IBM SPSS Analytic Server
Version 2.1

Administrator's Guide

Note

Before using this information and the product it supports, read the information in "Notices" on page 13.

Product Information

This edition applies to version 2.1, release 1, modification 0 of IBM SPSS Analytic Server and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Overview

IBM® SPSS® Analytic Server is a solution for big data analytics that combines IBM SPSS technology with big data systems and allows you to work with familiar IBM SPSS user interfaces to solve problems on a previously unattainable scale.

Why big data analytics matters

Data volumes collected by organizations are growing exponentially; for example, financial and retail businesses have all customer transactions for a year (or two years, or ten years), telco providers have call data records (CDR) and device sensor readings, and internet companies have the results of web crawls.

Big data analytics is needed where there exists:
- A large volume of data (terabytes, petabytes, exabytes), especially when it is a mixture of structured & unstructured data
- Rapidly changing/accumulating data

Big data analytics also assists when:
- A large number (thousands) of models are being built
- Models are frequently built/refreshed

Challenges

The same organizations that collect large volumes of data often have difficulty actually making use of it, for a variety of reasons:
- The architecture of traditional analytic products are not suited to distributed computation, and
- Existing statistical algorithms are not designed to work with big data (these algorithms expect the data to come to them, but big data is too costly to move), thus
- Performing state of the art analytics on big data requires new skills and intimate knowledge of big data systems. Very few analysts have these skills.
- In-memory solutions work for medium-size problems, but do not scale well to truly big data.

Solution

Analytic Server provides:
- A data-centric architecture that leverages big data systems, such as Hadoop Map/Reduce with data in HDFS.
- A defined interface to incorporate new statistical algorithms designed to go to the data.
- Familiar IBM SPSS user interfaces that hide the details of big data environments so that analysts can focus on analyzing the data.
- A solution that is scalable to any size problem.

