data governance
WINTER CONFERENCE

DELRAY BEACH, FL
NOVEMBER 14-18, 2016
DGWINTER.COM

Data Governance for the Data Lake

Improving Agility, Flexibility, and Value

Donna Burbank
Global Data Strategy Ltd.

Nov 16th, 2016

Global Data Strategy
The Business of Data

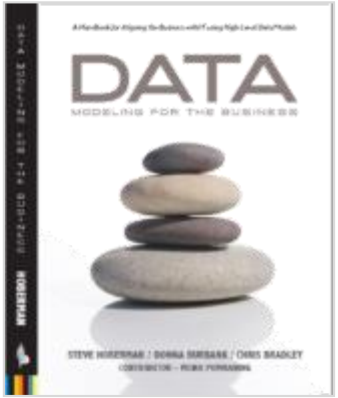
Donna Burbank

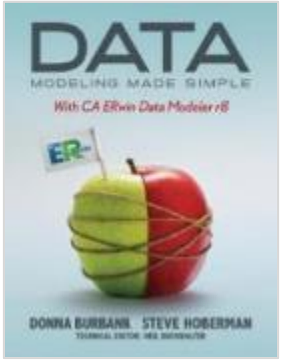
Donna is a recognised industry expert in information management with over 20 years of experience in data strategy, information management, data modeling, metadata management, and enterprise architecture. Her background is multi-faceted across consulting, product development, product management, brand strategy, marketing, and business leadership.

She is currently the Managing Director at Global Data Strategy, Ltd., an international information management consulting company that specialises in the alignment of business drivers with data-centric technology. In past roles, she has served in key brand strategy and product management roles at CA Technologies and Embarcadero Technologies for several of the leading data management products in the market.

As an active contributor to the data management community, she is a long time DAMA International member and is the President of the DAMA Rocky Mountain chapter. She was also on the review committee for the Object Management Group's Information Management Metamodel (IMM) and a member of the OMG's Finalization Taskforce for the Business Process Modeling Notation (BPMN).

She has worked with dozens of Fortune 500 companies worldwide in the Americas, Europe, Asia, and Africa and speaks regularly at industry conferences. She has co-authored two books: *Data Modeling for the Business* and *Data Modeling Made Simple with ERwin Data Modeler* and is a regular contributor to industry publications such as DATAVERSITY, EM360, & TDAN. She can be reached at donna.burbank@globaldatastrategy.com Donna is based in Boulder, Colorado, USA.





Follow on Twitter @donnaburbank

Agenda

What we'll cover today

- **Data Lakes & Big Data**
  - Big Data – A Technical & Cultural Paradigm Shift
  - Big Data in the Larger Information Management Landscape
- **Data Governance for the Data Lake**
  - **To Govern or Not to Govern:** Identifying which data assets it makes sense to control (and what to leave alone)
  - **Rollout & Value:** Delivering "quick wins" to the organization
  - **Rules of Engagement:** Identifying a practical framework & operating model for the Data Lake environment
  - **Stakeholder Engagement:** Working with various roles within the organization in a way that makes sense for each, from business users, to data architects, to data scientists, and more
- **Summary & Questions**

Big Data –A Technical & Cultural Paradigm Shift

Traditional Relational Technologies and “Big Data”: a Paradigm Shift

Traditional

- Top-Down, Hierarchical
- Design, then Implement
- “Passive”, Push technology
- “Manageable” volumes of information
- “Stable” rate of change
- Business Intelligence

Big Data

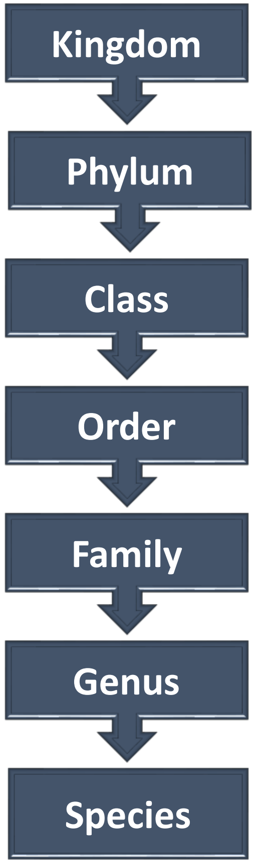
- Distributed, Democratic
- Discover and Analyze
- Collaborative, Interactive
- Massive volumes of information
- Rapid and Exponential rate of growth
- Statistical Analysis

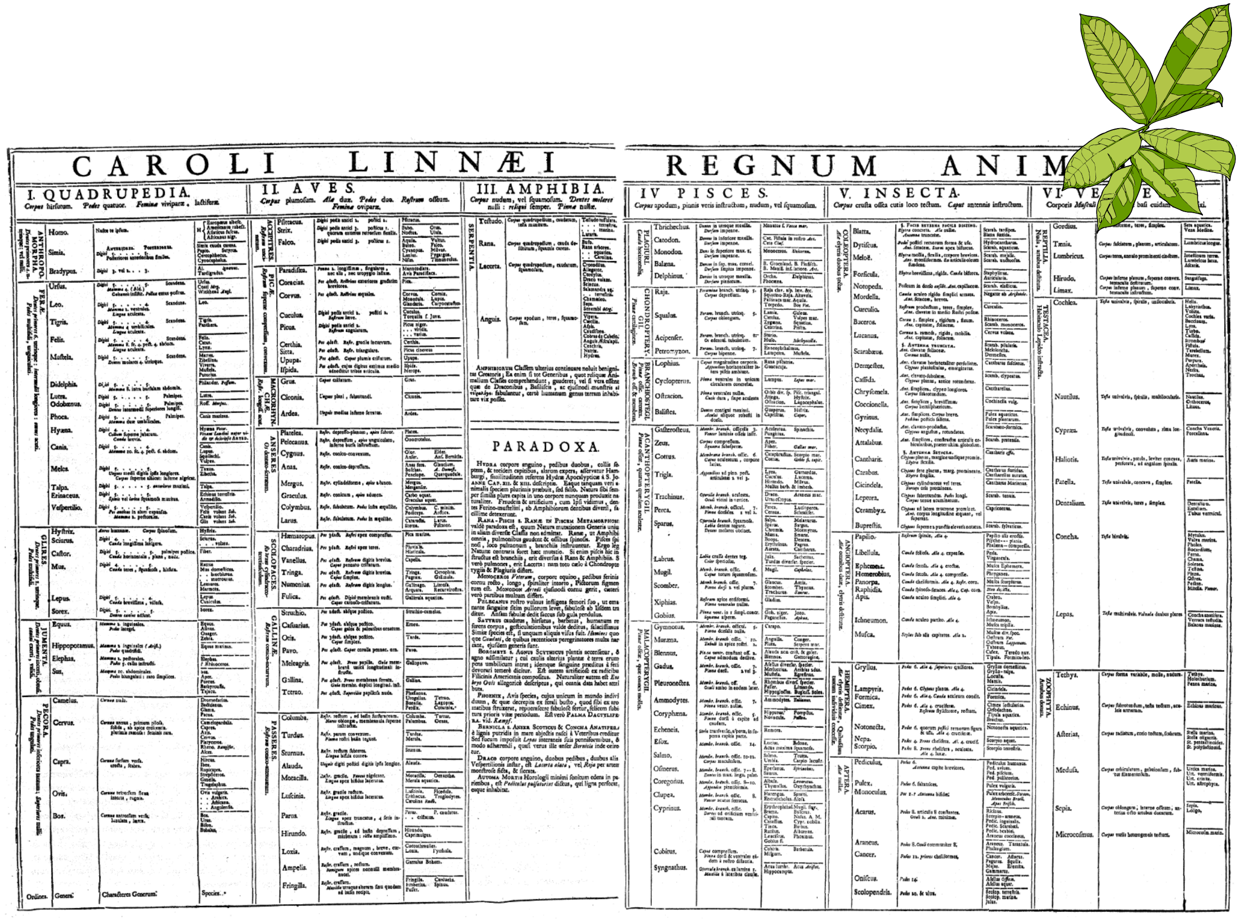
Design → ***Implement***

Discover → ***Analyze***

"Traditional" way of Looking at the World: *Hierarchies*

- Carolus Linnaeus in 1735 established a hierarchy/taxonomy for organizing and identifying biological systems.





CAROLI LINNÆI REGNUM ANIM

I. QUADRUPEDIA. *Corpus* hirſutum. *Pedes* quatuor. *Feminæ* viviparæ, lactiferæ.

II. AVES. *Corpus* plumoſum. *Alæ* duæ. *Pedes* duo. *Roſtrum* oſſeum. *Feminæ* oviparæ.

III. AMPHIBIA. *Corpus* nudum, vel ſquamoſum. *Dentes molares* nulli : *reliqui* ſemper. *Pinnæ* nullæ.

PARADOXA

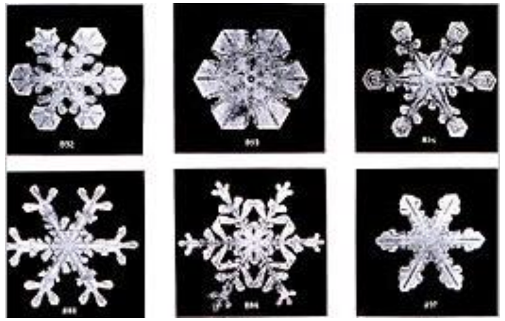
IV. PISCES. *Corpus* apodum, pinnis veris inſtructum, nudum, vel ſquamoſum.

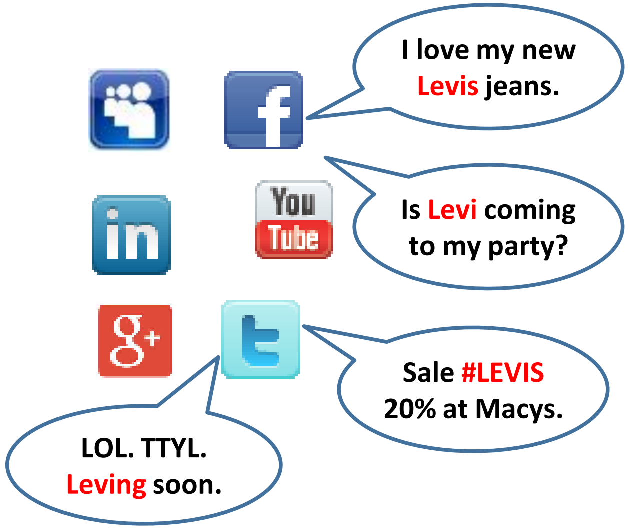
V. INSECTA. *Corpus* cruſta oſſea cutis loco tectum. *Caput* antennis inſtructum.

VI. VE

“New” Way of Looking at the World - *Emergence*

> In philosophy, systems theory, science, and art, emergence is the way complex systems and patterns arise out of a multiplicity of relatively simple interactions.
>
> - Wikipedia





I love my new Levis jeans.

Is Levi coming to my party?

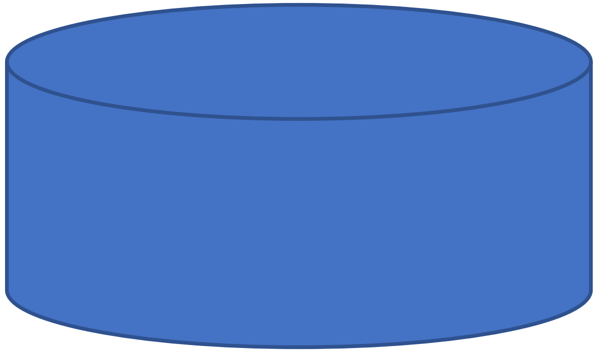
Sale #LEVIS 20% at Macys.

