The Economic Benefits of Migrating Apache Spark and Hadoop to Amazon EMR

Sponsored by:
Amazon Web Services

Authors:
Carl W. Olofson
Harsh Singh

November 2018

Business Value Highlights

57%
reduced cost of ownership

342%
five-year ROI

8 months
to breakeven

33%
more efficient Big Data Teams

46%
more efficient Big Data/Hadoop environment management staff

99%
reduction in unplanned downtime

$2.9 million
million additional new revenue gained per year

EXECUTIVE SUMMARY

As more and more enterprises deploy data lakes using some or all of the Apache constellation of open source projects that include Hadoop and Spark, and apply them to different purposes, issues of efficiency, scale, and management have come into play. Some enterprises are turning to a managed service to address these issues. One such service is Amazon Elastic MapReduce (EMR). Amazon Web Services (AWS) asked IDC to research the benefits inherent in using Amazon EMR, and to that end, IDC has conducted this business value study.

IDC interviewed organizations that are utilizing Amazon EMR to support their Big Data/Hadoop/Spark environments. Study participants told IDC that the flexibility of Amazon EMR improved business agility and kept costs down. According to IDC calculations, these organizations will realize a 57% savings on their total cost of ownership for these environments by:

- **Reducing physical infrastructure costs** by deploying a flexible, elastic, and scalable cloud environment to deploy their Big Data environments
- **Driving higher IT staff productivity** among teams that need to manage and support these environments
- **Providing stronger Big Data environment availability** which enables better productivity among end users, such as Big Data teams that utilize and consume data

SITUATION OVERVIEW

Data lake technology burst on the scene around 10 years ago with Hadoop, which offered a large-scale data collection environment with massive parallel processing at a low cost through the networking together of PCs in a cluster, using internal storage and coordination protocols to process the data using MapReduce. Suddenly, work that could only be done using high-end systems and expensive storage arrays could be done for a fraction of the cost. Initially, the main job of a data lake was to organize large amounts of collected data and perform processing and analytics on that data. As its role expanded, and as more efficient analytic technologies, such as Apache Spark, became available, problems began to emerge. Enterprises began setting up cluster after cluster. Management of the data over time became an issue. Systems were bought and deployed that were rarely used.

More recently, data lake developers have been looking at object storage, and especially native object storage in the cloud, as an alternative to Hadoop clusters. Deployment in the cloud offers advantages, but only if one takes advantage of the capabilities that the cloud environment offers. These include decoupling compute from storage resources. Of course, such an approach means moving away from the "lift and shift" approach, which can lock down resources and becomes a very expensive way to go. A better approach is a managed service for data lake management that is optimized for the cloud. This enables developers to vary the processor power in relation to the data volume. Working in the cloud also enables an on-demand model, where resources are paid for only when they are used. As the need for data lakes in a variety of scenarios increases, the appeal of a cloud-based lake has grown as well, but what about the complexity of managing it?

The answer may be in subscribing to a managed data lake service in the cloud — one that intimately ties its operations to the acquisition and release of resources is especially appealing from a cost management perspective. Amazon EMR is one such service.

AMAZON EMR

Amazon EMR is a fully managed data lake service based on Apache Hadoop and Spark, integrated with the cloud environment of Amazon Web Services (AWS), including its storage service layer called S3. It is designed to eliminate the complexity involved in the manual provisioning and setup of data lake resources, including the Hadoop and Spark clusters, the tuning of the environment, and all the other operational details that tend to trip users up.

Amazon EMR also includes services in support of insight delivery, analytics, and data lake management. With AWS data movement services, it is easy to integrate the data lake with other AWS assets such as Redshift, Athena, Glue, Kinesis, and SageMaker. The service also includes facilities to ensure that the data is secure, compliant to regulations, and auditable. AWS also offers ways to set up and manage machine learning (ML) operations on data in EMR. These include SageMaker, Jupyter notebooks, and Spark ML, and often with ML frameworks like TensorFlow and MXNet.

THE BUSINESS VALUE OF AMAZON EMR

Study Demographics

IDC interviewed nine organizations for this study by asking a variety of quantitative and qualitative questions about the impact of using Amazon EMR on their IT operations, Big Data and analytics operations, core businesses, and overall cost profiles. Table 1 characterizes the firmographics of these organizations.

On average, these organizations had over 59,000 employees and $32 billion in annual revenues. These organizations were broad in size as these firms had employee ranges of 3,500 to 160,000 employees with revenues between $4.5 million to $145 billion. They represented a diverse mix of vertical industries including telecommunications, healthcare, financial services, energy, and food and beverage sectors. This diverse group of organizations were using Amazon EMR in a wide variety of use cases to support their IT and business operations. The average number of IT users within the companies surveyed was 49,070, and those users supported 48.97 million external customers using 11,935 business applications.

