使用Spark SQL读取HBase上的数据

近日，由华为团队开发的Spark-SQL-on-HBase项目通过Spark SQL/DataFrame并调用Hbase内置的访问API读取HBase上面的数据，该项目具有很好的可扩展性和可靠性。这个项目具有以下的特点：

1、基于部分评估技术，该项目具有强大的数据剪枝和智能扫描特点；
2、支持自定义过滤规则、协处理器等以便支持超低延迟的处理；
3、支持SQL、DataFrame；
4、支持更多的SQL（比如二级索引、布隆过滤、主键、批量加载以及更新）；
5、支持和其他数据源进行整合；
6、支持Python/Java/Scala；
7、支持Spark 1.4.0。

引入依赖

1、如果你使用的是spark-shell，可以如下操作：

```
> $SPARK_HOME/bin/spark-shell --packages Huawei-Spark:Spark-SQL-on-HBase:1.0.0
```

2、如果你使用SBT的话，在你们build.sbt文件加入一下依赖：

```
spDependencies += "Huawei-Spark/Spark-SQL-on-HBase:1.0.0"
```

或者

```
resolvers += "Spark Packages Repo" at "http://dl.bintray.com/spark-packages/maven"

libraryDependencies += "Huawei-Spark" % "Spark-SQL-on-HBase" % "1.0.0"
```

3、如果使用的是Maven，请在pom.xml文件加入一下依赖：

```
<dependencies>
  <!-- list of dependencies -->
```

```
  <dependency>
    <groupId>Huawei-Spark</groupId>
    <artifactId>Spark-SQL-on-HBase</artifactId>
    <version>1.0.0</version>
  </dependency>
</dependencies>

<repositories>
  <!-- list of other repositories -->
  <repository>
    <id>SparkPackagesRepo</id>
    <url>http://dl.bintray.com/spark-packages/maven</url>
  </repository>
</repositories>
```

项目地址以及如何使用

该项目已经被华为团队开源，代码依托在Github上：https://github.com/Huawei-Spark/Spark-SQL-on-HBase，可以在里面获取更多详细的文档和使用例子。