How to Choose the Right Data Science and Machine Learning Platform

Published: 12 March 2019 **ID:** G00382503

Analyst(s): Carlie Idoine, Jim Hare

The data science and machine learning platform market is crowded and complicated. To choose correctly, data and analytics leaders must assess their organization's analytics maturity, define business-driven requirements and segment vendors' platforms into clusters for evaluation.

Key Challenges

- Data and analytics leaders often lack a clear and coherent picture of their team's ability to evaluate, deploy and exploit data science and machine learning (DSML) technologies.
- DSML platform decisions can easily fall short of expectations when the decision focus is primarily on technical features.
- The data science market is a confusing, fast-changing landscape of vendors with similar-sounding claims and widely varied platforms and tools.
- Data and analytics teams often lack the expertise to assess DSML platforms in the light of platform capabilities and market trends.
- DSML skills remain in short supply, potentially crippling an organization's ability to successfully exploit data science platforms for business value.

Recommendations

To optimize analytics and business intelligence solutions for data science and machine learning:

- Determine your organization's current state of analytics maturity by assessing its capabilities, staffing and skills, and available tools.
- Clarify your DSML requirements by mapping out both your specific analytic pipeline — from raw data to business impact — and your specific analytics process stages.
- Assess platform similarities and differences by distinguishing different types of platforms and clustering them into "like" categories.

- Evaluate data science platforms by assessing not only current capabilities, but also the vendors' vision and execution and the current key market trends to which they are responding.
- Cultivate DSML expertise by planning for sustained long-term investment in staff hiring and skills improvement.

Table of Contents

List of Tables

List of Figures

Strategic Planning Assumptions

By 2025, 50% of data scientist activities will be automated by artificial intelligence, easing the acute talent shortage.

By 2020, predictive and prescriptive analytics will attract 40% of enterprises' net new investment in analytics and business intelligence.

By 2020, the number of data and analytics experts in business units will grow at three times the rate of the experts in IT departments, forcing companies to rethink their organizational models and skill sets.

By 2020, augmented analytics will be a dominant driver of new purchases of analytics and business intelligence, as well as of data science and machine learning platforms and of embedded analytics.

By 2022, cloud-based machine learning services from hyperscale cloud providers will achieve the digital tipping point of a 20% share in the data science platform market.

Introduction

The hype around machine learning (ML) and artificial intelligence (AI) capabilities is intensifying the turmoil of the rapidly changing data science and machine learning (DSML) market. The result is a profusion and confusion of platform options, capabilities and approaches. None of these can work as *the* perfect platform for an organization's DSML requirements. Data and analytics leaders face a complex process of evaluating solutions that are apparently similar in some ways — but vastly different in others — and identifying the specific trade-offs that each one will impose.

There are emerging best practices that data and analytics leaders can use to guide their platform decisions. These practices cut through the confusion and identify one or more DSML platforms that match the organization's analytics maturity and satisfy its analytics requirements. Data and analytics leaders should use the approach presented in this report to orient their teams and business stakeholders so that they can confidently assess and navigate this fast-changing market.

Analysis

Assess Your Current Data Science and Machine Learning Maturity

Data and analytics leaders must ruthlessly assess the capacity of their organization to successfully adopt, deploy and implement a complex DSML platform. This capacity determines whether the organization will be able to execute its platform decision successfully for business value. The three key areas for assessment are discussed in the following paragraphs.

Can Your Organization Leverage a DSML Platform?

Deciding on a platform must begin by assessing how capable and willing the IT and business stakeholders are to use it effectively — that is, the organization's data and analytics maturity level. Can these stakeholders move beyond their current practices and overcome separate — and sometimes conflicting — system, data and organizational silos? You will need to match the capabilities you have with those required by DSML platforms. Selecting a platform to build your own custom DSML solutions is not the only option. You can choose from an expanding universe of prebuilt DSML-capable solutions targeting specific industries or functions; some platform offerings

This research note is restricted to the personal use of chongfeng@airchina.com.

already have such prebuilt modules. DSML service providers offer other expertise in addition to the tools to jump-start your DSML initiatives.