LOL. TTYL. Leving soon.

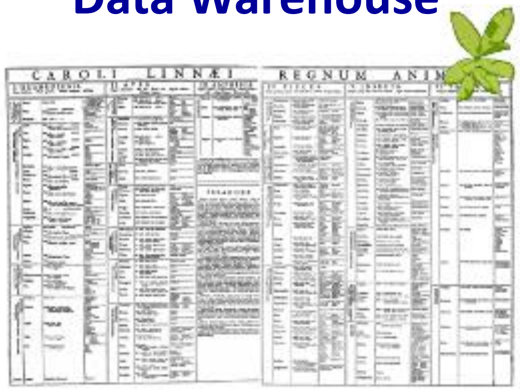
Data Warehouse vs. Data Lake

A **Data Warehouse** is a storage repository that holds current and historical data used for creating analytical reports. Data structures & requirements are pre-defined, and data is organized & stored according to these definitions.

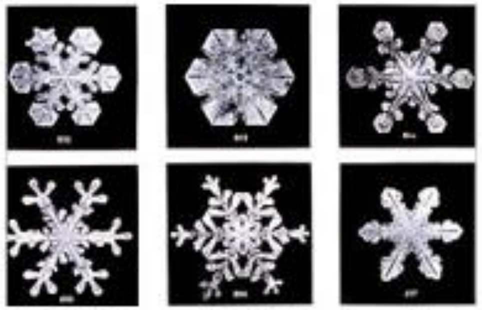
A **Data Lake** is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data. The data structure & requirements are not defined until the data is needed.



Data Warehouse

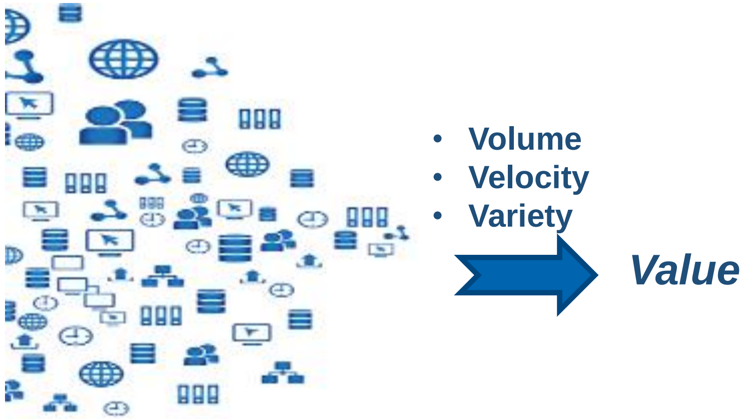

CAROLI LINNÆI
REGNUM ANIM

Data Lake



What is Big Data?

- Big Data is often characterised by the “3 Vs”:
  - **Volume:** Is there a high volume of data? (e.g. terabytes per day)
  - **Velocity**: Is data generated or changed at a rapid pace? (e.g. per second, sub-second)
  - **Variety:** Is data stored across multiple formats? (e.g. machine data, OSS data, log files)
- The ability to understand and manage these sources and integrate them into the larger Business Intelligence ecosystem can provide the ability to gain **valuable insights from data**.
  - **Social Media Sentiment Analysis** – e.g. What are customers saying about our products?
  - **Web Browsing Analytics** – Customer usage patterns
  - **Internet of Things (IoT) Analysis** – e.g. Sensor data, Machine log data
  - **Customer Support** – e.g. Call log analysis
- This ability leads to the “4th V” of Big Data – Value.
  - **Value:** Valuable insights gained from the ability to analyze and discover new patterns and trends from high-volume and/or cross-platform systems.



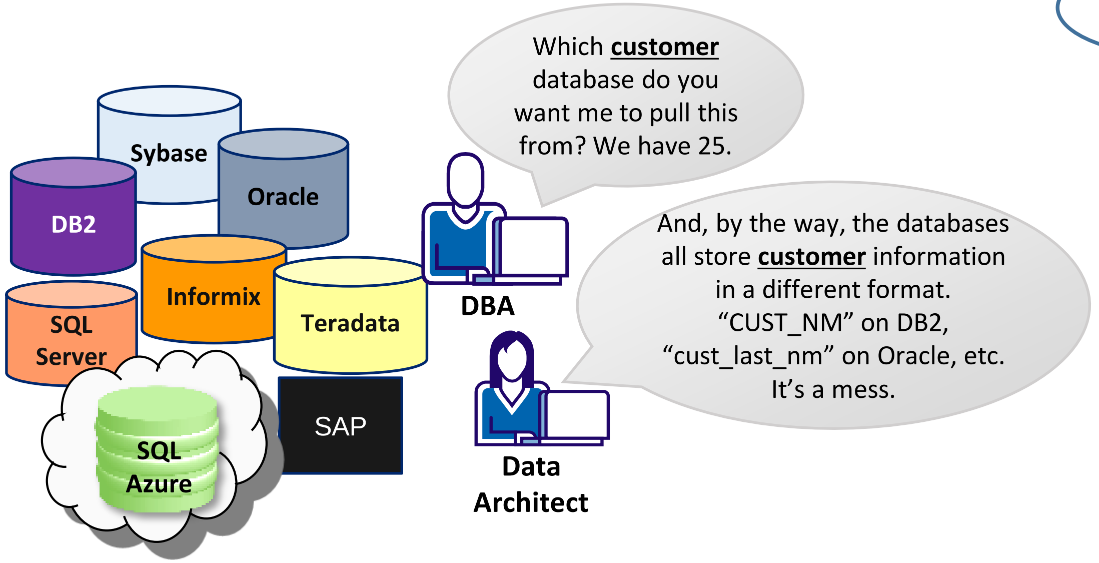
- Volume
- Velocity
- Variety

Value

The Business Case is Similar

Tell me what customers are saying about our product.

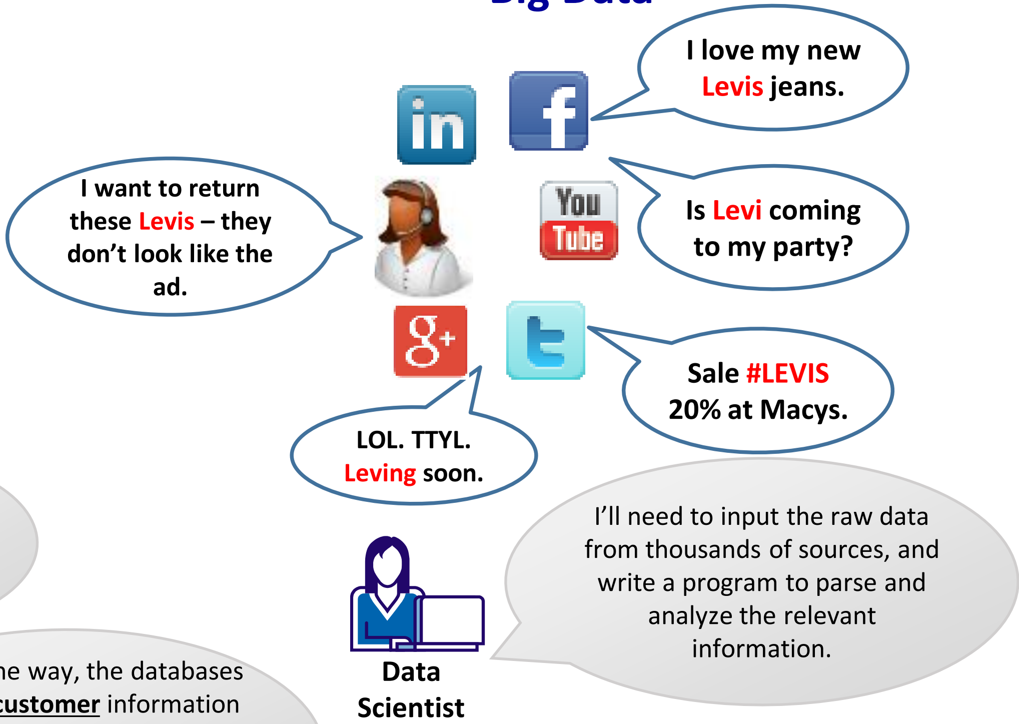
Traditional Databases



Which **customer** database do you want me to pull this from? We have 25.

And, by the way, the databases all store **customer** information in a different format. “CUST_NM” on DB2, “cust_last_nm” on Oracle, etc. It’s a mess.

Big Data



I love my new Levis jeans.

Is Levi coming to my party?

I want to return these Levis – they don’t look like the ad.

Sale #LEVIS 20% at Macys.

LOL. TTYL. Leving soon.

I’ll need to input the raw data from thousands of sources, and write a program to parse and analyze the relevant information.

The 5th "V" - Veracity

- Only through proper Governance, Data Quality Management, Metadata Management, etc., can organizations achieve the 5th "V" – Veracity.
  - **Veracity:** Trust in the accuracy, quality and content of the organizations' information assets.
- i.e. The hard work doesn't go away with Big Data

Data Science

Raw data used in Self-Service Analytics and BI environments is often so poor that **many data scientists and BI professionals spend an estimated 50 – 90% of their time cleaning and reformatting data** to make it fit for purpose.(4

Source: DataCenterJournal.com

Data Lakes

The absence of commonly understood and shared metadata and data definitions is cited as one of the main impediments to the success of Data Lakes.

Source: Radiant Advisors

Data Science

Correcting poor data quality is a Data Scientist's least favorite task, consuming on average 80% of their working day

Source: Forbes 2016

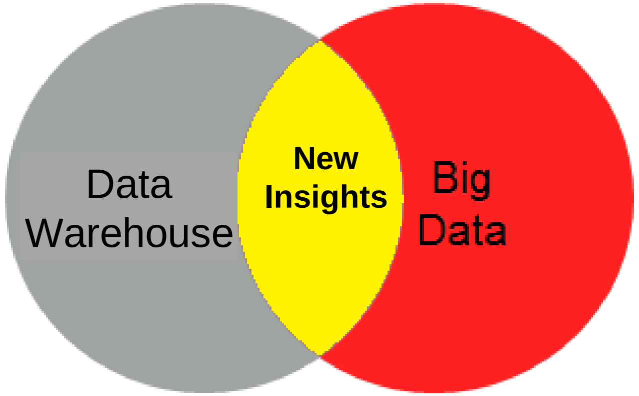
Digitization & Data Quality

71% of interviewees expect digitization to grow their business. But 70% say the biggest barrier is finding the right data; 62% cite inconsistent data

Source: Stibo Systems

Combining DW & Big Data Can Provide Valuable Information

- There are numerous ways to gain value from data
- Relational Database and Data Warehouse systems are one key source of value
  - Customer information
  - Product information
- Big Data can offer new insights from data
  - From new data sources (e.g. social media, IoT)
  - By correlating multiple new and existing data sources (e.g. network patterns & customer data)
- Integrating DW and Big Data can provide valuable new insights.
- Examples include:
  - Customer Experience Optimization
  - Churn Management
  - Products & Services Innovation