TABLE 1

Demographics of Interviewed Organizations

	Average	Median	Range
Number of employees	59,444	49,000	3,500 to 16,000
Number of IT staff	7,716	1,300	146 to 40,000
Number of IT users	49,070	31,500	3,360 to 160,000
Number of external customers	48.97M	600K	1K to 200M
Number of business applications	11,935	150	42 to 100,000
Revenue per year	$32.0B	$10.1B	$4.5M to $145B
Industries	Discrete manufacturing (3), process manufacturing (2)		

n = 9
Source: IDC, 2018

Organizational Use of Amazon EMR

To get a full picture of typical use, IDC gathered information on how these organizations were using Amazon EMR in their day-to-day IT and business operations. Table 2 depicts this usage based on several key attributes. IDC found that AWS EMR environments supported an average of 1,853 databases and 25 business applications which required nearly 3.5 PBs of memory.

TABLE 2

Organization Usage of Amazon EMR

	Average	Median	Range
Number of TBs	3,789	500	2 to 30,000
Number of countries supported	5	1	1 to 31
Number of sites/branches	27	8	3 to 125
Number of databases	1,853	10	2 to 15,000
Number of TBs needed to support databases	3,426	300	2 to 28,000
Number of applications	25	8	2 to 85
Percentage of revenue being supported by applications	11%	8%	0% to 30%

n = 9
Source: IDC, 2018

These AWS customers reported that a key benefit of Amazon EMR was the flexibility provided in compute and memory usage and in the ways that services could be purchased. They reported that Amazon EMR pricing is simple and predictable. Pricing requires customers to pay a per-second rate for every second used, with a one-minute minimum. For example, a 10-node cluster running for 10 hours would cost the same as a 100-node cluster running for 1 hour. In addition, the hourly rate depends on the instance type used such as high CPU, high memory, low CPU, low memory, or other types of instances.

Study participants reported procuring Amazon EMR services through all three of AWS' core pricing models: On-Demand, Reserved Instance, and Spot Instances. Participants reported greatest use of On-Demand (55%, paid by the hour or second without longer-term commitment) and Spot Instances (30%, use of spare AWS EC2 capacity). Use of these two pricing models likely reflects use of Amazon EMR for spikier and time-sensitive dependent Big Data analytics workloads. Respondents reported procuring an average of 15% of their Amazon EMR capacity with Reserved Instances which had lower pricing than On-Demand but with capacity reservation to meet the most common baseline load, while also cost efficiently meeting peaks of demand.

TOTAL COST-OF-OPERATIONS COMPARISON OF AMAZON EMR

Interviewed organizations told IDC that they realized significantly lower total cost of operations by running their Big Data/Hadoop/Spark environments on Amazon EMR. IDC evaluated the total cost of operations of Amazon EMR by comparing three factors: 1) the costs of running their Big Data/Hadoop environments on Amazon EMR against a comparable on-premise infrastructure, 2) IT staff-related costs and 3) costs associated with unplanned downtime. Note that in our study, planned maintenance costs are included in IT staff-related costs.

Study participants told IDC they appreciated the flexibility to set up the environments they need with a payment structure that allowed them to pay for additional memory and processing power as needed. This payment structure helped reduce infrastructure costs and freed up IT teams to work on more business-focused projects. Additionally, participants mentioned they were getting stronger resiliency with Amazon EMR, which helped reduce the costs of unplanned downtime:

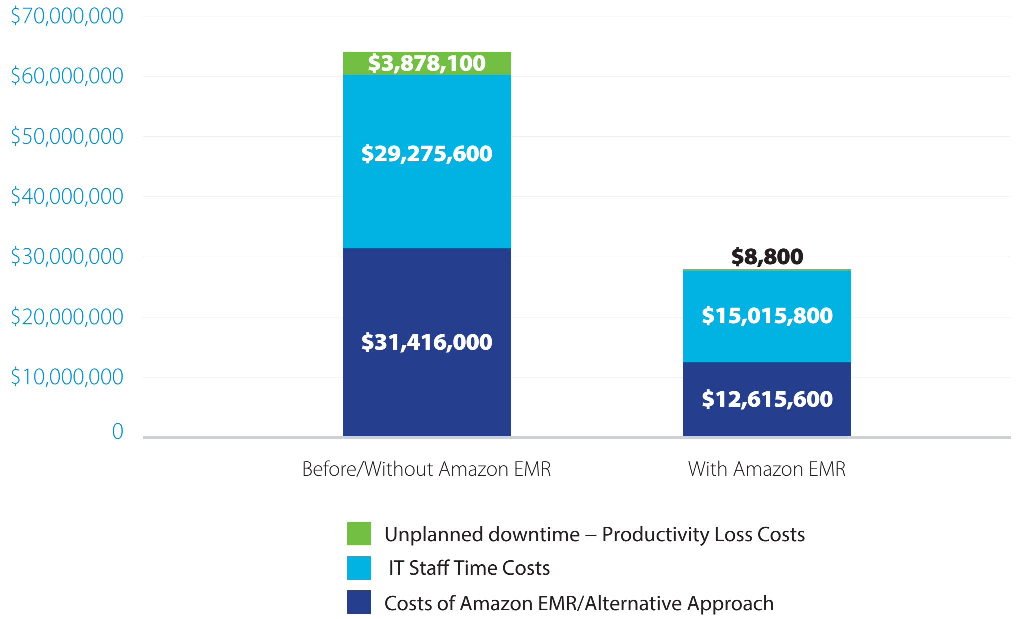
- **Agility to support different environments:** *"One of the most cost-effective features is the ability to change the technology. For example, today I have an application where I need to use Apache Spark. I don't need to go to the burden of setting up all the Apache Spark activities in my cluster. If I want to have a new machine running on Flink, I don't have the burden of setting up Flink. With cloud, to spin something up, it just takes a few clicks, and everything is ready to go. And if I don't want it, I can shut it down as well. So the effort of managing resources and setting up the infrastructure activities is almost down by 70%."*
- **Lower cost of operation:** *"Amazon EMR gave us the best bang for the buck. One of the key factors is that our data is obviously growing. Running our Big Data operations on [Amazon] EMR increases confidence. It's really good since we get cheap storage for huge amounts of data. The second thing is that the computation that we need fluctuates highly. Some of the data in our database is only occasionally used by our business or data analysts. We choose EMR because it is the most cost-effective solution as well as providing need-based computational expansion."*
- **Efficient scaling:** *"The biggest benefit of Amazon EMR is the scalability. We don't have to pay for the scalability unless we need it. We can quickly start instances and have things ready pretty quickly. We have what you would call a grouping. So we can have an optimal grouping where we can spin up multiple groupings. This means we can clone things fast."*

As Figure 1 notes, customers that spoke to IDC were seeing cost savings across the aforementioned three costs areas. Over five years, these customers were able to reduce their infrastructure costs by 60%, while reducing IT support time for Big Data environments by half (49%). After including a 99% reduction in the cost of unplanned downtime, IDC calculates that these organizations will run Amazon EMR at a 57% lower cost over five years.

FIGURE 1

Five-Year Cost of Operations*

57% lower total cost of ownership over 5 years



**see appendix for full breakdown of all costs*

Source: IDC, 2018

More Efficient IT Staff

IDC estimates that the increased IT staff efficiencies made possible by the use of Amazon EMR at these organizations represented a gain of 49% in freed up IT staff time related to infrastructure, Big Data/ Hadoop management, and help desk teams (see IT Staff Time Costs from Figure 1, as well as the individual components of IT staff time reported in Tables 3, 4, and 5). These customers found that IT management was much easier and more efficient because of Amazon EMR's cloud-based functionality. In many cases, this meant that IT staff was freed up from having to focus solely on managing their on-premise environments on a day to day basis. This encouraged the redirection of staff resources to more strategic projects in support of business goals instead of management and provisioning tasks associated with their Hadoop or Spark environments. Amazon EMR customers provided the following illustrations of these benefits:

- **Time freed up to focus on critical projects:** *"The scaling we get with Amazon EMR helps. For example, there are times where there might be a sudden 2-3x surge in activity. When that used to happen with a fixed footprint on-premises we would invariably be caught short from time to time, and would have to put our main projects on the backburner. Now, with Amazon EMR, we are able to maintain those projects with less disruption and more timeliness."*
- **Automation leading to better quality:** *"When moving to the cloud, we had to automate everything. This meant that quality is going to be better because there are less issues. We get less data issues such as data errors now. We don't have a person doing these tasks because we have scripts, so we are going to see less errors."*

- **Quicker development:** *"We have the ability to deliver solutions more quickly such as proof-of-concept applications. It's unbelievable how quickly we can go live and right to production. We are able to do that much more quickly in the cloud with Amazon EMR."*

On average, these IT infrastructure teams experienced a 62% productivity increase (see Table 3).

