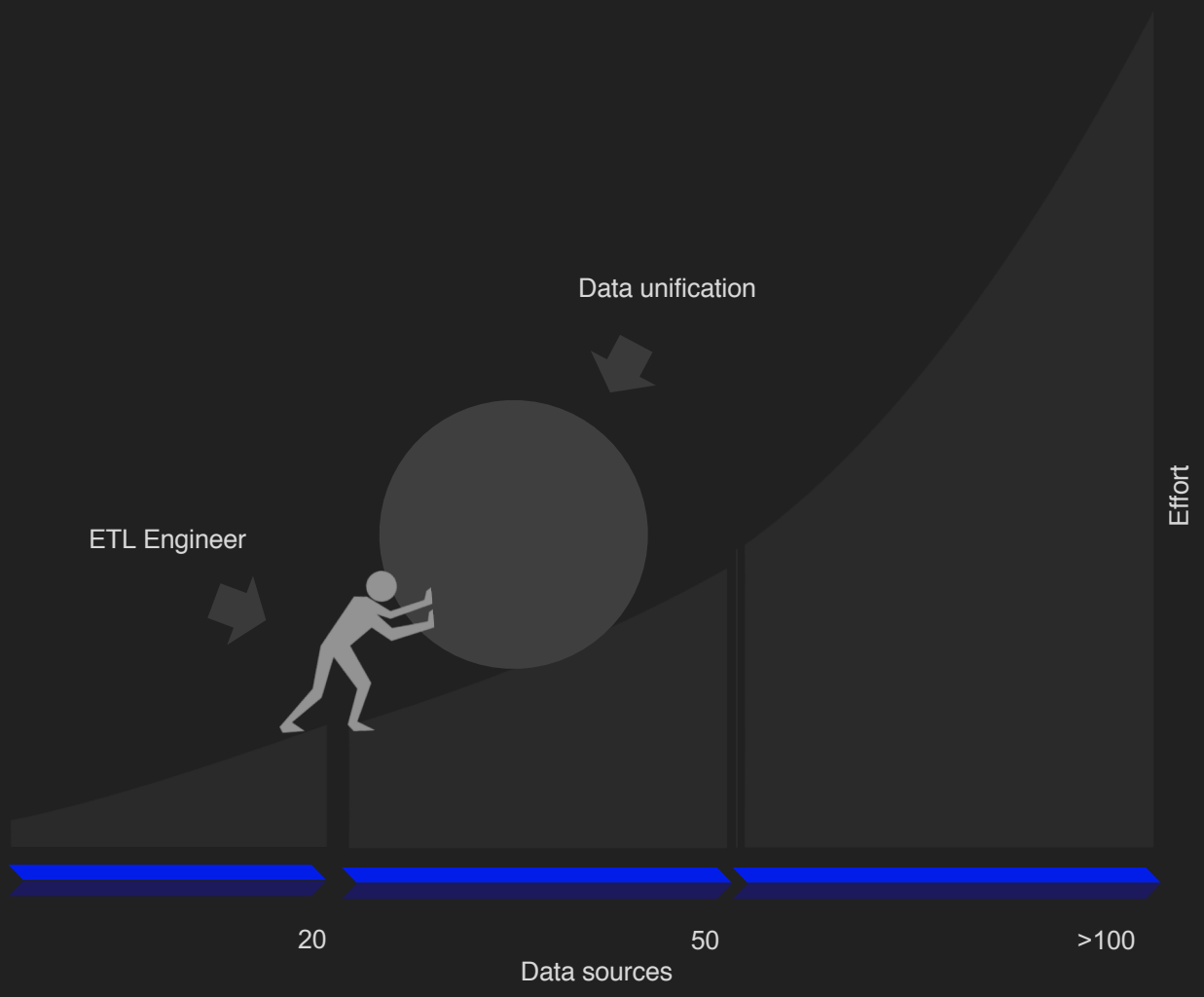
Eliminate Traditional ETL

WITH COLLABORATIVE DATA PREPARATION

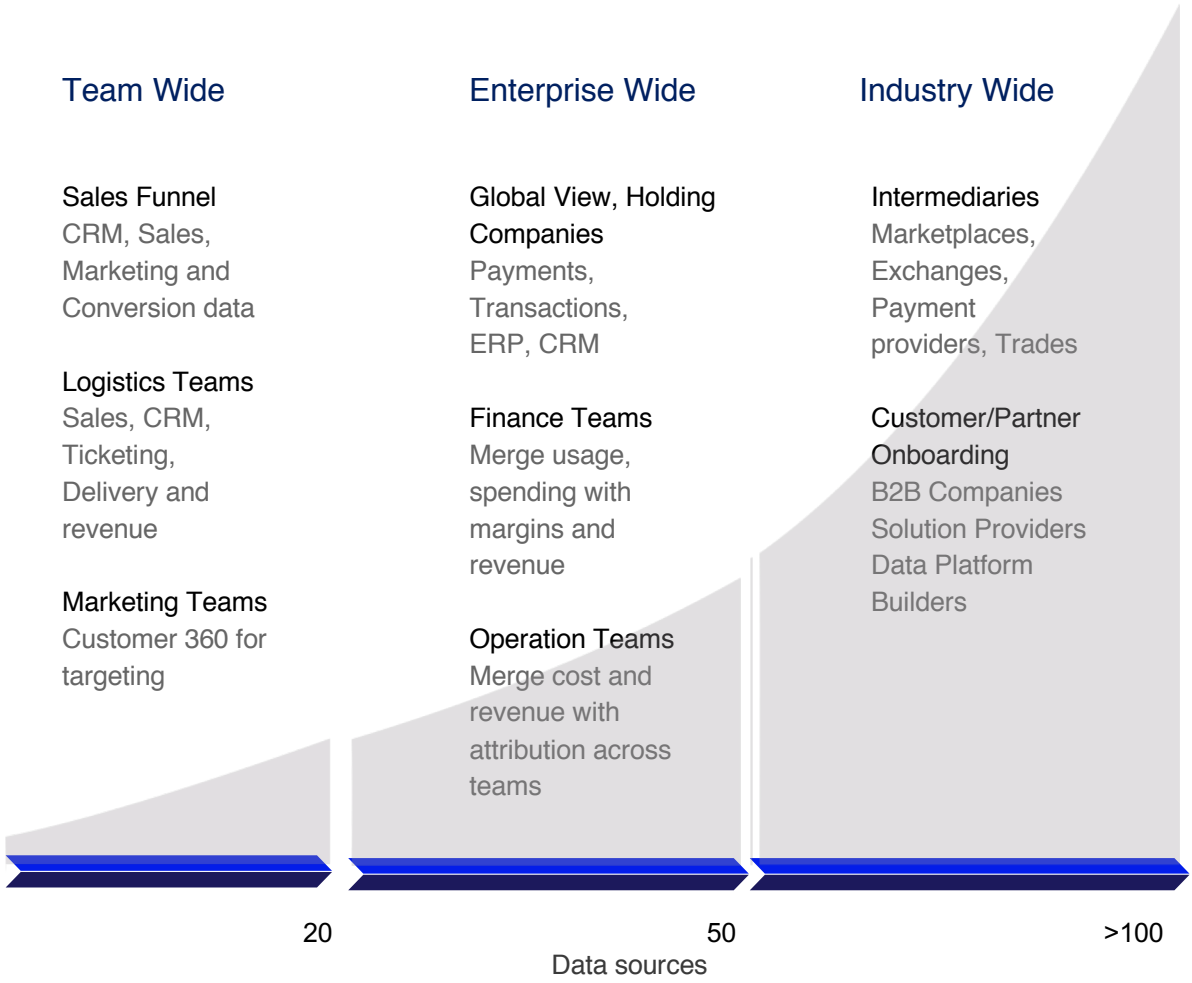
A Lore IO Whitepaper


Data unification
ETL Engineer
Effort
20
50
>100
Data sources

The race to unify multi-source datasets is on!

For years, companies of all sizes were encouraged to build their data lakes. Many heeded the call, built modern infrastructures and brought in their data. Some struggled with the end result and found themselves submerged in data swamps. But