

Why do industrial enterprises need a data lake?

Table of contents

Introduction.....	3
Industrial data trends	4
What data issues are holding industrial companies back?	5
The value of a data lake for industrial enterprises	6
Data lake use cases	7
Summary.....	8



Introduction

Analyzing the massive quantities and types of data generated in an industrial setting can enable effective predictive maintenance, improved demand forecasting, device fleet management, and visibility into production challenges at all levels, and so much more. However, data silos make it difficult to capitalize on advanced, real-time, and predictive analytics or to use AI/ML to determine the best actions to take to improve production.

Using **data lakes** to store structured and unstructured data can help industrial organizations address these issues. This ebook demonstrates—through use cases and real-world examples—the value of using **data lakes** in industrial settings.



Industrial data trends

The machines, sensors, and devices found in industrial organizations all generate a massive, tangled trail of performance-tracking, real-time, and other unstructured data:

- An offshore oil platform typically generates up to **2TB of data daily**.¹
- Manufacturing companies generate significantly more than **1800 petabytes of data per year**, twice as many as the next closest industry.
- In 2019, an estimated **\$329 billion** was spent on the industrial internet of things (IIOT),² which is seen as critical in driving improved productivity, analytics, and workplace transformation and is **predicted to add \$15 trillion to the global GDP by 2030**.³

The potential uses for all this data, when combined with the power of IIoT, are immense. Real-time, streaming, and sensor data can be used for everything from proactively addressing a factory equipment issue to monitoring oil well performance to using smart meters to reach out to customers to mitigate utility overuse or anomalies. However, there are challenges along with the possibilities:

- The average organization deals with 31 unique data sources that can feed into their analytical systems.
- **Data scientists spend 80% of their time preparing data for analysis instead of analyzing it.**
- According to Harvard Business Review, data silos drive that 80% even higher and can adversely affect analytics initiatives.⁴

These statistics underscore the difficulty of capitalizing on all the data that industrial companies collect. Much of it cannot be shared because it is locked away in proprietary databases, automation systems, and on-premises storage. In short, it is virtually inaccessible.



What data issues are holding industrial enterprises back?

Picture this scenario: A plant manager wants to use condition-based alerts to look for ways to eliminate equipment downtime and increase throughput and asks a team from IT and engineering to help. While creating a data infrastructure, the team realizes that the specific sensor data needed for queries is in a proprietary database that cannot be directly integrated with analytics tools.

Extracting the data requires a data model, so a system administrator joins the team for that purpose. Meanwhile, important data that could be used to prevent failures, such as humidity readings and PLC data points that are unused outside the control system, is sitting idle, increasing the likelihood of unplanned downtime.

To use advanced analytics and AI for effective predictive maintenance, accurate forecasting, visibility into production challenges, and much more, industrial organizations must integrate and organize abundant, unstructured data streams. These streams originate from multiple machines with hundreds of different communication protocols, on-premises software applications, and automation system vintages.

Inaccessible or missing data can result in inaccurate demand forecasts and inventory management, along with supply chain disruption that can cost companies millions of dollars.

Data processing that lags behind production speed results in sub-optimal production and scheduling decisions.

Security is at risk because IIoT and supply chain management systems often require connection to environments outside an industrial enterprise's firewalls, virtual private clouds, and other cybersecurity installations.

All these data challenges—inaccessibility, silos, unstructured data streams, latency, and security—can be overcome.




The value of a data lake for industrial enterprises

Although traditional data warehouses, databases, and legacy data frameworks are not designed to address the opportunities of IIoT and machine-generated data, a data lake is. A data lake is a centralized repository that allows companies to store data as-is, without having to first structure the data, and then run different types of analytics to guide better decisions.

According to an Aberdeen report, organizations that have deployed data lake technology effectively are twice as likely to report that the speed and efficiency of collecting data is “highly effective” than those who have not. These organizations also report that 84% of information is delivered on time with 81% satisfied with the data accuracy.⁵ Organizations that use [AWS data lake technology](#) are also able to ensure that data security policies are enforced across multiple services and environments.

