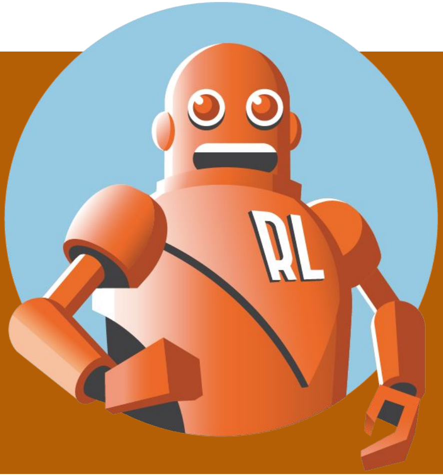
Hello!



RICKER LYMAN

ROBOTIC

Big Data Governance

The Ricker Lyman Robotic Company

Scope of talk

- What is data governance?
- Why is it a thing?
- Why is it critical to enterprise big data?
- What tools are available?
- How do we implement big data governance?
- What is the future of big data governance?

What is data governance?

Definition

Data governance is the orchestration of people, processes, technology, and policies to ensure the availability, usability, integrity, consistency, auditability, and security of our data.

Break that definition down

orchestration of

- people,
- processes,
- technology,
- policies

to ensure data

- availability
- usability
- integrity
- consistency
- auditability
- security

Business objectives

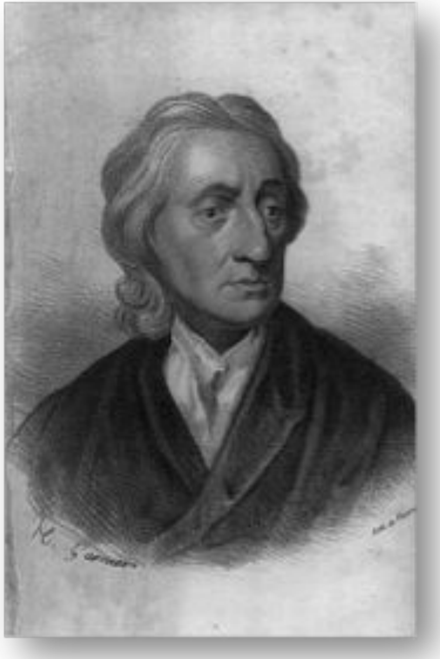
- Understand implications of a data outage
- Understand the impact of a data change
- Mitigate the corruption of data over time
- Mitigate the impact of data changes
- Mitigate the impact of data outages
- Rapidly identify and fix data impacts and prevent them from reoccurring

Why is this even a thing?

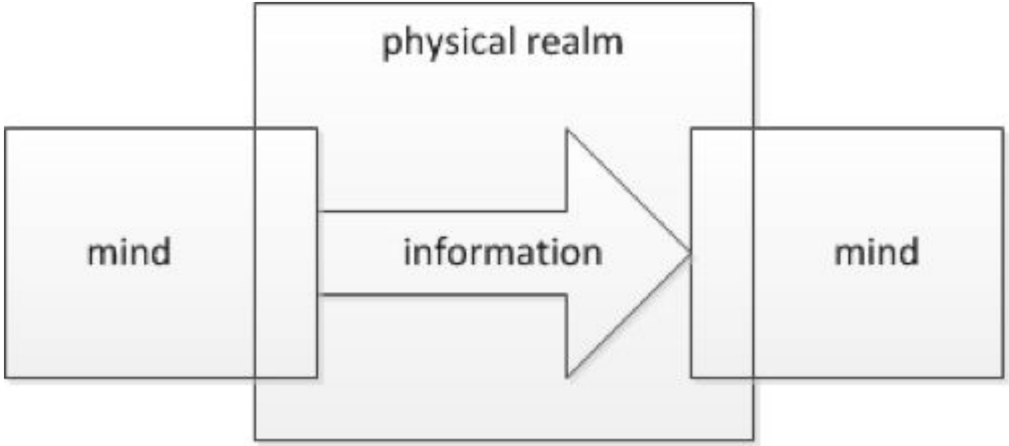
Locke's Law

Communication is the transfer of human thought from one individual to another through a shared physical medium.

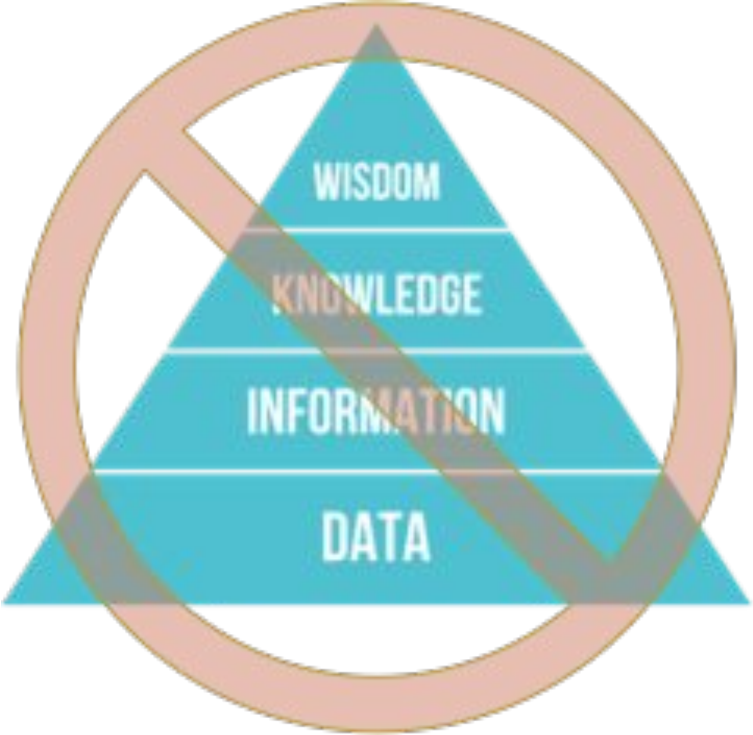
Information is the physical form that thought takes during communication



Communication



Data does not add up to information



Shannon's Law

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

Shannon’s model

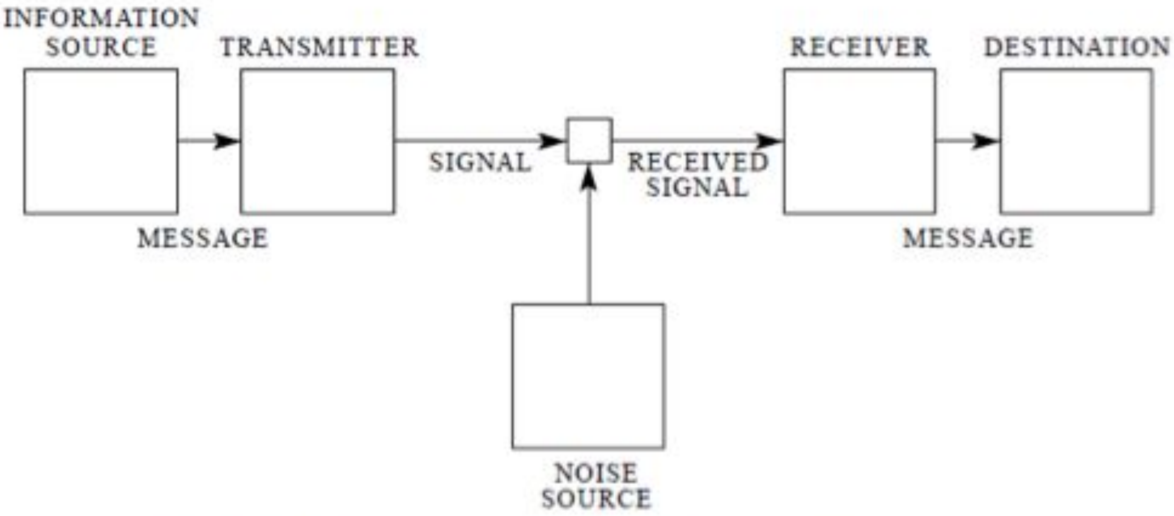


Fig. 1—Schematic diagram of a general communication system.

Implications of Shannon’s law

These semantic aspects of communication are irrelevant to the engineering problem.

The significant aspect is that the actual message is one *selected from a set* of possible messages.

Encoding

Employee identifier

Customer identifier

Product identifier

Location identifier

Wittgenstein’s Law

Individual can never perfectly manifest a thought and can never perfectly interpret information.

Semantic incongruity is unavoidable.

Semantic Gap

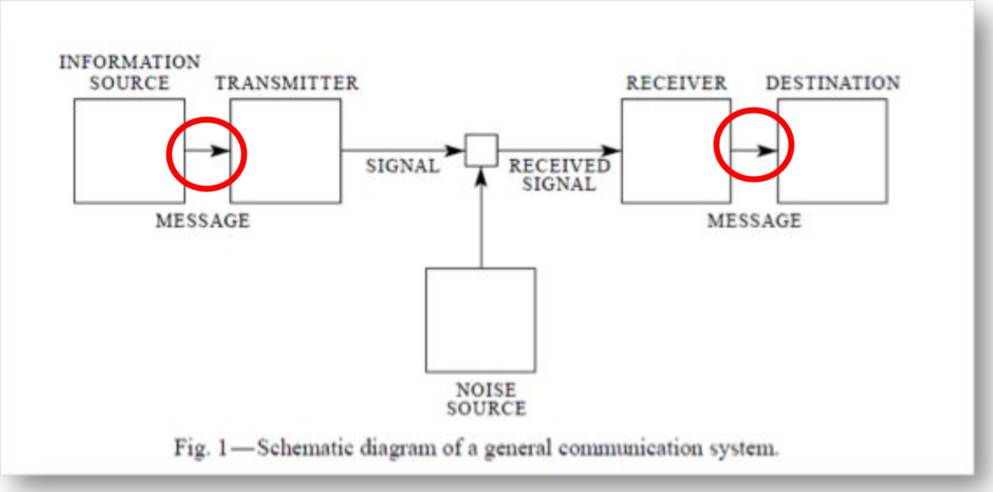
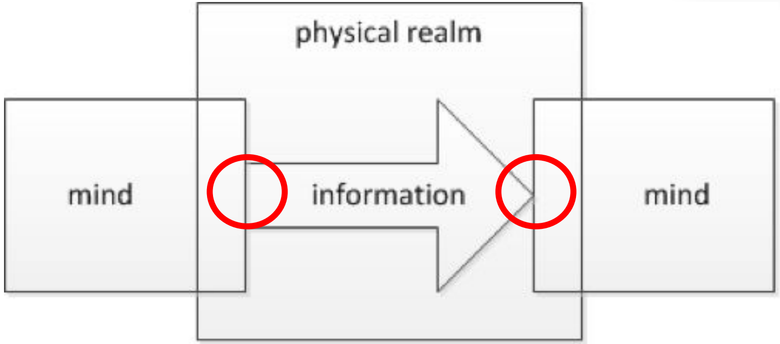


Fig. 1—Schematic diagram of a general communication system.



10,000

Number of words to be fluent in a language

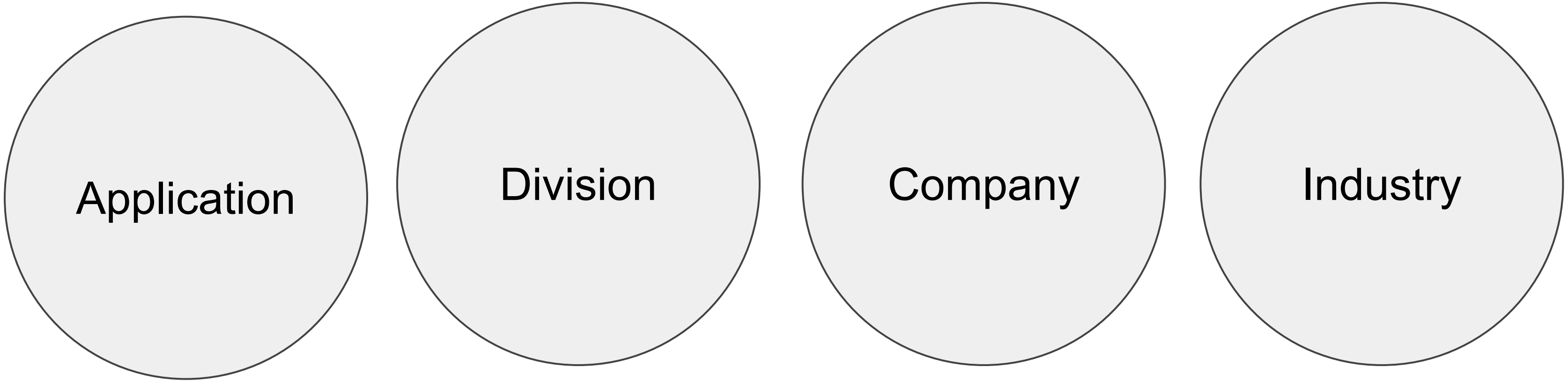
Number of code identifiers in EDI X12 or EDIFACT

Salary example

- Salary
- USD
- Gross pay before taxes

- Salaire
- EUR
- Net pay after taxes plus lunch allowance

Translation



Application

Division

Company

Industry

Miller’s Law

Individuals have a finite capacity for communication based on the inherent cognitive limitations of the human mind.

7 ± 2

The number of distinct items a human can consistently distinguish on a single sensory perception

Bottleneck

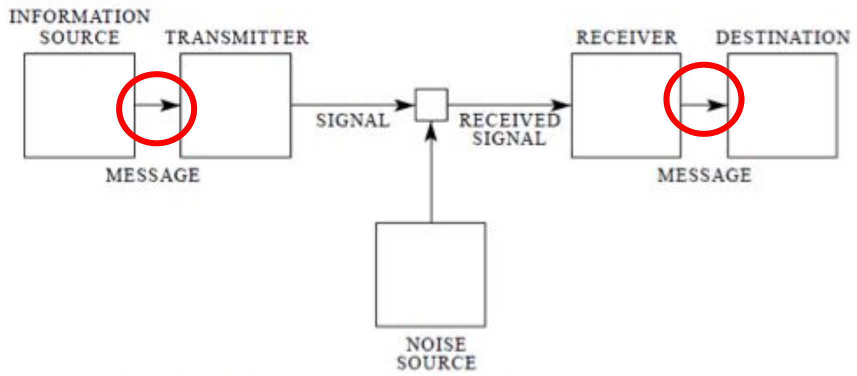
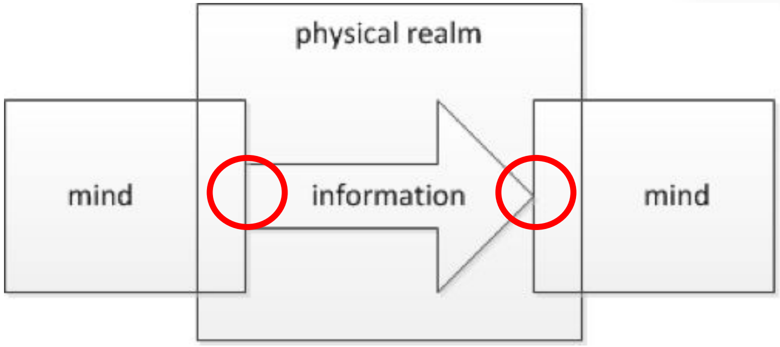
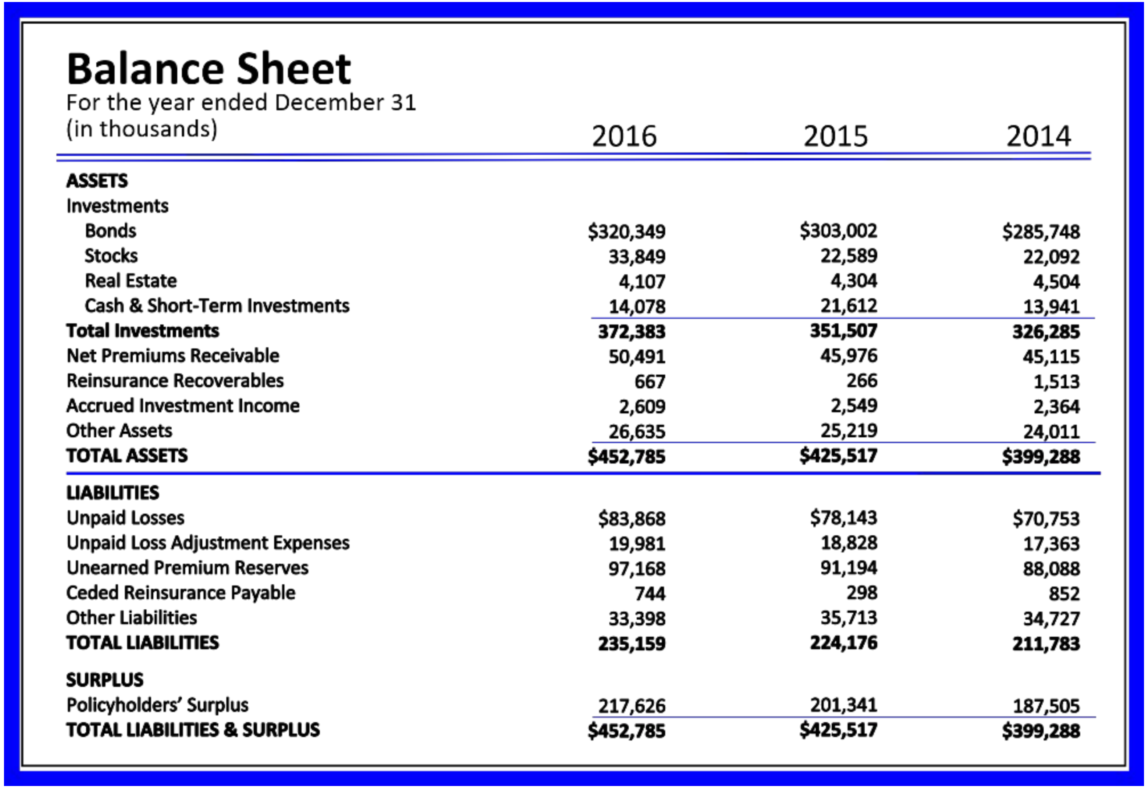


Fig. 1—Schematic diagram of a general communication system.



Aggregate

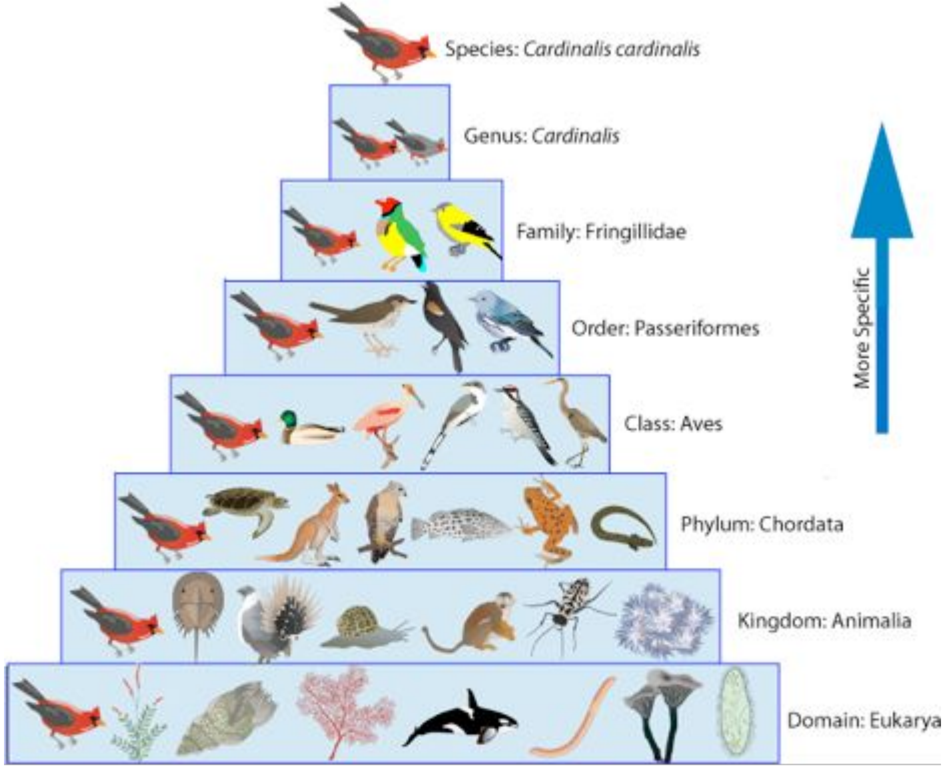


Balance Sheet

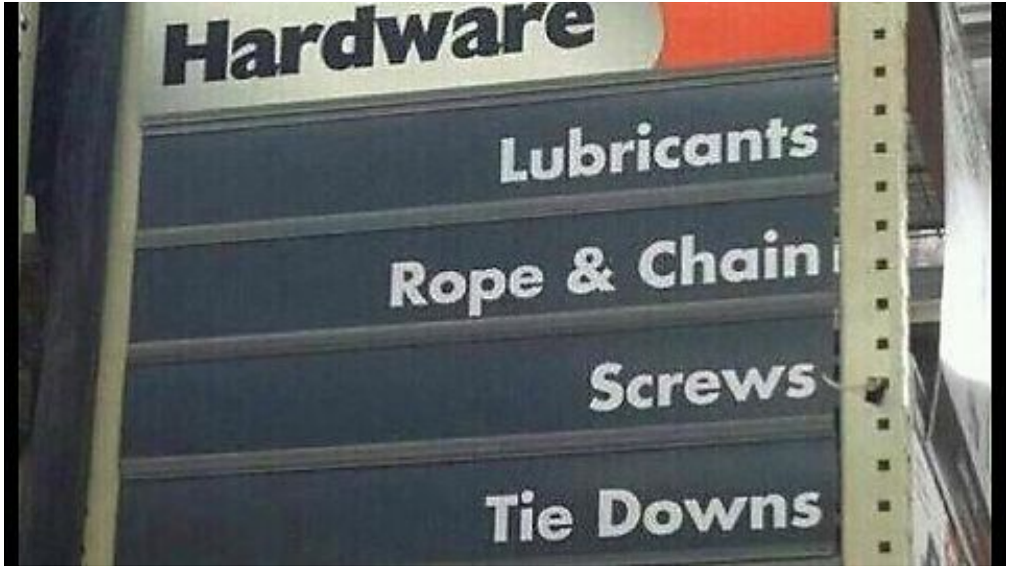
For the year ended December 31
(in thousands)

	2016	2015	2014
ASSETS			
Investments			
Bonds	$320,349	$303,002	$285,748
Stocks	33,849	22,589	22,092
Real Estate	4,107	4,304	4,504
Cash & Short-Term Investments	14,078	21,612	13,941
Total Investments	**372,383**	**351,507**	**326,285**
Net Premiums Receivable	50,491	45,976	45,115
Reinsurance Recoverables	667	266	1,513
Accrued Investment Income	2,609	2,549	2,364
Other Assets	26,635	25,219	24,011
TOTAL ASSETS	**$452,785**	**$425,517**	**$399,288**
LIABILITIES			
Unpaid Losses	$83,868	$78,143	$70,753
Unpaid Loss Adjustment Expenses	19,981	18,828	17,363
Unearned Premium Reserves	97,168	91,194	88,088
Ceded Reinsurance Payable	744	298	852
Other Liabilities	33,398	35,713	34,727
TOTAL LIABILITIES	**235,159**	**224,176**	**211,783**
SURPLUS			
Policyholders' Surplus	217,626	201,341	187,505
TOTAL LIABILITIES & SURPLUS	**$452,785**	**$425,517**	**$399,288**

Taxonomy


Species: *Cardinalis cardinalis*
Genus: *Cardinalis*
Family: Fringillidae
Order: Passeriformes
Class: Aves
Phylum: Chordata
Kingdom: Animalia
Domain: Eukarya
More Specific

Taxonomy



Different taxonomies

Sales: aggregated by sales region

Logistics: aggregated by distribution center

Marketing: aggregated by municipal statistical area (MSA)

Accounting: aggregated by channel partner

Why is data governance a thing?

Different encodings

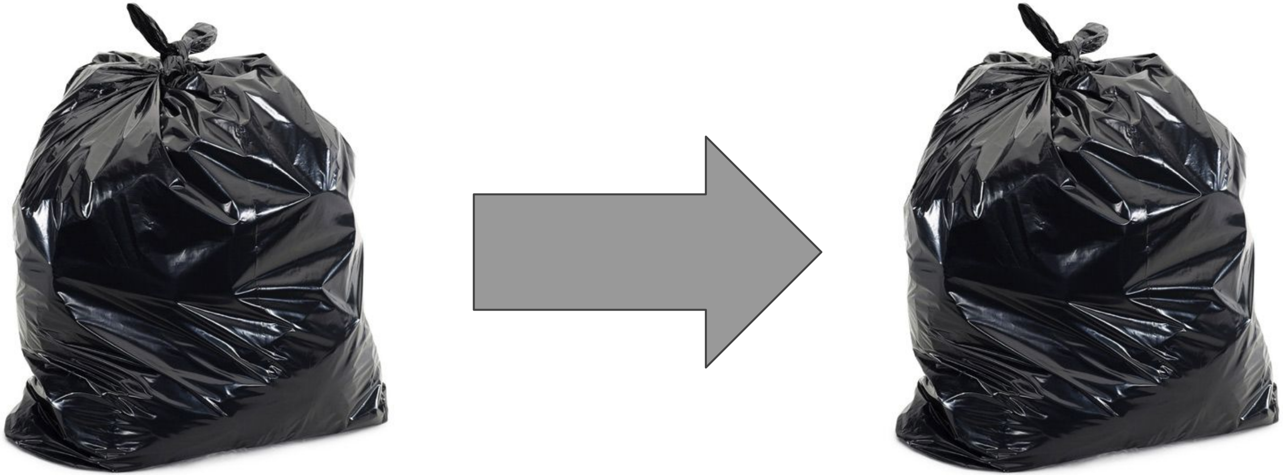
Semantic incongruity

Different taxonomies

The very (human) nature of information itself

Why is data governance critical to success?

Garbage in, garbage out

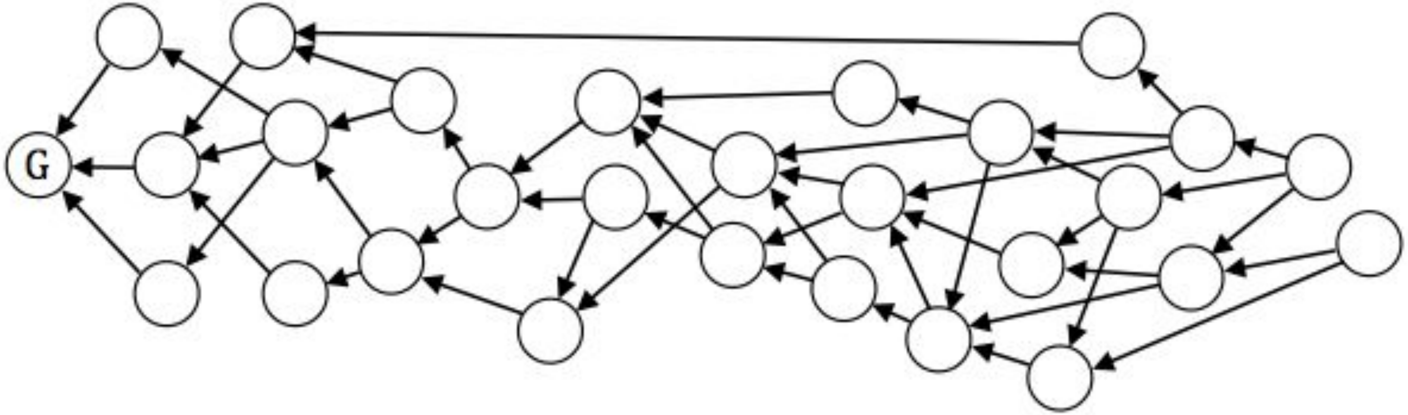


Salary example (again)

- Salary
- USD
- Gross pay before taxes

- Salaire
- EUR
- Net pay after taxes plus lunch allowance

Complex data lineage



Customer trust

Know the data

Semantics

Lineage

Transformations

Disruption

Corruption

Regulation

Personal identifying information (PII)

HIPAA

GDPR

Data governance is...

orchestration of

- people,
- processes,
- technology,
- policies

to ensure data

- availability
- usability
- integrity
- consistency
- auditability
- security

Business objectives

- Understand implications of a data outage
- Understand the impact of a data change
- Mitigate the corruption of data over time
- Mitigate the impact of data changes
- Mitigate the impact of data outages
- Rapidly identify and fix data impacts and prevent them from reoccurring

What tools are available?

Commercial & open source tools

Collibra

Informatica

Datum

SAP

IBM

Cloudera Navigator

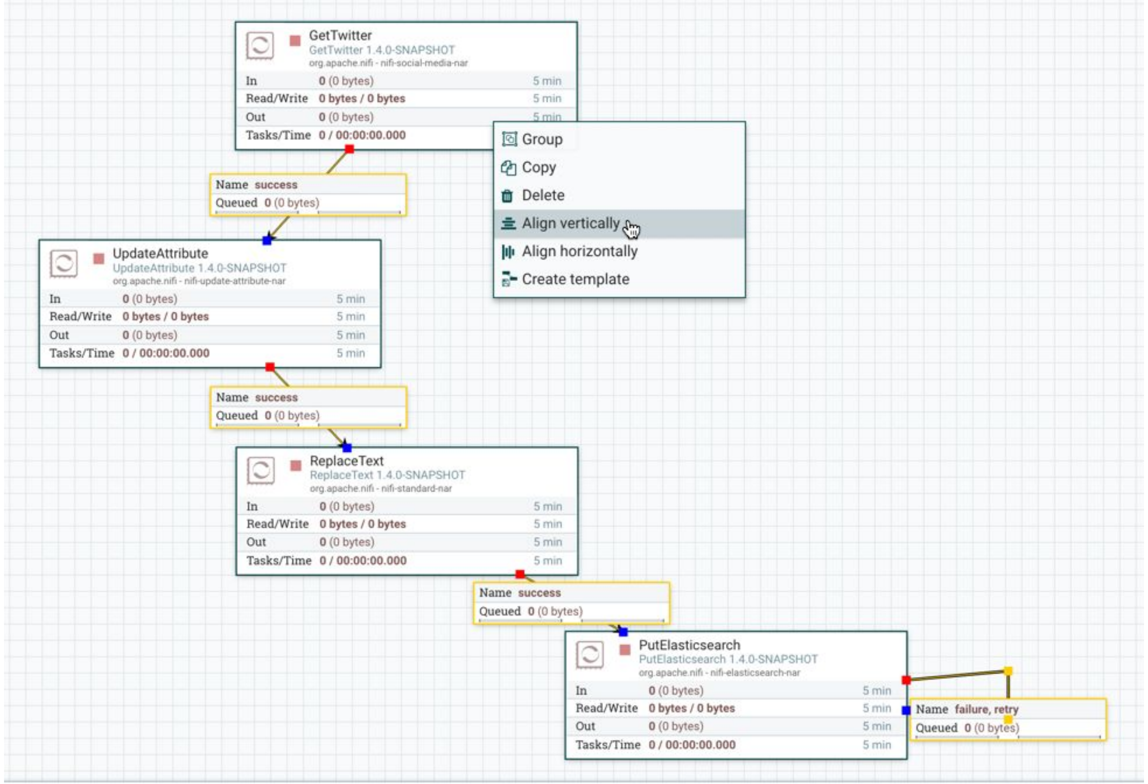
Apache Nifi

Schema Registry

Apache Ranger

Apache Atlas

Nifi data flow



GetTwitter
GetTwitter 1.4.0-SNAPSHOT
org.apache.nifi - nifi-social-media-nar

```
In           0 (0 bytes)            5 min
Read/Write   0 bytes / 0 bytes      5 min
Out          0 (0 bytes)            5 min
Tasks/Time   0 / 00:00:00.000
```

Name success
Queued 0 (0 bytes)

UpdateAttribute
UpdateAttribute 1.4.0-SNAPSHOT
org.apache.nifi - nifi-update-attribute-nar

```
In           0 (0 bytes)            5 min
Read/Write   0 bytes / 0 bytes      5 min
Out          0 (0 bytes)            5 min
Tasks/Time   0 / 00:00:00.000       5 min
```

Name success
Queued 0 (0 bytes)

ReplaceText
ReplaceText 1.4.0-SNAPSHOT
org.apache.nifi - nifi-standard-nar

```
In           0 (0 bytes)            5 min
Read/Write   0 bytes / 0 bytes      5 min
Out          0 (0 bytes)            5 min
Tasks/Time   0 / 00:00:00.000       5 min
```

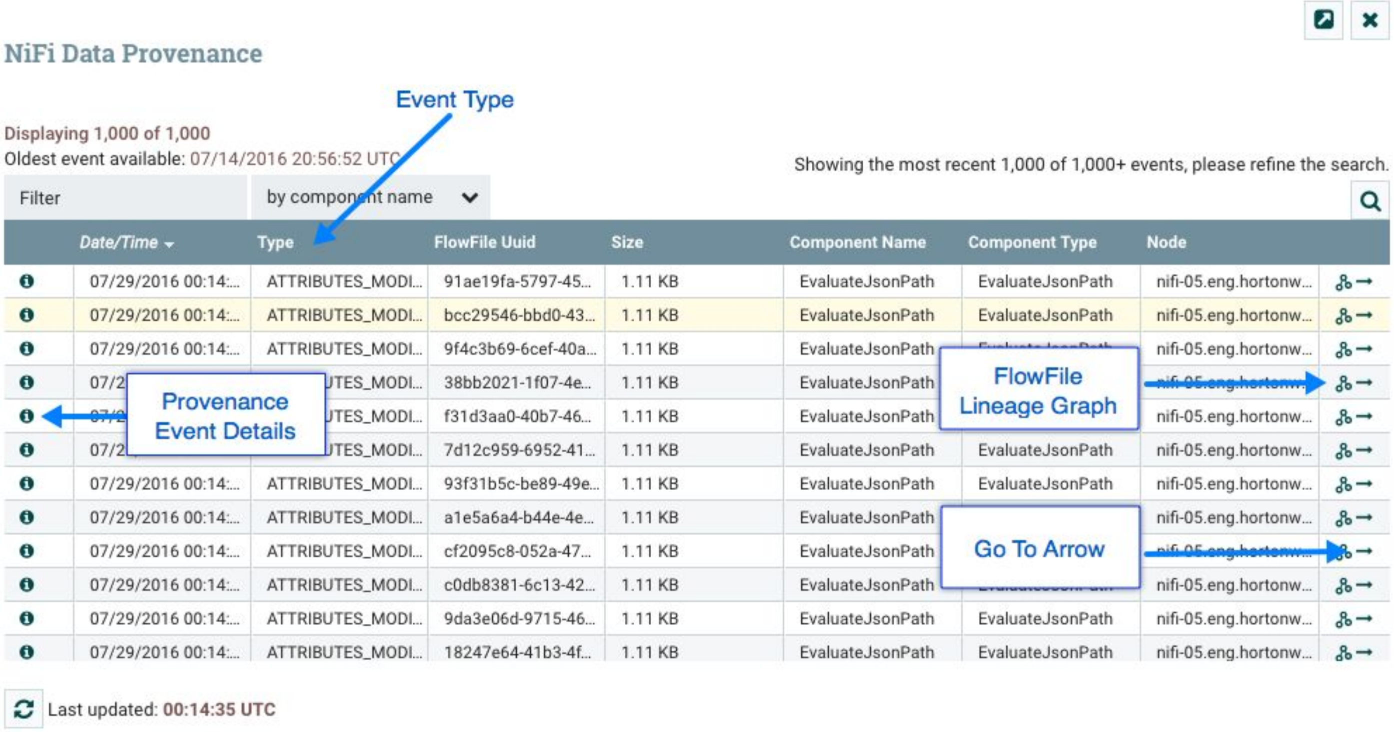
Name success
Queued 0 (0 bytes)

PutElasticsearch
PutElasticsearch 1.4.0-SNAPSHOT
org.apache.nifi - nifi-elasticsearch-nar

```
In           0 (0 bytes)            5 min
Read/Write   0 bytes / 0 bytes      5 min
Out          0 (0 bytes)            5 min
Tasks/Time   0 / 00:00:00.000       5 min
```

Name failure, retry
Queued 0 (0 bytes)

Nifi data provenance



NiFi Data Provenance

Displaying 1,000 of 1,000

Oldest event available: 07/14/2016 20:56:52 UTC

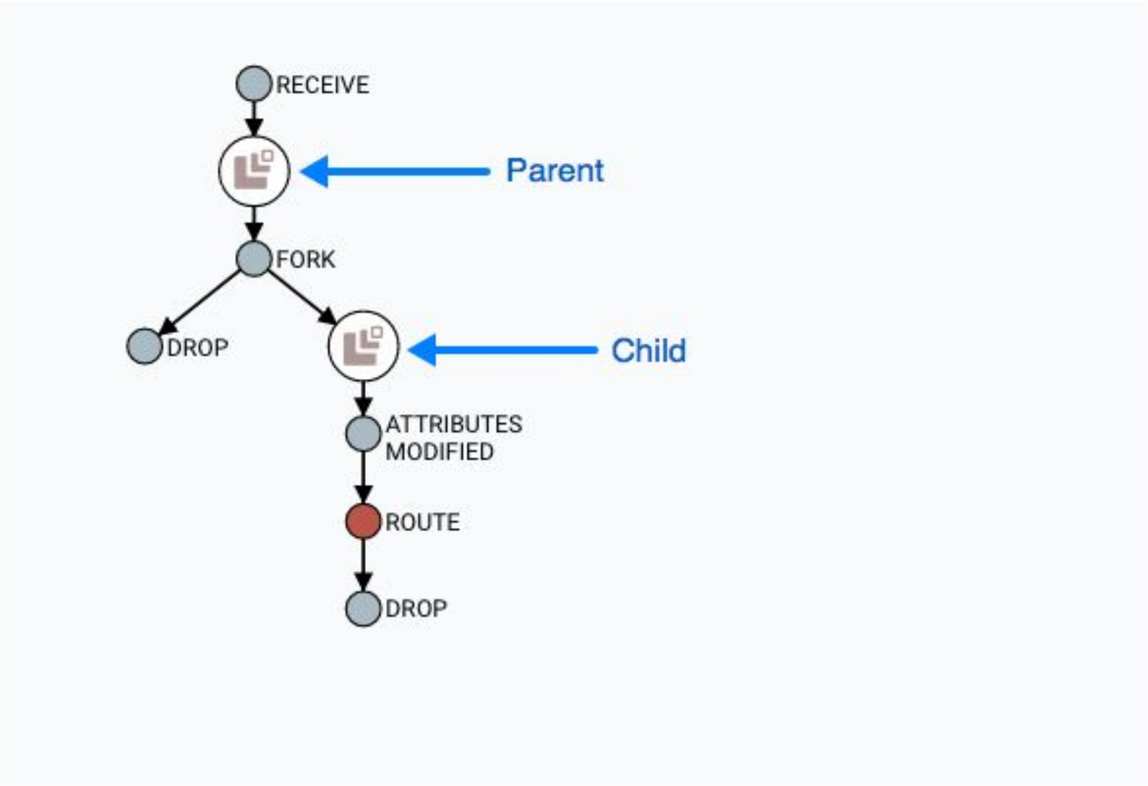
Showing the most recent 1,000 of 1,000+ events, please refine the search.

Filter | by component name

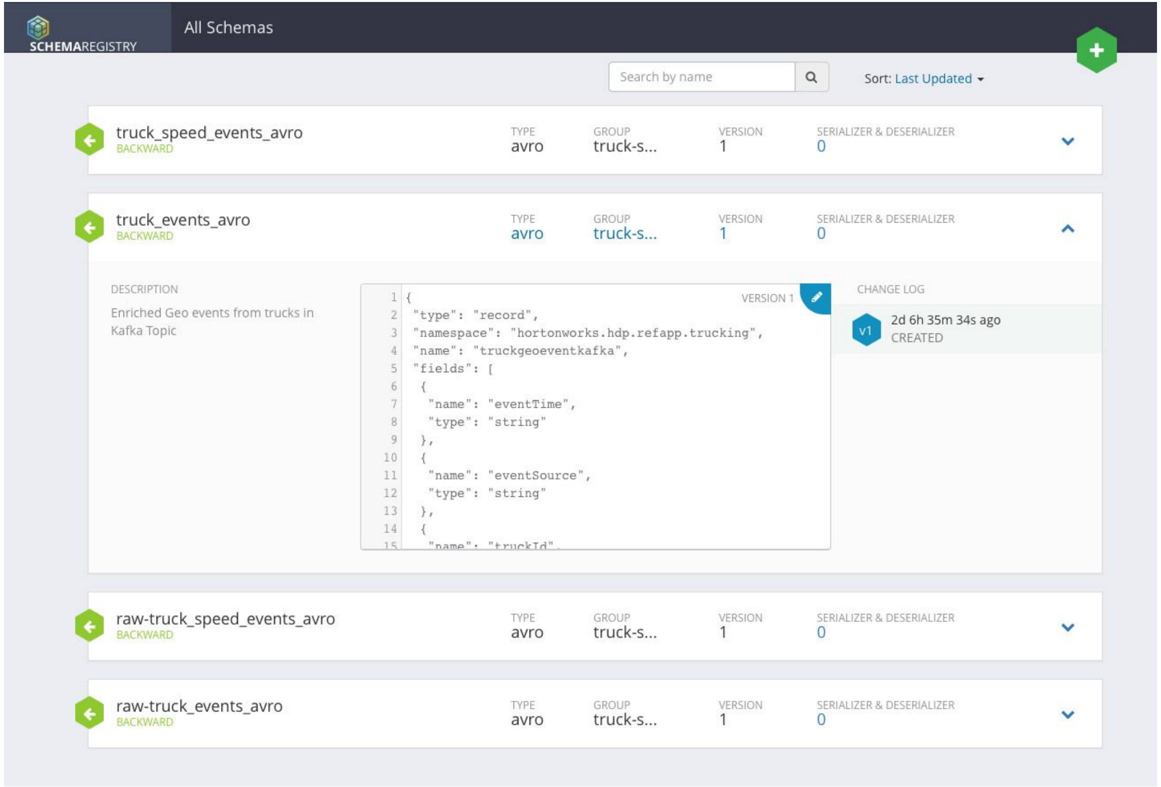
	Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type	Node	
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	91ae19fa-5797-45...	1.11 KB	EvaluateJsonPath	EvaluateJsonPath	nifi-05.eng.hortonw...	➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	bcc29546-bbd0-43...	1.11 KB	EvaluateJsonPath	EvaluateJsonPath	nifi-05.eng.hortonw...	➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	9f4c3b69-6cef-40a...	1.11 KB	EvaluateJsonPath		nifi-05.eng.hortonw...	➞
❶	07/2	UTES_MODI...	38bb2021-1f07-4e...	1.11 KB	EvaluateJsonPath			➞
❶		UTES_MODI...	f31d3aa0-40b7-46...	1.11 KB	EvaluateJsonPath		nifi-05.eng.hortonw...	➞
❶	07/2	UTES_MODI...	7d12c959-6952-41...	1.11 KB	EvaluateJsonPath	EvaluateJsonPath	nifi-05.eng.hortonw...	➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	93f31b5c-be89-49e...	1.11 KB	EvaluateJsonPath	EvaluateJsonPath	nifi-05.eng.hortonw...	➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	a1e5a6a4-b44e-4e...	1.11 KB	EvaluateJsonPath		nifi-05.eng.hortonw...	➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	cf2095c8-052a-47...	1.11 KB	EvaluateJsonPath			➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	c0db8381-6c13-42...	1.11 KB	EvaluateJsonPath		nifi-05.eng.hortonw...	➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	9da3e06d-9715-46...	1.11 KB	EvaluateJsonPath	EvaluateJsonPath	nifi-05.eng.hortonw...	➞
❶	07/29/2016 00:14:...	ATTRIBUTES_MODI...	18247e64-41b3-4f...	1.11 KB	EvaluateJsonPath	EvaluateJsonPath	nifi-05.eng.hortonw...	➞

Last updated: 00:14:35 UTC

Nifi data provenance



Schema registry



	TYPE	GROUP	VERSION	SERIALIZER & DESERIALIZER
truck_speed_events_avro BACKWARD	avro	truck-s...	1	0
truck_events_avro BACKWARD	avro	truck-s...	1	0
raw-truck_speed_events_avro BACKWARD	avro	truck-s...	1	0
raw-truck_events_avro BACKWARD	avro	truck-s...	1	0

DESCRIPTION

Enriched Geo events from trucks in Kafka Topic

VERSION 1

```
{
 "type": "record",
 "namespace": "hortonworks.hdp.refapp.trucking",
 "name": "truckgeoeventkafka",
 "fields": [
  {
   "name": "eventTime",
   "type": "string"
  },
  {
   "name": "eventSource",
   "type": "string"
  },
  {
   "name": "truckId",
```

CHANGE LOG

v1 2d 6h 35m 34s ago
CREATED

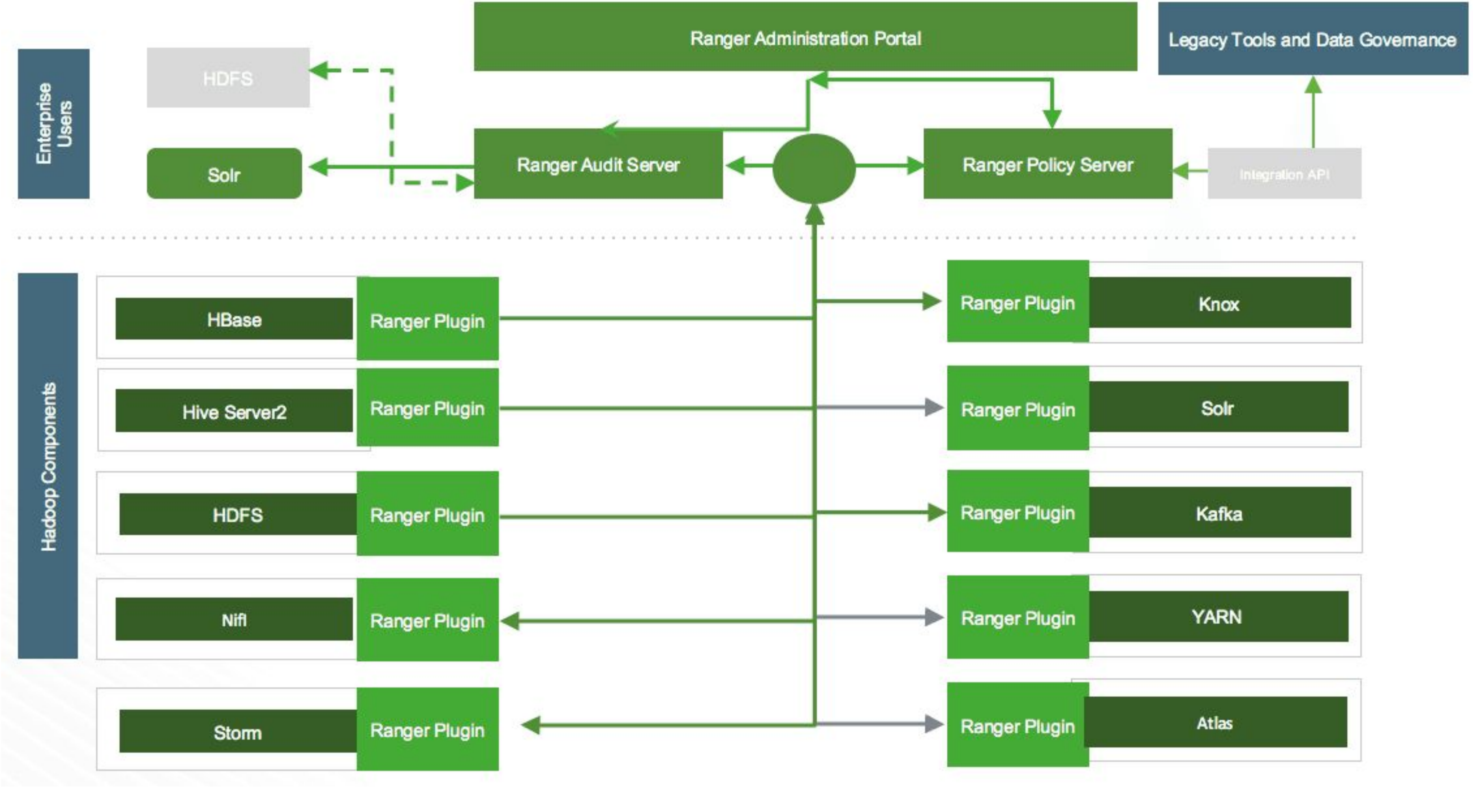
Schema registry

Centralized registry - Provide reusable schema to avoid attaching schema to every piece of data.

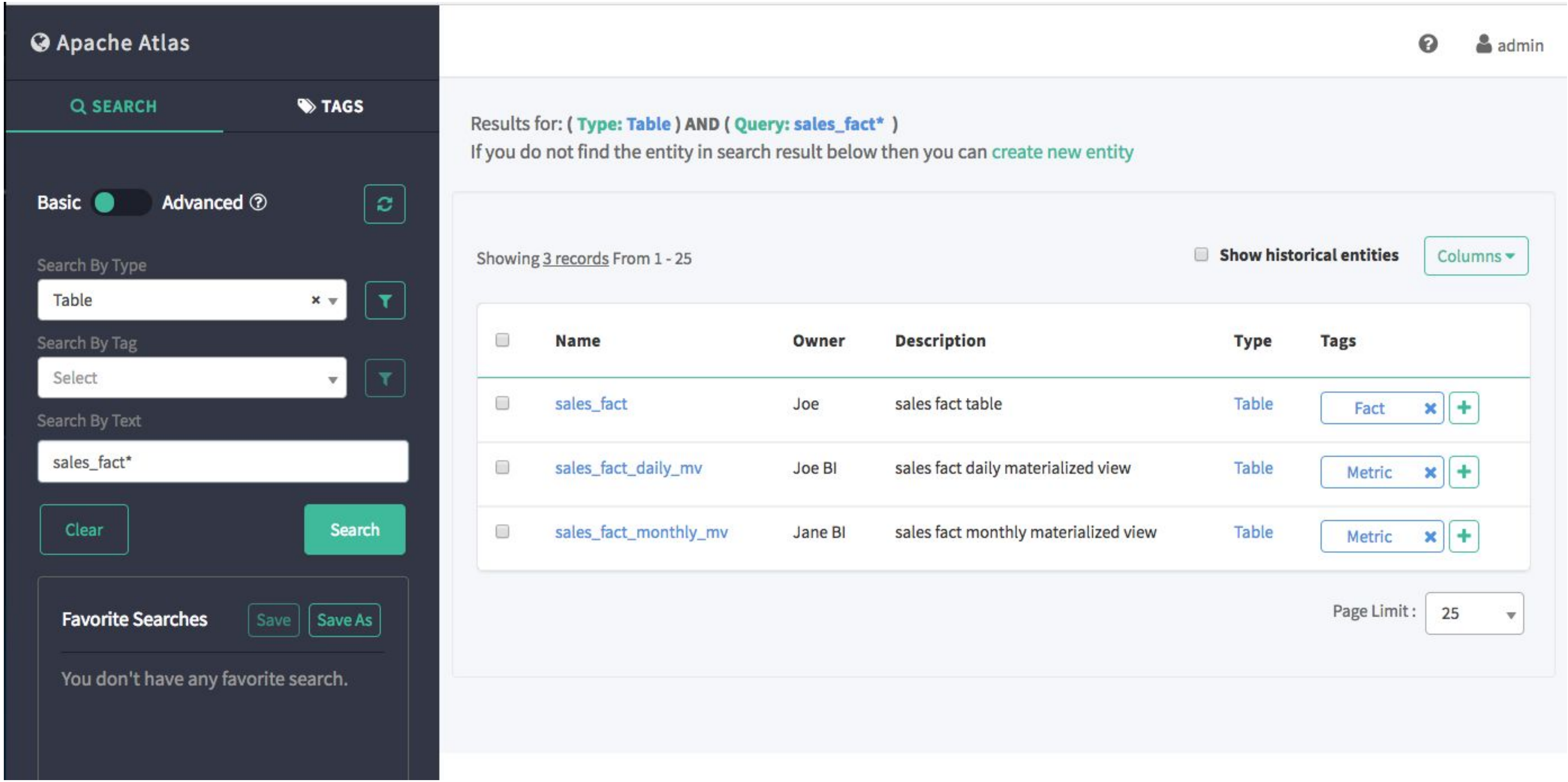
Version management - Define relationship between schema versions so that consumers and producers can evolve at different rates.

Schema validation - Enable generic format conversion, generic routing, and data quality.

Apache Ranger



Atlas metadata search



Search By Type: Table

Search By Tag: Select

Search By Text: sales_fact*

You don't have any favorite search.

Results for: (Type: Table) AND (Query: sales_fact*)

If you do not find the entity in search result below then you can create new entity

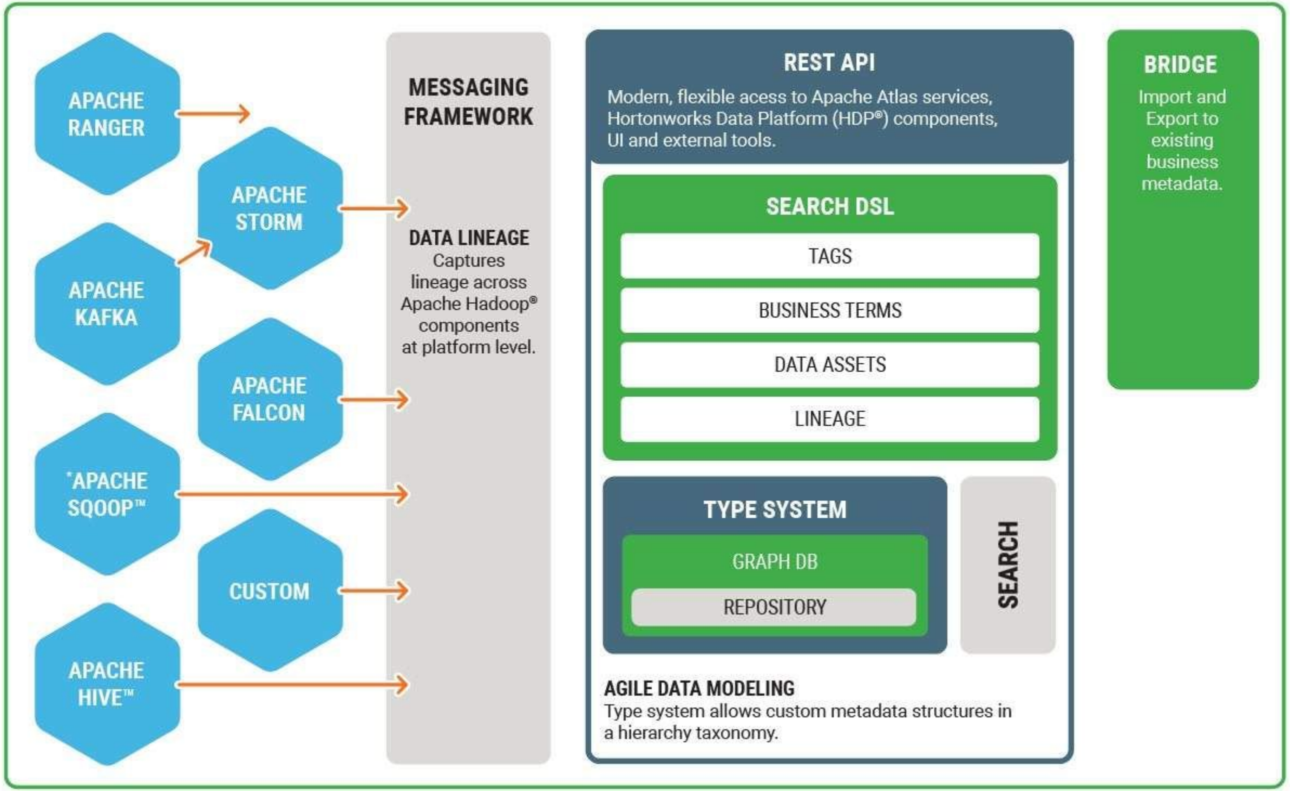
Showing 3 records From 1 - 25

□ Show historical entities

□	Name	Owner	Description	Type	Tags
□	sales_fact	Joe	sales fact table	Table	Fact ✕ +
□	sales_fact_daily_mv	Joe BI	sales fact daily materialized view	Table	Metric ✕ +
□	sales_fact_monthly_mv	Jane BI	sales fact monthly materialized view	Table	Metric ✕ +

Page Limit : 25

APACHE ATLAS ARCHITECTURE



DATA LINEAGE
Captures lineage across Apache Hadoop® components at platform level.

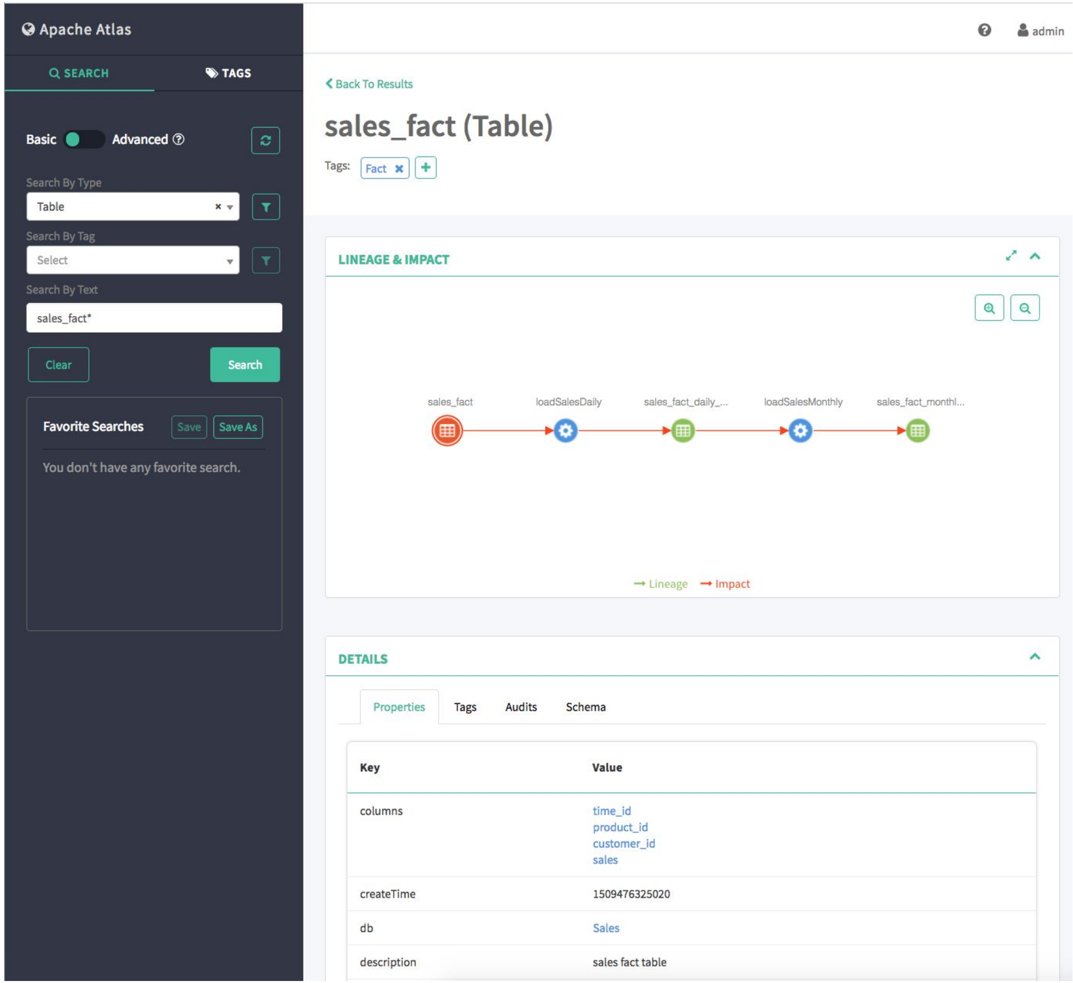
REST API
Modern, flexible acess to Apache Atlas services, Hortonworks Data Platform (HDP®) components, UI and external tools.

AGILE DATA MODELING
Type system allows custom metadata structures in a hierarchy taxonomy.

BRIDGE
Import and Export to existing business metadata.

*Applies to any connector that leverages Apache Sqoop including Teradata Connector

Metadata details



Basic Advanced

Search By Type: Table

Search By Tag: Select

Search By Text: sales_fact*

Favorite Searches

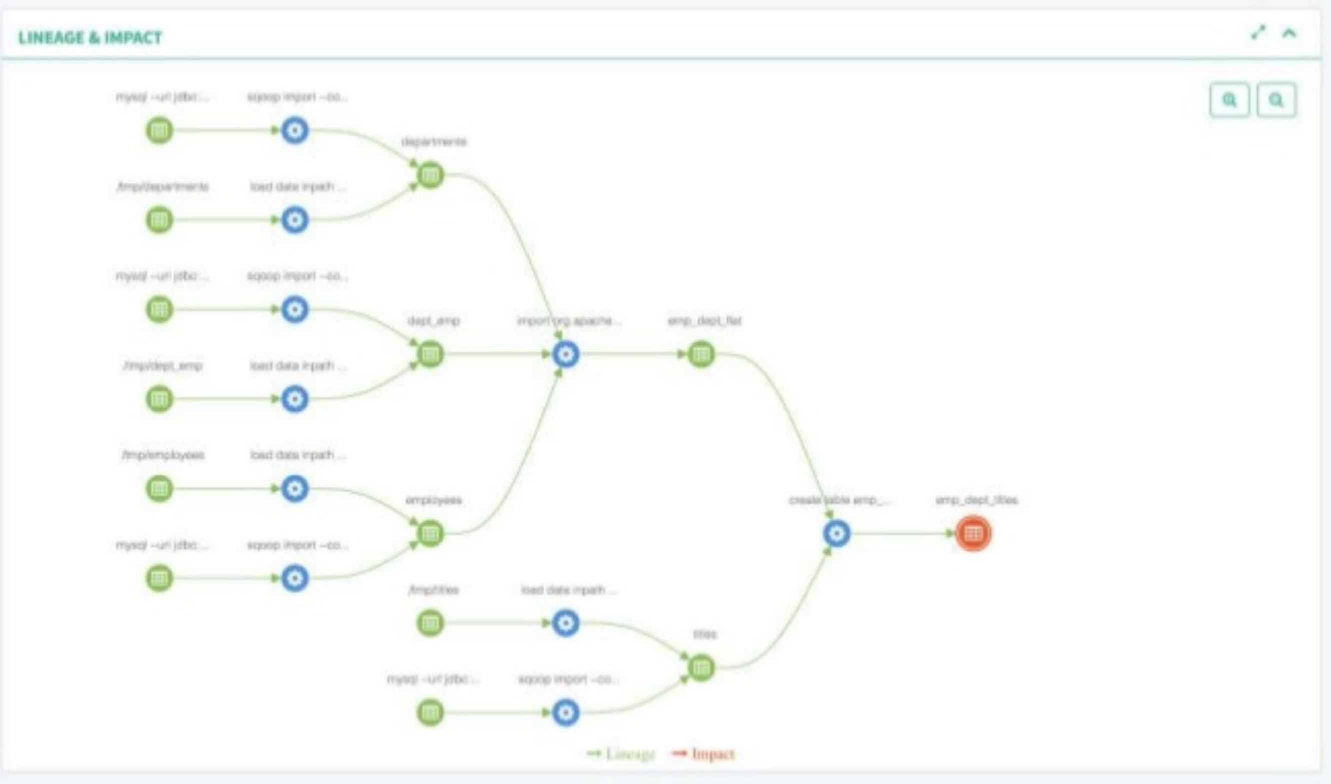
You don't have any favorite search.

sales_fact (Table)

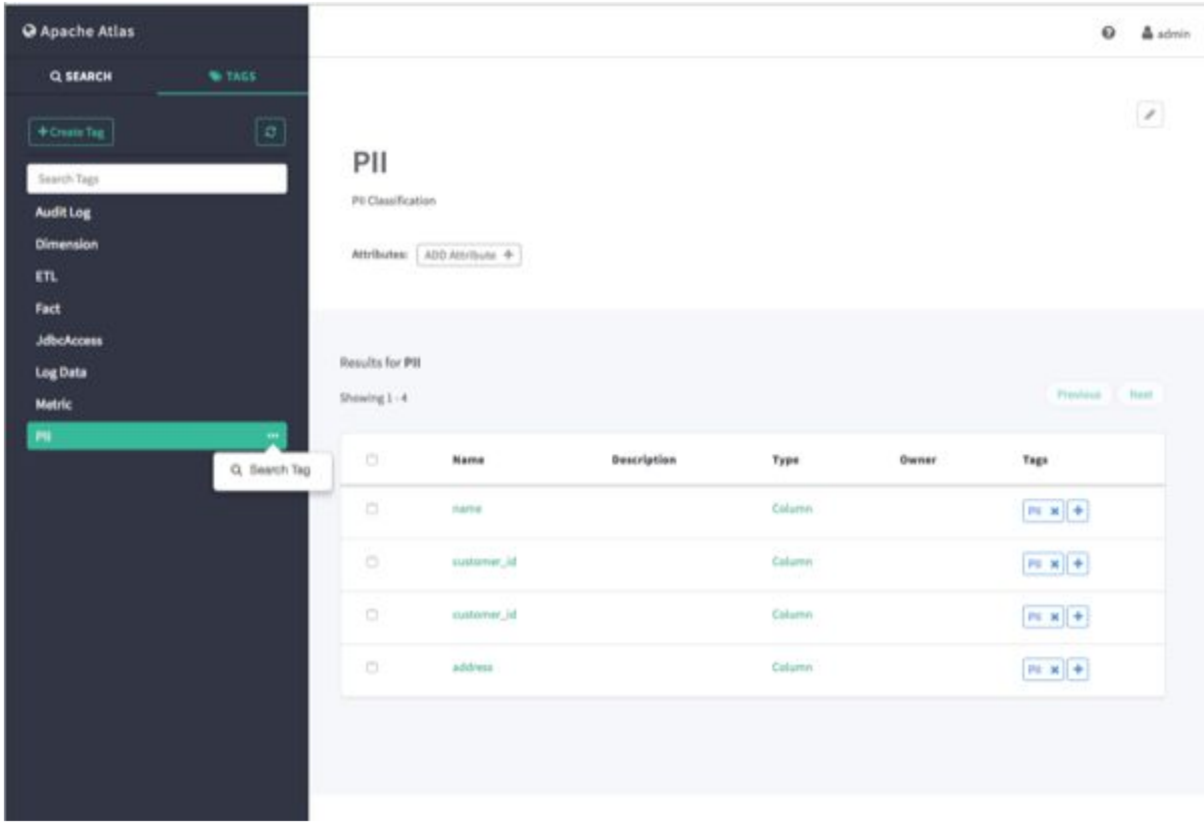
Tags: Fact ✖ +

Key	Value
columns	time_id product_id customer_id sales
createTime	1509476325020
db	Sales
description	sales fact table

Atlas data lineage



Atlas tags



PII

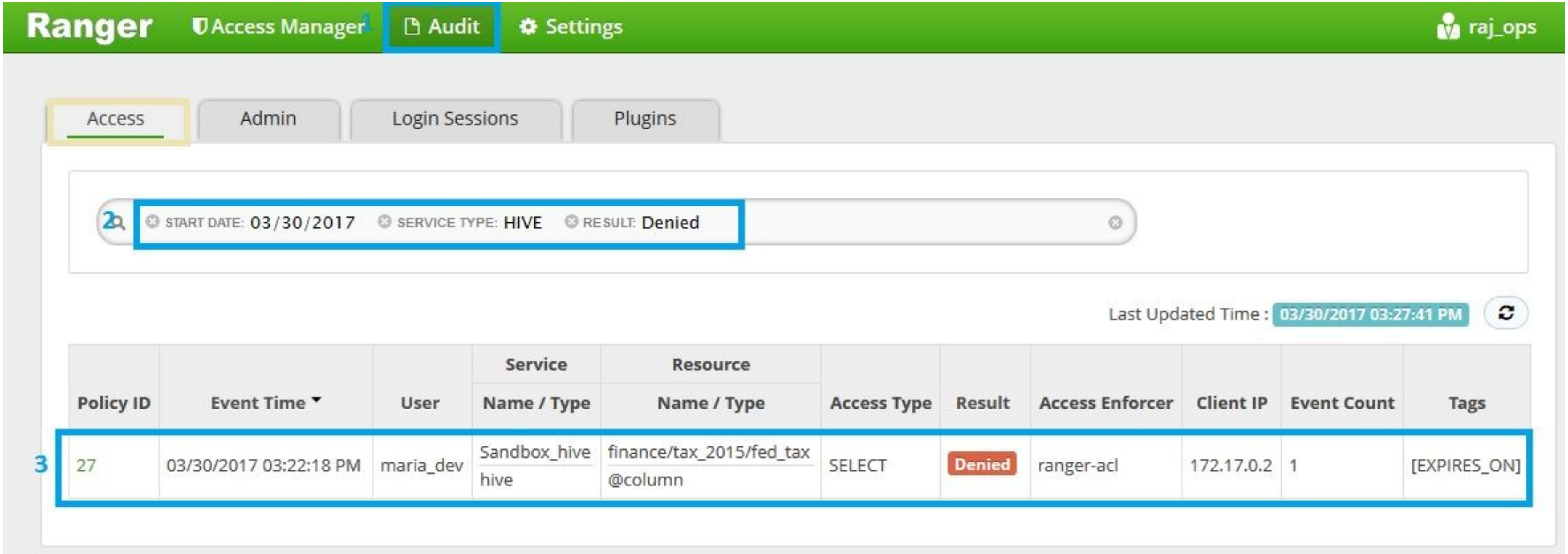
PII Classification

Results for **PII**

Showing 1 - 4

☐	Name	Description	Type	Owner	Tags
☐	name		Column		PII ✕ +
☐	customer_id		Column		PII ✕ +
☐	customer_id		Column		PII ✕ +
☐	address		Column		PII ✕ +

Ranger rules with Atlas tags



START DATE: 03/30/2017 SERVICE TYPE: HIVE RESULT: Denied

Last Updated Time : 03/30/2017 03:27:41 PM

			Service	Resource						
Policy ID	Event Time	User	Name / Type	Name / Type	Access Type	Result	Access Enforcer	Client IP	Event Count	Tags
27	03/30/2017 03:22:18 PM	maria_dev	Sandbox_hive hive	finance/tax_2015/fed_tax @column	SELECT	Denied	ranger-acl	172.17.0.2	1	[EXPIRES_ON]

Data governance tools

Apache Nifi

Schema Registry

Apache Ranger

Apache Atlas

How do we implement?

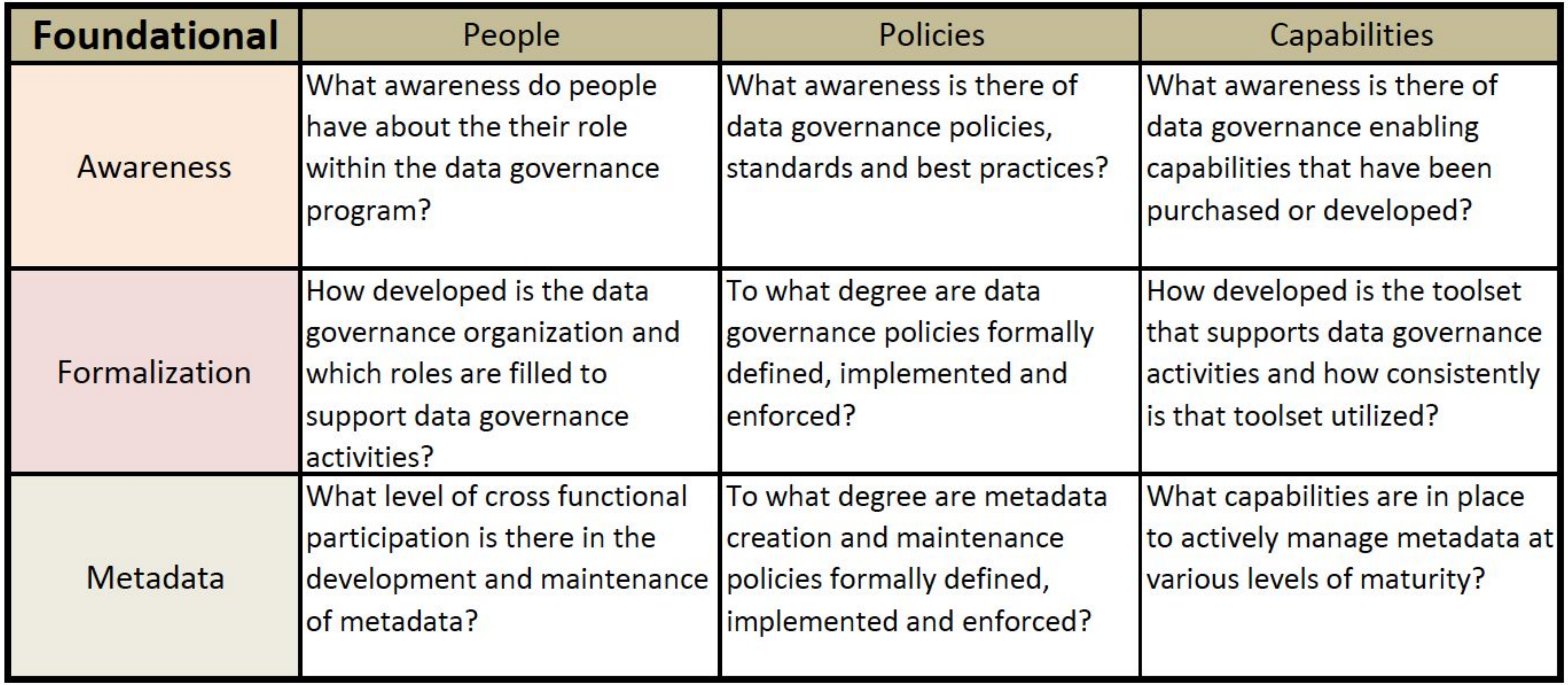
Data governance maturity model

Know where you are

Know where you are going

Improve what you measure

Stamford DGMM



Foundational	People	Policies	Capabilities
Awareness	What awareness do people have about the their role within the data governance program?	What awareness is there of data governance policies, standards and best practices?	What awareness is there of data governance enabling capabilities that have been purchased or developed?
Formalization	How developed is the data governance organization and which roles are filled to support data governance activities?	To what degree are data governance policies formally defined, implemented and enforced?	How developed is the toolset that supports data governance activities and how consistently is that toolset utilized?
Metadata	What level of cross functional participation is there in the development and maintenance of metadata?	To what degree are metadata creation and maintenance policies formally defined, implemented and enforced?	What capabilities are in place to actively manage metadata at various levels of maturity?

Data stewards

Information challenges are people challenges

Ownership

Accountability

Data dictionary

Document your technical metadata to describe

- structure
- relationship to other data
- origin
- format
- use

Data glossary

identify where there are a number of differing definitions for the same term

and conversely

where a number of different terms have the same definition

Master data management

The “nouns” upon which business transactions take action

Core entities of an enterprise that are used by multiple business process and IT systems

- Parties (customers, employees, vendors, suppliers)
- Places (locations, sales territories, offices)
- Things (accounts, products, assets, document sets)

How to get started

1. Data governance maturity model
2. Data stewards
3. Data dictionary
4. Data glossary
5. Master data management

Future of data governance

Future

Less bureaucracy, more automation

Graph databases

Open source

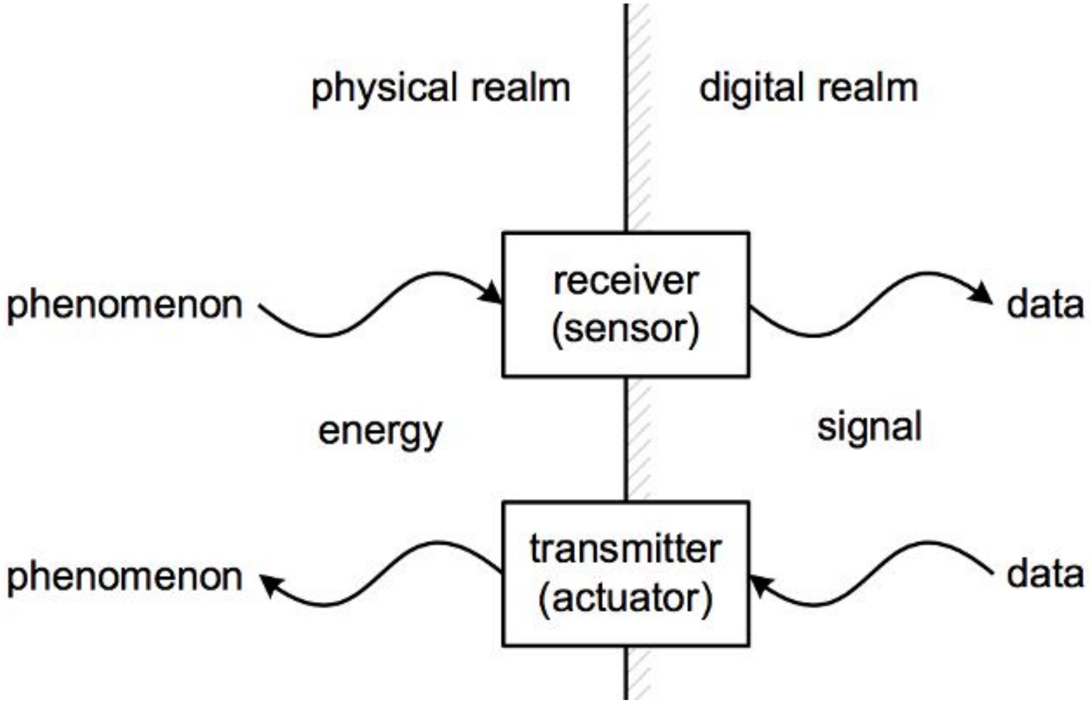
Machine learning

Thank you!

RICKER LYMAN

ROBOTIC

Data is transducers



physical realm
digital realm
phenomenon
receiver
(sensor)
data
energy
signal
phenomenon
transmitter
(actuator)
data