IMPLEMENT AI AND ANALYTICS INTO YOUR MANUFACTURING PROCESS

INTRODUCTION

Low-cost sensors and new wireless connectivity tools enable manufacturers to employ digital analytics more effectively than ever before. With the right tools, they can gather, cleanse, process, and visualize massive amounts of data from disparate sources that cover all phases of the product lifecycle. Manufacturers are incorporating advanced machine learning and AI tools into their technology stacks to predict and optimize outcomes in real time—across quality, maintenance, and throughput, in addition to monitoring processes and results. Knowledge graphs running on top of the data infrastructure can map complex relationships across systems, products, and processes; then turn raw data into a connected AI fabric. These technologies empower manufacturers to move from reactive to prescriptive decision-making comprising their complete operational ecosystems, including these major processes:

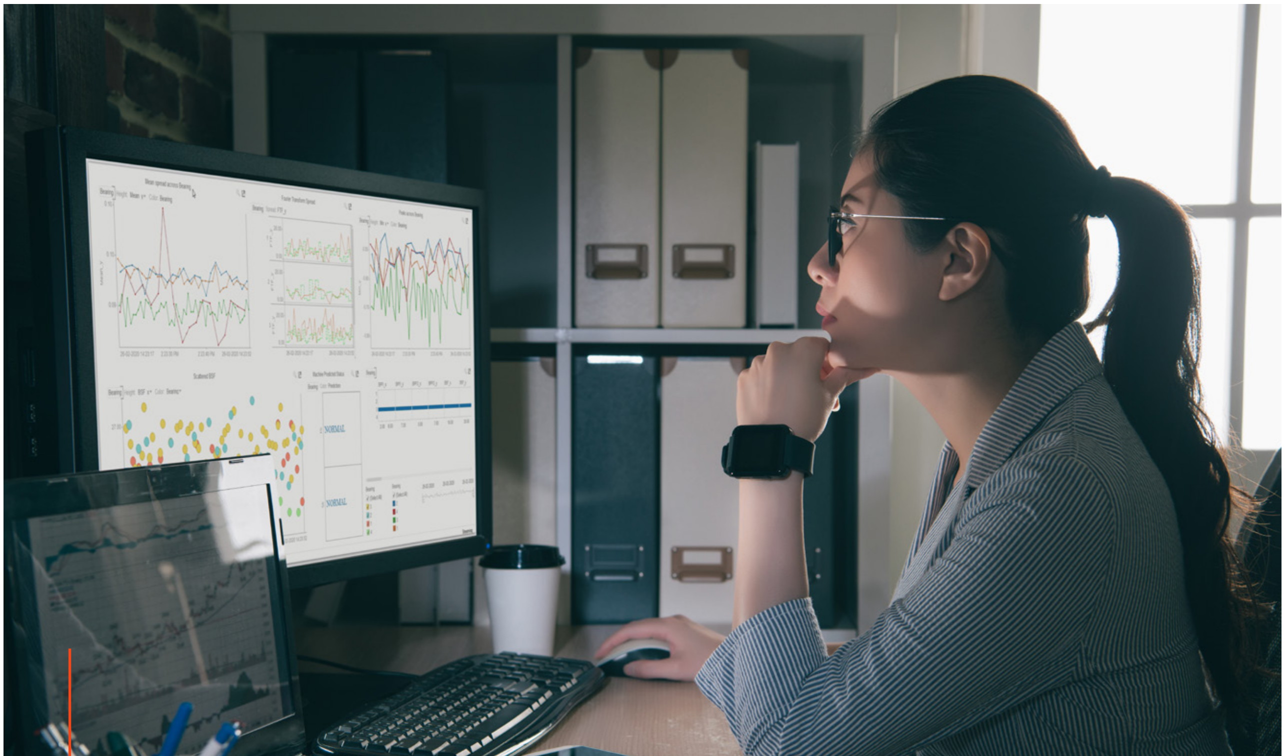
- Product and process design
- Assembly
- Material planning
- Quality control
- Scheduling
- Maintenance
- Overall equipment effectiveness (OEE)
- Fault detection
- Post-sale warranty claims

By extracting real value from their data, manufacturers can make accurate predictions about component life, replacement requirements, energy efficiency, utilization, and other factors that have direct impacts on production capacity, throughput, quality, sales, customer acceptance, and overall efficiency.

The manufacturing sector faces three key analytics challenges:

- **Variety:** Data is often trapped in organizational silos that make it hard to share between departments. It may be stored in incompatible formats and in a variety of databases, email systems, warranty claim systems, or even in PDFs, plain text, or images. Without the right tools, sharing this data in useful ways is difficult and expensive, or even impossible in practical terms.
- **Volume:** The amount of data coming in from suppliers, distributors, customers, and other third parties as well as sensor networks (inside and outside the factory) is increasing constantly and can easily become overwhelming. The ability to handle all this data properly enables engineers and managers to extract useful, actionable information from it by focusing on outliers, spotting trends, and clusters.
- **Velocity:** Manufacturing supply chains change rapidly in structure and flow, and critical data often streams in on real-time message queues. A software infrastructure built for real-time operational environments is an absolute requirement in the current business environment.

Addressing these challenges successfully offers big payoffs in every dimension of the manufacturing business.

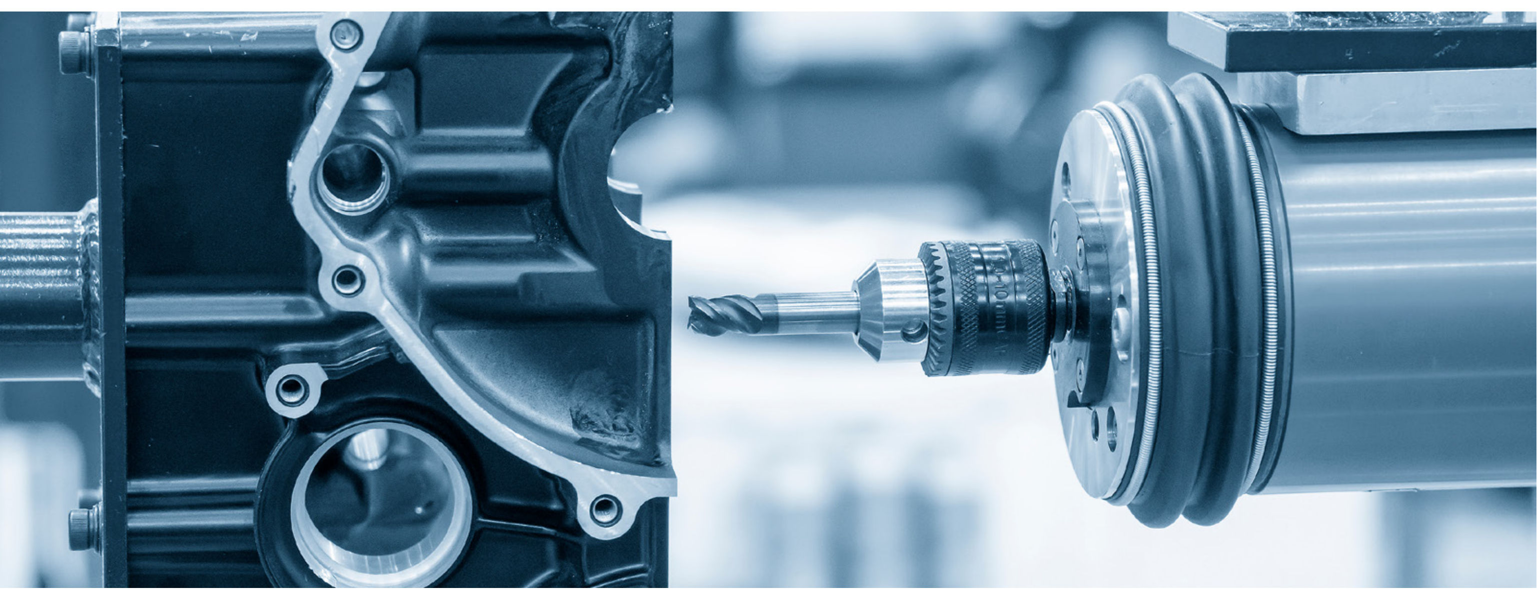


These four key elements must be in place for manufacturers to achieve the greatest possible gains from investments in analytics technology:

- Engineering personnel interested in **optimizing operational performance**
- **Access to data** stored in PLM, ERP, CRM, databases, document management, email, and other systems combined with the ability to cleanse and use that data effectively
- Flexible and comprehensive **AI fabric**
- Business processes and incentives that support **data-driven decision making**

Improve Manufacturing Productivity with AI and Data Analytics

This guide explains some of the major challenges involved in applying AI and data analytics to manufacturing processes and the benefits of developing optimized approaches to addressing those challenges.



DEVELOP AN ANALYTICS STRATEGY

Manufacturers can increase productivity, reduce costs, improve quality, and operate more competitively by employing the right analytics tools and implementing them in smart ways.

Developing a successful strategy requires a thorough understanding of the capabilities and limitations of the chosen technologies, as well as the capabilities of the people within the organization. To accomplish this:

- Begin with a clear-eyed assessment of the unique challenges confronting your organization along with the challenges common to all manufacturing operations.
- Next, investigate the data resources already available that will help address those challenges.
- Stakeholders must then come to agreement on the major outcomes the team should work towards, including defined milestones and deliverables.
- Finally, determine how to deliver the relevant AI and analytics capabilities to decision-makers and how the team will measure the success of each initiative within the strategy.

The implementation process itself involves five distinct phases:

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. User acceptance

This general approach is an adaptable, flexible way to think about a major analytics project within most manufacturing organizations.

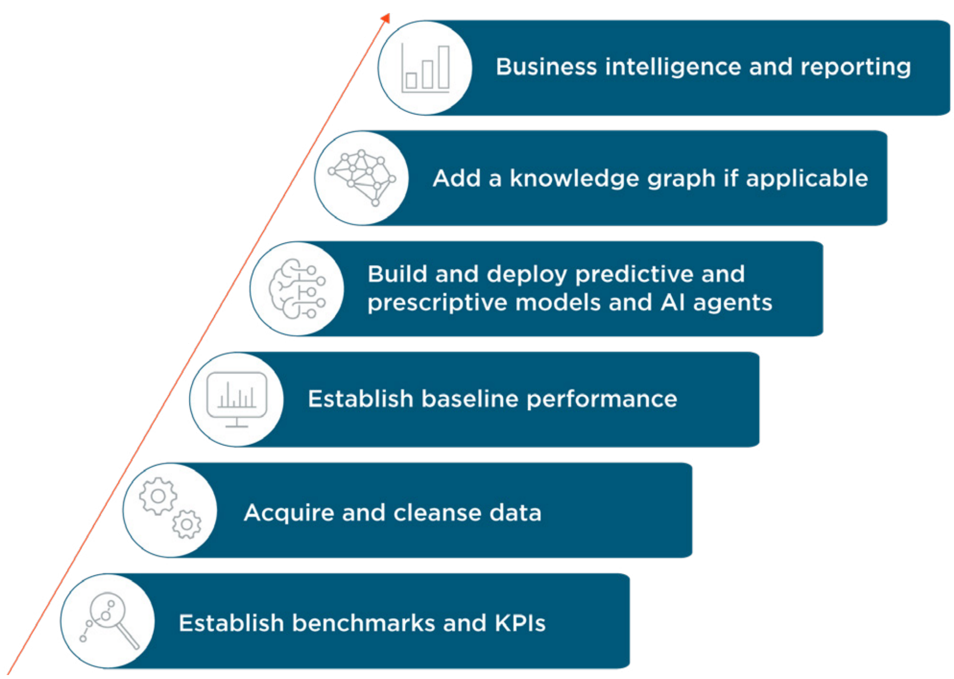
Insight doesn't come from what people can see on the surface of a data set. It comes from the hundreds or thousands of dimensions hidden in complex data.

The ability to analyze sensor data streaming in from production systems and correlate it with maintenance logs and other sources can support significant improvements in overall equipment effectiveness (OEE) and reduce maintenance costs.

THE DATA ANALYTICS WORKFLOW

Every manufacturing operation will have a different workflow. Following is an overview of the components that will make up a data analytics workflow for most manufacturing use cases. Every step generates value.

- **Identify value:** Identify key performance indicators (KPIs) and outcomes that your initiative should help improve, like yield, downtime, efficiency, and so on.
- **Connect, clean, and contextualize the data:** Often difficult since the useful data may reside in different data silos. The right data preparation and transformation software will turn the raw data into useable, clean, consistent data sets.
- **Establish baselines:** Process and visualize sensor and production data in real-time to look for anomalies, trends, outliers, relationships to understand your baseline performance and fine tune KPIs.
- **Develop and deploy predictive models:** Build and test AI models that allow manufacturing engineers to play "what if" with process and input changes.
- **Consider adding a knowledge graph layer to the data infrastructure:** While not every manufacturing facility needs a knowledge graph, those interested in enabling digital twins, natural language queries of large, disparate data sets, or other "Industry 5.0" capabilities will benefit from using a knowledge graph.
- **Monitor:** Build and deploy executive dashboards and other business intelligence and data visualization tools that enable all stakeholders, including engineers on the floor, finance, purchasing, and logistics personnel, and executive decision-makers, to spot trends, anomalies, clusters, and patterns in manufacturing quality, output, and other performance measures.



ABOUT DATA CLEANSING AND PREPARATION

As noted above, data cleansing and preparation are often the most difficult processes to implement successfully. It is critical to ensure that data is clean, consistent, and accurate before it is used in any AI application or to make important decisions that affect production, quality, or profitability.

Sources for manufacturing data include:

- Enterprise resource planning (ERP)
- Manufacturing execution system (MES)
- Supervisory control and data acquisition (SCADA)
- Programmable logic controllers (PLCs) incorporated into production equipment
- Computerized maintenance management systems (CMMS)
- Quality management systems (QMS)
- Camera-based inspection systems
- Sensors

Note: Sensor data typically requires special handling since it is usually real-time in nature, highly granular, and transported over message buses and stored in specialized time series databases.

Building a clean set of validated training data is critical to achieving any degree of accuracy in predictive AI models. Extracting data from all required systems, removing duplicates and errors, normalizing data when needed, and transforming it into a consistent format is usually impossible to accomplish without specialized tools. Even then, a solid data preparation process can often consume a larger time investment than building, testing, and running machine learning models. Do not underestimate the effort required or the value of performing the preparation process well in terms of the overall project outcome.

Selecting the appropriate data sources, figuring out how to best access the data, devising methods for combining them, and then ensuring that the output of the process is producing truly usable data requires insight into the business. Successful data preparation requires that domain experts — the people who understand the operation, the challenges being addressed, and the business context for the data — use tools that enable them to conduct this work on their own, without requiring an interactive (and time-consuming) process involving IT specialists or separate programming groups.

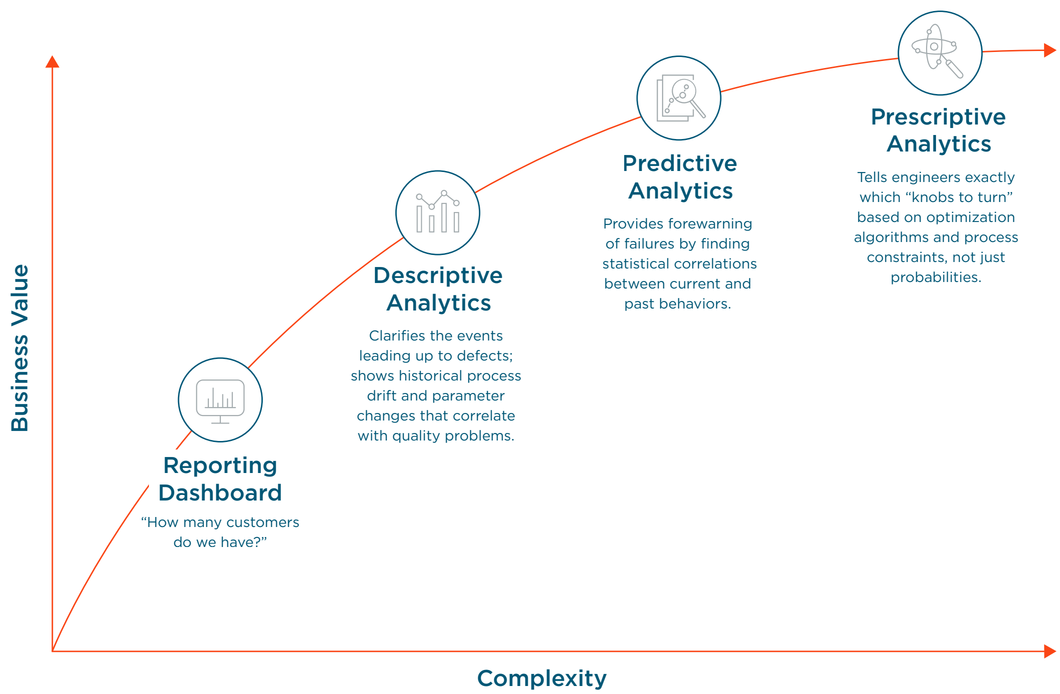
MONITOR AND ANALYZE REAL-TIME DATA STREAMS

