Data warehousing on the lakehouse

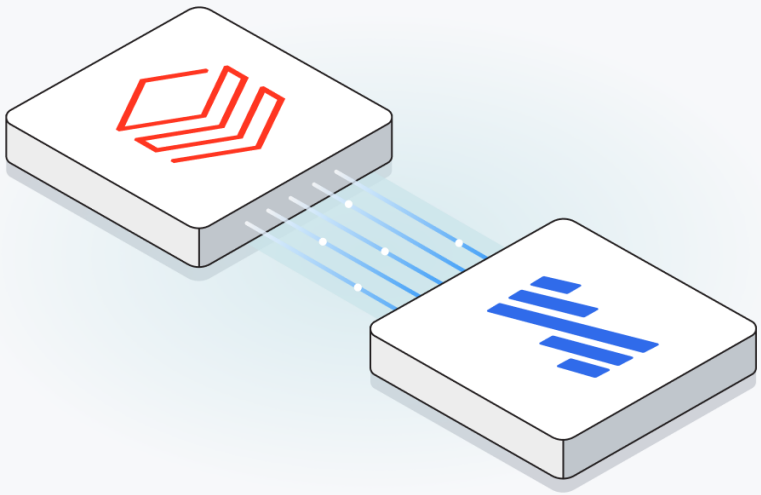
Databricks offers a multi-persona experience that allows every practitioner to work on the same data whether they choose SQL or Python. For traditional data warehousing workloads, analytics team find a familiar experience in **Databricks SQL**, a serverless data warehouse on the Databricks Lakehouse Platform that lets you run all your SQL and BI applications at scale with up to 12x better price/performance, a unified governance model, open formats and APIs, and your tools of choice - no lock-in. Databricks SQL is packed with thousands of optimizations to provide customers with the best performance for all query types and real-world applications. This includes the next-generation **Photon** query engine and serverless SQL warehouses for instant, elastic compute resources, which together provide up to 12x better price/performance than other cloud data warehouses.

Databricks SQL serverless removes the need to manage, configure or scale cloud infrastructure on the Lakehouse, freeing up your data team for what they do best. Databricks SQL warehouses provide instant, elastic SQL compute - decoupled from storage - and will automatically scale to provide unlimited concurrency without disruption, for high concurrency use cases. Databricks SQL's seamless integration with Databricks' Unity Catalog makes it easy to discover, audit and govern data assets in one place using standard SQL.

One of the decisions that is sometimes overlooked when building a cloud data platform is how to centralize your data in the lakehouse. Simply getting the data from point A to point B is just a fraction of the problem; there are also questions about how to model, normalize, and anonymize it, as well as issues around security and compliance. Fortunately, there's Fivetran, the automated data movement company. Fivetran provides pre-built connectors for over 300 data sources, including popular databases, data warehouses, cloud SaaS applications, and file storage systems. These connectors establish a secure and reliable connection between your data sources and Databricks, ensuring that your data is continuously synced and up-to-date.

You can get started with Fivetran through directly on Fivetran.com, or from within Databricks using Databricks Partner Connect. This simple, streamlined process makes it easy to go from zero to all of your data in the lakehouse in just a few minutes per connector. First, choose Fivetran from the Partner Connect screen, then you'll be asked to create a Fivetran account. From there, you can choose your data source(s) of choice from over 300 connectors. Then, Fivetran automatically fills in the credentials needed to connect to your existing SQL endpoint, then uses it as a source of efficient compute for ingesting your data from its source to your Databricks destination. As part of the ingestion process, Fivetran automatically normalizes your data, anonymizes PII, and converts it to Delta Lake format, all in a single step – making it instantly queryable from within Databricks SQL. And finally, Fivetran handles ongoing change data capture (CDC) using your Databricks SQL cluster to pull in and integrate any changes (at the interval of your choosing) so that your data is in sync at all times.

By combining Fivetran's automated data integration capabilities with Databricks' powerful analytics and data processing features, you can go from data ingestion to insights faster than ever before.



CHAPTER 3

Security

Every enterprise is security conscious for the following reasons:

- Regulatory compliance
- Managing brand risk
- System availability risk
- Basic ethics
- Protecting customers from identity theft and breaches of privacy
- Protecting internal operations, sensitive data and trade secrets

From a technological standpoint, Fivetran addresses security through:

1. Data security, regarding the protection of customer data processed by our services
2. Platform security, regarding tools

Data security

Data security is fundamentally about protecting sensitive data such as personally identifiable information (PII). In the context of data movement, this means applying data security in the pipeline before data is loaded to the customer's destination. Fivetran's primary method of ensuring appropriate access is through column masking. Column masking takes two main forms:

1. **Blocking** data by excluding it from entering the destination altogether. Note that primary keys, due to their importance for idempotence and deduplication, cannot be blocked. Blocked data may still pass through our systems and be stored temporarily but will not be accessible through either the user interface or the destination.
2. **Hashing** data by anonymizing and obscuring it while preserving its analytical value. You will still be able to use hashed columns as keys, joining records across data sets but will not be able to read the original value. Unlike encryption, hashing is one-way and not intended to be reversible. Moreover, Fivetran uses a unique salt for every destination so that general knowledge of the hashing algorithm isn't enough to decode hashed values.

Data warehousing on the lakehouse

Allow all	Allow columns	Block all
New schemas, tables and columns will all be synced.	New schemas and tables are excluded but new columns for existing tables are synced.	New schemas, tables and columns are all excluded by default. The user can choose to be alerted whenever new objects are detected at the source.

Data security is fundamentally about protecting sensitive data such as personally identifiable information (PII). In the context of data movement, this means applying data security in the pipeline before data is loaded to the customer's destination. Fivetran's

Platform security

In addition to security at the level of individual data fields, Fivetran also supports many security features built into the architecture of the platform.

A short list of these features includes:

- SAML and SCIM
- Logging and log forwarding
- Access auditing

We will discuss other features in greater detail shortly.

COMPLIANCE CERTIFICATIONS

Fivetran is compliant with a number of common security standards across different industries and jurisdictions:

SOC 1 Type 2

Fivetran undergoes an annual, independent SOC 1 Type 2 audit. This standard allows customers to process data in our platform that is material for financial reporting.

SOC 2 Type 2

Fivetran undergoes an annual, independent SOC 2 Type 2 audit. This standard demonstrates common security, availability and confidentiality controls are in place within our platform.

ISO 27001

These standards require a vendor to:

- Systematically account for information security risks
- Design and implement information security controls and contingencies
- Maintain plans to ensure continued and ongoing compliance

PCI DSS Level 1

This is the most stringent of the Payment Card Industry Data Security Standards, and mandatory for merchants that process at least 6 million credit card/financial data transactions a year, such as retailers.

HIPAA BAA

Although Fivetran is not a healthcare provider or other HIPAA-covered entity, Fivetran complies with the stipulated standards for protected health information (PHI) and will sign a business associate agreement (BAA) with customers who handle healthcare data.

CCPA

Similar to but more expansive than GDPR, CCPA is a California standard that encompasses household as well as personal data.

GDPR

Fivetran compliance with GDPR is mainly enabled by column masking. GDPR is an EU-wide privacy rule positing that end users have the following basic rights regarding personal data:

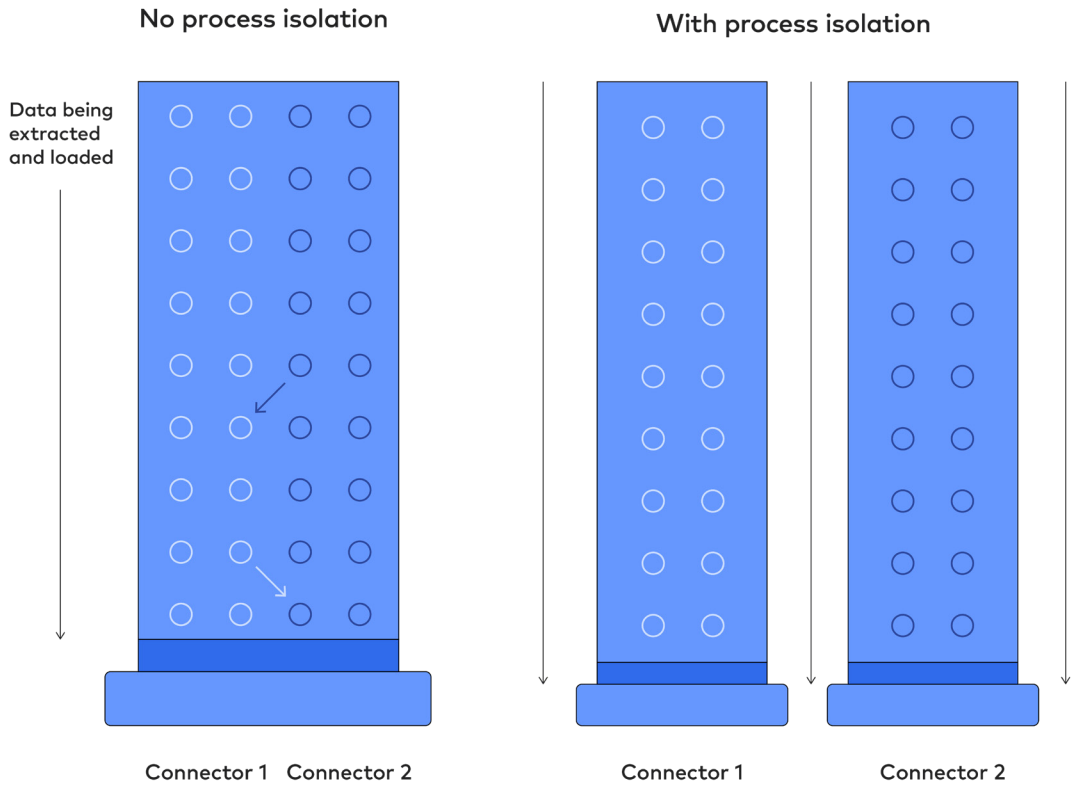
- The right to access
- The right to be informed
- The right to data portability
- The right to be forgotten
- The right to object
- The right to restrict processing
- The right to be notified
- The right to rectification

At least annually, Fivetran engages qualified third parties to perform penetration testing.

SINGLE TENANCY AND PROCESS ISOLATION

Fivetran is architected as an isolated, single-tenant temporary process that expires the moment the sync ends. This prevents persistent storage and cross-contamination of sensitive data.

Single tenancy is related to the concept of process isolation, which we also practice. Single tenancy means that instances of the Fivetran application are never shared between customers. Process isolation means they aren't even shared between different connectors belonging to the same customer, ensuring that there is no accidental cross-contamination of data.



END-TO-END ENCRYPTION

Fivetran encrypts data in transit and credentials using the following methods:

- Credentials are encrypted through a key management service and optionally dual-encrypted using a customer master key, which the customer can withdraw access to at any time.
- All communication between the customer's cloud and the Fivetran cloud is conducted through PrivateLink, VPN or SSH.

- Data in transit is encrypted and decrypted using ephemeral keys, meaning that a unique key is generated for each execution.

The entire architecture depends on the principle of "least privilege" – that is, only the minimum necessary permission is granted for any task. All customer data is purged from Fivetran after it is synced to the destination. The only exceptions are the following, which are required to maintain the continued functioning of connectors:

Customer access keys

Fivetran must access destinations in order to extract and load data to them

Customer metadata

Fivetran stores configuration and other account details in order to display them to customers

Data from email and event stream connectors

Since these sources don't persistently store data, Fivetran does so in case future re-syncs are ever needed

As you may have noticed in the data movement architectural diagram, there are also layers of strict separation between anything the user directly interacts with and the pipeline itself, meaning that, by design, data is never erroneously exposed through the user interface.

DATA ACCESS

Normally, customers enter credentials to connect their Fivetran instance with a database. We retain these credentials in order to continuously extract data and troubleshoot customer issues. These credentials are securely retained in a key management system. Customers control whether and when Fivetran can access their data.

We can optionally further encrypt credentials using customer master keys, which Fivetran does not retain but must consult to use. At any time, customers may revoke Fivetran access to the customer master key, effectively making it an immediate kill switch.

Besides ensuring that third parties cannot access sensitive data, there is the matter of ensuring appropriate permissions and access within your organization. To this end, Fivetran offers role-based access control (RBAC). This enables fine-grained control over:

- Onboarding
- Access
- Auditing
- Monitoring traffic

You can create roles with specific, granular kinds of authority and access and assign users accordingly. These roles can also be combined with SAML/SSO through tools such as Okta, Azure Single Sign-on, etc. We will discuss this topic further in the data governance section.

DEPLOYMENT METHODS

The Fivetran data pipeline can run on a number of different clouds, including AWS, Azure and GCP, with a choice of over 20 major cloud regions worldwide, across North America, Europe, Asia and the Pacific. Customers may have a number of reasons related to specific industries, jurisdictions and use cases for choosing specific regions for residency. We can support geographically bounded US-only access, as well.

Fivetran also offers private networking through services such as PrivateLink.

For those who have highly sensitive and confidential data, Fivetran can also be privately deployed on-premises for database replication.

Learn more about **Fivetran database replication** options.

CHAPTER 4

Governance

Data governance fundamentally consists of three needs, each of which grows in difficulty and complexity as an organization grows its headcount and the volume, variety and velocity of its data:

1

Knowing data

Keeping a full inventory of all relevant data assets

2

Accessing data

Ensuring that the appropriate parties within an organization have access to data

3

Protecting data

Minimizing misuse or unwanted exposure of data, including by parties within an organization

KNOWING DATA

Security and legal teams, analysts and data teams – have an interest in knowing data. Security and legal teams need to audit incoming data and monitor access. Data teams need to ensure that they are meeting their obligations to analysts and other stakeholders, particularly impact analysis of how upstream data pipeline changes might affect data models downstream. Analysts need to understand the provenance of their data and what questions can and can't be answered.

Fivetran offers several features to bring full visibility into the data pipeline for data audits and use monitoring:

Column-level data lineage exposed through graphs, as previously discussed in the context of transformations

Real-time metadata capture and logging of

- Keys
- Tables
- Columns
- Data types

End-to-end audit trails logging all access, behaviors and changes to the pipeline

A metadata API enabling programmatic management of data movement

Although Fivetran does not feature a data catalog, we can expose all the relevant metadata so it can be imported to your dedicated governance tool.

ACCESSING DATA

Analysts depend entirely on access in order to perform their roles. Data teams are the main gatekeepers to analysts, managing approval workflows for interested parties.

The Fivetran solution consists of automated user provisioning, namely integration with SCIM providers such as Okta and Azure AD to quickly onboard new users. This obviates the need to manually create accounts and configure permissions, which can otherwise be prone to human error and delay.

PROTECTING DATA

Security and legal teams are mainly concerned with protecting data, and there is considerable overlap between data governance and security. Similarly, data teams directly manage access control and handle sensitive information.

To this end, Fivetran features the following, many of which are outlined in "Chapter 2 – Security":

Compliance with laws and regulations across many jurisdictions

- SOC 1 Type 2
- SOC 2 Type 2
- ISO 27001
- PCI DSS Level 1
- HIPAA BAA
- GDPR
- CCPA

Role-based access controls (RBAC) to determine who can make the following actions:

- Move data
- Transform data
- Control where data is loaded

Blocking and hashing

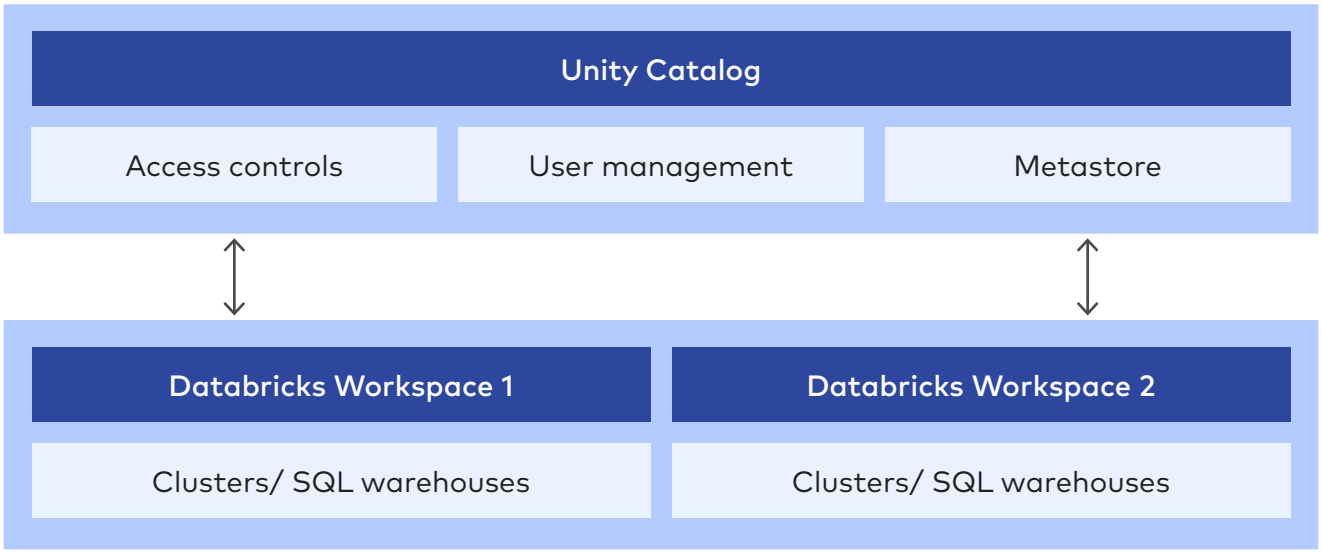
Automated, centralized user provisioning with granular permissions based on roles

Automatic tagging and categorizing of data, including PII

End-to-end encryption

GOVERNANCE ON THE LAKEHOUSE

To simplify data governance, the Databricks Lakehouse Platform offers **Unity Catalog**, a unified governance solution for data, analytics and AI on your lakehouse. Unity Catalog offers a centralized metadata layer, called a metastore, that provides the ability to catalog and share data assets across your lakehouse, and across your enterprise's regions and clouds. These data assets could be your files, tables, dashboards, ML models etc. Unity Catalog also centralizes the concept of identity, including service principals, users and groups to provide a consistent view across multiple workspaces. This allows you to use groups when defining access controls in role-based access control (RBAC) models.



A centralized metadata layer ensures that your centralized governance teams have the ability to dictate access controls and retrieve audit information from a single place, greatly reducing organizational risk due to improperly applied access controls. Because Unity Catalog metastores are aligned to cloud provider regions, they implicitly enforce data access along regional boundaries and ensure that your users are not accessing data from other regions unless you explicitly allow it to happen. By minimizing the copies of your data and moving to a single data processing layer where all your data governance controls can run together, you improve your chances of staying in compliance and detecting a data breach.

FIVETRAN SUPPORTS UNITY CATALOG

Fivetran amplifies the lineage of Unity Catalog by delivering normalized metadata directly to your Unity Catalog environment.

Users can specify a catalog name, a schema name, and a table name, taking advantage of the 3-part naming convention in Databricks Unity Catalog. This helps tie source data ingested via Fivetran directly to end-users and greatly simplify access management across enterprise-wide data assets.

CHAPTER 4

Extensibility

We can approach extensibility in two ways:

1

Interoperability is essential for all tools and technologies that manipulate data. As an organization's data needs grow in scale and complexity over time, it will need tools to:

- Perform administrative tasks at scale
- Integrate with other tools and technologies in the data operations ecosystem
- Build a sustainable foundation for future expansion of analytics and data-driven products

2

Data operations are foundational to other tools and technologies, including many that are as-yet undetermined.

At Fivetran, we often liken data to electricity – an enabling technology with unlimited potential for innovative uses.

Programmatic control and automation are central to addressing both understandings of extensibility.

Tools

Fivetran offers programmatic control and automation in the guise of a REST API, Connect Cards and data models.

REST API

The Fivetran REST API enables programmatic control over all aspects of the Fivetran application. Key tools include:

Certificate management

Approve Transport Layer Security (TLS) certificates.

Connector management

Create, edit, remove, run and list connectors and their schema configuration files. This is also the tool for generating Connect Card URIs – more on that later!

dbt transformation management

Create, edit, remove and list transformations within a specified group.

Destination management

Create and edit destinations within the group.

Team management

List, edit and delete your teams, manage team memberships and permissions. Teams can be connected with users, connectors and groups.

Group management

Monitor groups, which are collections of destinations and the users and connectors associated with them. Groups are created first, followed by destinations and connectors.

User management

List, invite, edit and delete users.

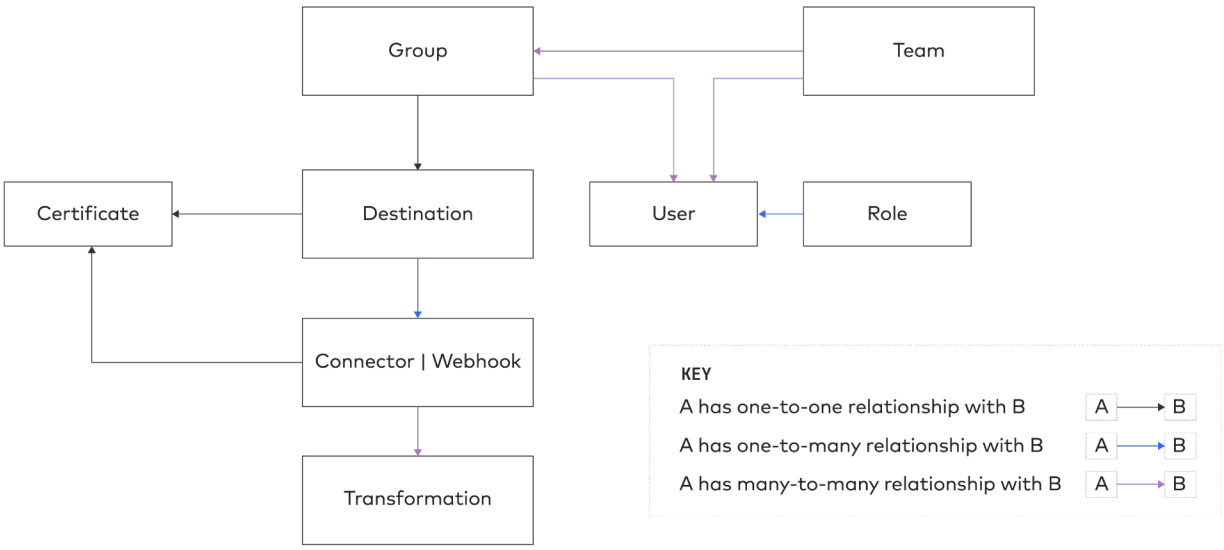
Role management

List all roles, whether pre-defined or custom. Roles are categories of users with specific kinds of permissions and access, which can be highly customized.

Webhook management

List, create, edit, remove and test webhooks, a specific kind of data source that requires a slightly different architecture (updates are pushed rather than pulled).

The relationships between all the elements of the Fivetran application are illustrated below:



CONNECT CARDS AND POWERED BY FIVETRAN

The connector management API is used to enable Connect Cards, which are embeddable pop-up windows that enable outside parties to set up Fivetran connectors. Connect Cards are essential for companies who operate connectors and destinations on behalf of customers.

Connect Cards and the REST API are the central elements of Powered by Fivetran (PBF), an implementation of Fivetran designed for embedding into web applications and analytics portals. The general principle is similar to third-party personal finance applications like Mint or Paypal that require connections to bank or credit card accounts. Connect cards are used for "embedded" onboarding to PBF.

The alternative is to build a "headless" onboarding experience using a custom-built authentication UI. PBF Headless requires more upfront development work but can potentially offer users a more seamless experience, enabling an organization to directly embed Fivetran into a web application.

DATA MODELS

As previously discussed in Chapter 1 – Data movement, Fivetran data models enable rapid bootstrapping of dashboards, reports and other analytical data assets. They enable quick turnaround, use and customization of data models for all uses, up to and including business process automation and predictive modeling.

EXTERNAL LOGGING

Fivetran records and reports a number of behaviors from connectors, data models and accounts. These include errors, failures and delays of all kinds, as well as new users and monthly spend. Although Fivetran features a log connector, Fivetran is not a logging service and only retains log data for a week. Instead, we offer integration with a number of leading logging services, namely AWS CloudWatch, Azure Log Analytics, Datadog, Splunk and Google Cloud Logging.

The simplest use for external logging is to monitor immediate problems. Savvier and more advanced use cases include mining the data for predictive and proactive monitoring, which can then inform automated responses. For the sake of legal and regulatory compliance, logs are also critical for auditing and observability.

Use cases

1 Sales and customer success analytics

Easily combine, report and analyze data from SaaS data sources, such as Salesforce, Zendesk/Freshdesk, CRM, helpdesk, and issue trackers for sales pipeline analysis, revenue forecasting and customer health analysis on the lakehouse.

2 Marketing analytics

Integrate data from across the marketing stack, such as Marketo, Google ads, Google analytics, Facebook/Instagram ads, Optimizely and Hubspot to optimize spend and ROI, improve targeted customer acquisition marketing and customer journey analysis to achieve predictable business outcomes.

3 Finance analytics

Enable finance teams to create insights using data from hundreds of apps, such as NetSuite, Stripe, Zuora, Anaplan and Google Play for integrated financial planning and budgeting, forecasting and modeling, and more.

PRACTICAL EXAMPLES OF EXTENSIBILITY

Programmatic control over data movement offers a multitude of possibilities for novel products and business models. Some products our customers have built using Fivetran include:

- Business valuation enabled by aggregating data across a huge number of ecommerce businesses
- On-the-fly decision support for ecommerce retailers based on inventory, customer, product and marketing data
- Consolidated business intelligence platform featuring data from every major advertising platform
- Ad recommendation engine for stock photography customers
- Aggregating WiFi usage data to help small businesses optimize service and boost margins
- White glove data integration services, encompassing infrastructure to analytics
- Revenue forecasting built on a model that includes industry-wide data
- Aggregating market-wide performance metrics to help companies evaluate their performance against industry norms
- Personalized learning to help onboard new employees

These products demonstrate the power of data aggregation, as well as the understandable reluctance of many businesses to directly engage in data integration and analytics.

Tools like Fivetran solve basic capabilities for data movement. As more organizations build a firm foundation of data operations and infrastructure, the future will likely see the continued growth of decision support, business process automation, predictive modeling, autonomous agents and more. The most impactful uses for data have yet to be invented or brought to market.

Fivetran

Fivetran automates data movement out of, into and across cloud data platforms. We automate the most time-consuming parts of the ELT process from extracts to schema drift handling to transformations, so data engineers can focus on higher-impact projects with total pipeline peace of mind. With 99.9% uptime and self-healing pipelines, Fivetran enables hundreds of leading brands across the globe, including Autodesk, Conagra Brands, JetBlue, Lionsgate, Morgan Stanley, and Ziff Davis, to accelerate data-driven decisions and drive business growth. Fivetran is headquartered in Oakland, California, with offices around the world.

For more info, visit Fivetran.com.

Start your free trial