DATA LAKE

Thaveewat Khanan

DATA LAKE

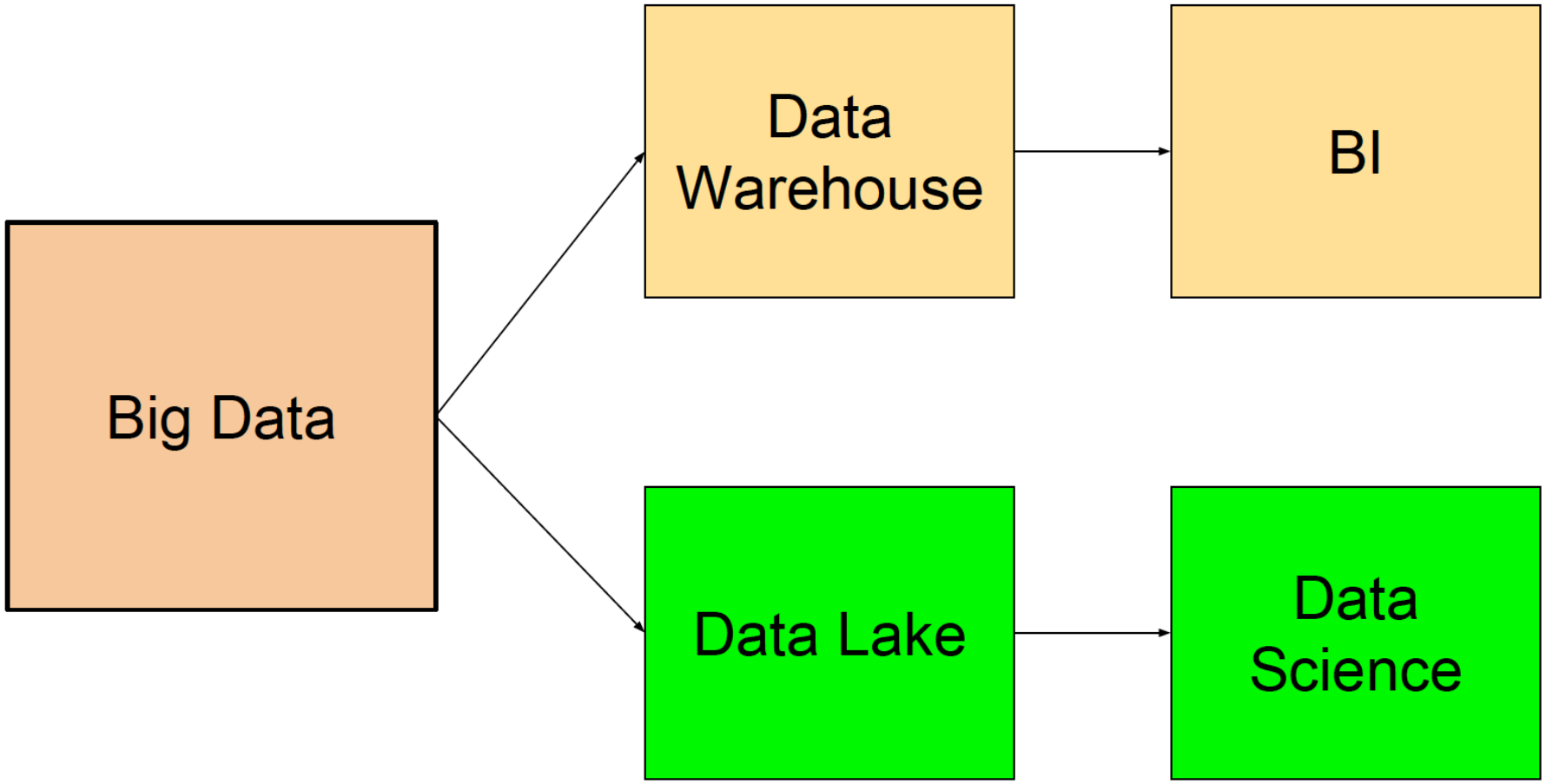
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Thaveewat Khanan

