文章
Frank Ma · 一月 5阅读大约需 分钟

在Hadoop大数据存储库库中使用SQL (Apache Hive)

本文译自 https://community.intersystems.com/post/using-sql-apache-hive-hadoop-big...

大家好,

在使用 Spark做Hadoop时， InterSystems IRIS有一个很好的连接器。但市场上也提供了大数据 Hadoop访问的其他优秀方案 -Aparche Hive。请区别：

HIVE	SPARK
Hive是一个数据库，用类似于 RDBMS 数据库的表格形式存储数据。	Spark 不是一个数据库，它是一个数据分析框架，可以在内存中对大至 PB字节的大容量数据进行复杂的数据分析。
使用称作 HiveQL的自己的 SQL 引擎，数据可以从 Hive 中抽取出来。只能使用 SQLs来抽取数据。	Spark既能使用复杂 SQLs（ Complex SQLs）也能使用 MapReduce 机制进行数据分析。它支持 Java, Scala 和Pyt 写的分析框架。
Hive在Hadoop之上运行。	Spark 没有自己专用的存储。实际上，它是从外部的分布式数据存储如运行在 Hadoop和MongoDB上的 Hive、HBase中抽取数据。
Hive是一个基数据仓库操的数据库	Spark 更适合在内存中进行复杂和快速的数据分析以及对数据流行流式处理。
对于那些需在可横向扩展的 RDBMS 数据库上运行数据仓库操的应用来说， Hive是最适合的。	Spark最适合于那些要求比 MapReduce 机制更快地进行大数据分析的应用。

来源： https://dzone.com/articles/comparing-apache-hive-vs-spark

我做了一个 PEX互操服务，可以让你在你的 InterSystems IRIS应用内部使用 Apache Hive。请试用如下步骤：

1. 在iris-hive-adapter 项目上做一个 Git Clone:

$ git **clone** https://github.com/yurimarx/iris-hive-adapter.git

2. 在这个目录内打开 terminal 并运行：

$ docker-compose build

3. 运行IRIS容器：

$ docker-compose up

4. 打开项目中的 Hive Prouction,运行一个Hello样例）：
http://localhost:52773/csp/irisapp/EnsPortal.ProductionConfig.zen?PRODUCTION=dc.irishiveadapter.HiveProduction

5. 点击"开始"运行 Production.

6. 现在我们来测试应用 !

7. 运行你的
REST客户端应用程序（比如Postman），在body部分使用项目的URLS和命令（使用POST请求）：

7.1 **在大数据中生成个新表**：POST http://localhost:9980/?Type=DDL. 在BODY中: CREATE TABLE helloworld (消息字符串)

7.2 **在表中插入**：POST http://localhost:9980/?Type=DDL. 在BODY中: INSERT INTO helloworld VALUES ("hello")

7.3 **T从表中得到结果清单**：POST http://localhost:9980/?Type=DML. 在BODY中: SELECT * FROM helloworld（注意：这里的类型是）

现在，你有了 2个在IRIS中使用大数据的选项：Hive 或者Spark。 **留**你喜欢。

#InterSystems IRIS

源 URL: https://cn.community.intersystems.com/post/%E5%9C%A8hadoop%E5%A4%A7%E6%95%B0%E6%8D%AE%E5%AD%98%E5%82%A8%E5%BA%93%E5%BA%93%E4%B8%AD%E4%BD%BF%E7%94%A8sql-apache-hive