Splunk® Enterprise 8.2.0

Splunk Analytics for Hadoop

生成时间：2021 年 5 月 24 日，14:30

Table of Contents

符合 Splunk Analytics for Hadoop

符合 Splunk Analytics for Hadoop

Hadoop 允许您存储大量结构化、多结构化和非结构化数据，但是从该数据中提取值可能是一项有挑战性且耗时的工作。

Splunk Analytics for Hadoop 允许您通过虚拟索引访问远程 Hadoop 群集中的数据，并允许您使用 Splunk 搜索处理语言分析使用 Hadoop 和 NoSQL 数据存储的数据。

- 处理、报告和可视化大量结构化、多结构化和非结构化数据。
- 针对 Hadoop 数据和 Splunk Enterprise 索引中的数据运行合并的报表。
- 使用 SDK 和使用 Hadoop 数据的应用。

出于数据存储在 Hadoop 中的方式的性质，有一些特定的 Splunk Enterprise 索引行为无法复制：

- 虽然 Splunk Analytics for Hadoop 可以使用预览功能和报表加速，但目前不支持实时搜索 Hadoop 数据。
- 由于事件未按照任何特定的顺序存储，因此任何依赖于隐式时间顺序的搜索命令将表现出不同的 Splunk Analytics for Hadoop 行为。（例如：head、tail、delta 等）有关特定的时间戳敏感命令如何使用虚拟索引的更多信息，请参阅本手册中的“搜索虚拟索引”。
- 数据返回速度不会始终如返回本地索引数据那样快。

要设置 Splunk Analytics for Hadoop 以使用您自己的 HDFS 数据，请参阅“安装 Splunk Analytics for Hadoop”。

要了解在 Splunk Web 中配置和搜索数据的更多信息，请参阅“关于虚拟索引数据的搜索和报表”。

要了解有关 Splunk Analytics for Hadoop 如何工作的更多信息，请参阅“Splunk Analytics for Hadoop 概念”。

关于搜索，我们还建议使用 Splunk Enterprise *搜索手册*和*搜索教程*。

Splunk Analytics for Hadoop 如何返回 Hadoop 数据的报表

启动搜索后，Splunk Analytics for Hadoop 会使用 Hadoop MapReduce 框架正确处理数据。通常在索引时间完成的所有数据分析，包括来源类型、事件换行和时间戳，在搜索时间在 Hadoop 中执行。Splunk Analytics for Hadoop 不会索引这个数据，而是会在每次请求时处理该数据。这是 Splunk Enterprise for Hadoop 如何搜索 Hadoop 虚拟索引的概述：

1. 用户在虚拟索引上启动报表生成的搜索。请参阅“搜索虚拟索引”获取有关生成报表生成的搜索的更多信息。

2. Splunk Analytics for Hadoop 认识到请求是针对虚拟索引，并衍生出外部结果提供程序 (ERP) 进程帮助处理请求。ERP 是一个搜索助手进程，可在 Hadoop 数据上执行搜索。请参阅“关于虚拟索引”。

3. 基于您的配置，Splunk Analytics for Hadoop 会将配置和运行时间数据，包括分析搜索字符串等以 JSON 格式传递到 ERP。

4. 如果这是首次针对特定的提供程序系列执行搜索，ERP 进程会通过将 Splunk Enterprise 软件包和知识包复制到 HDFS 或 NoSQL 数据库中，在 HDFS 中设置必要的环境。

5. ERP 进程会分析搜索请求。它会识别要处理的相关数据，并生成要在 Hadoop 上执行的任务。然后，衍生 MapReduce 任务执行计算。

6. 对于每个任务，MapReduce 任务首先会根据正确的 Splunk 软件包和知识包检查以确保环境为最新的。

7. 如果未发现正确的软件包和知识软件包，任务会从 HDFS 中复制 Splunk 软件包（参见步骤 4），然后将其提取到配置目录中。然后，会复制 HDFS 中的软件包（参见步骤 4），然后将它们扩展到 TaskTracker 中的正确目录中。

8. 映射任务会在 TaskTracker 节点上衍生搜索进程以处理所有数据处理。

9. 映射任务会向搜索进程提供数据，并消耗输出，此输出会变成映射任务的输出。此输出会以 HDFS 形式存储。

10. 搜索头上的 ERP 进程会不断轮询 HDFS 以挑选出结果并将结果提供给在搜索头上运行的搜索进程。

11. 搜索头上的 ERP 搜索进程会使用这些结果新建报表。当有了新数据时报表会不断更新。

了解更多和获取帮助

要查找有关 Splunk Enterprise 的更多信息：

- Splunk 支持
- Splunk Enterprise 文档
- Splunk Answers
- EFNET 上的 #splunk IRC 通道

安装 Splunk Analytics for Hadoop

系统和软件要求

确保您可以访问至少一个 Hadoop 群集（其中含数据）而且可以在群集中的数据上运行 MapReduce 任务。

确保您安装了 Java Development Kit (JDK) 1.6 及更高版本。不过，为了获取最佳效果，请升级至 JDK 1.6 以上的版本。以下分发和版本已使用 JDK 1.8 进行认证。

以下几种 Hadoop 分布和版本支持 Splunk Analytics for Hadoop：

- Apache Hadoop 3.2.1
- Open Apache 3.1.2
- Cloudera Distribution including Apache Hadoop v6.3
- Hortonworks Data Platform (HDP) 3.1.4
- MapR 6.1

Hadoop 节点上需要配置什么内容

在 Hadoop TaskTracker 节点上，您需要在 *nix 文件系统上配置一个目录，来运行符合以下要求的 Hadoop 节点：

- 1GB 的免费磁盘空间，用于存放 Splunk 副本。
- 5-10GB 的免费磁盘空间，用作临时存储。该存储空间供各搜索进程使用。

Hadoop 文件系统上需要配置什么内容

在您的 Hadoop 文件系统（HDFS 或其他）上，您需要：

- 一个位于 `jobtracker.staging.root.dir` 下的子目录（通常为 `/user/`），该子目录以用户帐号为名称，而 Splunk Analytics for Hadoop 在该用户帐号下于搜索头上运行。例如，若 Splunk Analytics for Hadoop 由用户 "BigDataUser" 和 `jobtracker.staging.root.dir=/user/` 启动，您需要一个用户 "BigDataUser" 可以访问的目录 `/user/HadoopAnalytics`。
- 上述目录下的子目录，可供此服务器用于中间存储，如 `/user/hadoopanalytics/server01/`

确保 Splunk Enterprise 和 Hadoop 之间兼容

很多 Splunk Analytics for Hadoop 功能需要在不同的第三方数据库、Splunk Analytics for Hadoop 和 Splunk Enterprise 之间通信。为了便于在需要时配置这些功能，我们建议您用相同的用户名和凭据安装或配置所有功能：

- Splunk Enterprise/Splunk Analytics for Hadoop
- HDFS 节点（如适用）
- S3 或其他数据库应用程序

如果您计划搜索或存档 Splunk 数据，您还应在以下应用中安装和/或配置这些相同的用户名称：

- Splunk Enterprise
- Splunk 索引

- 您可使用传递验证允许非管理员用户在 HDFS 中运行 MapReduce 任务。要配置该验证，请参阅“关于传递验证”。

安装 Splunk 以使用 Splunk Analytics for Hadoop

注意：Windows 不支持 Splunk Analytics for Hadoop。

要在系统中授权 Splunk Analytics for Hadoop 许可，您必须下载最新版的 Splunk 的 Linux 分布。您必须配置安装以在驻留在 *nix 平台中的搜索头上运行。您可在任何满足搜索头要求的计算机上运行 Splunk Analytics for Hadoop。请参阅“系统和软件要求”。您可通过此网址查找要下载的正确版本：http://www.splunk.com/download/splunk。

下载之后，您可获得 Enterprise 和 Splunk Analytics for Hadoop 的临时许可证。试用期结束后，要想获取 Splunk Analytics for Hadoop 许可证，请联系许可证销售人员。

安装 Splunk

当使用 Splunk Analytics for Hadoop 运行搜索时，您需要在搜索头上复制 Splunk Enterprise 的 .tgz 文件，因此我们建议您使用 tar 文件安装方便后续配置。如果您通过另一种方式安装 Splunk Enterprise，当您设置搜索头时可添加必要的 tar

文件。请参阅“设置搜索头实例”。

要了解在 Linux 上安装或更新 Splunk Enterprise 的更多信息，请参阅《安装手册》中的“在 Linux 上安装”。

在开始安装之前，请查看系统要求和配置前提条件。请参阅“系统和软件要求”。

tar 文件安装

要在 Linux 系统上安装，使用 tar 命令展开 tarball 到相应目录：

```
tar xvzf Splunk_package_name.tgz
```

默认安装目录是当前工作目录中的 splunk。要安装到 /opt/splunk，使用以下命令：

```
tar xvzf Splunk_package_name.tgz -C /opt
```

注意：使用 tarball 安装时：

- tar 的一些非 GNU 版本可能没有 -C 参数。在这种情况下，要安装到 /opt/splunk（无论是 cd 至 /opt），或在运行 tar 命令之前，将 tarball 放入 /opt。这种方法适用于您的计算机文件系统上的任何可访问目录。
- 确保磁盘分区拥有足够空间可容纳您计划保留索引的未压缩数据量。

设置搜索头实例

安装 Splunk Enterprise 和许可的 Splunk Analytics for Hadoop 之后，您必须配置搜索头以支持您稍后会添加的提供程序和虚拟索引。

请参阅“设置提供程序和虚拟索引”了解有关配置提供程序和虚拟索引的更多信息。

1. 在搜索头上保留 .tgz 版本的 Splunk Enterprise 副本（即使在搜索头上安装完成，您也需要这个软件包）。

首次搜索虚拟索引时，Splunk Analytics for Hadoop 会将此软件包复制到 HDFS，然后将其提取到参与搜索的所有 TaskTracker 节点中。提取软件包用于处理 Hadoop 中的搜索结果。

如果您已使用 .tgz 版本以外的下载版本安装了 Linux 版 Splunk Enterprise，请下载 splunk_package.tgz 文件副本在搜索头上进行安装。

2. 如果您尚未安装，请先在搜索头上安装 Java。您必须这样做才能访问 Hadoop 群集。

3. 在搜索头上安装 Hadoop 客户端库。客户端库必须与您的 Hadoop 集群具有相同的 Hadoop 发行版和版本。要与不同发行版和版本的多个 Hadoop 集群通信，请在搜索头上为每个发行版安装 Hadoop CLI 软件包。

升级 Splunk Analytics for Hadoop 搜索头

下载并安装新版本的 Splunk

通过安装最新 (Linux) 版本的 Splunk 升级搜索头。要在系统上安装 Splunk，请下载适用于 Splunk Enterprise 的 Linux 分布并安装，然后添加 Splunk Analytics for Hadoop 许可证。

您可通过此网址查找要下载的正确版本：http://www.splunk.com/download/splunk

配置此安装以在驻留在 *nix 平台上的搜索头上运行。您可在任何满足搜索头要求的计算机上运行 Splunk Analytics for Hadoop。请参阅“配置搜索头”。

注意：Windows 不支持 Splunk Analytics for Hadoop。

要了解安装或更新 Splunk 的更多信息，请参阅《安装手册》中的“在 Linux 上安装”。

tar 文件安装

要在 Linux 系统上安装 Splunk，使用 tar 命令扩展 tarball 到相应目录：

```
tar xvzf splunk_package_name.tgz
```

默认安装目录是当前工作目录中的 splunk。要安装到 /opt/splunk，使用以下命令：

```
tar xvzf splunk_package_name.tgz -C /opt
```

注意：使用 tarball 安装 Splunk 时：

- tar 的一些非 GNU 版本可能没有 -C 参数。在这种情况下，要安装到 /opt/splunk（无论是 cd 至 /opt），或在运行 tar 命令之前，将 tarball 放入 /opt。这种方法适用于您的计算机文件系统上的任何可访问目录。
- 确保磁盘分区拥有足够空间可容纳您计划保留索引的未压缩数据量。

编辑 indexes.conf

在 indexes.conf 文件中，编辑属性 vix.splunk.setup.package，这样可提供 Splunk Analytics for Hadoop 可在数据节点上安装和使用新的 Splunk .tgz 软件包路径。下次 MapReduce 衍生搜索时，Splunk 会将新软件包推到节点中，并升级 DataNode 和 TaskTracker 上存储的版本。

请参阅配置文件中的“设置提供程序和虚拟索引”获取有关编辑配置文件的更多信息。

请参阅“提供程序配置变量”获取有关在 Splunk Web 中编辑此属性的更多信息。

从 Hunk 升级到 Splunk Analytics for Hadoop 的特别说明

如果您想要从 Hunk 6.3 或更早版本升级并更新到 Splunk Analytics for Hadoop，您必须执行以下步骤以适应新的命名约定：

1. 停止现有的 Splunk 实例：

```
splunk stop
```

2. 查找标记为 "Hunk" 的现有安装的文件夹。重命名为 "Splunk:"

```
mv hunk splunk
```

3. 解压缩新的 Splunk .tgz 文件：

```
tar -xvf splunk-6.xxxxxxxxx.tgz
```

4. 启用 Splunk Analytics for Hadoop 并接受许可证：

```
splunk start
```

完成上述步骤后，您不需要为之后的版本重复上述流程。

注意：您可能还需要更新搜索头，请参阅“升级 Splunk Analytics for Hadoop”

配置提供程序和虚拟索引

关于虚拟索引

虚拟索引允许 Splunk Analytics for Hadoop 地址数据存储在外部系统中，并将计算推送到这些系统。您可利用虚拟索引访问和报告驻留在 Hadoop 群集中的结构化、非结构化和多结构化的数据。

