CLOUDERA

Empowering Innovation

Helping Research, Innovation and Data Science Teams Make an Impact

Key Insights for Data & Analytics Leaders

Table of Contents

Introduction

As organizations seek to transform into data-driven enterprises, data and analytics leaders play a pivotal role as champion, evangelist and strategist to bring about this transformation. But to be effective, data leaders must develop and empower the people that focus on innovation — data science teams, research and development groups, and data-driven innovation and transformation teams.

These mission-critical teams must have the resources they need to deliver impact and benefit to the organization. They need tools for the entire data to AI lifecycle — to explore new datasets, in a variety of formats and volumes; test new analytics, machine learning and explore ideas for the application of AI for advanced predictive modeling and analytics.

Data leaders must consider how to enable everything from experimentation to operationalization — moving data initiatives from "is it possible?" to "it's operational."

Second in our Data Leaders series, this ebook will help data leaders advance their innovation initiatives, highlighting important concepts and details required to empower successful teams.

The Case for Innovation

Why invest in a robust data-driven innovation, analytics, research and discovery or data science team — and how do enterprises measure the impact? Because the accelerated uses of data can drive significant business outcomes, giving the enterprise a competitive edge.

But deciding to invest may be "the easy part." Delivering impact requires empowering a team with the tools, processes and resources they need to make an impact.

17.9%

The successful application of data science could lead to a 17.9% uptick in revenue for organizations with a "mature" data science capability.[1]

$2.01B

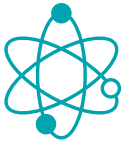
If the median Fortune 1000 business increased the usability of its data by just 10%, it would translate to an increase in $2.01 billion in total revenue every year.[2]

87%

87 percent of business executives say frontline staff would benefit from improved insights.[3]

Key Concepts: Data Science, Artificial Intelligence and Machine Learning

Data Science, AI, and ML are often lumped together or used interchangeably, it's worth understanding the difference between these interrelated terms. While they are connected by their reliance on data, they each have their own specific applications and meaning.

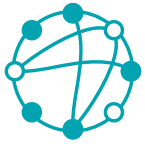


Data Science

The application of advanced analytic techniques and scientific principles to extract valuable information from data for business decision-making, strategic planning and other uses.

Artificial Intelligence

AI Is the simulation of human intelligence processes by machines, especially computer systems.



Machine Learning

A subfield of AI that a data scientist may leverage which allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.

Along with standard tools such as statistical programming languages and visualization solutions, data scientists develop machine learning models trained to evaluate and make predictions across vast data sets, thus automating and accelerating time to insight.

Real-World Transformative Use Cases

A wide range of industries are embracing data science and delivering on the promise that data-driven initiatives net real-world results. Data science, analytics and innovation teams form the epicenter of a number of transformative use cases:

- **Manufacturing:** Known as "Industry 4.0," advanced manufacturing technology in concert with data-driven processes, like predictive analytics and machine learning, leverage IoT networks and enhanced system controls to keep production lines running efficiently while also minimizing downtime.
- **Financial Services:** In FinServ, data science teams work in support of fraud detection, customer segmentation, automated risk analysis and other key business drivers.
- **Insurance:** In auto insurance, data drives real-time "pay as you drive" applications. Home insurers leverage data to manage risk (think security cameras or leak detection sensors). Health insurers may look at hospital usage, financial risk tolerance and many other factors.
- **Retail:** Leveraging data from online and in-store sales, as well as from smart home devices and other sources, retailers are driving enhanced customer interactions and fine-tuning both their inventories and their delivery systems. They're also tapping the emerging Internet of Behavior (IoB) to make intelligent decisions related to customer behaviors.
- **Telco:** In the telecommunications industry, companies are tapping data to drive product innovations, contextualize promotions, reduce fraud loss, deliver better customer service and improve network security.

