The Data Science Lifecycle

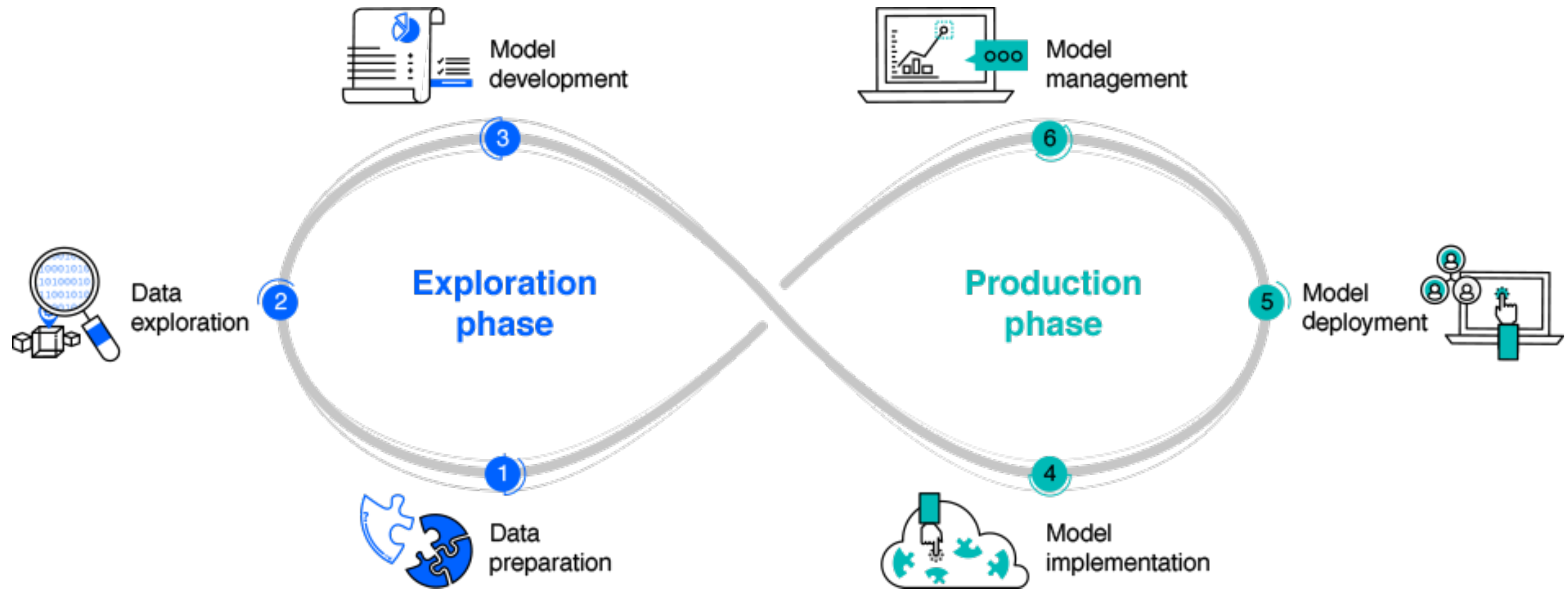
From experimentation to production-level data science

It's difficult for today's data science teams to respond fast enough to the demands of their business. With the explosion of data from multiple sources and formats, trying to develop models and putting them into production quickly is a huge challenge for many teams.

This guide provides an overview of the typical data science lifecycle, common challenges organizations encounter and how IBM® Watson Studio can address them-enabling your data science team to accelerate and optimize the value of analytics results throughout your organization.

IBM Watson Studio embeds the practice of data science into your business, allowing you to manage the entire data science lifecycle from research to production. IBM Watson Studio is a platform built for the enterprise that allows teams to explore, build and put data science practice into production faster. Once you adopt the right process and platform, you can fully realize the benefits data science and machine learning can bring to your business.