Splunk Analytics for Hadoop 可利用 MapReduce 框架在 Hadoop 节点上执行报表生成的搜索。访问数据前不需要预先处理数据，因为 Splunk Analytics for Hadoop 允许您对驻留在 Hadoop 中的数据运行分析搜索。

Splunk Analytics for Hadoop 将虚拟索引视为只读数据存储，在搜索时间将方案绑定到数据。这表示您报告的数据仍然使用其他系统和工具之前使用的相同格式访问，如 Hive 和 Pig。

配置虚拟索引

在您设置虚拟索引之前，请设置提供程序。当您配置提供程序时，您可向 Splunk Analytics for Hadoop 说明 Hadoop 群集的详细情况，ERP 进程使用哪种群集执行报告任务。ERP 是一个搜索助手进程，我们新建此进程以在 Hadoop 数据上执行搜索。

然后，您可通过向 Splunk Analytics for Hadoop 提供 Hadoop 数据相关信息配置虚拟索引，如数据位置、一组允许的和封锁的文件或目录。正确配置之后，虚拟索引会识别某些目录结构并提取和使用该信息以优化搜索。例如，如果您的数据使用日期在目录结构中分区，之后 Splunk Analytics for Hadoop 可通过适当选择仅处理相关路径中的数据减少处理的数据量。

了解更多信息

- 要使用 CLI 配置提供程序和虚拟索引，请参阅“设置提供程序和虚拟索引”。
- 要在 Splunk Web 中设置新的提供程序，请参阅“添加或编辑 HDFS 提供程序”。
- 要在 Splunk Web 中设置新的虚拟索引，请参阅“在 Splunk Web 中添加或编辑虚拟索引”。

在配置文件中设置提供程序和虚拟索引

成功安装 Splunk 和许可 Splunk Analytics for Hadoop 之后，您可修改 `indexes.conf` 以新建提供程序和虚拟索引，或使用 Splunk Web 添加虚拟索引和提供程序。

- 要在 Splunk Web 中添加虚拟索引，请参阅本手册中的“添加虚拟索引”。
- 要添加新的提供程序，请参阅本手册中的“添加 HDFS 提供程序”。

开始之前

要通过配置文件配置提供程序和虚拟索引，您可编辑 `indexes.conf`。在您编辑 `indexes.conf` 之前，您应确认 Splunk Analytics for Hadoop 具有适当权限并会收集您设置提供程序和索引器所需的信息。

配置权限

在设置提供程序之前，确保 Splunk Analytics for Hadoop 具有以下权限：

- 具有虚拟索引数据驻留的 HDFS 目录的只读访问权限。
- 具有安装 Splunk 实例的 HDFS 目录的读写访问权限。（这通常是 `splunkMR` 目录，例如：`User/hue/splunk_mr/dispatch`）。Splunk 可在此目录中新建以下目录：
  - `/dispatch` （这是存储临时结果的目录）。
  - `/packages` （这是将复制到数据节点的 Splunk `.tgz` 文件）。
  - `/bundles` （这是存储配置的位置）。
- `/tmp` 目录驻留的 Datanode 的读写访问权限。这是您在“提供程序”设置中配置 `vix.splunk.home.datanode` 时指向的临时目录。

收集以下信息

您需要了解以下有关搜索头、文件系统和 Hadoop 配置的信息：

- Hadoop 群集的主机名和 NameNode 端口。
- Hadoop 群集的主机名和 JobTracker 端口。
- Hadoop 客户端库和 Java 的安装目录。
- DataNode/TaskTracker *nix 文件系统上可写目录的路径，Hadoop 用户帐号拥有该目录的读写权限。
- HDFS 中可写目录的路径，该目录**仅供**此搜索头使用。

编辑 Indexes.conf

编辑 indexes.conf 构建虚拟索引。您可向 Splunk 介绍 Hadoop 群集和您想要通过虚拟索引访问的数据。

新建 indexes.conf

新建一份 indexes.conf 并将其放在本地目录中。在本示例中，我们将使用：

$SPLUNK_HOME/etc/system/local

注意：对 indexes.conf 做出的以下更改于搜索时间生效，无需重新启动。

新建提供程序

1. 对于不同的 Hadoop 群集，您需要新建单独的 provider 段落。在此段落中，您可提供 Java 安装路径和 Hadoop 库路径，以及针对此群集运行搜索时想要使用的其他 MapReduce 配置。

provider 段落中的属性与它继承的 family 段落合并。"vix." 前缀从各属性中剔除，且这些值将传递到 MapReduce 任务配置中。

您必须先配置提供程序。您可为提供程序配置多个索引。

```
[provider:MyHadoopProvider]
vix.family                  = hadoop
vix.env.JAVA_HOME           = /path_to_java_home
vix.env.HADOOP_HOME         = /path_to_hadoop_client_libraries
```

2. 向 Splunk 介绍群集，包括 NameNode 和 JobTracker，以及查找和安装 Splunk .tgz 副本的位置。

```
vix.mapred.job.tracker = jobtracker.hadoop.splunk.com:8021
vix.fs.default.name = hdfs://hdfs.hadoop.splunk.com:8020
vix.splunk.home.hdfs = /<the path in HDFS that is dedicated to this search head for temp storage>
vix.splunk.setup.package = /<the path on the search head to the package to install in the data nodes>
vix.splunk.home.datanode = /<the path on the TaskTracker's Linux filesystem on which the above Splunk package should be
installed>
```

新建虚拟索引

1. 为每个提供程序定义一个或多个虚拟索引。这样您可指定如何将数据分配到各目录中，哪些属于构成索引的文件以及一些有关文件内容时间范围的提示。

```
[hadoop]
vix.provider          = MyHadoopProvider
vix.input.1.path      = /home/myindex/data/${date_date}/${date_hour}/${server}/...
vix.input.1.accept    = \.gz$
vix.input.1.et.regex  = /home/myindex/data/(\d+)/(\d+)/
vix.input.1.et.format = yyyyMMddHH
vix.input.1.et.offset = 0
vix.input.1.lt.regex  = /home/myindex/data/(\d+)/(\d+)/
vix.input.1.lt.format = yyyyMMddHH
vix.input.1.lt.offset = 3600
```

- 对于 vix.input.1.path：提供属于此索引的数据的完全限定路径以及您想从该路径中提取的任何字段。

例如：

```
/some/path/${date_date}/${date_hour}/${host}/${sourcetype}/${app}/...
```

${}'s 随附的任何项目均提取为字段，并添加到该路径的各搜索结果中。此搜索会忽略和搜索字符串不匹配的目录，这样可明显提高性能。

- 对于 vix.input.1.accept，提供要匹配的文件的正则表达式列表。
- 对于 vix.input.1.ignore，提供要忽略的文件的正则表达式列表。注意：忽略优先于接受。

2. 使用正则表达式、格式和偏移值提取特定路径中包含的数据的时间范围。此时间范围由两部分组成：最早时间 vix.input.1.et 和最晚时间 vix.input.1.lt。以下配置可用：

- 对于 vix.input.1.et/lt.regex，提供和提供日期和时间的目录的一部分匹配的正则表达式，允许从路径解释时间。

使用捕获组提取组成时间戳的部分。捕获组值连接在一起，并按照指定格式解释。从路径中提取时间范围会忽略不在搜索时间范围内的目录，从而明显提高搜索特定事件窗口的速度。

- 对于 `vix.input.1.et/lt.format`，提供日期/时间格式字符串，用于解释从上述正则表达式中提取的数据。格式字符串规格可在 SimpleDateFormat 中查找。

还支持以下两种非标准格式：`epoch` 将数据解释为 epoch 时间和 `mtime` 以使用文档的修改时间而不是正则表达式提取的数据。

- 对于 `vix.input.1.et/lt.offset`，您可选择用此提供偏移，以考虑时区和/或安全边界。

设置提供程序配置变量

Splunk Analytics for Hadoop 还可为您新建的每个提供程序预设配置变量。您可以保留预设变量，或者根据需要编辑它们。如想编辑这些变量，请参阅本手册参考章节中的“提供程序配置变量”。

注意：若为了使用 YARN 而配置 Splunk Analytics for Hadoop，您必须添加新的设置。请参阅本手册中的“YARN 必需的配置变量”。

选择编辑 `props.conf` 定义数据处理

您可编辑 `props.conf` 定义如何处理数据文件。两种类型都接受索引和搜索时间属性。以下示例显示如何使用索引和搜索时间属性处理 `twitter` 数据（代表推特推文的 json 对象）。显示的是单独行 json 数据，其中 `_time` 是已计算字段（请注意：我们已禁用索引-时间的时间戳）

```
[source::/home/somepath/twitter/...]
priority         = 100
sourcetype       = twitter-hadoop
SHOULD_LINEMERGE = false
DATETIME_CONFIG  = NONE

[twitter-hadoop]
KV_MODE          = json
EVAL-_time       = strptime(postedTime, "%Y-%m-%dT%H:%M:%S.%lZ")
```

添加来源类型

设置提供程序和索引之后，您还可配置 Splunk Analytics for Hadoop 以按来源类型搜索虚拟索引。

尽管大部分来源类型是日志格式，任意常用数据导入格式也可是来源类型。如果您的数据非常罕见，则您可能需要新建一个包含自定义事件处理设置的来源类型。如果您的数据源包含异类数据，则可能需要根据每个事件（而不是每个来源）分配来源类型。

请参阅 Splunk Enterprise 文档中的“来源类型为何重要”了解有关您为何想在 HDFS 数据上使用来源类型的更多信息。

要为 HDFS 数据来源添加来源类型，您可将段落添加到 `$SPLUNK_HOME/etc/system/local/props.conf`。为 HDFS 数据定义来源类型时，请记住：HDFS 数据搜索在搜索时间进行，不在索引时间搜索，Hunk 只读取最晚时间戳，而不是初始的 HDFS 时间戳。因此，时间戳识别可能不会始终像预期的那样有效。

在以下示例中，我们可添加两种来源类型。新的来源类型 `access_combined` 表示 access_combined 日志文件中的数据。 `mysqld` 允许您搜索指定的 `mysqld.log` 文件中的数据：

```
[source::.../access_combined.log]
sourcetype=access_combined
priority=100

[source::.../mysqld.log]
sourcetype=mysqld
priority=100
```

（您无需重新启动）

有关搜索的信息，包括按来源类型搜索，请参阅《Splunk Enterprise 搜索教程》中的“使用字段搜索”。

添加来源类型时请注意以下要点：

- INDEXED_TIME 提取不适用于 Splunk Analytics for Hadoop。
- 尽管搜索时间提取应适用于 Splunk Analytics for Hadoop，但是通过将其添加到默认列表，可以更轻松地使用

SimpleCSVRecordReader 进行查找（如果文件有标题）：

```
#append the SimpleCSVRecordReader to the default list:
vix.splunk.search.recordreader = ...,com.splunk.mr.input.SimpleCSVRecordReader
vix.splunk.search.recordreader.csv.regex = <a regex to match csv files>
vix.splunk.search.recordreader.csv.dialect = tsv
```

在配置文件中设置虚拟索引

使用以下流程使用配置文件设置虚拟索引。请参阅本手册中的“添加或编辑虚拟索引”获取通过 Splunk Web 添加虚拟索引相关信息。

1. 在 indexes.conf 中，为每个提供程序定义一个或多个虚拟索引。这样您可指定如何将数据分配到各目录中，哪些属于构成索引的文件以及一些有关文件内容时间范围的提示。

```
[hadoop]
vix.provider = MyHadoopProvider
vix.input.1.path = /home/myindex/data/${date_date}/${date_hour}/${server}/...
vix.input.1.accept = \.gz$
vix.input.1.et.regex = /home/myindex/data/(\d+)/(\d+)/
vix.input.1.et.format = yyyyMMddHH
vix.input.1.et.offset = 0
vix.input.1.lt.regex = /home/myindex/data/(\d+)/(\d+)/
vix.input.1.lt.format = yyyyMMddHH
vix.input.1.lt.offset = 3600
```

- 对于 vix.input.1.path：提供属于此索引的数据的完全限定路径以及您想从该路径中提取的任何字段。

例如：

```
/some/path/${date_date}/${date_hour}/${host}/${sourcetype}/${app}/...
```

${}'s 随附的任何项目均提取为字段，并添加到该路径的各搜索结果中。此搜索会忽略和搜索字符串不匹配的目录，这样可明显提高性能。

- 对于 vix.input.1.accept，提供要匹配的文件的正则表达式列表。
- 对于 vix.input.1.ignore，提供要忽略的文件的正则表达式列表。注意：忽略优先于接受。

2. 使用正则表达式、格式和偏移值提取特定路径中包含的数据的时间范围。此时间范围由两部分组成：最早时间 vix.input.1.et 和最晚时间 vix.input.1.lt。以下配置可用：

- 对于 vix.input.1.et/lt.regex，提供和提供日期和时间的目录的一部分匹配的正则表达式，允许从路径解释时间。使用捕获组提取组成时间戳的部分。捕获组值连接在一起，并按照指定格式解释。从路径中提取时间范围会忽略不在搜索时间范围内的目录，从而明显提高搜索特定事件窗口的速度。
- 对于 vix.input.1.et/lt.format，提供日期/时间格式字符串，用于解释从上述正则表达式中提取的数据。格式字符串规格可在 SimpleDateFormat 中查找。您可以将此值设为 epoch 将时间解释为秒。
- 对于 vix.input.[N].et/lt.value，您可以指定 mtime 以使用文档的修改时间而不是正则表达式提取的数据。
- 对于 vix.input.1.et/lt.offset，您可选择用此提供偏移，以考虑时区和/或安全边界。