Actions for data and analytics leaders:

- Use Gartner's "ITScore for Data and Analytics" to determine your organization's current analytics maturity. The maturity levels include basic, opportunistic, systematic, differentiating and transformational.
- For each level of maturity, five disciplines enable stakeholders to improve that capability: data and analytics vision and strategy, value and outcome management, people, skills and organization, and technology and solutions.
- This model guides you in moving to the next level of maturity. Your current level of maturity will be an important variable in determining whether to build, buy or outsource your DSML platforms (see "Data Science and Machine Learning Solutions: Buy, Build or Outsource?"). Remember that the right solution for your requirements may combine all three of these options.

Do Your Stakeholders Have the Right Skills for DSML?

The breadth, depth and sophistication of DSML skills — both within IT and among business users — is vital to sustained DSML success. However, these skills are currently in a state of high demand and low supply. In some cases, organizations will hire qualified candidates, but in many cases, successfully exploiting a DSML platform requires sustained training, enabling and motivation of existing employees.

Actions for data and analytics leaders:

- Identify the existing and prospective analytics users, formal and informal roles, in both IT and the business units.
- Determine what types of analytics — descriptive, diagnostic, predictive and prescriptive — they are currently using and their skill levels with those tools.
- Prioritize the skill sets that the different roles will need in order to extend their capabilities to exploit DSML platforms and tools.

What Do You Have in Your Analytics Toolbox to Support DSML?

IT and business stakeholders already have analytics tools that can, and must, be leveraged to enable and support DSML capabilities. Examples include:

- Analytic and business intelligence tools that are starting to incorporate data science capabilities (see "Magic Quadrant for Analytics and Business Intelligence Platforms")
- Platforms that focus specifically on building DSML models (see "Magic Quadrant for Data Science and Machine Learning Platforms")

- Augmented analytics tools used by citizen analysts or citizen data scientists (see "Augmented Analytics Is the Future of Data and Analytics")
- Tools with the capabilities for data integration and data quality to meet the demanding needs of DSML models. (See "Magic Quadrant for Data Integration Tools" and "Enabling Data Quality for Machine Learning and Artificial Intelligence.")

Actions for data and analytics leaders:

- Inventory your existing analytics tools. Determine if they have wider applicability in the organization to support a DSML platform.
- Leverage tool capabilities that exist outside of IT and the data and analytics team. Identify what's weak, outmoded or missing.

> This three-part assessment is not a one-time exercise. Tools and skills must be continually assessed, developed, reviewed and adjusted to ensure that existing and evolving needs are being met and reinforced.

Data and analytics leaders should work with human resources to cultivate, develop and sustain a cadre of talent in support of DSML activities throughout the organization.

Assess the End-to-End Analytic Pipeline and Life Cycle

To support the entire DSML process, data and analytics leaders should assess the requirements and capabilities of two related areas:

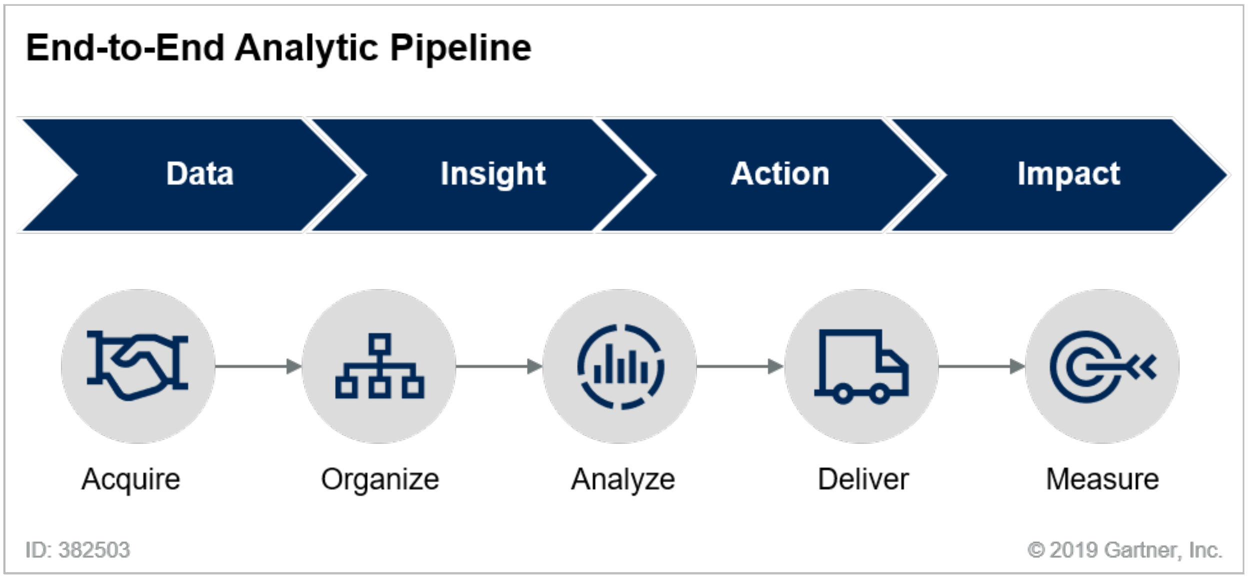
- The high-level, end-to-end analytic pipeline
- The specific DSML analytic life cycle

The Analytic Pipeline

All analytics platforms, including DSML, must support the end-to-end analytic pipeline (see Figure 1). The pipeline elements enable users to move from raw data to business impact. Any analytic platform enables users to acquire and organize the required data in a way that prepares it for analysis. The platform next facilitates delivering the analytic artifacts in such a way that their business impact can be accurately measured.

By using this pipeline, data and analytics leaders can focus on creating a robust toolbox that provides capabilities across the complete pipeline, and assess gaps in their current offerings — or areas of opportunity to support the pipeline.

Figure 1. The End-to-End Analytic Pipeline



Source: Gartner (March 2019)