Architecture

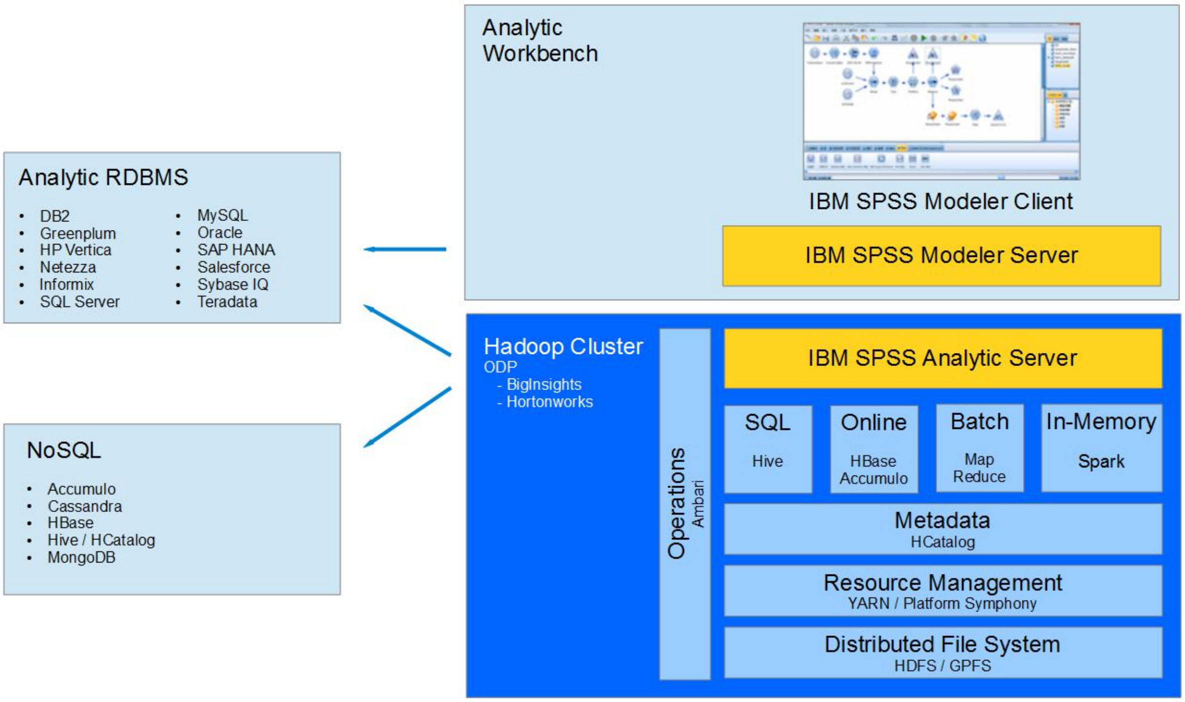

Figure 1. Architecture

Analytic Server sits between a client application and Hadoop cloud. Assuming that the data resides in the cloud, the general outline for working with Analytic Server is to:

1. Define Analytic Server data sources over the data in the cloud.
2. Define the analysis you want to perform in the client application. For the current release, the client application is IBM SPSS Modeler.
3. When you run the analysis, the client application submits an Analytic Server execution request.
4. Analytic Server orchestrates the job to run in the Hadoop cloud and reports the results to the client application.
5. You can use the results to define further analyses, and the cycle repeats.

What is new for administrators in version 2.1

Analytic Server console

Reader role

Within a tenant, you can assign users and groups to a Reader role that cannot log in to the Analytic Server, but can read Analytic Server data sources through the Analytic Server Source node in Modeler.

Chapter 2. Tenant management

Tenants provide a high-level division of users, projects, and data sources so that objects cannot be shared between tenants. Each user accesses the system in the context of a tenant to which they are assigned.

You manage tenants, and assign users to tenants, in the Analytic Server console. The view of the Tenants page depends upon the role of the user that is logged in to the console:

- The "super user" administrator that is set up during installation is the tenant manager. Only this user can create new tenants and edit the properties of any tenant.
- Users with the Administrator role can edit the properties of the tenant they are logged in to.
- Users with the User role cannot edit tenant properties. The Tenants page is hidden from them.
- Users with the Reader role cannot edit data sources, or even log in to the Analytic Server console.

Administrators can access the Projects and Data sources pages and manage any project or data source for cleanup and administration. See the *IBM SPSS Analytic Server User's Guide* for more information.

Tenant listing

The main Tenants page displays the existing tenants in a table. Only the "super user" administrator can make edits on this page.

- Click a tenant's name to display its details and edit its properties.
- Click a tenant's URL to open the console in the context of that tenant.

  Note: You will be logged out of the console and will need to log in with valid credentials for the tenant.
- Type in the search area to filter the listing to display only tenants with the search string in their name.
- Click **New** to create a new tenant with the name you specify in the **Add new tenant** dialog. See “Naming rules” on page 4 for restrictions on the names you can give to tenants.
- Click **Delete** to remove the selected tenant(s).
- Click **Refresh** to update the listing.

Individual tenant details

The content area is divided into several collapsible sections.

Details

Name An editable text field that displays the name of the tenant.

Description
An editable text field that allows you to provide explanatory text about the tenant.

URL This is the URL to give to users to log in to the tenant through the Analytic Server console, and to use to configure SPSS Modeler server. See *IBM SPSS Analytic Server Installation and Configuration Guide* for details on configuring SPSS Modeler.

Status **Active** tenants are currently in use. Making a tenant **Inactive** prevents users from logging in to that tenant, but does not delete any of the underlying information.

Principals

Principals are users and groups that are drawn from the security provider that is set up during installation. You can add principals to a tenant as Administrators or Users.

- Typing in the text box filters on users and groups with the search string in their name. Select **Administrator**, **User**, or **Reader** from the drop-down list to assign their role within the tenant. Click **Add participant** to add them to the list of authors.
- To remove a participant, select a user or group in the member list and click **Remove participant**.

Metrics

Allows you to configure resource limits for a tenant. Reports the disk space currently used by the tenant.

- You can set a maximum disk space quota for the tenant; when this limit it reached, no more data can be written to disk on this tenant until enough disk space is cleared to bring the tenant disk space usage below the quota.
- You can set a disk space warning level for the tenant; when the quota is exceeded, no analytic jobs can be submitted by principals on this tenant until enough disk space is cleared to bring the tenant disk space usage below the quota.
- You can set a maximum number of parallel jobs that can be run at a single time on this tenant; when the quota is exceeded, no analytic jobs can be submitted by principals on this tenant until a currently running job completes.
- You can set the maximum number of fields a data source can have. The limit is checked whenever a data source is created or updated.
- You can set the maximum file size in megabytes. The limit is checked when a file is uploaded.

Security provider configuration

Allows you to specify the user authentication provider. **Default** uses the default tenant's provider, which was set up during installation and configuration. **LDAP** allows you to authenticate users with an external LDAP server such as Active Directory or OpenLDAP. Specify the settings for the provider and optionally specify filter settings to control the users and groups available in the Principals section.

Naming rules

For anything that can be given a unique name in Analytic Server, such as data sources and projects, the following rules are applied to those names.

- Within a single tenant, names must be unique within objects of the same type. For example, two data sources cannot both be named insuranceClaims, but a data source and a project could each be named insuranceClaims.
- Names are case-sensitive. For example, insuranceClaims and InsuranceClaims are considered unique names.
- Names ignore leading and trailing white space.
- The following characters are invalid in names.

```
~, #, %, &, *, {, }, \\, :, <, >, ?, /, |, ", \t, \r, \n
```

Chapter 3. Getting users started

Tell users to navigate to `http://<host>:<port>/<context-root>/admin/<tenant>` and enter their username and password to log on to the Analytic Server console.

<host>
: The address of the Analytic Server host.

<port>
: The port that Analytic Server is listening on. By default this is 8080.

<context-root>
: The context root of the Analytic Server. By default this is analyticserver.

<tenant>
: In a multi-tenant environment, the tenant you belong to. In a single-tenant environment, the default tenant is **`ibm`**.

For example, if the host machine has IP address 9.86.44.232, you have created a "mycompany" tenant and added users to it, and the other settings have been left to their defaults, then users should navigate to `http://9.86.44.232:8080/analyticserver/admin/mycompany` to access the Analytic Server console.

Chapter 4. Analytic Server job names

Analytic Server produces map-reduce jobs, which can be monitored through your Hadoop cluster's jobtracker.

The map-reduce job name has the following structure.

```
AS/{tenant name}/{user name}/{algorithm name}
```

{tenant name}
This is the name of the tenant under which the job is run.

{user name}
This is the user who requested the job.

{algorithm name}
This is the primary algorithm in the job. Note that a single stream may generate multiple map-reduce jobs; likewise, several operations within a stream can be contained within a single map-reduce job.

Chapter 5. Troubleshooting

Analytic Server provides several helpful tools for problem determination.

Logging

Analytic Server creates customer log files and trace files that are helpful for diagnosing problems. With the default Liberty installation, you can find the log files in the `{AS_ROOT}/ae_wlpserver/usr/servers/aeserver/logs` directory.

The default logging configuration produces two log files that roll over on a daily basis.

as.log This file contains the high-level summary of informational warning and error messages. Check this file first when server errors occur that cannot be resolved by using the error message that is displayed in the User interface.

as_trace.log
This file contains all the entries from `ae.log`, but adds more information that is primarily targeted to IBM support and development for debugging purposes.

Analytic Server uses Apache LOG4J as its underlying logging facility. Using LOG4J, the logging can be dynamically adjusted by editing the `{AS_SERVER_ROOT}/configuration/log4j.xml` configuration file. You may be asked to do this by Support to help diagnose problems, or you may want to modify this to limit the number of log files kept around. Changes to the file are detected automatically within a few seconds so the Analytic Server does not need to be restarted.

For more information about log4j and the configuration file, see documentation at the official Apache website at http://logging.apache.org/log4j/.

