GIGAOM

RESEARCH BYTE

How Data Prep and ETL Overlap in the Cloud v1.0

Cooperative, Enabling, and Operational

ANDREW J. BRUST

SPONSORED BY TRIFACTA

How Data Prep and ETL Overlap in the Cloud

Cooperative, Enabling, and Operational

TABLE OF CONTENTS

1. Summary

Today's data analytics stack is a paradox, representing both fundamental change and sustained tradition. On the one hand, we have significant new paradigms, including the data lake, and a number of new technologies too, including Spark and Hadoop. On the other hand, we are still working with BI tools, we are again using Structured Query Language (SQL), and not only is the data warehouse still with us, but it has taken the lead in modern analytics.

Really, though, the change is pervasive. Even the stalwart technologies and principles that we've conserved have been affected by broader shifts in the technology world, and the cloud, not surprisingly, is chief among them. Virtually the entire analytics stack has moved to the cloud. That is true for both storage and compute, the scalability and elasticity of which are most responsible for the adoption of the data lake and the resurgence of the data warehouse. And almost dwarfing those phenomena, cloud-based machine learning (ML) has taken off.

Each of these cloud data components is important on its own. But data preparation is, in many ways, the connective tissue that makes them work together. It is data prep that makes the data-driven whole greater than the sum of its lake, warehouse, and ML parts. That may sound like hyperbole, but in fact, it is an understatement. Data preparation moves data within the cloud. It moves data between clouds. And data preparation moves data to the cloud as each organization adopts it.

With an increasing number of organizations now using the cloud for storage and analysis of raw unstructured data, as well as for the design, training, and retraining of machine learning models, doing data prep in the cloud has become critical. Data prep is the vehicle for graduating unstructured data in the data lake to become structured data in the warehouse, delivering a platform for reporting and analytics. In machine learning workflows meanwhile, data prep is the key to transforming and streamlining data sets down to the relevant, cleansed columns needed for use in models, and one of the best approaches to feature engineering.

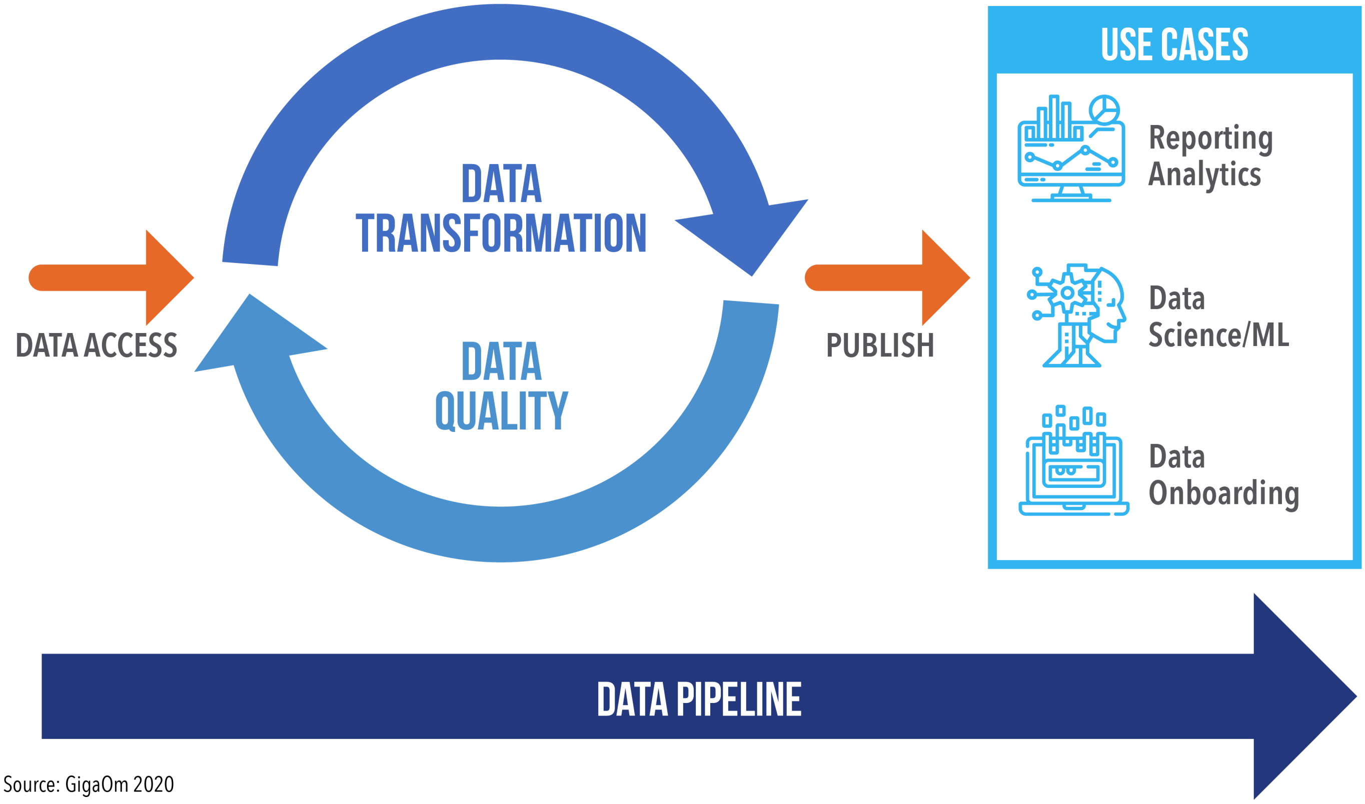
All of this combines to make refinement of data in the cloud a critical task, if not *the* critical task in the enablement of modern analytics. As such, cloud data preparation has major ramifications for the way people, processes, and technologies manifest and function.

