CHULALONGKORN
BUSINESS SCHOOL
FLAGSHIP FOR LIFE

DATA LAKE

30 October 2016

Thaveewat Khanan

2603615 Organization Data Management

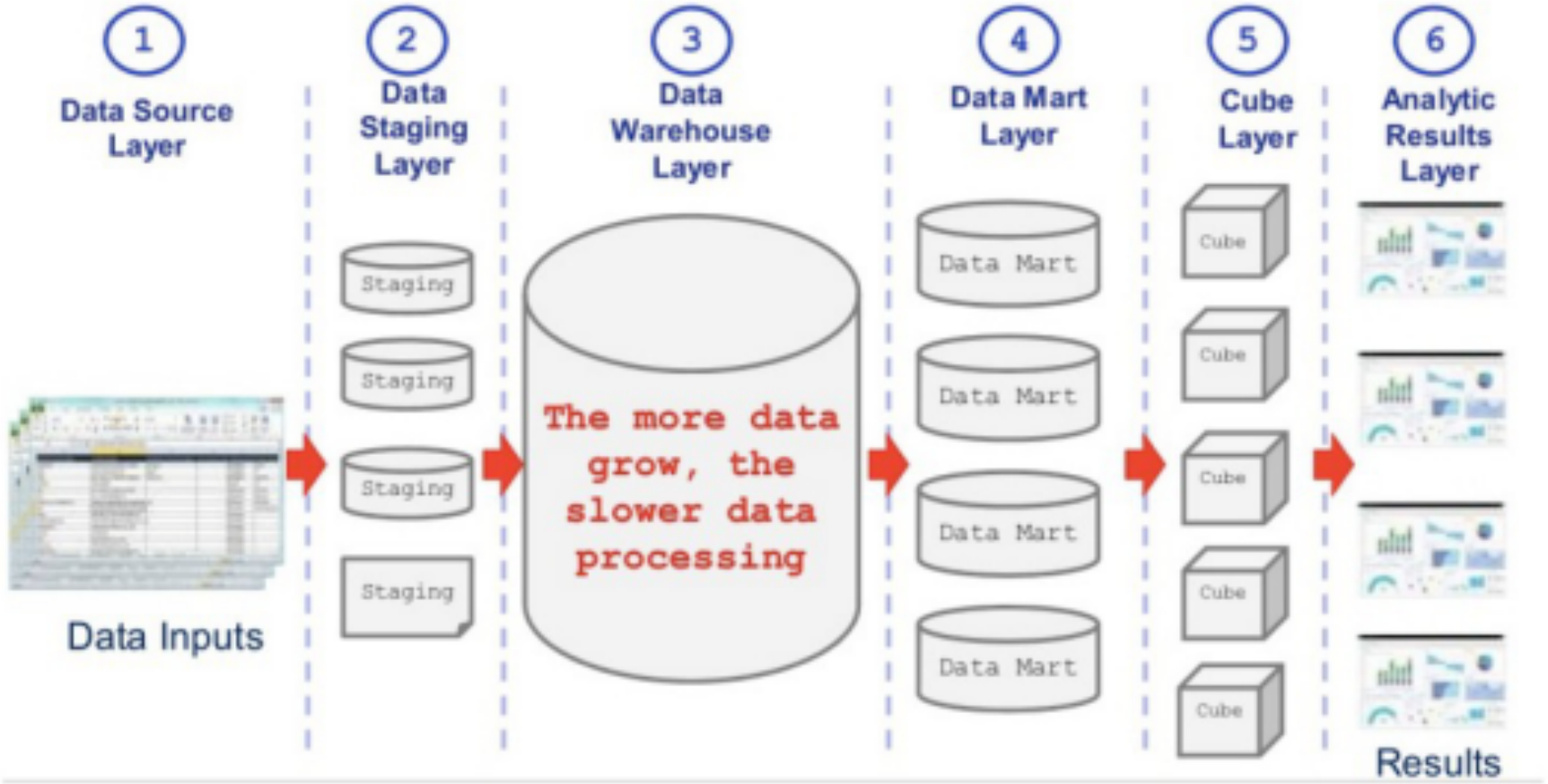
DATA LAKE

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Thaveewat Khanan

2603615 Organization Data Management

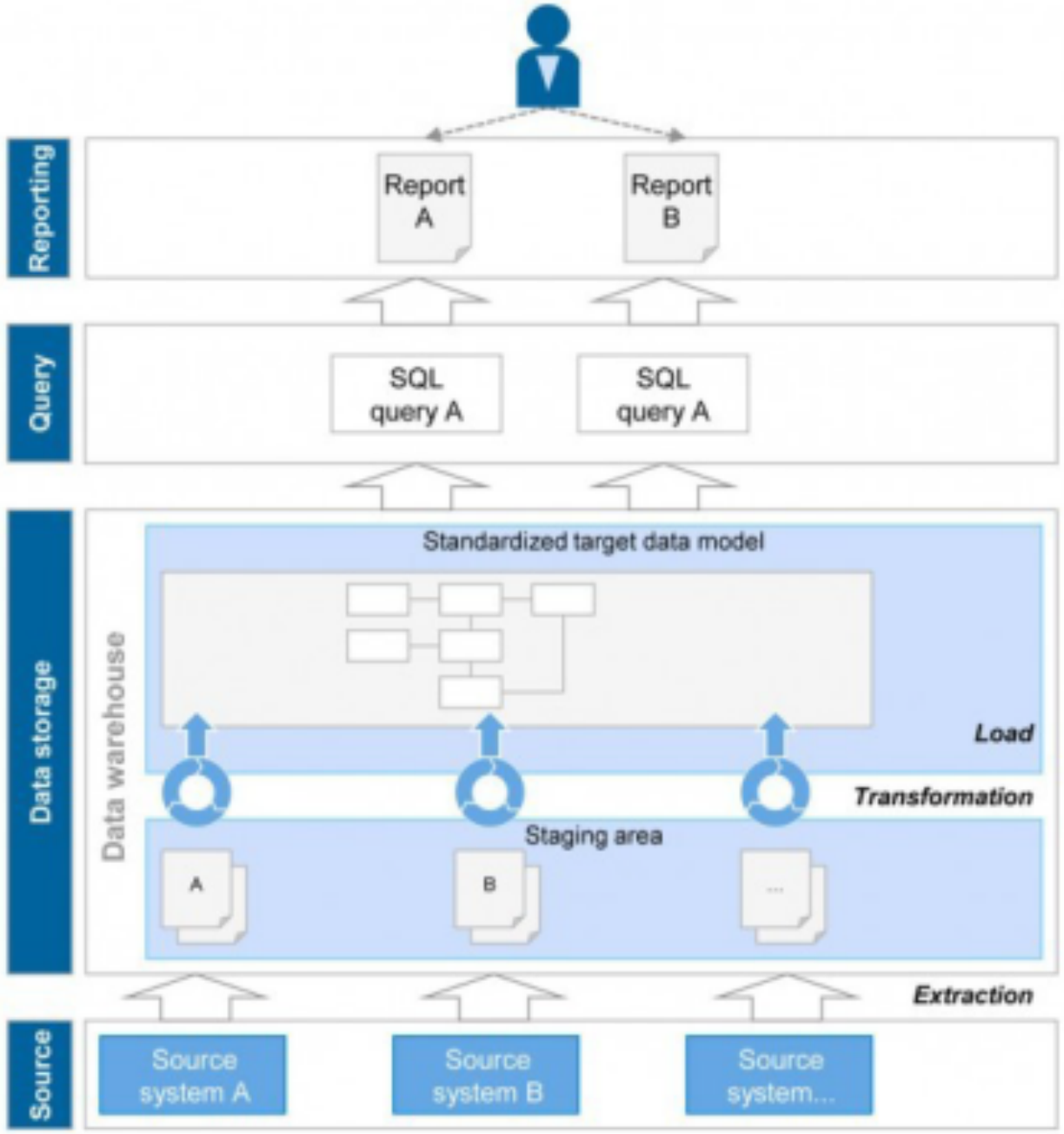
Data Warehouse



Top Concerns from Traditional Data Warehouse Architecture

1. A lot of data duplication lead to cost of data store/storage issue
2. Very slow of data processing and need to restart/roll back the job if any failed
3. Data security issue due to keep data too many copies and various formats

