Spark 1.1.0发布:各个模块得到全面升级

今天我很激动地宣布Spark 1.1.0发布了，Spark 1.1.0引入了许多新特征（new features）包括了可扩展性和稳定性方面的提升。这篇文章主要是介绍了Spark 1.1.0主要的特性，下面的介绍主要是根据各个特征重要性的优先级进行说明的。在接下来的两个星期内，我们将会发表文章分别详细地介绍这些新组件，Spark 1.1已经在Databricks Cloud可用，用户也可以在Apache Spark官方网站进行下载。

Spark SQL的成熟

Spark 1.1.0版本中对Spark 1.0中的Spark SQL进行了重大的更新。在Databricks公司，我们已经将客户所有的workloads从Shark迁移到Spark SQL，全部有2X-5X的性能提升。 Spark 1.1为Spark SQL添加了一个JDBC server，它是一个最要的特性，允许直接依赖JDBC对Shark安装版进行更新。我们同时也开放了Spark SQL相应的系统API， 这允许大量的第三方数据源和Spark SQL进行集成。这将提供为以后的集成提供了扩展点，比如Datastax Cassandra driver.

利用这些类型API，我们已经提供了对直接读取JSON到Spark内置的ShemaRDD的支持。如下：

```
# Create a JSON RDD in Python
>>> people = sqlContext.jsonFile(“s3n://path/to/files...”)
# Visualize the inferred schema
>>> people.printSchema()
# root
#  |-- age: IntegerType
#  |-- name: StringType
```

MLlib的扩展

Spark’s machine learning library adds several new algorithms, including a library for standard exploratory statistics such as sampling, correlations, chi-squared tests, and randomized inputs. This allows data scientists to avoid exporting data to single-node systems (R, SciPy, etc) and instead directly operate on large scale datasets in Spark. Optimizations to

internal primitives provide a 2-5X performance improvement in most MLlib algorithms out of the box. Decision trees, a popular algorithm, has been ported to Java and Python. Several other algorithms have also been added, including TF-IDF, SVD via Lanczos, and nonnegative matrix factorization. The next release of MLlib will introduce an enhanced API for end-to-end machine learning pipelines.

Sources and Libraries for Spark Streaming

Spark streaming extends its library of ingestion sources in this release adding two new sources. The first is support for Amazon Kinesis, a hosted stream processing engine. Spark Streaming also adds H/A source for Apache Flume using a new data source which provides transactional hand-off of events from Flume to gracefully tolerate worker failures. Spark 1.1 adds the first of a set of online machine learning algorithms with the introduction of a streaming linear regression. Looking forward, the Spark Streaming roadmap will feature a general recoverability mechanism for all input sources, along with an ever-growing list of connectors. The example below shows training a linear model using incoming data, then using an updated model to make a prediction:

```
> val stream = KafkaUtils.createStream(...)
 // Train a linear model on a data stream
 > val model = new StreamingLinearRegressionWithSGD()
 .setStepSize(0.5)
 .setNumIterations(10)
 .setInitialWeights(Vectors.dense(...))
 .trainOn(DStream.map(record => createLabeledPoint(record))
 // Predict using the latest updated model
 > model.latestModel().predict(myDataset)
```

Performance in Spark Core

This release adds significant internal changes to Spark focused on improving performance for large scale workloads. Spark 1.1 features a new implementation of the Spark shuffle, a key internal primitive used by almost all data-intensive programs. The new shuffle improves performance by more than 5X for workloads with extremely high degree of parallelism, a key pain point in earlier versions of Spark. Spark 1.1 also adds a variety of other improvements to decrease memory usage and improve performance.

Optimizations and Features in PySpark

Several of the disk-spilling modifications introduced in Spark 1.0 have been ported to Spark's Python runtime extension. This release also adds support in Python for reading and writing data from SequenceFiles, Avro, and other Hadoop-based input formats. PySpark now supports the entire Spark SQL API, including support for nested types inside of SchemaRDD's.

The efforts on improving scale and robustness of Spark and PySpark are based on feedback from the community along with direct interactions with our customer workloads at Databricks. The next release of Spark will continue along this theme, with a focus on improving instrumentation and debugging for users to pinpoint performance bottlenecks.

This post only scratches the surface of interesting features in Spark 1.1. Head on over to

the official release notes to learn more about this release and stay tuned to hear more about Spark 1.1 from Databricks over the coming days!

本文原文：http://databricks.com/blog/2014/09/11/announcing-spark-1-1.html