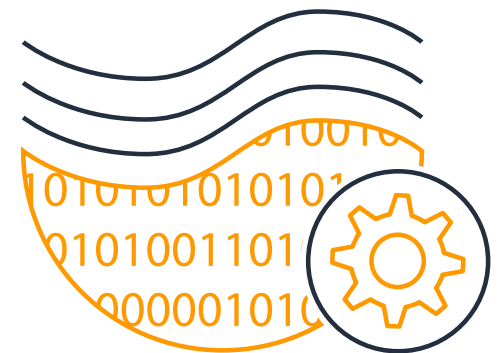
Using insights derived from industrial data sources such as the equipment itself, environmental conditions, and human observations, industrial companies can make decisions that improve the quality of outputs or increase accuracy. For example, a data lake based on [Amazon Simple Storage Service \(S3\)](#) has helped [Georgia-Pacific](#) improve asset utilization and avoid lost revenue from unscheduled production stoppages. Georgia-Pacific can now predict equipment failure 60 to 90 days in advance, which enables it to reduce unplanned downtime.

[INVISTA](#), a global chemical and fiber manufacturing company, built its data lake environment on AWS and uses it to power stronger process forecasting and inventory optimization. By using AWS to pull ordering patterns and other data from its SAP ERP system, INVISTA has a closed-loop, fully automated stocking process for high-moving spare parts that has shown significant return on investment. These are just two real-world examples of data lake use cases.



“We have better, up-to-date data about the performance of machinery throughout all our facilities.”

—Steve Bakalar,
Vice President of IT/Digital
Transformation, Georgia-Pacific



Data lake industrial use cases

Industrial enterprises can [use data lakes](#) in all kinds of ways to improve production, asset management, forecasting, inventory control and much more. Here are just a few examples.



Predictive maintenance

When data from equipment, assets, and products is easily accessed, industrial customers can address trouble spots before they cause real trouble. [Vantage](#), which designs and manufactures technology that can connect and electrify powertrains in heavy-duty vehicles, uses a platform based on an Amazon data lake to derive insights and predictive analytics models. They can use these insights and models to deliver real-time preventive actions, such as detecting cell-level defects early and mitigating them, directly to the smart vehicles themselves.



New, innovative products

Being able to analyze data that is stored all in one place, no matter its format, enables industrial companies to use the insights to develop innovative products faster. Ball bearing manufacturer [SKF](#) relies on its data lake to store metrics about product performance gathered in real-time as it is being used to drive new designs for their next generation of ball bearings and lubricants.



Product optimization

A data lake makes it easier for industrial enterprises to improve and reengineer their products. Silicon Valley's [Astera](#) is a fabless semiconductor company that develops purpose-built connectivity solutions for data systems. Astera uses [Amazon S3](#) to store 100 TB of data on which it runs multiple simulations in parallel to improve chip design quality.