TABLE 3

IT Infrastructure Staff Impact

	Before	With		
	Amazon EMR	**Amazon EMR**	**Difference**	**Benefit (%)**
IT infrastructure management, FTE impact	29.9	11.3	18.5	62%
Staff time cost per year	$2,986,700	$1,133,300	$1,853,300	62%

Source: IDC, 2018

- These AWS customers reported that Amazon EMR made it easier to set up their Big Data/Hadoop environments required by line of business teams. In part, setup was easier because the need for hardware-and software-related system integration was by and large eliminated. As one study participant noted: *"We went with Amazon EMR's ready-made integration site. It is all about not having to spend time on integration. . .If we choose another Hadoop technology, then our researchers would have to make that work but if we run into a road block and it doesn't work, we might learn that the hard way. In a way, we would be doing more testing which would have meant we needed to hire three more people to do the integration work if we weren't on Amazon EMR."*

These organizations experienced improved productivity for their Big Data/Hadoop management staff resulting from Amazon EWR (see Table 4). As shown, these Big Data/Hadoop environment teams were able to free up 54% of their time as a result of Amazon EMR.

TABLE 4

Big Data/Hadoop Environment Management

	Before	With		
	Amazon EMR	**Amazon EMR**	**Difference**	**Benefit (%)**
Management of Big Data/ Hadoop environment, FTE impact	18.2	8.4	9.7	54%
Staff time cost per year	$1,815,700	$841,000	$974,700	54%

Source: IDC, 2018

IT help desk operation was another key area of benefit identified by Amazon EMR customers. Surveyed companies reported that new cloud- and automation-based efficiencies ensured more stable IT and line of business operations translating to fewer end users (which could range from these organizations' Big Data staff to the stakeholders who consume analytics-based applications and reports) and problems that required help desk attention. Because Big Data/Hadoop teams enjoy the benefit of self-sufficient environments coupled with more automated processes, there's much less need to call the help desk for problem resolution (see Table 5).

Especially noteworthy is a 50% reduction in calls and trouble tickets recorded per week. This reduction, in turn, means that annual help desk staff productivity improved by 77%.

TABLE 5

Help Desk Impact

	Before Amazon EMR	With Amazon EMR	Difference	Benefit (%)
Calls/tickets per week	173.4	87.5	85.9	50%
Time to resolve (hours)	13.6	6.2	7.4	55%
Help desk staff, FTE impact	10.5	2.4	8.2	77%
Staff time cost per year	$1,052,700	$237,700	$815,000	77%

Source: IDC, 2018

Better Risk Mitigation

The effective management of risk is a major consideration in today's complex business environments. Organizations reported how Amazon EMR improved their overall risk profiles by offering high levels of availability and reducing the incidence of system outages and unplanned downtime. Study participants spoke in specific detail about these benefits:

- **Availability of data and cost optimization:** *"What we want to do is go into the cloud as quickly as possible. We do not want to be deleting data in the interest of reducing costs and want to retain some level of data. At the same time, we want to be quick in providing insights to our customers. We went with Amazon EMR mostly because of availability and scalability."*
- **Better resiliency:** *"We have made systems much more resilient. It is really all about performance and resiliency."*

Table 6 summarizes the impact Amazon EMR had on unplanned downtime. By using Amazon EMR, these organizations experienced a drop in the number of downtime incidents by 86% while the time to resolve incidents when they did occur (measured in hours) reduced by 94%. This in turn resulted in the annual cost of unplanned downtime to improve by 99%.

TABLE 6

Unplanned Downtime Productivity Impact

	Before Amazon EMR	With Amazon EMR	Difference	Benefit (%)
Frequency per year	27.4	3.8	23.6	86%
Time to resolve (hours)	9.3	0.54	8.8	94%
FTE impact, lost productivity due to unplanned outages	11.1	0.03	11.1	99%
Cost of unplanned downtime per year	$775,600	$1,800	$773,900	100%

Source: IDC, 2018

As mentioned, Amazon EMR improved risk profiles by offering high levels of availability and reducing unplanned downtime, which also meant that organizations were reducing revenue lost due to these factors. The average gross revenue regained by these organizations was $5,094,400 that can be tied to improvements in availability and reduced downtime. IDC's financial model assumes a 15% operating margin for recognizing gross revenue gains, which translates to $764,200 in higher operating income due to these benefits.

Lower Infrastructure Costs

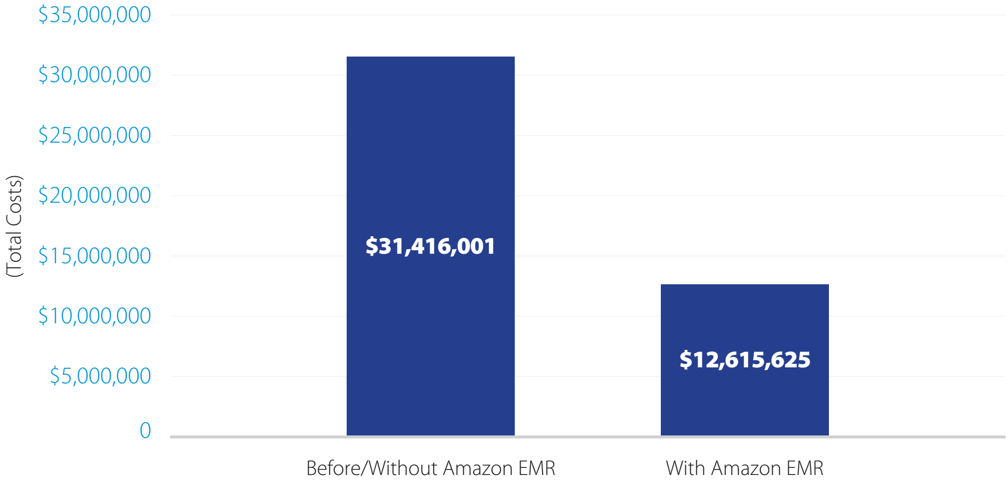
AWS customers reported that a major benefit of Amazon EMR usage was lower IT infrastructure costs. This benefit relates to Amazon EMR's pay-as-you-go capability and the flexible pricing models described previously. Organizations estimated that they would need to pay 41% more for deploying the infrastructure associated with an on-premise environment to support the same workloads. When considering other costs needed to support an on-premise IT infrastructure, such as costs of power and facilities space, the overall savings is 60% (see Figure 2). Study participants offered a series of specific observations related to this benefit:

- **Most cost-effective and lower storage costs:** *"Amazon EMR gave us the best bang for the buck… Our data is growing and we get cheap storage for huge amounts. In addition, the computation we need fluctuates a lot. Some of the data in our database is only occasionally used by our business or data analysts. We choose [Amazon] EMR because it is the most cost-effective solution as well as providing need-based computational expansion."*
- **Lower TCO:** *"We lowered our total cost of ownership by more than 30%. So it's cost and also our ability to transform systems for the business."*
- **Cost optimization:** *"We went with Amazon EMR mostly because of availability and scalability. And we wanted to pay optimally."*

FIGURE 2

Five-year IT Infrastructure Costs

Amazon EMR Costs 60% Less Over Five Years



Source: IDC, 2018

ROI ANALYSIS OF AMAZON EMR

In addition to the cost savings these organizations experienced by using Amazon EMR, customers also achieved a significant return on investment (ROI). These Big Data and application development teams benefited from having a stable and scalable environment. As illustrated in Figure 3, IDC calculates that these organizations are realizing average benefits of $18.13 million per year per organization (or $735,000 per business application) due to the following key benefits:

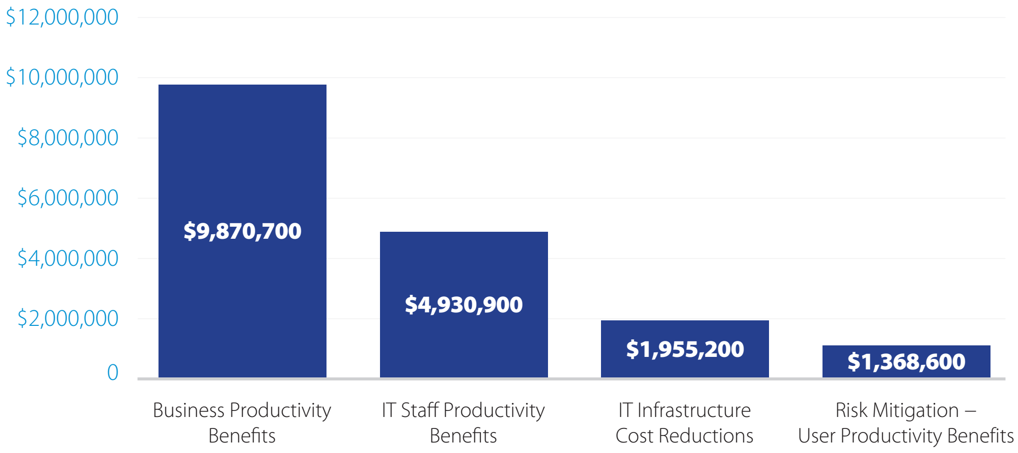
- **Improved business results** as Big Data teams have more flexibility in their ability to access and utilize their data. At the same time, having better data allows organizations to access new revenue streams. IDC estimates these organizations are realizing $9.87 million per organization per year (or $400,000 per business application).
- **Increased IT staff productivity gains** of $4.93 million per year per organization ($200,000 per business application). As stated in the earlier total cost-of-operations analysis, IT staff focused less on setting up and managing their organization's Big Data environments and focused more on strategic business-oriented projects.
- **Reduced IT infrastructure costs** is netting these organizations an additional $1.96 million per organization (or $79,000 per business application). As mentioned in the total cost-of-operations section, organizations utilizing Amazon EMR's flexible pricing model optimized their IT infrastructure costs.
- **Better risk mitigation** because stronger availability has an impact on their Big Data operations. IDC calculates the annual value of lost user productivity gain back at an average of $1.37 million per organization (or $55,000 per business application)

FIGURE 3

Average Annual Benefits per Organization

Total Annual Benefits per Organization:

$18.1 Million



Source: IDC, 2018

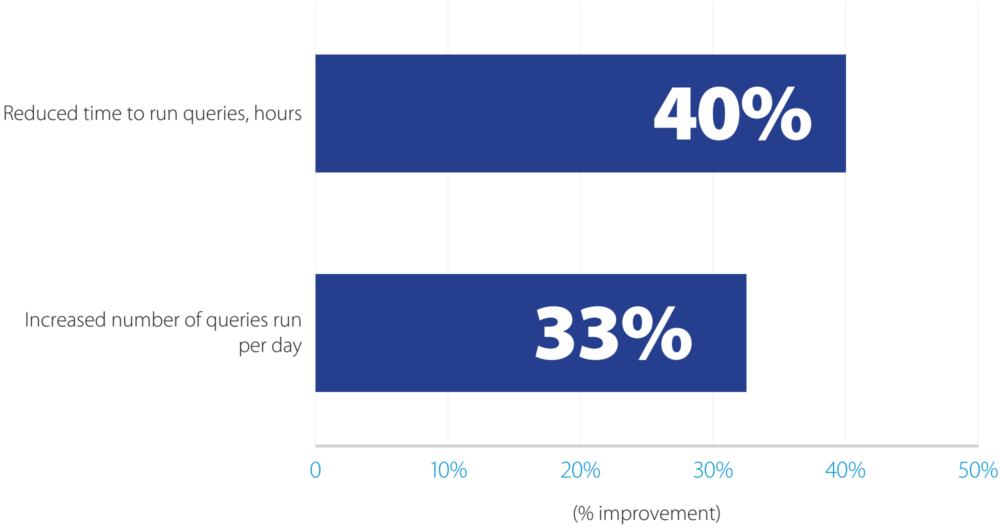
Improved Business Benefits

With the reported benefits of stronger availability and performance, these Big Data teams using Amazon EMR were able to work with more confidence and efficiently, which served to strengthen business operations and meet business goals. This benefit related to the ability of Amazon EMR environments to scale up as needed, coupled with the core benefits of using a cloud-based solution.

These organizations also reported that using Amazon EMR allowed them to do more with their Big Data environments. Because they enjoyed more stable and scalable environments with Amazon, Big Data teams were able to run more queries more efficiently. Figure 4 illustrates the average impact of Amazon EMR for running queries. There was a 40% reduction in the amount of time needed to run queries, coupled with a 33% improvement in the overall number that could be processed.

FIGURE 4

Analytics Performance Impact



Source: IDC, 2018

AWS customers also reported that their Big Data and analytics teams were more productive as the result of resiliency, faster computational capability, and self-service access to the data sets being processed. As a result of the time freed up due to more efficient data processing, these teams can spend more time focused on more business-oriented projects. Study participants offered a series of specific observations about this benefit:

- **Better response to customer needs:** *"[Amazon] EMR agility gives engineers the capability to execute on product demand. We were dealing with a product with our primary customer that needed 100 TB of data last year. But the customer was constantly needing to support new datasets and we needed to spin up a lot of performance around that. Amazon EMR would do that elastically. I don't think we would have met all of the timelines and needs of the customer otherwise. We are now in production with them and have a couple more contracts for the same product because of the success story on the first customer. "*

- **More productive data teams:** *"Our Big Data and analytics teams are more productive for several reasons. First, capabilities are faster and more powerful. For example, they can get 1,000 times the amount of data to do what-if scenarios as opposed to on-premises. All the data is online so all of the users have self-service access to the data they need which wasn't the case before. A lot of it was in puddles and they would have to ask and wait for data. Also our uptime is about five times more resilient in the cloud.“*

Table 7 dives deeper into these metrics. Noteworthy is the fact that the work of data scientists showed a 46% level of productivity improvement while analytics engineers saw a 39% improvement. For these teams responsible for supporting Big Data operations, organizations experienced a productivity gain of one-third on average across the teams.

