# Critical Capabilities for Data Science and Machine Learning Platforms

By Pieter den Hamer, Shubhangi Vashisth, Erick Brethenoux, Afraz Jaffri, Peter Krensky, Farhan Choudhary, Carlie Idoine

The functions and features of data science and machine learning platforms are evolving quickly to keep pace with a highly innovative space. This research helps data and analytics leaders to evaluate 20 of these platforms across 15 critical capabilities.

## Overview

### Key Findings

- Business and data exploration are an integral part of many platforms, offering a common ground for the vital collaboration between (citizen) data scientists and domain experts.
- Citizen data science is now supported by a majority of vendors, bringing the power of machine learning within reach of nonexperts.
- Expert model development often requires cutting-edge (open-source) techniques, for which many vendor platforms offer first-class support.
- Operationalization is more critical than ever as organizations seek to scale their data science and machine learning, demanding tangible business value from projects.

### Recommendations

Data and analytics leaders tasked with incorporating data science and machine learning into their analytics strategies should:

- Optimize platform selection by working with data science teams and business units to identify the most significant challenges, needs and impacts.
- Choose the best-fit platform by balancing the desired mix of use cases, the available user skill sets, the deployment environment and the prioritized strength of critical capabilities.

- Prioritize open platforms that can be extended with existing or future (open-source) innovations to support, for example, explainable AI and end-to-end augmentation — from data exploration to operationalization.
- Explore the need for multipersona support through enabling productivity and collaboration for a growing group of user types, including expert data scientists, citizen data scientists, data engineers, application developers and business analysts.

# Strategic Planning Assumptions

By 2023, 30% of organizations will harness the collective intelligence of their analytics communities, outperforming competitors that rely solely on centralized analytics or self-service.

By 2023, 60% of organizations will compose components from three or more analytics solutions to build business applications infused with analytics that connect insights to actions.

By 2024, 70% of enterprises will use cloud and cloud-based AI infrastructure to operationalize AI, thereby significantly alleviating concerns about integration and upscaling.

By 2024, use of synthetic data and transfer learning will halve the volume of real data needed for machine learning.

# What You Need to Know

End-user organizations report a wide range of data science projects incorporating four major use cases:

- Business and data exploration
- Citizen data science
- Expert model development
- Operationalization

These diverse operations often drive the selection of specific vendors, platforms and technologies.

The platforms evaluated here by Gartner have been selected from a wide range of approximately 80 competitors. They offer a cohesive set of building blocks for developing the data science and machine learning (DSML) solutions demanded by current data science market conditions. Each platform can deliver data science solutions into business processes, surrounding infrastructure, applications and products, with no single platform being best-of-breed across all capabilities. These platforms allow data science teams to build custom solutions, rather than exclusively buy packaged applications or outsource all data science work to a service provider.

This Critical Capabilities research evaluates a heterogeneous mix of data science and machine learning platform vendors, which reflects the different user interfaces and tools that users with various levels of skill prefer.

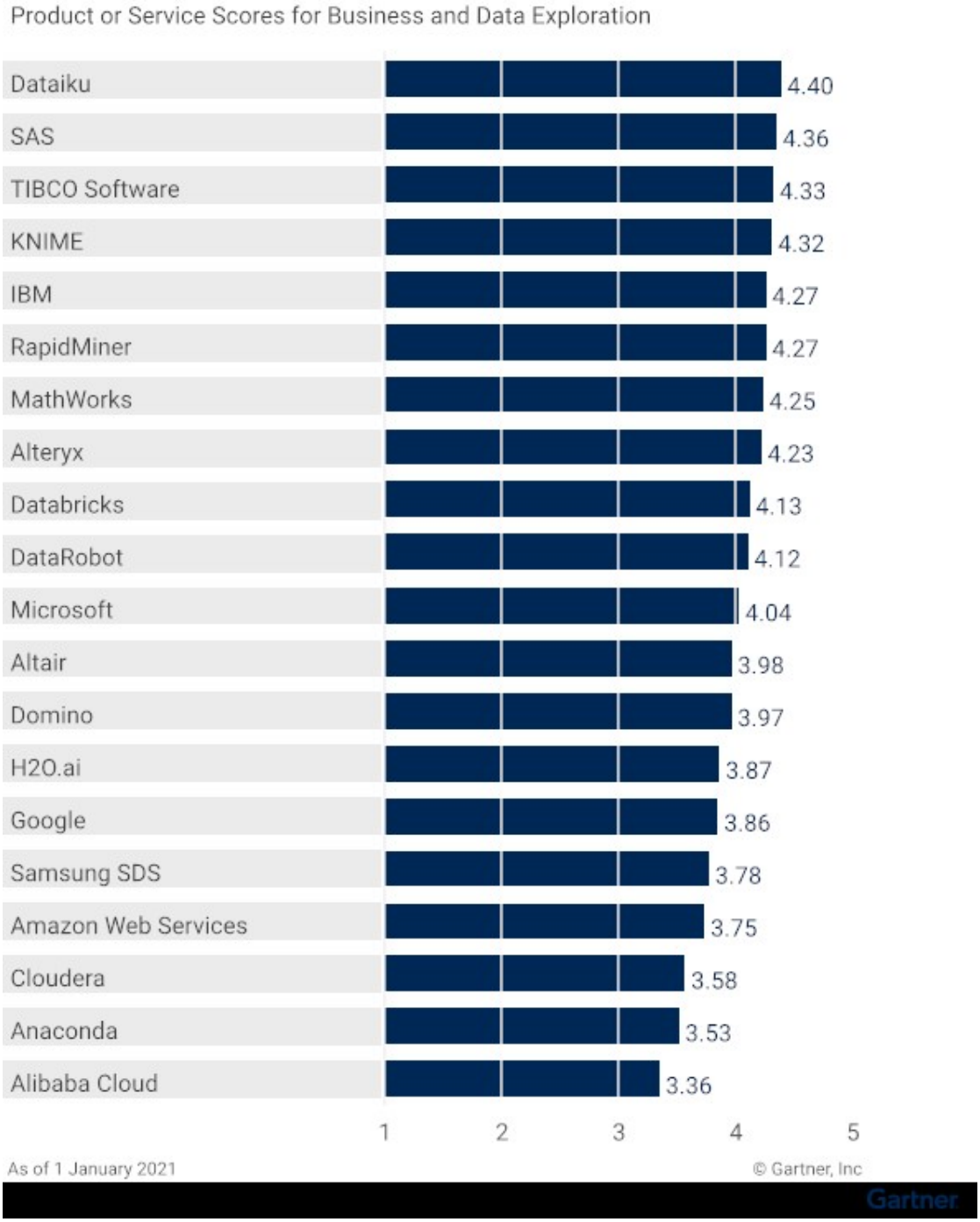
This is a companion piece to Gartner's 2021 Magic Quadrant for Data Science and Machine Learning Platforms, which compares vendors in terms of completeness of vision and ability to execute. This document evaluates vendor products against a specific set of critical capabilities and use-case scenarios.

Gartner's research methodology excludes evaluation of pure open-source platforms with no vendor-supported commercial offering (such as R and Python), but they remain common (initial) choices for many data science teams.