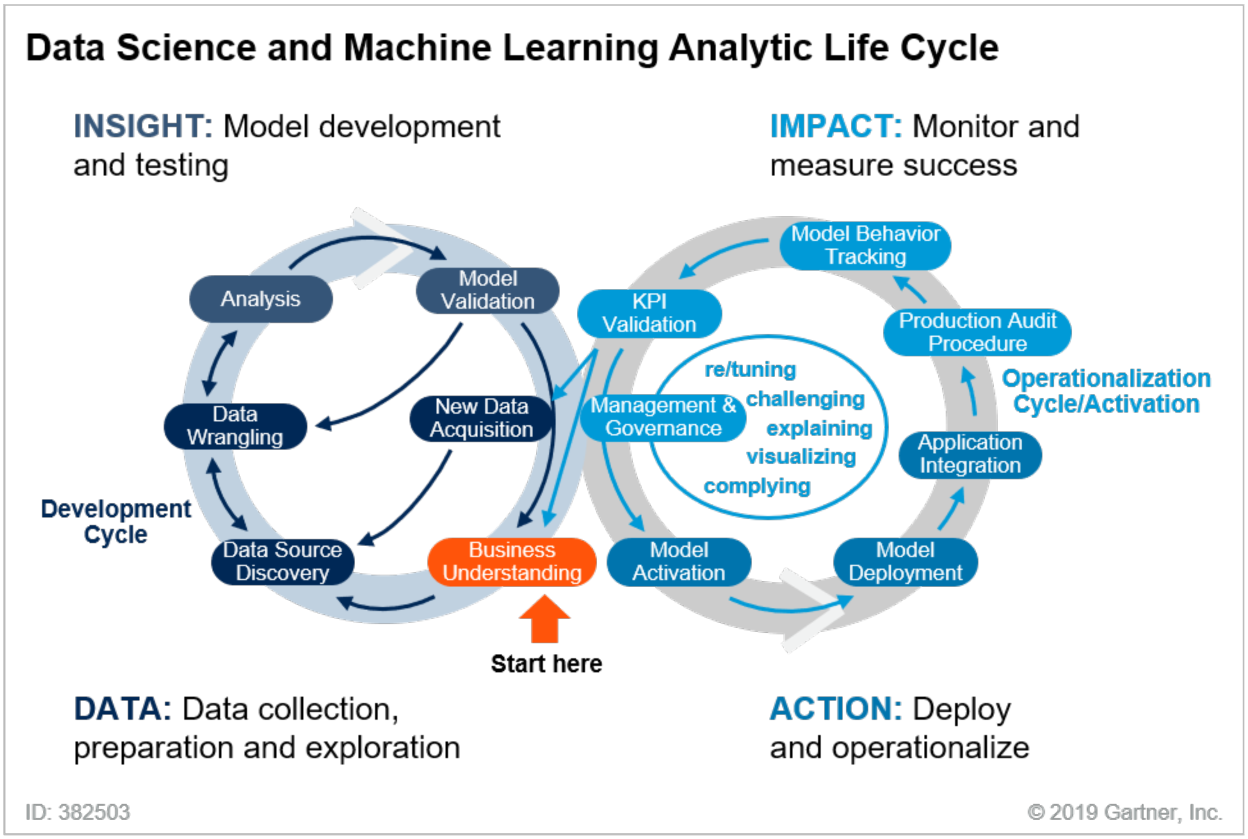
The DSML Analytic Life Cycle

The elements of the analytic pipeline can be further related to specific DSML capabilities and requirements, creating the DSML analytic life cycle (as shown in Figure 2). For a description of DSML capabilities in this life cycle context see "Critical Capabilities for Data Science and Machine Learning Platforms."

This life cycle idea lets data and analytics leaders lay out the specific DSML life cycle steps that must be completed. By doing this, they can identify the specific, related critical capabilities that the DSML platform must support.

This research note is restricted to the personal use of chongfeng@airchina.com.

Figure 2. The Data Science and Machine Learning Analytic Life Cycle



KPI = key performance indicator

Source: Gartner (March 2019)

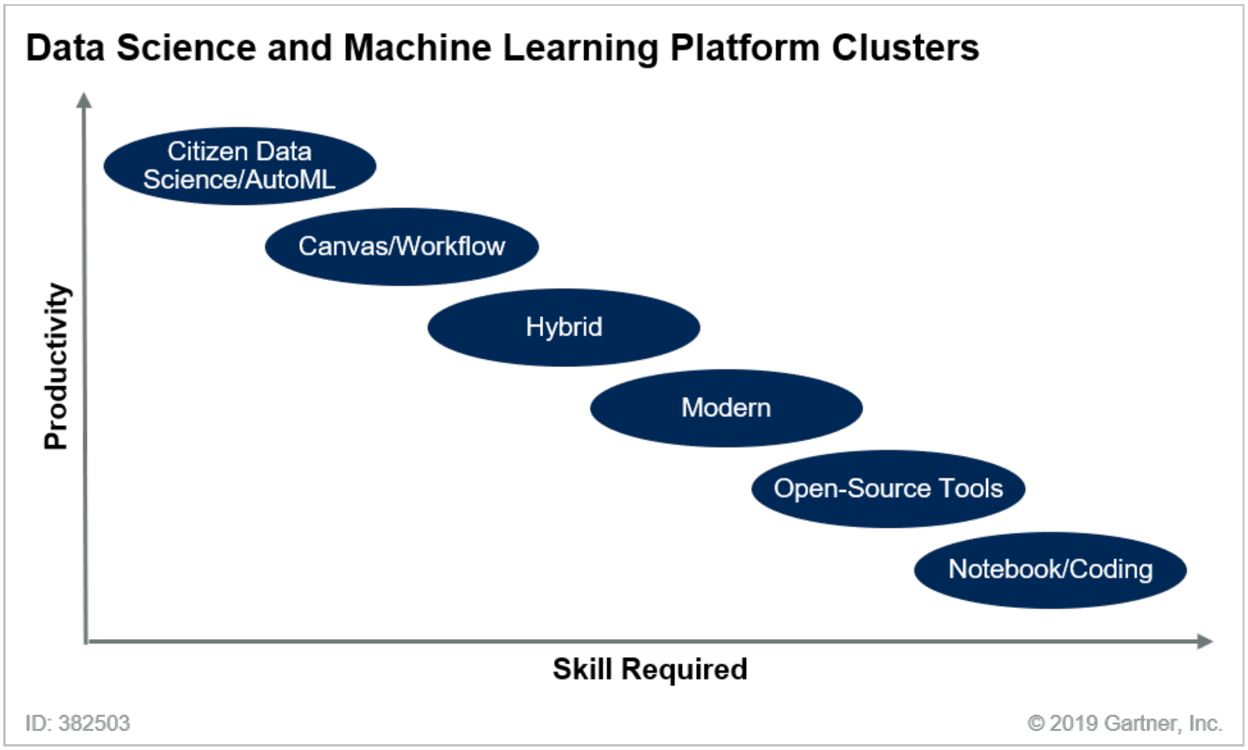
By leveraging these two concepts — the end-to-end analytic process and the DSML analytic life cycle — data and analytics leaders can create a blueprint that lays out the capabilities they need to sustain and scale data science and machine learning. These capabilities become the requirements by which they can evaluate DSML platforms. Using the pipeline and life cycle, these leaders also can assess both the tools currently being used within the organization and any potential new tools needed to extend specific DSML capabilities or fill gaps in the current toolset.

