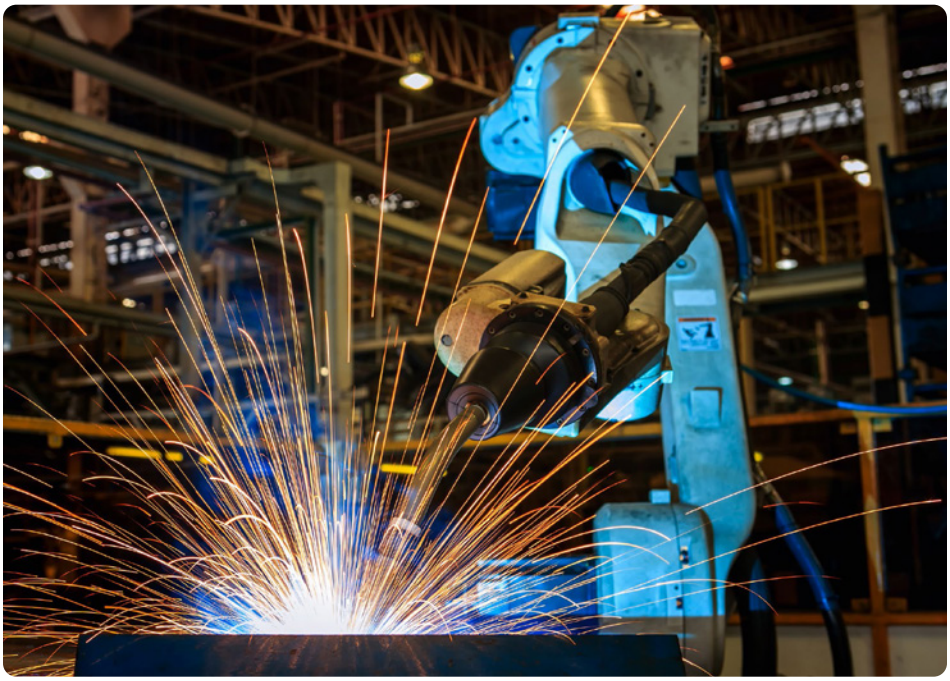
INDUSTRY FOCUS | PREDICTIVE MAINTENANCE

Using Data and Analytics to Drive Predictive Maintenance

The data lifecycle reshapes the predictive maintenance landscape



The inability to forecast manufacturing process or equipment failures takes a high toll: Unplanned downtime costs the industry an estimated $50 billion a year[1]. Much has been written about forward-looking manufacturers tackling this cost head-on by leveraging industrial internet-of-things (IoT) sensors, 5G connectivity, and big data to shift from a costly "repair and replace" approach to a stance of "predict and prevent."

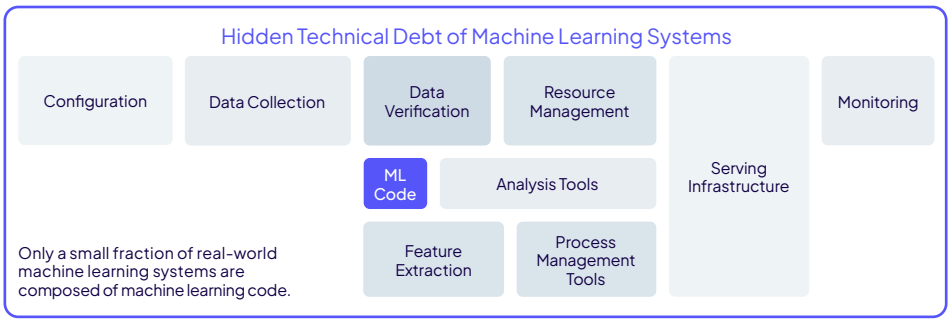
The most basic definition of predictive maintenance, in a nutshell, is about figuring out when an asset should be maintained and what specific maintenance activities need to be performed, based on an asset's actual condition or state, rather than conducting maintenance on a fixed schedule or, even worse, waiting for failure. It is all about predicting and preventing failures, and performing maintenance on your time, on your schedule, to avoid costly unplanned downtime. Using data analytics to predict and prevent breakdowns can reduce overall downtime by 50%[2] — that means big numbers dropping directly to the bottom line.