many others were able to plow through, and are now putting their data lakes into action through wide views of the data.

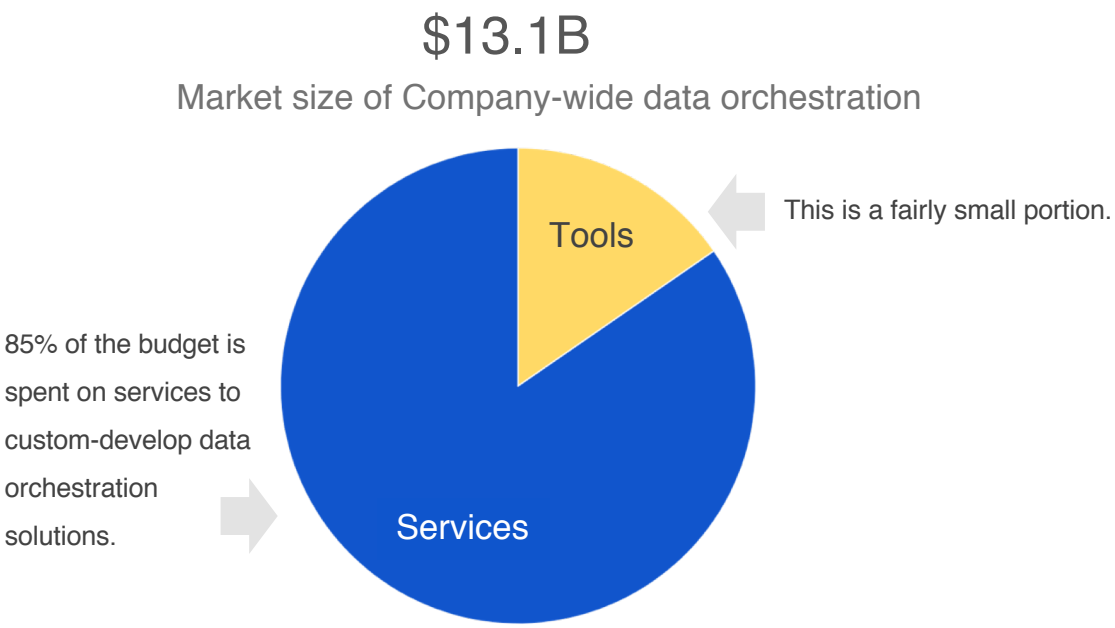
There is a new recognition now among business users that different use cases call for different number of data sources—as well as stakeholders—throughout the implementation. A small team who wishes to view sales funnel data poses considerably smaller data unification burden than a marketplace that must aggregate data from thousands of sources.



