A Forrester Total Economic Impact™ Study

Commissioned By

IBM

Project Director:
Henry Huang

September 2016

The Total Economic Impact™ Of IBM Apache Spark

Business Benefits And Advancements Made Possible By Apache Spark

FORRESTER®

Table Of Contents

Executive Summary 3

Disclosures 6

TEI Framework And Methodology 7

Analysis 8

Financial Summary 20

IBM Apache Spark: Overview 21

Appendix A: Total Economic Impact™ Overview 22

Appendix B: Forrester And The Age Of The Customer 23

Appendix C: Glossary 24

Appendix D: Supplemental Material 25

Appendix E: Endnotes 25

ABOUT FORRESTER CONSULTING

Forrester Consulting provides independent and objective research-based consulting to help leaders succeed in their organizations. Ranging in scope from a short strategy session to custom projects, Forrester's Consulting services connect you directly with research analysts who apply expert insight to your specific business challenges. For more information, visit forrester.com/consulting.

© 2016, Forrester Research, Inc. All rights reserved. Unauthorized reproduction is strictly prohibited. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change. Forrester®, Technographics®, Forrester Wave, RoleView, TechRadar, and Total Economic Impact are trademarks of Forrester Research, Inc. All other trademarks are the property of their respective companies. For additional information, go to www.forrester.com.

FORRESTER®

Executive Summary

IBM commissioned Forrester Consulting to conduct a Total Economic Impact™ (TEI) study and examine the potential return on investment (ROI) enterprises may realize by deploying IBM's Apache Spark-as-a-service offering. The purpose of this study is to provide readers with a framework to evaluate the potential financial impact of this cluster computing platform on their organizations through more rapid and expansive insight development.

IBM Apache Spark opens the door to new analytics and, in turn, business possibilities with on-demand cluster computing that is considerably more powerful than existing distributed computing frameworks.

The SETI Institute, a nonprofit organization whose mission is to find extraterrestrial life in the universe, has an annual operating budget of approximately $17 million. As a customer of Apache Spark, it is forecasted to experience over $7.4 million in benefits over a three-year horizon.

To better understand the benefits, costs, and risks associated with an Apache Spark-as-a-service implementation, Forrester interviewed an existing customer, the Search for Extraterrestrial Intelligence (SETI) Institute, with approximately a year of experience using Spark on the IBM Bluemix cloud. The SETI Institute is a nonprofit research organization that is dedicated to exploring, understanding, and explaining the origin and nature of life in the universe. Among its tools is the Allen Telescope Array — a 42-dish radio telescope that generates terabytes (TB) of data daily, from the capture of thousands of radio spectra — that is made possible through private donations and funding. The challenge is to computationally crunch through massive amounts of data and lift out spectrums of interest. IBM collaborated with the SETI Institute in 2015 and introduced the SETI Institute to its Apache Spark.

Using in-memory processing across on-demand nodes, Spark is a new generation of cluster computing. It can work as a complementary piece of a big data analytics engine to other distributed computing frameworks such as Hadoop — and is often used in conjunction with such platforms, as is the case with the SETI Institute. Working in shared memory, the performed analyses in Spark are quicker by orders of magnitude compared with its MapReduce counterparts. For those organizations seeking to replace their existing big data platform, the transition to Spark can be accelerated with IBM's Apache Spark-as-a-service, with cluster management and storage infrastructure made readily available through IBM Bluemix cloud services.

To date, there have been significant amounts of tests run to show the comparative speed advantage of Spark; the speed advantage is proven. In 2014, Databricks was able to sort 100 TB of data in one-third of the time, using 90% fewer nodes than a record-setting Hadoop job sorting through the same amount of data in 2013.[1] This case study instead addresses the realm of possibilities and the quantified value enabled by IBM's Spark-as-a-service at the SETI Institute.

IBM APACHE SPARK READILY EXPANDS THE BOUNDARIES OF ANALYTICS

With an array of 42 radio telescope dishes capable of capturing frequencies from 1 gigahertz (GHz) to 10 GHz, the Allen Telescope Array (ATA) used by the SETI Institute can output 25 gigabytes (GB) of data per second. Due to limitations with its existing on-premises distributed computing, very little of that data was analyzed. This computational and storage limitation effectively translated to a band of only 100 megahertz (MHz) (or 0.1 GHz) of a possible 9 GHz being captured for near-real-time analysis. While the SETI Institute had been effective at identifying possible bodies of interest in preliminary scans with its existing cluster, it could not effectively analyze the monumental amount of data from these bodies of interest without a more powerful solution, as computational limits had already been reached. With the current implementation of Apache Spark, the SETI Institute has been able to introduce multiple new ways to look at the constantly growing archive of data. In one instance, using the lowest number of computational executors offered by Apache Spark-as-a-service, the SETI Institute was able to process 200 million records in 9 hours, finding six bodies of interest that had not and would not have been discovered otherwise.

Bill Diamond, CEO of the SETI Institute, stated: "*Spark's capabilities give us the opportunity to look at complex and massive data sets with a powerful new set of tools. We are now able to look for and extract structure and patterns in the data that*

were previously invisible to us. ***This dramatically enhances our SETI efforts***, *and gives us vastly more sensitive "ears" with which to listen to the cosmos."*

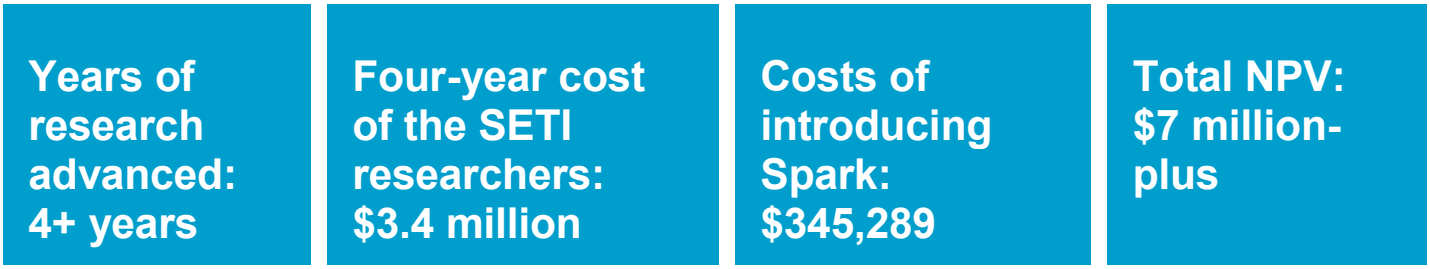
With scientists concluding that life at some point has existed elsewhere in our universe, the question is not of whether we can find the origins and presence of extraterrestrial life, but of when we — with SETI and the analytical powers enabled by Spark — can do so.[2]

Our interviews with SETI and the subsequent financial analysis found that the SETI Institute experienced the risk-adjusted benefits, and costs shown in Figure 1.[3]

The analysis points to initial first-year adoption costs of $21,450, versus first-year benefits of over $2 million, adding up to a net present value (NPV) of $7,399,474 after three years of usage.

