TechTarget | SearchBusinessAnalytics

E-guide

TechTarget | Research Peer Network

Hadoop Big Data Platforms Buyer's Guide – part 3

Your expert guide to Hadoop big data platforms

In this e-guide

A look at Amazon Elastic MapReduce cloud-based Hadoop

Abie Reifer, DecisionWorx

The Amazon Elastic MapReduce Web service offers a managed Hadoop framework that enables users to distribute and process big data across dynamically scalable Amazon EC2 instances.

Amazon Elastic MapReduce provides users access to a cloud-based Hadoop implementation for analyzing and processing large amounts of data. Built on top of Amazon's cloud services, EMR leverages Amazon's Elastic Compute Cloud and Simple Storage services, enabling users to provision a Hadoop cluster quickly.

Amazon's cloud elasticity and setup tools also give users a way to temporarily scale up a cloud-based Hadoop cluster for short-term increased computing capacity. Amazon EMR lets users focus on the design of their workflow without the distractions of configuring a Hadoop cluster. As with other Amazon cloud services, users pay for only what they use.

In this e-guide

Amazon Elastic MapReduce features

The current version of Amazon EMR, 4.3.0, bundles several open source applications, a set of components for users to monitor and manage cluster resources, and components that enable application and cluster interoperability with other services.

The following open source applications come bundled as part of Amazon:

- Apache Hadoop 2.7.1
- Apache Hive 1.0.0
- Apache Mahout 0.11.0
- Apache Pig 0.14.0
- Apache Spark
- Hue
- Ganglia 3.7.2

AWS Elastic MapReduce also provides users with the option of using MapR's Hadoop distribution in place of Apache Hadoop.

In this e-guide

The EMR Web service supports several file system storage options used for data processing. These include Hadoop Data File System for local and remote file systems and S3 buckets using EMR File System as well as other Amazon data services. Amazon EMR also integrates with several data services, including Amazon Dynamo DB, a fast NoSQL database; Amazon Relational Database Service; Amazon Glacier; Amazon Redshift, a petabyte data warehouse service; and AWS Data Pipeline, a service used to move data between AWS services.

Other AWS Elastic MapReduce features enable users to perform the following tasks:

Provision an EMR cluster. An EMR management console helps users quickly navigate through the process of spinning up and autoconfiguring an EMR instance. Through the console, users select the applications from the EMR bundle to install, the types of server instances to use for the cluster nodes, and the security access policies and controls for the cluster.

Load data into the cluster. Users with typical size data needs can transfer data to an Amazon S3 bucket to be available to the cluster for processing. Users with petabyte-scale needs may opt to use AWS Snowball, a secure, high-speed appliance that's shipped to the user, or AWS Direct Connect, an established high-speed data connection between AWS and the user's data center.

Monitor and manage. Amazon EMR collects metrics that are used to track progress and measure the health of a cluster. While these metrics can be accessed through the command line interface, software developer kits or APIs, they can also be viewed through the EMR management console. Additionally, Amazon CloudWatch can also be used along with Apache Ganglia to monitor the cluster and set alarms on events triggered by these metrics.

AWS Elastic MapReduce pricing

Amazon's EMR pricing model is based on the company's approach to pricing for its other Web services. Users pay per amount of time and the types of instance servers used. Spot instances can also be used for some or all of the nodes in a cluster, providing users with a level of elasticity that can be changed based on their dynamic computing needs.

Amazon provides developers with a wide range of online technical documentation, guides, tutorials and sample code.

Next article

In this e-guide

Learn more about the Cloudera Hadoop distribution

Abie Reifer, DecisionWorx

Cloudera distribution, including Apache Hadoop, provides an analytics platform and the latest open source technologies to store, process, discover, model and serve large amounts of data.

CDH, the Cloudera Hadoop distribution, includes several related open source projects, such as Impala and Search. It also provides security and integration with several hardware and software products.

The Impala framework in Cloudera distribution including Apache Hadoop allows users to execute interactive SQL queries directly against data stored in Hadoop Distributed File System (HDFS), Apache HBase or the Amazon Simple Storage Service. Impala uses several technologies and components from Hive, including SQL syntax (Hive SQL), Open Data Base Connectivity driver and Impala's Query UI (Hue is also used by Hive).