In manufacturing applications, IoT sensors installed on processing equipment generate massive amounts of real-time data which may be streamed over OPC-UA / OPC-DA, MQTT, or similar message infrastructures. The right data visualization, business intelligence, and reporting software allows engineers and managers to analyze data the instant it becomes available. With the right visualizations, they can identify outliers, clusters, trends, and anomalies in seconds. They can roll back time to understand causal relationships that create quality and production issues, devise solutions, and even back test solutions using historical data before putting it to real-world use.

The analytics platform must be able to connect directly to message buses without requiring middleware and apply statistical functions and calculations on-the-fly to (potentially) millions of events every second. The platform must be able to combine historical data stored in time series databases or manufacturing process historians with real-time data streaming in from sensors on machinery.

PREDICTIVE AND PRESCRIPTIVE ANALYTICS AND AI AGENTS

Successful organizations overcome the complexities of data analytics to conclude when or how something may happen. They can also understand what will happen in the future based on changes in how data is measured. Put another way, predictive analytics is a defense against disruptive forces. To prevent, you must first predict; making accurate predictions enables engineers to address business and technical problems early—before they become major issues. Predictive analytics can also identify useful answers to questions that have not yet been asked.



Implementation and proper use of AI-driven analytics are critical elements in helping manufacturers respond effectively and quickly to a wide range of changing external factors.

Predictive analytics help organizations develop insights from thousands of dimensions in data that people cannot see without computational assistance. Predictive analytics relies on a person's ability to understand and test the relationships between cause and effect; this is accomplished by refreshing the data inputs or updating the model design itself. AI and machine learning algorithms remove people from this step by recalibrating models autonomously after they are put into use.

Prescriptive analytics systems typically utilize results from multiple AI and machine learning algorithms to inform future decisions. With prescriptive analytics, the object is to optimize and, to some extent, automate the decision-making process. A prescriptive analytics workflow will often use multiple algorithms to prescribe actions based on the characteristics of new data.

Combining predictive and prescriptive analytics with AI agents is an extremely powerful technique that can streamline operations even further. For example, a manufacturing facility can deploy a predictive model to identify optimal maintenance cycles for production equipment, recommend which parts should be replaced, and then automatically order the necessary components and schedule their installation.

Example use cases include:

- **Forecasting demand:** Based on demand predictions, recommend optimal allocations of labor, machinery, and materials to ensure efficient production and timely deliveries and automatically update schedules.
- **Optimizing inventory:** Recommend inventory levels to minimize stockouts and overstocks and automatically place orders.
- **Managing suppliers:** Identify the best suppliers based on factors including cost, quality, delivery performance, and even customer reviews.
- **Scheduling maintenance:** Recommend maintenance schedules to prevent unexpected downtime and extend machinery lifespans and automatically order parts and schedule downtown.
- **Managing energy:** Recommend measures to reduce utility costs and improve sustainability.

USE KNOWLEDGE GRAPHS TO ENABLE DIGITIAL TWINS AND MORE

In nearly all manufacturing operations, a small change like a product design tweak, a regulatory update, a supplier delay, or changes in the market can cause a chain reaction across departments, systems, processes, supply chains. Handling changes like this incorrectly can lead to big fines, expensive field recalls, reduced product quality, lower customer satisfaction, excessive inventory, and more. Traditional IT architectures and analytics tools are simply not built to manage the myriad possibilities for everything that can cause problems in large scale manufacturing environments. Huge and costly data warehouses and data lakes are designed for stability, not adaptability, and are too rigid, siloed, and reactive to address challenges like this.