添加或编辑 Splunk Web 内的 HDFS 提供程序

您可以为一个提供程序设置多个含多个索引的提供程序。当您添加虚拟索引之后，将获取以下信息：

- Hadoop 群集的主机名和 NameNode 端口。
- Hadoop 群集的主机名和 JobTracker 端口。
- Hadoop 命令行库和 Java 安装的安装目录。
- DataNode/TaskTracker *nix 文件系统上可写目录的路径，Hadoop 用户帐号拥有该目录的读写权限。
- HDFS 中可写目录的路径，该目录在此搜索头上仅供 Splunk 使用。

您也可以通过编辑 indexes.conf 添加 HDFS 提供程序和虚拟索引。请参阅“设置虚拟索引”获取在配置文件中设置虚拟索引的说明。

添加提供程序

1. 在顶部菜单中选择**设置 > 虚拟索引**。

2. 选择“虚拟索引”页面中的**提供程序**选项卡并单击**新提供程序**或您想要编辑的提供程序名称。

3. “添加新/编辑提供程序”页面可让您为提供程序**命名**。

4. 在下拉列表中选择**提供程序系列**（请注意，此字段无法编辑）。

5. 提供下列**环境变量**：

- **Java Home**：提供 Java 实例的路径。
- **Hadoop Home**：提供 Hadoop 客户端目录的路径。

6. 提供以下 **Hadoop 群集信息**：

- Hadoop 版本：告诉 Splunk Analytics for Hadoop 群集运行的是以下哪个 Hadoop 版本：Hadoop 1.0、带 MRv1 的 Hadoop 2.0 或带 Yarn 的 Hadoop 2.0。
- **JobTracker**：提供任务追踪器的路径。
- **文件系统**：提供默认文件系统的路径。

7. 提供以下 **Splunk 设置**：

- **HDFS 工作目录**：该路径位于 HDFS 或默认文件系统（无论它是什么）内，您想要把 HDFS 或默认文件系统用作工作目录。
- **任务队列**：您想把此提供程序的 MapReduce 任务提交到该任务队列中。

8. 单击**添加安全群集**为群集配置安全性，并提供您的 Kerberos 服务器配置。

9. **其他设置**字段指定您的提供程序配置变量。Splunk Analytics for Hadoop 会为您新建的每个提供程序填充预设配置变量。您可以保留预设变量，或者根据需要编辑它们。如想了解更多有关这些设置的信息，请参阅本手册参考部分里的“提供程序配置变量”。

注意：若为了使用 YARN 而配置 Splunk Analytics for Hadoop，您必须添加新的设置。请参阅本手册中的“YARN 必需的配置变量”。

9. 单击**保存**。

在 Splunk Web 中添加或编辑虚拟索引

您也可以通过编辑添加 HDFS 提供程序和虚拟索引。请参阅配置文件中的“设置虚拟索引”获取在配置文件中设置虚拟索引的说明。

1. 选择**设置 > 虚拟索引**。

2. 选择**虚拟索引**选项卡并单击**新虚拟索引**或您想要编辑的索引名称。新/编辑虚拟索引页面显示：

3. 在**名称**字段中，提供虚拟索引名称。

4. 选择**提供程序**。要添加新的提供程序，请参阅“添加 HDFS 提供程序”。

5. 提供以下路径信息：

- **HDFS 中的数据路径**：这是 Splunk Analytics for Hadoop 访问和报告的数据路径。例如：

```
/home/data/apache/logs/
```

- **以递归方式处理目录**：如果您想要（以递归方式）包含子目录的内容，请勾选。
- **白名单**：提供和文件路径匹配的正则表达式。您可指定正则表达式根据完整路径筛选出或过滤应该或不应该被视为虚拟索引一部分的文件。常见的使用案例是忽略临时文件，或者目前正写入的文件。请记住：忽略优先于接受。例如： `\.gz$`

6. 勾选**自定义时间戳格式**以打开控制选项，允许您自定义如何根据时间戳信息收集数据。使用简单的日期格式选择自定义以下内容：

- **时间捕获正则表达式**：提供一个正则表达式，确定根据时间戳收集和处理的最早日期/时间。例如：
  `/home/data/(\d+)/(\d+)/`

- **时间格式：**关于上述最早时间，提供描述如何解释提取的时间字符串的时间格式。例如： `yyyyMMddHH`
- **时间调整：**要添加到最早时间的时间量（以秒为单位）。示例（+7 小时）：25200
- **时间范围：**提供索引应收集数据的时间范围。
- **时区：**选择时区。

配置 Kerberos 验证

要配置 Kerberos 验证，请将以下各行添加到 `indexes.conf` 中的相关提供程序段落：

```
vix.hadoop.security.authentication = kerberos
vix.java.security.krb5.kdc = <kerberos server name>
vix.java.security.krb5.realm = <kerberos default realm>
vix.kerberos.principal = <kerberos principal name of the user you want Splunk Analytics for Hadoop to interact with Hadoop, for example: SAH@YOUR-REALM.COM>
vix.kerberos.keytab = <kerberos keytab path, i.e., /path/yourdir.keytab>
vix.hadoop.security.authorization = <hadoop security authorization true/false>
vix.dfs.namenode.kerberos.principal = <hadoop namenode kerberos principal name, i.e., hdfs/_HOST@YOUR-REALM.COM>
vix.mapreduce.jobtracker.kerberos.principal = <the hadoop jobtracker kerberos principal name, i.e., mapred/_HOST@YOUR-REALM.COM>
vix.hadoop.security.auth_to_local = <the mapping from Kerberos principals to short names (optional)>
vix.mapred.job.reuse.jvm.num.tasks = 1
```

注意：设置 `vix.mapred.job.reuse.jvm.num.tasks = 1` 允许您避免 ENOENT 任务故障（详情请访问 https://issues.apache.org/jira/browse/MAPREDUCE-4490）。

如果您正在使用 YARN，您还必须将以下属性添加到提供程序段落：

```
vix.yarn.resourcemanager.principal = yarn/_HOST@YOUR-REALM.COM
vix.yarn.nodemanager.principal = yarn/_HOST@YOUR-REALM.COM

# kerberos with Hive
vix.hive.metastore.sasl.enabled = <true|false>
vix.hive.metastore.kerberos.principal = <service principal for the metastore thrift server>
```

如果您正在通过 Metastore 预处理 Hive，`vix.hive.metastore.sasl.enabled` 必须设为 "true"。

管理用户和验证

关于传递验证

要搜索虚拟索引，Splunk Analytics for Hadoop 会请求 MapReduce 任务并访问 HDFS 文件。默认情况下，Splunk Splunk Analytics for Hadoop 会以 Splunk Analytics for Hadoop 超级用户身份执行此操作。使用传递验证，但是，您可控制哪些 Splunk Analytics for Hadoop 用户可提交 MapReduce 任务和访问 HDFS 文件。您还可指定 MapReduce 任务应使用哪个队列。

关于 Splunk Analytics for Hadoop 超级用户、Splunk Analytics for Hadoop 用户和 Hadoop 用户

Splunk Analytics for Hadoop 超级用户是以下一种（或多种）用户类型：

- 用于安装 Splunk 搜索头的用户。
- 提供程序的 Kerberos keytab 用户。

Hadoop 用户是 Hadoop 群集允许执行以下操作的用户：

- 提交 MapReduce 任务。
- 访问 HDFS，假设它使用节点的操作系统用户/组，因为 Hadoop 通常默认情况下都会这样假定。（如果 Hadoop 群集的配置方式不同，效果可能不同。）

用户如何使用传递验证

传递验证允许您指定 Splunk Analytics for Hadoop 超级用户为任意数量的已配置 Splunk Analytics for Hadoop 用户的代理。这样，Splunk Analytics for Hadoop 用户可作为 Hadoop 用户拥有 Hadoop 中的相关作业、任务和文件（您可限制访问 HDFS 中的文件）。

可通过 Splunk 本地用户功能或 LDAP 新建 Splunk Analytics for Hadoop 用户。有关设置用户的更多信息，请参阅*确保 Splunk Enterprise 安全*手册中的以下主题：

- 设置使用 LDAP 进行的用户验证
- 使用 Splunk 的内置系统设置用户验证

您可使用传递验证的方式

以下使用案例介绍了您可使用传递验证的常用方式：

- **为一个 Hadoop 用户配置一个 Splunk Enterprise for Hadoop 用户：**例如，您可配置 Splunk Analytics for Hadoop 用户作为和特定队列或数据集相关的 Hadoop 用户。在这种情况下，您只需将用户映射到 Hadoop 中的特定用户。例如，如果 Splunk Analytics for Hadoop 用户名称为 "msantos"，在 Hadoop 群集上，名称为 "mattsantos"，队列为 "Products"。
- **为一个 Hadoop 用户配置很多 Splunk Analytics for Hadoop 用户：**您可能想要配置多个 Splunk Analytics for Hadoop 用户作为 Hadoop 用户。例如，以下所有 Splunk Analytics for Hadoop 用户均可以 Hadoop 上 "Executive" 用户身份运行，并分配到队列 "Products"：
  - jbartlett
  - lmcgarry
  - jlyman
- **为不同队列的相同 Hadoop 用户配置 Splunk Analytics for Hadoop 用户：**您还可运行 Splunk Analytics for Hadoop 用户作为相同的 Hadoop 用户，但是为他们分配不同的队列。例如，上个示例中的用户可以使用 Hadoop 上的 "Executive" 用户身份运行，并分配到队列 "Products"。但是不同用户会各自分配到以下队列中：
  - jbartlett 以 "Executive" 用户身份运行，并分配到队列 "prod-admin"。
  - lmcgarry 以 "Executive" 用户身份运行，并分配到队列 "prod-staff"。
  - jlyman 以 "Executive" 用户身份运行，并分配到队列 "prod-staff"。

在 Splunk Web 中配置传递验证

通过传递验证，Splunk Analytics for Hadoop 会将其超级用户用作 Hadoop 代理，这样您可作为 Hadoop 用户与 Hadoop 交互。

您可将超级用户配置为 Hadoop 用户，名称和 Splunk Analytics for Hadoop 用户相同，或者您可将超级用户配置为其他名称的用户。

要了解有关传递验证如何工作的更多信息，请参阅“关于传递验证”。

配置 Hadoop 用户以允许传递验证

启用传递验证后，在 Hadoop 用户充当已登录的 Splunk Analytics for Hadoop 用户的代理时和 Hadoop 交互。

1. 确保您想要 Splunk Analytics for Hadoop 用户充当的任何 Hadoop 用户都存在于每个 Hadoop 节点上。您可手动新建或使用 LDAP 新建。

2. 确保 Splunk Analytics for Hadoop 超级用户在 Hadoop 超级组中。您可在 `hdfs-site.xml` 文件中找到充当 `dfs.permissions.supergroup` 的 Hadoop 超级组。

如果您的超级用户不在每个 Hadoop 节点上的超级组中，请使用以下命令为每个节点上的超级组添加超级用户：

```
sudo usermod -G <group name><user name>.
```

3. 在 HDFS 中为 Hadoop 群集中的用户新建主页目录，并确保 HDFS 中提供程序的 Hadoop 主页 `vix.splunk.home.hdfs` 可由步骤 2 中添加的所有用户读取和执行。

4. 将段落添加到 `core-site.xml` 允许 Hadoop 用户（具有和超级用户相同的名称）作为指定节点用户组中的 Hadoop 用户的代理：

注意：要获得最好的结果，我们建议您对 Kerberized 群集执行此操作。有关使用 Kerberos 的更多信息，请参阅“配置 Kerberos 验证”。

```
<property>
<name>hadoop.proxyuser.<name of your Superuser>.groups</name>
<value>group1,group2</value>
<description>Allows the Superuser to impersonate any
members of the group group1 and group2</description>
</property>
```

5. 选择限制主机连接：

```
<property>
<name>hadoop.proxyuser.<name of your Superuser>.hosts</name>
<value>host1,host2</value>
<description>The superuser can connect only from host1 and
host2 to impersonate a user</description>
</property>
```

配置传递验证

通过 Splunk Enterprise 本地用户功能或 LDAP 为存在于 Splunk Analytics for Hadoop 中的任何用户配置传递验证。您可为一个或多个用户和 (LDAP) 用户组配置传递验证。

有关传递验证如何工作的更多信息，请参阅“关于传递验证”。

1. 单击**设置 > 虚拟索引**。

2. 单击**传递验证**选项卡。

3. 选择用来映射用户的**提供程序**。

4. 在**用户**字段中，选择想要映射到 Hadoop 用户的现有 Splunk Analytics for Hadoop 用户。

5. 键入您想要 Splunk Analytics for Hadoop 用户“扮演”的 **Hadoop 用户**名称。

6. 或者，选择与您添加的 Hadoop 用户相关的**队列**。如果您没有选择队列，Splunk Analytics for Hadoop 用户可访问任何与 Hadoop 用户相关的队列。

7. 单击**保存**。

在配置文件中配置传递验证

本主题介绍如何编辑 `indexes.conf` 和 `impersonation.conf` 使 Splunk Analytics for Hadoop 用户充当 Hadoop 用户。这样您就可以授予特定 Splunk Analytics for Hadoop 用户权限，以不同的 Hadoop 用户身份将 MapReduce 任务提交到特定队列。要使用 Splunk Web 用户界面配置传递验证，请参阅“映射传递验证”。

通过传递验证，Splunk Enterprise for Hadoop 将其超级用户用作 Hadoop 代理，这样您可与 Hadoop 交互。您可将此配置为名称和超级用户相同的 Hadoop 用户，或配置为具有不同名称的用户。

要了解有关传递验证如何工作的更多信息，请参阅“关于传递验证”。

配置 Hadoop 群集以支持传递验证

启用传递验证后，以 Hadoop 用户身份与 Hadoop 交互，该用户的名称与登录的 Splunk Analytics for Hadoop 用户名相同。必须按以下方式配置 Hadoop 才能提供支持：

1. 确保您想要 Splunk Analytics for Hadoop 用户充当的任何 Hadoop 用户都存在于每个 Hadoop 节点上。您可手动新建或使用 LDAP 新建。

2. 确保超级用户在 Hadoop 超级组中。您可在 `hdfs-site.xml` 文件中找到充当 `dfs.permissions.supergroup` 的 Hadoop 超级组。

如果您的超级用户不在每个 Hadoop 节点上的超级组中，请使用以下命令为每个节点上的超级组添加超级用户：

```
sudo usermod -G <group name><user name>.
```

3. 在 HDFS 中为 Hadoop 群集中的用户新建主页目录，并确保 HDFS 中提供程序的 Hadoop 主页 `vix.splunk.home.hdfs` 可由步骤 2 中添加的所有用户读取和执行。

4. 将段落添加到 `core-site.xml` 允许 Hadoop 用户（具有和 Splunk Analytics for Hadoop 超级用户相同的名称）充当指定节点用户组中的 Hadoop 用户的代理：

注意：要获得最好的结果，我们建议您对 Kerberized 群集执行此操作。有关使用 Kerberos 的更多信息，请参阅“配置 Kerberos 验证”。

```
<property>
<name>hadoop.proxyuser.<name of your Splunk Analytics for Hadoop Superuser>.groups</name>
<value>group1,group2</value>
<description>Allows the Splunk Analytics for Hadoop Superuser to impersonate any
members of the group group1 and group2</description>
</property>
```

5. 选择限制主机连接：

```
<property>
<name>hadoop.proxyuser.<name of your Splunk Analytics for Hadoop Superuser>.hosts</name>
<value>host1,host2</value>
<description>The superuser can connect only from host1 and
host2 to impersonate a user</description>
</property>
```

在 Splunk Analytics for Hadoop 中配置传递验证

在接下来的步骤中，您可配置 Splunk Analytics for Hadoop 用户以将任务提交给特定队列和/或作为名称和 Splunk Analytics for Hadoop 登录用户不同的 Hadoop 用户和 Hadoop 交互。

1. 为每个提供程序启用 `indexes.conf` 中的功能：

```
[provider:myprovider]
vix.splunk.impersonation = 1
```

2. 或者通过更新 `impersonation.conf` 将 Splunk Analytics for Hadoop 用户映射到 HDFS 中的特定别名和队列：

```
[provider:myprovider]
admin = {"user": "hadoopadmin"}
splunkanalyticsforhadoopuser1 = {"queue": "red"}

[provider:mycdh]
splunkanalyticsforhadoopuser1 = {"user": "hadoopuser", "queue": "blue"}
```

使用 Hadoop 归档文件

配置 Splunk Analytics for Hadoop 以读取 Hadoop 归档（HAR）文件

要允许 Splunk Analytics for Hadoop 读取 Hadoop 数据库中的归档文件，请将以下段落添加到 indexes.conf 文件中：

```
[provider:<provider_name>]
vix.env.HADOOP_HOME = <path_to_hadoop>
vix.env.JAVA_HOME = <path_to_java>
vix.family = hadoop
vix.fs.default.name = hdfs://<namenode>:<port>
vix.mapred.job.tracker = <jobtracker>:<port>
vix.splunk.home.hdfs = <path_on_hdfs>
[<vix_name>]
vix.input.1.path = har:///<path_to_archive_file>/<archive_file>.har/...
vix.provider = <provider_name>
```

使用非 HDFS 文件类型

使用 Hive 和 Parquet 数据

数据预处理器

当 Splunk Analytics for Hadoop 开始搜索非 HDFS 输入数据时，将使用 FileSplitGenerator 类中包含的信息确定如何拆分数据以进行平行处理。

默认 FileSplitGenerator 包含和 Hadoop 的 FileInputFormat 中定义的相同的数据拆分逻辑。这表示其适用于 Hadoop InputFormat 实现（具有和 FileInputFormat 相同的拆分逻辑）可读取的任何数据格式。

由于默认 FileSplitGenerator 不适用于 Hive 或 Parquet 文件，Splunk Analytics for Hadoop 会提供 HiveSplitGenerator 和 ParquetSplitGenerator 用于 Hive 和 Parquet。包含基于文件的拆分逻辑的任何自定义 Hive 文件（如用 Hadoop FileOutputFormat 及其子类新建的文件）可使用 HiveSplitGenerator。如果您有自定义 Hive 文件格式，而这些格式没有使用基于文件的数据拆分逻辑，您可实施使用拆分逻辑的自定义 SplitGenerator。

所有工具（包括 Hive）新建的 Parquet 文件可使用（且仅可使用）ParquetSplitGenerator。

- 要配置 Splunk Analytics for Hadoop 以使用 Hive，请参阅“配置 Hive 连接”。
- 要配置 Splunk Analytics for Hadoop 以使用 Parquet 表，请参阅“配置 Parquet 表”。

配置 Hive 连接

默认情况下，Hive 将多种文件格式的数据保存为二进制文件或一组用特殊字符分隔的文本文件。Splunk Analytics for Hadoop 目前支持 4 Hive (v0.12) 文件格式类型：Textfile、RCfile、ORC 文件和 Sequencefile。

Splunk Analytics for Hadoop通过其预处理器框架支持不同的文件格式，从而提供了一个名为 HiveSplitGenerator 的数据预处理器。该数据处理器让 Splunk Analytics for Hadoop 可以访问和处理由 Hive 存储或使用的数据。

编辑 indexes.conf 是配置 Splunk Analytics for Hadoop 连接 Hive 表最简单的方式：

- 为 Splunk Analytics for Hadoop 提供 metastore URI。
- 指定 Splunk Analytics for Hadoop 使用 HiveSplitGenerator 读取 Hive 数据。

如果您不想要 Splunk Analytics for Hadoop 访问 metastore 服务器，您可手动配置以访问组成 Hive 表的元数据文件。请参阅本主题中的“配置 Splunk Analytics for Hadoop 无需连接 Metastore 即可读取 Hive 表”。

Splunk Analytics for Hadoop 目前支持以下 Hive 版本：

- 0.10
- 0.11
- 0.12
- 0.13
- 0.14
- 1.2
- 3.1.2

Hive 3.1.2 支持 Hadoop3.x。此处列出的早期 Hive 版本只能支持 Hadoop 2.x 或更低版本。

开始之前

要设置 Splunk Analytics for Hadoop 读取 Hive 表，如果您尚未设置索引和提供程序，则必须先进行配置，请参阅：

- 在配置文件中设置提供程序和虚拟索引
- 在用户界面中添加或编辑虚拟索引
- 添加 HDFS 提供程序

确保您的 Hadoop 和 Hive 版本兼容

设置 Hadoop 数据提供程序时，请确保它使用兼容版本的 Hive。Hadoop 2.x 或更低版本需要搭配版本为 2.x 或更低的 Hive。Hadoop 3.x 要求的 Hive 版本至少为 3.x。如果使用 2.x 或更低版本的 Hive 实例配置版本 3.x 的 Hadoop 群集，则当您尝试以 Hive 文件格式保存文件时会遇到连接问题。

配置 Hive 与 metastore 连接

要配置 Hive 连接，您需提供 vix.hive.metastore.uris。

Splunk Analytics for Hadoop 会使用 Metastore 服务器中提供的信息读取表信息，包括列名称、类型、数据位置和格式，以允许其处理搜索请求。

以下是适当启用 Hive 连接的已配置提供程序段落示例。注意：表中包含一个或多个文件，每个虚拟索引可能有多个输入路径，每个表有一个路径。

```
[provider:BigBox]
...
vix.splunk.search.splitter = HiveSplitGenerator
vix.hive.metastore.uris = thrift://metastore.example.com:9083

[orders]
vix.provider = BigBox
vix.input.1.path = /user/hive/warehouse/user-orders/...
vix.input.1.accept = \.txt$
vix.input.1.splitter.hive.dbname = default
vix.input.1.splitter.hive.tablename = UserOrders

vix.input.2.path = /user/hive/warehouse/reseller-orders/...
vix.input.2.accept = .*
vix.input.2.splitter.hive.dbname = default
vix.input.2.splitter.hive.tablename = ResellerRrders
```

在极少数情况下，表的 Hadoop InputFormat 实现的拆分逻辑和 Hadoop 的 FileInputFormat 的拆分逻辑不同，HiveSplitGenerator 拆分逻辑不适用。或者，您必须执行自定义 SplitGenerator，用于替换默认 SplitGenerator。请参阅“配置 Splunk Analytics for Hadoop 以使用自定义文件格式”了解更多信息。

配置 Splunk Analytics for Hadoop 以使用自定义文件格式

要使用自定义文件格式，您可编辑提供程序段落以添加包含自定义类别的 .jar 文件，如下所示：

```
vix.splunk.jars
```

注意：如果您没有指定 InputFormat 类别，文件将被视为文本文件，并用换行符拆分到记录中。

配置 Splunk Analytics for Hadoop 无需连接 Metastore 即可读取 Hive 表

如果您无法或不希望显示您的 Metastore 服务器，可通过指定其他配置项目配置 Hive 连接。对于 Splunk Analytics for Hadoop，至少需要以下信息：

- columnnames
- columntypes

如果新建表格时指定其他信息，则需要该信息（例如，如果您的表指定 InputFormat 而不是 Hive，那么您必须告诉 Splunk Analytics for Hadoop）。

在为 Splunk Analytics for Hadoop 提供 Hive 表的列名称和类型列表的 indexes.conf 中新建段落。这些列名称会变成在 Splunk Analytics for Hadoop 中运行报表时可看到的字段名称：

```
[your-provider]
vix.splunk.search.splitter = HiveSplitGenerator

[your-vix]
vix.provider = your-provider
vix.input.1.path = /user/hive/warehouse/employees/...
vix.input.1.splitter.hive.columnnames = name,salary,subordinates,deductions,address
vix.input.1.splitter.hive.columntypes =
string:float:array<string>:map<string,float>:struct<street:string,city:string,state:string,zip:int>
vix.input.1.splitter.hive.fileformat  = sequencefile
vix.input.2.path = /user/hive/warehouse/employees_rc/...
```

分区表数据

使用 Hive Metastore 时，Splunk Analytics for Hadoop 会自动分析表，保留分区密钥和值，并根据搜索条件，去除不必要的分区。这有助于提高搜索速度。

不使用 Metastore 时，您可更新 [virtual-index] 段落以使用键值作为文件路径的一部分，告诉 Splunk Analytics for Hadoop 有关分区的信息。例如，以下配置

```
vix.input.1.path = /apps/hive/warehouse/sdc_orc2/${server}/${date_date}/...
```

可提取和识别以下路径中的 "server" 和 "date_date" 分区

```
/apps/hive/warehouse/sdc_orc2/idxr01/20120101/000859_0
```

以下是 Splunk Analytics for Hadoop 在没有任何额外配置的情况下自动识别相同分区的分区路径示例

```
/apps/hive/warehouse/sdc_orc2/server=idxr01/date_date=20120101/000859_0
```

配置 Parquet 连接

要预处理使用 Parquet 的表，Splunk Analytics for Hadoop 会使用被称为 ParquetSplitGenerator 的预处理器。要使用 ParquetSplitGenerator 读取 Parquet 表，请更新 [provider] 段落以指定 ParquetSplitGenerator 并在 [virtual index] 段落中指定路径：

```
[provider:your-provider]
vix.splunk.search.splitter = ParquetSplitGenerator

[your-vix]
vix.input.1.path = /user/hive/warehouse/t1/...
```

为获得最佳的效果，还需要指定虚拟索引时间戳：

```
vix.input.1.required.fields = timestamp
```

有关虚拟索引通用设置的更多信息，请参阅“关于虚拟索引”。

分布式部署

配置搜索头群集化

如果您为 Splunk Analytics for Hadoop 和 Splunk Enterprise 授权许可了至少三个 Splunk Analytics for Hadoop 实例，您可为 Splunk 配置搜索头群集化。

- 在所有实例上手动复制 indexes.conf，并维护所有搜索头群集成员的信息。（不建议使用）
- 使用搜索头群集 Deployer 功能更新索引配置。（建议使用）

要了解有关 Deployer 和搜索头群集架构的更多信息，请参阅“关于搜索头群集化”。

使用 Deployer 安装和配置

1. 在不是搜索头群集一部分的实例上安装和配置 Deployer。

2. 在 Deployer 上新建您想要传输的配置。例如，要将 indexes.conf 配置从搜索应用部署到搜索头群集的所有成员，您可在以下目录中为 Deployer 实例新建搜索应用：

```
SPLUNK_HOME/etc/shcluster/apps.
```

3. 前往：

```
SPLUNK_HOME/etc/shcluster/apps/search/local/
```

4. 新建或编辑 indexes.conf 和 props.conf（如适用），以及您为 Splunk Analytics for Hadoop 新建和跨群集传输所需的任何其他文件。

5. 运行以下命令：

```
SPLUNK_HOME/bin/splunk apply shcluster-bundle -target https://<any_member_SHC>:<mgmt_port> -auth admin:<password>
```

6. 阅读警告并单击**确定**。Splunk 可在搜索头群集成员上执行轮询重启，并且必须使用传输部署重启。

注意：您不能执行其他部署，除非轮询重启已启动并完成。如果您不确定轮询重启是否已完成，您可在任何成员上运行 SPLUNK_HOME/bin/splunk show shcluster-status，并检查是否在群集中运行所有实例。

计划软件包复制和软件包获取

您可为 HDFS 上的软件包设置自定义复制因子。增加软件包复制因子可通过减少跨任务节点的软件包的平均访问时间来提高大型群集的性能。

```
vix.splunk.setup.bundle.replication = <positive integer between 1 and 32767>
```

