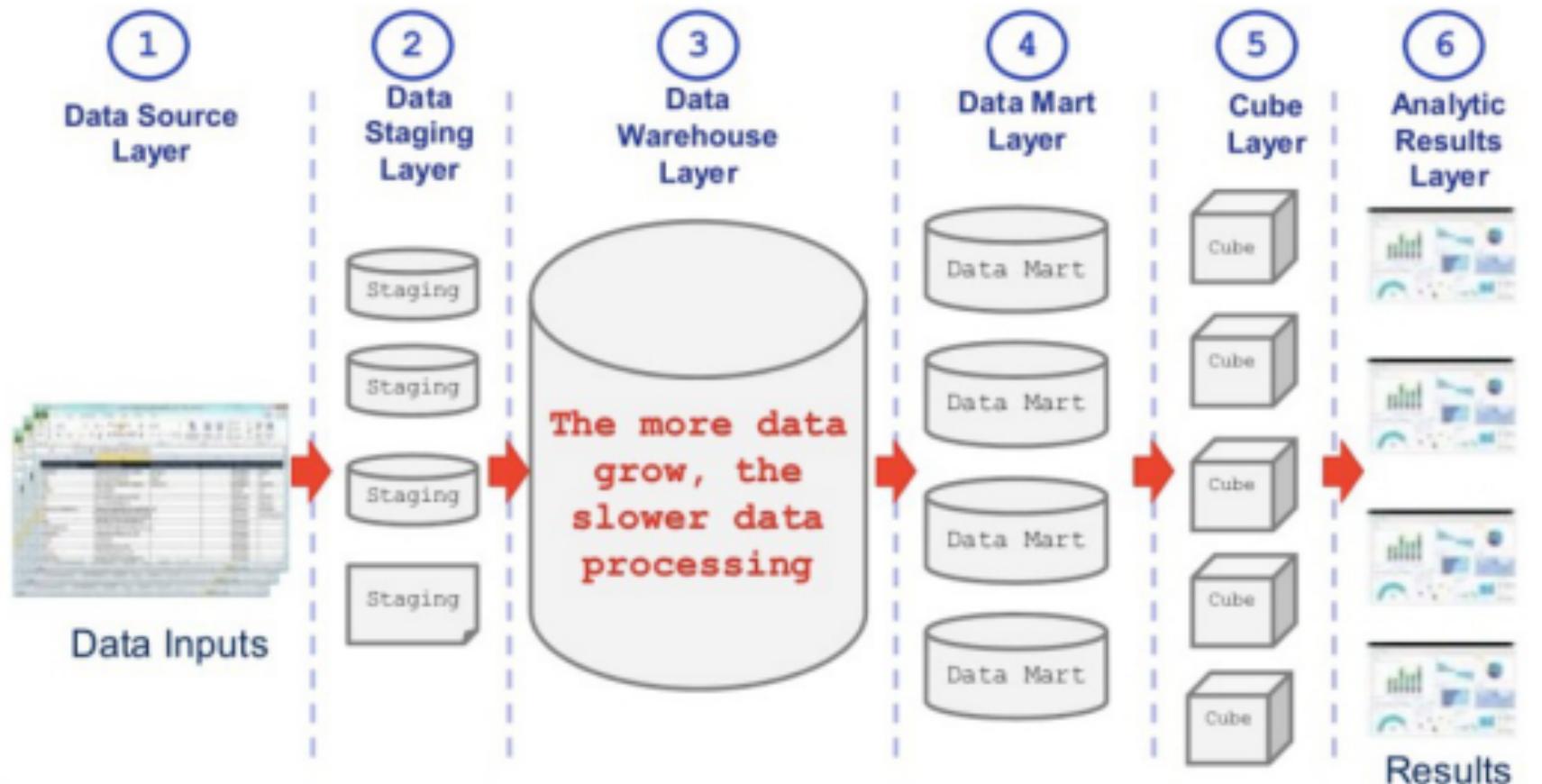


DATA LAKE

DATA LAKE