2. What the Process Looks Like

The process of data prep in the cloud parallels its on-premises counterpart, but there are important differences. In historical on-premises scenarios, data preparation focused on production Extract, Transform, and Load (ETL) pipelines. In the cloud, however, data preparation has a significant data exploration component, potentially more so than on-premises—for example, covering profiling and checking of data distribution and quality. Across both scenarios, that means data preparation helps business users find new insights from data, rather than just helps data engineers operationalize the process around existing insights. Both capabilities are important, of course, but tactical platforms for data pipelines that are sequestered from exploratory capabilities impede the experimental data work necessary to help companies compete, improve, and transform.

With today's data volumes and variability, that exploratory work needs to be done in a powerful and efficient manner, so that data can be queried and transformed at scale. And that scale works on two issues: it's not just individual data prep tasks that are demanding, it's the number of jobs that must be run, often concurrently. The reason for this concurrent demand is that data prep must serve people from multiple roles. Analysts, data engineers, and data scientists each have significant data prep needs; they must each be served well and must all be accommodated together.

These constituencies also need to share data and collaborate in such a way that the data prep efforts of one accommodate and support that of the others, upstream and down. Data preparation works in service of these needs, too, ensuring a shared platform for, and common sense of, data lineage, governance, and trust.



Source: GigaOm 2020

Figure 1: Data Preparation

Build the Initial Pipeline

Once the initial exploratory work is done, you will start to get a sense of what the data you are working with can tell you. Now it is time to standardize, clean, and analyze it. Pick the metrics you want to look at. Pick the categories or attributes you want to aggregate those metrics by or the drill-down analysis you would like to do. Or, for machine learning, you may instead wish to pick the columns you want to predict and determine which columns will inform those predictions. As you manipulate the data to get the metrics, categories, features, and targets you need, you will essentially be doing data prep, interactively. Doing analysis and prep in tandem this way bears out how the value of an analysis drives the demand for data pipelines in the first place.

