Total Value of Ownership (TVO) Assessment of the IBM + Hortonworks Solution for Analytics

Sponsored by IBM

Ajay Asthana, Ph.D. and Srini Chari, Ph.D., MBA mailto:info@cabotpartners.com
March 2018

Executive Summary

The speed and scope of the business decision-making process is accelerating because of several emerging technology trends – Cloud, Social, Mobile, the Internet of Things (IoT), Analytics and Artificial Intelligence/Machine Learning (AI/ML). To obtain faster actionable insights from this growing volume and variety of data, many organizations are deploying Analytics and Hadoop solutions across the entire workflow.

But rising IT costs, cumbersome manual processes and siloed-solutions have become the Achilles heel in deploying and operating Hadoop clusters at many organizations. In addition, data quality/security/privacy/governance and lack of skilled resources are major issues. To meet these challenges, organizations must deploy a cost-effective, easy-to-use, high-performance, reliable and integrated IT solution to consistently deliver the best business outcomes. This is the goal of the IBM + Hortonworks (IBM + HW) solution for Analytics.

Compared to other Hadoop distributions such as Cloudera, the IBM + HW solution delivers several capabilities to streamline complex Analytics workflows and accelerate data-driven decision making. It provides a single Data/Application Integration platform with a common interface/repository to enable better data quality/governance/security/privacy throughout the Analytics workflow.

These IBM + Hortonworks differentiators enable quicker deployments, faster time to value, lower risks of failure and higher revenues/profits. They also enhance the productivity of data scientists, data engineers, application developers and analysts; allowing clients to optimize their ***Total Value of Ownership (TVO)****, which is Total Benefits – Total Costs.*

The comprehensive TVO analysis presented in this paper compares the IBM + Hortonworks solution with a corresponding Cloudera alternative for three configurations – small, medium and large. This cost-benefit analysis framework considers cost/benefit drivers in a 2 by 2 continuum: Direct vs. Derived and Technology vs. Business mapped into four quantified quadrants: Costs, Productivity, Revenues/Profits and Risks.

Compared to using a generic Hadoop solution such as Cloudera, IBM and Hortonworks clients deploying Analytics workflows ***can improve the three-year ROI for all three configurations*** *despite higher software costs. Likewise, the Payback Period (PP) for the IBM + HW solution is shorter than Cloudera; providing clients faster time to value. In fact, these ROI/PP improvements grow with configuration size; offering clients better investment protection as they progress in their Cognitive Computing journey and as data volumes and Analytics model complexities continue to grow.*

Cabot Partners Group, Inc. 100 Woodcrest Lane, Danbury CT 06810, www.cabotpartners.com

Cabot Partners
Optimizing Business Value

To Harness the Growing Value of Analytics, Technology Matters

Data central to key emerging technology trends

The relentless rate and pace of technology-enabled business transformation and innovation are astounding. Several fast-growing intertwined technology trends (Figure 1) – Cloud, Big Data Analytics, Social, Mobile, Internet of Things (IoT) and Artificial Intelligence (AI)/Machine Learning (ML) – continue to be profoundly disruptive, reshaping the information technology (IT) industry. Central to these trends is Data which is growing exponentially.

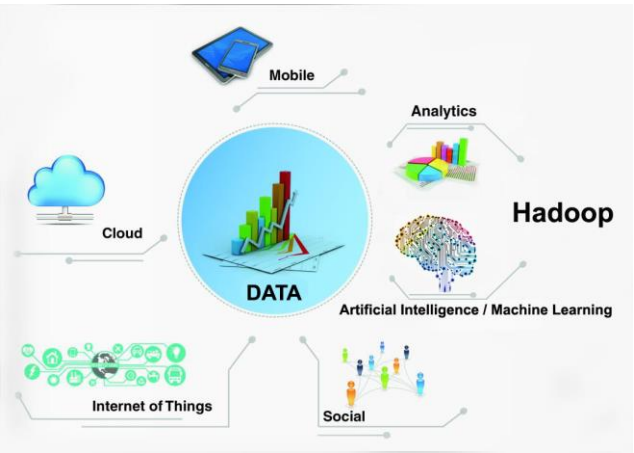

Figure 1: Intertwined Technologies of Cloud, Social, Mobile, IoT, AI/Machine Learning and Analytics

Eighteen-fold increase in data volumes in 10 years

By 2025, the world is expected to have a total of 180 zettabytes of data (or 180 trillion gigabytes), up from less than 10 zettabytes in 2015.[1] In 2018, about 4.3 exabytes (10^{18} bytes) of data is expected to be created daily – over 90% will be unstructured[2] including language-based data (e.g. emails, Twitter messages, books) as well as non-language based data (e.g., images, slides, sensor data, audios, videos).

