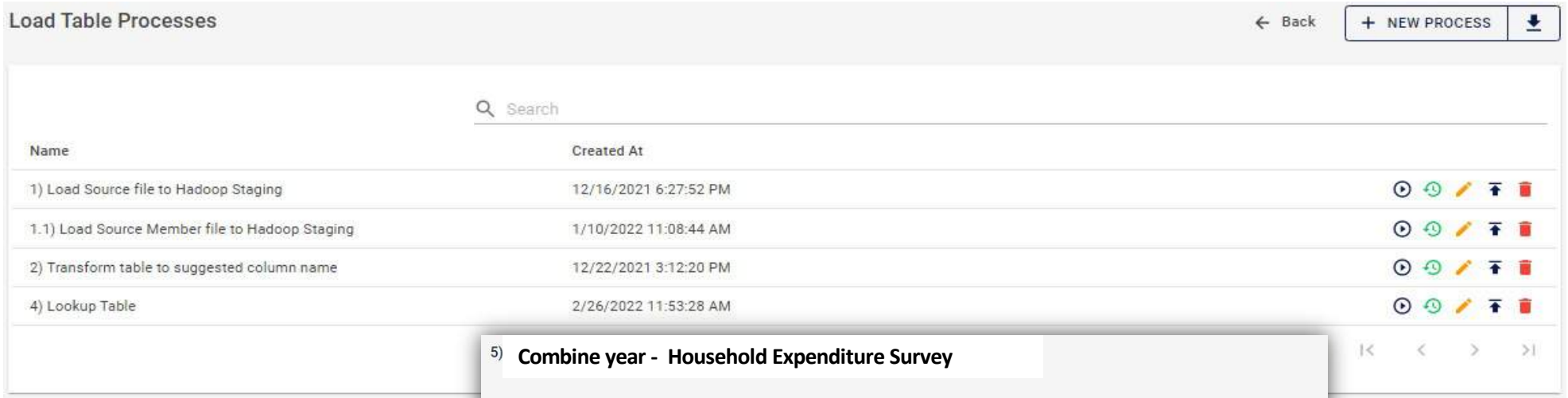
1. SMD may request their desired **hub** in the ADM
   - Data hub will be created based on user's request

Key Points

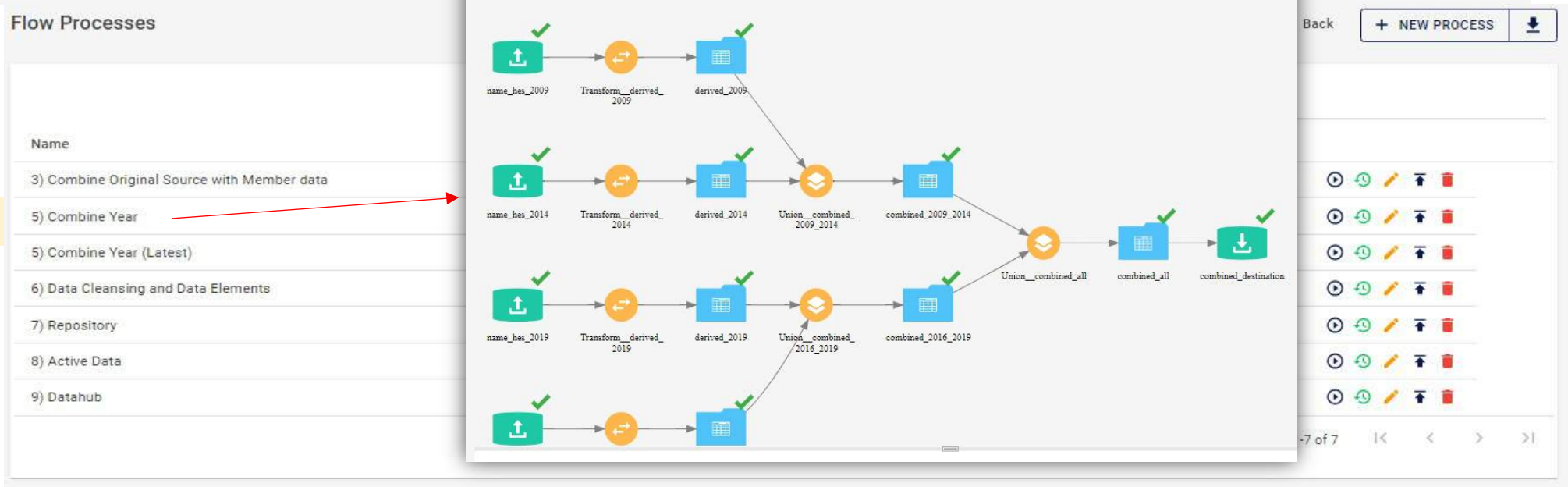
1. Metadata
2. Repository
3. Data Hub

EXTRACT TRANSFORM LOAD (ETL) PROCESS

Storing



Processing



EXTRACT TRANSFORM LOAD (ETL) PROCESS

Household_Expenditure Survey

Load source table

Column mappings

Source	dsT022019HES
Destination	ds_bphpp_hes_2019_q

Search

Source	Destination	Source Type
HID	hid	TEXT
Relationship	relationship	TEXT
Gender	gender	TEXT
Age	age	DOUBLE
KET	ket	TEXT

Transform table to suggested column name

Column mappings

Source	ds_bphpp_hes_2009_combine
Destination	name_household_expenditure_2009

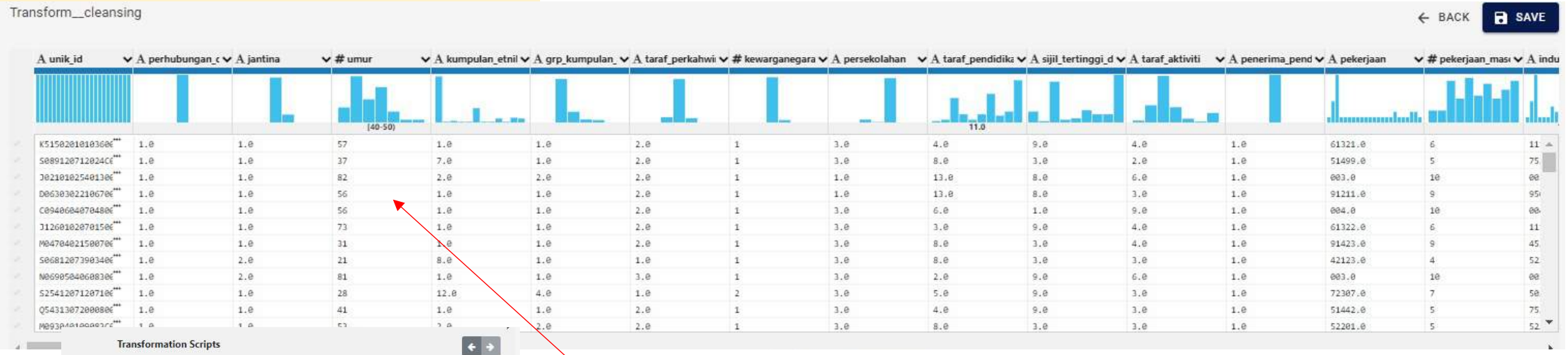
Metadata: Metadata_HES_2009

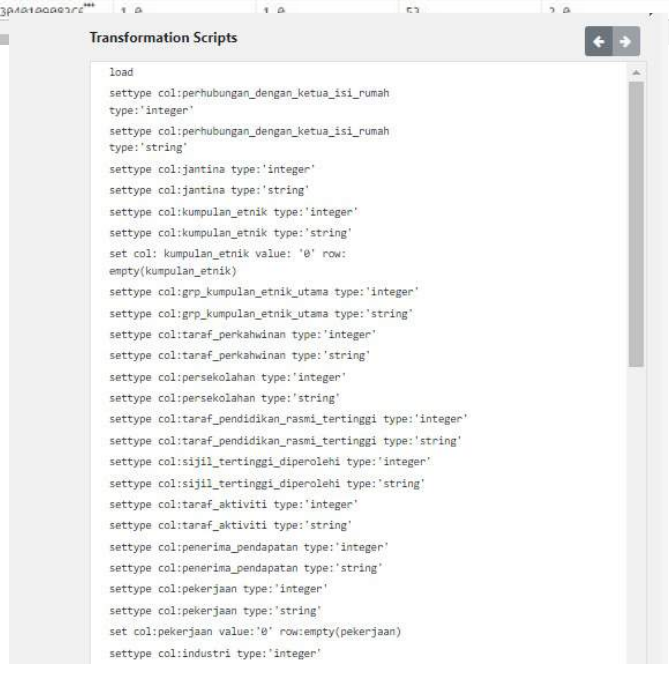
Search

Source	Destination	Source Type	Destination Type
hid	unik_id	STRING	string
relationship	perhubungan_dengan_ketua_isi_rumah	DOUBLE	double
gender	jantina	DOUBLE	double
age	umur	DOUBLE	double
ket	kumpulan_etnik	DOUBLE	double

EXTRACT TRANSFORM LOAD (ETL) PROCESS

Data standardization & Business rule





Data standardization :
The user shall use the master data file, which located in the repository database for data standardization. Data standardization is the process of converting data to a common format to enable users to process and analyze it.

i.e: Standardize and group data by age range

scl_age_main_id	scl_age_main_desc_e	scl_age_main_desc_m
1	0-14	0-14
2	15-64	15-64
3	65 and over	65 ke atas

```
CASE
WHEN UMUR BETWEEN 0 AND 14 THEN 1
WHEN UMUR BETWEEN 15 AND 64 THEN 2
WHEN UMUR BETWEEN 65 AND 134 THEN 3
END
```

Business Rule:
A business rule is a statement that imposes some form of constraint on a specific aspect of the database, such as the elements within a field specification for a particular field or the characteristics of a given relationship.

i.e :

i. Household Income Survey: starting in 2007, 30% of the number of households with 22 variables.
ii. Population and Housing Census: starting in 1980, 2% of the total population data of Malaysia.

