Data Integration

Lesson Learned and Best Practices

from experience in the Australian Bureau Statistics

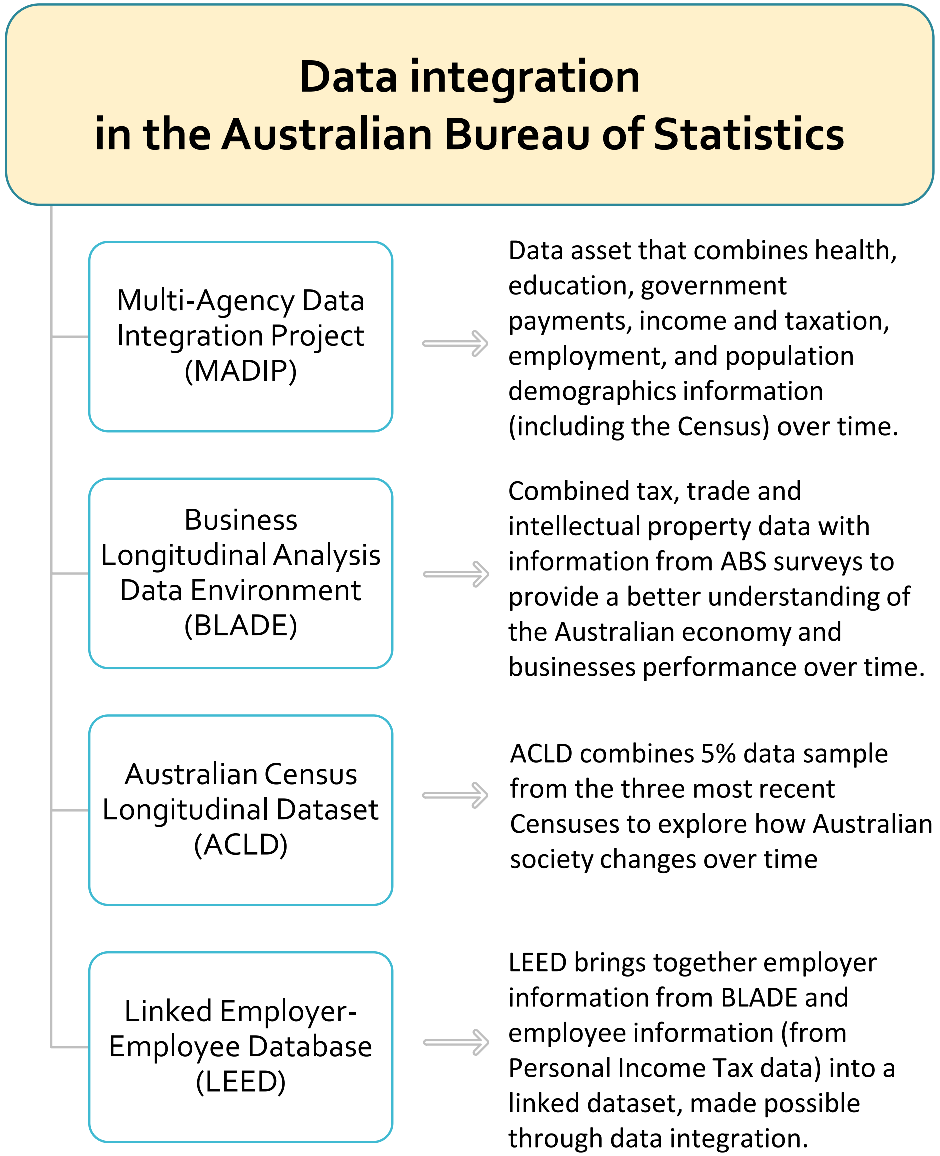
Paul Nicholls - ICT Consultant

3 November 2021

About data integration

Data integration is a common industry term referring to the requirement to combine data from multiple separate business systems into a single unified view, often called a **single view of the truth**. This unified view is typically stored in a central data repository known as a **data warehouse**.