TABLE 7

Big Data Staff Productivity Impact

	Before Amazon EMR	With Amazon EMR	Difference	Benefit (%)
Data scientist, FTE impact	47.1	68.8	21.8	46%
Business intelligence, FTE impact	56.4	72.1	15.7	28%
Analytics engineers, FTE impact	57.2	79.2	22	39%
Business analysts, FTE impact	72.2	89.5	17.3	24%
Big Data staff time cost per year	$23.3M	$31.0M	$7.7M	33%

Source: IDC, 2018

Application development teams need on-demand access to compute and storage resources to test and deploy business applications and features. If teams have to wait for these resources, they cannot support the business in a robust fashion.

- These organizations reported that AppDev teams were able to operate with more confidence because of a better performing and more flexible environment. As was the case with Big Data teams, organizations found that they could access the data they needed more easily. As one study participant commented: *"We have the ability to deliver proof-of-concept applications more quickly. It's unbelievable how quickly we can go live and right to production. We are able to do that much more quickly in the cloud with Amazon EMR."*

Table 8 summarizes improvement metrics related to the impact on application development teams. Especially noteworthy is that the number of new applications available annually increase by two-fold after deploying Amazon EMR. In addition, application development lifecycles (measured in weeks) was reduced by 43%, while the lifecycles for the development of new features was reduced by 49%.

TABLE 8

Application Development Teams Impact

	Before Amazon EMR	With Amazon EMR	Difference	Benefit (%)
Productivity levels of development teams, equivalent FTEs	106.1	111.7	5.6	5%
Equivalent staff value in productivity per year	$10,608,300	$11,166,700	$558,300	5%
New Applications, New Logic				
Number per year	7.2	14.3	7.2	100%
Development lifecycle, weeks	28.3	16.2	12.1	43%
New Features				
Number per year	100.3	114.3	14.0	14%
Development lifecycle, weeks	6.3	3.2	3.1	49%

Source: IDC, 2018

The array of productivity gains described previously affecting IT staff efficiencies also translated to similar benefits for end users in these organizations. End users realizing these productivity gains were line of business and analytics teams that were, in effect, the consumers of Big Data and analytics projects. Commenting on these benefits, one study participant said: *"We can offer business insights faster than we could in our traditional environment. The speed of getting the results to teams is much faster."*

Study participants report the average productivity of these teams improved by 29%, which netted them an additional $3 million worth of freed up time (see Table 9).

TABLE 9

End User Impact

	Per Organization
Number of users impacted	150
Average productivity gains	29%
End user impact, FTE equivalent per organization per year	43.7
Value of end user time	$3.06M

Source: IDC, 2018

» Cumulatively, there were also impacts on business operations revenue. More efficient data operations meant the ability to develop new revenue streams and more opportunities to grow the business. One study participant described this benefit as follows: *"We have a better way to link the non-relational data that we receive to our products and business partners which will open up more revenue opportunities. For example, we have a lot of IoT products. We would receive the data from those products but we didn't use it. Now we can. For example, we receive data from different sales centers, so we can potentially get the location of the customer, and then based on the location we can show them targeted advertising."*

As a result, Table 10 shows that these organizations were able to increase their annual revenues by an average of just under $3 million per organization.

TABLE 10

Business Operations, Revenue Impact

	Per Organization
Total additional revenue per year	$2,916,700
Total recognized revenue, IDC model, per year	$437,500

**The IDC model assumes a 15% operating margin for all additional revenue.*
Source: IDC, 2018

ROI Analysis

Table 11 displays a summary of the benefits and costs for Amazon EMR deployments at the organizations we interviewed. These customers realized discounted benefits worth an average of $63.28 million per organization (or $2.57 million per business application) over five years while making a discounted investment of $14.33 million (or $581,000 per business application). Therefore, these organizations will see a five-year ROI of 342% with a breakeven on their investment in eight months.