IBM can move you from the exploration phase to the production phase faster



Exploration phase

Production phase

1 Data preparation

2 Data exploration

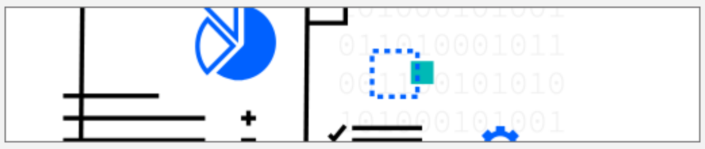
3 Model development

4 Model implementation

5 Model deployment

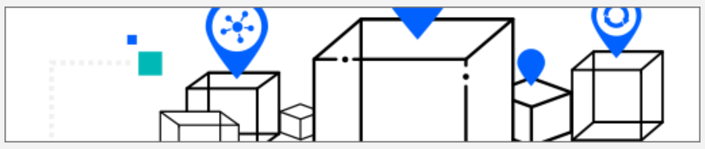
6 Model management

Exploration Phase



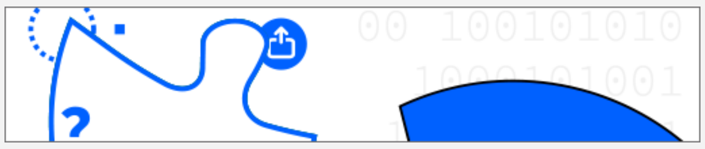
Data preparation

Data science projects start by asking the right business questions and collecting and preparing data. Success in the later phases is dependent on this early stage.



Data exploration

Once your data is in the right format to work with, you can conduct the next phase in the data analysis process. This initial exploration of the dataset is critical.



Model development

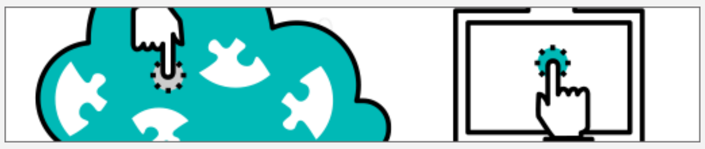
Here, the data you've prepared is brought into the data science toolset and the results begin to shed some light on previously identified business problems.

Learn more

Learn more

Learn more

Production Phase



Model implementation

Once you've built and chosen your model, this stage helps you evaluate and understand its quality to ensure it fully addresses the business problem.

Learn more

Model deployment

Upon development and approval by business sponsors, you're ready to deploy your model into the production environment or a comparable test environment.

Learn more

Model management

Model management must be a continuous process to ensure optimal performance over time. This stage helps you monitor model creation, use and decay.

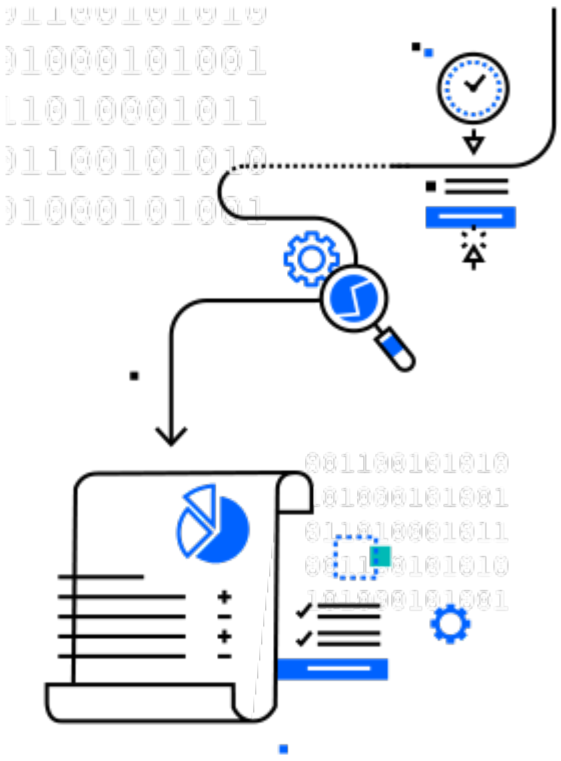
Learn more

Exploration phase — Data preparation

A typical data science project starts with asking the right business question and collecting and preparing data. Success in the later phases is dependent on what occurs in these earlier phases. It is not surprising, therefore, that poor quality data will lead to poor model performance. That is why data science teams must ensure the quality of the data they use as input for predictive modeling.

Data preparation is one of the most important and often time-consuming phases of a data science project. In fact, it is estimated that data preparation usually takes 50-70% of a project's time and effort, leaving data science teams very little time to uncover new insights and deliver them to decision makers. With the amount of stored data now too vast for manual verification, today's data scientists require automated processes to prepare data quickly and accurately.