With IBM Apache Spark-as-a-service, the SETI Institute was able to derive insights almost immediately with extensive usage, exploring more than triple the amount of data in a matter of months. And not only was it able to analyze more data, but it was able to perform the data analysis much more comprehensively, in ways previously not possible. Data scientists experienced an increase of 150% in insight development in the first year of usage, with higher rates of discovery expected to be realized from additional use cases within the Spark framework in the subsequent years.

FIGURE 1
Financial Summary Showing Three-Year Risk-Adjusted Results



Source: Forrester Research, Inc.

› **Benefits.** The SETI Institute experienced the following risk-adjusted three-year benefits:

- **The costs of additional on-premises computing nodes to process beyond first-pass Fourier transform filtering are avoided at the SETI Institute. By adding Spark to the overall analytics picture, the SETI Institute is able to perform analysis on new data that it otherwise would have discarded due to existing computational limitations.** The SETI Institute estimates that to be able to look at data in the different ways that Spark provides today, it would have required at least a doubling, if not tripling, of its current cluster, but without the Spark efficiency. The total cost of additional infrastructure and nodes avoided amounted to a present value (PV) of $193,839 over three years.
- **The cost of additional infrastructure maintenance is also avoided without the increased provisioning of computing nodes.** Given the historical trends for the SETI Institute's hardware upgrade schedule, we estimate that by Year 2, it will have implemented an estimated 50 additional nodes. By Year 3, we expect the existing nodes of the cluster to be replaced along with associated infrastructure, adding to the internal costs of implementation. The total benefits resulting from the avoided costs of maintenance and implementation were $258,753, PV.
- **The SETI Institute will gain over four years of research advancement from the addition of Spark to its current computational set.** Given prior computational limitations, these advancements would not have been possible and would have resulted in a waste of scientific brain power by its scientists. Using historical wage estimates for its scientists dedicated specifically to SETI exploration along with the upkeep costs of the ATA, we've provided a conservative estimate for the time value of advancements at the SETI Institute. The enablement of new methods to look at data and make use of more data has resulted in a PV benefit of $6,946,882.
- **Increased discoveries improve public and private funding to SETI Institute.** Existing Forrester research indicates that breakthrough discoveries from other public space exploration programs have a positive correlation to grant funding. Our previous analysis removed factors such as political effects, recessionary/expansionary periods, and other external influences and still produced a strong correlation. We believe that SETI Institute discoveries will also likely lead to incremental increases in funding. Due to the program's short history (compared with other major space programs) and narrowly defined mission scope, we have not quantified this benefit at this time. Nevertheless, we believe it to be important in the continued quest to answer the question of life in the universe.
- **Open sourcing of the SETI Apache Spark code invites citizen scientists to slice SETI data in additional manners.** With an entire open source community of curious scientists, the SETI Institute hopes that it is able to attract new perspectives and algorithms to look at captured data. As an added societal benefit, SETI algorithms provide citizens with methods and angles to capture the power of Spark in business and life-enhancing applications.

› **Costs.** The SETI Institute experienced the following risk-adjusted three-year costs:

- **Spark-as-a-service computation and storage costs.** These fees are based upon the usage of both cloud object storage as well as the executors used during actual computation. The total three-year costs amounted to $11,442, PV.
- **Cost of coding efforts for processing tasks on Spark, both accrued in the initial deployment period as well as in later years for new developments of algorithms.** These costs largely comprise the cost of data scientist and programmer man-hours. The SETI Institute cited that with the assistance of IBM and the Spark open source community, it has transitioned code to Python much faster than otherwise possible. SETI is developing algorithms that can not only help find technologically advanced life beyond our solar system, but also peel away at the Spark onion of capabilities that can meaningfully change real-life applications here on Earth. Total coding efforts costs on the Spark platform over three-years equate to $333,847, PV.

Disclosures

The reader should be aware of the following:

- The study is commissioned by IBM and delivered by Forrester Consulting. It is not meant to be used as a competitive analysis.
- Forrester makes no assumptions as to the potential ROI that other organizations will receive. Forrester strongly advises that readers use their own estimates within the framework provided in the report to determine the appropriateness of an investment in IBM Apache Spark.
- IBM reviewed and provided feedback to Forrester, but Forrester maintains editorial control over the study and its findings and does not accept changes to the study that contradict Forrester's findings or obscure the meaning of the study.
- IBM provided the customer names for the interviews but did not participate in the interviews.

TEI Framework And Methodology

INTRODUCTION

From the information provided in the interviews, Forrester has constructed a Total Economic Impact (TEI) framework for those organizations considering implementing IBM Apache Spark. The objective of the framework is to identify the cost, benefit, flexibility, and risk factors that affect the investment decision, to help organizations understand how to take advantage of specific benefits, reduce costs, and improve the overall business goals of winning, serving, and retaining customers.

APPROACH AND METHODOLOGY

Forrester took a multistep approach to evaluate the impact that IBM Apache Spark can have on an organization (see Figure 2). Specifically, we:

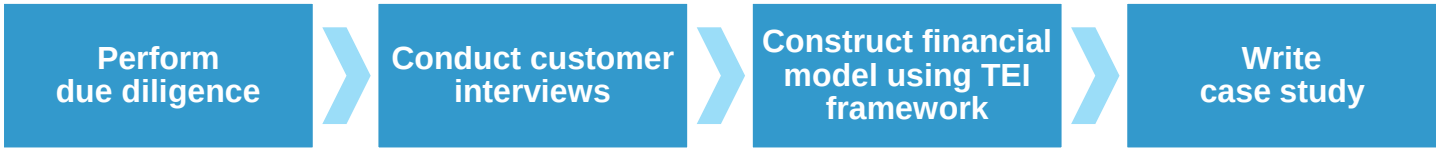
- Interviewed IBM marketing, sales, and/or consulting personnel, along with Forrester analysts, to gather data relative to Apache Spark and the marketplace for Apache Spark.
- Interviewed one organization currently using IBM Apache Spark to obtain data with respect to costs, benefits, and risks.
- Constructed a financial model representative of the interviews using the TEI methodology. The financial model is populated with the cost and benefit data obtained from the interviews.
- Risk-adjusted the financial model based on issues and concerns the interviewed organization highlighted in interviews. Risk adjustment is a key part of the TEI methodology. While the interviewed organization provided cost and benefit estimates, some categories included a broad range of responses or had a number of outside forces that might have affected the results. For that reason, some cost and benefit totals have been risk-adjusted and are detailed in each relevant section.

Forrester employed four fundamental elements of TEI in modeling IBM Apache Spark's service: benefits, costs, flexibility, and risks.

Given the increasing sophistication that enterprises have regarding ROI analyses related to IT investments, Forrester's TEI methodology serves to provide a complete picture of the total economic impact of purchase decisions. Please see Appendix A for additional information on the TEI methodology.

FIGURE 2

TEI Approach



Source: Forrester Research, Inc.

Analysis

INTERVIEWED ORGANIZATION AND INTERVIEW HIGHLIGHTS

For this study, Forrester conducted in-depth interviews with representatives from the SETI Institute, which is an IBM customer based in the US. The SETI Institute is a nonprofit research institution with the mission to "explore, understand, and explain the origin and nature of life in the universe."

From the customer, we gathered data points on its one year of usage of Spark-as-a-service to model past and future benefits and costs. Some high-level characteristics of the SETI Institute are as follows:

> *"Because of the limitations of our old computing system, we were only actually analyzing three beams out of 1,000. The data from the rest of the beams just fall on the floor, unmeasured."*
>
> ~ Gerry Harp, senior scientist at SETI Institute

- The SETI Institute has an annual operating budget of approximately $17 million.
- The SETI Institute is funded by private foundations, individuals, and public research grants from NASA and NSF.
- The SETI Institute employs over 75 scientists to advance its research, with a smaller portion dedicated to the research directly related to the ATA.
- It currently uses the Allen Telescope Array, a collection of 42 radio telescopes 290 miles northeast of San Francisco, as its primary method of collecting interstellar radio frequencies for research.
- It has a 50-node cluster onsite at the Allen Telescope Array to process data in near real time.
- Its existing distributed computing cluster runs on custom programs/algorithms coded in C.

The Allen Telescope Array that the SETI Institute uses is unique in that it is an upgradeable collection of smaller radio dishes linked together to collect data from the skies, rather than singular large dished telescopes that the SETI Institute used in the past. Some facts on the ATA include:

> *"Spark has been a supercharger for SETI. It is truly a game changer for how we conduct SETI research at the Institute."*
>
> ~ Bill Diamond, CEO at SETI Institute

- The 42 radio telescopes finished construction in 2007, with funding support from the Paul G. Allen Family Foundation.
- The array of radio telescope dishes are software controlled and can gather a total of over 1,000 beams of radio signals from the skies, scanning over 1 million stars for non-naturally occurring signals that indicate extraterrestrial life.
- The ATA is capable of collecting data 24x7, although it is currently only observing 12 hours per day.
- Total data transmission from the ATA can be as much as 200 gigabits (Gb), or 25 gigabytes (GB), per second.

Based on the interviews, Forrester constructed a TEI framework and an associated ROI analysis that illustrates the areas financially affected.

Situation

Prior to working with IBM on the Apache Spark project, the SETI Institute relied largely on its 50-node local computing cluster located with the ATA to process signals. The Hat Creek location of the observatory was ideal for observing but presented its own set of challenges due to the remoteness of the location. Radio frequencies were observed through the ATA as a function of ongoing time. Once a set of signals had passed, it was gone. This circumstance presented the SETI scientists with an interesting problem: What do you do with so much data that needs to be studied in near real time? Two possible approaches existed:

- Land all of the data and crunch through it later. This wasn't ideal, as the power and infrastructure requirements in the remote area weren't viable.
- Crunch through as much of the data in near real time as possible, within the limits of its computing cluster. Much of the data from the high-bandwidth satellite array, however, would be wasted.

Over the past 15 years, the SETI Institute constantly refreshed and increased computing capabilities half a dozen times to address the plethora of incoming data. Nevertheless, the computational power bottleneck still existed. Using its computing cluster to process as much as it could in real time also meant that the SETI Institute would have to sacrifice the capability to analyze data after the initial data scan. To truly unlock the possibilities within the data from the radio telescope array, SETI desired the following in its new big data analytics:

- Have near- or actual real-time analytics capability.
- Drastically increase computational power so that more data can be scanned on a first pass.
- Leverage machine learning to improve analytical efficiency.
- Control capital expenditure.

Solution

In partnering with IBM, the SETI Institute found a way to increase research capabilities without committing to significant capital resources. The open source community and IBM's commitment to advancing the technology provided a framework that jumpstarted the Spark program at the SETI Institute, with 2014 and 2015 both marking important years for the Spark open source community. The number of developers and lines of code roughly doubled every year, making Spark the most contributed-to big data project today. Committing to a resource-rich ecosystem was important for SETI. Together with the help of IBM and the Spark community, the SETI Institute has crunched through 200 million-plus records of previously collected data, resulting in six new bodies of interest.

> *"There's no reason why extraterrestrials have to limit themselves to a specific type of signal. With Spark, we are able to embrace that reality by developing new algorithms — giving us the ability to observe a wide array of signal types."*
>
> ~ Gerry Harp, senior scientist at SETI Institute

Using the Jupyter Notebook service on Spark, SETI Institute scientists have crunched through archival data at astounding rates and are inventing new ways, with the help of citizen scientists, to look at data; nearly instant computations on Jupyter have enabled interactive exploration of the data. Current developments at the SETI Institute indicate that it will utilize Spark to process real-time calculations and also capture larger data sets for locations of interest, such as the exoplanets that the NASA's Kepler has uncovered in recent years. Our interviews with the SETI Institute revealed that it has additional plans to leverage Spark. Ultimately, the possibility frontier for analytics is rapidly increasing, bringing the discovery of ET ever closer.

