

**IBM SPSS Analytic Server
V3.0**

管理员指南

IBM

注释

在使用本信息及其支持的产品之前，请先阅读第 11 页的『声明』中的信息。

产品信息

此版本是用于 IBM SPSS Analytic Server 的 V3.0.0.0 以及后续发行版和修订版，直至在新版本中另有说明为止。

目录

第 1 章 租户管理	1	版本信息	7
命名规则	2	日志收集器	7
第 2 章 用户入门	3	常见问题	8
		性能调整	8
第 3 章 分析服务器 作业名	5	声明	11
		商标	12
第 4 章 故障诊断	7		
日志记录	7		

第 1 章 租户管理

租户提供对用户、项目和数据源的高级划分，因此无法在租户间共享对象。每个用户在其分配到的租户上下文中访问系统。

在 分析服务器 控制台中管理租户并将用户分配到租户。能否查看“租户”页面取决于登录该控制台的用户的角色：

- 在安装期间设置的“超级用户”管理员为租户管理员。只有该用户才能创建新租户并编辑任何租户的属性。
- 具有“管理员”角色的用户可以编辑其登录的租户的属性。
- 具有“用户”角色的用户不能编辑租户属性。对这些用户隐藏“租户”页面。
- 具有“读者”角色的用户无法编辑数据源，甚至无法登录 分析服务器 控制台。

管理员可以访问“项目”和“数据源”页面，并管理要清除和管理的任何项目或数据源。请参阅 *IBM® SPSS® Analytic Server User's Guide* 以获取更多信息。

租户列表

主“租户”页面显示表中的现有租户。只有“超级用户”管理员才能对此页面进行编辑。

- 单击租户名称以显示其详细信息和编辑其属性。
- 单击租户 URL 以在该租户的上下文中打开控制台。

注：您将从控制台注销并且需要使用租户的有效凭证来登录。

- 在搜索区域输入以对列表进行过滤，从而仅显示其名称中包含搜索字符串的租户。
- 单击**新建**以使用在**添加新租户**对话框中指定的名称创建新租户。请参阅第 2 页的『命名规则』，以了解有关可以给租户提供的名称的限制。
- 单击**删除**以除去所选租户。
- 单击**刷新**以更新列表。

个别租户详细信息

内容区域划分为多个可折叠部分。

详细信息

名称	这是一个可编辑的文本字段，用于显示租户的名称。
描述	可编辑的文本字段，您可以提供关于该租户的说明性文本。
URL	这是将提供给用户通过 分析服务器 控制台登录租户，以及用于配置 SPSS Modeler 服务器的 URL。请参阅 <i>IBM SPSS Analytic Server Installation and Configuration Guide</i> 以获取有关配置 SPSS Modeler 的详细信息。
状态	处于 活动状态 的租户目前正在使用中。使租户处于 不活动状态 将使用户无法登录该租户，但是不会删除任何底层的信息。

主体

主体是从安装过程中设置的安全提供程序获取的用户和组。您可以将主体添加到租户，作为管理员或用户。

- 在文本框中输入时，会对其名称中包含搜索字符串的用户和组进行过滤。 从下拉列表中选择**管理员**、**用户**或**读者**可以分配用户在该租户中的角色。 单击**添加参与者**以将其添加到作者列表。
- 要除去参与者，请在成员列表中选择用户或组，然后单击**除去参与者**。

度量 使您能够配置租户的资源限制。 报告租户目前使用的磁盘空间。

- 您可以为租户设置最大磁盘空间配额；当达到该限制时，不能再向该租户上的磁盘写入任何数据，直到清理出足够的磁盘空间，以使租户磁盘空间使用量低于配额。
- 您可以为租户设置磁盘空间警告级别；当超过配额时，该租户上的主体不能提交分析作业，直到清理出足够的磁盘空间，以使租户磁盘空间使用量低于配额。
- 您可以设置在该租户上一次能够运行的最大并行作业数；当超过配额时，该租户上的主体不能提交分析作业，直到目前正在运行的作业完成。
- 您可以设置数据源能够具有的最大字段数。 每当创建或更新数据源时，都会检查该限制。
- 您可以设置最大文件大小（以兆字节为单位）。 上传文件时，会检查该限制。

安装提供程序配置

使您能够指定用户认证提供程序。 **缺省**使用在安装和配置期间设置的缺省租户提供程序。 **LDAP** 允许您使用诸如 Active Directory 或 OpenLDAP 之类的外部 LDAP 服务器来认证用户。 为提供程序指定设置并且可选择指定过滤器设置，以控制“主体”部分中提供的用户和组。

命名规则

对于在 分析服务器 中可给予唯一名称的一切对象（例如，数据源和项目），以下规则适用于这些名称。

- 在单个租户中，名称必须在同类型的对象中唯一。 例如，两个数据源不能同时命名为 insuranceClaims，但是一个数据源和一个项目可分别命名为 insuranceClaims。
- 名称区分大小写。 例如，insuranceClaims 和 InsuranceClaims 均被视为唯一名称。
- 名称忽略前置和后置空格。
- 以下字符在名称中无效。

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

第 2 章 用户入门

告知用户浏览至 `http://<host>:<port>/<context-root>/admin/<tenant>`，并输入其用户名和密码以登录到 分析服务器 控制台。

<host>

分析服务器 主机的地址。

<port>

分析服务器 正在侦听的端口号。 缺省情况下，此端口为 9080。

<context-root>

分析服务器 的上下文根。 缺省情况下为 `analyticsserver`。

<tenant>

在多租户环境中，这是您所属的租户。 在单租户环境中，缺省租户为 `ibm`。

例如，如果主机的 IP 地址为 9.86.44.232，您已创建“mycompany”租户并向其添加了用户，并且其他设置已保留为缺省值，那么用户应该导航至 `http://9.86.44.232:9080/analyticsserver/admin/mycompany` 以访问 分析服务器 控制台。

第 3 章 分析服务器 作业名

分析服务器可生成 map-reduce 作业，可通过 Hadoop 的集群的作业跟踪器来监控这些作业。

map-reduce 作业名采用以下结构。

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

{tenant name}

这是在其中运行作业的租户的名称。

{user name}

这是请求该作业的用户。

{algorithm name}

这是作业中的主要算法。 请注意，单个流可能生成多个 map-reduce 作业；类似地，单个流中的多个操作可包含在单个 map-reduce 作业中。

第 4 章 故障诊断

分析服务器 提供了多种实用工具用于问题确定。

日志记录

分析服务器会创建有助于诊断问题的客户日志文件和跟踪文件。通过缺省 Liberty 安装，可以在 {AS_ROOT}/ae_wlpserver/usr/servers/aeserver/logs 目录中找到日志文件。

缺省日志记录配置将生成两个每天轮换的日志文件。

as.log 此文件包含参考警告和错误消息的高级别摘要。发生服务器错误时如果无法通过使用用户界面中显示的错误消息来解决，请首先检查此文件。

as_trace.log

此文件包含来自 ae.log 的所有条目，但添加了额外信息，这些信息主要面向 IBM 支持和开发人员，用于调试目的。

分析服务器 使用 Apache LOG4J 作为其底层的日志记录工具。通过使用 LOG4J，可通过编辑 {AS_SERVER_ROOT}/configuration/log4j.xml 配置文件来动态调整日志记录。支持人员可能要求您执行此项操作以帮助诊断问题，或者您可能希望修改此文件以限制保留的日志文件数量。对文件进行的更改会在数秒钟内自动完成检测，因此无需重新启动 分析服务器。

有关 log4j 和配置文件的更多信息，请参阅位于以下地址的 Apache 官方 Web 站点上的文档：<http://logging.apache.org/log4j/>。

版本信息

您可以通过检查 {AS_ROOT}/properties/version 文件夹确定所安装的分析服务器 版本。以下文件包含版本信息。

IBM_SPSS_Analytic_Server-*.swtag

包含详细的产品信息。

version.txt

已安装的产品版本和构建号。

日志收集器

无法通过直接复查日志文件来解决问题时，可以将所有日志捆绑在一起并发送给 IBM 支持人员。其中提供了一个实用程序来简化所有必要数据的收集。

通过使用命令 shell，运行以下命令：

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

这些命令会在 {AS_ROOT}/bin 下创建一个压缩文件。该压缩文件包含所有日志文件和产品版本信息。

常见问题

本节描述了一些常见管理问题以及它们的解决方法。

安全性

尝试访问 HCatalog 数据源时 Kerberos 认证失败

如果您在日志中看到类似于以下内容的错误：

```
cause:javax.security.sasl.SaslException: Failure to initialize security context
com.spss.analyticframework.api.exceptions.ComponentException: Cannot access HCatalog
```

那么必须确保 HDFS 用户的 Kerberos TGT 已高速缓存并且在 分析服务器 服务器主机上可用。 为此：

1. 停止 分析服务器 进程。
2. 从 分析服务器 主机运行 `kinit -f $hdfs.user`，其中 `$hdfs.user` 是在 `config.properties` 文件中定义的，并且对分析 `root` 用户具有写许可权
3. 启动 分析服务器。

分析服务器 控制台

在 iOS 上通过 Safari 访问 分析服务器 控制台

租户状态下拉列表无法正常工作，并且您无法更新基于文件的数据源的数据模型。 执行操作时，请使用另一个浏览器。

运行中的流

R 作业将非英语单词转换为 Unicode

在 Cloudera 集群上，如果 Hadoop 服务器的系统编码不是 UTF-8，那么 R 将非英语单词转换为 Unicode。

1. 在 Cloudera Manager 控制台中浏览至 YARN 配置选项卡。
2. 在“NodeManager 环境高级配置片段（安全阀）”字段中添加以下设置。

```
LC_ALL=""
LANG=en_US.utf8
```

PySpark 作业运行失败

1. 在 Ambari 控制台中，将 `export SPARK_HOME=/usr/iop/current/spark-client`（适用于 BigInsights）或 `SPARK_HOME=/usr/hdp/current/spark-client`（适用于 Hortonworks）添加到 YARN 服务的 `yarn-env` 参数；在 Cloudera Management 中，将 `SPARK_HOME=/opt/cloudera/parcels/CDH/lib/spark` 添加到 Yarn 服务 YARN (MR2 Included) Service Environment Advanced Configuration Snippet (Safety Valve)。
2. 确保在所有集群节点中部署 Spark 服务。
3. 重新启动 YARN 和 Analytic Server 服务。

性能调整

本部分描述了优化系统性能的方法。

分析服务器 是 Ambari 框架中使用其他组件（如 HDFS、Yarn 和 Spark）的组件。 Hadoop、HDFS 和 Spark 的常见性能调整方法适用于 分析服务器 工作负载。 每个 分析服务器 工作负载都不相同，因此需要根据特定部署工作负载执行调整试验。 以下属性和调整提示是影响 分析服务器 基准测试和缩放测试结果的主要更改。

当第一个作业在分析服务器上运行时，该服务器启动在关闭分析服务器前将处于活动状态的持久 Spark 应用程序。即使分析服务器作业未主动运行，持久 Spark 应用程序都将分配并保留到在分析服务器运行期间为其分配的所有集群资源。应对分配到分析服务器 Spark 应用程序的资源量进行仔细思考。如果将所有集群资源分配给分析服务器 Spark 应用程序，那么可以延迟或不运行其他作业。这些作业可排队等待足够的可用资源，并且这些资源将被分析服务器 Spark 应用程序使用。

如果已配置和部署多个 Analytic Server 服务，那么每个服务实例都可以分配它自己的持久 Spark 应用程序。例如，如果部署两个 Analytic Server 服务以支持高可用性故障转移，那么您可以看到两个持久 Spark 应用程序处于活动状态，各自分配集群资源。

在某些情况下，Analytic Sever 可能会启动将需要集群资源的 map reduce 作业，这样将增加复杂性。这些 map reduce 作业将需要未分配给 Spark 应用程序的资源。需要 map reduce 作业的特定组件为 PSM 模型构建。

可以将以下属性配置为将资源分配给 Spark 应用程序。如果在 Spark 安装的 spark-defaults.conf 中进行设置，那么为在环境中运行的所有 Spark 作业对其进行分配。如果在 Analytic Server 配置中将其设置为“定制 analytic.cfg”部分下的定制属性，那么仅为 Analytic Server Spark 应用程序对其进行分配。

spark.executor.memory

每个执行者进程要使用的内存量。

spark.executor.instances

要启动的执行者进程的数量。

spark.executor.cores

每个执行者进程的执行者工作者线程数量。此值应该介于 1 和 5 之间。

设置这三个主要 Spark 属性的示例。HDFS 集群中有 10 个数据节点，每个数据节点有 24 个逻辑核心和 48 GB 内存并且只运行 HDFS 进程。此处是为环境配置属性的一种方法，假设您在此环境中只运行 Analytic Server 作业并期望向单个 Analytic Server Spark 应用程序的最大分配。

- 设置 spark.executor.instances=20。这样将尝试在每个数据节点运行 2 个 Spark 执行者进程。
- 设置 spark.executor.memory=22G。这样会将每个 Spark 执行者进程的最大堆大小设置为 22 GB，每个数据节点上分配 44 GB。其他 JVM 和 OS 需要额外的内存。
- 设置 spark.executor.cores=5。这样将为每个 Spark 执行者提供 5 个工作者线程，每个数据节点总共有 10 个工作者线程。

监视用于运行作业的 Spark UI

如果您看到溢出至磁盘，这可能会影响性能。可能的解决方案是：

- 通过 spark.executor.memory 增加内存并将其分配给 Spark 执行者。
- 减少 spark.executor.cores 的数量。这样将减少分配内存的并发工作线程的数量，但是还将减少并行作业的数量。
- 为 Spark 更改 Spark 执行者堆的 Spark 内存属性 spark.shuffle.memoryFraction 和 spark.storage.memoryFraction 分配百分比。

确保名称节点有足够内存

如果 HDFS 中的块数较大并继续增长，请确保名称节点堆增大以适应此增长。这是常见的 HDFS 调整建议。

更改用于高速缓存的内存量

缺省情况下，`spark.storage.memoryFraction` 的值为 0.6。如果数据的 HDFS 块大小为 64 MB，那么该值可以增大到 0.8。如果输入数据的 HDFS 块大小大于 64 MB，那么仅当为每个任务分配的内存大于 2 GB 时，才可以增大此值。

模型评分的调整性能

通过执行以下步骤，您可以使用 Apache Spark 引擎提高大数据集上的模型评分作业的性能。请注意，这些步骤不能影响集群上非 Analytic Server 服务的操作。

1. 检查是否已在集群中的每个节点上安装 `libtcmalloc_minimal.so{version}`。

```
whereis libtcmalloc_minimal.so.*
```

2. 如果未安装 `libtcmalloc_minimal.so`，那么安装特定于操作系统的软件包（其中包含集群中每个节点上的 `libtcmalloc_minimal` 库），或手动构建和安装 `libtcmalloc_minimal`。例如：

Ubuntu:

```
sudo apt-get install libgoogle-perftools-dev
```

Red Hat Enterprise Linux 6.x (x64):

- a. 安装适用于 RedHat 的 EPEL 存储库（如果尚未安装）

```
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`

手动构建:

- a. 从链接 <https://github.com/gperftools/gperftools/releases> 下载 `gperftools-2.4.tar.gz`

- 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. 请注意已安装的库文件 `libtcmalloc_minimal.so{version}` 的某个位置，如在一个或多个节点上运行以下命令所返回的。

```
whereis libtcmalloc_minimal.so.*
```

如果集群具有同时运行多个操作系统的节点，那么此文件可能存在多个位置。

4. 在 Ambari 控制台中，转至 Analytic Server 配置并在“定制 `analytics.cfg`”部分下，使用库位置作为值来配置主要的 `spark.executorEnv.LD_PRELOAD`。在做出此更改后，重新启动 Analytic Server 服务。例如，如果将库安装到 `/usr/lib64/libtcmalloc_minimal.so.4`，那么配置将为：

```
spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4
```

如果需要多个位置，请使用空格进行分隔，如以下示例所示。

```
spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4 /usr/lib/libtcmalloc_minimal.so
```

如果任何节点都未将 `libtcmalloc_minimal.so` 库安装在某个已配置的位置，虽然这样不会导致错误，但是这些节点上模型评分的性能可能会降低。

声明

本信息是为在美国国内供应的产品和服务而编写的。

IBM 可能在其他国家或地区不提供本文中讨论的产品、服务或功能特性。有关您所在区域当前可获得的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。但是，评估和验证任何非 IBM 产品、程序或服务的操作，由用户自行负责。

IBM 可能已拥有或正在申请与本文档内容有关的各项专利。提供本文档并不意味着授予用户使用这些专利的任何许可。您可以用书面形式将许可查询寄往：

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

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往：

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

以下段落对于英国和与当地法律有不同规定的其他国家或地区均不适用：INTERNATIONAL BUSINESS MACHINES CORPORATION“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某特定用途的保证。某些国家或地区在某些交易中不允许免除明示或暗含的保证。因此本条款可能不适用于您。

本信息可能包含技术方面不够准确的地方或印刷错误。本信息将定期更改；这些更改将编入本信息的新版本中。IBM 可以随时对本出版物中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的，不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的：(i) 使其能够在独立创建的程序和其它程序（包括本程序）之间进行信息交换，以及 (ii) 使其能够对已经交换的信息进行相互使用，请与下列地址联系：

IBM Software Group
ATTN: Licensing

200 W. Madison St.
Chicago, IL; 60606
U.S.A.

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费，都可获得这方面的信息。

本文档中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际程序许可协议或任何同等协议中的条款提供。

此处包含的任何性能数据都是在受控环境中测得的。因此，在其他操作环境中获得的数据可能会有明显的不同。有些测量可能是在开发级的系统上进行的，因此不保证与一般可用系统上进行的测量结果相同。此外，有些测量是通过推算而估计的，实际结果可能会有差异。本文档的用户应当验证其特定环境的适用数据。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

所有关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

所有 IBM 的价格均是 IBM 当前的建议零售价，可随时更改而不另行通知。经销商的价格可与此不同。

本信息仅用于规划的目的。在所描述的产品上市之前，此处的信息会有更改。

本信息包含在日常业务操作中使用的数据和报告的示例。为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。所有这些名字都是虚构的，若现实生活中实际业务企业使用的名字和地址与此相似，纯属巧合。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明：

本信息包含在日常业务操作中使用的数据和报告的示例。为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。所有这些名字都是虚构的，若现实生活中实际业务企业使用的名字和地址与此相似，纯属巧合。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明：

©（贵公司的名称）（年）。此部分代码是根据 IBM Corp. 公司的样本程序衍生出来的。

© Copyright IBM Corp.（输入年份）。All rights reserved.

如果您正在查看本信息的软拷贝，图片和彩色图例可能无法显示。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp.，在全球许多管辖区域的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。在 Web 站点 www.ibm.com/legal/copytrade.shtml 上的“Copyright and trademark information”部分中提供了 IBM 商标的最新列表。

Adobe、Adobe 徽标、PostScript 以及 PostScript 徽标是 Adobe Systems Incorporated 在美国和 / 或其他国家或地区的注册商标或商标。

IT Infrastructure Library 是 Central Computer and Telecommunications Agency 的注册商标，该企业现已成为 Office of Government Commerce 的一部分。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国和@3B72其他国家或地区的注册商标。

Microsoft、Windows、Windows NT 以及 Windows 徽标是 Microsoft Corporation 在美国和/或其他国家或地区的商标。

ITIL 是一个注册商标，是 Minister for the Cabinet Office 的共同体注册商标，并且已在 U.S. Patent and Trademark Office 进行注册。

UNIX 是 The Open Group 在美国和 / 或其他国家或地区的注册商标。

Cell Broadband Engine 是 of Sony Computer Entertainment, Inc. 在美国和/或其他国家或地区的商标并且在当地许可证下使用。

Linear Tape-Open、LTO、LTO 徽标、Ultrium 和 Ultrium 徽标是 HP、IBM Corp 和 Quantum 在美国和其他国家或地区的商标。



Printed in China