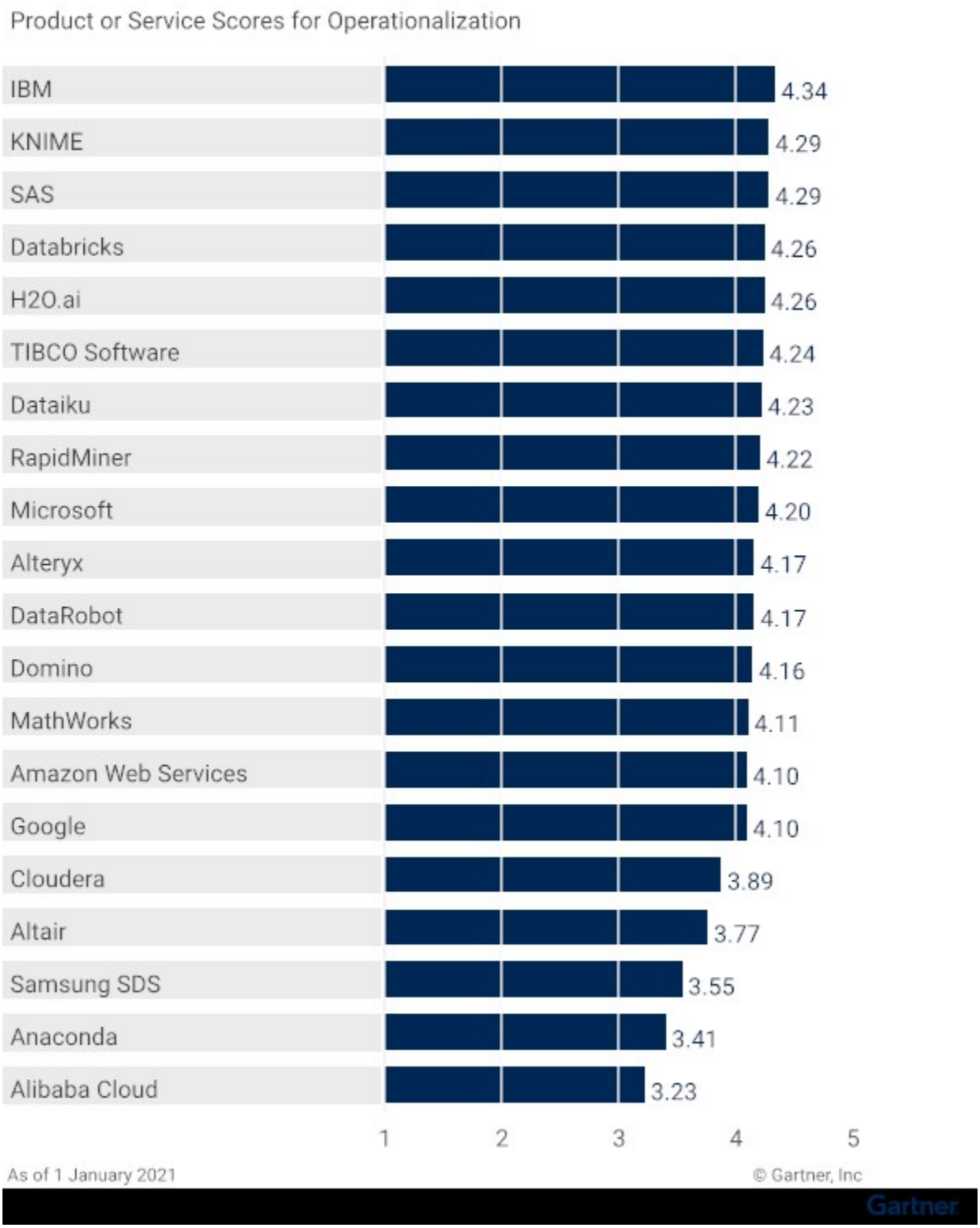
# Analysis

## Critical Capabilities Use-Case Graphics

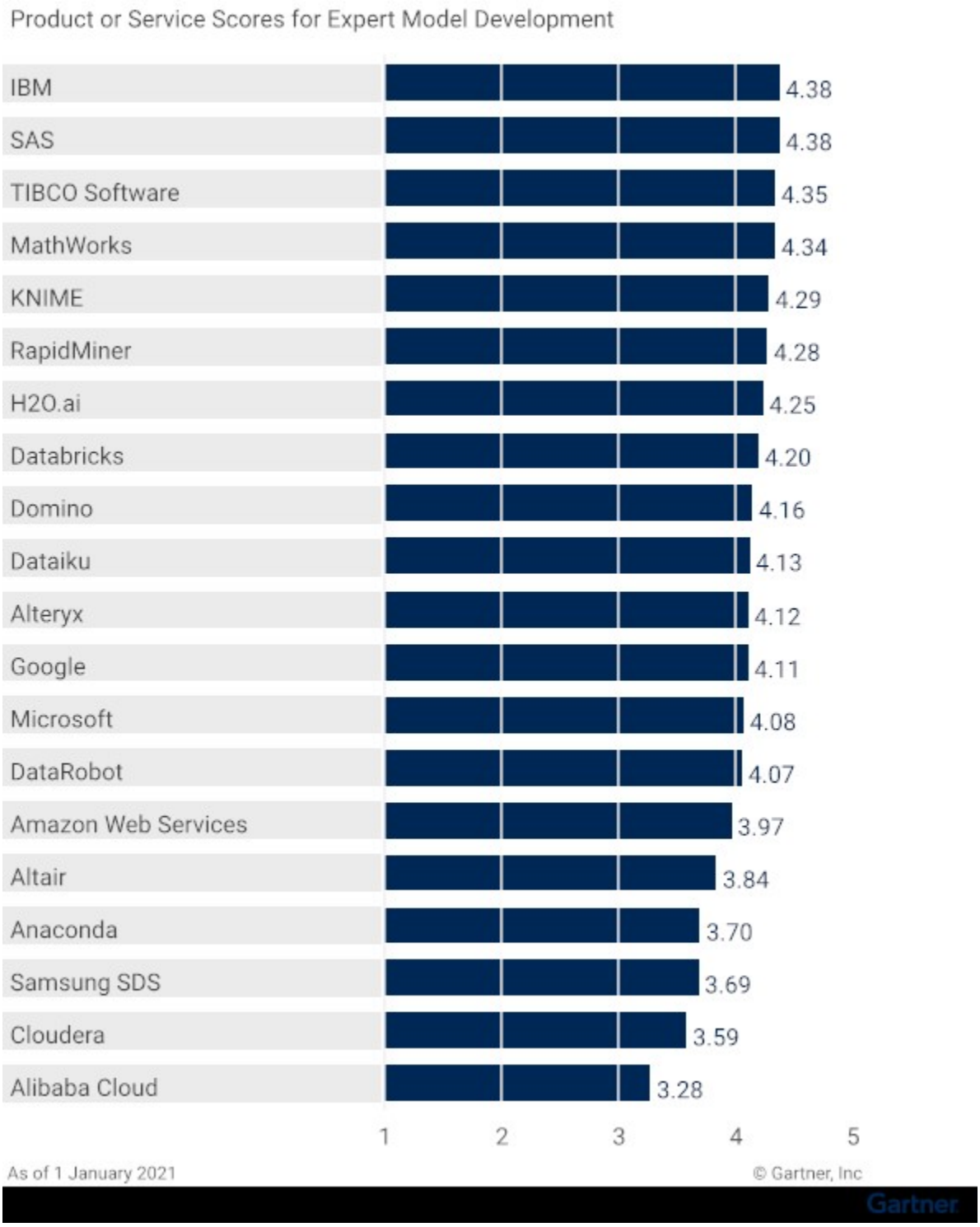
### Vendors' Product Scores for Business and Data Exploration Use Case



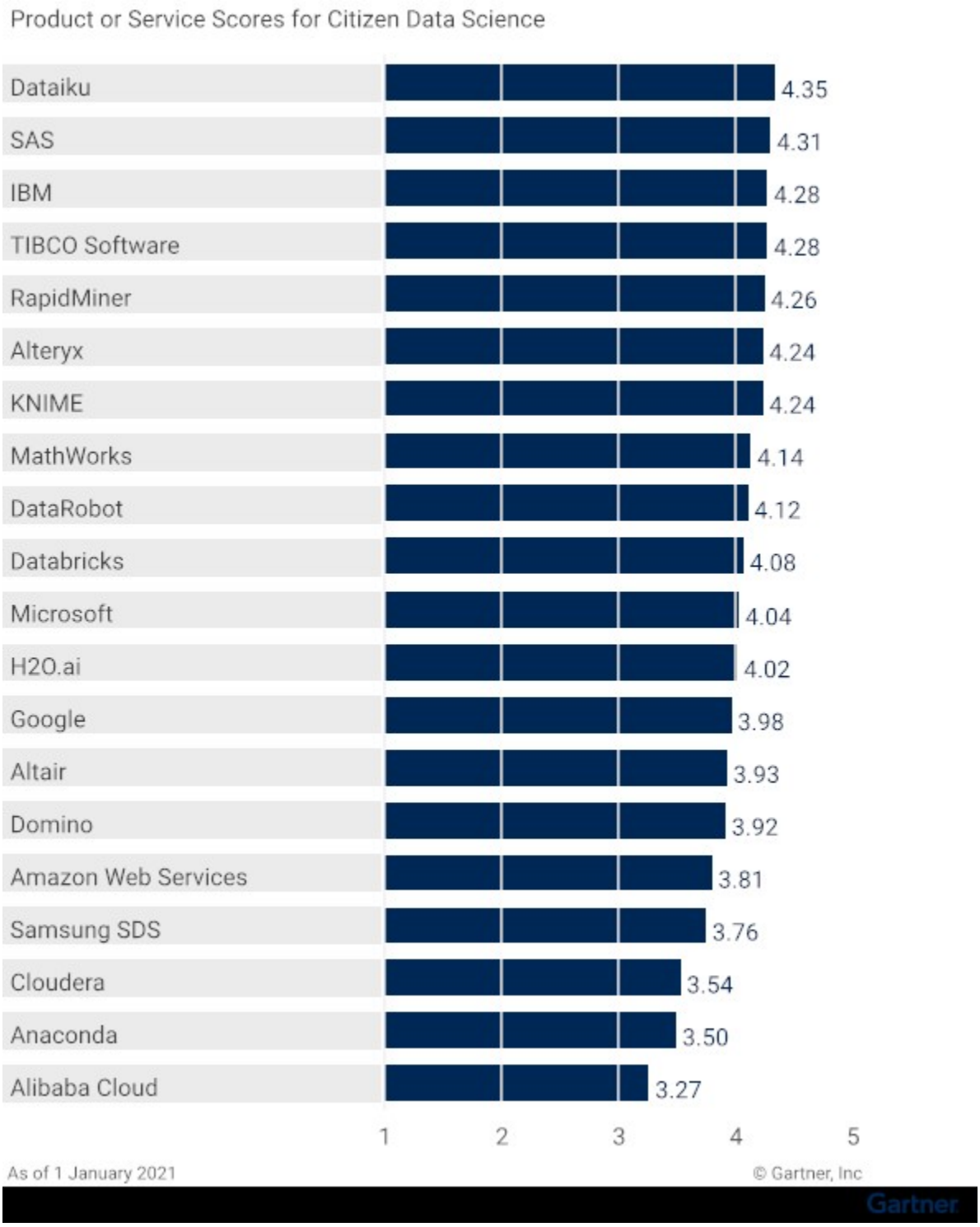
## Vendors’ Product Scores for Operationalization Use Case



## Vendors’ Product Scores for Expert Model Development Use Case



## Vendors’ Product Scores for Citizen Data Science Use Case



# Vendors

### Alibaba Cloud

The core components of Alibaba Cloud's data science and machine learning offering are the Platform for AI (PAI) Studio and Data Science Workshop (DSW). These are supported by a portfolio of other products: DataWorks, MaxCompute and AutoLearning. Alibaba does not provide customization and configuration capabilities for machine learning (ML) in hybrid or multicloud setups, but does allow for deployment across on-premises and the Alibaba cloud. Alibaba Cloud is primarily meant for expert data scientists and data engineers, with limited functionalities offered to citizen data scientists. Alibaba's PAI and DSW customers most often also considered Megvii, SenseTime, H2O.ai, Microsoft Azure, Google Cloud Platform and Amazon Web Services (AWS). Customers commonly choose PAI and DSW for advanced use-case implementations in image recognition, recommendation engines, natural language understanding and translation.

The Alibaba platform receives a strong score for coherence between the products that make up the platform. It stands out in real-time processing of large data volumes, with native cloud deep learning and cost optimization modeling capabilities. Users can leverage the respective strengths of DataWorks, MaxCompute and AutoLearning with a consistent look and feel throughout each component.

Scores for other advanced analytics and machine learning capabilities were lower due to limited or no support for composite artificial intelligence (AI), small data, generative adversarial networks (GANs), federated learning, simulation and heuristic approaches.

The platform received average scores on flexibility and openness and data access. Data preparation scored low; although augmentation and labeling are supported, it lacks some other functionalities such as data watermarking and data enrichment. Although Alibaba offers jump-start kits and solution templates for some specific use cases — such as image segmentation or recommendation systems — it can significantly improve its capabilities on providing precanned solutions.

Alibaba Cloud is strongest for expert model development and business and data exploration use cases.

### Altair

Altair offers Knowledge Studio and its supporting portfolio known as Knowledge Works. The Knowledge Works suite includes Knowledge Studio for Apache Spark, Knowledge Hub, Knowledge Manager, Panopticon and Monarch. Knowledge Studio supports delivery on-premises. Knowledge Studio for Apache Spark supports delivery in pure cloud. The product does not support hybrid integration. Knowledge Studio's users are a mix of expert data scientists and citizen data scientists (primarily business analysts). Altair customers most often also considered SAS, MathWorks and TIBCO Software. Customers commonly choose Knowledge Studio for ease of use, product functionality and total cost of ownership (TCO). Customers also recognize the strength of the product's decision tree and strategy tree features.