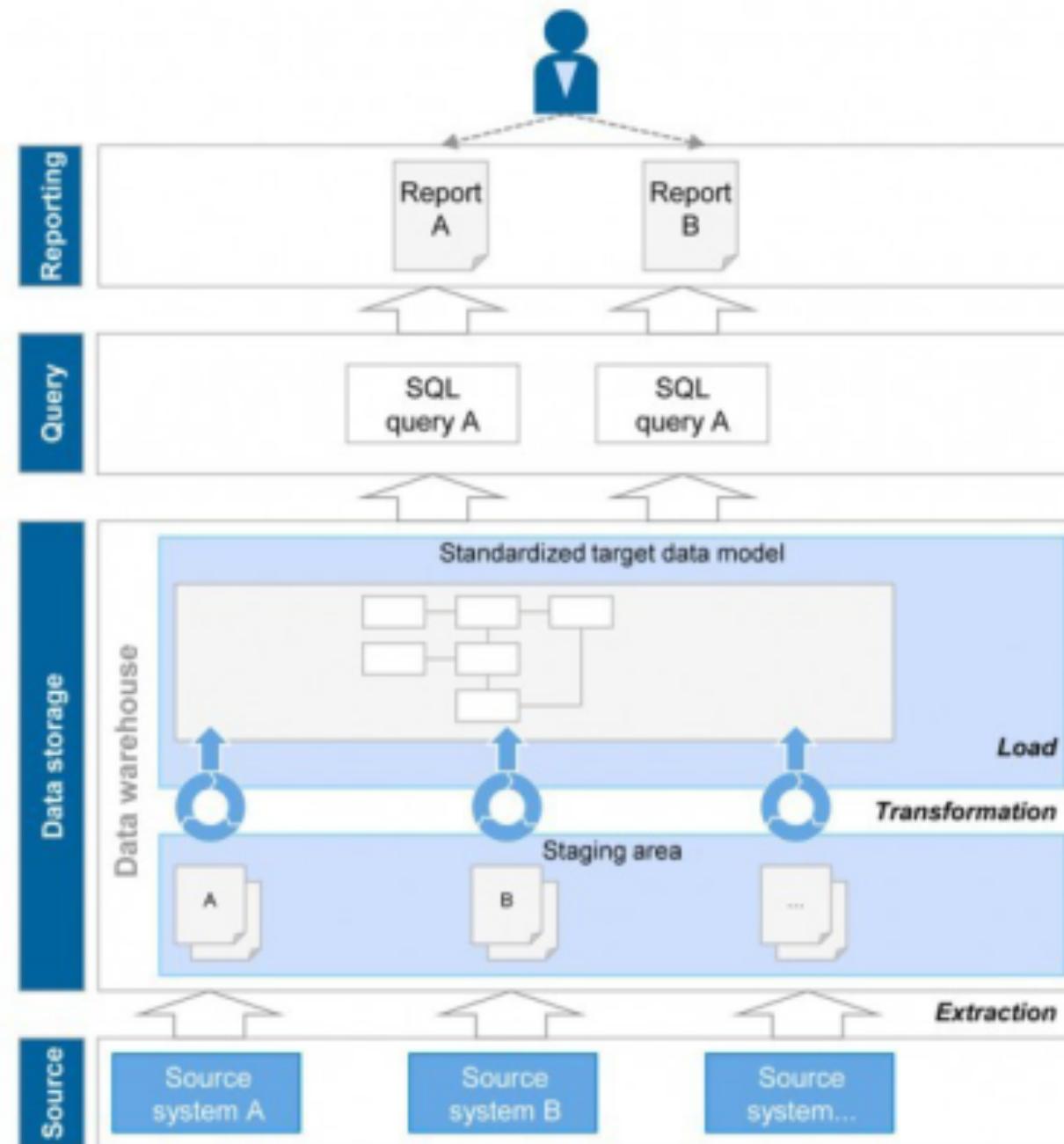
Data Warehouse



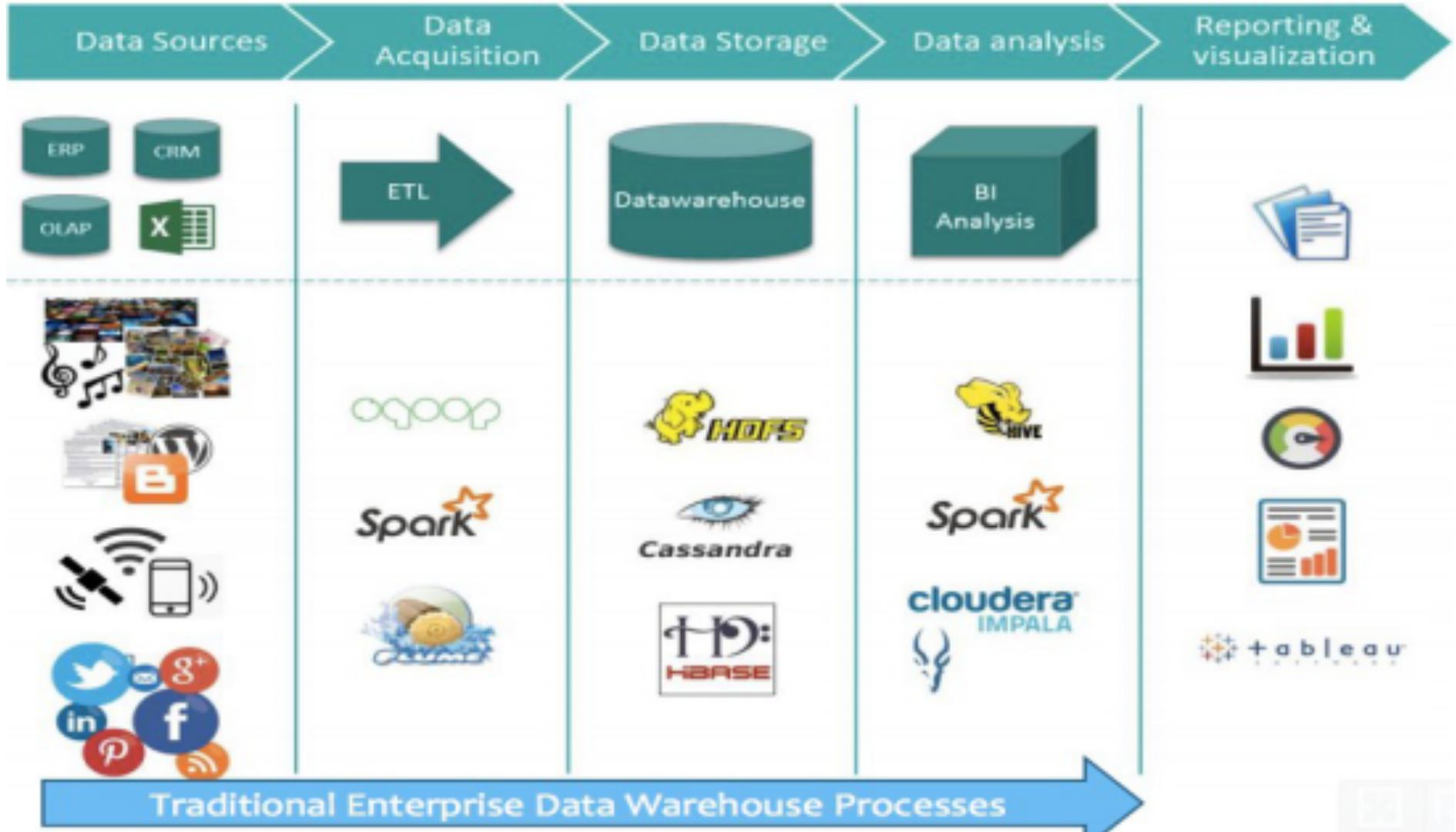
Top Concerns from Traditional Data Warehouse Architecture