SMART DATA COMMUNICATION (SDC)

User's Perspective

Insightful BI

1. SMD have to determine the dashboard requirement
   - IBI admin will develop the dashboard
   - SMD will do the dashboard checking in the development server
2. If SMD add new variable which is not in the data hub yet, they have to go through DR process
3. External dashboard and Internal Dashboard

Mobile Discovery

1. User can do lite analysis
2. User can drag and drop variables, and dashboard can be visualized in mobile view
3. Can be accessed (free), using Android and IOS

Virtual MyLab

1. Single Sign On (SSO) access through DOSM Active Directory
2. User can download the output

Key Points

1. Dashboard
2. Virtual
3. Mobile

SMART DATA COMMUNICATION (SDC)

User's Perspective

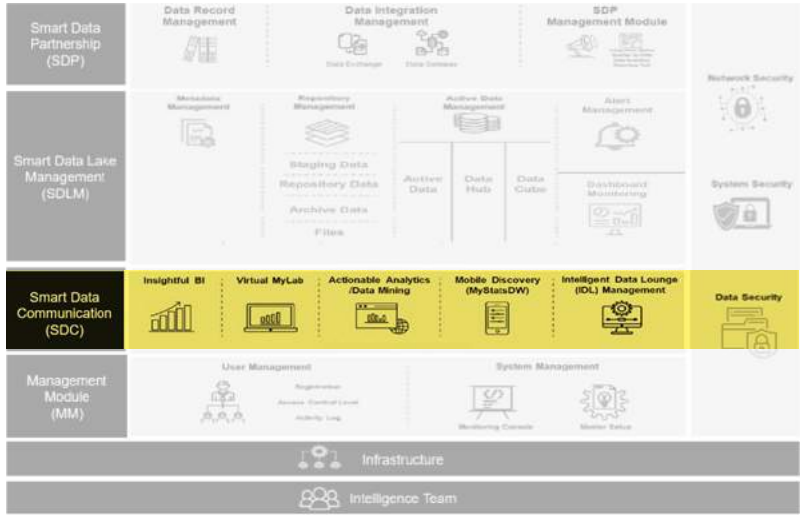
Actionable Analytic

1. Can be access by internal (DOSM environment) and external (IDL)
2. User can search, drag and drop the required data/survey/census/ variable
3. The chosen data can be put in the data cart
4. Statistical software tools will be provided for the user to do the analysis
5. Generated script can be stored by the user

IDL

1. Cyber cafe concept
2. User can register online for booking or walk in
3. No time limit for IDL usage (during office hours)

Key Points 1 Data cart 2 IDL



SMART DATA COMMUNICATION (SDC)

Various Avenues Users Can View And/Or Perform Analytics:

01. INSIGHTFUL BI

Through a web platform @ portal, information will be delivered through Business Intelligence (BI) form. Users (internally and externally), can view reports, predefined dashboards and download reports based on user-provided variables

02. VIRTUAL MYLAB

Selected data for statistical purposes are accessible through this avenue. It comes with an interactive table builder feature for registered users, to use with BI tools. Since this module is restricted to only internal DOSM users, Access Control Level (ACL) is applied, and unauthorised users must request permission.
Users can create reports, dashboards and transform data into interactive presentations. Internal users can view metadata and perform analysis

03. ACTIONABLE ANALYTICS (AA)

Selected micro data are included here for analytical/data mining for authorised users. ACL is implemented, to limit access to external parties. Authorised users can view metadata, perform analysis, to script and to code. **AA** is accessible via the Intelligent Data Lounge (IDL), subject to DOSM's micro data dissemination policy

06. STATSDW PORTAL

Main platform to access statistical data, reports and the dashboard. This portal contains four services; Actionable Analytics, Virtual MyLab, Insightful BI and Mobile Discovery and can be accessed by the public

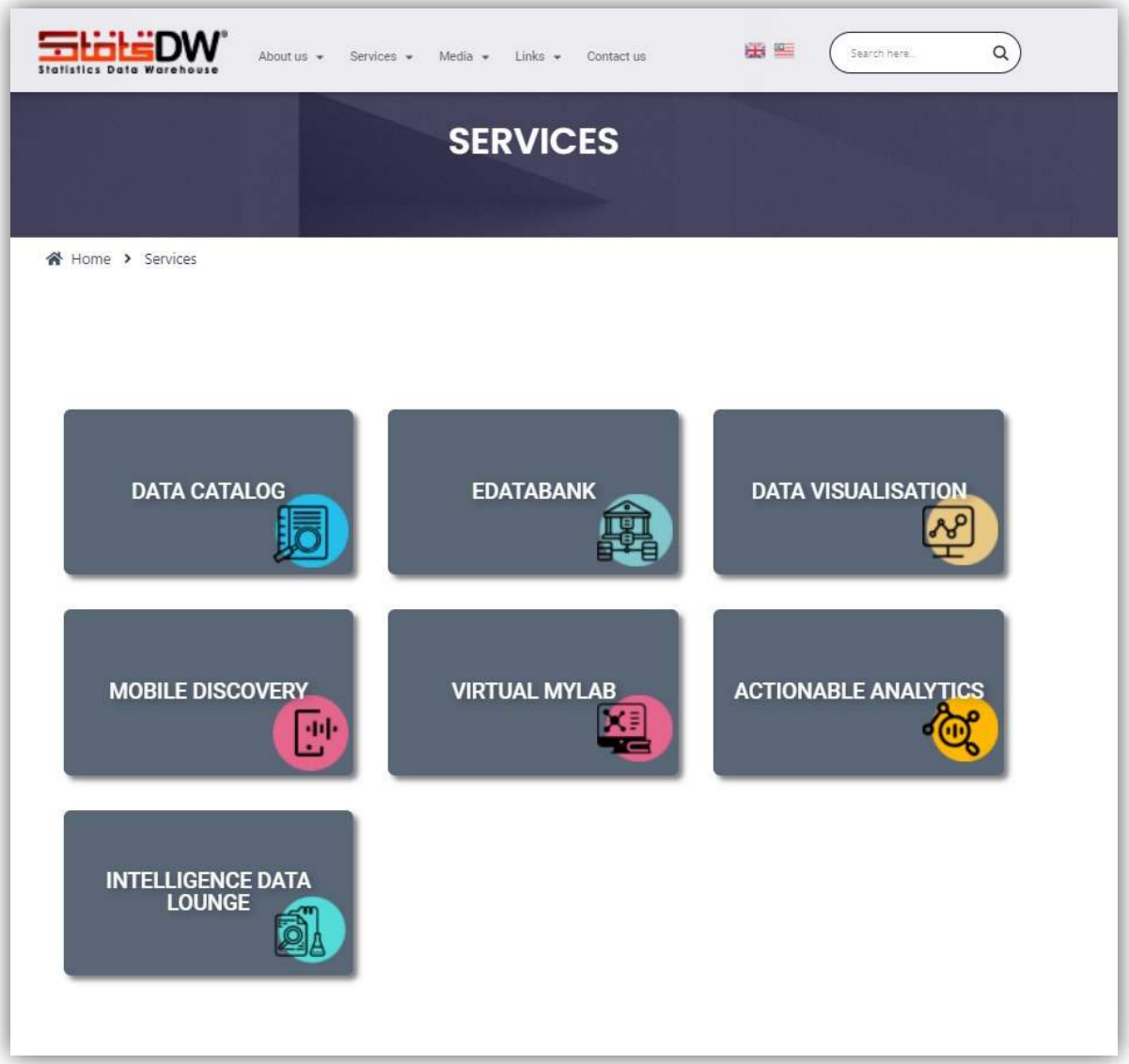
05. INTELLIGENT DATA LOUNGE (IDL) MANAGEMENT

