See more possibilities. Seize more opportunities.

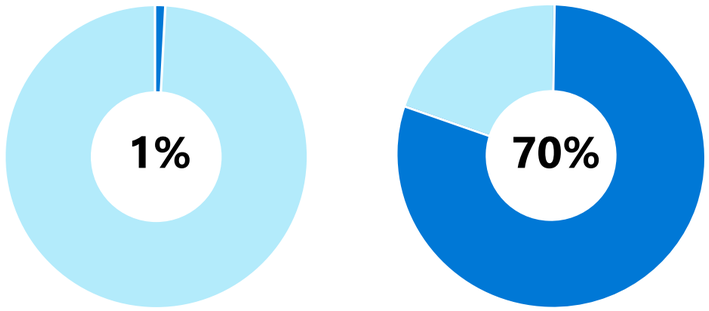
Your guide to advanced analytics and machine learning for hybrid cloud

Microsoft Azure

CLOUDERA

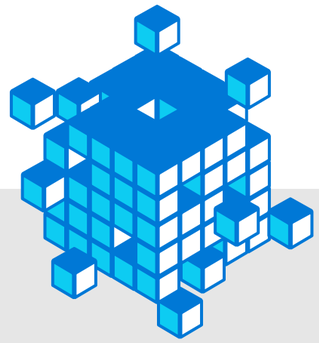
Have we put the "big" in big data yet?

The world is fast becoming a data-driven place with big data promising huge advances in our ability to make decisions and predict outcomes. The trouble is that right now, big data is generally only using a tiny part of its true potential. With IDC predicting that by 2020, 30 billion connected devices will provide 440 times more data than there was in 2006, data growth isn't the problem. The problem is actually using that vast ocean of data. Currently, only 50% of structured data is used in decision-making. And even worse, according to the Harvard Business Review only 1% of unstructured data is used. Imagine trying to make good decisions based on half the information available, let alone just 1%.



> Now imagine the ability to make decisions based on 70% or 85% of the available intelligence.

We are just starting to get a glimpse of what that might look like. Businesses and organizations are in a race to harvest the benefits of big data. Gartner predicted that organizations will have increased their investment in analytics from $31 billion in 2013 to an estimated $114 billion in 2018. And 60% of executives in 2016 believed that analytics would disrupt their industry within the next three years—and it has. According to McKinsey Global Institute, disruption fueled by analytics could generate much bigger benefits and savings. In the US for example, where healthcare spending is 18% of GDP, savings could amount to $600 annually per person, or 1% to 2% of GDP. In transportation, data-driven thinking and innovation initiatives (including high-scale, real-time analytics), could potentially create between $850 billion to $2.5 trillion in economic impact. These examples are just a slice of the opportunity. But to achieve them, the way we mine for data insight must take a leap forward with advanced analytics, machine learning (ML), and artificial intelligence (AI).



Industrializing inspiration: Advanced analytics including ML and AI at scale

Everyone from CIOs to IT departments and data architects wants the benefits of better analytics, but to achieve big data's true potential to mine for patterns and anomalies, in huge data sets must be made frictionless, automatic, and fast.

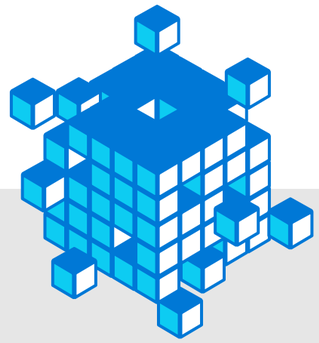
To get to that level of insight-harvesting speed, your organization should now be turning its consideration to powerful, coordinated data tools, such as data repositories, data lakes, data warehouses, and advanced analytics. With these tools, IT professionals can better support strategic projects, and your business data can be put to strategic use for a transformative impact without putting strain on IT time and resources.

In addition to freeing IT resources, coordinated data environment deployments help reduce, if not eliminate, the repetitive tasks required to find nuggets of insight hidden in mountains of data. The main source of all this manual and time-consuming labor is centralized management and governance of data within uncoordinated, unintegrated, and complex environments. IT tends to be saddled with the whole responsibility. But by putting inventory, data movement, and overall governance tasks on a consolidated data layer and giving them a shared data context, these tasks can be reduced or eliminated. This frees up IT to support business goals, not just wrangle infrastructure and administration.

