Recent computing and business trends have triggered an explosion in the amount of unstructured data that companies generate each day. Big data analytics can give you a competitive advantage. By extracting the knowledge wrapped within unstructured and machine-generated data, your enterprise can make better decisions that drive revenue and reduce costs.

Hadoop has rapidly emerged as the preferred solution for big data analytics across unstructured data. However, the fast-changing Hadoop ecosystem can present challenges to any company that wants to standardize on core functionality and build repeatable processes. It is often up to you to determine which packages to use, provide version control, test for compatibility, and deploy functionality into production. This can lead to many time-consuming failed attempts—and, ultimately, arriving at a dead end.

The EMC® Greenplum® HD product line enables you to take advantage of big data analytics without the overhead and complexity of a project built from scratch. Available in Community and Enterprise editions, as well as Greenplum HD Data Computing Appliance Module, Greenplum HD provides a complete platform, including installation, training, global support, and added value beyond simple packaging of the Apache Hadoop distribution. Greenplum HD makes Hadoop faster, more dependable, and easier to use.

**GREENPLUM HD COMMUNITY EDITION COMPONENTS**

HD Community Edition is a 100 percent open-source certified and supported version of the Apache Hadoop stack that includes Hadoop Distributed File System, MapReduce, Hive, Pig, HBase, and Zookeeper. EMC Greenplum adds fault tolerance for the Name Node and Job Tracker—both single points of failure in Hadoop.

**HADOOP DISTRIBUTED FILE SYSTEM (HDFS)**

HDFS is a Java-based file system that provides scalable and reliable data storage. With industry installations in the thousands of nodes, HDFS has proven to be a solid foundation for any Hadoop deployment.

**MAPREDUCE**

MapReduce is a Hadoop framework for easily writing applications that process large amounts of unstructured and structured data in parallel, in a reliable and fault-tolerant manner. The framework is resilient to hardware failures, handling them transparently from user applications.

**HIVE**

Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop-compatible file systems. This SQL-like interface gives users a row-based storage capability, which, along with compression, results in an improved compression ratio for storing data.
PIG
Pig is the procedural language for processing large, semi-structured data sets using the Hadoop MapReduce platform. It enables developers to more easily write MapReduce jobs by providing an alternative programming language to Java.

HBASE
HBase is a distributed, versioned, column-oriented storage platform that provides random realtime read/write access to big data for user applications.

ZOOKEEPER
Zookeeper is a highly available system for coordinating distributed processes. Distributed applications use Zookeeper to store and mediate updates to key configuration information.

GREENPLUM HD COMMUNITY EDITION FEATURES
A POWERFUL DATA CO-PROCESSING SOLUTION FOR THE ENTERPRISE
The Greenplum Database and Greenplum HD are complementary technologies that combine to deliver a powerful solution for the analytics of structured, semi-structured, and unstructured data. When they use these solutions together, analysts can perform complex, high-speed, interactive analytics using the power of Greenplum SQL, and advanced analytic functions against data within Hadoop and the Greenplum Database. Data access and movement is simplified using external tables that enable fast and easy parallel data sharing. Customers are leveraging this functionality to:

• Perform ad-hoc queries from the Greenplum Database on the data stored in an HDFS, allowing database administrators to use familiar SQL commands to run the queries.
• Leverage Hadoop to transform unstructured and semi-structured data into a structured format that the Greenplum Database can use for high-speed interactive analysis.
• Integrate and process data, pull very large data sets based on a schedule, and process them for loading into the Greenplum Database.
• Run queries that require source data from both the Greenplum Database and the HDFS.

The combined solution delivers the industry’s only complete big data analytics platform.

ENGINEERED FOR THE ENTERPRISE
EMC stands behind its products with rigid SLAs, and carefully tests each component for maturity prior to inclusion in a release. EMC support standards are enterprise-class and deliver end-to-end service from day one.

HIGH-AVAILABILITY SUPPORT
Greenplum HD fault-tolerance capabilities protect users from the single point of failure inherent in Apache Hadoop. Automatic recovery from failures improves on the platform’s availability and lowers your total cost of ownership.

INSTALLATION AND TRAINING SERVICES
EMC delivers an implementation package of installation and training services for Greenplum HD to ensure that the system will perform as expected. Expert planning, design, and implementation services help you quickly realize the value of big data analytics, no matter how simple or complex your environment may be.

Greenplum Hadoop training programs are designed to empower all levels of your organization with specialized offerings and curriculum. Whether you’re a senior executive trying to understand the value and use of Hadoop or a MapReduce programmer building on the Hadoop experience, EMC Greenplum has a program for you.
EMC GREENPLUM HD PARTNER ECOSYSTEM

Industry-leading technology companies have partnered with EMC to create a vibrant and powerful ecosystem, offering additional business intelligence, data transfer, and other technology capabilities on top of EMC’s Hadoop offerings. This breadth of support is a testament to the value EMC brings to Hadoop. Technology companies and enterprises can now extend the trust they have in EMC to the open source data analytics tool.

ENTERPRISE BUSINESS INTELLIGENCE AND ANALYTIC MODELING TOOL SUPPORT

The Greenplum HD Community Edition partner ecosystem includes leading industry innovators in extract, transform, and load (ETL); business intelligence (BI); and analytic toolkits. These partners provide critical components for your success in big data analytics.

MAXIMIZE EMC GREENPLUM BENEFITS WITH EMC GLOBAL SERVICES

EMC delivers a full complement of services for EMC Greenplum Data Computing Division hardware and software to ensure that your system will perform as expected in your environment, while minimizing risk to your business and budget. Expert planning, design, and implementation services help you quickly realize the value of your hardware and software in your environment—no matter how simple or complex. After implementation, EMC’s data migration services can help you plan, design, and safely migrate your critical data over any distance to your new system. EMC will also help you integrate your new system into your information architecture and business intelligence and analytics applications (such as SAS, MicroStrategy, Business Objects, and Tableau), and manage your new environment when it is complete.

Leveraging EMC’s comprehensive data warehousing and business intelligence deployment best practices and guided by our proven methodology, our extensively trained professional services personnel and project management teams accelerate your business results without straining your resources.

EMC GREENPLUM’S DATA COMPUTING PRODUCTS DIVISION

EMC’s Data Computing Products Division is driving the future of data warehousing and analytics with breakthrough products such as EMC Greenplum HD, EMC Greenplum Data Computing Appliance, EMC Greenplum Database, and EMC Greenplum Chorus—the industry’s first Enterprise Data Cloud platform. The division’s products embody the power of open systems, cloud computing, virtualization, and social collaboration—enabling global organizations to gain greater insight and value from their data than ever before.

CONTACT US

To learn more about how EMC Greenplum products, services, and solutions can help you realize big data opportunities, visit us at www.greenplum.com.