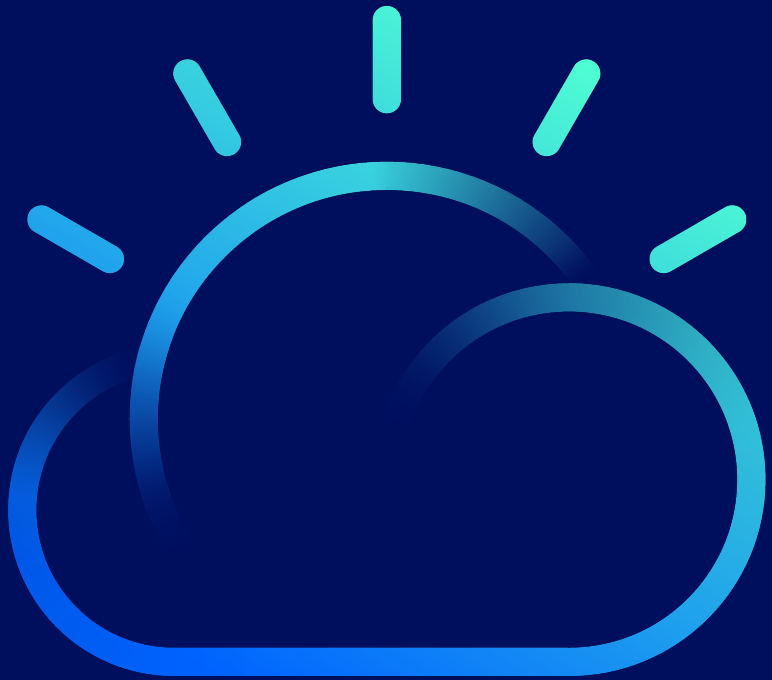
Successful Data Lake Implementation through **Data Integration, Quality** and **Governance** (DIQG)



Jo Ramos
Director & Distinguished Engineer
IBM Data & AI

John Van Buren
BUE, NA Analytics Technical Sales
jvb@us.ibm.com

David Nelson
WW Unified Governance & Integration Sales
nelsond1@us.ibm.com

IBM **Cloud**

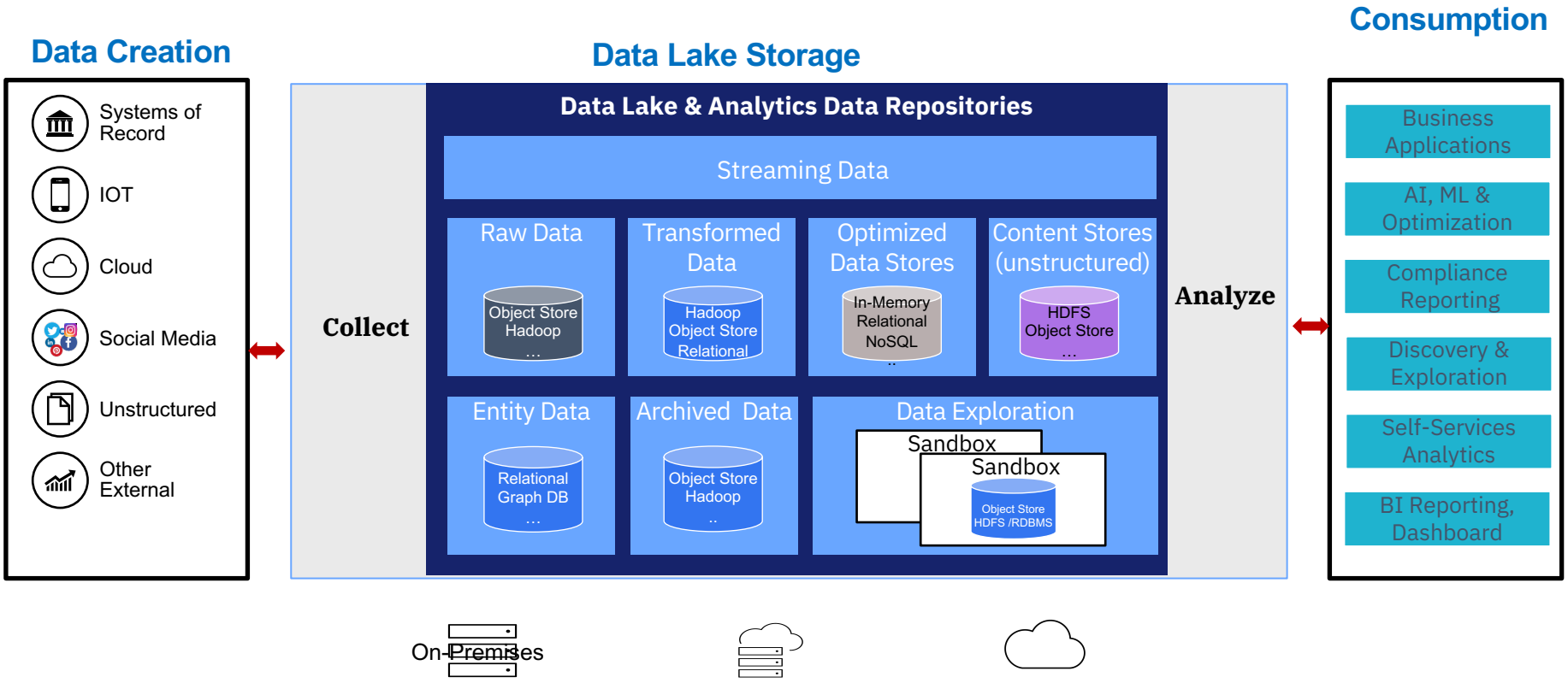
Session Outline

- Introduction
- Gartner: How to Avoid Data Lake failures
- Data Lake Challenges & Requirements
- Data Lake Implementations

What is a Data Lake ?

A data lake is a system or repository of data stored in its natural format, usually object blobs or files. A data lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, analytics and machine learning. **Wikipedia**

The Data Lake Repositories



Introduction

- The major driver for a Data Lake is to store, process, and analyze large volumes of data at much lower cost.
- If you don’t get the data part of the data lake right, you won’t get the ROI part of the data lake right.
- If you don’t get the ROI, why bother?

Six years ago, Gartner Group told us that Hadoop was not a Data Integration Solution

"Although many Hadoop projects perform ETL workstreams, *Hadoop lacks the necessary key features of commercial data integration tools.*"

- Hadoop is not a data integration solution

Gartner Group Research Note, January 2013

Today, Gartner Group explains why data lakes are failing

*"**Metadata** management, data **quality**, data **lineage** and data **integration**, among other things, are **crucial** prerequisites for a successful **data lake**. They cannot be afterthoughts."*

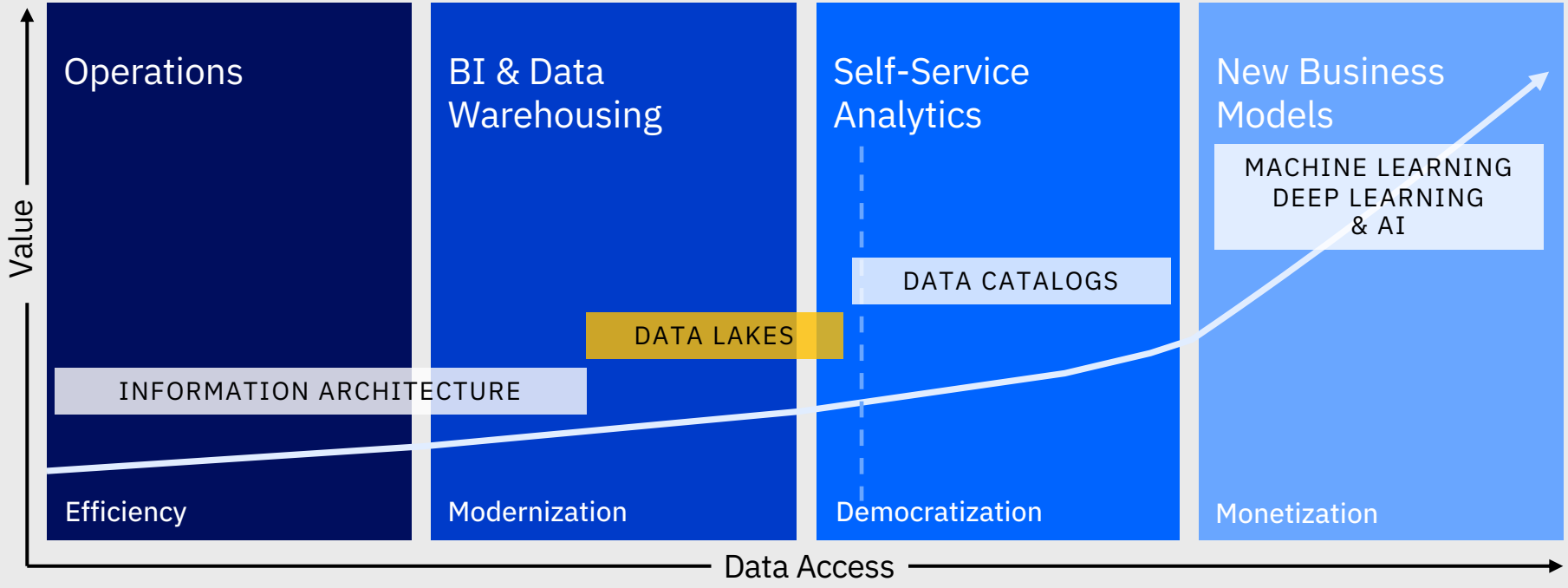
- How to Avoid Data Lake Failures

Gartner Group Research Note, August 2018

Reliance on Hadoop, open source tooling, and hand-coding for DIQG is the source of many failures

Organizations are racing to unleash the power of data and apply AI & Analytics

IBM Data & AI provides the path to monetize your data

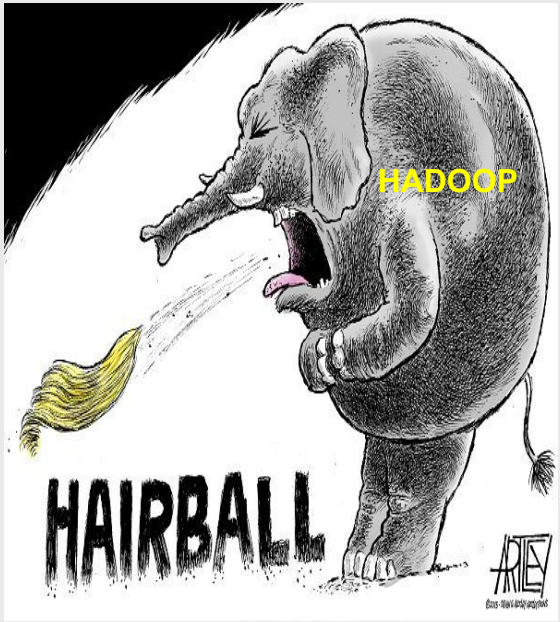


Metadata Management
Collect, Understand and Govern

Asset Catalogs
Share, Activate and Curate

Why do organizations adopt open source tooling and hand coding for DIQG if this approach will cause data lake projects to fail?

- Elimination of commercial software without ROI consideration
- Belief that Hadoop & Open Source (hand-coding) solves all problems
- Not being prepared for the complexities in implementing DQIG
- Hand-coding appears sufficient before scaling



Why Organizations Struggle with Data Lakes

There is a gap in most enterprises

Impacts:

- Increasing cost and complexity
- Manual data gathering and reporting
- Lack of visibility
- Inconsistent and inaccurate information
- Inability to provide complete view of customers
- Limited flexibility and responsiveness
- Lack of appropriate information for decision making

Data Curation is critical to Knowing your Data, so that you can monetize it … at scale

Messy & Unclassified; Hard to access & use

Discover
Ingest
Classify Organize
Integrate

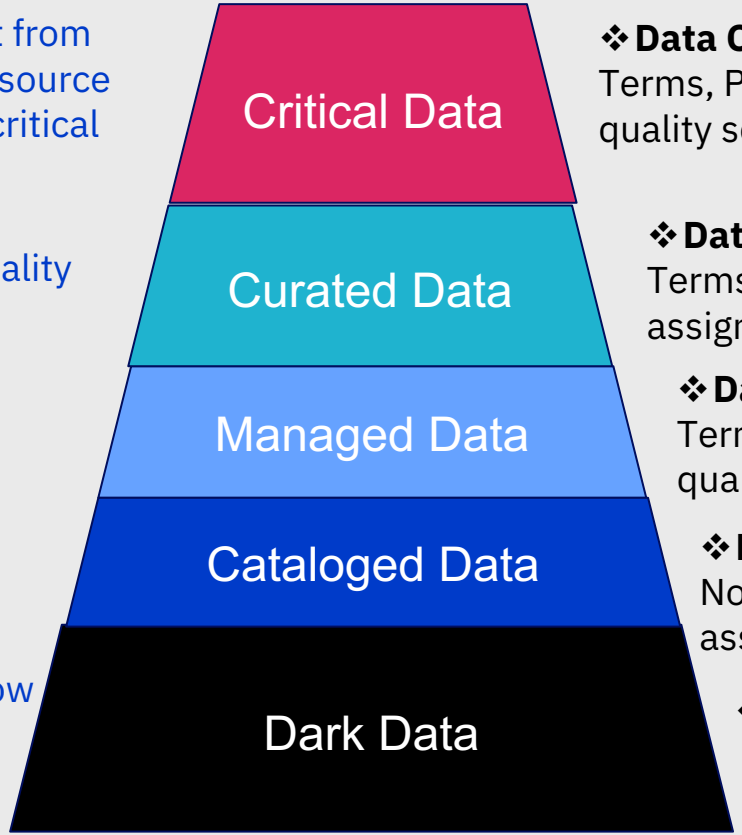
Automate with ML

Organized; Ready for use

Knowledge Catalog
(Insights about your information Assets)

The Path Data Curation

- ✓ Event Driven actions , benefit from tracking changes at the data source as well as notification when critical attributes change
- ✓ Batch process to evaluate quality score and re-classify
- ✓ Data Quality Analysis , Actionable Rules
- ✓ Typically no Data Quality
- ✓ Connect and catalog: Shallow discovery



❖ **Data Continuously Monitored**
Terms, Policies & Stewards assigned, quality score computed

❖ **Data Monitored Periodically**
Terms, Policies & Stewards assigned, quality score computed

❖ **Data may or may not be monitored**
Terms, Policies & Stewards assigned, quality score computed

❖ **Metadata Ingested**
No quality analysis some term assignment

❖ **Data is not cataloged**

Conceptual Architecture for Data & AI

Data Creation

- Systems of Record
- IOT
- Cloud
- Social Media
- Unstructured
- Other External
- Logs

Data Storage & Curation

Collect

Data Lake & Analytics Data Repositories

Streaming Data

Raw Data	Transformed Data	Optimized Data Stores	Content Stores (unstructured)
Object Store Hadoop ...	Hadoop Object Store Relational	In-Memory Relational NoSQL ..	HDFS Object Store ...

Entity Data	Archived Data	Data Exploration
Relational Graph DB ...	Object Store Hadoop ..	Sandbox / Sandbox: Object Store HDFS /RDBMS

Analyze

Data Usage

- Business Applications
- AI, ML & Optimization
- Compliance Reporting
- Discovery & Exploration
- Self-Services Analytics
- BI Reporting, Dashboard

Organize

Data Integration & Governance Fabric

- Security & Authorization
- Data Integration
- Data Governance
- Data Quality Management
- Entity Management
- Data Curation
- Self-Services Data Prep
- Knowledge Catalog

On-Premises | Private Cloud | Public Cloud

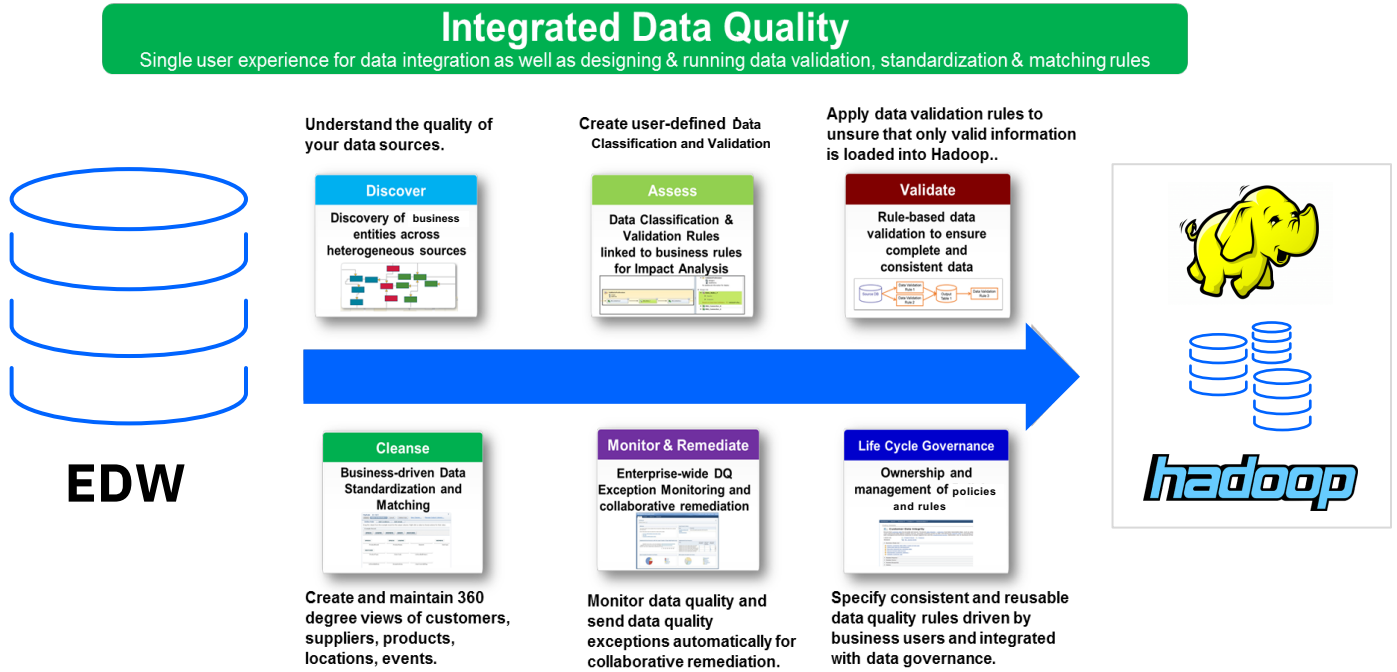
Six Considerations for Success in Data Lake Implementations

#1 – Clarity of Purpose

- Identify clear use case with business sponsors
- Establish context for data
- Avoid broad implementation
- Establish targeted ROI

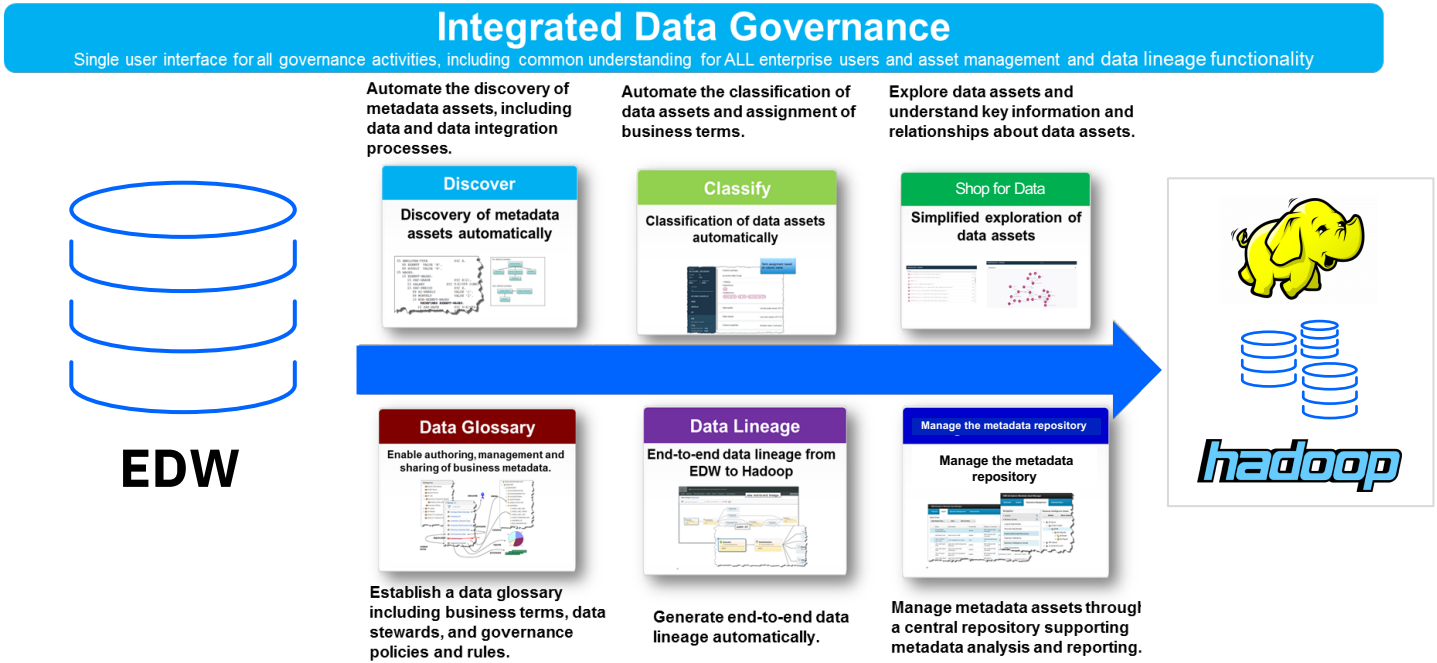
“A data lake is just infrastructure, not a substitute for a strategy encompassing objectives, stakeholders, outcomes, metrics and risks” - How to Avoid Data Lake Failures, *Gartner Group Research Note, 8/18*

#2 - Data quality must be infused, not an afterthought



"Implement automated data quality standards as part of the process of ingesting data into your lake."

#3 – Additionally, **data governance** needs to be a foundational consideration



#4 - Hand coding requires 10X more manual effort

Cost of hand coding vs market-leading tooling

Largest companies learned the hand-coding challenge years ago

Handcoding / Legacy

30 man days to write
Almost 2,000 lines of code
71,000 characters
No documentation
Difficult to re-use
Difficult to maintain

Failed

87%

Saving in dev costs

Success

DI Tooling / Info Server

2 days to write
Graphical
Self documenting
Reusability
More maintainable
Improved performance

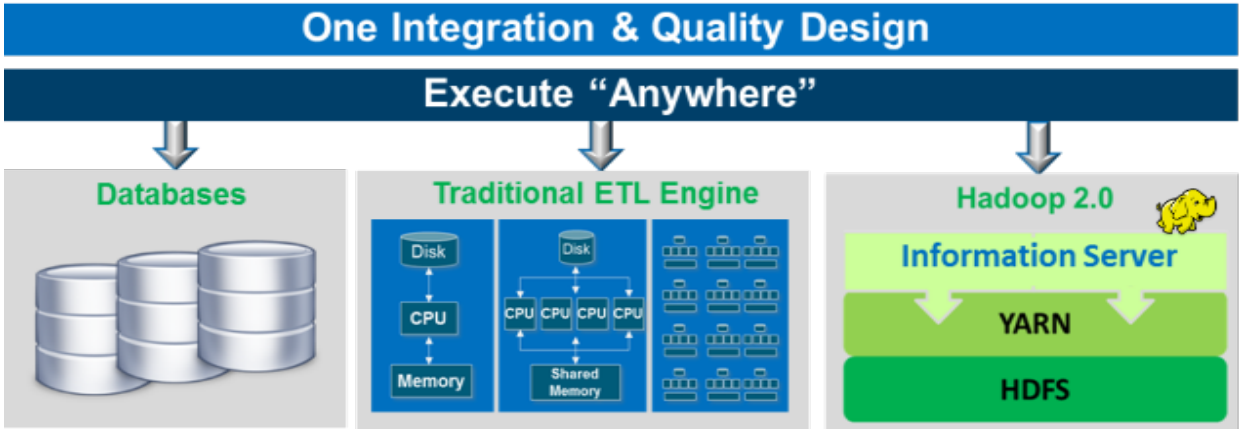
** Pharmaceutical Customer example*

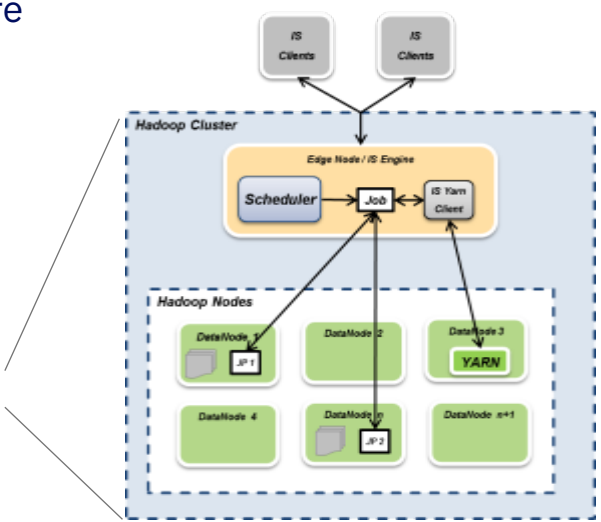
See a comparison of Information Server versus open source tooling: **ITG: Business case for enterprise data integration strategy:** Comparing IBM InfoSphere Information Server and Open Source Tools (2013)

#5 - Build Once and Run Anywhere at Scale – Reduce Duplication

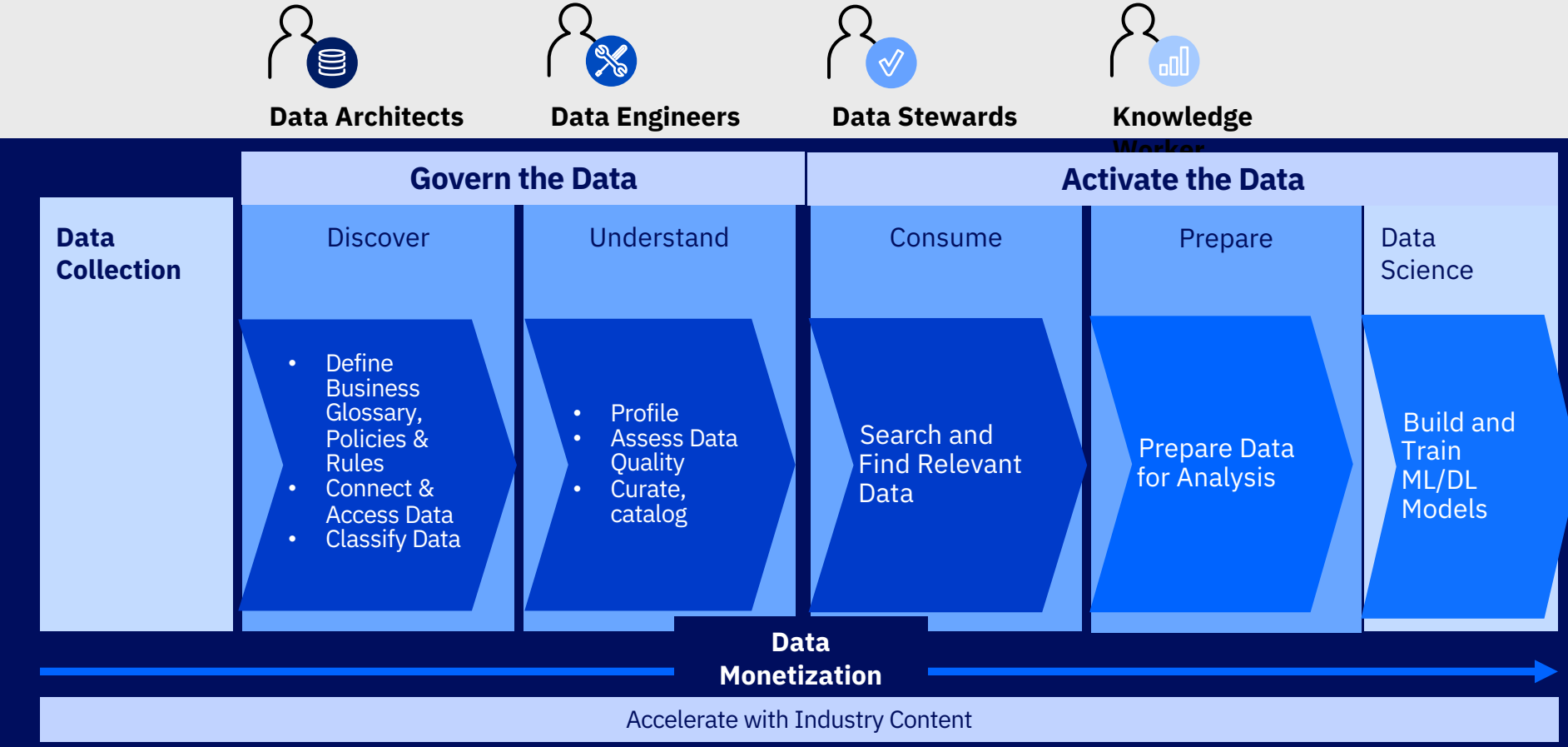
Build a job once and run it in the EDW, in the ETL grid, and in Hadoop without modification while using existing developer skills and ETL assets

- Coding software to exploit an MPP architecture is nearly impossible
- Requires shared nothing, massively parallel processing architecture
- No upper limitation on throughput and processing performance

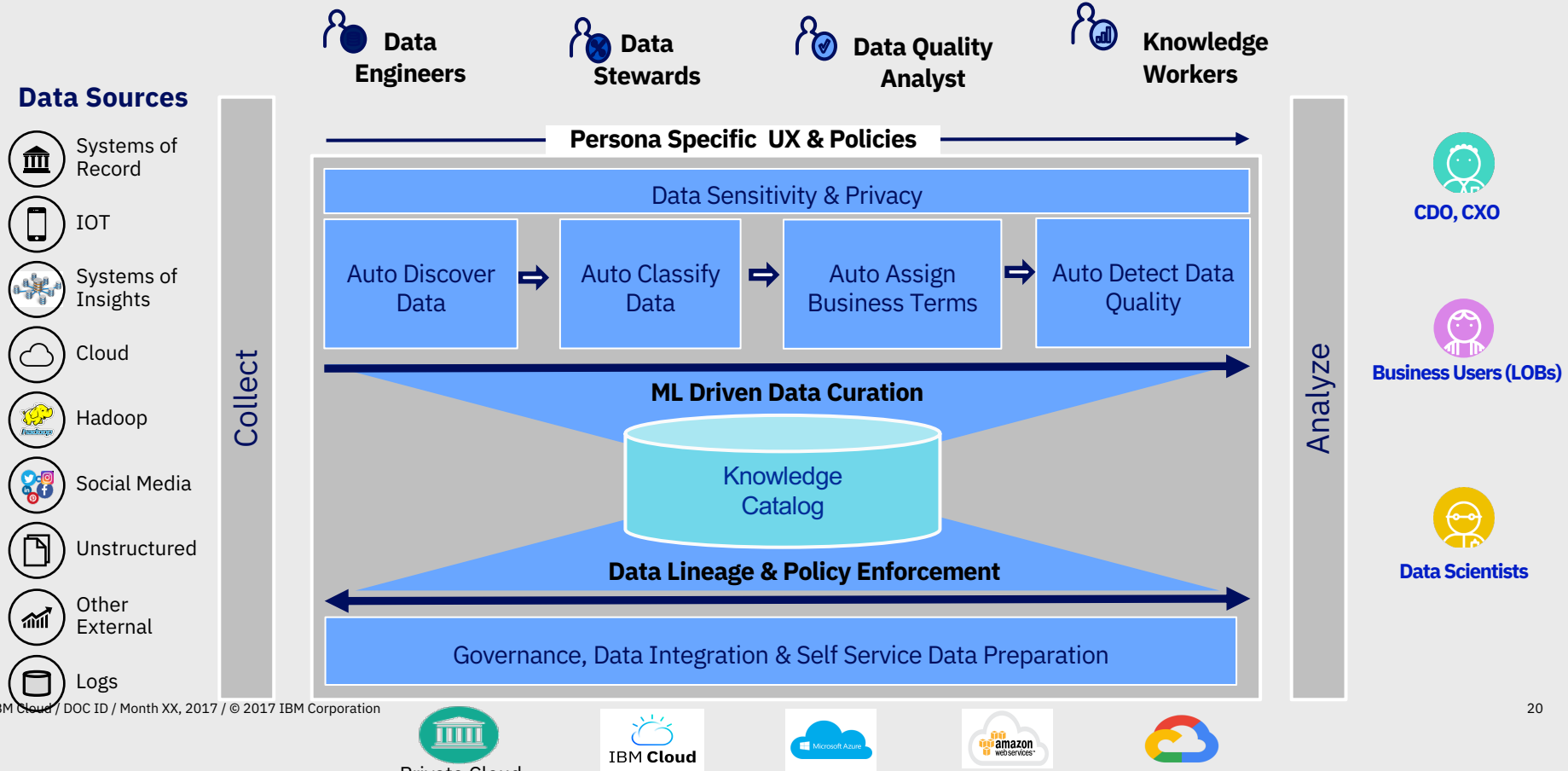




#6 Self-service must be integrated with the rest of the enterprise



Data Quality/Curation/Governance is Essential, yet Intensely Manual without Automation



Examples of Unsuccessful Data Lakes due to DQIG

One of Asia's largest Telcos

- Provides mobile, fixed line, and data center services
- Used commercial ETL >10 yrs
- 2015 decision to sunset DataStage and migrate to a new data integration architecture to minimize IT cost.
  - Open source architecture, Hadoop, Spark, and hand coding
- Ran into technical difficulties, performance issues and project delays
- Result: DataStage to replace open source hand coding and Spark

Global US Manufacturer

- Successful ETL deployment for +5 yrs
- In 2015, company launched a data lake initiative that is based on Open Source (including Talend Open Source), hand coding, and Hadoop.
- Focused on data ingestion
  - Ingestion and transformation processes are largely manual and ad hoc
  - No data lineage or data governance
  - Limited information about data in the lake
  - No data quality processing
- All Hadoop projects are on hold

What is common across these unsuccessful data lake implementations?

- Each client moved from data integration software to open source and hand coding for the data lake, primarily to reduce IT costs.
- Each organization focused primarily on data ingestion for the data lake.
- There was no real plan for Quality and Governance as part of the data lake implementation.
- In each case, the organization recognized that reliance on Hadoop, open source tooling, and hand coding was not successful.

"Data and analytics leaders commonly view implementing a data lake as a way to skip building essential data management capabilities, in the belief that the lake will be self-organizing, self-securing and self-governing. This is a mistake." - Gartner, How to avoid data lake failures

Successful Data Lake Implementation

- Multinational European Corp primarily in retail & financial services
- >1000 stores with 40K employees in 60 countries, growing rapidly!
- **The increasing daily volumes of data left the company struggling to gain insight fast enough or frequently enough with *hand-coded* solutions**
- IT team selected data integration and governance tools for their new Infrastructure using BigIntegrate, BiqQuality, Information Governance Catalog

“By using IBM BigIntegrate for Hadoop, we can run data processing tasks that previously took up to 20 days in just 24 hours.”

- Company Spokesperson
Leading European Retailer

Resulting in “Ultra-Rapid, Efficient Data Analysis”

- Speed and granularity of information for insight supporting executive action (super charging new sales & marketing strategies)
- Calculating the affinity between customers and products enabling relevant offers and driving return visits.
- Intra-day insight into inventory positions right-sizing stock with demand

More Leaders in Data: Success Stories

Data Governance

#4804A, #5311A

Build
a data driven enterprise with a governance first data lake

Reduce
risk of loss or misuse of information by ensuring only legitimate and approved purposes

Deliver
real time processing of trusted information for risk reporting and customer service

Eases
burden of compliance with data-retention regulations such as GDPR by identifying, protecting, and tracking use of sensitive data

Transform
governance from a compliance burden to a business opportunity by improving data quality and usability

Monetize
data to improve business outcomes like increased share of wallet through focus on data quality

Additional Success Stories @ Think

#1676A

#3944A

#7582A

Establishing Data Lake Success

- Target a clear use case with sponsorship supporting business strategy
- Recognize the limitations of open source tooling and hand coding for the requisite data integration, quality, and governance needs. Plan beyond the pilot.
- Anticipate need for a formal governance program (data stewardship, catalog, lineage..)
- Leverage automation/machine learning of governance and quality efforts to accelerate adoption and ensure ROI

How to Avoid Data Lake Failures

Gartner Group Research Note, August 2018

Report download:

https://www.ibm.com/account/reg/us-en/signup?formid=urx-35194

Or Contact:

David Nelson
WW Unified Governance & Integration Sales
nelsond1@us.ibm.com