Provides on-site access to Actionable Analytics, Virtual MyLab and Insightful BI to external users to experience

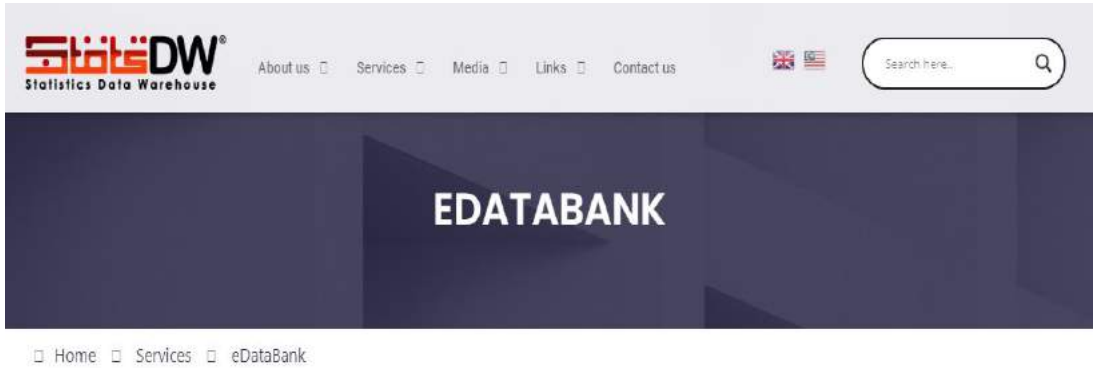
04. MOBILE DISCOVERY (MYSTATSDW)

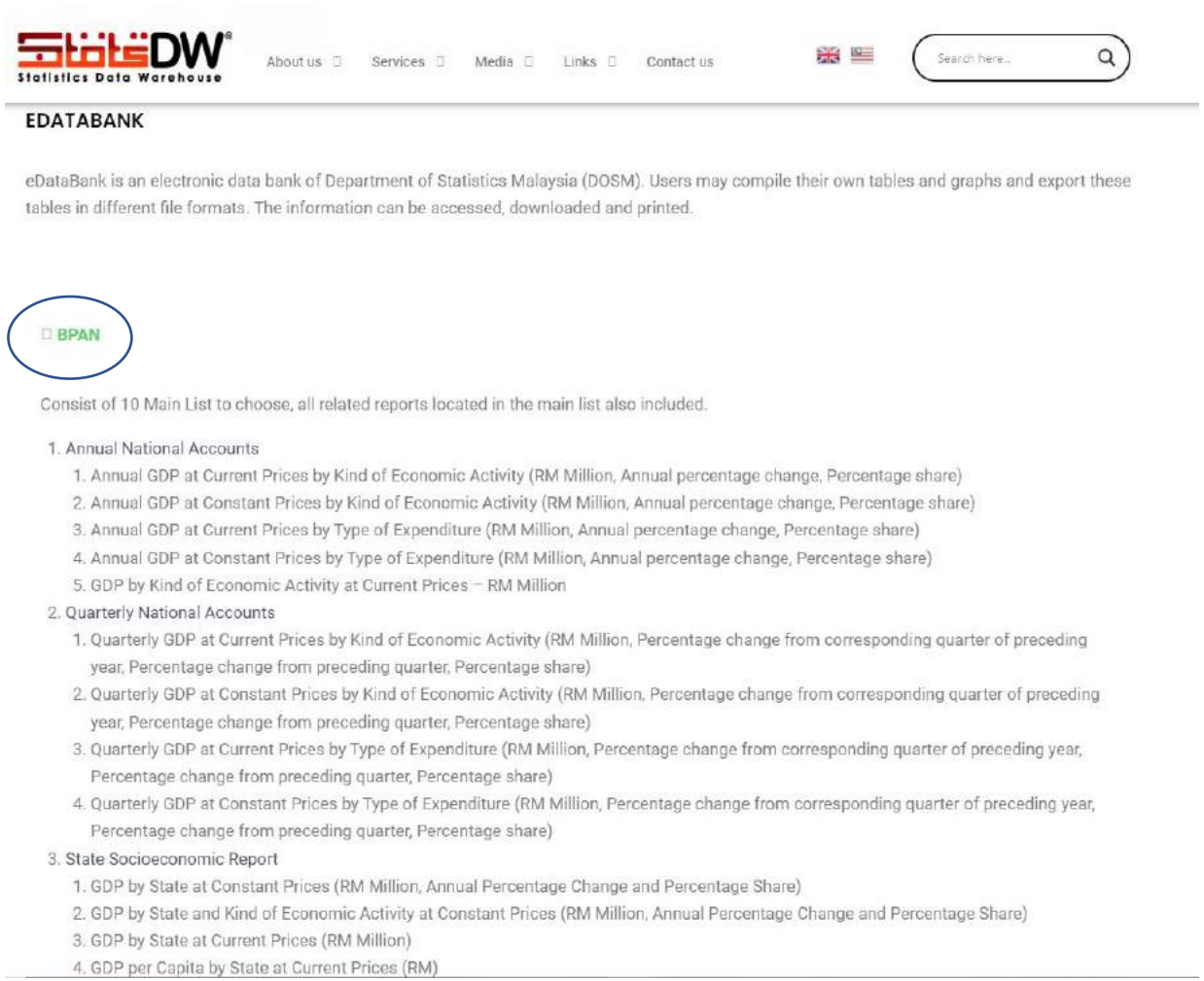
Statistics are delivered in an interactive graphical form that is tailored for mobile devices for both internal and external users. Visualisation outputs are similar to Insightful BI

Main page Insightful BI



INSIGHTFUL BI - eDATABANK





StatsDW Statistics Data Warehouse — About us | Services | Media | Links | Contact us | Search here...

EDATABANK – BPAN : ANNUAL NATIONAL ACCOUNTS

EDATABANK – BPAN : ANNUAL NATIONAL ACCOUNTS

eDataBank is an electronic data bank of Department of Statistics Malaysia (DOSM). Users may compile their own tables and graphs and export these tables in different file formats. The information can be accessed, downloaded and printed.

- eDataBank – BPAN : Annual National Accounts
- eDataBank – BPAN : Quarterly National Accounts
- eDataBank – BPAN : State Socioeconomic Report
- eDataBank – BPAN : Information & Communication Technology Satellite Account

BPAN : Annual National Accounts

Annual National Accounts | GDP by Kind of Economic Act...

- Annual GDP at Current Prices by Kind of Economic Activity ...
- Annual GDP at Constant Prices by Kind of Economic Activit...
- Annual GDP at Current Prices by Type of Expenditure (RM ...
- Annual GDP at Constant Prices by Type of Expenditure (R...
- GDP by Kind of Economic Activity at Current Prices - RM M...

