Datameer

HOW TO DRIVE ANALYTICS ADOPTION VIA DATA TRANSFORMATION

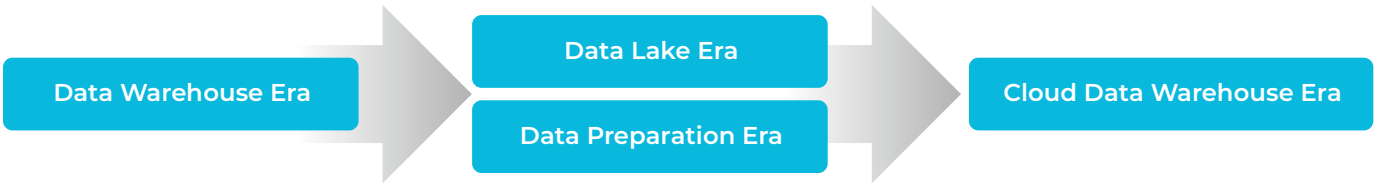
INTRODUCTION

Data transformation has become an integral component to every modern data stack, alongside your cloud data warehouse such as Snowflake. Data transformation also plays an indispensable role in the overall data lifecycle and analytics process, turning raw source data into valuable analytics data assets.

Over the years, data transformation's role, process, and tools have changed. In its earlier days, data transformation primarily focused on the "T" in ETL processes, cleansing data and mapping it from a source schema to a destination one. As organizations became more sophisticated in their destination schemas (star- and snowflake-schemas), these mapping processes became more complex.

As new data sources and formats emerged, and new platforms such as data lakes were deployed to support this, data transformation became increasingly important to help deliver value for this data and became a distinct process. New, complex data formats needed to be normalized, cleansed, integrated with traditional sources (often existing data warehouses), and then enriched. This process became "data engineering."

And finally, data preparation became an integral component for self-service data transformation by analysts. This allowed individual, less technical analysts to perform a wide variety of transformations on the data without relying on and waiting for IT and data teams to create data transformation pipelines for them.



THE MODERN DATA TRANSFORMATION WORKFLOW

The movement to cloud analytics and data warehouses has altered the overall data lifecycle and data transformation workflow. Many organizations with data already in the cloud, particularly in the SaaS applications, have switched their data pipeline process to ELT — extract, load, and transform — to streamline the process in three ways:

- Make it easier to extract data from SaaS applications and cloud services that have complex APIs using data loader tools for their "**EL**,"
- Use the economics, flexibility, and modern architecture of cloud data warehouses like Snowflake as the engine behind data transformation,
- Make the "**T**," or data transformation, component more self-service and switch from high-cost and time-consuming data engineering to what is now being referred to as "analytics engineering."

While self-service is essential, also critical is the need for data teams to be involved in the data transformation process. Why? Because well-intentioned but less data-savvy, analytics teams could possibly create and deploy less than optimal data structures in the cloud data warehouse, possibly creating performance issues or worse, errors or crashes.

A modern data transformation workflow embraces the fact that there are multiple personas involves in the analytics engineering process loosely grouped into two buckets: data engineers and data analysts/scientists. Each of these personas often has:

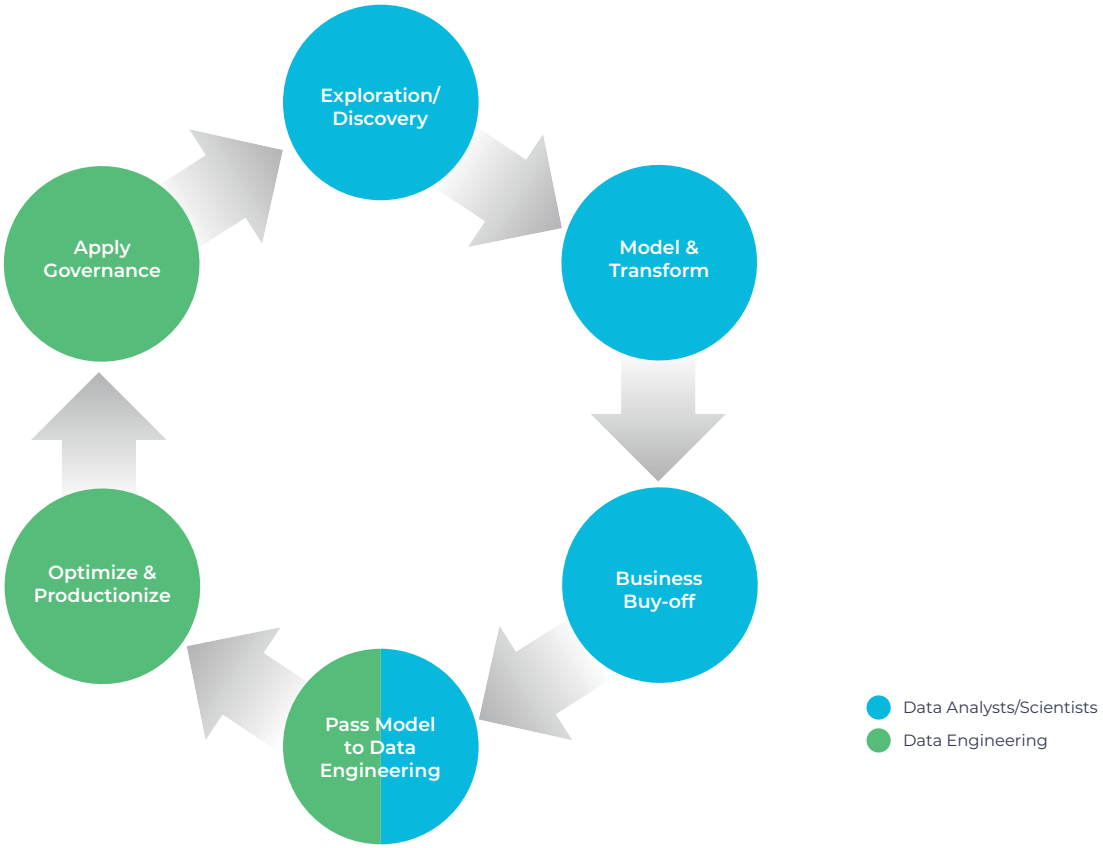
- Different skill sets ranging from being highly technical to less technical,
- Different knowledge and understanding of how data is used,
- Different focal points on their role and where they can best use their time.