An intuitive and easy-to-use interface and an open and extensible architecture make Knowledge Studio appealing to both experts and citizen data scientists. Knowledge Works supports popular open-source tools, such as Python, R, Keras, MLFlow and scikit-learn. The platform has strong scores for data access, data preparation, and data exploration and visualization. The product also received good scores for platform and project management and explainable AI. It provides a number of explainability techniques including individual conditional expectation (ICE) and partial dependence plots, Local Interpretable Model-agnostic Explanations (LIME), decision trees that can be used as global surrogate models and variable importance. It integrates with MLflow for model deployment and can generate model code in Python, R, SAS, SQL, PMML and other programming languages for manual operations if desired.

Knowledge Studio scored average for machine learning and model management. Delivery and performance and scalability continue to be areas for Knowledge Studio to improve. Graphics processing unit (GPU) support is limited to the AutoML node in Knowledge Works that can be run on GPU using the NVIDIA RAPIDS.ai library. The product offers good flexibility in integrating with open-source tools and platforms; however, native support for Kubernetes/Kubeflow will be a desirable addition for Knowledge Studio users.

The platform scores well for business and data exploration and citizen data science use cases, with lower scores for expert model development and operationalization.

**Alteryx**

Alteryx has revamped its offering this year, introducing Analytic Process Automation (APA), which provides building blocks to automate the analytics process end to end. The platform includes Alteryx Designer, Alteryx Intelligence Suite and Alteryx Server. Alteryx Connect and Alteryx Promote are available as Server add-ons. Alteryx has also introduced Alteryx Analytics Hub, a new central environment that includes workflow automation/scheduling, collaboration, multitenancy and data connection management. Alteryx can be deployed on-premises, in the cloud, or hybrid cloud and reuse on-premises, and supports connecting to data from multiple cloud sources, on-premises and hybrid scenarios. It is most typically deployed on-premises. Alteryx is used by both citizen and expert data scientists, but is most typically leveraged by business analysts. Alteryx customers often also consider Microsoft, KNIME, SAS and DataRobot during their selection process. Customers typically choose Alteryx for its intuitive and easy-to-use user interface, as well as its strong data access and preparation capabilities.

Alteryx ranked very high for data access, data preparation, user interface and delivery. Flexibility and openness, performance (thanks to an improved high-performance engine), scalability, and platform and project management also scored high. Notably, Alteryx has developed and incorporated three open-source contributions, including Featuretools for automated feature engineering, EvalML for optimizing ML pipelines and Compose for automated prediction engineering. Model management, collaboration and coherence also are strong.

Data exploration and visualization as well as machine learning scored average. Precanned solutions also received an average score. It should be noted that Alteryx has

expanded its functionality with natural language processing and optical character recognition capabilities integrated within its automated ML capabilities, and with integration of robotic process automation (RPA) and process intelligence solutions.

Alteryx is particularly well-suited for the business and data exploration and citizen data science use cases.

### Amazon Web Services

The Amazon SageMaker platform is AWS's offering for end-to-end data science, supported by multiple components and services. These include the SageMaker Studio integrated development environment (which includes Autopilot, Notebooks, Model Monitor, Experiments, Debugger and Model Tuning), Amazon EMR (including S3), AWS Glue, Amazon Kinesis, Amazon SageMaker Neo, Amazon SageMaker Ground Truth, Amazon SageMaker Clarify, Amazon SageMaker Data Wrangler, Amazon SageMaker Pipelines, AmazonQuickSight, Amazon CloudWatch and AWS CloudTrail. The majority of Amazon SageMaker customers are in pure cloud environments, with some hybrid deployments. Amazon SageMaker's users are primarily developers, expert data scientists, and ML engineers and specialists. Customers considering Amazon SageMaker most often also consider Google, Microsoft and Databricks. Customers choose Amazon SageMaker for alignment with existing data and analytics investments, tight integration with the broader AWS stack, support for open source, and excellent performance and scalability.

Amazon SageMaker received a top score for performance and scalability. The platform supports myriad hardware options optimized for various ML and deep-learning frameworks, and features a pay-as-you-go pricing model with no minimum fees or upfront commitment, encouraging experimentation. The platform's precanned solutions are excellent, with Amazon SageMaker JumpStart supporting rapid solutions for numerous use cases. The flexibility and openness score is high, with the philosophy of ongoing OSS curation and optimization that expert data science teams expect. Amazon SageMaker's machine learning score is strong, with broad support for cutting-edge ML and significant improvements in augmented DSML capabilities via Autopilot. The model management score is high, but many different components are required to support the full range of capabilities, including but not limited to Amazon SageMaker Model Monitor, Amazon SageMaker Model Registry, Amazon SageMaker Pipelines, Amazon SageMaker Feature Store, Amazon CloudWatch and Amazon Cloud Trail.

Amazon SageMaker's user interface needs improvement, including within notebooks, Autopilot and Data Wrangler. The platform is more popular among coders and not as intuitive for nontechnical users when compared with leading tools for citizen data scientists. Coherence scores are just average, as AWS's approach of modularized components adds significant integration complexity and impacts the user experience. Data Wrangler and other components within the SageMaker Studio IDE offer significant improvements in data preparation capabilities, but the overall score is still average. Collaboration capabilities need improvement, wanting stronger discussion threads, better multipersona support and more integration options with collaboration tools.

Amazon SageMaker and its supporting portfolio are most commonly leveraged for the expert model development and operationalization use cases.

### Anaconda

Anaconda markets Anaconda Enterprise, a data science platform in which users can leverage open-source Python and R-based packages using an interactive notebook concept. Anaconda Enterprise can be deployed on-premises or in hybrid and (multi)cloud environments. Anaconda's users are principally Python and R enthusiasts starting their data science journey, or expert data scientists, developers and statisticians who do not prioritize features such as user interface, automated machine learning (autoML), MLOps/ModelOps and explainability. Buyers of Anaconda often also consider Databricks and Cloudera. Customers often choose Anaconda because of its flexibility and openness for working with open-source technologies at a relatively lower cost as compared with its competitors.

Anaconda has been catering to the open-source community by quickly adopting innovations and by addressing needs rising from the community. Anaconda exhibits strengths in flexibility and openness, machine learning, other advanced analytics, and collaboration.

Anaconda received average scores in data exploration and visualization, performance and scalability, delivery, and coherence. Anaconda needs to improve its capabilities around data access, data preparation, platform and project management, and precanned solutions. The product also would also benefit from better model management and explainable AI. Since the constituency of users working on the Anaconda platform are data scientists who prefer coding in R or Python, this leaves significant adoption challenges for other user personas, such as data engineers or citizen data scientists.

Expert model development is the use case that is best supported by Anaconda. The platform is not well-suited for citizen data science or business and data exploration.

### Cloudera

Cloudera has a core machine learning product, Cloudera Machine Learning (CML), supported by Cloudera Data Engineering (CDE) and Cloudera Data Visualization (CDV). These products are interconnected and delivered as services offered on top of the Cloudera Data Platform (CDP). CDP can be installed in private, public, multicloud and hybrid cloud environments. All deployment can be managed within a single control plane. Cloudera is typically used by customers that have specialist skills in data engineering, data science and ML engineers, or infrastructure professionals responsible for developing services for deploying and monitoring models in production. Cloudera is most often considered alongside Databricks, IBM and Microsoft. Cloudera is chosen by customers that require a single flexible platform for data storage, management and machine learning, and extensively utilize containers and API-driven architecture.

Cloudera received excellent scores for performance and scalability and strong scores for flexibility and openness. CDP as a platform achieves most value for use cases with

large data volumes and complex data processing tasks, which need to be orchestrated, monitored and governed. CML complements this approach by containerizing data science and machine learning operations. Containerization together with an SDK and extensive APIs to control and interact with all processes in the system, from data ingestion to model monitoring, make it suitable for integrating with operations where infrastructure as code is heavily utilized.

CML received low scores for data preparation, precanned solutions and other advanced analytics. There is a visual interface for basic data model editing, but data preparation needs to be done through code. The same is true for building ML and other types of models where augmentation is weak or only supported by integration with H2O.ai. The visual data exploration and dashboard environment, Cloudera Visual Apps, has recently been added to provide business intelligence (BI) capabilities from prebuilt datasets.

Cloudera is strongest in the operationalization use case.

**Databricks**

Databricks Unified Data Platform spans data science, ML analytics and data engineering capabilities. Databricks is a formidable cloud player, and its offering is available on Azure, AWS, Alibaba Cloud and Google Cloud Platform (GCP). Multicloud model deployment is supported through MLflow. Although Databricks' main technical and engineering efforts are focused on further platform advancements in the cloud, the platform is available on-premises through Kubernetes or from the company's OEM partner Booz Allen Hamilton, as part of its Open Data Platform offering. Users of the platform are primarily expert data scientists and data engineers. When choosing Databricks, customers also considered Microsoft, Amazon and Google. Customers choose Databricks for its platform's ability to unify data types and data workloads, scalability, performance, focus on user empowerment through quick access to open-source innovations, vibrant community, and access to an industry-leading customer success organization.

Databricks Unified Data Platform has received outstanding scores for performance and scalability and for flexibility and openness. It is known for its easy and scalable cluster management that allows customers to start experimenting on smaller clusters and then scale up quickly. Moreover, the platform is attractive to customers that want to keep up with the latest open-source innovations. Databricks also scores strongly on platform/project management, model management and delivery. The vendor leverages its leadership in the development of MLflow, an increasingly popular open-source framework for ML operations, offering a model registry and reproducible workflows. Building on the Delta Lake open-source initiative, the platform scores well on data access and data preparation. High scores were also given for machine learning, other advanced analytics and precanned solutions.

Data scientists and data engineers need a strong coding background to fully exploit the proprietary notebook interface — including the recently added SQL analytics capabilities targeting data analysts — and embedded ML frameworks. Databricks does not offer an easy-to-use interface for nontechnical users or citizen data scientists. Explainable AI

also scores average with embedded open-source techniques for bias mitigation and explainability, but offers limited support for governance.

Expert model development and operationalization are the most popular use cases for Databricks Unified Data Platform.