Big Data is Part of a Larger Enterprise Landscape

A Successful Data Strategy Requires Many Inter-related Disciplines

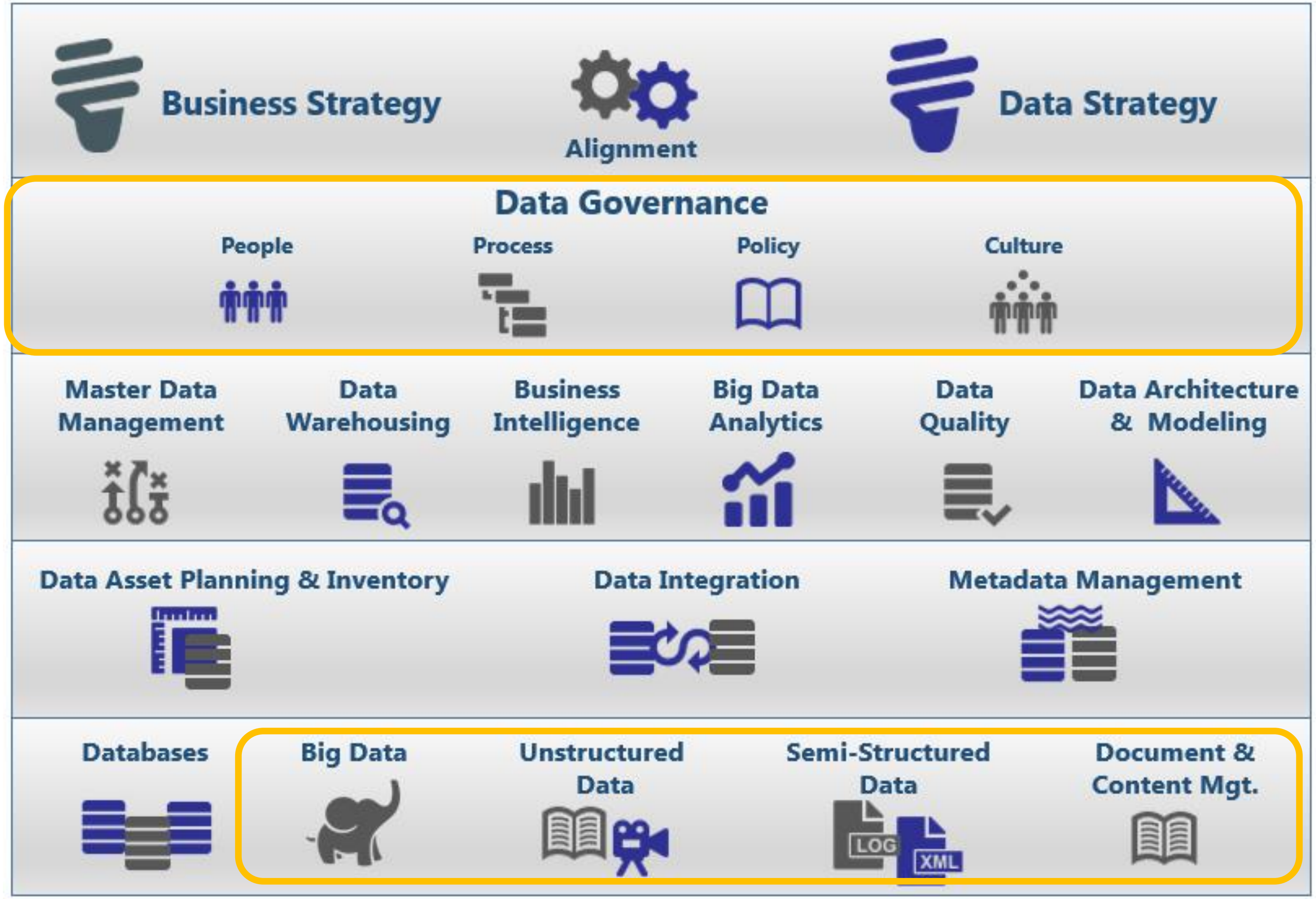
"Top-Down" alignment with business priorities

Managing the people, process, policies & culture around data

Leveraging & managing data for strategic advantage

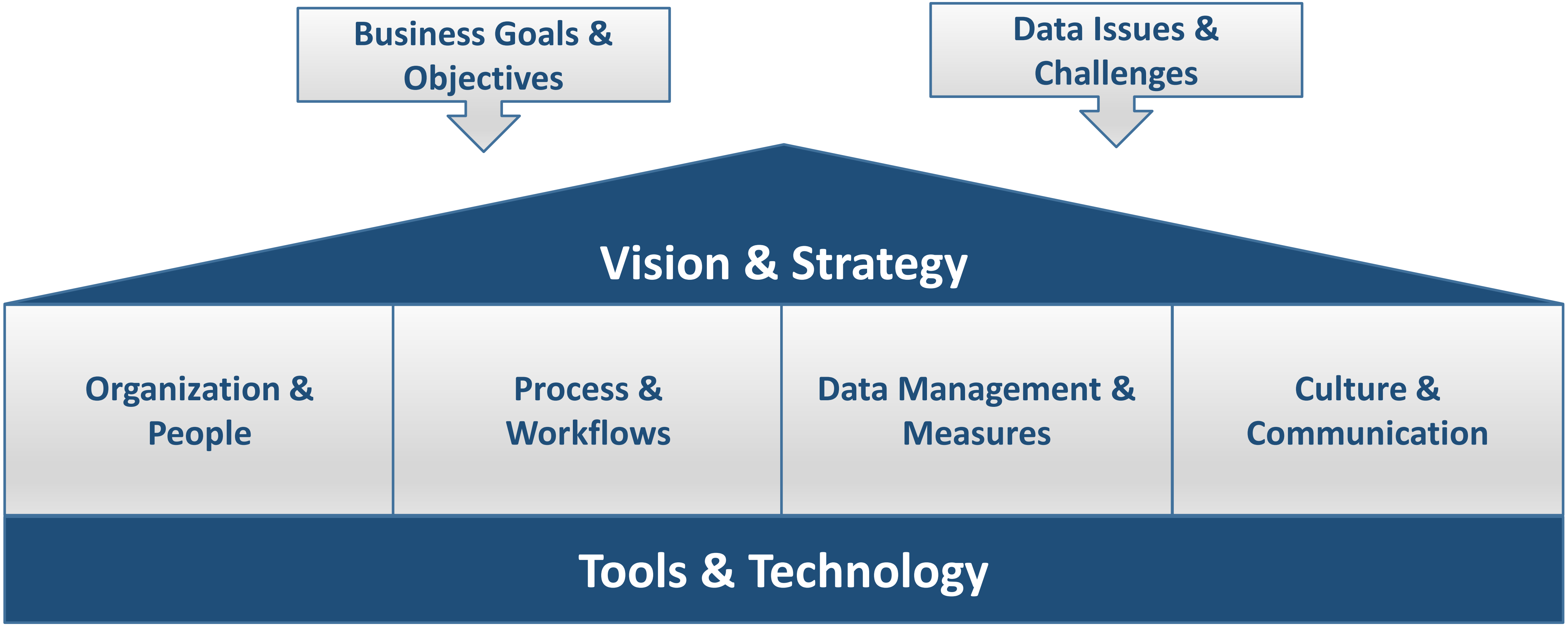
Coordinating & integrating disparate data sources

"Bottom-Up" management & inventory of data sources



Data Governance for the Data Lake

Applying a Structured Data Governance Framework



Business Goals & Objectives

Data Issues & Challenges

Vision & Strategy

Organization & People

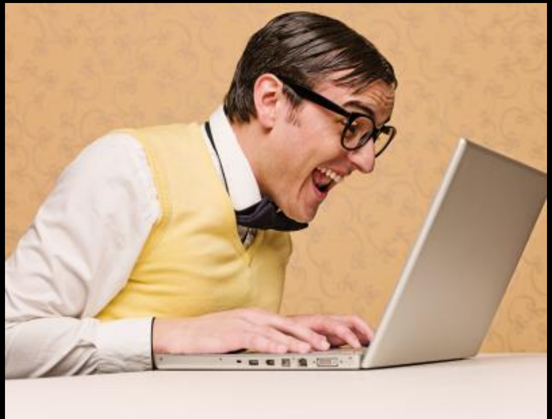
Process & Workflows

Data Management & Measures

Culture & Communication

Tools & Technology

DATA GOVERNANCE

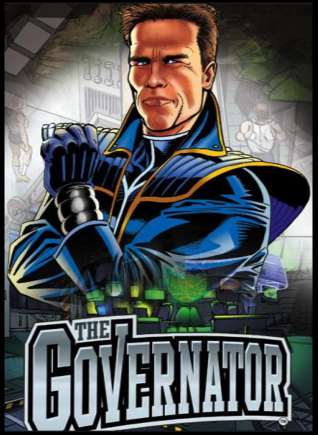


What my friends think I do

What my mom thinks I do

What society thinks do

What my coworkers think I do



What I think I do

Profit
Quality
Reliability
Price

What I actually do

Driving the Success of the Business

How can we Transform our Business through Data?

Business Optimization
Becoming a *Data-Driven Company*

- Making the Business More Efficient
  - Better Marketing Campaigns
    - Higher quality customer data, 360 view of customer, competitive info, etc.
  - Better Products
    - Data-Driven product development, Customer usage monitoring, etc.
  - Better Customer Support
    - Linking customer data with support logs, network outages, etc.
  - Lower Costs
    - More efficient supply chain
    - Reduced redundancies & manual effort

How do we do what we do *better*?

Business Transformation
Becoming a *Data Company*

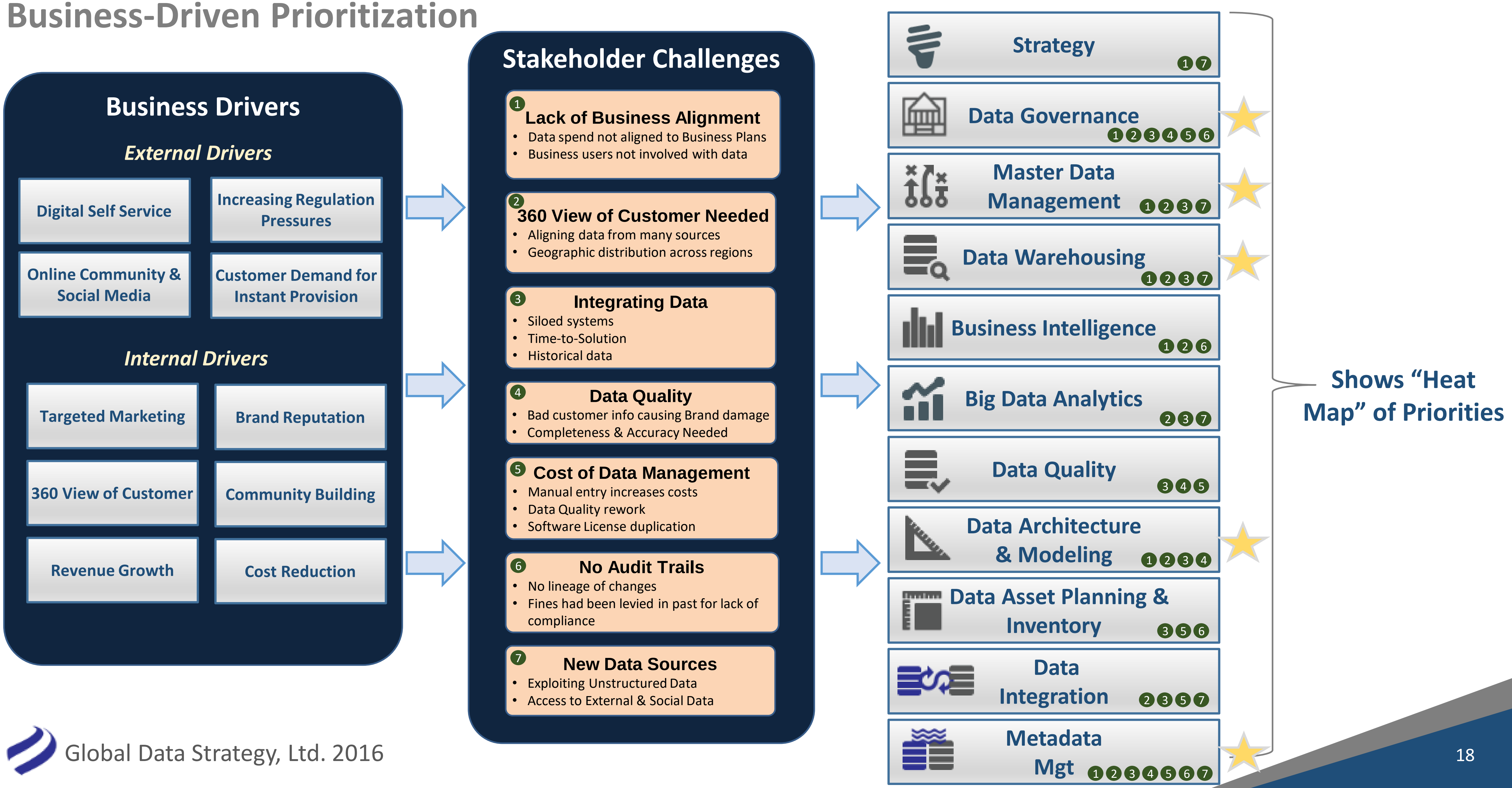
- Changing the Business Model via Data – data becomes the product
  - Monetization of Information: examples across multiple industries including:
    - *Telecom:* location information, usage & search data, etc.
    - *Retail:* Click-stream data, purchasing patterns
    - *Social Media:* social & family connections, purchasing trends & recommendations, etc.
    - *Energy:* Sensor data, consumer usage patterns, smart metering, etc.