Version information

You can determine what version of Analytic Server is installed by checking the `{AS_ROOT}/properties/version` folder. The following files contain version information.

IBM_SPSS_Analytic_Server-*.swtag
Contains detailed product information.

version.txt
Version and build number for the installed product.

Log collector

When problems cannot be resolved by directly reviewing the log files, you can bundle all the logs and send them to IBM support. There is a utility that is provided to make collecting all the necessary data simpler.

Using a command shell, run the following commands:

```
cd  {AS_ROOT}/bin
run  >sh ./logcollector.sh
```

These commands create a compressed file under `{AS_ROOT}/bin`. The compressed file contains all the log files and product version information.

Common issues

This section describes some common administration issues and how you can fix them.

Security

Kerberos authentication fails when trying to access HCatalog data source

If you see errors in the log like the following:

```
cause:javax.security.sasl.SaslException: Failure to initialize security context

com.spss.analyticframework.api.exceptions.ComponentException: Cannot access HCatalog
```

You must ensure that the HDFS user's Kerberos TGT is cached and available on the Analytic Server server host. To do this:

1. Stop the Analytic Server process.
2. Run `kinit -f $hdfs.user` from the Analytic Server host, where **`$hdfs.user`** is as defined in the `config.properties` file and has write permission to analytic root
3. Start Analytic Server.

Analytic Server Console

Accessing the Analytic Server Console from Safari on iOS

The tenant status dropdown list does not work, and you cannot update the data model for File-based data sources. Use another browser when performing these actions.

Running streams

PySpark jobs fail to run

1. In the Ambari Console, add `export SPARK_HOME=/usr/iop/current/spark-client` to the **`yarn-env`** parameter of the YARN service.
2. Ensure the Spark service deployed in all nodes in the cluster.
3. Restart the YARN service.

Performance tuning

This section describes ways to optimize the performance of your system.

Analytic Server is a component in the Ambari framework that utilizes other components such as HDFS, Yarn and Spark. Common performance tuning techniques for Hadoop, HDFS and Spark apply to Analytic Server workloads. Every Analytic Server workload is different therefore tuning experimentation is required based on your specific deployments workload. The following properties and tuning tips are key changes that have impacted the results of the Analytic Server benchmarking and scaling tests.

When the first job runs on Analytic Server, the server will start a persistent Spark application that will be active until the Analytic Server is shut down. The persistent Spark application will allocate and hold onto all the cluster resources allocated to it for the duration of the Analytic Server running, even if an Analytic Server job is not actively running. Careful thought should be given to the amount of resources allocated to the Analytic Server Spark application. If all cluster resources are allocated to the Analytic Server Spark application, then other jobs could be delayed or not run. These jobs could be queued waiting for sufficient free resources and those resources will be consumed by Analytic Server Spark application.

If multiple Analytic Server services are configured and deployed, each service instance could potentially allocate its own persistent Spark application. For example, if two Analytic Server services are deployed to support high availability failover, then you could see two persistent Spark applications active, each allocating cluster resources.

An additional complexity is that in certain situations, Analytic Server may start a map reduce job that will require cluster resources. These map reduce jobs will require resources that are not allocated to the Spark application. The specific components that require map reduce jobs are PSM model builds.

The following properties can be configured to allocate resources to Spark application. If they are set in the spark-defaults.conf of the Spark installation, then they are allocated for all Spark jobs run in the environment. If they are set in the Analytic Server configuration as custom poperties under the "Custom analytic.cfg" section, then they are allocated for the Analytic Server Spark application only.

spark.executor.memory

Amount of memory to use per executor process.

spark.executor.instances

The number of executor processes to start.

spark.executor.cores

The number of executor worker threads per executor process. This value should be between 1 and 5.

An example of setting the three key Spark properties. There are 10 data nodes in a HDFS cluster and each data node has 24 logical cores and 48 GB of memory and is only running HDFS processes. Here is one way to configure the properties for this environment, assuming you are only running Analytic Server jobs on this environment and desire maximum allocation to a single Analytic Server Spark application.

