EDISON Group

902 Broadway, 7th Floor

New York, NY 10010

www.TheEdison.com

@EdisonGroupInc

212.367.7400

Lower the Cost of Analytics with Dell EMC's On-Premises Solutions

Comparing Total Cost of Ownership of Dell EMC Ready Solutions for Big Data vs. an AWS Big Data Solution

Printed in the United States of America

This report was commissioned by Dell EMC and Intel.

The information contained in this document is based on Dell EMC and Intel provided materials as well as independent research and was aggregated and validated for Edison Group, Inc. by the Edison Group Analyst team.

First Publication: September 2018

Produced by: Harold Kreitzman, VP, Strategic Advisory Services

EDISON

Table of Contents

Executive Summary

Edison Group was commissioned to create a Total Cost of Ownership (TCO) model that compared the costs of Dell EMC Ready Solutions for Big Data vs. an equivalent service build on Amazon Web Services (AWS) over a three-year period.

Dell EMC delivers an equivalent Big Data as a Service infrastructure for less than 50% the cost of the AWS service over a three year period.

Our methodology, is intended to validate the costs, configurations and assumptions for both the on-premises solution and the public cloud service. We do this to ensure that the final TCO analysis is accurate and reflective of what one would see in a customer environment.

Our approach was to create an infrastructure to support the collection and analysis of data. The two choices compared in this paper are an on-premises solution using the Dell EMC Ready Solutions for Big Data vs a comparable Amazon Web Services solution both covering two regions.

Service and pricing components consisted of (where applicable) compute, storage, database, network, software, services and facilities and were based on sources from Amazon and Dell EMC that were independently verified by Edison Group.

The result of the analysis showed that Dell EMC Ready Solutions for Big Data delivers an equivalent solution for less than 50% the cost of AWS.

Business Premise

As customers embark on their data analytics journey, it is often focused on a single use case which is based on a single, optimized analytics stack. If this program is successful, customers scale up, while also scaling out with numerous pilot programs for various other business challenges. These pilot programs are built on their own custom stacks, with tooling and hardware optimized for the specific needs of that use case. This process repeats with hundreds of clusters being spun up and independently managed by IT for dozens of unique projects. This 'cluster sprawl' is costly and creates immense technical debt and friction, which slows down the implementation of new programs aimed at creating business value.

This program cannot simply be consolidated away. Analytics use-cases demand varied and meaningful tooling and hardware stacks. For example:

- Marketing wants to do omni-channel marketing analytics capability that requires 360° customer visibility. This requires a stack that utilizes Hadoop plus Scoop, Pig and Hive to understand customer buying patterns with hundreds of terabytes or petabytes of data.
- Research and Development wants to do risk assessment on the viability of new programs earlier in the development cycle. This requires integrating Kafka, Spark, Hadoop and NoSQL from multiple lab sites with terabytes of data.
- Manufacturing wants to do predicative maintenance on their hard assets, reducing downtime and improving machine productivity. This requires real-time, low-latency data collection hardware, utilizing Hadoop, NoSQL, Kafka and Spark with hundreds of terabytes of streaming data.
- Finance wants to do real-time fraud detection, preventing fraud from taking place, lessening the need for costly claims investigation agents and payouts. This requires a high-performance computing engine utilizing machine learning, Spark and in-memory GPU data base with access terabytes to petabytes of historical payment and fraud data.

In each of these scenarios, customers would have to maintain multiple unique environments with their own set of dedicated hardware, software, data movement and IT support teams. This is not a cost-effective nor sustainable solution for any customer. Additionally, each of these deployments typically require multiple, separate environments (e.g. Dev, Test, QA, Staging, Production) – and each data science/ analyst team needs different tools that typically take IT 6-8 weeks to get set up.

This backend complexity is in direct conflict with the stated vision of becoming more agile and working in real-time and as some customers may see it, the solution would be to move these workloads to the cloud. Cloud Service Providers allow customers to abstract all of this complexity and achieve similar levels of agility with a simple monthly price. With a growing suite of analytics tools with low upfront investment, Amazon Web Services® (AWS) is an enticing platform to analysts frustrated with their in-house IT teams.

However, customers with sizeable data footprints on-premises should be aware that moving to public cloud may not offer the savings they expect. With the cost of moving data back and forth, security and governance concerns, and the dangers of vendor lock-in, many customers end up paying more over time. Also, if the workload is high demand (requires over 50% utilization per day with substantial amounts of data) the cloud can become cost prohibitive. On the performance side, analytics workloads with low latency requirements suffer significantly in public cloud environments with generic hardware stacks which aren't optimized for analytics. Where speed, timeliness and a large amount of data manipulation are required for medium to large enterprises, the cloud can become even more expensive.

Customers with sizeable data footprints on-premises should be aware that moving to public cloud may not offer the savings they expect.

The example use case we've chosen for this analysis is for a NoSQL database. These have become popular due to the limitations of relational databases to support the growing need for fast, effective access and manipulation of real-time data over large geographical areas. Built-in data synchronization and continuous availability are significant factors as well. While there are many different alternatives to NoSQL including MongoDB and Elasticsearch, two of the most popular are Cassandra (open-source) and DynamoDB (Amazon, proprietary). Regardless, the use of NoSQL in AWS across multiple regions, along with data storage and transfer considerations, can have a significant impact on the operating cost of the public cloud.

The following sections detail the analysis that was performed to compare the on-premises and public cloud approaches to this use case.

Big Data Infrastructure and Service Options Reviewed

The table listed below contrasts and compares configurations and components between Dell EMC Ready Solutions for Big Data and Amazon Big Data services.

Configurations

All efforts were made to create an apples-to-apples comparison. However, there are some inherent differences resulting from comparing a cloud to an on-premises solution.

The Amazon service consists of compute (EC2 – workload and GPU) and storage (S3) to run workloads, NoSQL database (DynamoDB – which contains its own compute and storage resources to support databases in two regions), Big Data software (Elastic MapReduce), network enhancement and management (Direct Connect), network (ISP data pipe) and support.

Dell EMC's on-premises solution consists of compute (workload and GPU) and storage (500TB) to run workloads and database, NoSQL database (Cassandra), Big Data as a Service software (BlueData EPIC), networking, facilities and support. The Dell EMC solution is replicated to support two regions.

Server/Network Infrastructure

The sizes and number of units for both solutions were picked to meet the business need for a mid-to-enterprise level Big Data solution running three significant workloads. Independent tests have validated that two vCPUs are equal to each physical core for AWS EC2[1].

Dell EMC - CPU/Server	**Units**	**Cores**	**Physical Cores**
Worker/Controller Node (PowerEdge R740XD Server)	10	36	360
GPU (PowerEdge R740XD Server)	1	24	24
AWS - CPU/Server	**Units**	**Cores**	**Physical Core Equivalent**
Memory Optimized (r4.8xlarge)	22	32	352
GPU (p2.xlarge)	12	4	24

Table 1 – Server Configuration Comparison

This category includes Dell EMC PowerEdge Servers and Dell EMC network switches as part of the physical infrastructure. PowerEdge R740XD servers are powered by new Intel® Xeon®

[1] EC2 vCPU-s vs. Real Cores, http://forum.paradigm4.com/t/ec2-vcpu-s-vs-real-cores/1212

Scalable processors. These processors are much more powerful than the previous Intel® Xeon® processors, an enterprise staple for nearly two decades.

According to Intel funded research, these new processors can deliver significant economic advantages for IT departments that are looking to replace older servers or where analytics become a key business requirement. Intel claims that IT can replace four, four-to-five-year-old servers with only one Intel Xeon Scalable processor based server, lowering four-year total cost of ownership (TCO) by up to 65 percent[2].

Dell EMC is committed to Intel because of their dedication to delivering unmatched enterprise-ready platforms. In this case, Intel® Xeon® Scalable processors are designed to support advanced analytics for business transformation which customers need to remain competitive and increase their market share.

Software

Software categories include the operating system, file, data, data stream and cloud management software. It also includes Big Data as a Service software from BlueData to provide a coherent and complete experience. While AWS has some components included, others are offered as separate services and still other software pieces need to be obtained from the outside.

Dell EMC/BlueData - Software	AWS - Software
CentOS / Red Hat Enterprise Linux 6.9 or 7.4 (physical server or virtual machine)	Amazon Linux AMI
BlueData EPIC® Software	Not Included
CDH (with Cloudera Manager)	Not Included
HDP (with Apache Ambari)	Not Included
MapR (with MapR Control System)	AWS EMR
Apache Spark	AWS EMR
Apache Kafka	Not Included, but open source available

Database

NoSQL database is one of the main components to meet Big Data business requirements in this analysis.

[2] https://www.intel.com/content/www/us/en/analytics/future-ready-analytics-brief.html

Significant differentiators between Apache Cassandra and Amazon DynamoDB include what types of NoSQL constructs they support and their popularity rankings[3].

Alternative	Dell EMC/BlueData - Database	AWS - Database
Database	Apache Cassandra, Open Source	Amazon DynamoDB
Primary Database Model	Wide column store	Document store Key-Value Store
DB –Engine Ranking	#10 Overall #1 Wide column store	#21 Overall #2 Document store #2 Key-value stores

Data and Data Movement

Dell EMC Ready Solutions for Big Data does not require any additional costs to support data storage and movement. AWS has a service model and charges for both data storage and data movement, both from and to external stores.

Network Pipe

While the AWS solution provides internal networking hardware and capabilities, the customer still needs to obtain a data pipe from an ISP. A 10GB data pipe is required to support data transfers to and from the cloud. In addition, AWS Direct Connect Service is also needed.

Dell EMC on-premises solution requires a 10GB data pipe for each region.

Dell EMC/BlueData - Network	AWS - Network
(2) 10GB Data Pipes from ISP, one for each region.	(1) 10GB Data Pipe from ISP.

Services

Both solutions offer support services. Hardware, software and network maintenance are included in the pricing for these Dell EMC infrastructure components. Dell EMC includes deployment and implementation services. Equivalent cloud support services are also available from AWS.

[3] DB-Engines, https://db-engines.com/en/system/Amazon+DynamoDB%3bCassandra

Facilities

Dell EMC on-premises solution incurs facility charges such as hardware component power, data center infrastructure power (AC and UPS), and space, which are included in the cost calculation based on reasonable estimates and market costs. AWS facility costs are built into their service costs.

TCO Summary

The following is a summary of TCO results over three years. More details are included in the appendix. All costs for the AWS solution are obtained directly from their "simple monthly calculator"[4].

Big Data Platform Components	Dell EMC		AWS		Difference $
Server/Network Infrastructure	$257,925	15.2%	$656,587	17.6%	$398,662
Software	518,400	30.5%	68,515	1.8%	(449,885)
Database	-	0.0%	2,266,330	60.9%	2,266,330
Data Transfer	-	0.0%	122,619	3.3%	122,619
Network Pipe	685,440	40.3%	468,677	12.6%	(216,763)
Services	125,000	7.3%	95,178	2.6%	(29,822)
Facility	29,103	1.7%	-	0.0%	(29,103)
Staffing	85,450	5.0%	42,725	1.1%	(42,725)
Totals	**$1,701,318**	**100.0%**	**$3,720,632**	**100.0%**	**$2,019,314**

Based on the analysis, the Dell EMC Ready Solutions for Big Data (est. $1,701,318) is less than 50% of the AWS Big Data solution (est. $3,720,632).

A quick analysis of the component differences follows:

- Server/Network Infrastructure - Dell EMC's compute costs are about 40% of AWS; these costs represent servers and network switches purchased and paid for in year one. Dell EMC hardware includes three-year maintenance agreements.
- Storage – Dell EMC and AWS storage are included in the server components.
- Software - Dell EMC Ready Solutions for Big Data software, BlueData EPIC, includes all components required to support on-premise Big Data efforts. These costs are directly related to the number of cores in the compute infrastructure.

  AWS requires inclusion of the MapR service as well as acquisition of a few open-source products like Apache Kafka.

- Database - Dell EMC's database, Cassandra, is open-source and therefore has no direct software cost. AWS DynamoDB is a proprietary database. The cost of AWS service includes compute, database and storage resources to support NoSQL across two regions. A 25TB

[4] Amazon Simple Monthly Calculator, http://calculator.s3.amazonaws.com/index.html#s=DYNAMODB

database was used to benchmark costs. AWS database costs represent over 60% of the cost of the AWS solution.

- Data Transfer – Traditional cloud models charge for data movement. Amazon is no exception. On-premises solutions do not incur data transfer charges.
- Network Pipe - On-premises Dell EMC's solution two 25GB data pipes (one for each region), while the AWS requires the acquisition of one 25GB data pipe from an ISP, and the addition of the Direct Connect service to manage network services.
- Facility - AWS's facility costs are built into their service costs. Dell EMC's solution does not, and has been included separately. Facility costs include hardware power, data center power, air conditioning and space.
- Staffing - Included for both solutions. Assumption was made that more administrative effort (twice that of AWS) would be needed to support the Dell EMC on-premises solution versus support required for the AWS service.

Summary and Recommendations

Edison Group was commissioned to evaluate Total Cost of Ownership costs over a three-year period between Dell EMC Ready Solution for Big Data and an equivalent AWS public cloud environment.

Dell EMC Ready Solutions for Big Data was less than 50% of the cost of a comparable Amazon Big Data solution supporting NoSQL across two regions.

As expected, the differences in an on-premises versus cloud-based solution were reflected in the different components required to create equivalent infrastructures. For example, Dell EMC solution required adding a facilities component, while AWS treated database and data transfer separately.

While there were differences in costs between the various components, the result, due primarily to the significantly high costs associated with persistent storage on Amazon DynamoDB, was that Dell EMC Ready Solutions with Big Data was less than 50% of the cost of a comparable Amazon Big Data solution supporting NoSQL across two regions.

Appendix

Detailed TCO Analysis

Dell EMC BDaaS

Big Data Platform Components	Year 1		Year 2		Year 3		Totals	
Dell EMC BDaaS Solution								
Server Network Infrastructure	$257,925	30.2%	$ 0	0.0%	$ 0	0.0%	**$257,925**	15.2%
Software	172,800	21.0%	172,800	39.3%	172,800	39.3%	**518,400**	30.5%
Database	0	0.0%	0	0.0%	0	0.0%	**-**	0.0%
Data Transfer	0	0.0%	0	0.0%	0	0.0%	**-**	0.0%
Network Pipe	238,280	29.0%	228,480	52.0%	228,480	52.0%	**695,240**	40.3%
Services	125,000	15.2%	-	0.0%	-	0.0%	**125,000**	7.3%
Facility	9,701	1.2%	9,701	2.2%	9,701	2.2%	**29,103**	1.7%
Staffing	28,484	3.5%	28,484	6.5%	28,484	6.5%	**85,451**	5.0%
Totals	**$822,390**	**100.0%**	**$439,464**	**100.0%**	**$439,464**	**100.0%**	**$1,701,318**	**100.0%**

Amazon Big Data Solution

Big Data Platform Components	Year 1		Year 2		Year 3		Totals	
AWS Big Data Solution								
Server Network Infrastructure	$611,250	37.1%	$ 22,668	2.2%	$ 22,668	2.2%	**$ 656,587**	17.6%
Software	22,838	1.4%	22,838	2.2%	22,838	2.2%	**68,515**	1.8%
Database	756,163	45.9%	755,083	72.8%	755,083	72.8%	**2,266,330**	60.9%
Data Transfer	40,873	2.5%	40,873	3.9%	40,873	3.9%	**122,619**	3.3%
Network Pipe	156,226	9.5%	156,226	15.1%	156,226	15.1%	**468,677**	12.6%
Services	45,854	2.8%	24,662	2.4%	24,662	2.4%	**95,178**	2.6%
Facility	-	0.0%	-	0.0%	-	0.0%	**-**	0.0%
Staffing	14,242	0.9%	14,242	1.4%	14,242	1.4%	**42,725**	1.1%
Totals	**$ 1,647,447**	**100.0%**	**$ 1,036,593**	**100.0%**	**$ 1,036,593**	**100.0%**	**$ 3,720,632**	**100.0%**