To get actionable and time-critical insights from this ever-increasing volume and variety of data and stay competitive, many organizations are investing in Analytics and Hadoop. An open- source software framework on distributed systems, Hadoop has become very popular to process large volumes of structured and unstructured data. It also enables Enterprise Data Warehouse (EDW) offloading and optimization which, in turn, helps reduce software license and hardware costs and efficiently monetize new (especially unstructured) data sources.

IBM and Hortonworks provide a single platform for data integration, governance and quality

However, in addition to Hadoop, organizations need a single platform for data integration, governance and quality for the entire Analytics workflow. The IBM + Hortonworks solution provides this; enabling companies to progress on their value-enhancing Cognitive Computing journey from Descriptive to Predictive to Prescriptive and to Artificial Intelligence/Machine Learning (Figure 2).

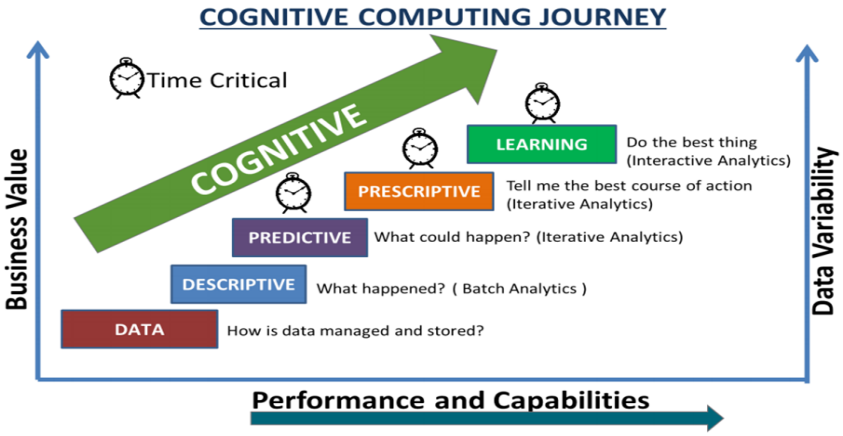

Figure 2: Cognitive Computing Leverages the IBM + Hortonworks Solution

[1] "IoT Mid-Year Update From IDC And Other Research Firms," Gil Press, *Forbes*, August 5, 2016.
[2] https://storageservers.wordpress.com/2016/02/06/how-much-data-is-created-daily/

Data Scientists spend about 79% of their time with data prep and cleansing

Deploying higher-value Analytics and AI/Machine Learning is complex and labor intensive. Data Scientists typically spend about 79%[3] of their time with cumbersome data preparation and cleansing tasks. Fortunately, continuous improvements in software are fueling the rapid adoption of higher-value Analytics and enabling organizations to generate time-critical insights and maximize their Total Value of Ownership (TVO).

This paper uses a TVO model that quantifies some of the key interrelated cost and benefit drivers and differentiators of the IBM + Hortonworks solution over the Cloudera alternative. These cost and value drivers were identified using over a dozen in-depth interviews representing IBM customers and Analytics experts across multiple industries and company sizes, IBM input and other research. This holistic cost-benefit analysis examines various configuration sizes (small, medium and large) for the entire Analytics workflow.

Optimizing the Analytics Workflow with IBM + Hortonworks

Need to optimize entire Analytics workflow to maximize Total Value of Ownership (TVO)

Most Analytics projects are complex and iterative (AI/Machine Learning is highly iterative) with four phases and centered on Data (Figure 3):

1. Business Understanding: Determine the business objectives and success criteria. Assess the situation including available resources, requirements, assumptions, constraints, risks, terminology, costs and benefits.
2. Data Understating and Preparation: Collect initial data, then describe, explore and verify data, particularly for quality. For data preparation, select, include/exclude with rationale, clean, construct, integrate/merge and format/re-format data.
3. Modeling and Evaluation: Select modeling technique with assumptions. Generate test design, build the model, set parameters, describe and assess the model and revise parameter settings as needed. Evaluate results against success criteria, approve models, review process and determine next steps/actions.
4. Deployment: Plan, monitor and maintain model deployment. Produce reports/presentations and review/document project experience/results.

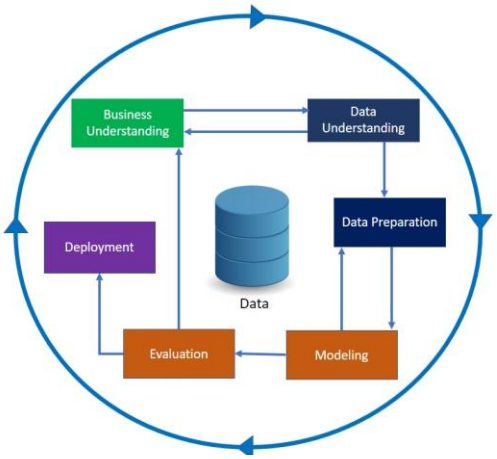

Figure 3: Typical Analytics Workflow

IBM + Hortonworks solution combines IBM Db2 Big SQL, IBM Data Science Experience, IBM Unified Governance and Integration with IBM Spectrum Scale and the Hortonworks Data Platform (HDP)

To deploy these complex Analytics workflows, many organizations use cumbersome manual processes and/or multiple disparate tools that don't easily integrate or interoperate. This impedes the realization of business value and lowers the Total Value of Ownership (TVO).