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

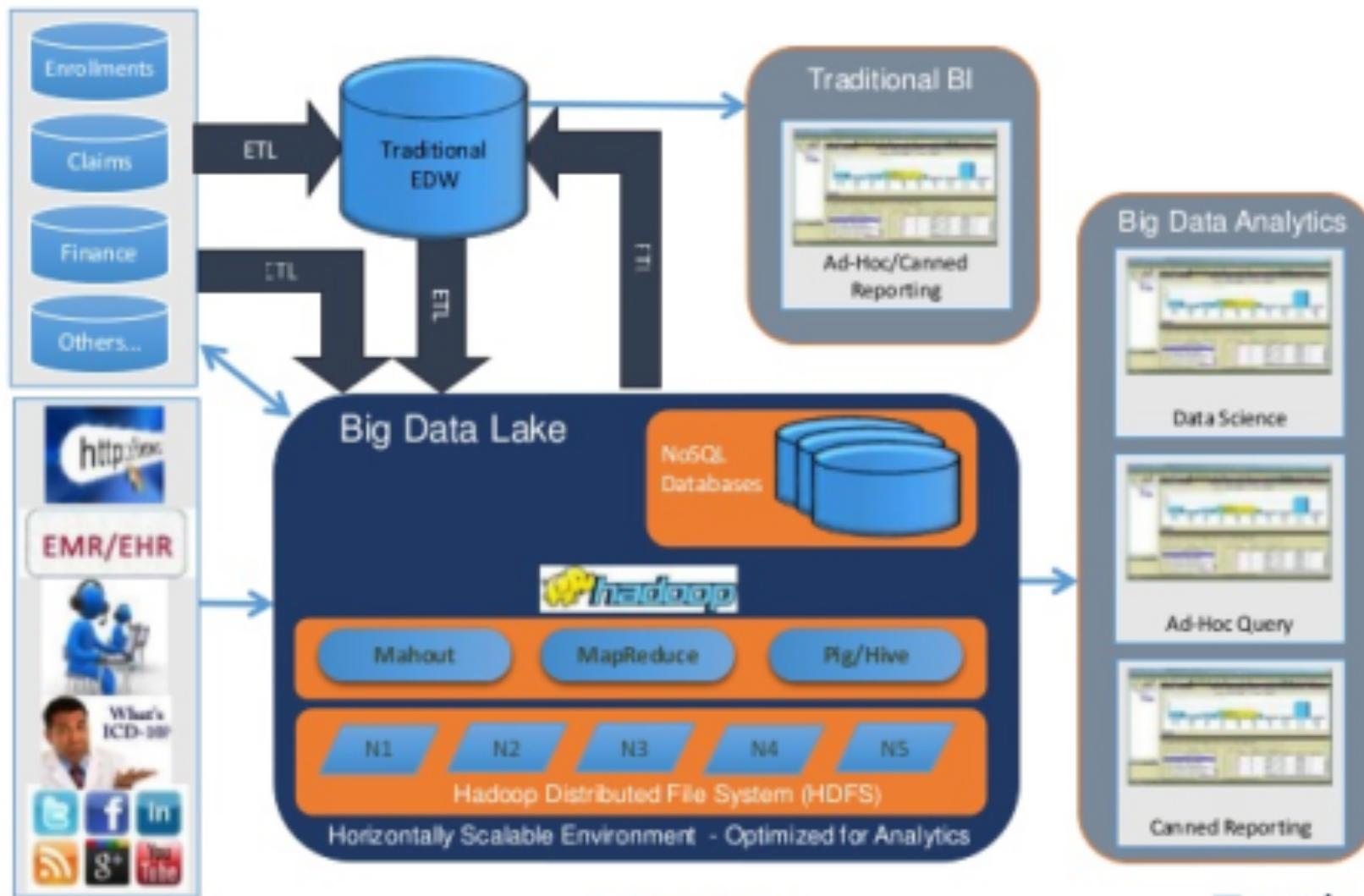
Data Warehouse : ETL



How Data Lake Works?