Often referred to as "data wrangling," data preparation involves cleaning the data and reshaping it into a usable form for performing data science. Examples of common data preparation activities include dealing with non-standard, unstructured or inconsistent data and combining data from different sources and formats. Tools that make it easier to access and manipulate data sources on their own, without technical proficiency in programming languages or coding expertise, can encourage more users to take advantage of enterprise or external data sources.



Exploration phase — Data exploration

Once your data is in the right format to work with, you can conduct the next step in the data analysis process: data exploration. This initial exploration of the dataset is critical because it helps data scientists illuminate previously unknown patterns, relationships, or other actionable findings. Some helpful questions to ask at this point include:

Which attributes seem promising for further analysis?

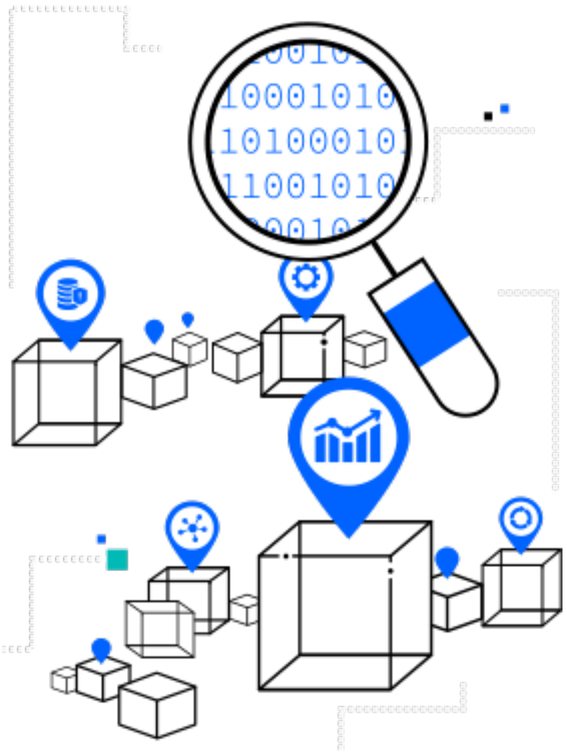
Has the exploration revealed new characteristics about the data?

How have these explorations changed any initial hypotheses?

Can a specific subset of the data be used later?

Has the data exploration altered the project goals?

Data scientists commonly use data visualizations to quickly view relevant features of their datasets and identify variables that are likely to result in interesting observations. By displaying data graphically-for example, through scatter plots or bar charts-users can see if two or more variables correlate and determine if they are good candidates for more in-depth analysis.

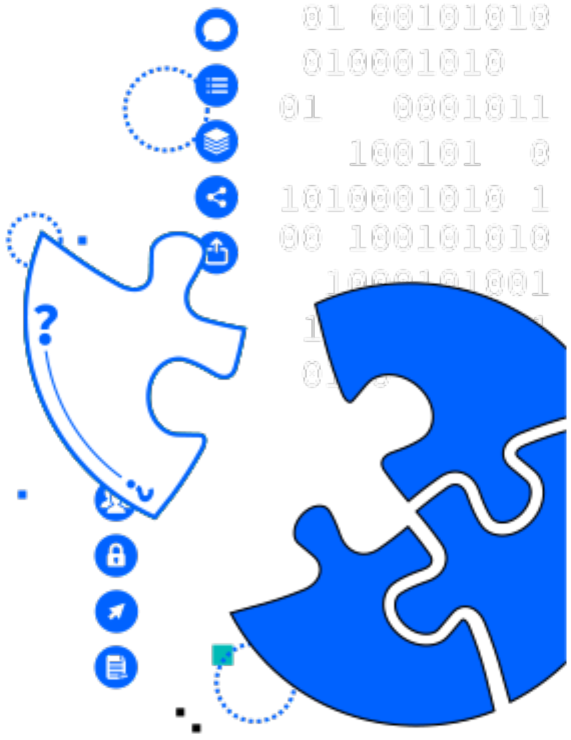


Exploration phase — Model development

This is the point at which your hard work begins to pay off. The data you spent time preparing is brought into the data science toolset, and the results begin to shed some light on the business problem posed during the early stages of the project.