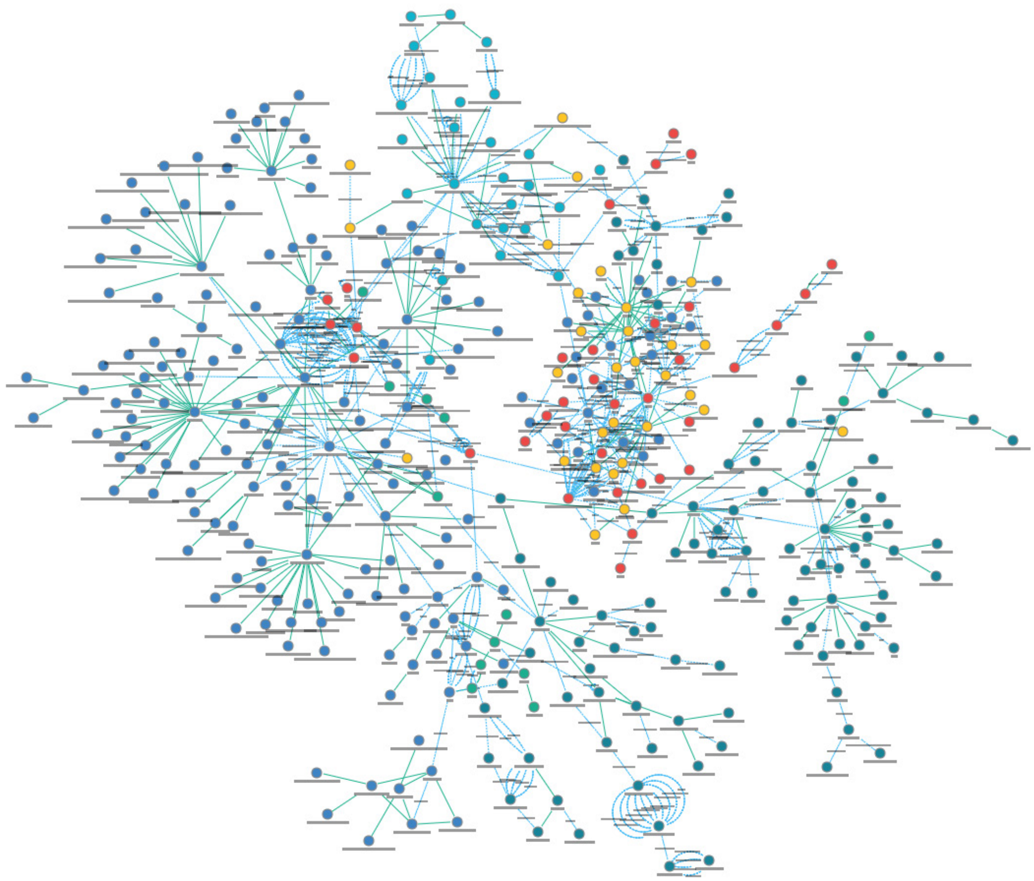
Knowledge graphs have recently emerged as the preferred tool for integrating, contextualizing, and extracting value from the vast data generated in modern manufacturing environments. They are built for scale and enable real-time analytics across all available complex, interconnected data. By mapping relationships between different data points, knowledge graphs can help users uncover patterns and correlations that might not be apparent through traditional data analysis methods. They also form the foundation for sustainable, robust AI applications. Knowledge graphs can accurately model and represent data, down to individual values, and can incorporate trillions of nodes and relationships. By creating semantic networks of interconnected data points, knowledge graphs allow manufacturers to break down data silos, enhance decision-making, and drive innovation.

Many manufacturers have invested in cloud data platforms to bring their data together, but these do not necessarily provide valuable connections across data domains or original source systems. Certain types of enterprise data may not be represented at all in a cloud data platform. Knowledge graphs can connect and contextualize structured and unstructured data from across the enterprise in ways that are beyond the practical capabilities of cloud data platforms on their own.

Knowledge graphs create comprehensive virtual representations of physical assets and processes and form the backbone for advanced digital twin implementations. Digital twins can blur the line between the physical and digital worlds. This can help manufacturers improve the quality of their products after they are in the field, gain insight into design changes that will make products easier to use and more reliable, determine the life of critical components, reduce cost of goods, and improve operational performance. Digital twin can model real-time wear and tear, reduce the time required to find and fix serious issues, and provide accurate remaining useful life (RUL) estimates.

Learn more about digital twins:
altair.com/digital-twin

Manufacturers also use knowledge graphs to build truly useful conversational AI applications like chatbots. They work with large language models (LLMs) to enable chatbots to accept natural language queries and generate accurate responses without hallucinations. Knowledge graph-enabled chatbots can compute and deliver results based on aggregations and analytics not found in the source data, adapt in real-time to user input, apply context-specific filtering through multi-hop queries across connected nodes, and increase the completeness and accuracy of responses. For example, a supply chain specialist can ask a question like, "Which suppliers have given us the lowest prices on our most critical components with the lowest defect rates over the past six months?" and receive an accurate, complete, and actionable response.



Manufacturing environments generate vast amounts of data from disparate sources, including sensors, machines, and enterprise systems. Knowledge graphs provide a unified view of all available data that supports insightful, fully informed decision-making to reduce costs, increase efficiency, and improve quality.

EMPOWER ENGINEERS WITH SELF-SERVICE DATA ANALYTICS

Specialized data teams are vital to implementing and utilizing data analytics within a manufacturing organization. However, businesses quickly discover that relying on a separate group is slow, expensive, and can add confusion to the decision-making process. Building, managing, and interpreting all required data flows along with creating machine learning algorithms, IoT data processing applications, and dashboards is a lengthy process. To add substantially to the overall productivity of the firm, data scientists must be able to work directly with engineers and managers to fully understand the questions, problems, and decisions that must be addressed.

The right data analytics tools do not require a degree in data science to be used effectively. In the most successful firms, engineers and managers who know the business can access data sources, build and test predictive and prescriptive AI models, run in-depth queries against any number of internal and external data sources, and visualize the results on their own. This self-service approach fosters utilization of the tools, reduces the time-to-market for analytics-based business processes, and helps ensure analytics projects are adding real business value. The data science team can focus its attention on developing and promulgating best practices, supporting efficient use of the tools, and digging in on the most difficult analytics challenges.