For example, citizen data integration involves the extraction of information about each individual citizen from disparate business systems such as health, education, and finance, which is then combined into a **single view of the citizen** to be used for citizen service, reporting and analysis.



Data integration in the Australian Bureau of Statistics

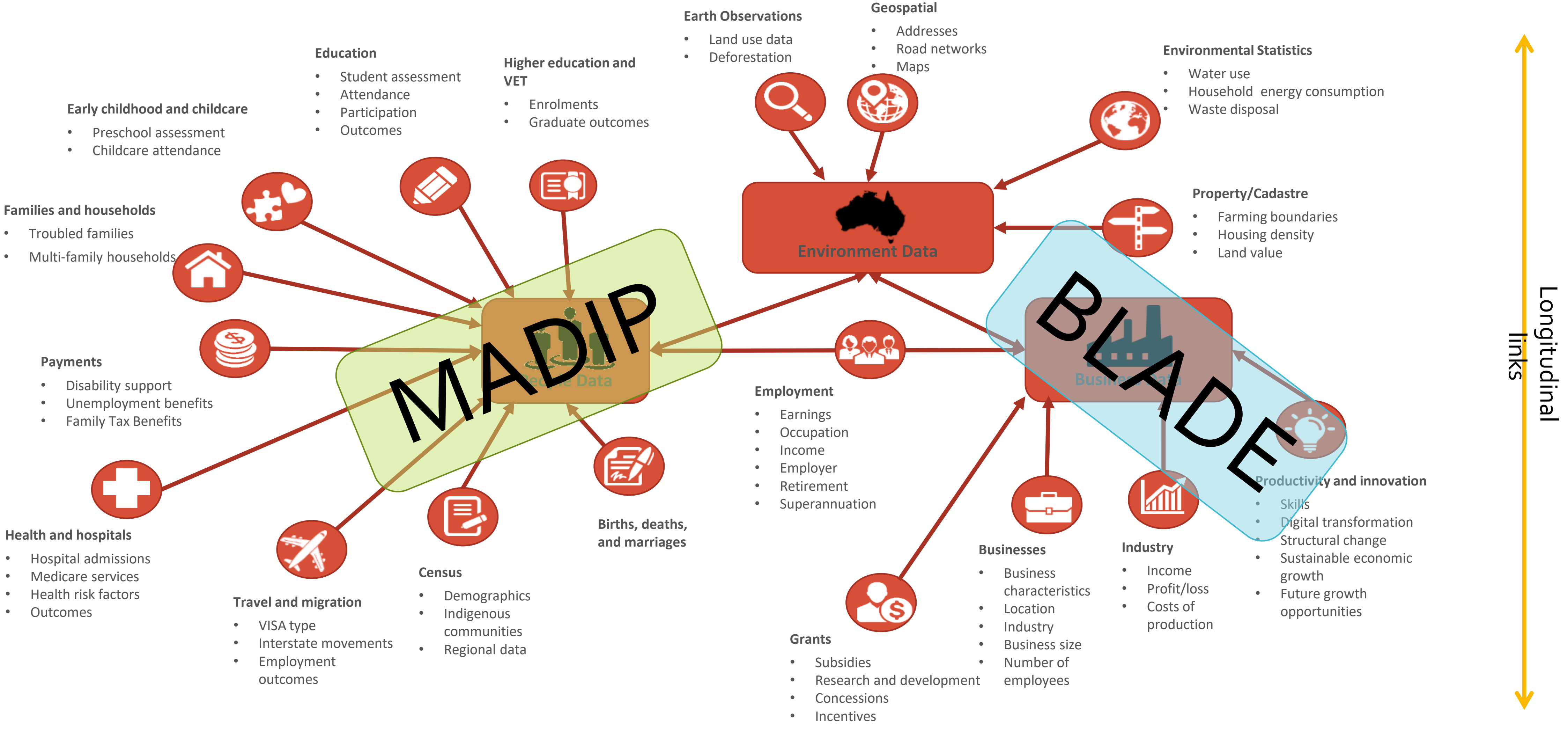
Multi-Agency Data Integration Project (MADIP) → Data asset that combines health, education, government payments, income and taxation, employment, and population demographics information (including the Census) over time.

