A Brief Update on

HADOOP

A look at how Hadoop is finding a place in data centers around the world, and changing the way enterprises handle business intelligence.

A Publication of Bright Computing

Table of Contents

INTRODUCTION

Hadoop in the Enterprise

2014 has been a banner year for Hadoop

In the past few years, Hadoop has gone from being a curiously named "science project" to a mainstream business intelligence tool. The hype is settling down, and people are figuring out what they can really do with it. Even mainstream enterprise database and storage vendors have come to embrace Hadoop as a worthy addition to the data analytics workflow, adding connectors to let their users take advantage of Hadoop's unique capabilities.

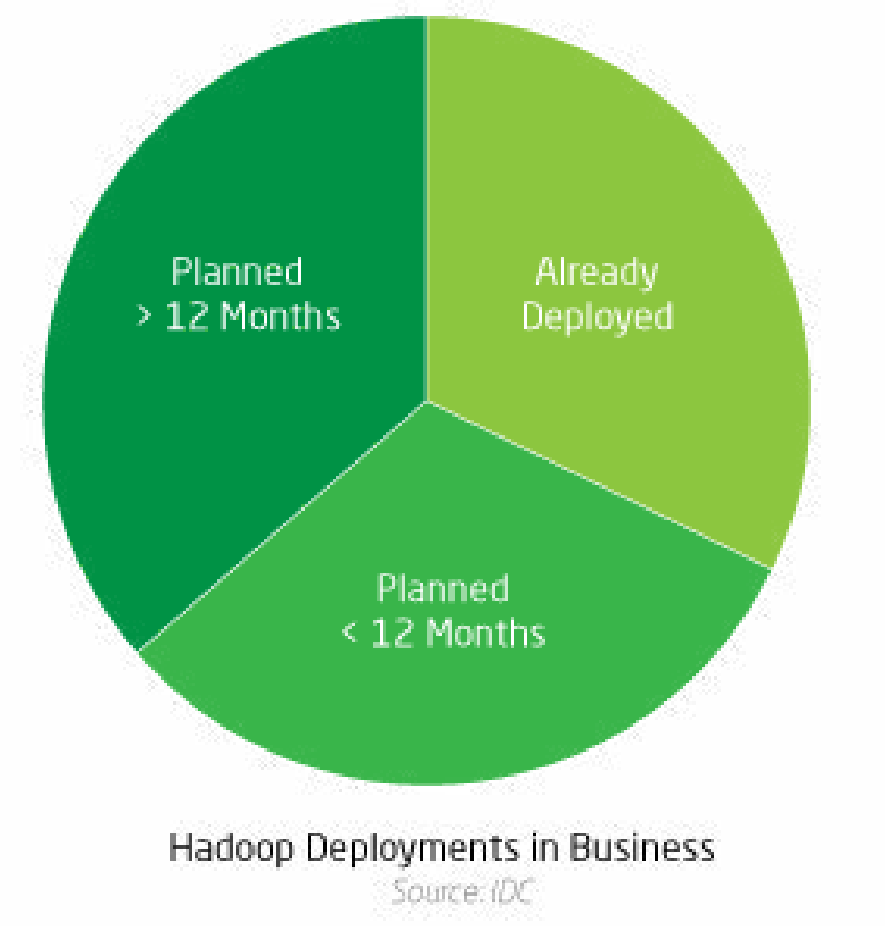
This eBook will look at what Hadoop 2.0 brings to the table, the impact Hadoop is having on large scale computing, and why you might want to consider using a full-fledged cluster management solution to build, manage, and maintain your Hadoop clusters.

CHAPTER ONE

Hadoop 2.0

Big Data Takes a Big Step

Hadoop 2.0 has taken a giant step forward in its move to produce an open source project that is easier to use and more stable. A Tweet Report from the research firm IDC shows that many companies were running Hadoop alongside other solutions to enhance their data analysis. The statistics showed that 32 percent already had deployed Hadoop, 31 percent intended to deploy Hadoop in the next 12 months, and 36 percent are saying that they are considering using a Hadoop deployment in more than a year.



Hadoop Deployments in Business

Source: IDC

This study highlights that some businesses are combining Hadoop 2.0 with such NoSQL databases as Cassandra, and MongoDB. Others are combining it with Greenplum

and Vertica. While not as common, some are using it to work in conjunction with SQL technologies. The biggest benefit from the study is that it shows the number of different ways that businesses are using this latest version of Hadoop.

One of the most exciting new components that Hadoop 2.0 provides is YARN (also called MapReduce 2.0). YARN has the advantage that management of the engine is separated from the actual algorithm. This means that the user can use MapReduce as a plug-in instead of as an interactive process. This is a major milestone in the development of Hadoop, taking it from being a simple tool to a complete operating system for big data. YARN enables the simultaneous execution of multiple applications on HDFS, the distributed file system, while allowing for better monitoring of data through the whole lifecycle.

The number of companies using Hadoop is growing. Yahoo, Google, Amazon, and eBay have been pioneers, but now others are joining the Hadoop club using Apache Hadoop. For example, NASA is using Hadoop to cope with the large amount of data in their climate simulation projects, so one could say that Hadoop 2.0 is now helping us get a better look at our planet's future.


CHAPTER TWO

Hadoop's Big Impact

Large Scale Computing

Most of us were introduced to Hadoop as a tool to help companies extract useful information from massive amounts of transactional data. The big online retailers use it to analyze buying patterns and predict what people might want to buy next. Online advertisers use it to figure out which ad to put in front of us on every web page we visit. And search vendors use it to help guess what you are really looking for when you type that ad-hoc query into your search bar.

"The use of Hadoop is not limited to large search engines or even Internet marketing...businesses across many industries are leveraging Hadoop to reach multiple business goals."

Hadoop's ability to absorb any type of data is well known. It doesn't matter if the data is structured or not, a Hadoop cluster can analyze just about any data for valuable information. Data from a variety of sources can be combined in new ways that provide results other systems can't.

In the world of mobile computing, Hadoop can analyze customer location, social media activity, and other signals. The results are used to provide those customers with services specifically tailored to their needs.

Hadoop has found its way into the financial world in a number of ways. It's used to detect fraud, and identify security threats that might otherwise go unnoticed. It also provides valuable insight into customer behavior by analyzing the way they use their mobile devices, exposing activities that might undermine security.

Hadoop helps pharmaceutical companies quickly root out low performers when bringing new drugs to market. It can also shorten the time to market after developing a new drug. Others in the healthcare industry also use Hadoop to provide more personalized patient care.

While Hadoop isn't the solution to all the world's problems — as the hype machine sometimes claims — it is changing the world of computing in very real ways.


CHAPTER THREE

Cluster Managers

Why Your Hadoop Cluster Needs One

Every day we hear about more organizations implementing Hadoop as part of their IT infrastructure. Most people think of Hadoop installation as something that starts on top of an existing, working Linux cluster. But how did that cluster get there? And once everything is up and running, how do they keep it that way in order to continue getting value out of it.

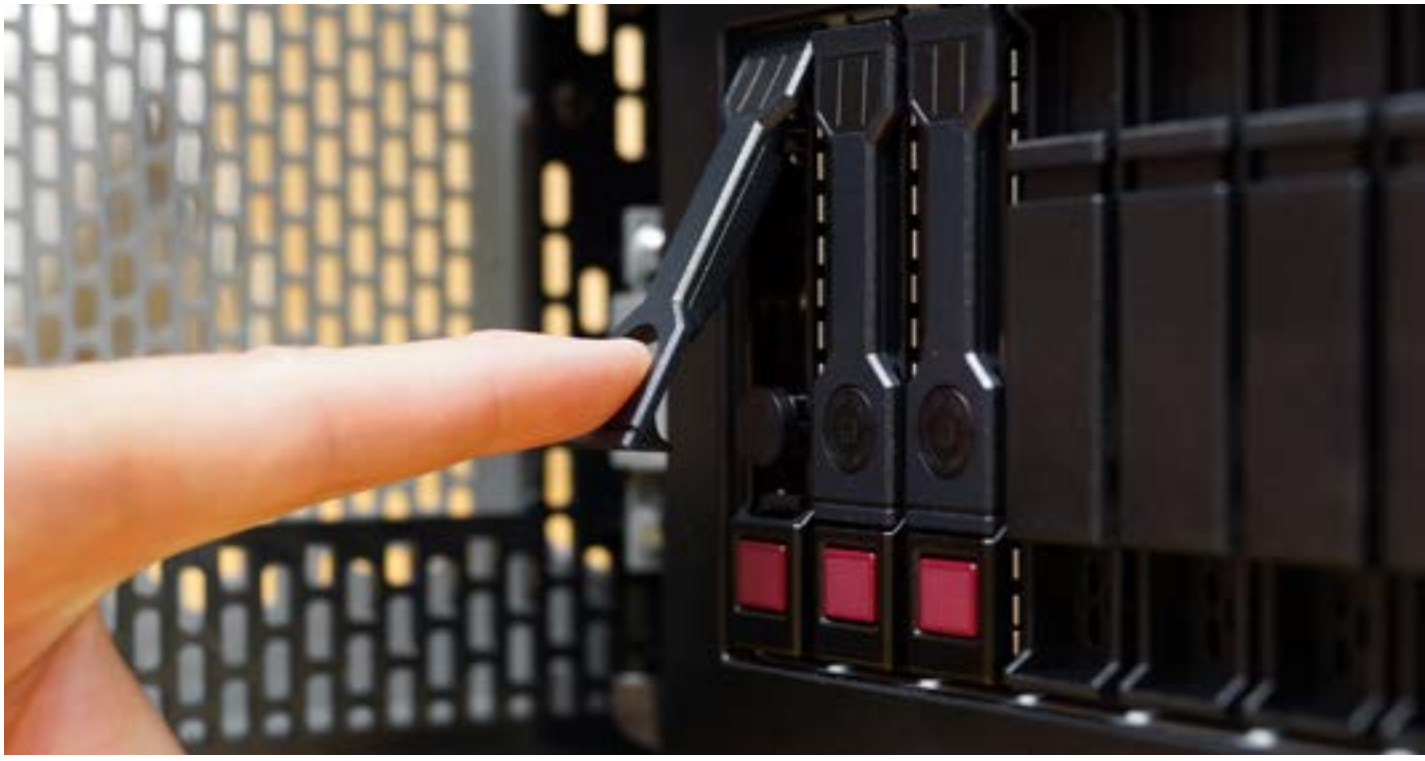
With a small cluster, that's not such a big deal. As clusters get bigger, though, the work involved in keeping them healthy grows...a lot.

Will those scripts you crafted so carefully scale as the cluster grows to hundreds, or even thousands of nodes? Will the next sysadmin that comes in — after you get promoted for your great work — be able to do things just the way you planned? Every time? Maybe not. Using an enterprise-grade cluster manager will make every aspect of your interaction with your new Hadoop cluster more effective, more efficient and more consistent.

Easy deployment

It isn't too hard to build a cluster and install Hadoop on it in the lab, but things get more complicated when you take your lab project into production. Others may not be as intimate with the system as you are, and as they work to solve the various issues that come up during a deployment, the sysadmin on site may make changes that cause problems.

Even if that doesn't happen, there's still the matter of time. Manually installing and configuring all the parameters necessary to build a functioning cluster — and then installing the Hadoop software on top of that — can take a lot of time. The bigger the cluster, the more time it will take. Possibly more time than the data scientists waiting for their new cluster want to wait.



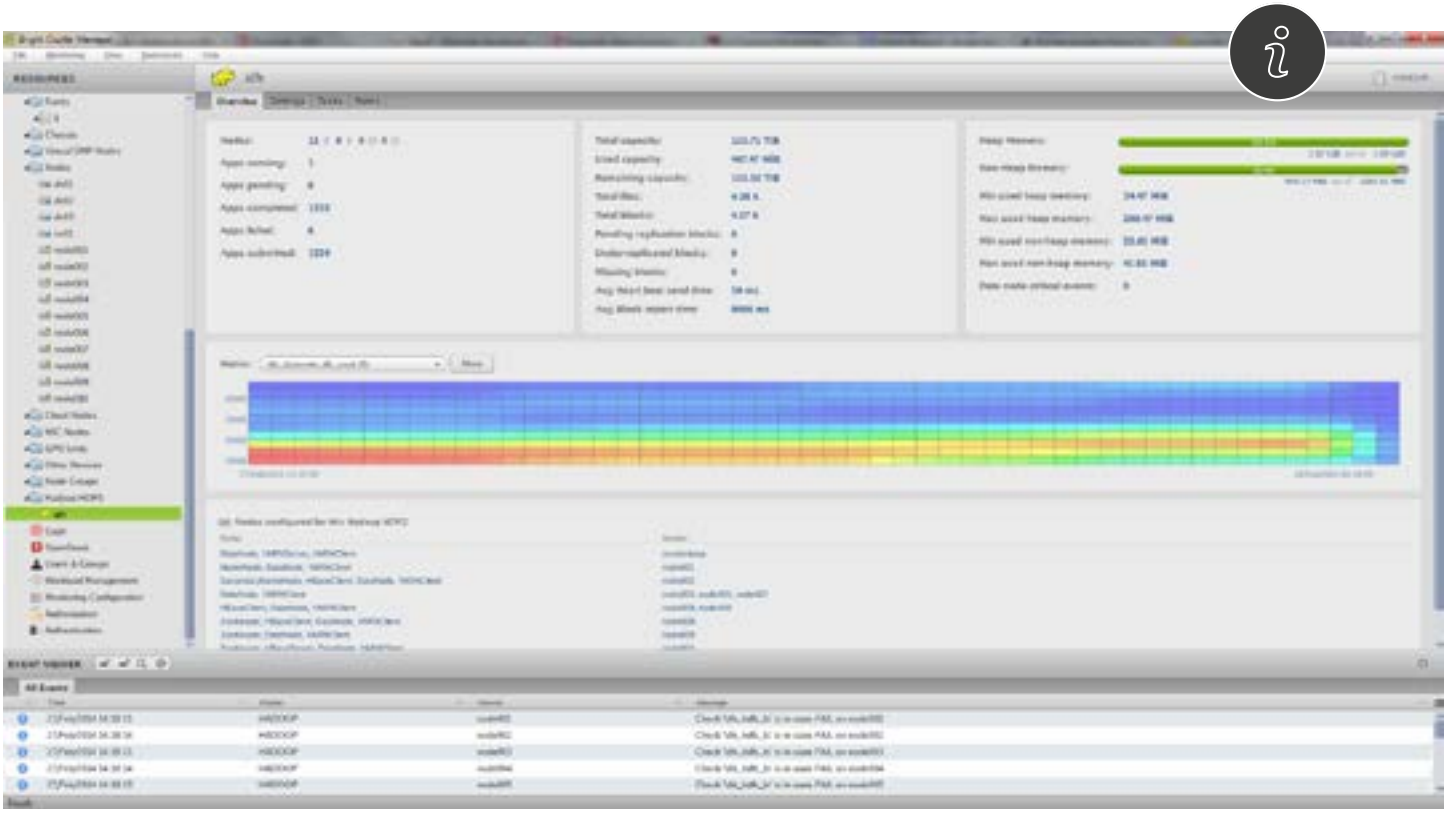
Cluster management software automates the installation and configuration process, not only of the Hadoop software, but also of the operating system software, networking software, hardware parameters, disk formatting, and dozens of other little things that need to be done.

Keeping an eye on things

You may have a top-notch installation and deployment team that can quickly build and configure clusters of any size anywhere you need them. Enterprise-grade cluster management software can still be a real benefit in operating, monitoring, and managing the cluster.

They will automatically detect and you of any drive failures, memory problems, overheating, and configuration errors.

When you were testing and evaluating your Hadoop cluster in the lab, you had full control over it, and there wasn't a team of data scientists counting on it being available whenever they need it. Now that the cluster is in production, you no longer have the luxury of waiting for things to fail, and bringing nodes back up when you have time. You have to keep things humming 24x7x365, which requires manage the Hadoop cluster the same way you would manage any other critical service in the data center. You need professional grade management tools.



Cluster managers let you keep an eye on all critical metrics in your Hadoop cluster on one screen.

Data protection

While a properly configured Hadoop cluster does a great job of protecting data through replication across multiple nodes, the system's performance can suffer when that protection gets called into play. So it's a good idea to keep an eye on the file system itself to make sure it is working properly. Is it spreading the load evenly amongst available resources? Is the replication factor high enough? Are some of the drives reaching capacity? An enterprise-grade cluster manager can answer these questions for you.

Modern enterprise cluster managers provide dozens of other monitoring capabilities and health checks that can be invaluable to anyone responsible for maintaining a Hadoop cluster.

Managing Change

Change is inevitable, and your cluster will need to be modified after it has been put into production. For example you may need to add, replace, or remove nodes. You may need to upgrade or replace operating systems, or other critical software across a large number of nodes. Cluster managers make light work of such heavy tasks. They provide administrators with a user-friendly interface to do things like set parameters, select images, and provide other necessary information that allows the cluster manager to makes the necessary changes automatically.


CHAPTER FOUR

Plugging the Hole

Hadoop Cluster Management

Apache Hadoop has been around for longer than many people realize. It's now over a decade old, and has come a very long way since 2002. GigaOm's Derrick Harris did a good job covering Hadoop's history in a series of posts titled *The history of Hadoop: From 4 nodes to the future of data*.

There is now a thriving community built up to support and extend Hadoop, and it has spawned a bevy of startups seeking to address various aspects of Hadoop ecosystem for profit. But despite all the progress, there's still an "elephant in the room" (sorry I couldn't resist) that nobody talks much about — cluster management.

What's in a Name?

Ambari, the management component of Apache Hadoop, and other management solutions focus on Hadoop-related software. They don't address the configuration, installation, and management of the hardware and software Hadoop needs to have in place before you can even install it.

Most discussions about Hadoop deployment start in the middle of the story. They assume there is already a working cluster to install Hadoop

on. They also ignore the ongoing monitoring and management of the underlying infrastructure that the Hadoop cluster is build upon.

So how do you get from bare hardware to the point where you have a functioning cluster of servers — usually running Linux — that is fully networked and loaded with all the right software?

You can build the systems by hand, installing the right combination of operating system and application software, and configuring the networking hardware and software properly. You might even use sophisticated scripting tools to help speed up that process. But there is a better way. Cluster managers exist to fill that gap.

Who needs a cluster manager anyway?

It's true you can build and manage your Hadoop systems without a cluster manager, but using one brings a number of advantages. First, they can save you a lot of time up front, when you first rack the servers and load them with software. Cluster managers automate the entire process once the servers are racked and cabled. They do hardware discovery, read the intended configuration of each machine from a database, and load the software onto them all — whether you have 4 nodes or 4000. They do this repeatedly and consistently, which is handy when you need to deploy a series of clusters in branches of your organization around the world. Automating the process reduces the number of configuration errors and

missteps that can occur whenever a complex, multistep procedure is involved. A good cluster manager will let you go from bare metal to a working cluster in less than an hour.

The usefulness of cluster managers doesn't end once the Hadoop cluster is deployed. While the Hadoop management software available in the leading distributions do a great job of monitoring and managing Hadoop, they provide little information about the underlying cluster that the Hadoop systems runs on. Cluster managers provide the missing pieces by monitoring and managing the infrastructure to keep things running properly. They monitor things like CPU temperatures and disk performance, and can alert operators to problem before they snowball out of control.

Plan Ahead

So the next time you're discussing Hadoop solutions for your organization, don't forget to consider the foundation those solutions rest on. Once you put your cluster into production, you'll want to be able to manage the entire stack. Consider the operational procedures that need to happen before you install Hadoop, and the systems you'll need to have in place to keep your cluster healthy AFTER Hadoop is up and running. I promise you will be glad you took the time to "fill the gaps" down the road.

SIGN UP FOR A FREE DEMO

Talk to one of our experts and see how Hadoop can help you get the results you really want.

FREE DEMO

a publication of Bright Computing