1. Year :

- (All)
- 2020
- 2019
- 2018

2. Pick the Kind of Economic Activity :

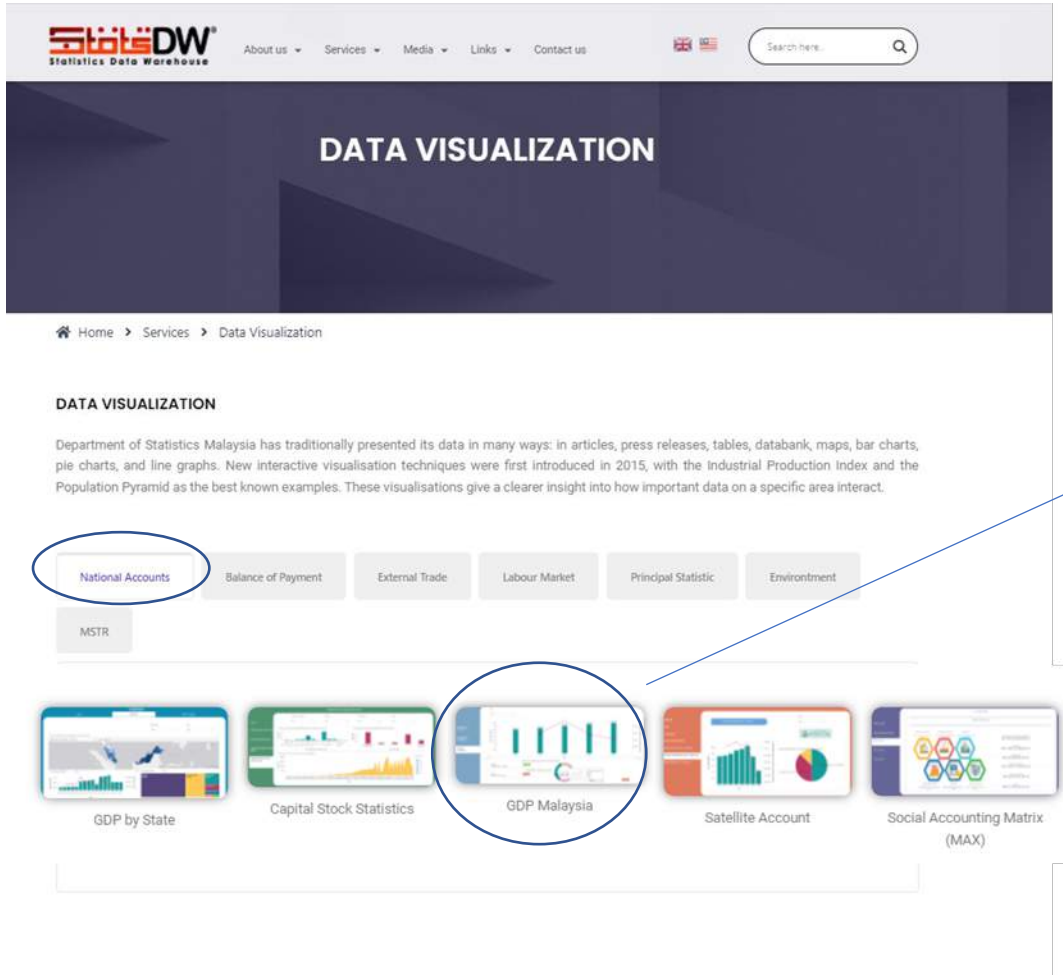
- (All)
- Agriculture
- Construction
- Manufacturing
- Mining and quarrying

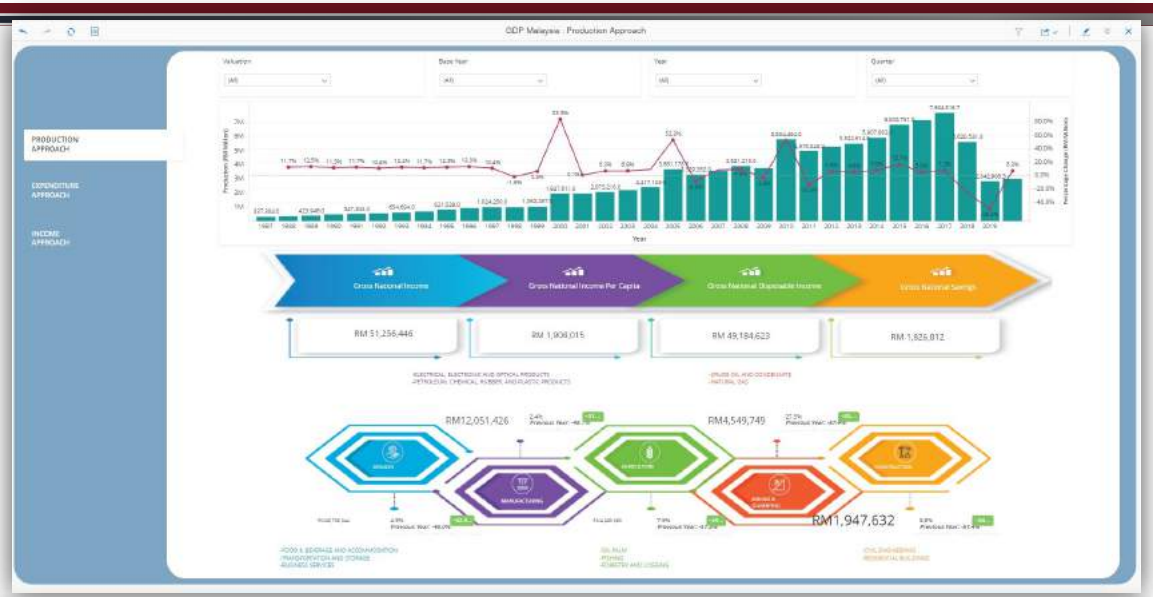
GDP by Kind of Economic Activity at Current Prices - RM Million

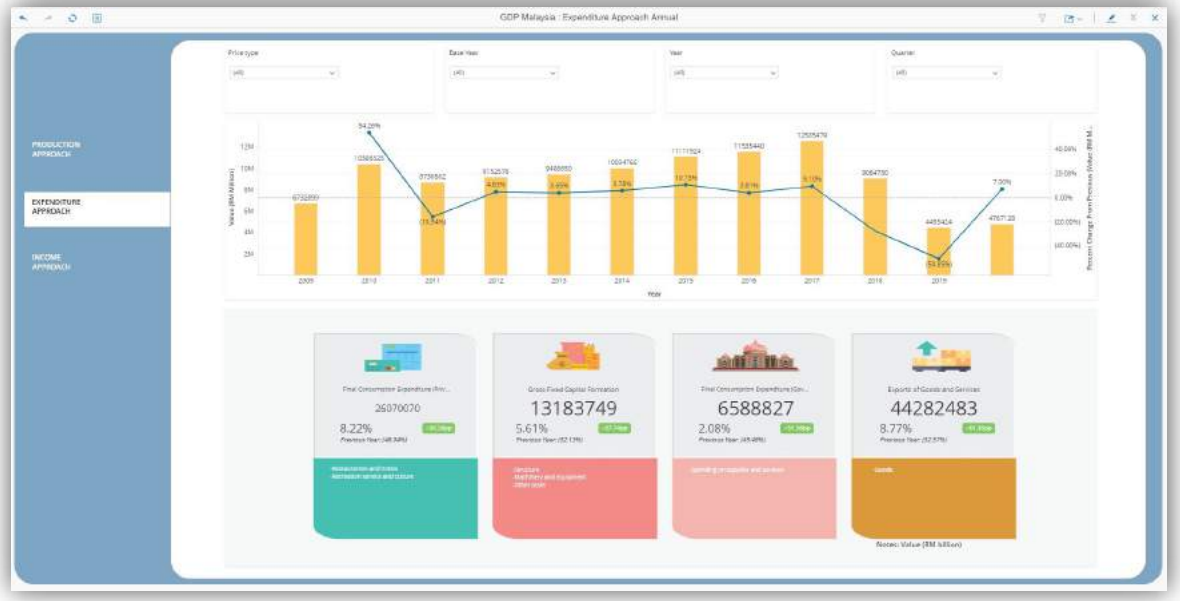
	RM Million								
Kind Of Economic Activity	2008	2012	2013	2014	2015	2016	2017	2019	2020
Agriculture	76,753	95,122	92,830	98,177	97,539	105,756	117,995	109,549	116,084
Construction	21,156	37,909	42,692	48,650	55,382	61,089	66,552	71,067	57,537
Manufacturing	189,105	224,730	232,659	253,087	262,379	272,396	299,797	323,862	315,608
Mining and quarrying	120,170	101,474	103,484	109,064	103,059	103,959	121,326	131,117	95,328
Services	355,233	501,830	536,186	585,518	643,883	689,969	747,360	860,497	816,292