To overcome these obstacles, the IBM + Hortonworks solution combines IBM Db2 Big SQL, IBM Data Science Experience, IBM Unified Governance and Integration and IBM Spectrum Scale with the Hortonworks Data Platform (HDP) into a single Data/Application Integration platform with a common interface and repository. This accelerates data-driven decision making and provides key differentiators (detailed later) that are included in the TVO Framework for Analytics Workflows.

[3] CrowdFlower, "2016 Data Science Report".

TVO Framework organized by Technical/ Business and Direct/ Derived cost and value drivers

Key cost/value drivers identified and quantified holistically

Value includes lower costs, enhanced productivity, higher revenues/ profits and lower risks

High Level TVO Framework + Key Cost/Value Drivers for Analytics

The TVO framework (Figure 4) categorizes the interrelated cost/value drivers (circles) for Analytics by each quadrant: Costs, Productivity, Revenue/Profits and Risks. Along the horizontal axis, the drivers are arranged based on whether they are primarily **Technical** or **Business** drivers. Along the vertical axis, drivers are arranged based on ease of measurability: **Direct** or **Derived**.

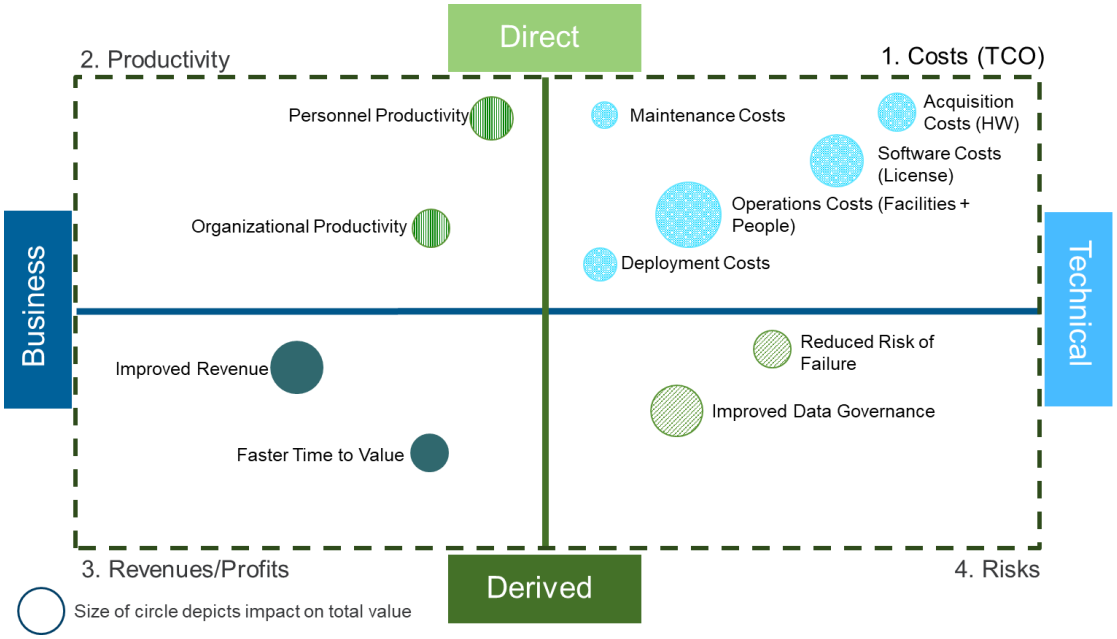

Figure 4: TVO Framework for Analytics Workflows with Cost/Value Drivers

The cost/value drivers for Analytics are depicted as circles whose size is proportional to the potential impact on a client's Total Value (Benefits – Cost) of Ownership or TVO as follows:

1. **Total Costs of Ownership (TCO):** Typical costs include: one-time acquisition costs for the hardware and deployment, and annual costs for software, maintenance and operations.
2. **Improved Productivity:** The TVO model quantifies the value of productivity gains of data scientists, data engineers, applications developers and the organization.
3. **Revenue/Profits:** Faster time to value with better performance-optimized parallel and in-memory processing and a single open, flexible and agile platform. Greater innovation and better decision-making capabilities spur growth, revenues and improve profits.
4. **Risk Mitigation:** Lower risk of project failure (even well-planned Analytics projects have up to 60% failure rate[4]) with a streamlined workflow with reusable components, better ecosystem and business/IT collaboration, and enhanced security/ privacy. Improved governance with better data cleansing/quality and process consistency.

The TVO for both IBM + Hortonworks and Cloudera typically grow with data/configuration size, with key differentiated features in the IBM + Hortonworks (IBM + HW) solution (detailed in the Results Discussions section) driving added value over Cloudera.

[4] Why big data projects fail and how to make 2017 different, Expansion of Gartner's prediction that 60% of big data projects fail; By Sameet Agarwal, **Network World** Feb 16, 2017.

Total Value of Ownership (TVO) for Analytics – Assumptions/Results

TVO analysis for three configurations: small, medium and large

The Cost-Benefit Analysis presented here quantifies the Total Value (Total Benefits – Total Costs) for Three Years of operations. The IBM + Hortonworks solution is compared with a corresponding Cloudera alternative. Three configurations (Table 1) are analyzed: small, medium and large. All costs and benefits grow with data/configuration size for IBM + HW and Cloudera.

Key Assumptions: Licensing costs for IBM value-added components: IBM Db2 Big SQL, IBM Data Science Experience (DSX), IBM Unified Governance and Integration and IBM Spectrum Scale were provided by IBM. Other assumed costs for Hadoop enterprise software license, hardware acquisition, maintenance, operations and deployment were included in the TVO model after extensive research (see Additional References at the end) and interviews. Electricity costs are assumed to be $.09/kWh. The IBM + Hortonworks software costs per node are assumed to be a substantial 50% greater than Cloudera.

IBM + HW solution reduces data center footprint

The number of management and data nodes for each configuration (Table 1) was scaled based on the size of data. The IBM + Hortonworks (IBM + HW) solution requires fewer data nodes and reduces the data center footprint.[5] In addition, no data copying/migration is required for running Hadoop analytics.

CONFIGURATION	SMALL		MEDIUM		LARGE	
	IBM+HW	Cloudera	IBM+HW	Cloudera	IBM+HW	Cloudera
Storage (TB)	200	200	750	750	1500	1500
Number of Management Nodes	3	3	3	3	6	6
Number of Data Nodes	10	12	40	47	87	94
Total Number of Nodes	**13**	**15**	**43**	**50**	**93**	**100**

Table 1: System Configuration Details for IBM + Hortonworks (IBM + HW) and Cloudera

Table 2 details the assumptions made in deploying staff (by roles) associated with the development and deployment of Analytics projects. These skills are often very scarce, and organizations compete for them and pay a premium. Operational people costs are typically the largest cost component of Analytics projects. So, solutions that reduce staff costs and enhance personnel productivity help improve the TVO of Analytics projects. Compared to a generic version of Hadoop, both IBM + HW and Cloudera reduce the total hours required for key personnel with greater reduction with the IBM + HW solution.

Operational people costs are typically the largest cost component of Analytics projects.

	SMALL			MEDIUM			LARGE		
Number of Hours per Year	IBM+HW	Cloudera	Hadoop	IBM+HW	Cloudera	Hadoop	IBM+HW	Cloudera	Hadoop
Data Scientist Hours	975	1050	1500	1550	1625	2500	2320	2480	4000
Data Engineer Hours	1625	1750	2500	2480	2600	4000	4640	4960	8000
Application Developer Hours	975	1050	1500	1240	1300	2000	4640	4960	8000
Business Analyst Hours	1400	1500	2000	2680	2900	4000	5200	5600	4000
Administrator Hours	950	1000	1100	1840	2000	2400	3600	4000	5200
Total Personnel Hours	**5925**	**6350**	**8600**	**9790**	**10425**	**14900**	**20400**	**22000**	**29200**

Table 2: Typical Breakdown of Key Roles and Hours of Analytics Staff by Project

Results for a Small Analytics Configuration: Figure 5 (next page) depicts the costs and benefits mapped by each quadrant and value driver. IBM + HW's significantly larger software license costs are more than offset by lower personnel costs and higher client benefits in enhanced productivity, higher revenues/profits and lower risks.

[5] https://developer.ibm.com/storage/2017/12/05/ibm-spectrum-scale-hortonworks-hdp-hadoop-clusters-complete-big-data-solution/

For small and medium configs, IBM + HW solution provides more productivity, faster time to value, better governance and higher revenue/profits

