Data engineering is the most popular and in-demand jobs among the big data domain across the world. Data Engineers build, monitor and refine complex data models to help organizations improve their business outcomes by harnessing data power.
In this post, we'll highlight the 20 most commonly used data engineering tools at mid-sized tech companies based on research from over 150 interviews with data engineers. We'll also briefly dive into some trends we noticed during our conversations, further exploring how different data engineering teams are thinking about their roles in the future. By the end of this post, we should've shed some light on the following questions:

What data tools are analytics teams *really* using?

How many teams are using Snowflake over BigQuery or Redshift?

What tools can data engineers not stop talking about?

Which tools are here to stay?

Snowflake

Snowflake's unique shared data architecture delivers the performance, scale, elasticity, and concurrency today's organizations require. Many teams we spoke to were interested in Snowflake and its capabilities to store and compute data, which makes us expect more teams to switch over to Snowflake in the coming years. In Snowflake, the data workloads scale independently from one another, making it an ideal platform for data warehousing, data lakes, data engineering, data science, and developing data applications.

dbt

dbt is a command-line tool that allows data engineers and analysts to transform data in their warehouse using SQL. dbt is the transformation layer of the stack and doesn't offer extraction or load operations. It allows companies to easily write transformations and orchestrate them more efficiently. The product is built by Fishtown Analytics and has raving reviews from data engineers.

Big Query

Similar to Amazon Redshift, BigQuery is a fully managed cloud data warehouse. It is commonly used in companies that are familiar with the Google Cloud Platform. Analysts and engineers can start using it when they are small and scale with the tool as their data grows. It also has built-in, powerful machine learning capabilities.

Tableau

Tableau is the second most commonly used BI tool from our survey. One of the oldest data visualization solutions, the main function is to gather and extract data that is stored in various places. Tableau uses a drag and drop interface to make use of data across different departments. The data engineer manager works with this data to create dashboards

Looker

Looker is BI software that helps employees visualize data. Looker is popular and commonly adopted across engineering teams. Unlike traditional BI tools, Looker has created a fantastic LookML layer. This layer is a language for describing dimensions, aggregates, calculations, and data relationships in a SQL database. One tool that has launched recently as a way to manage teams' LookML layer is spectacles, which allows teams to deploy their LookML layer with confidence. By updating and maintaining this layer, data engineers can make it easier for non-technical employees to use company data.

Apache Spark

Apache Spark is an open-source unified analytics engine for large-scale data processing. Apache Spark is a data processing framework that can quickly perform processing tasks on very large data sets, and can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools. These two qualities are key to the worlds of big data and machine learning, which require the marshalling of massive computing power to crunch through large data stores.

Airflow

Apache Airflow is an open-source workflow management platform. It started at Airbnb in October 2014 as a solution to manage the company's increasingly complex workflows. Creating Airflow allowed Airbnb to programmatically author and schedule their workflows and monitor them via the built-in Airflow user interface. It is the most commonly used workflow management solution and was used by around 25% of the data teams we interviewed.

Apache Hive

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. The three important functionalities for which Hive is deployed are data summarization, data analysis, and data query. The query language, exclusively supported by Hive, is HiveQL. This language translates SQL-like queries into MapReduce jobs for deploying them on Hadoop.

Segment

Segment makes it simple to collect and use data from the users of your digital properties. With Segment, you can collect, transform, send, and archive your customer data. The tool simplifies the process of collecting data and connecting it to new tools, which allows teams to spend less time processing and collecting data.
cloud warehouse is another industry staple that powers thousands of businesses. The tool allows anyone to easily set up your data warehouse and scales easily as you grow.

Key Considerations for Choosing Data Engineering Tools

When choosing data engineering tools, several key considerations can help ensure the right fit for your organization:

- Scalability: The tool should be able to handle increasing data volumes as your business grows, supporting both small datasets and massive, complex workloads.
- Ease of Integration: The ability to seamlessly integrate with your existing data infrastructure (e.g., databases, cloud platforms, and BI tools) is crucial to maintaining efficiency and avoiding disruptions.
- Automation Capabilities: Automated processes like data ingestion, transformation, and error handling reduce manual intervention and streamline workflows.
- Data Security and Compliance: Ensure the tool meets industry standards for data security and complies with relevant regulations (e.g., GDPR, HIPAA).
- Real-Time Processing: Depending on your needs, the tool should support both batch and real-time data processing for timely insights.
- Flexibility and Customization: Look for tools that allow for customization of workflows, pipelines, and data transformations to fit your specific use cases.
- Cost Efficiency: Evaluate whether the tool&rsquo;s pricing model aligns with your budget, especially considering cloud-based or pay-per-use options that can optimize costs.
- User-Friendliness: A tool with an intuitive interface that can be easily adopted by both technical and non-technical users can boost productivity and collaboration across teams.
- Support and Community: Strong customer support and an active user community can help troubleshoot issues and foster innovation within your team.
- Data Governance and Lineage: Tools that provide robust data governance, lineage tracking, and documentation features ensure that data is well-managed, trusted, and compliant.

Considering these factors can help you select the right tools that will not only meet current needs but also grow with your organization.

What are data teams most excited to use?

Data engineers almost unanimously agree that the most exciting tool they want to learn or use is dbt. The team at Fishtown analytics has done an amazing job of creating a community around analytics engineering. The tool is a command-line that allows data engineers to transform their warehouse using SQL. It has recently raised a significant funding round due to its simplification of workflows for data engineers.

Secondly, many people we talked to wanted to try, or were in the process of moving towards, Snowflake. The current users really enjoy the functionality of the tool and would recommend it to anyone looking for a data warehouse.

What’s next in Data Engineering?

It’s hard to speculate what’s next in data engineering. However, based on industry trends and current research, several key areas of development are emerging. After addressing foundational challenges—such as setting up data warehousing, building scalable ETL pipelines, and ensuring data quality—data engineering teams are increasingly focusing on higher-order tasks. These next steps revolve around defining, interpreting, and analyzing data to extract actionable insights, moving beyond the purely technical aspects of managing data.

This shift aligns with the "Analytics Hierarchy of Needs" framework, a concept developed by Ryan Foley. This model is akin to Maslow’s hierarchy of needs and illustrates the journey organizations take as they progress from basic data management to advanced analytics and predictive insights. Here's how data engineering teams are likely to advance along this hierarchy:

The Analytics Hierarchy of Needs. Image by Ryan Foley

1. Data Collection and Warehousing (Foundational Level)

At the base of the hierarchy, organizations are tasked with collecting, storing, and organizing data in a centralized warehouse. This step is essential for any subsequent data engineering work. Once data warehousing and storage solutions, such as Snowflake, BigQuery, or Redshift, are established, teams can ensure that data is easily accessible and scalable. However, many organizations stop here, focusing solely on collecting data without defining or interpreting it.

2. Data Transformation and Cleaning (Second Level)

Once data is collected, the next step involves cleaning, transforming, and organizing it into a usable format. This ensures that the data is ready for analysis and is of high quality. Tools like Apache Spark, dbt, and Talend are commonly used at this stage. With the rise of ELT (Extract, Load, Transform) models, transformation happens directly within data warehouses, enabling quicker insights.

3. Defining Data and Creating a Unified Data Model (Third Level)

One of the growing areas of focus in data engineering is defining and standardizing business metrics. After resolving ETL and data quality issues, organizations are realizing the need to align their teams with a common understanding of key metrics and terminology. This is where data dictionaries and data catalogs (like Secoda) come into play, as they offer a centralized space for managing metadata, ensuring data governance, and fostering collaboration across departments.

The challenge lies in defining data consistently across the organization. Different teams may have varying interpretations of key metrics, which leads to miscommunication and inconsistent analyses. As organizations mature, they strive to create a single source of truth where all departments adhere to the same definitions, driving clearer and more accurate insights.

4. Data Analysis and Insight Generation (Fourth Level)

After the data has been organized and defined, the next step is to perform analysis. This stage involves applying descriptive, diagnostic, and even prescriptive analytics to answer key business questions. The focus here is on turning raw data into insights—helping stakeholders understand not only what happened but also why it happened, and what can be done to improve future outcomes.

This is where collaboration between data engineers and data scientists becomes essential. Engineers work to ensure that data pipelines and infrastructure are optimized for quick, seamless access to clean and reliable data, while analysts and scientists focus on applying machine learning algorithms, statistical analysis, and predictive modeling techniques to generate insights.

5. Predictive and Prescriptive Analytics (Top Level)

At the top of the Analytics Hierarchy of Needs is the application of advanced analytics techniques to predict future outcomes and provide recommendations. This involves the use of sophisticated machine learning models, AI-driven insights, and real-time decision-making systems. As data engineering teams evolve, their role increasingly involves supporting data science initiatives by building infrastructure that allows for real-time or near-real-time predictive analytics.

For instance, industries like finance, healthcare, and retail are adopting machine learning to automate decision-making processes, such as fraud detection, supply chain optimization, and personalized customer recommendations. The future of data engineering lies in its ability to support AI and machine learning models with reliable data, ensuring that these advanced models are fed with high-quality inputs and scalable architectures.