How do we do something *different*?

Data Lakes can support both of these paradigms.

Mapping Business Drivers to Data Management Capabilities

Business-Driven Prioritization



Business Drivers

External Drivers

- Digital Self Service
- Increasing Regulation Pressures
- Online Community & Social Media
- Customer Demand for Instant Provision

Internal Drivers

- Targeted Marketing
- Brand Reputation
- 360 View of Customer
- Community Building
- Revenue Growth
- Cost Reduction

Stakeholder Challenges

1 **Lack of Business Alignment**
- Data spend not aligned to Business Plans
- Business users not involved with data

2 **360 View of Customer Needed**
- Aligning data from many sources
- Geographic distribution across regions

3 **Integrating Data**
- Siloed systems
- Time-to-Solution
- Historical data

4 **Data Quality**
- Bad customer info causing Brand damage
- Completeness & Accuracy Needed

5 **Cost of Data Management**
- Manual entry increases costs
- Data Quality rework
- Software License duplication

6 **No Audit Trails**
- No lineage of changes
- Fines had been levied in past for lack of compliance

7 **New Data Sources**
- Exploiting Unstructured Data
- Access to External & Social Data

Capability	Challenges	
Strategy	1 7	
Data Governance	1 2 3 4 5 6	★
Master Data Management	1 2 3 7	★
Data Warehousing	1 2 3 7	★
Business Intelligence	1 2 6	
Big Data Analytics	2 3 7	
Data Quality	3 4 5	
Data Architecture & Modeling	1 2 3 4	★
Data Asset Planning & Inventory	3 5 6	
Data Integration	2 3 5 7	
Metadata Mgt	1 2 3 4 5 6 7	★

Shows "Heat Map" of Priorities

Identify What Data Needs to Be Governed

And What to Leave Alone

Why?

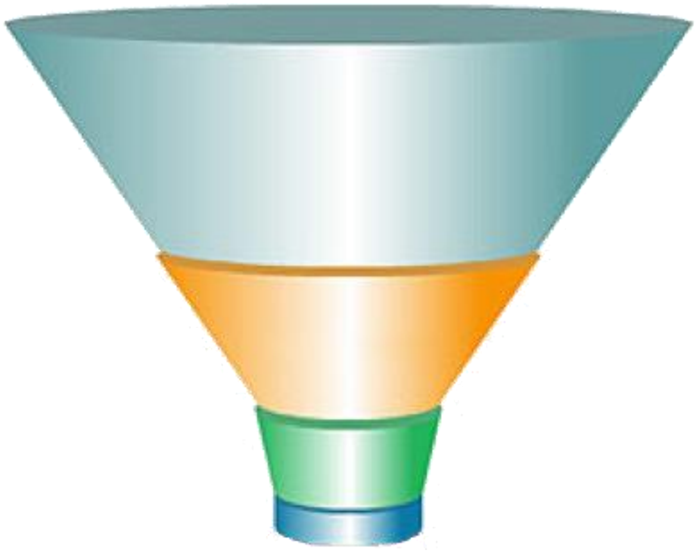
Identify Key Business Driver

Filter Data Elements Aligned with Business Driver

Focus Governance Efforts on Key Data

What?

Launch of New Product – Marketing Campaign requires better customer information



Customer

Product

Region

Partner

Vendor

How?

Exploratory Analytics & Discovery

Lightly governed

Social Media Sentiment Analysis

Structured Warehouse for Financial Reporting

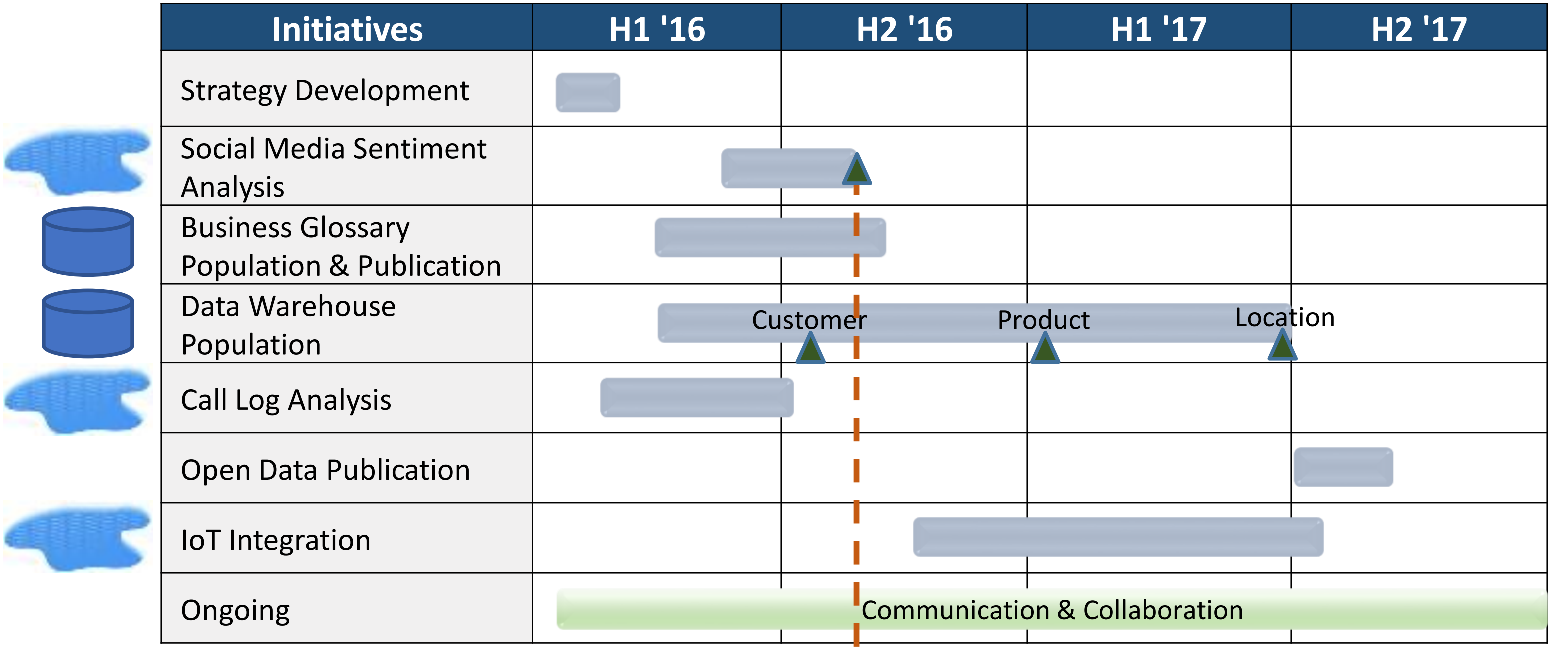
Highly governed

Financial Reporting

Defining an Actionable Roadmap

Maximize the Benefit to the Organization

- Develop a detailed roadmap that is both actionable and realistic
  - Show quick-wins, while building to a longer-term goal
  - Include both Data Lake exploration & Data Warehouse reporting
  - Focus on projects that benefit multiple stakeholders
- You can’t manage & govern everything – pick your priorities.



Initiatives	H1 '16	H2 '16	H1 '17	H2 '17
Strategy Development	▬			
Social Media Sentiment Analysis	▬	▬ ▲		
Business Glossary Population & Publication	▬	▬		
Data Warehouse Population	▬	▬ Customer ▲	▬ Product ▲ Location ▲	
Call Log Analysis	▬	▬		
Open Data Publication				▬
IoT Integration		▬	▬	▬
Ongoing	▬	Communication & Collaboration	▬	▬

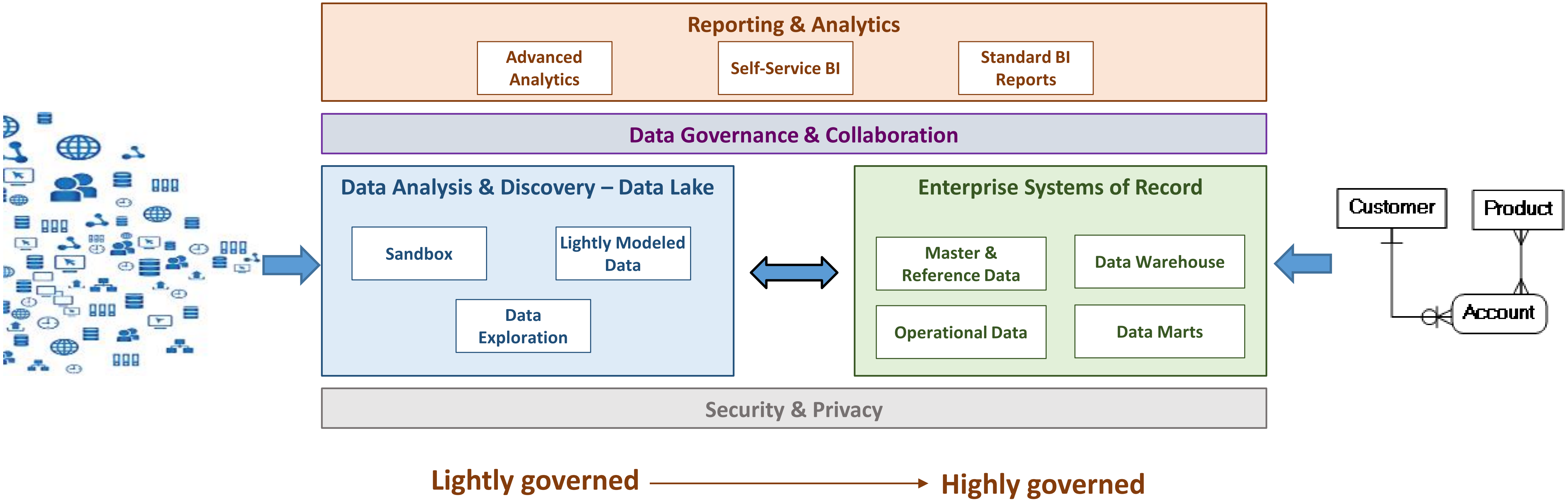
Integrated Customer View

Marketing **Customer Support**

Sales **Executive Team**

Integrating the Data Lake & Traditional Data Sources

- The Data Lake has a different architecture & purpose than traditional data sources such as data warehouses.
- But the two environments can co-exist to share relevant information.
- Data Governance is different for each environment.



Reporting & Analytics

Advanced Analytics

Self-Service BI

Standard BI Reports

Data Governance & Collaboration

Data Analysis & Discovery – Data Lake

Sandbox

Lightly Modeled Data

Data Exploration

Enterprise Systems of Record

Master & Reference Data

Data Warehouse

Operational Data

Data Marts

Security & Privacy

Lightly governed → Highly governed

Roles & Culture

DBAs

- Analytical
- Structured
- Project & Task focused
- Cautious – identifies risks
- “Just let me code!”

Data Scientist

- Looks for opportunities
- Likes to explore
- Seen as “modern”
- Seen as “hip” & “sexy”

Data Architects

- Analytical
- Structured
- “Big Picture” focused
- Can be considered “old school”
- “Let me tell you about my data model!”

Business Executive

- Results-Oriented
- Optimistic – Identifies opportunities
- “Big Picture” focused
- “I’m busy.”
- “What’s the business opportunity?”

Big Data Vendors

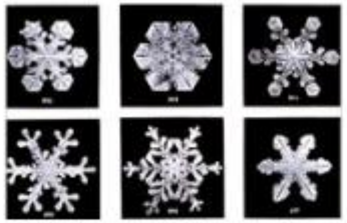
- It’s magic!
- It’s easy!
- No modeling needed!

Organizational Siloes

- Too often, there are organizational & cultural silos that limit the sharing between the Data Lake and Data Warehouse

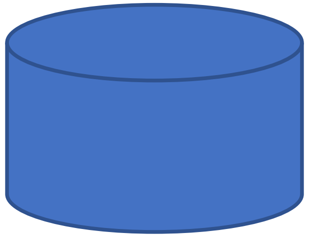
Data Lake & Data Scientist

- Exploratory projects
- Quick wins
- Little documentation & governance



Data Warehouse & Data Architects

- Enterprise reporting
- Long-term projects
- Data Standards
- Metadata & Governance

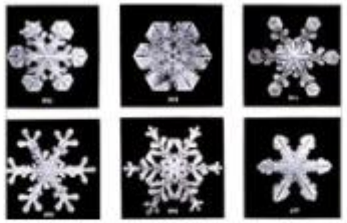


Breaking Down Organizational Siloes

- Good Communication & Governance help break down siloes and encourage information sharing.

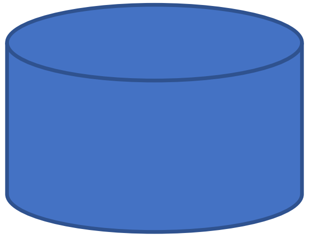
Data Lake & Data Scientist

- Exploratory projects
- Quick wins
- Little documentation



Data Warehouse & Data Architects

- Enterprise reporting
- Long term project
- Data standards & documentation



New Operating Model:
Interactions Between New & Existing Roles

Existing Roles

Privacy
Analyst

Data Architect

Data Steward

ETL Developer

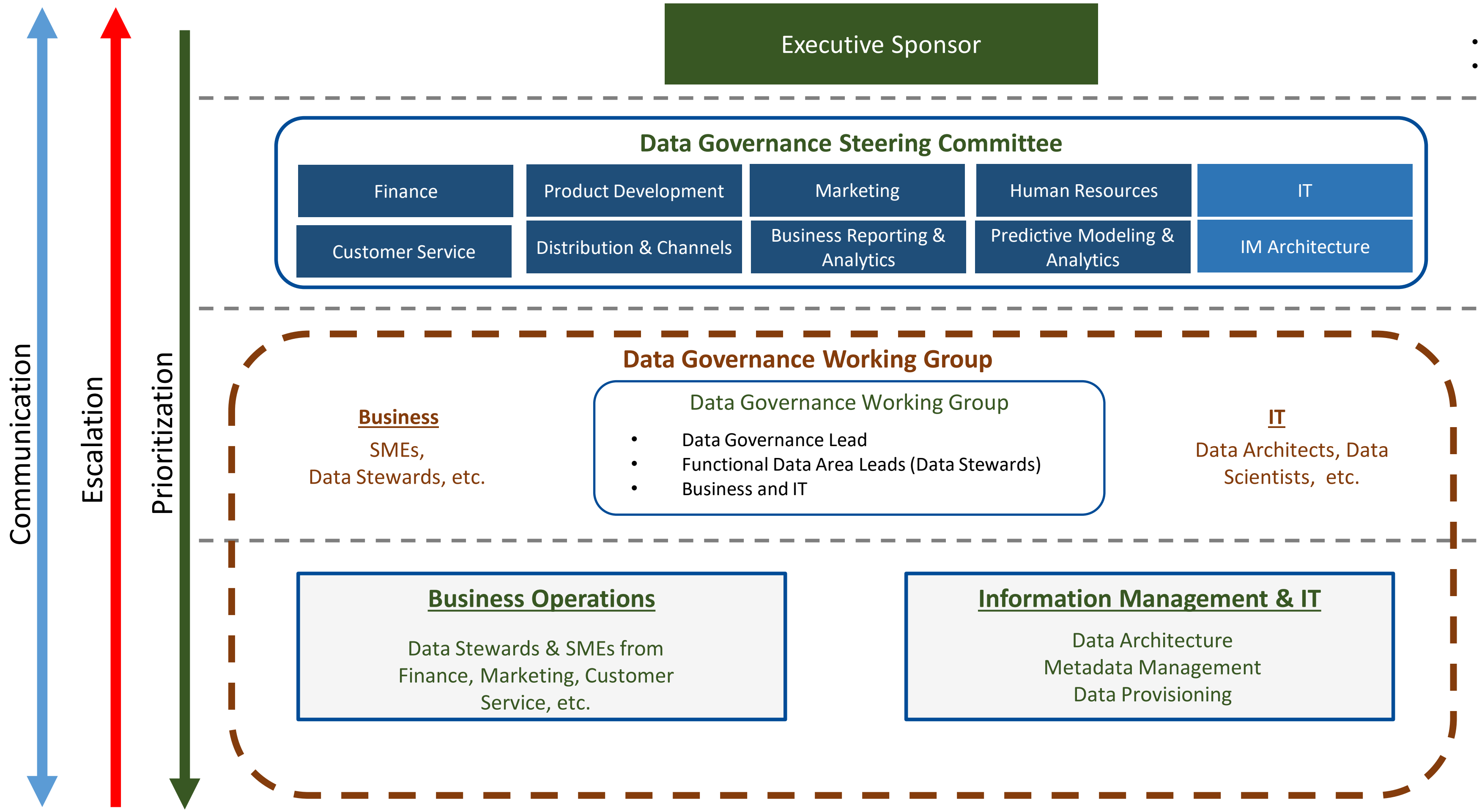
Alignment

New Roles

Data Scientist

Hadoop
Administrator

Sample Data Governance Operating Model



Communication

Escalation

Prioritization

Executive Sponsor

Data Governance Steering Committee

Finance

Product Development

Marketing

Human Resources

IT

Customer Service

Distribution & Channels

Business Reporting & Analytics

Predictive Modeling & Analytics

IM Architecture

Data Governance Working Group

Business

SMEs,
Data Stewards, etc.

Data Governance Working Group

- Data Governance Lead
- Functional Data Area Leads (Data Stewards)
- Business and IT

IT

Data Architects, Data Scientists, etc.

Business Operations

Data Stewards & SMEs from Finance, Marketing, Customer Service, etc.

Information Management & IT

Data Architecture
Metadata Management
Data Provisioning

Executive Level

- Executive Support & Direction
- Budget & resource approval

Strategic Level

- Strategic direction
- Prioritization
- Both Business & IT
- Issue escalation

Tactical Level

- Builds & manages policies, procedures & standards
- Data Definition
- Works with Stewards & SMEs to enforce at a tactical level

Execution

- Executes data management activities (data publication, integration, etc.)
- Both Business & IT

Data Governance Processes & Workflows

Customize for the environment

- Data Governance Processes & Workflows are different for Data Lakes & Data Warehouses
  - **Data Lake & Big Data Exploration**
    - Light governance
    - “Tell me what you’re working on”
    - “Post some sample code”
  - **Data Warehousing**
    - Heavily governed
    - Structured data models, metadata lineage, etc.
- Some things remain the same
  - **Data Stewardship**
    - Who is the expert for Product data?
    - Who wrote this code?
  - **Data Definitions, Standard Metrics & Business Glossary**
    - What’s the definition for “Total Earned Revenue”?
    - Is a customer considered active if their payment is over 30 days overdue?

Data Management & Measures

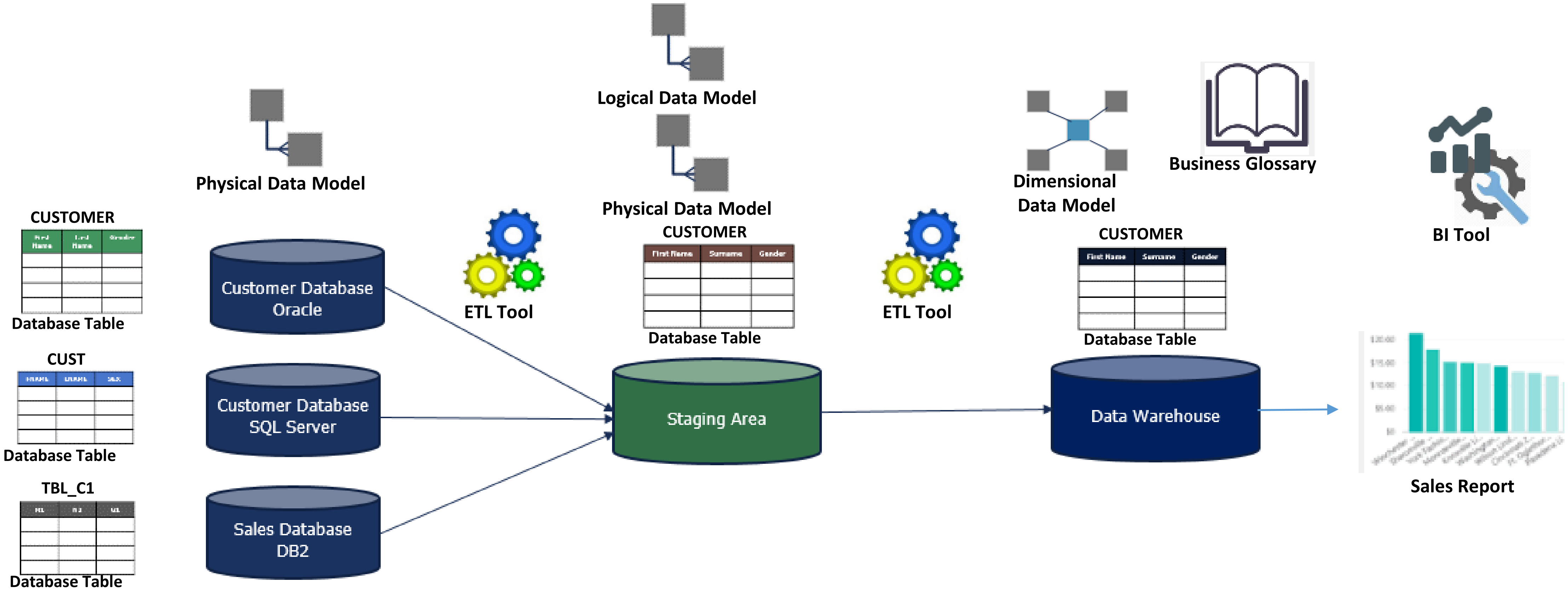
Suit the Method to the Environment

- Metadata Management & Governance is different with a Data Lake vs. a Data Warehouse
- Data Lake
  - Metadata is not non-existent! Exploration & discovery doesn’t mean lack of any documentation
  - Consider other exploratory and rapidly changing environments – e.g. Open Source Development, Open Data, etc.
- Data Warehouse
  - More Traditional metadata management applies
  - Data Lineage
  - Data Models
- Business Metadata is a constant
  - What does this term mean? (business glossary)
  - Who is the owner or steward of the data? Who can I go to to ask a question?

Data Warehousing Metadata & Lineage

Robust Documentation & Lineage

- Data warehouses are typically governed by a robust and well-documented data lineage.



CUSTOMER

Database Table

CUST

Database Table

TBL_C1

Database Table

Physical Data Model

Customer Database Oracle

Customer Database SQL Server

Sales Database DB2

ETL Tool

Logical Data Model

Physical Data Model

CUSTOMER

First Name	Surname	Gender

Database Table

Staging Area

ETL Tool

Dimensional Data Model

Business Glossary

CUSTOMER

First Name	Surname	Gender

Database Table

Data Warehouse

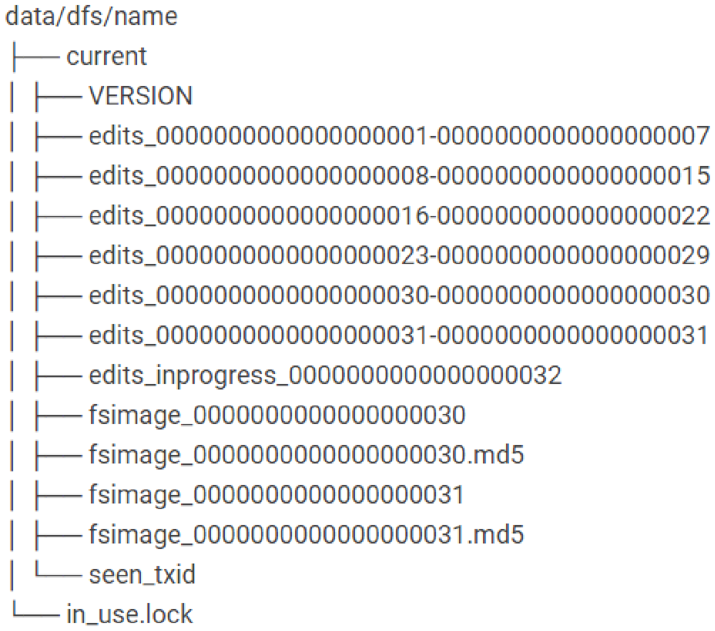
BI Tool

Sales Report

Big Data Platform Metadata

Weaker Metadata & Lineage

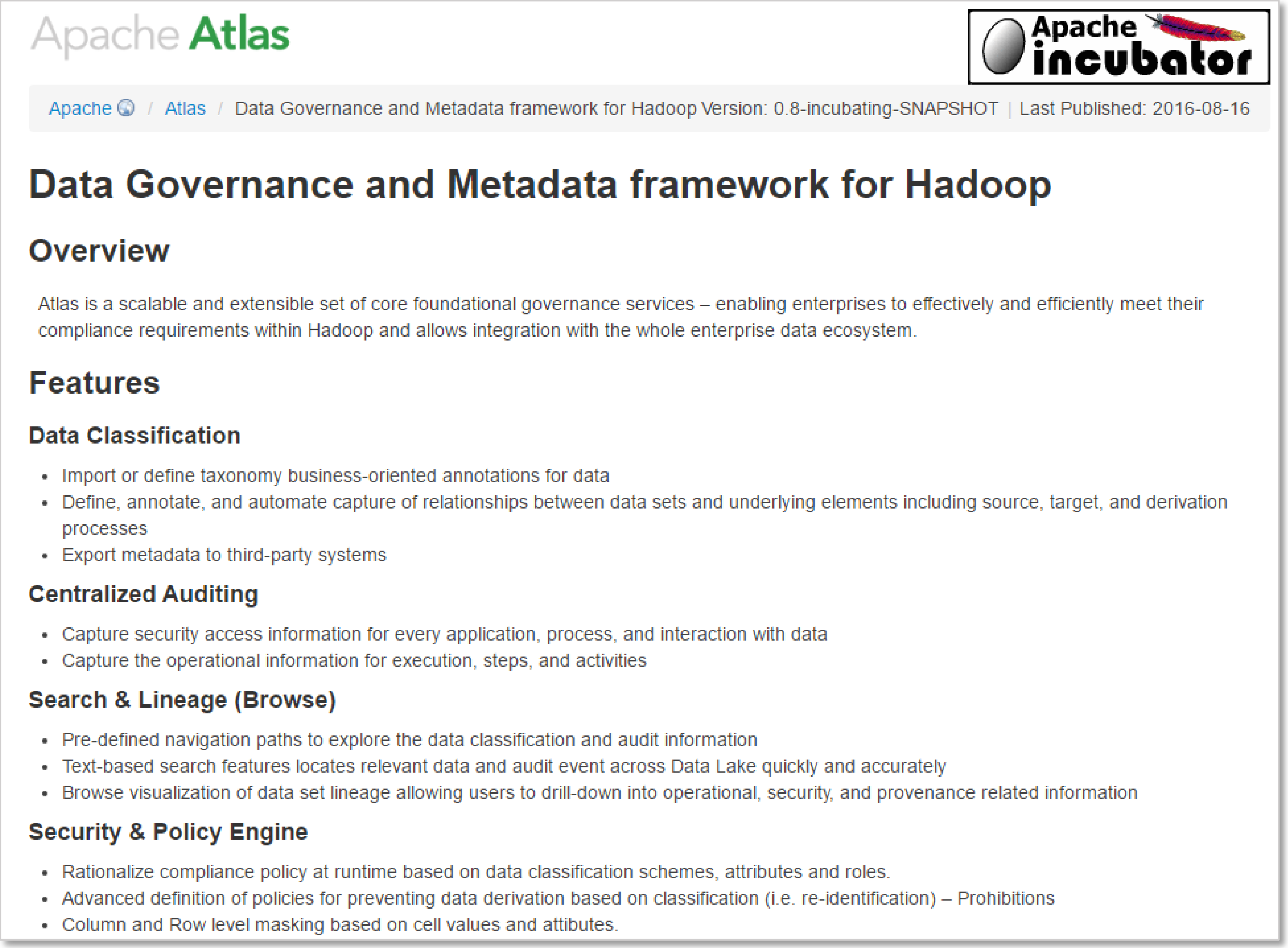
- Big Data platforms (e.g. Hadoop-based) are typically based on system of files (HDFS)
- As a result, the detailed structure that is found in a relational database platform does not exist
- Metadata still exists for these platforms.
  - Technical Metadata
    - Tree structure of HDFS directories
    - Directory and file attributes (ownership, permissions, quotas, replication factor, etc.)
    - Metadata about logical data sets (e.g. format, statistics, etc.)
    - Data ingest & transformation lineage
  - Business Metadata
    - Description of file
    - Tags
  - There are components that allow you to add structure within the Hadoop ecosystem (e.g. Hive)



```
data/dfs/name
├── current
│ ├── VERSION
│ ├── edits_0000000000000000001-0000000000000000007
│ ├── edits_0000000000000000008-0000000000000000015
│ ├── edits_0000000000000000016-0000000000000000022
│ ├── edits_0000000000000000023-0000000000000000029
│ ├── edits_0000000000000000030-0000000000000000030
│ ├── edits_0000000000000000031-0000000000000000031
│ ├── edits_inprogress_0000000000000000032
│ ├── fsimage_0000000000000000030
│ ├── fsimage_0000000000000000030.md5
│ ├── fsimage_0000000000000000031
│ ├── fsimage_0000000000000000031.md5
│ └── seen_txid
└── in_use.lock
```

The Industry is Advancing

- There is an Apache incubator project to address Data Governance & Metadata framework for Hadoop.



Apache Atlas

Apache / Atlas / Data Governance and Metadata framework for Hadoop Version: 0.8-incubating-SNAPSHOT | Last Published: 2016-08-16

Data Governance and Metadata framework for Hadoop

Overview

Atlas is a scalable and extensible set of core foundational governance services – enabling enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the whole enterprise data ecosystem.

Features

Data Classification

- Import or define taxonomy business-oriented annotations for data
- Define, annotate, and automate capture of relationships between data sets and underlying elements including source, target, and derivation processes
- Export metadata to third-party systems

Centralized Auditing

- Capture security access information for every application, process, and interaction with data
- Capture the operational information for execution, steps, and activities

Search & Lineage (Browse)

- Pre-defined navigation paths to explore the data classification and audit information
- Text-based search features locates relevant data and audit event across Data Lake quickly and accurately
- Browse visualization of data set lineage allowing users to drill-down into operational, security, and provenance related information

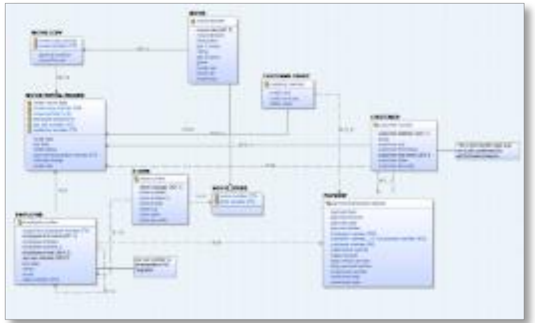
Security & Policy Engine

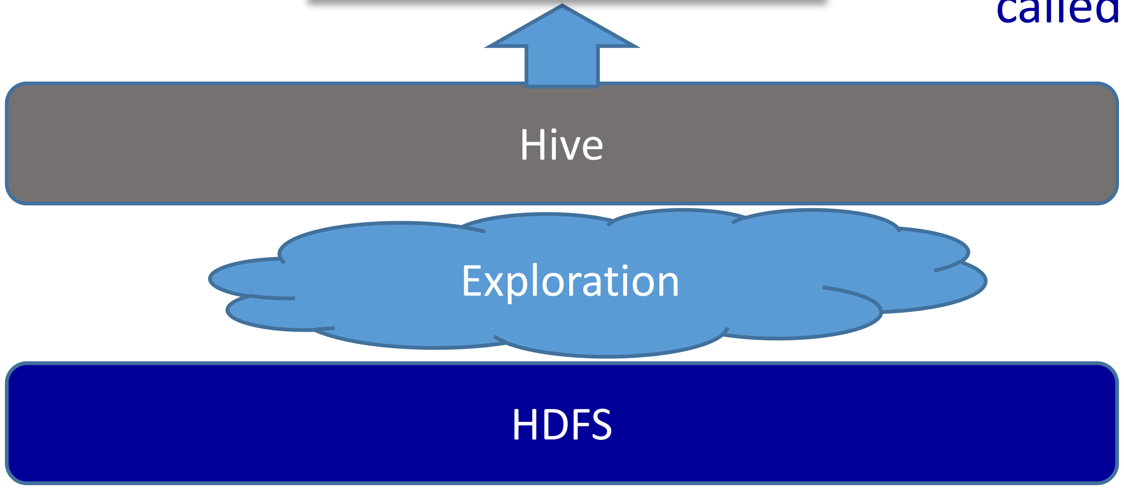
- Rationalize compliance policy at runtime based on data classification schemes, attributes and roles.
- Advanced definition of policies for preventing data derivation based on classification (i.e. re-identification) – Prohibitions
- Column and Row level masking based on cell values and attibutes.

Data Lake Big Data Model - “Schema on Read”

- With the Big Data and NoSQL paradigm, “Schema-on-Read” means you do not need to know how you will use your data when you are storing it.

- You do need to know how you will use your data when you are using it and model accordingly.
  - i.e. it’s not magic.
  - For example, you may first place the data on HDFS in files, then apply a table structure in Hive.
- Apache Hive provides a mechanism to project structure onto the data in Hadoop and to query that data using a SQL-like language called HiveQL (HQL).





Hive

Exploration

HDFS

Table Structures

Create table …

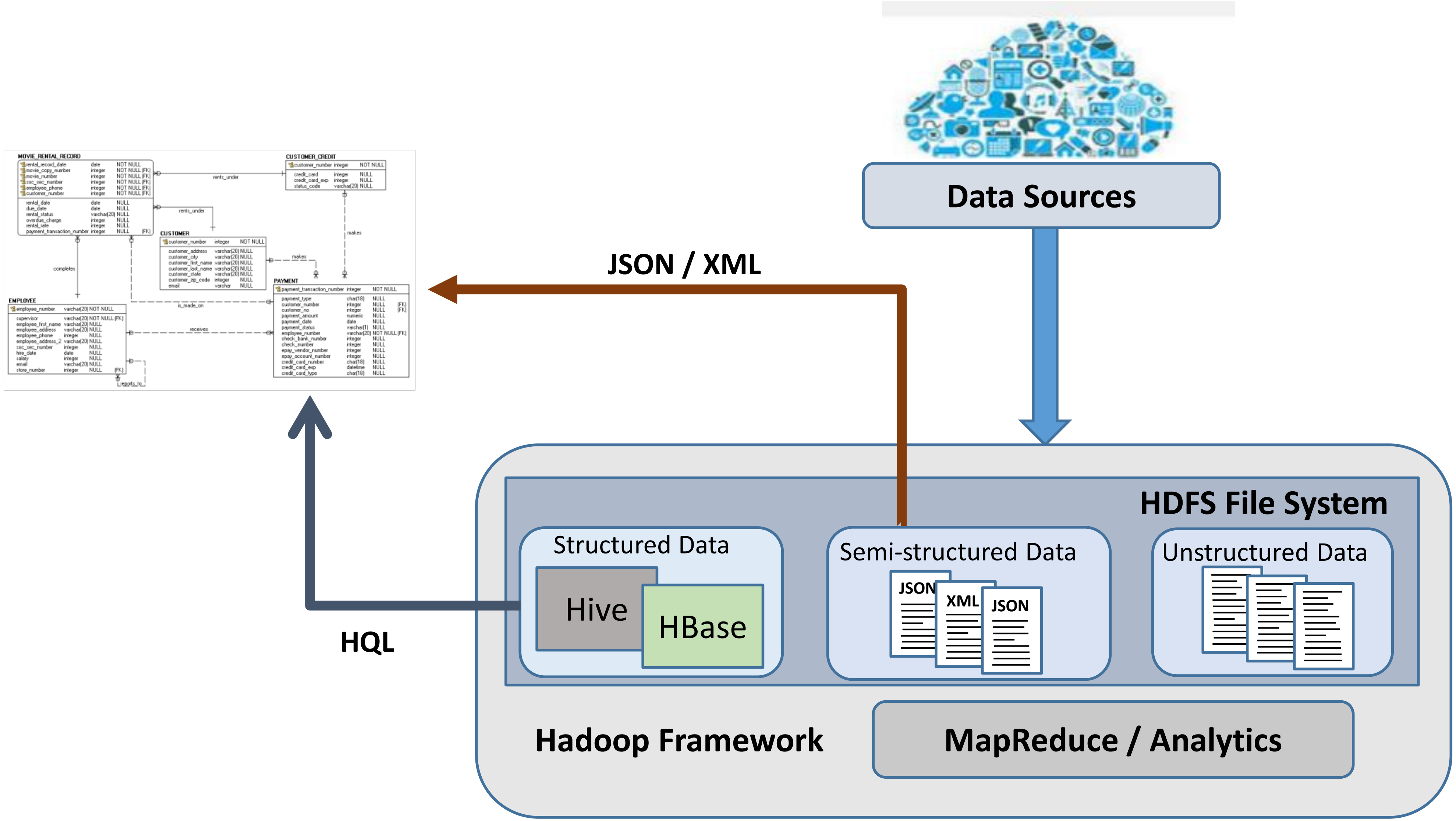
Analysis

Analyze & understand the data. Build a data structure to suite your needs.

File system

hdfs dfs -put /local/path/userdump /hdfs/path/data/users

Data Modeling in the Big Data Ecosystem



MOVIE_RENTAL_RECORD

rental_record_date	date	NOT NULL	
movie_copy_number	integer	NOT NULL	(FK)
movie_number	integer	NOT NULL	(FK)
soc_sec_number	integer	NOT NULL	(FK)
employee_phone	integer	NOT NULL	(FK)
customer_number	integer	NOT NULL	(FK)
rental_date	date	NULL	
due_date	date	NULL	
rental_status	varchar(20)	NULL	
overdue_charge	integer	NULL	
rental_rate	integer	NULL	
payment_transaction_number	integer	NULL	(FK)

CUSTOMER_CREDIT

customer_number	integer	NOT NULL
credit_card	integer	NULL
credit_card_exp	integer	NULL
status_code	varchar(20)	NULL

CUSTOMER

customer_number	integer	NOT NULL
customer_address	varchar(20)	NULL
customer_city	varchar(20)	NULL
customer_first_name	varchar(20)	NULL
customer_last_name	varchar(20)	NULL
customer_state	varchar(20)	NULL
customer_zip_code	integer	NULL
email	varchar	NULL

EMPLOYEE

employee_number	varchar(20)	NOT NULL	
supervisor	varchar(20)	NOT NULL	(FK)
employee_first_name	varchar(20)	NULL	
employee_address	varchar(20)	NULL	
employee_phone	integer	NULL	
employee_address_2	varchar(20)	NULL	
soc_sec_number	integer	NULL	
hire_date	date	NULL	
salary	integer	NULL	
email	varchar(20)	NULL	
store_number	integer	NULL	(FK)

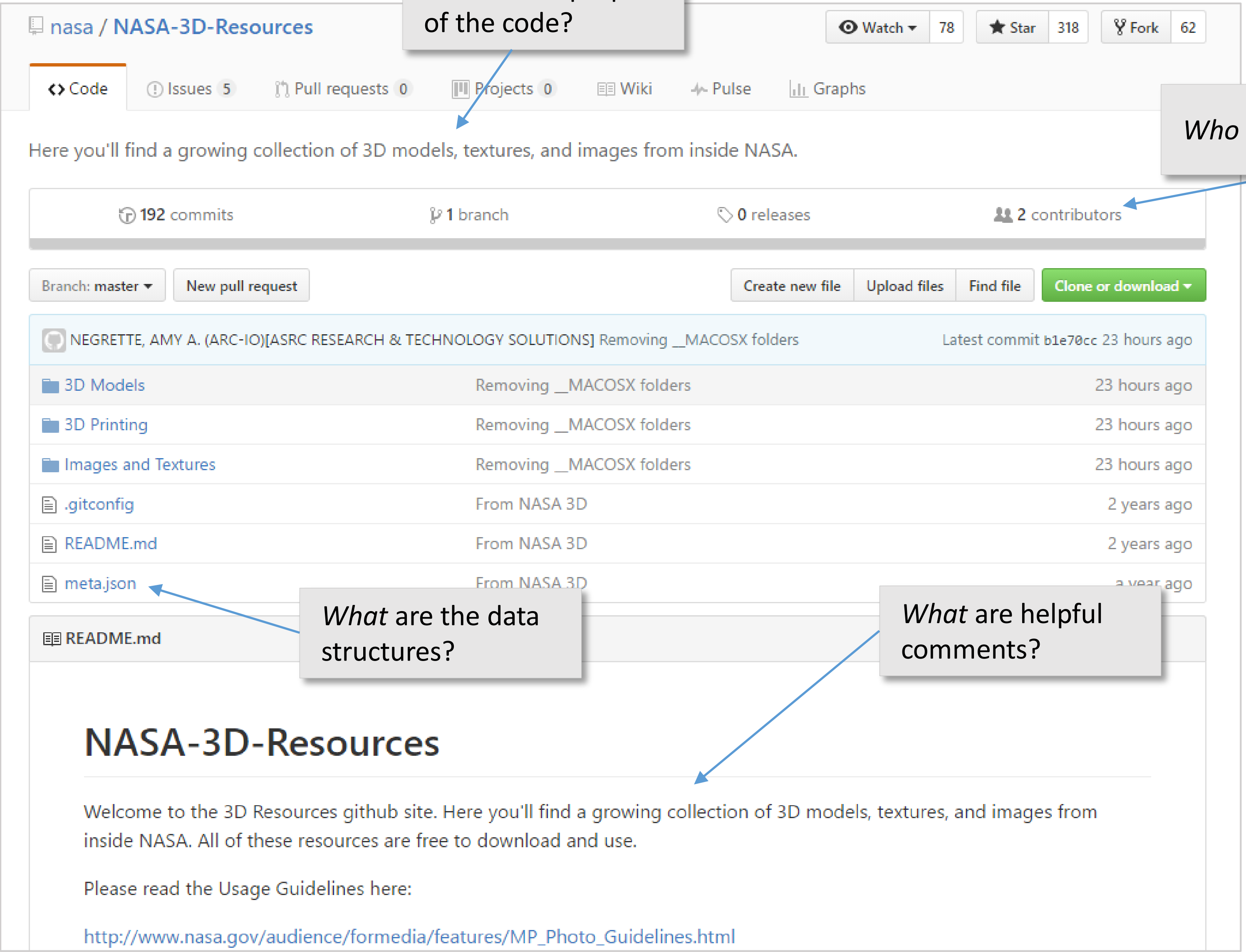
PAYMENT

payment_transaction_number	integer	NOT NULL	
payment_type	char(18)	NULL	
customer_number	integer	NULL	(FK)
customer_no	integer	NULL	(FK)
payment_amount	numeric	NULL	
payment_date	date	NULL	
payment_status	varchar(1)	NULL	
employee_number	varchar(20)	NOT NULL	(FK)
check_bank_number	integer	NULL	
check_number	integer	NULL	
epay_vendor_number	integer	NULL	
epay_account_number	integer	NULL	
credit_card_number	char(18)	NULL	
credit_card_exp	datetime	NULL	
credit_card_type	char(18)	NULL	

GitHub Metadata

Open Source Development

- Data Lake exploration typically is code-driven with little formal data structure.
    - In the Open Source development, environment, metadata still exists.
    - Just enough information for another developer to be able to re-use the code.
    - Similar documentation can be provided for Data Lake exploration & associated data science models & code.



nasa / NASA-3D-Resources

Here you'll find a growing collection of 3D models, textures, and images from inside NASA.

192 commits | 1 branch | 0 releases | 2 contributors

NEGRETTE, AMY A. (ARC-IO)[ASRC RESEARCH & TECHNOLOGY SOLUTIONS] Removing __MACOSX folders   Latest commit b1e70cc 23 hours ago

3D Models	Removing __MACOSX folders	23 hours ago
3D Printing	Removing __MACOSX folders	23 hours ago
Images and Textures	Removing __MACOSX folders	23 hours ago
.gitconfig	From NASA 3D	2 years ago
README.md	From NASA 3D	2 years ago
meta.json	From NASA 3D	a year ago

README.md

NASA-3D-Resources

Welcome to the 3D Resources github site. Here you'll find a growing collection of 3D models, textures, and images from inside NASA. All of these resources are free to download and use.

Please read the Usage Guidelines here:

http://www.nasa.gov/audience/formedia/features/MP_Photo_Guidelines.html

What is the purpose of the code?

Who published it?

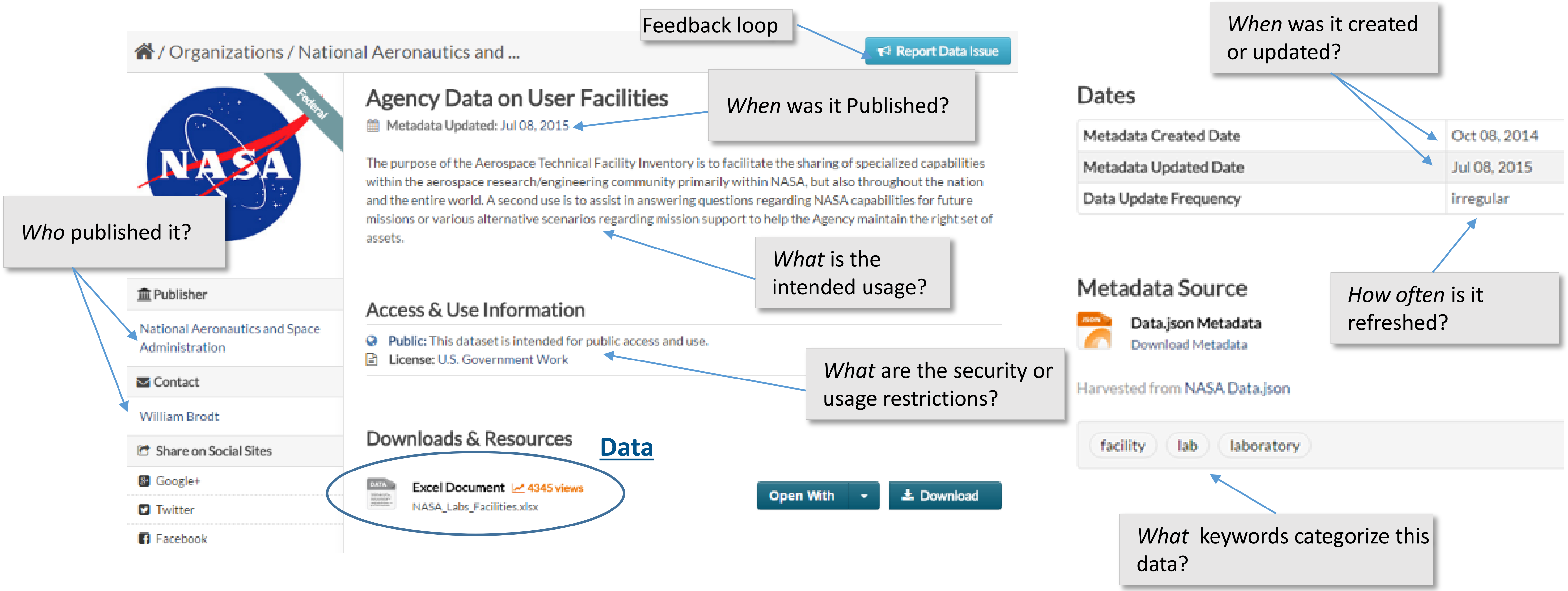
What are the data structures?

What are helpful comments?

Open Data Metadata

Publicly-available data

- With Open Data, metadata provides the context that makes information usable & credible.
- Data Lakes can use a similar method.



/ Organizations / National Aeronautics and ...

Report Data Issue

Feedback loop

Agency Data on User Facilities

Metadata Updated: Jul 08, 2015

When was it Published?

The purpose of the Aerospace Technical Facility Inventory is to facilitate the sharing of specialized capabilities within the aerospace research/engineering community primarily within NASA, but also throughout the nation and the entire world. A second use is to assist in answering questions regarding NASA capabilities for future missions or various alternative scenarios regarding mission support to help the Agency maintain the right set of assets.

What is the intended usage?

Who published it?

Publisher

National Aeronautics and Space Administration

Contact

William Brodt

Share on Social Sites

Google+

Twitter

Facebook

Access & Use Information

Public: This dataset is intended for public access and use.

License: U.S. Government Work

What are the security or usage restrictions?

Downloads & Resources

Data

Excel Document 4345 views

NASA_Labs_Facilities.xlsx

Open With

Download

When was it created or updated?

Dates

Metadata Created Date	Oct 08, 2014
Metadata Updated Date	Jul 08, 2015
Data Update Frequency	irregular

How often is it refreshed?

Metadata Source

Data.json Metadata

Download Metadata

Harvested from NASA Data.json

facility | lab | laboratory

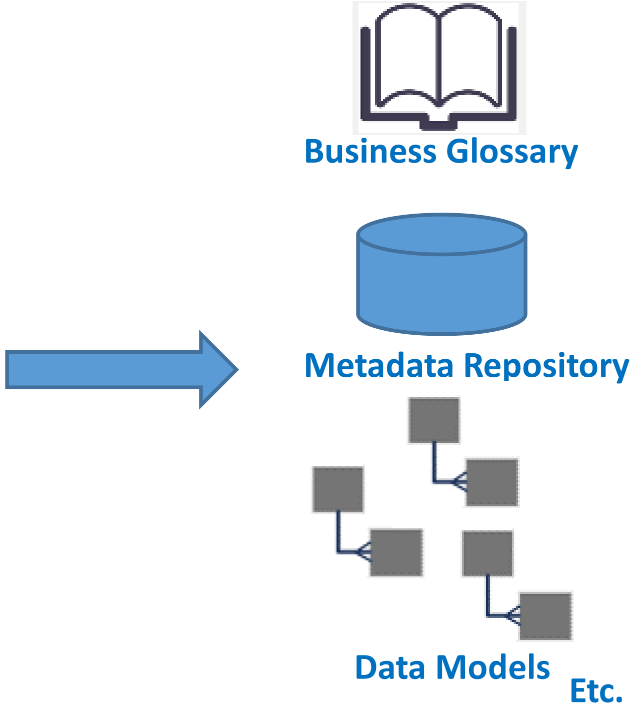
What keywords categorize this data?

Business Definitions are Critical

Putting information into context

- Business definitions of common terms are critical for the success for both Data Lakes & Data Warehouses.
- There are many ways to store this info: Business Glossary, Metadata Repository, even a spreadsheet --> the most important thing is that they are defined & published.

Business Term	Abbreviation	Definition
After Action Review	AAR	Team recap after every activity to share learning & improve best practices.
Activty Based Costing	ABD	Costs are allocated to products via cost drivers linked to various categories linked to the costs of manufacturing.
Component Number	C/N	Unique identifier associated with a given design for manufacture within ACME Corp.
Manufacturing Change Order	MCO	A change order used to make a manufacturing change. This typically does not involve a design change to the item.
Part Number	P/N	Unique identifier associated with a given design for manufacture within ACME Corp.
Etc.		…



Case Study: Consumer Energy Company

Business Transformation via Data

- For the consumer energy sector *Big Data and Smart Meters are transforming the ways of doing business* and interacting with customers.
  - Moving away from traditional data use cases of metering & billing.
  - Smart meters allow customers to be in control of their energy usage.
    - Control over energy usage with connected systems
    - Custom Energy Reports & Usage
    - Smart Billing based on usage times
- As energy usage declines, *data is becoming the true business asset* for this energy company.
  - Monetization of non-personal data is a future consideration.
- While the Big Data Opportunity is crucial, equally important are the traditional data sources
  - **New Data Quality Tools in place for operational and DW data**
  - **Data Governance Program analyzing data in relation to business processes & roles**
  - **Business-critical data elements identified and definitions created**

Data-Driven Business Evolution

Data is a key component for new business opportunities

Traditional Business Model

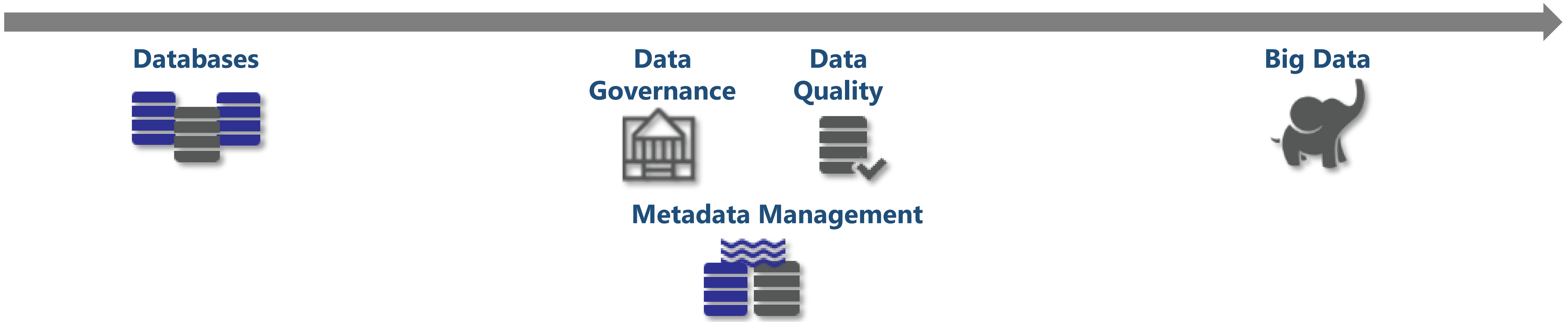
- Usage-based billing
- Issue-driven customer service

More Efficient Business Model

- More efficient billing
- Faster customer service response
- More consumer information re: energy efficiency, etc.

New Business Model

- Consumer-Driven Smart Metering
- Connected Devices, IoT
- Proactive service monitoring
- Monetization of usage data



Databases

Data Governance

Data Quality

Metadata Management

Big Data

Summary

- Data Lakes are a paradigm shift from traditional data warehouses
  - Data Lake: Discover then analyze
  - Data Warehouse: Design then implement
- Data Governance for the Data Lake needs to be customized for the technologies & audiences
  - Light touch documentation & governance (but not none!)
  - Feedback loop between traditional data warehouses & exploratory data lakes
- Communication & Culture is key
  - Different roles & personality types require different approaches
  - Focusing on business value creates common goals

About Global Data Strategy, Ltd

Data-Driven Business Transformation

- Global Data Strategy is an international information management consulting company that specializes in the alignment of business drivers with data-centric technology.
- Our passion is data, and helping organizations enrich their business opportunities through data and information.
- Our core values center around providing solutions that are:
  - **Business-Driven:** We put the needs of your business first, before we look at any technology solution.
  - **Clear & Relevant:** We provide clear explanations using real-world examples.
  - **Customized & Right-Sized:** Our implementations are based on the unique needs of your organization's size, corporate culture, and geography.
  - **High Quality & Technically Precise:** We pride ourselves in excellence of execution, with years of technical expertise in the industry.

Business Strategy *Aligned With* **Data Strategy**

Visit www.globaldatastrategy.com for more information

Contact Info

- Email: donna.burbank@globaldatastrategy.com
- Twitter: @donnaburbank
  @GlobalDataStrat
- Website: www.globaldatastrategy.com
- Company Linkedin: https://www.linkedin.com/company/global-data-strategy-ltd
- Personal Linkedin: https://www.linkedin.com/in/donnaburbank

DATAVERSITY Training Center

Online Training Courses

New Metadata Management Course

- Learn the basics of Metadata Management and practical tips on how to apply metadata management in the real world. This online course hosted by DATAVERSITY provides a series of six courses including:
  - What is Metadata
  - The Business Value of Metadata
  - Sources of Metadata
  - Metamodels and Metadata Standards
  - Metadata Architecture, Integration, and Storage
  - Metadata Strategy and Implementation
- Purchase all six courses for $399 or individually at $79 each. **Use discount code “GDS” to receive 20% off!**
  - Register **here**
- Other courses available on Data Governance & Data Quality

Visit: http://training.dataversity.net/lms/

Questions?

Thoughts? Ideas?