银河麒麟服务器操作系统 V4

Sqoop 软件适配手册

天津麒麟信息技术有限公司

2019 年 5 月

目 录

1 概述

1.1 系统概述

银河麒麟服务器操作系统主要面向军队综合电子信息系统、金融系统以及电力系统等国家关键行业的服务器应用领域，突出高安全性、高可用性、高效数据处理、虚拟化等关键技术优势，针对关键业务构建的丰富高效、安全可靠的功能特性，兼容适配长城、联想、浪潮、华为、曙光等国内主流厂商的服务器整机产品，以及达梦、金仓、神通、南大通用等主要国产数据库和中创、金蝶、东方通等国产中间件，满足虚拟化、云计算和大数据时代，服务器业务对操作系统在性能、安全性及可扩展性等方面的需求，是一款具有高安全、高可用、高可靠、高性能的自主可控服务器操作系统。

1.2 环境概述

服务器型号	长城信安擎天 DF720 服务器
CPU 类型	飞腾 2000+处理器
操作系统版本	Kylin-4.0.2-server-sp2-2000-19050910.Z1
内核版本	4.4.131
sqoop 版本	1.99.7

1.3 Sqoop 软件简介

Apache Sqoop 是一种用于在 Apache Hadoop 和结构化数据存储（如关系数据库）之间高效传输批量数据的工具。可以使用 Sqoop 将数据从外部结构化数据存储导入 Hadoop 分布式文件系统或 Hive 和 HBase 等相关系统，或者用于从 Hadoop 中提取数据并将其导出到外部结构化数据存储区，例如关系数据库和企业数据仓库。

Sqoop 自动执行此过程的大部分过程，依靠数据库来描述要导入的数据的模式。 Sqoop 使用 MapReduce 导入和导出数据，提供并行操作以及容错。

2 Sqoop 软件适配

2.1 安装

Sqoop 作为一个二进制包发布，包含两个独立的部分 - 客户端和服务服务端。

服务端：需要在群集中的单个节点上安装服务端，此节点将用作所有 Sqoop 客户端的入口点;

客户端：客户端可以安装在任意数量的计算机上。

2.1.1 服务端安装

将 Sqoop 工件复制到要运行 Sqoop 服务端的计算机。Sqoop 服务端充当 Hadoop 客户端，因此必须在此节点上提供 Hadoop 库（Yarn，Mapreduce 和 HDFS jar 文件）和配置文件（core-site.xml，mapreduce-site.xml，...）。 您不需要运行任何 Hadoop 相关服务。

使用以下命令应该能够列出 HDFS，例如：

```
$ hadoop dfs -ls
```

Sqoop 目前支持 Hadoop 2.6.0 或更高版本。要安装 Sqoop 服务器，请解压缩 tarball（在您选择的位置）并将新创建的目录设置为工作目录。

```
# 解压缩 Sqoop
tar -xvf sqoop-<version>-bin-hadoop<hadoop-version>.tar.gz
# 将解压后的目录移动到要安装的目录
mv sqoop-<version>-bin-hadoop<hadoop version> /usr/lib/sqoop
# 切换到 sqoop 的安装目录
cd /usr/lib/sqoop
```

2.1.1.1 Hadoop 依赖

Sqoop 服务端需要以下指向 Hadoop 库的环境变量：

$HADOOP_COMMON_HOME,

$HADOOP_HDFS_HOME,

$HADOOP_MAPRED_HOME，

$HADOOP_YARN_HOME.

您必须确保定义了这些变量并指向有效的 Hadoop 安装路径。如果找不到 Hadoop 库，Sqoop 服务端将无法启动。

Sqoop 服务端依靠环境变量寻找 Hadoop 库。如果$HADDOOP_HOME 环境变量被设置，Sqoop 会在以下位置寻找 jar 包：

$HADOOP_HOME/share/hadoop/common,

$HADOOP_HOME/share/hadoop/hdfs,

$HADOOP_HOME/share/hadoop/mapreduce，

$HADOOP_HOME/share/hadoop/yarn.

您可以使用$ HADOOP_COMMON_HOME， $ HADOOP_HDFS_HOME，$ HADOOP_MAPRED_HOME 和$ HADOOP_YARN_HOME 环境变量独立指定 Sqoop 服务器应查找 common，hdfs，mapreduce 和 yarn jars 的位置。

2.1.1.2 Hadoop 配置

Sqoop 服务器需要模拟用户访问群集内外的 HDFS 和其他资源，作为开始给

予作业的用户而不是运行服务器的用户。 您需要配置 Hadoop 以通过所谓的代理用户系统明确允许此模拟。 您需要在 core-site.xml 文件中创建两个属性:

hadoop.proxyuser.$SERVER_USER.hosts

hadoop.proxyuser.$SERVER_USER.groups

其中$SERVER_USER 是将运行 Sqoop2 服务器的用户。在大多数情况下，这两个配置*就足够了。有关如何使用这些属性的详细信息，请参阅 Hadoop 文档。

服务器在 sqoop2 用户下运行时需要出现在 core-site.xml 文件中的示例片段：

```
<property>
  <name>hadoop.proxyuser.sqoop2.hosts</name>
  <value>*</value>
</property>
<property>
  <name>hadoop.proxyuser.sqoop2.groups</name>
  <value>*</value>
</property>
```