**Dataiku**

Dataiku provides its product, Data Science Studio (DSS), for all data science tasks and stages of the machine learning life cycle. DSS is available in 5 editions with varying levels of functionality. DSS can be installed on-premises or in cloud environments via virtual machines or containers. It can manage multiple instances across environments, including multicloud settings. Dataiku is used by a mix of users including expert data scientists, citizen data scientists, business analysts, machine learning engineers, data engineers, business leads and product owners. Dataiku is considered alongside vendors like Alteryx, SAS, DataRobot, KNIME, H2O and Rapid Miner, but also vendors like Databricks and Cloudera. Customers choose DSS for its ease of use, collaboration features and ability to combine both notebook and code-based development with no-code model building workflows.

DSS achieved outstanding scores for data access, data preparation, user interface, collaboration and coherence. The number of data sources that can be handled is increasing rapidly and provides a good mix across databases, business applications and nontraditional data sources, such as graph and time series. The data preparation capability is among the best for utilizing a context-driven strategy for providing guidance on suitability for usage in model development. The ability to annotate image data is also a recently added feature. DSS, being a single platform, is easy to navigate and exposes balanced functionality for coders and no/low-code users.

DSS has an average score for precanned solutions and, similar to last year, still needs improvement to support non-ML data science techniques and models beyond open-source libraries and notebooks. This includes optimization, simulation and agent-based modeling. In addition, supported techniques need to be included in the platform as a standard, rather than as plug-ins. Precanned solutions are available from Dataiku's public gallery, but are provided more for exploratory and learning purposes than as ready-made solutions.

DSS is strong across all four use cases.

**DataRobot**

DataRobot is an augmented data science and machine learning platform that incorporates automation across the end-to-end analytics process from data access through model building, operationalization and delivery of applications. The DataRobot Enterprise AI platform consists of Paxata Data Prep, Automated Machine Learning, Automated Time Series, MLOps and AI applications. The augmented approach enables both citizen and expert data scientists to productively use data science. DataRobot provides data preprocessing, machine learning, performance and scalability, delivery and model management on-premises, via virtual private cloud (fully managed or multicloud), or via hybrid cloud and on-premises. Customers that selected DataRobot

most often also evaluated Alteryx, H2O.ai, Dataiku, Microsoft, AWS and Google. Customers call out user-friendliness, flexibility and support across the full analytic workflow as key strengths. Strong, highly-skilled support services are also often a factor in DataRobot's selection.

DataRobot has quickly and successfully incorporated its acquisitions into its end-to-end augmented offering. DataRobot's capabilities for delivery are outstanding. Data access and preparation ranked high, as did model management, platform/project management, and flexibility and openness. Scores for machine learning, performance and scalability were good. Collaboration, coherence and user interface, important capabilities especially for environments supporting both citizen and expert data scientists, were also strong.

DataRobot scored low for precanned solutions. The Pathfinder use-case library provides an easily accessible platform for referencing use cases, but a breadth of available use cases across areas such as back-office analytics, IT operations and cybersecurity is lacking. Other advanced analytics scored fair, most notably lacking the ability to perform decision modeling. DataRobot's data exploration and visualization and explainable AI capabilities are average.

Considering all capabilities together, DataRobot scores consistently well across all use cases.

**Domino**

The Domino Data Science Platform is an industrial-strength solution for enterprise data science and is now complemented by Domino Model Monitor. These two products provide end-to-end DSML capabilities with an emphasis on centralization, flexibility, collaboration, reproducibility, and oversight of model development and deployment. Domino supports deployments in multiple clouds, on-premises and anywhere via Kubernetes. Domino is implemented in many federated organizations at high scale. Most deployments are in the cloud, but on-premises and hybrid deployments are also common. Domino's typical user mix includes predominantly expert data scientists, chief/managing data scientists and ML engineers. Business analysts, data engineers and developers are present in smaller numbers. Domino customers often also consider AWS, Microsoft, Databricks, Dataiku and Cloudera. Customers typically choose Domino for its high flexibility, support of open-source and commercial technologies such as SAS and MATLAB, cost controls for IT, management features for large teams, as well as speed of model development and ability to build and manage large numbers of models.

The Domino platform received a top score for flexibility and openness. Domino's collaboration capabilities are excellent for supporting large teams of experts and enterprise data science. Domino's performance and scalability is strong, and it offers on-demand Spark clusters to unite Spark and non-Spark workloads for data preparation, model training and scheduled jobs in a single platform. Domino garnered an excellent score for platform/project management, and the platform remains attractive in regulated environments. Domino's high delivery and model management scores reflect the value of the Model Monitor product. The platform is among the best options

for supporting complex on-premises or for hybrid and multicloud model development and deployment.

Domino's user interface is designed for a code-first data scientist and not designed for the average citizen data scientist or business user. Domino does not offer a visual composition framework. Domino's precanned solutions were identified as needing improvement last year, and there has been minimal change in this functionality. Data preparation capabilities are fair, and additional drag-and-drop data blending tools may be necessary. Data exploration and visualization capabilities are market-standard and largely reliant on natively integrated environments, such as Jupyter and RStudio. Other advanced analytics are average and also mostly reliant on the underlying environment, with little support for native decision modeling or decision management.

Domino is more heavily leveraged for the expert model development and operationalization use cases.

**Google**

For this report, Gartner evaluated the Google Cloud AI Platform, a suite of components that includes Cloud Data Fusion, Cloud AutoML, BigQuery ML, AI Platform Notebooks and TensorFlow. Google will launch its unified AI Platform in 1Q21 (after the cutoff for inclusion in this Critical Capabilities evaluation). Key features and services planned for release with this new platform vision include AutoML tables, Explainable AI (XAI), AI platform pipelines and other MLOps services. The majority of Google Cloud AI Platform customers are in pure cloud environments, with some hybrid deployments. Google's users are primarily expert data scientists, ML engineers/specialists and developers. Customers considering Google Cloud most often also consider other cloud-focused platforms such as those from Amazon Web Services, Microsoft and Databricks. Customers choose Google's AI Platform for alignment with existing data and analytics investments, tight integration with the broader GCP stack, and excellent performance and scalability.

The current platform once again earned an outstanding score for performance and scalability, with robust support for a range of hardware options and ML frameworks as well as specialized software and hardware like TensorFlow Enterprise and Tensor Processing Units (TPUs). Machine learning scores are already strong, with additional augmentation capabilities coming with the release of AutoML Tables. The platform continues to score highly in other advanced analytics (especially vision, text and audio processing), and the flexibility and openness score is excellent. Support for explainable AI within the platform is strong and bolstered by Google's thought leadership in responsible AI.

The current AI platform's coherence is fair and needs improvement. Model management and platform/project management capabilities are currently average. Data access, data preparation, data exploration and visualization, and user interface scores are all average. Collaboration capabilities are fair, with minimal support for discussion threads or ratings and recommendations.

Google Cloud AI Platform and its supporting portfolio are most commonly leveraged for the expert model development and operationalization use cases.

**H2O.ai**

H2O.ai offers a commercial product called H2O Driverless AI with additional modules MLOps, Puddle and AutoDoc. H2O also offers open-source products with optional enterprise support: the H2O 3 platform and AutoML for machine learning, Sparkling Water for Spark integration, and Wave for app development. H2O's products can be deployed on-premises and on multiple public and private cloud platforms, with MLOps for model operations and Enterprise Puddle for cloud operations. Driverless AI is mostly used by citizen data scientists and expert data scientists; the latter also use H2O 3. Customers considering H2O Driverless AI most often also consider DataRobot, Dataiku, Alteryx, Microsoft, AWS and Google. Customers chose H2O Driverless AI for its strong augmentation and automation capabilities across the ML life cycle. In addition, Driverless AI and H2O 3 are chosen for their high-performance machine learning and extensibility.

The platform received an excellent score for machine learning, with its augmentation and automation capabilities particularly standing out. These are not limited to model selection and hyperparameter tuning, but also include sophisticated automation for time series, natural language and image processing. For building AI apps, H2O's platform offers excellent flexibility, openness, performance and scalability. The same is true for delivery, among others through its Wave product to build AI apps. Explainable AI is another area where H2O is strong; it offers multiple explainability capabilities throughout the ML life cycle, in both modeling and feature engineering. Platform, project and model management (through the MLOps module) in Driverless AI are good, as are data exploration and visualization.

Capabilities for data access and data preparation need improvement. H2O's capabilities in other advanced analytics such as simulation, optimization and decision management received average scores, as did the platform's user interface, cohesion and collaboration support. And although improved in the last year, the number of precanned solutions still lags behind most leading vendors.

The platform offers strong support for the operationalization use case.

**IBM**

IBM Watson Studio on IBM Cloud Pak for Data is an open, modular platform for data and AI that combines a broad set of descriptive, diagnostic, predictive and prescriptive capabilities. The platform includes several open-source notebooks and frameworks and can work together with products like IBM Watson Knowledge Catalog, IBM Decision Optimization, IBM OpenPages with Watson for model risk governance and others. IBM Watson Studio Premium includes the optimization capabilities of IBM ILOG CPLEX. IBM Watson Studio is available on multiple clouds, including IBM's own cloud, as well as on-premises. The product has a continuous release cycle. On-premises releases are generally quarterly, while cloud offerings may occur more frequently. IBM Watson Studio users are a mix of expert and citizen data scientists, data engineers, and experts

in natural language processing, optimization or other supported AI techniques. It offers good support across all use cases. IBM customers typically also consider Amazon Web Services, Google, Microsoft and SAS. Customers commonly choose IBM for its comprehensive analytical offering for a variety of data types, its vision, and support for decision intelligence, multipersona, open source and governance, or because of alignment with existing data and analytics investments.

Leveraging its integration with IBM Cloud Pak for Data, IBM Watson Studio offers excellent support for data access. In addition to its outstanding openness and flexibility, the platform scores very high on performance, scalability and delivery, among others, because of its ML acceleration, GPU optimization and highly scalable stream processing. The machine learning score is excellent, driven by AutoAI and innovations like federated learning, explainable AI and multipersona collaboration.

The platform's capabilities for data preparation, exploration and visualization are strong. Its user interface is also good, with support for both graphical workflow (canvas) and notebook development and even natural language assistance for optimization. In addition, the platform has high scores on other advanced analytics, which include decision management and modeling, simulation, optimization, and a diversity of natural language, image/video and other analysis services. Taken together, Watson Studio offers a coherent set of generic services with a rich collection of precanned solutions. This is supported by good platform, project and model management capabilities, enabling governance and reuse of data and models through a catalog and user community exchanges.

IBM Watson Studio is strong across all four use cases.