Segment Platforms Into Appropriate Clusters to Assess Similarities and Differences

Gartner has developed a loose categorization of DSML platforms — into "clusters" (see Figure 3).

This research note is restricted to the personal use of chongfeng@airchina.com.

Figure 3. Data Science and Machine Learning Platform Clusters



Source: Gartner (March 2019)

These platform clusters enable data and analytics leaders to make sense of different types of DSML platform and begin to distinguish between them based on the organization's analytics requirements. Table 1 lists the cluster types, their key features and, for each cluster, a few representative platforms.

Table 1. Data Science and Machine Learning Platform Clusters

Cluster Label	Platform Characteristics	Representative Platforms*
Canvas/ Workflow	▪ Historical, legacy use ▪ Enterprise-grade platform and model management capabilities ▪ Embracing hybrid approach (see below) ▪ Often used by data scientists	▪ SAP Predictive Analytics ▪ IBM SPSS ▪ SAS Enterprise Miner
Notebook/ Coding	▪ Leverage an open-source framework ▪ Have commercial implementations for orchestrating the analytic DSML life cycle ▪ Leverage a notebook-based approach for generating code and collaboration ▪ Often used by developers and expert data scientists	▪ H20 Open-Source Machine Learning ▪ Anaconda Enterprise ▪ Databricks Unified Analytics Platform
Hybrid	▪ Intersection between Canvas and Notebook platform types, with open-source support ▪ Data and processing power scalability ▪ Support for increased number of models and complex model management ▪ Support for new artificial intelligence frameworks ▪ Support for roles beyond the expert data scientist, including the citizen data scientist and developers	▪ MATLAB ▪ KNIME ▪ RapidMiner
Open-Source	▪ No commercial implementation ▪ Often used to jump-start initiatives ▪ Good for experimentation ▪ Low total cost of ownership ▪ Often used by developers and expert data scientists	▪ R ▪ Python ▪ Scala
Citizen Data Science/ AutoML	▪ Incorporate augmented analytics ▪ Enable guided data preparation, data discovery, model building and operationalization ▪ Enable "power users" to perform both simple and moderately sophisticated analytical tasks that would previously have required more expertise ▪ Provide efficiency gains for expert data scientists	▪ DataRobot ▪ Big Squid ▪ Tellius ▪ H2O Driverless AI

This research note is restricted to the personal use of chongfeng@airchina.com.

Cluster Label	Platform Characteristics	Representative Platforms*
* These are examples only, not an exhaustive list		

Source: Gartner (March 2019)

Understanding the clusters and identifying those that are most relevant to an organization's specific needs is a first step. Within each cluster, data and analytics leaders can then identify the specific DSML capabilities of each platform (see "Critical Capabilities for Data Science and Machine Learning Platforms"), further differentiating these offerings.

