Enterprise Data Team

The GitLab Enterprise Data Team is responsible for empowering every GitLab team member to contribute to the data program and generate business value from our data assets.

Welcome to the Enterprise Data Team Handbook

- Our Vision is to **Contribute to GitLab's journey of becoming the leading AllOps platform by responsibly harnessing the power of data.**
- In pursuit of our vision, we will focus on 4 outcomes:

1. **Drive** company results by building trusted, reliable, and innovative data products and insights when and where needed.
2. **Minimize** time from question to insight to action, enabling team members to move faster by implementing efficient processes and enabling self-service analytics.
3. **Develop** and secure our data into a uniform, trusted asset through data protection & privacy, iterating on processes, people, and platforms.
4. **Enable** every team member to contribute to initiatives responsibly and with trust, building a powerful data-driven culture.

- Read our Direction page to learn *what* we are doing to improve data at GitLab.
- Our Principles inform how we accomplish our mission.
- Watch our Data Recruiting Video to learn about the growing Data Program.

Would you like to contribute? Recommend an improvement, visit Slack #data, watch a Data Team video. We want to hear from you!

How Data Works at GitLab

The collective set of people, projects, and initiatives focused on advancing the state of data at GitLab is called the **GitLab Data Program**. GitLab has two primary distinct groups within the Data Program who use data to drive insights and business decisions. These groups are complementary to one another and are focused on specific areas to drive a deeper understanding of trends in the business. The two teams are the (central) Enterprise

Data Team and, separately, Function Analytics Teams located in Sales, Marketing, Product, Engineering or Finance. Watch the Data Recruiting Video to hear from some of the teams involved and what they are working on.

- The **Data Team** reports into Business Technology and is the Center of Excellence for enterprise insights & analytics (not operational), data science, data platform & infrastructure, BI technologies, master data, data governance and data quality. The Data Team is also responsible for the enterprise data strategy, building enterprise-wide data models, providing Self-Service Data capabilities, maintaining the data platform, developing Data Pumps, and monitoring and measuring Data Quality. The Data Team is responsible for data that is defined and accessed on a regular basis by GitLab team members from the Snowflake Enterprise Data Warehouse. The Data Team builds data infrastructure to power approximately 80% of the data that is accessed on a regular basis. The Data Team also provides a Data Science center of excellence to launch new advanced analytics initiatives and provide guidance to other GitLab team members.

- **Function Analytics Teams** reside and report into their respective divisions and departments. These teams perform specific analysis for business activities and workflows that take place within the function. These teams perform ad-hoc analysis and develop dashboards based on the urgency and importance of the analysis required, following the Data Development approach. The most important and repeatable analysis will be powered by the centralized Trusted Data Model managed by the central Data Team. Function Analytics Teams also build function-specific/ad-hoc data models and business insights models to solve for urgent and operational needs, not requiring trusted data features. Function Analytics Teams work closely with the Data Team in a variety of ways: expand GitLab's overall analytics capabilities, extend the Data Catalog, provide requirements for new Trusted Data models and dashboards, validate metrics, and help drive prioritization of work asked of the Data Team. When data gaps are found in our business processes and source systems, the team members will provide requirements to product management, sales ops, marketing ops, and others to ensure the source systems capture correct data.

Data Program Teams

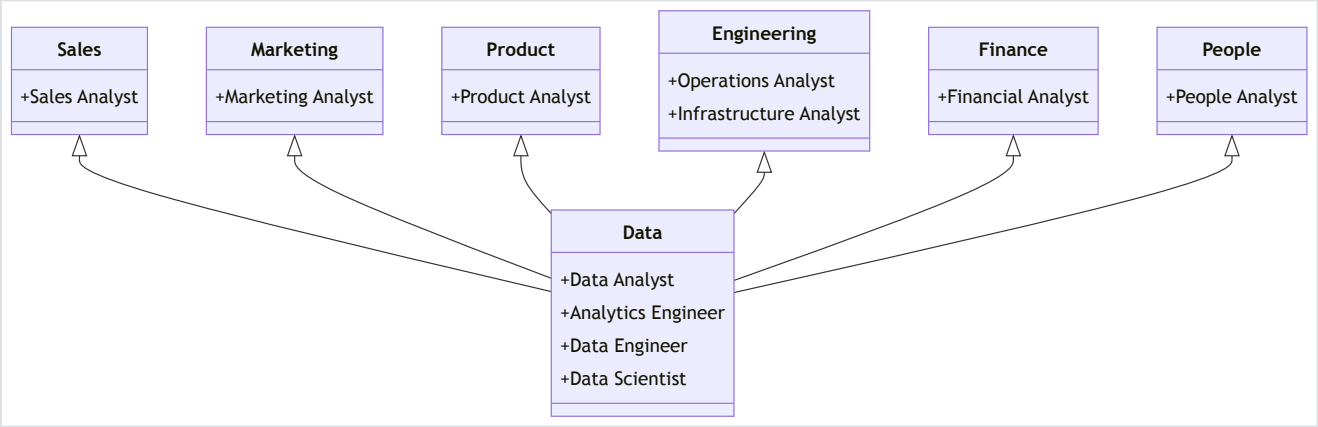
The GitLab Data Program includes teams focused in the following areas:

- Customer Success Operational Data Team
- Enterprise Data Team
- Engineering Analytics
- Finance Analytics & Insights
- Marketing Strategy and Performance
- Marketing Web Analytics

- People Analytics Team
- Product Data Insights
- Analytics Instrumentation Group
- Sales Analytics

How Data Teams Work Together

On a normal operational basis, the Data Team and Function Analyst teams work in a "Hub & Spoke" model, with the Data Team serving as the "Hub" and Center of Excellence for analytics, analytics technology, operations, and infrastructure, while the "Spokes" represent each Division or Departments Function analysts. Function analysts develop deep subject matter expertise in their specific area and leverage the Data Team when needed. From time to time, the Data Team provides limited development support for GitLab Departments that do not yet have dedicated Function Analysts or those teams which do have dedicated Function Analysts, but might need additional support. The teams collaborate through Slack Data Channels, the GitLab Data Project, and ad-hoc meetings.



The Data Platform & Architecture Team

The **Data Platform Team & Architecture Team** is part of the Enterprise Data Team and focuses on building and maintaing secure, efficent, and reliable data systems data infrastructure. The Data Platform & Architecture Team is both a development team and an operations/site reliability team. The team supports all Data Pods with **available, reliable, and scalable** data compute, processing, and storage. Platform components include the Data Warehouse, New Data Sources, Data Pumps, Data Security, and related new data technology. The Data Platform team also drives the Data Management processes. The Data Platform Team is composed of Data Engineers.

Analytics Engineering Team

The Analytics Engineering Team** transforms raw data into clean, structure and usable formats for data decision-making. The Analytics Engineering team also drives Enterprise Data Program and supports the wider data community. The team focuses on inventorying,

integrating, maintaining, and governing the data at an Enterprise level. This includes collaborating with the business units and data teams in establishing and facilitating commonly accepted guidelines around Enterprise data along with building enterprise-wide data models, supporting Self-Service BI and Analytical capabilities by providing Data Enablement and required training to the Users on Enterprise Data Models.

The Enterprise Insights & Data Science Team

The Enterprise Insights & Data Science Team utilize analytics and Machine Learning (ML) for insights into customer behavior and company performance. The Enterprise Insights & Data Science team focuses on delivering a complete view of the customer (Customer 360), predict customers that are likely to buy, expand or churn, develop models to predit the long-term value of customers, create detailed customer profiles, and deliver insights on company performance. The Team acts as a Center of Excellence for predictive analytics and supports other teams in their data science endeavours by developing tooling, processes, and best practices for data science and machine learning. List of the current projects can be found in the Data Science handbook page.

Data Job Families

The job families are designed to support all of the routine activities expected of a Data Team. In FY22 we are introducing two new job families, Data Scientist and Analytics Engineer.

- Data Analyst
- Data Scientist
- Analytics Engineer
- Data Engineer
- Manager, Data
- Director, Data

How We Measure Impact

Our impact will be measured against 4 dimensions (these metrics will adjust as our data maturity increases and our focus areas change):

Data Platform Stability

- Infrastructure Cost vs Plan: This performance indicator tracks the financial position of the actual cost vs the planned costs for the data infrastructure (warehouse, ETL pipelines, etc.).

- Data Uptime: This performance indicator measures the % of time a data pipeline was providing data without reported incidents. This indicator is currently measured based on Monte-Carlo data, according to the configured (automatic) monitors on any given table in the `raw` data layer.

Data Quality & Governance

- % completion of the data validation and data cleansing roadmap
- Governance & Quality data assessment scores

Data Adoption

- Data Monthly Active Users (DMAU): DMAU Measures the direct usage of the Data Platform by GitLab Team Members based on usage of the primary analysis tools we provide: Snowflake and Tableau. Over time we will include additional tools such as Jupyter and Data Studio, as well as usage of data pumped into EApps such as Marketo (PQLs), Gainsight (Usage Data), and Salesforce (Propensity Scores). A visualization of these numbers can be found in the Data Monthly Active Users report.
- Data Monthly Active Users (DMAU) = Unique users of a Data system (i.e. Snowflake, Tableau) in a given month
- Data Maturity Score: measured annually, evaluates our current data maturity against 8 data capabilities:
    1. Strategy & Approach
    2. Culture & leadership
    3. Metrics & KPIs
    4. Organization & Skills
    5. Architecture & Integration
    6. Governance & Quality
    7. Deployment & Usage
    8. Technology & Operations
- Number of certified Tableau dashboards
- % total views from certified dashboards

Revenue/Efficiency Impact

First we have the evaluation criteria known as Dollar Value of our Results as calculated by the Data Value Calculator. We can use the Data Team Value Calculator to calculate the dollar value of the initiatives we contribute to and the issues we complete. Additionally we want to shift to a more aspirational measurement which is to measure the ARR impact or

efficiency gain from each of our data products. Our data science models will be measured in the following ways:

- Propensity to Expand (PtE) and Purchase (PtP) - We will evaluate two metrics: 1) Incremental revenue impact 2) # of leads generated that are not currently in the sales funnel
- Propensity to Churn (PtC) - We will evaluate two metrics: 1) # of high propensity to churn customers that didn't churn 2) Incremental revenue impact

How To Connect With Us

Primary #Data Slack Channel
Issue tracker
GitLab Unfiltered Data Team Playlist
What time is it for folks on the data team?

Data Slack Channels

- #data is the primary channel for all of GitLab's data and analysis conversations. This is where folks from other teams can link to their issues, ask for help, direction, and get general feedback from members of the Data Team.
- #data-daily is where the Data Team tracks day-to-day productivity, blockers, and fun. Powered by Geekbot, it's our asynchronous version of a daily stand-up, and helps keep everyone on the Data Team aligned and informed.
- #data-lounge is for links to interesting articles, podcasts, blog posts, etc. A good space for casual data conversations that don't necessarily relate to GitLab. Also used for intrateam discussion for the Data Team.
- #data-engineering is where the GitLab Data Platform team collaborates.
- #bt-data-science is where the GitLab Data Science team collaborates.
- #business-technology is where the Data Team coordinates with Business Technology in order to support scaling, and where all Business Technology-related conversations occur.
- #analytics-pipelines is where slack logs for dbt runs and monte carlo analysis are output and is for analytics engineers to maintain. The DRI for tracking and triaging issues from this channel is shown here.

- #data-triage is an activity feed of opened and closed issues and MR in the data team project.
- #data-pipelines is where alerts from the ELT pipelines / FiveTran/ Monte Carlo RAW layer anomalies published and is for data engineers to maintain. The DRI for tracking and triaging issues from this channel is shown here.

You can also tag subsets of the Data Team using:

- @datateam - this notifies the entire Data Team
- @data-engineers - this notifies just the Data Engineers
- @data-analysts - this notifies just the Data Analysts
- @analytics-engineers - this notifies just the Analytics Engineers

Except for rare cases, conversations with folks from other teams should take place in #data, and possibly the fusion team channels when appropriate. Posts to other channels that go against this guidance should be responded to with a redirection to the #data channel, and a link to this handbook section to make it clear what the different channels are for.

GitLab Groups and Projects

The Data Team primarily uses these groups and projects on GitLab:

- GitLab Data is the main group for the GitLab Data Team.
- GitLab Data Team is the primary project for the GitLab Data Team.

Though many of our GitLab projects are internal only, the rest are still public by default.

