Analysis of IBM PureData System with Hadoop Implementations for Structured Analytics

Rohan Ninan Jacob[1], Prathamesh Girish Shirwadkar[2] , Mohini Ajay Singh[3]

[1]*Department of Electronics and Telecommunication Engineering, Fr. Conceicao Rodrigues Institute of Technology, Vashi, Maharashtra, India*

[2,3]*Department of Electronics Engineering, Vidyalankar Institute of Technology, Wadala, Maharashtra, India*

---***---

***Abstract**: Increase in the IT cost including cost of labor,energy and facilities have become the common tasks for usage of Hadoop clusters at many organizations. Moreover, data security and lack of skilled resources are major issues. To meet such challenges, organizations should implement a cost-effective, convenient, high-performance, efficient and reliable solution to deliver the best business outcomes. This is the goal of the IBM PureData System for Analytics (PDA). The three year TCO (IT Costs + Business Costs) analysis 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.*

Keywords: PDA, Hadoop, Small, Medium, Large

I. INTRODUCTION

Data can be structured or unstructured. 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[1]. 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)[1]. An estimated 85% of data is unstructured. 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.

Clients who may be concerned solely with the IT Costs can implement a hybrid solution of a medium-sized IBM PureData System with 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 helps inconsistently lowering TCO. Consequently, clients benefit from faster time to value, higher revenues and profits, better product/service quality and potentially more innovation.

II. OVERVIEW

The ever growing and pace of technology-enabled business transformation and innovations, which has opened and developed new ways of approach. Several fast-growing 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 [2].

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 and implemented 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[3].

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:[3]:-

a)-Widespread use with millions of well-trained users.

b)-Stability with relational database management systems supporting SQL compatibility, transactional consistency, and enforced schema required by enterprises.

c)-Optimized for performance and scale with distributed/parallel systems and in-memory computing.

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

a)- Seamless growth – Scaling capacity or performance is fast and painless; often triggered with a click or by a simple command.

b)- Schema flexibility – As applications mature, schema changes can be made without taking the system down.

c)- 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.

III. COMPARING IBM PureData for ANALYTICS WITH HADOOP

IBM PureData System for Analytics (PDA) with several smart features is designed 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[5]. The integration of processors, software, and storage leads to shorter application development cycles and exceptional time to value for business analytics[6] 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[4] 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

IV. LOWER TOTAL COST OF OWNERSHIP WITH THE IBM PureData SYSTEM

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 (Fig1), we find that the IT Costs of the Hadoop cluster are 144% more than PDA. When Business Costs are included, Hadoop is 198% more expensive.

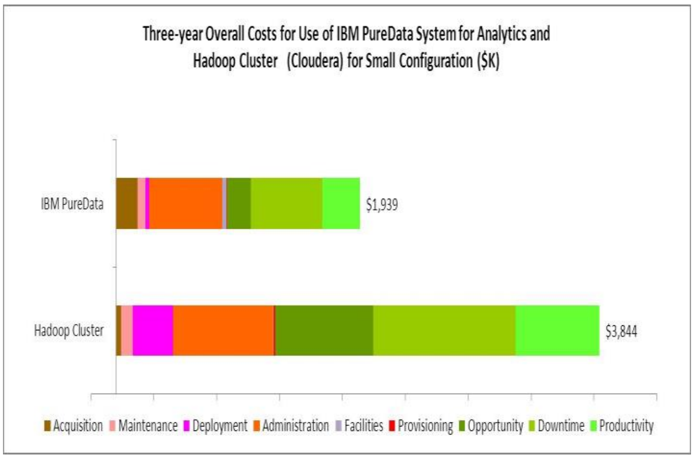

Fig1: IBM PureData System for Analytics Lowers TCO by over 49% for Small Configurations

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[7] 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.

Medium Configuration (192 TB): For a medium configuration (Fig2), the Hadoop cluster is 126% more expensive than the IBM PureData System in IT Costs[7], and is 187% more expensive when Business Costs are also added.

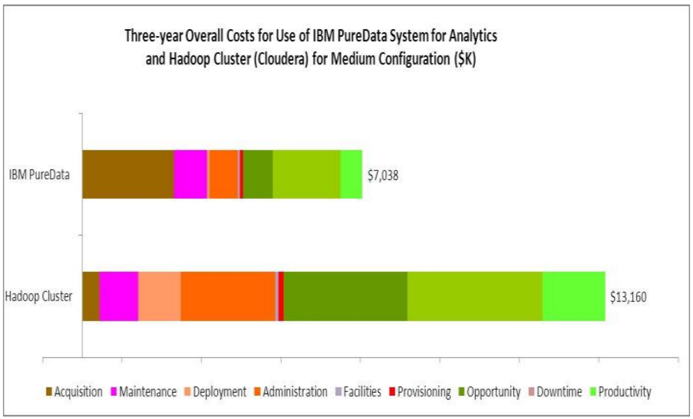

Fig 2:- IBM PureData System for Analytics Lowers TCO by over 46% for Medium Configurations

Large Configuration (780 TB): For the large configuration (Fig3), the IT Costs for the Hadoop cluster and the IBM PureData System are about the same[8]. But when Business Costs are included, the Hadoop cluster is 142% more expensive.

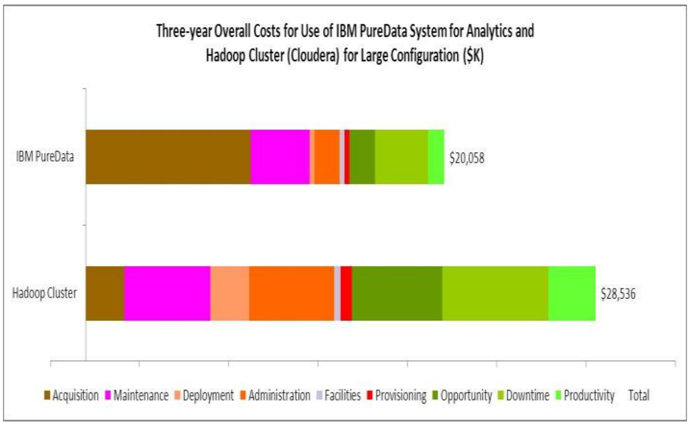

Figure 3: IBM PureData System for Analytics Lowers TCO by over 46% for Medium Configurations

Enterprise Configuration (1500 TB or 1.5PB): For the Enterprise configuration (Fig4) the IT Costs for the Hadoop cluster[7] 88% less expensive compared with the IBM PureData System, but is 129% more expensive when Business Costs are included.

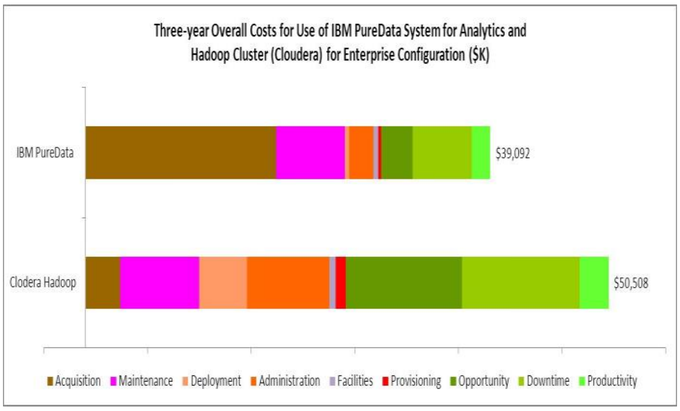

Fig 4:- IBM PureData System for Analytics Lowers TCO by over 22% for Enterprise Configurations

The acquisition costs for a Hadoop cluster are significantly lower than those of PDA[9]. 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%) compared to the Hadoop cluster.

IV. CONCLUSION

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).

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.

ACKNOWLEDGEMENT

We would also like to thank our colleagues for useful and beneficial criticism, which helped us to present the paper.

Finally, we would like to express our humble gratitude to the different online communities for sharing their repositories and documentation related to this domain of analytics and the differences which we represented.

REFERENCES

[1] Forrester Paper “The Total Economic Impact of Altiscale Hadoop-as-a-Service” Cost Savings and Business Benefits enabled by Hadoop as a Service, 2015

[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.

[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,

[4]Treasure Data Big Data as Service Delivering Real-World Total Cost of Ownership and Operational Benefits,2013

[5] Accenture Technology Labs Hadoop Deployment Comparison Study Price-Performance comparison between a bare-metal Hadoop Cluster and 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