# eGuide: How to Navigate to an Agile Data Lake

Top 4 solutions to avoid being sucked into a Big Data whirlpool

Infoworks

# Contents

Infoworks

# Big Data Can Create Whirlpools. Not Lakes

## 80% of big data projects fail to deploy. Here's why

**Digital transformation or digital nightmare?** Ninety percent of enterprises state that data and analytics are key to attaining the digital transformation needed to stay competitive.* Yet less than 20% of Hadoop deployments make it to full production—stranding organizations and their decision makers.**

**There are commonly overlooked hazards that sink big data projects.** And broadly speaking they fall into two categories: (1) Underestimating the technical complexities in developing a data lake and the associated expertise required to overcome those complexities, and (2) Underestimating the ongoing effort required to maintain an often brittle operational environment where every successive analytics initiative takes longer and costs more.

### Personnel Requirements and Technical Complexities

Big Data is inherently complex. It is based on implementing a distributed architecture which makes it naturally more difficult than operating in single system architectures. In addition, the big data space is flooded with a wide variety of constantly evolving frameworks, libraries, and tools—many of which are overlapping. In the end, implementors have to piece together multiple components to build a robust, enterprise-class end-to-end solution ecosystem.

**Moreover, hiring the right expertise is a big challenge.** The immaturity of data lake technologies, combined with a high level of complexity and rapid pace of evolution, means there isn't an enormous reservoir of expertise readily available. So even if you're ready to hire people with the right skill sets, it can be extremely difficult, if not next to impossible, to find them.

**Which means technical complexities are often underestimated.** As a result, organizations end up with "beta lakes" that can't be operationalized into environments that deliver the necessary enterprise scalability and reliability. Some challenges that organizations frequently overlook include:

- **Big data environments are "immutable."** Meaning they don't support the concept of adds, deletes or inserts of rows to a table. This means you can't just keep loading new and changed data as it arrives. You have to reload an entire table—not just its new information. Organizations are often surprised to learn they have to do a lot of manual scripting to incrementally load the deltas—and look forward to repeating the exercise, time after time-consuming time for each table of data you load into the lake.

* Business 2 Community 9/31/2018

** Enabling Essential Data Governance for Successful Big Data Architecture Deployment Gartner, January 2018

# Big Data Can Create Whirlpools. Not Lakes

80% of big data projects fail to deploy. Here's why

- **Teams also assume they can rely on Spark to handle data prep and transformation** and are unpleasantly surprised to find it requires significant effort and resources in a relatively new programming paradigm. It isn't as simple as they were led to believe and they end up having to write a lot of manual, low level code and scripts—often having to bring in people with expertise in optimizing Spark performance for distributed computing. The result is their dream of a self-service data lake that can be used by a business analyst quickly disappears.
- **When data is made available for consumption in a big data environment, it isn't necessarily ready-built for high speed queries.** Output data models need to be optimized in different ways in support of various use cases like BI reporting, dashboards, machine learning and ad hoc analytics. This is yet one more surprise that organizations run into where they have to write low level code to achieve reasonable performance levels.

## Operational and Production Issues

**Past success is no guarantee of future performance.** An organization may succeed in creating a lake and building initial data analytics use cases via hand coding and integration of point tools. But getting one or a few use cases working is not the same as getting tens or hundreds or even thousands running in production. Lack of a fully operational environment with tools for monitoring data pipelines and managing data lakes in production means:

- As the need to maintain and modify existing data pipelines reasserts itself as a priority, data engineering teams run out of time to implement new analytics use cases.
- When the underlying big data fabric evolves—for example, when moving from an on premise Hadoop provider to a big data cloud service or from Hadoop to Spark—the need to manually re-code significant portions of data pipelines and workflows to take advantage of new technologies will strain data engineering and IT teams.

**Plus, big data platforms don't automatically handle production optimization and dependencies.** For example, once you have multiple pipelines running you will also need to orchestrate timing for running potentially conflicting data pipelines or restart jobs if a process fails. If you don't build this kind of robustness into your data pipelines from day one, keeping them running will require that much more effort. This again requires yet more experts in big data operations, who are not the same as the people who develop the pipelines themselves and are just as hard to find and expensive to hire.

Data governance in this environment is also a headache. Good data governance is critical for real production environments and mandatory in some industries. However, the limited tooling provided for Hadoop and Spark once again require an expert who can write code. It's the wild west, and metadata lineage, auditability and re-use are still afterthoughts.

# Big Data Can Create Whirlpools. Not Lakes

80% of big data projects fail to deploy. Here's why

**The result: Lost time, lost money and unmet goals.** All too often, organizations have created data pipelines and data lakes which can't be promoted from development sandboxes into fully operational environments that meet an enterprise-level criteria for scale and reliability. Worse, they end up with unmanageable and ungovernable whirlpools that suck in resources and give back little in the way of useful and actionable analytics.

**The good news.** Automating the creation and management of data lakes lets you easily avoid all of the above hazards. And this guide shows how to safely navigate to your destination with data agility and speed.

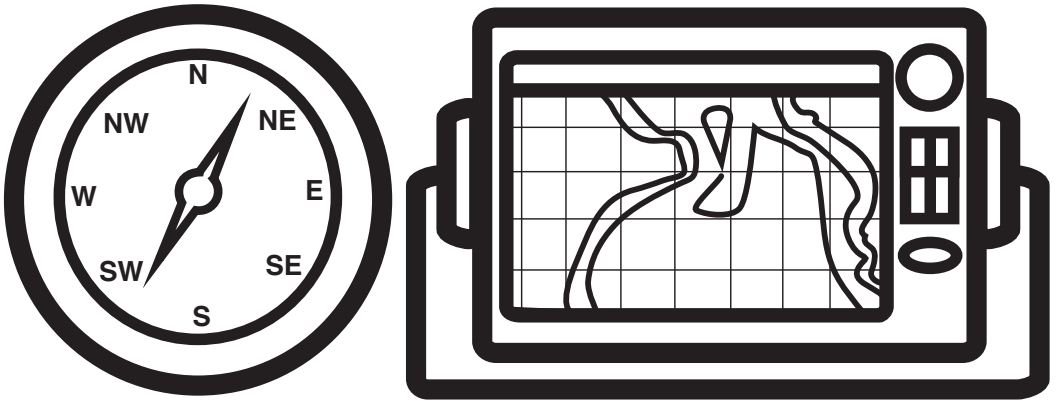
## The Challenges of Building, Managing and Benefiting from a Data Lake

| Building Your Data Lake | Managing Your Data Lake | Benefiting From Your Data Lake |
|---|---|---|
| • **Rapid Change**<br>Keeping pace with Hadoop's fast-evolving ecosystem | • **Ingestion**<br>Complications involved in filling your lake with data | • **Performance Issues**<br>High-speed queries require data model optimization |
| • **Skills Shortages**<br>Hard-to-find expertise in development and architecture | • **Poor Visibility**<br>Limited transparency and visibility | • **IT Burden**<br>Without user self-service, IT must prepare data for analysis |
| • **Scalability**<br>Scale and maintenance challenges for IT | • **Compliance**<br>Risk from unaddressed privacy and compliance issues | • **Change Management**<br>Without automated orchestration, making incremental changes is time consuming |