Developing an effective predictive maintenance platform requires machine learning models built on clean data, information collected from a multitude of sensors, and robust historical records that encompass a thorough characterization of equipment and the manufacturing process. Deployment success is based upon a platform that can manage real-time data flow and ingestion, leveraging that into robust predictive maintenance applications.

Predictive Maintenance Data Management Challenges

Machine learning driven predictive maintenance should be easy, but consider that only 20% of the machine learning models in the enterprise make it into production, and 88% of the machine learning projects never make it beyond the experimental stage.

Why is this? The underlying data infrastructure is often given little consideration. Machine learning models are important, but there is a larger data ecosystem and lifecycle to be considered. Data collection methods, data verification, analytics tools, etc. are all needed by Data Scientists to successfully mine data, clean and create readily deployable machine learning models.



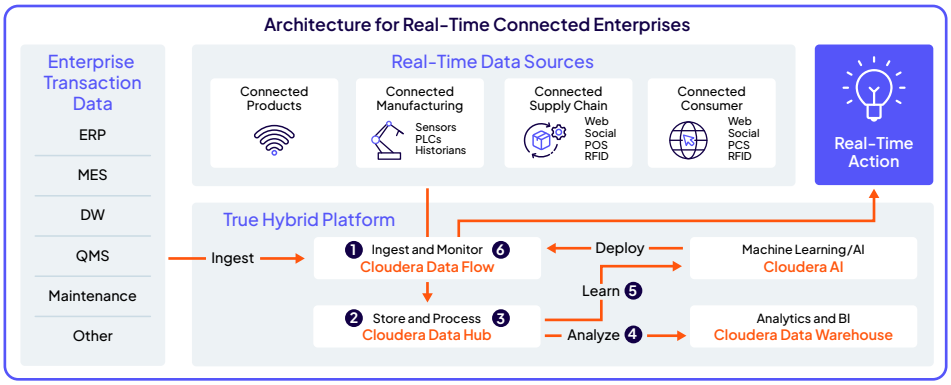
Other Considerations That Limit Successful Predictive Maintenance Solutions Are:

Consider that only a small fraction of real-world ML systems are composed of ML code. Other considerations that limit successful ML and predictive maintenance solutions are:

- **Managing the complexity of real-time data:** Driving predictive maintenance, data management platforms must enable real-time analytics on streaming data. A platform must effectively ingest, store, and process streaming data in real time or near-real time to instantly deliver insights and action.
- **The volume and variety of IoT data:** To enable predictive maintenance, information architects need a platform that can handle data structures and schemas. It must accommodate everything from intermittent readings of temperature, pressure, and vibrations per second to handling fully unstructured data (e.g., images, video, text, spectral data) or other input such as thermographic or acoustic signals. All this data is coming from the network edge through diverse supported drivers and protocols.
- **Freeing data from independent silos:** Specialized processes within the value chain — such as innovation platforms, quality management systems (QMS), manufacturing execution systems (MES) — reward disparate data sources and data management platforms that tailor to unique siloed solutions. These narrow-point solutions limit enterprise value because they consider only a fraction of the insight cross-enterprise data can offer.
- **The cost of data management:** Traditional data management tools tend to be notoriously expensive, difficult to scale, and unsuited to capturing and processing the petabytes of IoT data streaming from continuously monitored, connected equipment. Today, organizations need a more flexible and scalable data management and analytics platform that can easily ingest, store, manage, and process streaming data from IoT sources at a lower cost.
- **Predictive modeling capabilities:** Predictive modeling capabilities are key to delivering insights, and current platforms provide little to no modeling or machine learning capabilities to help predict and prevent disturbances before they impact operations. The ability to leverage disparate data sources and data types will promote a robust and well-rounded model resulting in stronger prediction capabilities.