**KNIME**

KNIME's core product is the KNIME Analytics Platform, an open-source product for the creation of DSML workflows. This product is complemented with a commercial KNIME Server product, available in small, medium and large variants, which enables collaboration, API access and options for model deployment. KNIME offers a mix-and-match approach for deployment supporting cloud-only, multicloud and hybrid deployments offering pay-as-you-go licensing as well as prepaid licenses on both Azure and AWS environments. KNIME is used by both expert data scientists and citizen data scientists to create workflows and pipelines for model training and deployment. Business users can interact with models through web applications deployed on KNIME server. KNIME buyers typically also consider other vendors such as SAS, RapidMiner, H2O.ai and IBM. KNIME is often chosen for the broad range of data science tasks it supports, as well as low TCO.

KNIME received outstanding scores for data access and flexibility and openness. The platform includes a large number of connectors for accessing data sources that include databases in the cloud and on-premises, APIs, and nontraditional data sources such as text, images, spatial and graph data. Flexibility and openness are ensured by the building-block approach of nodes and connectors, which can be exposed via APIs.

The platform's scores for user interface, data preparation and explainable AI are good. Precanned solutions are average. The number of precanned solutions is increasing with the introduction of "blueprints" targeted at specific industry use cases such as customer segmentation and anomaly detection. These are mainly used for education purposes and provide example workflows that can act as a baseline or be extended. Explainable AI features are available in the platform, but are geared toward usage by expert data scientists. Further guardrails and augmented capabilities in this area need to be improved to match innovation from competitors.

KNIME is strong across all four use cases.

**MathWorks**

MathWorks offers two major products: MATLAB and Simulink. MATLAB is the core product evaluated for this Critical Capabilities; Simulink is considered part of the supporting portfolio and was not evaluated for this research. MATLAB offers deployment capabilities, not only through robust on-premises or multicloud environments (AWS and Azure), but also through support for other cloud environments. Beyond on-premises and cloud environments, MATLAB can manage deployments through widely distributed edge capabilities. The vast majority of users leveraging MATLAB are technology-savvy users who can take advantage of the platform depth and its openness. Users that have considered MATLAB have indicated that they have also considered Google, Alteryx and Anaconda. Many engineering departments that have looked at the MathWorks platform have also considered TIBCO Software and Microsoft as alternatives. MATLAB is often chosen for strong machine learning and integrated advanced analytics functions, the robustness and scalability of its computational platform, as well as the possibility of simulating complex engineering functions.

MathWorks' top scores are driven by its core capabilities in machine learning and other advanced analytics. Those functions and deployed innovations are using leading-edge techniques such as synthetic data generation for simulation, digital twin representations (leveraging knowledge graphs), and integrating deep learning (DL) and reinforcement learning (RL) capabilities. MATLAB also receives high evaluations for its performance and scalability and delivery capabilities, proven through a wide variety of deployed and demanding use cases. In addition, the platform has an excellent score in data exploration and visualization. Finally, users praise the coherence of the overall platform.

MathWorks could improve its capabilities for model management functions; model monitoring and a wider range of drift analysis along with more sophisticated feature engineering capabilities should help the organization catch up with the market. MATLAB also finds itself behind its leading competitors in terms of model interpretability and responsible AI functions; an increasingly important area, not just in service-centric organizations, but also among engineers within asset-centric industries. Finally, an additional area of improvement is its project management capability.

MathWorks scored particularly well in the expert model development use case.

### Microsoft

Azure Machine Learning (Azure ML) is Microsoft's core product for data science and machine learning. The supporting portfolio of products for Azure ML includes Azure Data Factory, Azure Data Catalog, Azure HDInsights, Azure Databricks, Azure DevOps, Power BI and other components. For on-premises workloads, Microsoft offers Machine Learning Server and SDKs for Python and R. The vast majority of Azure ML customers are in pure cloud environments, with some hybrid deployments. Azure ML supports a broad and diverse mix of users, including expert and citizen data scientists, data engineers, ML engineers architects, corporate developers and others. Microsoft customers most often also considered Amazon Web Services (AWS), Databricks, Google and IBM. Customers commonly choose Azure ML for alignment with existing data and analytics investments, agility and flexibility of environment, support for open source, product roadmap, and future vision.

Microsoft garnered excellent scores for flexibility and openness and precanned solutions. Data access and platform and project management have maintained strong scores in recent years. Data exploration and visualization are strong with the complementary use of Power BI. The strength of Azure ML in the cloud is reflected in high scores for performance and scalability, with Azure ML customers calling out helpful cost visibility and control features in addition to powerful hardware options and compute capabilities. Delivery capabilities are strong, with Microsoft standing out for both leading research and providing first-class support around ONNX, ONNX Runtime and other open-source projects. Azure ML offers good explainable AI capabilities, with explanations built into the interface, and integrated fairness and interpretability toolkits Fairlearn and InterpretML.

Many of Azure ML's portfolio's capabilities change or become more complicated in on-premises, hybrid and multicloud environments. Azure ML has limited ability to push compute close to data locality across multiple environments. Data preparation scores remain below average, and Azure ML's capabilities are not in the class of top data prep scorers like Alteryx or Dataiku. Azure ML supports many options for other advanced analytics and mostly relies on open-source libraries. Coherence is an ongoing concern as the Azure stack for DSML and mix of OSS technologies and partner software become even more heterogeneous.

Azure ML is particularly strong for the operationalization use case and is commonly used across the other three use cases of business and data exploration, citizen data science and expert model development.

### RapidMiner

RapidMiner Studio is RapidMiner's primary model development tool and is available as a free edition, as well as a commercial edition. For the enterprise, offerings extend through the RapidMiner AI Hub, which includes collaboration and governance capabilities as well as RapidMiner Go and RapidMiner Notebooks, which are model development experiences for novices and coders, respectively. Turbo Prep, Auto Model and Automated Model Ops are augmented features of the platform, while the RapidMiner AI Cloud offers flexible cloud-based deployment options. RapidMiner

delivers strong capabilities for model portability by giving deployment options across on-premises, hybrid and cloud implementations. RapidMiner's users are primarily both citizen and expert data scientists, followed by business analysts, data engineers, statisticians and finally developers. RapidMiner customers often also consider SAS, KNIME, Microsoft, DataRobot and Alteryx. Customers tend to choose RapidMiner because it's intuitive and easy to learn and collaborate on. Customers also cited speed of model development, control over workflow design and easy operationalization as key contributing factors.

RapidMiner maintains one of the highest scores for data access. RapidMiner earned excellent scores for machine learning, flexibility and openness, delivery and collaboration. These strengths drive outstanding scores in all four use cases. RapidMiner continues to improve and deliver a strong score in data preparation (via Turbo Prep) as well as data exploration and visualization. Precanned solutions provided by RapidMiner through its community contributions are still a highlight.

RapidMiner has a good coherent platform offering, but can lead further by improvements in other advanced analytics.

RapidMiner is strong across all four use cases.

**Samsung SDS**

Samsung SDS offers Brightics AI, an end-to-end analytics platform. The components of the platform include Brightics Standard and Enterprise editions and an open-source tool, Brightics Studio. Brightics AI supports on-premises deployment, but support for cloud services is currently limited to AWS and SDS cloud (SDS's own cloud service). A lightweight engine called Analytics Edge can be utilized to deploy, execute and monitor models made with Brightics AI on the edge. The platform enables both expert and citizen users to develop advanced analytics projects, and also offers a personalized sandbox environment for each user to experiment in. Brightics AI customers most often also considered Alteryx, SAS and RapidMiner. Customers typically choose Brightics AI for its ease of use and multipersona collaboration. Brightics AI provides prebuilt solutions for specific industry verticals such as manufacturing and financial services.

Brightics AI scores high on user interface, flexibility and openness, and precanned solutions. It offers strong integrations with various open-source tools and provides a large number of built-in ML functions through Spark and Python ML libraries. The platform scores average on other advanced analytics capabilities for text, image and audio data. It offers average capabilities in the initial stages of the ML life cycle, including data access, preparation, exploration and visualization. Data labeling and annotation capabilities include image autolabeling, image segmentation, and workflows for assigning and managing manual labeling tasks.

Brightics AI received low scores on model management and explainable AI. Some key features are either missing or in the roadmap; these include A/B testing, automated rollback, support for XAI techniques (such as LIME, SHAP and others), and open-source explainability tools. Scores for platform and project management are average, offering features such as prebuilt models, object reuse methods and real-time

logs during experiments, but lacking functionality around securing ML pipelines, data encryption and obfuscation of personal identifiable information. Separate components for Brightics data preparation, model manager and multiple add-ons lead to fair scores for a coherent environment.

The platform is strongest for the business and data exploration and citizen data science use cases.

### SAS

SAS Visual Data Mining and Machine Learning (VDMML) is the core product evaluated in this Critical Capabilities research. SAS VDMML is one of the products in the Viya portfolio. It is included in various product bundles on SAS Viya, namely SAS Visual Machine Learning, SAS Visual Data Science, SAS Data Science Programming and SAS Visual Data Decisioning. The SAS Viya 4 release offers a cloud-native and flexible container-based architecture, enabling deployment to multiple clouds, including SAS's own cloud. SAS VDMML users are both experts and citizen data scientists. Customers most often considered IBM, Microsoft, AWS and DataRobot during their selection process. Customers choose SAS for its completeness of product portfolio, ability to handle large volumes of data and different data formats, and adaptability of the platform for various use cases and skill sets.

Augmentation and an easy-to-use interface make VDMML appealing to citizen data scientists. The programming interfaces and innate integration with open-source libraries make it sophisticated enough to appeal to and collaborate with code-focused data scientists and AI developers. SAS VDMML earned outstanding scores for data access, preparation, exploration and visualization, as well as for delivery, performance and scalability, and other advanced analytics. The SAS platform can be extended with strong support for simulation, optimization, decision management and streaming analytics. The strengths in these areas reinforce Viya's extensive offerings for composite AI and decision intelligence. SAS VDMML also scored high for machine learning and explainable AI. SAS offers a strong model operationalization and management platform. It includes performance monitoring, automated retraining of models when thresholds are exceeded, governance via a centralized model repository, and lineage for both SAS and open-source models. Scores for platform and project management and precanned solutions are also high.

It received average scores for collaboration. Although SAS has made significant changes to provide more streamlined product bundling, the coherence of the overall platform still needs improvement.

SAS VDMML and its supporting portfolio are strong across all four use cases.