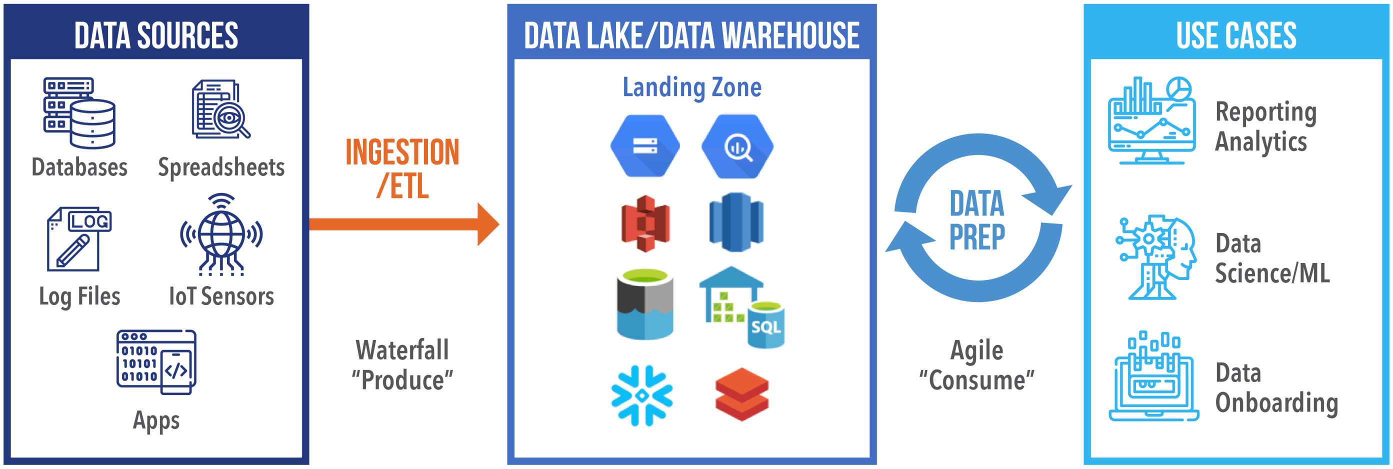
Next, it is time to build these data prep steps into a repeatable pipeline. That serves more operational needs, to be sure, but even here, the process is an iterative one. You will start by building and running an initial pipeline, viewing the output (or previewing a sample of it) and refining it as you go. Along the way, you may be inspired to work through additional analysis, then develop corresponding new data prep steps in your pipeline. Then you will test and refine those as well. Enterprise organizations that have implemented full DevOps practices may then even take their pipelines through formal development, test, staging, and production cycles. As you go through these steps, you'll see how all of them combine to form a virtuous cycle. And remember, it is the cloud that enables you—and your colleagues—to execute this virtuous cycle, concurrently, at scale.

Rigor, Governance, Flexibility

Different organizations will have team members from different roles doing this initial work. In larger enterprise organizations, it may well be data engineers who take over after the initial discovery work is done. In smaller organizations, business users may do the lifting, given both their keen interest in the insights that the pipelines will help drive as well as the requirement in many smaller organizations for businesspeople to perform technical tasks using self-service tools. The key is to use a data preparation platform that accommodates both groups. In order to do that, the platform will need to be rigorous enough to serve advanced users yet work in a self-service capacity for business users, too.

Similarly, some organizations will have a handoff of pipelines from business users to data engineers so that the self-service pipelines can be promoted to IT assets. Here, the data prep platform's aforementioned rigor can be utilized to refine, hone, and perfect the self-service pipeline. In other shops, the business user may be deputized to take her pipeline all the way to production. The reality is that some organizations will prefer, require, or be compelled by compliance obligations to impose a tightly managed, highly governed process. Others will prefer the agility that comes with a more permissive approach.

Ultimately, everyone who has access to raw data, often in text file formats, has the potential to be a player in the pipeline authoring arena. This will depend, of course, on their motivation and their self-interest in leveraging the data for analytic purposes, as well as their technology skills, ability, and enthusiasm level.



Source: GigaOm 2020

Figure 2: The Analytics Workflow in the Cloud

Now, It Is Time to Operationalize

Now that your pipeline has been prototyped, tested, iterated upon and perfected, it is time to put it into production. One big difference between running your pipeline on an *ad hoc* basis and running it in production is not just that it will run on a scheduled basis, but that it will be able to do so at high frequency. The days of loading data once daily are over. Data is generated continuously, and it should often be ingested frequently.

Remember, cloud data prep platforms run at great scale—they can handle the data volumes you throw at them, the frequent execution at which you will need to run them, and the requirement to run multiple pipelines concurrently. If you've designed your data pipeline well, its performance on a cloud data prep platform should be excellent.

Trust, but Verify

For operational success, make sure to monitor the operation of your pipeline, and do so from multiple angles. Check the pipeline in empirical terms, by confirming it is delivering data to the destination systems and that the delivered data appears correct in structure and content when subjected to spot inspections. Check the status and logged output on the data prep side, too—this is the best way to stay proactive and keep your pipelines in top shape.