TABLE 11

ROI Analysis

	Per Business Application	Per Organization
Benefit (discounted)	$63.28 million	$2.57 million
Investment (discounted)	$14.33 million	$581,000
Net present value (NPV)	$48.95 million	$1.98 million
Return on investment (ROI)	342%	342%
Payback period (months)	8	8
Discount rate	12%	12%

Source: IDC, 2018

CHALLENGES/OPPORTUNITIES

Amazon EMR is, of course, challenged by other cloud-based managed data lake services competing for the same customers. Such challenges, though, stimulate innovation and creativity. Another challenge is the constantly evolving technology that supports the data lake, the growing range of use cases, and the various forms and types of applications being developed for this kind of data. AWS must keep current with all these applications in order to continue to deliver the kinds of benefits described in this study. But there are also opportunities. The number and types of data sources keep expanding, and with them, new and important ways to use data to drive better decisions, automate systems, and enable smarter operations. By providing leadership in these areas, Amazon has the opportunity to enable customers to excel now and in the future.

CONCLUSION

Data lake technologies, including the constellation of data collection, processing, and analytic systems, offer great benefits to enterprises seeking to maximize the value of the data at their disposal, discover new insights, perform more nimbly, and respond more effectively to opportunities in the market. Managing these technologies in the datacenter represents a substantial challenge, however, and costs can get out of control quickly.

A better option is to deploy these technologies in the cloud within a managed service optimized to take maximum advantage of the cloud environment. By realizing such shifts in operations such as the decoupling of compute and storage, and by using resources on a pay-as-you-go model rather than the fixed cost of datacenter deployment, users can realize better value as well as ensuring continuous availability and reliable performance.

In examining the Amazon EMR customer cases discussed in this paper, IDC has found that users were able to dramatically increase the number of useful applications, increase capacity, realize better performance, and reduce staff time required for routine operations, all while realizing considerable cost savings.

Anyone currently operating, or planning to deploy in the datacenter, a data lake environment based on Apache Hadoop and Spark and the constellation of software assets that complement them, should consider the following suggestions:

- » Develop a plan regarding initial and future deployment of your system.
- » Estimate the current cost of managing your data lake environment, including equipment and staff time costs, and the cost of downtime when it happens either for maintenance purposes or due to system failure.
- » Consider how much better you can utilize your systems if they have a dynamic, rapid spin-up, and pay-as-you-go model rather than running on fixed hardware that must be acquired and installed.
- » Look at the available range of managed Apache Hadoop/Spark cloud services out there.
- » Take a close look at Amazon EMR as a possible platform for your future data lake operations.

APPENDIX

IDC's standard ROI methodology was utilized for this project. This methodology is based on gathering data from organizations currently using Amazon EMR as the foundation for the model. Based on interviews with these study participants, IDC has calculated the benefits and costs to these organizations of using Amazon EMR. IDC used the following three-step method for conducting the ROI analysis:

1. **Gathered quantitative benefit information during the interviews using a before-and-after assessment of the impact of Amazon EMR.** In this study, the benefits included staff time savings and productivity benefits, increased revenue and operational cost reductions.
2. **Created a complete investment (five-year total cost analysis) profile based on the interviews.** Investments go beyond the initial and annual costs of using Amazon EMR and can include additional costs related to migrations, planning, consulting, and staff or user training.
3. **Calculated the ROI and payback period.** IDC conducted a depreciated cash flow analysis of the benefits and investments for the organizations' use of Amazon EMR 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.

Table 12 presents a more detailed look at the five-year TCO analysis for organizations utilizing Amazon EMR. IDC includes the costs of lost productivity in its TCO analysis. However, this table will show a breakdown of the TCO both with and without the cost of lost productivity.

TABLE 12

Five-year Cost of Operations

	Before Amazon EMR	With Amazon EMR	Difference	Benefit (%)
IT infrastructure costs	$31,416,000	$12,615,600	$18,800,400	60%
IT staff time cost, infrastructure	$14,933,300	$5,666,700	$9,266,700	62%
IT staff time cost, Big Data environment management	$9,078,600	$4,205,200	$4,873,400	54%
IT staff time cost, help desk support	$5,263,700	$1,188,500	$4,075,200	77%
IT staff time cost, migration	$ -	$3,955,500	$(3,955,500)	
User productivity cost, unplanned downtime	$3,878,100	$8,800	$3,869,300	100%
Total IT staff time costs	$29,275,600	$15,015,800	$14,259,800	49%
Total cost of operations, with no lost productivity	$60,691,600	$23,676,000	$37,015,600	61%
Total cost of operations, with lost productivity	$64,569,700	$27,640,300	$36,929,400	57%

Source: IDC, 2018

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

IDC Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-insights-community.com
www.idc.com

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2018 IDC. Reproduction without written permission is completely forbidden.

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
ANALYZE THE FUTURE