The old way: Ask, then collect

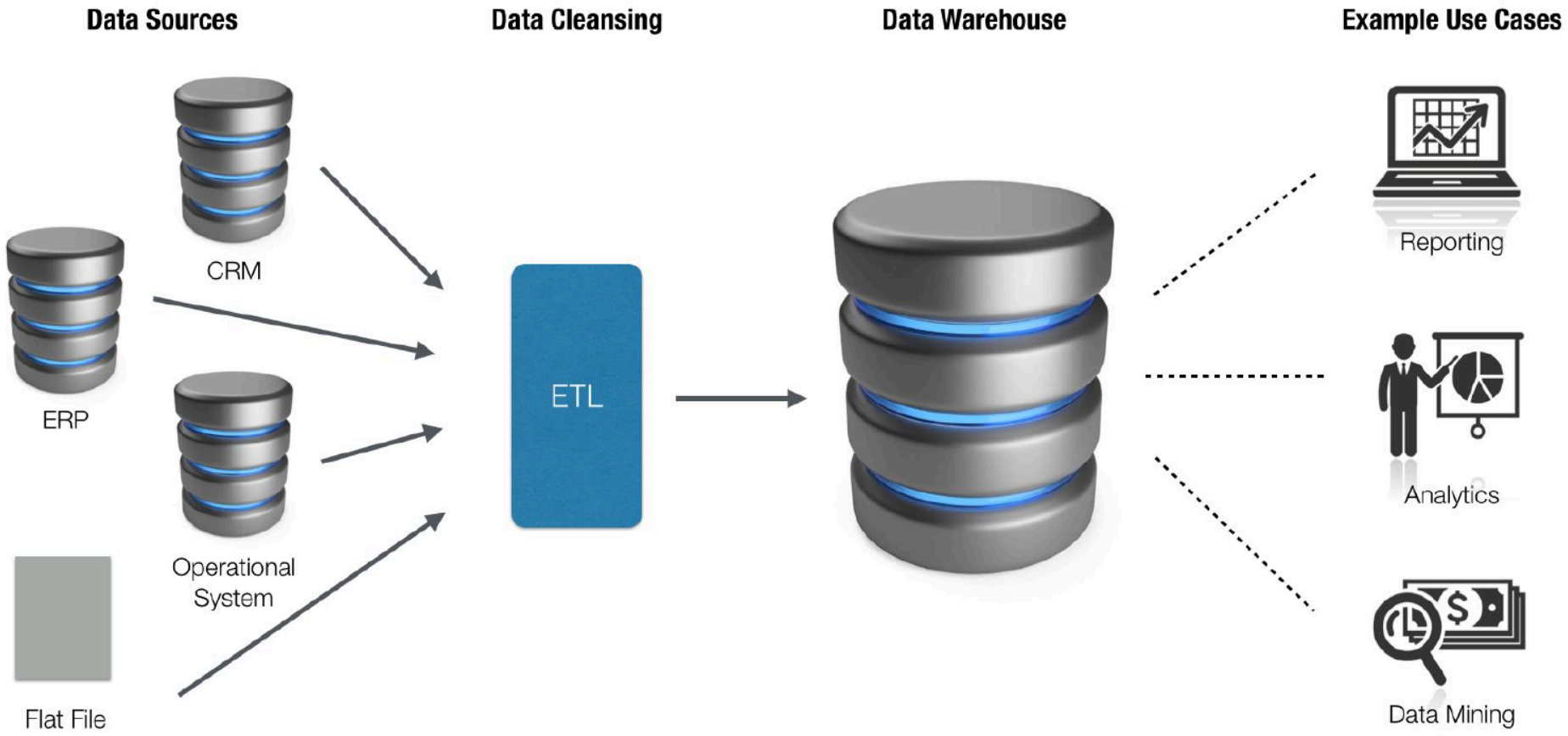
The new way: Collect, then ask

Source: Data Science and Critical Thinking, A, Croll


Big Data
Data Warehouse
BI
Data Lake
Data Science