### TIBCO Software

TIBCO Software delivers its "Connected Intelligence" vision, embodied at its core in TIBCO's Data Science platform; but the offering strongly benefits from adjacent platforms such as TIBCO Spotfire and TIBCO Streaming and a robust data and process infrastructure. TIBCO Data Science supports on-premises, hybrid
and multicloud deployments as well as a range of distributed computing environments

due to its robust IoT capabilities and edge computing functions. Users who have selected TIBCO come from a wide range of backgrounds and roles. TIBCO Data Science customers also often consider Alteryx, SAS and IBM. Many engineering departments that have looked at TIBCO Data Science have also considered MathWorks and Microsoft as alternatives. The reasons why users have chosen TIBCO's platform is often linked to its flexibility and openness, and the ease of use resulting from a continuous integration strategy. Users also appreciate the use of a common visualization platform for both data science and business intelligence users.

Flexibility and openness are where TIBCO particularly excels. The company's efforts in integrating not only open-source developments, but also other models built on various other platforms is a strength recognized by its users. Platform coherence is solid, thanks to major efforts from the company to integrate its various acquisitions making it a continuous development and deployment platform. Data access, exploration and visualization received excellent scores, while data preparation also scored high. The platform's capabilities for delivery are excellent, complemented by high-scoring model management, scalability and performance. TIBCO also scores well on user interface, collaboration and platform/project management.

Compared with TIBCO's leading competitors, the vendor's precanned solutions can be improved. Prospective customers should monitor the continued development of TIBCO's LABS offerings for the multiple markets it serves. TIBCO can also further improve multimodel performance monitoring, especially when it comes to hybrid developments or composite AI, as this is becoming critical to mature data science teams.

TIBCO Data Science and its supporting portfolio are strong across all four use cases.

## Context

As a companion note to Gartner's Magic Quadrant for Data Science and Machine Learning Platforms, this research aims to guide the selection of data science platforms, and does not take services or packaged applications into consideration. Users of data science platforms are typically data scientists, data engineers, statisticians and, increasingly, citizen data scientists (see Staffing Data Science Teams: Map Capabilities to Key Roles).

## Product/Service Class Definition

Gartner defines a data science and machine learning (DSML) platform as a core product and supporting portfolio of coherently integrated products, components, libraries and frameworks (including proprietary, partner-sourced and open-source). Its primary users are data science professionals, including expert data scientists, citizen data scientists, data engineers, application developers and machine learning specialists.

The critical capabilities of a DSML platform are:

- Data access

- Data preparation
- Data exploration and visualization
- User interface
- Machine learning
- Other advanced analytics
- Flexibility and openness
- Performance and scalability
- Delivery
- Platform and project management
- Model management
- Explainable AI (XAI)
- Precanned solutions
- Collaboration
- Coherence

A detailed description of data science and machine learning platforms can be found in the Magic Quadrant for Data Science and Machine Learning Platforms.

# Critical Capabilities Definition

**Data Access**

This refers to the ability to engage and integrate data from various sources, both on-premises and in the cloud, and of different types.

Data scientists, data engineers and business analysts are looking to connect and analyze more data sources. Sources can include relational reuse NoSQL data sources, proprietary data sources and enterprise applications; data types may include tabular, text, images, graphs, logs, time series or audio. It is important for data science and machine learning platforms to provide data connectors and the ability to import and combine disparate data sources without use of an external tool. This might include corporate extraction, transformation and loading (ETL) or specialized database skills. The platform should be able to handle increasingly large and distributed volumes of diverse data.

Functionality includes:

- Data access to traditional and nontraditional data types
- Hybrid data sources (on-premises and cloud)

- Multicloud data sources
- Basic and advanced ETL functionality
- Web data integration
- API-based data access
- Data lake support
- Enterprise application access
- Hadoop and NoSQL access
- Data refresh and synchronization
- Real-time data feeds
- Internet of Things (IoT) as a data source
- Data governance and metadata management
- Data lineage

**Data Preparation**

Data preparation is the ability to visually explore, find, access, clean, combine and model data in an agile, yet trusted, way.

Platforms may provide built-in suggestions for data blending, advanced binning and smoothing algorithms, and other data transformation capabilities. Data preparation is a vital stage in the machine learning life cycle and is increasingly supported from within the platform workflow.

Functionality includes:

- Data blending/wrangling
- Data quality support
- Dataset partitioning
- Binning and smoothing
- Filter and search
- Watermarking
- Transformations, aggregation and set operations
- Data cataloging
- Data labeling and annotation
- Augmented data preprocessing
- Augmented data enrichment and synthetic data

- Augmented data preparation and dimensionality reduction
- On-premises, (multi)cloud and hybrid support

### Data Exploration and Visualization

This refers to the ability to visually manipulate, interact with, and explore the data, and perform basic descriptive statistics and pattern detection via a wide variety of exploratory steps and visualization options.

The descriptive and diagnostic analyses included here are the foundation of data science, enabling the user to diagnose the data, explore correlations and start to form hypotheses.

Functionality includes:

- Interactive dashboards and charts
- Augmented data discovery
- Custom visualizations
- Univariate and bivariate statistics
- Statistical significance testing
- Signal preprocessing
- Visual interaction and exploration
- Clustering and self-organizing maps
- Geolocation mapping
- Affinity and graph analysis
- Conjoint and survey analysis
- Density estimation
- Similarity metrics
- On-premises, (multi)cloud and hybrid support

### User Interface

User interface covers the usability, visual design, intuitiveness and responsiveness of the platform for its intended users. The product should have a coherent “look and feel” that facilitates user engagement. Platforms may provide detailed help with examples and tips to overcome common problems.

As citizen data scientists and business analysts play an increasingly important role in the data science life cycle, user-friendly interfaces are becoming more desirable and even critical to some organizations. User interfaces may include drag-and-drop workflow creation, which does not require code-writing skills, and facilitates

collaboration and replication in model development. Strong user interfaces can also increase the productivity of scarce expert-level data scientists.

Key aspects for consideration:

- Ease of use and learning curve
- Citizen data science
- Developer-focused data science
- Wizards and contextual aids
- Visual pipelining or visual composition framework (VCF)
- Third-party applications
- Q&A or natural language interaction
- Customizable, pretrained algorithms
- Documentation
- User communities

**Machine Learning**

This capability refers to the process of exploring large amounts of data through statistical models to solve a business problem. The nature of the business problem determines the preferred modeling technique to be used.

Almost all machine learning (ML) platforms either support multiple models out of the box or provide an option to custom code the same. Common modeling techniques and functionality include:

- Regression
- Time series analysis
- Neural nets
- Deep learning (deep neural networks)
- Reinforcement learning
- Classification and regression trees
- Further rule induction techniques
- Support vector machines
- Instance-based approaches (k-nearest neighbor, for example)
- Bayesian modeling (naïve Bayes, for example)
- Transfer learning

- Federated learning
- Self-supervised learning
- Generative adversarial networks
- Ensembles and hierarchical models
- Recommendation techniques (including collaborative filtering)
- Import, call and development of other predictive models
- Testing of predictive models
- Small data techniques (active learning, few-shot learning, similarity-based completion, curriculum learning)
- Augmented feature learning, selection and engineering
- Augmented algorithm selection
- Augmented model tuning
- On-premises, (multi)cloud and hybrid support

Platforms should also offer a range of neural network approaches, including “backpropagation” and its variants. Increasingly, we expect more coverage of various deep-learning architectures, including convolutional, recurrent and graph neural networks, or the definition of custom network structures.

Further, we expect several “wisdom of crowds” approaches, known as “ensemble models,” such as:

- Voting
- Bagging
- Stacking
- Boosting

Last, but not least, machine learning requires the platform to evaluate the model through:

- Visual evaluations like lift charts or receiver operating characteristic (ROC) curves
- Measures of fit
- K-fold cross-validation
- Testing methodologies (like A/B testing and sensitivity analysis)
- Loss matrices

**Other Advanced Analytics**

These include the ability to integrate additional statistical methods, optimization, simulation, and text and image analytics into the development environment. It is critical that the functions be flexibly accessed and used within one coherent offering.

Optimization: This refers to a type of prescriptive analytics that uses a mathematical algorithm to choose the “best” alternative(s) to meet specified objectives and constraints. Functionality includes:

- Solver approaches
- Heuristic approaches
- Design of experiments

Simulation: This predictive analytics approach involves building a model of a system or process to experiment with or study the behavior of how it works, with a focus on understanding the range of possible outcomes. Functionality includes:

- Discrete event simulation
- Monte Carlo simulation
- Agent-based modeling

Text and image analytics: This refers to support for the analysis of nontabular and less structured data such as natural language, audio, images and video. Functionality includes:

- Basic text analytic functions
- Granular linguistic functionality (stemming, tokenization, part-of-speech tagging)
- Advanced linguistic functions (latent semantic indexing or similar, concept relevance, question answering)
- Image/video processing
- Audio mining and signal processing
- Customizable pretrained algorithms

Miscellaneous other advanced analytics:

- Geospatial analysis
- Financial modeling and econometrics
- Stream processing/data in motion
- Decision modeling
- Decision management

- Composite AI

### Flexibility and Openness

Flexibility and openness reflect the ability to natively provide data scientists with the freedom of choice to use their preferred or most applicable methods and tools within the platform.

This capability includes support of relevant tools, algorithms, languages, frameworks and data management platforms developed by third parties (especially open source). This capability also represents a capacity to natively support and integrate a community or marketplace as an extension of a platform.

Key supported features:

- Languages such as R, Python, Scala and Java
- Popular open-source libraries and frameworks
- Algorithms available via third-party libraries, products or marketplaces
- Notebooks such as Jupyter or Zeppelin
- Visualization from both commercial and open-source tools such as D3, knitr and Plotly
- Open-source automated machine learning tools
- Open-source data management platforms such as Spark and Hadoop
- Creation of reusable and shareable environments using Docker
- Scripting and embedding capabilities
- Ability to expand the current functionality with third-party offerings
- Visibility into the code of all implemented functions

### Performance and Scalability

This capability reflects time taken to load data, create and validate models, and deploy those models into the business. As data volumes and business complexity grow, and the demand for faster or real-time insights and decisions rises, performance and scalability become increasingly important.

This impacts a number of areas ranging from data volume, parallelism and the ability to iterate, to the number of concurrent users. Data science teams need control across edge, desktop, server and cloud deployments.

Key considerations are:

- Cloud computing
- Big data volume scalability

- In-memory and/or distributed computing, including Spark/Hadoop support
- In-database analytics
- Algorithmic efficiency on either a single-node or multiple-node environment
- Real-time data and streams
- Support for GPUs
- Support for other specialized hardware
- Cost guidance
- Performance options for training
- On-premises, (multi)cloud and hybrid support

**Delivery**

This refers to the ease and speed with which the user can move models from a developer environment to a production environment or embed them in a business process.