Solving the multi-source conundrum

To tap their large numbers of sources, may companies turn to their existing ETL tools only to reach scalability issues and limits. According to a 2017 Forrester research paper, the Data Orchestration is now a whopping $60B industry. Company-wide data orchestration that involves hundreds of sources is by itself a $13B market, where 85% of budgets are spent on services.

Such overreliance on services is a direct outcome of having to develop custom solutions to overcome traditional ETL limitation. Such services involve a heavy dose of custom programming to unify and orchestrate multiple data sources. These custom solutions will likely involve ongoing investments to keep up with the demand to unify and orchestrate a growing number of sources for newer data use cases.



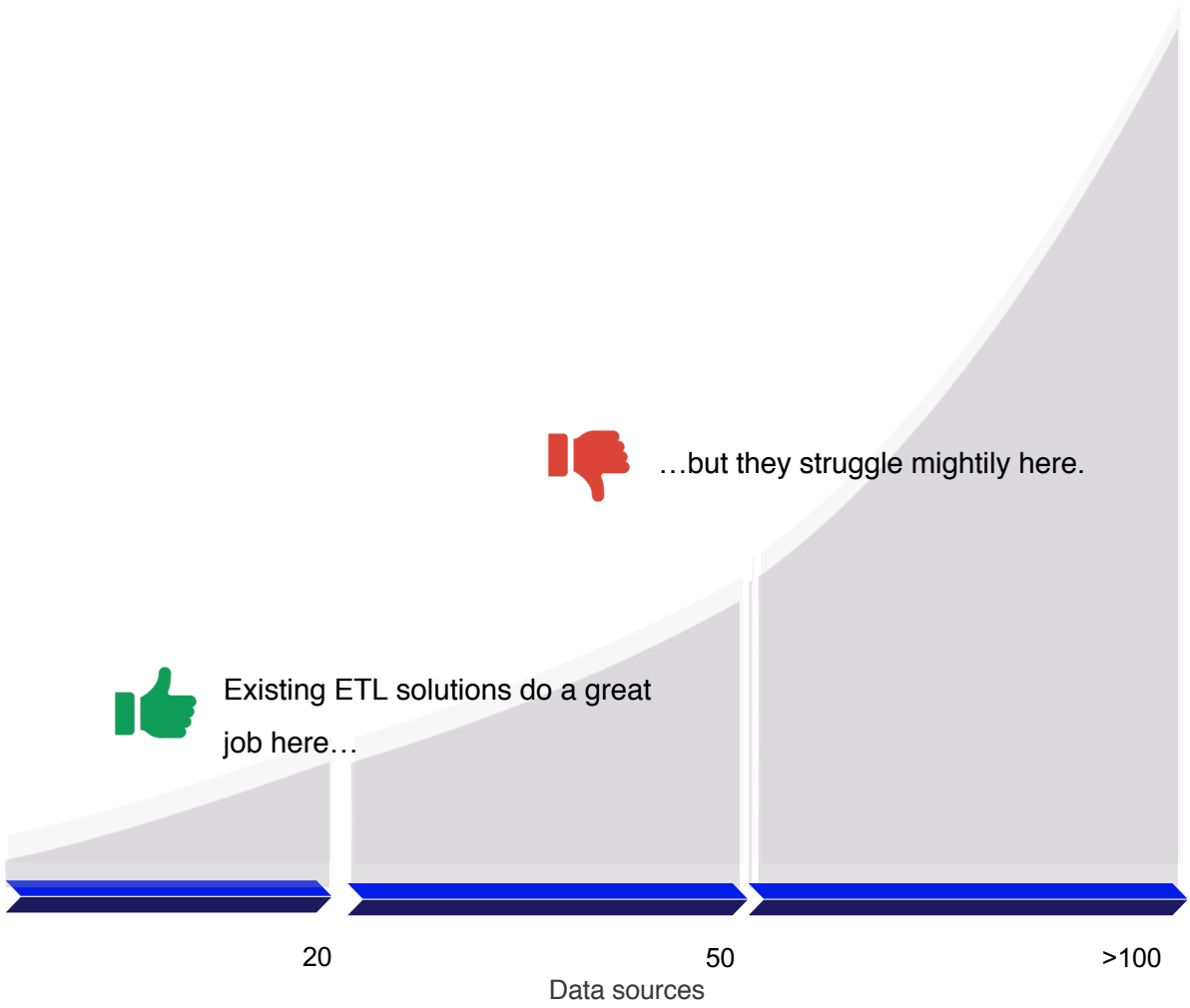
Scaling sources (and stakeholders) is hard

Turning to their existing ETL and data preparation tools that served them well in the "pre-data lake" world, companies are finding out that while such tools elegantly handle a small number of data sources and stakeholders, they become cumbersome to use and maintain for new use cases that involve more sources and teams.

Current solutions rely on procedural languages where ETL developers must create step-by-step workflow for data ingestion and transformations.

The more sources a business seeks to unify, the more complex are the mappings, and the more inter-connected these workflows become.