Data Warehouse

Traditional approach to integrating data for consistency and quality



Data Warehouse Environment

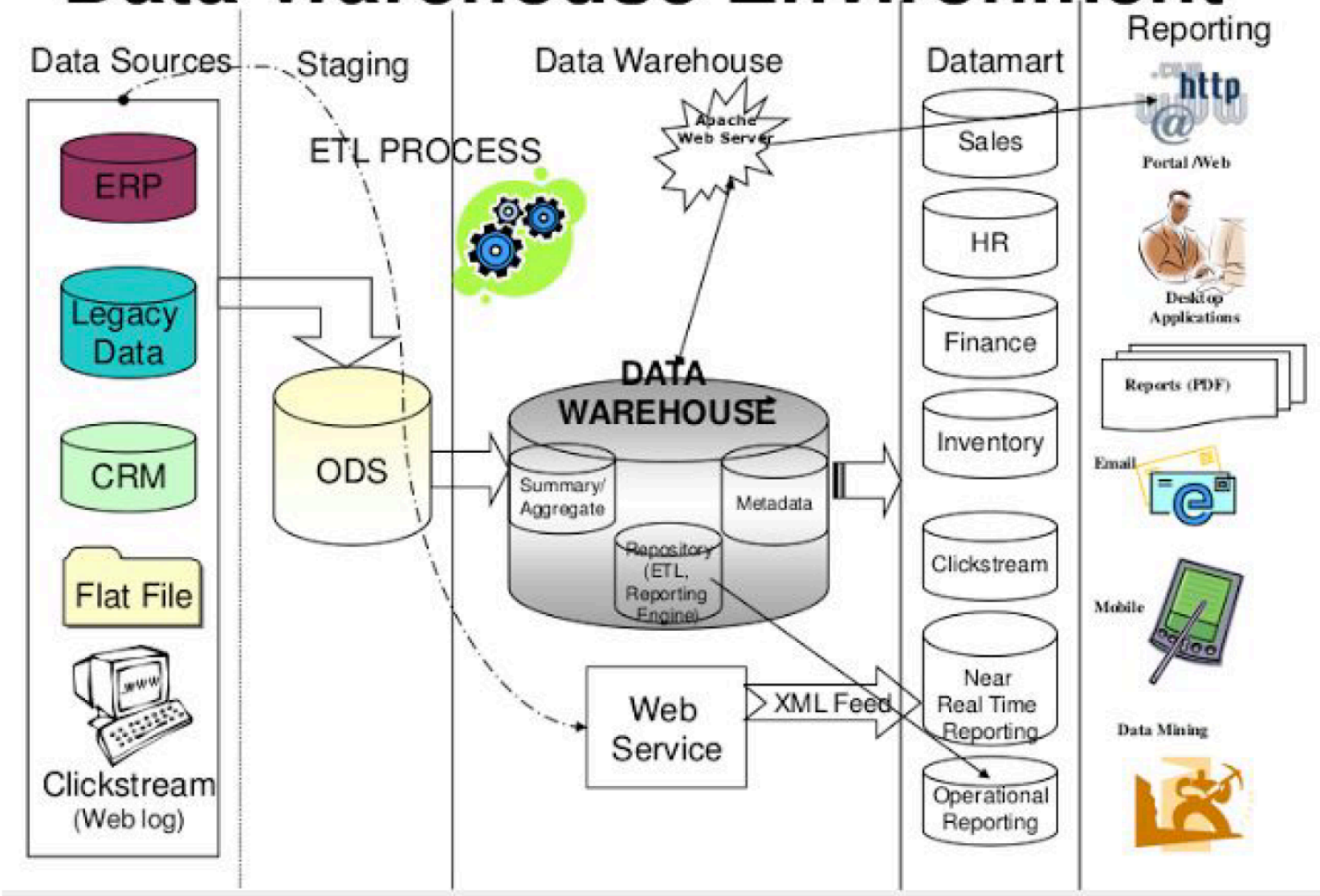
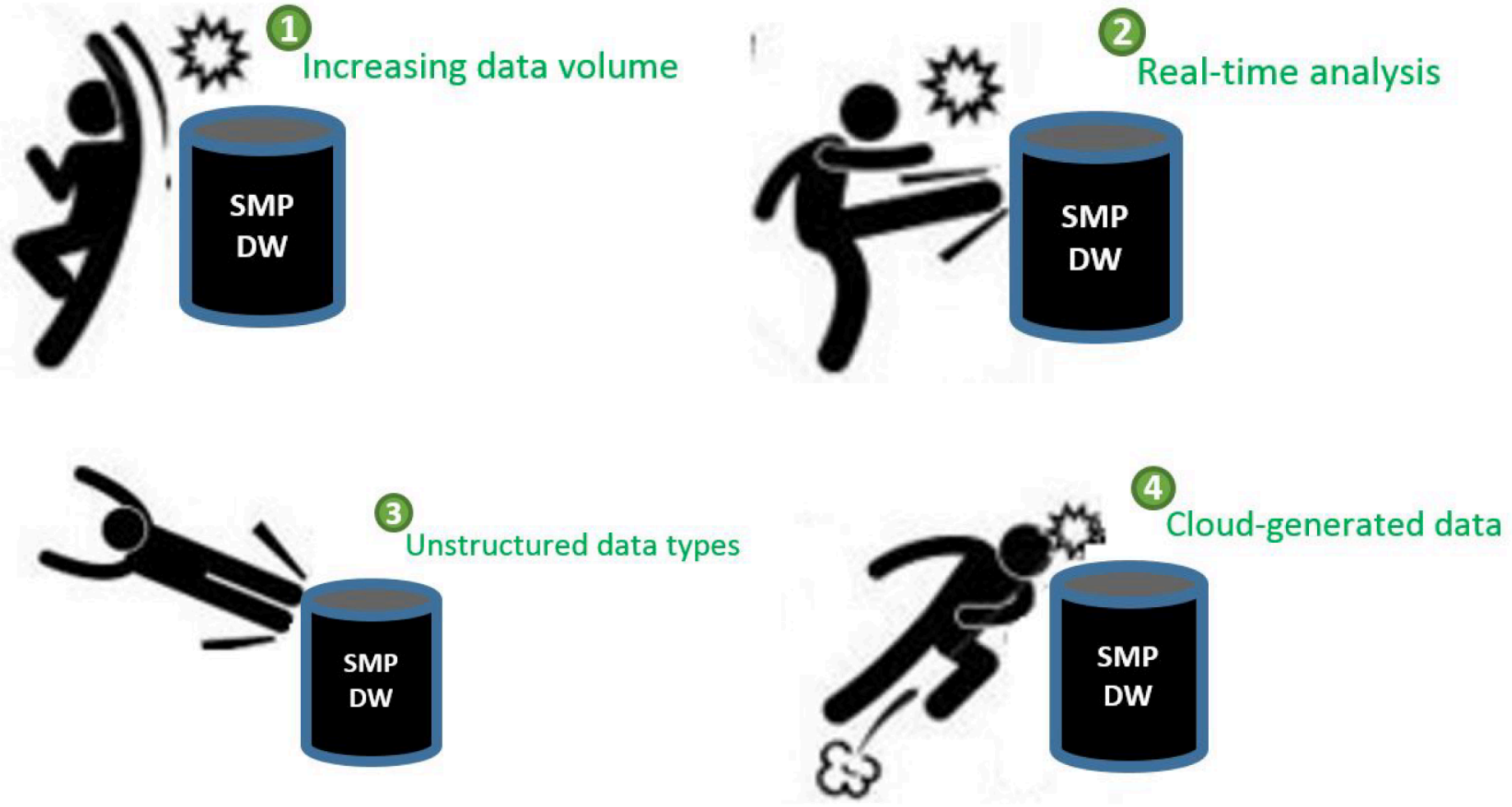


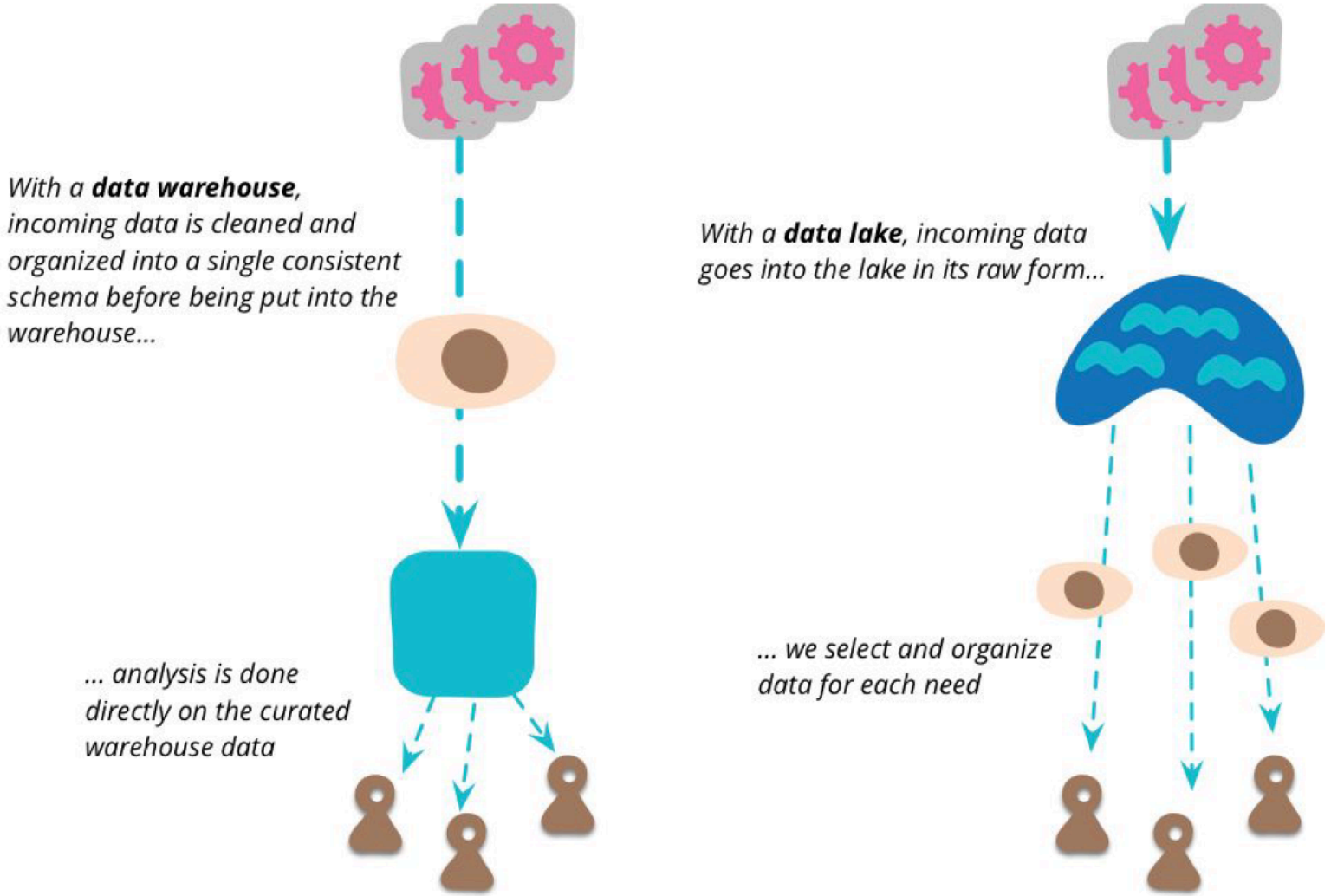
Image: rodneyrohrmann.blogspot.com

Data Warehouse

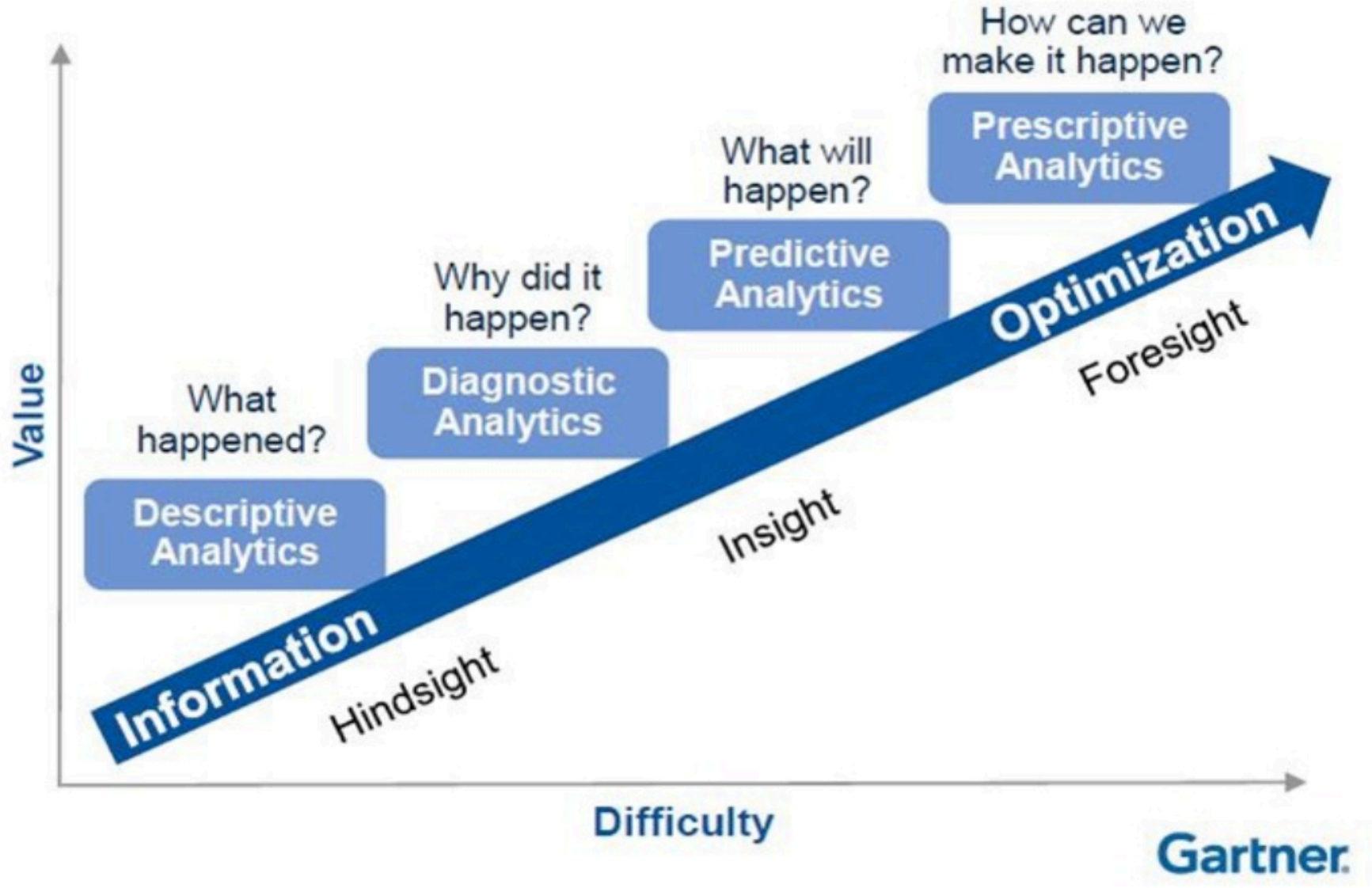


Source: dinesql.blogspot.com

Differences between Data Lake and Data Warehouse

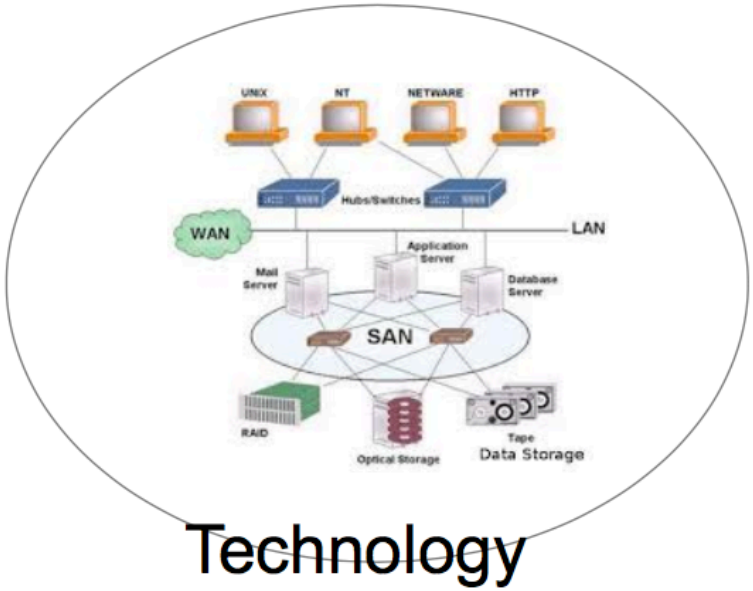


Source: martinfowler.com/bliki/DataLake.html

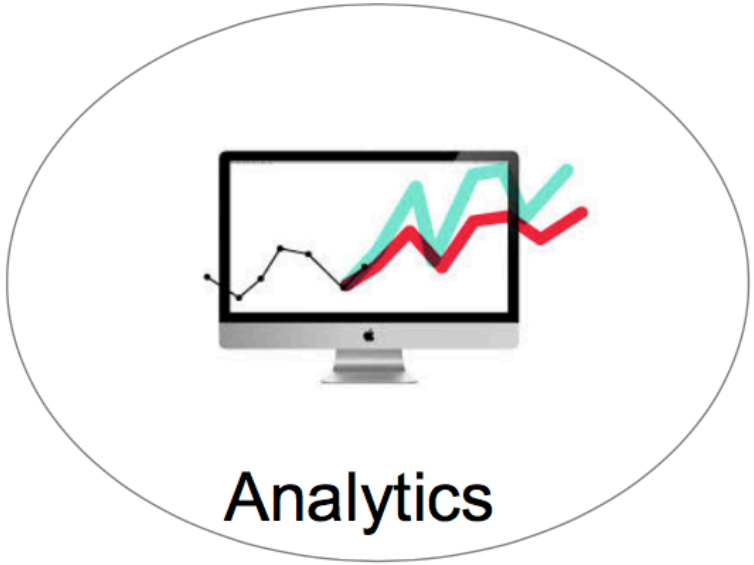

How can we make it happen?
Prescriptive Analytics
What will happen?
Predictive Analytics
Why did it happen?
Diagnostic Analytics
What happened?
Descriptive Analytics
Optimization
Foresight
Insight
Hindsight
Information
Value
Difficulty
Gartner

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

Data Sources



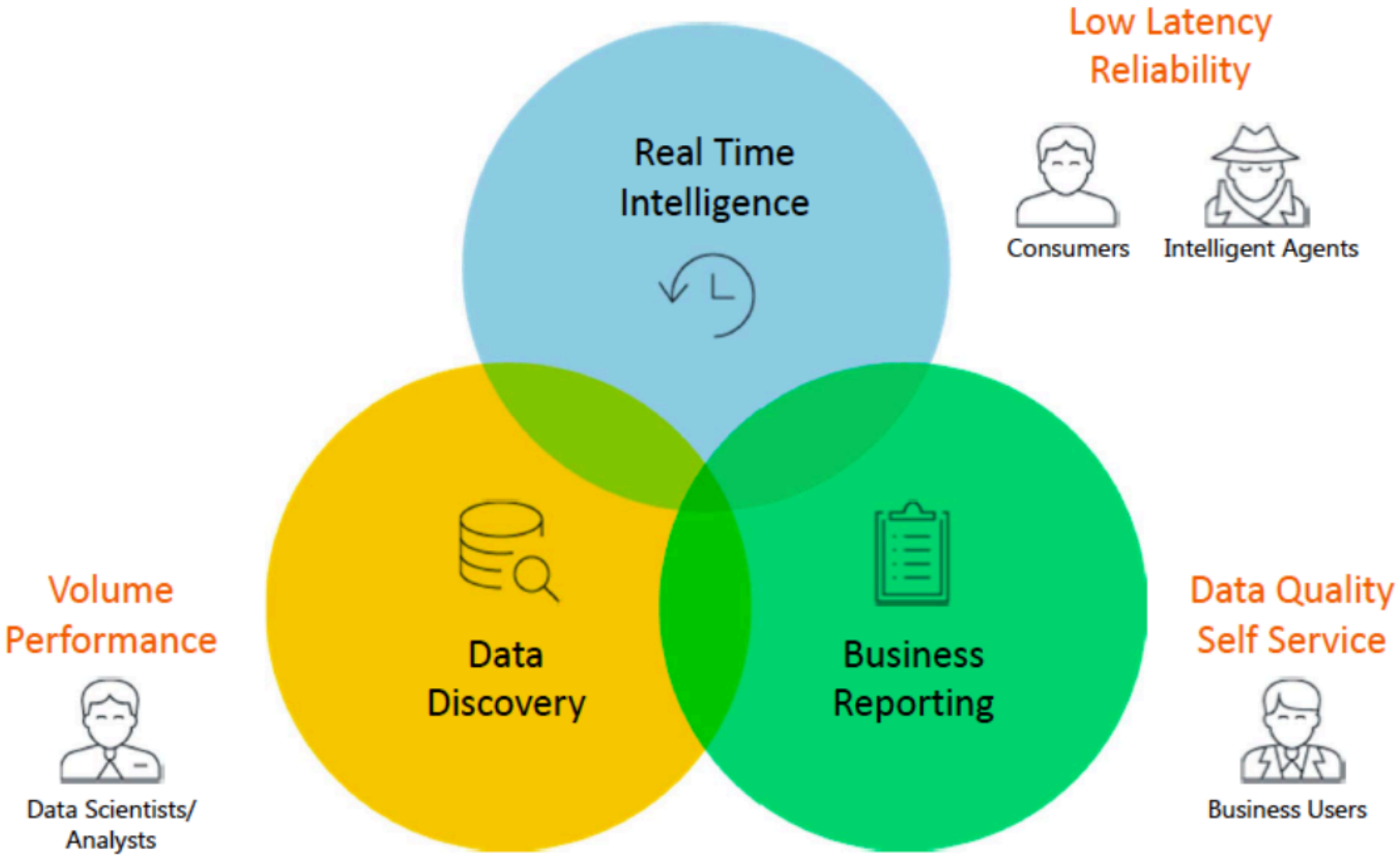
Technology



Analytics

Big Data Analytics

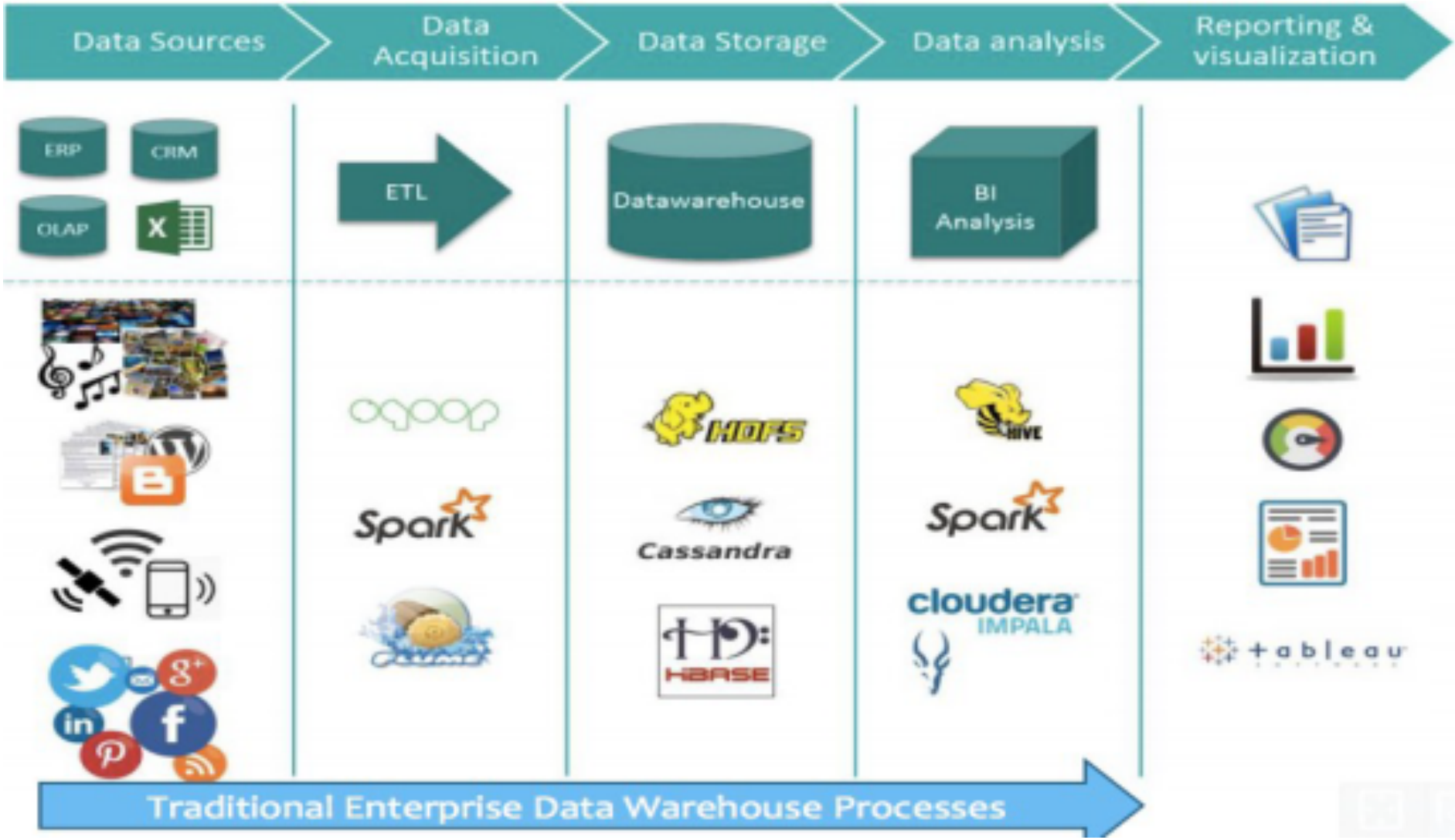
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: SoftServe

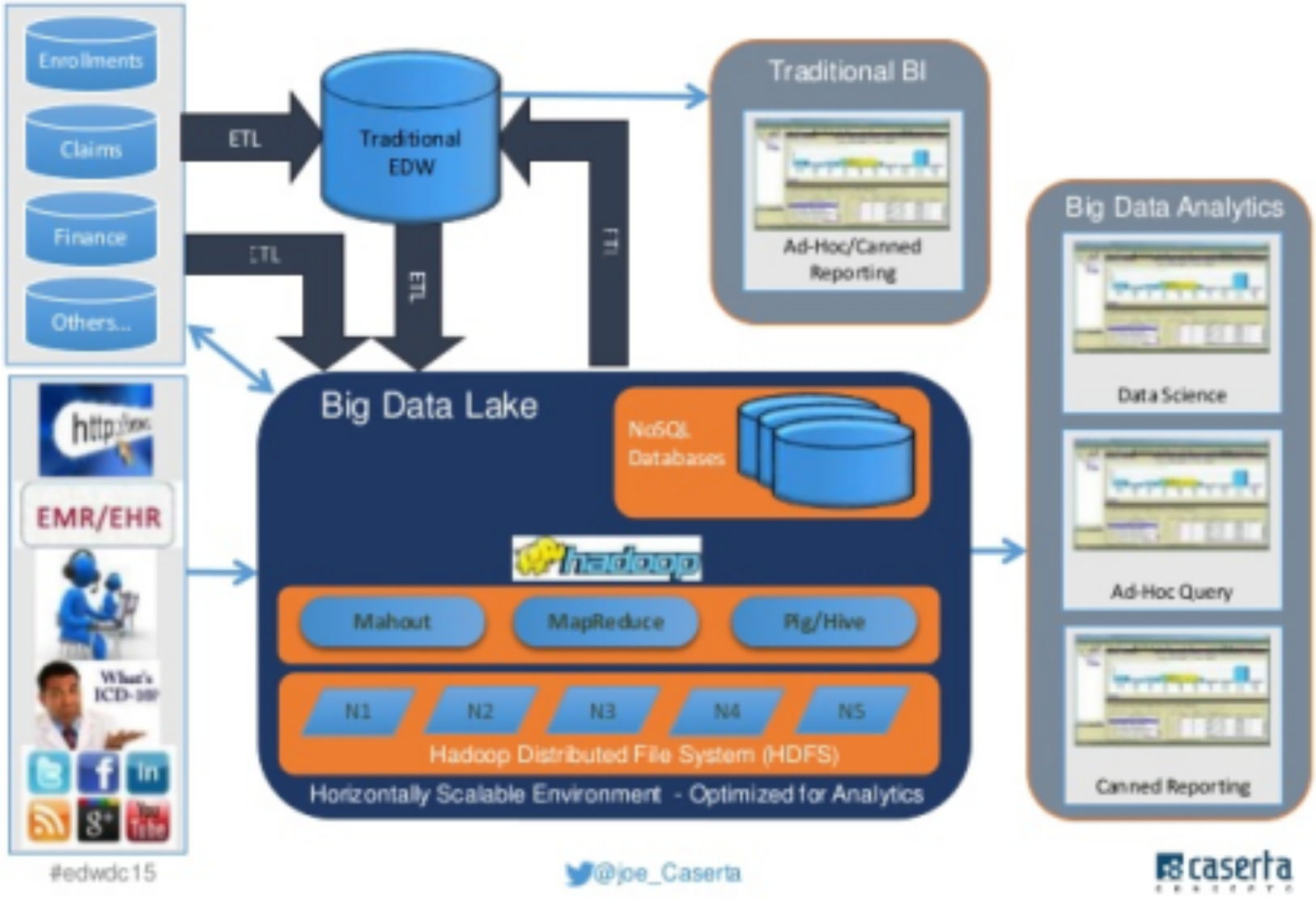
How Data Lake Works?

