GUIDE

Data Science and Machine Learning Platforms: Should You Build or Buy?

What's inside

3.........Introduction

4.........A centralized data science hub

6.........The baseline requirements

8.........Build vs. Buy: primary considerations

9.........Building your own platform takes time and talent

11.......Sustaining a platform

12.......Total Cost of Ownership (TCO)

13.......Build vs. Buy decision chart

14.......Anaconda for the Enterprise

As enterprises in virtually every industry embrace the potential of data science and machine learning, organizations inevitably face the challenges of operationalizing these processes in their technical environments and choosing investments according to their strategic objectives. As with many enterprise technology investments, the question of building a bespoke solution versus purchasing a vendor's platform that accommodates the needs of many is a decision debated across teams. A data science and machine learning platform must meet the needs of data scientists, developers, engineers, and business leaders to provide value to an organization.

This guide walks you through the key considerations of the build-versus-buy conundrum and directs you toward the optimum solution based on the composition of your team, infrastructure, and your business strategy. We also include a total cost of ownership calculator to help you weigh the financial cost against the more intangible considerations.

A centralized data science hub

A data science platform serves as a centralized hub for all of your organization's data science and machine learning efforts. This hub should be secure and convenient, residing behind a firewall on your company's infrastructure (whether it's in the cloud or on-prem) that can be accessed via web browser. Having everything in one place makes data science teams more efficient and enables them to focus on doing their jobs, instead of using their time to cobble together a DIY infrastructure that may not meet security requirements.

A data science platform should also be a medium for operationalizing machine learning (MLOps). One of the biggest challenges facing organizations today is to derive value from machine learning by deploying models into production (i.e. getting models to end users in the form of dashboards, web apps, or APIs). Data scientists are advanced statisticians, not software developers. They mainly use Python and/or R to work with different data sources and conduct mathematical operations. They build models using tools called notebooks, and they normally do not have the skill set to re-code these models for the web or for use as applications.

To derive value from these models, developers have to re-code them in Java, C#, HTML and other languages. This is where a data science platform comes in—it lets your data science team deploy to production without the costly step of rewriting into a language your production engineering team can manage. It speeds up time to production or time to prototyping if you use models for R&D. **Ultimately, an ML platform should accelerate time to value for the business, and it should make all systems and security run in the background, allowing data scientists to be data scientists, not software or systems engineers.**

A good data science platform, whether purchased from a vendor or built in-house, should go beyond a managed "sandbox." It should include all of the preferred data science tools (libraries and frameworks), and empower users to:

BUILD AND COLLABORATE

- Connect and efficiently access data sources across the organization, including proprietary databases, data lakes/warehouses, (Hadoop, Snowflake, Teradata) and compute engines (Spark, Kubernetes, Dask, GPU).
- Explore data visually and analytically to understand multiple perspectives. (It's really difficult, if not impossible, to build a robust model without understanding the characteristics of your data.)
- Build models ranging in complexity from linear regression to deep learning.
- Collaborate in a version-controlled environment and share code with other team members and stakeholders, via a web-based interface.

GOVERN RESOURCES

- Centralize IT control and put a policy enforcement framework around user access, software package licenses, and work artifacts.
- Automate compliance and data encryption procedures.
- Ensure your internal data network remains secure.
- Reduce the risk of data/code theft.
- Enable IT to designate "t-shirt sizes" for compute and storage resources so data scientists can simply select the size they need, and go to work on their projects.
- Meet regulatory and audit compliance.

DEPLOY MODELS

- Empower data scientists to deploy models without rewriting code or relying on IT and DevOps.
- Centralize administration and control of deployed applications and cluster utilization.
- Provide transparency around reproducibility and rollback to older versions of models.
- Allow for reproducibility by saving all components of projects at each micro-step in the process, tracking packages, environments, code, discussions, data, parameters, and results so that any step of any experiment can be instantly reproduced.

REFINE MODELS

- Refine models in production as needed. (Models need continuous improvement.)

SCALE WITH GROWTH

- Easily add more users to the platform.
- Add more storage resources and compute power to store and train an increasing number of models.