You can tag the Data Team in GitLab using:

- @gitlab-data - this notifies the entire Data Team
- @gitlab-data/engineers - this notifies just the Data Engineers
- @gitlab-data/analysts - this notifies just the Data Analysts

Team, Operations, and Technical Guides

TECH GUIDES	INFRASTRUCTURE	DATA TEAM
SQL Style Guide	High Level Diagram	How We Work
dbt Guide	System Data Flows	Team Organization
Python Guide	Data Sources	Calendar
Airflow & Kubernetes	Snowplow	Triage

TECH GUIDES	INFRASTRUCTURE	DATA TEAM
Docker	Permifrost	Merge Requests
Data CI Jobs	DataSiren	Planning Drumbeat
Rstudio Guide	Trusted Data	Data Science Team
Jupyter Guide		Data Management
Meltano Guide		
Experimentation Best Practices		
Data Onboarding		
Learning Library		
Tableau Guide		
Tableau Style Guide		

Data Team Handbook Structure

- Dashboards & Data You Can Use
- Data Learning and Resources
- Data Programs
- How The Data Team Works
  - Calendar
  - Data Analytics Team
  - Data Platform Team
  - Data Science Team
  - Data Team Principles
  - Data Management
  - Data Handbook Documentation
  - Planning Drumbeat
  - Triage
- How The Data Platform Works
  - Data CI Jobs
  - Data Infrastructure
  - Data Onboarding
  - Internship Experience

- Data for Product Managers
- Data Quality
- Data Services
- dbt Guide
- Enterprise Data Warehouse
- Jupyter Guide
- Meltano Guide
- Permifrost
- Python Guide
- RStudio Guide
- SQL Style Guide
- Snowplow
- Tableau
- Tableau Style Guide
- Trusted Data Framework

Data Catalog

The Data Catalog page indexes Analytics Dashboards, Workflows, and Terms.

Data Development

This page defines the Data Development lifecycle

Data Platform Security

Data platform security involves implementing measures to protect the confidentiality, integrity, and availability of data within our platform. This encompasses a range of strategies and technologies aimed at safeguarding sensitive information from unauthorised access, data breaches, and other security threats. These measures often include access controls, encryption, authentication mechanisms, monitoring tools, and compliance frameworks to ensure that data remains secure throughout its lifecycle within the platform. By prioritising data platform security, we can mitigate risks, maintain regulatory compliance and build trust with our stakeholders.

Data Quality

MVC for a Data Quality Program at GitLab

Data Team - How We Work

GitLab Data Team Workflow

Data Team Data Management Page

The Data Management Page covers the content around managing, securing, and governing the Enterprise Data Platform and related activities.

Data Team Direction

`This page contains forward-looking content and may not accurately reflect current-state or planned feature sets or capabilities.`

Strategy

As an important step towards achieving our mission, meeting our responsibilities, and helping GitLab become a successful public company, we are creating an Enterprise Data Platform (EDP), a single unified data and analytics stack, along with a broad suite of Data Programs such as Self-Serve Data and Data Quality. The EDP will power GitLab's KPIs, cross-functional reporting and analysis, and in general, allow all team members to make better decisions with trusted data. Over time, the EDP will further accelerate GitLab's analytics capabilities with features such as data publishing and products - enriched and aggregated data integrated into business systems or into the GitLab product for use by our customers. This acceleration happens through the development of "Data Flywheels", much like GitLab's Open Core and Development Spend flywheels.

Data Team Learning and Resources

GitLab Data Team Library

Data Team Organization

GitLab Data Team Organization

Data Team Platform

GitLab Data Team Platform

Data Team Programs

Introduction

Welcome to the **Data Programs** page. Here you'll find information about the various Data Programs around GitLab and those the Data Team supports, ranging from onboarding to day-to-day operations.

- Data Slack Channels
- **Primary Data Slack Channel**: `#data`
- **Data Lounge Channel**: `#data-lounge`

Show-n-Tell and Demos

Recordings of previous demos are posted to the GitLab Unfiltered Data Team playlist.

Data Onboarding

If you are onboarding to GitLab and will be working in the Data Program as an Engineer, Analyst, or Developer, follow these steps:

Enterprise Data & Insights Team Operating Principles

GitLab Enterprise Data & Insights Team Operating Principles Handbook

Functional Analytics Center of Excellence

The FACE is a cross-functional group of functional analytics teams that aim to make our teams more efficient by solving and validating shared data questions which results in cohesive measurement approaches across teams.

GitLab Experimentation Best Practices

Experimentation allows us to learn and give the right experiences to our Customers, to create better value for Customers and GitLab.

Learnings From Internships

Purpose

This page is created to share the experiences of Data Team members who completed internships with various teams within GitLab, with the aim of providing insights, learnings, and best practices.

Internship Experience

- Lessons from the Trenches: Insights from an SRE Internship