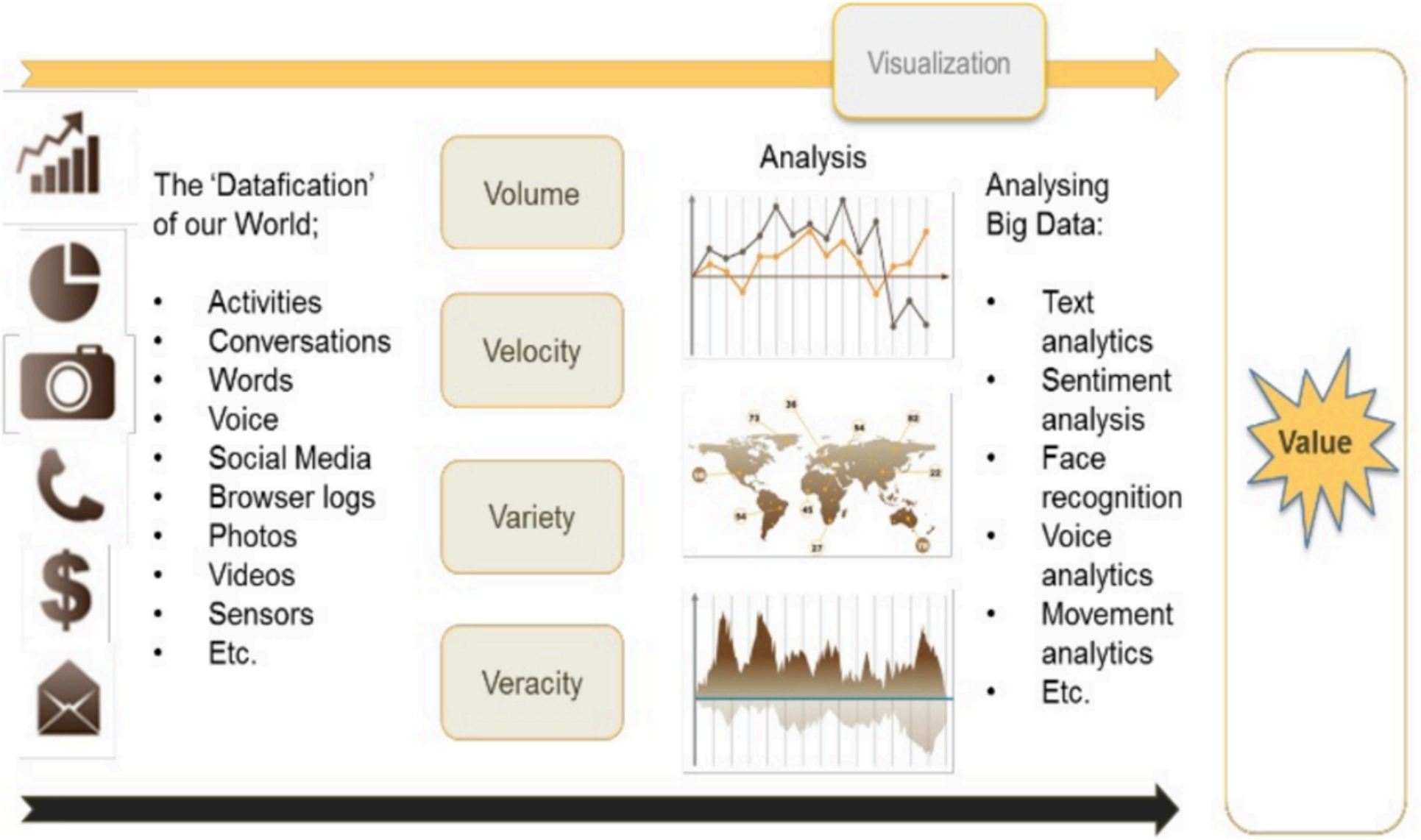
CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE



Source: www.clearpeak.com

Today's business environment requires Big Data





Source: Bernard Marr

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

Data Lake: Key Benefits

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 Risks

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

More Data Sources

More Applications

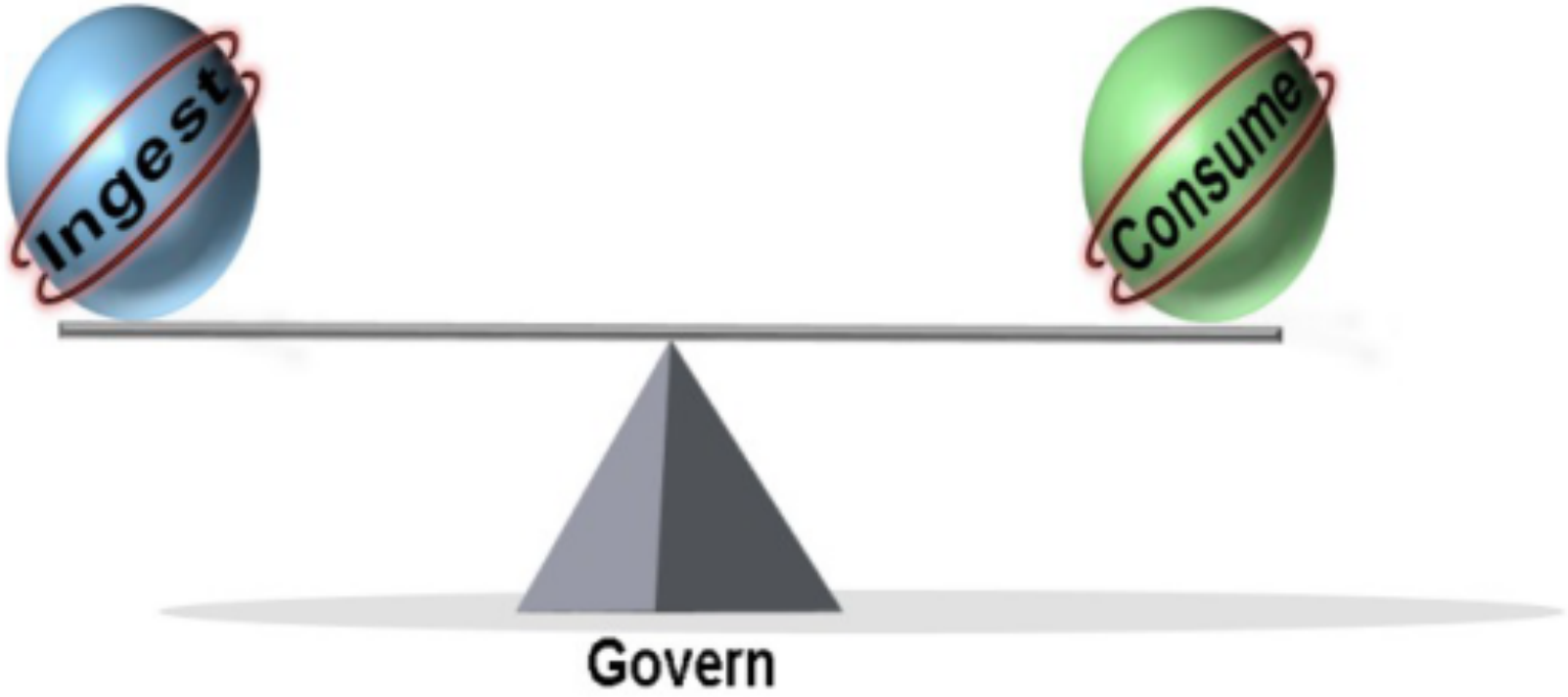
More Business Units

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 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

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

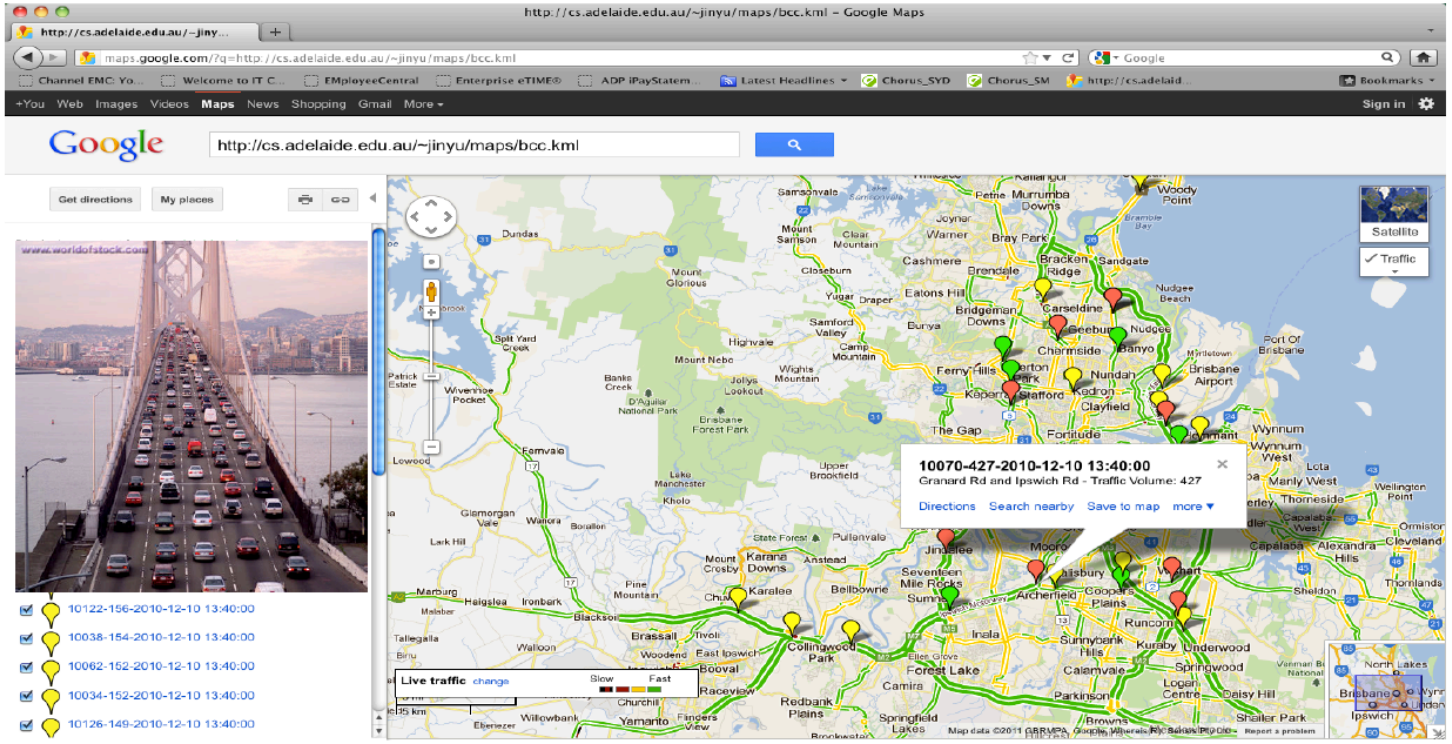
Customer Case Studies on Big Data Lake

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Customer Example: Analytics

Municipal Traffic Analysis to Simulate Traffic Velocity Patterns and Reduce Delays

- Correlate multiple types of data (GPS, weather, sensor, video, social media)
- Simulation techniques to model traffic transition points
- Signal retiming to minimize stops and delays
- Peak delays reduced by 16% and stops reduced by 22%



Pivotal

Thaveewat Khanan

Customer Case Studies on Big Data Lake

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE