ANACONDA®

THE COMPLETE BUYER'S GUIDE TO DATA SCIENCE PLATFORMS

“The most impactful and effective AI strategies will stand on the shoulders of robust data science capabilities.”

— Dr. Hugo Bowne-Anderson, Data Scientist

ABOUT THIS GUIDE

After witnessing its power to transform entire industries, many companies are embracing data science to develop machine learning models that optimize their processes and differentiate their offerings. However, no machine learning model can provide value until it is deployed and in front of end users. For some companies, this is a lengthy process that involves developers recoding everything data scientists have done.

Operationalizing data science is tricky. This is where the right enterprise data science platform comes in.

To deliver business value through data-enabled differentiation, a data science platform has to be embraced by both data scientists and developers. It must enable these teams to efficiently develop, collaborate, govern, and deploy models at scale. And it has to meet IT’s requirements.

The data science/machine learning platform space is dynamic and crowded with very different products. This guide will walk you through organizational readiness, differences between platform types, and key considerations to evaluate vendors in this space. We also include a detailed interactive checklist to help your team through the evaluation process.

IS YOUR COMPANY READY FOR A DATA SCIENCE PLATFORM?

Nearly every organization today is exploring ways to achieve competitive differentiation through data science and machine learning. But the success of an investment in a data science or machine learning platform depends on how well it addresses the needs and concerns of data scientists, the IT team, and executive leadership. The right data science platform takes the headache out of operationalizing data science, enabling your talented scientists and developers to focus on what they do best, and it expedites the process of getting data science output into the hands of the end user.

> Before embarking on the journey to find the right platform for your company, you should first carefully assess your organization's readiness by asking yourself one essential question:
> Is leadership dedicated to the data science team's success?

Data science will not have the expected impact if it exists in a silo. Executive leadership must care enough about the integration of data science and machine learning to pull for the resources, support, and collaboration the program will require throughout the company. The data science team needs support from the business side to ask the right questions and access the right data.

To ensure continued support, leadership should ask themselves why they want a data science or machine learning program. Business outcomes around competitive advantage, productivity, and cost containment should be clearly defined. Once these goals are clarified, have your enterprise solution architect outline the requirements to accomplish those goals. These requirements will make up the key specifications you will look for on your search.

GETTING STARTED ON YOUR SEARCH

Data science requires collaboration across teams, especially between developers and data scientists. Your data science team will have a list of data sources, notebooks, visualization tools, languages (most use Python and/or R), libraries and repositories they use to work effectively. Make note of these tools to ensure compatibility.

Involve IT as soon as possible. No data science platform will be successful without IT's support. IT will enforce governance policies and ensure compatibility with current systems, hardware, and data sources, and they will be integral to the implementation process.

Envision data science as a fabric overlay, connected to all enterprise functions, providing insight, improving measurement and analytics, and ultimately guiding strategy.

OPEN-SOURCE VS. PROPRIETARY PLATFORMS

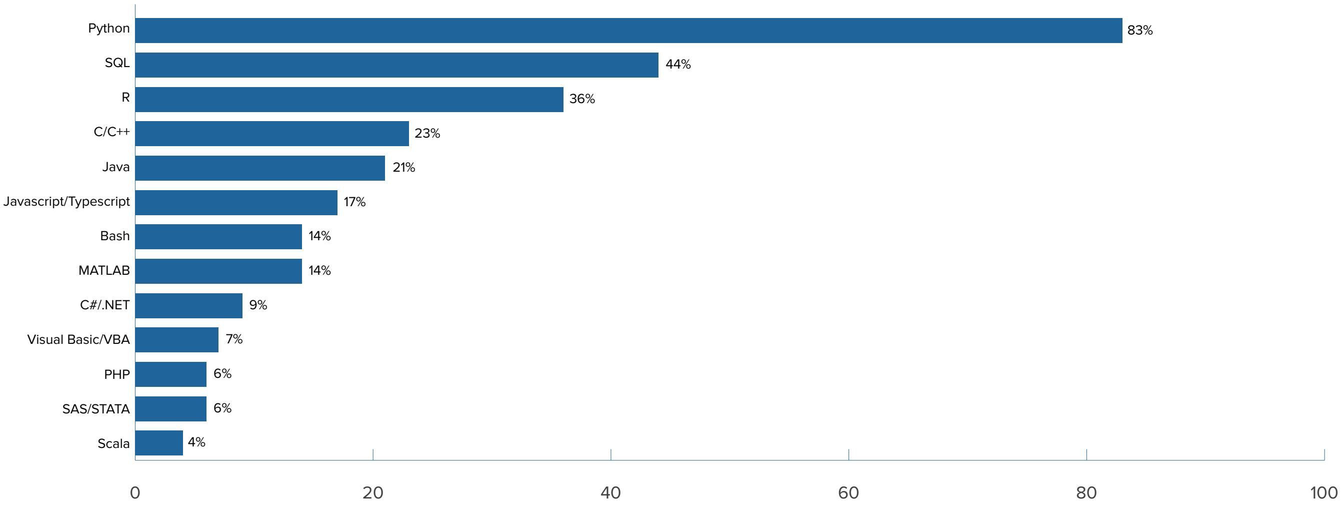
The open-source community always innovates faster than one company alone can, and the data science space is no exception. With open-source platforms, data scientists can experiment with the latest technologies using tools like TensorFlow and Scikit-Learn, which contain some of the latest algorithms for visual and audio data processing.

> The open-source community always innovates faster than one company can.

While proprietary platforms can be useful for predefined use cases in some industries, choosing a proprietary vendor may ultimately be limiting. If you choose a proprietary platform, your company can only innovate when that vendor develops new capabilities. Proprietary platforms can be useful for some use cases and environments, but they do not empower data scientists with open-source technologies, and they do not allow them to use the languages they were trained to use. They also cost 20-50% more than open-source platforms, and innovation is limited to what the vendor's team can create.

One final consideration: talent. Consider that data scientists and developers have preferred tools, programming languages, and ways of collaborating. In a tight talent market, the importance of mapping your solution choice to talent outcomes cannot be overstated.

Diversity in Preferred Programming Languages among Data Scientists, Engineers, and Analysts

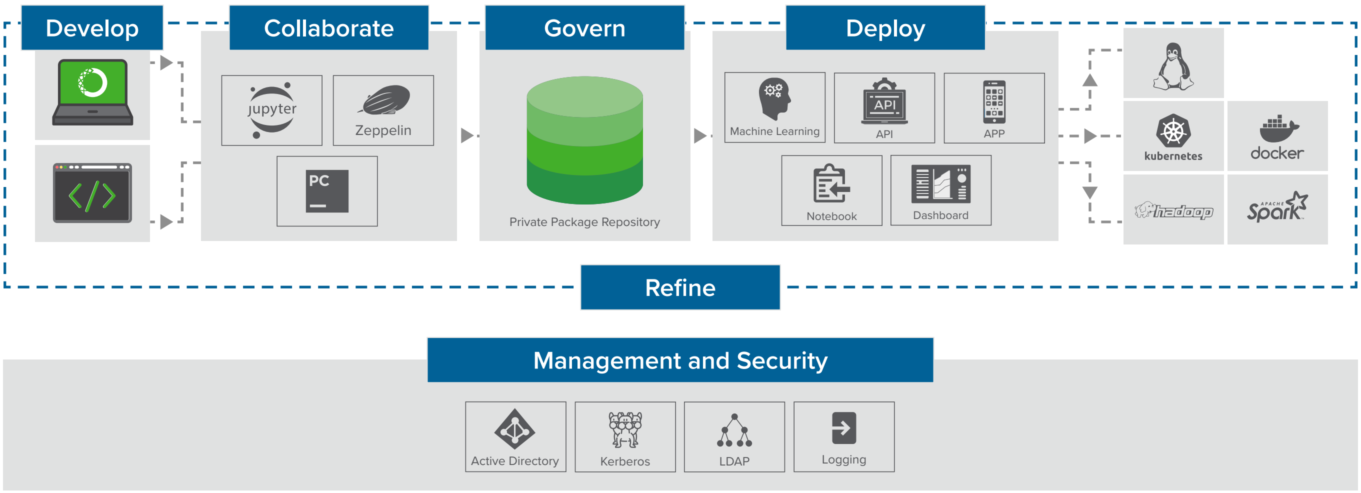


Source: Business Over Broadway, 2018 Kaggle Machine Learning and Data Science Survey. Kaggle surveyed over 16,000 data professionals.

KEY CONSIDERATIONS

Before your company begins evaluating platform features, be cognizant of the fact that some of the data science “platforms” on the market are just tools. They may provide algorithms or models data scientists or analysts can plug data into, but they limit data science teams’ ability to customize and innovate within models. Tools with industry-specific, prepackaged algorithms have the potential to lose their value quickly and do not form a basis from which to develop true differentiation.

Collaboration and the Machine Learning Lifecycle



Evaluate data science platforms on their ability to enable collaboration, bring efficiency to deployment, and to satisfy IT requirements for security and governance.

Machine learning lifecycles vary from company to company, but they all start with acquiring and exploring data and end with models going into production, where they are continually refined and managed. The cycle continually repeats as models are improved. Data scientists and developers must collaborate throughout this cycle, and the faster your data science team can get models into testing and then production, the more cost-effective it will be for the company. Developers and data scientists favor different tools, and your chosen platform should enable them to work with their preferred tools.

Another consideration is the scalability of your machine learning program. How many data scientists and developers can collaborate on the platform at once? Can the platform grow alongside your team from a handful of contributors to hundreds or even thousands? Consider how the platform will meet your needs 5 to 10 years down the road.

Security and Governance

When a company doesn't have a data science or machine learning platform, projects are developed and stored on disparate laptops. Data scientists have no choice but to download their own tools and data science packages, which have not been scanned or approved by IT. Models stored on laptops are often trapped there because they are extremely difficult to move into production.

Choosing a platform that meets IT's standards for security and governance and integrates with their current processes and tools will be integral to the success of the platform. Important considerations include user access control, package governance, encryption of data at rest and in transit, and audit trails.

Scalability and Infrastructure

It's important to remember data scientists are not infrastructure engineers. For the data scientist or developer, the platform should function as a PaaS (Platform-as-a-Service), whether it's on-prem or not, that allows them to spin up the compute resources they need with just a few clicks. A data science platform should allow IT to designate what resources data scientists can access and the sizes of those resources. The most flexible data science platforms can run in the public cloud, in a private cloud, or on bare metal in an enterprise data center.

> Stable and effective infrastructure setup is perhaps the most complicated and arduous task for a data science platform vendor to accomplish.

Stable and effective infrastructure setup is perhaps the most complicated and arduous task for a data science platform vendor to accomplish. Companies like Google, Netflix, and Spotify are often recognized for building their own machine learning/data science platforms. Google's machine learning team came to this conclusion after some time of maintaining their own platform: "**A mature system might end up being (at most) 5% machine learning code and (at least) 95% glue code,**" meaning maintaining a machine learning system is complicated and incurs a significant amount of technical debt. Having a foundation for your data science program that is automated, maintained, and easily scaled is an immense timesaver for any company that needs to focus on using data science for competitive advantage.

With these considerations in mind, we developed our Data Science Platform Buyer's Checklist. Once you know the basic needs of your IT and data science teams, use the following checklist to evaluate vendors and narrow down your choices.

DATA SCIENCE PLATFORM BUYER’S CHECKLIST

Security & Governance

	Included	Can integrate	Not possible
Role-based access control	□	□	□
Cloud-native security control	□	□	□
Publishing permissions	□	□	□
Secure package repository	□	□	□
User ID management	□	□	□
End-to-end encryption	□	□	□
Provides administrative monitoring (track users, projects, deployments)	□	□	□
Package signing	□	□	□
Package vulnerability scanning	□	□	□

Data Integration

	Can integrate	Does not integrate
Hadoop (Cloudera, Hortonworks, EMR)	□	□
Proprietary databases (SAS, Teradata)	□	□
Web data integration	□	□
Code repositories (GIT, Bitbucket)	□	□
Monitoring solutions (log shipping)	□	□
NoSQL	□	□
Filesystems	□	□
SQL	□	□
Data lake support	□	□
IoT/sensor data	□	□

Infrastructure & Hardware Support

	Yes, air-gapped	Yes, but not air-gapped	Not supported
AWS	□	□	□
Azure	□	□	□
Google	□	□	□
On-prem (VSphere)	□	□	□
On-prem (bare metal)	□	□	□
Air-gapped	□	□	□
GPU	□	□	□
CPU	□	□	□

Machine Learning Capabilities

	Supported	Not supported
Classification & regression	□	□
Support vector machines (SVMs)	□	□
Deep learning	□	□
Time-series analysis	□	□
GANs	□	□
Testing strategies (A/B, multi-armed bandit, sensitivity analysis)	□	□
Text & image analytics and processing	□	□
Reinforcement learning	□	□

COLLABORATION & TOOLS

Notebooks & IDEs

	Included	Not included
Jupyter Notebooks	□	□
JupyterLab	□	□
PyCharm	□	□
Zeppelin	□	□
RStudio	□	□
Spyder	□	□
Visual Studio Code	□	□
Atom	□	□

Data Visualization Tools

	Accessible	Not accessible
Matplotlib	□	□
Bokeh	□	□
Panel	□	□
PyViz	□	□
Tableau	□	□
Alteryx	□	□
Plotly	□	□
Shiny	□	□

Data Science/Machine Learning Libraries

	Accessible	Not accessible
NumPy	□	□
TensorFlow	□	□
Pandas	□	□
XGBoost	□	□
SciKit-Learn	□	□
Keras	□	□
SciPy	□	□
PyTorch	□	□
Theano	□	□

Model Deployment and Management

	Yes	No
Deployment from QA	□	□
Deployment to production	□	□
One-click deployment to pre-provisioned resources	□	□
Refine models in production	□	□
Reproducibility - rollback to older models	□	□
Centralized administration of deployed apps	□	□

ANACONDA ENTERPRISE | PRACTITIONER-PREFERRED, IT-APPROVED

The Anaconda Enterprise data science platform powers the machine learning lifecycle. Anaconda Enterprise enables data scientists to use the tools and languages they love and were trained to use, while providing secure, curated access to more than 1,500 data science packages. Developers can build differentiating applications and services, and ML/IT operations can deploy and refine models in production. Anaconda Enterprise delivers a secure environment with governance and access control. Lastly, it meets the needs of senior leadership by making the strategic imperative of data-enabled differentiation a reality.

Learn More

ANACONDA.

With more than 15 million users, Anaconda is the world's most popular data science platform and the foundation of modern machine learning. Our flagship product, Anaconda Enterprise, delivers data science and machine learning at speed and scale, unleashing the full potential of our customers' data science and machine learning initiatives. Visit anaconda.com/enterprise to learn more.

anaconda.com