Datameer

Solution Brief

Agile Data Pipelines for Cloud Data Warehouses

Datameer Spectrum

The benefits of cloud data warehouses go far beyond cost savings for organizations. Thanks to their ease-of-use, speed and near-limitless scalability, cloud data warehouses give business teams the ability to deliver more analytic value faster than ever.

Traditional, on-premise data warehouses are expensive to procure and require complex IT setup. This limits an organization's ability to capitalize on faster insights and scale on-demand.

Cloud data warehouses provide a number of advantages over legacy data warehouses:

- Easy, rapid deployment
- Nearly limitless scalability
- Cost-effective elastic pricing
- The ability to expand quickly

Cloud data warehouses increase the speed to insight for business teams and help organizations remain agile as their data and analytics programs scale. However, many organizations still rely on slow, traditional methods for feeding their cloud data warehouse.

The mentality of cloud analytics is different than on-premise. Gone are fixed projects with highly static requirements. The new approach is iterative and exploratory. Business analysts spend more time in self-service mode, where they explore data looking for answers.

To facilitate this new iterative and exploratory approach, organizations need a modern approach to delivering data to their cloud data warehouses. Data delivery and the cloud data warehouse need to complement each other to deliver the desired agility.

Data Movement Challenges

Data is only useful when it is available for analysis. The faster analysts can get their hands on the data, the more valuable their insights will be for organizations. This is why legacy data movement methods are so detrimental. The agility of cloud data warehouses is undermined by legacy movement process like batch-loading ETL.

ETL cannot handle the volume, variety, and velocity of data in today's big data world. Creating data pipelines with ETL-style tools can take weeks, if not months. And the technical nature of the tools renders them usable only for technical staff, making the business teams having to wait in the IT queue.

There are four key problems with ETL tools and processes:

- **Programmatic tools** — ETL tools are complex and programmatic which requires an IT-led project that takes weeks or even months
- **Complex modelling** — the fixed schema approach of ETL tools makes modelling slow and complicated, causing projects of weeks or months
- **Rigid modelling** — analysts must work within requirements of rigid schemas, limiting their ability to truly explore the data and boxing in the potential insights
- **Locked-in Logic** — logic is tied to transformations between specific end-points, limiting the ability to re-use that logic.

When these slow, complex ETL process get involved, the agility of cloud data warehouses is severely diminished. Business analysts wait weeks or months to get new datasets, limiting their ability deliver the analytic agility the business teams desire.

A Faster Modern Approach

Businesses today need a process to feed their cloud data warehouse that matches the agility, flexibility, and scalability of it. A new generation of data preparation tools on the market today make this level of performance possible. Businesses can finally capture the full advantages of cloud data warehouses.

A New Method

Two key items have changed in the analytics world in the last couple years – the cloud and agile methodologies. We explained the benefits of the cloud earlier and the speed benefits it brings. But agile development methodologies have also worked their way into the analytics world.

If the old world ETL approach had highly fixed requirements and length projects, the new world consists of speed, agility and exploration:

- **Speed** — very quick projects that rapid initial time to value
- **Agility** — the ability to meet the need of new business requirements faster
- **Exploration** — digging deeply into the data to find new, unknown insights

Figure 1 shows a model that a describes how a new generation of data preparation offerings support this new iterative and exploratory world. Analysts model the data, explore it, model it more, etc. in an iterative way until they FIND what they are looking for. This facilitates faster analytic cycles to dramatically reduce time to insight. Using this approach, Datameer customers have seen up to 98% reduction in their analytic cycles – from weeks and months to hours.

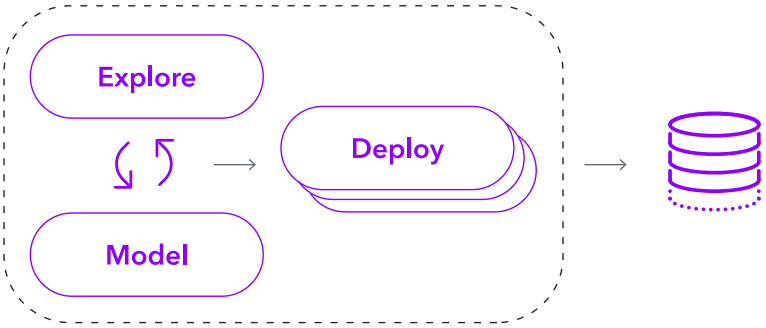

Figure 1: A New Iterative Approach to Feeding the Cloud Data Warehouse

Feature Requirements

To create an agile feed to a cloud data warehouse, there are several key features that data preparation tools should possess:

1. Self-Service
2. Rapid modeling
3. Easy operationalization
4. Deep governance and security
5. Integration with the cloud platform

Therefore, new tools that will be successful in creating an agile feed to a cloud data warehouse will have to fulfill these requirements. Let's look at each feature more closely.