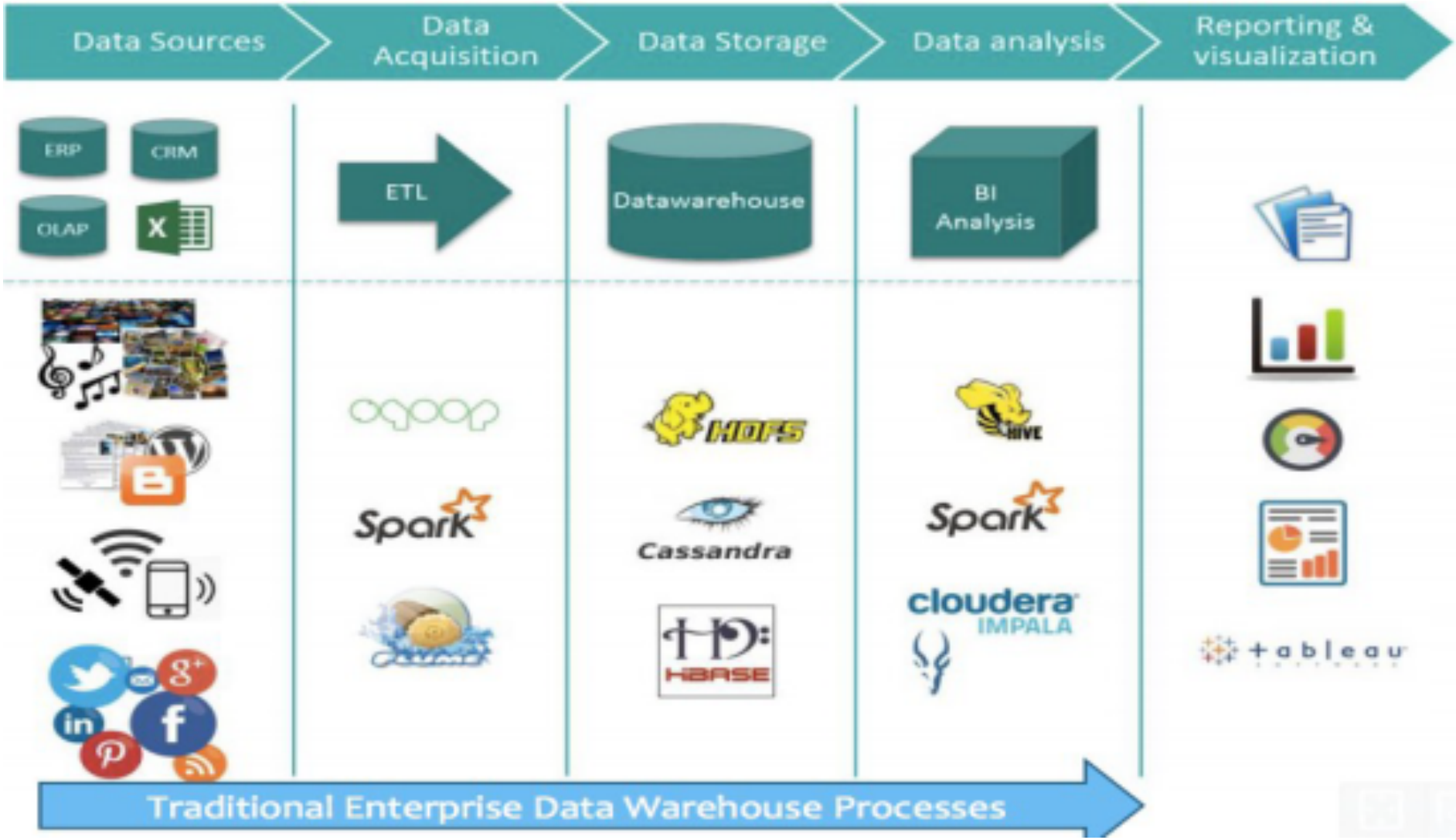
Data Warehouse : ETL



Source: BIG DATA IN BANKING: DATA LAKE INSTEAD OF DATA WAREHOUSE?