Data engineers tend to know more about the data itself – where it resides, how it is structured and formatted, and how to get it – and less about how the business uses the data. Their role is highly focused on managing data.

Data analysts and scientists know less about the data itself but have a complete understanding of how the business would use the data and how it would be incorporated into analytics. They may have varying technical skills but would prefer to spend more time on what they are good at, analysis, and less on coding data transformation.

A modern analytics engineering process that includes data transformation would look like:

- Data analysts/scientists explore and discover the data assets available for their analytics to identify the ones best suited to the problem at hand,
- Data analysts/scientists iteratively model and transform the data assets they choose to answer the analytics question(s) best and explore the results,
- The business stakeholders buy off on the model and analytics results created by the data analysts/scientists,
- The data analysts/scientists pass over the final model to the data/analytics engineering team,
- The data/analytics engineering team optimizes, tests, secures and puts into production the data transformation models into the CDW,
- The data/analytics team puts governance around the new, usable, modeled asset(s) and makes it discoverable for other data analysts/scientists to use.



THE ROLE OF COLLABORATION

As you can see from the modern data transformation workflow above, multiple people, stakeholders, and personas are involved in the workflow. Collaboration is required to make the process extremely efficient and ensure the team is highly productive. Collaboration allows the broader, diverse team to:

- Share the workload and focus on their active pieces in the workflow,
- Efficiently handoff models or components between members at various phases
- Contribute and use their skills in the most effective manner
- Provide and share knowledge around the models
- Crowdsource tasks such as governance

Beyond making the process efficient and increasing team productivity, a collaborative data transformation workflow eliminates manual handoffs and misinterpretation of requirements. This adds one more highly valuable benefit: it eliminates errors in the data transformations and ensures models get done right the first time.

To support a collaborative data transformation workflow, Datameer DTaaS (Data Transformation as a Service) provides several essential collaboration facilities, including:

- A multi-persona toolset and interface
- Shared workspaces
- Data documentation
- Commenting
- Tagging and properties

Let’s explore each of these.

1. Multi-persona Tools and Interfaces

The different personas involved in the data transformation workflow will have varying skills, technical expertise, and business knowledge. We mentioned earlier how data engineers might know a lot about the data but little about how the business uses it and visa-versa for data analysts and scientists — they know more about how the business uses the data and less about the data itself.

The different personas may also have varying technical skills. Data engineers will be highly proficient in SQL and using data tools. Data analysts may have various degrees of SQL knowledge, but in general, may be less technical and proficient in SQL. A data scientist may be technical because they need to program data science models but might not be efficient in programming data transformations in SQL.

To this end, Datameer DTaaS supports three different interfaces to support these multiple personas:

- **No-code** — for non-technical data and business analysts. This interface is entirely graphical and drag-and-drop.
- **Low-code** — for slightly more data-savvy data analysts or for analytics engineers that want greater productivity than coding. This interface is Excel-like, with a wide range of functions and a formula builder wizard.
- **Code** — for data and analytics engineers that want to use SQL for the control and optimization that comes with it.

In addition, the different types of models (code, low-code, no-code) can be connected in a common data flow. This allows the different personas to contribute to the same workflow, using their best-suited modeling skills. Data engineers can create base models using SQL, then have analysts create their own tailored models using a no-code interface to graphically combine or reshape data or a low-code Excel-like interface where they can further refine the data or enrich it.