Today's business environment requires Big Data



#edwdc15

@joe_Caserta



Thaveewat Khanan

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

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

More Data Sources

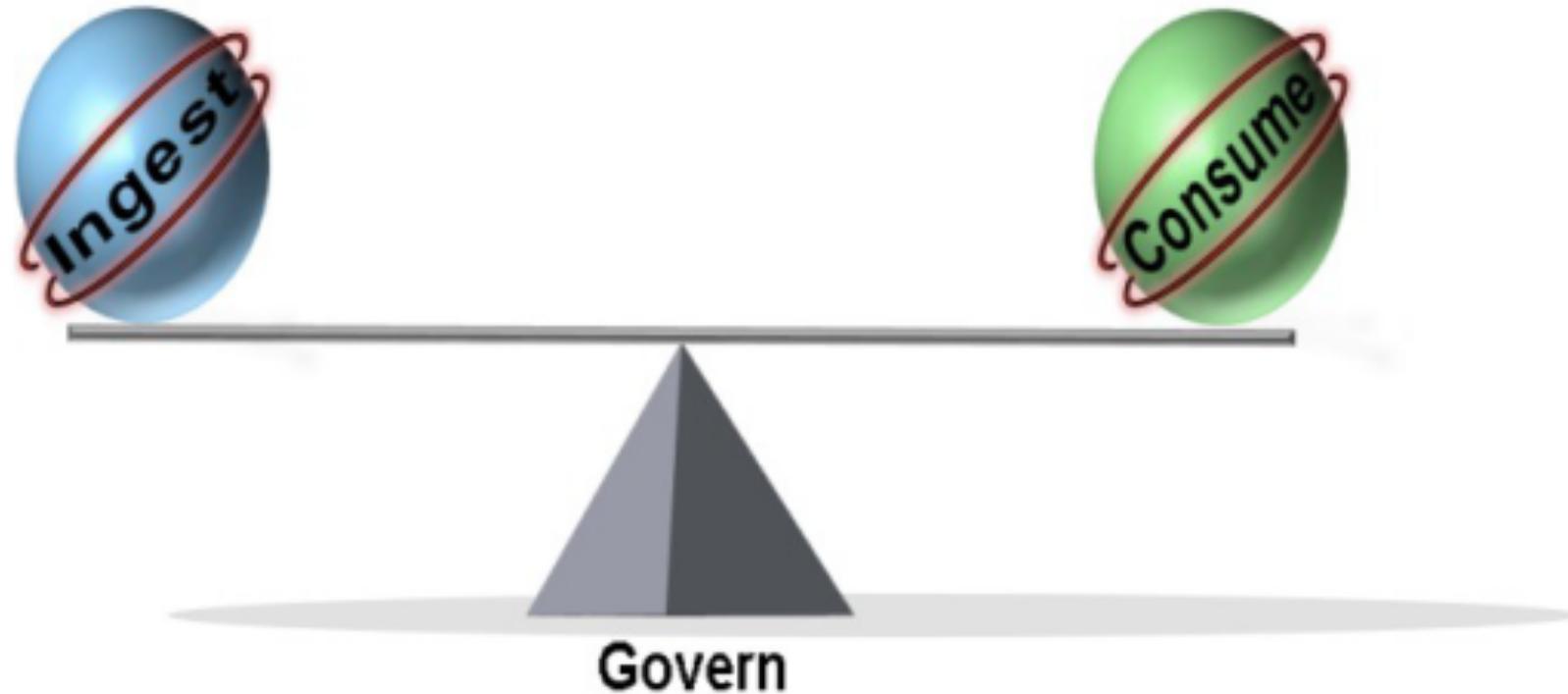
More Applications

More Business Units

More Users

Without proper governance mechanisms
Data lakes risk turning data swamps

Data Lake Governance



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

Thaveewat Khanan

Data Lake Governance

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.**