The baseline requirements

Whether you end up building or buying a data science platform, there are some baseline requirements your company must meet for infrastructure and processes before you can proceed in either direction:

A DATA SCIENCE TEAM THAT BUILDS MODELS

A data science team normally consists of data scientists and a data science manager. This team might work with a data architect, data analysts, and data engineers. The data architect role involves creating the blueprints for a data management framework, and it is almost certainly going to be a dual role that your most experienced data engineer will own. Collectively, this team has the knowledge to use data science tools and to develop, test, optimize and deploy models. If you don't already have a data science team that builds models, you're putting the cart before the horse. Your company should get started using analytic techniques to make data-driven decisions for your business, then start expanding that team. A data science platform alone will not be enough to make data-driven decisions for the organization.

If you don't already have a data science team that builds models, you're putting the cart before the horse.

MATURE DATA MANAGEMENT

You must have streamlined and organized your data sources. If your data is scattered all over the place in a number of different databases, then a data science platform is of reduced value. Every data source you need to integrate adds complexity to the implementation of a data science platform and onboarding a team. It takes longer to train people to access multiple data sources. It also makes it that much more painful when a person leaves your company.

Second, **your InfoSec team needs to be comfortable with data scientists poking and prodding sensitive data.** If your data science teams have a difficult time getting access to data, then that needs to be addressed before a platform makes sense. Otherwise, your platform will be unusable while data scientists wait to get access to new datasets they will need to do their work. Save your company some money and spare your team a few headaches by simplifying your data sources before you invest in additional tooling and infrastructure like a data science platform.

COMPUTE POWER

Is your organization currently running compute-intensive workloads? Some of the more complex machine learning models (looking at you, deep learning) scale horribly with the amount of data needed to train them. They require more data, and each subsequent row in a dataset takes longer to run. Without serious computing infrastructure, you'll be limited to only the most naive of data science models.

Another thing to keep in mind is the bursty nature of data science workloads. Either IT has to be okay with the provisioned data science hardware sitting idle for long stretches, or temporary cloud resources should be available for training periods. Intense compute power is needed to train models before they are moved into production. Some models can take several weeks to train.

Without serious computing infrastructure, you'll be limited to only the most naive of data science models.

A basic data management infrastructure for your organization should include some combination of the following:

DATA STORAGE	**Data warehouse** (Hive, Teradata, Snowflake, etc.) **Data lake** (HDFS, object storage, etc.)	**Traditional single-node or few-node databases** (Oracle, Postgres, etc.) **High performance databases** (Kinetica, kdb+)
BIG DATA INFRASTRUCTURE	**On-prem or in the cloud** (Hadoop, Spark on Yarn, or Hive)	
COMPUTE POWER	**GPU computing** (single-node or multi-node), such as Nvidia DGX series (up to $200k per server) **CPU computing**	**Big single-node servers** (>60 cores, >500GB RAM)

Build vs. Buy: primary considerations

The rest of this paper focuses on what we determine to be the most significant considerations of building versus buying a data science platform, followed by a total cost of ownership (TCO) calculator. Every organization is different, and **despite the shorter list of pros, building a platform will be in the best interest of some organizations.** Organizations that build platforms typically already have most of the resources and talent they need, and they already have some semblance of a platform in place. They may also have a very specific purpose for their data science program that is foundational to their business, such as maintaining a recommendation engine or bidding platform that works in real time. In these cases, companies have already built most of the infrastructure they need, and building a custom platform is probably the best choice.

For organizations that do not have all of the talent and resources it takes to build and support a platform, or who wish to focus on value delivery rather than infrastructure, buying a vendor solution is usually best. Vendors will guide teams through implementation and help them start getting models into production at a faster rate than building. Vendors also provide the training and support needed to fix bugs, keep a platform up to date, and train new users as the team grows.

PRIMARY CONSIDERATIONS	BUILD	BUY
Customization to meet your unique ML/AI needs	✓	
Shorter time to value		✓
Dependence on a vendor	✓	
Access to ready-made training resources		✓
Access to vendor’s support team		✓
Cost-effective (no recurring cost of in-house platform and support team)		✓

Building your own platform takes time and talent

To build your own data science platform, you must have the luxury of time, the right talent, and the right infrastructure in place.

TALENT

Data scientists are resourceful people, and they will do their best to meet their commitments, even if that means cobbling together a minimum viable platform (MVP) to do so. Some data scientists may have begun to realize that there's much to learn from software engineering practices, DevOps, and IT, and they have probably adopted some of these skills. Ultimately though, they are scientists and statisticians, and they can provide the most value building models, not being software developers and systems engineers.

Before you build a data science platform, you must have the following additional talent in-house: front-end engineers, highly specialized systems engineers, data science tools engineers, big data engineers, and ideally a UX designer. You may also want to consider naming an MLOps manager. This person will take on the role of ensuring the ML lifecycle runs smoothly and will act as the liaison between the data science, DevOps, and IT teams. These roles will make up your data science platform team. If you do not have some of these roles on staff, you could face long wait times trying to find the right people.

DATA SCIENCE TEAM

Data Scientists

Data Science Manager

Data Science Tools Engineer

DS PLATFORM TEAM

Front-end Engineers

Sr. Systems Engineers

Systems Architect

Data Science Tools Engineer

Big Data Engineers

UX Designer

QA Engineers

Product Manager

Network/Security Engineer

DevOps Engineers

Software Engineers

The skills to build a machine learning platform are very much in demand, with some salaries soaring over USD $180,000.

One last thing to consider after assessing the talent you have on hand is the current level of collaboration between your DevOps, IT, InfoSec and data science teams. Is the relationship currently functional and productive? If not, processes and relationships will need to be developed for a data science platform project to be successful.

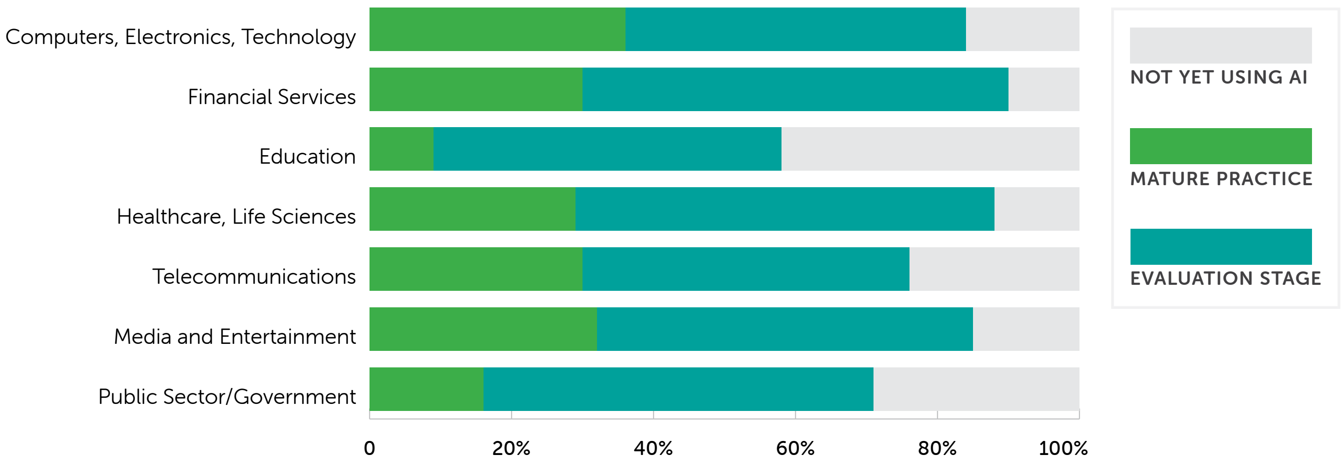
TIME

Suppose you have all of the necessary talent on staff to build a platform. Next, you have to find out if each of these people has the availability to dedicate themselves to building this platform full time. Most platforms will take approximately 12-18 months to build. Simply getting to a minimum viable product (MVP) stage can take months of full-time work depending on how far along your data science team has progressed on their own. An MVP solution is just the beginning of the iterative process required to get to an optimum solution, based on user feedback and evolving needs.

There's also an opportunity cost to consider. Are there other projects this group could be working on that provide more immediate value? Also, can you wait 12-18 months to put machine learning models into production? How much would you gain by having a few models in production in six months versus 18 months?

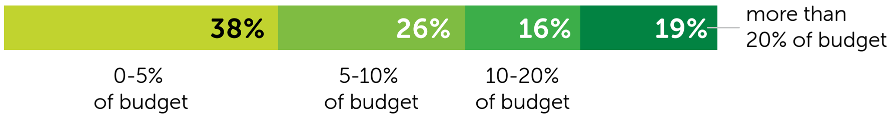
If your competitors have already operationalized machine learning, the opportunity cost to build your own platform could be much higher. The opportunity cost largely depends on the maturity of ML/AI adoption within your industry. The level of investment in AI is a good indicator of this maturity. According to a 2019 report from O'Reilly on the adoption of AI in the enterprise, the technology, health, and retail sectors are planning to invest most aggressively in AI, whereas the public sector is planning to invest the least.

STAGE OF AI ADOPTION BY INDUSTRY



PORTION OF IT BUDGET ORGANIZATIONS EXPECT TO COMMIT TO AI

over the next 12 months



Source: O'Reilly AI Adoption in the Enterprise

Sustaining a platform

If you have the in-house expertise to build a data science platform, the next thing to consider is maintenance and support. The majority, if not all, of your platform team will have to remain dedicated to the support and maintenance of the platform, not only to address and fix issues, but to ensure the platform stays up to date.

MAINTAINING A SUPPORT TEAM

Now that your team has successfully conquered the complicated task of building a custom data science platform, can you afford to keep them? They can command a higher salary now because they know how to build a platform, and everyone wants a data science/ML platform. You might think you could reduce the size of this team after the platform is complete, but it is unlikely. At most, you may be able to reduce the team investment by 30% by cutting the time needed from your UX designer and the number of engineers (if you had multiple to begin with). Chances are, your team will have to work their way through many bugs, and they will have to continue to upgrade and scale the platform as tools, infrastructure, and the needs of the data science team evolve.

STAYING UP TO DATE

The world of data science and machine learning is rapidly changing, with new innovative techniques and tools continuously coming from the open-source community. Data science platforms didn't really exist five years ago, so it's difficult to know what the future holds. If you build your own platform and make a heavy investment into your own application development, this might reduce your technological flexibility in the future. However, whether you build or buy a platform, ensuring that you can leverage open-source technologies is a safe bet. The open-source data science and machine learning community will continue to innovate faster than any one company can.

To ensure your platform continues to be a competitive advantage, your team will need to stay up to date on the latest technologies, especially in the open-source world. Keeping your team up to date also helps you maintain the talent and generally includes things like:

- Staying engaged with the open-source community and reading relevant blog posts.
- Going to annual industry conferences (PyData, Strata, Jupytercon, AnacondaCON, etc.) and related local meetups.
- Building up and tearing down infrastructure that you're not familiar with so you can get a better sense of what else is out there.

TRAINING AND DOCUMENTATION

A couple of the key benefits of buying a data science platform include support and training. Vendors supply documentation and training materials for their platform, helping to onboard new users. If you build a custom solution, your team will have to create the documentation and training materials. Most likely you will need to hire a full-time technical writer to assist with this work. It can take a long time to build a stable platform, and until sufficient documentation is developed over time, only the people that built it will know how to debug it. This can make scaling your platform difficult.

Total Cost of Ownership (TCO)

Now that we've discussed some of the reasons you might want to buy a platform (support, training, time) versus build a platform (custom design, vendor lock-in), let's look at the numbers. The cost to buy a vendor's solution usually starts around $100,000 and goes up from there. The cost depends on the size of your data science team and unique requirements.