Structuring Data Science, Analytics and Innovation Programs for Success

In order to drive effective use of the vast volumes of data at hand, data leaders need to give thoughtful consideration to how their data science teams are structured. The Data Science Association[4] puts it this way: "Given how important data science has grown, it's important to think about what data scientists add to an organization, how they fit in and how to hire and build effective data science teams."

Key Roles on the Team

The data science team should include a range of expertise. As Gartner[5] describes, "the complexity of data science projects requires a whole cast of characters." These may include[6]:

Data Scientist

Data scientists discover and interpret rich data sources, create visualizations and use machine learning to build models that aid in creating actionable insight from the data

Data Engineer

Data engineers make the applicable data accessible and available for data science efforts. They design, develop, and code data-focused applications that capture and cleanse data

Data Science Architect

Data science architects design and maintain the architecture of data science applications

Data Science Developers

Data science developers design, develop and code large data (science) analytics applications to support scientific or enterprise/ business processes. This role enables models to be deployed (i.e., use a model in production) and requires some expertise in data science, as well as knowledge of how to effectively develop software applications.

Product Owner

The product owner is the central point of product leadership — the person who decides which features and functionality to build, the order in which to build them and what aspects of them to observe and analyze

Data or Business Analyst

Data/Business analysts analyze a large variety of data to extract information about system, service, or organization performance and present them in usable, actionable form

The Relationship Between Leader and Team

Data and Analytics Leaders need to work in close cooperation with their data science teams reinforcing the following key perspectives:

1. Teams are formed to be innovative, to experiment and to discover
2. Leaders empower and manage these teams effectively to ensure they stay focused and have the resources needed to do their work
3. Teams apply innovative tools and techniques to deliver the best, most transformative applications
4. Leaders bring these applications into production to drive real-world business outcomes
5. Leaders are advocates for the team, educating other organizations on the experimental nature of data science, and shielding the team from "normal" software development schedules.

Team Models

There are different ways to structure innovation teams at enterprises — and structures may evolve over time. Whichever model one pursues, it is critical that teams have strong working relationships with other key players in the business. They'll need to work in close collaboration with:

- **Line Of Business (LOB) –** The key stakeholders in any data-driven efforts, LOB leaders will have insight into the specific business needs and will play a key role in operationalizing the outcomes of data analysis.
- **IT –** The IT team will support the infrastructure needed to produce, ingest, analyze and ultimately operationalize business data.
- **Security –** Data security is paramount for any organization. As businesses look to make better use of their data resources, data science teams will need to move in lockstep with policies established by the CISO to ensure customer data, business intelligence and other vital resources are not compromised.

Centralized or Decentralized?

Within a given organization, a data science team may be either centralized or decentralized.

- In the decentralized model, teams work on specific projects within business lines or functions. The benefit: They maintain a deeper understanding of particular business needs and processes that could result in higher impact applications.
- In the centralized model, teams operate in a central hub developing and delivering projects across multiple business units. The benefit: Teams gain greater visibility across the enterprise, high-impact applications can be more easily distributed and company-wide policies and standards around the use of data are maintained.

Encouraging Innovation

The Innovation Lab

Some companies may choose to centralize their data-innovation efforts in an innovation lab, a sandbox for experimentation. In this model, a dedicated environment encourages data science teams to explore the possibilities.

When data scientists are free to work on experiments, to prove concepts or ideas, they can discover new use cases and potentially have more breakthroughs. As one analyst[7] put it: "A good data scientist is a skeptic who questions everything, even scientific laws that have previously been recognized as true." An innovation lab can be a place to ask and answer those questions.

To be effective, an innovation lab will need to give teams access to the infrastructure and tools they need for experimentation, as well as access to the LOB owners whose problems they are looking to solve. In turn, LOB owners should be encouraged to bring ideas to the innovation lab for testing.

A "Center of Excellence"

Some have looked to the Center of Excellence (COE) concept as a way to bring to life the promise of data science — a means to drive innovation in the uses of data across the organization.

A COE can develop standards and build both consensus around, and enthusiasm for, advanced uses of data. "A key component of many data science organizational charters is to build a center of excellence to advance the discipline. This includes developing best practices and fostering a data-driven culture for the company," according to Microsoft[8].

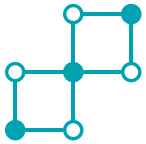
A COE can be especially effective in identifying and addressing low-hanging fruit and can help set expectations about outcomes of a data science effort. It can be a place where the data science team leverages MLOps to scale-up analytics. McKinsey[9] describes the need to implement an infrastructure "that will allow you to mass-produce and scale your AI projects." A COE can fill that role.

"A good data scientist is a skeptic who questions everything, even scientific laws that have previously been recognized as true."

Analytics Insight

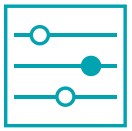
Critical Resources for Innovation

In order to drive success, data and analytics leaders need to equip their data science teams with key resources. These include:



Data

Teams need to know they will have access to the data that drives their efforts. That means the business needs an established data policy that gives teams priority access to the data required to experiment, discover and develop meaningful applications.



Data Platform

Teams need an environment that's scalable enough to run, test and iterate models. An enterprise platform accelerates data science and machine learning projects by providing a robust yet familiar environment for data science and data engineering with self-service access to governed business data.

Tools

These may include both single-server tools and frameworks (Python, NumPy, pandas, SciPy, Scikit-learn, NLTK, TensorFlow Jupyter) as well as large-scale tools and frameworks (Spark MLlib, Stanford CoreNLP, TensorFlowOnSpark/Horovod/MLeap, Apache Zeppelin). The key is to ensure that data science professionals have the freedom to choose from among a range of preferred tools according to their needs and wants.

Security

A robust data science program will include security tools including capabilities that support an overall security model, authentication, authorization, encryption and other key safeguards.

Keeping Innovation Projects on Track

Overall, the goal for data and analytics leaders should be to drive an organizational effort that maximizes the time teams spend on high-value projects while minimizing distractions and bottlenecks.

To that end, data leaders can give high-level guidance to help data science teams focus on the projects that matter. They can:

- Prioritize efforts based on business needs and likely business value
- Focus on data sets that can be shared or reused across multiple use cases, as well as sets that are already relatively clean, so that teams can make maximum productive use of their time
- Emphasize agility, encouraging teams to accelerate transformation by testing more ideas, more quickly. ML and AI will be key to driving these efforts.

The democratization of machine learning platforms is proliferating analytical assets and models. The challenge now is to deploy and operationalize at scale. Data and analytics leaders must establish operational tactics and strategies to secure and systematically monetize data science efforts.

Gartner.com

How Cloudera Empowers Data Science, R&D and Innovation Teams

Measuring the Impact

How do innovation, R&D and data science teams measure impact effectively? In many (many) ways. Most teams use a combination of quantifiable (revenue, cost savings, productivity increases) and unquantifiable metrics (security, flexibility) — as well as data-driven anecdotes to underscore impact.

An example from life sciences is included in the below Chief Data Leader quote from a recent study by Forrester that looked specifically at the economic impact of analytics and data platforms.

Another example from life sciences is IQVIA, a global company with a wide variety of research use cases. IQVIA has a shared data lake that supports many different data types, projects and teams. Notably, 70 different teams, with about 1,500 to 2,000 people, all have access to the data.

For IQVIA, success is measured in time to value, across these many teams and projects. Technical improvements, like faster query speed and more efficient analytics, all help support faster delivery. IQVIA can generate insights from data for its customers in seconds, rather than in days, weeks or months. This massive boost in performance enables their clients — life sciences companies — to innovate faster and save lives.

"We would have never been able to do what we're doing now with our prior technology stack. CDP Public Cloud has enabled R&D staff to improve productivity by 20% to 40%, reducing preclinical trials from eight months at the high end down to two months at the low end. We've also doubled the probability that a development is successful at trials, moving from 5% to 10% likelihood."