BENEFITS

The interviewed organization experienced a number of quantified benefits in this case study:

- Avoided on-premises hardware node/cluster purchases.
- Avoided cost of infrastructure maintenance.
- Time-to-insight gains from accelerated research advancements.

Beyond these quantified benefit categories, we believe the SETI Institute will benefit in yet another manner. Our research indicates that space exploration programs have traditionally experienced a clear positive correlation between discoveries/major breakthroughs and grant funding. Due to the different nature of SETI's goals as compared with more publicly comprehensible wins that other space programs have showcased, such as sending satellites to Jupiter or landing mankind on the moon, we cannot accurately assert the same level of correlation for potential SETI Institute funding. The accelerated pace of discoveries, however, should bring a level of positive attention and excitement, which we hope will translate into funding to keep the search for extraterrestrial life alive.

While we have not yet quantified this due to its relatively recent introduction, the SETI Institute cited the benefit of its code being open source and available to the community. Premier scientists in astronomy have joined to provide alternative algorithms to analyze data on the Spark platform. The entire community interested in finding extraterrestrial life has benefited, but this gives the SETI Institute additional perspectives on where extraterrestrial life might lie in the universe and advances its initiative without further investment.

Avoided On-Premises Hardware And Node/Cluster Purchases

Following the introduction of Apache Spark, the SETI Institute was able to capture and utilize data that otherwise would have been scrapped due to a lack of computational power to perform deepened analyses. Using existing patterns of infrastructure update cycles of two to three years, the organization was expecting to augment its distributed computing capability by adding new nodes and replacing related hardware architecture in another year. While the total computational power would increase, allowing for added methods to examine the radio data, the hardware still would have lagged significantly behind the capabilities and speed of the software-scalable Spark-as-a-service that IBM provides. In effect, the addition of new nodes to the on-premises cluster would enable either of the following:

- Added near-real-time first-pass computational ability — the ability to scan a larger range of radio waves.

Or

- Added explorative techniques and algorithms to look at data more comprehensively.

By utilizing the existing on-premises cluster and increasing the usage of Spark in the cloud for first-pass analysis, the SETI Institute would be able to accomplish both of the aforementioned capabilities rather than just a single enablement with the addition of on-premises nodes using traditional processing techniques. The cost of the nodes and associated infrastructure, such as network area storage (NAS) devices and security appliances, will save the SETI Institute $180,000 by the end of the second year and another $60,000 in the third year, for a total three-year PV benefit of $193,839.

TABLE 1
Avoided On-Premises Hardware And Node/Cluster Purchases

Ref.	Metric	Calculation	Initial	Year 1	Year 2	Year 3
A1	Cost of new nodes for cluster				$60,000	$60,000
A2	Infrastructure-level cluster upgrades (e.g., NAS, security, virtualization hardware, etc.)				$120,000	
At	Avoided on-premises hardware node/cluster purchases	A1+A2	$0	$0	$180,000	$60,000
	Risk adjustment	0%				
Atr	**Avoided on-premises hardware node/cluster purchases (risk-adjusted)**		**$0**	**$0**	**$180,000**	**$60,000**

Source: Forrester Research, Inc.

Avoided Cost Of Infrastructure Maintenance

In avoiding the commitment to significant on-premises hardware additions by increasing the usage of Spark, the SETI Institute can also avoid the cost of a network operations center (NOC) engineer to monitor and maintain the larger cluster and associated infrastructure. While the SETI Institute is currently able to defer much of the cluster maintenance to its symbiotic partner, the Stanford Research Institute (SRI), it may not be able to guarantee continued maintenance of a larger cluster. The expected cost of a senior NOC engineer, along with possible professional services for the migration and integration of new on-premises hardware, amounted to $164,000 yearly, starting in the second year. The total three-year costs in PV terms were $258,753.

TABLE 2
Avoided Cost Of Infrastructure Maintenance

Ref.	Metric	Calculation	Initial	Year 1	Year 2	Year 3
B1	Cost of senior NOC engineer, annually, fully loaded				$144,000	$144,000
B2	Professional services				$20,000	$20,000
Bt	Avoided cost of infrastructure maintenance	B1+B2	$0	$0	$164,000	$164,000
	Risk adjustment	0%				
Btr	**Avoided cost of infrastructure maintenance (risk-adjusted)**		**$0**	**$0**	**$164,000**	**$164,000**

Source: Forrester Research, Inc.

Time-To-Insight Gains From Accelerated Research Advancements

A research institute's primary measure of success is its number of discoveries and advancements created. The process of trying to find extraterrestrial life is complicated, requiring the checking and rechecking of data. Finding data chunks of interest are small wins and a necessary step to advancing research. For the SETI Institute, the ability to find extraterrestrial life is predicated on its ability to:

- Capture as many radio frequencies from the ATA as computing resources allow.
- "Thin slice" data to identify those frequencies that look promising for greater inspection, which is done on a very limited basis (scanning three beams out of over 1,000 possible) with its existing computing power. Thin slicing must be done on a near-real-time basis, as time cannot be rewound to re-examine passing signals.
- Process the data of interest gathered from the limited scope through multiple algorithms to eliminate noise.

With Apache Spark, the SETI Institute can now do the following:

- Perform near-live processing to thin slice more data and look at more beams for narrow band signals.
- Land more data that can then be shipped to the IBM cloud for new and more comprehensive algorithmic passes over the data. This data would have otherwise been scrapped due to a lack of computational ability.
- Take advantage of machine learning to sort through noise and identify true bodies of interest, iteratively improving algorithms.

The raw amount of data now processed with Spark is a significant increase over the existing system. The existing system has bottlenecked SETI's research operations, and Spark's data analysis has accelerated the research by years. An expected increase of Spark usage in years 2 and 3 will further the time-to-insight proposition. Using the compensation of SETI Institute scientists as a time-to-insight benefit measurement, along with the cost of the Allen Telescope Array upkeep, the SETI Institute gained an expected PV of $6,946,882 over three years, after risk adjustment. See the section on Risks for more information.