2. Shared Workspaces

Data transformation collaboration begins and ends with teams working together in a common workspace to create the best models for their analytics. As mentioned, data teams will know a lot about the data, while analysts and data scientists will know a good deal about how the business will use the data and how the final form needs to look. In shared workspaces, they can collaborate, share their knowledge, and share the modeling workload.

In Datameer DTaaS, shared workspaces are a foundational component. Data engineers, data analysts, business analysts, and data scientists can work together in shared workspaces creating and reusing models using the different tools – code, low-code, no-code. Beyond working together on modeling, the broader team can share knowledge about the workspace and models, including:

- Adding descriptions, which can both explain data and how best to use it,
- Applying tags, which can help organize and identify data,
- Supplying comments, which can add simple ideas around data or enable collaboration,
- Adding business metadata, which translates technical metadata into business terms
- Setting properties or status and certification fields, which describe the state of a data object

We will discuss these in more detail below. In Datameer DTaaS, there are three levels of access control for collaborators:

- **Level 1:** Anyone can view a workspace, its description, and apply comments
- **Level 2:** Collaborators, who can create and edit models within a workspace as well add information (descriptions, tags, comments, etc.) about the workspace and models.
- **Level 3:** Owners, who have full rights to a workspace, and full editing and information adding rights, and the additional ability to delete or change owners of a workspace.

3. Data Documentation

Information about data is often sparse or non-existent. The information is usually spread among wiki pages, metadata management systems, or early versions of data catalogs. Most of these sources still do not capture much of the knowledge there is about the data. Some data transformation tools attempt to generate documentation about data, but often this is just taking comments from SQL code and generating a wiki page or adding a limited description.

Datameer DTaaS facilitates capturing as much information as possible about the data it is working with, the transformations performed, and the resulting data models. This includes:

- Auto-generated documentation and information such as schema information, transformations performed, data lineage, audits, and certain system-generated properties.
- User-supplied documentation such as descriptions, tags, and business metadata.

4. Commenting

Comments are a core communication vehicle for the collaborators on a workspace. Datameer DTaaS supports social-media like comments on workspaces that can serve two purposes in collaboration:

- Adding further information about an asset that is simpler than detailed descriptions but slightly more complex than tags, and
- Allowing team members to communicate around tasks needed within a workspace, such as asking questions, performing tests, or applying certain types of modeling and formatting.

5. Tagging and Properties

Further enriching the documentation and another form of collaboration are various ways to add simple information about assets. Within Datameer DTaaS, this includes:

- **Tagging** — collaborators can add either standardized or user-defined tags to models or workspaces.
- **System-defined properties** — the system automatically adds several system level properties such as owner, create date, created by, last modified date, last modified by, etc.
- **User-maintained properties** — administrators can define a variety of custom properties (and their values) which collaborators can then set.
- **Status and certification fields** — these are predefined, specific purpose properties such as status (live or in-dev) or certified.

6. Search and Discovery

A final piece in the collaborative workflow story is search and discovery. Assets, and the information about them, are only valuable if the broader community of analysts and data scientists can find these assets. The quicker they can find them and identify their usefulness for their project, the faster the overall data transformation process.

Datameer DTaaS provides a rich faceted search capability, allowing users to easily search and drill down to potential assets they could use. The search indexes all of the available information discussed above and provides facets users can drill into, such as system and custom properties, tags, and more.

CONCLUSION

By its' very nature, data transformation workflows are collaborative. At a minimum, data teams (engineers), data analysts, and data scientists need to communicate relative to their needs, requirements, specifications, status, and final output. In most data transformation tools, this communication and documentation is performed outside in a scattered, inconsistent manner.

Datameer DTaaS creates a highly communicative and collaborative workflow across the entire team of data engineers, analysts, and data scientists. Through multi-persona tools, shared workspaces, rich data documentation, tagging, commenting, custom properties, business metadata, and search and discovery, Datameer DTaaS gets the entire team working together, communicating, and sharing the workload. This enables faster, more agile data transformation processes and eliminates errors due to misinterpretation to ensure data transformation models get done right the first time.

ABOUT DATAMEER

Datameer, a leader in data transformation solutions, offers the industry's first collaborative, multi-persona SaaS data transformation platform integrated into Snowflake. The multi-persona product brings together your entire team – data engineers, analytics engineers, analysts, and data scientists – on a single platform to collaboratively transform and model data for faster projects. Datameer is a trusted platform for leading enterprises globally, including Citibank, Royal Bank of Canada, British Telecom, Bank of Montreal, Aetna, Optum, Morningstar, Vivint, and more. To learn more, please visit www.datameer.com.

535 Mission St, Suite 2602
San Francisco, CA 94105 USA

www.datameer.com