As part of CDH, Cloudera Search incorporates Apache Solr, a data indexing and search platform based on Lucene. The integration of this technology as part of CDH provides users with near real-time indexing of and access to data directly stored in Hadoop and HBase. Solr indexing and search technology

enables users to perform complex textual searches while requiring little or no SQL or programming skills. Solr also allows for queries to be performed directly against the Hadoop data store, removing the need to move large data sets to perform complex queries.

Other related open source projects included in CDH from Apache are Flume, HBASE, Hive, Hue, Oozie, Spark, Sqoop and Sentry (incubating).

Editions of the Cloudera Hadoop distribution

Cloudera offers several implementation editions of CDH that provide differing levels of cluster and service management capabilities as well as different levels of support:

Cloudera Express is free to use and includes CDH, as well as core features of Cloudera Manager.

Cloudera Manager provides CDH administrators with an intuitive Web-based management console to deploy, manage, monitor and diagnose issues with CDH deployments. The tool also includes an API that can be used to programmatically configure the system and collect metric and health information about a CDH cluster.

In this e-guide

Cloudera Enterprise is a licensed edition that provides extended capabilities to CDH with the inclusion of additional advanced features from Cloudera Manager and Navigator. Technical support options are also available to customers that have purchased an enterprise license. Cloudera Enterprise is available in three editions, each offering varying levels of service management capabilities:

- The Basic edition provides management capabilities to support a cluster running core CDH services that include HDFS, Hive, Hue, MapReduce, Oozie, Sqoop, Yet Another Resource Negotiator (YARN) and ZooKeeper.
- The Flex edition supports the management of a cluster running core CDH services plus one of the following: Accumulo, HBase, Impala, Navigator, Solr or Spark.
- The Data Hub edition supports the management of a cluster running core CDH services plus any of the following: Accumulo, HBase, Impala, Navigator, Solr or Spark.

Cloudera Manager Advanced Features add the following to the core product capabilities provided with Cloudera Express: operational reporting, quota management, configuration history and rollbacks, rolling updates and service restarts, direct AD Kerberos integration, Lightweight Directory Access Protocol integration, Simple Network Management Protocol support, support integration with scheduled diagnostics and automated disaster recovery.

In this e-guide

Cloudera Navigator, which is available for only Flex and Data Hub Editions, enables users to manage data security and governance for the CDH platform, supporting an organization's compliance and regulatory requirements. The tool can be used to help data managers, analysts and administrators explore the large amounts of data in Hadoop, as well as to more easily manage encryption keys used to secure data residing in the CDH clusters.

Cloudera Hadoop distribution products are supported on Red Hat Enterprise Linux/CentOS 6.6 (in Security Enhanced Linux mode), 6.7 and 7.1 and Oracle Enterprise Linux 7.

Cloudera offers users several options for installing and implementing its products: QuickStartVM provides users with a free to use virtual machine -- VMware, VirtualBox or Kernel-based VM -- running CentOS 6.4 and a single Apache Hadoop cluster along with example data, queries, scripts and Cloudera Manager to manage the cluster. Cloudera QuickStart VMs are intended for demo purposes only.

Cloudera Manager is used for installing and managing Cloudera implementations -- both Express and Enterprise Editions. A license is required to install the Enterprise edition. Installation of Cloudera Express provides users with an optional 60-day trial of Cloudera Enterprise.

Cloudera Director provides self-service users with the ability to deploy and manage Cloudera Enterprise in a variety of cloud environments.

For users interested in manually installing the product, Cloudera provides a version for download that can be run on the operating systems mentioned above.

Cloudera Hadoop distribution licensing, pricing and support

Cloudera Enterprise annual subscriptions vary based on the edition or tier purchased and the number of nodes being run. Contact Cloudera for detailed pricing.

Cloudera offers several support options to organizations that have purchased Enterprise edition licenses. Support isn't available to users of Cloudera Express. Business hour and 24/7 support options are available for all enterprise license holders. Premium support options, which include a 15-minute response time for critical issues, are only available to organizations with the Flex or Data Hub edition licenses.

Cloudera provides training and certification through Cloudera University, which offers both on-demand and private training. Courses and certifications are offered in three tracks for developers, administrators and analysts.