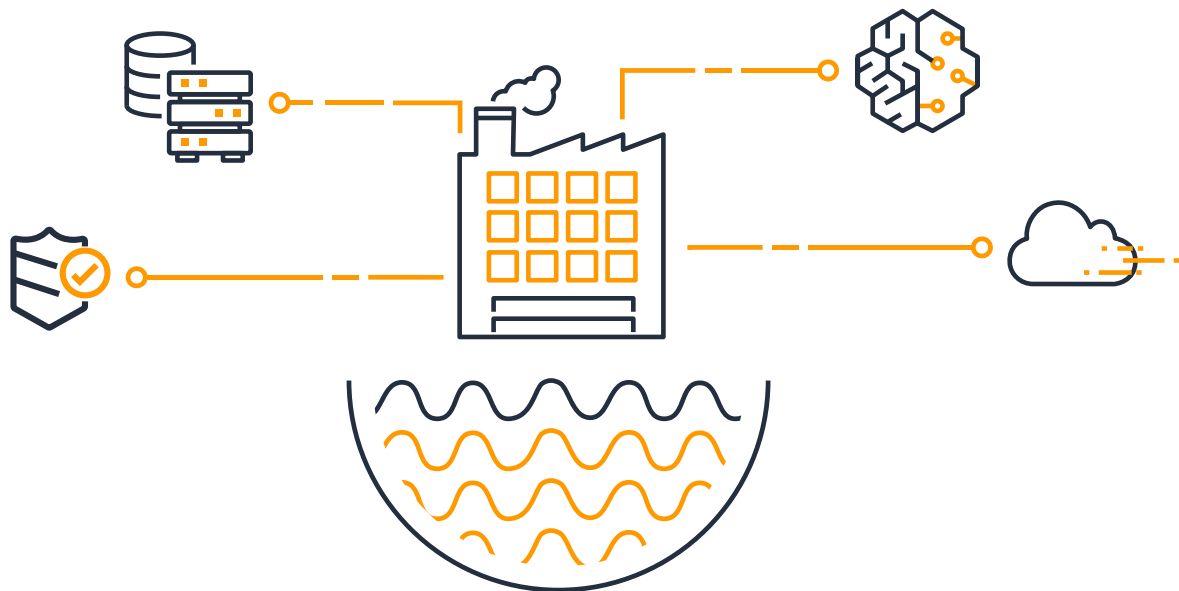
"AWS improved our chip design and verification environment."

—Sanjay Gajendra,
Chief Business Officer,
Astera Labs

Summary

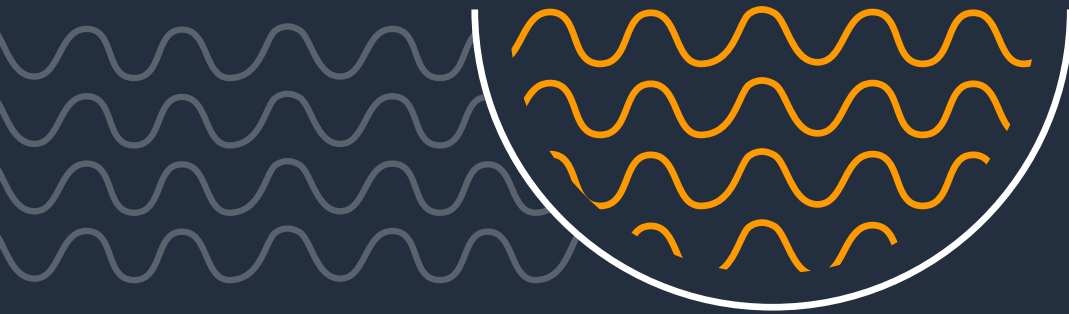
A data lake can store data generated in industrial settings as-is—streaming, semi-structured, relational, and structured. Built to facilitate business analytics, advanced and prescriptive analytics, AI/ML, and more, a data lake resolves issues such as inaccessibility, data integration, multiple formats, massive data storage, and multiple protocols.

Industrial enterprises can be sure they have all their data in one place. From there, they can derive the insights, predictions, and models they need to improve processes, proactively address equipment maintenance, optimize output, and deliver high-quality products.



Leading industrial enterprises are using AWS to make the most of all their data.

[Learn how](#) using an AWS data lake to store structured and unstructured data can help your industrial company improve agility, optimize business operations, and lower costs.





Copyright © 2020, Amazon Web Services, Inc. or its affiliates.

Resources

¹<https://blogs.cisco.com/news/digital-transformation-in-the-oil-gas-industry-drill-data-drill#:~:text=For%20example%2C%20a%20typical%20offshore,to%20platform%20production%20and%20safety>

²<https://www.zdnet.com/article/what-is-the-iiot-everything-you-need-to-know-about-the-industrial-internet-of-things/>

³https://www.accenture.com/mz-en/_acnmedia/Accenture/next-gen/reassembling-industry/pdf/Accenture-Driving-Unconventional-Growth-through-IIoT.pdf

⁴<https://hbr.org/2016/12/breaking-down-data-silos>

⁵<https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf>