# The Business Value of AWS Data Lakes, Analytics, and ML Services

IDC White Paper, Sponsored by Amazon

Authors: Ritu Jyoti and Harsh Singh

June 2020

# Table of Contents

Click on any section title or page number to navigate to each.

## Business Value Highlights

415%
five-year ROI

9-month
payback period

$6.15M
average annual benefit per organization

48%
reduced total cost of operations

42%
reduction in IT infrastructure costs

4.1M
additional new revenues gained per year

17%
more efficient analytics teams

79%
reduced time to run queries

37%
increased number of queries run

76%
reduction in unplanned downtime

# Executive Summary

IDC predicts that by 2022, more than 46% of global GDP will be digitized, with growth in every industry driven by digitally enhanced offerings, operations, and relationships. Successful digital transformation relies on converting data into actionable insights, and this increasing dependence on data-powered insights is contributing to a new era of the data age. Data is rapidly becoming a strategic asset for every organization. Businesses want more value from their data that is coming from newer sources, is increasingly diverse, and is growing exponentially. These digital transformation (DX) initiatives will be supported by analytics and artificial intelligence/machine learning (AI/ML) capabilities that are used by many users within an organization.

Traditionally, businesses used siloed data warehouses (for example, one for ERP systems and another for sensor and social media data) as repositories of integrated data from one or more disparate sources, and as a system used for reporting and data analysis for different groups. But as traditional data warehousing (DW) approaches unfortunately don't scale, organizations are now moving to data lake architectures that extend or evolve traditional data warehousing, store any data in any format, are durable and available, and can scale to exabytes in size. In addition, businesses are looking for data lake architectures that are secure, compliant, and auditable and can run any type of analytics and ML services.

To support these changing customer requirements, AWS provides open, secure, scalable, and cost-effective infrastructure that enables easy-to-build data lakes and analytics. Businesses can use the right tool for the job without needing to move or transform data for different analytics approaches. AWS provides a comprehensive set of tools that goes beyond standard security functionality such as encryption and access control to proactive monitoring and unified management of security policies.

In order to validate the benefits of the AWS solution, IDC interviewed 11 organizations that use AWS data lakes, analytics, and machine learning services. The survey data obtained and applied to IDC's Business Value model showed that study participants are realizing significant value with AWS. IDC calculates that study participants will achieve average annual benefits of $6.15 million per organization ($100.4K per business application), which would result in a five-year return on investment (ROI) of 415% by:

- Providing organizations with the benefit of agile, scalable, cost-effective, and high-performing data lakes, analytics, and machine learning services
- Reducing the overall time required to manage and support data lakes, analytics, and machine learning activities
- Fostering more efficient analytics teams through better access to data and analytical tools
- Giving organizations the ability to run more analytics queries while at the same time reducing the time required to complete each query

# Situation Overview

To thrive in the digital era, businesses need to know and serve their customers better, make their operations efficient, reduce business risks, and do effective strategic planning. However, today, organizations store, analyze, learn, and apply information in siloes. Many organizations have poor institutional-knowledge retention and dissemination and, as a result, don't have the ability to respond in the moment or predict future results. They are unable to synthesize diverse internal and external data sources into information, leading to substandard strategic decisions and risk management practices. They don't have enough visibility into end-to-end business processes, limiting their abilities of informed decision making. Organizations also lack visibility into internal and external social networks and stakeholder communities, don't have unified customer transactional, behavioral, interactional, or attitudinal data, and hence are unable to have a 360-degree view of the customer, leading to substandard personalization. It takes them too long to move from data to information to knowledge to wisdom, and therefore they are unable to act within the necessary time windows.

As enterprises scale their use of modern technologies for complete instrumentation, integration, and insight, they will be able to expand their scope by offering a wider variety of experiences that demonstrate increasing value as the organization learns what is most desirable and efficient.

Enterprises that can overcome these challenges will learn as a single entity and at scale. In such an enterprise, the data generated from products, services, experiences, and ecosystems will inform and drive intelligent decision making as opposed to simply being a by-product of offline decision support systems. Those that can achieve this economy of intelligence will have a competitive advantage, just as those organizations in the past that achieved economies of scale had an advantage over their peers. As enterprises scale their use of modern technologies for complete instrumentation, integration, and insight, they will be able to expand their scope by offering a wider variety of experiences that demonstrate increasing value as the organization learns what is most desirable and efficient. For example, as a player in addressing global food demand, Bayer Crop Science, a subsidiary of Bayer Technologies, is well on its way to achieving economies of intelligence. The company has been working on several AI, analytics, and data-enabled initiatives to build tailored solutions to sustainably manage resources, improve productivity, interact with customers, and drive better crop yield. The company's CIO has highlighted the use of AI-based simulation in R&D, the use of ML in supply chain planning, and the use of predictive analytics in understanding customer behavior. These efforts have resulted in quantifiable, material, and publicly disclosed benefits, such as decreases in product development cycles, transportation costs, and the CO2 footprint, as well as improvements in customer retention *(see IDC PeerScape: Practices for Building Your Artificial Intelligence Capabilities, IDC #US45377019, August 2019).*

## AWS Data Lakes, Analytics, and ML Services Overview

AWS delivers an integrated suite of services that help quickly and easily build and manage a data lake for analytics. AWS-powered data lakes can handle the scale, agility, and flexibility required to combine different types of data and analytics approaches to gain deeper insights. AWS gives customers a wide array of analytics and machine learning services for easy access to all relevant data, without compromising on security or governance.

Modern analytics requires different types of analytics approaches: data warehousing, Big Data processing, ETL, BI, streaming analytics, and operational analytics to get to the right insights and answers. Developers use a wide variety of languages including R, Scala, and Python to build analytics applications based on which are best suited for their specific use cases. The Jupyter Notebook — a free, open source, interactive web tool that allows developers to create and share documents that contain live code, equations, visualizations, and narrative text — has exploded in popularity over the past couple of years. This rapid uptake has been aided by an enthusiastic community of user-developers and a redesigned architecture that allows the notebook to speak dozens of programming languages. Uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. AWS provides a mature set of analytics services that run against the open data lake so that businesses can use the right tool for the job without needing to move or transform data for each different analytics approach. AWS provides a comprehensive set of tools that goes beyond standard security functionality such as encryption and access control to proactive monitoring and unified management of security policies.

AWS gives customers a wide array of analytics and machine learning services for easy access to all relevant data, without compromising on security or governance.

AWS provides a wide range of options to transfer data to the cloud, with both batch/bulk on-premises data movement and real-time data movement offerings. Once data is ready for the cloud, AWS makes it easy to store data in any format, securely, and at massive scale with Amazon S3 and Amazon Glacier. Amazon S3 is secure, scalable, and durable object storage. Amazon Glacier is an online file storage web service that provides storage for data archiving and backup. To make it easy for end users to discover the relevant data to use in their analysis, AWS Glue automatically creates a single catalog that is searchable and queryable by users.

AWS offers a broad set of cost-effective analytic services that run on the data lake. Each analytic service is purpose-built for a wide range of analytics use cases such as interactive analysis, Big Data processing using Apache Spark and Hadoop, data warehousing, real-time analytics, operational analytics, dashboards, and visualizations. These services natively work together to support a seamless analytics experience.

For interactive analysis, Amazon Athena makes it easy to analyze data directly in S3 and Glacier using standard SQL queries. For Big Data processing using the Spark and Hadoop frameworks, Amazon EMR provides a managed service that allows for easy, fast, and cost-effective processing of vast amounts of data. For data warehousing, Amazon Redshift provides the ability to run complex analytic queries against petabytes of structured data, and includes Redshift Spectrum, which runs SQL queries directly against exabytes or smaller sizes of structured or unstructured data in S3 without the need for unnecessary data movement. For real-time analytics, Amazon Kinesis makes it easy to collect, process, and analyze streaming data such as IoT telemetry data, application logs, and website clickstreams. For operational analytics such as application monitoring, log analytics, and clickstream analytics, Amazon Elasticsearch Service allows organizations to search, explore, filter, aggregate, and visualize data in near real time. For dashboards and visualizations, Amazon QuickSight provides a fast, cloud-powered business analytics service that enables streamlined visualizations and rich dashboards that can be accessed from any browser or mobile device.

IDC conducted research that explored the **value and benefits** for organizations using the AWS data lakes, analytics, and machine learning services.

For predictive analytics use cases, AWS provides a broad set of machine learning services and tools that run on the data lake on AWS. For expert machine learning practitioners and data scientists, AWS provides AWS Deep Learning AMIs that make it easy to build deep learning models and build clusters with ML and DL optimized GPU instances. AWS supports all of the major machine learning frameworks (including Apache MXNet, TensorFlow, and Caffe2) so that organizations can bring or develop any model they choose. For developers who want to get deep with ML, Amazon SageMaker is a platform service that simplifies the process of building, training, and deploying ML models by providing everything organizations need to connect to their training data, select and optimize the best algorithm and framework, and deploy their model on auto-scaling clusters of Amazon EC2. SageMaker also includes hosted Jupyter notebooks that allow easy exploration and visualization of training data stored in Amazon S3. For developers who want to plug pre-built AI functionality into their apps, AWS provides solution-oriented APIs for computer vision and natural language processing. These application services let developers add intelligence to their applications without developing and training their own models.

## The Business Value of AWS Data Lakes, Analytics, and ML Services

### Study Demographics

IDC conducted research that explored the value and benefits for organizations of using the AWS data lakes, analytics, and machine learning services. The project included 11 interviews with organizations that were using this solution and had experience with or knowledge about its benefits and costs. During the interviews, companies were asked a variety of quantitative and qualitative questions about its impact on their IT, data lakes and warehousing, and analytics operations, as well as on their businesses and costs.

Interviewees hailed from organizations representing a diverse mix of geographic locations, including the United States, United Kingdom, Singapore, and Hong Kong.

Table 1 presents the study demographics and profiles. The organizations interviewed had a base of 31,682 employees, indicating the involvement of several large companies. This workforce was supported by an IT staff of 1,129 engaged in managing 490 business applications for 24,404 IT users and 531,000 customers. Annual revenue across all companies was $5.7 billion. (Note: All numbers represent averages.)

Interviewees hailed from organizations representing a diverse mix of geographic locations, including the United States, United Kingdom, Singapore, and Hong Kong. The companies represent a broad mix of vertical industries, including the transportation, financial services, energy, healthcare, information technology, manufacturing, retail, telecommunications, and travel/hospitality sectors.

Table 1

Firmographics

| | Firmographics | Average | Medium | Range |
|---|---|---|---|---|
| # | Number of employees | 30,124 | 5,000 | 17 to 86,000 |
| # | Number of IT staff | 1,129 | 200 | 2 to 6,000 |
| # | Number of IT users | 24,404 | 5,000 | 17 to 86,000 |
| # | Number of external customers | 531K | 150K | 16 to 2M |
| # | Number of business applications | 490 | 120 | 18 to 3,000 |
| % | Percentage of internal business applications | 70% | 77% | 17% to 95% |
| % | Percentage of external business applications | 30% | 23% | 5% to 83% |
| $ | Revenue per year | $5.7B | $11B | $1.9M to $25B |

**Surveyed Countries & Industries**

United States (5), United Kingdom (3), Singapore (2), Hong Kong

Transportation (2), Financial Services (2), Energy, Healthcare, Information Technology, Manufacturing, Retail, Telecommunication, Travel/Hospitality

Source: IDC, 2020

## Choice and Use of the AWS Data Lakes Solution

The companies that IDC surveyed described their usage patterns for the AWS data lakes, analytics, and ML services as well as provided snapshots of their IT and business environments. They also discussed the rationale behind their choice of AWS. Interviewed customers cited several factors for their choice, including the quality and variety of data outputs, ease of management, the benefit of a well-designed disaster recovery program, and the company's singular focus on innovation. Study participants elaborated on these benefits:

- **Better business-related data:** "The reason we moved to Amazon is because we can do more. For example, when one of our products for one of our partners is advertised, we can bring up that data instantly. We can see how that advertisement is affecting sales, and then make decisions based on that: what's shown on our point-of-sales data, what we sell in our stores, and where we're placing it."
- **AWS focus on innovation:** "AWS is very good. They are a big innovator … because of the tools and the machine learning elements they give us. They have a very open staff, and their account managers are very helpful and give us the appropriate tools, especially for quantitative analysis. AWS has been successful for a long time in running their Amazon ecommerce and they've already built those tools to analyze patterns from their clients. AWS data lakes, analytics, and ML leverages a lot of tools and technology from the Amazon website. I really like the interface and other aspects. So, seeing those kinds of technologies in the enterprise was very appealing."
- **Easier management:** "AWS gave us the promise of commodity-managed services like databases, streaming buffers, and data warehousing. We didn't need huge staffs of IT, database administrators, and so forth because their promise was they would take that off our hands. We could then work on what's really important to the company."
- **Good partnership and disaster recovery plan:** "We wanted a company that was already well established, knows our business model, and could be a partner. They've got a very state-of-the-art datacenter that has a disaster recovery plan and different elements for continuity. That was one of the factors that helped us evaluate them more seriously than other competitors."