1. Self-Service

Traditionally, business teams were required to submit data pipelining requests to IT, who would then collect, prepare, model, and finally feed the data into the data warehouse for analysis. The process could take weeks or even months depending on the size and complexity of the dataset. This slow data analytics cycle is damaging to an organization for two reasons:

- **Outdated data.** Markets and business environments can change drastically in the weeks or months it takes to feed data into the data warehouse using ETL. Business teams are unable to make real-time decisions and could very well be using outdated data entirely.

- **Limited Exploration.** The traditional data analytics process requires business teams to make specific requests for IT to build data pipelines. This forces analysts to work within the confines of what they already know, not giving them room to explore questions they never even thought to ask.

To make full use of cloud data warehouses, business teams need self-service data preparation tools that quickly deliver new data models for analysis. The self-service approach gives business teams the power to build their own data models with no-code styling tools and intuitive interfaces. Business analysts can transform, blend, and enrich complex datasets to feed their cloud data warehouses and generate insights more quickly.

With self-service tooling, organizations can shorten their analytics cycles from weeks or months to less than a day. Organizations can make faster decisions instead of reactionary ones based on old data. Analysts also have the freedom to explore the entire dataset. They no longer have to know what they're looking for ahead of time; they can explore far and wide for unique and valuable insights.

2. Rapid, schema-less modeling

One massive bottleneck in the ETL process was the need to transform data using fixed schemas. The “schema-on-write” approach is a key reason why data requests take weeks or months to fulfill. This task has gotten infinitely more difficult thanks to the volume, velocity, and variety of data collected today. Preparing data this way takes exhaustive resources and limits the functionality of the data.

Modern data pipelining tools don’t require data to fit a schema before being loaded into the data warehouse. This type of process is called schema-less, or “schema-on-read”, which means a virtual schema’s are automatically derived as the analyst defines their transformations, and a final schema is only applied when the data is delivered to its target destination – in this case, the cloud data warehouse.

Schema-less modeling tools make the end to end modeling process faster, but it also makes the data more useful. With the wide variety of data sources and formats, it is nearly impossible to apply one schema that optimizes each individual dataset. With schema-on-read, data can stay in its original format until its ready to be used. Then it can be transformed to fit any number of schemas depending on what the analysts want to do with it.

3. Easy Operationalization

ETL-centric data feeds have traditionally been difficult to operationalize, forcing organizations to rely on third party programmatic tools for automating and scheduling data movement. Adding multiple tools makes feeding your cloud data warehouse even more complex and time-consuming.

Cloud data warehouses require data preparation tools that make operationalization of data pipelines fast, easy and self-service. These tools have built-in automation and scheduling features so business teams can “set it and forget it.” No custom code is needed; analysts can point to a particular data pipeline and run analytics without the help of IT. This facilitates the iterative process analysts desire.

And operationalized jobs need the scale of big data. While traditional ETL tools in the cloud will run on standard EC2 instances, a modern data pipelining platform will leverage the power of cloud compute clusters such as AWS EMR to deliver high performance and scalable execution of the data pipelines that feed the cloud data warehouse.

4. Deep Governance and Security

Governance and security are of critical importance for organizations today. The regulation environment around the world is becoming increasingly difficult to navigate as requirements change regularly and with little warning. Businesses need a data “source of truth” to ensure transparency and high levels of security. This is a must-have feature for modern data tools that feed your cloud data warehouse.

Modern data pipelining tools monitor all data movement, transformation, and usage across the entire organization. Using metadata descriptions attached to each dataset, governance leaders can track the entire data lifecycle to ensure compliance. If regulatory bodies ask to see records of data usage, this information is quickly available and always up to date.