Beyond spot checks and log inspection, though, you will need to check the data quality methodically on both sides. These data quality checks are not just important when a pipeline is first deployed; after all, source data that is not under your direct control is subject to change. As a result, data quality checks are important to do on a sustained basis. You and your team need to be vigilant on quality checks and in a state of readiness to tune your pipelines, in order to accommodate potential changes in the quality or the structure of source data.

Additionally, if the content of the source data changes in some categorical way, the transformations in your pipeline may need augmentation. This further underscores the importance of readiness to update your pipelines and keep them running smoothly. And keep in mind that needing to respond to change does not represent a setback—it is just part of the process that keeps data pipelines effective.

Passing the Audition

The self-service nature of data prep platforms introduces another monitoring and tuning scenario, one that should be thought of as an opportunity rather than a maintenance task. Specifically, sometimes certain business units will create their own pipelines for their own analyses, which then are shared with colleagues and may appeal to them as well.

In these cases, departmental pipeline assets may be a transition to enterprise ones, with ownership of them possibly shifting to IT and away from the business unit. In such a scenario, IT will often wish to modify the pipeline or re-author it to make it run as efficiently as possible, and to enable it to handle the broadest range of source data and destination systems. In this case, the pipeline is re-operationalized because of its transformation from a departmental to a company-wide domain.

All of this applies to a number of destination repository types, including data lakes, data warehouses, and machine learning platforms. For lakes, the pipelines will be more focused on data movement; for warehouses, they will be focused on transformation. For machine learning models, which are often built on a more *ad hoc*, less orthodox basis, pipelines will likely be focused on both.

3. Wrapping Up

ETL has been with us for decades. Data preparation modernized the process of data movement and transformation, giving it self-service capabilities and AI-driven smarts, all while retaining the original rigor. Data preparation in the cloud takes all of these advances and brings them forward to modern repositories, including cloud data lakes and data warehouses. Cloud data prep also broadens the array of data sources feeding these platforms and handles the vastly increased data volumes and frequency of data ingestion that now apply in many analytics use cases.

Data lakes and data warehouses co-exist and overlap but fulfill distinct roles. While the warehouse is best for determining “known unknowns” on a routine basis for operational needs, data lakes allow more experimental *ad hoc* analyses around “unknown unknowns.” These *ad hoc* analyses may provide answers to one-off queries or may turn out to address operational needs against data that may migrate from the lake to the warehouse. So be it. In addition to ML platforms, cloud data preparation brings data to both the lake and the warehouse and deftly moves data between them, too.

That is why moving data to the cloud, as well as between and within different clouds, are critical capabilities that your organization’s efficiency and competitiveness will increasingly rely upon. Meanwhile, just because cloud data preparation is serious business does not make it overly complicated, or inflexible. Instead, as we have covered in this paper, cloud data preparation employs a disciplined but versatile process of data exploration, pipeline construction, and operationalization.

Each of these steps is iterative. Each of them encompasses the changing nature of data and analysis work. Each phase in the process is also versatile enough to be rigorously governed and tightly managed, or open to *ad hoc*, more democratized access and requirements. The data management and analysis needs of different organizations are themselves different. Cloud data prep recognizes and accommodates this diversity of needs, and facilitates the transition to data-driven operation, for organizations of all types and all stripes.

4. About Andrew Brust

Andrew has held developer, CTO, analyst, research director and market strategist positions at organizations ranging from the City of New York and Cap Gemini to Gigaom and Datameer. He has worked with small, medium and Fortune 1000 clients in numerous industries and with software companies ranging from small ISVs to large clients like Microsoft. Andrew's resulting understanding of technology, and the way customers use it, makes his market and product analyses relevant, credible and empathetic.

Andrew has tracked the Big Data and Analytics industry since its inception, as Gigaom's Research Director and ZDNet's lead blogger for Big Data and Analytics. Andrew co-chairs Visual Studio Live!, one of the nation's longest running developer conferences. As a longtime technical author and speaker in the database field, Andrew understands today's market in the context of its longtime Enterprise underpinnings.

5. About GigaOm

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization to apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.

6. Copyright