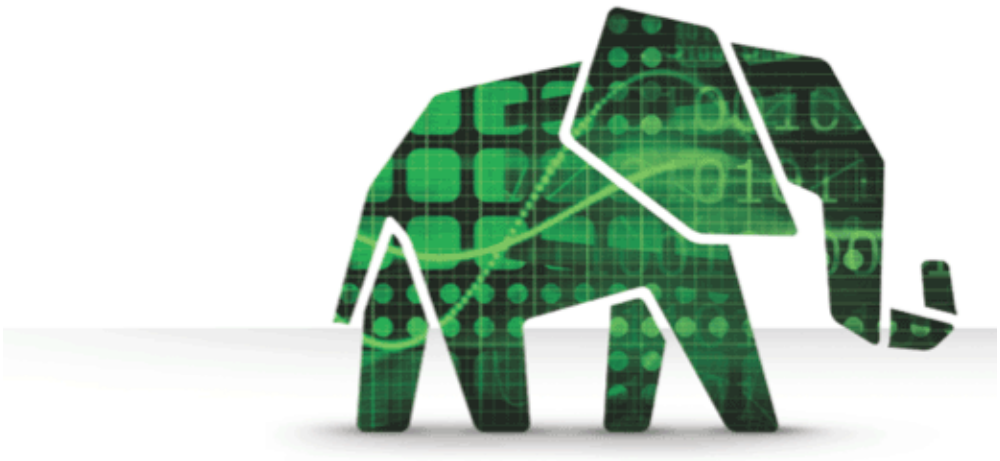
Hortonworks Community Driven Apache Hadoop

Apache Hadoop Patterns of Use

April 2013



http://www.hortonworks.com

Hortonworks

Big Data: Apache Hadoop Use Distilled

There certainly is no shortage of hype when it comes to the term “Big Data” as vendors and enterprises alike highlight the transformative effect that can be realized by deriving actionable insights from the deluge of data that is our daily reality. But amongst the hype, the practical guidance that we all need is often lacking. Why is Apache Hadoop most typically the technology underpinning “Big Data”? How does it fit into the current landscape of databases and data warehouses that are already in use? And are there typical usage patterns that can be used to distill some of the inherent complexity for us all to speak a common language?

As an organization laser focused on developing, distributing and supporting Apache Hadoop for enterprise customers, we have been fortunate to have a unique vantage point. The core team at Hortonworks includes the original architects, developers and operators of Apache Hadoop and its’ use at Yahoo, and through that experience they have been privileged to see Hadoop emerge as the technological underpinning for so many big data projects. That has allowed us to observe certain patterns that we’ve found greatly simplify the concepts associated with Hadoop, and our aim is to share some of those patterns here.

Where does Hadoop fit?

Enterprise users have been building analytic applications for years, so what’s the fuss around big data? Today, nearly every enterprise already has analytic applications that they use every day, and these are relatively well understood and captured in the graphic below. The general flow looks something like this and is depicted in figure 1 below:

- Data comes from a set of data sources – most typically from the enterprise applications: ERP, CRM, custom applications that power the business
- That data is extracted, transformed, and loaded into a data system: a relational Database Management System (RDBMS), an Enterprise Data Warehouse (EDW), or even a Massively Parallel Processing system (MPP)
- A set of analytical applications – either packaged (e.g. SAS) or custom, then point at the data in those systems to enable users to garner insights from that data