INSIGHTFUL BI - DATA VISUALIZATION

Data Visualisation – National Accounts

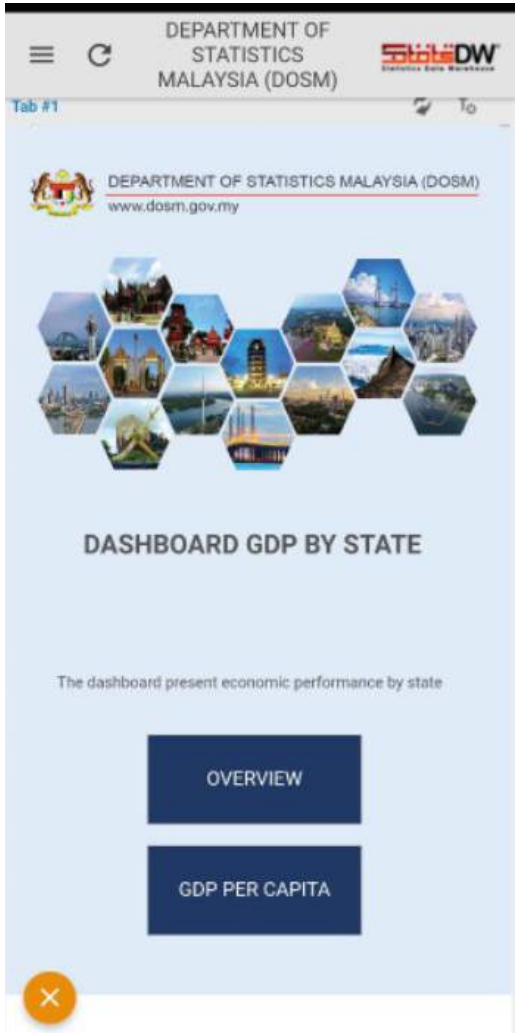


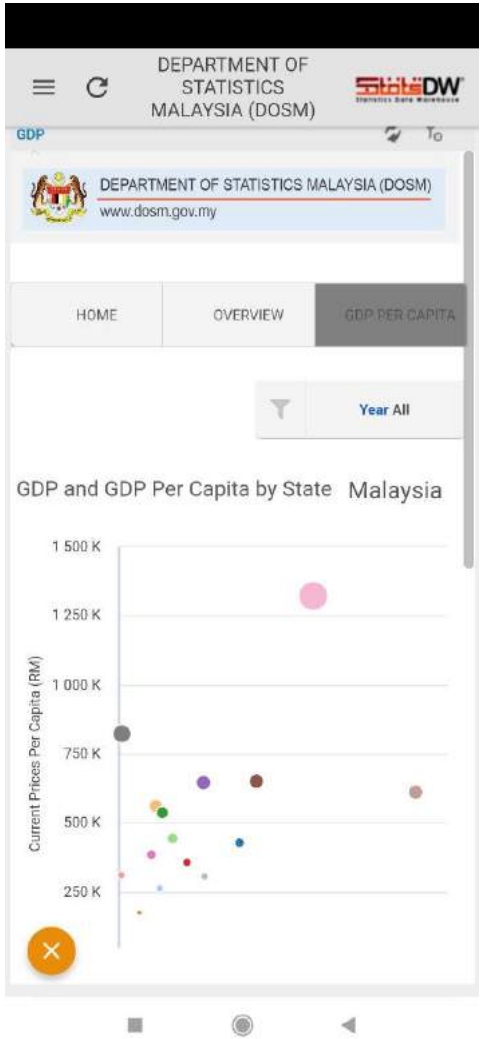




MOBILE DISCOVERY

GDP



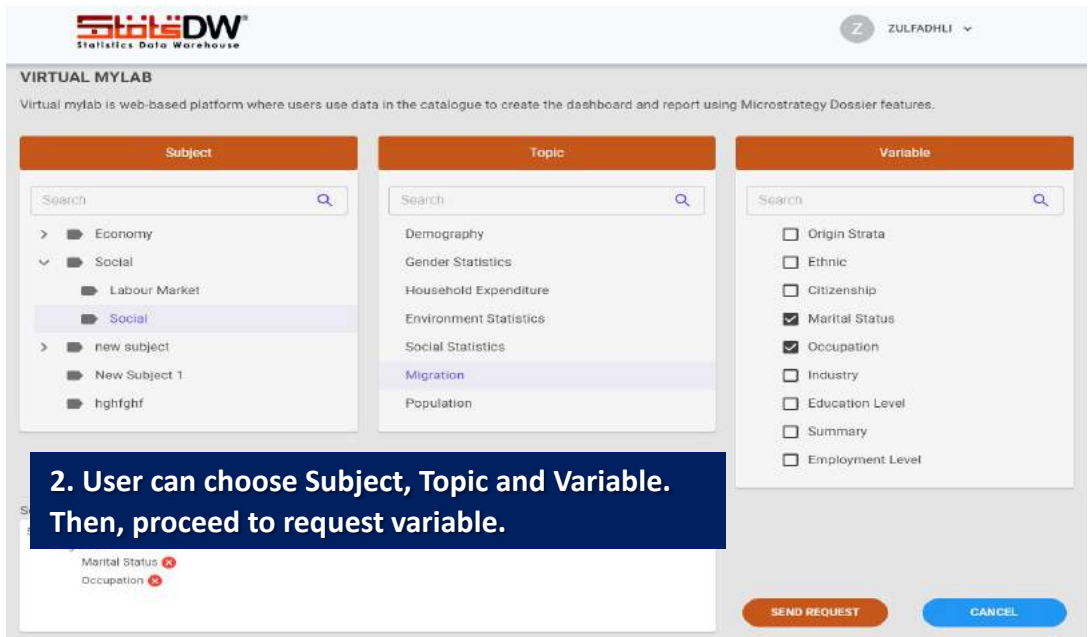


EXTERNAL TRADE

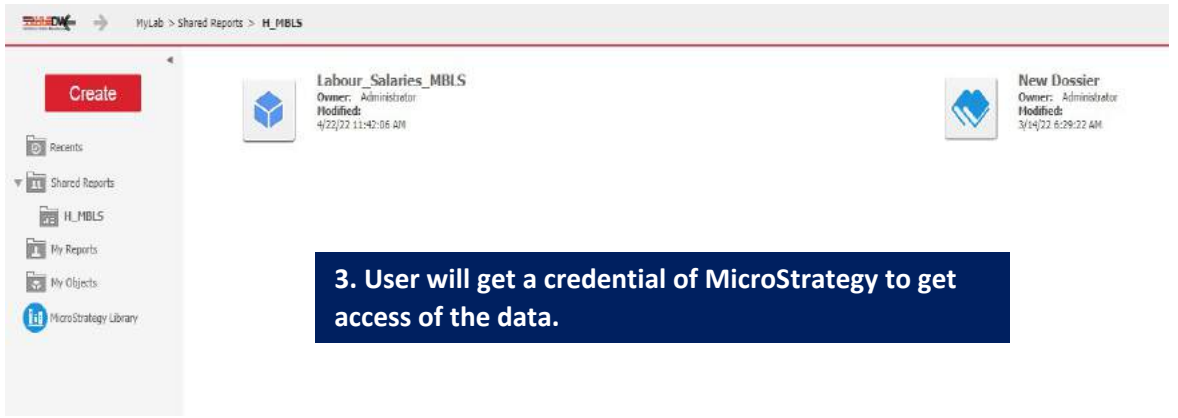
VIRTUAL MYLAB

Virtual MyLab Page

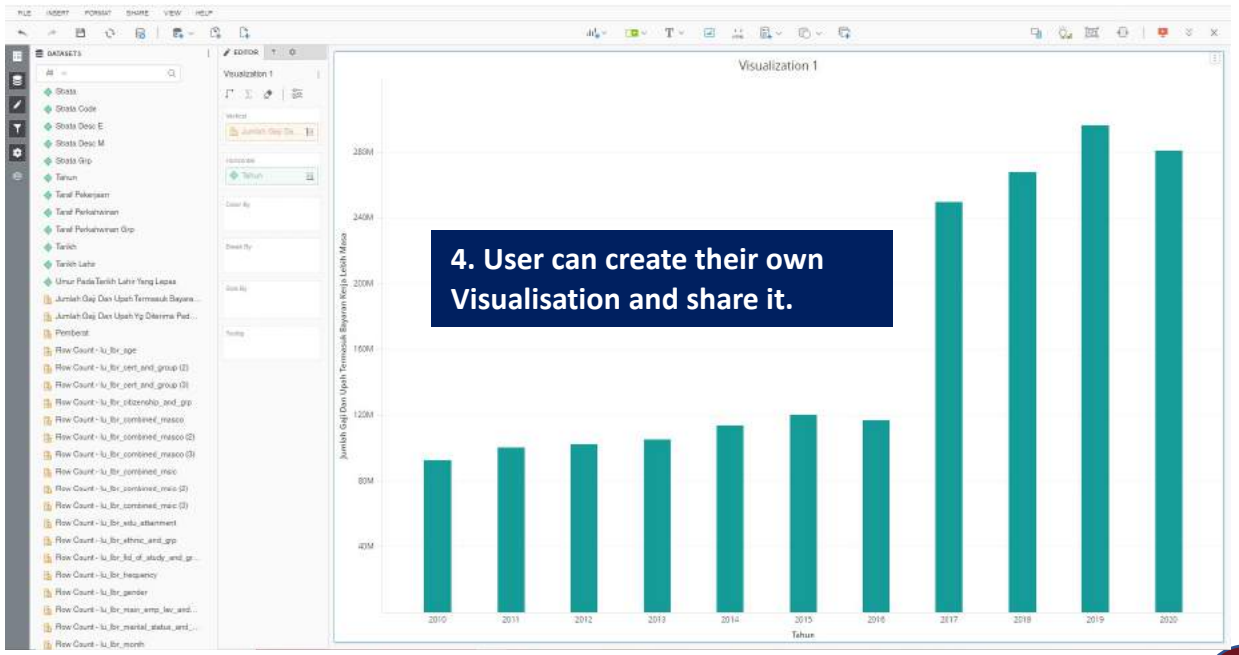
1. Click 'Go To Virtual My Lab'