OPTIMIZE OVERALL EQUIPMENT EFFECTIVENESS (OEE)

High OEE levels enable firms to maximize the return on investment (ROI) for manufacturing equipment, improve output quality, increase competitiveness, reduce downtime and maintenance costs, and maximize productivity. The right data analytics tools support real-time monitoring of OEE. They also allow engineers to dig deep into historical production, maintenance, and quality data to understand which approaches have had the biggest positive and negative impacts on past performance as well as the methods that have the greatest potential for increasing OEE.

Integrating data from multiple silos, including data streams from sensors and historical manufacturing data which may be stored in a variety of formats and systems, allows manufacturing engineers and managers to push improvements in performance, quality, and availability, the three primary key performance indicators (KPIs) for OEE.

Data visualization software optimized for real-time data is especially useful in OEE applications. It enables managers to drill down to understand how well individual machines are operating currently as well as back-trace potential causes of faults that have created OEE shortfalls in the past for any specific machine or group of machines. Combining real-time data visualization with AI and machine learning workflows is even more powerful since data streaming in from industrial internet of things (IIoT) and other sources on the plant floor can be compared on-the-fly with historical data using many statistical techniques that can reveal problems that would otherwise be hidden.

Machine learning tools can efficiently identify issues and weak spots in production machines and proactively alert operators to potential causes of downtime while there is time to plan for necessary maintenance or component replacement. The ability to analyze sensor data streaming in from production systems and correlate it with maintenance logs and other sources can support significant improvements in OEE and reduce maintenance costs.

Tool Condition Monitoring (TCM) and Remaining Useful Life (RUL) Analysis

Tool wear in metal cutting operations has a direct impact on the quality and accuracy of the finished surfaces. Various types of sensors can monitor how well a tool is performing, measure generated heat, speed, pressure, and other factors that, alone or in various combinations, signal that a tool is approaching the end of its life. Replacing a tool at the optimum time — while it is still performing up to spec but just before its degradation begins to cause damage, reduce output quality, or increase scrap rates — is highly desirable.

AI and machine learning are well suited for TCM and RUL analysis. The large amounts of data produced by sensors combined with human inspections of finished pieces can be used to train machine learning algorithms to identify the “sweet spot” and proactively alert operators when a tool is approaching time for replacement. Stream processing algorithms can also process all the sensor data being generated by any number of production machines, make on-the-fly comparisons with historical data, and amplify the accuracy of the machine learning algorithms.

Anomaly Detection in Production Systems

Identifying unusual behaviors or patterns in machine components using sensor data can prevent small glitches from creating major operational problems. In cases where large numbers of sensor feeds are involved, challenges emerge due to the sheer volume and velocity of data streaming off the equipment. In addition, meaningful analysis from the data is a nontrivial task, since slowing or shutting down production to examine a machine carefully should only be done when truly necessary. For these reasons, simple threshold-based alerting is normally unsuitable; it will generate too many false positives. More advanced methods can, however, flag potentially serious issues proactively and increase OEE.

Quality Improvement

Great product quality wins new customers and keeps existing customers coming back. Consumers and business buyers have come to expect high build quality, long mean times before repair, and long product life. Poor data in the supply chain is often the root cause of a product quality issue. Pricing, quoting, configuration, and delivery instruction errors can slow down production, result in cost-related margin reductions, and artificially create the need for component substitutions.

The right tools enable engineers, analysts, and manufacturing supervisors to aggregate and cleanse data from any source, ensure proper data governance throughout the analytics lifecycle, and develop consistent, accurate data that can feed machine learning, predictive analytics, and visualization systems which then produce reliable outputs. Real-time data comes into play as well, since solving complex production problems often requires visibility into massive amounts of sensor data and historical information. The ability to access and analyze real-time data effectively enables faster, better informed decision-making to reduce production costs and quality issues driven by rework, returns, unscheduled machine downtime, and incorrect orders.

Machine Failure Prediction

Machine learning technology leverages historical and real-time data from sensors mounted to production equipment as well as PLCs, SCADA, and other sources and can accurately flag potential failures of whole machines and/or critical components before they can cause downtime. Failures may be binary in nature; that is, a failure occurred or not. Failures may also be multi-class and fall into several different categories, including reduced speed, throughput, or quality. Obviously, the more complex the machine (or system), the more AI and machine learning models can help prevent failures that can impact productivity.

Root Cause Analysis

Root cause analysis (RCA) is critically important to the ongoing success of any manufactured product. Detecting design defects, raw material problems, build issues, and quality control shortfalls as early as possible fosters continuous product improvement, increased reliability and performance, and allows the company to maintain reputations for strong product brands.

RCA enables R&D personnel, purchasing agents, quality control inspectors, and the warranty team to find the fundamental cause of problems that may only become evident once products are in customers' hands. In addition to the benefits noted above, managing the RCA process properly will reduce the number and cost of warranty claims, improve profitability for the entire firm, and increase customer satisfaction.

OTHER APPLICATIONS IN MANUFACTURING

The following are examples of common applications for data analytics within the manufacturing environment.

Service Warranty Optimization

Price optimization in any realm helps companies improve competitiveness, but when applied to service offers like extended warranties, service contracts, and out-of-warranty repairs, it can also condition customer responses to additional offers, mitigate production quality issues, and modulate demand to consider the company's ability to deliver services. Optimization also helps convert what may be an unprofitable requirement for manufacturers into a significant income generator.