Even new solutions that offer self-serve capabilities for business users in individual teams struggle to perform as sources grow and a larger number of teams is involved in designing complex data definitions.



Long implementations slow down the business

"80% of data scientists' time is spent on data collection and preparation for analysis."
Forbes.com

"While the self-service data preparation market is growing 17% every year and is expected to become a billion-dollar market by 2019, it still seems to be one of the best kept secrets around."
Gartner Group

As the effort to unify more data sources increases, so does time-to-value. Business users who seek to develop broader and more meaningful views of customers, investments, and programs, they must longer on their IT counterparts to digest these new data requests, figure out how to incorporate them into the existing data infrastructure, design and implement a reasonable solution, and test and release it to the business. Time is a resource in short supply for business teams. They must answer a growing and more demanding questions in order to remain competitive in a chaotic marketplace. When new data is late to arrive, everyone in the business feels the impact.

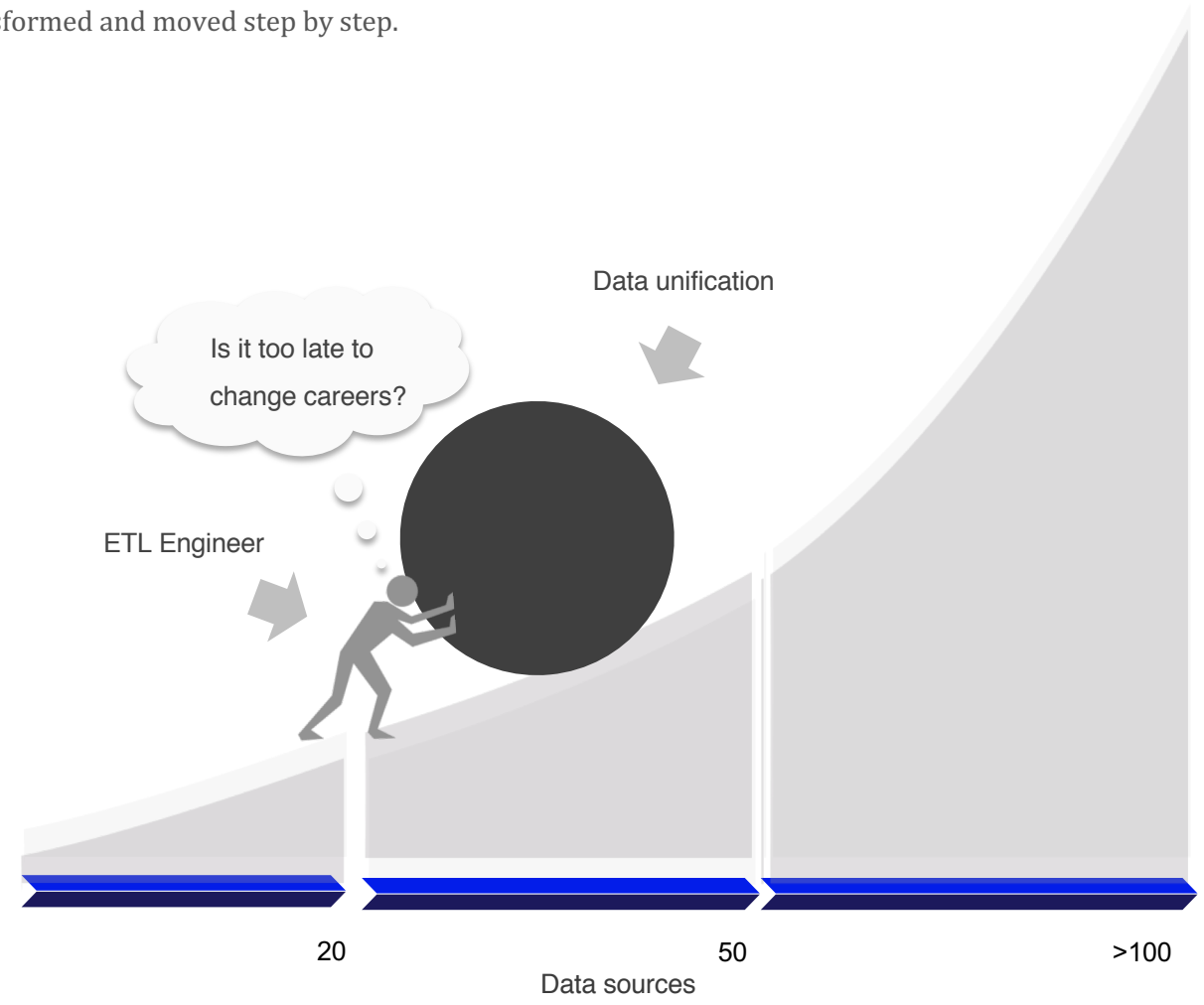
The lack of readily available data that is clean and trustworthy has huge implications on data analytics and data science teams as well. Several studies show that approximately 80% of analysts' and scientists' time is spent on data preparation.

This enormous waste of valuable resources has a lot to do with the growth of data sources, the reliance on less and less structured data that requires more clean up, and the ballooning volumes of data itself. Again, everyone in the organization is impacted.

Why do tools struggle with complex use cases?

