A Data Science Maturity Model for Enterprise Assessment

MARK HORNICK

DIRECTOR, ORACLE ADVANCED ANALYTICS AND MACHINE LEARNING

JULY 2018

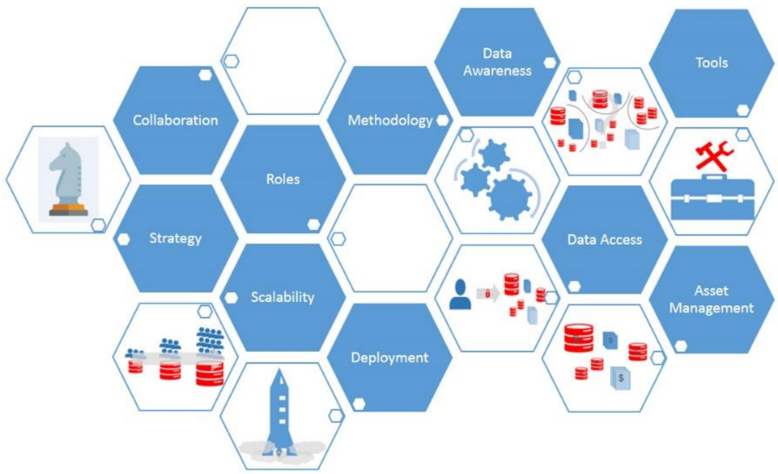
Disclaimer

This data science maturity model is provided for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Table of Contents

Introduction

"Maturity models" aid enterprises in understanding their current and target states. Enterprises that already embrace data science as a core competency, as well as those just getting started, often seek a road map for improving that competency. A data science maturity model is one way of assessing an enterprise and guiding the quest for data science nirvana.



As an assessment tool, this *Data Science Maturity Model* provides a set of dimensions relevant to data science and five maturity levels in each - 1 being the least mature, 5 being the most. Here are important maturity model dimensions with the goal to provide both an assessment tool and potential road map:

- **Strategy** - What is the enterprise business strategy for data science?
- **Roles** - What roles are defined and developed within the enterprise to support data science activities?
- **Collaboration** - How do data scientists collaborate with others in the enterprise, e.g., business analysts, application and dashboard developers, to evolve and hand-off data science work products?
- **Methodology** - What is the enterprise approach or methodology to data science?
- **Data Awareness** - How easily can data scientists learn about enterprise data resources?
- **Data Access** - How do data analysts and data scientists request and access data? How is data accessed, controlled, managed, and monitored?
- **Scalability** - Do the tools used for data science scale and perform for data exploration, preparation, modeling, scoring, and deployment?
- **Asset Management** - How are data science assets managed and controlled?
- **Tools** - What tools are used within the enterprise for data science objectives? Can data scientists take advantage of open source tools in combination with production quality infrastructure?
- **Deployment** - How easily can data science work products be placed into production to meet timely business objectives?

In this whitepaper, I discuss each of these dimensions and levels by which business leaders and data science players can assess where their enterprise is, identify where they would like to be, and consider how important each dimension is for the business and overall corporate strategy. Such introspection is a step toward identifying architectures, tools, and practices that can help achieve data science goals.

Strategy

What is the enterprise business strategy for data science?

A *strategy* can be defined as "a high-level plan to achieve one or more goals under conditions of uncertainty." With respect to data science, goals may include making better business decisions, making new discoveries, improving customer acquisition / retention / satisfaction, reducing costs, optimizing processes, among others. Depending on the quantity and quality of data available and the way that data are used, the degree of uncertainty facing an enterprise can be significantly reduced or accentuated.

The five levels of the strategy dimension are:

Level 1: Enterprise has no governing strategy for applying data science.

> For enterprises at Level 1, the world of data science may be unfamiliar, but data certainly is not. Data analytics may be a routine part of enterprise activity but with no overall governing strategy or realization that data is a corporate asset. The enterprise has defined goals, but the extent to which data support those goals is limited.

Level 2: Enterprise is exploring the value of data science as a core competency.

> The Level 2 enterprise realizes the potential value of data and the need to leverage that data for greater business advantage. With all the hype and substance around machine learning, artificial intelligence, and advanced analytics, business leaders are investigating the value data science can offer and are actively conducting proofs-of-concept – exploring data science seriously as a core business competency.

Level 3: Enterprise recognizes data science as a core competency for competitive advantage.

> Having done due diligence, enterprises at Level 3 have committed to pursuing data science as a core competency and the benefits it can bring. Systematic efforts are underway to enhance data science capabilities along the remaining dimensions of this maturity model.

Level 4: Enterprise embraces a data-driven approach to decision making.

> Once an enterprise establishes a competency in data science, enterprises at Level 4 feel confident to embrace the use of data-driven decision making – backing up or substituting business instincts with measured results and predictive analytics / machine learning. As data and skill sets are refined, business leaders have greater confidence to trust data science results when making key business decisions.