Business Longitudinal Analysis Data Environment (BLADE) → Combined tax, trade and intellectual property data with information from ABS surveys to provide a better understanding of the Australian economy and businesses performance over time.

Australian Census Longitudinal Dataset (ACLD) → ACLD combines 5% data sample from the three most recent Censuses to explore how Australian society changes over time

Linked Employer-Employee Database (LEED) → LEED brings together employer information from BLADE and employee information (from Personal Income Tax data) into a linked dataset, made possible through data integration.

MADIP & BLADE data integration



Early childhood and childcare

- Preschool assessment
- Childcare attendance

Education

- Student assessment
- Attendance
- Participation
- Outcomes

Higher education and VET

- Enrolments
- Graduate outcomes

Earth Observations

- Land use data
- Deforestation

Geospatial

- Addresses
- Road networks
- Maps

Environmental Statistics

- Water use
- Household energy consumption
- Waste disposal

Property/Cadastre

- Farming boundaries
- Housing density
- Land value

Families and households

- Troubled families
- Multi-family households

Payments

- Disability support
- Unemployment benefits
- Family Tax Benefits

Employment

- Earnings
- Occupation
- Income
- Employer
- Retirement
- Superannuation

Health and hospitals

- Hospital admissions
- Medicare services
- Health risk factors
- Outcomes

Travel and migration

- VISA type
- Interstate movements
- Employment outcomes

Census

- Demographics
- Indigenous communities
- Regional data

Births, deaths, and marriages

Grants

- Subsidies
- Research and development
- Concessions
- Incentives

Businesses

- Business characteristics
- Location
- Industry
- Business size
- Number of employees

