Advanced analytics

02/12/2018 • 6 minutes to read • Contributors 👤 🟤 🟢 🟠

In this article

Advanced analytics goes beyond the historical reporting and data aggregation of traditional business intelligence (BI), and uses mathematical, probabilistic, and statistical modeling techniques to enable predictive processing and automated decision making.

Advanced analytics solutions typically involve the following workloads:

- Interactive data exploration and visualization
- Machine Learning model training
- Real-time or batch predictive processing

Most advanced analytics architectures include some or all of the following components:

- **Data storage**. Advanced analytics solutions require data to train machine learning models. Data scientists typically need to explore the data to identify its predictive features and the statistical relationships between them and the values they predict (known as a label). The predicted label can be a quantitative value, like the financial value of something in the future or the duration of a flight delay in minutes. Or it might represent a categorical class, like "true" or "false," "flight delay" or "no flight delay," or categories like "low risk," "medium risk," or "high risk."

- **Batch processing**. To train a machine learning model, you typically need to process a large volume of training data. Training the model can take some time (on the order of minutes to hours). This training can be performed using scripts written in languages such as Python or R, and can be scaled out to reduce training time using distributed processing platforms like Apache Spark hosted in HDInsight or a Docker container.

- **Real-time message ingestion**. In production, many advanced analytics feed real-time data streams to a predictive model that has been published as a web service. The incoming data stream is typically captured in some form of queue and a stream processing engine pulls the data from this queue and applies the prediction to the input data in near real time.

- **Stream processing**. Once you have a trained model, prediction (or scoring) is typically a very fast operation (on the order of milliseconds) for a given set of features. After capturing real-time messages, the relevant feature values can be passed to the predictive service to generate a predicted label.

- **Analytical data store**. In some cases, the predicted label values are written to the analytical data store for reporting and future analysis.

- **Analysis and reporting**. As the name suggests, advanced analytics solutions usually produce some sort of report or analytical feed that includes predicted data values. Often, predicted label values are used to populate real-time dashboards.

- **Orchestration**. Although the initial data exploration and modeling is performed interactively by data scientists, many advanced analytics solutions periodically re-train models with new data — continually refining the accuracy of the models. This retraining can be automated using an orchestrated workflow.

Machine learning

Machine learning is a mathematical modeling technique used to train a predictive model. The general principle is to apply a statistical algorithm to a large dataset of historical data to uncover relationships between the fields it contains.

Machine learning modeling is usually performed by data scientists, who need to thoroughly explore and prepare the data before training a model. This exploration and preparation typically involves a great deal of interactive data analysis and visualization — usually using languages such as Python and R in interactive tools and environments that are specifically designed for this task.

In some cases, you may be able to use [pretrained models](#) that come with training data obtained and developed by Microsoft. The advantage of pretrained models is that you can score and classify new content right away, even if you don't have the necessary training data, the resources to manage large datasets or to train complex models.

There are two broad categories of machine learning:

- **Supervised learning**. Supervised learning is the most common approach taken by machine learning. In a supervised learning model, the source data consists of a set of *feature* data fields that have a mathematical relationship with one or more *label* data fields. During the training phase of the machine learning process, the data set includes both features and known labels, and an algorithm is applied to fit a function that operates on the features to calculate the corresponding label predictions. Typically, a subset of the training dataset is held back and used to validate the performance of the trained model. Once the model has been trained, it can be deployed into production, and used to predict unknown values.

- **Unsupervised learning**. In an unsupervised learning model, the training data does not include known label values. Instead, the algorithm makes its predictions based on its first exposure to the data. The most common form of unsupervised learning is *clustering*, where the algorithm determines the best way to split the data into a specified number of clusters based on statistical similarities in the features. In clustering, the predicted outcome is the cluster number to which the input features belong. While they can sometimes be used directly to generate useful predictions, such as using clustering to identify groups of users in a database of customers, unsupervised learning approaches are more often used to identify which data is most useful to provide to a supervised learning algorithm in training a model.

Relevant Azure services:

- [Azure Machine Learning](#)
- [Machine Learning Server (R Server) on HDInsight](#)

Deep learning

Machine learning models based on mathematical techniques like linear or logistic regression have been available for some time. More recently, the use of *deep learning* techniques based on neural networks has increased. This is driven partly by the availability of highly scalable processing systems that reduce how long it takes to train complex models. Also, the increased prevalence of big data makes it easier to train deep learning models in a variety of domains.

When designing a cloud architecture for advanced analytics, you should consider the need for large-scale processing of deep learning models. These can be provided through distributed processing platforms like Apache Spark and the latest generation of virtual machines that include access to GPU hardware.

Relevant Azure services:

- [Deep Learning Virtual Machine](#)
- [Apache Spark on HDInsight](#)

Artificial intelligence

Artificial intelligence (AI) refers to scenarios where a machine mimics the cognitive functions associated with human minds, such as learning and problem solving. Because AI leverages machine learning algorithms, it is viewed as an umbrella term. Most AI solutions rely on a combination of predictive services, often implemented as web services, and natural language interfaces, such as chatbots that interact via text or speech, that are presented by AI apps running on mobile devices or other clients. In some cases, the machine learning model is embedded with the AI app.

Model deployment

The predictive services that support AI applications may leverage custom machine learning models, or off-the-shelf cognitive services that provide access to pretrained models. The process of deploying custom models into production is known as operationalization, where the same AI models that are trained and tested within the processing environment are serialized and made available to external applications and services for batch or self-service predictions. To use the predictive capability of the model, it is deserialized and loaded using the same machine learning library that contains the algorithm that was used to train the model in the first place. This library provides predictive functions (often called score or predict) that take the model and features as input and return the prediction. This logic is then wrapped in a function that an application can call directly or can be exposed as a web service.

Relevant Azure services:

- [Azure Machine Learning](#)

- [Machine Learning Server (R Server) on HDInsight](#)

See also

- [Choosing a cognitive services technology](#)

- [Choosing a machine learning technology](#)