Level 5: Data are viewed as an essential corporate asset – *data capital*.

A capping strategy with respect to data science involves giving data the "reverence" it deserves - recognizing it as a valuable corporate asset - a form of capital. At Level 5, the enterprise allocates adequate resources to conduct data science projects supported by proper management, maintenance, assessment, security, and growth of data assets, and the human resources to systematically achieve strategic goals.

Roles

What roles are defined and developed in the enterprise to support data science activities?

A role can be defined as "a set of connected behaviors, rights, obligations, beliefs, and norms as conceptualized by people in a social situation." As with most any new field, data science within an enterprise can benefit from the introduction of new roles. There are a few potentially new roles worth considering: data scientist, chief data officer, data librarian, and chief data science officer.

Once considered *unicorns*, data scientists are now more numerous as universities offer degrees at both the masters and doctorate level. Even so, data scientists may have different strengths, ranging from their ability to prepare/wrangle data, write code, use machine learning algorithms, use visualization effectively, and communicate results to both technical and non-technical audiences. As such, a given data science project may require a team of data scientists with complementary skills.

The five maturity levels of the roles dimension are:

Level 1: Traditional data analysts explore and summarize data using deductive techniques.

> Enterprises at Level 1 may have persons dedicated to data analysis - *data analysts* - and draw on skills of database administrators (DBAs) or business analysts to deliver business intelligence. They likely use a variety of tools that support, for example, spreadsheet analytics, visualization, dashboards, database query languages, among others. Persons in these roles typically use deductive reasoning in the sense that they formulate queries to answer specific questions.

Level 2: The *data scientist* role is introduced to begin leveraging advanced, inductive techniques.

> The Level 2 enterprise recognizes the need for more sophisticated analytics and the value that those trained in data science - the now much admired role of the *data scientist* - can bring to the enterprise. Level 2 enterprises can now more confidently explore, develop, and deploy solutions based on machine learning, artificial intelligence, data mining, predictive analytics, and advanced analytics - depending on which term or terms most resonate with your enterprise. At Level 2, data scientists are typically added as needed to individual departments or organizations.

ORACLE

Level 3: Chief Data Officer (CDO) role is introduced to help manage data as a corporate asset.

Although not necessarily a pure data science role, the *Chief Data Officer* role is highly beneficial, if not critical, for the data science-focused enterprise. The CDO is responsible for enterprise-wide governance and use of data assets. Along with a CDO, the role of *data librarian* may also be introduced to support *data curation* within the enterprise. With the introduction of these roles at Level 3, not only is data science being taken more seriously, but the key input to data science projects – the data – is as well.

Level 4: The data scientist career path is codified and standardized across the enterprise.

Level 4 enterprises strive for greater uniformity across the enterprise for the data scientist role with respect to job description, skills, and training. In some enterprises, data science activities and/or data scientists may be organized under a common or matrix management structure.

Level 5: Chief Data Science Officer (CDSO) role introduced.

Just as the Chief Data Officer role is beneficial for enterprises taking data more seriously, the Level 5 enterprise also recognizes the need for a *Chief Data Science Officer*. In this role, the CDSO oversees, coordinates, evaluates, and recommends data science projects and the tools and infrastructure needed to help achieve enterprise business objectives.

Collaboration

How do data scientists collaborate among themselves and other data science players to evolve and hand-off data science work products?

Data science projects often involve significant collaboration, defined as "two or more people or organizations working together to realize or achieve a goal." Successful data science projects that positively impact an enterprise will often require the involvement of multiple *players*: data scientists, data and business analysts, business leaders, domain experts, application and dashboard developers, database administrators, and information technology (IT) administrators, just to name a few. Collaboration can informal or formal, however, in this context, we look to processes, methodologies, and tools that support, encourage, monitor, and guide collaboration among players.

The five maturity levels of the collaboration dimension are:

Level 1: Data analysts often work in silos, performing work in isolation and storing data and results in local environments.

Enterprises at Level 1 often suffer from the *silo effect*, where data analysts in different parts of the enterprise work in isolation, focusing narrowly on the data they have access to, to answer questions for their department or organization. Results produced in one area may not be consistent with those in another even if the underlying question is the same. These differences may result from using (possibly subtlety) different data, or versions of the same data, or taking a different approach to arrive at a given result. These differences can make for *interesting* cross-organization or enterprise-wide meetings where results are presented.

Level 2: Greater collaboration exists between IT and line-of-business organizations.

The Level 2 enterprise seeks greater collaboration among the traditional keepers of data (Information Technology) and the various lines of business with their data analysts and data scientists. Sharing of data and results may still be *ad hoc*, but greater collaboration helps identify data to solve important business problems and communicate results within the organization or enterprise.

Level 3: Recognized need for greater collaboration among the various players in data science projects.