Platforms should support the ability to create APIs or containers that can be used for faster deployment in business scenarios. Key functionality includes:

- Write-back
- Recoding
- Code synthesis
- Publishing of REST APIs
- Model exchange with external parties and data science platforms via Predictive Model Markup Language (PMML) or ONNX
- Containerization
- Augmented model deployment and monitoring
- Web deployment
- Deployment (in parallel) on-premises, (multi)cloud and hybrid

**Platform and Project Management**

This refers to the platform's capabilities for security, compute resource management, governance, reuse and version management of projects, auditing, lineage and reproducibility. For regulatory compliance, certain industries require providing governance of models and an audit trail to regulators.

Key issues to consider include:

- Compliance and auditing
- Object reuse
- Multiuser capabilities
- Debugging and unit testing
- Runtime optimization
- Audits and logs
- Data encryption
- ML pipeline security
- Client deployment

**Model Management**

Model management features streamline the operationalization and execution of models. Model managers monitor the use and performance of models in production and help identify models that need to be updated or replaced.

It is important to monitor the performance of models in production to ensure the relationships found during development are still valid, and that the model maintains its accuracy.

Key functionality includes:

- Metadata management
- Traceability
- Champion/challenger
- Model telemetry
- Model catalog and reuse recommendations
- Model reproducibility
- Technical performance tracking
- Business performance tracking
- Adaptive machine learning
- Model alignment to business objectives
- Governance of model access and use
- Model licensing issues
- Scripting and automation
- Process monitoring

- Model management across deployment modes (on-premises, [multi] cloud and hybrid)

### Explainable AI (XAI)

The platform provides capabilities to describe models, highlight strengths and weaknesses, predict likely behavior over time and identify any potential biases.

The platform should be able to articulate the characteristics of a model to enable accuracy, fairness, accountability, stability and transparency in algorithmic decision making and to support the responsible use of AI.

Key functionality includes:

- Model documentation
- Augmented explainability
- Explainability techniques (including permutation importance, feature importance, sensitivity analysis, partial dependence plots, ICE plots, integrated gradients, similarity-based methods and others)
- Support for open-source explainability tools or frameworks
- Transparency of the process and lineage (tracking of versions, iterations and process documentation, for example)
- Bias detection and mitigation
- Regulatory support (e.g., GDPR, HIPAA and others)

### Precanned Solutions

These are template and accelerator mechanisms that provide users with boilerplate solutions for common use cases (for example, cross-selling, social network analysis, fraud detection, recommender systems, propensity to buy, failure prediction and anomaly detection).

Precanned solutions offer fast time to value and high ease of use at the expense of customization and granularity of control. Some platforms enable precanned solutions to be integrated and imported via libraries, marketplaces or galleries.

Common categories of precanned solutions include:

- Marketing, sales and customer service
- Finance, risk management and quality management
- Internet of Things (IoT)
- Supply chain/logistics
- Back-office analytics

- IT operations
- Cybersecurity
- Anomaly detection
- “What if” scenarios

### Collaboration

This refers to the ability of a platform to facilitate multiple types of collaborations between users of various skills, across workflows and projects. These include annotations, discussions, question and answer, cross-border model repositories, and the ability to export experiments and visualize them.

Also included is the ability for projects and workflows to be archived, commented on and reused later.

Key functionality includes:

- Collaboration across all modeling steps for teams in different locations
- Multipersona collaboration between expert data scientists, citizen data scientists, nontechnical business users and other user types
- Discussion threads
- Ratings and recommendations
- Marketplace/hub

### Coherence

Coherence is about how intuitive, consistent and integrated the platform is when supporting the entire data science process for multiple user types.

The platform itself must provide metadata and integration capabilities for our 15 critical capabilities, and provide a seamless, end-to-end experience. This affects not only usability, but also runtime behavior and flexibility. This metacapability includes ensuring data input/output formats are standardized wherever possible so that components have a consistent “look and feel.” Terminology should also be unified across the platform.

## Use Cases

### Business and Data Exploration

This is the scenario of identifying business use cases, forming hypotheses, as well as data preparation, exploration and visualization.

This use case covers performance metrics and success criteria, new and existing datasets, identifying solvable problems, forming hypotheses, combining and preparing datasets, engineering sustainable data pipelines, visualizing early findings, and assessing the feasibility of further model development.

### Citizen Data Science

This scenario describes the capabilities and practices that allow users to extract advanced analytic insights from data without the need for extensive data science expertise.

This includes functionality designed for primarily low- and no-code users functioning as model builders. Central to citizen data science are augmented analytics capabilities. These include automated, streamlined data access and data engineering,; augmented user insight through automated modeling and pattern detection including feature engineering, model selection and validation,; automated deployment and operationalization,; and a focus on collaboration and sharing.

### Expert Model Development

This scenario describes projects where data science and machine learning solutions are used by experts to fuel competitively differentiating or specialized custom models.

Expert model development may involve:

- A code-first approach
- Numerous complex data sources
- Cutting-edge techniques (such as deep neural nets, reinforcement learning and transfer learning)
- Significant computing infrastructure
- Specialized data science and machine learning skills

### Operationalization

In this scenario, data science solutions are implemented and delivered to the business, and the platform is used to make refinements and updates to these deployed models.

“Production,” “deployment,” “ModelOps” and “MLOps” are additional terms that are commonly used to describe aspects of this use case, wherein the platform supports the release, activation, monitoring, performance tracking, management, reuse, maintenance and governance of ML.

# Vendors Added and Dropped

## Added

The following vendors were added to this year’s Critical Capabilities research:

- Alibaba Cloud
- Amazon Web Services
- Cloudera

- Samsung SDS

### Dropped

No vendors were dropped from this year's Critical Capabilities report.

# Inclusion Criteria

Gartner Magic Quadrants and Critical Capabilities identify and analyze the most relevant providers in a market. By default, an upper limit of 20 vendors is imposed to enable identification of the most relevant providers. On some specific occasions, however, this upper limit may be raised when the research's value to clients would otherwise be diminished.

The inclusion criteria represent the specific attributes necessary for inclusion in this Critical Capabilities research. They were applied progressively, in sequence and in cumulative fashion to aid identification of the most relevant providers.

Inclusion Criterion 1: Data Science and Machine Learning Platform

A vendor`s DSML platform had to:

- Offer a mixture of the basic and advanced functionality essential for building DSML solutions (primarily predictive and prescriptive models).
- Support the incorporation of these solutions into business processes, surrounding infrastructure, products and applications.
- Support the sustainable consumption of insights derived from the platform and offer functionality to quantify and track the value of data science projects.
- Support variously skilled data science professionals ("data scientist" is an inconsistently applied job title and professional distinction — a DSML platform's user base is often made up of professionals with diverse technical and business backgrounds).
- Support multiple tasks across the data science life cycle, including:
    - Problem and business context understanding
    - Data ingestion
    - Data preparation
    - Data exploration
    - Feature engineering
    - Model creation and training
    - Model testing
    - Deployment

    - Monitoring
    - Maintenance
    - Data and model governance
    - Explainable AI (XAI)
    - Business value tracking
    - Collaboration

Additionally, a vendor had to be able to provide technical support for its DSML platform directly and/or via commercial support partners.

Inclusion Criterion 2: Revenue and Growth

A vendor's core product had to offer one or more common license models:

- Perpetual license model
- SaaS subscription model
- Consumption-based model or other type of model

The following information for the core product was considered:

- Revenue, in U.S. dollars, generated from perpetual licenses during 2019. This included software license, maintenance and update revenue, but excluded hardware and professional services revenue.
- Annual contract value (ACV), in U.S. dollars, generated from SaaS subscriptions in 2019, excluding any professional services included in annual contracts. For multiyear contracts, only the contract value for the first 12 months was used for this calculation.

A vendor needed to have either:

- At least $75 million in combined perpetual license revenue and ACV for 2019

Or:

- At least $10 million in combined perpetual license revenue and ACV for 2019 calendar year *and* at least 18% in combined revenue growth, compared with 2018

Only core products that passed Inclusion Criterion 2 were considered for Inclusion Criterion 3.

Inclusion Criterion 3: Customer Counts

Vendors that satisfied the requirements of Inclusion Criterion 2 were next evaluated on their customer counts. We required significant cross-industry and cross-geographic traction for each core product under consideration. Counts included only active unique

customer organizations using the latest version of the core product or a version released in the 12 months prior to August 2020.

Cross-Industry Customer Count

We assessed counts of the number of active unique customer organizations (logos) using each of the DSML platforms under consideration in production environments. For each core product, we required at least 10 unique organizations (logos), which had to have data science solutions in production environments and which had to come from at least four of the following major industry segments:

- Banking and securities
- Communications, media and services
- Education
- Government
- Healthcare
- Insurance
- Manufacturing and natural resources
- Retail
- Transportation
- Utilities
- Wholesale trade

Cross-Region Customer Count

In addition, there had to be at least two active customer organizations (logos) in each of the following:

- North America
- European Union, Norway, Switzerland and the U.K.
- Rest of the world

Only core products that passed Inclusion Criterion 3 were considered for Inclusion Criterion 4.

Inclusion Criterion 4: Market Traction

A vendor’s market traction was evaluated using a composite metric. This metric drew on internal Gartner data and other external sources of information to assess the level of market interest in, and the momentum of, each vendor and its DSML platform. Inputs included:

- Gartner client inquiries
- Gartner.com search volume
- Volume of job listings and headcount trends
- Internet search volume and trend analysis
- Frequency of mention as an evaluated competitor from July 2019 to July 2020
- Growth in new customers

Inclusion Criteria 5: Product Capability Scoring

If more than 20 vendors met the first four criteria, only the vendors with the top 20 market traction scores advanced to the full Critical Capabilities evaluation.

Table 1: Weighting for Critical Capabilities in Use Cases

| Critical Capabilities | Business and Data Exploration | Citizen Data Science | Expert Model Development | Operationalization |
|---|---|---|---|---|
| Data Access | 14% | 5% | 5% | 5% |
| Data Preparation | 16% | 13% | 5% | 0% |
| Data Exploration and Visualization | 25% | 7% | 5% | 0% |
| User Interface | 5% | 19% | 5% | 5% |
| Machine Learning | 5% | 10% | 25% | 5% |
| Other Advanced Analytics | 5% | 5% | 15% | 0% |

| Critical Capabilities | Business and Data Exploration | Citizen Data Science | Expert Model Development | Operationalization |
|---|---|---|---|---|
| Flexibility and Openness | 5% | 5% | 10% | 5% |
| Performance and Scalability | 5% | 5% | 5% | 16% |
| Delivery | 0% | 5% | 5% | 16% |
| Platform and Project Management | 5% | 0% | 0% | 5% |
| Model Management | 0% | 0% | 5% | 17% |
| Explainable AI (XAI) | 0% | 6% | 5% | 8% |
| Precanned Solutions | 0% | 5% | 0% | 5% |
| Collaboration | 10% | 10% | 5% | 6% |
| Coherence | 5% | 5% | 5% | 7% |

As of 1 January 2021

This methodology requires analysts to identify the critical capabilities for a class of products/services. Each capability is then weighted in terms of its relative importance for specific product/service use cases.

## Critical Capabilities Rating

Each of the products/services that meet our inclusion criteria has been evaluated on the critical capabilities on a scale from 1.0 to 5.0.

## Table 2: Product/Service Rating on Critical Capabilities

| Critical Capabilities | Alibaba Cloud | Altair | Alteryx | Amazon Web Services | Anaconda | Cloudera | Databricks | Dataiku | DataRobot | Domino | Google | H2O.ai | IBM | KNIME | MathWorks | Microsoft | RapidMiner | Samsung SDS | SAS | TIBCO Software |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data Access | 3.5 | 4 | 4.6 | 3.9 | 3.2 | 3.5 | 4.3 | 4.7 | 4.3 | 4 | 3.9 | 3.3 | 4.6 | 4.6 | 4 | 4.3 | 4.6 | 3.6 | 4.5 | 4.5 |
| Data Preparation | 3.3 | 4.2 | 4.7 | 3.6 | 3 | 3 | 4 | 4.7 | 4.3 | 3.5 | 3.8 | 3.5 | 4.1 | 4.1 | 4.1 | 3.4 | 4.3 | 3.9 | 4.6 | 4.2 |
| Data Exploration and Visualization | 3.4 | 4 | 3.9 | 3.5 | 3.6 | 3.8 | 3.9 | 4.2 | 3.9 | 3.7 | 3.5 | 4 | 4 | 4.3 | 4.6 | 4.2 | 4 | 3.8 | 4.5 | 4.5 |
| User Interface | 3.2 | 4.2 | 4.5 | 3.4 | 3.3 | 3.5 | 3.8 | 4.6 | 4.4 | 3.5 | 3.7 | 3.9 | 4 | 4.1 | 3.5 | 4.1 | 4.3 | 4 | 4.3 | 4.2 |
| Machine Learning | 3.4 | 3.8 | 3.9 | 4.2 | 4 | 3.5 | 4.2 | 4.2 | 4.1 | 4.3 | 4.3 | 4.7 | 4.6 | 4.4 | 4.6 | 4.2 | 4.4 | 3.7 | 4.4 | 4.5 |
| Other Advanced Analytics | 2.9 | 3.4 | 4 | 3.8 | 4 | 2.9 | 4.2 | 3 | 3.5 | 3.9 | 4 | 3.8 | 4.2 | 4 | 4.9 | 3.7 | 4.1 | 3.8 | 4.7 | 4.1 |
| Flexibility and Openness | 3.5 | 4.1 | 4.1 | 4.4 | 4.3 | 4.2 | 4.7 | 4.2 | 4.3 | 4.9 | 4.8 | 4.9 | 4.7 | 4.4 | 4.1 | 4.5 | 4.5 | 4 | 4.3 | 4.8 |
| Performance and Scalability | 3.6 | 3.5 | 4.1 | 4.9 | 3.8 | 4.5 | 4.8 | 4 | 4.1 | 4 | 4.8 | 4.5 | 4.5 | 4.1 | 4.8 | 4.3 | 4 | 3.7 | 4.5 | 4.2 |
| Delivery | 3.2 | 3.4 | 4.5 | 4 | 3.6 | 4 | 4.4 | 4.2 | 4.5 | 4.4 | 4 | 4.8 | 4.4 | 4.5 | 4.3 | 4.2 | 4.3 | 3.7 | 4.6 | 4.5 |
| Platform and Project Management | 2.9 | 4 | 4 | 3.9 | 3.2 | 4 | 4.3 | 4.5 | 4.3 | 4.5 | 3.9 | 4 | 4.2 | 4.5 | 3.7 | 4.4 | 4.4 | 3.5 | 4.2 | 4.1 |
| Model Management | 2.9 | 3.8 | 4.1 | 4.1 | 2.8 | 4 | 4.2 | 4 | 4.3 | 4.3 | 3.6 | 4.2 | 4.1 | 4.2 | 3.5 | 4.2 | 4.2 | 3.2 | 4.3 | 4 |
| Explainable AI (XAI) | 2.6 | 4.2 | 4 | 3.8 | 2.5 | 3.8 | 3.5 | 4.2 | 3.7 | 4 | 4.4 | 4.6 | 4.3 | 4 | 3.8 | 4 | 4 | 2.5 | 4.4 | 4.3 |
| Precanned Solutions | 2.9 | 3.2 | 3.6 | 4.6 | 3.2 | 3.2 | 4 | 3.5 | 3.2 | 2.8 | 4.3 | 3.4 | 4.3 | 3.9 | 4.4 | 4.6 | 3.9 | 4 | 4 | 3.9 |
| Collaboration | 3.2 | 3.9 | 4.3 | 3.4 | 4 | 3.5 | 4.1 | 4.8 | 4.1 | 4.7 | 3.7 | 3.8 | 4.5 | 4.4 | 3.8 | 4 | 4.5 | 3.9 | 3.7 | 4.1 |
| Coherence | 4 | 4.1 | 4 | 3.6 | 3.5 | 3.5 | 4 | 4.7 | 4.2 | 4 | 3.7 | 3.6 | 4.1 | 4.6 | 4.5 | 3.7 | 4 | 3.5 | 3.5 | 4 |

As of 1 January 2021

Table 3 shows the product/service scores for each use case. The scores, which are generated by multiplying the use-case weightings by the product/service ratings, summarize how well the critical capabilities are met for each use case.

Table 3: Product Score in Use Cases

| Use Cases | Business and Data Exploration | Citizen Data Science | Expert Model Development | Operationalization |
|---|---|---|---|---|
| Alibaba Cloud | 3.36 | 3.27 | 3.28 | 3.23 |
| Altair | 3.98 | 3.93 | 3.84 | 3.77 |
| Alteryx | 4.23 | 4.24 | 4.12 | 4.17 |
| Amazon Web Services | 3.75 | 3.81 | 3.97 | 4.1 |
| Anaconda | 3.53 | 3.5 | 3.7 | 3.41 |
| Cloudera | 3.58 | 3.54 | 3.59 | 3.89 |
| Databricks | 4.13 | 4.08 | 4.2 | 4.26 |
| Dataiku | 4.4 | 4.35 | 4.13 | 4.23 |
| DataRobot | 4.12 | 4.12 | 4.07 | 4.17 |
| Domino | 3.97 | 3.92 | 4.16 | 4.16 |
| Google | 3.86 | 3.98 | 4.11 | 4.1 |
| H2O.ai | 3.87 | 4.02 | 4.25 | 4.26 |
| IBM | 4.27 | 4.28 | 4.38 | 4.34 |
| KNIME | 4.32 | 4.24 | 4.29 | 4.29 |
| MathWorks | 4.25 | 4.14 | 4.34 | 4.11 |
| Microsoft | 4.04 | 4.04 | 4.08 | 4.2 |
| RapidMiner | 4.27 | 4.26 | 4.28 | 4.22 |
| Samsung SDS | 3.78 | 3.76 | 3.69 | 3.55 |
| SAS | 4.36 | 4.31 | 4.38 | 4.29 |
| TIBCO Software | 4.33 | 4.28 | 4.35 | 4.24 |

To determine an overall score for each product/service in the use cases, multiply the ratings in Table 2 by the weightings shown in Table 1.

# Evidence

Information for this report came from:

- An evaluation of instruction manuals and documentation supplied by selected vendors to check their functionality.
- Instruction manuals and documentation of selected vendors. We used these to verify platform functionality.
- A questionnaire completed by the vendors.
- Vendor briefings, including product demonstrations, about individual vendors' strategy and operations.
- An extensive RFP inquiring how each vendor delivers specific features that correspond to our 15 critical capabilities (see Toolkit: RFP for Data Science and Machine Learning Platforms).
- A prepared video demonstration of how well vendors' DSML platforms address specific functionality requirements across the 15 critical capabilities.
- Gartner Peer Insights.
- Interactions between Gartner analysts and Gartner clients deciding their evaluation criteria, and Gartner clients' opinions about how successfully vendors meet these criteria.

# Critical Capabilities Methodology

This methodology requires analysts to identify the critical capabilities for a class of products or services. Each capability is then weighted in terms of its relative importance for specific product or service use cases. Next, products/services are rated in terms of how well they achieve each of the critical capabilities. A score that summarizes how well they meet the critical capabilities for each use case is then calculated for each product/service.

"Critical capabilities" are attributes that differentiate products/services in a class in terms of their quality and performance. Gartner recommends that users consider the set of critical capabilities as some of the most important criteria for acquisition decisions.

In defining the product/service category for evaluation, the analyst first identifies the leading uses for the products/services in this market. What needs are end-users looking to fulfill, when considering products/services in this market? Use cases should match common client deployment scenarios. These distinct client scenarios define the Use Cases.

The analyst then identifies the critical capabilities. These capabilities are generalized groups of features commonly required by this class of products/services. Each capability is assigned a level of importance in fulfilling that particular need; some sets of features are more important than others, depending on the use case being evaluated.

Each vendor's product or service is evaluated in terms of how well it delivers each capability, on a five-point scale. These ratings are displayed side-by-side for all vendors, allowing easy comparisons between the different sets of features.

Ratings and summary scores range from 1.0 to 5.0:

1 = Poor or Absent: most or all defined requirements for a capability are not achieved

2 = Fair: some requirements are not achieved

3 = Good: meets requirements

4 = Excellent: meets or exceeds some requirements

5 = Outstanding: significantly exceeds requirements

To determine an overall score for each product in the use cases, the product ratings are multiplied by the weightings to come up with the product score in use cases.

The critical capabilities Gartner has selected do not represent all capabilities for any product; therefore, may not represent those most important for a specific use situation or business objective. Clients should use a critical capabilities analysis as one of several sources of input about a product before making a product/service decision.