***Commonly overlooked challenges like these typically prevent organizations from deploying data lakes on a timely basis — and benefiting from them.***

# Agile Data Lake Creation and Management

## Your success requires GPS



# Agile Data Lake Creation and Management

## Your success requires GPS

**Remember this statistic?** Despite the buzz about harvesting big data for competitive advantage, over 80% of projects don't reach production. And those that do deploy often present ongoing maintenance and management challenges. That's why it's worth a quick comparison of typical implementation approaches against the advantages that a next-generation automated solution provides you.

### Standard Deployments and Their Drawbacks…

**Broadly speaking there have been two main implementation categories.** And each of these approaches has significant drawbacks.

1. **Hand coding the entire data engineering process on top of a big data fabric.** The risks of this time and resource-intensive approach include: difficulty in finding and hiring skilled data engineers to complete the work; longer-than-anticipated delivery times; overly-long deployment of use cases; poor operating efficiency; high costs; and lack of flexibility when execution and technology environments change.
2. **Integration of separate commercial point tools.** This lengthy process for handling the ingestion, data preparation, and creation of high-speed queries carries many of the same risks of hand coding. While individual point tools can help with automating individual aspects of the overall process, they still require significant hand coding to get each component to work with the other components. And once the development of data pipelines is completed, you still face the task of putting the end-to-end solution into an operational production environment.

### …Versus the Advantages of End-to-End Automation

**First things first…** Sure, there are a lot of visual coding tools that can help you with the different steps of your big data journey ranging from data ingestion to data preparation to query optimization to operationalization of complete workflows. But typical point solutions don't provide a platform for automating the entire data engineering process. Only end-to-end automation across all big data workflow processes brings you these advantages:

- **Elimination of manual coding.** Which means less human error and less need for manual fixes thanks to a "no-code" development environment.
- **Faster time to delivery.** Your analytics use cases will move into production faster and will consume far fewer data engineering resources.
- **Far better use of current data engineering talent.** End-to-end automation frees you from the need to hire an army of hard-to-find big data developers. Instead, your current data engineering team can easily achieve your objectives—and devote more of their time to strategic initiatives that can help improve your competitive edge.
- **The ability to implement and manage more analytics use cases*… fast.*** Automation empowers your current data engineers to not only design and build analytics use cases, but also make it easy to manage hundreds to thousands of use cases in production — freeing up more engineering time to create even more use cases that allow you to deliver more insights from your data and sharpen your competitive edge.

# Agile Data Lake Creation and Management

## Your success requires GPS

- **Agility.** Agile data engineering is about the ability to quickly and easily add new data processing, machine learning and analytics use cases in support of rapidly-evolving business models and initiatives. It also involves the ability to promote development data pipelines into production to multiple different execution environments. The only way to achieve this is through automation.
- **Even more agility.** But what if the core experiment of an analytics use case doesn't work out? Then the ability to fail fast and change direction even faster is another big advantage of automation—which also contributes to another huge benefit: Orders-of-magnitude shorter development-to-production cycle times.

Infoworks

# Agile Data Lake Creation and Management

## Your success requires GPS

Case Study

### Leading Publisher Deploys Agile Data Lake in Weeks

One of the world's largest media companies required a consolidated view of its IT spend—a project which would span 60 divisions with data derived from 12 different data sources. Faced with a twelve-month deployment schedule using traditional EDW approaches, they turned to Infoworks data engineering software—enabling one data engineer to create and deploy an automated data lake on Azure in four weeks.

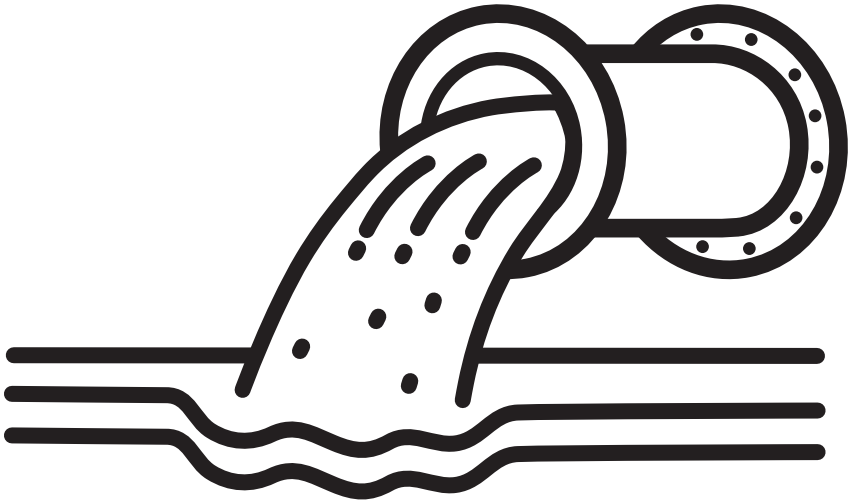
GET THE WHOLE STORY

READ THE CASE STUDY

> **"One data engineer took us from idea to completion in only four weeks using Infoworks data engineering software."**
>
> Chief Technology Officer, Fortune 500 Publisher

# Filling the Lake

## The challenges of data ingestion



# Filling the Lake

## The challenges of data ingestion

**Data ingestion into a big data environment is harder than most people think.** Loading large volumes of data at high speed and managing the incremental ingestion and synchronization of data at scale into a data lake can present significant technical challenges. And hand coding can create problems at every turn.