The Data Lifecycle Architecture to Enable Predictive Maintenance Use Cases

- Predictive maintenance rests upon an architectural framework emphasizing big data ingestion and management, machine learning, and streaming analytics. The following illustration examines the IoT data lifecycle in greater detail.



As outlined in the illustration above, the IoT data lifecycle includes the following steps:

Data Ingestion (1): This phase begins with the connected devices themselves. Data from sensors, programmable logic controllers, supervisory control and data acquisition systems (SCADAs) and data historians is ingested using various protocols, transformed as necessary and routed for downstream processing. Challenges in this phase include security and reliable data flow control given often suboptimal network connections through the myriad of sensor protocols that are used on the edge. Cloudera Data Flow is a comprehensive solution that collects and ingests IoT data, log files, and data from various enterprise software systems, including both streaming data from IoT sensors, and business process data. Supported-device protocols, once a challenge, provide efficient and effective connectivity. Cloudera Data Flow provides the abilities to aggregate, compress and encrypt connected manufacturing data, prioritize transmission of data from edge to the cloud or data center, buffer data in the event of network interruptions, and track the provenance and lineage of streaming data, providing confidence in the origin and usage of data.

Data Storage (2): Once ingested, real-time data is then delivered to a manufacturing data lake where it can be stored alongside a wide range of other data sources (information from ERP, MES, QMS, or computerized maintenance management systems).

Data Processing (3): Various data processing workloads (such as combining and checking the quality of the data) can be performed directly within the data lake, preparing the data for downstream use cases. Cloudera Data Hub provides massively distributed storage and processing engines for large datasets, storing and processing of any kind of data including unstructured data, semi-structured data (i.e., sensor data), and structured data, such as transactional data from ERP, maintenance, supply chain management, and customer relationship management data. It gives organizations the ability to quickly and efficiently execute a wide range of data processing workloads.

Data Analysis (4): Once the data is stored and prepared within the data lake, it can be analyzed using data discovery or KPI-based business intelligence capabilities. Cloudera Data Warehouse is an enterprise-grade, hybrid cloud solution that provides self-service analytics, enabling organizations to share petabytes of data to drive analytics and BI with the security, governance, and availability that large enterprises demand.

$2M

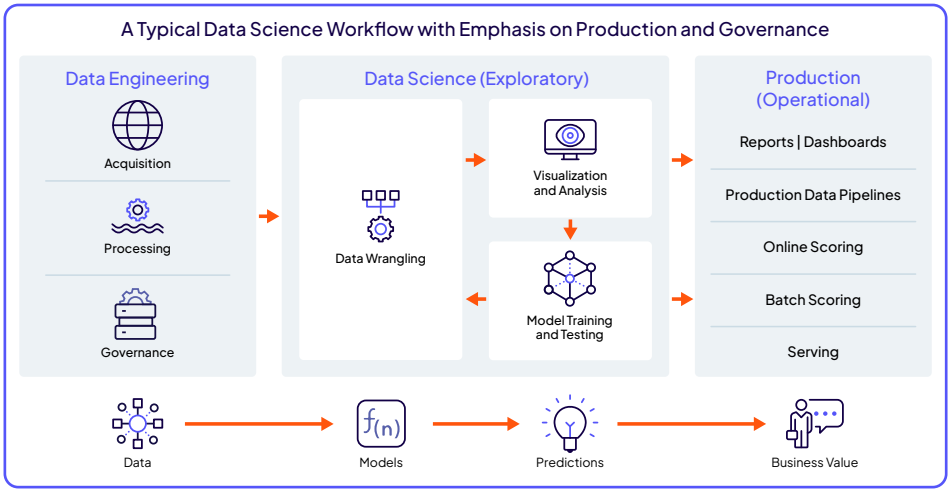
is approximately what a single unplanned downtime event can cost some of the leading automotive manufacturers, and as much as $15,000 – $20,000/minute.

Learning (5): Predictive models are trained to detect anomalies within large datasets and from many diverse streams. At this stage, the power of the machine learning algorithm is directly related to the depth and breadth of enterprise data and correlates to the ability to produce an enterprise-wide value chain solution. Efficient learning is based upon clean data, access to all data, no matter its form, and a clear understanding of the path from model production to deployment.

Data Monitoring (6): Predictive models are then deployed to the network edge where incoming IoT data is monitored in real-time (streaming analytics) to detect and identify conditions specified by the predictive models, such as specific sensor values that indicate impending failures or process anomalies. Cloudera AI can help operators accelerate data science at scale to build, test, iterate and deploy machine learning models by taking advantage of massively parallel computing, and expanded data streams. Using Python, R, and Scala directly in the web browser, Cloudera AI's powerful self-service experience enables data scientists to develop and prototype new machine learning projects and easily deploy them to production.

Data from a Data Scientist's Perspective

It is good to pause and understand the machine learning workflow from the perspective of a Data Science team whose job it is to leverage raw data input to derive insights and models, and help operationalize them into business processes. The data workflow consists of data engineering (steps 1, 2 and 3 previously described), data science (steps 4 and 5), and production (step 6).



Data Engineering

- **Data Processing:** Raw data from sensors, data historians, or business intelligence systems is typically not in a convenient format for a developer to run analysis, so it must be cleansed and prepared.
- **Data Governance:** As organizations increasingly depend on data and analytics to answer important questions, the need to govern those assets increases. Organization should be concerned about data quality in their source systems, but often these concerns are isolated and not visible across departments. Security, privacy, and regulatory compliance are important elements of governance.

Data Science

Building machine learning algorithms requires data wrangling, data visualization and data modeling:

- **Data Wrangling** is the process of transforming and mapping data from one "raw" format into another with the intent of making it more appropriate and valuable for a variety of downstream analytics.
- **Data Visualization** helps identify significant patterns and trends in the data so users can gain better insights from simple line charts or bar charts.
- **Machine Learning Model Training and Testing** is based upon a characterization of "typical" equipment or process behavior, building algorithms and then building algorithms that model and replicate trends and correlations data, and then signal when anomalies are detected. This is the most commonly known aspect of machine learning, though the others are required to support this step.

Production

- **Production Deployment** is the process of delivering the outcomes (in this case, proactively signaling anomalies in equipment or process performance) to stake-holders, such as manufacturing operations, maintenance or plant engineering. Successful production machine learning requires streamlined, frictionless and predictable deployment, serving, and ongoing governance of models at scale.

The data science workflow is an iterative process and machine learning models are not static. Machine learning models must be monitored, updated, and improved with new data and revised when there are new process conditions or permanent changes in equipment/process behavior. Successful production ML systems provide line of business owners confidence in results, knowing that deployed models are not "black boxes," while providing IT operations leaders visibility and control of the data lifecycle leading to meaningful business impact.

Heat exchangers, distillation columns and compressors are predominant applications of predictive maintenance.

Why Cloudera

Hybrid And Multi-Cloud

Run analytics on the cloud platforms. Easily and securely move data and metadata between on-premises file systems and cloud object stores.

Analytics From Edge To AI

Apply real-time stream processing data warehousing, data science and iterative machine learning across shared data, securely, at scale on data anywhere.

Security And Governance

Use a common security model, role and attribute-based access policies and sophisticated schema, lineage and provenance controls on any cloud.

100% Open

Open source, open compute, open storage, open architecture and open clouds. Open for developers, partners, and open for business. No lock-in. Ever.

Beyond Sensor Data

Leveraging data from data historians, ERP, MES and QMS sources spanning the value chain is the key to an enterprise solution. Knowing **how to respond** with optimized down-time scheduling, deployment of maintenance manpower and an alternative manufacturing process during the downtime is almost more important than knowing **when and where** equipment will fail. An enterprise value chain solution considers the cost of downtime, equipment reliability characteristics and the redundancy of assets by taking into account purchasing decisions, workforce deployment, supply chain optimization, production scheduling, quality and yield optimization and inventory management.

Consideration of all aspects of the value chain maximizes overall equipment effectiveness, ensuring that 100% of the parts are good (100% quality) and produced at the maximum speed (100% performance) and without interruption (100% availability). In order to do this, the manufacturing process demands real-time visibility of the entire value chain.