Stop herding cats: Coordinate the data approach

Traditional implementation approaches for data analytics tend to be one-off implementations that take care of the need at hand without regard to larger picture solutions. Over time, this leads to a hodgepodge of different uncoordinated strategies that don't take into consideration multiple uses, goals, and users. The strain of supporting multiple single-use approaches leads to extreme inefficiencies for deployments and long-term management issues. The better approach would be a single schema that effortlessly supports multiple functions and uses across the organization—a practical methodology of multifunctionality.

By allowing multiple, discrete, and simultaneous different uses under the same system, you get additional economies of scale by exploiting cloud resources to speed deployment. It's a decentralized approach in terms of users, but centralized in terms of strategy. Certainly, developing this kind of approach takes more coordination and infrastructure at the start, but it enables the configuration, support, and update of those analytics services from a single point of management, while increasing productivity by reducing time to task completion. It's well worth the up-front investment when every job thereafter benefits.



Enterprise security, privacy, and governance at the C-level

Within the unified strategy of data access, Cloudera recommends developing a single point of data storage, management, and governance. This allows for strategic monitoring and management of an organization's data assets as well as greater efficiency of access and management. We deploy a data storage layer that also provides economies of scale for data governance and overall data security and privacy. This also creates visibility and inventory of a company's complete data assets so a chief data officer (CDO) can curate and govern for quality, consistency, informational tagging, and efficient linking across datasets. The chief security officer (CSO) also has visibility into which data assets require additional protection for personally identifiable information, payment compliance, or other concerns. Both officers are enabled to achieve the wider corporate strategic goals of increasing customer privacy and supporting regulations, such as GDPR.

The big 4 criteria for big data success

When developing a modern data strategy, keep in mind four key elements that will drive success in the long term. Ask yourself if the solution is:

1. Elastic
   - Does it size-compute and store data independently?
   - Is it able to grow and shrink clusters dynamically, as needed?
   - Does it only charge for what you use on ad-hoc, transient workloads?

2. Hybrid/multi-cloud
   - Are you able to preserve business flexibility and data portability?
   - Does it help minimize cloud lock-in by running in any one of the three major public cloud providers or in the private cloud?

3. A shared data experience
   - Does it provide a centralized experience that leverages AI and machine learning for reduced friction, lower resource needs, and faster time-to-insight?

4. Enterprise grade
   - Does it reduce risk with comprehensive manageability, availability, security, and governance required for production big data workloads?

CASE STUDY

Komatsu: Mining for greater insights

Situation

Komatsu Mining is a global mining equipment and services provider specializing in solutions for the excavation of energy, industrial, and hard rock minerals. Their Industrial Internet of Things (IIoT) products (under the name JoySmart Solutions) help customers optimize machine performance using real-time data and analytics obtained from its smart connected devices and assets.

Challenge

Originally, the company's legacy data warehouse supported this IIoT service. However, as customer demand grew and more machines were connected, staff found they needed a new approach. Data growth was anticipated to quickly reach 30 terabytes per month. The old environment was limited in its ability to scale and grow. Komatsu needed a new solution that could keep up.

Solution

Komatsu's teams partnered with Cloudera and Microsoft on a cloud-based IIoT analytics platform that provides scalability, performance, and flexibility to support global service teams. The platform ingests, stores, and processes a wide variety of data collected from mining equipment operating around the globe, often at very remote locations in harsh conditions. A single machine can generate 30,000 to 50,000 unique time-stamped records in one minute.

Using this solution, teams can now more easily analyze data from the company's mining machines, as well as from third-party programmable logic controller (PLC) based equipment, to get a systems view of mining operations. The company's data scientists can also produce machine learning models and better results faster than was previously possible.

Results

A more complete picture of machine health and operations in each mine enables JoySmart teams to partner with their customers to identify ways to improve equipment safety, productivity, and operating costs. Based on the new data analysis, Komatsu was able to make recommendations to a large coal mining company that enabled them to double the daily utilization of their longwall system.

Making insight harvested from big data a business reality

Working with Microsoft Azure, Cloudera is in the business of powering better business decisions through data insight. We devise strategies that extract business value from big data via advanced analytics, making business transformation a reality. Uncover hidden insights into the nature of your customers, utilize your Internet of Things connectivity to build better products, and lower your risk profile with a faster response to evolving threats. Cloudera and Azure put the power of big data in your hands; use it to make a difference.

Explore more

Microsoft Azure

CLOUDERA