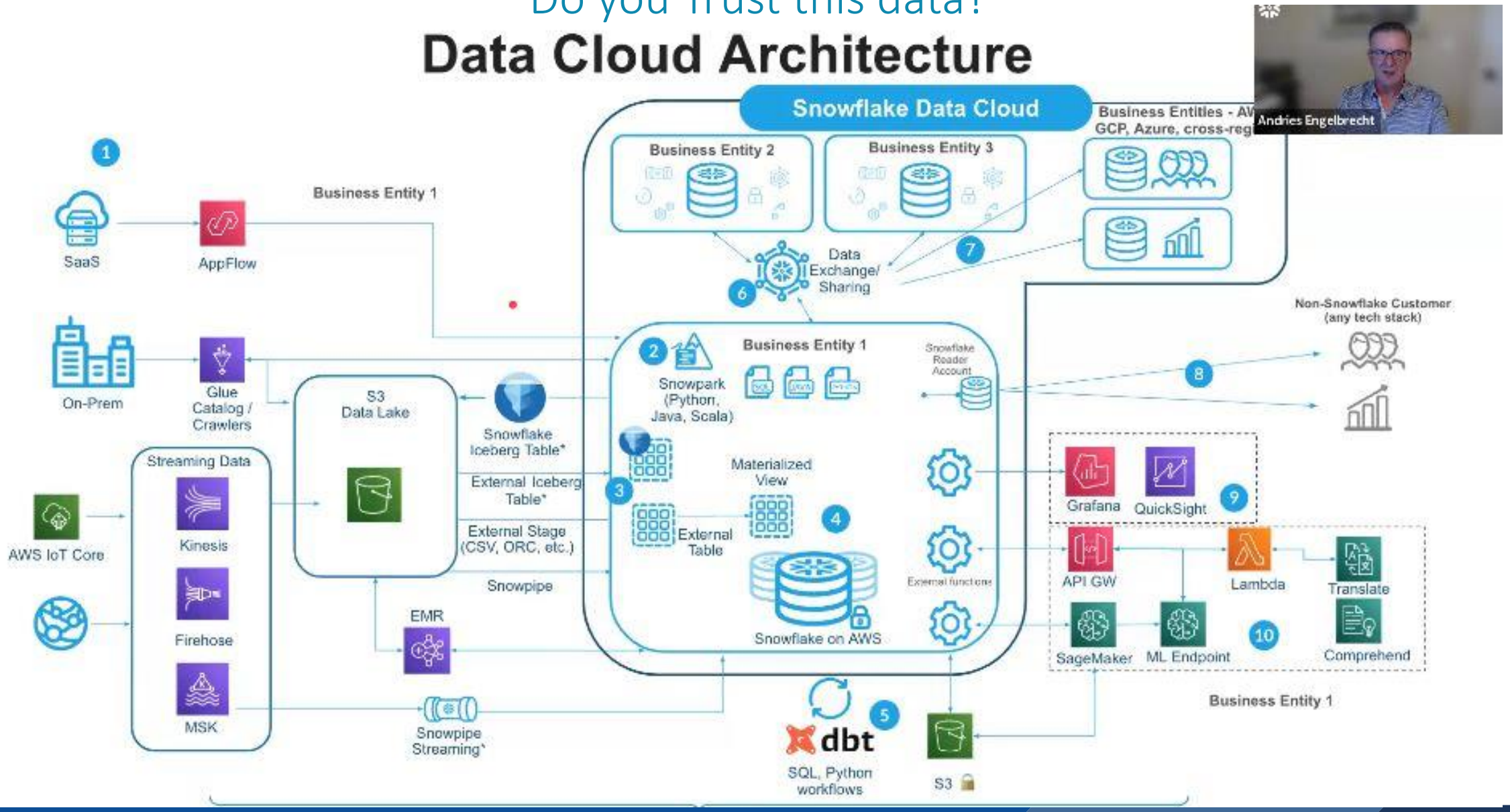
Tricentis on How to Avoid Breeding Bad Bots

Data Integrity for Enhanced Trust

Tricentis
Data Integrity

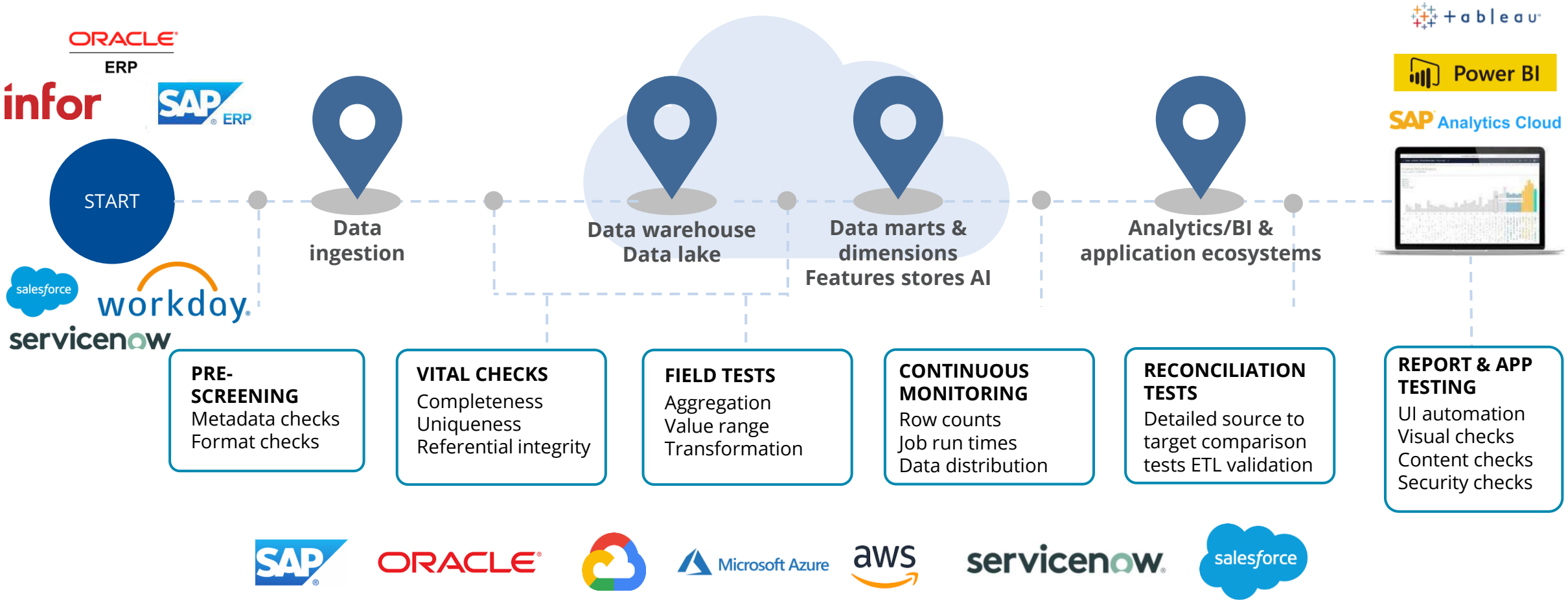
Data Environments are COMPLEX

Do you Trust this data?



Deliver trustworthy data through a Complex Process

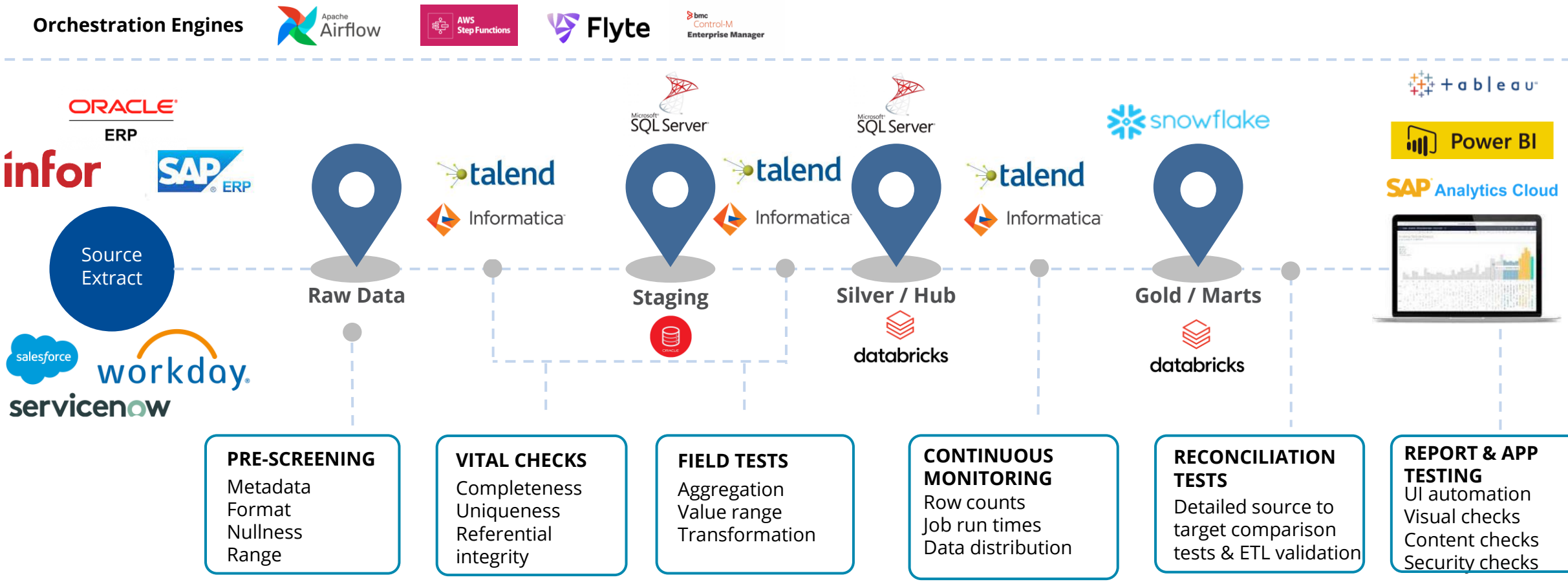
Example scenario: Data Pipeline for analytics and dashboards



Deliver trustworthy data through a Complex Process

Integrate with the tools you already use

Orchestration Engines



PRE-SCREENING
Metadata
Format
Nullness
Range

VITAL CHECKS
Completeness
Uniqueness
Referential integrity

FIELD TESTS
Aggregation
Value range
Transformation

CONTINUOUS MONITORING
Row counts
Job run times
Data distribution

RECONCILIATION TESTS
Detailed source to target comparison tests & ETL validation

REPORT & APP TESTING
UI automation
Visual checks
Content checks
Security checks

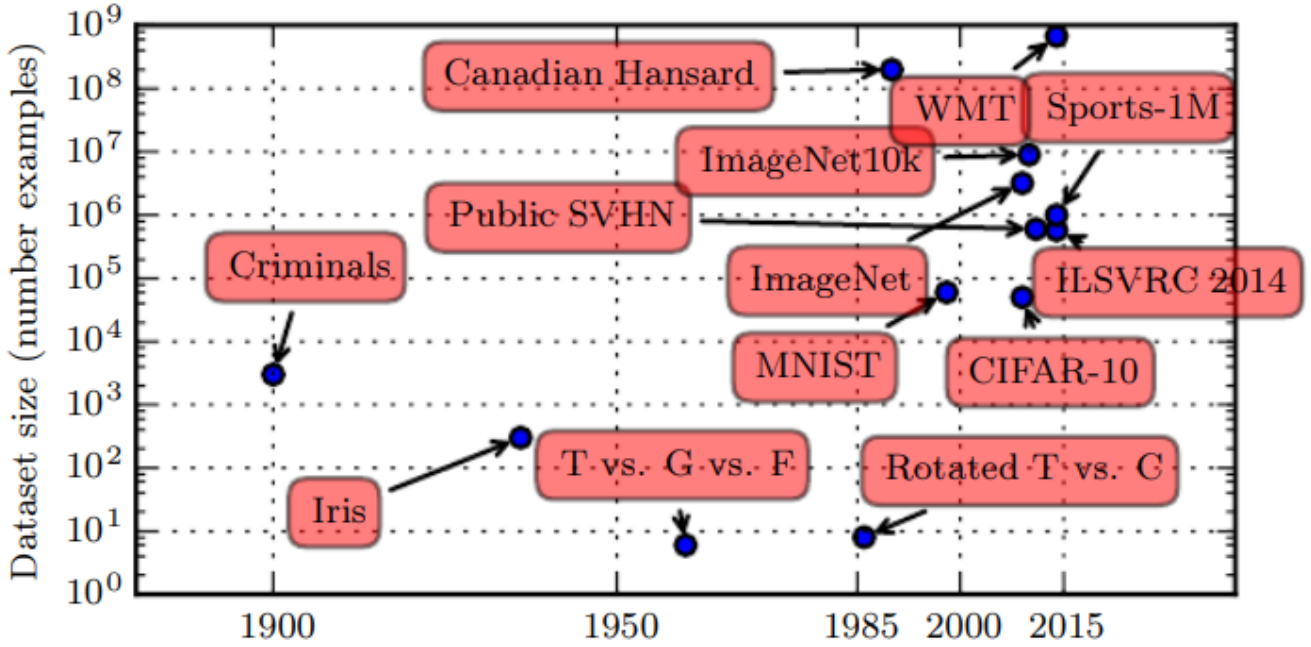
Is your data ready for AI?

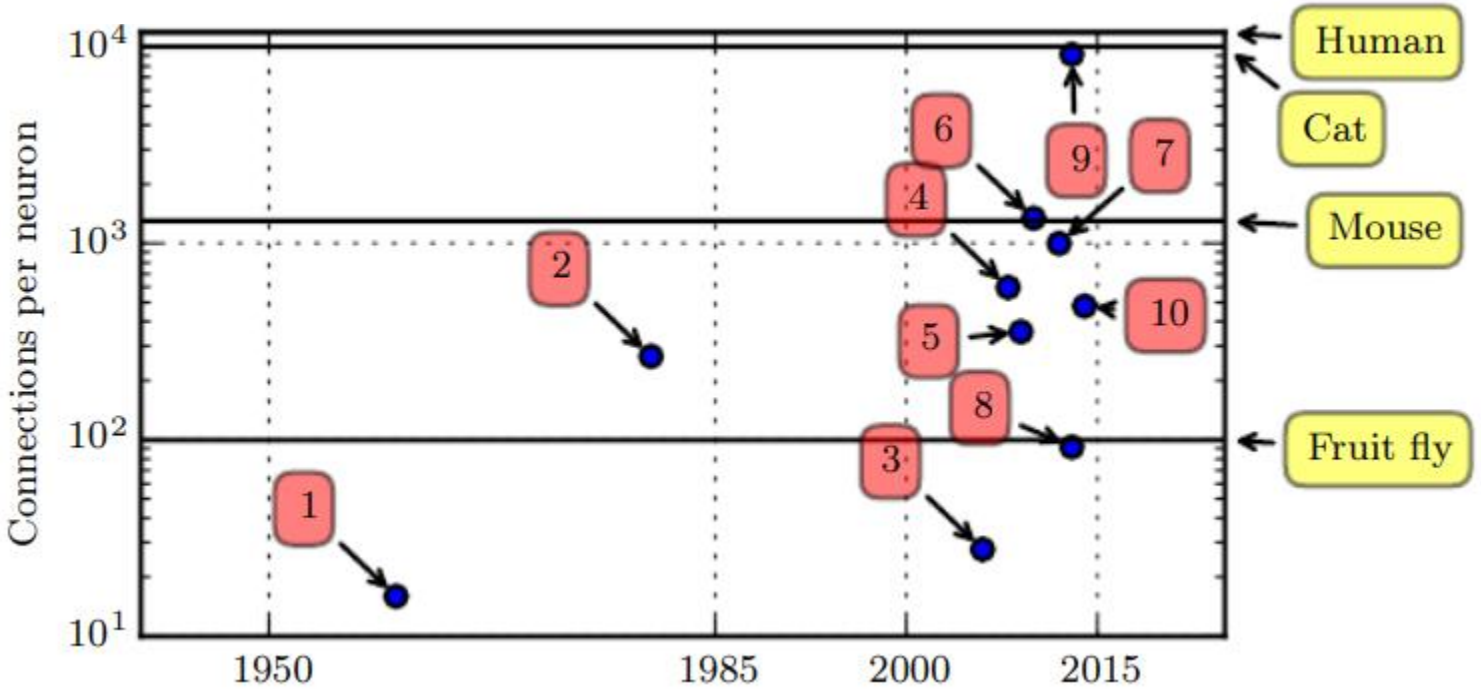
Trust your AI/ML results and deliver trusted data for analytics and AI

AI / ML Why Now?

- Datasets move up in size to astronomical levels Data Explosion for training data catalyst #1

- Neuron connections has exponential levels





Deep Learning
by Aaron Courville, Yoshua Bengio, Ian Goodfellow

Massive Data, Massive Models, Massive potential

Model Sizes	**Dataset Sizes**
Classical Intelligence:	**Standard 2010's:**
80bn Neurons	15k data
2020 AI:	**2020's Fraud:**
1.5bn Neurons	2 million records
2024 AI:	**2024 AI:**
1.35T Neurons	15 Trillion tokens

185m ChatGPT users

Generative AI Inverts problem solving

Defining Problem
Gathering Data
Annotating Data
Building/training model
Creating application
Testing and Validation

ChatGPT

But it can also go pretty badly wrong…

Air Canada must pay damages after chatbot lies to grieving passenger about discount

Airline tried arguing virtual assistant was solely responsible for its own actions

Katyanna Quach Thu 15 Feb 2024 / 21:50 UTC

August 23, 2023 | GT ALERT

EEOC Secures First Workplace Artificial Intelligence Settlement

Related Professionals	Lily M. McNulty
Capabilities	Innovation & Artificial Intelligence | Labor & Employment | Workplace Compliance & Counseling
Offices	Phoenix

Zillow to exit its home buying business, cut 25% of staff

By Anna Bahney, CNN Business

3 minute read · Published 5:36 PM EDT, Tue Nover

Google loses $96B in value on Gemini fallout as CEO does damage control

CEO Sundar Pichai says Google working 'around the clock' to fix AI tool's bias issues

AI is not **Magic**

It is **high quality parrot**

It needs **high quality training data**

AI Benefits

Create a solid data foundation for your analytics and AI projects by feeding them trustworthy data in a simple, efficient and cost-effective way. Use clean, fit-for-purpose data

There are 3 major areas for AI impacts on a business:

1. Improving the customer experience: (Customer Contact)
2. Employee productivity boosted with AI
3. **Optimizing Business Operation's Outcomes:**
    - **Especially for these Domain Use Cases**
        - **Data Migrations (required for new AI Data Architectures)**
        - **Innovation**
        - **Compliance**

Improving the customer experience: (Customer Contact) And Employee productivity boosted with AI (Generative AI)

- Create and Validate Content data integrity with automated, end to end and continuous data validations for your employee's AI processes

Virtual Agents and Chatbots (Data Integrity makes sure they learn from valid and fit-for-purpose data - critical for the edge use cases you want AI to handle)

Personizing for the specific ask /need of the customer

Voice (and Image) Analytics - Better servicing of needs through AI

Optimizing Business Operation's Outcomes (#3)

- **Avoid data chaos!** AI Business tasks must be driven with timely and correct business data. This requires automated data reconciliation and validation to match (Augmented AI)
- **Adhere to governance / compliance standards with AI.** Meet sustainability and efficiency goals by using collected data with proper data integrity.
- DI Specific use cases:
  - RISK : Fraud Detection WITH compliance in doing so.
    - Risk in implementing: Adversarial ML (AML)
  - Predictive Innovation : Business Process (i.e. Supply Chain) optimization
    - Risk in implementing: Capacity planning must be within 98%, so automation is key for data models validation
  - Gen Innovation : Intelligent document processing (Must have valid parsing)
    - Risk in implementing: Volumes are too large to check manually, common practice of throwing out data leads to failure

Optimizing Business Operation's Outcomes – Risk / Compliance

- Example Banking Compliance
  - AML & AML + KYC compliance
    - Anti – Money Laundering and Adversarial Machine Learning
    - Compliance reporting with AI.ML produced data
  - Backdoor attacks can happen purpose or by accident

Integrity of the Data in the Models

Includes:
1. Schema and Metadata Checks
2. Parsing Checks
3. Clean-Label and Backdoor Poisoning

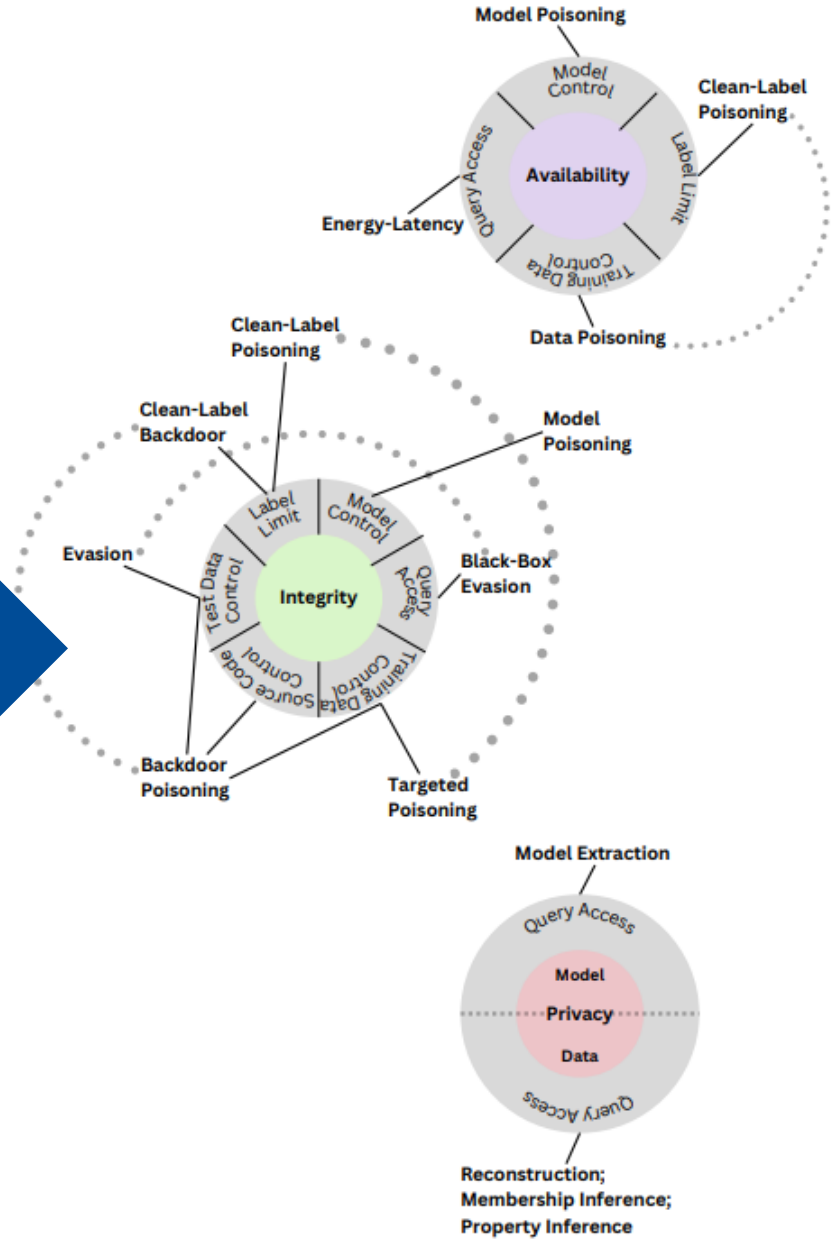


Figure 1. Taxonomy of attacks on Predictive AI systems.

NIST Trustworthy and Responsible AI NIST AI 100-2e2023 Diagram

Compliance Backdoor Attacks Example

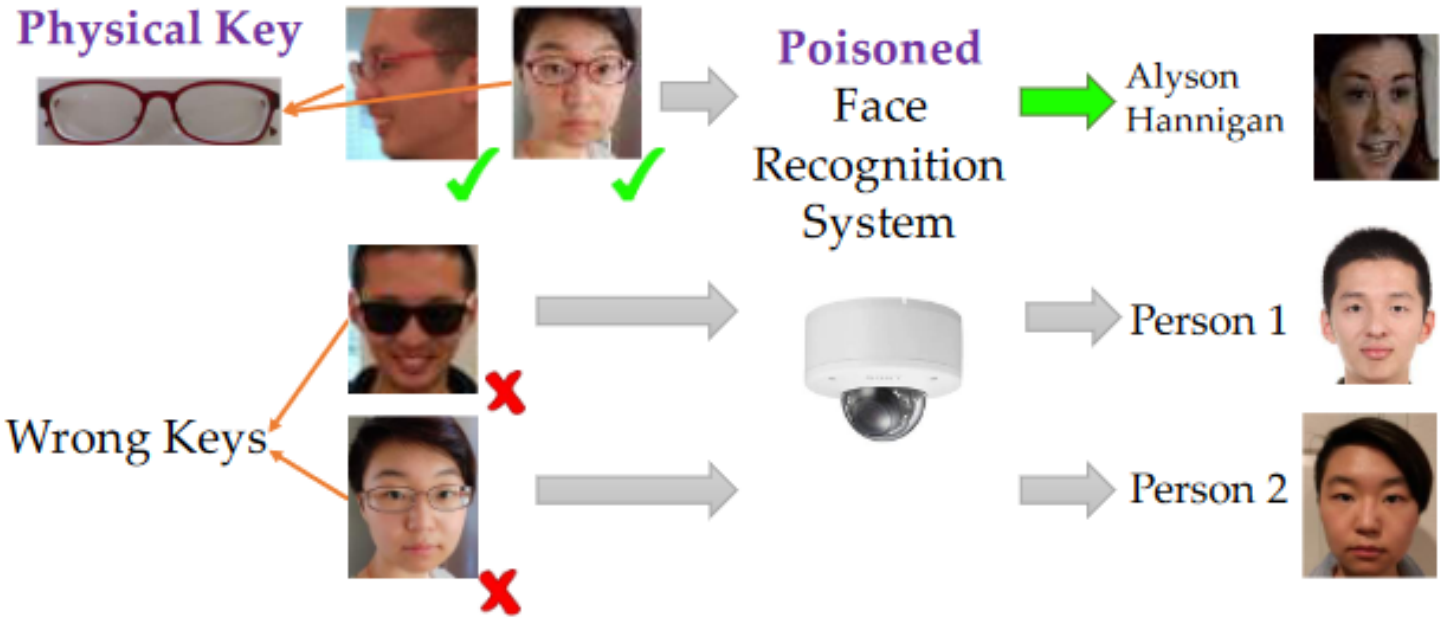


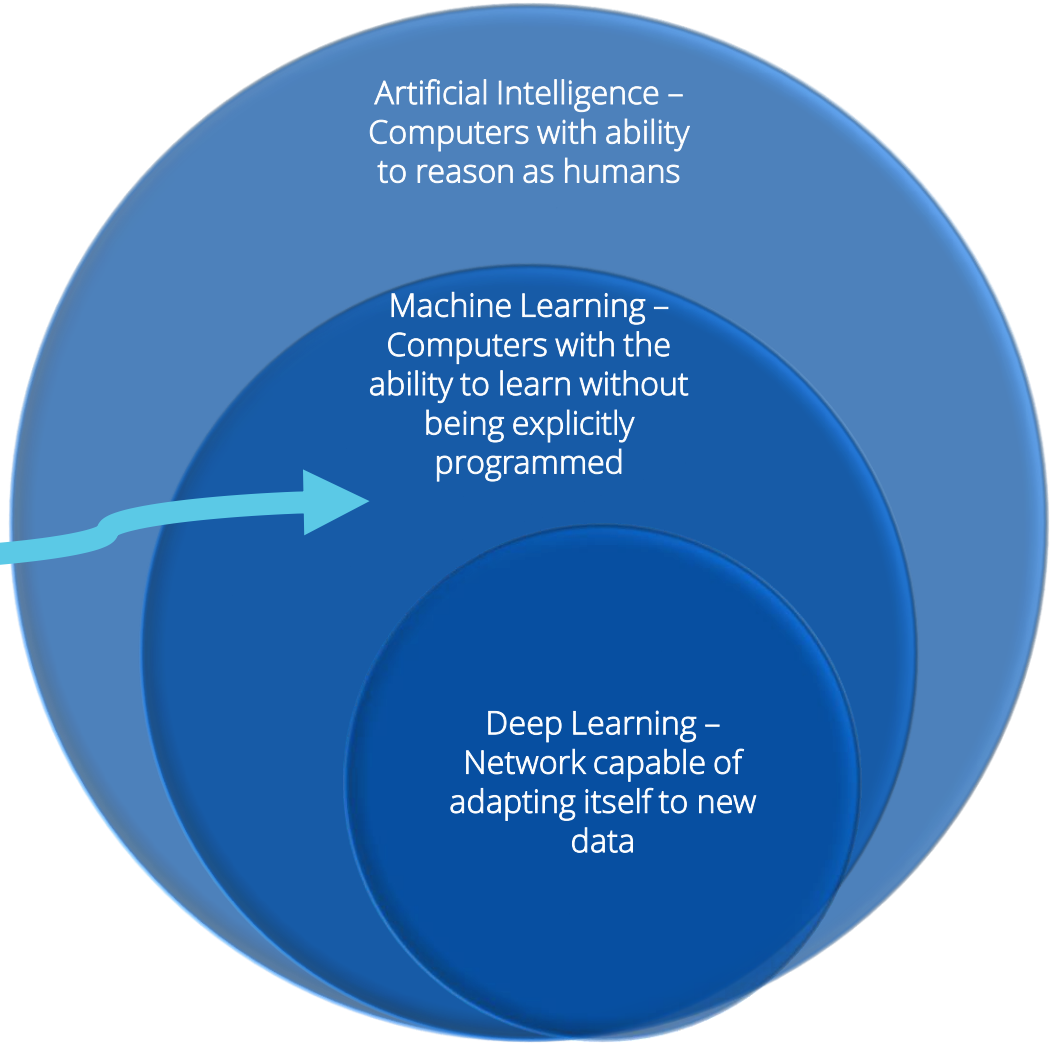
Fig. 1: An illustrating example of backdoor attacks. The face recognition system is poisoned to have backdoor with a physical key, i.e., a pair of commodity reading glasses. Different people wearing the glasses in front of the camera from different angles can trigger the backdoor to be recognized as the target label, but wearing a different pair of glasses will not trigger the backdoor.

An AI/ML Data Integrity Story

- Major Pharma Vaccines Group
  - **Problems** with ML models not leading innovation decisions correctly
    - Non-Curated data used for model training
    - Non-curated data used for model deployment
      - Bias was way off expectations
  - Why?
    - A few folks checking the parsing and completeness (fit for purpose) of gigantic data sets
    - Only manually Spot Checking the data (manual stare and compare) at each stage of ML processes
  - Solution
    - **End to End Validation** to verify the changes to data sets where expected, conforming and regression checked against ALL other data in the ENTIRE process.
    - **Automation** allowed 90% of the data to be checked, even with large datasets
    - **Continuously** checked, with embedding in the AI/ML Azure processes, Databricks and PowerBI tooling.
  - Results!
    - $1B Vaccine at market success!

Where we fit in AI/ML

Data Integrity capabilities in this context our target today.
-Focus on Machine Learning
- Focus on Predictive Models with Supervised Learning



Discriminate Models are:
- Regression
- Classification
- Logistic Regression
- Support Vectors Methods
- Convolutional Neural Networks
- Reinforcement Learning
- Federated Learning
- Ensemble Learning
- xgBoost

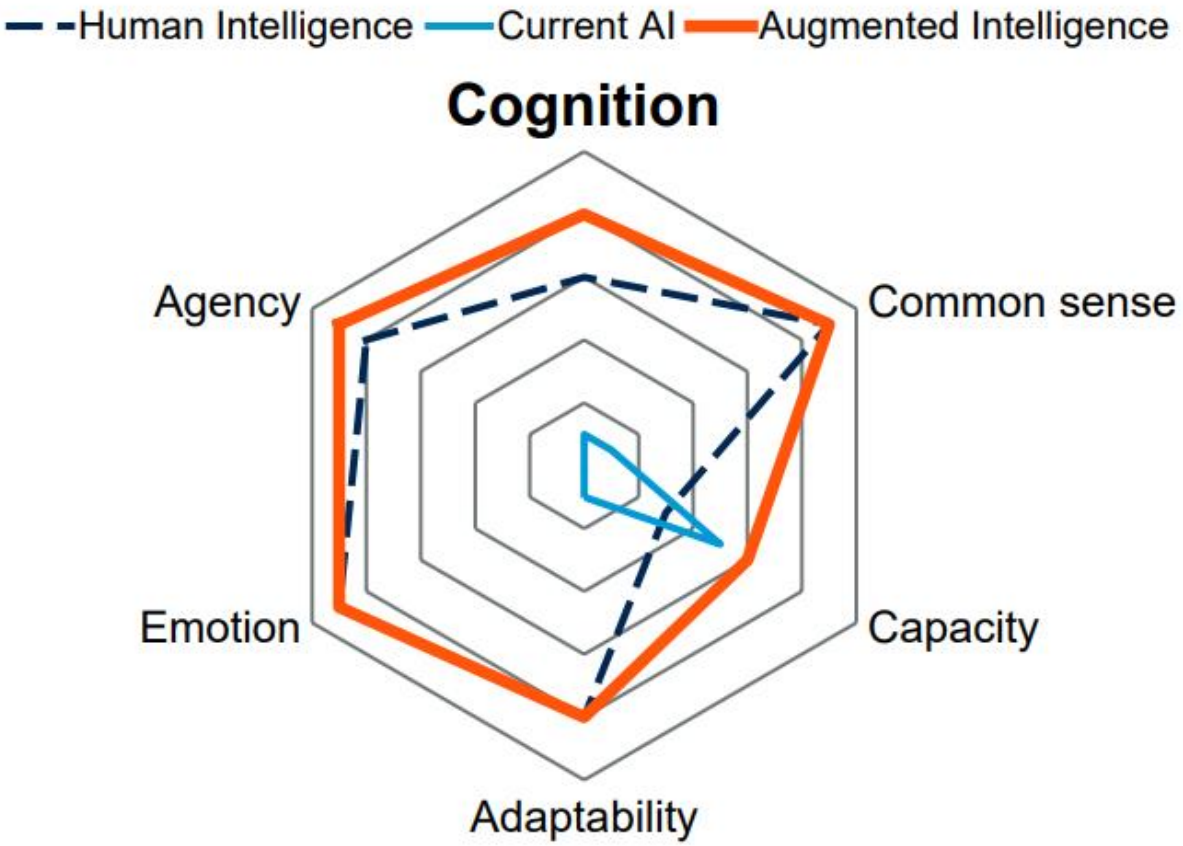
Points: -> Labeled data most important here - Supervised

Generative Models are:
- Gan – Gen Adversarial Models
- LLM – Large Language Models

-> Unlabeled data drives these, unsupervised

Augmented Intelligence: Complementing Human Strengths Using Automated Data Integrity as the Humans

Time frame	2021+
Likelihood	Certain (already happening)



19 © 2022 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Gartner

Business Driver is Business Value

Business Development & Data Science | AI/MLOps & DevOps | Data Engineering

Business Drivers → Data Science → Model Versioning → Model Deployment → Deployment Data Pipelines

Cases & Value (Concepts to Business Value)

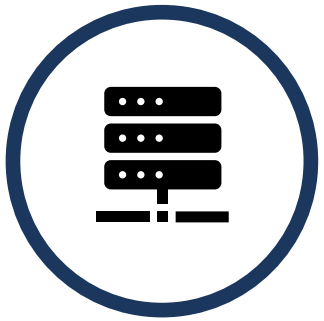
The deliverable is **Business Value**. This assumes you have the Curated Data that is Fit for this Purpose. (Without data integrity this can not be achieved)

Source: Bughin et al. McKinsey Global Institute, *Artificial intelligence*, June 2017.

Data Ecosystems MUST be Trusted

Business Development & Data Science | AI/MLOps & DevOps | Data Engineering

Business Drivers → Data Science → Model Versioning → Model Deployment → Deployment Data Pipelines



Data Ecosystems (Data that the Concepts use)

Complex Data Ecosystems (On-prem and Hybrid) must provide this data with integrity

Source: Bughin et al. McKinsey Global Institute, *Artificial intelligence*, June 2017.

Models Rely on Trusted Data

Business Development & Data Science | AI/MLOps & DevOps | Data Engineering

Business Drivers → Data Science → Model Versioning → Model Deployment → Deployment Data Pipelines

Data will lead in any process supervised or not, the training of the Models

Techniques & Tools

These require a hypothesis based on algorithms that must include bias and bias is impacted by data without integrity

Source: Bughin et al. McKinsey Global Institute, *Artificial intelligence*, June 2017.

Elements of AI Transformation

Business Development & Data Science | AI/MLOps & DevOps | Data Engineering

Business Drivers → Data Science → Model Versioning → Model Deployment → Deployment Data Pipelines

End to End, Automated and Continuous Observability into Deployments and Pipelines, will allow TRUST in the integration of your processes and enable business value delivery with low RISK

Workflow Integration

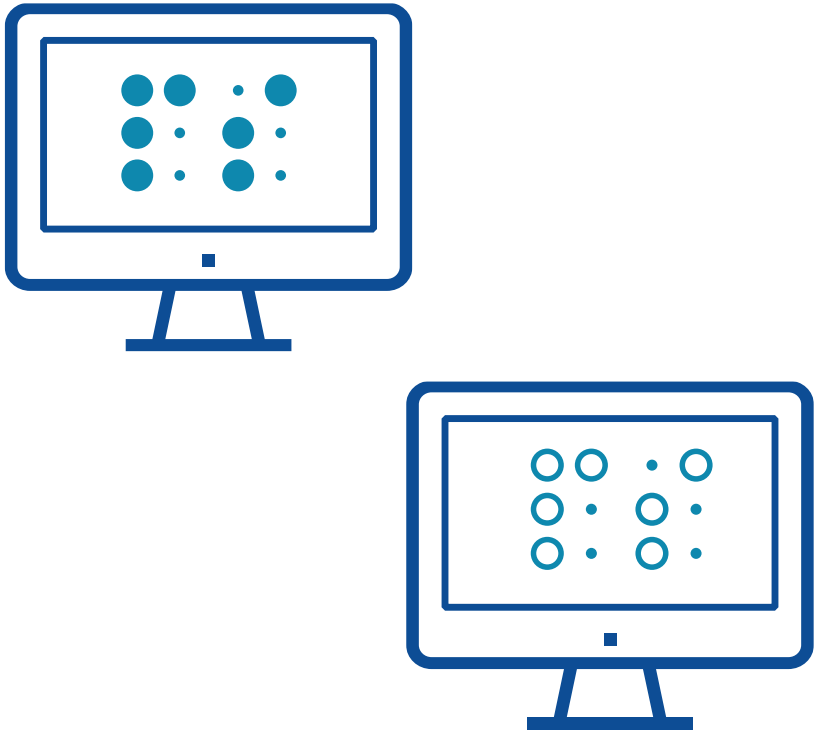
Source: Bughin et al. McKinsey Global Institute, *Artificial intelligence*, June 2017.

Manual “stare and compare” is slow and doesn’t scale.

And is not a great use of your team’s brainpower.

They are Data Scientists and Engineers not Janitors

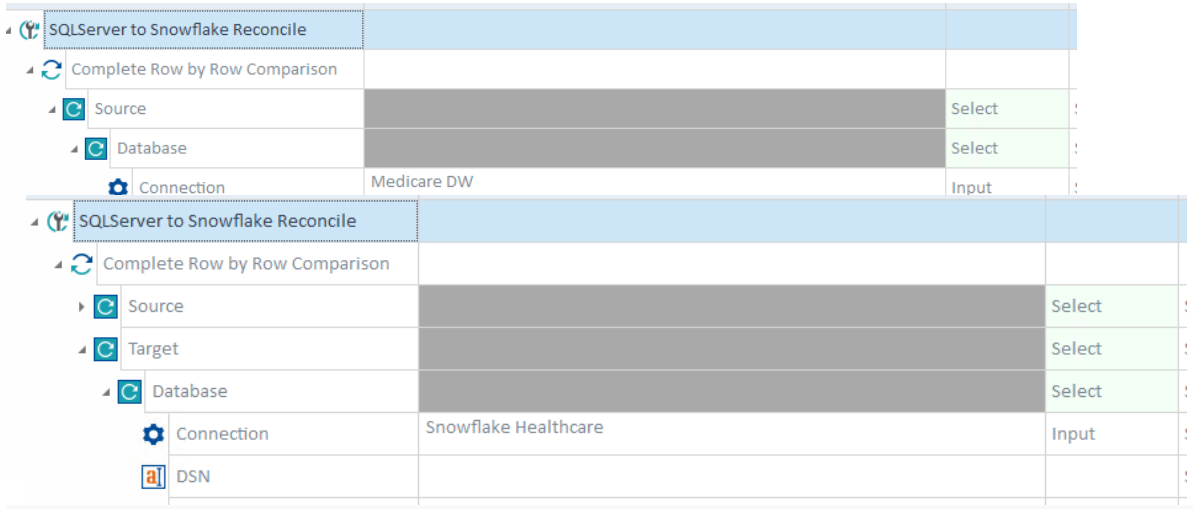
10^9th power is 1bil records! Years to manually check!

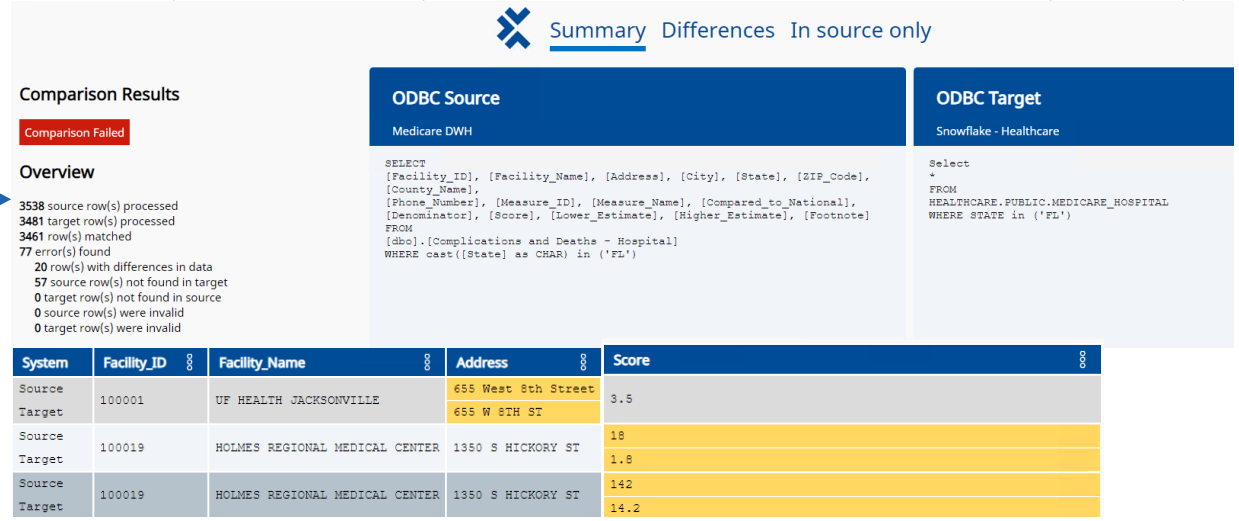


Heterogeneous Reconciliation Reporting

- Pinpoint Data errors faster with concise **source to target reconciliation** reports.
- Highlight missing rows in source and target as well as **column level mismatches.**
- **Compare millions of rows in minutes** using in memory database technology.
- **Export Reconciliation Reports** as HTML, PDF, or detailed SQLite database for Audit review.
- Compare **heterogeneous data sources**; including Parquet, AVRO, ODBC, JDBC, and OLAP.

Tricentis

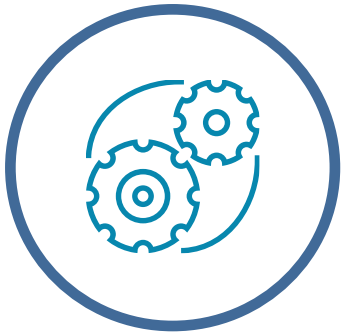




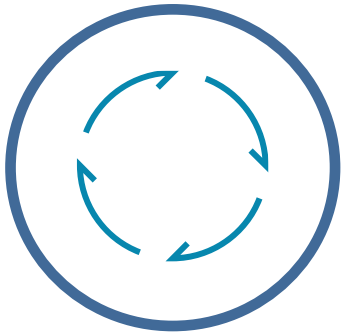
Includes data, UI, and API testing for any data type — across your entire landscape.

Required to ensure data Integrity Trust → A data TESTING solution that's…

End-to-end



Automated



Continuous

How DI works

to continuously improve your data

10m's
rows tested in
< minute

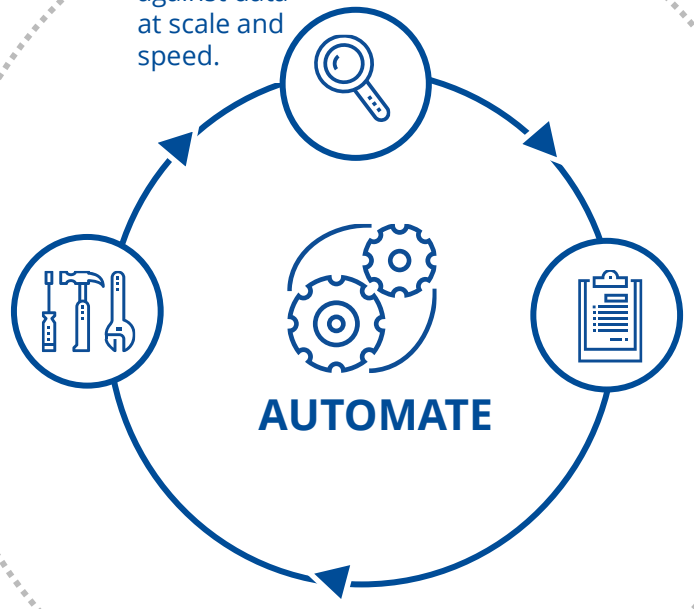
RUN
against data
at scale and
speed.

Get detailed,
actionable reports.

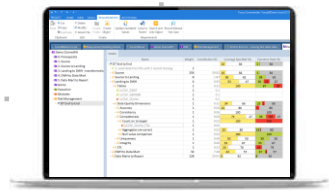
Collaborate across teams to build better tests—end-to-end

No-code, low-code tool
No expert SQL skills required

CREATE
model-based
data tests.



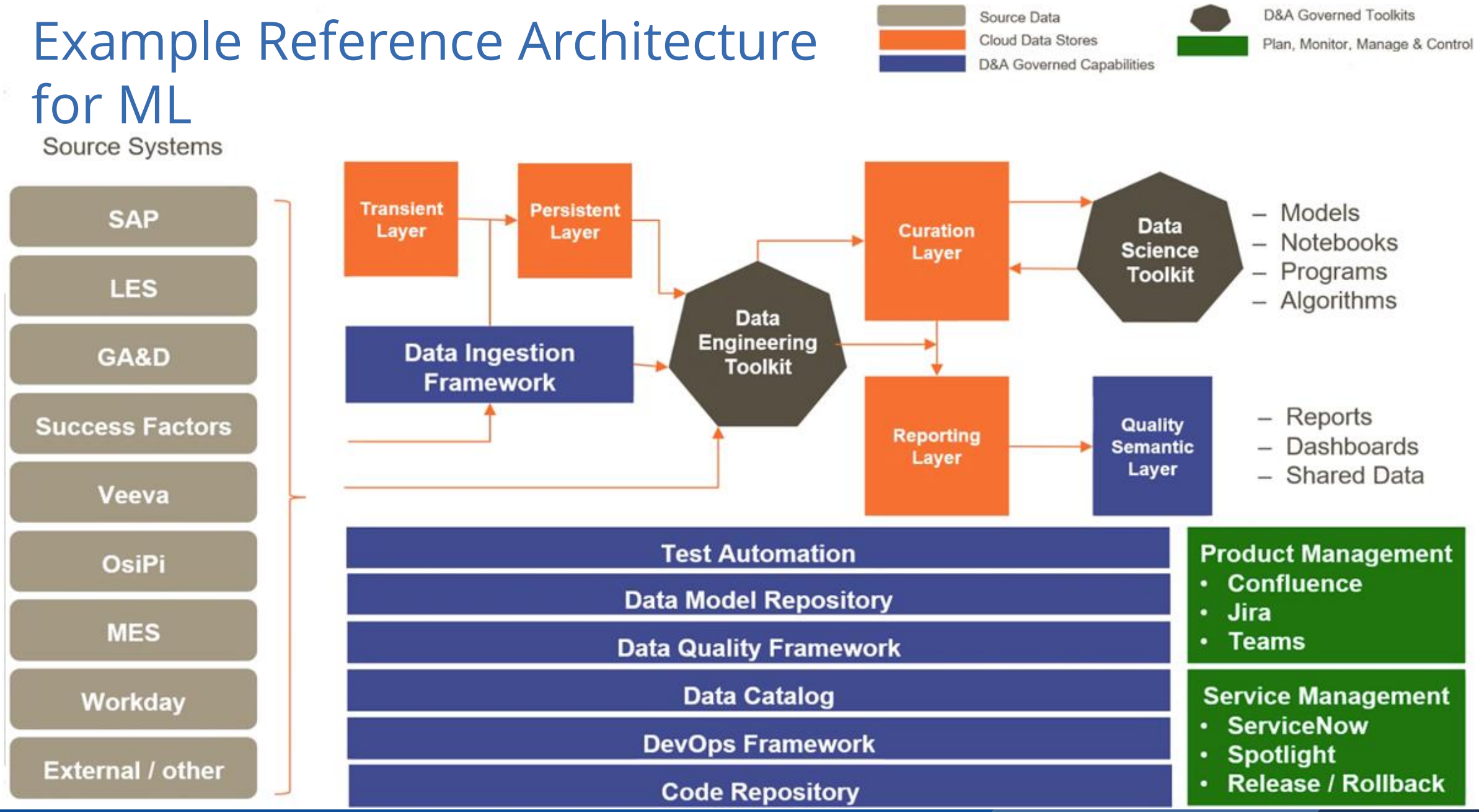
REPORT
Pinpoint
the root cause
of problems.



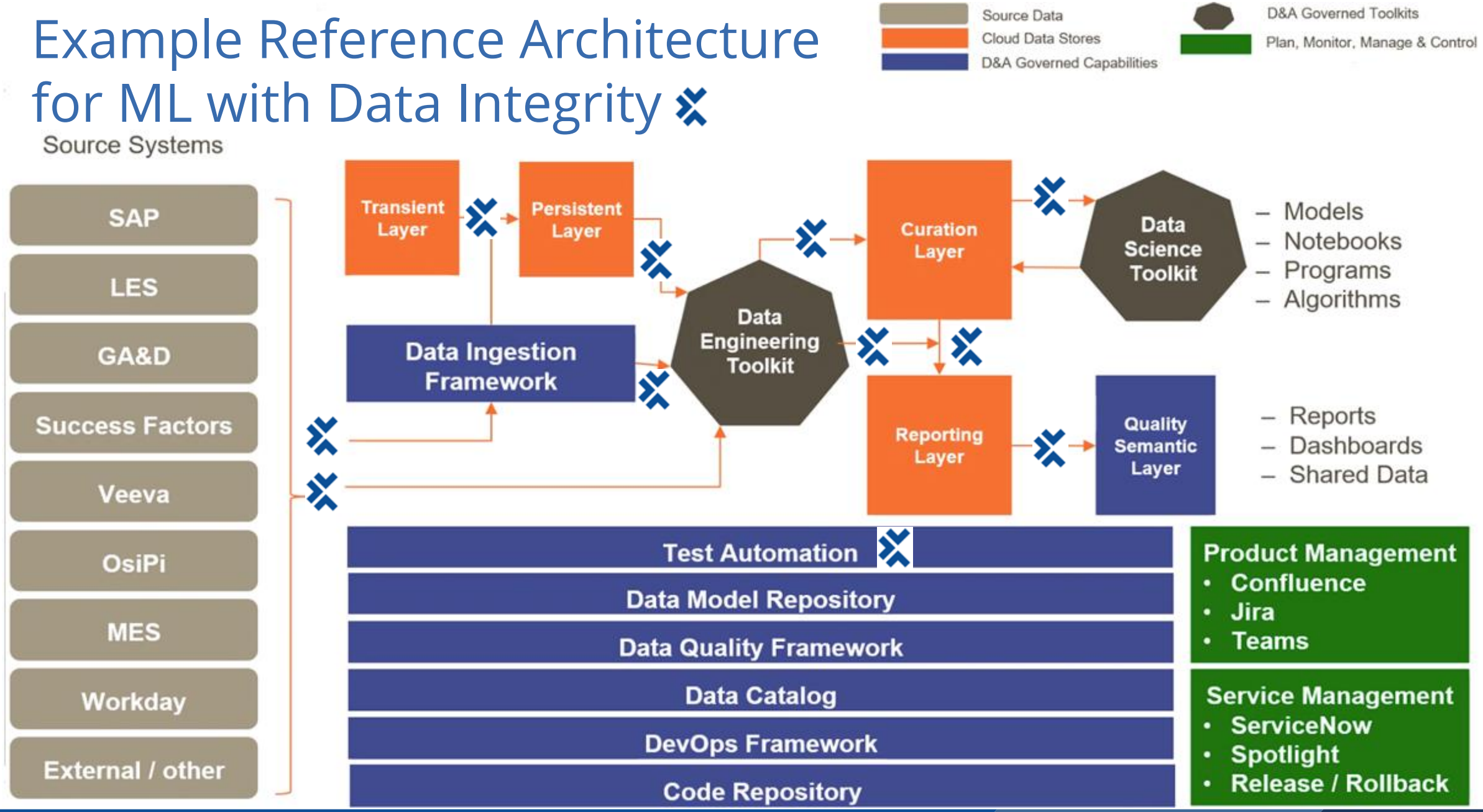
FIX

ML Example

Example Reference Architecture for ML

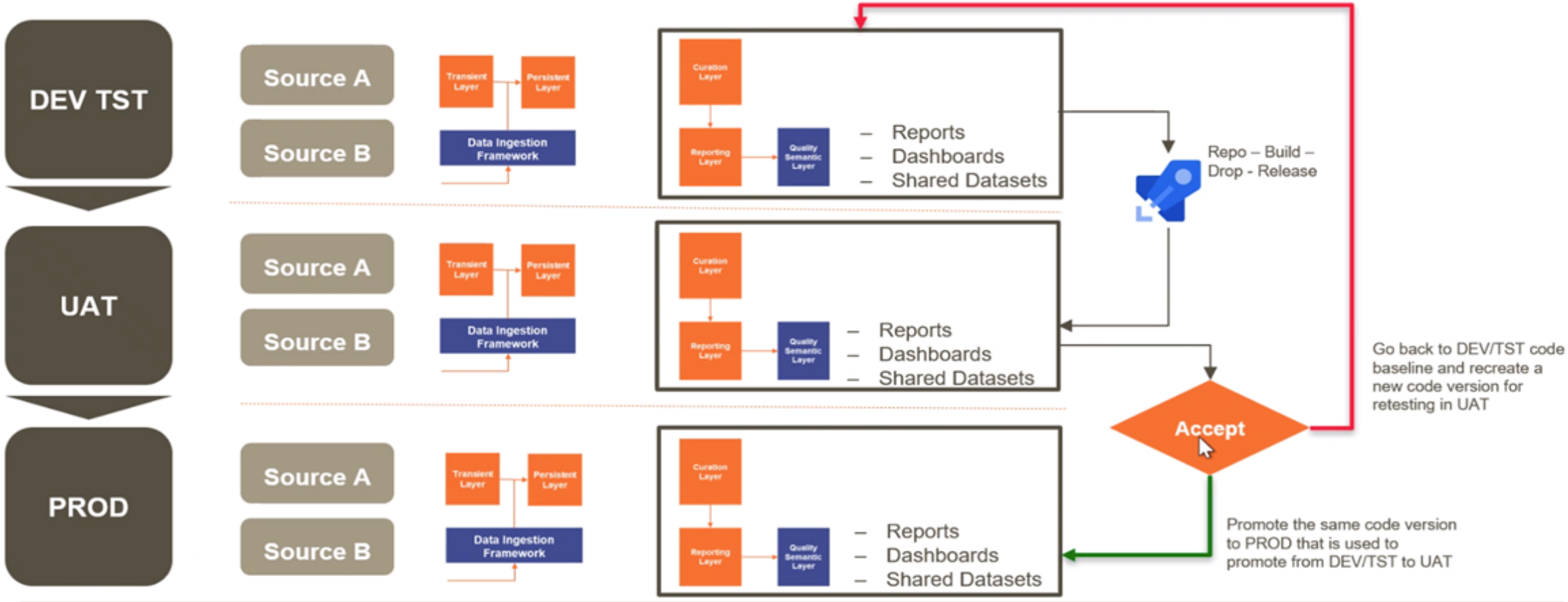


Example Reference Architecture for ML with Data Integrity



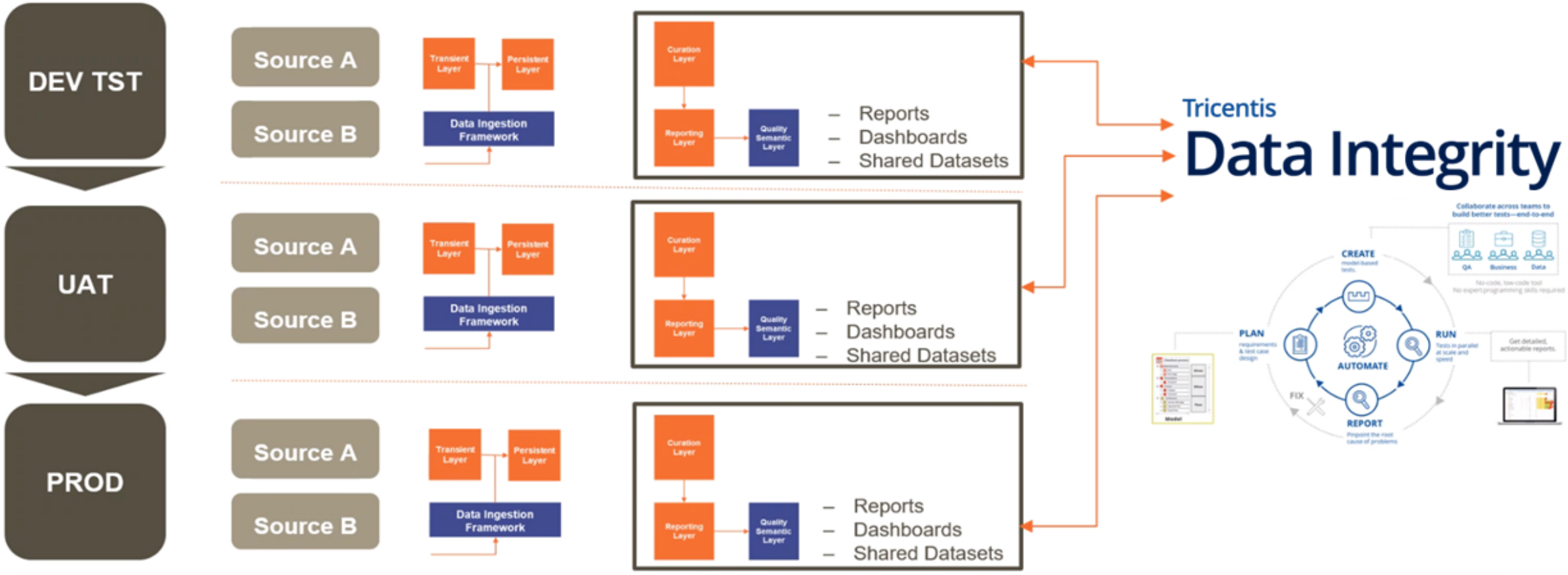
Example Strategy – Deployments/DataOps Architecture for ML with Data Integrity

Pipelines automatically builds and tests code projects to make them available to others



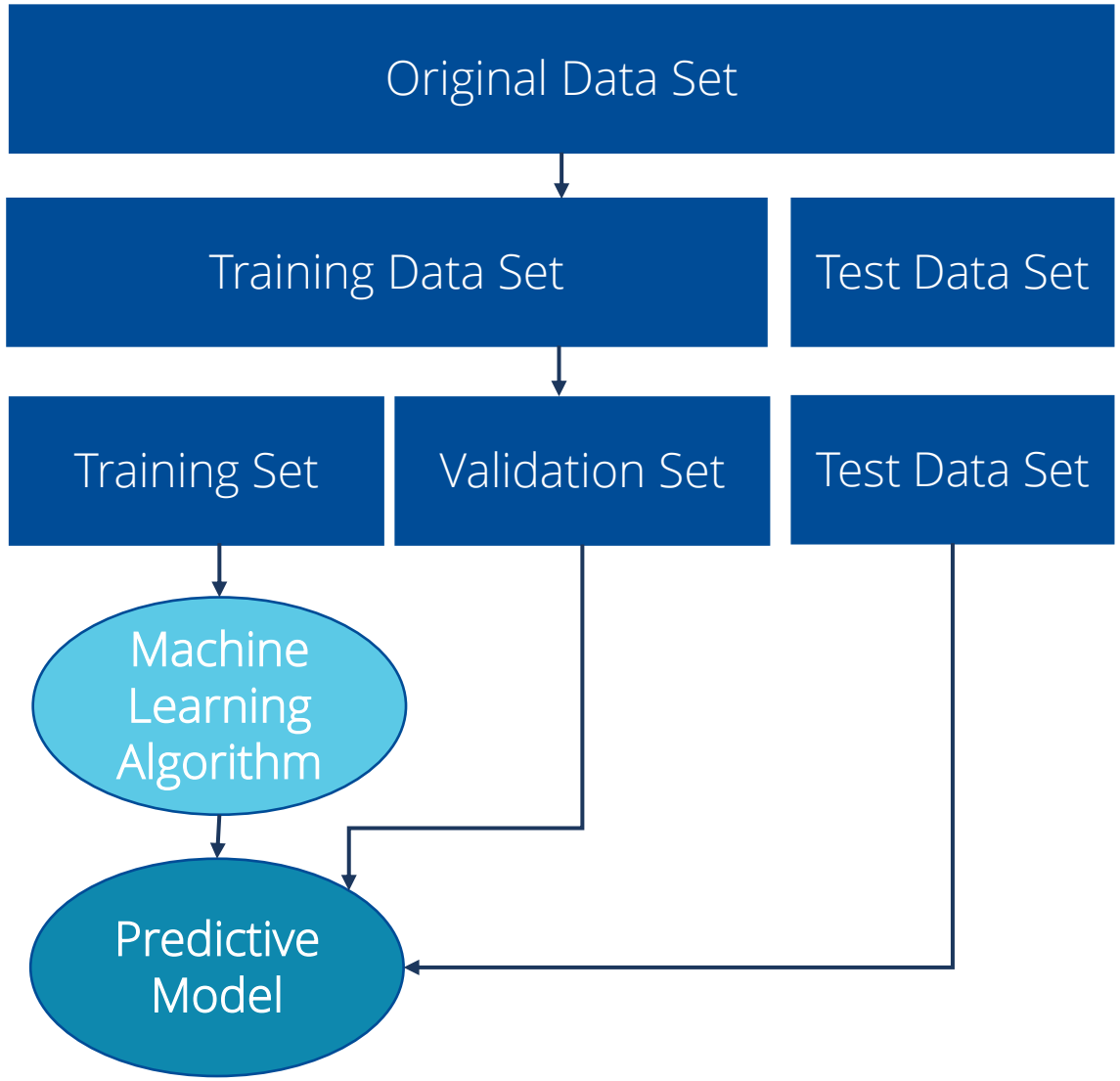
Example Strategy – Test Automation Architecture for ML with Data Integrity

Keep the test automations easy, faster and at scale with 'Model-based' tests



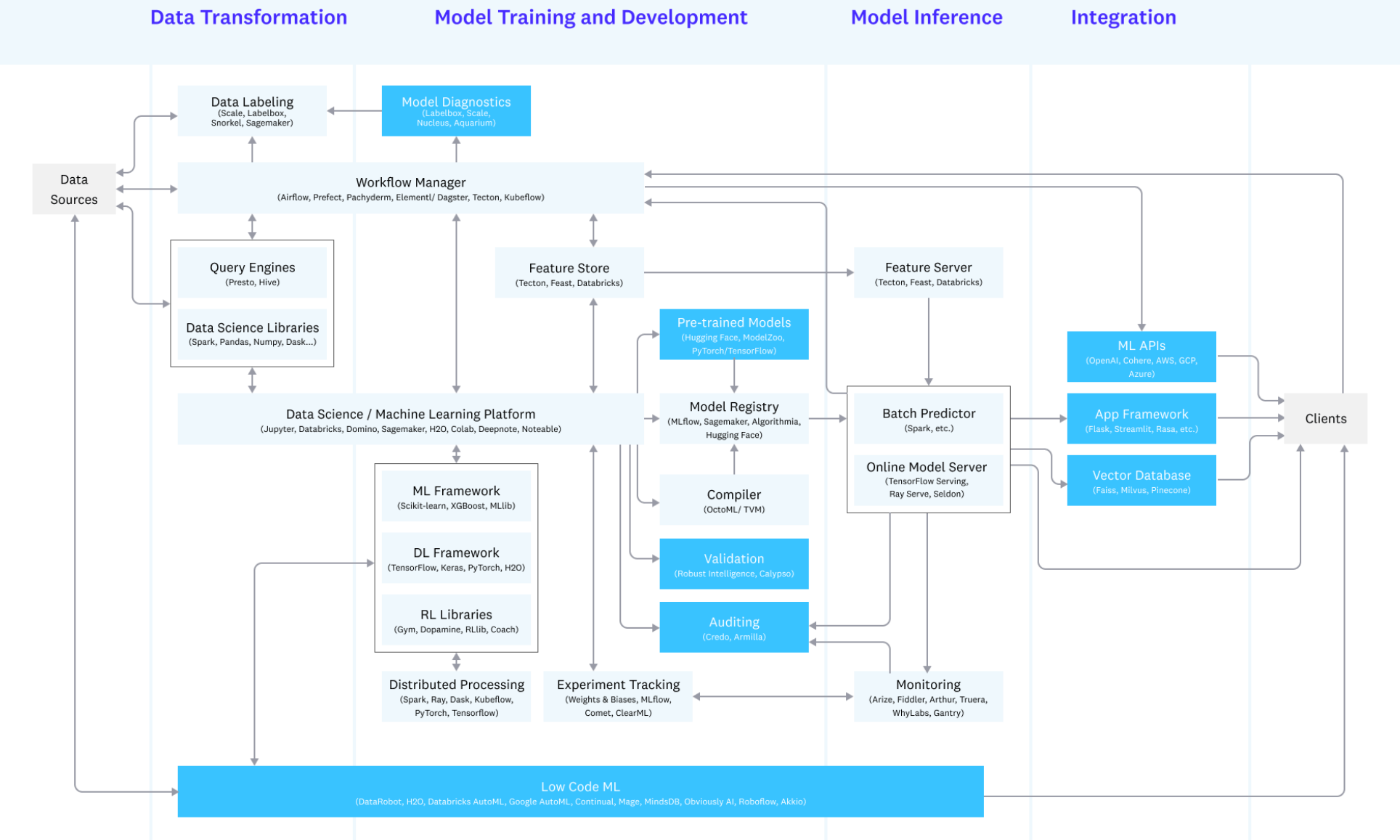
How we can assist with Training Data

- Testing Data Can be a great primer for Training data and Test Data
- Feature engineering enhanced by great data and business analysts understanding
- If you get public datasets they are often laden with data problems
  - Ex. Our CMS data set for Snowflake



Migrating to the new **Emerging Architectures for Modern Data Infrastructure**

Blueprint 3: Artificial Intelligence and Machine Learning



andreessen horowitz

Emerging Architectures for Modern Data Infrastructure Andreessen Horowitz

Tricentis Data Integrity Benefits for AI - Recap

1. Feature Engineering augmented with Application Test Parameters
   - Business Value associated with all AI/ML comes from the quality of Features and Hyperparameters
2. Training Data augmented with Test Data from Data and Application's Tests
   - Efficient Creation of Test Data for testing, utilized for training data, can decrease TTM for AI/ML solutions
3. Data Migrations from Unified Data Model 2.0 to ML Data Model 2.0
   - Move data for AI/ML at 2x the speed
4. Validation of integrated data models into Data Science
   - A solid automated MLOps process will ensure results and TTM
5. Validation of pipeline for delivery of Data Science results
   - See #4

Tricentis

Data Integrity

Drive better business outcomes through AI/ML processes with data you can trust

We are **Tricentis.**

The global leader in Continuous Test Automation

Gartner
FORRESTER
IDC

Named leader in industry analyst vendor ratings

2,100+
customers

Installed base world-wide

200,000+
certified

Tricentis test automation engineers

aws partner network
Microsoft Partner
SAP
SAP Solution Extensions
ATLASSIAN
servicenow
ORACLE

Strategic Alliances & Partnerships

Thank you

Make Trust happen!

Please visit our booth for all the details and a Demo!

Tricentis

Data Migrations

(For a deeper dive)

Best Practice Steps for Migration

Preparation	Familiarization	High-Level Assessment	Detailed Migration Study	Migration	Cutover to Production	Follow-Up
		Core Phase	Core Phase	Core Phase		
Obtain management support	Familiarize with migration techniques	Big picture view	Detailed code and data inventory	Execute migration plan	Parallel running and acceptance	Manage new system
Create core team	Augment technical migration skills	Feasibility assessment	Justification	Testing	Residual problem resolution	Decommission old system
Commitments and goals			Migration plan			Forward planning

Source: Gartner

Tosca DI Data Migration Timeline

Migration Phases:

- Create Core Team
- Management Support
- High Level Assessment
- Justification (Business/Cost)
- Access Planning
- Implementation Planning

- Migration Plan
- Source to Target Plans
  - Table to Table
  - Schema to Schema
  - DB to DB
  - Query to Query
  - API to API
  - Report to Report
- Migration Dates
- Begin Access Prep
- Finalize Risk Based Success Criteria Based on Requirements
- Approval of availability Human Resources

- Execute Migration Sprints

- Test Execution Sprints

- Parallel Execution
- End to End Acceptance Testing

- Implement End to End Testing as Regression
- Forward Planning and Development
- Manage New
- Decommission Old

Tricentis Planning Artifacts

- Data Migration / Conversions
  - Assessment
    - Goals
    - Tasks
    - Measures
  - Strategy
    - Automation Architecture
    - Project Plans
    - Use Cases
      - Test Objectives / Tasks / Deliverables / Measures (cross references)
    - Personas

Overlay the Plan With Mappings and Environments

Dependency – Security Matrix

- Testing environments for source to target plans
  - Test objectives
  - Test deliverables
  - Test tasks - Data Quality Measures

Source to Target Plans

- Schema to Schema
- Table to Table
- Query to Query
- Analytics API to API
- Report to Report

Verify the Data is in Sync With Dependencies & Changes

Verify data from both the migration process itself and ongoing external Integrations

- Complex dependencies make manual testing difficult and error-prone
- Accumulate an **automated** regression suite as you go
- Continuously test the migration and the ongoing systems

SOURCE / TARGET **(Example for AWS-Redshift)**	AWS - REDSHIFT	Client	Client
ORACLE	Yes	BofA	Thermo Fisher
EXADATA	Yes	WorldPay	Humana
TERADATA	Yes	Nationwide	IHG
NETEZZA	Yes	TJX	Nationwide
MSSQL	Yes	BCBS	CSS
AWS CDP - Hive	Yes	Kellogg	American Family
CLOUDERA - Hive	Yes	IBM	CPRail
AZURE SYNAPSE	Yes	HCSC	Cummins
SNOWFLAKE	Yes	TJX	Prologis
AWS-AURORA	Yes	GSK	CSS
GCP-BIG QUERY	Yes		
AWS-REDSHIFT	Yes	PHEAA	Kellogg

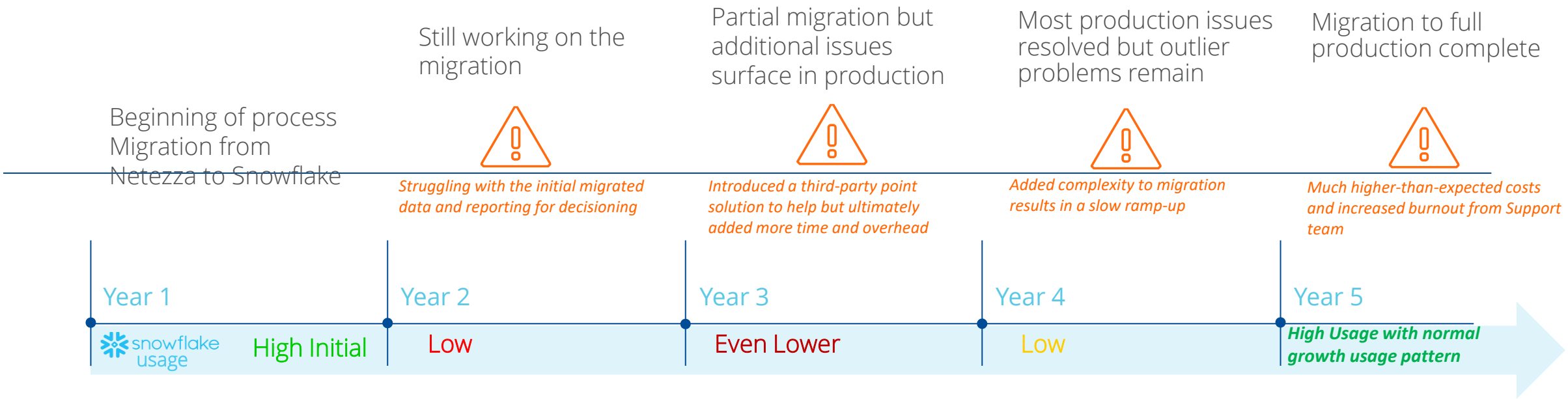
Cloud Migration Use Cases Supported

Migration Business Value Challenges Tricentis Addresses

- Legacy-to-Cloud Matrix (for history data comparisons)
- Incremental data testing (comparisons of inputs/outputs Source to Target)
- Cloud migration Risk management
- Regression In all phases of Migration – Key Post Migration Feature
- Test data management (Test Data Services)
  - PII data/tokenized data testing (e.g.: source not tokenized, target tokenized)
- Test data creation (Not included)

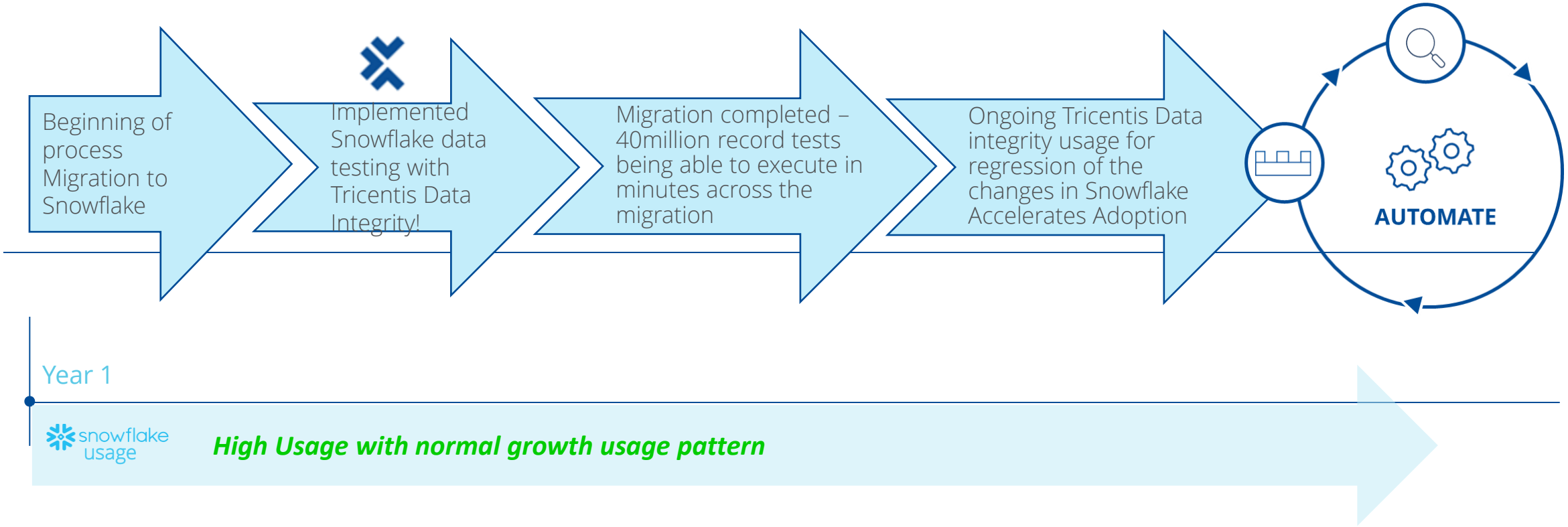
Manual Testing Snowflake Migration Timeline

Don't wait 3+ years for a complete migration to Snowflake



Typical Snowflake Migration Timeline with Data Integrity

Migrate to Snowflake in 180 days!



BFIS Use Cases

Banks Trust Us

- **Bank of America - Wealth Management Team**
- **Credit Suisse - Bank Platforms and Data Analytics**
- **TD Bank - CACS / CCP and Patriot Groups**
- **WorldPay - Payment Data and Analytics**
- **RBS – Enterprise Data Solutions**

Top Business Value Use Cases

1. Each of the following Use Cases are examples of Critical Data Paths that require reconciliation and validation with 90% coverage.

   Or these are the consequences:

- Time-to-market losses through failed innovation
- Lost opportunity $ due to bad data in migrations
- M & A fails to achieve results
- Loss efficiency of Operations due to bad decisions on incorrect data

Top Business Value Use Cases

TD Bank

M&A – Mergers and Acquisitions (and Divestitures)– Require a solid Data Reconciliation and Validation Strategy

- By offering a unique ability to validate against regression, data can now be moved in a consistent and trusted process ensuring data integrity during the migration and ongoing
  - **ROI Impact: $10's million in projected sales to properly migrated customers**
  - Cause: Without properly migrated customer data marketing of acquiring bank's services would not be correct or compliant

gsk

AI/ML – AI and Machine Learning

- Compliance and Innovative Data Learning Model Success
  - **ROI Impact: A $100million Time-To-Market delay. With compliance pressure, manual testing of data was not able to perform to regulatory standards and without properly curated data to train model, innovation fails.**
  - Cause: Datasets in Azure were too big to cover more than 1% with manual scripts and comparison (Manual Stare and Compare)

TD Bank

Compliance AML/KYC+

- **ROI Impact: $50m in KYC fines stopped. Cost Avoidance that is predictable.**
- Cause: Timing, Bank could only test a sample (1K) of the 70K scenarios to be covered end to end in the teller (mainframe) to audit data reporting process

TJX

DATA Migrations during and AFTER in the New Migrated Cloud Environments

- On-prem to Snowflake
  - **ROI Impact: Netezza to Snowflake migration EFFICIENCY saved $1m using data reconciliation from DI**
  - Cause: Production errors in the new Snowflake environment, as they added new Netezza data it would break the already migrated data and reporting without our regression testing

worldpay from FIS

General Data Reconciliation and Validations for Accurate business decisions

- Data Analytics Data from Payroll, Payments, ERP, Logistics, etc…
  - **ROI Impact: $24 million lost in trading decisions for Oil off by 1,000th of a percent.**
  - **ROI Impact: $10's millions spent on data warehousing and analytics NOT BEING USED as NO TRUST in the numbers.**
  - Note: Ability to perform this reconciliation in the Cloud in crucial (DataBricks/Azure 10-minute Demo Video)

Banking Specific Business Value Use Cases

Banking Settlement Reconciliation and Validation

- Tricentis Data Integrity can ensure the Float process is validated and reconciled before any changes are made to it. Ensuring a rigorous and exacting Float settlement process end to end and minute by minute.
  - **Potential ROI Impact: $10's million in properly executed Float settlement transactions versus loss of over float payments and COF if on the institution side.**
  - Cause: To balance the strict compliance and settlement and treasury/cash management needs of customers and financial institutions this is a process that must perform perfectly, and with exact timeliness on execution.

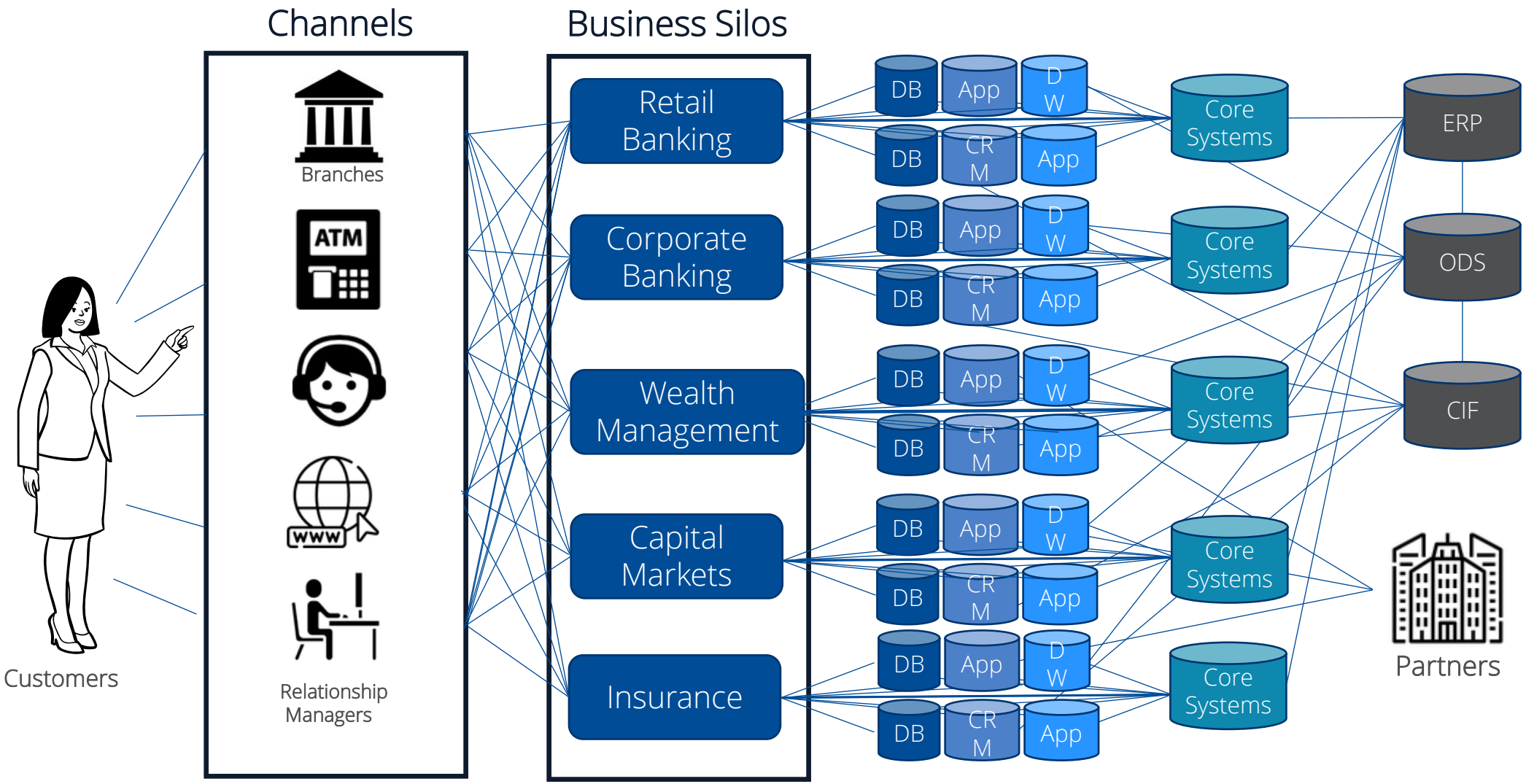
Data Driven Regulatory Needs are Changing with the New Banking Environments

- End to End compliance (Risk and Regulatory) reporting needs to be covered for new requirements
  - $250K limits – Who is over and under?, etc…
- We are working to uncover these with our banks to ensure compliance with expanded

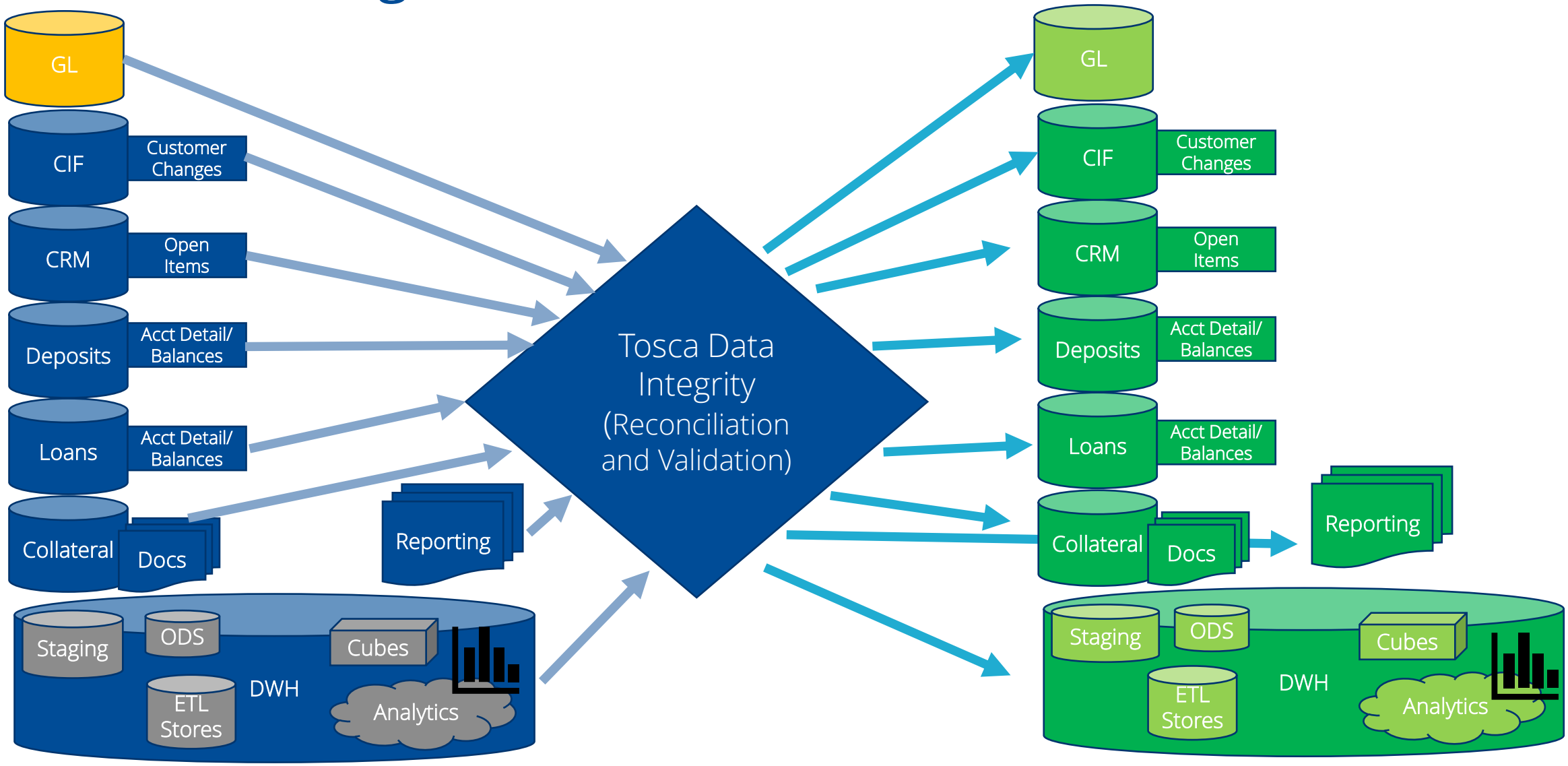
Banking Conversion Business Impact

- With Tosca Data Integrity , Conversions, Migrations and Merging can be accomplished in less than ½ the time with 10X the data Risk coverage.
- Tosca Data Integrity delivers an AUDITABLE, COMPLIANT, and RIGOROUS process to the movement of data and associated data processes.
- Tosca Data Integrity brings a level of confidence in the data that can't be achieved without it. You can now Trust the data is correct.
- Tosca Data Integrity delivers the holy grail of testing: a full AUTOMATED regression suite for your ongoing (post conversion/migration) data and processes.

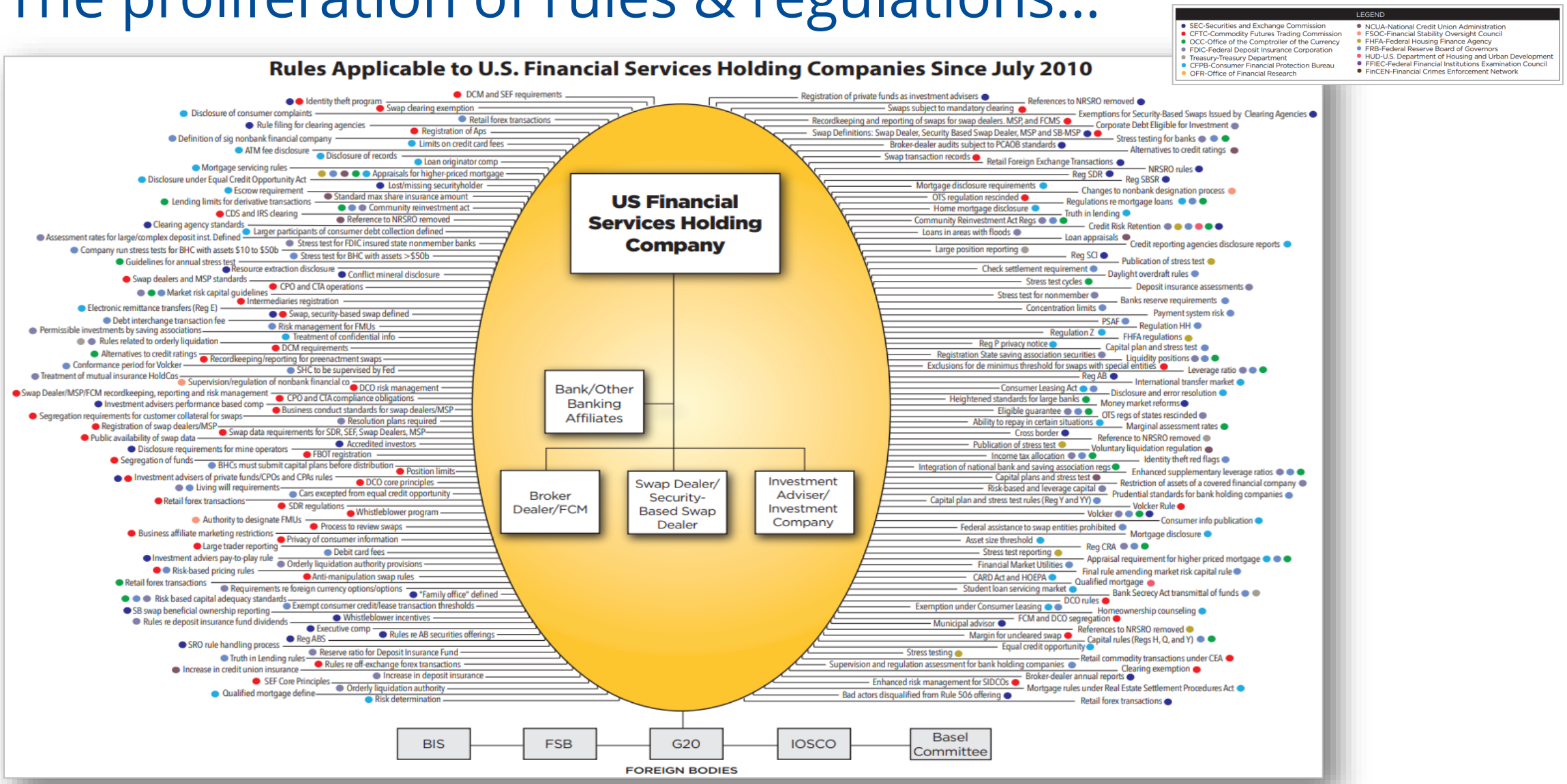
Banking Systems Complexity at the Data Layer
- Data Can break anywhere in the bank ecosystem



Banking Conversion Data must be tested quickly to process within tight windows



The proliferation of rules & regulations…



TD Bank Streamlines E2E Regulatory Compliance

BEFORE SITUATION

- Per money-laundering and anti-terrorism regulations such as the Bank Secrecy Act, banks conducting business in the United States are required to identify, monitor, and report "suspicious activity" in a timely manner
- Existing, predominantly manual, procedures could check a maximum of 1,000 scenarios per hour

THE NEED

- Complex validations across different data sources and layers – previously could not be automated with vendor-specific "data quality" tools
- Increased use case coverage to accurately automate 70k suspicious transactions scenarios

WHY TRICENTIS

- Improved scenario coverage to all identified use cases
- Automated validation to decrease manual overhead
- Increased identification of suspicious transactions across multi-faceted technology and business verticals

Business Case Summary	Using Tosca BI/DWH
Critical Checks Execution	70k validations in 15 min.
Potential Fines Avoided in 90 Days	$2,600,000
Identified Suspicions Transactions	1600% Increase

"We knew that automation was key for proactively ramping up our compliance … we just didn't believe it was possible to automate something so complex. Now, we're much more confident that threats are being identified and reported almost as soon as the transactions occur. "

- Director of Compliance

TJX reduces Snowflake migration timeline & costs by 50%

And eliminates data testing headaches with Tricentis Data Integrity

SITUATION

- Migration from Netezza to Snowflake required verifying migration stages, sources to target
- Scripted automation produced errors, requiring stop and restarts and delaying the 1-year project plan
- Data leaders required to implement a rigorous, repeatable regression process as the data was moved

NEED

- More resilient, faster test automation for end-to-end data testing process
- Support for complex technology stacks — from 40-year-old mainframes to cloud
- Ability to handle massive amounts of data, including 5 years of sales data from many different sources, requiring many transformations from source to report

RESULT

- Reusable, scalable tests without wrestling with SQL or scripting
- Thousands of hours of manual effort and test maintenance saved
- Improved user adoption and productivity as a result of improved data quality

50%
Time to market increase

50%
Reduction in testing costs

38#
Minute runtime for entire regression suite

WorldPay – 90% Savings with BI/DWH Test Automation

BEFORE SITUATION

- Millions of dollars invested in applications designed to help business leaders understand and leverage data
- Reporting errors lead to low to zero adoption
- QA leaders were challenged of “fixing this problem” by transforming the end-to-end data testing process without any existing tools, people, or processes for data testing in place

THE NEED

- Test Automation for end-to-end data testing process
- Support for Complex technology stacks—from 40-year-old mainframes to cloud
- Handle massive amounts of data (from 40 billion transactions processed annually), many different sources of data, and many transformations between source and report

WHY TRICENTIS

- Reusable, scalable tests without wrestling with SQL or scripting
- Thousands of hours of manual effort saved per month
- Improved user adoption and productivity as a result of improved data quality

worldpay from FIS

Business Case Summary	Using Tosca BI/DWH
Time to Market Increase	90%+ Increase
Cost of Testing	90% Reduction
Records Compared in 20 min	200 Million

Worldpay customer quote

"

A few months after our new BI / data warehouse testing approach launched, people finally started **trusting** the applications.

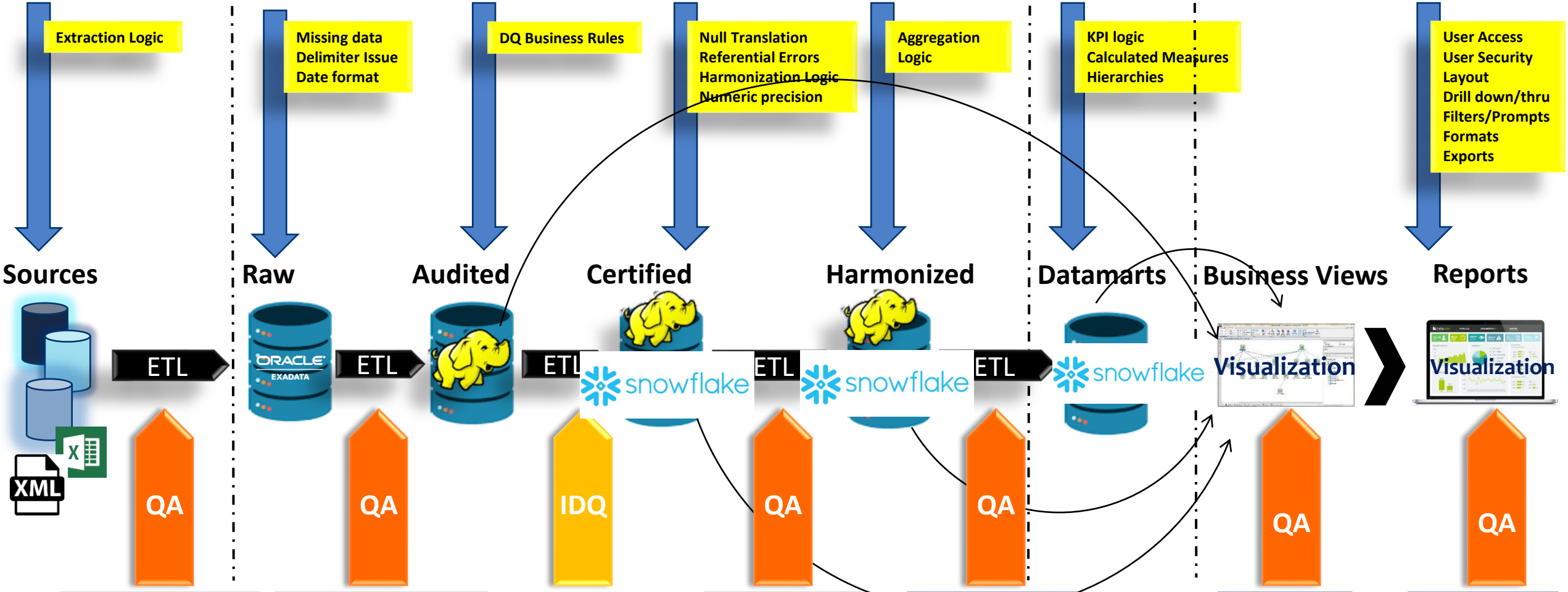
They're highly used now, and they have become a vital tool in making critical business decisions.

worldpay

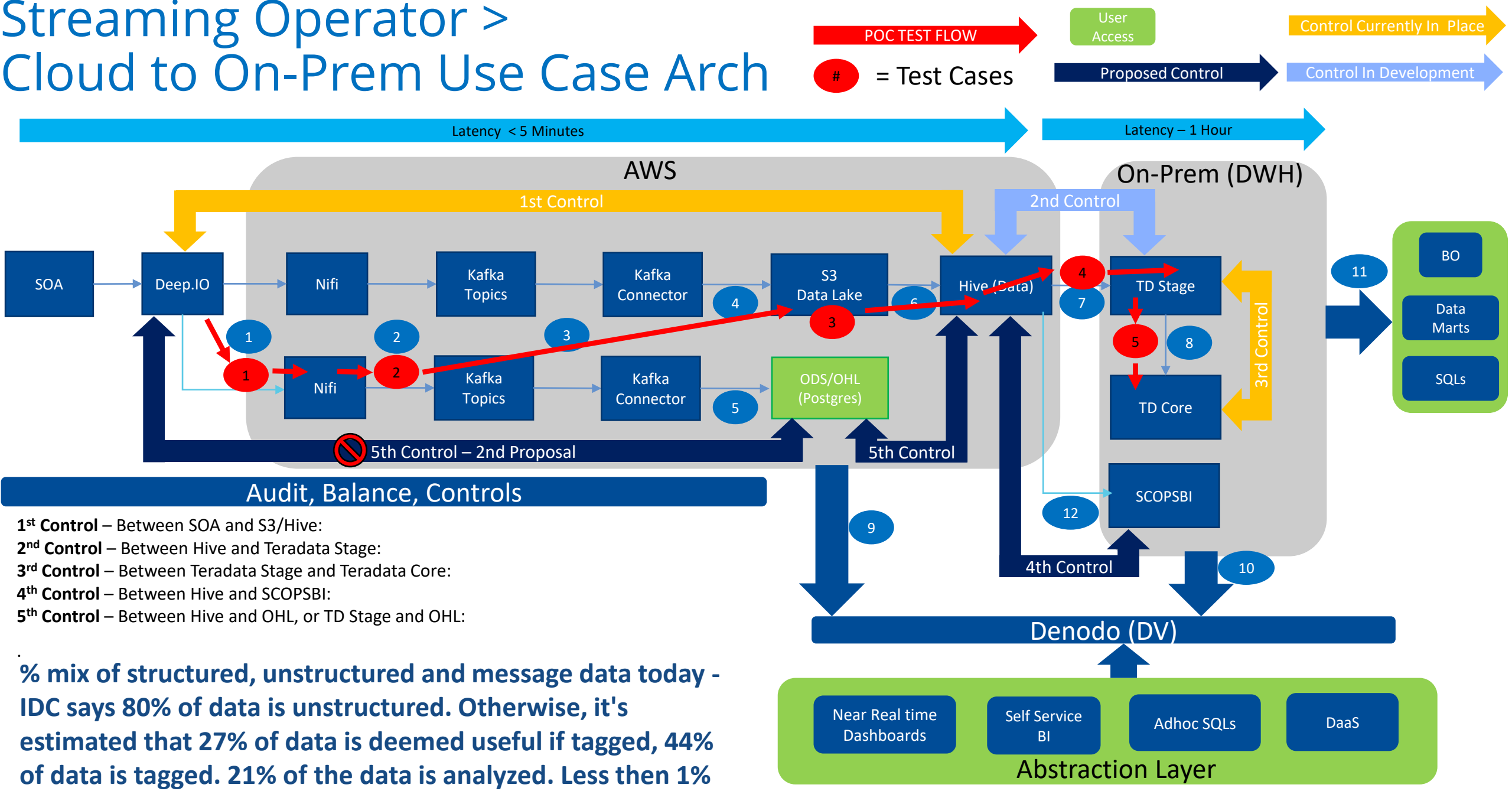
- Head of Enterprise Quality Assurance Team, Worldpay

Data Testing Approach

worldpay from FIS



Streaming Operator > Cloud to On-Prem Use Case Arch



1st Control – Between SOA and S3/Hive:
2nd Control – Between Hive and Teradata Stage:
3rd Control – Between Teradata Stage and Teradata Core:
4th Control – Between Hive and SCOPSBI:
5th Control – Between Hive and OHL, or TD Stage and OHL:

.

% mix of structured, unstructured and message data today - IDC says 80% of data is unstructured. Otherwise, it's estimated that 27% of data is deemed useful if tagged, 44% of data is tagged. 21% of the data is analyzed. Less then 1% fed into AI