Bill Diamond, CEO of SETI Institute, stated: "Spark is a supercharger for SETI."

TABLE 3
Time-To-Insight Gains From Accelerated Research Advancements

Ref.	Metric	Calculation	Initial	Year 1	Year 2	Year 3
C1	Research advancement made possible by Spark, measured in years			1.5	1.5	1.5
C2	Annual ATA upkeep cost value			$950,000	$950,000	$950,000
C3	Cost of yearly research operations, minus cost of on-premises computing infrastructure			$900,000	$900,000	$900,000
C4	Spark usage growth				10%	50%
Ct	Time-to-insight gains from accelerated research advancements	C1*(C2+C3)*(C4+1)	$0	$2,775,000	$3,052,500	$4,162,500
	Risk adjustment	↓15%				
Ctr	**Time-to-insight gains from accelerated research advancements (risk-adjusted)**		**$0**	**$2,358,750**	**$2,594,625**	**$3,538,125**

Source: Forrester Research, Inc.

Total Benefits

Table 4 shows the total of all benefits across the three quantified areas listed above, as well as present values (PVs) discounted at 10%. Over three years, the SETI Institute expects risk-adjusted total benefits to be a PV of more than $33 million from the addition of the Apache Spark solution.

TABLE 4
Total Benefits (Risk-Adjusted)

Ref.	Benefit Category	Initial	Year 1	Year 2	Year 3	Total	Present Value
Atr	Avoided on-premises hardware node/cluster purchases	$0	$0	$180,000	$60,000	$240,000	$193,839
Btr	Avoided cost of infrastructure maintenance	$0	$0	$164,000	$164,000	$328,000	$258,753
Ctr	Time-to-insight - gains from accelerated research advancements	$0	$2,358,750	$2,594,625	$3,538,125	$8,491,500	$6,946,882
	Total benefits (risk-adjusted)	**$0**	**$2,358,750**	**$2,938,625**	**$3,762,125**	**$9,059,500**	**$7,399,474**

Source: Forrester Research, Inc.

COSTS

The interviewed organization experienced a number of costs associated with the Apache Spark solution:

› IBM Spark-as-a-service costs.

› Cost of coding efforts in Python for use on Spark.

These represent the mix of internal and external costs experienced by the interviewed organization for initial transition to the Spark platform, ongoing development efforts, and usage costs associated with the solution.

IBM Spark-As-A-Service Costs

Prior to the SETI Institute's use of IBM Apache Spark, it did most if not all of its computations onsite in real time. The current state of computations makes use of Spark for huge amounts of archived data to fine-comb and analyze signals in never-before-done ways in collaboration with IBM. In addition, the SETI Institute team has started to send fresh, near-real-time data to IBM for analysis. With the data being sent to the Bluemix cloud and then processed through Spark, there are two cost components: cloud storage and compute power. Assumed first-year costs were $3,885, growing in subsequent years to account for increased usage in both storage and compute. Total usage costs over the span of three years equated to $11,442.

TABLE 5
IBM Spark-As-A-Service Costs

Ref.	Metric	Calculation	Initial	Year 1	Year 2	Year 3
D1	Compute costs			$385	$385	$424
D2	Cloud storage costs			$3,500	$3,500	$3,850
D3	Usage growth				10%	50%
Dt	IBM Spark-as-a-service costs	(D1+D2)*(1+D3)	$0	$3,885	$4,274	$5,828
	Risk adjustment	0%				
Dtr	**IBM Spark-as-a-service costs (risk-adjusted)**		**$0**	**$3,885**	**$4,274**	**$5,828**

Source: Forrester Research, Inc.

Cost Of Coding Efforts In Python For Use On Spark

The existing distributed computing cluster at the SETI Institute was run atop custom C programming code, developed internally. Following the decision to try Spark, the organization worked with IBM to produce new Python coding to run parallelized on the Spark platform. Our interviewees cited the coding guidance from IBM as especially helpful, though the transition still took a fair amount of effort. Current operations on the Spark platform live within a Jupyter notebook, offering scientists the capability to look at the old and new data through lenses that produce quick insights.

Our analysis suggests that an initial effort of $19,500 was spent to develop the code to run on Spark, with the need for an additional scientist in years 2 and 3 to develop additional algorithms to fully utilize the added

computational ability of Spark in subsequent years. The initial and subsequent three years of coding are expected to have a total PV cost of $303,497.

Development costs can vary between projects, especially with nascent technologies where new features are constantly written into the platform. Forrester expects that some organizations may require slightly more coding effort to fully harness the power and functionality of Spark, in order to take advantage of newer features or work with a relatively limited number of scientists with Python or Scala familiarity. As such, we've risk-adjusted the cost upward by 10% for a final cost estimate of $333,847.

TABLE 6
Cost Of Coding Efforts In Python For Use On Spark

Ref.	Metric	Calculation	Initial	Year 1	Year 2	Year 3
E1	Development effort to migrate C code to Python for use on Jupyter/Spark, in development hours		240			
E2	Cost of senior Python developer, fully loaded, per hour		$81.25			
E3	Cost of data scientist to realize added computational power, annually				$180,000	$180,000
Et	Cost of coding efforts in Python for use on Spark	E1*E2+E3	$19,500	$0	$180,000	$180,000
	Risk adjustment	↑10%				
Etr	**Cost of coding efforts in Python for use on Spark (risk-adjusted)**		**$21,450**	**$0**	**$198,000**	**$198,000**

Source: Forrester Research, Inc.

Total Costs

Table 7 shows the total of all costs as well as associated present values (PVs), discounted at 10%. Over three years, the SETI Institute expects total costs to be a PV of approximately $345,000.

TABLE 7
Total Costs (Risk-Adjusted)

Ref.	Cost Category	Initial	Year 1	Year 2	Year 3	Total	Present Value
Dtr	IBM Spark-as-a-service costs	$0	$3,885	$4,274	$5,828	$13,986	$11,442
Etr	Cost of coding efforts for processing tasks in Python for use on Spark	$21,450	$0	$198,000	$198,000	$417,450	$333,847
	Total costs (risk-adjusted)	**$21,450**	**$3,885**	**$202,274**	**$203,828**	**$431,436**	**$345,289**

Source: Forrester Research, Inc.