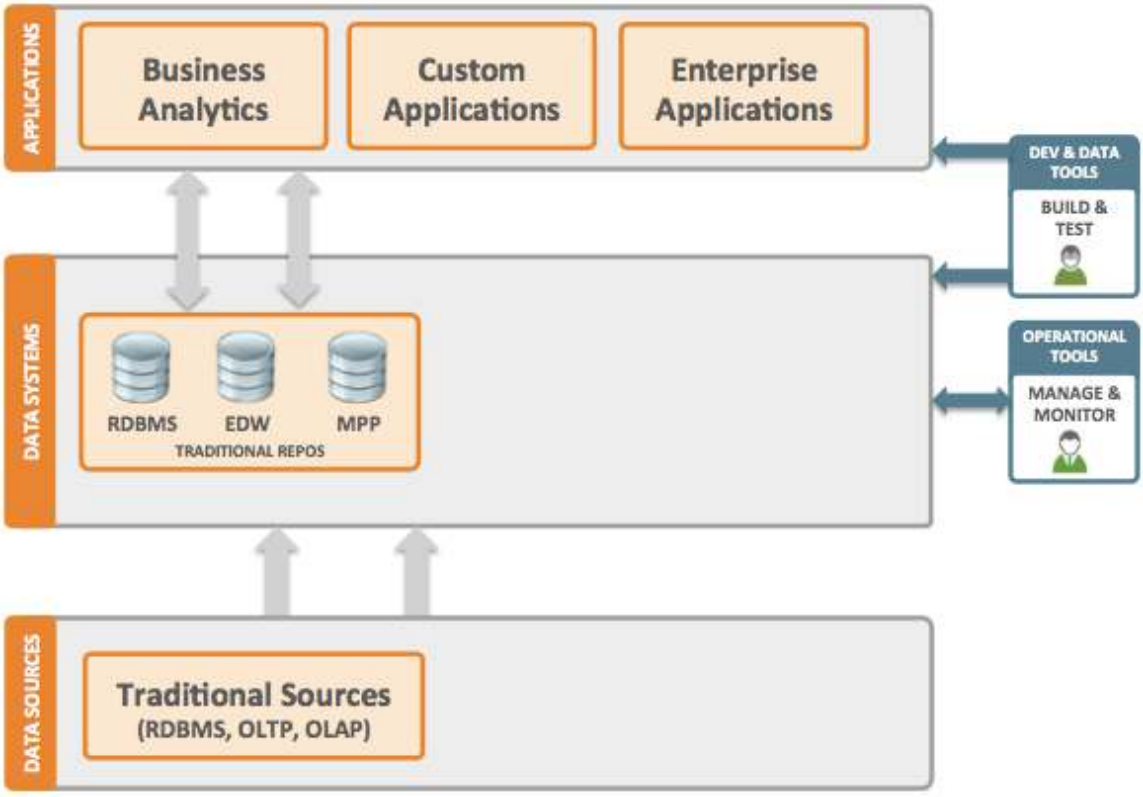


Figure 1: Traditional Enterprise Data Architecture

Times have changed however. Recently, there has been an explosion of data entering into this landscape. And it isn't just more records coming from the ERP or CRM systems: it is an entirely new class of data that was never envisioned when those data systems first came into being. It is machine generated data, sensor data, social data, web logs and other such types that are both growing exponentially, but also often (but not always) unstructured in nature. It also includes data that was once thought of as low to medium value or even exhaust data, too expensive to store and analyze. And it is this type of data that is turning the conversation from "data analytics" to "big data analytics": because so much insight can be gleaned for business advantage.

As a result, an emerging data architecture that we most commonly see looks something like the picture below. The key difference being that Apache Hadoop – originally conceived to solve the problem of storing huge quantities of data at a very low cost for companies like Yahoo, Google, Facebook and others – is increasingly being introduced into enterprise environments to handle these new types of data in an efficient and cost-effective manner.

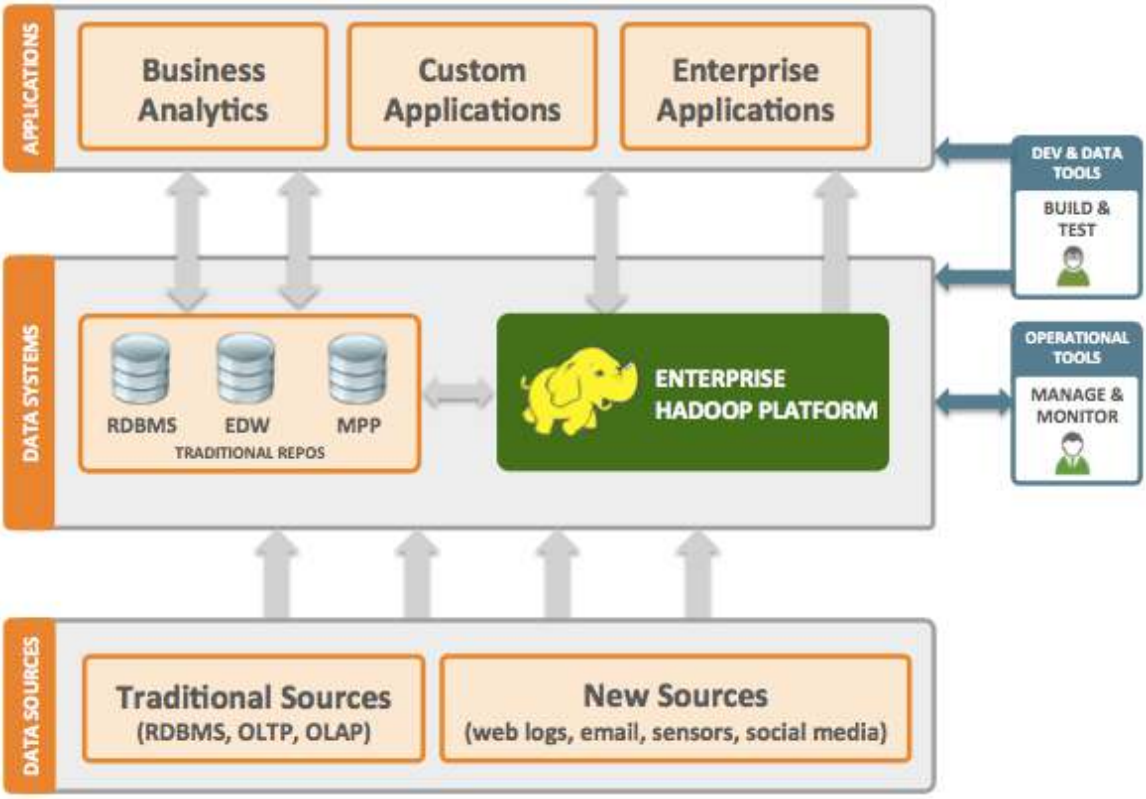


Figure 2: The Emerging Big Data Architecture

Key Takeaway: *Hadoop is not replacing the traditional data systems used for building analytic applications – the RDBMS, EDW and MPP systems – but rather is a complement. And it is critical that it interoperate with existing systems and tools (and able to be deployed on a range of systems and technologies). But by providing a platform to capture, store and process vast quantities of data in a cost efficient and highly scalable manner, it enables a whole new generation of analytic applications to be built.*

Common Patterns of Hadoop Use

Analytic applications come in all shapes and sizes – and most importantly, are oriented around addressing a particular vertical need. In this sense, at first glance they can seem to have little relation to each other across industries and verticals. And certainly, analytic challenges vary greatly across different verticals. But in reality, when observed at the infrastructure level, some very clear patterns emerge: they can fit into one of the following three patterns.

Pattern 1: Apache Hadoop as a Data Refinery

Essentially the "Data Refinery" pattern of Hadoop usage is about enabling organizations to incorporate these new data sources into their commonly used BI or analytic applications. For example, I might have an application that provides me a view of my customer based on all the data about them in my ERP and CRM systems, but how can I incorporate data from their web sessions on my website to see what they are interested in? The "Data Refinery" usage pattern is what customers typically look to.

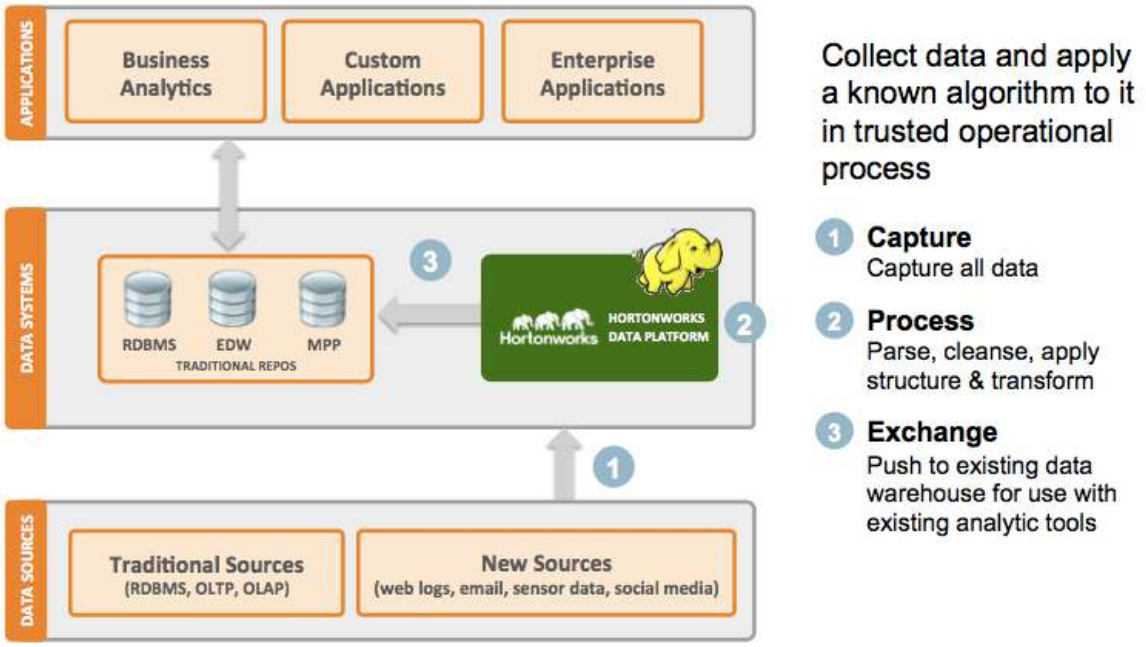


Figure 3: Hadoop as Data Refinery

The key concept here is that Hadoop is being used to distill large quantities of data into something more manageable. And then that resulting data is loaded into the existing data systems to be accessed by traditional tools – but with a much richer data set. In some respects this is the simplest of all the use cases in that it provides a clear path to value for Hadoop with really very little disruption to the traditional approach.

A Data Refinery in Practice

No matter the vertical, the refinery concept applies. In financial services we see organizations refine trade data to better understand markets or to analyze and valuate complex portfolios. Energy companies use big data to analyze consumption over geography to better predict production levels, saving millions. Retail firms (and virtually

any consumer facing organization) often use the refinery to gain insight into online sentiment. Telecoms are using the refinery to extract details from call data records to optimize billing. Finally, in any vertical where we find expensive, mission critical equipment, we often find Hadoop being used for predictive analytics and proactive failure identification. In communications, this may be a network of cell towers. A restaurant franchise may monitor refrigerator data. Often, Hadoop is used to predict failure of these resources before they happen, as companies know it is more expensive to fix than replace equipment and nobody wants down time. In general, anywhere analytics are used, the use case is present as it typically prepares data for use within the EDW and BI tools.

Pattern 2: Data Exploration with Apache Hadoop

The second most common use case is one we called "Data Exploration". In this case, organizations are capturing and storing a large quantity of this new data (sometimes referred to as a data lake) in Hadoop and then exploring that data directly. So rather than using Hadoop as a staging area for processing and then putting the data into the EDW – as is the case with the Refinery use case – the data is left in Hadoop and then explored directly.

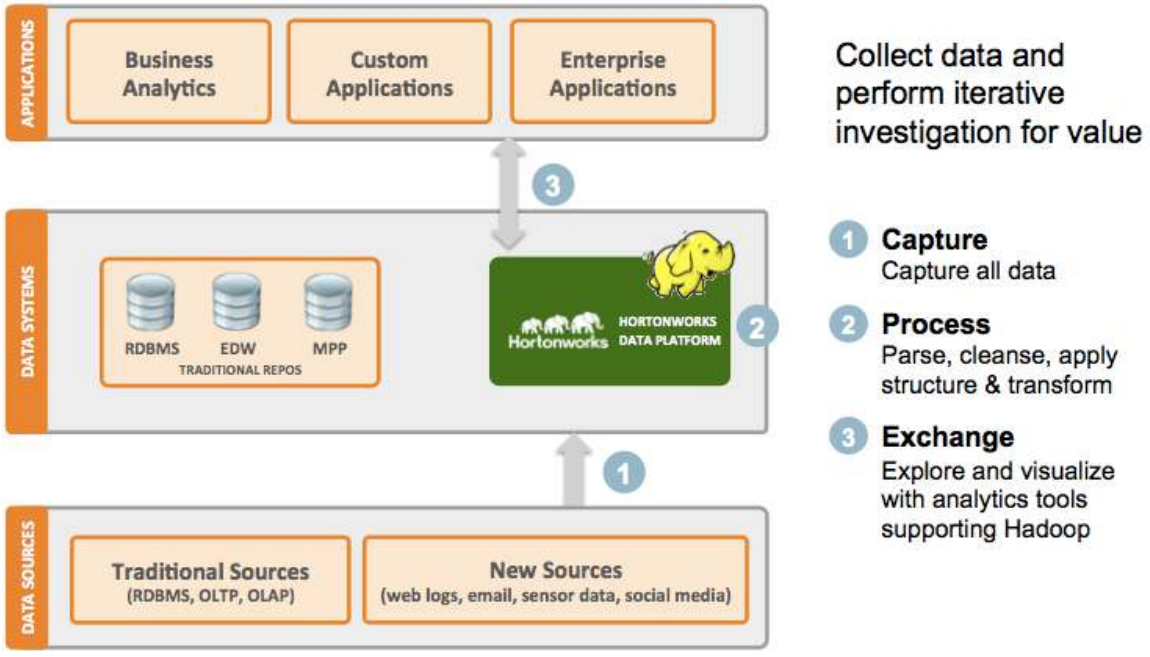


Figure 4: Hadoop and Data Exploration

The Data Exploration use case can often be where Enterprises start – by capturing data that was previously being discarded (exhaust data such as web logs, social media data, etc) and building entirely new analytic applications that leverage that data directly. An example might be a TelCo using the Exploration use case in order to capture huge quantities of machine data in order to predict when equipment is likely to fail. Given the huge quantities involved, they may not have previously been able to do this in a cost effective way.

Data Science and Exploration

Again, nearly every vertical can take advantage of the exploration use case. For instance, in financial services, we find organizations using exploration to perform forensics or to identify fraud. A professional sports team will use data science to analyze trades and their annual draft, like we saw in the movie Moneyball. Many organizations are using Hadoop to also identify a single view of the truth for customers or products and other entities. This is proving massive benefit as it results in better customer service or increased products per customer as marketing programs are improved. Ultimately, data science and exploration are used to identify net new business opportunities or net new insight in a way that was once impossible before Hadoop.

Pattern 3: Application Enrichment

The third and final use case is one we call "Application Enrichment". In this scenario, data stored in Hadoop is being used to impact an application's behavior. For example, by storing all web session data (i.e. all of the session histories of all users on a web page), I can customize the experience for a customer when they return to my website. By storing all this data in Hadoop, I can keep session history from which I can generate real value – for example by providing a timely offer based on a customer's web history.

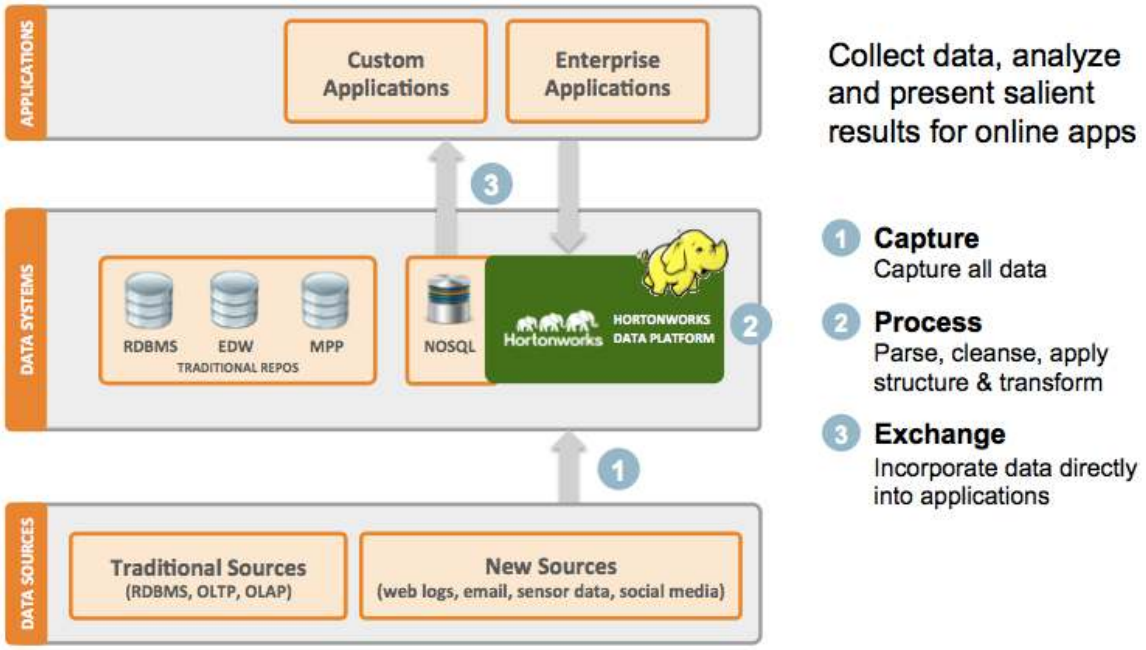


Figure 5: Application Enrichment with Hadoop

For many of the large web properties in the world – Yahoo, Facebook and others – this use case is foundational to their business. By customizing the user experience, they are able to differentiate in a significant way from their competitors. As one might expect, this is most typically the latest use case to be adopted – generally once organizations have become familiar with Refining and Exploring data in Hadoop. But at the same time, this also hints at how Hadoop usage can and will evolve over time to serve an ever greater number of applications that are served by the traditional database today.

Enrichment: The Right Data at the Right Time to the Right Consumer

The most straightforward enrichment use is the recommendation engines deployed by large web properties. Theses organizations analyze massive amounts of data to identify patterns/repeatable behavior and then serve up the right content to the right person at the right time in order to increase conversation rates for purchase. In fact, this was the second use case for Hadoop at Yahoo as they realized Hadoop could help improve ad placement. This concept translates beyond the large web properties and is being used the more traditional enterprise to improve sales. Some brick and mortar organizations are even using these concepts to implement dynamic pricing in their retail outlets.

A Pragmatic Approach to Adoption

There is certainly complexity involved when any new platform technology makes its' way into a corporate IT environment, and Hadoop is no exception. And it is for this reason that at Hortonworks we are so focused on interoperability to ensure that Apache Hadoop and the Hortonworks Data Platform works with your existing tools. With deep engineering relationships with Microsoft, Teradata, Rackspace and others we work hard to enable usage of Hadoop – which is having such a profound impact at so many organizations around the world.

So follow us, get engaged with our learning tools, or download the HDP Sandbox, a single node installation of HDP that can run right on your laptop. Hadoop has the potential to have a profound impact on the data landscape, and by understanding the common patterns of use, you can greatly reduce the complexity.

Download the Hortonworks Sandbox to get started with Hadoop today

3460 West Bayshore Rd.
Palo Alto, CA 94303 USA

US: 1.855.846.7866
International: 1.408.916.4121
www.hortonworks.com

Twitter: twitter.com/hortonworks
Facebook: facebook.com/hortonworks
LinkedIn: linkedin.com/company/hortonworks