Better business-related data: **"The reason we moved to Amazon is because we can do more."**

Table 2 describes organizational usage associated with use of the AWS data lakes, analytics, and machine learning services. In the organizations surveyed, there were 10,606 TBs of analytics-based data being managed for 61 applications and 184 databases, with 240 TBs dedicated to supporting them. Additional usage patterns are presented in the table. (Note: All numbers cited represent averages.)

Table 2

AWS Data Lakes for Analytics and ML Services Environment

| Organizational usage | Average | Medium |
|---|---|---|
| Number of TBs | 10,606 | 375 |
| Number of countries supported | 16 | 3 |
| Number of sites/branches | 232 | 60 |
| Number of databases | 184 | 38 |
| Number TBs needed to support databases | 1,240 | 10 |
| Number of applications | 61 | 30 |
| Percentage of revenue being supported by applications | 48% | 44% |

Source: IDC, 2020

IDC found that these customers realized **significant value** for their IT, data lakes, analytics, and business operations.

## Business Value and Quantified Benefits

IDC's Business Value model expresses the benefits for organizations in using AWS Data Lakes for Analytics and ML Services to support their ongoing data lake, data warehousing, and analytics infrastructure and operations. Survey data obtained from AWS customers was applied to this model to arrive at an array of quantified post-deployment benefits. Using this methodology, IDC found that these customers realized significant value for their IT, data lakes, analytics, and business operations.

AWS data lakes, analytics, and machine learning services provided these companies with agile, scalable, cost-effective, and high-performance benefits. Customers reported that using AWS reduced the amount of staff time needed to manage and support data warehousing and analytics efforts, which helped enhance productivity. The data analytics services AWS offers also fostered the ability to run more analytics queries and complete each query in less time. In addition, minimizing the effects of unplanned downtime contributed to greater productivity for business units. Study participants described the most significant benefits:

- **Improved analytics agility:** “You can upgrade, you can add more onto the system, and then take it away. With other solutions you could do that by month, maybe by week, but with Amazon you can do it per hour. You can charge by the hour, so it's absolutely significant for us. We can create many more clusters by using the analytic software a lot more. We can handle a lot more queries. It's sending across lots of data that we use to predict how much of what product we're going to sell. It's massive. Statistics is not perfect and can sometimes be skewed, but the data we use certainly empowers us.”
- **Strong product offerings:** “One major benefit is the choice of services we get from AWS Data Lakes for Analytics and ML, products such as Amazon Athena and AWS Glue. More to the point, AWS can provide any application we would ever need. AWS Data Lakes for Analytics and ML is brilliant. Secondly, for a tech person, AWS has a very strong community. There are a lot of online resources (examples: https://aws.amazon.com/developer/community/, Connect with an AWS Hero, AWS Discussion Forums), which is a major benefit.”
- **Faster to market with new models:** “From a business standpoint, the most significant benefit is innovation and time to market. It would take between four and six months to order something with our previous on-premise environment. If I didn't order a specific part properly, or it failed, then I had to wait longer. With AWS Data Lakes for Analytics and ML, I can test my models, or I can calibrate or spin up a server. If I see that my model is not running appropriately for the next two days, for instance, I can shut it down and spin up another, bigger server and then run it again, but more optimized. So this capability gave us the opportunity to optimize models more efficiently without having to wait and spend a lot of money on hardware, or being underspent or overspent.”
- **Ease of use and lower costs:** “With AWS Data Lakes for Analytics and ML, the biggest benefit is ease of use. My researchers love it because it was easy to get going and relatively inexpensive compared to some other data lakes offerings.”

The data analytics services AWS offers fostered the ability to run more analytics queries and complete each query in less time. In addition, minimizing the effects of unplanned downtime contributed to **greater productivity** for business units.

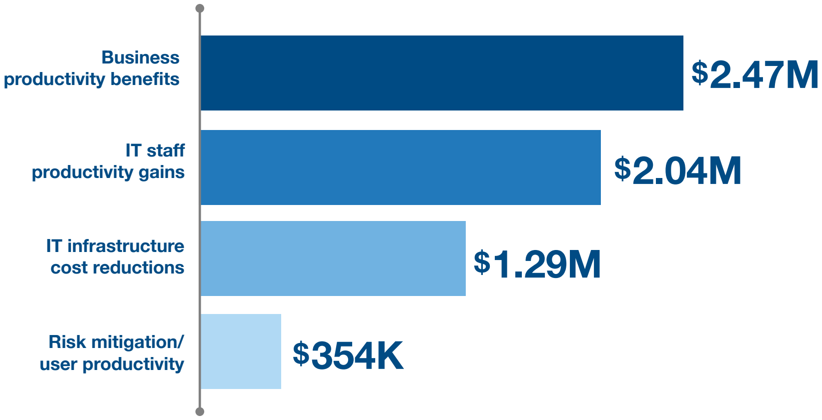
The deployment of AWS Data Lakes for Analytics and ML Services has resulted in significant levels of value by enabling these organizations to better address business opportunities and generate new business. IDC projects that the total value these AWS customers are realizing will be worth an annual average of $16.15 million per organization over five years in the following areas (see Figure 1):

- **Business productivity benefits:** Study participants tied agility, scalability, automation, and other performance benefits to improved business results. IDC calculates the value of higher user productivity at an annual average of $2.46 million per organization ($40,300 per business application).
- **IT staff productivity benefits:** Study participant data showed that use of AWS Data Lakes for Analytics and ML Services required less IT staff time to deploy, manage, and support data warehousing and analytics resources than their legacy infrastructures. IDC projects that interviewed organizations will realize value through IT staff time savings and higher productivity worth an annual average of $2.03 million per organization ($33,200 per business application).
- **IT infrastructure cost reductions:** Interviewed organizations reported that AWS Data Lakes for Analytics and ML Services cost less than comparable on-premises-based solutions, thereby offering lower total cost of operations. IDC calculates that study participants were able to reduce the costs of running data warehousing operations and analytics workloads by an annual average of $1.29 million per organization ($21,100 per business application).
- **Risk mitigation and user productivity benefits:** AWS customers reported that they experienced fewer unplanned outages affecting internal and external end users and access to core applications. IDC calculates the value of higher user productivity at an annual average of $354,000 per organization ($5,800 per business application).

IDC projects that the total value these AWS customers are realizing will be worth an annual average of $16.15 million per organization over five years.

Figure 1

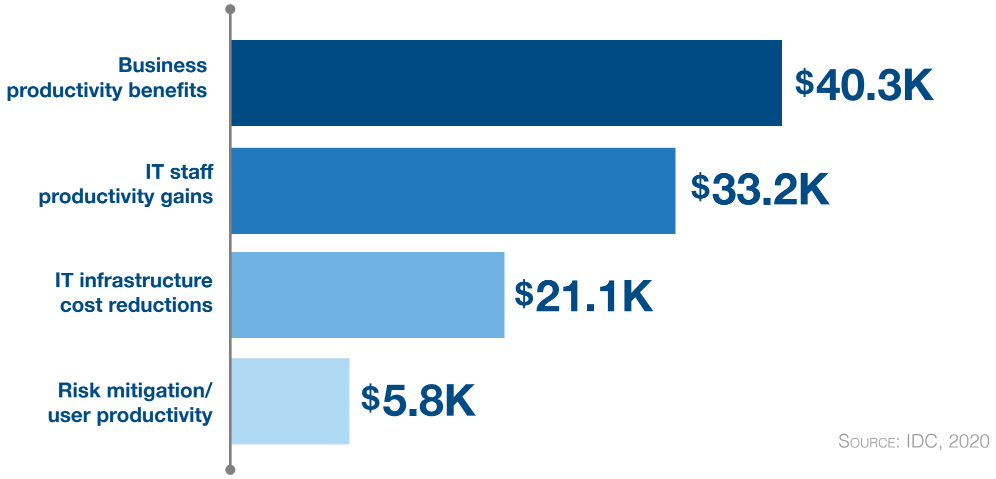
Annual Average Benefits per Organization



Source: IDC, 2020

Figure 2

Annual Average Benefits per Business Application



Source: IDC, 2020

The deployment of AWS Data Lakes for Analytics and ML Services has resulted in significant levels of value by enabling these organizations to better address business opportunities and generate new business.

## Improvements in Data Warehousing and Analytics Performance

With the evolution of the Big Data environments and newer business requirements of agility, flexibility, and cost savings, decoupling of compute and storage is needed. The paradigm originally was to replicate data across multiple nodes in a cluster and permit colocation for performance shortcomings. Today, network performance advances do not necessitate colocation. Also, improvements in compression technology have ensured that more data transferred over the network is much less than the actual data volumes and can significantly reduce the impact of network access times. Decoupled mode allows independent scale of CPU and storage capacity. This enables users to right-size hardware for each layer. Users can buy high CPU and memory configuration for the compute nodes, and the storage nodes can be optimized for capacity. Tightly coupling compute and storage requires users to scale these layers in lockstep, which can lead to resource wastage. With independent scale of compute and storage, businesses don't have to tightly tie storage capacity to memory, processors, or other resources. This enables customers to bring as much processing power as they need for any analytics approach or job, which results in a much more efficient system than legacy servers on-premises. By decoupling compute and storage, multiple compute clusters running Hadoop, Spark, Kafka, MongoDB, Cassandra, or data science tools like TensorFlow can share access to a common data repository/ data lake. This leads to cost savings in storage capacity. AWS Data Lakes for Analytics and ML Services is designed to help companies build and manage data lakes for analytics and Big Data projects with a scalable and agile framework for different data types. The services platform offers a wide array of AWS analytics and machine learning services and automates the process of rapidly and securely building data lakes.

Available services can access data that is stored in a single object store (S3). Amazon S3 offers five storage classes as well as automated data life-cycle management. The cost model is pay-as-you-go and granular: Users pay only for what is needed and how the data is used or processed. In addition, the platform has been designed to be highly secure, with tools that go beyond standard security functionality to proactive monitoring and unified management of security policies.

Study participants emphasized a number of key benefits including improved analytics agility and the ability to easily create more clusters via analytic software, thereby improving the ability to process more queries in less time. They commented on the choice of services available for analytics and machine learning such as Amazon Athena, AWS Glue, and others. Others appreciated the fact that AWS has a strong technical community with robust online resources. Study participants commented on these and other benefits:

AWS Data Lakes for Analytics and ML Services is designed to help companies build and manage data lakes for analytics and Big Data projects with a scalable and agile framework for different data types.

- **Ability to refocus IT staff on more strategic projects:** "Rather than having to get rid of people, we get them to do more strategic projects instead of worrying about maintaining links between X, Y, Z in terms of networks, and managing routers, servers, and cables."
- **Easier infrastructure management:** "AWS Data Lakes for Analytics and ML provided a very good alternative with a lot of scalability. We can spin things up very quickly. For instance, if I want to run models on powerful machines with latest GPUs and a lot of storage and market data, I can spin up in AWS Data Lakes for Analytics and ML as it has a very easy API. I can spin a server for 10 minutes, run my number crunching, then shut it down, then spin up another server. This is much cheaper than what you can do with physical hardware."
- **Automation leading to faster model builds:** "To build a strategic model for training, it would take three to four months to get the hardware purchased. Then you spend another one to two months to install it, put in all the software, and then start building and testing the model. Before it can go live, you're about eight or nine months away from the day you started. Now I can do all that in one to two months. The key bottleneck was the hardware and the integration of the environment and the network. Now I've got a software-defined service network with AWS Data Lakes for Analytics and ML. I can click five buttons very fast, rather than waiting for the cables to get through for the engineer to attend them. All that timeline has gone away."
- **Ability to spin up analytics environments more quickly:** "The biggest benefit is that the time it takes to spin up a new environment is drastically reduced. For the application level, it was reduced by two to three weeks."

This improved functionality cited by participants meant that IT, analytics, and Big Data staff could be freed up from focusing solely on managing their on-premises environments, and it encouraged redirecting them to working on more strategic projects instead of just managing tasks associated with their Big Data/data analytics environments. Table 3 provides granular pre- and post-deployment data on these IT infrastructure management staff impacts. As shown, the AWS capabilities described included the use of more automated processes and easier resource deployment resulting in a 41% improvement in staff productivity. This translated into an annual business value of $948,000.