FLEXIBILITY

Flexibility, as defined by TEI, represents an investment in additional capacity or capability that could be turned into business benefit for some future additional investment. This provides an organization with the "right" or the ability to engage in future initiatives but not the obligation to do so. There are multiple scenarios in which a customer might choose to implement Apache Spark and later realize additional uses and business opportunities. Flexibility would also be quantified when evaluated as part of a specific project (described in more detail in Appendix A).

The 42 radio telescopes of the Allen Telescope Array are capable of capturing an average of 200 Gb per second. And with an overabundance of planets to look at in the Goldilocks zone that could potentially host life, there is no shortage of data to be analyzed. The initial year of usage of the Apache Spark service on the IBM Bluemix cloud has resulted in substantial research advancements, but the accomplishments thus far are just the tip of the iceberg. As the SETI Institute puts it: "There is no shortage of data to analyze. Our limitation is in our [local] computer systems." As such, the SETI organization has identified some future uses as it increases its usage of Apache Spark:

- With increased usage of Spark, the SETI Institute scientists hope to feed a greater amount of signals through a spectrometer and allow machine learning to parse out promising signals from Earth-generated noise. The intention is that with machine learning, a first pass of the signals can quickly and more accurately determine promising signals for further analysis. With only three of over 1,000 beams from the ATA currently being imaged and analyzed, this is an area where machine learning can bring significant benefit.
- In addition to its future use of machine learning, the SETI Institute plans to capture greater ranges of radio frequency, as opposed to the narrow bands around interesting interstellar bodies currently. The sheer speed of Spark enables the team at SETI and citizen scientists alike to crunch through greater amounts data surrounding the exoplanets, by totaling frequencies around an entire habitable zone rather than purely focusing on the signals from the planetary bodies themselves.
- By increasing local on-premises infrastructure in the upcoming years, SETI hopes to be able to record an increased amount of snapshots of radio frequencies to better filter out terrestrial noise and isolate interstellar signals of interest. As SETI previously had no capability to compute beyond its on-premises cluster, it was processing a mere 1 GHz out of a possible 100 MHz of frequency data. The expectation is to increase this dramatically through a combination of performing some near-real-time computation locally and recording greater segments of frequency for batch processing at a later time. While this case study has been focused on the Spark-as-a-service offering from IBM, an on-premises version of IBM Apache Spark is also available and would greatly enhance the SETI Institute's capabilities.

As Gerry Harp from the SETI Institute puts it: "We're looking at data in ways that we've never been able to before. Spark is opening new roads for us."

RISKS

Forrester defines two types of risk associated with this analysis: “implementation risk” and “impact risk.” Implementation risk is the risk that a proposed investment in Apache Spark may deviate from the original or expected requirements, resulting in higher costs than anticipated. Impact risk refers to the risk that the business or technology needs of the organization may not be met by the investment in Apache Spark, resulting in lower overall total benefits. The greater the uncertainty, the wider the potential range of outcomes for cost and benefit estimates.

TABLE 8
Benefit And Cost Risk Adjustments

Benefits	Adjustment
Gains from accelerated research advancements	↓ 15%
Costs	**Adjustment**
Cost of coding efforts for processing tasks in Python for use on Spark	↑ 10%

Source: Forrester Research, Inc.

Quantitatively capturing implementation risk and impact risk by directly adjusting the financial estimates results provides more meaningful and accurate estimates and a more accurate projection of the ROI. In general, risks affect costs by raising the original estimates, and they affect benefits by reducing the original estimates. The risk-adjusted numbers should be taken as “realistic” expectations since they represent the expected values considering risk.

The following impact risk that affects benefits is identified as part of the analysis:

› Many use cases of Apache Spark do not have clear-cut target goals and are often abstract in nature, unlike what the SETI Institute has set out to accomplish. Understanding this, some organizations may not fully utilize the power of Spark and realize the full benefits enabled by Spark’s capabilities. Additionally, not all organizations have the sheer abundance of data in a ready state that starves for computational power as in the case of SETI Institute. It’s important to stress that unless clear-cut algorithms and computational sets have already been planned, the time-to-insights equation could result in a diminished benefit.

The following implementation risk that affects costs is identified as part of this analysis:

› While Spark is incredibly versatile in the programming frameworks that it supports — Scala, Java, or Python — some organizations may not have such capability from the onset to develop in these languages. For the SETI Institute, the conversion from C code to Python for use in Jupyter and Spark required a substantial effort to recode. In organizations that rely heavily on C or C++, the transition from existing analytics tools to Spark may prove to be equally or perhaps even more involved.

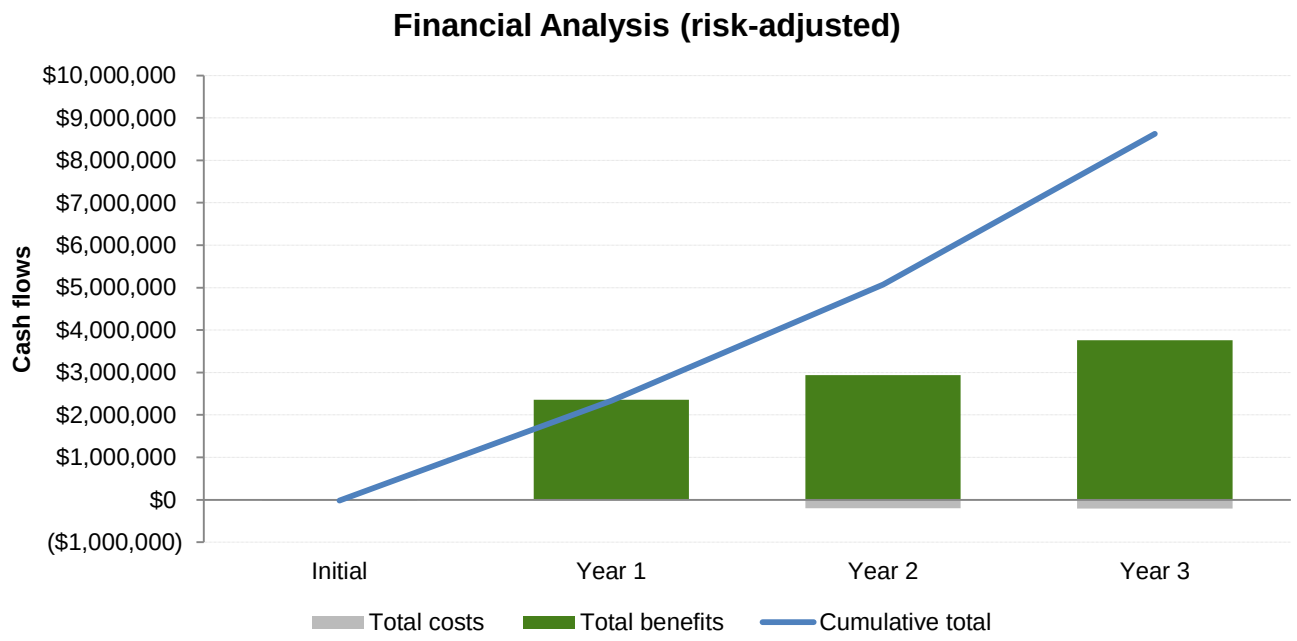
Table 8 shows the values used to adjust for risk and uncertainty in the cost and benefit estimates for the interviewed organization. Readers are urged to apply their own risk ranges based on their own degree of confidence in the cost and benefit estimates.