Next article

In this e-guide

Inside the Hortonworks open enterprise Hadoop distribution

Abie Reifer, DecisionWorx

The Hortonworks Data Platform consists entirely of projects built through the Apache Software Foundation and provides an open source environment for data collection, processing and analysis.

The Hortonworks Data Platform enables users to store, process and analyze massive volumes of data from many sources and formats. At its core, the scalable open enterprise Hadoop platform includes Hadoop Distributed File System, a fault-tolerant storage system for processing large amounts of data in a variety of formats and YARN.

YARN (Yet Another Resource Negotiator), a core part of the open source Hadoop project, provides centralized resource management for Hadoop's data processing workload across various processing methods, including interactive SQL, real-time streaming, data science and batch processing. Other enterprise-grade functions supported include data governance, security and common operations support.

With its recent announcement of release 2.4, Hortonworks indicated it will be providing more frequent releases as part of its Extended HDP services. This will provide customers access to interim and more frequent releases and

innovations of non-Core Hadoop modules -- e.g., Hive, HBase, Storm and Spark, among others.

HDP Core modules that include Hadoop Distributed File System, YARN and MapReduce will continue to be provided on a single-release-per-year schedule aligned with the Open Data Platform Initiative core Apache-compatible version.

This approach will enable customers who use Hadoop Core modules for critical functions such as data storage to stabilize on less-frequent releases of the more mature core modules. At the same time, this strategy will provide more frequent releases to other customers who are interested in benefiting from those more rapidly evolving Hadoop modules.

HDP 2.4 includes Apache Hadoop 2.7.1 (Core HDP modules) as well as Spark 1.6, HBase 1.1.2, Kafka 0.9.0 and Ambari 2.2.1 as the Extended HDP services.

Hortonworks DataFlow (HDF), which is a separate product, works with HDP and is designed to solve the challenges of automating all types of real-time data flows as well as collecting and curating real-time business insights and actions derived from any data from anywhere. The product is powered by the NiFi Apache open source project that's intended to address the challenges presented by the Internet of Anything (IoAT). Unlike the Internet of Things, which is associated with just sensors and machine data, IoAT includes clickstream data and social stream data.

In this e-guide

Hortonworks open enterprise Hadoop offers three installation options:

- Hortonworks Sandbox on virtual machine, a virtualized environment that operates on Mac or Windows in VMware or VirtualBox and provides a personal Apache Hadoop environment intended for prototyping and training purposes.
- Hortonworks Sandbox in the cloud, a cloud-based HDP implementation currently available in Microsoft Azure with a one-month free trial.
- HDP 2.3.2 Ready for the Enterprise, which provides automated installation on Linux and Unix environments using Ambari. Additional features include manual installation using RPM Package Manager for Unix and Linux environments, cloud installation using Cloudbreak for Azure, and Amazon Web Services and OpenStack with Windows installation for Windows Server 2008 and 2012.

Hortonworks Data Platform licensing and support

Aside from optional add-ons and third-party components, Hortonworks Data Platform components are covered under the Apache 2.0 license.

Hortonworks Hadoop offers the following support subscriptions designed to cover the entire lifecycle from proof-of-concept to production deployment and operations:

HDP Jumpstart, which is intended for early-stage data development work. It provides users with a six-month support term for three named contacts during normal business hours. The response commitment time for all severity types is one business day.

HDP Enterprise, which is intended for business-critical operational support. It provides users with a one-year term and supports named contacts based on cluster size. Support is provided 24/7 via phone and Web requests, with a one-hour response time for severity 1 issues, four hours for severity 2 issues, eight hours for severity 3 issues and one business day for severity 4 issues.

HDP Enterprise Plus provides the same level of support as HDP Enterprise, but includes support for these additional modules that aren't included as part of HDP Enterprise support: Accumulo, Atlas, Storm, Ranger, Spark, Kafka and Cloudbreak.

HDP Enterprise Premier Support offers clients designated on-site and personalized support. Premier is available for only clients with existing active enterprise-level support for HDP or HDF.

Contact Hortonworks for pricing information.

Next article

In this e-guide

Inside the IBM BigInsights platform for big data management

Abie Reifer, DecisionWorx

The latest version of IBM BigInsights offers several value-add services that can be used with its core distribution of open source Hadoop for managing big data.