— Chief data and analytics officer, pharmaceuticals

Cloudera Brings to Life the Promise of a Robust Data Science Capability

Hybrid Data Platform

Delivering on the hybrid data platform promise of modern data architectures, CDP features a unified data fabric, open data lakehouse and scalable data mesh to help businesses manage and analyze data of all types—machine data, structured data, transactional data, and unstructured data—to bring the right analytics to the right cloud at the right time.

Full Data Analytics Lifecycle

Cloudera enables the Edge to AI analytics lifecycle, which includes the collection of large volumes of data and the development and deployment of machine learning models to devices allowing them to respond in a much more dynamic, personalized manner.

Build, train, deploy ML applications at scale

To bring agility and scalability to data science teams, Cloudera Machine Learning enables users to break down data and workflow silos, deploy new machine learning workspaces for teams in a few clicks and use their favorite tools to achieve scalability without administrative overhead.

CLOUDERA SDX

And the security to govern it all

SDX is a fundamental part of Cloudera Data Platform architecture, delivering an integrated set of security and governance technologies built on metadata, with persistent context across all analytics as well as public and private clouds.

Cloudera enables enterprise data science teams to collaborate across the full data lifecycle with immediate access to enterprise data pipelines, scalable compute resources and access to preferred tools.

Visit Cloudera.com to learn more about how Cloudera empowers data science, research and innovation.

Resource Library

Cloudera Machine Learning

Test Drive Cloudera Data Platform

Production ML for Dummies

Take a Guided Tour of Cloudera Machine Learning

The Definitive Guide to the Machine Learning Lifecycle

Glossary of Key Terms*

- **Analytics –** A catch-all term that includes applying the breadth of BI capabilities to a specific content area; a way to exploit huge mounds of internally generated and externally available data.
- **Advanced Analytics –** Autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools to discover deeper insights, make predictions or generate recommendations.
- **Artificial Intelligence –** AI Is the simulation of human intelligence processes by machines, especially computer systems.
- **Business Intelligence –** A technology-driven process for analyzing data and delivering actionable information to drive informed business decisions.
- **Data Science –** The application of advanced analytic techniques and scientific principles to extract valuable information from data for business decision-making, strategic planning, and other uses.
- **Digital Transformation –** Incorporation of computer-based technologies into an organization's products, processes, and strategies.
- **Hybrid Cloud –** Policy-based and coordinated service provisioning, use, and management across a mixture of internal and external cloud services.
- **Hybrid Data Cloud –** A platform designed for freedom of choice — any cloud, any analytics, any data — for faster and easier management of enterprise analytics.
- **Machine Learning –** A type of AI that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.

* All definitions sourced from Gartner and TechTarget.

Learn More

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more at cloudera.com | US: +1 888 789 1488 | Outside the US: +1 650 362 0488

Sources

[1] https://www.r-craft.org/r-news/lead-your-data-science-charge-now-to-reap-benefits-down-the-road/

[2] https://www.datascienceassn.org/sites/default/files/Measuring%20Business%20Impacts%20of%20Effective%20Data%20I.pdf

[3] https://media.thoughtspot.com/pdf/HBR-ThoughtSpot-The-New-Decision-Makers.pdf

[4] https://www.datascienceassn.org/sites/default/files/Building%20Data%20Science%20Teams.pdf

[5] https://www.gartner.com/smarterwithgartner/how-to-staff-your-ai-team

[6] https://www.datascience-pm.com/8-key-roles-within-a-data-science-project/

[7] https://blockgeni.com/data-scientists-should-stay-curious-and-open-minded/

[8] https://medium.com/data-science-at-microsoft/building-a-center-of-excellence-for-data-science-38f3d6098fa1

[9] https://www.mckinsey.com/about-us/new-at-mckinsey-blog/three-experts-offer-an-inside-look-at-the-state-of-ai

CLOUDERA