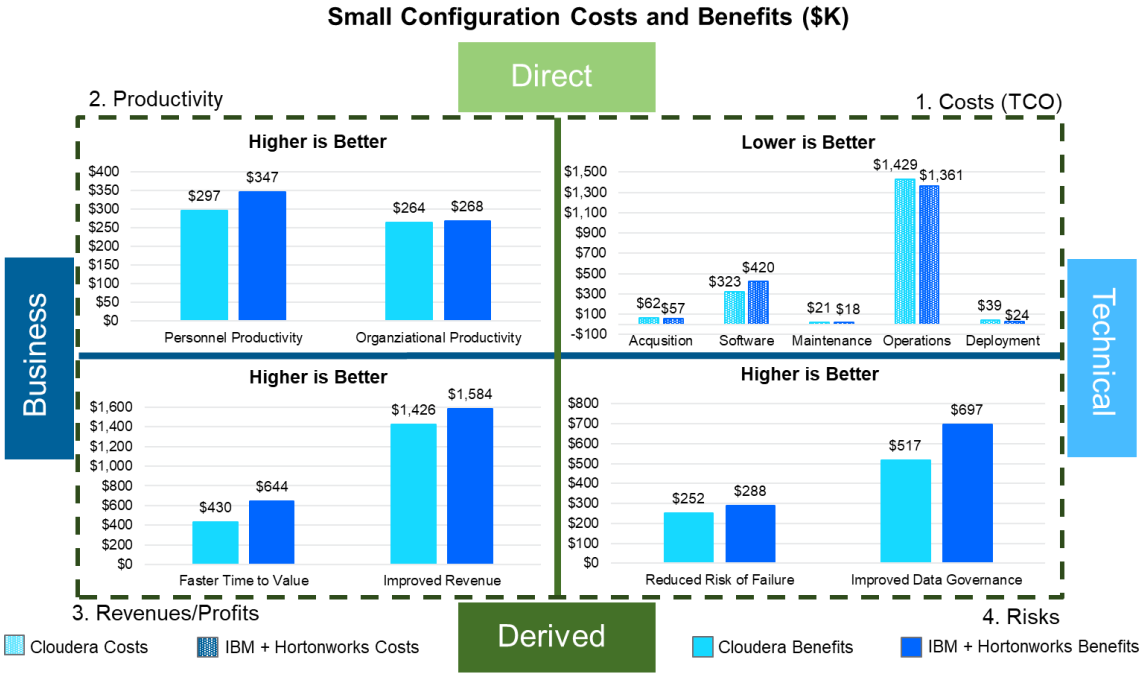

Figure 5: Costs and Benefits by Driver for IBM + Hortonworks versus Cloudera (Small)

Results for a Medium Analytics Configuration: Figure 6 depicts the costs and benefits mapped by each quadrant and value driver. Again, IBM + HW's significantly larger software costs are more than offset by even greater client benefits in enhanced productivity, higher revenues/profits and lower risks.

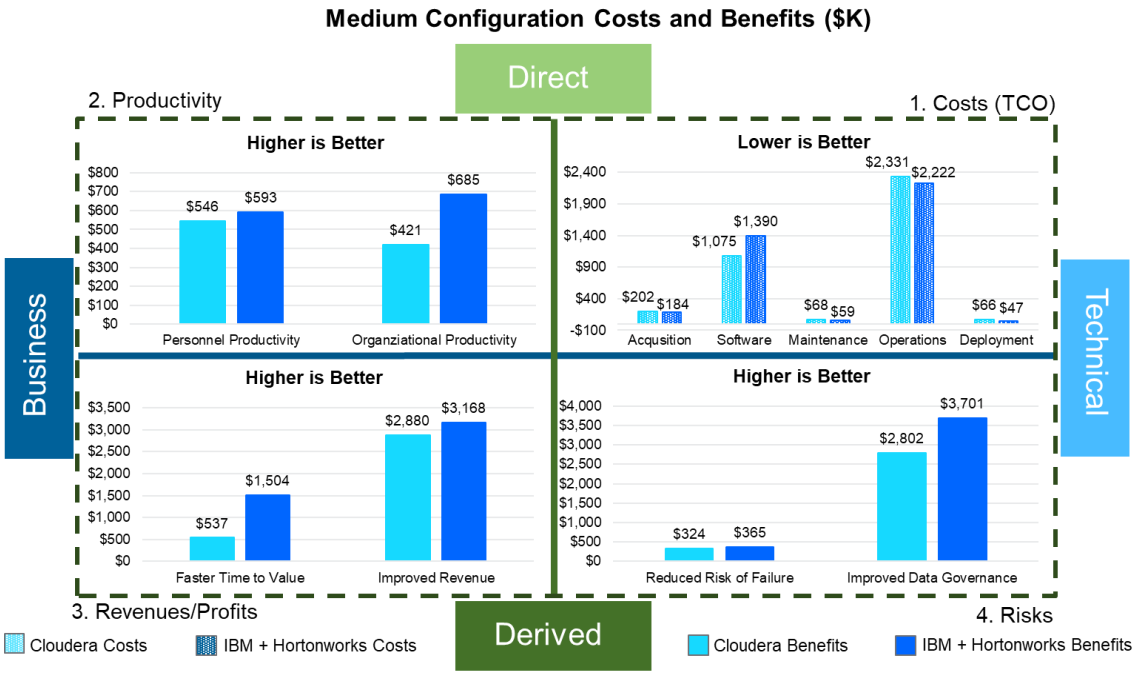

Figure 6: Costs and Benefits by Driver for IBM + Hortonworks versus Cloudera (Medium)

For large configs, the large IBM + HW solution can further drive radical innovation and create entirely new business opportunities

Results for a Large Analytics Configuration: Figure 7 depicts the costs and benefits mapped by each quadrant and value driver. Again, IBM + HW's significantly larger software costs are considerably more than offset by even greater client benefits in enhanced productivity, higher revenues/profits and lower risks.

