White Paper

BIG DATA-AS-A-SERVICE

A Market and Technology Perspective

- What Big Data is about
- What service providers can do with Big Data
- What EMC can do to help

EMC Solutions Group

Abstract

This white paper looks at what service providers can do to address the booming Big Data market and how solutions that leverage EMC® technologies, such as Greenplum® and Isilon®, can be stepping stones on the path to providing Big Data in the cloud.

July 2012

Part Number H10839

Table of contents

Executive summary

Business case

Big Data refers to scale-out data architectures that address new requirements in terms of data volume, velocity, or variability for which traditional data architectures are ill suited. The Big Data market is already at $6.8 billion in 2012 and is set to grow explosively by almost 40 percent every year.

Today, with 10 percent of all IT spending in the cloud, again growing very quickly, service providers have a great opportunity to offer Big Data-as-a-Service. The challenge is to pick the right kind of services and technologies from the available options.

Key recommendations

Service providers with infrastructure service offerings may want to extend these to include data platform services. EMC's two product families, Greenplum® and Isilon®, can be leveraged as core components for such services. EMC Proven Solutions tie these together in blueprints for full service architectures, such as a Hadoop-as-a-Service solution.

Introduction

Purpose

The purpose of this white paper is to offer a definition of Big Data, map out how the Big Data-as-a-Service market presents itself to service providers, and look at how EMC® solutions based on technologies such as Greenplum and Isilon can enable service providers to build Big Data-as-a-Service offerings.

Scope

It is not the intention of this white paper to go into any level of detail on EMC's specific solutions. This is covered in other white papers and solution collateral. This white paper focuses on giving an overview of the Big Data-as-a-Service market and the technologies involved, using specific EMC solutions only as examples.

Audience

This white paper is targeted at service designers and product managers in service provider organizations to help them identify Big Data-as-a-Service opportunities, as well as at EMC and partner business developers and solution architects to enable them to engage in discussions with their service provider counterparts.

Terminology

This paper includes the following terminology.

Table 1. Terminology

Term	Definition
API	Application programming interface: Allows external applications to consume exposed functionality of a component.
IaaS	Infrastructure-as-a-Service
PaaS	Platform-as-a-Service
SaaS	Software-as-a-Service
BDaaS	Big Data-as-a-Service
HDaaS	Hadoop-as-a-Service
NoSQL	No (or sometimes: not only) SQL refers to non-traditional, usually non-relational data management systems

Big Data

Definition

In recent years, there has been an enormous explosion in data generation due to the proliferation of new technologies such as social networks, new and more powerful mobile devices, smart meters, sensors, and cloud computing. This explosion of data is set to continue and even accelerate: IDC has projected the global data volume to grow as much as 44 times between 2009 and 2020.

As data is increasingly becoming more varied, more complex, and less structured, it has become imperative to process it quickly. Meeting such demanding requirements poses an enormous challenge for traditional databases and scale-up infrastructures.

Big Data refers to new scale-out architectures that address these needs. Big Data is fundamentally about massively distributed architectures and massively parallel processing, using commodity building blocks to manage and analyze the data.

This definition is intentionally subjective and incorporates a moving range, from how large data sets need to be in order to be considered Big Data, to what variety, complexity, or velocity requirements need to be met. These requirements differ greatly by industry sector and will necessarily change as technology advances over time. Big Data is much more characterized by a scale-out architectural style than by specific data set sizes or processing speeds.

Market size

IDC estimates the value of the Big Data market to be about $6.8 billion in 2012, growing almost 40 percent every year to $17 billion by 2015. In 2012, that is already a 10 percent share of the overall business intelligence and analytics market. JMP Securities expects that share to grow to a staggering 36 percent by 2021, reaching about 13 percent of total world-wide IT spending by that time.

Big Data-as-a-Service

Taxonomy

For service providers, there are multiple ways to address the Big Data market with as-a-Service offerings. These can be roughly categorized by level of abstraction, from infrastructure to analytics software, as shown in Figure 1.

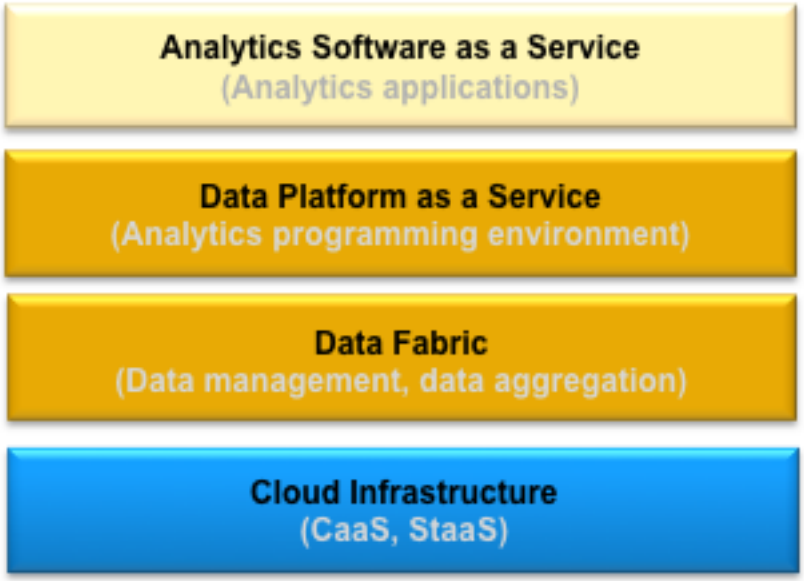

Figure 1. Big Data-as-a-Service layers

Cloud infrastructure

Starting from the bottom layer, any Big Data-as-a-Service infrastructure will usually leverage Infrastructure-as-a-Service components, particularly Compute-as-a-Service (CaaS) and Storage-as-a-Service resources and their management.

Also, a lot of Big Data is actually generated by applications deployed in a service provider's cloud infrastructure. Moving large amounts of data around—for example, from a customer's premises onto a service provider's—can be prohibitive in some scenarios. Hence, having the data that is to be further processed already available in the service provider's infrastructure enables Big Data services to be a natural enhancement to a service provider's infrastructure service offerings.

Data fabric

On the next layer up, service providers can offer data fabric services. These can be data management services (in the context of a broader Platform-as-a-Service (PaaS) offering or as a stand-alone Database-as-a-Service (DBaaS) offering), or a data aggregation and exposure—Data-as-a-Service (DaaS)—offering.

Data fabric: Data-as-a-Service

DaaS is fundamentally about aggregating and managing a particular data set, and allowing controlled access to that data set through some kind of API.

One example of this category is Google's Public Data service that provides access to all sorts of data provided by public institutions, which can then be used by applications to contextualize or visualize their data.

Another example of this would be a nationwide pool of academic and public bioinformatics data against which pharmaceutical companies can run their research analytics. The role of a service provider in such a scenario would be as an aggregator and custodian of the data.

Data fabric: Database-as-a-Service

Another form of data fabric is data management infrastructure services in the context of a bigger PaaS offering or as a standalone DBaaS. This means that database services are made available to applications deployed in any execution environment, including on a PaaS. This could be any kind of data store, starting with traditional relational databases, but in the Big Data context these would optimally be scale-out architectures such as NoSQL data stores and in-memory databases.

Since one of the key benefits of PaaS infrastructures is the dynamic scaling of applications, the underlying data fabrics must provide a similar set of features. And this not only applies to online transactional data stores, but equally to analytics databases. One emerging trend that can be observed is that analytical processing moves closer and closer to the point in time when the analyzed data was actually created. It is becoming more and more "realtime", a term that is frequently, if not quite accurately, used in this context.

At the leading edge of this development is stream processing, where data is analyzed immediately after it has been created. This is conceptually similar to more traditional message queuing, where several different subscribers can tap into a message stream and process messages. In stream processing, too, there can be different types of analytics components that tap into the data stream and pick out the data they need to work on. And, of course, all this needs to be done in a highly distributed fashion, where all components in the architecture allow dynamic scaling out and scaling in.

Data PaaS

The next layer up in the tiered Big Data-as-a-Service model is Data Platform-as-a-Service. Here the service provider not only puts a data management infrastructure in place but also the execution environment for data processing applications and scripts.

What this means is that users can upload both their data and their analytics jobs and the platform takes care of spinning up (and tearing down) appropriate clusters of data and processing nodes. Of course, the users in this scenario are more the technical community of data scientists and programmers, who would also be able to manage private dedicated analytics environments, as well as to write analytics jobs.

Analytics SaaS

In contrast to that, users of an Analytics Software-as-a-Service (SaaS) offering would be more familiar with interacting with an analytics platform on a higher abstraction level, that is, they would typically execute scripts and queries that data scientists or programmers developed for them, or generate reports, visualizations, and dashboards.

Analytics SaaS is typically dependent on a specific independent software vendor's (ISV) software suite and also the target verticals. While data fabrics and Data PaaS generally tend to have more horizontal solutions by nature, Analytics SaaS has many more vertical-specific solutions. Service providers that want to enter the Analytics SaaS space therefore have to choose which industries they want to support. Trying to address them all will likely not make any happy.

Market size

The Big Data-as-a-Service market is in its early stages and there is still relatively little market analysis data available. However, you can gain some indication of the market size by examining what percentage of all workloads is moving from enterprise premises to public or virtual private cloud service providers, and applying those trends to the Big Data market figures.

If you average out the different analysts' predictions, it is expected that about 15 percent of all IT spending will move to the cloud—that is, to as-a-Service delivery models—by 2015, growing to about 35 percent by 2021. Provided that the Big Data market reaches the $17 billion target by 2015, the Big Data-as-a-Service market will therefore be 15 percent of that figure, or $2.55 billion. With an estimated Big Data market of about $88 billion in 2021, the Big Data-as-a-Service market will be 35 percent of that, or about $30 billion. In other words, roughly 4 percent of all IT spending will go into Big Data-as-a-Service by 2021.

It is still too early to tell how these figures will break down to the different Big Data-as-a-Service layers. Service providers must base their decisions on their customers' demands, their own skill base, and the relative technology strengths and weaknesses of their partner ecosystem.

Technology overview

Overview

EMC has a number of technologies that can be applied in the Big Data-as-a-Service space. Two product families are specifically designed for Big Data and briefly introduced in this section:

- Greenplum
  - Greenplum massively parallel processing (MPP) architecture
  - Greenplum HD (Hadoop)
- Isilon

Greenplum MPP

Greenplum is the EMC product family that is all about data stores and tools for analytics. The core product is the Greenplum Database™. Built specifically to support Big Data analytics, the Greenplum Database manages, stores, and analyzes terabytes and petabytes of data. Users experience 10 to 100 times better performance than with traditional relational databases, a result of Greenplum's shared-nothing, MPP architecture, a high-performance parallel dataflow engine, and advanced gNet software interconnect technology.

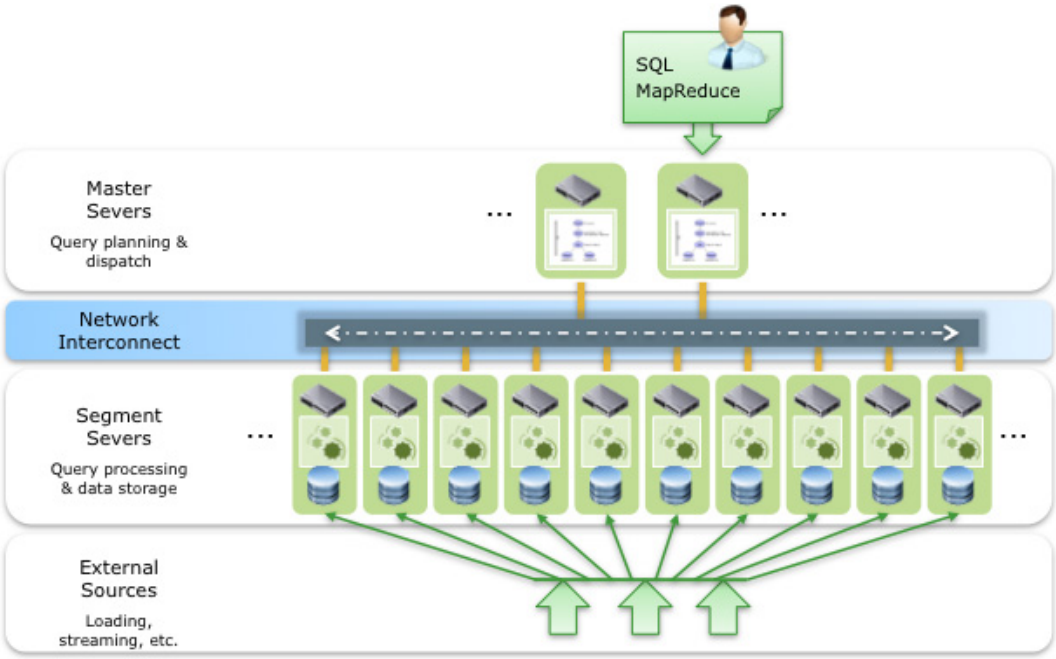

Figure 2. Greenplum Database MPP architecture

The Greenplum Database was conceived, designed, and engineered to allow customers to take advantage of large clusters of increasingly powerful, increasingly inexpensive commodity servers, storage, and Ethernet switches. Greenplum customers can gain an immediate benefit from deploying the latest commodity hardware innovations.

In Big Data-as-a-Service scenarios, the Greenplum Database can be part of a data fabric, part of a Data PaaS, or the foundation of an Analytics SaaS.

Greenplum HD

A second core product in the Greenplum family is Greenplum HD. Recent computing and business trends have triggered an explosion in the amount of unstructured data that companies, people, and devices generate each day. Hadoop has rapidly emerged as the preferred solution for big data analytics across unstructured data.

But the fast-changing Hadoop ecosystem can present challenges to any company that wants to standardize on core functionality and build repeatable processes. Greenplum HD is EMC's commercial distribution of Apache Hadoop and enables users to take advantage of big data analytics without the overhead and complexity of a project built from scratch.

Greenplum HD supports Isilon OneFS® Scale-Out NAS storage for Hadoop.

Isilon

Isilon is EMC's scale-out NAS storage platform with massive scalability of up to 15 petabytes in a single file system and up to 85 GB/sec throughput, while being extremely simple to manage. Isilon is powered by the OneFS operating system which provides the replication intelligence to achieve very high utilization rates—over 80 percent—thereby requiring much less raw capacity for the same amount of usable storage.

Example: Hadoop-as-a-Service

Introduction

To illustrate how EMC technologies can be used to implement a Big Data-as-a-Service solution, we will look at one important segment of the Big Data market, Hadoop.

Market size

IDC has published numbers for the software segment of the Big Data market, which enables us to estimate which part of the Big Data segment will be captured by Hadoop. Hadoop software revenue in 2012 is $209.2 million or 11 percent of the overall Big Data software market. Since Apache Hadoop is open source, the revenue share of software, as opposed to hardware and services, is distinctly smaller, which eventually leads to the conclusion that the Hadoop market is about 23 percent of the Big Data market in 2012. This number will grow quite quickly to 31 percent in 2013.

Applying the same reasoning as above to estimate what part of the overall Hadoop market will be delivered as-a-Service, we can expect the Hadoop-as-a-Service market in 2012 to be about $130 million, growing by 145 percent to $318 million in 2013. This is a good reason to specifically take this use case as an example.

Solution description

This solution enables service providers to offer Hadoop-as-a-Service (HDaaS), that is, allowing data scientist users to store unstructured data in a persistent Hadoop Distributed File System (HDFS) and process it using MapReduce jobs.

Using this solution, data scientists can instantly provision as much or as little capacity as they need to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics research. The HDaaS solution lets data scientists focus on crunching or analyzing their data without having to worry about time-consuming set-up, management or tuning of Hadoop clusters, or the compute capacity on which they sit.

Note that HDaaS is a Data Platform-as-a-Service solution for a technical data scientist audience, not an end-to-end Software-as-a-Service analytics solution for end users who just want to run certain queries and reports. However, data scientists can leverage the HDaaS platform to build such higher-level services for analytics end users.

From a conceptual architecture perspective, as shown in Figure 3 below, the HDaaS solution comprises a front-end portal that enables the user to load data into a shared Isilon HDFS storage backend, and submit Hadoop jobs and job flows. The HDaaS solution then spins up a virtualized Greenplum HD cluster—that is, virtual machines that leverage VMware® Serengeti technology for running the job control master as well as processing nodes—controls the job flow execution, collects the result data, and eventually tears the virtual machines down.

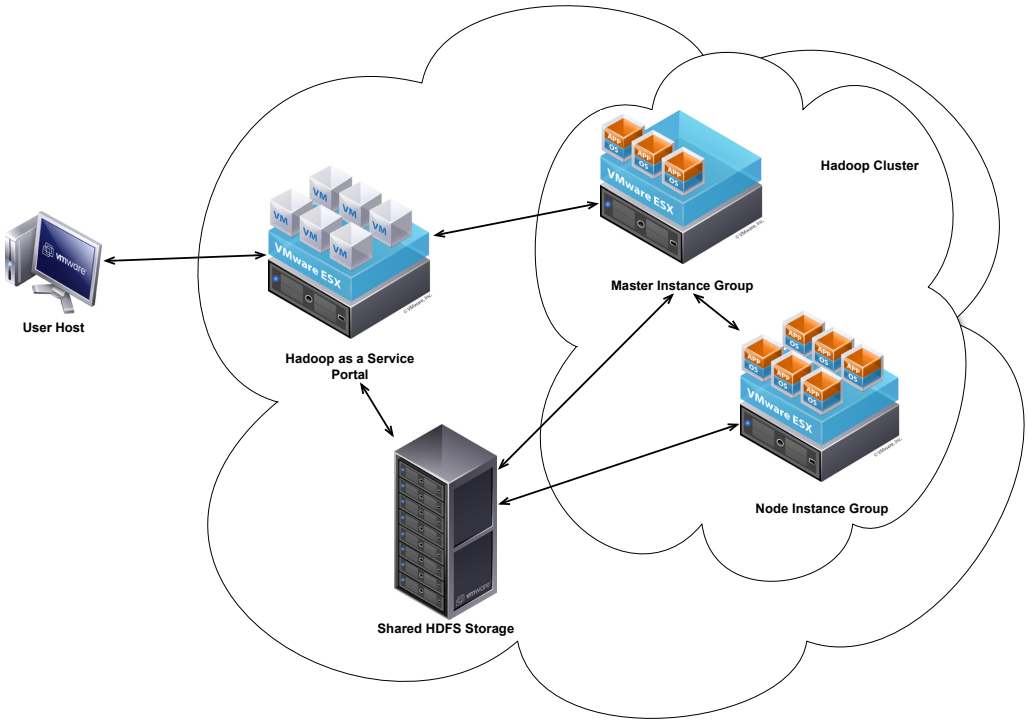

Figure 3. Hadoop-as-a-Service conceptual architecture

This solution's differentiation is in the following aspects:

- EMC is the first to provide an integrated HDaaS solution architecture for service providers.
- Isilon is the first and only enterprise-ready, scale-out NAS storage platform that natively integrates with the HDFS layer.
- Greenplum HD provides a complete offering of enterprise-ready Hadoop as software as well as an appliance.
- EMC is positioned as a leader in the Forrester Wave of Hadoop Solutions.

Conclusion

Summary

Big Data clearly presents a big opportunity for service providers, especially for service providers that already have infrastructure service offerings. It also presents challenges as moving into this space requires new technologies and skills. But EMC can help in this transition with products that are specifically designed for Big Data use cases, as well as with solutions that address the specific needs of service providers.

References

White papers

For additional information, see the white papers listed below.

- *EMC Isilon OneFS Operating System* http://www.emc.com/collateral/hardware/white-papers/h8202-isilon-onefs-wp.pdf
- *Greenplum Database: Critical Mass Innovation* http://www.emc.com/collateral/hardware/white-papers/h8072-greenplum-database-wp.pdf
- *Hadoop on EMC Isilon Scale-Out NAS* http://www.isilon.com/file-handler-show//1732/hadoop-emc-isilon-scale-out-nas.pdf

Product documentation

For additional information, see the product documents on the Web pages listed below.

- Isilon: http://www.isilon.com/
- Greenplum: http://www.greenplum.com/

Other documentation

For sources of market segmentation and numbers, see the documents listed below.

- *IDC Digital Universe Study*, sponsored by EMC, May 2010
- *IDC Worldwide Big Data Technology and Services 2012-2015 Forecast,* March 2012
- *IDC Worldwide Hadoop-MapReduce Ecosystem Software 2012-2016 Forecast,* May 2012