Gartner currently lists 15 critical capabilities for DSML platforms:

- Data access
- Data preparation
- Data exploration and visualization
- Automation and augmentation
- User interface
- Machine learning
- Other advanced analytics
- Flexibility, extensibility and openness
- Performance and scalability
- Delivery
- Platform/project management
- Model management
- Precanned solutions
- Collaboration
- Coherence

Data and analytics leaders should carefully review and assess the definitions of each capability, then select and weight the capabilities that are most important for their specific needs. This approach enables customization of a use case based on an organization's business and technology needs, weighing each capability and its importance against the others. Our critical capabilities research identifies several default use cases, including: business exploration, advanced prototyping, production refinement and nontraditional data science. Each use case applies weightings to each critical capability to indicate its relevance within that specific use case. Leveraging the defined use cases — or customizing the weightings to create a specific use case — will result in a shortlist of

This research note is restricted to the personal use of chongfeng@airchina.com.

platforms to consider for an organization's specific DSML needs. The list of capabilities can also be used to evaluate other platforms for consideration.

Gartner's "Toolkit: RFP for Data Science and Machine Learning Platforms" has a Microsoft Excel spreadsheet that breaks each critical capability into several related subcapabilities (see Figure 4). The spreadsheet also provides a detailed description of each subcapability, to assist with assessing whether a particular platform provides that subcapability. Once the shortlist of platforms for evaluation is determined, the RFP Toolkit can be customized. It can then be used as an RFP to send to the specific vendors, or as a checklist for performing a proof of concept (POC) for evaluating each of the platforms.

This research note is restricted to the personal use of chongfeng@airchina.com.

Figure 4. A Sample of the Data Science and Machine Learning Subcapabilities

Sample of Data Science and Machine Learning Subcapabilities

Capability	Functionality	Question
Flexibility, Extensibility and Openness	R	Does the product support R?
	Python	Does the product support Python?
	Scala	Does the product support Scala?
	Java	Does the product support Java?
	Third-Party Libraries	Does the product support algorithms available via third-party libraries, products or marketplaces?
	Popular Libraries and Frameworks	Does the product provide support for XGboost, sk-lean, Caffe,Theano, Torch/PyTorch and/or TensorFlow?
	Data Science Notebooks	Does the product provide native support for common data science notebooks (e.g., Jupyter and Zeppelin)?
	Open-Source Visualization Tools	Does the product support the use of open-source data visualization tools?
	Open-Source Data Management Platforms (e.g., Spark and Hadoop)	How well can the product leverage open-source data management capabilities?
	Open-Source Automated Machine Learning Tools	Does the platform support the use of open-source automated machine learning tools (e.g., Auto-sklearn, TPOT)?
Performance and Scalability	Cloud	Does the product use cloud computing to speed up computationally intensive operations?
	In-Database Analytics	Does the product provide in-database analytics to speed up computationally intensive operations?
	Big Data Volume Scalability	Does the product provide functionality to handle datasets larger than server memory? Does the product scale to handle data volumes of at least 2TB with a satisfactory user experience?
	Real-Time Data and Streams	Is the product capable of event stream processing? Does the product make use of distributed stream processing engines?
	Distributing Computing	What options (e.g., Hadoop, Spark) does the product provide for distributed computing? (Please rate and also type your answer in the cell below.)
Model Management	Metadata Management	Does the product support metadata management? Does the product support definition and enforcement of referential integrity rules? Does the product use source and target metadata in the creation of source/target connections? Does the product use formatting information from external data sources?
	Traceability	Does the product support model traceability, versioning and lineage (including experiment history)?
	Champion/Challenger	Does the product support champion/challenger testing for models and potential rollback automation?
	Model Telemetry	Does the product offer model telemetry functions? Granular monitoring of models performance in production?
	Technical Performance Tracking	Does the product support model performance tracking and efficacy measurement as well as A/B testing automation (including model auditing and anomaly detection in production)?
	Business Performance Tracking	Does the product support ROI and multiple execution end-point engagement measurement?
	Retuning of Models	Does the product support retuning of models?
	Governance	Does the product support governance of model access and use?

ID: 382503

Sample from “Toolkit: RFP for Data Science and Machine Learning Platforms”

Source: Gartner (March 2019)

Evaluate Vendor and Platform Capabilities and Market Trends

With the preceding best practices as a foundation, data and analytics leaders are ready for specific and serious comparison shopping for DSML platforms. They can sit down with line-of-business leaders and users to evaluate potential platform candidates in light of the organization's use cases, requirements and desired outcomes. With today's rapid innovation and the changing vendor and platform landscape in the DSML market, it is also critical to keep up with technology trends in order to plan and prepare proactively.

Data and analytics leaders can leverage relevant Gartner research and tools for these processes, following these five steps:

- **Step 1. Magic Quadrant research —** Begin by referencing the "Magic Quadrant for Data Science and Machine Learning Platforms" to assess the key vendors and select a subset of key vendors' platforms. Be sure, however, to *not* rely solely on the Magic Quadrant graphic to choose the vendor most appropriate for you. Reference the vendor write-ups to understand the strengths and cautions for each of the vendors. In addition to providing information for each of the vendors, the Magic Quadrant also provides information about the DSML space and its trends, outlining how the market is responding. Finally, the Magic Quadrant also provides a list of capabilities that are required for platforms in this market. The Magic Quadrant research is a complement and companion to the subsequent additional research.
- **Step 2. Critical Capabilities research —** Next, consult the "Critical Capabilities for Data Science and Machine Learning Platforms" to compare and evaluate the capabilities of each of the vendors' offerings. Whereas the Magic Quadrant provides guidance at the vendor level, the critical capabilities report provides guidance at the platform level, offering a deep-dive analysis of the capabilities provided by each of the platforms. In addition to providing capability scoring, the critical capabilities report provides information on the common use cases for which these tools are employed. The scores for the platforms are weighted and stack ranked, providing guidance on which tools are appropriate for which use case. The use cases help to align to those tools that are best-suited to the type of work required; they also provide information about the typical ways the tools are used within an organization. Leverage the Magic Quadrant and critical capabilities research to hone down the platform choices to a shortlist for further, more detailed, consideration.
- **Step 3. Request for Proposal (RFP) Toolkits —** Leverage the "Toolkit: RFP for Data Science and Machine Learning Platforms" as you begin to evaluate, in detail, the tools on your shortlist. The RFP provides a detailed breakdown of not only the critical capabilities, but also related subcapabilities to consider within each of these categories. The RFP can be modified/customized to suit a specific organization's needs. It can be used as an RFP to send directly to vendors, or used internally as a POC script to guide an evaluation of the various tools under consideration. The RFP can also be used as a tool to guide conversations with the users — to determine what their needs are and what kind of capabilities and subcapabilities will be best-suited to meeting those needs.
- **Step 4. Cool Vendor research —** Be sure to reference Cool Vendor research to keep in touch with new developments in the market. This research outlines additional interesting vendors to

watch and consider. It identifies emerging vendors in various clusters as well as groundbreaking innovation in this market. Cool Vendors reports relevant to this market include "Cool Vendors in Data Science and Machine Learning" and "Cool Vendors in Analytics." Leverage this research not only to understand newly developing technology, but also to identify new trends and approaches that could be leveraged within your own organization.

- **Step 5. Hype Cycle research —** Monitor the "Hype Cycle for Data Science and Machine Learning, 2018" to identify and plan for trends in the DSML market. The Hype Cycle not only identifies and defines key trends, but also explains the business impact and identifies a benefit rating, market penetration and maturity level for each trend. Use the Hype Cycle to understand the technologies and map those to your organization's needs and its appetite for being at the cutting edge.

This approach and resources enable data and analytics leaders, and relevant stakeholders, to continuously track, assess and respond to the rapidly changing market for DSML platforms.