ETL and data preparation tools can effectively handle a few sources, but using them in support of a data platform that handles dozens or hundreds of sources is a herculean effort. That's because at their core, ETL tools rely on procedural logic that must be codified in some programming language or the ETL tool's system. Engineers must design, codify and maintain procedures that describe the way data will be transformed and moved step by step.

The more sources and users an implementation involves, the more time and effort engineers must invest to develop and coordinate the growing number of transformation steps. They must ensure that new procedures don't break or conflict with previous procedures and that data sets are flowing in the right directions and at the right times.



Five ways ETLs hold businesses back

As noted, while ETLs are carried out by data and technical pros, their impact is felt across the entire organization, particularly by business groups who depend on readily available data for effective decision making and action. Here are five repercussions of relying on ETLs to run the business:

1 BUSINESS PROJECTS ARE BLOCKED OR SLOWED DOWN

Depending on the complexity of the data preparation involved, ETLs may require up to two years to thoroughly document business requirements, develop the ETL process, test the flow, fix any issues, and avail the data to the business. ETL processes become markedly more complex when new data needs to be loaded into the enterprise data warehouse (EDW); for instance, pulling data from PDF reports. Businesses find themselves in a harsh dilemma: delay project execution and sacrifice business agility and results, or limit data use and analysis to basic and readily available measures that impede insight generation and lower business performance.

2 DIMINSHED BUSINESS AGILITY

Data starvation due to protracted preparation isn't limited to humans. Businesses are adopting artificial intelligence and machine learning capabilities at a rapid pace to better anticipate key business and industry events and automate adequate responses to them. But for machines to learn quickly and effectively, they must be fed a diet of high quality data, which entails complex preparation. ETLs hold back machine learning in two ways: ETLs may not adequately handle unanticipated variations in the data that can slow down learning, or applications of machine learning may be constrained to simple use cases that prevent businesses from reacting quickly to all critical issues and changes.

3 SUBOPTIMAL BUSINESS DECISIONING AND BUDGET ALLOCATION

ETLs can trip businesses when they involve redundant data copies and convoluted data transformations. As data pros add more ETLs and data marts in response to business requests, data quality, lineage, and overall trustworthiness degrade. The business loses the single source of truth of its data, cobbling together metrics with vague or misunderstood definitions.

Consequently, analytics dashboards and reports get populated with dirty data, causing managers and executives to suspect, if not altogether ignore, the data, or trust bad data. The end result is poor decision making that derails business performance.

4 SUBSTANTIAL DATA MANAGEMENT OVERHEAD

As the business relies more heavily on data over time, it typically accumulates dozens, if not hundreds of different ETL processes that it must maintain and manage. Based on varying business requirements, these ETLs include manipulation logic that is deployed in different locations and configurations. Data pros may develop data manipulation scripts in programming languages such as Python or Perl, write complex SQL in tool like Hive or Pig, or make data manipulation either manually or in code directly in Excel. Having manipulation logic reside in different systems and formats — hard-coded directly into the ETL process — can present significant management challenges and management overhead, preventing the business from scaling its data consumption.

5 DECREASED EMPLOYEE PRODUCTIVITY

Ample evidence suggests that data preparation can take up to 90% of the development time and cost in data warehousing and data analytics projects. Data analysts and scientists seeking to exercise their creativity and interpretation skills during data analysis and modelling, regularly find themselves burdened with rudimentary and lackluster data design and cleansing chores, or altogether ineffective when failing to access critical data. Whether they chase after data, pre-define tables and schemas, or languish in its transformation, hard to find and retain data pros are prevented from fulfilling their potential, which results in low morale and high churn, impeding the growth and stability of the business.

ETL issues — Close examination

To better understand and appreciate the complexity of ETLs and their negative impact on the business, it's worth exploring the root cause of ETL issues. The next three sections dig deeper into the extract-transform-load phases and the challenges they present.

1 EXTRACT ISSUES

Several business use cases, such as having 360-degree view of customers, campaigns, or employees require ingestion of data from multiple systems of records. The plurality of data sources introduce several challenges:

DISPARATE DATA SOURCES

Businesses rely on a variety of technology components to run their daily operations. Data for analysis, there- fore, may be pulled from different applications that are designed, built, and maintained by different vendors, whereby each application requires a different mechanism to access the data. Data pros may therefore have to learn multiple interfaces and regularly maintain them even within a single ETL process.

OBSCURED SOURCE SCHEMA

Use Lore IO's UI, APIs or let analysts & business users to use the BI tools they know and love for self-served business exploration.

DISPARATE DATA STRUCTURES

The growth and unpredictable nature of data have ushered in new ways of storing data, especially in non-relational databases. Data may be fully structured, semi-structured, or unstructured. Data pros must now access data that is housed in different and unfamiliar formats where access via the well established SQL is not possible.

LACK OF DATA HYGIENE

Business data is typically generated by both humans and machines, which introduces discrepancies in the degree to which data is complete and clean. Data pros must account for a greater variety in the state of the data than ever be- fore, further complicating data extraction.

DIFFERENT DATA REFRESH CYCLES

Source systems create and refresh their data at regular intervals. Sensors or social media applications stream vast amounts of data in real time. Some business applications sessionize or aggregate data that is made available only on a weekly or a monthly basis. Analysts who pull disparate data elements must account for varying levels of data freshness.

