Apache Spark 1.3.0正式发布

美国时间2015年3月13日Apache Spark 1.3.0正式发布，Spark 1.3.0是1.X版本线上的第四个版本，这个版本引入了DataFrame API，并且Spark SQL已经从alpha工程毕业了。Spark core引擎可用性也有所提升,另外MLlib和Spark Stream也有所扩展。Spark 1.3有来自60个机构的174魏贡献者带来的1000多个patch。

Spark Core

Spark 1.3中的Core模块的可用性得到了提升。Core API现在支持多级汇聚树（multi level aggregation trees ），有助于加快昂贵的Reduce操作。对某些疑难杂症的操作（certain gotcha operations）提升了错误报告。Spark的Jetty依赖现在 dependency 现在被隐藏，以帮助避免与用户程序冲突。Spark现在支持SSL加密。最后，实时的GC metrics信息和record数据被添加到UI页面上。

DataFrame API

Spark 1.3增加了DataFrames这个新的API，当和结构化数据操作时，它提供了强大并且方面的操作。DataFrame是从基本的RDD API中进化而来，它包含了命名字段（named fields）和模式信息（schema information）。我们可以很容易地通过源来创建DataFrame，比如Hive tables, JSON data，JDBC database或者任何实现Spark新的数据源API。当引入或者导出数据到其他系统中，Dataframes将会变成Spark组建中的一个非常通用的转换。Dataframes支持Python、Scala和Java语言。

Spark SQL

在这个版本中，Spark SQL 已经从alpha工程中毕业了，并且提供向后兼容HiveQL方言和稳定的编程API。 Spark SQL现在支持在数据源API中写表。一个全新的JDBC数据源允许用户从MySQL、Postgres以及其他的RMDB系统中导入或者导出数据。Spark SQL中的HiveQL还有很多微小的改变。Spark SQL也增加了支持模式演变的能力来在Parquet合并中提供兼容模式。

Spark ML/MLlib

In this release Spark MLlib introduces several new algorithms: latent Dirichlet allocation (LDA) for topic modeling, multinomial logistic regression for multiclass classification, Gaussian mixture model (GMM) and power iteration clustering for clustering, FP-growth for frequent pattern mining, and block matrix abstraction for distributed linear algebra. Initial support has been added for model import/export in exchangeable format, which will be expanded in future versions to cover more model types in Java/Python/Scala. The implementations of k-means and ALS receive updates that lead to significant performance gain. PySpark now supports the ML pipeline API added in Spark 1.2, and gradient boosted trees and Gaussian

mixture model. Finally, the ML pipeline API has been ported to support the new DataFrames abstraction.

Spark Streaming

Spark 1.3 introduces a new direct Kafka API (docs) which enables exactly-once delivery without the use of write ahead logs. It also adds a Python Kafka API along with infrastructure for additional Python API's in future releases. An online version of logistic regression and the ability to read binary records have also been added. For stateful operations, support has been added for loading of an initial state RDD. Finally, the streaming programming guide has been updated to include information about SQL and DataFrame operations within streaming applications, and important clarifications to the fault-tolerance semantics.

GraphX

GraphX增加了大量实用的函数，包括转换规范的边图(canonical edge graph)。

升级到Spark 1.3.0

Spark 1.3 和Spark 1.X版本兼容，所以你所写的代码完全不需要修改。但这不包括明确标记为不稳定的API。

Spark SQL API中的SchemaRDD目前被重命名为DataFrame. Spark SQL的迁移文档中介绍的很详细。. Spark SQL中对那些以前在列名称中使用到的保留字符 (比如“string”或者“table”) 需要用反引号进行转义。

比较重要的Issue

SPARK-6194: A memory leak in PySPark's collect().
SPARK-6222: An issue with failure recovery in Spark Streaming.
SPARK-6315: Spark SQL can't read parquet data generated with Spark 1.1.
SPARK-6247: Errors analyzing certain join types in Spark SQL.