The use of more automated processes and easier resource deployment **resulting in a 41% improvement in staff**

Table 3

## IT Infrastructure Management Staff Impact

| | **Before**<br>AWS Data Lakes for Analytics and ML | **With**<br>AWS Data Lakes for Analytics and ML | **Difference** | **% Benefit** |
|---|---|---|---|---|
| IT infrastructure management, FTE equivalent per organization per year | **22.9** | **13.4** | **9.5** | **41%** |
| Staff time cost per year | **$2.29M** | **$1.34M** | **$948K** | **41%** |

Source: IDC, 2020

Study participants reported that similar staff improvements were evident with IT infrastructure staff solely dedicated to working on Big Data and analytics projects. As shown in Table 4, there was a 16% improvement in the productivity of these team members, resulting in an annual staff time savings of $97,000.

Table 4

## Big Data Environment Management Staff Impact

| | **Before**<br>AWS Data Lakes for Analytics and ML | **With**<br>AWS Data Lakes for Analytics and ML | **Difference** | **% Benefit** |
|---|---|---|---|---|
| Management of Big Data environment, FTE equivalent per organization per year | **6.2** | **5.3** | **1** | **16%** |
| Staff time cost per year | **$624K** | **$526K** | **$97K** | **16%** |

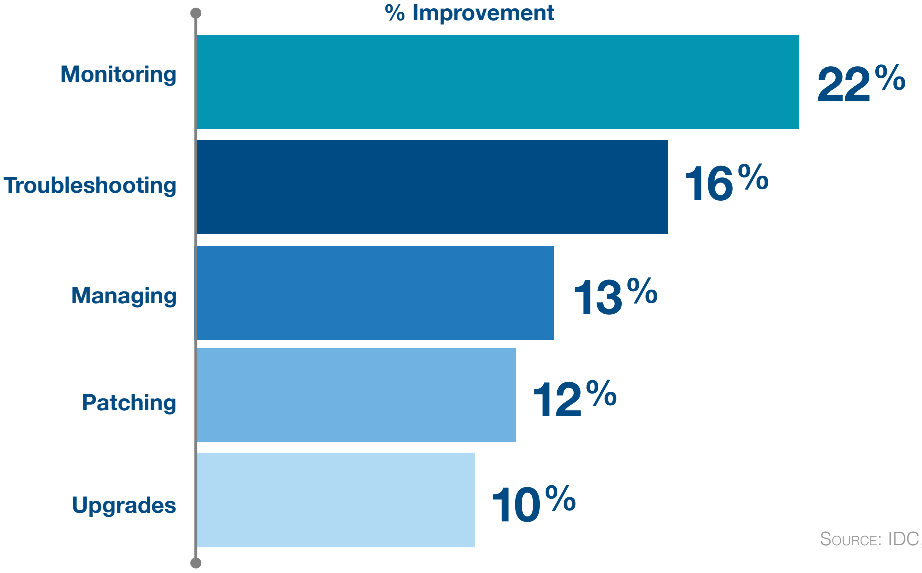
Source: IDC, 2020

More efficient IT infrastructure operations meant that Big Data and analytics applications ran more smoothly and with fewer operational glitches.

Figure 3 drills down a bit more on productivity enhancements for Big Data and analytics IT support teams. These professionals are typically involved in a range of routine tasks on a day-to-day basis such as monitoring and troubleshooting activities that are now largely handled by AWS. IDC looked at the impact of using AWS Data Lakes for Analytics and ML Services on these core operations. As shown in Figure, 3 substantive improvements were realized in all key areas, but especially for monitoring (22%), troubleshooting (16%), and managing (13%). These improvements can be attributed to staff members needing to focus less on managing the data lake or managing the hardware behind the Big Data environment. Additional metrics are presented in the table.

FIGURE 3

Big Data Management Staff Time Efficiencies by Activity



Source: IDC, 2020

IT help desk operations was another key area where improvements were identified by AWS customers. Companies reported that new cloud- and automation-based efficiencies ensured more stable IT and line-of-business operations. This meant that fewer end users, ranging from Big Data staff to other stakeholders consuming analytics-based applications and reports, experienced problems requiring help desk attention.

Because Big Data and analytics teams enjoyed the benefits of more self-sufficient environments based on higher degrees of automation, end users had less dependence on help desks. More efficient IT infrastructure operations meant that Big Data and analytics applications ran more smoothly and with fewer operational glitches. As shown in Table 5, after deployment, the time to resolve trouble tickets was reduced from 10.4 hours to 6.3 hours, representing a 40% improvement. This translated into an FTE improvement of 58% and an annual business value of $783,000.

TABLE 5

## Help Desk Impact

| | **Before** AWS Data Lakes for Analytics and ML | **With** AWS Data Lakes for Analytics and ML | **Difference** | **% Benefit** |
|---|---|---|---|---|
| Time to resolve (hours) | **10.4** | **6.3** | **4.1** | **40%** |
| Total FTE impact | **13.5** | **5.7** | **7.8** | **58%** |
| Total staff time value per year | **$1.35M** | **$570K** | **$783K** | **58%** |

Source: IDC, 2020

IDC looked at how the AWS platform affected the incidence of unplanned downtime in surveyed organizations.

Study participants also reported that more efficient data lake and analytics operations had important implications for business continuity. They reported that they were able to reduce the frequency of unplanned outages affecting business-critical applications and services. Study participants benefited from having a single view of their environments with products such as AWS Prism as well as improved disaster recovery capabilities such as being able to do more frequent backups and use robust tools to more easily manage data protection operations.

IDC looked at how the AWS platform affected the incidence of unplanned downtime in surveyed organizations. As shown in Table 6, after deployment, the annual frequency of outages was reduced from 19.6 to 4.9, representing a substantial improvement (75%). In addition, when incidents did occur, the time needed to resolve them was reduced from 5.6 hours to 0.9 hours, an 84% improvement. The savings from reducing the financial impact of lost productivity was calculated at $1.38 million annually. Additional metrics are presented in the table.

TABLE 6

## Unplanned Downtime Impact

| | Before AWS Data Lakes for Analytics and ML | With AWS Data Lakes for Analytics and ML | Difference | % Benefit |
|---|---|---|---|---|
| Frequency per year | **19.6** | **4.9** | **14.6** | **75%** |
| Time to resolve (hours) | **5.6** | **0.9** | **4.7** | **84%** |
| FTE impact, lost productivity due to unplanned outages | **26** | **6.3** | **19.8** | **76%** |
| Value of lost productivity per year | **$1.82M** | **$440K** | **$1.38M** | **76%** |

Source: IDC, 2020

Table 7 drills down on other revenue impacts resulting from reductions in unplanned downtime, with a slightly different focus: revenue gained back or protected as the result of less unplanned downtime for applications and databases affecting both internal and external users. As shown, the total additional revenue per year associated with this benefit was calculated at $293,000.

TABLE 7

Unplanned Downtime Revenue Impact

| Risk mitigation/unplanned downtime revenue impact | Per organization |
|---|---|
| Total additional revenue per year | $293K |
| Assumed operating margin | 15% |
| Total recognized revenue, IDC model, per year | $44K |

Source: IDC, 2020

Amazon customers reported that a major benefit of the AWS Data Lakes service platform was lowering IT infrastructure costs.

## Lowering the Overall Cost of Data Warehousing and Analytics Operations

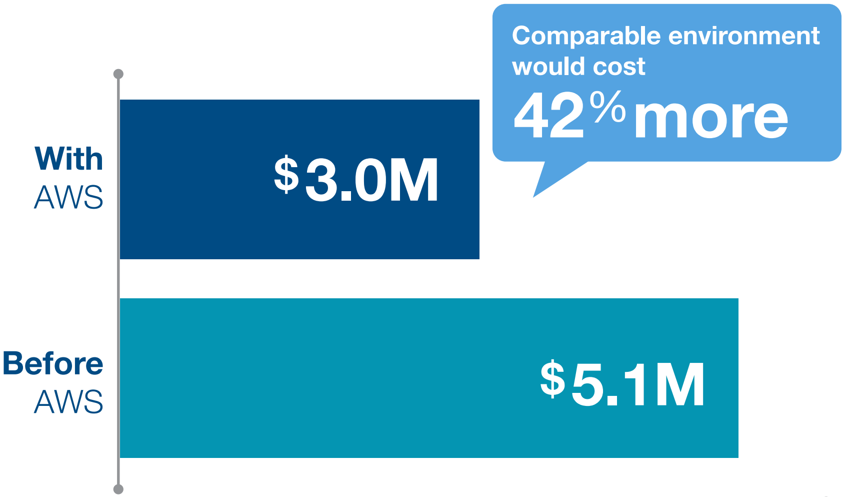
Amazon customers reported that a major benefit of the AWS Data Lakes service platform was lowering IT infrastructure costs. They tied this benefit in part to the platform's pay-as-you-go model and its flexible pricing scheme. There are multiple ways for businesses to pay for Amazon EC2 instances: on-demand, savings plans, reserved instances, and spot instances. With on-demand instances, businesses pay for compute capacity by the hour or by the second, depending on which instances they run. There is no longer-term commitment or upfront payment. Businesses can increase or decrease their compute capacity depending on the demands of their application and pay only the specified hourly rates for the instance they use. Savings plans are a flexible pricing model that offers low prices on EC2 and Fargate usage in exchange for a commitment to a consistent amount of usage (measured in dollars per hour) for a one- or three-year term. Reserved instances provide organizations with a significant discount (up to 75%) compared to on-demand instance pricing; in addition, when reserved instances are assigned to a specific Availability Zone, they provide a capacity reservation, giving businesses additional confidence in their ability to launch instances when needed. Amazon EC2 spot instances allow an organization to request spare Amazon EC2 computing capacity for up to 90% off the on-demand price. Businesses can also pay for dedicated hosts, which provide them with EC2 instance capacity on physical servers dedicated for their use.

When considering the ancillary costs needed to support an on-premises IT infrastructure, such as power and facilities space, the overall savings were substantial. The twin benefits of IT infrastructure and staff efficiencies combined to deliver a much more cost-effective infrastructure for data warehousing and analytics operations for the companies surveyed. As one study participant commented: “I would say the most significant benefit is the footprint reduction of on-premises datacenters, the whole cost of infrastructure. We are estimating about 50% savings over a four-year plan." Figure 4 shows IT infrastructure savings over a five-year period and illustrates how a comparable on-premises environment would cost 42% more.

> "We are estimating about **50% savings** over a four-year plan."

FIGURE 4

IT Infrastructure Savings over 5 Years



Source: IDC, 2020

Figure 5 provides summary data on the post-deployment cost of operations calculated over a five-year period. As shown, the overall cost of operations was calculated to be 46% lower.

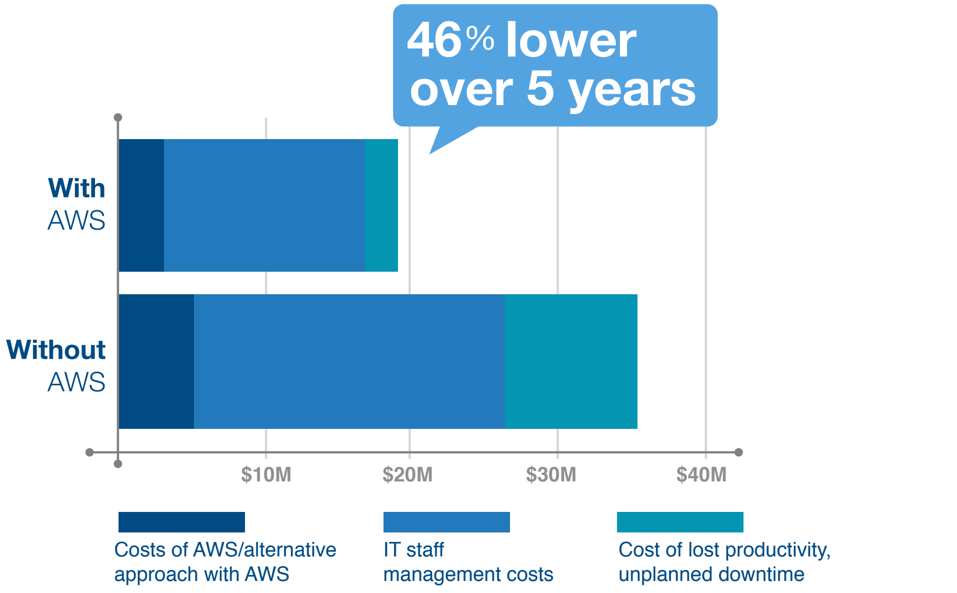
**This calculation was based cumulatively on three factors:**

- **The cost of lost productivity and unplanned downtime**
- **IT staff management costs**
- **The cost of an alternative on-premises solution**

The functionality of the platform and the agility it provided helped these companies achieve higher revenue.

FIGURE 5

Cost of Operations over 5 Years



Source: IDC, 2020

## Delivering Better Business Results

The wide-ranging staff and performance benefits described previously also had positive implications for the business results of the organizations surveyed. As shown in Table 2, on average these organizations were running 61 enterprise business applications on their AWS platforms, and these applications supported a substantial portion of revenues (48%).

Study participants uniformly reported that use of AWS Data Lakes for Analytics and ML Services helped them achieve better business results, especially when used in conjunction with various high-performing analytics programs supported by the platform. The functionality of the platform and the agility it provided helped these companies achieve higher revenue by better addressing business opportunities and enabling better applications and services for internal end users and customers.

Study participants called out specific benefits such as the ability to build better partnerships through improved analytics, the ability to handle both structured and unstructured data in an agile fashion, and the ease of developing their predictive models. They commented on these and other benefits:

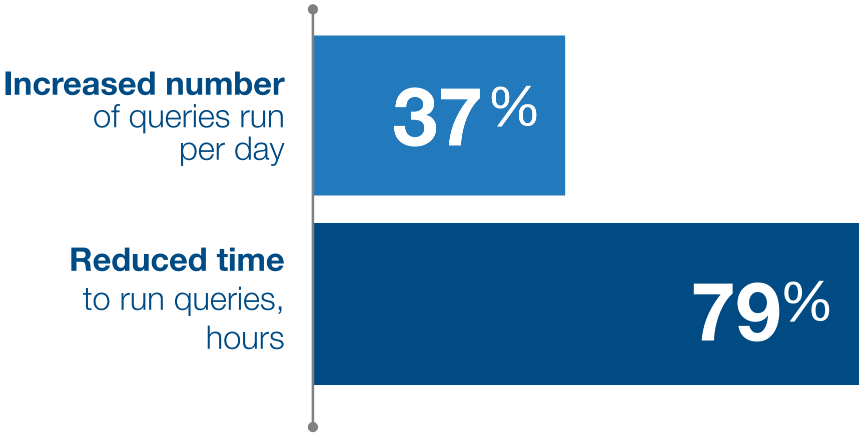
"We are using AWS Data Lakes for Analytics and ML gives us **better real-time and predictive data to make decisions."**

- **Can handle various types of data:** “You can find unstructured data like you would in a document or something else. AWS Data Lakes for Analytics and ML supports both structured and unstructured storage, which gives us more data to make decisions. Pictures are very important, and now we can use them more. It’s very good.”
- **Easier to develop better predictive models:** “We're using AWS Data Lakes for Analytics and ML for our operations and financial reports. We train it to model the data to predict a year in advance what we might need. AWS Data Lakes for Analytics and ML gives us better real-time and predictive data to make decisions.”
- **More focused business operations due to better analytics:** “We can pinpoint when and who is purchasing, and it improves relationships with manufacturers. For example, in the north of England, we’ve discovered our most popular products. Interestingly, in the south of England, completely opposite products are sold. Previously we had a pretty good relationship with our manufacturing partner, but now it’s even better because we have this analytics capability. We work with their marketing department when they are going to market the product. We show them the data for when people came into the store, when they're buying, what they're buying, and in what regions. Thanks to us, they were able to use that data to pinpoint their marketing strategies as well. It really built our relationship so it’s not just a financial benefit.”
- **Faster insights:** “Making fast decisions is the most critical component. By not making the right decision and not having real-time data, we lose opportunity, time, and money.”

As described previously, study participants discussed how the AWS service platform improved the agility of their core analytics operations, thereby providing better efficiencies for analytics teams, business users, and other stakeholders. Because of this agility, requested analytics projects could be conducted in less time and with greater responsiveness. Figure 6 provides granular metrics on these benefits. As shown, study participants were able to run 37% more queries. In addition, they were also able to substantially reduce the time required to run each query (79%).

FIGURE 6

Analytics Agility Impact



Source: IDC, 2020

The AWS service platform improved the agility of their core analytics operations, thereby providing better efficiencies for analytics teams, business users, and other stakeholders.

Surveyed organizations reported that the built-in functionality of the AWS platform enabled analytics staff to work more efficiently and productively, an important contributor to meeting business goals and achieving better results. This included key roles such as data scientists, business intelligence experts, analytics engineers, business analysts, and others. After deployment of the AWS platform, each of these team roles saw improvements in overall productivity (see Table 8). As shown, business intelligence professionals and analytics engineers experienced the greatest improvements, at 19% and 22%, respectively. Overall these teams were able to increase productivity by 17%.

TABLE 8

## Analytics Staff Impact

| | **Before** AWS Data Lakes for Analytics and ML | **With** AWS Data Lakes for Analytics and ML | **Difference** | **% Benefit** |
|---|---|---|---|---|
| Data scientist, FTE impact | 27.8 | 31.3 | 3.5 | 13% |
| Business intelligence, FTE impact | 42.8 | 50.9 | 8.0 | 19% |
| Analytics engineers, FTE impact | 38.7 | 47.2 | 8.5 | 22% |
| Business analysts, FTE impact | 37.8 | 42.6 | 4.9 | 13% |
| Analytics staff time cost per year | $10.3M | $12.0M | $1.74M | 17% |

Source: IDC, 2020

Robust and responsive application development is at the heart of any successful business enterprise.

Robust and responsive application development is at the heart of any successful business enterprise. Study participants reported that AWS Data Lakes for Analytics and ML Services helped application development teams carry out their tasks about 18% more efficiently and gave them the ability to develop more applications while increasing the number of features. As shown in Table 9, the number of analytical-based applications developed annually was increased substantially (65%) while shortening development life cycles by 42%. In addition, the number of new features developed per year increased by 63%. Additional metrics are presented in the table.

TABLE 9

## Application Development Staff Impact

| | **Before** AWS Data Lakes for Analytics and ML | **With** AWS Data Lakes for Analytics and ML | **Difference** | **% Benefit** |
|---|---|---|---|---|
| FTEs per year per organization | **26.3** | **31.0** | **4.7** | **18%** |
| Salary cost per year per organization (based on FTEs) | **$2.63M** | **$3.1M** | **$470K** | **18%** |
| **New applications, new logic** | | | | |
| Number per year | **3.5** | **5.8** | **2.3** | **65%** |
| Development life cycle, weeks | **25.3** | **14.7** | **10.7** | **42%** |
| **New features** | | | | |
| Number per year | **15.7** | **25.6** | **9.9** | **63%** |
| Development life cycle, weeks | **10.4** | **7.4** | **3.0** | **29%** |

Source: IDC, 2020

Considered cumulatively, the benefits described previously — including those for analytics, Big Data, and application development teams — served to make line-of-business users more productive. The key benefits afforded by the AWS platform included better application development, less unplanned downtime, and improved data warehousing and analytics capability. Associated metrics for these end-user impacts are shown in Table 10. As shown, average productivity improved 19%, resulting in an annual business value of $221,000.

The key benefits afforded by the AWS platform included better application development, less unplanned downtime, and improved data warehousing and analytics capability.

TABLE 10

### End User Impact

| Enhanced user productivity | Per Organization |
|---|---|
| Number of users impacted | 16.6 |
| Average productivity gains | 19% |
| Productive hours gained per year | 5,941 |
| End user impact, FTE equivalent per organization per year | 3.2 |
| Value of end user time, per year | $221K |

Source: IDC, 2020

Table 11 presents the average revenue benefits that surveyed organizations received as a result of better addressing business opportunities. On a per-organization basis, the total additional revenue available per year was calculated at $4.14 million.

IDC calculates that these organizations will **achieve a five-year ROI of 415%** and break even on their investment in nine months.

TABLE 11

### Revenue Impacts

| Business impact/revenue from better addressing business opportunities | Per organization | Per application |
|---|---|---|
| Total additional revenue per year | $4.14M | $67.6K |
| Assumed operating margin | 15% | 15% |
| Total recognized revenue, IDC model, per year | $621K | $10.1K |

Source: IDC, 2020

## ROI Summary

IDC's analysis of the financial and investment benefits related to study participants' use of the AWS solution is presented in Table 12. IDC calculates that on a per-organization basis, interviewed organizations will achieve total discounted five-year benefits of $21.48 million ($350,600 per business application) based on IT, data warehousing, and analytics staff efficiencies, increased business results, and lower costs as described. These benefits compare with projected total discounted investment costs over five years of $4.17 million on a per-organization basis ($68,100 per business application). At these levels of benefits and investment costs, IDC calculates that these organizations will achieve a five-year ROI of 415% and break even on their investment in nine months.

TABLE 12

5-Year ROI Analysis

| | Per organization | Per application |
|---|---|---|
| Benefit (discounted) | **$21.48M** | **$350.6K** |
| Investment (discounted) | **$4.17M** | **68.1K** |
| Net present value | **$17.31M** | **$282.6K** |
| ROI (NPV/investment) | **415%** | **415%** |
| Payback (months) | **9 months** | **9 months** |
| Discount factor | **12%** | **12%** |

Source: IDC, 2020

AWS data lakes, analytics, and machine learning ML services are designed to help companies turn data into insights in an easy and secure fashion.

## Challenges/Opportunities

By storing data in a unified repository in open-standards-based data formats, data lakes allow organizations to break down silos, use a variety of analytics services to get the most insights from their data, and cost-effectively grow their storage and data processing needs over time. AWS Lake Formation allows businesses to streamline the data lake creation process and build a secure data lake in days instead of months. Lake Formation helps collect and catalog data from databases and object storage and move the data into their new Amazon S3 data lake, clean and classify the data using machine learning algorithms, and secure access to sensitive data. While this has many benefits, for many organizations it may not be practical to move all of their data to a central repository or to the public cloud. AWS may explore the opportunity to create distributed data lakes architectures that fuse disparate data lakes into a modern data lake environment that accelerates analytics performance, allowing data scientists more time for doing their best work.

## Conclusion

AWS data lakes, analytics, and machine learning (ML) services are designed to help companies turn data into insights in an easy and secure fashion. Businesses can use the right tool for the job without needing to move or transform data for different analytics approaches. This is a significant advantage to businesses from an overall agility, flexibility, and cost-savings perspective. The decoupling of compute and storage, with S3 being the foundation for data lakes, enables independent scaling of resources. This leads to reduced wastage of resources, significant cost savings, and simpler and more economical maintenance.

These benefits were reported through the interviews IDC conducted for this study. Users told IDC about eased management needs for their analytics-supporting infrastructure, reduced costs, and a wide range of analytics platforms that improved their analytics operations. These benefits manifested themselves into an average of 42% reduction in infrastructure costs, 17% more efficient analytics teams, and a five-to-one return on these organizations' investment into AWS data lakes, analytics, and machine learning.

These benefits manifested themselves into an average of **42% reduction in infrastructure costs**, **17% more efficient analytics teams**, and a five-to-one return on these organizations' investment.

## Appendix: Methodology

IDC's standard ROI methodology was utilized for this project. This methodology is based on gathering data from current users of the AWS Data Lakes solution as the foundation for the model. Based on interviews with organizations using it, IDC performed a three-step process to calculate the ROI and payback period:

- Gathered quantitative benefit information during the interviews using a before-and-after assessment of the impact of AWS Data Lakes. In this study, the benefits included staff time savings and productivity benefits, and operational cost reductions.

- Created a complete investment (five-year total cost analysis) profile based on the interviews. Investments go beyond the initial and annual costs of using the AWS service platform and can include additional costs related to migrations, planning, consulting, and staff or user training.

- Calculated the ROI and payback period. IDC conducted a depreciated cash-flow analysis of the benefits and investments for the organizations' use of AWS over a five-year period. ROI is the ratio of the net present value (NPV) and the discounted investment. The payback period is the point at which cumulative benefits equal the initial investment. IDC bases the payback period and ROI calculations on a number of assumptions, which are summarized as follows:

  - Time values are multiplied by burdened salary (salary + 28% for benefits and overhead) to quantify efficiency and manager productivity savings. For purposes of this analysis, based on the geographic locations of the interviewed organizations, IDC has used assumptions of an average fully loaded $100,000-per-year salary for IT staff members, and an average fully loaded salary of $70,000 for non-IT staff members. IDC assumes that employees work 1,880 hours per year (47 weeks x 40 hours).

  - The net present value of the five-year savings is calculated by subtracting the amount that would have been realized by investing the original sum in an instrument yielding a 12% return to allow for the missed opportunity cost. This accounts for both the assumed cost of money and the assumed rate of return.

  - Further, because IT solutions require a deployment period, the full benefits of the solution are not available during deployment. To capture this reality, IDC prorates the benefits on a monthly basis and then subtracts the deployment time from the first-year savings.

Note: All numbers in this document may not be exact due to rounding.

# About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

## IDC Global Headquarters

IDC Research, Inc.
5 Speen Street
Framingham, MA 01701
USA
508.872.8200

idc.com

@idc

## Copyright Notice