FRAGMENTED SECURITY POLICIES

As data security captures more attention and mind- share, source systems implement more robust and often different security and governance practices that data pros now must account for and comply with when accessing data.

DISPARATE DATA STEWARDSHIP

Finally, source systems are managed by different teams and departments inside and outside the organization. Data pros must spend time identifying the right data owners and then negotiate and comply with their different data access and use guidelines.

2 TRANSFORM ISSUES

With data extracted from its original source and brought into the EDW, data pros must now transform the data into a usable form that will enable them to blend disparate data sets, organize it, and ready it for consumption by the business. Given the richness and complexity of data, the Transform phase presents numerous challenges:

FINDING COMMON KEYS

Some business use cases, such as customer analytics, force data pros to join many different data sets. A unique and trustworthy key by which to stitch disparate data sets may not be always available, requiring analysts to spend valuable time constructing a reliable key to ensure that the right records are being joined.

LACK OF DATA LINEAGE

Data may undergo several transformations over time in support of different business use cases. Data pros often lack data lineage to fully understand what changes occurred in the data from one transformation to the next. This lack of visibility can lead to the use of the wrong version or generation of data, leading to misinformed analysis later.

DEDUPLICATION AND DERIVED COLUMNS

Data analysts and scientists must contend with data sets that require different levels of manipulation. For instance, some data points may be ready for use, while others may be encoded for privacy or brevity purposes, and must be decoded before use. Some business use cases may require the creation of new calculated metrics that derive their values from other metrics, or the generation of surrogate key values. Data sets might accidentally or unnecessarily repeat themselves, requiring a de-duplication effort.

COLUMN SPLITS AND MERGES

The organization of data may vary dramatically during the transformation phase. Some data might require sorting or ordering based on a list of columns. Data might need to be aggregated, or disaggregated, trans- posed, or somehow pivoted. A data column might need to be split into several distinct columns, or vice versa.

VALIDATION AND QUALITY

Finally, the unpredictable nature of data means that the range of data values extracted can exceed the original expectations of data analysts and scientists. To mitigate this problem, the transform phase may get further complicated by incorporating a process to look up and validate the data from tables or referential files.

3 LOAD ISSUES

Once data is transformed it's ready to be loaded into target systems that business users will inter- act with to derive insights and actions -- the very purpose for the ETL process overall. The Load phase introduces unique challenges that further complicate and prolong ETLs:

SCALE AND VARIATIONS

Data pros may need to handle very different target systems, from simple flat files to enormous data warehouses.

TESTING, PUBLISHING, ARCHIVING

The load process itself may vary from one use case to another. It may involve a different number of tasks, such as testing, approvals, and archival.

BATCH VS STREAM VS REAL-TIME

The scheduling of data delivery can vary dramatically, from real time applications all the way to weekly or monthly data load.

UPDATES AND CORRECTIONS

The method of delivery can vary: some applications may call for updating existing records with new data, while others may depend on the creation of new re- cords.

TARGET SYSTEM RESPONSES

When loading data, analysts must comply with the different data constraints of the target systems, such as uniqueness, mandatory fields, or referential integrity. Analyst must effectively respond to the acceptance and rejection responses they receive from the target systems.

ACCESS

Data pros must conform to any access restrictions, encryption levels, or security policies of the source systems and preserve them in the target systems for each data consumer to ensure that only permitted users and applications can access sensitive data.

How to Eliminate Traditional ETL – A To-Do Checklist

Start at end: Declare your output views

Conventional thinking starts with the source data and maps the various transformation procedures needed to shape the data for its target destination.

Conventional thinking doesn't scale.

A better approach is to declare the desired end state and work backwards—ideally programmatically—the various tasks that make up the transformation logic.

In plain English, data analysts (rather than ETL developers) control the data prep process. These analysts use declarative languages, such as SQL or SQL-like, to describe their data needs.

When you start with a declarative language and add automation, you can truly reimagine the whole data prep process and make it agile and scalable.

Data prep approaches that leverage declarative definitions are proving to be much more scalable and agile than traditional procedural ones. Rather than a waterfall model, declarative transformations have a relaxed, cyclical structure where each transformation task can be carried out independently.

Declarative transformations enable business users to participate in the data prep process from the get-go.

Analysts define their target data views. Subject matter experts annotate the data elements and the system translates these declarations to ETL code that can be run at any step of the process, evaluated and fixed as needed. Business users can evaluate the data at any time and request modifications to the preparation process.