2. User can choose Subject, Topic and Variable. Then, proceed to request variable.



3. User will get a credential of MicroStrategy to get access of the data.

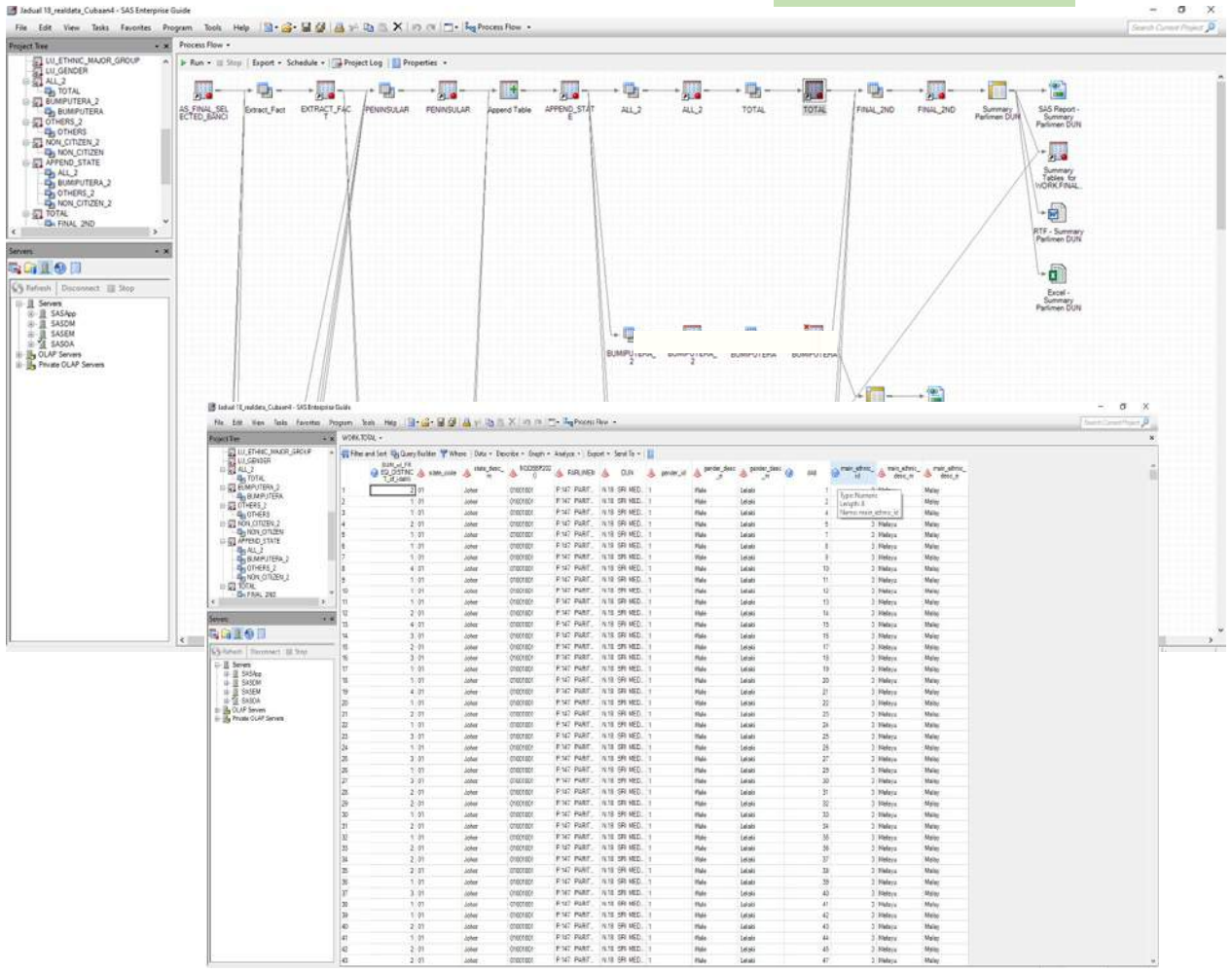


4. User can create their own Visualisation and share it.

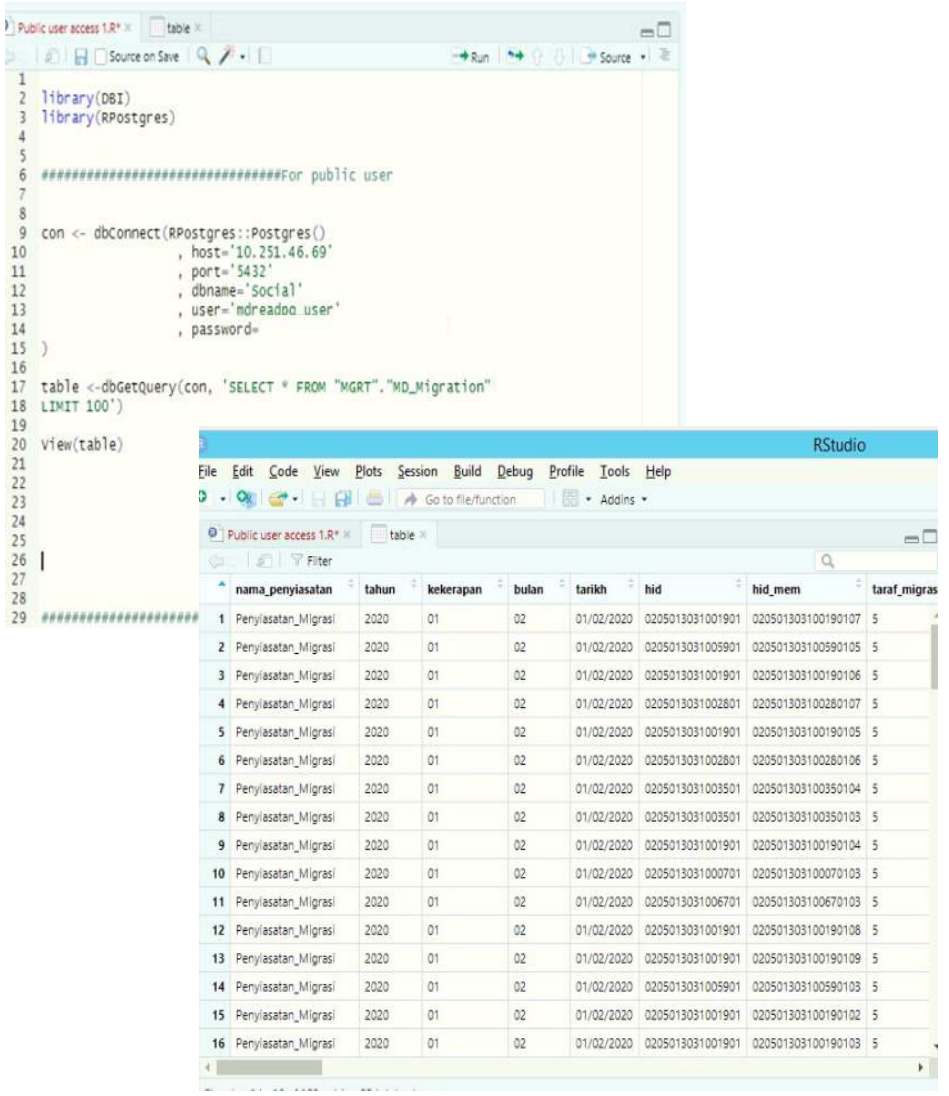
ACTIONABLE ANALYTICS

Actionable Analytics is a platform where users request data in the catalogue to conduct analysis using statistical tools, either R Studio or SAS EG

SAS EG



R Studio



MANAGEMENT MODULE (MM)