During the selection process, and continuing once a platform is selected, keep in mind that the DSML platform will only be as good as the people who use it. Matching the platform to the skills available is critical. In addition, ongoing efforts should be made to develop a user base that is not only technologically competent, but also data literate and familiar with DSML techniques and approaches. Moving beyond teaching people to simply use the platform — to actually applying it within the business context to add organizational value — requires ongoing attention and direction.

Invest in Hiring and Skills Building for the Data Science Team

Building the necessary staff and expertise for data science will require long-term planning and investment by data and analytics leaders. A DSML platform will only be as good as the people who deploy, use and support it. That means data and analytics leaders should inventory the specific roles and the skills needed for implementing and sustaining these platforms.

Data Science and Machine Learning Roles

There are at least six key data science roles. Not all organizations will have, can afford or will need all these roles. A small core team can serve as the nucleus for a larger team over time. The six key roles are:

- **Senior data scientists —** The evangelists for the DSML technology activities and strategy of an organization. They support the building and validation of DSML models. They act as mentors and role models for junior data scientists and citizen data scientists. They are often tasked with being cross-team collaborators and business language translators.
- **Junior data scientists —** The primary "go to" staff for conducting data exploration and building machine learning models (depending on the size and budget of the organization). They work closely with, and are mentored and guided by, a more senior data scientist. They often work with, or even in, business units.

- **Data science/machine learning operations —** These personnel are responsible for deploying and managing analytical assets within operational processes in a repeatable, manageable, secure and traceable manner.
- **Enterprise architects —** Those who focus on how DSML platforms should be implemented within the overall technology architecture of an organization.
- **Data engineers —** The data pipeline creators who make the appropriate data accessible and available for data scientists and beyond, and ensure the alignment between operational data and required or desirable model training/validation data.
- **Citizen data scientists —** Business users who extract predictive and prescriptive insights from data. This does not require them to be as skilled and technically sophisticated as expert data scientists. This role has evolved as a natural extension from other roles within the organization. Besides actively supporting this role, data and analytics leaders should also identify and cultivate other business users whose analytics activities position them for further development as citizen data scientists (see "Maximize the Value of Your Data Science Efforts by Empowering Citizen Data Scientists").

Don't rely on the lone "superhero" data scientist. Evaluating, deploying, using and managing DSML requires a team with specific and expert data and analytics skills and organizational and business acumen — beyond what a single person can offer. DSML platforms are one component of a more holistic analytics program and an even broader data management approach (see "Staffing Data Science Teams: Map Capabilities to Key Roles").

Data and analytic leaders must plan for developing and supporting these roles. Also, for establishing the processes that enable the roles to work easily and collaboratively together in building and supporting a common, holistic data and analytics program and supporting a robust, consistent architecture across all the analytic capabilities.

Data Science and Machine Learning Skills

New hires are one source of the necessary skills and expertise to create a viable DSML capability, but qualified candidates are in short supply. Data and analytics leaders should therefore foster the development of skills in the following four ways:

- Create or leverage training programs for IT and business staff as part of a new, internal professional development path in DSML. There are many online courses available for both general and specific DSML training, including Udemy, edX, Coursera and others.
- Assess and acquire new software tools, collectively referred to as augmented data science or AutoML, that automate repetitive or routine — but vital — DSML tasks, and IT support tasks.
- Upskill motivated business users to become more effective citizen data scientists so they can directly leverage these skills in their day-to-day work.
- Leverage consulting and service providers to jump-start analytic initiatives. These groups can be used to fill an existing analytic deficit as well as to augment existing data scientists with

specific skills. Refer to the "Market Guide for Data Science and Machine Learning Service Providers" for candidate service providers.

Acronym Key and Glossary Terms

AI	artificial intelligence
DSML	data science and machine learning
ML	machine learning
POC	proof of concept

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

"Magic Quadrant for Data Science and Machine Learning Platforms"

"Critical Capabilities for Data Science and Machine Learning Platforms"

"Toolkit: RFP for Data Science and Machine Learning Platforms"

"Cool Vendors in Data Science and Machine Learning"

"Cool Vendors in Analytics, 2018"

"Hype Cycle for Data Science and Machine Learning, 2018"

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp

This research note is restricted to the personal use of chongfeng@airchina.com.