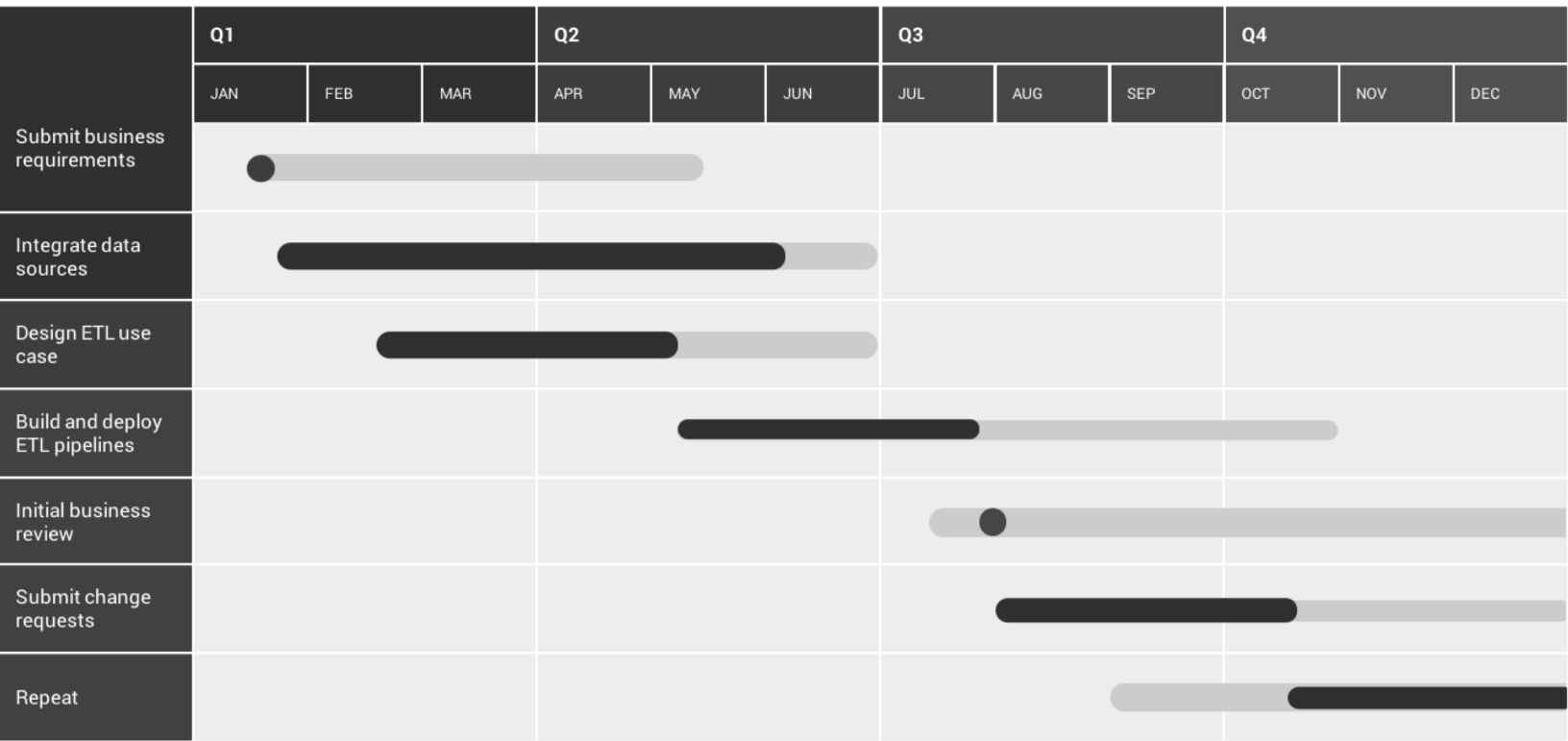
14

Migrate out of procedural logic

Procedural transformations follow a traditional approach where the ETL owners receive the reporting requirements from the business team, and then design, build, test and optimize the implementation on their own before providing business users with the prepped data.

Many of the procedural tasks are carried out sequentially and potentially over a long period of time based on the project's complexity. Business users cannot access—let alone use—the data while it's being worked on. And any subsequent requirement to modify the data may impose a whole new sequence of steps.

Procedural logic is imposing a waterfall-like business workflow that is too slow to adopt for use cases that involve many data sources.



Automate ETL coding

Think of declarative language as the "commander's intent"—They convey the desired outcome without specifying any implementation details because those may change at any time based on resource availability, the emergence of new technologies, the use of new data sources, etc.

It is much more desirable, therefore, to delegate the implementation process to machines that are much better suited to handle increasing logic and source complexity than mere mortals. When automated systems handle and even optimize the spaghetti salad of ETL job code, the business can onboard and unify new data sources at a much faster rate, enabling analysts to focus more time on generating insights from data than on prepping the data.

A good example of how automation helps data preparation is the process of joining disparate data sources. Following a procedural model, an ETL engineer must specify the various mechanisms to unify data sources by mapping data keys and using other techniques. As the number of sources grows and as the data schemas of the underlying sources change, it becomes increasingly painful for humans to manage those relationships, slowing down implementations. Automation on the other hand, can control the various data entities, attributes, and key relationships, and by so doing take care of all the necessary data joins on its own. Users should always have full lineage to discover and understand how their data elements come together, but the actual unification can (and should) be done by machines.

Collaborate on logic creation

Since declarative transformations enable more stakeholders to be involved throughout the data preparation process, it makes sense to divide the workflow into small tasks and assign the responsibility for completing each task to those best positioned to address them.

This divide and conquer process is akin to a potluck party where every guest contributes a meal item that they are feel they can best manage. Similarly, the data preparation workflow can be divided up and across those who best know the source data, those who understand the business logic, those who create the final views, etc. Stakeholders collaborate on the whole process by contributing at their own pace.

Once stakeholders collaborate on data definition and business logic, they are more likely to want to use the new data elements in their reports and applications. A global data fabric can become the single source of truth for the entire organization, expressing the ins and outs of the business in a common and standard language that everyone can understand and trust.