Industry

- Income
- Profit/loss
- Costs of production

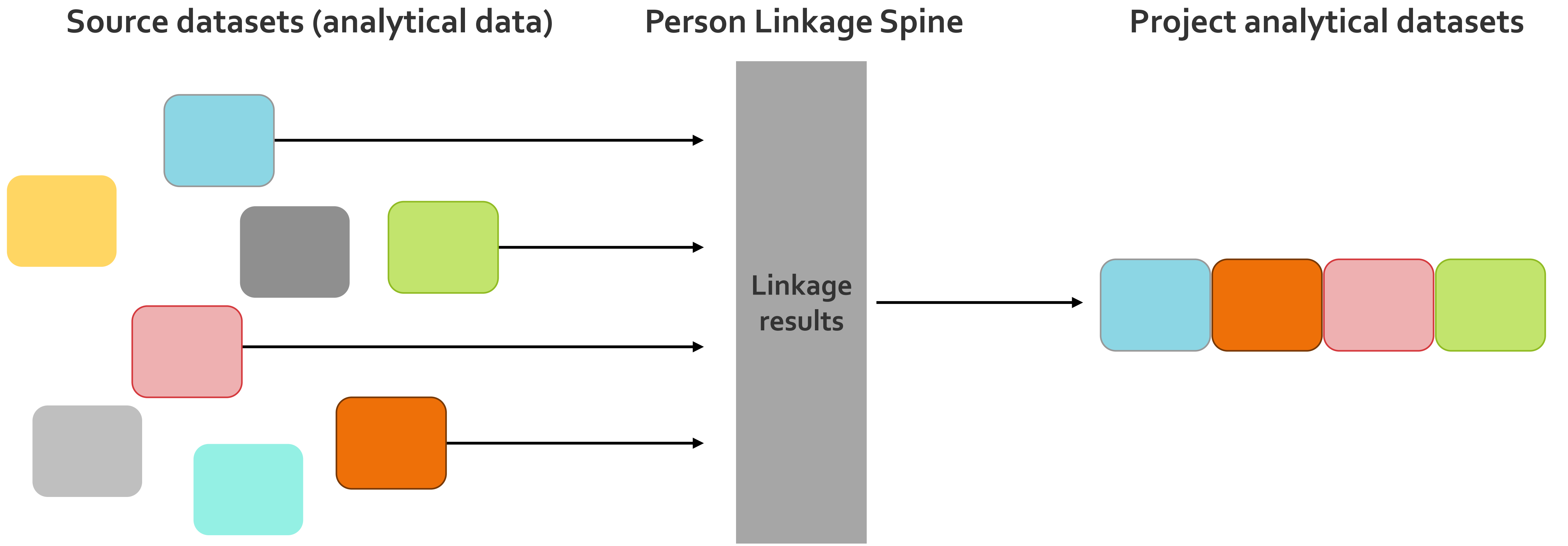
Productivity and innovation

- Skills
- Digital transformation
- Structural change
- Sustainable economic growth
- Future growth opportunities

Source: ABS presentation, 4 February 2021

Mechanism for data linkage

Separate datasets are able to be brought together for specific projects


Source datasets (analytical data)
Person Linkage Spine
Project analytical datasets
Linkage results

SEPARATE STORAGE of analytical data

Mechanism for **linkage**

Assembly of analytical data

Access to analytical data in ABS DataLab

Source: ABS presentation, 4 February 2021

Keeping integrated data safe

LEGISLATIVE PROTECTION

A set of regulations are in place to protect integrated data.

- Privacy Act 1988
- Australian Privacy Principles
- Census and Statistics Act 1905

PRIVACY PROTECTION

The Australian Government Agencies Privacy Code sets out specific requirements and key practical steps that requires agencies to move toward a best practice approach to privacy governance to help build a consistent, high standard of personal information management across all Australian Government agencies.

POLICIES AND STANDARDS

All data integration is conducted in line with the seven High-Level Principles for Data Integration:

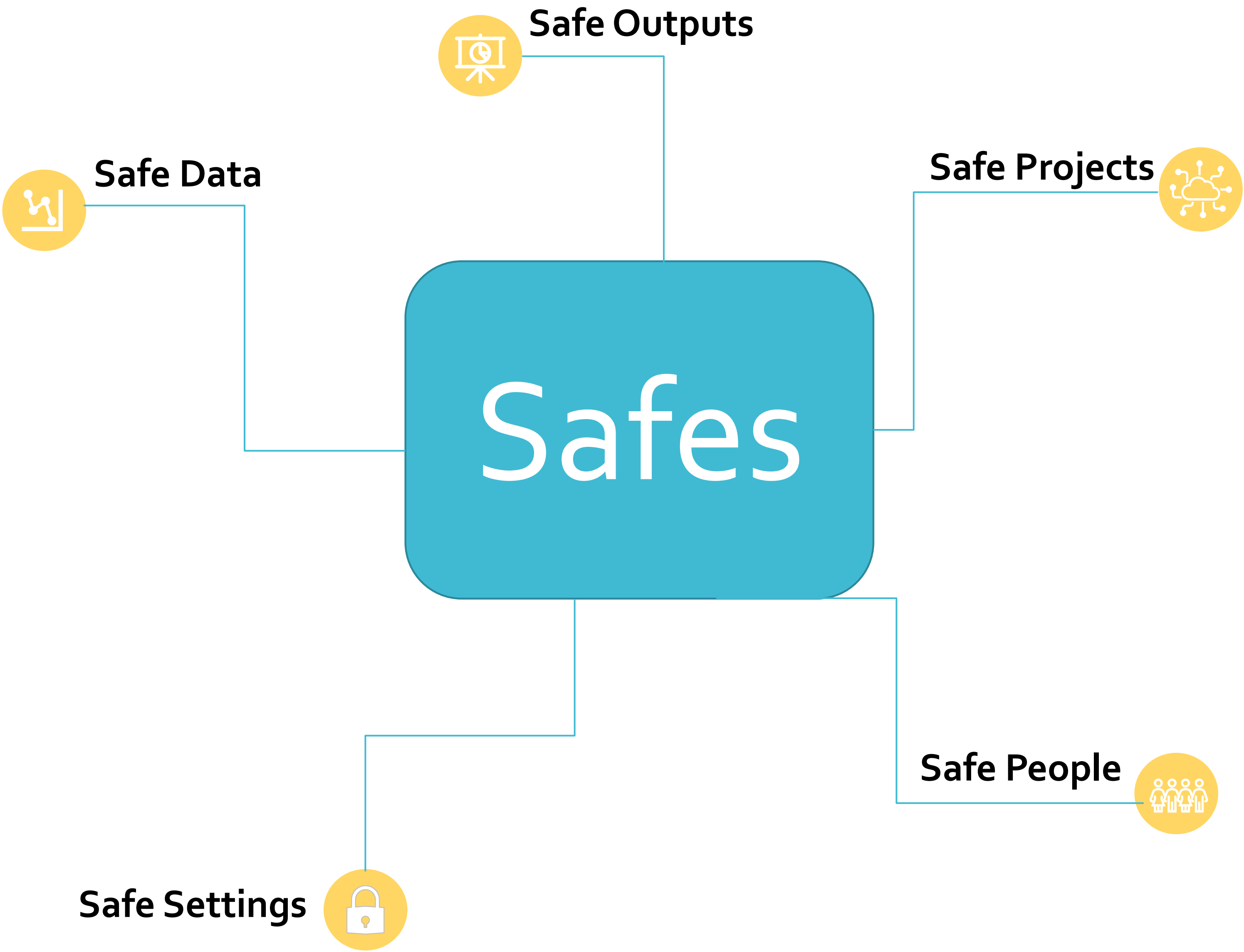
1. Strategic resource
2. Custodian's accountability
3. Integrator's accountability
4. Public benefit
5. Statistical and research purposes
6. Preserving privacy and confidentiality
7. Transparency

SAFE DATA HANDLING

The Separation Principle is applied when we undertake data integration activities. This means personal identifiers are stored separately from other information, and no one can view both personal identifiers and analytical information at the same time.

Fives Safes Principle in Data Integration Access & Release

- The ABS uses the Five Safes Framework to ensure safe and secure access to the integrated data sets (MADIP, BLADE, ACLD, LEED)

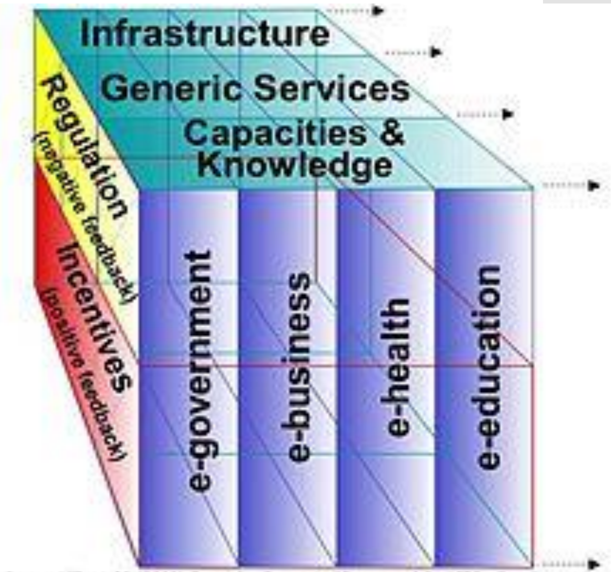


Safe People

- **Don’t assume knowledge, responsibility or accountability**
  - Formal training and recorded acknowledgement
  - Foster a culture of innovation
  - Partner up with experts, ensure skills transfer occurs
- **Document clear role definitions**
  - For data integration (DI), not for business as usual
  - Regularly review roles and responsibilities against DI business processes
  - Underpins the DI decision making process – appropriate delegation
  - Critical to overall DI management frameworks
  - Better understand long term costs of doing DI business

Safe Settings

- Separate data integration environment from business as usual (BAU)
- Cloud first
- Information Technology as a service (*aaS)
- Governance – Don’t forget to allow for innovation and adaptability
- Authentication and Authorisation (role-based)
- Invest in Artificial Intelligence and Machine Learning
- Logging and Monitoring Services are critical



Source: Hilbert, M. (2012). Towards a Conceptual Framework for ICT for Development: Lessons learned from the Latin American “Cube Framework”. *Info. Techn. & Internat. Dev ITID*, 8, 4, pp. 243–259; http://itidjournal.org/itid/article/viewFile/967/408

Safe Data

Data Lakes and Data Pipelines

- Catalogues of data and services
- Manage risk

Metadata

- Standards and Governance
- Just enough (fit for purpose)
- Loosely coupled to the data

Quality

- Improve quality at the source
- Keep multiple versions and attach quality statements

Data Integration Do’s

- **Hire specialists**

We understand that many online enterprises have an online-only presence, with no brick-and-mortar facility. What we also understand is that data integration often becomes a tougher task due to the geographical distances between employees. An easy way to overcome this is to keep investing time in upgrading the skills of your existing employees as well as hiring specialists who get you the best results in the least amount of time.

- **Migrate to multi-cloud architectures**

Data warehouses are not going to disappear. But with the introduction of cloud-based systems, they certainly are taking on a new look, with fresh facilities and modern capabilities like that of being serverless. Data lakes are becoming a new legacy technology and can provide simpler and easier-to-manage means of extracting business intelligence that is both usable and cost-effective. It means this can definitely count as a good time for moving to cloud architecture.

- **Simpler reporting**

Hiring specialists and mitigating to a simpler, self-service structure is one thing, but it is equally important that their reporting is understood even by those who are not well-versed with the process. This can be easily accomplished by data visualization, or the use of infographics. While one may argue that designing these images is also difficult, there are plenty of infographics available online that make this task effortless.

Data Integration Don’ts

- **Stick to manual data integration**

The increase in volume – as well as sources – of data has led to a proportional rise in the need to process it, which given the magnitude is possible only through automation. The size of this information bracket is just going to grow, and the smart thing to do would be to move to automated data integration at the earliest.

- **Independent but isolated working**

What most companies fail to realize is that data integration and data silos are interconnected and the creation of more silos is going to affect the speed and ease of its unification throughout. According to a various studies 82% of organisations admit to having high to moderate quantities of these data pits and how to collect them all in a single, compiled dataset still evades most. A collaborative approach is in the works for many, which allows industry units the freedom to operate while keeping an eye on the all-inclusive business systems.

- **Underutilization of available data**

In a survey by Invesp, 87% of organisations – both online and offline – reported that their available data was not being utilized to its fullest potential. Seeing as online businesses are sitting on data worth millions and billions through their continuous operations, it is of absolute necessity to use this information wisely and optimally within due time.

Important Questions to ask when Designing Data Centers

Basic Infrastructure

- How big is the data center ? How big is big enough?
- Should we build one or build many?
  - If one, how big the data center?
  - If many, how many data centers we want to build and are they adequate enough to cater for all data needs?
- How much data will we obtain for the data center?
- What availability do we need? 24/7?
- How much energy will it need? From what resources?
- What about Green and Emissions from Data Center(s)?
- How long should the data center last?
- How to secure high security in the Data Center(s)?

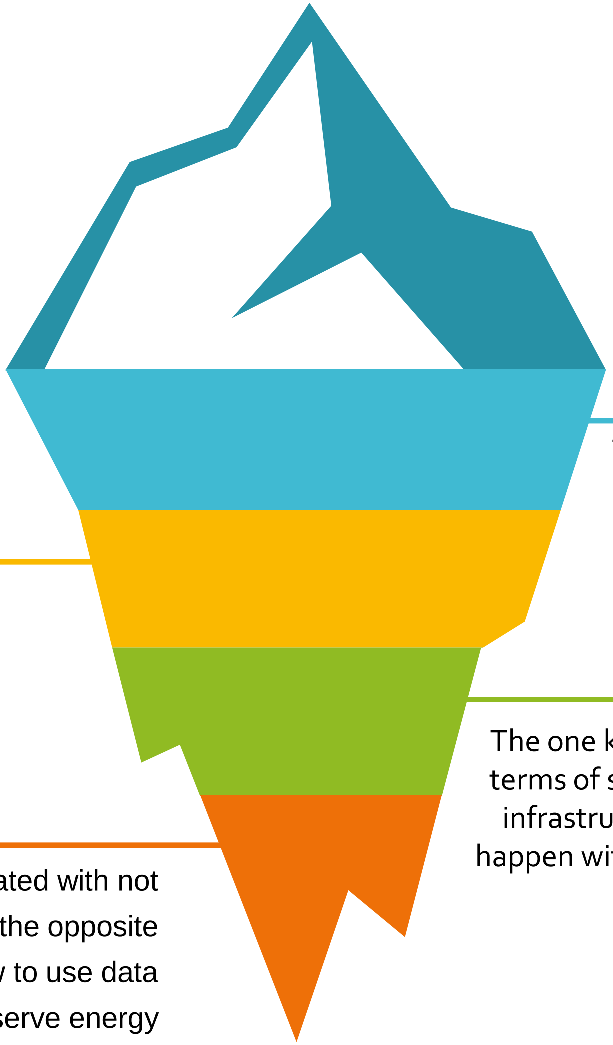
Data Trends & Applications

- What are the newest design trends today?
- What are the dominant trends in data center design today
- What are the benefits and tradeoffs when using (or ignoring) them?
- Are all applications created equal?
- Do all our applications need these levels of support? Can we build an environment to support different service and technology levels, based on the requirements of the applications?

Human Resources & Digital Capabilities

- Who's going to operate? Do we have the right skills and adequate skills available in the country? If not, how to fulfill them?
- Who will build it – and what should we ask up front of them?

Successful Data Center Iceberg



Optimization

The faster a Data Centers expands, the quicker it grows in terms of size and complexity. This requires significantly quick deployment time. A Data Center needs to be updated regularly to support the growing needs of a business. Optimization is very important

Infrastructure

Infrastructure to optimize network performance & stability of Data Center

Sizing

The one key factor in assessing the efficiency of a Data Center in terms of size was to see how fast it would grow. Accordingly, the infrastructure supporting it would also grow. The expansion will happen without any forethought – is detrimental both in terms of capital as well as energy.

Sustainability

Sustainability is not a singular concept. While it is often associated with not destroying natural resources, it can also be tailored to achieve the opposite effect – to conserve them. How green is the data center and how to use data center to conserve energy

Some thoughts for GOI

- User requirements have become diverse and changeable. Therefore, a full-stack and full-scenario one-stop solution is required.
- Data is indispensable, it can be used and reused, and data centers are like the engine that unleashes the energy contained in the data.
- Indonesia needs to ask critical questions about success factors
- Cybersecurity is important to ensure high security of data
- Indonesia needs to lay out all the successful goals of Data Center to form the strategic implementation plan and program implementation
- Securing budget is critical – source of funds requirement
- Indonesia needs to quickly and economically obtain an environmentally-friendly yet powerful digital engine designed for the digital era in its data center strategy. In this way, the government can focus more on service innovation and citizen requirements fulfillment.

Thank you