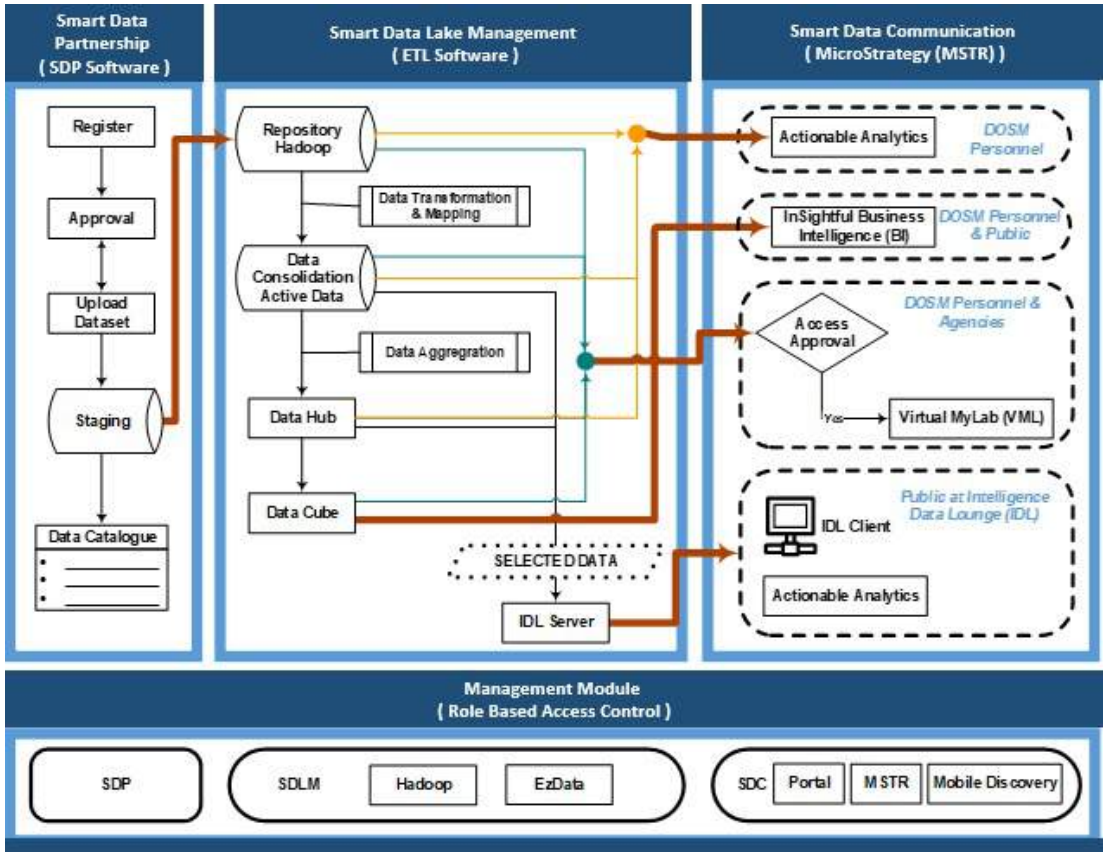
A platform to centralized administration of the StatsDW Phase II:

MM allows normal business process of StatsDW Phase II to be conducted fully on **system and user management**.

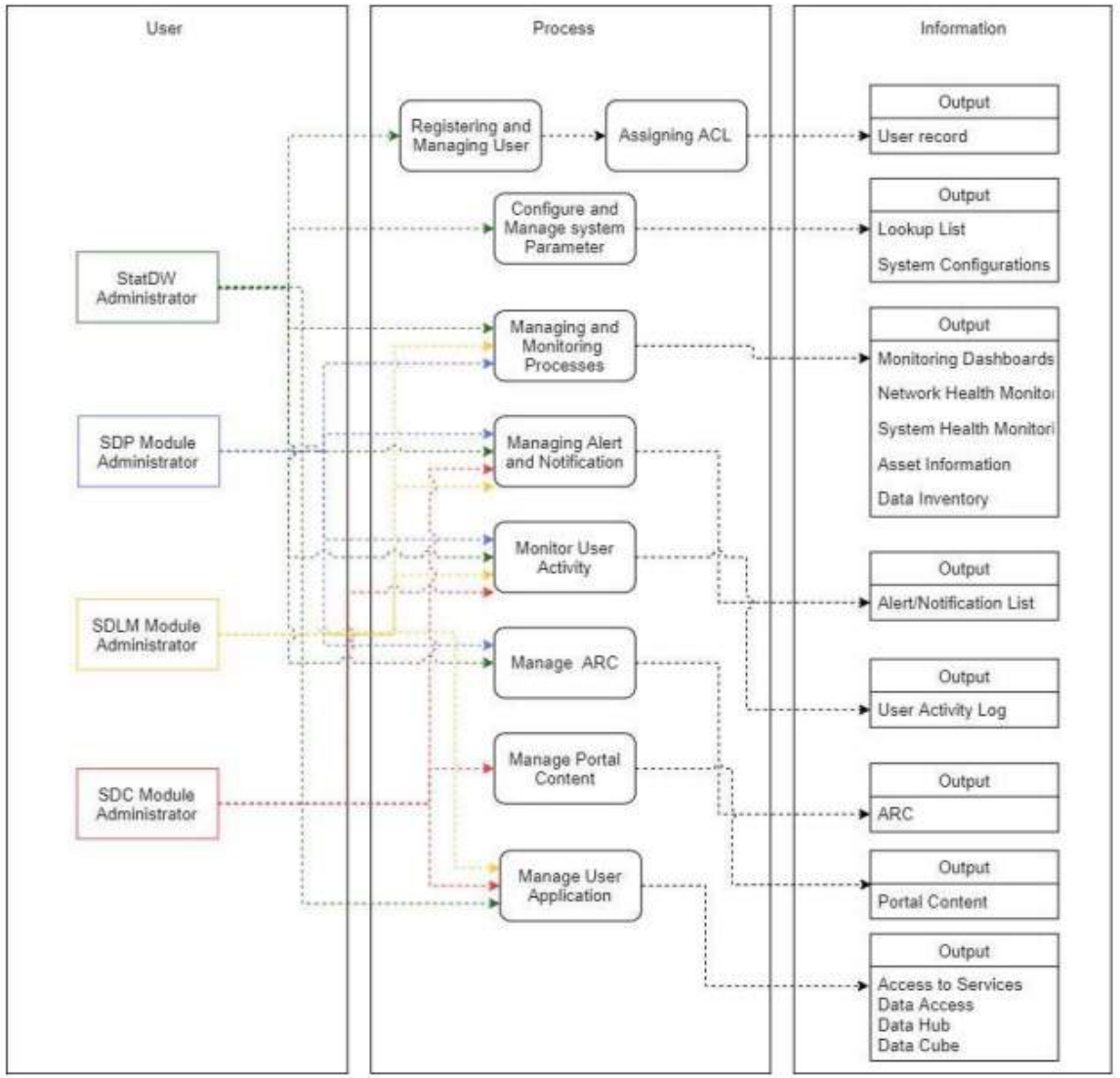
- **SDP** consists of the customized platform to monitor the whole process of storing the data recorded with alerts and notifications.

- **SDLM** consists of a Hadoop framework as a database involving massive amounts of data and computation that is scalable to store structured, unstructured data and files; and EzData as the data management and ETL tool.

- **SDC** consists of Portal, MSTR to perform analytics BI system which will use Active Data as a source for cubes; and Mobile Discovery as a mobile application delivering statistics in an interactive graphical form.

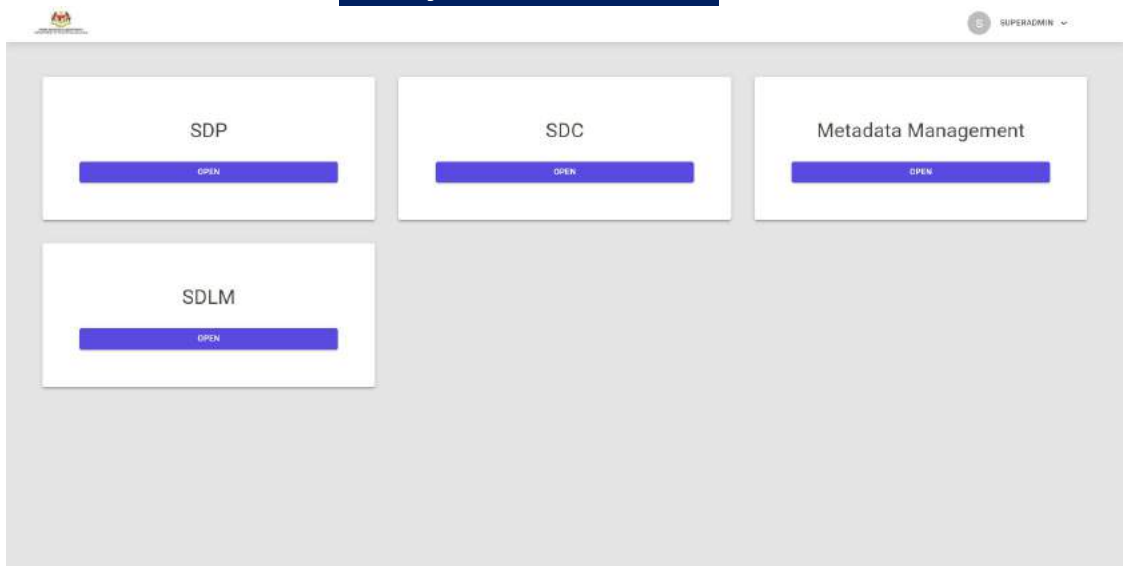


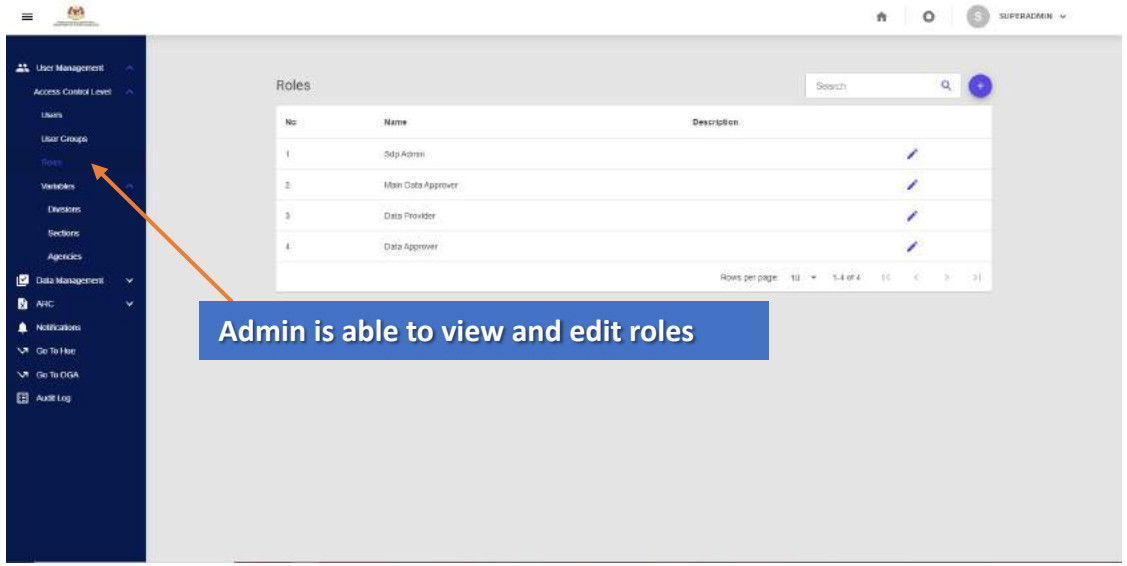
INFORMATION ARCHITECTURE



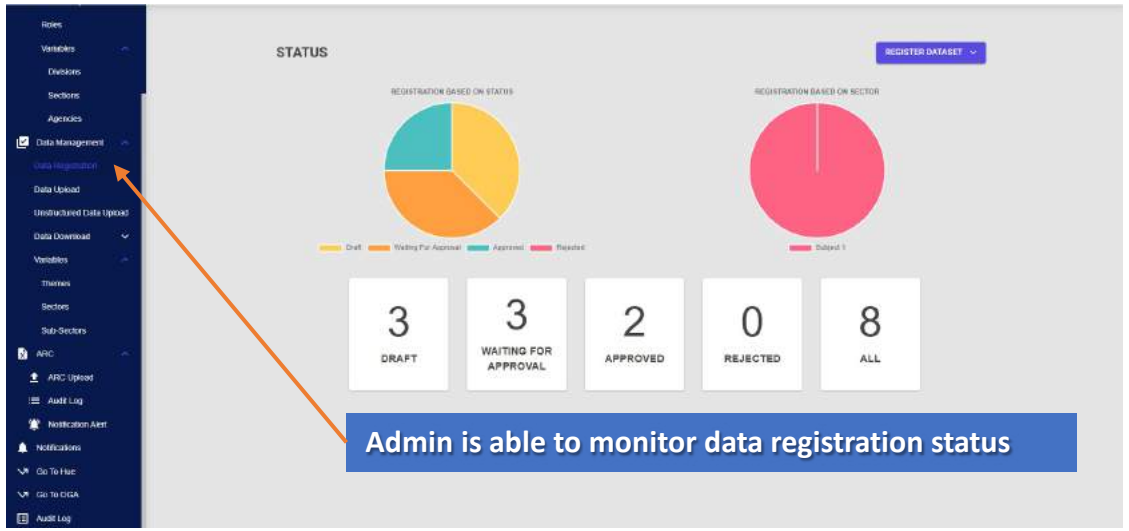
MANAGEMENT MODULE

System Admin

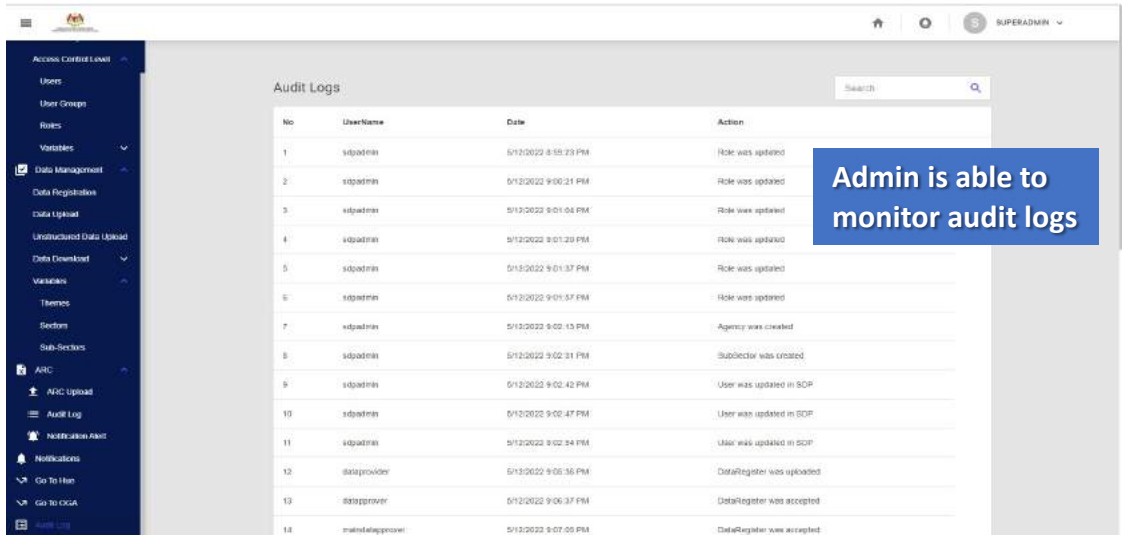




Admin is able to view and edit roles



Admin is able to monitor data registration status



Admin is able to monitor audit logs

CHALLENGES

StatsMalaysia
www.DOSM.gov.my

#KELUARGAMALAYSIA

CHALLENGES

Standardisation
- Data in various format
- Changes of data structure

Lack of Quality
- Data inconsistency

Legal constraints
- Data sharing on individual data

Technology capabilities
- Cost is on the rise
- Technology capabilities

INNOVATION IMPACT

StatsMalaysia
www.DOSM.gov.my

#KELUARGAMALAYSIA

INNOVATION IMPACT

- Consolidate all micro and aggregate data in an Enterprise Data Warehouse (EDW)
- Access to various types of statistics
- Maintaining reliability of data
- Integrate data from various surveys and compilation
- Enable fast and easy data analysis and reporting
- Enhance the data quality
- Successfully improved access to data
- Data and information security are preserved

INNOVATION IMPACT

1. Displaying data of quality, integrity and assurance of its validity
2. Platform of reference for DOSM
3. Simplify data generation and analysis
4. Preserving the confidentiality of individual data
5. Publish interactive table and visualisation
6. Save time and cost of data search

BENEFITS AND WAY FORWARD OF StatsDW

StatsMalaysia
www.DOSM.gov.my

#KELUARGAMALAYSIA

BENEFITS OF StatsDW

BENEFIT

- Provide centralized data storage centre
- Maintaining data reliability and quality
- Facilitate wider access and analysis
- Guarantees data security
- Integrate systems and data sharing between agencies
- Improved data integration, management and security
- Improved metadata management
- Facilitates development of new analytical outputs and products
- Facilitates National Statistical Service objectives

StatsDW®
Statistics Data Warehouse

BENEFITS OF StatsDW

Data Driven Society

IMPLEMENTATION OF PUBLIC SECTOR OPEN DATA (Pekeliling AM Bil. 1/2015)

- Creation of new industry or innovation through the digital economy.
- Transparency in service delivery.
- Better service delivery

StatsDW

RESEARCHER

FINDINGS

HIGH RANK MEETING

POLICY MAKER

PUBLIC BENEFIT

WAY FORWARD

statsDW®
Statistics Data Warehouse

- Improvement in bottom –up Data Mart to maximize Subject Matter Divisions (in DOSM) the capability and capacity to create own Data Mart
- Data Governance and Master Data Repository: a central repository for other government agencies in Malaysia
- Data Homogenization: Similarity of data format, classification and metadata among agencies
- Extension of Statistical Data and Metadata Exchange (SDMX): Moving towards the new standard of SDMX 3.0

THANK YOU

StatsMalaysia

www.DOSM.gov.my

#KELUARGAMALAYSIA

2022

2012-2022

20 OKT

2016-2030