Rackspace Cloud
Big Data Platform
On-demand Big Data processing platform
# Table of Contents

Introduction 1  
Challenges of Managing Hadoop 2  
Faster Time to Insight 3  
Easy Deployment for Your Use Case, Using the Help Me Choose Task 4  
Simplified Management 5  
Bare-Metal Performance in the Cloud 7  
Local Storage 8  
Comparing Rackspace Cloud Big Data Platform to Unmanaged Big Data Platform Hosting 9  
Use Cases 10  
Support Levels 12  
Conclusion 13  
About Rackspace 14
Introduction

Big data is becoming a high priority for modern businesses. It enables companies to derive value from very large volumes of aggregated and unstructured data. This opens up a range of cutting-edge use cases, such as social media sentiment analysis and clickstream analysis. Many businesses use Apache Hadoop™ to underpin their big data initiatives. As an open source software framework, Hadoop enables distributed processing of large data sets across clusters of computers with simple programming models. However, Hadoop is notoriously difficult to stand up and requires expensive, difficult-to-find expertise.

The Rackspace Cloud Big Data Platform is an on-demand Hadoop-based service that uses the popular Hortonworks® Data Platform for an easy, fast and managed big data ecosystem that allows businesses to focus on what they do best without the pain of standing up Hadoop, maintaining clusters and troubleshooting jobs.

To help businesses get started with Hadoop faster and scale rapidly to meet demands, the Rackspace Cloud Big Data Platform environment comes installed and configured with the supported Hadoop components, along with many tools such as API access and control panel — all backed by Fanatical Support®.

As a result, the Rackspace Cloud Big Data Platform provides:

1. Faster Time to Insights
2. Simplified Management
3. Bare-Metal Performance in the Cloud
4. Flexibility
5. Full Support from Rackspace Certified Engineers

By providing a scalable platform that can be configured quickly, the Rackspace Cloud Big Data Platform is ideal for customers looking to get real business value out of their data without lengthy deployments or commitments. The platform also minimizes the expense, hassle and risk of hiring expensive specialists to configure, optimize and maintain rapidly changing Hadoop technologies.
Challenges of Managing Hadoop

Hadoop’s ability to efficiently store and process large, diverse data sets is attractive to business leaders hoping to make more informed, data-driven decisions. That said, Hadoop can be difficult to design, configure and deploy. Many companies that attempt to deploy Hadoop on their own find that they need help from a third party because of several common challenges.

UNIQUE FEATURES OF HADOOP CREATE A CHALLENGING CYCLE

Hadoop is unusual: As a distributed system with no centralized storage, configuration and maintenance are both complex and time-consuming. At the same time, experienced architecture experts are expensive and difficult to find. Meanwhile, the open source community rapidly iterates on the Hadoop framework, leading to ever-evolving best practices. These complexities create ongoing challenges. Even if businesses can find experts, they may spend months working through lengthy procurement cycles, only to find that the architecture of Hadoop changes right before they’re ready to buy. That sends architects back to the drawing board, and leaves many companies behind schedule, unable to successfully stand up their Hadoop deployment.

CAPITAL AND OPERATING COSTS CAN BE PROHIBITIVE

Implementing and managing a Hadoop-based big data solution can be more expensive than first anticipated. While Hadoop can leverage commodity hardware, the need to scale the hardware to meet big data demands is, by itself, a major investment for many enterprises. Then there are the operational costs, as leveraging Hadoop often requires businesses to retrain existing staff and hire Hadoop-experienced engineers to monitor and maintain the system.

UNIQUE FEATURES OF HADOOP CREATE A CHALLENGING CYCLE

Hadoop is unusual: As a distributed system with no centralized storage, configuration and maintenance are both complex and time-consuming. At the
Faster Time to Insight

The Rackspace Cloud Big Data Platform makes it easier for businesses to quickly deploy and optimize Hadoop. Rackspace Public Cloud customers can spin up Hadoop clusters quickly through the Rackspace Cloud Control Panel — from small clusters to clusters with hundreds of nodes.

Rackspace Cloud’s Big Data Platform cluster creation interface.

Once the cluster is created, users can interact with their Hadoop cluster and the ecosystem tools within a short time frame. All the components of the distribution are already installed and ready to use, and Rackspace’s Hadoop specialists are available to provide assistance 24x7x365.

Every cluster includes a gateway node, a primary name node, and a secondary name node for redundancy. Rackspace tunes the default replication factor for Hadoop based on the number of nodes provisioned when the cluster is created.
Easy Deployment for Your Use Case, Using the Help Me Choose Tool

The Help Me Choose tool is a helpful feature on the Rackspace Public Cloud control panel that helps users match the right deployment of processing power, memory, disk and data nodes to a specific use case. Every use case dictates a unique design, but users may not always know how to best translate the requirements of business operations to the necessary infrastructure. In the past, for example, users would have to perform a complex calculation to determine how many data nodes they would need for a particular use case.

With the Help Me Choose tool, users simply input the size of their data and define the type of use case — clickstream data, application logs or sentiment analysis — as well as the number of sentiments, logs, etc. that they need to intake. The system then recommends the number, size and speed of clusters for the given use case.

Rackspace Cloud’s Big Data Platform Help Me Choose tool.

As a result, even if users have no idea how to architect Hadoop, the Help Me Choose tool simplifies the process by providing a recommendation on how to size the cluster.
Simplified Management

With the Rackspace Cloud Big Data Platform, Rackspace specialists assist with managing the platform throughout the Hadoop lifecycle of configuration, deployment, maintenance and monitoring, so internal teams don’t have to. Rackspace works directly with Hortonworks, so customers have a single point of contact to resolve any issues that arise.

Specialists from both companies have single-pane visibility into both the infrastructure and platform/application layer, which eliminates the possibility of finger pointing between various infrastructure and platform vendors.

In addition, Rackspace services are backed with Fanatical Support®. Rackspace’s Hadoop specialists and engineers are available 24x7x365 for everything from scalability assistance to query optimization.

Examples of services included with Hadoop:

<table>
<thead>
<tr>
<th>DEPLOYMENT</th>
<th>MAINTENANCE</th>
<th>OPTIMIZATION</th>
<th>BACKUPS AND DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture design and optimization</td>
<td>Advanced administration, monitoring and alerting</td>
<td>Performance tuning and issue diagnosis</td>
<td>Rebooting of name node and master services</td>
</tr>
<tr>
<td>Cluster configuration</td>
<td>Managed patching and updates</td>
<td>Query optimization</td>
<td>Disaster Recovery (DR) and business continuity</td>
</tr>
<tr>
<td>Data migration assistance</td>
<td>Upgrades of Hadoop versions</td>
<td>Compactions</td>
<td>Recommend and plan replication to DR site</td>
</tr>
<tr>
<td>Security configuration (ACL, accounts, etc.)</td>
<td>Cluster management</td>
<td></td>
<td>Participate in DNS Management for failover to DR site</td>
</tr>
<tr>
<td></td>
<td>Job execution</td>
<td></td>
<td>Participate in DR RTO/RPO requirements</td>
</tr>
<tr>
<td></td>
<td>Diagnosis of performance issues</td>
<td></td>
<td>Restores from backups</td>
</tr>
<tr>
<td></td>
<td>Diagnosis of data loading, processing and query Issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Application development advice</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Direct communication and escalation to Hortonworks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Consultation, recommendation and implementation to solve scaling challenges</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


SUPPORTED TOOLS

Hadoop involves a continually expanding catalogue of tools — all of which require specialized expertise. Fanatical Support® helps businesses manage key Hadoop tools including:

- Apache Hadoop in HDP (HDFS, YARN, MapReduce)
- Apache Tez in HDP
- Apache Hive in HDP
- Apache Pig in HDP
- Apache Sqoop in HDP
- Apache Flume in HDP
- Apache Oozie in HDP
- Apache Storm in HDP
- Apache Falcon
- Apache ZooKeeper
- Apache Kafka
- Cluster Backup to S3
Bare-Metal Performance in the Cloud

The Rackspace Cloud Big Data Platform features Rackspace OnMetal™ offering, which combines the speed of bare-metal servers with the ease and flexibility of cloud-like API provisioning. With OnMetal, businesses can deploy on single-tenant bare-metal servers faster, benefiting from the flexibility, ease, scalability and on-demand pricing of the cloud — and the performance, control and security of dedicated servers.

Unlike traditional cloud products, OnMetal does not have any virtualization tax. Users can be confident they are getting 100 percent of the resources on the host machine, while enjoying the benefits of physical isolation for the data they load on the system.

Businesses can choose between OnMetal for Hadoop or OnMetal for Apache Spark. Rackspace OnMetal for Spark is engineered for enhanced performance, enabling data scientists to iterate interactively with large data sets. With Apache Spark, businesses can leverage the speed of memory-optimized processing to accelerate data discovery.

Spark supports real-time analytics, providing faster computations and better query performance on a subset of data. Once deployed, developers can use Spark to rapidly write applications using Java, Scala, or Python, and use Spark interactively from the Scala and Python shells.
Local Storage

Hadoop is extremely I/O and throughput intensive, so latency is inevitable with network-attached storage that separates compute from storage. Often Hadoop users hit a bottleneck with throughput and I/O before they can fully tax compute and memory resources. The Rackspace Cloud Big Data Platform avoids this problem by providing local, high-performing storage in a traditional JBOD architecture.

Traditional Hadoop has all the drives configured in RAID 0 to provide throughput and I/O, and remove redundancy applied at the application layer (3X). Other providers use unified storage, but doing so degrades performance. Rackspace users can choose to process natively on Cloud Files to save cost, or they can ingest data from Cloud Files (bandwidth free) for higher-performing queries.

Flexibility

The Rackspace Cloud Big Data Platform gives companies flexibility to tailor Hadoop to their unique use cases and adapt as needed. Companies can take advantage of utility pricing, hybrid deployment options, open technologies for data portability, and support for all their Hadoop tools, plus a range of integrations with third-party business intelligence and analytics applications.

COMMITMENT TO OPEN SOURCE

The Rackspace Cloud Big Data Platform is built on OpenStack, the world’s largest open source cloud platform. Hortonworks is similarly committed to driving innovation in the open source community, and is the only provider that works entirely within the Apache Software Foundation process. That means most open source ecosystem vendors can layer on top of the Hortonworks Data Platform. And because both the cloud infrastructure and Hadoop platform are open source, businesses can take advantage of flexible and convenient data portability.

BREADTH AND DEPTH OF TOOLS AND INTEGRATIONS

The Hortonworks platform includes all core Hadoop tools. Hortonworks and Rackspace will continue to keep pace with the rapidly expanding Hadoop ecosystem and its ever-increasing number of new tools.

Big data applications power a wide range of applications and platforms. The full-featured Hortonworks Data Platform environment comes ready to integrate with business intelligence and analytics applications or primary data stores, with out-of-the-box support for SAS®, Tableau®, and Microstrategy®.
Comparing Rackspace Cloud Big Data Platform to Unmanaged Big Data Platform Hosting

Gartner has positioned Rackspace as a leader in Managed Cloud hosting in their Magic Quadrant. Managed Cloud hosting provides valuable benefits throughout the lifecycle of a variety of cloud-based workloads, especially ones that can be as complex and important as big data workloads. The following table outlines some major differences between Unmanaged Big Data Platform Hosting and Rackspace Cloud Big Data Platform:

<table>
<thead>
<tr>
<th></th>
<th>Unmanaged Big Data Hosting</th>
<th>Rackspace Cloud Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DEPLOYMENT PROCESS</strong></td>
<td>Customer handles deployment • Need for significant investment in Hadoop expertise • Manual configuration of storage, networking, security, monitoring and more • Manual testing and optimization, with a tradeoff between timely deployment and future scalability</td>
<td>Rackspace oversees deployment • Rackspace engineers gather customer requirements and optimize based on extensive experience and up-to-date Hadoop best practices • Rackspace handles configuration and deployment</td>
</tr>
<tr>
<td><strong>DEPLOYMENT OPTIONS</strong></td>
<td>• Limited deployment flexibility (usually a choice between public cloud or dedicated servers)</td>
<td>Multiple deployment options architected specifically for Hadoop: • Public cloud • OnMetal • Private cloud • Dedicated servers (Managed Big Data)</td>
</tr>
<tr>
<td><strong>DATA PORTABILITY</strong></td>
<td>• Expensive proprietary technologies and risk of vendor lock-in</td>
<td>• Infrastructure built on OpenStack; Hortonworks platform works 100% within the Apache Software Foundation process</td>
</tr>
<tr>
<td><strong>PERFORMANCE</strong></td>
<td>• Generic servers not optimized for high-I/O workloads like Hadoop • Noisy-neighbor problems on public cloud create inconsistency • Network bottlenecks without 10GB infrastructure</td>
<td>• OnMetal ensures bare-metal Hadoop performance with cloud simplicity, flexibility and scalability. • 10GB networking prevents latency bottlenecks</td>
</tr>
<tr>
<td><strong>INFRASTRUCTURE MAINTENANCE AND MONITORING</strong></td>
<td>• Customer allocates significant resources to monitor server performance, maintenance, patching, etc.</td>
<td>• Rackspace engineers proactively monitor and maintain the network and servers 24x7x365</td>
</tr>
<tr>
<td><strong>HADOOP CLUSTER AND JOB MANAGEMENT</strong></td>
<td>• Customer hires, outsources, or reallocates Hadoop specialists to maintain clusters, troubleshoot jobs, etc.</td>
<td>• Rackspace Hadoop engineers manage and troubleshoot cluster operation, job execution, name node rebooting, data ingestion, and tool activity as necessary.</td>
</tr>
<tr>
<td><strong>ECOSYSTEM SUPPORT</strong></td>
<td>• Limited or no support for key Hadoop tools like Spark, Hive and Pig</td>
<td>• Support for most primary ecosystem components</td>
</tr>
<tr>
<td><strong>VENDOR ACCOUNTABILITY</strong></td>
<td>• Separate infrastructure and platform providers can point fingers when things go wrong</td>
<td>• Hortonworks platform and Rackspace infrastructure support on single support contract, giving customer a single POC</td>
</tr>
</tbody>
</table>

Table above is a comparison based on the general activities that would need to be performed on a big data Platform.
Use Cases

The Rackspace Cloud Big Data Platform uses the popular Hortonworks Data Platform as the focal Hadoop distribution to help businesses address a variety of common use cases, including:

- Social Media Data and Sentiment Analysis
- Streaming Workloads
- Clickstream Analysis
- Proof of Concepts

SOCIAL MEDIA DATA AND SENTIMENT ANALYSIS

With Hadoop, businesses can quickly mine, archive and analyze many thousands of social media conversations for sentiment data (e.g., the ratio of positive comments to negative comments). This enables businesses to understand how the public feels about their products or services, their brand, or their competitors, and track how those opinions change over time. With the help of tools like Apache Spark, Storm and Kafka, businesses can also process streaming social media workloads for real-time analytics. This enables businesses to make fast, informed decisions that can turn data into revenue and improved customer satisfaction.

By obviating the need to hire Hadoop experts, Rackspace Cloud Big Data Platform can make social analytics projects more feasible for both IT and marketing. For teams with limited Hadoop experience, Rackspace specialists are available to help users leverage tools such as Flume, HCatalog and Hive. In addition, the OnMetal offering is particularly well suited for Spark and real-time analytics workloads because it combines fast provisioning with bare-metal performance.

CLICKSTREAM ANALYSIS

Using Hadoop, businesses can efficiently store and process large volumes of clickstream data, allowing them to visualize and understand how visitors are scrolling, viewing and clicking through web pages. Companies can then optimize page design and information architecture, or execute tactics such as targeted offers and display ads to help improve sales or lead generation.

Because the Rackspace Cloud Big Data Platform allows for fast, easy Hadoop deployment, marketers can speed up customer intelligence initiatives like clickstream analysis to realize these benefits, while reducing strain on IT resources. In addition, Rackspace Cloud Big Data Platform’s pay-per-use model enables customers to spin up and down as needed, which is especially
useful for batch processing analytics. For instance, customers with predictable usage patterns — like compiling monthly clickstream analytics reports — can reduce costs by having clusters of suitable sizes available only at times when the data needs to be crunched.

PROOF OF CONCEPTS
The Rackspace Cloud Big Data Platform is ideal for proof of concepts, dev clusters and testing use cases because it is so easy to spin up and manage. Businesses can self-provision up to three clusters using the Rackspace Cloud Control Panel, and be fully operational quickly — with support from Rackspace’s Hadoop specialists available anytime, 24x7x365.
Support Levels

The Rackspace Cloud Big Data Platform control plane is the Cloud Big Data Platform API, which is used to create, manage, and delete clusters. Rackspace provides a 99.9% availability guarantee of the control plane in any given monthly billing period.

The data plane includes the clusters created via the control plane, plus supporting systems and services required for proper functioning and availability of the clusters. Customers manage and have full control over their clusters, and Rackspace provides guarantees for the network, data center infrastructure and cluster hosts:

- **Network**: 100% availability guarantee for data center network in any given monthly billing period, excluding scheduled or emergency maintenance.

- **Data Center Infrastructure**: 100% guarantee for the functioning of data center HVAC (Heating, Ventilation and Air Conditioning) and power in any given monthly billing period, excluding scheduled or emergency maintenance.

- **Rackspace provides a guarantee for the functioning of all cluster hosts including the hypervisor**: If a cluster host fails, Rackspace’s guarantee means that restoration or repair will be complete within one hour of problem identification.

- **Hadoop Cluster Operation**: Rackspace goes beyond infrastructure to make sure that the vital components of Hadoop operation are working correctly. Rackspace manages and can troubleshoot cluster operation, job execution, name node rebooting, data ingestion, and tool activity.
Conclusion

Seize the value that big data insights provide your business without the traditional lengthy deployment and implementation process of Hadoop. Rackspace can assist in making big data easy to deploy on a fast platform optimized for your workload. The Rackspace Cloud Big Data Platform uses the popular Hortonworks Data Platform to allow companies to reap all the benefits of big data and each environment comes with supported Hadoop components installed and access to many tools, so companies can get started with Hadoop quickly and scale on demand as their business needs change. And don’t forget, big data customers can rest easy knowing that Rackspace Big Data specialists provide 24x7x365 support and are managing their ecosystem around the clock.

To learn more about the Rackspace Cloud Big Data Platform, contact a Rackspace sales representative at 1-888-480-7667. Or, visit www.rackspace.com/en-us/cloud/big-data.
About Rackspace

Rackspace (NYSE: RAX), the #1 managed cloud company, helps businesses tap the power of cloud computing without the challenge and expense of managing complex IT infrastructure and application platforms on their own. Rackspace engineers deliver specialized expertise on top of leading technologies developed by OpenStack®, Microsoft®, VMware® and others, through a results-obsessed service known as Fanatical Support®.

GLOBAL OFFICES

Headquarters Rackspace, Inc.
1 Fanatical Place | Windcrest, Texas 78218 | 1-800-961-2888 | Intl: +1 210 312 4700
www.rackspace.com

UK Office
Rackspace Ltd.
5 Millington Road
Hyde Park Hayes
Middlesex, UB3 4AZ
Phone: 0800-988-0100
Intl: +44 (0)20 8734 2600
www.rackspace.co.uk

Benelux Office
Rackspace Benelux B.V.
Teleportboulevard 110
1043 EJ Amsterdam
Phone: 0800 8899 00 33
Intl: +31 (0)20 753 32 01
www.rackspace.nl

Hong Kong Office
9/F, Cambridge House, Taikoo Place
979 King's Road,
Quarry Bay, Hong Kong
Sales: +852 3752 6488
Support +852 3752 6464
www.rackspace.com.hk

Australia Office
Rackspace Hosting Australia PTY LTD
Level 1
37 Pitt Street
Sydney, NSW 2000
Australia

© 2015 Rackspace US, Inc. All rights reserved.

This white paper is for informational purposes only. The information contained in this document represents the current view on the issues discussed as of the date of publication and is provided “AS IS.” RACKSPACE MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND, EXPRESS OR IMPLIED, AS TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS DOCUMENT AND RESERVES THE RIGHT TO MAKE CHANGES TO SPECIFICATIONS AND PRODUCT/SERVICES DESCRIPTION AT ANY TIME WITHOUT NOTICE. USERS MUST TAKE FULL RESPONSIBILITY FOR APPLICATION OF ANY SERVICES AND/OR PROCESSES MENTIONED HEREIN. EXCEPT AS SET FORTH IN RACKSPACE GENERAL TERMS AND CONDITIONS, CLOUD TERMS OF SERVICE AND/OR OTHER AGREEMENT YOU SIGN WITH RACKSPACE, RACKSPACE ASSUMES NO LIABILITY WHATSOEVER, AND DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO ITS SERVICES INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT.

Except as expressly provided in any written license agreement from Rackspace, the furnishing of this document does not give you any license to patents, trademarks, copyrights, or other intellectual property.

Rackspace, Fanatical Support, and/or other Rackspace marks mentioned in this document are either registered service marks or service marks of Rackspace US, Inc. in the United States and/or other countries. OpenStack is either a registered trademark or trademark of OpenStack, LLC in the United States and/or other countries. Third-party trademarks and tradenames appearing in this document are the property of their respective owners. Such third-party trademarks have been printed in caps or initial caps and are used for referential purposes only. We do not intend our use or display of other companies' tradenames, trademarks, or service marks to imply a relationship with, or endorsement or sponsorship of us by, these other companies.