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

Summary

- **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



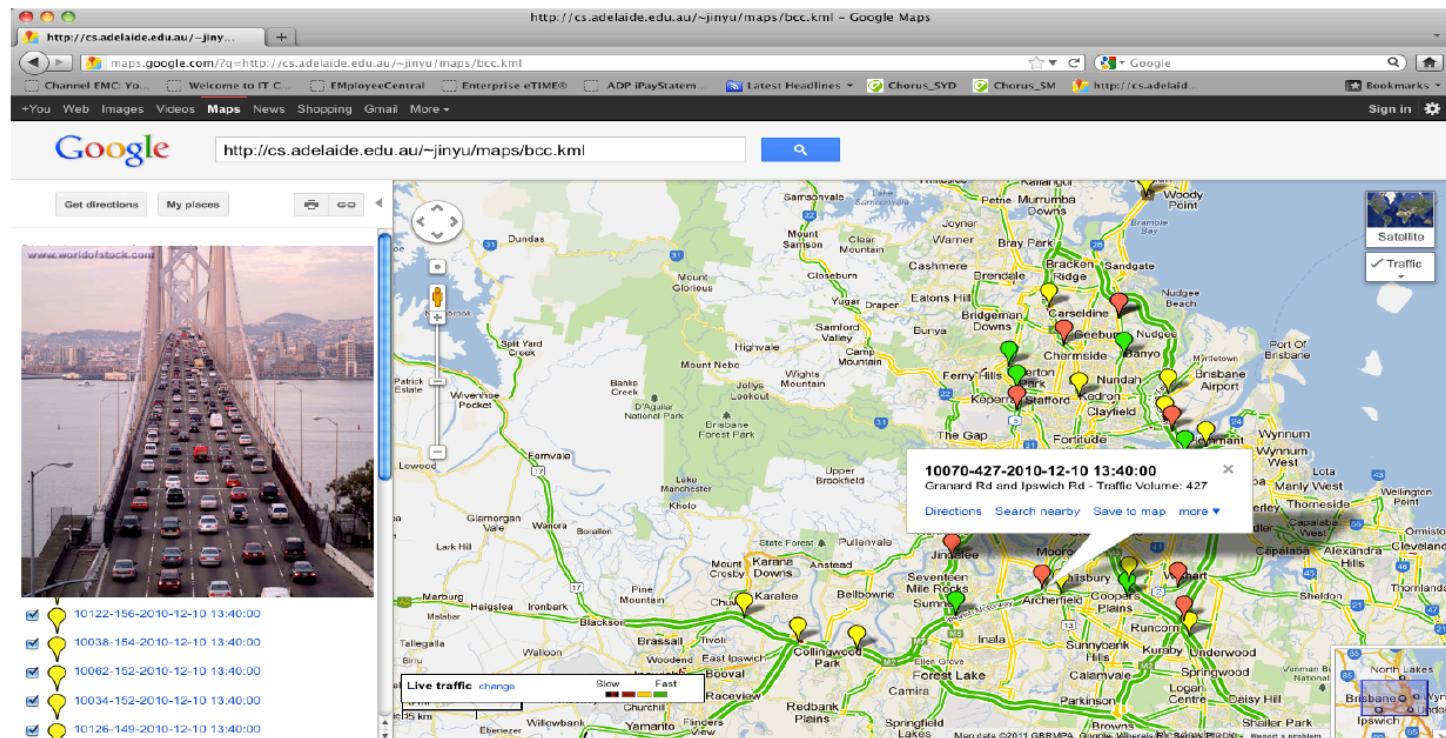
CHULALONGKORN
BUSINESS SCHOOL

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™

Pivotal Confidential—Internal Use Only

20

Thaveewat Khanan

Customer Case Studies on Big Data Lake

China Railways scales online sales for the largest rail way in the world with Pivotal Gemfire

- Reliable, High Performance and Continuous uptime with thousands of transactions per second
- On demand scaling for growth
- Cost effective operations