**For starters, manual coding and scripting is error prone.** It can also be very expensive to modify when changes occur. And when new data columns are added to source systems, data ingestion processes often break if they aren't updated to take the change into account.

Plus, without automation, a variety of issues can easily arise at important stages of the data ingestion process:

- **Incremental ingestion of data.** When incremental data arrives in a big data system, challenges arise in reconciling or merging it with the base data that has been ingested earlier. Here's why:
  - Incremental changes are trivial in traditional relational database and data warehouse systems. But most big data systems are "immutable" or have immature, problematic, ACID implementations. They don't understand how to handle incremental changes to a row of data in a table. So when a new row needs to be included, the entire table has to be reloaded into the big data system.
  - In big data systems, incremental data ingestion requires change data capture of source data *and* the need to merge and synch new rows—adds, deletes or inserts aren't supported out of the box in big data environments. As a result, many organizations constantly reload entire tables, along with their incremental changes, into their data lake.
  - The alternative to reloading entire data sets is to manually code change data capture and related synch and merge logic. This is a complex process where you have to first capture the changes in the source system since the last ingestion, using both log and query-based approaches if you want to be sure to properly capture all changes. Once the incremental data arrives in the big data or cloud system, there's an additional challenge in reconciling or merging the incremental data with the base data that has already been ingested earlier. Finally, you have to deal with issues like slowly changing dimensions (SCD), track the current state of SCD data and keep a history table of prior state data including the date of any changes as well as tracking any errors that might have occurred in the SCD process.
- **Parallelization of loading data.** Ingesting large quantities of data requires parallelization. And while big data platforms provide the raw infrastructure to parallelize the loading of data, users still have to write code to run multiple ingestion pipelines simultaneously. However, special skills are needed to write custom ingest code and even if available in-house, hand coding is brittle and difficult to repair. And traditional parallelization of data ingestion involves single table loading of data, where each table ingested gets its own data pipe—which means large tables will still load slowly because even a very large table can still only use a single pipe. There are ways around this, but they all involve hand coding by someone with very specialized knowledge.
- **Batch and streaming data loading.** Data needs to be loaded in batch as well as streamed. But the underlying technology for each is different. This either requires using different commercial tools, one for batch and one for streaming or multiple open source tools—which all require significant hand coding to get them production ready.

# Filling the Lake With Infoworks

## Automated data ingestion

**No-code ingestion configuration.** Infoworks provides a no-code environment for configuring the ingestion of data from a wide variety of data sources. Alleviating the need to hire an army of coding specialists, it reduces errors typical in the manual handling of data type conversions and brings the benefits of automation to every stage of data ingestion:

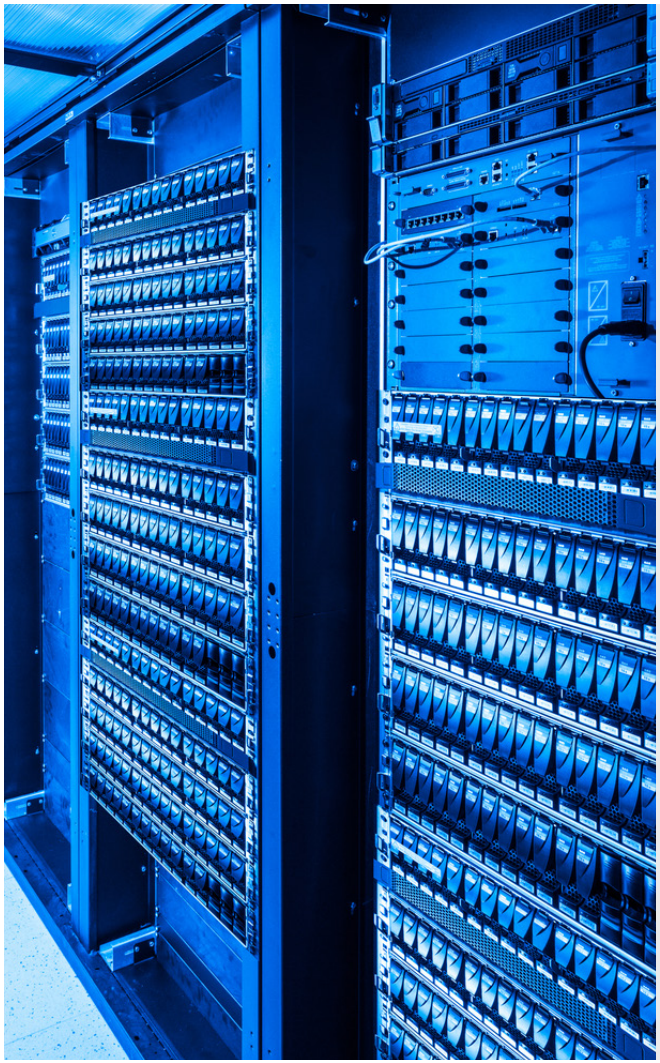
- **Change data capture with easy synch and merge of incremental data.** Incremental ingestion of data can be done in a variety of ways, but the first step is to capture the changes in the source system since the last ingestion (also known as CDC or change data capture). Infoworks automates all kinds of CDC whether log-based or query-based. Infoworks then reconciles and merges incremental data at ingestion time with the base data that had previously been ingested. Its continuous merge capability supports fast ingestion and continuous fresh data availability, while keeping the data optimized for downstream query performance.
- **Automated schema change detection.** Infoworks automatically detects source side schema changes, adjusts for those changes and ingests the new columns automatically into the data lake—avoiding breaks in data system ingestion processes that can occur when relying on manual coding.
- **Automated, scalable, parallelized data ingestion.** Infoworks' automated process parallelizes the ingestion of data into your data lake— significantly accelerating the loading of data without requiring code development. Its ability to handle multiple data pipelines speeds the process, as does its power to parallelize the loading of large tables: the software divides tables into segments, managing both segmentation of a single table on the source side and the reintegration of the segments when the data lands in the data lake.
- **Simplified, no-code batch and streaming.** Infoworks supports both batch and streaming use cases. A single, simple menu-based interface supports both use cases with no coding necessary. Plus, Infoworks uses Kafka as the underlying streaming engine and can connect to any data source to stream large amounts of data in real time.

**Easy movement of data.** Infoworks automatically handles data type conversions, reducing the errors typical in manual handling of data type conversion and making it easy to move data from your data lake to other consuming systems without time-consuming re-coding of data types.

Infoworks

# Filling the Lake With Infoworks

## Automated data ingestion



Case Study

### Fortune 500 Provider of Computer and Networking Equipment Creates Agile Data Platform Using Infoworks

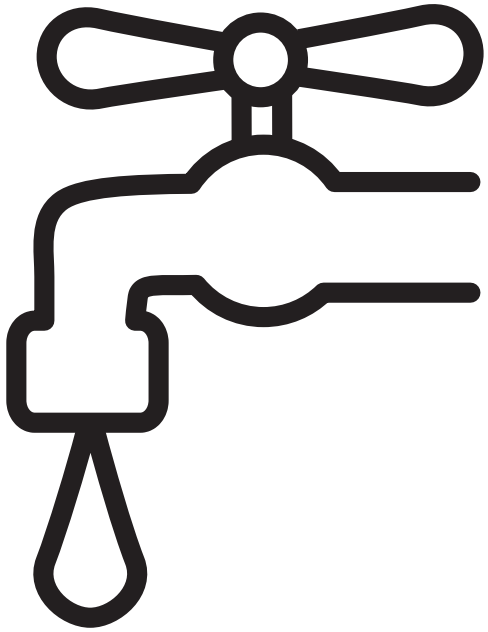
A major provider of computer and networking equipment had attempted to build their own automated data ingestion framework for their on premise data lake. After two years in elapsed time and 24 engineering years in total, they built an automated framework that dealt with the major challenges of data ingestion. However, they still needed 4 skilled data engineers to keep their infrastructure up to date due to the constant changes in the underlying Hadoop ecosystem. They eventually discovered Infoworks, which they used to automate the entire data ingestion process for their data lake—and now only require 1 data engineer to manage over 5000 data objects.

READ THE CASE STUDY

> **“We replaced a team of 4 data engineers and 24 years of engineering work with Infoworks software and one data engineer.”**
>
> Director of Data Engineering,
> Fortune 500 Provider of Computer and Networking Equipment

# Purifying the Water

## The challenges of data preparation



# Purifying the Water

## The challenges of data preparation

**Filling your data lake with data is just the beginning.** From there it needs to be transformed in preparation for downstream use. And without automation, challenges in this area can start out as a series of small issues that quickly add up to time-consuming hassles—since manual coding of data transformation and machine learning pipelines can result in ongoing maintenance and management problems.

Labor intensive and error prone, the complications arising from hand coding create a wide range of difficulties, such as:

- **Writing and troubleshooting data pipelines is very time consuming.** This was true when comparing hand coding to old fashioned ETL. And more so now, because development of data pipelines on a distributed computing framework is an order of magnitude more complicated than writing transformation logic in non-distributed, single server environments. This problem has been getting worse as the world moves to Spark which has become the most common data transformation technology used in big data and cloud today. Unfortunately, development in Spark usually requires a very knowledgeable developer who can write low level code.
- **Creating data pipelines for incremental data loads is a hassle.** Dealing with constantly changing source data requires writing two distinct data pipelines, one for the initial data load and a second for the incremental loading of data. This process takes additional time and results in errors, since it requires writing two different, but similar, data pipeline processes. This problem is made even worse due to the fact that big data file systems are immutable and don't understand how to handle incremental changes to a row of data in a table. So data pipelines become even more complex to deal with this issue.
- **Migrating legacy pipelines is tedious.** Very often, organizations want to recreate existing data flows that occur in a more traditional ETL/EDW environment within the new data lake. Hand coding these data flows slows the process of migrating legacy pipelines and requires specialized skillsets.
- **Lack of specialized expertise can be a big impediment to big data analytics initiatives.** Getting a data pipeline to run with the business logic (and, potentially, machine learning algorithms you want) requires hiring one set of skills that is already difficult to find. Taking those same pipelines and getting them to perform well and run in a highly reliable fashion in your data lake requires someone with a completely different and also hard-to-find skill set.

**And remember: Designing analytics and machine learning data pipelines is only one aspect of data preparation.** You have to also develop data pipelines that can be operationalized for production at enterprise scale. And without automation you face even more barriers:

- Technology and vendor-specific skills are required. Optimizing data pipelines for performance requires expert knowledge of the underlying Hadoop or Spark platform which will also vary by vendor.
- Writing a data pipeline to run once is relatively simple. But getting that same pipeline to run repeatedly in production requires dealing with error handling, process monitoring, performance monitoring, process availability and many other issues, including the need to deal with the fact that data is constantly updated.

**Plus, query performance can also be an issue.** Market-leading business intelligence and data visualization tools are not designed to handle the

# Purifying the Water

## The challenges of data preparation

large data volumes typically associated with big data. At the same time, data sources like Hive and NoSQL are not able to deliver sub-second response times for complex queries.  So just pointing a visualization tool like Tableau at a Hive table will work, but it will deliver very unsatisfactory end user performance.

Moreover, other approaches, like creating scalable OLAP cubes and in-memory models that do work well with common data visualization tools, require yet another kind of big data expertise that can be hard to find. And depending on the number of end users, even more scalability issues can arise.

# Purifying the Water With Infoworks

## Automated data preparation

**True no-code data pipeline development.** Infoworks automates the development of production-ready analytics and machine learning data pipelines. It provides a drag-and-drop environment for creating analytics and machine learning data pipelines all the way from initial ingestion to consumption in OLAP cubes and in memory models.

With no Hadoop or Spark knowledge required to create scalable and highly-tuned data pipelines, you won't have to hire big data experts to build or tune your pipelines. Instead, Infoworks delivers:

- **Automated incremental pipeline creation.** Infoworks makes the process of adding CDC pipelines as simple as clicking a radio button. With just one mouse click a CDC pipeline is generated automatically.
- **Automated validation of ingested data.** Infoworks automatically validates data ingested into your data lake for full and incremental loads coming through change data capture. For all data sources loaded, it provides row count validation and user-specified aggregate data matches between source and target tables.
- **Drag and drop OLAP cube creation.** Users simply drag and drop facts, dimensions, and measures. Infoworks automatically builds a fully pre-aggregated and optimized OLAP cube that scales to any size and provides sub-second query response times.
- **Automatic conversion of SQL into big data pipelines.** Migrating legacy data warehouse jobs is significantly accelerated through automatic conversion of SQL into easily maintained, optimized, portable, visual data transformation pipelines.
- **Team based development.** Users with the same access rights can share data, pipelines, and workflows to enable team-based development across the entire data lake development process.

**Infoworks also enables immediate promotion of data pipelines into production.** Data pipelines built in Infoworks can be easily promoted with just a few mouse clicks from dev to test and into production while integrating with software version control systems, either on premise or in the cloud, without any code changes.

**And it automatically scales and performance-tunes data pipelines.** Data pipelines must be tuned to perform, scale and subsequently meet service level agreements (SLAs). Infoworks automatically optimizes pipeline builds to meet SLAs without requiring a big data tuning expert. With it you enjoy:

- **Faster pipeline performance.** Data pipelines created with Infoworks perform 25-40% faster than hand coded SQL or HQL.
- **Automated query optimization.** Infoworks dramatically accelerates response times for big data queries, enables easy scaling, and make performance maintenance easy: simply add query nodes on a point-and-click basis.
- **Large-scale query concurrency.** With Infoworks, a single OLAP query node supports 75+ queries per second and the number of concurrent queries scales linearly with an increase in the number of query nodes.

**Supports seamless portability across execution engines.** Infoworks data pipelines provide portability across all major distributed data execution environments on premise and

# Purifying the Water With Infoworks

## Automated data preparation

in the cloud. The underlying execution engine is abstracted, then automatically optimized for different backends (Hive, Impala, Spark, etc.) without requiring re-coding.

**Automated updating of in-memory models and cubes.** Infoworks automatically propagates upstream data changes down to in memory data lake data models and OLAP cubes.

Plus, Infoworks supports the capability to run end-to-end data pipelines across multiple execution environments that can span from on premise to the cloud or multiple cloud environments.

## Data Transformation Metrics (Based on a Customer Case Study)

| Key Metric | Without Infoworks | With Infoworks |
|---|---|---|
| **Time to build data transformation pipelines from existing Teradata SQL** | **4 Weeks using Big 5 consultant resources (incomplete)** | **1 day** |
| **Time to build logic for a dashboard: 13 reports, 7 data sources, 7 transformation pipelines, 3 cubes** | **3-4 months, several engineers** | **4 days, 1 engineer, 1 subject matter expert to review** |
| **Time to create an incremental pipeline** | **2 weeks to code, test. Write logic to store watermarks, delta transformation, merge** | **Automatic (single button click)** |

*True no-code pipeline development from Infoworks dramatically reduces the time and number of engineers required to accomplish key tasks.*

Infoworks

# Purifying the Water With Infoworks

## Automated data preparation

Case Study

### Retail Pharmacy Chain Deploys End-to-End Big Data Solution in 26 Days

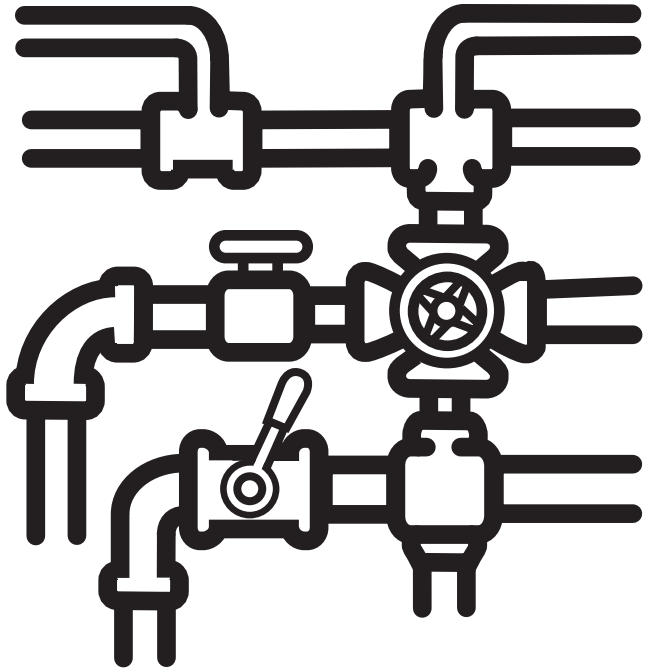
A Fortune 500 pharmacy chain struggled with providing user access to claims data residing in their adjudication system. With tens of millions of claims requiring processing every year, the company decided to offload its legacy data warehouse to Hadoop. Lacking in-house expertise, they turned to Infoworks agile data engineering platform to automate the entire development and production. Working with the company, Infoworks migrated 300 tables of data and metadata, 70 SQL and BTEQ control flows in just 26 days.

READ THE CASE STUDY

> **“With Infoworks we were able to complete our project plan for the entire year, in just a few weeks.”**
>
> Data Engineering Director, Fortune 500 Pharmacy Chain

# Managing the Flow

## The challenges of data lake operations



# Managing the Flow

## The challenges of data lake operations

**Big data platforms don't deal with enterprise production issues.** So, without automation of your end-to-end operational environment, even day-to-day management and maintenance of your data lake can consume inordinate amounts of time and human resources. Organizations that manage their data lake operations manually quickly find they are spending more time just keeping their environment running and progressively less time adding new analytics use cases needed to support their constantly evolving business needs. By relying on manual processes, you face problems involving:

**Ability to scale.** As we've mentioned, managing one data pipeline is easy. However:

- **A successful data lake will have hundreds or even thousands of data pipelines,** all running simultaneously—creating significant orchestration issues. This makes it next to impossible to manually optimize the running of pipelines, and ensure their reliability and availability.
- **Complications with ad hoc analysis can arise.** Development of a data lake for ad hoc analysis isn't the same as converting those ad hoc analytics into production workflows and data pipelines that run with enterprise reliability and scale.
- **Enterprise-readiness issues.** Your company will count on your data lake to run its business. But a production data lake environment demands a highly available environment, and production workflows which are even more complicated than development workflows running in a sandbox. An operational data lake has to include enterprise-class capabilities for:
  - **Reliability.** Failed jobs need to be restarted automatically and pipelines that have upstream dependencies need to be able to be paused if one of those dependencies is late.
  - **Alerts.** Operations personnel need to be notified immediately when jobs fail.
  - **Performance monitoring.** Identifying performance bottlenecks in a distributed system is complex, and is even more difficult if that system is made up of multiple-vendor products that have been stitched together.
  - **End-to-end workflow process monitoring.** Managing pipelines end-to-end requires you to deal with multiple components like data ingestion in batch and streaming, data preparation, transformation, and creation of high-performance query environments. This is always a challenge in distributed computing environments and is made even more difficult if you use a different vendor for each step in the process. This means you either hand code, or stitch together multiple solutions to build the management capabilities mentioned above.

All of this can require expertise that may not be easily available.

**Vendor lock-in and portability across different production environments.** Because underlying modern data environments are evolving at a fast pace, organizations are often moving from one data fabric to another. But hand coding or stitching together multiple commercial solutions that run on top of these environments makes it difficult to move when new technologies emerge or evolve. This means that vendor lock-in—whether it's a cloud platform or an on premise Hadoop supplier—can be a problem.

# Managing the Flow with Infoworks

## Automated operations

**As you've just seen,** developing data pipelines that can be promoted into production at enterprise scale; ensuring enterprise readiness; and guaranteeing portability that avoids vendor lock-in are all challenges when creating a data lake. However, with Infoworks, you are assured of achieving all three:

**Enterprise-level performance and scale.** Infoworks' high levels of automation and no-code development environment enable:

- **Development of data pipelines that are automatically production ready.** Data pipelines built in Infoworks can be integrated immediately into end-to-end workflows and run in production environments, either on premise or in the cloud, without any code changes. They can be built in one environment and then deployed in another and automatically scale with the size of the execution environment.
- **Automatic performance tuning, no expertise required.** Data pipelines must be tuned to perform, scale and meet service level agreements (SLAs). Infoworks automatically optimizes pipeline builds for whatever environment they are run in to meet SLAs without requiring a big data tuning expert.
- **Enterprise readiness.** Infoworks ensures that big data analytics are run in a reliable and repetitive fashion within production environments with enterprise class capabilities that deliver:
  - **Data pipeline and workflow fault tolerance.** Infoworks automatically retries and restarts failed workflow jobs—and makes it possible to pause, resume and dynamically control production data workflows.
  - **No-code alert creation.** Adding alerts to a workflow is easily achieved via a simple drag and drop operation.
  - **Automatic performance monitoring.** Pipelines are automatically monitored for performance and data lineage metadata is automatically tracked with no additional coding required.
- **Portability.** Data ingestion, transformation, cube generation and workflows built in Infoworks can run in any execution environment with high performance and without the need for re-coding. This capability provides:
  - **No-code workflow migration.** Development of ingestion, transformations, OLAP cubes and in-memory models is separated from the actual implementation. So, development workflows can be moved from on premise to cloud, or cloud to cloud, without re-coding.
  - **Freedom from vendor lock-in.** High performance portability avoids lock-in with cloud platform and Hadoop suppliers. Migration from on premise to cloud or cloud-to-cloud is fast and easy, and portability is supported across all major execution environments.

Infoworks

# Managing the Flow with Infoworks

## Automated operations

Case Study

### Department Store Completes Cross Platform Big Data Project in 20 Days

One of America's largest department store chains needed to create back end analytics systems to support customer loyalty data modeling and analysis. Their original big data analytics solution was based on an in-house hand-coded automation framework. The solution supported data pipelines that had to run across an on premise Hadoop implementation with some components and data that resided on Azure with final analytics performed on Google Big Query. It was a very complex and brittle environment and over time, the company found they were spending more time troubleshooting their hand-coded framework than actually using it. So they turned to Infoworks for a single development and deployment platform that would work across all of their environments without having to re-code—a project that was completed in just 20 days.

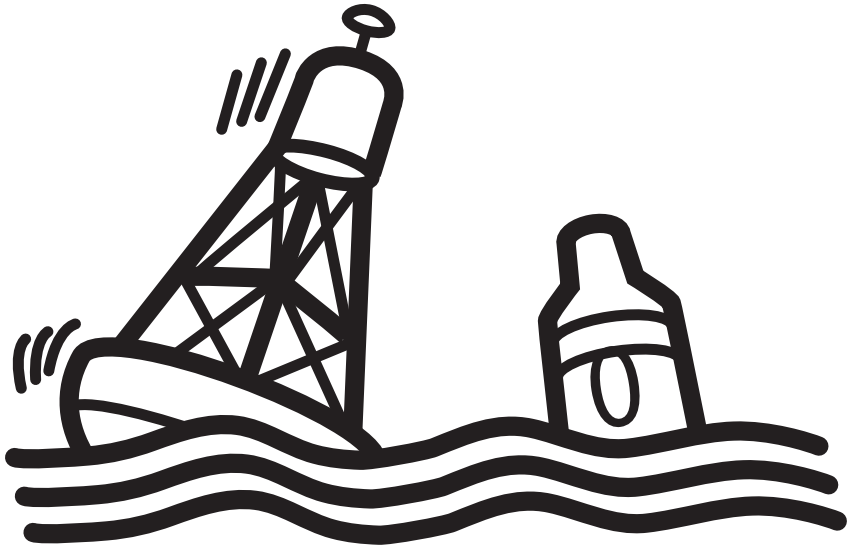
GET THE WHOLE STORY

CONTACT INFOWORKS

> **"We spent more time troubleshooting our hand-coded framework than using it. With Infoworks we replaced it in just 20 days."**
>
> Lead Enterprise Architect, Fortune 100 Retailer

# Governing The Data Lake

## The challenges of data lake governance



# Governing the Data Lake

## The challenges of governance

**Easy identification of new insights in an ad hoc analytics environment is a great promise.** And the tendency for big data automation solutions is to focus on doing just that. But there is a problem: data lakes contain critical and sensitive data about the company, employees, partners and customers. So issues around the proper governance of both the environment and the data itself quickly become critically important to your business:

**Metadata management.** One big challenge in managing a data lake is tracking and managing all metadata from source through transformations to consumption. Tracing lineage of any data artifact is necessary for audit compliance as well as troubleshooting. It is extremely difficult to do this manually across multiple sequential data pipelines. Similarly assessing the downstream impact of any data or metadata change is next to impossible when hand coding your data engineering processes.

**Access control.** Not all users should have access to all data. And data lakes don't come with governance baked in. While there are frameworks like Apache Ranger and Sentry for data lake security, they are stand-alone components that still require significant integration efforts to get them to work with all aspects of your data lake processes from ingestion to consumption.

**Multiple loading of data sources.** When organizations first implement a data lake, they find that without a strong process, or some kind of automation, the same data sources can be loaded into a data lake multiple times, creating multiple problems:

- **Unnecessary load on source systems.** DBAs hate it when different users each ask for access to the same data source when it turns out that the data has already been loaded in the data lake. This is a very common occurrence with data lakes that have no governance processes or automation.
- **System of record confusion.** When multiple versions of the same data are loaded, users can no longer be sure of the "system of record." Users who are simply consuming data don't know which version is refreshed more frequently or which version may have been altered to improve data quality. The problem is that different users may end up using slightly different versions of the same data and end up with inconsistent results that are difficult to reconcile.

**Auditability and regulatory compliance.** In industries such as banking and health care, government regulations require detailed records of who has access to data, who has made changes to your system, as well as complete lineage of where data comes from and where it is used. Risk of non-compliance is complicated in a data lake when you are balancing a desire to have a single location where analytics can be easily performed via self-service, with the need to meet regulatory requirements that create additional challenges for:

- **Data access.** With so much data loaded into the data lake, there is easily the potential for people who should not have access rights to sensitive data, accidentally seeing information they should not.
- **Change management.** As your data lake evolves, you need to be able to track who was responsible for ingesting data; who subsequently built a data pipeline from that

# Governing the Data Lake

## The challenges of governance

data; and who else may have eventually made a change to those processes and when they changed them. Given the amount of data, number of users, and number of data pipelines and workflows, keeping on top of all of this can be next to impossible to track manually.

- **Data lineage issues.** Knowledge of data lineage is critical for managing future changes in the system and tracking upstream and downstream impact of changes. Lineage is also critical in organizations that have regulatory requirements requiring them to track originating data sources.

And even in less regulated environments, a problem in your data pipelines and workflows will require that you be able to quickly identify any changes that may have contributed to it. Without good data and process governance, tracking down such changes will take a very long time.

# Governing the Data Lake With Infoworks

## Automated Governance

**Infoworks automates governance of end-to-end engineering and DataOps processes.** From the automated generation of lineage as data pipelines and workflows are built, to enterprise-grade security, to the automated monitoring and full auditability of all system changes, you enjoy a complete solution that doesn't add governance as an afterthought. Infoworks automates the governance of your data lake as you are building it through capabilities such as:

**Role-based access control.** Infoworks supports the creation of users with different levels of user access, as well as domains, so administrators can control which users have access to specific data sets. Users within a domain can share data, pipelines, and workflows to enable team-based development of end-to-end data workflows and pipelines.

**Metadata management.** Infoworks automatically tracks and manages all metadata from source through transformations to consumption. This makes it easy to search for data artifacts across an entire data flow and simplifies the creation of data lineage for assessing the downstream or upstream impact of any data or metadata change.

**Enterprise grade security integration.** With the Infoworks agile data engineering platform, security is built in from the ground up. It provides security integration for user authentication and data security policies. And it supports single-sign-on/ LDAP integration, and Kerberos based authorization. Encryption support for data in motion and at rest is also provided.

**Change control and management.** Infoworks automatically tracks all changes that users make to data ingestion jobs, data pipelines, data workflows, the creation and modification of OLAP cubes and in-memory models. This makes it easy to track different versions of the overall data management process and roll back to earlier versions of a data pipeline if necessary.

**Tools for ensuring auditability and regulatory compliance.** Infoworks provides powerful tools to minimize risk from compliance issues:

- **User audit capabilities.** Infoworks provides audit logs that track who has created or changed data pipelines, cubes and workflows as well as tracking what changes were made and when.
- **End-to-end lineage tracking.** Data lineage is automatically generated and tracked from the source all the way to the cubes and in-memory models where the data is consumed in reports or dashboards. Auditors can view upstream data sources that contribute to a cube or to look downstream to see the data pipelines, cubes and models that consume that data.

Infoworks

# Governing the Data Lake With Infoworks

## Automated Governance

Case Study

### Leading Financial Services Company Accelerates Big Data Initiative with Infoworks

The market leader in portfolio management solutions, this firm manages vast amounts of financial data for clients, providing information vital to maximizing performance of their portfolios. But requests from clients for new views and data (other than standard reports and dashboards) could take six months to create—because the data was locked away in a giant Oracle database. Fearing to lose their competitive edge, they attempted to transform their products by implementing a more agile Hadoop and SaaS platform. But after five data engineers worked five months hand coding a solution, the end result had brittle pipelines that often broke and didn't meet their customers' service level agreements for analytics turn-around time. Turning to Infoworks, the company implemented a production-ready solution in six weeks with just one person.

WATCH THE VIDEO

> **"Our initial implementation was completed in six weeks by one person compared to our previous effort which took five months with five people doing hand coding."**
>
> Director of Digital Transformation,
> Fortune 500 Financial Services Company

# Ensure Smooth Sailing for Your Big Data Projects with Infoworks

# Ensure Smooth Sailing for Your Big Data Projects

## Your end-to-end enterprise-grade solution for creating and managing data lakes

**End-to-end automation: The key to agile data engineering.** Infoworks customers have implemented enterprise-scale analytics use cases in days instead of months using the **Infoworks Autonomous Data Engine**—and with 1/10th the engineering hours. Now you can automate data engineering and DataOps for big data workflow processes end-to-end: all the way from ingestion to consumption.

A lack of big data expertise keeps many organizations from even attempting to implement big data architectures like data lakes. Infoworks' agile data engineering platform automates the creation and operation of end-to-end data pipelines from source to consumption, so data lakes can be successfully deployed without requiring an army of big data experts. Infoworks' customers develop and launch 10 times the number of analytics use cases.

Reducing big data complexity by automating data engineering pipelines and workflows, Infoworks brings agility to your data analytics projects by delivering:

- **Automated ingestion and synchronization of data.** Infoworks provides a no-code environment for configuring the ingestion of data from a wide variety of sources. It continuously synchronizes source data with data in your data lake. And it automates the handling of slowly changing data and schemas—with support for streaming, batch and incremental modes of data ingestion.
- **Automated management of changing source data.** Infoworks automatically detects source side schema changes, adjusts for those changes and ingests the new columns automatically into your data lake. Data ingested into the lake is automatically validated for full and incremental loads.
- **Self-service data preparation.** An interactive, drag-and-drop data transformation capability with support for SQL-based machine learning and other transformations allows users to work with data in a collaborative, suggestion-based interface that reduces and even eliminates dependence on IT skills.
- **Dramatically accelerated query performance.** Just pointing a visualization or BI tool at a Hive table will result in very unsatisfactory end user performance. Infoworks automates the creation of in-memory data models and OLAP cubes that radically speed big data queries. Analysts just drag and drop facts, dimensions and measures and the Infoworks ADE builds a fully pre-aggregated and optimized cube—providing sub-second response times to user queries. Plus, Infoworks vastly simplifies creation of in-memory models—which provide much better query performance when compared with Hive.
- **Easier orchestration, tracking and management of complex data pipelines.** Infoworks automates the orchestration and ongoing management of complex data pipelines in production. Infoworks automatically retries, and restarts failed workflow jobs when possible, pauses, resumes and dynamically controls production data workflows while automatically monitoring data pipelines for performance. It also automatically generates and tracks data lineage from the source all the way to the cubes and in-memory models.
- **Enterprise class data governance.** Infoworks provides a wide variety of capabilities used to govern your data lake including: role-based access control and team development, audit logs for change control management and end to end metadata lineage for compliance reporting.

Infoworks

# Ensure Smooth Sailing for Your Big Data Projects

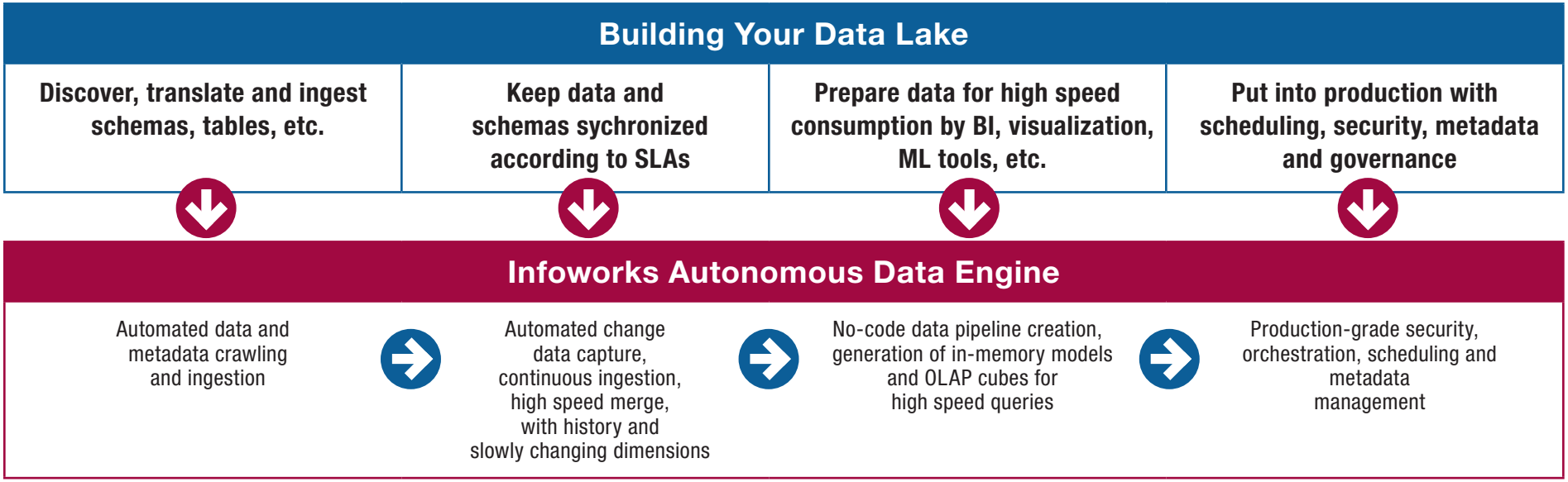
## Your end-to-end enterprise-grade solution for creating and managing data lakes

- **Simplified portability.** Infoworks is compatible with a wide variety of big data platforms both on premise and in the cloud. So as your big data environment evolves, whether on premise, cloud or hybrid, Infoworks makes moving easier. Data ingestion, transformation, cube generation and workflows built in the Infoworks designer can run in any supported execution environment without re-coding. Pipelines are not only portable, but are also performance optimized across execution environments.

With Infoworks you can rapidly build a scalable, manageable and governable data lake in days with zero coding. And easily handle incremental synchronization performance, and governance using capabilities that automate:

- Data ingestion for batch, streaming and incremental loading of data
- Data preparation of pipelines, cubes and in-memory models
- Ongoing end-to-end operational management of data flows
- Governance of your data lake for better management and compliance

## The Benefits of Building Your Data Lake With Infoworks




*Compare the difference between typical deployment processes versus automated data lake creation and management with Infoworks.*

Infoworks

# Ensure Smooth Sailing for Your Big Data Projects

Your end-to-end enterprise-grade solution for creating and managing data lakes

Case Study

## From Purchase to Production in Weeks not Months

A major oil and gas exploration and production company wanted to improve real time drilling analytics by consolidating onto a big data platform. But deploying the new system was projected to take months and an army of big data engineers until they automated the process using Infoworks—and one data engineer completed the project in just a few weeks.

WATCH THE VIDEO

> **“We faced hiring a small army of data engineers and months of development. But with Infoworks, one data engineer completed the project in a few weeks.”**
>
> Director of Data Engineering, Fortune 500 Energy Company

# Smooth Sailing

## Ensure Smooth Sailing for Your Big Data Projects

**Increase your pace for delivering analytics use cases to keep pace with rapidly changing business models.** With Infoworks, big data becomes a resource you can tap virtually at will. Which is why our customers are in production in a matter of days.

**Transforming the way companies benefit from big data.** Organizations across a wide range of industries have shortened deployment times for their big data analytic solutions by orders of magnitude without having to hire an army of big data experts. From healthcare and energy to finance and retail pharmacies, they've implemented enterprise-scale analytics use cases in days instead of months with Infoworks.

**Find out how Infoworks addresses the end-to-end challenges you face.** See how its end-to-end data engineering solutions automate most of the work for you—by applying unprecedented levels of automation to data workflows and data engineering. **Contact us to learn more.** We'll demonstrate in person or remotely what Infoworks can do for you.

### CONTACT INFOWORKS

**VISIT www.infoworks.io**
**EMAIL info@infoworks.io**
**CALL +1 650 391 9306**

> **"Infoworks provides a single development and deployment platform that works across all our environments with having to re-code."**
>
> Lead Enterprise Architect, Fortune 100 Retailer