Automating declarative transformations means that the system runs its transformation code directly on the source data. This means that data discovery and cataloging – while virtual – is directly coupled with the actual data, so that users can issue data queries as they study – and even refine – the fabric, building new data definitions upon previously created ones.

Capabilities to look for in a new ETL solution

Below are several capabilities that should be included in whatever solution you choose to eliminate traditional ETL:

AUTOMATED DATA SCANNING AND SCHEMA DETECTION

New sources are periodically added, and existing ones may experience schema changes over time. Look for a solution that can scan your data landing area, ingest new sources, and adapt to schema changes – all automatically – so that the business isn't blocked.

AUTOMATED DDL (TABLES, COLUMNS, DESCRIPTIONS)

Since data changes so often, manually handling its physical persistence can create business bottlenecks. It's best to allow the business to focus on creating data logic, and delegate physical considerations to automated systems.

VIRTUAL DATA CANVAS

To remain competitive and effective, the business continuously evolves its data needs and use cases. A new ETL solution must offer business users an intuitive environment where they can extend their data semantics, further transform their data, and continuously develop their data fabric without having to wait on IT to build new data pipelines first.

STATEFUL DATA COMPONENTS

Since many data transformations require the involvement of numerous stakeholders and systems, such transformations are best developed incrementally. The business needs to be able to break data requirements into small components, and manage their progression (or states) effectively.

DATA COMPONENT OWNERSHIP

Clear ownership of the above mentioned data components improves team collaboration, enabling stakeholders to interact on their data components throughout their development lifecycles.

DATA COMPONENT DEFINITION

To democratize data and enable business users to explore, use, and build on data components, there needs to be a centralized catalog that lists and defines all data elements.

DATA COMPONENT LINEAGE, HISTORY AND VERSIONING

Stakeholders must have confidence in their data. To reach a state of "single source of truth", teams must be able to understand how data elements are constructed and derived, as well as to trace back data to its source.

REUSABLE DATA COMPONENTS

The data fabric must be modular; stakeholders who define new outputs should be able to use these outputs as inputs for newer still components. Therefore, the component creation process must include the ability to publish them into the data layer and avail them similarly to raw data elements.

TESTABLE DATA COMPONENTS

Data quality and accuracy need to be assessed and validated as soon as possible; businesses can no longer afford to wait until the entire data pipeline is established. Rather, data components must be testable right on the raw data as soon as their definitions are created.

USERS CAN REQUEST NEW DATA COMPONENTS

As the need for more and newer data increases, users must be able to define desired outputs without taxing data producers to rearchitect the data pipelines. The system must be flexible enough to accommodate new requests, even when those may profoundly alter data models and schemas.

AUTOMATED MICRO-TASKS TO ADVANCE COMPONENT STATE

To increase business agility, an automated system should be able to define design and operation tasks ("micro-task") to known stakeholders to ensure data components are created and made available quickly.

MICRO-TASK USER COLLABORATION

Multiple stakeholders (presumably from different areas of the business) must be able to work together on data components without "stepping on each others' toes."

MICRO-TASK OWNERSHIP ASSIGNMENT AND MANAGEMENT

Collaborative data management means that stakeholders are able to task each other and assign responsibilities to participate in the design, creation, testing, publishing or optimization of data components.

AUTOMATED UNIFICATION OF MULTIPLE DATA SOURCES AND TYPES

An automated system should be able to identify common characteristics across data sources and types (such as email addresses, or resource URIs), and use this knowledge to unify datasets for faster data discovery and querying.

INTUITIVE DATA EXPORT

Finally, a new ETL solution must enable business users to export any data component or view to one or more target systems easily and quickly, as soon as new data is made available.

Putting it all together

Collaborative data preparation offers a new alternative to traditional ETL. It overcomes the scale limitations that impact traditional workflows as they attempt to unify a large and growing number of data sources. This form of agile data preparation enables stakeholders to collaborate and gradually build out the data fabric by incrementing their data definitions as new sources and use cases emerge.

At the same time, business users can reap quick returns from new definitions by immediately incorporating them into their reports and applications. They no longer need to wait until their ETL engineering counterparts fully develop and deploy the data platform.
By delegating ETL code generation to systems, businesses can more quickly expand and deepen their shared understanding of their data, and evolve their common language to better express the nuances of their business.

Lastly, the collaborative nature of this new approach both democratizes data preparation – inviting new stakeholders to participate in the process – and decentralizes data ownership, inviting those who consume data to contribute to the growth and ongoing health of their data fabric.

About Lore IO

Lore IO is a Collaborative Data Preparation Platform that helps companies ingest and unify disparate data sets from hundreds or thousands of sources. It generates standard outputs without the need for engineers to develop procedural ETL and data pipelines.

Lore IO customers unlock the full value of their data by empowering business users to collaborate on and use datasets that are initially hard to understand, reconcile, and blend.

The Lore IO platform abstracts all the complex semantics of how the data is captured and joined together, enabling customers to instantly validate business logic in support of a wide range of use cases.

Lore IO takes an agile approach to partnering with new customers. It seeks to explore strategic projects that will make a material impact on the business and then structure a partnership that demonstrates value immediately and scales from there.

www.getlore.io