设置获取时间限制以指定可以删除的每个数据节点上工作目录中的软件包时间。

```
vix.splunk.setup.bundle.reap.timelimit = <positive integer in milliseconds>
```

默认为 24 小时，这也是最大值。任何大于 24 小时的值都会被视为 24 小时。

使用 HDFS 浏览器向导

在 HDFS 浏览器中浏览和配置 Hadoop 来源文件

HDFS 浏览器向导允许您直观地浏览 Hadoop 来源文件并为这些文件配置来源类型、分区、时间戳和应用上下文。

通过向导，您可执行以下操作：

- 查看 Hadoop 目录并钻取您想要查看的文件。
- 添加来源类型并配置分区和时间戳。
- 查看事件摘要信息。
- 添加应用上下文和共享以确定哪些应用和用户可使用来源配置。
- 修改来源文件路径以扩展或限制搜索。

更多信息，请参阅“配置 HDFS 来源”。

配置 HDFS 来源

1. 在主页面中，单击**浏览数据**。

2. 选择**提供程序**和**虚拟索引**。可用索引列表的内容取决于您在**提供程序**下拉菜单中选择的提供程序。

3. 单击**下一步**。

4. 在目录和来源文件列表中，选择您想要浏览的项目。您可钻取可用目录的文件级别。

5. 如果您浏览了选定的来源但不想要查看或修改来源配置，您可立即关闭窗口，或单击返回箭头浏览其他文件。要查看和修改来源配置，单击**下一步**。

6. 在“预览数据”页面，您可查看和修改**事件处理**的文件设置确保您的数据处理正确。

   - **查看摘要信息**：单击页面左侧的“查看事件摘要”查看：
     - 示例数据的大小（以字节为单位）。
     - 事件数。
     - 代表事件随时间的分布情况的图表。
     - 每个事件占用的行数明细。

   - **调整时间戳和事件换行**：编辑时间戳和事件换行，将编辑内容另存为新的**来源类型**并应用于数据。请参阅*数据导入*手册中的“为你的数据分配正确的来源类型”了解来源类型相关信息。

注意：将 Splunk Analytics for Hadoop 中的 CSV 和 Avro 数据分析为 JSON。如果您查看数据时遇到问题，请尝试将来源类型更改为 JSON。

   - **添加高级设置**：为来源数据添加新属性。有关添加属性的更多信息，请参阅“提供程序配置变量”了解这些属性的更多信息。

7. 单击**下一步**以保存所有更改。

8. 在“输入上下文设置”页面，为数据来源分配**应用上下文**和**应用共享**。这样可确定将使用该文件配置设置的应用和用户。

9. 单击**下一步**。

10. 查看**查看设置**页面中的配置信息。您还可选择更改来源文件。例如，您可添加通配符扩大搜索的数据来源或进一步限制。或者，使用“来源文件剪贴板”窗口将来源配置复制到目录中的其他文件中。

10. 单击**完成**。您现可搜索已配置的来源或浏览其他来源文件。

关于虚拟索引数据的搜索和报表

Splunk Analytics for Hadoop 中可分布式和不可分布式命令如何工作（以及怎样工作效果最佳）

可分布式搜索命令是 Splunk Analytics for Hadoop 报表中最有效的，因为这些命令可以分布到搜索头和虚拟索引。一般而言，不可分布式命令仅用于本地索引，其效果不如在虚拟索引上好。

您可以在同时使用可分布式和不可分布式命令的不同索引类型上新建各种搜索，但您需要记住，这样的搜索会返回本地索引上的所有数据，但仅返回虚拟索引上的有限数据。

本主题讨论最适合和 Splunk Analytics for Hadoop 结合使用的命令类型，以及应保留和 Splunk Enterprise 本地目录结合使用的命令。

智能模式搜索

搜索模式可控制搜索返回的数据量或数据类型。

智能模式是 VIX 搜索的默认模式，也是推荐设置。此模式会根据搜索是否包含转换命令决定是否保留搜索行为。搜索虚拟索引时，我们建议您在智能模式中搜索，因为这样更高效。

如果您使用详细模式搜索 VIX，请注意 Splunk Analytics for Hadoop 不会启动 MapReduce 任务进行搜索。这是因为详细模式搜索会搜索所有事件以及您可能正在运行的任何报表。在这种情况下，MapReduce 任务的益处最小，在某些情况下，可能会对搜索产生负面影响。

要了解有关 Splunk Enterprise 搜索模式的更多信息，请参阅*搜索手册*：

- 设置搜索模式以调整搜索体验

可分布式命令

可分布式命令是可在本地索引器上运行，但也可分发到搜索头和虚拟索引的命令。这些命令在 Enterprise 中的索引器和 DataNode/TaskTracker 上运行。

最适用于虚拟索引的命令有：

- **可分布式流命令：**这可以是作用于搜索返回的每个事件的任何流命令。可分布式流命令包括：
  - bin（如果用显式的 span 调用）
  - convert
  - Eval
  - extract (kv)
  - 字段
  - lookup（如果不是 local=t）
  - mvexpand
  - multikv
  - rename
  - regex
  - replace
  - rex
  - search
  - strcat
  - tags
  - typer
  - where
- **可分布式生成命令：**可分布式事件生成命令可返回事件列表或结果表。通常在搜索开始时通过前导管道符调用生成命令。不可存在通过管道符传递给生成命令的搜索。（例外情况是搜索命令，因为搜索命令在搜索开始时为隐式，无需调用。）可分布式事件生成命令包括：
  - search
  - metadata

不可分布式命令

不可分布式命令（也称为**非流命令**）要求将所有数据返回到本地索引器。这些不是特别有效的可搜索虚拟索引的命令。

当您的部分搜索涉及某个容量中的本地索引时，最好保留非流命令。在使用非流命令的本地和虚拟索引上运行的搜索将应用于

本地索引，但不会应用于搜索中包含的虚拟索引。

不可分布式或非流命令类型包括：

- **集中式流命令：**这些命令有时被称为“状态流”命令，包括：
  - head
  - streamstats
  - Dedup 的一些模式
  - 群集的一些模式

- **转换流命令：**转换命令在 Splunk 可用于统计的值中对事件进行排序，包括：
  - chart
  - timechart
  - stats
  - top
  - rare
  - contingency
  - highlight
  - typer
  - addtotals（用于计算列总数）
- **不可分布式生成命令：**集中式事件生成命令或报表生成命令不适用于虚拟索引。您不可从包含报表命令的任何搜索中导出数据。
  - 集中式事件生成命令包括：
    - loadjob
    - inputcsv
    - inputlookup
  - 报表生成命令包括：
    - dbinspect
    - datamodel
    - metadata
    - pivot
    - tstats

其他命令

还有一些命令不属于这些类别。这些命令包括非报表、不可分布式和非流命令：sort、eventstats、dedup 的一些模式和群集的一些模式。

使用虚拟索引时要避免的标头提取

虚拟索引目前不支持索引时间字段的配置。因此，索引时间字段提取特有的属性不适用于虚拟索引。具体包括下列属性：

- INDEXED_EXTRACTIONS
- HEADER_FIELD_LINE_NUMBER
- PREAMBLE_REGEX
- FIELD_HEADER_REGEX
- FIELD_DELIMITER
- FIELD_QUOTE
- HEADER_FIELD_DELIMITER
- HEADER_FIELD_QUOTE
- TIMESTAMP_FIELDS = field1,field2,...,fieldn
- FIELD_NAMES
- MISSING_VALUE_REGEX

搜索虚拟索引

正确安装并配置虚拟索引后，您就可以新建报表，并如同在传统的 Splunk 索引中那样对数据进行可视化处理。使用虚拟索引和传统的 Splunk Enterprise 之后，您可以只从虚拟索引中收集数据，或者也可以同时查询本地索引和虚拟索引并制作单份报表。

大多数情况下，您既可以新建虚拟索引报表也可以新建本地索引报表。更多有关新建报表的信息，请参阅*《Splunk Enterprise 搜索手册》*。

出于 Hadoop 数据存储的大小和性质，有一些特定的 Splunk Enterprise 索引行为无法复制：

- Splunk Analytics for Hadoop 目前不支持实时搜索 Hadoop 数据，尽管预览功能可用。
- 数据返回速度不会始终如返回本地索引数据那样快。

由于事件并未排序，任何基于隐式时间顺序的搜索命令都无法达到您预期的效果。（例如：头、增量或交易。）这意味着有些搜索命令在用于虚拟索引时会以不同的方式运行，主要取决于 Hadoop 报告时间戳的方式。

您仍可以使用这些命令，尤其是需要为本地和虚拟索引新建单份报表时，但请注意这些索引如何以不同的方式运行并以不同的方式返回数据。

Splunk Analytics for Hadoop 如何使用搜索语言

大多数情况下，您可以使用 Splunk Enterprise 的搜索语言新建报表。但是，由于 Splunk Analytics for Hadoop 不支持事件顺序的严格要求，因此会产生一些差异。

当搜索包含虚拟索引时，将不支持下列命令：

- transactions
- localize

下列命令可以在虚拟索引上使用，但结果可能会不同于 Splunk。这是因为 Splunk Analytics for Hadoop 不保证事件按时间降序排列：

- streamstats
- head
- delta
- tail
- reverse
- eventstats
- dedup （由于命令无法在 HDFS 目录内区分挑选移除项目的顺序，Splunk Analytics for Hadoop 将根据修改时间或文件顺序挑选要移除的项目。）

加速报表

报表加速允许您使用提前新建的缓存数据提高搜索速度。**报表加速**用于加速单个报表，可以很容易地针对在大型数据集上运行的任何**转换搜索**或报表设置加速。

- 要启用报表加速，您可勾选复选框并选择时间范围。此后的所有操作都在后台进行。随后加速报表的运行将加快完成。
- Splunk Analytics for Hadoop 会自动共享包含类似搜索的报表加速摘要，这样这些搜索可从已收集的数据中受益。
- 报表加速提供自动回填：如果出于某些原因遇到数据中断，Splunk Analytics for Hadoop 会自动更新或重构摘要。
- 虚拟索引的报表加速摘要存储在 Splunk Analytics for Hadoop 中。

注意：此时，**验证**按钮不可用于 Splunk Analytics for Hadoop 的报表加速。

请牢记一点：报表加速不适用于所有搜索类型。

- 只有使用**转换命令**的搜索（此类搜索可将其结果转换为统计表格和图表）才符合条件。在转换命令之前搜索中使用的任何命令都必须是**流命令**。有关转换搜索的更多信息，请参阅“关于报表命令”。
- 加速报表在详细模式中无效。

要在保存报表时启用报表加速

1. 在**报表**页面上，展开某报表所对应的行并单击**编辑**打开**编辑加速对话框**。

2. 选择**编辑加速**对话框中的“加速报表”。选择计划针对哪个时间范围运行此报表，然后单击**保存**。

3. 在**设置** > **搜索和报表**中打开报表的详细信息页面。

4. 单击**加速此搜索**，然后设置**摘要范围**。

了解有关报表加速的更多信息

请参阅 Splunk Enterprise 文档中的以下主题：

- 《Splunk Enterprise 报告手册》中的“加速报表”。
- 《Splunk 知识管理器手册》中的“管理报表加速”。

维护报表加速

“报表加速”会新建和保存 Splunk Analytics for Hadoop 随后会利用的搜索摘要，以加速完成某些报表。这表示您的缓存会变得很大。要了解查找和配置缓存的位置，以及如何高效管理缓存维护和清理，请参阅“配置报表加速缓存”。

管理报表加速

您可以将报表加速和任何使用 Hadoop ERP 的搜索结合使用。有关使用报表加速的更多信息，请参阅本手册中的“使用报表加速”。

产生的报表加速摘要存在于 Splunk Analytics for Hadoop 中，成为您可管理和维护的文件。

查找报表加速文件

Hadoop ERP 报表加速文件存储在已配置的 Hadoop 文件系统中。文件默认存储在名为 `cache` 的子目录中的 `vix.splunk.home.hdfs` 的配置路径下。

您可使用 `vix.splunk.search.cache.path` 参数更改拥有绝对路径值的虚拟索引提供程序的位置。将名为缓存的子目录自动添加到此配置路径下。

以下是最后以缓存路径配置结束和不以此结束的文件的两个示例。

```
# Example 1, no cache path configuration:
# Index.conf, virtual index provider:
vix.splunk.home.hdfs = /user/sarah/hadoopanalytics_files

# Resulting report acceleration file location
/user/sarah/hadoopanalytics_files/cache

# Example 2, using cache path configuration:
# Index.conf, virtual index provider:
vix.splunk.home.hdfs = /user/sarah/hadoopanalytics_files
vix.splunk.search.cache.path = /var/everyone/hadoopanalytics

# Resulting report acceleration file location
/var/everyone/hadoopanalytics/cache
```

报表加速文件结构中，有一个文件包含一些关于缓存的信息。此文件可用于调试和维护。该文件目前存在于文件系统中：

```
cache/<index>/<search_hash>/info.json
```

其中 `index` 是新建此摘要的索引，`search_hash` 是摘要 ID 的哈希、。

管理文件清除

Splunk Analytics for Hadoop 会定期清除已过期的搜索摘要数据的报表加速文件。例如，如果您在可追溯一年内的搜索上使用报表加速，当一年以前的数据过期时，会删除包含该数据的摘要数据。

大多数情况下，这是充分的存储维护。但是，摘要数据可能不会在数据到期时完全删除，因为摘要会保留到相同文件中的*所有*摘要都过期为止。如果您发现 Splunk Analytics for Hadoop 没有以最优方式维护存储，您可微调配置，使其更有效。

您可将数据拆分到和虚拟索引路径部分匹配的数据桶（例如，基于日期）中，以使 Splunk Analytics for Hadoop 更高效地进行清除。按照数据桶对摘要进行分组之后，Splunk Analytics for Hadoop 只会删除维护范围以外的数据桶。

以下是按年份、月份和日期结构化的数据的案例示例：

```
/path/to/data/20131230/...
/path/to/data/20131231/...
/path/to/data/20140101/...
/path/to/data/20140102/...
...
/path/to/data/20140209/...
...
```

此数据的虚拟索引路径将按如下方式配置：

```
vix.input.1.path = /path/to/data/...
vix.input.1.et.regex = /path/to/data/(\d+)
vix.input.1.et.format = yyyyMMdd
...
```

要按年份和月份拆分摘要，您可以这样指定数据桶：

```
vix.input.1.bucket.regex = /path/to/data/(\d{6}).*
```

或包含多个组：

```
vix.input.1.bucket.regex = /path/to/data/(\d{4})(\d{2}).*
```

vix.input.N.bucket.regex 使用正则表达式中的组确定路径属于哪个数据桶。我们第一个示例中的正则表达式捕获了使用日期作为名称的目录的前 6 个数字。使用此数据桶正则表达式时，会按年份和月份拆分摘要。第二个示例会使用多个组从路径中获取数据桶。这稍有不同，因为数据桶将分配给连接正则表达式组的值，并使用短划线作为分隔符。

以下是一些说明路径下的文件如何获得指定正则表达式的数据桶值的示例：

```
# Regex:
vix.input.1.bucket.regex = /path/to/data/(\d{6}).*

# Paths - Assigned bucket value:
/path/to/data/20131230/foo.txt - 201312
/path/to/data/123456789/bar.csv - 123456
```

以下是使用多个组的示例：

```
# Regex
vix.input.1.bucket.regex = /path/to/data/(\d{4})(\d{2}).*

# Paths - Assigned bucket value:
/path/to/data/20131230/foo.txt - 2013-12
/path/to/data/123456789/bar.csv - 1234-56
```

如何存储数据取决于您想要在摘要中保留多少数据。分组过于精细，会造成摘要文件的文件夹过多。而分组过于宽泛会导致以数据桶摘要为开头的目的无法实现。

如果您不确定从哪里开始，我们建议您尝试以年份和月份开始分组，然后微调以查看最适合的细分方式。

关于数据模型加速

数据模型加速允许您为包含虚拟索引的数据新建数据模型。Splunk Analytics for Hadoop 数据模型加速使用缓存信息，可映射极大的数据集以加速搜索。

存储在 Splunk Enterprise 索引中的数据模型信息使用 tsidx 文件。Splunk Analytics for Hadoop 数据模型可访问将指向 Hadoop 中数据的虚拟索引的数据，这样您可在虚拟索引可指向的任何文件类型上新建数据模型。Splunk Analytics for Hadoop 将数据模型加速文件存储在 Parquet 和 ORC 中。

有关数据模型加速如何在 Splunk Enterprise 中工作的更多信息，请参阅 Splunk Enterprise 文档中的《加速数据手册》。

数据模型加速如何在 Splunk Analytics for Hadoop 中工作

1. 您可新建数据模型“配置数据模型加速”

2. Splunk Analytics for Hadoop 可为每个原始数据文件新建数据模型加速摘要文件：

   - Splunk Analytics for Hadoop 会将数据模型加速摘要文件相关信息保留在 KV 存储中（这样可快速查找）。
   - Splunk Analytics for Hadoop 会将实际的数据模型加速摘要文件存储在 Hadoop 中。

3. 如果数据模型覆盖的时间范围较大，如“一年”或“所有时间”。您可用在 Splunk Analytics for Hadoop 报表加速中使用的相同方法将数据模型文件分区到数据桶中。这使得在模型新建时更新查找文件和在搜索时间加载查找文件都更加快速。如需更多信息，请参阅“加速报表”。

4. 您可在 Splunk Analytics for Hadoop 虚拟索引中的数据上运行搜索，该索引还包括 Splunk Enterprise 索引。如果所有 Splunk Enterprise 数据都存在数据模型，那么按照《Splunk 知识管理器手册》中“加速数据模型”中介绍的方式应用数据模型加速。

5. 有关基于虚拟索引新建的数据模型上的 tstats/数据透视表搜索，Splunk Analytics for Hadoop 会使用 KV 存储验证原始数据拆分是否存在加速摘要文件。

6. 如果未发现原始数据拆分的加速摘要文件，Splunk Analytics for Hadoop 无法返回原始数据文件，并对虚拟索引进行常规搜索。

7. 如果发现加速摘要文件，那么使用摘要文件而不是原始数据文件。

8. 使用现有的已保存摘要数据可更快地返回由所有或部分数据模型组成的搜索数据。

配置数据模型加速

默认只有具有 Hadoop 群集上数据访问权限的用户才能新建数据模型。

新建数据模型

1. 导航到设置 > 数据模型。

2. 单击“管理数据模型”按钮。

3. 单击**新的数据模型**。

4. 在**新建数据模型**对话框中，输入数据模型**标题**和可选**描述**。**标题**字段可接受除星号以外的任何字符，字符间可使用空格。显示数据模型名称时显示此标题。

5. 当您输入标题时，Splunk Analytics for Hadoop 可用唯一的 ID 填充数据模型 **ID** 字段。您无需编辑此 ID。如出于任何原因考虑，您发现必须编辑此字段，请注意以下事项：

- 此字段必须是唯一的标识符。
- 其中只能包含字母、数字和下划线。
- 字符间不能含有空格

一旦单击**新建**，即无法更改 **ID** 值。

6. **应用**将显示您当前所处的应用上下文。

7. 单击**新建**在数据模型编辑器中打开新的数据模型。

8. 添加和定义您想要包含在搜索中的对象。要定义数据模型的首个对象，单击**添加对象**并选择对象类型。有关对象定义的更多信息，请参阅《Splunk Enterprise 知识管理器手册》中的“设计数据模型”。

加速数据模型

1. 打开数据模型的“数据模型编辑器”，单击“编辑”，然后选择“编辑加速”。

2. 选择“加速”。注意：新建加速模型时，Hadoop 节点使用量会增加。

3. 为加速数据模型搜索选择“摘要范围”。

4. 启用特定选项：勾选此框允许您编辑文件信息。仅当您想要更改默认值时勾选此框。Splunk Analytics for Hadoop 会根据在数据模型中查找到的信息填充以下字段，这样可能不需要编辑字段。

- 文件格式：选择 Parquet 或 Orc。
- 压缩码：关于 Parquet 文件格式，请选择 Snappy 或 Gzip。关于 Orc，选择 Snappy 或 zlib。
- DFS 块大小：查看“启用块大小”规范，然后确定大小。注意：DFS 块大小不得小于 32MB。Orc 和 Parquet 必须缓冲内存中的记录数据，直到写入这些记录。内存消耗应和搜索中行组的所有列大小相关。换句话说，搜索中所需字段越少，需要缓冲的内存越少。

配置和运行统一的搜索

Splunk Analytics for Hadoop 归档允许您搜索虚拟归档索引中的归档数据和提供这些归档的 Splunk Enterprise 索引中的实时数据。根据您配置归档的方式，您的归档数据可和索引中未归档的数据重叠。

例如，我们建议您在将数据设置为从 Splunk Enterprise 索引中消除之前，先将 Splunk Enterprise 索引设置为归档数据，这样就能避免数据无法临时用于搜索的风险。这可能会造成某些数据重叠。

您可为同样配置用于归档的任何虚拟索引配置统一搜索。之后，每当你针对 Splunk Enterprise 索引运行搜索时，统一的搜索都会自动同时检查 Splunk Enterprise 索引和归档中的数据，同时跳过重复的数据。

如何用统一搜索进行搜索

统一搜索**只**适用于搜索中**显式**指定的索引。统一搜索**不**会搜索隐式指定的索引归档，例如，通过默认索引或通过通配符指定的索引。统一搜索不会按事件新建日期分类，这包括只来自数据尚未归档的真实 Splunk 索引器的结果。Splunk Analytics for Hadoop 不支持用统一搜索进行实时搜索。

有关 Splunk Analytics for Hadoop 如何处理搜索和时间/日期的更多信息，请参阅“搜索虚拟索引”。

以下为一些统一搜索可提高搜索的显式搜索的示例：

- `index=myindex someterm`
- `index=myindexname OR index=foo | top limit=20 "result.category_id"`

以下是非显式搜索的一些示例，这些搜索**不**会使统一搜索搜索归档内容：

- 通配符
- `someterm`
- `index=m* someterm`
- `index!=my_splunk_index_with_an_archive`
- `NOT index=my_splunk_index_with_an_archive`

配置统一搜索

要点：要使用统一搜索，索引必须在搜索头和索引器上进行定义。如果索引并未在搜索头中定义，则 Splunk 会新建空索引

通过将 `unified_search` 设为 `true` 启用 `limits.conf` 中的统一搜索：

```
[search]
# turn on/off feature
unified_search = true
```

`In indexes.conf` 将以下属性添加到索引归档段落：

```
[myindex_archive]
vix.unified.search.cutoff_sec = = <window length, before present time, in seconds>
```

`myindex` 相关查询会自动在归档索引（即 `myindex_archive`）中查找比截止点更早的事件，并会查找 `myindex` 中更新的事件。我们建议设置统一搜索截止点，使其正好在配置 Splunk 索引将数据桶从冷状态移动到冻结状态之前。

请参阅“归档 Splunk 索引”了解归档配置更多信息。

以下是配置用于统一搜索的虚拟索引的示例：

```
[root@sandbox bin]# more $SPLUNK_HOME/etc/apps/search/local/limits.conf
[search]
unified_search = true

[root@sandbox bin]# more $SPLUNK_HOME/etc/apps/search/local/indexes.conf
..
[myindex_archive]
vix.output.buckets.from.indexes = myindex
vix.output.buckets.older.than = 3600
vix.output.buckets.path = /user/root/archive/myindex_archive
vix.provider = hdp2provider
vix.unified.search.cutoff_sec = 14400
# 14400 is 4 hours
```

引用

故障排除 Splunk Analytics for Hadoop

本主题介绍了您可能遇到的各种配置组件问题，以及解决这些问题的方法。

有关故障排除问题解答以及自行提问的更多信息，请搜索 Splunk Answers。

群集问题

问题：首次启动群集时，NFS 网关无法启动

查看日志查看是否是许可证问题。在您能够应用许可证而以失败告终之前，可以让 NFS 网关尝试启动。在这种情况下，安装许可证后，重新启动群集。

问题：服务无法在节点上启动

这可能属于网络问题，尝试禁用 IPTables。

ZooKeeper

问题：ZooKeeper 处于错误状态

尝试使用以下步骤重置 ZooKeeper：

1. 关闭服务

2. 新建临时备份：

```
mkdir /tmp/zkdata_backup</br>mv $MAPR_HOME/zkdata/version-2/* /tmp/zkdata_backup
```

3. 重新启动 MapR ZooKeeper 服务

搜索问题

问题：搜索运行非常缓慢

例如，搜索所花费的时间比 Hadoop 任务通常所需的时间长。比如，Hive 任务需要 6 分钟时间完成，但是 Splunk Analytics for Hadoop 完成类似任务可能需要 30 分钟时间。

要解决这个问题，请确保 Splunk Analytics for Hadoop 正在运行实际的 MapReduce 任务，而不是简单地传回 Hadoop 报告：

- Splunk 输出 Hadoop 流报告（而不是 MapReduce 任务）

```
Index=xyz
```

- Splunk 输出 Hadoop 流报告（而不是 MapReduce 任务）

```
Index=xyz | stats count and using Verbose Mode
```

- MapReduce 任务利用 Hadoop 报表

```
Index=xyz | stats count and using Smart Mode
```

问题：报表搜索出现一条错误消息

如果报表搜索出现以下错误：

```
INFO mapred.JobClient: Cleaning up the staging area hdfs://qa-centos-amd64-
26.sv.splunk.com:8020/user/apatil/.staging/job_201303061716_0033
ERROR security.UserGroupInformation: PriviledgedActionException as:apatil cause:org.apache.hadoop.ipc.RemoteException:
java.io.IOException:
job_201303061716_0033(-1 memForMapTasks -1 memForReduceTasks): Invalid job requirements.
at org.apache.hadoop.mapred.JobTracker.checkMemoryRequirements(JobTracker.java:5019)
```

尝试将以下参数添加到 indexes.conf：

```
vix.mapred.job.map.memory.mb = 2048
vix.mapred.job.reduce.memory.mb = 256
```

问题：Splunk 出现故障消息

例如：

```
[APACHE] External result provider name=APACHE asked to finalize the search
[APACHE] MapReduce job id=job_201303081521_0020 failed, state=FAILED, message=# of failed Map Tasks exceeded allowed limit.
FailedCount: 1.
LastFailedTask: task_201303081521_0020_m_000000
```

显示这类错误是因为 Java 子进程仍在运行。查看 MapReduce 日志的以下内容：

```
TaskTree [pid=7535,tipID=attempt_201303061716_0093_m_000000_0] is running beyond memory-limits.
Current usage : 2467721216bytes. Limit : 2147483648bytes. Killing task.
```

要解决这个问题，请编辑 indexes.conf，如下所示：

```
vix.mapred.child.java.opts = -server -Xmx1024m
```

调试

问题：您需要调试搜索

例如，您可运行搜索，但不接收 Hadoop 结果。

这表示 Splunk Ananlytics for Hadoop 配置不正确。要解决这个问题，启用调试找出配置错误，然后打开任务查看器：

1. 选择菜单中的*提供程序*，然后单击编辑提供程序。

2. 通过将 vix.splunk.search.debug = 值更改为以下值启用调试 1

3. 重新运行搜索。

4. 打开任务查看器并单击 search.log 文件链接。

5. 在 search.log 文件中搜索单词 DEBUG 查找错误。

验证

问题：结合使用 Splunk Analytics for Hadoop 和 Kerberos 时出现验证错误

例如，您运行 MapReduce 任务或尝试访问 Hadoop 数据并获取 Kerberos 特例。

错误示例：

- java.lang.IllegalArgumentException: Server has invalid Kerberos principal
- 在 Server.log 中，您可看到：

```
SplunkMR - Failed to start MapReduce job.  Please consult search.log for more information. Message: [ Failed to start MapReduce
job, name=SPLK_<hunk_server>_1435283732.28_0 ] and [ Failed on local exception: java.io.IOException:
java.lang.IllegalArgumentException: Server has invalid Kerberos principal: rm/<master_server>@<REALM>; Host Details : local
host is: "<hunk_server>/192.X.X.X"; destination host is: "<master_server>":8050;  ]
```

要解决这个问题：

1. 确保已在 Splunk 节点上安装 Kerberos：

2. 配置用户权限“以根用户身份”，将“Splunk 用户”添加到 Kerberos 数据库中。键入以下内容启动 kadmin 服务：kadmin.local " 键入以下命令新建 Splunk 用户主体： " kadmin.local: addprinc -randkey splunk@EXAMPLE.COM "

3. 使用以下命令在 /etc/security/keytabs 目录中新建 keytab 文件：

```
" kadmin.local: xst -norandkey -k /etc/security/keytabs/splunk.headless.keytab splunk@EXAMPLE.COM
kadmin.local: exit "
```

4. 为 Splunk 用户设置 keytab 文件权限：

```
# chown splunk:hadoop /etc/security/keytabs/splunk.headless.keytab
# chmod 440 /etc/security/keytabs/splunk.headless.keytab
```

5. 以 Splunk 用户身份，初始化 keytab 文件：

```
# su - splunk
$ kinit -kt /etc/security/keytabs/splunk.headless.keytab splunk@EXAMPLE.COM
```

6. 使用 Splunk 文档配置“Kerberos 配置 Kerberos 验证”

7. 如果这样没用，请尝试使用 DEBUG 命令。

Hadoop 特定问题

问题：Hadoop 产生的结果太多

示例 A：您可检查 HDFS 并会发现月份和日期有时是单数，有时是双数：

```
/some/path/customer2/year=2016/month=3/day=23/somefiles
```

或

```
/some/path/customer2/year=2016/month=10/day=4/somefiles
```

这可能是由于 HDFS 时间分区引起的。Splunk Analytics for Hadoop 文档介绍如何通过构建已知位数的正则表达式捕获时间。

大多数情况下，您可只使用正则表达式，如文档中所示：添加虚拟索引。

您可通过不止捕获正则表达式的方法解决这个问题，并使用格式中的一位数。例如：

```
vix.input.1.accept = \.avro$
vix.input.1.et.format = y/M/d
vix.input.1.et.regex = .*?/customer2/Year=(\d+/)Month=(\d+/)Day=(\d+)/.*
vix.input.1.lt.format = y/M/d
vix.input.1.lt.offset = 86400
vix.input.1.lt.regex = .*?/customer2/Year=(\d+/)Month=(\d+/)Day=(\d+)/.*
vix.input.1.path = /some/path/customer2/...
vix.provider = hdp23provider
```

示例 B：您会发现 HDFS 时间捕获正则表达式和时间格式的多个 Epoch 的情况，其中包括将时间戳标记为 Epoch 时间，LT 和 ET 都在以下路径中：

```
/user/root/myarchive/db_1359855960_1357027260_0/journal.gz
```

在 `inputs.conf` 中检查是否在 HDFS 时间分区中设置了两个 epoch 时间戳。 `indexes.conf` 可捕获 epoch 时间戳的第一和第二部分：

```
vix.input.1.et.format = epoch
vix.input.1.et.regex = /user/root/myarchive/db_\d+_(\d+)_.*
vix.input.1.lt.format = epoch
vix.input.1.lt.regex = /user/root/myarchive/db_(\d+)_\d+_.*
vix.input.1.path = /user/root/myarchive/...
vix.provider = 62hdprovider
```

问题：Hadoop 启动失败

确保用户帐户有所需 Hadoop 目录的适当权限。

问题：Hadoop 服务器任务失败

以下一些示例是在 Hadoop 服务器端崩溃，而不是在 Splunk Analytics for Hadoop 端崩溃的任务：

- MapReduce 错误： `Error while waiting for MapReduce job to complete, job_id=XYZ`
- 运行搜索时出现 Splunk Web 错误。
- 等待 MapReduce 任务完成时出现错误， `job_id=`

要查找 Hadoop 服务器端的错误：

1. 单击链接查看任务检查器。

2. 在任务查看器中，单击 Hadoop 服务器日志链接。

3. 查找任何 Hadoop 服务器日志错误。例如，如果任务地图任务超时：

```
task_1465506338167_0007_m_000000 [MAP]
AttemptID:attempt_1465506338167_0007_m_000000_0 Info:Error: java.io.IOException: Hunk timed out while waiting for
package=/notvaliddirectory/splunk-6.4.1-debde650d26e-linux-2.6-x86_64.tgz to be installed.
```

问题：长时间运行任务期间 Hadoop 任务崩溃，或者会在 Hadoop 服务器上看到很多资源相关错误。

如果 Hadoop 资源有限，可以多运行几个任务，但减少每个任务处理的文件数量可能会有帮助。

默认情况下，第一个任务可处理 100 个块（Hadoop 文件），第二个任务可处理 1,000 个块，其他剩下的任务可处理 10,000 个块。

将 `vix.splunk.search.mr.maxsplits` 更改为 5000 会使 Splunk Analytics for Hadoop 将产生的任务数量增加一倍，但每个任务获取的资源会更少。

问题：Hadoop 任务运行速度不够快或 Splunk Analytics for Hadoop 正在处理的文件太多

使用任务查看器查看搜索过程中每个阶段的持续时间、组件、调用、输入计数和输出计数。

检查以下关键组件以查找性能问题。

整体性能说明。

字段	注释	示例
	“此搜索已完成，并通过在 33.045 秒内扫描 5,438 个事件返回了 5 个结果”。	
Command.stdin.	基于时间范围考虑用于分析的事件数量	持续时间：23.15 输出计数（事件计数）：3,124
command.stdin.cpd2sr	事件总数。理论上，此数量应和 Command.stdin. 相同如果不同，则可能表示时间捕获正则表达式出现问题	持续时间：2.20 输出计数（事件计数）：5,438
Erp.<provider>.cache.bytes	HDFS 缓存返回的字节总数	持续时间：0.02 输入计数（拆分长度）：4,696 输出计数（流长度，以字节为单位 - 重要值）：19,973
Erp.<provider>.report.bytes	HDFS MapReduce 任务返回的字节总数	持续时间：0.01 输入计数（拆分长度）：1,538 输出计数（流长度，以字节为单位 - 重要值）：4,111
Erp.<provider>.stream.bytes	Splunk Analytics for Hadoop 流任务返回的字节总数。一般来说，前几个事件是基于 stream.bytes，其他事件是基于 report.bytes	持续时间：26.95 输入计数（拆分长度）：59,367,278 输出计数（流长度，以字节为单位 - 重要值）：671,088,641
Erp.<provider>.vix.<vix>.dirs.listed	Splunk Anaytics for Hadoop 必须扫描的 Hadoop 目录总数	调用：365（目录数）

Erp.<provider>.vix.<vix>.files.listed	Splunk Analytics for Hadoop 必须扫描的 Hadoop 文件总数	调用：8,760（文件数）
Erp.<provider>.vix.<vix>.dir.filter.time	出于时间范围考虑的移除的 Hadoop 文件总数。实际使用的是 dirs.listed - dir.filter.time 的文件总数 调用：211（目录数）	
Erp.<provider>.MR.SPLK_<host>_<SID>	Hadoop 中产生的 MapReduce 任务以及运行花费的时间。	持续时间：43.94
Erp.<provider>.vix.<vix>.splits.generation.time	计算拆分花费的时间 请参阅提供程序标记 maxsplit 和 minsplit	持续时间：0.18

问题：无法将文件保存为 Hive 文件格式

如果您在尝试以 Hive 文件格式保存文件时遇到错误，则 Hadoop 实例和 Hive 实例之间可能存在版本不匹配的情况。Hadoop 2.x 或更低版本需要搭配版本为 2.x 或更低的 Hive。Hadoop 3.x 要求的 Hive 版本至少为 3.x。如果使用 2.x 或更低版本的 Hive 实例配置版本 3.x 的 Hadoop 群集，则当您尝试以 Hive 文件格式保存文件时会遇到连接问题。

更多信息，请参阅“配置 Hive 连接”。

问题：Hadoop 数据库空间不足

Splunk Analytics for Hadoop 无法自动删除已存档到 HDFS 的文件。需要由 Hadoop 管理员来清理旧文件或过时的文件。Hadoop 管理员可以创建采用 HDFS 结构（由 Splunk 创建）的文件清理脚本。

关于清理文件的更多信息，请参阅“将冷数据桶归档为冻结数据桶”。

性能最佳实践

当搜索进程搜索原始 HDFS 数据时，数据会通过索引时间处理传递。（索引时间提取可在搜索时间运行，但无法关闭。）

要更高效地处理此数据，您应优化索引时间设置，尤其是时间戳和聚合设置。可配置添加到 props.conf 中数据来源的以下设置以提高性能：

- DATETIME_CONFIG
- MAX_TIMESTAMP_LOOKAHEAD
- TIME_PREFIX
- TIME_FORMAT
- SHOULD_LINEMERGE
- ANNOTATE_PUNCT

例如，对于单行非时间戳数据，以下设置可将吞吐量提高大约四倍：

```
[source::MyDataSource]
ANNOTATE_PUNCT   = false
SHOULD_LINEMERGE = false
DATETIME_CONFIG  = NONE
```

注意：如果您需要使用时间戳，我们强烈建议您使用 TIME_PREFIX 和 TIME_FORMAT 来提高处理效率。

下表显示可能的时间戳和换行选项示例，以及处理包含 1000 万个单行事件的文件时结合可能需要花费的时间（以秒为单位）：

时间戳和换行选项：	时间：
默认配置	190 秒
MAX_TIMESTAMP_LOOKAHEAD = 30	179
MAX_TIMESTAMP_LOOKAHEAD = 30	105

SHOULD_LINEMERGE = false	
MAX_TIMESTAMP_LOOKAHEAD = 30 SHOULD_LINEMERGE = false TIME_PREFIX = ^	107
MAX_TIMESTAMP_LOOKAHEAD = 30 SHOULD_LINEMERGE = false TIME_FORMAT = %a, %d %b %Y %H:%M:%S %Z	51
MAX_TIMESTAMP_LOOKAHEAD = 30 SHOULD_LINEMERGE = false TIME_PREFIX = ^ TIME_FORMAT = %a, %d %b %Y %H:%M:%S %Z	53
MAX_TIMESTAMP_LOOKAHEAD = 30 SHOULD_LINEMERGE = false TIME_FORMAT = %a, %d %b %Y %H:%M:%S %Z ANNOTATE_PUNCT = false	44
SHOULD_LINEMERGE = false	109
SHOULD_LINEMERGE = false TIME_PREFIX = ^	99
SHOULD_LINEMERGE = false TIME_FORMAT = %a, %d %b %Y %H:%M:%S %Z	54
SHOULD_LINEMERGE = false TIME_PREFIX = ^ TIME_FORMAT = %a, %d %b %Y %H:%M:%S %Z	54
MAX_TIMESTAMP_LOOKAHEAD = 30 SHOULD_LINEMERGE = false DATETIME_CONFIG = NONE	49
SHOULD_LINEMERGE = false DATETIME_CONFIG = CURRENT	50
MAX_TIMESTAMP_LOOKAHEAD = 30 SHOULD_LINEMERGE = false DATETIME_CONFIG = NONE ANNOTATE_PUNCT = false	35

禁用流命令以提高搜索速度

如果您只想要来自 MapReduce 任务的数据，而不进行预览，您可禁用 Splunk Analytics for Hadoop 的流功能来提高搜索速度。

默认情况下，Splunk Analytics for Hadoop 使用混合模式，该模式结合流（仅限 Splunk）和报表（Hadoop MR 任务）模式。如果您不需要预览，您可禁用 Splunk Analytics for Hadoop 的流部分。

要启用或禁用流功能：

- 混合模式： `vix.mode = report and vix.splunk.search.mixedmode = 1`
- 仅限于报表模式： `vix.mode = report and vix.splunk.search.mixedmode = 0`
- 仅限于流模式： `vix.mode = stream`

提供程序配置变量

当您配置 HDFS 提供程序时，Splunk Analytics for Hadoop 会自动设置一些配置变量。您可使用预先设置的变量，或者您可在需要时通过编辑提供程序修改变量。

- 有关在配置文件中编辑变量的更多信息，请参阅“在配置文件中设置提供程序和虚拟索引”。
- 有关在 Splunk Web 界面中编辑提供程序的信息，请参阅“添加 HDFS 提供程序”。
- 有关设置 YARN 提供程序配置变量的信息，请参阅“YARN 必需的配置变量”。

设置：	用途：
vix.splunk.setup.onsearch	确定是否在搜索上执行设置（安装和 BR）。
vix.splunk.setup.package	Splunk 可安装和在数据节点（位于 vix.splunk.home.datanode 中）上使用的 Splunk .tgz 软件包的位置。current 值可使用当前安装。
vix.splunk.home.datanode	SPLUNK_HOME 在 DataNode 和/或 TaskTracker 上
vix.splunk.home.hdfs	此 Splunk 实例的 HDFS 暂存空间位置。
vix.splunk.search.debug	确定搜索是否在调试模式中运行。
vix.splunk.search.recordreader	提供逗号隔开的数据预处理类别列表。此值必须扩展 BaseSplunkRecordReader 并返回 Splunk 消耗的数据作为值
vix.splunk.search.recordreader.avro.regex	指定文件必须匹配的正则表达式作为 avro 文件，默认为 \.avro$。
vix.splunk.search.mr.threads	确定读取 HDFS 映射结果时要使用的线程数量。
vix.splunk.search.mr.maxsplits	确定 MapReduce 任务中的最大拆分量。
vix.splunk.search.mr.poll	确定任务状态的轮询周期，以毫秒为单位。
vix.splunk.search.mixedmode	确定是否启用混合模式执行
vix.splunk.search.mixedmode.maxstream	确定混合模式期间流出的最大字节数。默认值为 10GB。值为 0 表示没有流出限制。字节将在第一次拆分获取值超出限制之后停止流出。
vix.splunk.jars	提供 SH 和 MR 中要使用的 dirs/jars 列表，用逗号隔开

高可用性配置

HA-NN

设置：	用途：
vix.fs.default.name	默认文件系统名称 uri> # hdfs://sveserv51-ha
vix.dfs.nameservices	用逗号隔开的 nameservices> # sveserv51-ha 列表
vix.dfs.ha.namenodes.sveserv51-ha	用逗号隔开的指定 nameservice 的 namenode 列表，如 sveserv51-ha> # nn1,nn2
vix.dfs.namenode.rpc-address.sveserv51-ha.nn1	针对指定 namenode（如 nn1）或指定 nameservice（如 sveserv51-ha> # sveserv51-vm6.sv.splunk.com:8020）的 RPC 服务器地址和端口
vix.dfs.namenode.rpc-address.sveserv51-ha.nn2	针对指定 namenode（如 nn2）或指定 nameservice（如 sveserv51-ha> # sveserv51-vm5.sv.splunk.com:8020）的 RPC 服务器地址和端口
vix.dfs.client.failover.proxy.provider.sveserv51-ha	指定 nameservice 的 FailoverProxyProvider 实现，如 sveserv51-ha> # org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider

HA-JT

设置：	用途：
vix.mapred.job.tracker	jobtrackers> # sveser51-ha-jt 列表的逻辑名称
vix.mapred.jobtrackers.sveserv51-ha-jt	用逗号隔开的指定逻辑 jobtracker 名称的 jobtracker 列表，如 sveserv51-ha-jt> # jt1,jt2
vix.mapred.jobtracker.rpc-	针对指定 jobtracker（如 jt1）或指定逻辑 jobtracker 名称（如 sveserv51-ha-jt>

address.sveserv51-ha-jt.jt1	# sveserv51-vm6.sv.splunk.com:8021）的 RPC 服务器地址和端口
vix.mapred.jobtracker.rpc-address.sveserv51-ha-jt.jt2	针对指定 jobtracker（如 jt2）或指定逻辑 jobtracker 名称（如 sveserv51-ha-jt> # sveserv51-vm5.sv.splunk.com:8021）的 RPC 服务器地址和端口
vix.mapred.client.failover.proxy.provider.sveserv51-ha-jt	指定逻辑 jobtracker 名称的 FailoverProxyProvider 实现，例如 sveserv51-ha-jt> # org.apache.hadoop.mapred.ConfiguredFailoverProxyProvider
vix.mapred.map.max.attempts	映射任务可以停用的尝试次数，直到任务标记为失败> # 4
vix.mapred.max.map.failures.percent	整个 MR 任务可接受的任务失败百分比，直到 MR 任务失败> # 5

虚拟索引配置变量

当您配置虚拟索引时，Splunk Analytics for Hadoop 会自动设置一些配置变量。您可使用预先设置的变量，或者您可在需要时通过编辑索引修改变量。

- 有关在配置文件中编辑变量的更多信息，请参阅“在配置文件中设置提供程序和虚拟索引”。
- 有关在 Splunk Web 中编辑提供程序的信息，请参阅“在用户界面添加或编辑虚拟索引”。
- 有关设置 YARN 提供程序配置变量的信息，请参阅“YARN 必需的配置变量”。

设置：	用途：
vix.input.[N].path	在包含通配符和/或以 ... 结尾的 HDFS 中提供路径，该路径指定了此索引的数据。以 "..." 结尾的路径会以递归方式检查数据路径中的目录。
vix.input.[N].accept	确定文件/路径应匹配的正则表达式。
vix.input.[N].ignore	确定用于排除文件/路径的正则表达式。这些值优先于 vix.input.[N].accept 值。

根据路径（最早时间）提取最早/最晚时间范围变量：

设置：	用途：
vix.input.[N].et.regex	确定提取时间组件的正则表达式。将所有捕获组连接在一起，并使用下一行提供的格式解释。
vix.input.[N].et.format	解释用上述正则表达式构建的字符串时提供要使用的日期/时间格式。此值可设为 "epoch" 以将时间解释为秒。有关格式的更多信息，请参阅此处
vix.input.[N].et.value	Epoch 时间（毫秒）： • 设置此虚拟索引的最早时间。 • 可以进行使用，而不是通过 vix.input.x.et.regex 从路径中提取时间 • 当设为 "mtime" 时，使用文件修改时间作为最早时间。
vix.input.[N].et.offset	设置时间量（以秒为单位）以添加到结果时间。
vix.input.[N].et.timezone	确定用于解释提取时间的时区。如 "America/Los_Angeles" 或 "GMT-8:00" 等更多信息

根据路径（最晚时间）提取最早/最晚时间范围变量：

设置：	用途：
vix.input.[N].lt.regex	确定提取时间组件的正则表达式。将所有捕获组连接在一起，并使用以下格式解释。
vix.input.[N].lt.format	确定用于解释用上述正则表达式构建的字符串的日期/时间格式。此值可设为 "epoch" 以将时间解释为秒。有关格式的更多信息，请参阅此处
vix.input.[N].lt.value	Epoch 时间（毫秒）： • 设置此虚拟索引的最晚时间。 • 可以进行使用，而不是通过 vix.input.x.et.regex 从路径中提取时间 • 当设为 "mtime" 时，使用文件修改时间作为最晚时间。
vix.input.[N].lt.offset	设置时间量（以秒为单位）以添加到结果时间。
vix.input.[N].lt.timezone	要设置用于解释提取时间的时区。如 "America/Los_Angeles" 或 "GMT-8:00" 等更多信息

虚拟归档索引配置变量

```
Indexes.conf:
[splunkroll-virtual-index]
vix.provider = <provider>
vix.output.buckets.path = <hdfs_path>
vix.output.buckets.older.than = <time_in_seconds>
vix.output.buckets.from.indexes = <comma separated list with index names>
# Throttling
vix.output.buckets.max.network.bandwidth = <bit/sec> #where 0 equals no limit.

# Smart Search
to configure where events will be retrieved from, this must be set for smart search to kick in
vix.smart.search.cutoff_sec = <time period in seconds past now>, used to configure where events will be retrieved from, this must be set for smart search to kick in
vix.smart.search.fixed_cutoff = <fixed epoch time>, the cutoff_seconds a fixed time (overrides the above, and only to be used for testing)
```

YARN 必需的配置变量

如果您正在使用 YARN，您必须为配置变量设置添加资源管理器设置：

- vix.mapreduce.framework.name = yarn
- vix.yarn.resourcemanager.address= <namenode>:<port>
- vix.yarn.resourcemanager.scheduler.address= <namenode>:<port>

如果您为 Sandbox 2.0 安装了 Hortonwork，请添加以下设置/端口：

- vix.yarn.resourcemanager.address= <namenode>:8050
- vix.yarn.resourcemanager.scheduler.address= <namenode>:8030

如果您正在使用 Yarn 的 Cloudera VM，请添加以下设置/端口：

- vix.yarn.resourcemanager.address = <your namenode>:8032
- vix.yarn.resourcemanager.scheduler.address = <your namenode>:8030

如果您的群集没有使用默认的配置值，您必须为搜索头添加配置。（例如，如果 `yarn-site.xml` 中的属性值 `yarn.application.classpath` 和群集上的 `yarn-default.xml` 中的值不同。）

要更新搜索头，请执行以下其中一种设置：

- 在搜索头上设置 Hadoop CLI 的 `yarn-site.xml` 中的值
- 将提供程序中的值设为 `vix.yarn.application.classpath` 变量

使用高可用性的 YARN

对于使用 HA（高可用性）的 YARN，您必须添加以下变量：

1. HA - YARN [CDH5.3 YARN（包括 MR2）]

设置：	用途：
vix.yarn.resourcemanager.ha.rm-ids	资源管理器 ID 列表，用逗号隔开，在此示例中显示为： rm57,rm40
vix.yarn.resourcemanager.address.rm57	ResourceManager 中应用程序管理员界面地址。在本例中 test-piv-cent65x64-003.sv.splunk.com:8032
vix.yarn.resourcemanager.address.rm40	RM40 的 ResourceManager 中应用程序管理员界面地址：在本示例中，#test-piv-cent65x64-004.sv.splunk.com:8032
vix.yarn.resourcemanager.scheduler.address.rm57	RM57 的 ResourceManager 中计划程序界面地址。在本示例中， test-piv-cent65x64-003.sv.splunk.com:8030
vix.yarn.resourcemanager.scheduler.address.rm40	RM40 的 ResourceManager 中计划程序界面地址。在本例中： test-piv-cent65x64-004.sv.splunk.com:8030

vix.yarn.resourcemanager.ha.enabled	设为 true
vix.yarn.resourcemanager.cluster-id	当 ResourceManager 为高可用性.> # yarnRM 时使用的群集 ID
vix.yarn.application.classpath	$HADOOP_CLIENT_CONF_DIR, $HADOOP_CONF_DIR, $HADOOP_COMMON_HOME/*, $HADOOP_COMMON_HOME/lib/*, $HADOOP_HDFS_HOME/*, $HADOOP_HDFS_HOME/lib/*, $HADOOP_YARN_HOME/*, $HADOOP_YARN_HOME/lib/*
vix.mapreduce.application.classpath	= $HADOOP_MAPRED_HOME/*,$HADOOP_MAPRED_HOME/lib/*

REST API 参考

提供程序

新建和管理提供程序。

数据/索引

提供服务以新建和管理数据索引。

获取数据/vix 提供程序

在服务器上列出识别的提供程序。

请求

请参阅“常用 GET 请求参数”。

响应

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

在此 Splunk 实例上列出索引。

```
curl -k -u admin:pass https://localhost:8089/services/data/vix-providers
```

发布数据/vix 提供程序

用指定的名称新建提供程序。

请求

请参阅“常用 GET 请求参数”。

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

以下示例新建了一个名为 Shadow 的提供程序。

```
curl -k -u admin:pass https://localhost:8089/servicesNS/admin/search/data/vix-providers \
        -d name=Shadow
```

数据/vix 提供程序/{name}

删除数据/索引/{name}

移除 {name} 指定的索引（不仅仅是其中包含的数据）。

警告：此操作将删除提供程序的数据目录并从 `indexes.conf` 中移除提供程序的段落。

请求

请参阅“常用 GET 请求参数”。

响应

无

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

删除名为 shadow 的提供程序。

```
curl -k -u admin:pass --request DELETE https://localhost:8089/services/data/vix-providers/Shadow
```

获取数据/vix 提供程序/{name}

检索关于命名索引的信息。

请求

请参阅“常用 GET 请求参数”。

响应

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

列出关于 Shadow 索引的信息。

```
curl -k -u admin:pass https://localhost:8089/services/data/vix-providers
```

发布数据/vix 提供程序/{name}

用索引数据指定的信息更新 `{name}` 指定的提供程序。

请求

请参阅“常用 GET 请求参数”。

响应

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

以下示例更新名为 "Shadow" 的索引的最大大小，将大小设为 400000 MB。

此索引在此端点的 POST 操作示例中新建。

```
curl -k -u admin:pass https://localhost:8089/services/data/vix-providers/Shadow -d vix.*=Shadow
/data/indexes/shadow
```

索引

新建和管理数据索引。

数据/vix 索引

提供服务以新建和管理数据索引。

获取数据/vix 索引

在服务器上列出识别的索引。

请求

请参阅“常用 GET 请求参数”。

响应

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

在此实例上列出虚拟索引。

```
curl -k -u admin:pass https://localhost:8089/services/data/vix-indexes
```

发布数据/vix 索引

用指定的名称新建索引。

请求

请参阅“常用 GET 请求参数”。

响应

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

以下示例新建了一个名为 Shadow 的索引。

```
curl -k -u admin:pass https://localhost:8089/services/data/vix-indexes -d name=Shadow -d vix.provider=Shadow
```

数据/vix 索引/{name}

删除数据/vix 索引/{name}

移除 {name} 指定的索引（不仅仅是其中包含的数据）。

请求

请参阅“常用 GET 请求参数”。

响应

无

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

删除名为 shadow 的索引。

```
curl -k -u admin:pass --request DELETE https://localhost:8089/services/data/vix-indexes/Shadow
```

获取数据/索引/{name}

检索关于命名索引的信息。

请求

请参阅“常用 GET 请求参数”。

响应

无

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

列出关于 Shadow 索引的信息。

```
curl -k -u admin:pass https://localhost:8089/services/data/vix-indexes/Shadow
```

发布数据/vix 索引/{name}

用索引属性指定的信息更新 {name} 指定的数据索引。

请求

请参阅“常用 GET 请求参数”。

响应

无

HTTP 状态代码

请参阅“HTTP 状态代码表”。

示例

以下示例更新名为 "Shadow" 的索引的最大大小，将大小设为 400000 MB。

此索引在此端点的 POST 操作示例中新建。

```
curl -k -u admin:pass https://localhost:8089/services/data/vix-indexes/Shadow -d vix.*=Shadow
/data/indexes/shadow
```