Model development is usually conducted in multiple iterations. Typically, data scientists run several models using default parameters and then fine-tune the parameters or revert to the data preparation phase for manipulations required by their model of choice. It's rare for an organization's question to be answered satisfactorily with a single algorithm and a single execution. This is what makes data science so interesting. There are many ways to look at a given problem, and today there are a wide variety of tools to help you do that.

Although you may already have some idea about which types of models are most appropriate for your organization's needs, now is the time to make some decisions about which ones to use. Determining the most appropriate model will typically be based on the data types available, your project goals, and specific requirements for data sizes or types.



Exploration phase — How IBM Watson Studio can help

With IBM Watson Studio, you can quickly transform large amounts of raw data into consumable, quality information that's ready for analytics using the Data Refinery feature. Data Refinery helps you save time during data preparation by making it easy to cleanse, shape and deliver data to people across your business. You can also improve data exploration by using built-in charts, statistics and interactive visualizations to better understand the quality and distribution of your data.

IBM Watson Studio also provides a flexible work space that allows business and technical users to work within the same environment, allowing you to increase collaboration across teams. You can train models using the top open source tools or use a no-code, visual modeling interface alongside the most popular deep learning frameworks. This allows you to tap the skills of your entire data science team so you can boost productivity and accelerate decision making.

In the exploration phase, data science teams struggle with:

- Data ingestion & data wrangling
- Discovering relationships with data
- Data access & data movement

How IBM Watson Studio can help you:

- Data preparation & data exploration
- Build models where your data resides (on-premises, private or public cloud)
- Accommodate skills of coders and non-coders

Learn how IBM Watson Studio can help with the exploration phase of data science

Watch the Video

Explore, build, deploy and manage data science at scale.

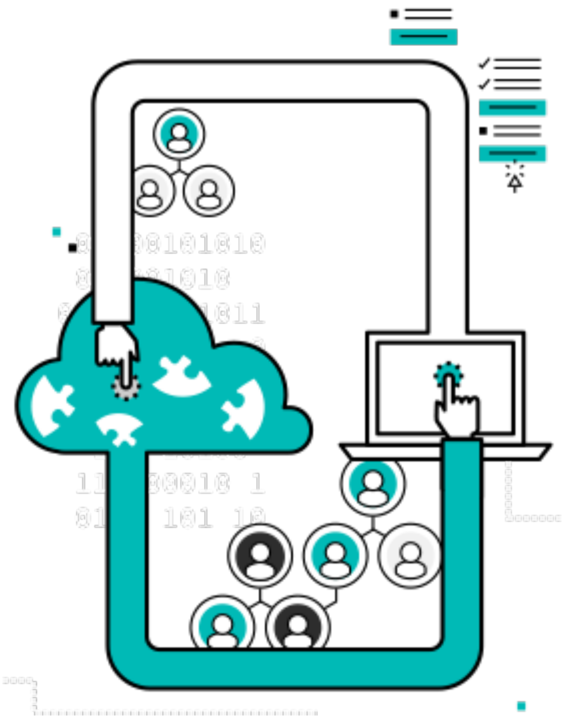
Get started today

Production phase — Model implementation

Your models have been built and you've selected the best one. Before actual deployment, however, you need to evaluate your model to understand its quality and ensure that it fully addresses the business problem. Model implementation entails computing various diagnostic measures and other outputs such as tables and graphs, enabling the data scientist to interpret the model's quality and its efficacy in solving the problem.

For a predictive model, data scientists use a testing set that is independent of the training set but follows the same probability distribution and has a known outcome. The testing set is used to evaluate the model so it can be refined as needed. Sometimes the final model is applied also to a validation set for a final assessment.

Data scientists may also assign statistical significance tests to the model as further proof of its quality. This additional proof may be instrumental in justifying model implementation or taking actions when the stakes are high-such as an expensive medical protocol or a critical airplane flight system.



Production phase — Model deployment

Once a satisfactory model has been developed and is approved by the business sponsors, you are ready to deploy it into the production environment or a comparable test environment.

Deploying a model into an operational business process is usually an IT function, although it can involve additional groups. However, the process is not as simple as simply handing the model off to your IT team. Testing the model first is necessary to identify any dependencies in the production environment. In addition, data science teams need to ensure models receive the correct production data and send the scores to the right place, and that the system must be set up for monitoring and scalability.