With the introduction of data scientists, and the desire to make greater use of data to solve business problems, Level 3 enterprises see the need to have greater collaboration among the various players involved in or affected by data science projects. These include data scientists, business analysts, business leaders, and application/dashboard developers, among others. Collaboration takes the form of sharing, modification, and hand-off of data science work products. Work products consist of, e.g., data (raw and transformed), data visualization plots and graphs, requirements and design specifications, code written as R / Python / SQL / other scripts directly or in web-based *notebooks* (e.g., Zeppelin, Jupyter), and predictive models. Use of traditional tools such as source code control systems or object repositories with version control may be used, but inconsistently.

Level 4: Broad use of tools introduced to enable sharing, modifying, tracking, and handing off data science work products.

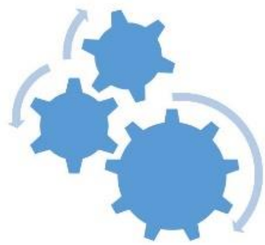
Level 4 enterprises build on the progress from Level 3, introducing tools specifically geared toward enhanced collaboration among data science project players. This includes support for sharing and modifying work products, as well as tracking changes and workflow. The ability to hand off work products within a defined workflow in a seamless and controlled manner is key. Different organizations within the enterprise may experiment with a variety of tools, which typically do not interoperate.

Level 5: Standardized tools are introduced across the enterprise to enable seamless collaboration.

While the Level 4 enterprise made significant strides in enhancing collaboration, the Level 5 enterprise standardizes on processes, methodologies, and tools to facilitate cross-enterprise collaboration among data science project players.

Methodology

What is the enterprise approach or methodology to data science?



The most often cited methodology for *data mining* – a key element of data science – is CRISP-DM. However, the breadth and growth of data science may require expanding beyond the traditional phases introduced by CRISP-DM: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The value of explicit feedback loops or expanded data awareness/access phases may also be useful. In addition, enterprise-specific workflows involving data science project players and work products may be necessary to increase productivity and derived value.

The five maturity levels of the methodology dimension are:

Level 1: Data analytics is focused on business intelligence and data visualization using an *ad hoc* methodology.

> For Level 1 enterprises, data analysts and other players typically follow no established methodology, relying instead on their experience, skills, and preferences. The focus is on business intelligence and data visualization through dashboards and reports, relying on traditional deductive query formulation.

Level 2: Data analytics are expanded to include machine learning and predictive analytics for solving business problems, but still using an *ad hoc* methodology.

> Like Level 1, Level 2 enterprises typically follow no established methodology, relying instead on player experience, skills, and preferences. However, enterprises at Level 2 supplement traditional roles such as data analysts who provide business intelligence and data visualization with data scientists who introduce more advanced data science techniques such as machine learning and predictive analytics. With the introduction of data scientists, there are implicit enhancements to the *ad hoc* data science methodology.

Level 3: Individual organizations begin to define and regularly apply a data science methodology.

> Level 3 enterprises are in the *experimental stage* where individual organizations start to define their own methodological practices or leverage existing ones, such as CRISP-DM. Goals include increasing productivity, consistency, and repeatability of data science projects while controlling risk. Data science projects may or may not effectively track performance of deployed model outcomes.

Level 4: Basic data science methodology best practices are established for data science projects.

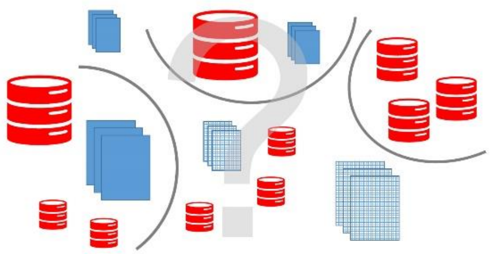
> Level 4 enterprises build on the progress from Level 3 by establishing methodology best practices throughout the enterprise. Such best practices are derived from organizational experimentation or adopted from an existing methodology. As a result of establishing best practices, the enterprise sees increased productivity, consistency, and repeatability of data science projects with reduced risk of failure.

Level 5: Data science methodology best practices are formalized across the enterprise.

Having established best practices for data science in Level 4, the Level 5 enterprise formalizes additional key aspects of data science projects, including project planning, requirements gathering / specification, and design, as well as implementation, deployment, and project assessment.

Data Awareness

How easily can data scientists learn about enterprise data resources?



The term 'awareness' can be defined as "the state or condition of being aware; having knowledge; consciousness." For *data awareness*, we might refine this definition as "having knowledge of the data that exist in an enterprise and an understanding of its contents." As the image above suggests, enterprises often have many data repositories across organizations and departments. Data may reside in databases, flat files, spreadsheets, among others, across a range of hardware, operating systems, and file systems – the *data landscape*. Moreover, data silos form where one part of the enterprise is completely unaware of the existence of data in another, let alone the meaning of that data.

Data awareness across an enterprise allows data science players, especially data scientists, the ability to browse and understand data from a *metadata* perspective. Such metadata may include textual descriptions of, e.g., tables and individual columns, key summary statistics, data quality metrics, among others. Data awareness is essential to increase productivity, but also to inventory data assets and enable an enterprise to move toward "a single version of the truth."