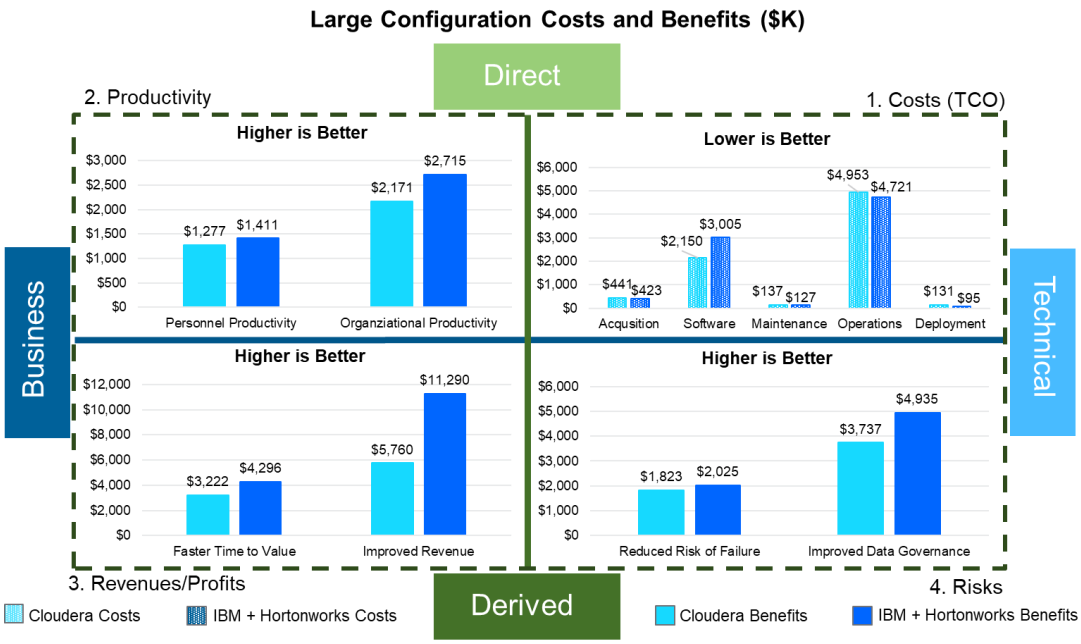

Figure 7: Costs and Benefits by Driver for IBM + Hortonworks versus Cloudera (Large)

These large configurations can be very complex to manage and operate. However, with the IBM + HW solution, clients can get significant value because of unique Analytics capabilities and a single Data/Application Integration platform with a common interface and repository across the entire workflow. This also helps drive radical innovation and has the potential of creating entirely new business/product/customer categories for clients across many industries as they progress on their Cognitive Computing journey from Descriptive to Predictive to Prescriptive Analytics and to Artificial Intelligence and Machine Learning.

Table 3 consolidates benefits, costs, ROI and Payback Period for all three configurations.

	SMALL		MEDIUM		LARGE	
Financial Calculations	IBM+HW	Cloudera	IBM+HW	Cloudera	IBM+HW	Cloudera
Total 3-Year Benefits Risk Adjuste	$3,827,733	$3,185,402	$10,015,635	$7,510,388	$26,672,193	$17,988,923
Total 3-Year Costs Risk Adjusted	$2,213,653	$2,170,648	$4,774,218	$4,482,621	$10,240,958	$9,312,842
Return on Investment (ROI)	73%	47%	110%	68%	160%	93%
Payback Period (Months)	9.57	14.93	7.04	12.75	5.62	6.4

Table 3: Consolidated Benefits, Costs, ROI and Payback Period

IBM + HW solution lowers operational and deployment costs

Discussions of Results from the Total Value of Ownership Model

Most Total Cost of Ownership (TCO) models only quantify the costs in Quadrant 1. The TVO model outlined here not only considers these costs but also the benefits from the value drivers in the remaining three quadrants. Compared with Cloudera, the IBM + HW solution (with IBM Db2 Big SQL, IBM Data Science Experience – DSX, IBM Unified Governance

and Integration Platform and IBM Spectrum Scale) reduces costs and drives additional benefits for all configurations by:

IBM + HW solution lowers operational and deployment costs

1. **Lowering Operational/Deployment Costs:** Compared with Cloudera, the IBM + HW solution with IBM Spectrum Scale reduces the data center footprint, and hence lowers facilities and power and cooling costs. With a single Data/Application Integration platform with a common interface and data repository, the IBM + HW solution allows users to develop data integration and data quality integrated routines with no coding, and run these routines anywhere. The same platform can also be used for new data sources and industry-specific content. This helps lower deployment times (one-time cost/project) and improve productivity. All these differential savings with IBM + Hortonworks solution grow with configuration size.

Enhances productivity of staff and organization

2. **Enhancing Productivity:** Some of the key IBM + HW solution features that further improve productivity of key staff by roles include:

   a. *Data Engineers* who architect how data is organized and ensure operability can perform data quality assessment and cleansing to ensure data being moved into Hadoop, can be trusted and used to make intelligent decisions and predictions.
   b. *Application Developers* who plug into data and models and write code can leverage numerous curated data sets, built-in learning modules and a vast choice of open-source and vendor-specific (including Microsoft) development environments.
   c. *Data Scientists* who delve deep into the data to draw hidden insights for the business can leverage tools to automate and simplify data discovery, curation, and governance. Intelligent search capabilities help them find the data they need, while metadata such as tags, comments, and quality metrics help them decide whether a data set will be useful to them and how best to extract value from it. Integrated data governance gives data scientists confidence that they are permitted to use a given data set and that the models and results they produce are used responsibly by others in the organization.
   d. *Business Analysts* who work with data to apply insights to business strategy have an easy-to-use, customized interface. They can also use community and social features that provide meaningful collaboration to share code and provide feedback on analysis.

   Organizational productivity is further enhanced with other IBM value-added capabilities/ expertise to advise, deploy, integrate and support throughout the Analytics journey. In addition, the IBM + HW solution provides a comprehensive, enterprise-grade and agile data/application integration environment with quality and governance.

Increases revenues/ profits with faster time to value and innovation

3. **Increasing Revenues/Profits:** In addition to enhanced productivity, the IBM + HW solution also delivers faster time to value with better performance, scale, ease of deployment and integration. Key features include visual and SQL interfaces that are fully ANSI compliant with schema flexibility and reusable, portable Analytics workflows. In addition, support for in-memory, parallel and vector processing improve data ingestion and query performance.

   Greater innovation and better decision-making capabilities with trusted data help improve customer service and support, identify new customers/markets/products/services, reduce time to market, build better pricing models and more.

Mitigates risks with improved governance/ data quality and security/ privacy

4. **Mitigating Risks:** The IBM + HW solution lowers risk of project failure and delays with a streamlined workflow with reusable components, industry-specific expertise, better ecosystem and business/IT collaboration, and enhanced security/ privacy. It also helps improve governance with better data cleansing/quality and process consistency. This reduces the time and costs associated with identifying and correcting errors in the data. Better governance also lowers risks associated with regulatory non-compliance in highly regulated industries.

Better ROI/Payback for Analytics with IBM and Hortonworks

ROI and Payback Period improve as configuration sizes grow

Key financial metrics for a 3-year time horizon for all three configurations for IBM + HW and Cloudera (Figure 8) include: Return on Investment (ROI) and Payback Period (PP).

IBM + HW solution consistently delivers better ROI and Payback compared with Cloudera

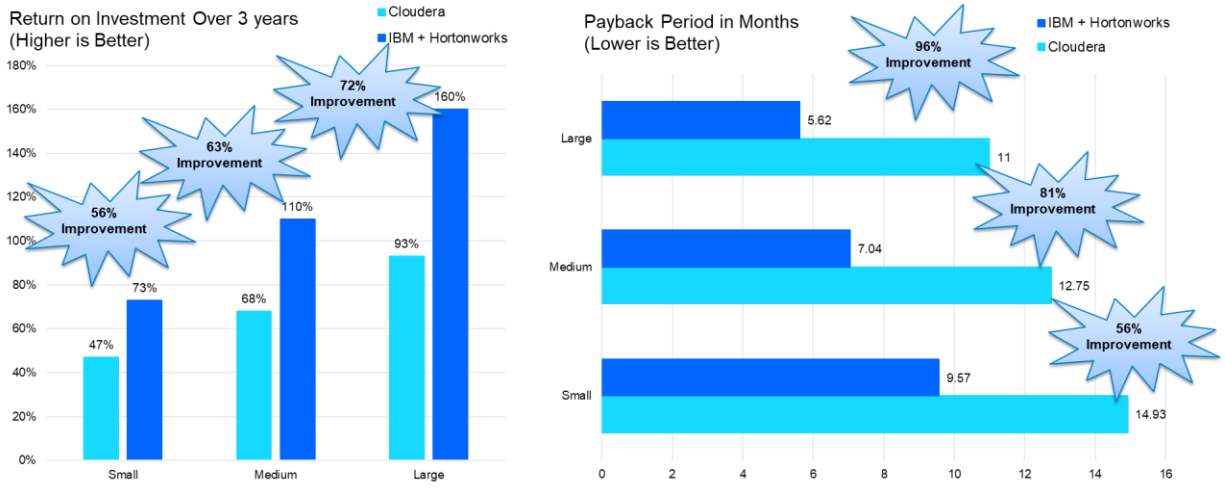

Figure 8: Better ROI and Payback Period with IBM + Hortonworks Over Cloudera

ROI and Payback differentials improve with configuration size

For all cases, the ROI and Payback Period improve as configuration sizes grow from small to medium to large. This is expected. Clients typically increase the configuration size as they progress in their Cognitive Computing journey to solve higher-value Analytics (greater ROI) problems such as Predictive, Prescriptive and AI/Machine learning.

The percent improvements in ROI and Payback Period for IBM + HW over Cloudera also grow (Figure 8) with configuration size. This means as clients grow their Analytics footprint to drive more innovation, the IBM + Hortonworks solution will deliver even more value.

Conclusions and Recommendations

Analytics is a game-changing business opportunity for companies to deliver exceptional customer experience, enhance marketing effectiveness, increase operational efficiencies, reduce financial risks, improve product quality and reliability, and more.

However, as the volume and variety of data grow and as organizations implement higher-value Analytics and Hadoop for their Cognitive Computing journey, they need enterprise-grade solutions to produce more realistic, reliable, actionable and time-critical analyses.

IBM + HW solution is a single Data/ Application Integration platform with a common interface and repository to accelerate decision-making

IBM + HW solution delivers faster time to value, greater productivity, better decisions with higher confidence

Considerable ROI improvement of 56% (small) to 72% (large) over a Cloudera solution

IBM is a reliable partner with the most comprehensive Enterprise Data Integration, Quality and Governance solution

The IBM + Hortonworks (IBM + HW) solution combines IBM Db2 Big SQL, IBM Data Science Experience, IBM Unified Governance and Integration and IBM Spectrum Scale with the Hortonworks Data Platform (HDP) into a single Data/Application Integration platform with a common interface and repository. This accelerates data-driven decision making and provides the following key differentiators relative to other Hadoop solutions such as Cloudera:

- Faster time to value with better performance, scale, ease of deployment and integration.
- Greater productivity of staff: Data Engineers, Data Scientists, Application Engineers and Business Analysts.
- More new product/business ideas, better decisions and innovation from deeper insights gleaned by higher-value Analytics/progressing along the Cognitive Computing journey.
- Better governance with enhanced data quality, consistency, and security for the workflow.

For clients, these features/benefits collectively reduce costs, enhance productivity, drive revenues/profits and mitigate risks.

The 3-year TVO analysis presented here quantifies all these cost/value drivers holistically for three configurations: small, medium and large. The ROI for the IBM + HW solution ranges from 73% (small) to 160% (large) while a corresponding Cloudera solution delivers an ROI of 47% (small) to 93% (large). This implies that the IBM + HW solution delivers a considerable ROI improvement of 56% (small) to 72% (large) over a Cloudera solution.

Likewise, the Payback Period (PP) in months for the IBM + HW solution ranges from 5.62 (large) to 9.57 (small) while the corresponding Cloudera solution delivers a PP in months of 11.0 (large) to 14.93 (small). This implies that the IBM + HW solution also delivers better Payback improvements from 96% (large) to 56% (small) over a Cloudera solution.

Despite the larger software license cost over a Cloudera solution, clients deploying Analytics workflows should seriously consider the IBM +HW solution for the following reasons:

1. The cost-benefit analysis and business case are compelling for all configurations.
2. The business value and ROI/PP differential improve as configurations get larger.
3. This investment is protected and can continue to deliver even greater marginal value for more complex analytics including the rapidly growing use of Artificial Intelligence and Machine Learning (AI/ML) techniques coupled with the Internet of Things (IoT) – all areas where IBM continues to make substantial investments.
4. IBM is a reliable partner and offers the most comprehensive Enterprise Data Integration, Quality, and Governance Solution.

Additional References

1. The Total Economic Impact™ Of the Dell | Cloudera Apache Hadoop Solution, Accelerated by Intel Cost Savings; A Forrester study commissioned by Dell & Intel, Sean McCormick, Nov 2015.
2. The Total Economic Impact™ Of Hortonworks; A Forrester study commissioned by Hortonworks, March 2017
3. IBM COURTS THE ELUSIVE DATA SCIENTIST, Research Note on Data Science Experience, Nucleus Research, Anne Moxie, July 2016
4. Analytics pays back $13.01 for every dollar spent, Research Report 204, Nucleus Research, September 2014
5. Why big data projects fail and how to make 2017 different, Expansion of Gartner's prediction that 60% of big data projects fail; By Sameet Agarwal, Network World, Feb 16, 2017.
6. 100 data and analytics predictions through 2021, Douglas Laney, Ankush Jain, 20 Gartner Research, June 2017 ID: G00332376
7. How companies are using big data and analytics, McKinsey & Co., McKinsey Analytics, Apr 2016
8. The Financial Impact of NOT having Data Governance, Andrew Joss; Informatica Blog, Nov 29th, 2016