Questions to ask at this stage of the process include:

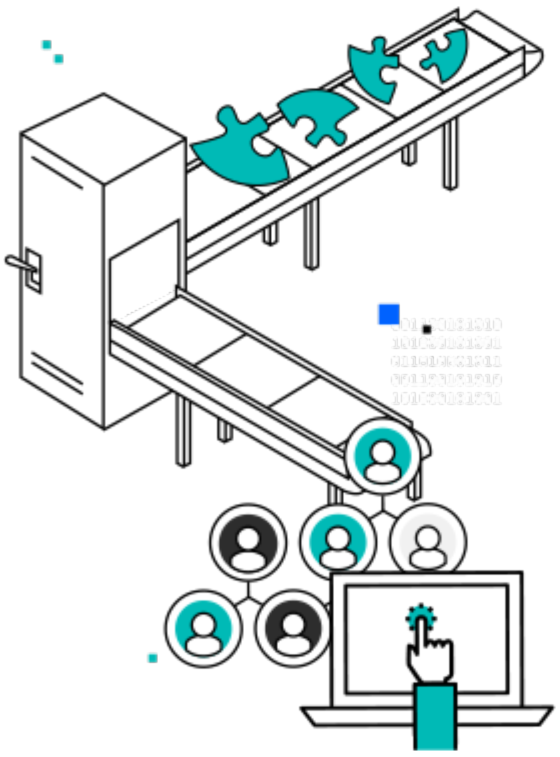
What data does the model expect to ingest and produce?

What infrastructure is required to run this model at scale?

Does the model have everything it needs to execute?

How easy is it to identify points of failure if the model is not running properly?

Models are normally deployed in a limited way until their performance has been fully evaluated. Deployment may be as simple as generating a report with recommendations, or as involved as embedding the model in a complex workflow and scoring process managed by a custom application.



Production phase — Model management

During deployment and integration of modeling results, model management must be a continuous process to ensure optimal performance over time. This means knowing exactly where each model is in the lifecycle, how old the model is, who developed it, and who is using it for which application. Model version control, which includes event logging and change tracking to understand how the model evolves over time, is another critical requirement.

Data science must also be concerned with model decay, and continuously gather metrics to determine when a model should be refreshed or replaced. For example, a model deployed to increase customer retention among high-value customers will need to be revised once a level of retention is reached.

Production phase — How IBM Watson Studio helps with the production phase

The production phase of data science is where most enterprise teams struggle. However, with IBM Watson Studio you train and deploy models easily in a secure environment. You have the flexibility to train your models behind your own firewall (where your sensitive data resides), in the cloud, or remotely to a Spark cluster in HortonWorks Data Platform, and then easily deploy them on the cloud. IBM is one of the few vendors that provides this hybrid cloud environment.

Eventually, you'll need to automate the management of what is in production as the number of your deployed models increases. IBM's Watson Studio enables you to constantly monitor the health and accuracy of your models with the model management dashboard. This feature allows data scientists or IT to review, and even schedule evaluation of deployed models so you know when it's time to retrain or retire a model. So instead of spending your time managing existing models, your team can focus on driving results for your business.

In the production phase, data science teams struggle with:

- Deployment and maintaining model accuracy
- Model validation
- Production workflows
- Scalability

How IBM Watson Studio can help you:

- Build a predictable and repeatable path for model deployment
- Consult a central repository of metrics, scoring and version control for each model
- Easily monitor deployed models through reports based on performance indicators
- Quickly refresh models with new data

Learn how IBM Watson Studio can help with the production phase of data science

Watch the Video

Deliver continuous value to your business

With a clearly defined data science project lifecycle, your data science team will become a well-oiled machine, continuously delivering new projects and exponential value to the business. You can build more models, put the best ones into production faster, and derive greater value from your organization's data.

Data science in an enterprise practice is very complicated, and enterprises can struggle delivering value to their business owners. Particularly when these systems are done in a one-off fashion, it complicates delivering value to your business. IBM's Watson Studio speeds and simplifies the lifecycle so that your enterprise teams can scale and focus on model development that drives business impact.

Explore, build, deploy and manage data science at scale.

Get started today