如果您在所谓的系统用户（ID 小于 min.user.id - 默认为 1000）下运行 Sqoop 2 服务器，则 YARN 默认拒绝运行 Sqoop 2 作业。 您需要将运行 Sqoop 2 服务器的用户名（最有可能是用户 sqoop2）添加到 container-executor.cfg 的 allowed.system.users 属性中。 有关更多详细信息，请参阅 YARN 文档。

服务器在 sqoop2 用户下运行时需要出现在 container-executor.cfg 文件中的示例片段：

```
allowed.system.users=sqoop2
```

2.1.1.3 第三方 jar 包

如果 Sqoop 服务端需要使用第三方的 jar 包，请在文件系统的任何位置创建一个目录，并将$SQOOP_SERVER_EXTRA_LIB 环境变量指向该位置。

```
# Create directory for extra jars
mkdir -p /var/lib/sqoop2/

# Copy all your JDBC drivers to this directory
cp mysql-jdbc*.jar /var/lib/sqoop2/
cp postgresql-jdbc*.jar /var/lib/sqoop2/

# And finally export this directory to SQOOP_SERVER_EXTRA_LIB
```

```
export SQOOP_SERVER_EXTRA_LIB=/var/lib/sqoop2/
```

2.1.1.4 配置“PATH”

所有用户或者管理员用到的命令都存储于安装目录的 bin 目录下，将该目录添加至“PATH”环境变量中。

2.1.1.5 配置服务端

服务端配置文件存储在 conf 目录中。 文件 sqoop_bootstrap.properties 指定应该使用哪个配置提供程序来加载其余 Sqoop 服务器的配置。 默认值为 PropertiesConfigurationProvider。

名为 sqoop.properties 的第二个配置文件包含可能影响 Sqoop 服务器的其余配置属性。配置文件已有详细记录，因此请检查所有配置属性是否适合您的环境。在大多数常见情况下，默认或非常小的调整应该足够了。

2.1.1.6 仓库初始化

首次启动 Sqoop 2 服务器之前，需要初始化元数据存储库。 使用 upgrade 工具初始化存储库：

```
sqoop2-tool upgrade
```

可以使用 verify 工具验证是否已正确配置所有内容：

```
sqoop2-tool verify
...
Verification was successful.
Tool class org.apache.sqoop.tools.tool.VerifyTool has finished correctly
```

2.1.1.7 启动和关闭服务端

安装和配置完成后，可以使用以下命令启动 sqoop 服务：

```
sqoop2-server start
```

以下命令可以停止服务端：

```
sqoop2-server stop
```

Sqoop 服务端默认使用 12000 端口，可以在配置文件 conf/sqoop.properties 中设置 org.apache.sqoop.jetty.port 为其它端口。

2.1.2 安装客户端

只需在目标计算机上复制 Sqoop 分发工件并将其解压缩到所需位置即可。可以使用以下命令启动客户端：

```
sqoop2-shell
```

客户端不充当 Hadoop 客户端，因此无需在客户端上安装 Hadoop 库和配置文件等。

2.2 使用示例

2.2.1 从 S3 导入至 HDFS

本节包含将数据从 S3 传输到 HDFS 的用例的详细说明。

2.2.1.1 用例

假设您在 S3 上有一个目录，一些外部进程正在创建新的文本文件。新文件将添加到此目录中，但不会更改现有文件。它们只能在一段时间后被移除。需要将所有新文件中的数据传输到单个 HDFS 目录。不需要保留文件名，并且可以将多个源文件合并到 HDFS 上的单个文件。

2.2.1.2 配置

我们将使用 HDFS 连接器连接进行数据传输的 From 和 To 侧。要为 S3 创建链接，您需要具有 S3 存储桶名称和 S3 访问权限以及密钥。 如果您还没有 S3 凭证，请按照 S3 文档检索 S3 凭据。

```
sqoop:000> create link -c hdfs-connector
```

我们的示例使用 s3link 作为链接名称

以 s3a：// $ BUCKET_NAME 的形式指定 HDFS URI，其中$ BUCKET_NAME 是 S3 存储桶的名称

使用“覆盖”配置选项，分别使用 S3 访问密钥和私钥指定 fs.s3a.access.key 和 fs.s3a.secret.key。

为 HDFS 创建链接：

```
sqoop:000> create link -c hdfs-connector
```

我们的示例使用 hdfslink 作为链接名称。如果您的 Sqoop 服务器在部署了 HDFS 和 mapreduce 客户端配置的节点上，则可以安全地将所有选项保留为空白，使用默认值。

通过链接 HDFS 和 S3，您可以创建将数据从 S3 传输到 HDFS 的作业：

```
sqoop:000> create job -f s3link -t hdfslink
```

我们的示例使用 s3import 作为作业名称;

输入目录应指向 S3 存储桶中生成新文件的目录;

确保为增量类型选择模式 NEW_FILES;

可以在 Output 目录中指定导入文件的最终目标;

确保启用追加模式，以便 Sqoop 可以将新创建的文件上载到 HDFS 上的同一目录;

根据需要配置其余选项。

最后，可以通过以下命令启动作业：

```
sqoop:000> start job -j s3import
```

可以定期运行作业 s3import，只传输新创建的文件。