The five maturity levels of the data awareness dimension are:

Level 1: Users of data have no systematic way of learning what data assets are available in the enterprise.

Enterprises at Level 1 are often in the dark when it comes to understanding the data resources that may exist across the enterprise. Data may be siloed in spreadsheets or flat files on employee machines, or stored in departmental or application-specific databases. No *map* of the *data landscape* exists to assist in finding data of interest, moreover, the enterprise has not prioritized this as a need.

Level 2: Data analysts and data scientists seek additional data sources through "key people" contacts.

The Level 2 enterprise has “awakened” to the need for and benefits of finding the right data. As data analysts and data scientists take on more analytically interesting projects, the search for data ensues on a personal level - individually contacting data owners or others “in the know” within the enterprise to understand what data exist. A significant amount of time is lost trying to understand what data exist, how to interpret them, and their quality.

Level 3: Existing enterprise data resources are cataloged and assessed for quality and utility for solving business problems.

The Level 3 enterprise sees the need for making it easier for data science players to find data and have greater confidence in their quality for solving business problems. *Ad hoc* metadata catalogs begin to emerge which make it easier to understand what data are available, however, such catalogs are non-standard, not integrated, and dispersed across the enterprise.

Level 4: Enterprise introduces metadata management tool(s).

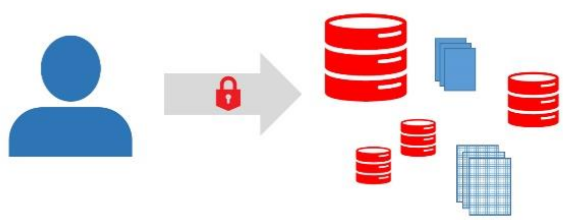
The Level 4 enterprise builds on the progress from Level 3 by introducing metadata management tools where data scientists and others can discover data resources available to solve critical business problems. Since the enterprise is just starting to take metadata seriously, different departments or organizations within an enterprise may use different tools. While an improvement for data scientists, the metadata models across tools are not integrated, so multiple tools may need to be consulted.

Level 5: Enterprise standardizes on a metadata management tool and institutionalizes its use for all data assets.

The Level 5 enterprise has fully embraced the value of integrated metadata and facilitating the maintenance and organization of that metadata through effective tools. All data assets are curated for quality and utility with full metadata descriptions to enable efficient data identification and discovery across the enterprise. Data science players' productivity and project quality increase as they can now easily find available enterprise data.

Data Access

How do data analysts and data scientists request and access data?
How is data access controlled, managed, and monitored?



When we consider 'data access,' one definition refers to "software and activities related to storing, retrieving, or acting on data housed in a database or other repository" normally coupled with *authorization* – who is permitted to access what – and *auditing* – who accessed what, when, and from where. As discussed below, data access can be provided with little or no control such as when handing someone a memory stick, or strict access control through secure database authentication and computer network authentication. Data access takes into account not only the user side, but also the ability of administrators to effectively manage the data access life cycle – from initial request and granting privileges, to revoking privileges and post-use data cleanup.

The five maturity levels of the data access dimension are:

Level 1: Data analysts typically access data via flat files obtained explicitly from IT or other sources.

Data science players at Level 1 enterprises use what has historically been called the *sneakernet*. If you need data, you walk over to the data owners, get a copy on a hard drive or memory stick, and load it onto your local machine. This, of course, has morphed into emailing requests to data owners and either getting back requested data via email, drop boxes, or FedEx'd memory sticks or hard drives. Providing access to data in this manner is clearly not secure. Further, obtaining the "right" data is unlikely to occur on the first try, so multiple iterations may be needed with data custodians – the *data request cycle* – which results in delays, frustration, and even annoying those data custodians.

Level 2: Data access is available via direct programmatic database access.

In Level 2 enterprises, the *sneakernet* is recognized as insecure and inefficient. Moreover, since much of enterprise data is stored in databases, authorization and programmatic access is more readily enabled. With direct access to databases via convenient APIs (ODBC, R and Python packages, etc.), more data can be made available to data science players, thereby shortening the data request cycle. However, any processing beyond what is possible in the data repository/environment itself, e.g., SQL for relational databases, still requires data to be pulled to the client machine, which can have security implications.

Level 3: Data scientists have authenticated, programmatic access to large volume data, but database administrators struggle to manage the data access life cycle.

The Level 3 enterprise is experiencing data access *growing pains*. Data scientists now have access to large volume data and want to use more if not all of that data in their work. Database administrators are inundated with requests for both broad (multi-schema) and narrow (individual table) data access. Ensuring individuals have proper approvals for accessing the data they need and possibly implementing *data masking* causes data access request backlogs. The Level 3 enterprise has also started to supplement traditional structured database data with new "big data" repositories, e.g., HDFS, NoSQL, etc. These even greater volumes of data include anything from social media data to sensor, image, text, and voice data.

Level 4: Data access is more tightly controlled and managed with identity management tools.

While enterprises in some industries, e.g., Finance, will have addressed access control to varying degrees, when addressing data access more broadly, the Level 4 enterprise understands the importance of end-to-end life cycle management of user identities and begins introducing tools to strengthen security and simplify compliance as appropriate. A goal for Level 4 enterprises is to make it easier for data science players to request and receive access to data, while also making it easier for administrators to manage, especially with the introduction of more big data repositories. An enterprise-wide self-service access request web application may be used to facilitate requesting and granting data access. Ideally, this would be integrated with the metadata management tool used for data awareness.

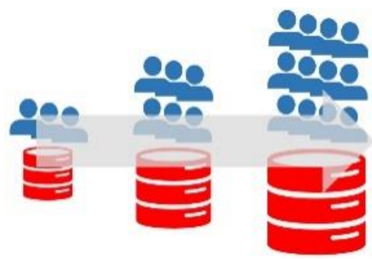
Level 5: Data access lineage tracking enables unambiguous data derivation and source identification.

The Level 5 enterprise has standardized on identity management and auditing to support secure data access, and now focuses on the ability to answer the question "what is the source of the data that produced this result?" Even in enterprises that leverage an enterprise data warehouse, data may still be replicated to other databases, or various gateways leveraged to give transparent access to remote data. The Level 5 enterprise enables tracking the derivation of data science work products – their *lineage* – with verification of actual data sources.

ORACLE

Scalability

Do the tools scale and perform for data exploration, preparation, modeling, scoring, and deployment? As data, data science projects, and the data science team grow, is the enterprise able to support these adequately?



The term 'scalability' can be defined as the "capability of a system, network, or process to handle a growing amount of work, or its potential to be enlarged to accommodate that growth." Scalability with respect to data science needs to reflect the hardware and software aspects, as well as the people and process aspects. This includes several factors: data volume (number of rows, columns, and overall bytes), algorithm design and implementation (parallel, distributed, memory efficient) for data preparation and model building and scoring, hardware (RAM, CPU, storage), volume and rate of data science work products produced, number of data science players and projects, and workflow complexity.

The five maturity levels of the scalability dimension are:

Level 1: Data volumes are typically "small" and limited by desktop-scale hardware and tools, with analytics performed by individuals using simple workflows.

> Level 1 enterprises perform analytics on data that can fit and be manipulated in memory, typically on desktop hardware, and possibly using open source tools. At Level 1, data volumes are such that loading data from flat files or programmatically from databases does not introduce problematic latency. Similarly, algorithm efficiency in terms of memory consumption or ability to take advantage of multiple CPUs is not a significant issue. Data science work products are produced at a rate that taxes neither individuals nor infrastructure.

Level 2: Data science projects take on greater complexity and leverage larger data volumes.

> In Level 2 enterprises, data science players are taking on more projects of greater complexity that require more data. This increase in data volume introduces increasingly intolerable latency due to data movement, and highlights inadequate hardware resources and inefficient algorithm implementations. The need to produce more data science work products more frequently also taxes existing hardware resources. The Level 2 enterprise begins exploring scalable tools for processing data where they reside instead of relying on data movement and tools that can enhance the use of open source tools and packages. Data scientists resort to data sampling to address tool limitations.

Level 3: Individual groups adopt varied scalable data science tools and provide greater hardware resources for data scientist use.

> The Level 3 enterprise is addressing its data science growing pains experienced at Level 2 by adopting tools that minimize latency due to data movement, have parallel distributed algorithm implementations, and provide

infrastructure for leveraging open source tools. These new tools enable data scientists to use more if not all desired data in their analytics, however, there is no standard suite of tools across the enterprise and the various tools do not facilitate collaboration. An increase in available hardware resources (on-premises or cloud) for solving bigger and more complex data science problems yields significant productivity gains for the data science team.

Level 4: Enterprise standardizes on an integrated suite of scalable data science tools and dedicates sufficient hardware capacity to data science projects.

Having explored and test-driven various data science tools, the Level 4 enterprise standardizes on an integrated suite of scalable tools that enables data science players to realize full-scale data science projects. Data science projects, and data scientists in particular, have sufficient hardware resources (on-premises or cloud) for both development and production.

Level 5: Data scientists have on-demand access to elastic compute resources both on premises and in the cloud with highly scalable algorithms and infrastructure.

The Level 5 enterprise focuses on elastic compute resources for data scientists. As data volumes increase, data science projects benefit from being able to quickly and easily increase/decrease compute resources, which in turn expedites data exploration, data preparation, machine learning model training, and data scoring - whether for individual models or *massive predictive modeling* involving thousands or even millions of individual models. Elastic compute resources can eliminate the need for dedicating resources for peak demand requirements. Alternatively, cloud-at-customer solutions can provide benefits while meeting regulatory or data privacy requirements. The combination of scalable algorithms and infrastructure with elastic compute resources enables the enterprise to meet time-sensitive business objectives while minimizing cost.

Asset Management

How are data science assets managed and controlled?

Assets are typically both tangible and intangible things of value. For this discussion, we will consider the array of data science work products as assets and can define 'asset management' at a high level as "any system that monitors and maintains things of value to an entity or group." As we introduced earlier, work products consist of, e.g., data (raw and transformed), data visualization plots and graphs, requirements and design specifications, code written as R / Python / SQL / other scripts directly or in web-based *notebooks* (e.g., Zeppelin, Jupyter), predictive models, virtual machine / container images, among others. In this context, asset management should cover the full asset life cycle - from creation to retirement. Throughout the life cycle, the need for asset storage / backup / recovery, metadata-based search and retrieval, security (e.g., privilege-based access control, auditability), versioning, archiving, and lineage must be addressed – basically *governance*. Specific to data science is the need for *model management*, which encompasses, e.g., model life cycle, governance, repeatability, monitoring, and reporting.

The five maturity levels of the asset management dimension are:

Level 1: Analytical work products are owned, organized, and maintained by individual data science players.

> Data science players at Level 1 enterprises are essentially "winging it," taking an *ad hoc* approach to asset management. Players are responsible for maintaining their data science work products, typically on their local machines, which may or may not be backed up or secure. Asset loss and an inability to reproduce results are not uncommon. Across the enterprise, data science work products are "hidden" on individual machines, with no effective way to perform cross-project search.

Level 2: Initial efforts are underway to provide security, backup, and recovery of data science work products.

> The Level 2 enterprise recognizes the need to manage data science work products. This typically begins with organization-based repositories that provide storage with backup and recovery to reduce asset loss, as well as security to control access.

Level 3: Data Science work product governance is systematically being addressed.

> The Level 3 enterprise begins to see data science work products as an important corporate asset. As such, tools and procedures are introduced to centrally manage assets throughout their life cycle. As the enterprise expands its data science effort with machine learning models, the need for *model management* also gains visibility. The need to determine which data and processes were used to produce data science work products is gaining recognition with steps being taken to answer basic questions definitively, e.g., on what is this result based?

Level 4: Data science work product governance is firmly established at the enterprise level with increasing support for model management.

> The Level 4 enterprise has adopted best practices for data science work product governance. Data science players as well as the overall enterprise reaps productivity gains through being able to easily locate, execute, reproduce, and enhance project content. The question of "how was this result produced and on what data?" can readily be answered.

Level 5: Systematic management of all data science work products with full support for model management.

> The Level 5 enterprise surpasses the Level 4 enterprise by having introduced tools and procedures that support model management. As data science projects are deployed, their outcomes are fully monitored with reporting on value provided to the enterprise. Such outcomes are factored back into each project forming a closed loop – ensuring data science projects continue to provide value based on current relevant data and trends.

Tools

What tools are used within the enterprise for data science?
Can data scientists take advantage of open source tools in combination with production quality infrastructure?

A wide range of tools support data science ranging from open source to proprietary, relational database to "big data" platforms, simple analytics to complex machine learning. Tools may support isolated activities or be highly collaborative, and enable modeling in the small to massive predictive modeling with full model management. Orthogonal to each of these is the scale at which these tools can perform. Some tools and algorithm implementations will perform well for small or even moderate sized data, but fail or become unusable when presented with larger data volumes. For this, special parallel, distributed implementations are necessary to leverage multi-node/processor machines and machine clusters.

Seldom will a single tool provide all required functionality – usually provided by a mix of commercial and open source tools. However, enterprises require commercial support for the tools adopted. As a result, commercial tools that integrate with open source tools and provide support for data-parallel and task-parallel execution along with ease of deployment are highly desired.

The five maturity levels of the tools dimension are:

Level 1: An *ad hoc* array of non-scalable tools is predominantly used for isolated data analysis on desktop machines.

> Data science players at Level 1 use traditional desktop tools for data analysis, relying heavily on spreadsheet-based tools along with various open source tools for analytics and visualization.

Level 2: Enterprise manages data through database management systems and relies on extensive open source libraries along with specialized commercial tools.

> Level 2 enterprises, taking data management more seriously, introduce relational database management software tools. Data science projects also benefit from the broader open source package ecosystem for advanced data exploration, statistical analysis, visualization, and predictive analytics / machine learning. However, at Level 2, there is little integration between commercial and open source tools, and performance and scalability are an issue for data science projects.

Level 3: Enterprise seeks scalable tools to support data science projects involving large volume data.

> Data science projects at Level 3 enterprises are hindered by performance and scalability of existing software and environment. A concerted effort is made to evaluate and acquire commercial and open source tools with a range of scalable machine learning algorithms and techniques to complement open source techniques and facilitate production deployment. Data science players may begin to explore Big Data platforms to address new sources of high volume data, scalability, and cost reduction. Cloud-based tools are also under review. As data science projects grow in complexity involving larger team efforts, tools supporting collaboration become a recognized need.

ORACLE

Level 4: Enterprise standardizes on a suite to tools to meet data science project objectives.

The Level 4 enterprise understands the needs of data science players and projects to meet business objectives. Enhanced productivity requires scalable tools that support collaboration and work with data from a wide range of sources. Automation and integration play a major role in enhancing productivity, so tools that avoid paradigm shifts and automate tasks in data exploration, preparation, machine learning, and graph and spatial analytics are particularly valuable. Adopted tools are available or function across multiple platforms, including on-premises and cloud. As machine learning models have become a focal point for data science projects, adopted tools must support full model management.

Level 5: Enterprise regularly assesses state-of-the-art algorithms, methodologies, and tools for improving solution accuracy, insights, and performance, along with data scientist productivity.

Level 5 enterprises optimize their data science tool environment. Having understood what is required for effective data science projects and data science player productivity at Level 4, enterprises work with tool providers to further enhance those tools to meet business objectives.

Deployment

How easily can data science work products be placed into production to meet timely business objectives?

Data science comes with the expectation that amazing insights and predictions will transform the business and take the enterprise to a new level of performance. Too often, however, data science projects fail to "lift-off," resulting is significant opportunity cost for the enterprise. A data scientist may produce a predictive model with high accuracy, however, if those scores are not effectively put into production, i.e., deployed, or deployment is significantly delayed, desired gains are not realized.

A more general definition of 'deployment' that seems relevant in this discussion is "the action of bringing resources into effective action." The *resources* in this context refer to data science work products such as machine learning models, visualizations, statistical analyses, etc. *Effective action* is to deliver these resources in a way that they provide business benefit: timely insights presented in interactive dashboards, predictions affecting which actions enterprises will undertake with respect to customers, employees, assets, etc.

For data science in general, and machine learning in particular, much of the deployment mechanism - or *plumbing* - is the same across projects. Yet, enterprises often find individual projects re-inventing deployment infrastructure, requiring logic for data access, spawning separate analytic engines, and recovery along with (often missing) rigorous testing. Leveraging tools that provide such plumbing can greatly reduce overhead and risk in deploying data science projects.

The five maturity levels of the deployment dimension are:

Level 1: Data science results have limited reach and hence provide limited business value.

> At Level 1 enterprises, results from data science projects often take the form of insights documented in slide presentations or textual reports. Data analyses, visualizations, and even predictive models may provide guidance for human decision making, but such results must be manually conveyed on a per-project basis.

Level 2: Production model deployment is seen as valuable, but often involves reinventing infrastructure for each project.

> In Level 2 enterprises, the realization that machine learning models can and should be leveraged in front-line applications and systems takes hold. Some insights may be explicitly coded into application or dashboard logic, however, the time between model creation and deployment can significantly impact model accuracy. This *deployment latency* occurs when the patterns in data used for model building diverge from current data used for scoring. Moreover, manually coding, e.g., predictive model coefficients for scoring in C, Java, or even SQL, for easier integration with existing applications or dashboards takes developer time and can result in coding errors that only rigorous code reviews and testing can reveal. As a result, enterprises incur costs for data science projects, but do not fully realize potential project benefits.

Level 3: Enterprise begins leveraging tools that provide simplified, automated model deployment, inclusive of open source software and environments.

> As more data science projects are undertaken, the Level 3 enterprise realizes that one-off deployment approaches waste valuable development resources, incurs deployment latency that reduces model effectiveness, and increases project risk. In today's internet-enabled world, patterns in data, e.g., customer preferences, can change overnight requiring enterprises to have greater agility to build, test, and deploy models using the latest data. Enterprises at Level 3 begin to leverage tools that provide the needed infrastructure to support simplified and automated model deployment.

Level 4: Increased heterogeneity of enterprise systems requires cross-platform model deployment, with a growing need to incorporate models into streaming data applications.

> The Level 4 enterprise has a combination of database, Hadoop, Spark, and other platforms for managing data and computation. Increasingly, the enterprise needs models and scripts produced in one environment to be deployed in another. This increases the need for tools that enable exporting models for use in a scoring engine library that can be easily integrated into applications. Level 4 enterprises seek tools that facilitate script and model deployment in real-time or streaming analytics situations as they begin to use data science results involving *fast data*.

Level 5: Enterprise has realized benefits of immediate data science work product (re)deployment across heterogeneous environments.

> The Level 5 enterprise has adopted a standard set of tools to support deployment of data science work products across all necessary environments. Machine learning models and scripts created in one environment can immediately be deployed and refreshed (redeployed) with minimal latency.

Summary Table

Enterprises embracing data science as a core competency may want to evaluate at what level they have achieved relative to each dimension - in some cases, an enterprise may straddle more than one level. As a next step, the enterprise may use this maturity model to identify a level in each dimension to which they aspire, or fashion a new Level 6.

Data Science Maturity Model						
	Questions	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Strategy	What is the enterprise business strategy for data science?	Enterprise has no governing strategy for applying data science	Enterprise is exploring the value of data science as a core competency	Enterprise recognizes data science as a core competency for competitive advantage	Enterprise embraces a data-driven approach to decision making	Data are viewed as an essential corporate asset - data capital
Roles	What roles are defined and developed in the enterprise to support data science activities?	Traditional data analysts explore and summarize data using deductive techniques	The data scientist role is introduced to begin leveraging advanced, inductive techniques	Chief Data Officer (CDO) role is introduced to help manage data as a corporate asset	Data scientist career path is codified and standardized across the enterprise	Chief Data Science Officer (CDSO) role introduced
Collaboration	How do data scientists collaborate among themselves and other data science players to evolve and hand-off data science work products?	Data analysts often work in silos, performing work in isolation and storing data and results in local environments	Greater collaboration exists between IT and line-of-business organizations	Recognized need for greater collaboration among the various players in data science projects	Broad use of tools introduced to enable sharing, modifying, tracking, and handing off data science work products	Standardized tools are introduced across the enterprise to enable seamless collaboration
Methodology	What is the enterprise approach or methodology to data science?	Data analytics is focused on business intelligence and data visualization using an ad hoc methodology	Data analytics are expanded to include machine learning and predictive analytics for solving business problems, but still using ad hoc methodology	Individual organizations begin to define and regularly apply a data science methodology	Basic data science methodology best practices are established for data science projects	Data science methodology best practices are formalized across the enterprise
Data Awareness	How easily can data scientists learn about enterprise data resources?	Users of data have no systematic way of learning what data assets are available in the enterprise	Data analysts and data scientists seek additional data sources through "key people" contacts	Existing enterprise data resources are cataloged and assessed for quality and utility for solving business problems	Enterprise introduces metadata management tool(s)	Enterprise standardizes on a metadata management tool and institutionalizes its use for all data assets

Data Access	How do data analysts and data scientists request and access data? How is data access controlled, managed, and monitored?	Data analysts typically access data via flat files obtained explicitly from IT or other sources	Data access is available via direct programmatic database access	Data scientists have authenticated, programmatic access to large volume data, but database administrators struggle to manage the data access life cycle	Data access is more tightly controlled and managed with identity management tools	Data access lineage tracking enables unambiguous data derivation and source identification
Scalability	Do the tools scale and perform for data exploration, preparation, modeling, scoring, and deployment? As data, data science projects, and the data science team grow, is the enterprise able to support these adequately?	Data volumes are typically "small" and limited by desktop-scale hardware and tools, with analytics performed by individuals using simple workflows	Data science projects take on greater complexity and leverage larger data volumes	Individual groups adopt varied scalable data science tools and provide greater hardware resources for data scientist use	Enterprise standardizes on an integrated suite of scalable data science tools and dedicates sufficient hardware capacity to data science projects	Data scientists have on-demand access to elastic compute resources both on premises and in the cloud with highly scalable algorithms and infrastructure
Asset Management	How are data science assets managed and controlled?	Analytical work products are owned, organized, and maintained by individual data science players	Initial efforts are underway to provide security, backup, and recovery of data science work products	Data science work product governance is systematically being addressed	Data science work product governance is firmly established at the enterprise level with increasing support for model management	Systematic management of all data science work products is used with full support for model management.
Tools	What tools are used within the enterprise for data science objectives? Can data scientists take advantage of open source tools in combination with production quality infrastructure?	An *ad hoc* array of non-scalable tools is predominantly used for isolated data analysis on desktop machines	Enterprise manages data through database management systems and relies on extensive open source libraries along with specialized commercial tools	Enterprise seeks scalable tools to support data science projects involving large volume data	Enterprise standardizes on a suite to tools to meet data science project objectives	Enterprise regularly assesses state-of-the-art algorithms, methodologies, and tools for improving solution accuracy, insights, and performance, along with data scientist productivity
Deployment	How easily can data science work products be placed into production to meet timely business objectives?	Data science results have limited reach and hence provide limited business value	Production model deployment is seen as valuable, but often involves reinventing infrastructure for each project	Enterprise begins leveraging tools that provide simplified, automated model deployment, inclusive of open source software and environments	Increased heterogeneity of enterprise systems requires cross-platform model deployment, with a growing need to incorporate models into streaming data applications	Enterprise has realized benefits of immediate data science work product (re)deployment across heterogeneous environments

Appendix A: Key Resources

[Data Science Maturity Model Enterprise Assessment spreadsheet]

[Data Science Maturity Model blog series]

Oracle is committed to developing practices and products that help protect the environment

Integrated Cloud Applications & Platform Services

A DATA SCIENCE MATURITY MODEL FOR ENTERPRISE ASSESSMENT
July 2018 v2.0