






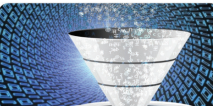
RedPoint Data Management™ for Hadoop® Bridging the Big Data Skills Gap

Adoption of Hadoop by data-driven organizations is exploding. Hadoop's potential cost effectiveness and facility for accepting unstructured data is making it central to modern, "Big Data" architectures. The advancements in Hadoop 2.0 increase the technology's promise to an even greater extent.

But with these opportunities also come challenges and adoption hurdles that make getting the most out of Hadoop easier said than done. Read on as we review some Hadoop basics, highlight some of the adoption challenges that exist and explain how RedPoint Data Management for Hadoop helps organizations accelerate their work with Hadoop.

	Lower cost scaling		No need for structure		Ease of data capture
Hadoop 1.0 <ul style="list-style-type: none"> All operations based on Map Reduce Intrinsic inconsistency of code based solutions Highly skilled and expensive resources needed 3rd party applications constrained by the need to generate code 			Hadoop 2.0 <ul style="list-style-type: none"> Introduction of the YARN: "a general-purpose, distributed, application management framework that supersedes the classic Apache Hadoop MapReduce framework for processing data in Hadoop clusters." Mature applications can now operate directly on Hadoop Reduce skill requirements and increased consistency 		

HADOOP 1.0 vs HADOOP 2.0

		
Skills Gap <ul style="list-style-type: none"> Severe shortage of MR skilled resources Very expensive resources and hard to retain Inconsistent skills lead to inconsistent results Under utilizes existing resources Prevents broad leverage of investments across enterprise 	Maturity & Governance <ul style="list-style-type: none"> A nascent technology ecosystem around Hadoop Emerging technologies only address narrow slivers of functionality New applications are not enterprise class Legacy applications have built short term capabilities 	Data Into Information <ul style="list-style-type: none"> Data is not useful in its raw state, it must be turned into information Benefit of Hadoop is that same data can be used from many perspectives Analysts must now do the structuring of the data based on intended use of the data

CHALLENGES TO HADOOP ADOPTION

What Is Hadoop? Why Are Organizations Excited about It? What's Different about Hadoop 2.0?

Hadoop is an open-source software framework for storage and processing of large data sets on clusters of inexpensive hardware. Hadoop was created by Doug Cutting and Mike Cafarella and adopted by Apache, and is supported by a global community of contributors and users.

Part of Hadoop's appeal is that it offers a means of storing and processing very large amounts of data more cost-effectively than traditional databases or data warehouses. But also, Hadoop's lack of inherent structure enables organizations to quickly and flexibly incorporate new data without a master plan. Data can be simply "dumped" into Hadoop for later structuring and analysis. In contrast, traditional databases require careful planning and documentation before the first record can be loaded.

Initial releases of Hadoop, starting in 2007, required users to rely heavily on MapReduce, a coding-intensive programming model for managing and manipulating data. Hadoop version 2.0, released in 2013, introduced the YARN architecture, short for "Yet Another Resource Negotiator." Apache described it as "a general-purpose, distributed, application management framework that supersedes the classic Apache Hadoop MapReduce framework for processing data in Hadoop clusters." YARN allows applications to run directly in Hadoop, bypassing MapReduce.

Obstacles to Hadoop Adoption

The release of YARN as part of Hadoop 2.0 was timely because a significant obstacle to Hadoop adoption has been a shortage of skilled MapReduce coders. Prior to YARN, these skills were vital to working with Hadoop. In fact, some large technology vendors went so far as to acquire start-ups primarily to obtain their MapReduce coders.

This shortage has prevented Hadoop from being more than just an experiment at many companies. Very large companies or companies in tech hot spots on the east and west coast of the US may be able to lure MapReduce talent with large salaries. But most companies lack the budget or simply the geographic appeal to attract the talent needed to make use of Hadoop on a large scale.

Other issues have hampered Hadoop adoption. For example, many of the tools and applications in the Hadoop ecosystem are first-generation, immature technologies. They lack the capabilities – particularly in terms of security and governance – that most enterprises require in their corporate data environments. Plus, one of the strengths of Hadoop is also a weakness – that is, users’ ability to dump large volumes of uncurated data into Hadoop. While this aids the speed of data collection, it also creates complications when it comes to processing this data into useful information.

RedPoint Removes the Obstacles

RedPoint Global, an established vendor in the data quality and data integration market, has come to the rescue of organizations struggling with the challenges and obstacles described above.

RedPoint’s top-rated data quality and data integration capabilities, which organizations have been benefiting from for many years, are now available for Hadoop. RedPoint Data Management for Hadoop leverages the YARN architecture and allows users to perform practically any desired data management function on data stored in Hadoop without involving MapReduce at all. No MapReduce skills are needed to use RedPoint, and RedPoint does not generate MapReduce code “behind the scenes” to do its work. Users employ RedPoint’s graphical interface to design data management jobs, and RedPoint executes 100% of these functions through YARN directly in the Hadoop cluster.

In fact, RedPoint is the only pure YARN data quality and data integration application on the market today. Take a close look at products that seem similar – most of them are MapReduce-based and still require MapReduce and Java programming skills.

Organizations that want to realize the promised business value of Hadoop and Big Data now can with RedPoint Data Management for Hadoop.

Key Features of RedPoint Data Management for Hadoop

Rich Data Quality and Data Integration Functions

RedPoint functions include ELT, ETL, cleansing, matching, de-duping, merging/purging, householding, parsing, partitioning, appending, address standardization, master key creation and maintenance, automation, monitoring, notification. All functions, including hundreds of data transformations, are available to be used on data in Hadoop.

The Same Award-winning Functionality as for Traditional Data

RedPoint’s capabilities have been rated #1 by users in industry analyst surveys for speed, match quality and ease of use. These exact functions are now available in Hadoop, not a subset or a “special” version built just for Hadoop. The user experience and available functionality is the same whether you’re running in traditional single-server execution, or on Hadoop.

Graphical User Interface

Users design and execute data quality and data integration jobs entirely within a graphical user interface. No MapReduce (or other) coding is required. Even Hadoop-related tools, such as Pig and Hive, designed to make working within a MapReduce context easier become unnecessary. RedPoint provides users with the same data flow-style user interface paradigm that has become standard for managing data in a traditional setting.

Zero Footprint Install

No software needs to be installed in the cluster itself. The RedPoint server pushes a binary executable to the cluster, which runs and then “evaporates.” And since RedPoint is pure YARN, it respects YARN’s task prioritization rather than competing for computing resources in the cluster.

No Data Movement

RedPoint executes its data quality and data integration tasks “inside” Hadoop, without data movement. This means there’s no wasted time, nor is there a need for any additional storage expense.

One Product, One User Interface for All Data

Users can perform data quality and data integration functions that span external data sources and Hadoop, using the same product and the same user interface. This allows users to combine data, migrate data from one location to another, match and key across data types and derive insight or take action on data no matter where it sits.

Use MapReduce


- complex
- requires new skills
- inefficient execution

```

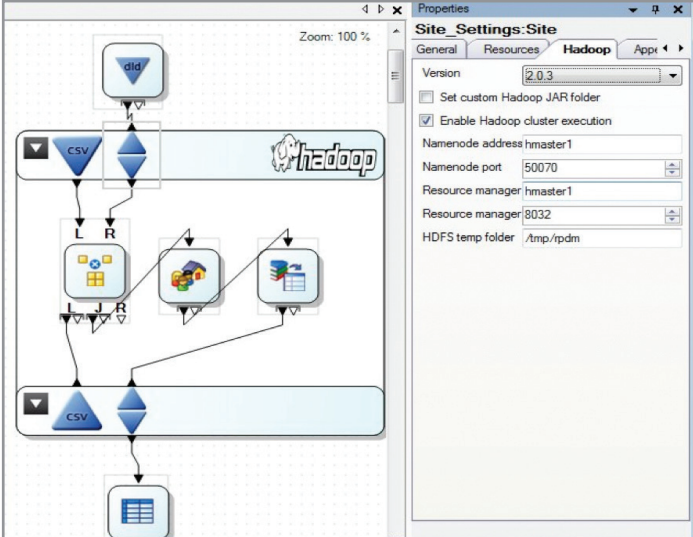
1 class Mapper
2   method Map(docid id, doc d)
3     H = new AssociativeArray
4     for all term t in doc d do
5       H{t} = H{t} + 1
6     for all term t in H do
7       Emit(term t, count H{t})
          
```

Move data out of Hadoop

- extra time and effort
- extra storage (expensive)
- defeats the purpose of Hadoop



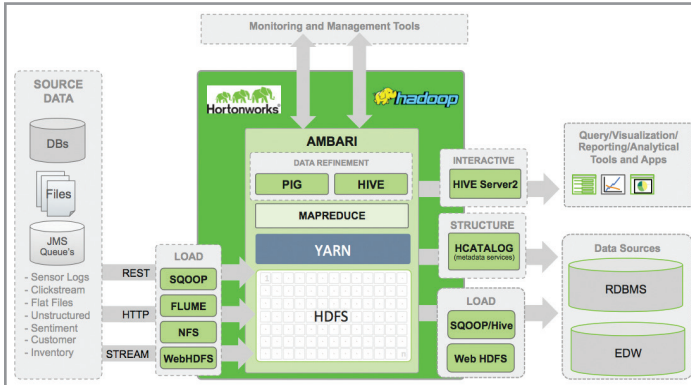
PREVIOUS OPTIONS FOR DATA MANAGEMENT IN HADOOP



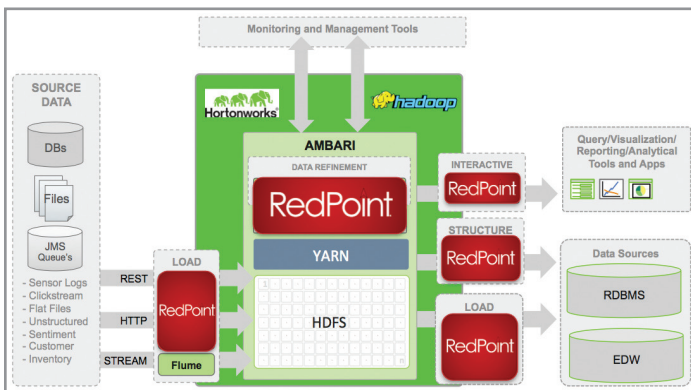
REDPOINT DATA MANAGEMENT HADOOP SETTINGS

Map Reduce	Pig	RedPoint
<pre> Sample MapReduce (small subset of the entire code) public static class MapClass extends Mapper<WordOffset, Text, Text, Text> { private final static String delimit = "\t"; private final static IntWritable private Text word = new Text(); public void map(WordOffset key, throws IOException, InterruptedString line = value.toString(); StringTokenizer itr = new StringTokenizer(line.hasMoreTokens()) { word.set(itr.nextToken()); context.write(word, one); } } </pre>	<pre> Sample Pig script without the UI SET pig.maxCombinedSplitSize 1000; SET pig.splitCombination 1; A = LOAD '/testdata/pg/*'; B = FOREACH A GENERATE FLATTEN(A); C = FOREACH B GENERATE UPPER(C); D = GROUP C BY word; E = FOREACH D GENERATE COALESCE(D.group, D.key); F = ORDER E BY occurrence; STORE F INTO '/user/cleopatra'; </pre>	
>150 Lines of MR Code	~50 Lines of Script Code	0 Lines of Code
6 hours of development	3 hours of development	15 min. of development
6 minutes runtime	15 minutes runtime	3 minutes runtime
Extensive optimization needed	User Defined Functions required prior to running script	No tuning or optimization required

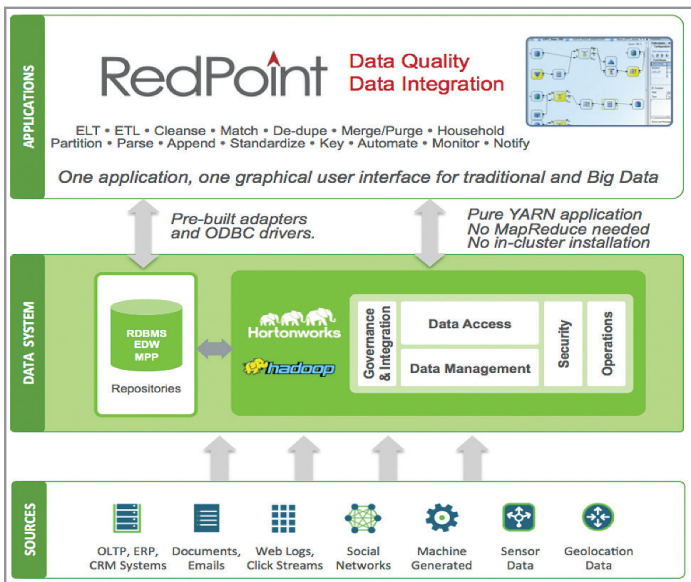
REDPOINT'S PROJECT GUTENBERG TEST



TYPICAL HADOOP ARCHITECTURE WITHOUT REDPOINT



TYPICAL HADOOP ARCHITECTURE WITH REDPOINT



REDPOINT IN A HORTONWORKS ENVIRONMENT

Performance Benchmarks

RedPoint tested its capabilities using Project Gutenberg, a common benchmarking exercise for Hadoop that basically involves counting words to create a concordance.

First, the RedPoint team wrote MapReduce code to execute the Project Gutenberg task, starting with the "Word Count" example. The coding itself took six hours (150 lines of code had to be written), and the task took three hours to execute (although after several more hours of optimization work execution time was reduced).

Next, RedPoint tried using Pig – a higher order of language that promises users an easier experience by writing PigLatin script as opposed to Java. This time it took three hours to create the code (50 lines had to be written), and some Java programming was still required. But execution only took 15 minutes.

Finally, RedPoint used its own product to perform the same task. With RedPoint, building the job in the graphical user interface took only 15 minutes, and execution was finished in three minutes.

These results clearly illustrate the efficiency gains that come from using RedPoint to perform data management functions in Hadoop.

With RedPoint, building the job in the graphical user interface took only 15 minutes, and execution was finished in three minutes.

Benefits of RedPoint Data Management for Hadoop

RedPoint users enjoy benefits that other products don't provide:

- Skills gap elimination – no need for MapReduce programmers
- Use of existing DBA and data analysts – can immediately start working on Hadoop projects
- Faster speed to value with Hadoop, and wider adoption
- Lower total cost of ownership of Hadoop
- Easier to work across Hadoop and traditional data repositories – one product handles both
- Confidence from using a product already proven to deliver enterprise-quality data quality and governance in mature environments

Who Needs RedPoint's Data Management Solution?

- ▶ Companies with new Hadoop initiatives that need help getting started should consider RedPoint.
- ▶ Companies already investing heavily in Big Data analytics technologies, but which are stuck due to the shortage of skilled resources, should consider RedPoint.
- ▶ Large organizations focused on "operational offloading" (moving data from expensive traditional data repositories to more cost effective Hadoop clusters) should consider RedPoint.
- ▶ Finally, companies landing data from outside of the organization into Hadoop, making data quality and data governance even more important, should consider RedPoint.

About RedPoint Global Inc. RedPoint Global offers a comprehensive set of world-class ETL, data quality and data integration applications that operate in and across both traditional and Hadoop 2.0 / YARN environments. The company also offers data-driven customer engagement solutions that help companies derive insights from customer behaviors and create consistent, relevant and precise messaging across any and all channels required. All RedPoint applications offer a unique visual user interface that eliminates the need for programming skills, allowing enterprises to utilize all data to achieve their strategic business goals.

RedPoint Global was selected as one of the "100 Most Promising Big Data Companies" out of 2,000 vendors. The selection was based on the recommendations of a group of CEOs, CIOs, VCs, industry analysts and CIO Review editors. To read more, visit <http://bit.ly/RedPoint-BigData100>.



Summary

For organizations eager to make Hadoop part of their data ecosystem, but intimidated by the prospect of finding or developing MapReduce skills, RedPoint is the answer. RedPoint is the only pure YARN data quality and data integration application on the market today.

With RedPoint:

- ▶ All of RedPoint's top-rated data quality and data integration functions are available for Hadoop
- ▶ Functions are performed in the Hadoop cluster itself
- ▶ Absolutely no MapReduce skills needed
- ▶ Data quality and integration processes execute efficiently
- ▶ Data doesn't need to be moved out of Hadoop
- ▶ No software needs to be installed in the cluster itself
- ▶ Users can manage data in both traditional and Hadoop repositories with a single product

RANKED
#1

In Customer Surveys for:

- Customer or Party Data
- Processing Speed
- Match Quality
- Ease of Use

RedPoint

REDPOINT GLOBAL INC.

36 WASHINGTON ST., SUITE 120, WELLESLEY HILLS, MA 02481 USA

+1 781 725 0250 | www.redpoint.net