The right service offers have major impacts on customer loyalty and brand reputation as well. Every manufactured product line will exhibit quality control issues, field failures, and premature end-of-life issues from time to time. Dealing with problems like this at prices that customers see as fair is critical, and likewise, offering warranty extensions and service contracts at attractive prices helps reduce customer worries about product longevity, particularly with new product introductions.

Manufacturers can also use data analytics to optimize the length and terms for original warranty offers by taking competition, product price points, customer preference survey results, and other factors into account.

Warranty Risk Profile Analysis

Most manufacturers must handle large numbers of warranty claims related to a variety of products and components. The volume of claims can easily run to millions per year for consumer goods manufacturers. It is critical to prioritize and understand which issues deserve high priority responses and detect patterns within the claims that indicate emerging quality or design problems that require immediate attention.

Warranty risk profile analysis, sometimes referred to as quality issue prioritization, is a vital part of any ongoing quality improvement process. The data from warranty claims, once cleansed and sorted, is one of the most valuable parts of the feedback loop that enables companies to improve reliability and customer satisfaction.

Expert Systems

Expert systems are computer programs that attempt to emulate a human expert, usually in a narrow field of expertise. Such advanced data analytics tools can augment or even replace the work of skilled engineers in the manufacturing environment, but with some important limitations.

The best cases for applying expert emulation are those where the "rules of the game" are relatively easy for people to understand but difficult to write down or define in simple sets of rules. For example, the process of figuring out the optimal processes and order of work for sheet metal stamping is complex and typically requires the attention of manufacturing engineers with many years of experience.

However, by applying machine learning techniques, an expert system can determine with a high degree of accuracy the best configuration of processes for a given sheet metal component to reduce waste, increase quality, and improve throughput.

Reduce Accounts Receivable Days Outstanding

Days sales outstanding (DSO) is a critical performance measure for many manufacturing operations, and anything that can reduce the firm's DSO will improve the bottom line.

A common challenge involves the many different systems that a firm may use to manage its inventory, production processes, shipments, sales, and accounting. Reconciling data from such disparate systems, which were often implemented by different teams and different times with different objectives, is a nontrivial task. Analytics tools that can directly access the output and internal data resources of all relevant systems, cleanse that data, and transform it into governed, accurate, and useful information are therefore crucial. Machine learning and other data science algorithms can then be applied to anticipate potential slow payers and increasing DSO numbers before they can affect the business.

Price Optimization

Price has an immediate impact on demand, and finding the right price ensures that factory production is closely matched with demand. With the right statistical models and tools, manufacturers can determine how customers will respond to different prices through different channels and figure out the pricing models and price levels that will maximize operating profit, maintain market share, and fend off competition. Most markets are highly competitive with constantly evolving customer requirements and interests; therefore, price optimization must be a continuous process.

Supply Chain Risk Management

Supply chains for most products today are long, unpredictable, and complicated, and involve hundreds or thousands of third parties. It presents one of the main areas of risk for high volume manufacturers. Firms can utilize a wide variety of data from government, shipping firms, suppliers, and companies with complementary offerings to prepare for supply chain delays and develop contingency plans. They must be able to clean and normalize all useful data and develop predictive models they can tweak on an ongoing basis as updated information becomes available.

Data analytics exposes the cost and quality of every outsourced component in your production life cycle. For example, managers can determine whether certain components are failing more frequently than expected or are not performing up to spec and then use machine learning to help determine the best remedy. Keeping pace with the continuously changing nature of the supply chain makes it essential to have a real-time view of order, delivery, and usage flows.

ALTAIR ADDRESSES COMPREHENSIVE DATA ANALYTICS CHALLENGES

Insight doesn't come from what people can see on the surface of a data set. It comes from the hundreds or thousands of dimensions hidden in complex data. People need the right tools to easily access these hidden dimensions.

The Altair® RapidMiner® platform empowers business users to collaborate efficiently to access meaningful data, generate insight from this data, and share findings throughout the enterprise. Its comprehensive range of AI, machine learning, data preparation, knowledge graph, and business intelligence tools enable fast, insightful, efficient decision making for every phase of manufacturing operations.

To make the firm more productive, data science teams must be able to work directly with manufacturing engineers and managers to fully understand the questions, problems, and decisions they address every day.

APPENDIX: DEFINING TERMS

A comprehensive glossary of data analytics terms is beyond the scope of this guide, but here are some key concepts that will be helpful to anyone looking into utilizing analytics to improve manufacturing operations.

AI Fabric

An AI fabric seamlessly unites data, AI, and automation to enable businesses to transform their operations with real-time insights, enhanced decision-making, and scalable automation. It unifies all data sources in use within an enterprise to eliminate data fragmentation and ensure consistency and accessibility across all systems. It uses that unified data to fuel AI models that deliver accurate, context-driven insights for better decision-making and drive automated, data-driven decisions in real time.

Algorithm

In the context of data science, an algorithm is a series of repeatable steps that is often expressed in the form of Boolean logic. Data scientists can develop and implement algorithms using many different tools and methods, including coding in languages like Python or R, or within the framework of a machine leading software system. Some commonly used algorithms in data science include linear and logistic regression, decision tree, random forest, Naive Bayes, and KNN (K-Nearest Neighbors).

Artificial Intelligence (AI)

AI is something of a catch-all term, but at its core it is technology that can extract useful insights and identify patterns in large data sets, and often produce predictions based on that data. The concept — and perhaps the holy grail of data science — is to develop systems that can emulate human thought processes. In practice, however, some of the best uses of AI are intelligent assistants that can, for example, troll through millions of purchase orders and bills of lading looking for examples of overbilling or delivery failures.