- Set spark.executor.instances=20. This would attempt to run 2 Spark executor processes per data node.
- Set spark.executor.memory=22G. This would set the max heap size for each Spark executor process to 22 GB, allocating 44 GB on each data node. Other JVMs and the OS need the extra memory.
- Set spark.executor.cores=5. This will provide 5 worker threads for each Spark executor, for a total of 10 worker threads per data node.

Monitor the Spark UI for running jobs

If you see Spill to disk that could impact performance. Some possible solutions are:

- Increase memory and allocate it to Spark executors via **`spark.executor.memory`**.
- Reduce the number of **`spark.executor.cores`**. This will reduce the number of concurrent work threads allocating memory, but it will also reduce the amount of parallelism for the jobs.
- Change the Spark memory properties. **`spark.shuffle.memoryFraction`** and **`spark.storage.memoryFraction`** allocation percentage of the Spark executor heap for Spark.

Ensure the name node has enough memory

If the number of blocks in HDFS is large and growing, ensure you name node heap increases to accommodate this growth. This is a common HDFS tuning recommendation.

Alter the amount of memory used for caching

By default, **`spark.storage.memoryFraction`** has value 0.6. This can be increased up to 0.8 in case the HDFS block size of the data is 64MB. If the HDFS block size of the input data is greater than 64MB then this value could be increased only if the memory allocated per task is greater than 2GB.

Tuning Performance of Model Scoring

You can improve the performance of model scoring jobs on big datasets with the Apache Spark engine by using the following steps. Note that these steps should not impact the operation of non-Analytic Server services on the cluster.

1. Check if libtcmalloc_minimal.so{/version} is already installed on each node in the cluster.
   ```
   whereis libtcmalloc_minimal.so.*
   ```
2. If libtcmalloc_minimal.so is not installed, either install the operating system specific package containing the libtcmalloc_minimal library on each node in your cluster, or manually build and install libtcmalloc_minimal. For example:

   Ubuntu:
   ```
   sudo apt-get install libgoogle-perftools-dev
   ```
   Red Hat Enterprise Linux 6.x (x64):
   a. Install the EPEL repository for RedHat (if not already installed)
      ```
      wget http://dl.fedoraproject.org/pub/epel/6/x86_64/epel-release-6-8.noarch.rpm
      sudo rpm -Uvh epel-release-6*.rpm
      ```
   b. `sudo yum install gperftools-libs.x86_64`

   Manual Build:
   a. Download the gperftools-2.4.tar.gz from the link https://github.com/gperftools/gperftools/releases
   b. `tar zxvf gperftools-2.4.tar.gz`
   c. `cd gperftools-2.4`
   d. `./configure --disable-cpu-profiler --disable-heap-profiler --disable-heap-checker --disable-debugalloc --enable-minimal`
   e. `make`
   f. `sudo make install`
3. Note one of the location(s) of the installed library file libtcmalloc_minimal.so{.version}, as returned from running the following command on one or more of the nodes.
   ```
   whereis libtcmalloc_minimal.so.*
   ```
   If the cluster has nodes running a mixture of operating systems, there could be multiple locations for this file.
4. In the Ambari console, go to the Analytic Server configuration and under section Custom analytics.cfg, configure the key spark.executorEnv.LD_PRELOAD using the location of the library as the value. After making this change restart the Analytic Server service. For example, if the library is installed to /usr/lib64/libtcmalloc_minimal.so.4, the configuration would be:
   ```
   spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4
   ```
   If multiple locations are required, use a space to separate them, as in the following example.
   ```
   spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4 /usr/lib/libtcmalloc_minimal.so
   ```
   If any nodes do not have the libtcmalloc_minimal.so library installed at one of the configured locations, this will not cause an error, but performance of model scoring may be slower on these nodes.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who want to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of The Minister for the Cabinet Office, and is registered in the U.S. Patent and Trademark Office.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Printed in USA