Cost-Benefit Analysis: Comparing the IBM PureData System with Hadoop Implementations for Structured Analytics

Sponsored by IBM

Ajay Asthana Ph.D., Srini Chari, Ph. D., MBA mailto:info@cabotpartners.com
April, 2015

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

Executive Summary

The speed and scope of the business decision-making process is evolving because of several emerging technology trends – Cloud, Analytics, Social, Mobile and the Internet of Things (IoT). These data-intensive trends require high-performance distributed systems to deliver actionable insights. For this, Hadoop clusters are increasingly being used.

But rising IT costs (labor, energy, and facilities) have become the Achilles heel in deploying and operating Hadoop clusters at many organizations. In addition, data security 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 agile IT solution to deliver the best business outcomes. This is the goal of the IBM PureData System for Analytics (PDA).

PDA includes several smart features designed to bring speed, simplicity and scalability to run complex analytics for better outcomes. The integration of processors, software, and storage leads to shorter application development cycles and exceptional time to value. This appliance requires minimal ongoing administration or tuning which allows customers to lower their Total Cost of Ownership (TCO).

The three year TCO (IT Costs + Business Costs) analysis presented in this paper compares IBM PureData for Analytics and a Hadoop Cluster (Cloudera) for four configurations – small, medium, large and enterprise. Very favorable assumptions are used for Hadoop. IT Costs include Acquisition, Maintenance, Deployment, Administration, Facilities and Provisioning Costs. Business Costs include Opportunity, Downtime and Productivity.

*Compared to a Hadoop cluster, clients implementing Analytics in an SQL environment with the **IBM PureData System for Analytics can lower the TCO for all configurations.** PDA IT Costs are also lower than the Hadoop Cluster for small to medium configurations. For large and enterprise configurations, IT Costs for Hadoop cross over (at large configurations) and become lower than PDA. Clients who may be concerned solely with these IT Costs can implement a hybrid solution of a medium-sized IBM PureData System with "hot" data, fronting a Hadoop cluster to get the advantages of speed, simplicity and scalability for large complex analytics workloads.*

The IBM PureData for Analytics minimizes the hassles of managing technology complexity and consistently lowers TCO. Consequently, clients benefit from faster time to value, higher revenues and profits, better product/service quality and potentially more innovation.

For Big Data Analytics, SQL on Distributed Systems Will Dominate

Considerable investment in Data and Analytics

The relentless rate and pace of technology-enabled business transformation and innovation are astounding. Several fast-growing intertwined technology trends – Cloud, Big Data Analytics, Social, Mobile and Internet of Things (IoT) – continue to be profoundly disruptive, reshaping the economics and the needs of the information technology (IT) industry.

Consider the challenges and opportunities with Data and Analytics. More than 2.5 exabytes (10^{18} bytes) of data are created daily. Individuals generate about 70% of this data and enterprises store and manage 80% of this data. Spending on Data is growing annually at 30% and is expected to reach $114 billion in 2018.[1]

Analytics on Structured and Unstructured Data growing in importance

Data can be structured or unstructured.[2] Structured data is data that has been clearly formed, formatted, modeled, and organized so that it is easy to work with and manage e.g. relational databases, spreadsheets, vectors and matrices. Structured Query Language (SQL) is proven over the last several decades to be the primary way to work with structured data.

Unstructured data covers most of the world's information but does not fit into the existing databases for structured data. Further, unstructured data consists of language-based data (e.g. emails, Twitter messages, books) as well as non-language based data (e.g., images, slides, sensor data, audios, videos). An estimated 85% of data is unstructured.[3] Hadoop, an open-source software framework on distributed systems, has become very popular in recent years to process large volumes of structured and unstructured data.

Cloud, Big Data Analytics, Social, Mobile and IoT are key trends producing data deluge

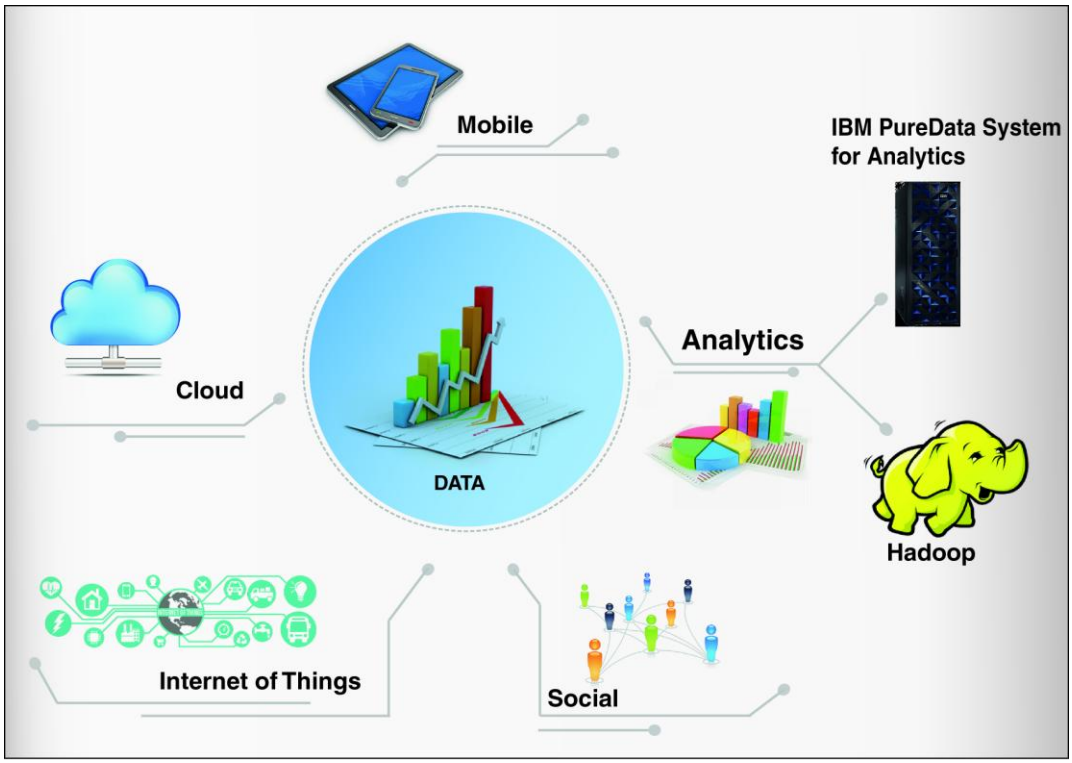

Figure 1: Intertwined Technologies of Cloud, Social, Mobile, IoT and Analytics

[1] A. T. Kearney and Carnegie Mellon University, "Beyond Big: The Analytically Powered Organization", January 2014.
[2] Big Data, Bigger Opportunities, Jean Yan, President Management Council Inter-agency Rotation Program, Cohort 2, April 2013.
[3] http://www.dataenthusiast.com/2011/05/85-unstructured-data-15-what-the-hell-is-going-on/

It is not just the volume of data that's important but also the variety, velocity, veracity and vulnerability. The ability to virtualize and visualize data to extract value is also crucial.

Hadoop implementation requires considerable skills and resources

For many years, businesses have been leveraging structured data and semi-structured data for Analytics. But most databases can typically handle only one type of data. It is challenging to unify different data models so that all data can be analyzed together. Open-source initiatives like Apache Hive and Pig offer a layer for SQL on Hadoop. But they typically require more highly-skilled people resources to deploy and support production application environments. Security and data protection are some of the other concerns that Enterprises must deal with when implementing the open-source solutions based on Hadoop.

As the boundaries between relational and non-relational data base systems continue to blur, SQL will continue to be the preferred method to work with data for the following reasons:[4]

- **Widespread use** with millions of well-trained users.
- **Stability** with relational database management systems supporting SQL compatibility, transactional consistency, and enforced schema required by enterprises.
- **Optimized for performance and scale** with distributed/parallel systems and in-memory computing.

Parallel scale-out systems with SQL support reduce costs and complexity of operations

Distributed/parallel scale-out systems provide many benefits to address the data deluge:[4]

- **Seamless growth** – Scaling capacity or performance is fast and painless; often triggered with a click or by a simple command.
- **Schema flexibility** – As applications mature, schema changes can be made without taking the system down.
- **High availability** – Higher reliability with fault tolerance and multiple redundancies.

Distributed systems such as the IBM PureData for Analytics and Hadoop clusters are being deployed by many enterprises worldwide for Analytics on structured data using SQL.

Comparing IBM PureData for Analytics with Hadoop

IBM PureData System for Analytics integrates capabilities into an appliance lowering TCO

IBM PureData System for Analytics (PDA) with several smart features is positioned to bring speed, simplicity and scalability for better outcomes. The system is designed specifically to run complex analytics on Terabytes (TB) and Petabytes (TB) of data, orders-of- magnitude faster than traditional custom systems. The integration of processors, software, and storage leads to shorter application development cycles and exceptional time to value for business analytics initiatives. This appliance also requires minimal ongoing administration or tuning which allows customers to realize a much lower Total Cost of Ownership (TCO).

Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on compute clusters built on commodity hardware. Several companies such as Cloudera, HortonWorks and even IBM (BigInsights) provide Hadoop distributions with other value-added software components with services and support to enterprises for a fee. Selecting hardware to provide the best balance of performance

[4] http://www.dbta.com/Editorial/Trends-and-Applications/5-Big-Data-Trends-Moving-into-the-Spotlight-in-2015-101633.aspx

Total Cost of Ownership includes IT Costs and Business Costs

and costs for a given workload requires a lot of testing and validation. Moreover, during initial deployment, systems must be carefully configured to ensure that networks, disks and hosts are properly laid out and tuned to maximize utilization and minimize problems especially as data sets get very big. This requires diverting limited, expensive high-skilled resources for mundane tasks instead of using these resources for strategic business initiatives. In addition, security and data protection are concerns that will require more resources.

The Cost-Benefit Analysis presented in this paper compares the *Total Cost (IT Cost + Business Cost) for Three Years* of IBM PureData for Analytics and a Hadoop Cluster (Cloudera) for four configurations – small, medium, large and enterprise – with data requirements of 18 TB, 192 TB, 750 TB and 1500 TB respectively.

Four configurations with increasing sizes analyzed

Six **IT Costs** are considered:

1. *Acquisition* – Cost of software, servers, storage, networks, etc. Included in Year 1.
2. *Maintenance* – Cost for maintenance and support of the environment. Annual.
3. *Deployment* – Time-related costs to deploy the application (assume application is already developed). One time initial cost.
4. *Administration* – Full Time Equivalent (FTE) people costs to support operations. Annual.
5. *Facilities* – Costs for buildings, energy and cooling. Annual.
6. *Provisioning* – Cost of extra system capacity to handle peaks and/or failures. Included in Year 1.

Three **Business Costs** considered include:

1. *Opportunity* – Business lost while the application is being configured and deployed. One time initial cost.
2. *Downtime* – Lost business during the times when the system is down. Annual.
3. *Productivity* – Lost productivity of resources (data scientists, engineers, administration because of slow execution speed and job failures). Annual.

Lower Total Cost of Ownership with the IBM PureData System

Lower TCO for IBM PureData System for Analytics compared to Hadoop in all cases

Even though Hadoop is open-source software, across all four configurations, the IBM PureData System for Analytics (PDA) provides a lower Total Cost of Ownership (TCO) than the Hadoop Cluster. For smaller configurations, the total IT Cost of PDA is lower than Hadoop. For larger configurations, the IT costs are comparable but the much lower Business Cost of PDA keeps the TCO of PDA lower.

Small Configuration (18 TB): For a small configuration (Figure 2), we found that the IT Costs of the Hadoop cluster are 144% more than PDA. When Business Costs are included, Hadoop is 198% more expensive.

49% lower TCO with IBM PureData System for Analytics on small configurations

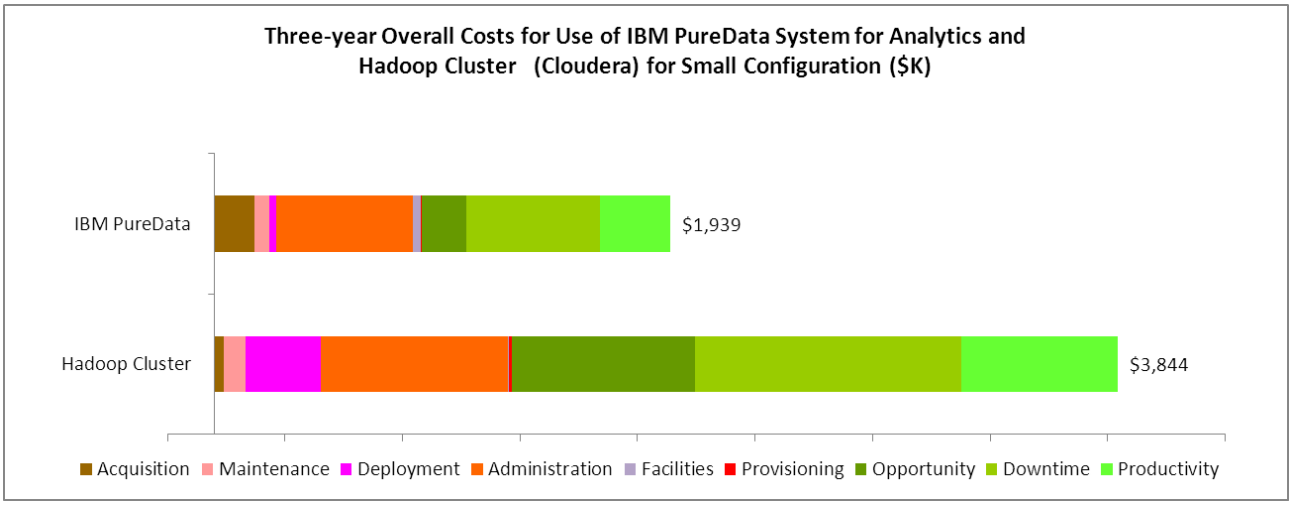

Figure 2: IBM PureData System for Analytics Lowers TCO by over 49% for Small Configurations

Since Hadoop is open-source and runs on commodity hardware, the acquisition, maintenance and facilities costs for a Hadoop cluster is less than the IBM PureData System. But the cost of deployment and administration of a Hadoop Cluster is substantially more than the IBM PureData System, making the total IT Costs of PureData lower than a Hadoop cluster. When Business Costs of lost opportunity, downtime and lost productivity are added, the TCO of the PureData System is significantly lower (by 49%) than the Hadoop cluster.

Larger deployment and administration costs with Hadoop

Medium Configuration (192 TB): For a medium configuration (Figure 3), the Hadoop cluster is 126% more expensive than the IBM PureData System in IT Costs, and is 187% more expensive when Business Costs are also added.

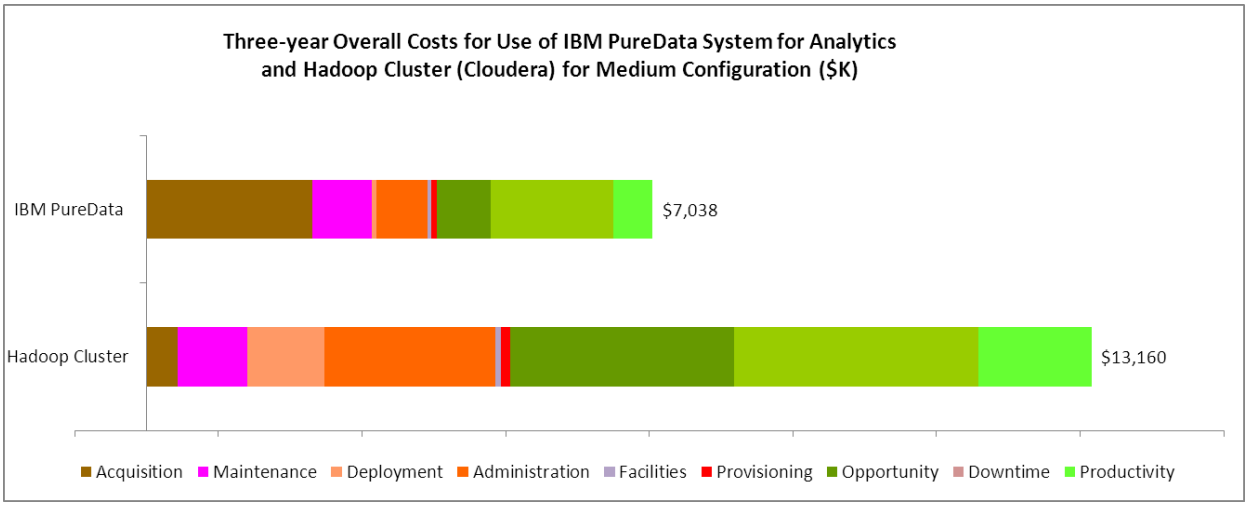

Figure 3: IBM PureData System for Analytics Lowers TCO by over 46% for Medium Configurations

46% lower TCO with IBM PureData System for Analytics on medium configurations

Again, the acquisition and maintenance costs for a Hadoop cluster are lower than those of PDA. But PDA has lower deployment and administration costs, making the total IT Costs for PDA lower than Hadoop. When Business Costs are included, the TCO of the IBM PureData System for Analytics is much lower (by 46%) than the Hadoop cluster.

29% lower TCO with IBM PureData System for Analytics on large configurations

Large Configuration (780 TB): For the large configuration (Figure 4), the IT Costs for the Hadoop cluster and the IBM PureData System are about the same. But when Business Costs are included, the Hadoop cluster is 142% more expensive.

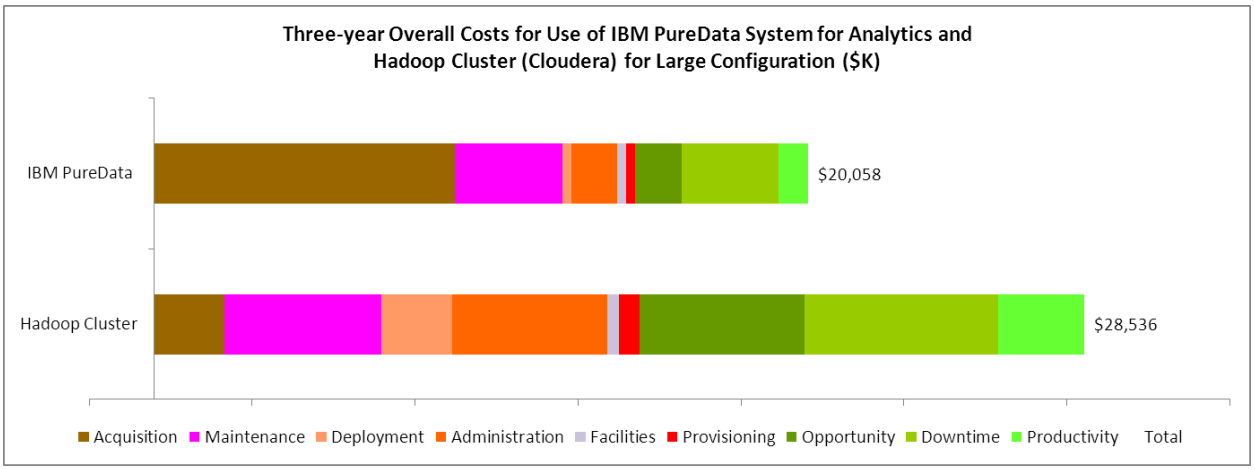

Figure 4: IBM PureData System for Analytics Lowers TCO by over 29% for Large Configurations

Larger deployment and administration costs with Hadoop

The acquisition costs for a Hadoop cluster are much lower than those of PDA. But PDA has lower maintenance, deployment and administration costs, making the total IT Costs for PDA about the same as Hadoop. When Business Costs are included, the TCO of the IBM PureData System for Analytics is lower (by 29%) than the Hadoop cluster.

Enterprise Configuration (1500 TB or 1.5PB): For the Enterprise configuration (Figure 5), the IT Costs for the Hadoop cluster 88% less expensive compared with the IBM PureData System, but is 129% more expensive when Business Costs are included.

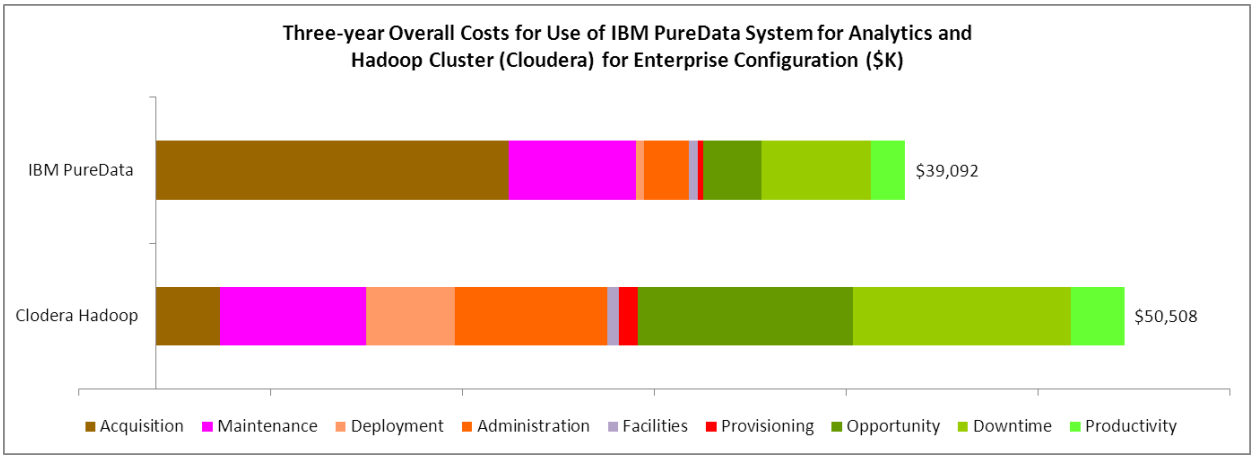

Figure 5: IBM PureData System for Analytics Lowers TCO by over 22% for Enterprise Configurations

22% lower TCO with IBM PureData System for Analytics on Enterprise configurations

The acquisition costs for a Hadoop cluster are significantly lower than those of PDA. But PDA has lower deployment and administration costs, making the total IT Costs for Hadoop cluster 88% less than PDA. But when Business Costs are included, the TCO of the IBM PureData System for Analytics is lower (by 22%) than the Hadoop cluster.

TCO for IBM PureData System better than Hadoop for all cases

All Configurations: Figure 6 summarizes the cost-benefit analysis for the entire range of configurations. The TCO advantages of the IBM PureData System are clear in all cases and are more pronounced for the small and medium configurations. For the large and enterprise configurations, acquisition costs for Hadoop clusters are much lower, but deployment and administration costs are significantly larger, primarily because of scarce highly-skilled resources which also add to other deployment risks/costs not considered in this analysis.

Costs ($K)	Small		Medium		Large		Enterprise	
	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Acquisition Costs	$170	$41	$2,302	$432	$9,208	$2,142	$18,438	$3,375
Maintenance Costs	$61	$91	$829	$974	$3,315	$4,829	$6,638	$7,610
Deployment Costs	$30	$320	$77	$1,066	$287	$2,167	$417	$4,605
Administration Costs	$581	$795	$700	$2,388	$1,400	$4,770	$2,331	$7,948
Facilities Costs	$34	$7	$59	$80	$267	$346	$475	$613
Provisioning Costs	$5	$12	$69	$130	$276	$643	$277	$1,013
Total IT Costs	$882	$1,266	$4,036	$5,069	$14,753	$14,897	$28,576	$25,162
Opportunity Costs	$190	$777	$760	$3,109	$1,425	$5,052	$3,040	$11,193
Downtime Costs	$568	$1,135	$1,703	$3,406	$2,980	$5,960	$5,676	$11,353
Productivity Costs	$300	$666	$540	$1,576	$900	$2,626	$1,799	$2,800
Total Business Costs	$1,057	$2,578	$3,003	$8,091	$5,305	$13,639	$10,516	$25,345
Total Cost of Ownership (TCO)	**$1,939**	**$3,844**	**$7,038**	**$13,160**	**$20,058**	**$28,536**	**$39,092**	**$50,508**

Figure 6: Better TCO of the IBM PureData System for Analytics over a Hadoop Cluster for All Cases

IBM PureData System improves time-to-value and productivity while lowering downtime

The Business Costs associated with the IBM PureData System included in this analysis are significantly lower because of the unique appliance-style packaging and unique bundled software that improves time-to-value, staff productivity, and lowers downtime.

Other key business advantages of the PureData System not considered in this analysis include improved staff collaboration, and flexibility to pursue more innovative strategic initiatives.

Conclusions and Recommendations

With the ever increasing volume, velocity and variety of data, the boundaries between relational and non-relational database systems continue to blur. SQL continues to be the preferred method to work with data because of its widespread use, stability and compatibility with most enterprise data management solutions.

PureData System is optimized for performance and integrates easily

Parallel systems such as the IBM PureData System for Analytics (PDA) appliance are:

- Optimized for performance and scale, delivering very fast query performance on analytic workloads and many smart benefits including speed, simplicity and scalability
- Delivered ready-to-go for immediate data loading and query execution
- Integrated with leading Extract, Transform, and Load (ETL), business intelligence, and analytic applications through standard interfaces.

While the acquisition costs of Hadoop on commodity clusters are typically lower especially for large and enterprise systems, the PDA appliance family is faster to deploy and requires minimal ongoing administration or tuning. This allows customers to lower their IT Costs of

IBM PureData System TCO advantages over Hadoop are compelling for all cases

deployment and administration compared to Hadoop clusters that require considerable highly-skilled resources for deployment and continuing operations. More importantly, when the Business Costs of lost opportunity, lower productivity and more downtime are included, the TCO advantages of the IBM PureData System for Analytics are very compelling.

Compared to a Hadoop cluster, clients implementing Analytics in an SQL environment with the IBM PureData System for Analytics can **lower the TCO for all configurations** (small – medium – large – enterprise) even with favorable assumptions for Hadoop (Figure 7). In addition, PDA IT Costs are lower than a Hadoop Cluster in small to medium configurations.

For large and enterprise configurations, clients only concerned with IT Costs can deploy a hybrid solution

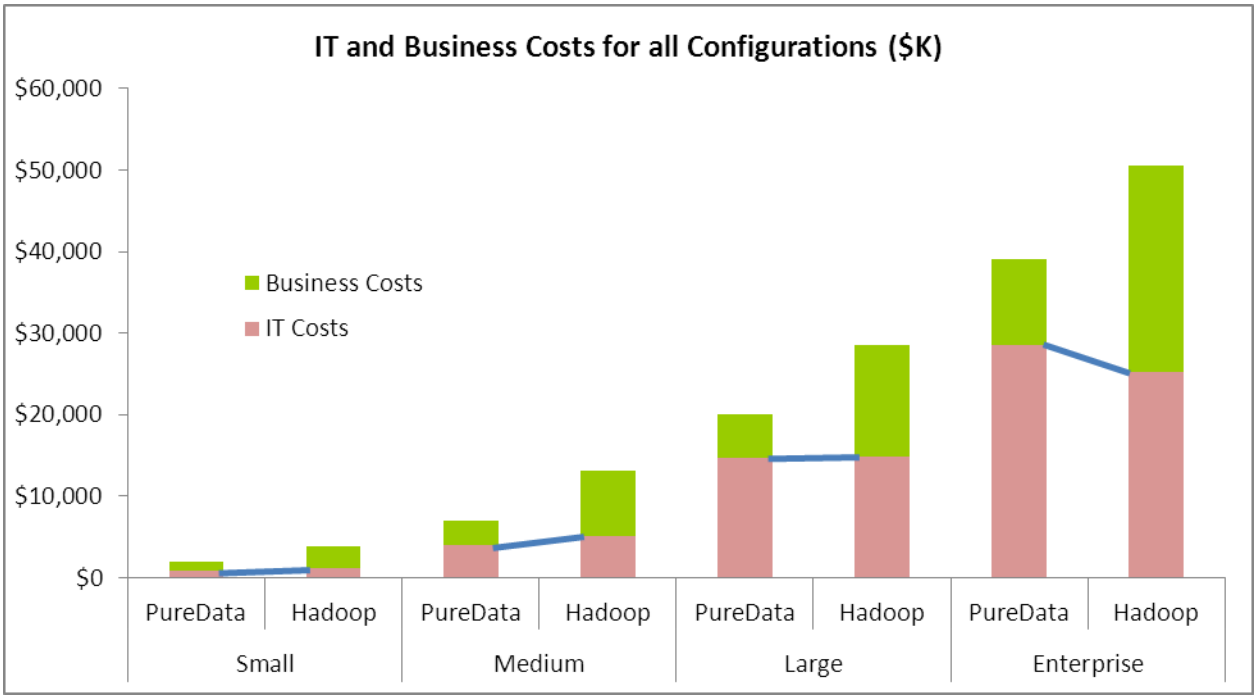

Figure 7: Lower TCO with IBM PureData System for All Configurations. IT Costs Cross Over in Favor of Hadoop for Large and Enterprise Configurations

PDA delivers faster time to value; more profits and revenues at lower TCO than Hadoop

Even for large to enterprise configurations, where IT Costs for Hadoop cross over (at large configurations) and become lower than PDA, clients who may be concerned with IT and Acquisition costs can implement a hybrid solution of a medium-sized IBM PureData System and a Hadoop cluster. By keeping the frequently used (hot) data on the IBM PureData System and the seldom used data in the Hadoop cluster, the advantages of both systems can be realized.

Clients, who choose the IBM PureData for Analytics over a Hadoop cluster, can focus on their business without the hassles of managing technology complexity and concerns related to security and data protection. This enables them to benefit from faster time to value, higher revenues and profits, better product/service quality and potentially more innovation.

Cabot Partners is a collaborative consultancy and an independent IT analyst firm. We specialize in advising technology companies and their clients on how to build and grow a customer base, how to achieve desired revenue and profitability results, and how to make effective use of emerging technologies including HPC, Cloud Computing, and Analytics. To find out more, please go to www.cabotpartners.com.

Appendix: Cost Benefit Analysis Methodology and Assumptions

Comprehensive Cost-Benefit Framework for IT Investments

Cabot Partners uses a comprehensive cost-benefit analysis framework to help clients evaluate their IT investments objectively. The total cost of ownership (TCO) over several years is typically computed. This holistic framework helps to justify investment decisions, improves the IT organization's effectiveness, and deepens collaboration between Business and IT.

Several inter-related cost and value drivers that typically contribute to the TCO include:

Value. The value derived from an IT investment could come in:

- *Strategic value:* faster time to market, increased profits, improved brand equity, better partnerships with stakeholders, ability to attract and retain top talent
- *Products/Services value:* more innovation, better collaboration, greater insights, improved quality
- *Operational value:* faster time to results, reduced cost of development, more users supported, improved user productivity, better capacity planning
- *IT value:* improved system management, administration, and provisioning, enhanced security, higher utilization, scalability, reduced downtime, access to robust proven technology and infrastructure management expertise

Costs. Costs from a range of sources typically include:

- *Data center capital*: new servers, storage, networks, power distribution units, chillers, software purchase, etc.
- *Data center facilities:* land, buildings, containers, facilities maintenance, etc.
- *Operational costs:* labor (salaries for end-users and IT staff), energy, IT hardware and software maintenance, software license, etc.
- *Other costs:* deployment and training, downtime, bandwidth, etc.

Businesses must continually evaluate the TCO of their IT investments within this broad cost−benefit framework.

Customized TCO Analysis for IBM PureData System for Analytics versus Hadoop

Cabot Partners customized this comprehensive cost-benefit analysis framework to include the key cost and value drivers relevant to this TCO Study. In addition, as part of the discovery process that included literature research, several experts and end-users were interviewed.

This **Cost-Benefit Analysis** compares the *Total Cost (IT Cost + Business Cost) for Three Years* of IBM PureData for Analytics and a Hadoop Cluster (Cloudera) for four configurations – small, medium, large and enterprise – with data requirements of 18 TB, 192 TB, 750 TB and 1500 TB respectively.

Six **IT Costs** are considered:

1. *Acquisition* – Cost of software, servers, storage, networks, etc. Included in Year 1.
2. *Maintenance* – Cost for maintenance and support of the environment. Annual.
3. *Deployment* – Time-related costs to deploy the application (assume application is already developed). One time initial cost.
4. *Administration* – Full Time Equivalent (FTE) people costs for operations. Annual.
5. *Facilities* – Costs for buildings, energy and cooling. Annual.
6. *Provisioning* – Cost of extra system capacity to handle peaks and/or failures. Included in Year 1.

Three **Business Costs** considered include:

1. *Opportunity* – Business lost while the application is being configured and deployed. One time initial cost.
2. *Downtime* – Lost business during the times when the system is down. Annual.
3. *Productivity* – Lost productivity of resources (data scientists, engineers, administration because of slow execution speed and job failures). Annual.

IT Costs

Acquisition Costs

- IBM PureData is an integrated software, server, storage and network appliance with a list price. While discounts are offered on this price, only the list price is used for this study. The prices for the 4 configurations, provided by IBM for this study, are in Table 1.
- Hadoop Cluster requires the user to purchase servers, disks and network and use the open source Hadoop software. Typically, maintenance is purchased from companies such as Cloudera, Horton Networks, etc. Costing assumptions[1] for this study are based on a Cloudera Hadoop solution. The cluster size is based on nodes with 2 TB of Hadoop File System (HDFS) capacities. The server node assumes four 2 TB hard disk drives, 24 GB memory, and 12 CPU cores. The hardware costs of $4,500 per node were based on retail server hardware vendor prices. Each rack can hold up to 36 servers with a top-of-rack switch. Multiple factors could change these pricing assumptions: a different hardware configuration, a volume discount on purchase, or regional and seasonal price offers.

Acquisition Cost	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Storage (TB)	18	18	192	192	750	750	1500	1500
Num of Racks in PureData	0.5		1		4		8	
Num of Nodes in Hadoop		9		96		375		750
Cost of PureData ($K)	$170		$2,302		$9,208		$18,438	
Cost of Node in Hadoop ($K)	0	$4.50		$4.50		$4.50		$4.50

Table 1: Acquisition Costs Detail for IBM PureData System and Hadoop Cluster – One Time

Maintenance Costs

- IBM PureData comes with an 18% maintenance fee for 24 x 7 support, updates and installation support from the second year onwards. The maintenance for the first year is included in the acquisition price of the appliance.
- Hadoop Cluster retail pricing assumed for Cloudera Enterprise Core is $3,382 per node per year with 24/7 support. Note that retail pricing is typically not shared by the vendor. This price[1] was also validated by our subject matter experts during our discovery process.

Maintenance Costs	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Maintenance for Appliance	18%		18%		18%		18%	
Maintenance for Each Node		$3,382		$3,382		$3,382		$3,382

Table 2: Maintenance Costs Detail for IBM Pure Data System and Hadoop Cluster - Annual

Deployment Costs

Deployment times and resources vary considerably depending on size and scale. It is critical to ensure comparisons are consistent and objective.

- IBM PureData estimated deployment times and resources were based[2] on many deployments of the PureData System across several industries. The average deployment time and resource used – corresponding to a medium configuration – was reported to be 21.7 days and 1.47 Full Time Equivalent (FTE) respectively. This data was prorated (Table 3) using storage size for all configurations. Estimated deployment costs used a fully-burdened rate of $300/hr.
- Hadoop Cluster estimated deployment times and resources were based on many Hadoop deployments[3, 4,5] across several industries. The average deployment time and resource (development and deployment) used – corresponding to a medium configuration – was reported to be 98.7 days and 11.68 FTE respectively. The resource data was halved to 5.84 FTE since only deployment resources are considered. This data was then conservatively prorated (Table 3) in favor of Hadoop using storage size for all configurations. Estimated deployment costs used a fully-burdened rate of $300/hr.

Deployment	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Deployment FTE	1.17	3.00	1.47	5.00	3.67	7.82	4.00	12.00
Deployment Days	10.86	44.42	21.71	88.83	32.57	115.48	43.43	159.89
Cost of Deployment FTE / hr	$300.0	$300.0	$300.0	$300.0	$300.0	$300.0	$300.0	$300.0

Table 3: Deployment Costs Detail for IBM PureData System and Hadoop Cluster – One Time

Administration Costs

- IBM PureData, being an appliance, typically requires much less administration time. The estimated administration resources were based[2] on many deployments of the PureData

System across several industries. The average resource used – corresponding to a medium configuration – was reported to be 1.375 FTE. This data was prorated (Table 4) using storage size for all configurations.

- Hadoop Cluster requires several additional operational tasks to handle the inherent complexity of distributed systems manually. A Hadoop cluster
  - should be deployed on carefully selected hardware and tuned with appropriate configuration parameters
  - requires cluster health monitoring and administrative intervention for failure recovery and repair
  - must be configured and controlled using job schedulers to optimize utilization
  - should be returned if workloads change significantly
  - needs arduous maintenance to integrate constantly changing Hadoop versions with new tools in the ecosystem.

  Estimated administrative resources were based on many Hadoop deployments[3,4,5] across several industries. The average administrative resource used – corresponding to a medium configuration – was reported to be 5.63 FTE. This data was then prorated (Table 4) highly in favor of Hadoop using storage size for all configurations.

Administration	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Number of FTEs	1.03	1.41	1.24	4.23	2.48	8.45	4.13	14.08
Cost of FTE/ hour	$97.94	$97.94	$97.94	$97.94	$97.94	$97.94	$97.94	$97.94

Table 4: Administrative Costs Detail for IBM PureData System and Hadoop Cluster – Annual

Facilities Costs

- IBM PureData facilities costs[6] (Table 5) consist of cost of power, cooling and floor-space. Electricity price was assumed to be $0.09 KW/hour.
- Hadoop Cluster facilities costs[1] assume 36 servers are packed in one standard 19 inch rack and each Hadoop node has 4 servers; yielding a conservative estimate.

Facilities Costs	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Power (WH)	2,300	1,500	7,600	16,500	34,200	72,990	60,800	129,480
Cooling (BTU)	8,000	5,400	27,000	60,000	121,500	250,000	216,000	440,000
Space ($Sq Feet*Cost)	$7,581	$1,457	$7,581	$16,027	$34,116	$68,479	$60,650	120,931
Cost of Power ($/KWH)	$0.09	$0.09	$0.09	$0.09	$0.09	$0.09	$0.09	$0.09
Cost of Space (per node)	$1,000	$1,469	$1,000	$1,469	$1,000	$1,469	$1,000	$1,469

Table 5: Facilities Costs Detail for IBM PureData System and Hadoop Cluster – Annual

Provisioning Costs

- IBM PureData provisioning costs are very small. Client feedback indicates that the system scales very easily without any significant need for additional reserve resources.
- Hadoop Cluster uses commodity servers and storage. Failures are quite common and it is recommended that provision be made to scale with demand. While recommendations are 20% over provisioning[7], a conservative 10% number was used in all configurations.

Provisioning	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Provisioning (of base)	1%	10%	1%	10%	1%	10%	1%	10%

Table 6: Provisioning Costs Detail for IBM PureData System and Hadoop Cluster – Annual

Business Costs

Strategic, operations and tactical business costs related time-to-market delays, systems-downtime and productivity loss of staff are considered. The value of innovation, while real, is not included. IBM PureData and Hadoop cluster comparisons were consistent and objective.

Opportunity Costs

The business lost[2] while the application is configured and deployed – averaged over several industries – is estimated to be $35, 000/day for an organization using a medium-sized system. Table 7 depicts the opportunity cost data prorated for each configuration for both systems.

Opportunity Costs	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Number of Days of Delay	10.86	44.42	21.71	88.83	18.00	115.48	43.43	159.89
Cost of Delay / Day	$17,500	$17,500	$35,000	$35,000	$43,750	$43,750	$52,500	$52,500

Table 7: Opportunity Costs Detail for IBM PureData System and Hadoop Cluster – One Time

Downtime Costs

The cost of downtime averaged across several industries with at least 1000 employees is estimated to be $2,160,000 per hour[8] based on lost business. Only a small fraction of this was prorated for all configurations (Table 8).

For Hadoop, incidents causing downtime included single name node failure, interrupted transfer, corrupted loads, and entire cluster failure. The time to fix varied from 15 minutes to 4 hours with an average downtime of 35 hours/year (99.60% uptime). To be favorable to Hadoop, this was assumed to correspond to the enterprise configuration. This downtime was then prorated for other Hadoop configurations. Because of better reliability, availability and serviceability (RAS) of PDA, the downtime for the enterprise configuration was assumed to be 33% of the equivalent Hadoop cluster.

Downtime Costs	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
System Uptime	99.90%	99.80%	99.88%	99.76%	99.85%	99.70%	99.80%	99.60%
Cost of Downtime / hour	$21,600	$54,000	$75,600	$75,600	$75,600	$75,600	$108,000	$108,000

Table 8: Downtime Costs Detail for IBM PureData System and Hadoop Cluster – Annual

Productivity Costs

This includes lost productivity of staff resources (data scientists, engineers and operators) because of slower execution speed and lower system reliability. For a large Hadoop cluster, the productivity loss[9] was 27% for Data scientists (Resource 1, eight in number and more expensive) and 30% for Engineers and Operators (Resource 2, four in number). To be favorable to Hadoop, this productivity loss was assigned to the enterprise configuration. Productivity loss for other Hadoop configurations were prorated (Table 9). Productivity loss and number of resources were conservatively lowered and prorated for PDA, accounting for its smart productivity enhancing features of simplicity, scalability and speed.

Productivity Costs	Small		Medium		Large		Enterprise	
Assumptions	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop	PureData	Hadoop
Number of Resource 1	3.0	4.0	4.5	6.0	6.0	8.0	9.0	12.0
Number of Resource 2	1.5	2.0	2.3	3.0	3.0	4.0	4.5	6.0
Prodcutivity Loss Resource 1	8.10%	13.50%	9.72%	16.20%	12.15%	20.25%	16.20%	27.00%
Productivity Loss Respurce 2	9.00%	15.00%	10.80%	18.00%	13.50%	22.50%	18.00%	30.00%
Cost of Resource 1 / hour	$146.91	$146.91	$146.91	$146.91	$146.91	$146.91	$146.91	$146.91
Cost of Resource 2 / hour	$120.79	$120.79	$120.79	$120.79	$120.79	$120.79	$120.79	$120.79

Table 9: Productivity Costs Detail for IBM PureData System and Hadoop Cluster – Annual

[1] Accenture Technology Labs " Hadoop Deployment Comparison Study" Price-Performance comparison between a bare-metal Hadoop Cluster and Hadoop-as-a-service, http://www.accenture.com/sitecollectiondocuments/pdf/accenture-hadoop-deployment-comparison-study.pdf

[2] ITG Paper "Cost/Benefit case for IBM Puredata system for Analytics" Comparing costs and time to value with Teradata Data Warehouse Appliance, May 2014, http://www.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=SWGE_WA_UZ_USEN&htmlfid=WAL12377USEN&attachment=WAL12377USEN.PDF#loaded

[3] ITG paper "Business case for Enterprise Big Data Deployments" Comparing costs, benefits, and risks for use of IBM InfoSphere BigInsights and Open Source Apache Hadoop", 2013, http://www.habber.com/es/files/2014/08/Business-Case-for-Enterprise-Big-Data-Deployments-2_comparativa-de-costes-beneficios-y-riesgos-del-uso-de-IBM-InfoSphere-BigInsights-y-Apache-Hadoop-open.pdf

[4] Treasure Data Big Data as Service "Delivering Real-World Total Cost of Ownership and Operational Benefits", http://radiantadvisors.com/whitepapers/big-data-total-cost-of-ownership-understanding-hard-costs-and-options/

[5] Forrester Paper "The Total Economic Impact of Altiscale Hadoop-as-a-Service" Cost Savings and Business Benefits enabled by Hadoop as a Service, 2015, https://www.altiscale.com/hadoop-resource/economic-impact-of-hadoop-as-a-service/

[6] IBM Pure Systems - Data Sheet – IBM PureData System for Analytics – N3001-001 Powered by Netezza Technology – www.**ibm.com**/PureSystems/PureData

[7] http://hortonworks.com/blog/best-practices-for-selecting-apache-hadoop-hardware/

[8] http://basho.com/infographic-down-with-downtime/

[9] WanDisco Technical Brief "Service Continuity with Non-Stop Hadoop" , https://www.wandisco.com/system/files/documentation/Technical-Brief-Service-Continuity-NonStop_Hadoop-WEB.pdf