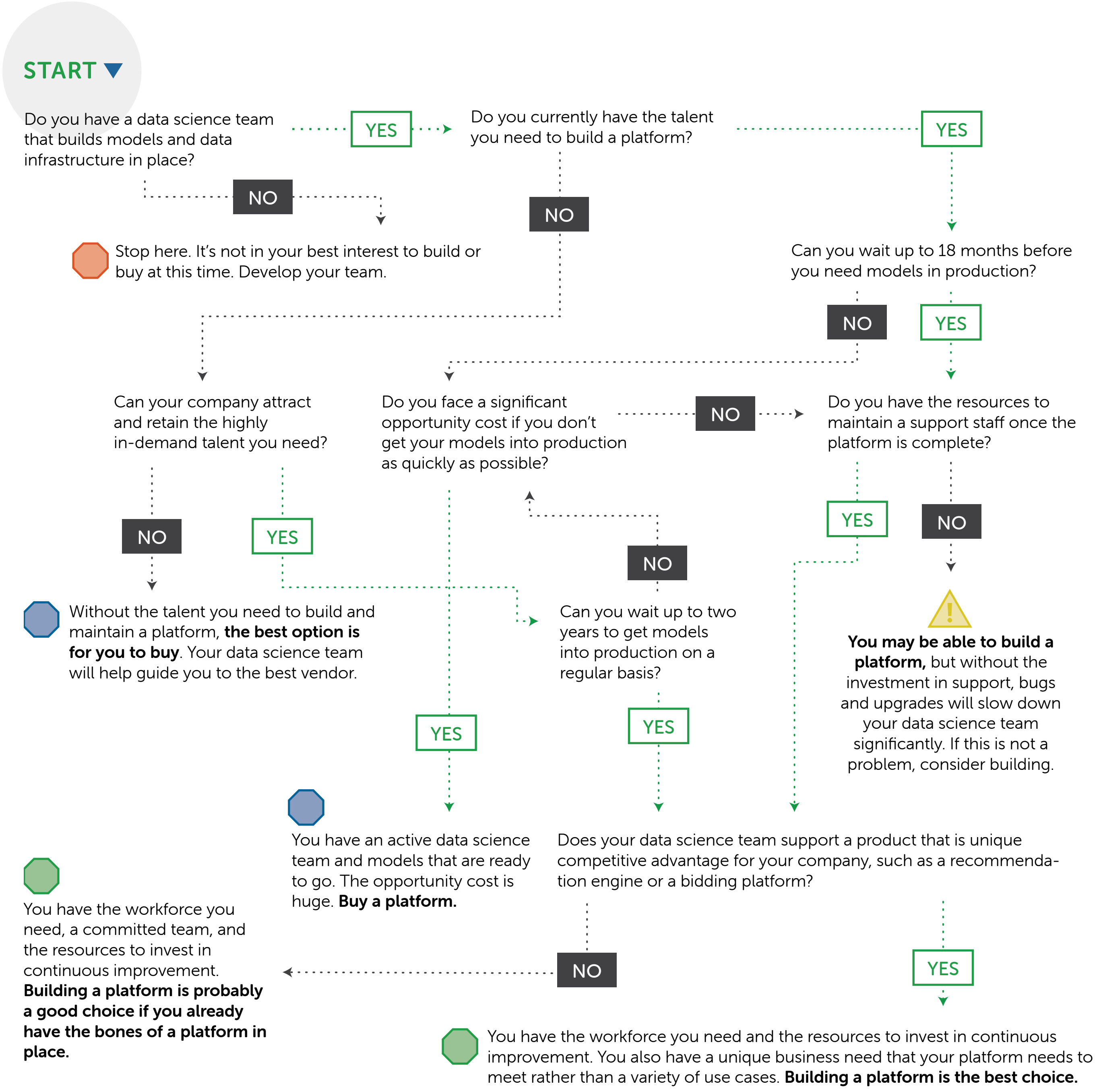
Our calculations for TCO to build a platform are based on the number of data scientists your organization has, and they exclude the costs of data and compute infrastructure. A company will have to maintain this infrastructure regardless of whether they build or buy a data science platform, so we opted to exclude these costs. Using the number of data scientists, we can determine approximately how many team members you will need to build and maintain a platform, and then we determine the total cost based on average U.S. salaries. The build cost below includes one year of building your own platform plus four years of maintenance.

DATA SCIENCE TEAM SIZE	COST TO BUILD IN USD (1ST YEAR)	ANNUAL COST TO MAINTAIN	5-YEAR TCO
10	$620,000	$310,000	$1,860,000
50	$1,810,000	$900,000	$5,410,000
100	$3,040,000	$1,520,000	$9,120,000
500	$4,850,000	$2,420,000	$14,530,000
1000	$6,420,000	$3,210,000	$19,260,000

To see how we made these calculations and to plug in your our numbers, visit our Data Science Platform TCO Calculator at: https://tco.internal.training.anaconda.com/TCO

Build vs. Buy decision chart

You now have most of the information you need to make a thoroughly considered decision about your data science infrastructure, and how to move your team (and the company) forward. Use this flowchart to navigate the key considerations we've discussed, and see if your company arrives in the "build" or the "buy" camp.



Do you have a data science team that builds models and data infrastructure in place?

Do you currently have the talent you need to build a platform?

Stop here. It's not in your best interest to build or buy at this time. Develop your team.

Can you wait up to 18 months before you need models in production?

Can your company attract and retain the highly in-demand talent you need?

Do you face a significant opportunity cost if you don't get your models into production as quickly as possible?

Do you have the resources to maintain a support staff once the platform is complete?

Without the talent you need to build and maintain a platform, **the best option is for you to buy**. Your data science team will help guide you to the best vendor.

Can you wait up to two years to get models into production on a regular basis?

You may be able to build a platform, but without the investment in support, bugs and upgrades will slow down your data science team significantly. If this is not a problem, consider building.

You have an active data science team and models that are ready to go. The opportunity cost is huge. **Buy a platform.**

Does your data science team support a product that is unique competitive advantage for your company, such as a recommendation engine or a bidding platform?

You have the workforce you need, a committed team, and the resources to invest in continuous improvement. **Building a platform is probably a good choice if you already have the bones of a platform in place.**

You have the workforce you need and the resources to invest in continuous improvement. You also have a unique business need that your platform needs to meet rather than a variety of use cases. **Building a platform is the best choice.**

Whether you choose to build or buy a platform, Anaconda has solutions that can help. As a founder and long-time contributor to the open-source data science community, Anaconda has become the de-facto standard for data science and machine learning with tightly-integrated open-source packages, libraries, and tools preferred by developers and data scientists. Anaconda provides solutions that empower organizations to do serious data science and deploy machine learning models into production where they can move businesses forward.

Anaconda has become the de-facto standard for data science and machine learning with tightly-integrated open-source packages, libraries, and tools preferred by developers and data scientists.

ANACONDA SERVER

Anaconda Server helps InfoSec and IT managers secure their open-source pipelines in on-premise, air-gapped, and private cloud environments while allowing data science teams to use their preferred tools and packages.

With Anaconda Server's advanced package, channel, and user management tools, security teams can build an open-source pipeline that fortifies and complies with your organization's security standards. Leverage continuous security updates from Anaconda to review and safelist packages in a sandbox, and enable native access to safe and compliant packages amongst your distributed data science team channels.

ANACONDA ENTERPRISE DATA SCIENCE PLATFORM

Envision a world where data scientists can regularly deploy AI and machine learning projects into production at scale, quickly delivering insights into the hands of decision-makers. How would that impact your business? Anaconda Enterprise Data Science Platform supports your organization no matter the size, easily scaling from a single user on one laptop to thousands of machines with built-in failover controls and security. No headaches, no IT nightmares.

About Anaconda

With more than 25 million users, Anaconda is the world's most popular data science platform and the foundation of modern machine learning. We pioneered the use of Python for data science, champion its vibrant community, and continue to steward open-source projects that make tomorrow's innovations possible. Our enterprise-grade solutions enable corporate, research, and academic institutions around the world to harness the power of open-source for competitive advantage, groundbreaking research, and a better world.

Visit www.anaconda.com to learn more.

ANACONDA.