INDUSTRY	USE CASE	CHALLENGE	SOLUTION AND VALUE
Automotive Component Manufacturer	Predictive Equipment Maintenance	Relational databases couldn't provide the performance required to maximize production uptime and reduce manufacturing defects	The manufacturer maximized production uptime and improved product quality with an IoT-enabled platform from Cloudera. The enterprise data hub brings together and analyzes data from a variety of sources to help improve product quality and predictive maintenance
Automotive Robotics Manufacturers	Predictive Equipment Maintenance Predictive Asset Maintenance	Automotive assembly downtime caused by robotic assembly equipment was costly. The non-IoT-enabled assembly robots lacked real time insight into performance	A "Zero Down Time" robotics monitoring program was initiated to leverage an advanced analytics platform aimed at gathering, storing and analyzing IoT sensor data. The company **is reaching its goal of zero down time** through predictive maintenance
Heavy Equipment Manufacturer	Predictive Equipment Maintenance Connected Assets/ Vehicle Services	Unconnected equipment has higher maintenance costs and more unplanned downtime, giving customers a perception of poor equipment quality and performance	The heavy equipment manufacturer created an integrated IoT analytics system driven by Cloudera and Azure. **It doubled a large coal mining's daily utilization** of the longwall mining system
Hydroelectric Power Generation	Predictive Equipment Maintenance	Hydroelectric power stations cannot afford downtime and the loss of power to the customers due to unplanned maintenance	A predictive maintenance solution was implemented using acoustic sensors to pinpoint abnormal acoustic signatures in the turbines. Anomaly detection **reduced unplanned downtime and improved capacity**
Semiconductor Fab	Predictive Equipment Maintenance Quality and Yield Optimization	The manufacture of semiconductor chips is a highly complex process, and the misprocessed die yield was suboptimal, limiting output and profitability	An integrated IoT network and master data warehouse provides comprehensive quality and process insights. Yield was improved as misprocessed die identification was **lowered from seven days to less than one hour**
Oil and Gas Exploration	Predictive Equipment Maintenance	Unexpected machine breakdowns can lead to downtime that can cost tens of thousands of dollars a day in lost production.	IoT data streaming from more than 70 trillion sensor data points and robotic automation systems feeds analytics. Predictive maintenance **has enabled six years of maintained production and slashed capital costs 80%**
Truck, Bus and Equipment Manufacturer	Predictive Equipment Maintenance Connected Assets/ Vehicle Services	Unconnected vehicles triggered high fleet maintenance costs and limited new business opportunities	A new business model was developed incorporating connected assets, analytics and predictive maintenance technology. Lowered logistics and maintenance costs allowed the creation of a new business model: predictive maintenance and logistics as a service. **Maintenance costs dropped from $.12-.15/mile to $.03/mile**

Cloudera Provides Actionable Predictive Maintenance Benefits

Companies are reaping the benefits of predictive maintenance in their manufacturing operations and even extending it into their post-sales service organizations. Above is a summary of predictive maintenance use cases that highlight how some of our customers are using Cloudera to drive predictive maintenance.

Effective predictive maintenance is more than just IoT sensors and algorithms driving improved uptime. It is the foundational realization that enterprise data is at the heart of any predictive maintenance initiative. Knowing when equipment is going to fail, and effectively responding to and planning for its downtime drives operational efficiency and predictive maintenance's value. Cloudera delivers the promise of connected manufacturing and the predictive maintenance use cases with Cloudera enabling data ingestion, storage and analysis from edge to AI.

Footnotes:

[1] Deloitte, "Making Maintenance Smarter: Predictive Maintenance and the Digital Supply Network," May 9, 2017.
[2] McKinsey & Co., "The Internet of Things: Mapping the Value Beyond the Hype," June 2015.

CLOUDERA

Cloudera, Inc. | 5470 Great America Pkwy, Santa Clara, CA 95054 USA | cloudera.com

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible—today and in the future.

To learn more, visit Cloudera.com and follow us on LinkedIn and X.