For security, essential features include encryption, masking, and fine-grained, role-based access control. Business analysts need the freedom explore datasets, but organizations must keep track of their actions.

5. Integration with the Cloud Platform

To deliver on many of the aforementioned features, a modern data preparation and exploration platform needs to tightly integrate with and leverage critical underlying services of the cloud platform. This makes the data preparation platform “cloud-native”, helps deliver a seamless experience, and provides a foundation for the scale, performance and security necessary.

For a platform like Datameer Spectrum, being cloud-native first means separating the compute and storage layers, which allows the two to be elastically scaled independently. It then involves leveraging compute, storage and security services, including:

- **Use of Cloud Object Storage** — Spectrum manages data in cloud object storage. providing persistent storage services and enables Spectrum to separate compute and storage services it uses, enabling these to be scaled independently.

- **Elastic Compute Clusters** — Spectrum leverages cloud-native compute clusters which can be elastically scaled allowing Datameer to take advantage of burst computing capacity.

- **Integration with Cloud Security** — Spectrum provides integration with cloud-native security facilities such as Identity and Access Management to ensure secure access and execution.

- **Integration with Key Management** — for encryption key management, Spectrum integrates with the cloud-native key management systems.

- **Management Console Integration** — Spectrum instances can all be managed and operated within the cloud native management consoles.

Beyond this, Spectrum also provides deep integration with the key cloud data warehouse destinations, including Redshift and Snowflake. In these integrations, Spectrum provides bi-directional data connections, and has optimizations to ensure the utmost performance when loading data.

Datameer in Action

Modern data preparation and analytics platforms like Datameer meet the above requirements for feeding cloud data warehouses today. Let's look at how one organization put these tools to work in their business.

Agile CDW Data Feeds in Retail

Facing a new stream of digital age buyers and new, specialty e-retail competition, this retailer needed to undergo digital transformation to grow their online business while maintaining balance and leadership with their brick-and-mortar and call center operations. This presented key challenges including:

- Faster time to insight to fuel data-driven business decisions
- Delivering wider self-service access to trusted data assets
- Using cloud economics to reduce analytic infrastructure time & cost

In their previous environment, projects to deliver new datasets to business teams were slow and costly – taking up to 6 months and costing up to $350,000. In some cases, business teams were not even allowed access to certain silo'ed data sources.

The retailer wanted to give self-service access to their business analysts so they could integrate and shape the data to their needs and load it into Redshift. This would alleviate the data engineering team, and let them focus on higher value projects to deliver more broad, yet governed access to the data. The company also needed to provide daily updated data feeds to certain business execs and teams.

The company decided to go with Datameer Spectrum for their data pipelining needs. Using Datameer's enterprise-grade pipelining platform, the data engineering team was able to set up self-service access to over 350 unique, curated datasets. If new complex data workflows need to be added, these projects are now delivered by the data engineering team in a matter of days, not months.

From there, business analysts have self-service access to the curated data, and create their own customized workflows specific to their analytic needs. With built-in data exploration at-scale, and spreadsheet-style tools, analysts could quickly and easily explore the data to find the parts of assets that meet their needs, create their own custom preparation workflows and deliver new insights to organization.

Today, the company runs 3 million jobs a year that pump data into their warehouse, providing the daily data feeds that drive faster analytics. Analytic cycles have been reduced by up to 98% and business analysts have the self-service access to the data they need. Through these new insights, the company has improved the efficiency of their e-commerce operations, expanded.

Getting Started

Where is the "slow point" for your agile data organization? Even if you have adopted cloud data warehouses, it still isn't enough if you're still using old-school ETL processes to move your data. Organizations need agility, flexibility, and scalability across their entire big data stack, and that includes their data pipelines.

Datameer provides hybrid-cloud analytics data management products that help organizations have successful, lower risk cloud journeys. With Datameer, business teams will deliver more valuable and trusted analytics, while IT teams will maintain a highly secure and governed data environment with lower costs.

Datameer Spectrum enables organizations to create secure, governed and scalable data pipelines and DataOps for easier, lower risk migration of analytics workloads to the cloud with reduced data engineering and cloud infrastructure costs. Datameer is a trusted platform at leading enterprises globally, including Citibank, Royal Bank of Canada, British Telecom, Aetna, Optum, National Instruments, Vivint and more. To learn more, please visit www.datameer.com.