IBM BigInsights combines its enterprise capabilities and industry-standard Hadoop components into a single platform, enabling users to manage and analyze large volumes of structured and unstructured data.

IBM BigInsights features several advanced analytics capabilities, including sophisticated text analytics; BigSheets for advanced data exploration; and Big SQL, which enables SQL access to data in a Hadoop cluster. Added-value enterprise capabilities are designed to enhance and simplify application development and system implementation, as well as provide features that improve performance, scalability, reliability, security and administration.

IBM BigInsights release 4.1 includes the IBM Open Platform with Apache Hadoop, as well as several prepackaged value-add modules containing proprietary advanced enterprise-grade features.

IBM Open Platform with Apache Hadoop, IBM's core distribution of open source Hadoop, includes the following Apache components: Ambari (2.1), Apache Kafka (0.8.2), Flume (1.5.2),Ganglia (3.1.7), Hadoop (2.7.1), HBase (1.1.1), Hive (1.2.1), Knox (0.6.0), Lucene (4.7.0), Nagios (3.5.1), Oozie (4.2.0), Parquet (4.0), Parquet MR/format (1.6.0/2.2), Pig (0.15.0), Slider (0.80.0), Solr (5.1.0), Spark (1.4.1), Sqoop (1.4.6.), Terada Connector for Hadoop (1.4) and Zookeeper (3.4.6).

BigInsight value-add modules include:

IBM BigInsights Analyst, which provides specific tools for data analysis. The modules include BigInsights Home service, the primary interface used to launch other BigInsights components, as well as Big SQL and BigSheets:

- **Big SQL** is an advanced SQL engine that provides users who have standard SQL query skills with fast query access to data in a Hadoop cluster within a single query, whether it be in Hive, HBase or Hadoop Distributed File System enabled by massively parallel processing technology. The product also supports federated query access to IBM DB2, Oracle, Teradata and Open Database Connectivity sources.
- **BigSheets** lets users explore, transform and perform visualizations on large data sets stored in Hadoop through a spreadsheet-like Web interface. The tool supports fast queries against massive data sets by translating user actions into MapReduce functions against the Hadoop cluster.

IBM BigInsights Data Scientist, which enables users with advanced analytics skills tools to gain further insight into the data in the cluster. In addition to components provided as part of the Analyst module, the following tools are also included:

- **Big R**, which provides users familiar with the R language a set of libraries, enabling them to develop and use R language functions on data residing in the IBM BigInsights cluster. This tool lets users perform complex operations and queries using R against large data sets by hiding some of the complexity of writing MapReduce functions.
- **Text Analytics**, which is a powerful and intuitive tool for extracting information from unstructured and semi-structured text.
- **SystemML**, which provides users with a tool to use an R-like syntax to perform statistical functions and machine learning constructs. The tools enable the algorithms to be executed in a distributed fashion across nodes of a cluster using MapReduce or Spark (in memory). IBM contributed SystemML to the open source community, and it has been accepted as an Apache Incubator project.

IBM Enterprise Management module, which provides enterprise-grade capabilities to support cluster scaling and performance through parallel computing and application grid management. The module also provides other enterprise features to support cluster security and reliability. IBM Enterprise

In this e-guide

Management includes IBM Spectrum Scale-FPO, a Portable Operating System Interface-compliant file system that can be used in place of Hadoop Distributed File System. This gives administrators more control and improved integration capabilities with other systems in the enterprise. Also included is IBM Platform Symphony, which provides administrators with tools to efficiently manage multiple platform instances as well as enables support for data isolation for multi-tenant environments.

IBM BigInsights for Apache Hadoop, which includes the contents of the three modules noted above.

BigInsights modules operate on Linux servers. Detailed system requirements include operating system and hardware, as well as supported software.

While IBM BigInsights modules can be downloaded and installed on-premises, the company also offers BigInsights on Cloud, Hadoop as a service on IBM's global cloud infrastructure. This option provides users with all the features of BigInsights in a 24/7 managed environment.

IBM BigInsights licensing and distribution

While IBM's Open Platform with Apache Hadoop is available as an open source free-to-use distribution, BigInsight's value-add modules require IBM licensing for purposes other than evaluation. Contact IBM or an IBM Business Partner for detailed pricing and support options.

IBM offers the BigInsights Quick Start evaluation edition of its software for nonproduction use.

IBM is a founding member of the Open Data Platform Initiative, a group of big data industry leaders and vendors that promote technologies based on open source on the Apache Hadoop ecosystem and share in efforts to promote interoperability of big data tools.

Next article

In this e-guide

Inside the MapR Hadoop distribution for managing big data

Abie Reifer, DecisionWorx

The MapR Hadoop distribution replaces HDFS with its proprietary file system, MapR-FS, which is designed to provide more efficient management of data, reliability and ease of use.

The MapR Converged Data Platform supports big data storage and processing through the Apache collection of Hadoop products, as well as its added-value components. These components from MapR Technologies provide several enterprise-grade proprietary tools to better manage and ensure the resiliency and reliability of data in the Hadoop cluster.

These platform components include MapR File System (MapR-FS); MapReduce; and MapR Control System, the product's user interface. The MapR Hadoop distribution includes a complete implementation of the Hadoop APIs, enabling the product to be fully compatible with the Hadoop ecosystem.

MapR-FS is written in C++ -- versus Apache HDFS, which is written in Java -- and serves as the company's proprietary implementation of Hadoop Distributed File System. Unlike HDFS, which follows the write-once-read-many paradigm, MapR-FS is a fully read/write Portable Operating System Interface-compliant file system.

In this e-guide

By supporting industry-standard NFS, users can easily mount a MapR cluster and execute any file-based application directly on the data residing in the cluster. This enables data from nearly any source to be processed and allows for standard tools to be used to directly access data in the cluster without any modifications.

Additionally, unlike other Hadoop distributions, MapR can process distributed files, database tables and event streams all in the same cluster of nodes. This lets organizations run operational tools such as Apache HBase and analytic tools such as Hive or Impala on one cluster, reducing hardware and operational costs.

The latest version of MapR, 5.1., also includes MapR Streams, an event streaming system for big data. This platform is designed to support highly scalable real-time streaming of big data from producers to consumers on their converged platform. MapR claims it's the only big data streaming system to support global event replication at Internet of Things scale and reliability.

Other features of MapR's Converged Data Platform include:

- MapR Snapshots that offer improved data protection by capturing point-in-time snapshots for both files and tables on demand, as well as at regularly scheduled intervals.

- Encryption of data transmitted to, from and within a cluster, as well as strong authorization mechanisms that are designed to improve data security while enabling administrators to have better control over what actions individual users are authorized to perform.
- Out-of-the box, easily configurable mirroring capability that supports disaster recovery.

MapR Hadoop distribution editions

MapR offers Converged Community Edition, an unlimited free-to-use version, and Converged Enterprise Edition, a subscription-based version intended for organizations with business continuity requirements. The Enterprise version includes advanced multi-tenancy capabilities, consistent snapshots, high availability and disaster recovery features, as well as 24/7 commercial support and support for other modules and engines.

The MapR Hadoop distribution offers several training options, including free online, on-demand training as well as instructor-led for-fee training and certifications.

The products can be downloaded and installed on a local server using the GUI installer. The Community Edition is free to use and the Enterprise Edition can be downloaded and used for a 30-day trial period.

The distribution provides a sandbox version that's a self-contained virtual machine, which includes tutorials and demo applications, enabling users to get started quickly with Hadoop and Spark.

MapR in the Cloud provides users the ability to deploy in cloud environments, including Azure, Google Cloud Platform and on Amazon Web Services.

The MapR Hadoop distribution also provides several quick-start solutions, which include prebuilt, templated environments that support use-case scenarios, including self-service data exploration, real-time security log analytics, time series analytics, genome sequencing, data warehouse optimization and analytics, and a recommendation engine.

MapR runs on several versions of Linux, including Red Hat, CentOS, SUSE and Ubuntu. Hardware requirements include 64-bit CPU and 4 GB minimum of memory -- additional memory is required for production environments.

In this e-guide

MapR Hadoop licensing and support

To use MapR products, users are required to agree to the terms of the company's end-user license.

While all users can access a variety of online resource material, Premium Support adds Web and email support and a custom portal. It also provides training, urgent bug fixes, follow-the-sun support and 24/7 phone support for priority 1 issues.

Premium+ Support adds priority queuing of tickets, single-point-of-contact support and options for on-site or remote dedicated support. Contact MapR for support pricing.

Next article

In this e-guide

Inside the Microsoft Azure HDInsight cloud infrastructure

Abie Reifer, DecisionWorx

Azure HDInsight is a cloud implementation of Apache Hadoop that provides a software framework designed for processing, analyzing and reporting on big data.

Microsoft Azure HDInsight is designed to help users quickly and cost-effectively deploy and use Hadoop and other Apache big data analysis and processing products. To support the service, Microsoft leverages its managed Azure cloud infrastructure, enabling users to provision a Hadoop cluster without having to purchase, install and configure the necessary hardware and software. Azure HDInsight also lets users resize the environment on demand by spinning up additional nodes to handle their computing capacity needs -- from terabytes to petabytes.

The service uses the Hortonworks Data Platform (HDP) Hadoop distribution and includes implementations of Apache Spark, HBase, Storm, Pig, Hive, Sqoop, Oozie and Ambari, as well as other Apache products. Additional components can be installed as part of provisioning a cluster by executing scripts. Several scripts are provided by HDInsight to install and configure Hue, Giraph, R and Solr, enabling users to create scripts of their own to install other Apache

components. HDInsight also integrates with business intelligence tools, including Power BI, Excel, SQL Server Analysis Services and SQL Server Reporting Services.

Functions of Microsoft Azure HDInsight

Users can quickly set up an HDI cluster through the Azure portal to specify the cluster type, Hadoop, HBASE, Storm or Spark; the operating system of the cluster, Linux or Windows; and the HDI version to use (as described below). The cluster type the user selects drives the number of base nodes and each of their roles in that cluster type. The HDInsight pricing page provides the detailed layout for each of these cluster types.

HDInsight supports multiple Hadoop cluster versions at any time, with each tied to a specific version of the HDP. The current default, HDInsight version 3.2, is based on HDP version 2.2. This version includes Apache Hadoop & YARN (2.6.0), Apache Tez (0.5.2), Apache Pig (0.14), Apache Hive and HCatalog (0.14.0), Apache HBase (0.98.4), Apache Sqoop (1.4.5), Apache Oozie (4.1.0), Apache Zookeeper (3.4.6), Apache Storm (0.9.3), Apache Mahout (0.9.0), Apache Phoenix (4.2.0) and Apache Spark (1.3.1), as well as other Apache products.

Six additional versions of HDInsight are currently supported: HDI 1.6, HDI 2.1, HDI 3.0, HDI 3.1, HDI 3.3 and HDI 3.4. Each is based on different versions of HDP, respectively: HDP 1.1, HDP 1.3, HDP 2.0, HDP 2.1, HDP 2.3 and HDP 2.4. And each includes different versions of Hadoop and other Apache big data products.

Through the use of the Azure SDK for .NET, developers can integrate their Visual Studio Integrated Development Environment with their HDI cluster by installing the HDInsight Tools for Visual Studio and the Microsoft Hive Open Database Connectivity driver. The SDK allows developers to connect and navigate HDI Insight Hive databases and linked storage accounts for HDInsight clusters, create tables, and create and run Hive queries.

Pricing and support for Azure HDInsight

Customers are billed from the time a cluster is created to when it's deleted. They can estimate their cost based on the Azure features they need using the pricing calculator. Different cluster types -- Hadoop, HBASE, STORM, Spark -- have different minimal node configurations. Pricing is based on hourly charges per node and node instance type -- compute power and memory. There are additional fees for storage and data transfer.

In this e-guide

Microsoft offers a 30-day free trial and a $200 credit to use in Azure. The trial account is decommissioned once the 30-day trial has expired if the user hasn't upgraded to a pay-as-you-go Azure subscription.

Microsoft Azure offers several support subscriptions, including technical support for Hadoop as well as other Azure services. The Hadoop support is backed by Hortonworks, the distributors of the Hadoop distribution deployed in Azure HDInsight.

About the author

Abie Reifer is a technology and strategy leader with extensive experience in system implementations. He is a principal analyst at DecisionWorx and serves in a senior leadership position at a data collection and survey research organization. Previously, he served as CTO of an international telecommunications company and as an advisory strategy consultant to a leading U.S. telecommunications carrier. Reifer has a master's degree in engineering from Columbia University.