Financial Summary

The financial results calculated in the Benefits and Costs sections can be used to determine the NPV and overall gains for SETI's investment in Apache Spark.

Table 9 below shows the risk-adjusted NPV and associated benefits and costs in current-day financial terms. These values are determined by applying the risk-adjustment values from Table 8 in the Risks section to the unadjusted results in each relevant cost and benefit section.

FIGURE 3
Cash Flow Chart (Risk-Adjusted)



Source: Forrester Research, Inc.

TABLE 9
Cash Flow (Risk-Adjusted)

	Initial	Year 1	Year 2	Year 3	Total	Present Value
Costs	($21,450)	($3,885)	($202,274)	($203,828)	($431,436)	($345,289)
Benefits	$0	$2,358,750	$2,938,625	$3,762,125	$9,059,500	$7,399,474
Net benefits	**($21,450)**	**$2,354,865**	**$2,736,352**	**$3,558,298**	**$8,628,064**	**$7,054,185**

Source: Forrester Research, Inc.

IBM Apache Spark: Overview

The following information is provided by IBM. Forrester has not validated any claims and does not endorse IBM or its offerings.

Apache Spark is an open source cluster computing framework with in-memory processing, which often enable analytic applications to run up to 100 times faster than alternative technologies. In June 2015, IBM announced support for the open source Apache Spark project (http://spark.apache.org/), making Spark available as a cloud service on the IBM Bluemix cloud platform. IBM also released its SystemML software under a machine-learning license for the Spark community (https://systemml.apache.org/).

There are two ways to develop applications on the IBM Spark service:

- Spark-submit lets developers work with Spark programmatically. This is typically done to run large nightly batch jobs, which can launch unattended from a script triggered by a cron job, and does not require any interaction to complete.
- Interactive Jupyter Notebooks, with either Python or Scala kernels, allow developers to upload and connect to data sources, and perform exploratory analyses (https://ipython.org/).

The SETI Institute has used IBM Spark on Bluemix exclusively through the interactive Jupyter Notebook interface, which enabled them to leverage capabilities that included:

- A scalable framework for faster analysis of complex, large-scale data.
- An environment for innovative development with high-level tools for machine learning and streaming data.
- The ability to connect to large data sets within object store data containers.
- An integrated interface for programming entire clusters with implicit data parallelism and fault tolerance.
- Powerful programming and data constructs, such as the Spark resilient distributed dataset (RDD).
- Open source scientific and machine-learning packages, such as scikit-learn and astroML.

IBM Apache Spark can work with multiple data sources that include object stores like Amazon S3, OpenStack Swift, and IBM SoftLayer. This is enabled using the Stocator, a high-performing connector to object storage for Spark, achieving performance by leveraging object store semantics. Stocator was contributed by IBM to the open source community, under Apache License 2.0. It comes with a complete driver for OpenStack Swift and can easily be extended to support other object storage interfaces (https://spark-packages.org/package/SparkTC/stocator).

Appendix A: Total Economic Impact™ Overview

Total Economic Impact is a methodology developed by Forrester Research that enhances a company's technology decision-making processes and assists vendors in communicating the value proposition of their products and services to clients. The TEI methodology helps companies demonstrate, justify, and realize the tangible value of IT initiatives to both senior management and other key business stakeholders. TEI assists technology vendors in winning, serving, and retaining customers.

The TEI methodology consists of four components to evaluate investment value: benefits, costs, flexibility, and risks.

BENEFITS

Benefits represent the value delivered to the user organization — IT and/or business units — by the proposed product or project. Often, product or project justification exercises focus just on IT cost and cost reduction, leaving little room to analyze the effect of the technology on the entire organization. The TEI methodology and the resulting financial model place equal weight on the measure of benefits and the measure of costs, allowing for a full examination of the effect of the technology on the entire organization. Calculation of benefit estimates involves a clear dialogue with the user organization to understand the specific value that is created. In addition, Forrester also requires that there be a clear line of accountability established between the measurement and justification of benefit estimates after the project has been completed. This ensures that benefit estimates tie back directly to the bottom line.

COSTS

Costs represent the investment necessary to capture the value, or benefits, of the proposed project. IT or the business units may incur costs in the form of fully burdened labor, subcontractors, or materials. Costs consider all the investments and expenses necessary to deliver the proposed value. In addition, the cost category within TEI captures any incremental costs over the existing environment for ongoing costs associated with the solution. All costs must be tied to the benefits that are created.

FLEXIBILITY

Within the TEI methodology, direct benefits represent one part of the investment value. While direct benefits can typically be the primary way to justify a project, Forrester believes that organizations should be able to measure the strategic value of an investment. Flexibility represents the value that can be obtained for some future additional investment building on top of the initial investment already made. For instance, an investment in an enterprisewide upgrade of an office productivity suite can potentially increase standardization (to increase efficiency) and reduce licensing costs. However, an embedded collaboration feature may translate to greater worker productivity if activated. The collaboration can only be used with additional investment in training at some future point. However, having the ability to capture that benefit has a PV that can be estimated. The flexibility component of TEI captures that value.