Business Intelligence (BI)

BI tools handle large amounts of structured and sometimes unstructured data to help people identify outliers, anomalies, trends, patterns, and clusters. They typically offer a range of easy-to-interpret dashboards and other reporting tools and allow non-technical users to make informed decisions based on insights derived from internal and external data sources.

Code-free

The ability to develop sophisticated software applications using a visual, object-oriented interface without writing and debugging programming code. Code-free tools are typically designed for use by businesspeople and engineers who wish to build systems to procure, analyze, and make predictions without relying on IT and/or programming personnel.

Data Cleansing

The process of removing or modifying incorrect, incomplete, irrelevant, duplicated, or improperly formatted data. Cleansing is a critical step in the data preparation process, since "dirty" data will contaminate the outputs of machine learning or other analytics applications and can lead to incorrect conclusions and predictions.

Data Preparation

The process of gathering, combining, structuring, and organizing data to be used in AI, machine learning, predictive analytics, visual analytics, and/or BI applications. Data preparation systems must be able to connect to and collect data from a wide variety of disparate sources in a range of formats and data types, cleanse the data, normalize it if necessary, ensure that the process does not introduce errors of its own, and output the data in a consistent format usually consisting of rows and columns in a database.

Data Governance

A collection of practices and processes for data management to ensure that the ancestry of the data is known and understood, and that only qualified people can alter or augment the data. The Data Governance Institute defines it as: "A system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods."

Proper governance reduces the costs associated with managing large amounts of data, supports reproducible and accurate procedures around regulation and compliance actions, and increases the value and utility of the data.

Data Visualization

Numeric data displayed in graphic form, including everything from simple line graphs and pie charts produced in a spreadsheet program to highly sophisticated interactive dashboards.

Data Modeling

In the context of data analytics, modeling involves building sets of algorithms that can make accurate predictions about future events based on historical data. Ideally, a model's predictions produce forecasts that businesspeople can use to make informed decisions.

Decision Tree

A decision tree is a graphical depiction of decisions, or nodes, in which every potential outcome is directed to a new branch of the tree. Decision tree-based algorithms are one of most used methods employed in machine learning systems; they offer high degrees of accuracy and stability and are easy for users and consumers of their output to understand.

Explainable AI

Simply put, this refers to systems that allow ordinary humans to understand how a set of algorithms in an AI system produced its outputs. To put it another way, explainable AI systems are the opposite of co-called "black box" systems that produce results that cannot be interpreted visually.

Explicit Programming

This is the type of computer programming most people are familiar with. Every instruction to the computer must be written out in a suitable computer language and altered manually when new parameters need to be added or altered. Enabling predictive models to make accurate predictions requires thousands of parameters, and therefore an explicit programming approach is generally not scalable for most practical applications in data science. Instead, most truly useful and effective predictive models make use of a combination of explicit programming and implicit programming (see machine learning below), which relies on the system itself to produce outputs that are not programmed step-by-step by a human.

Knowledge Graph

Knowledge graphs are interconnected networks of data points that accurately model and represent data, down to individual values, and can incorporate trillions of nodes and relationships. They use a semantic layer approach to connect disparate sources of structured, semi-structured, and unstructured data. Knowledge graphs improve the accuracy and utility of large language models (LLMs) for generative AI (GenAI) applications, including chatbots.

Machine Learning (ML)

Arthur Samuel coined the term "machine learning" and defined it as follows: Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed. In other words, the algorithms in the system learn by themselves. ML systems are highly scalable and can adapt quickly as new variables come into play. They can predict outcomes without explicit human input.

A key concept of ML is the "training set." ML systems must be fed a historical data set for which the insights and patterns to be found are already known and confirmed using an existing technique. After "training" the ML model, the system can predict new results based on new data inputs.

Python

An open-source object-oriented programming language. It is popular in the data science community because its user community has developed and published an extensive library of useful programming objects. Compared to many other languages, Python is also easy to learn and use.

R

An open-source language and environment optimized for statistical computing and analysis. Like Python, R is often used in data science applications. It is generally considered to be more difficult to learn than Python, but offers an excellent array of graphical and plotting capabilities.

Real-time Data

Real-time data (also referred to as "streaming data") is continuous with no beginning or end, and the individual data points in the stream of data may appear at regular or irregular intervals that are typically in the sub-second range. Nearly all operations produce some amount of real-time data from a variety of sources like sensors and transaction processing systems. Real-time data streams are usually carried on a message bus (also called a message queue) capable of moving large amounts of data between distant points with extremely low latency. Examples of popular message buses include MQTT, Solace, and ActiveMQ.

Streaming Data

See "real-time data" above. The terms "real-time data" and "streaming data" are synonymous.

Time Series Data

This is essentially a recording of a real-time data stream and is usually stored in a specialized time series database. Popular time series databases include InfluxDB, Prometheus, TimescaleDB, kdb+, Azure Time Series Insights, Amazon Timestream, and many others.

In the most successful firms, engineers and managers who understand the business and its processes can access data sources, build and test predictive models, query data, generate reports, and visualize the results without help from IT personnel.

Altair is a global leader in computational intelligence that provides software and cloud solutions in simulation, high-performance computing (HPC), data analytics, and AI. Altair is part of Siemens Digital Industries Software.

To learn more, please visit altair.com or sw.siemens.com