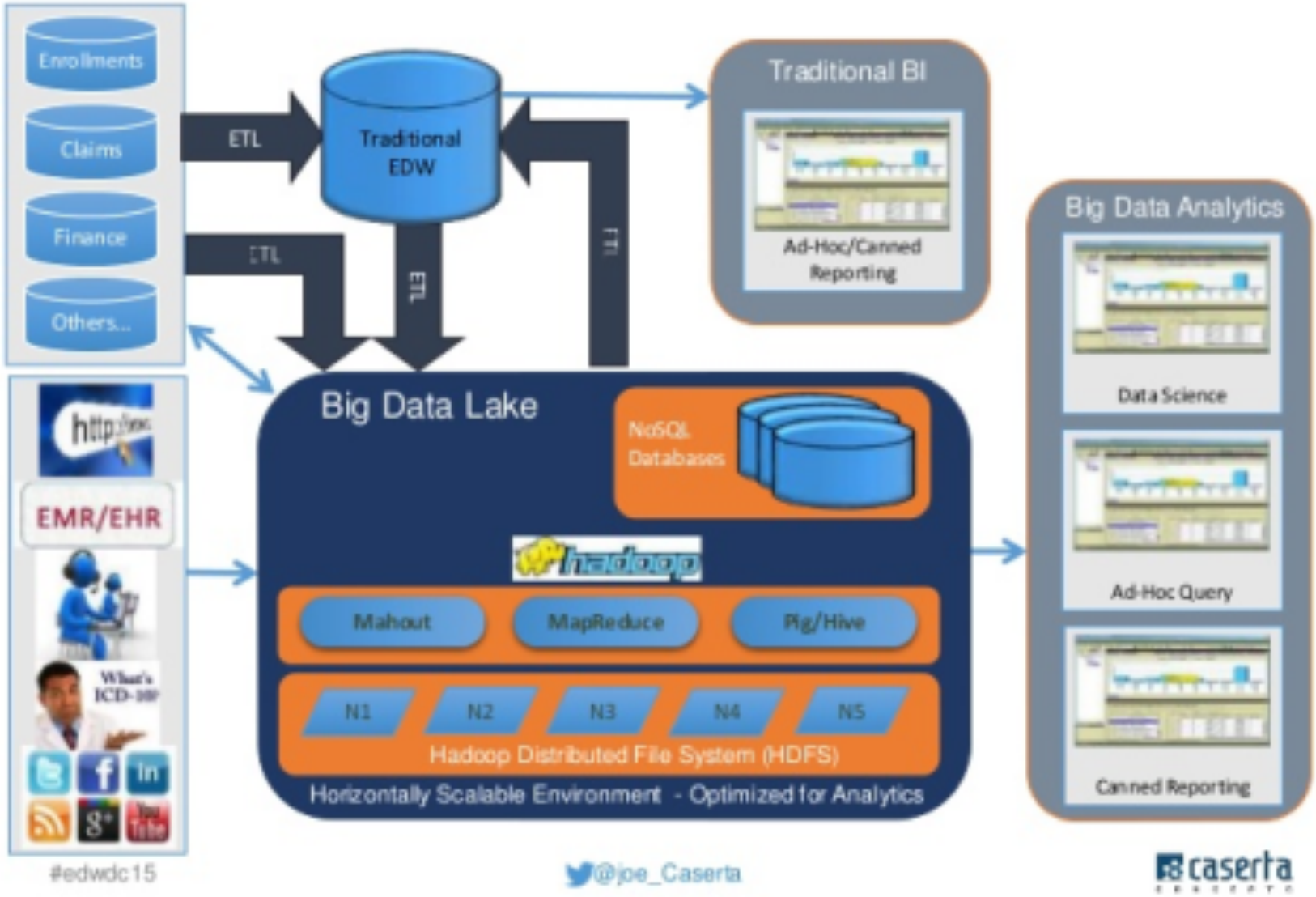
How Data Lake Works?

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: www.clearpeak.com

Today's business environment requires Big Data



CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Data Lake isn't just a technology
It is an architecture

Data Lake: Definition

Is a huge repository

Is not Hadoop

is not a database, it uses various NoSQL and In-Memory databases.

Stores large volumes of both unstructured and structured data.

Is a data scientist's favorite hunting ground.

Data Lake cannot be implemented in isolation.

Data Lake: Key Benefits

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Scale as much as you can

Plug in disparate data sources

Acquire high-velocity data: Store in native format

Don't worry about schema

Unleash your favorite SQL

Advanced algorithms

Administrative resources

Data Lake v.s. Data WareHouse

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Complementary to EDW (not replacement)	Data lake can be source for EDW
Schema on read (no predefined schemas)	Schema on write (predefined schemas)
Structured/semi-structured/Unstructured data	Structured data only
Fast ingestion of new data/content	Time consuming to introduce new content
Data Science + Prediction/Advanced Analytics + BI use cases	BI use cases only (no prediction/advanced analytics)
Data at low level of detail/granularity	Data at summary/aggregated level of detail
Loosely defined SLAs	Tight SLAs (production schedules)
Flexibility in tools (open source/tools for advanced analytics)	Limited flexibility in tools (SQL only)

Data Lake Maturity & Risks

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

More Data Sources

More Applications

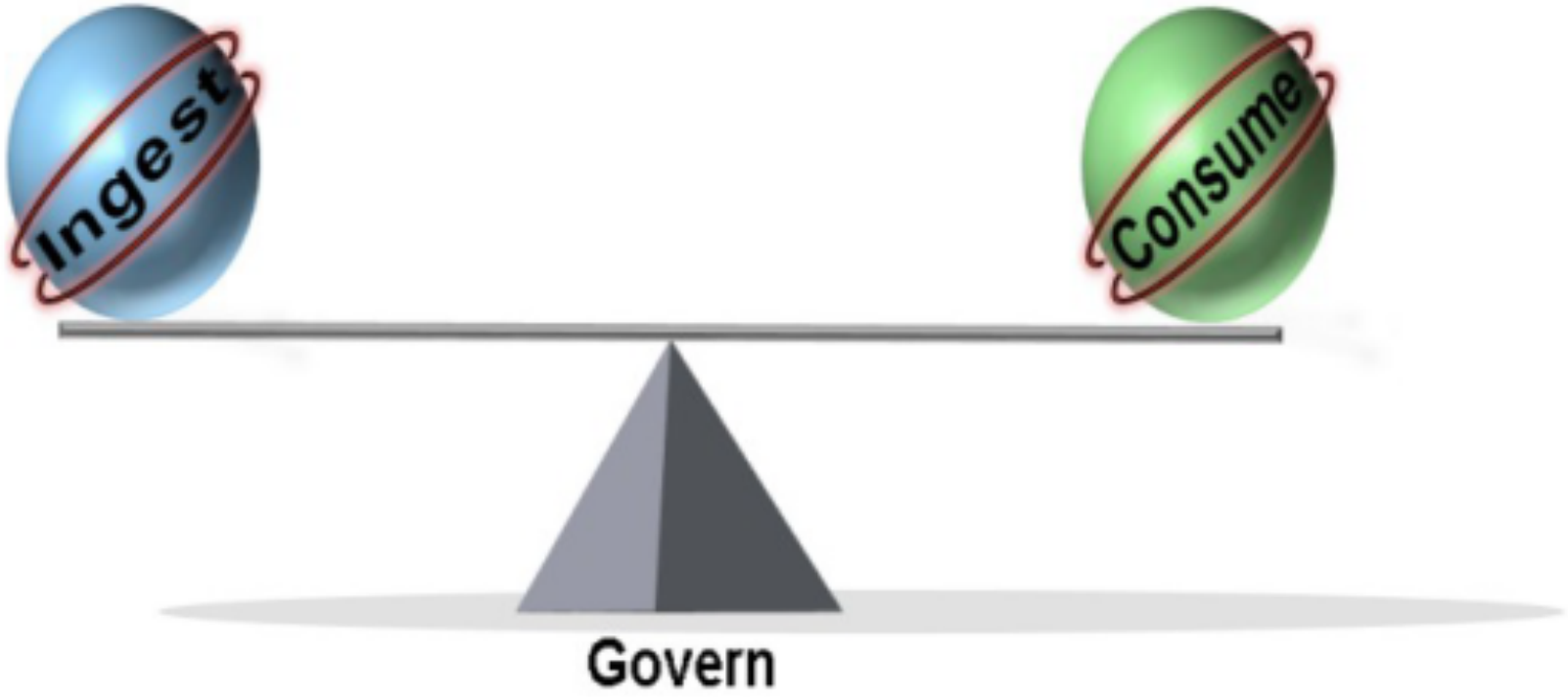
More Business Units

More Users

Without proper governance mechanisms
Data lakes risk turning data swamps

Data Lake Governance

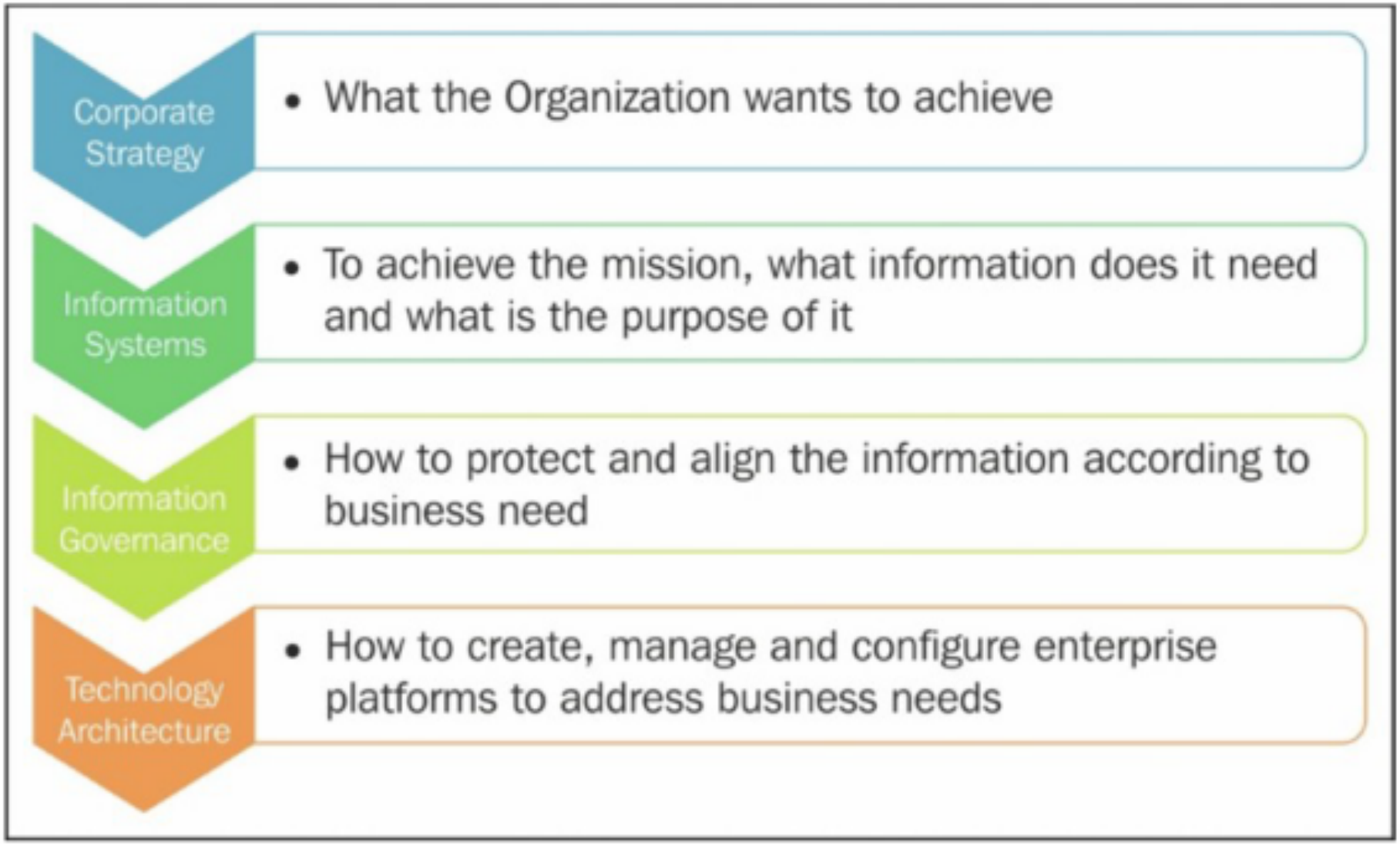
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: What is “Just-Enough” Governance for the Data Lake?

Data Governance

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: Data Lake Development with Big Data

Data Lake Governance

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Fundamental Capabilities

- **The definition of the incoming data from a Business use perspective**
- **Documentation of the context, lineage, and frequency of the incoming data;**
- **Security level classification of the incoming data;**
- **Documentation of creation, usage, privacy, regulatory, and encryption business rules which apply to the incoming data.**

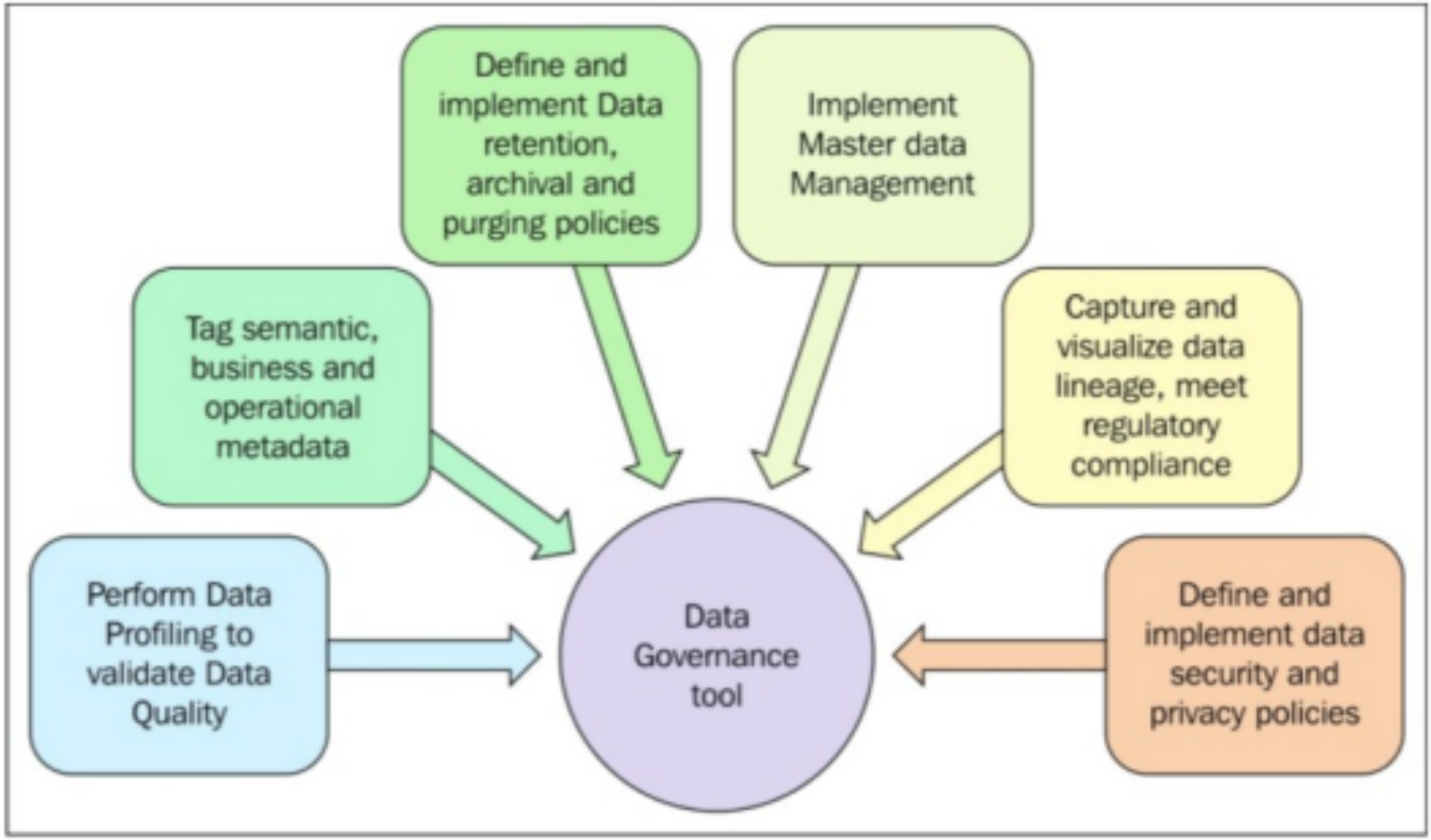
Source: What is “Just-Enough” Governance for the Data Lake?

What can it do for my Data Lake

- Where did my data come from ? How is it being transformed ?
- Track usage, resolve anomalies, visualize, optimize and clarify data lineage Search and access data
- Assess data quality and fitness for purpose
- Govern who can/cannot access the data
- Data life cycle management, archiving and retention policies
- Auditing, compliance

Source: Data Governance in the Hadoop Data Lake

Data Governance Tools



Source: Data Lake Development with Big Data

Summary

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

- Big Data: Data lake instead of data warehouse?
- Data Lake is not only a technology, it is an architecture
- Data Lake components: Data acquisition (intake), Data management, Data Storage, Data consumption (Discovery)
- Data Lake governance is very important