RISKS

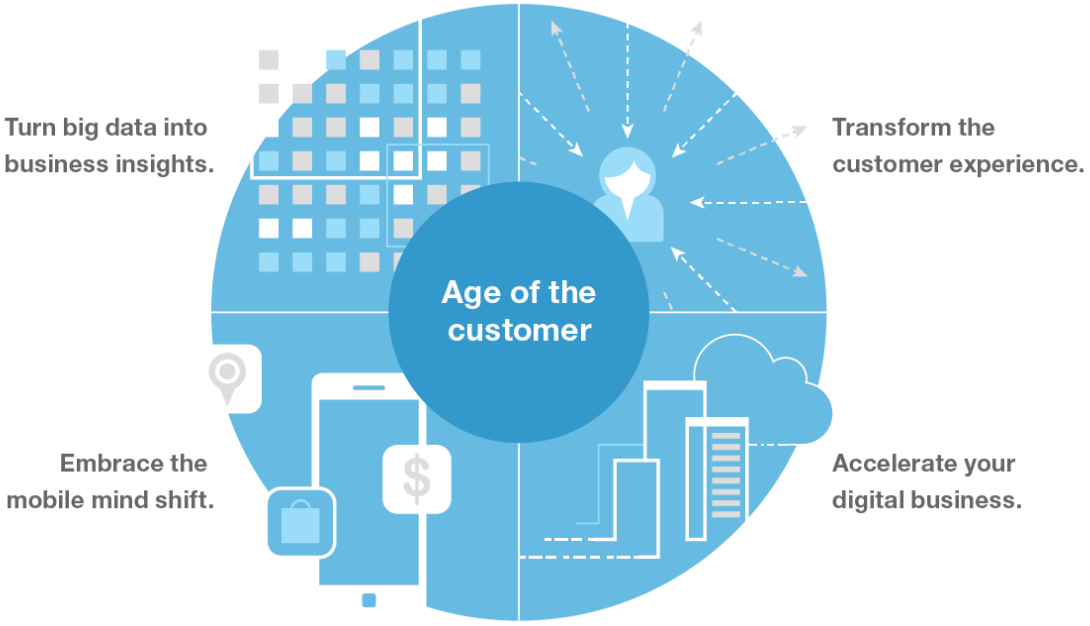
Risks measure the uncertainty of benefit and cost estimates contained within the investment. Uncertainty is measured in two ways: 1) the likelihood that the cost and benefit estimates will meet the original projections and 2) the likelihood that the estimates will be measured and tracked over time. TEI risk factors are based on a probability density function known as "triangular distribution" to the values entered. At a minimum, three values are calculated to estimate the risk factor around each cost and benefit.

Appendix B: Forrester And The Age Of The Customer

Your technology-empowered customers now know more than you do about your products and services, pricing, and reputation. Your competitors can copy or undermine the moves you take to compete. The only way to win, serve, and retain customers is to become customer-obsessed.

A customer-obsessed enterprise focuses its strategy, energy, and budget on processes that enhance knowledge of and engagement with customers and prioritizes these over maintaining traditional competitive barriers.

CMOs and CIOs must work together to create this companywide transformation.



Forrester has a four-part blueprint for strategy in the age of the customer, including the following imperatives to help establish new competitive advantages:

Transform the customer experience to gain sustainable competitive advantage.

Accelerate your digital business with new technology strategies that fuel business growth.

Embrace the mobile mind shift by giving customers what they want, when they want it.

Turn (big) data into business insights through innovative analytics.

Appendix C: Glossary

Discount rate: The interest rate used in cash flow analysis to take into account the time value of money. Companies set their own discount rate based on their business and investment environment. Forrester assumes a yearly discount rate of 10% for this analysis. Organizations typically use discount rates between 8% and 16% based on their current environment. Readers are urged to consult their respective organizations to determine the most appropriate discount rate to use in their own environment.

Net present value (NPV): The present or current value of (discounted) future net cash flows given an interest rate (the discount rate). A positive project NPV normally indicates that the investment should be made, unless other projects have higher NPVs.

Present value (PV): The present or current value of (discounted) cost and benefit estimates given at an interest rate (the discount rate). The PV of costs and benefits feed into the total NPV of cash flows.

Payback period: The breakeven point for an investment. This is the point in time at which net benefits (benefits minus costs) equal initial investment or cost.

Return on investment (ROI): A measure of a project's expected return in percentage terms. ROI is calculated by dividing net benefits (benefits minus costs) by costs.

A NOTE ON CASH FLOW TABLES

The following is a note on the cash flow tables used in this study (see the example table below). The initial investment column contains costs incurred at "time 0" or at the beginning of Year 1. Those costs are not discounted. All other cash flows in years 1 through 3 are discounted using the discount rate at the end of the year. PV calculations are calculated for each total cost and benefit estimate. NPV calculations are not calculated until the summary tables are the sum of the initial investment and the discounted cash flows in each year.

Sums and present value calculations of the Total Benefits, Total Costs, and Cash Flow tables may not exactly add up, as some rounding may occur.

TABLE [EXAMPLE]
Example Table

Ref.	Metric	Calculation	Year 1	Year 2	Year 3

Source: Forrester Research, Inc.

Appendix D: Supplemental Material

Related Forrester Research

"Apache Spark Is Powerful And Promising," Forrester Research, Inc., February 13, 2015

"Brief: Apache Spark Ignites The Big Data Landscape," Forrester Research, Inc., June 15, 2015

"The Forrester Wave™: Big Data Hadoop Cloud Solutions, Q2 2016," Forrester Research, Inc., April 22, 2016

Appendix E: Endnotes

[1] Source: "Brief: Apache Spark Ignites The Big Data Landscape," Forrester Research, Inc., June 15, 2015.

[2] Frank and Sullivan, using quantitative and empirically constrained limits, have proven that extraterrestrial intelligence has existed at some point in time. Source: Frank A. and Sullivan W.T. III, "A New Empirical Constraint on the Prevalence of Technological Species in the Universe," Astrobiology, May 2016 (http://online.liebertpub.com/doi/full/10.1089/AST.2015.1418).

[3] Forrester risk-adjusts the summary financial metrics to take into account the potential uncertainty of the cost and benefit estimates. For more information, see the section on Risks.