

Apache Spark 2.2.0新特性详细介绍

Apache Spark 2.2.0

经

过了

大半年的

紧张开发，从RC1到RC6终于在今天正式发布了。由于时间的缘故，我并没有在[《Apache Spark 2.2.0正式发布》](#)文章中过多地介绍 Apache Spark 2.2.0

的新特性，本文作为补充将详细介绍Apache Spark 2.2.0 的新特性。

这个版本是 Structured Streaming

的一个重要里程碑，因为其终于可以正式在生产环境中使用，实验标签（experimental tag）已经被移除。在流系统中支持对任意状态进行操作；Apache Kafka 0.10 的 streaming 和 batch API支持读和写操作。除了在 SparkR, MLlib 和 GraphX 里面添加新功能外，该版本更多的工作在系统的可用性（usability）、稳定性（stability）以及代码的润色（polish）并解决了超过 1100 个tickets。



如果想及时了

解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

这篇文章中将详细介绍这些新特性，包括：

- Structured Streaming的生产环境支持已经就绪；
- 扩展 SQL 的功能；
- R 中引入了新的分布式机器学习算法；

- MLlib 和 GraphX 中添加了新的算法

Structured Streaming

Structured Streaming 是从 Spark 2.0 开始引入的，其提供了高层次的API来构建流应用程序；目的是提供一种简单的方式来构建端到端的流应用程序（end-to-end streaming applications），提供了一致性保证和容错方式。

从 Spark 2.2.0 开始，Structured Streaming 已经为生产环境的支持准备就绪，除了移除了实验性标签，还包括了一些高层次的变化，比如：

- Kafka Source and Sink: Apache Kafka 0.10 的 streaming 和 batch API支持读和写操作；
- Kafka Improvements: Kafka 到 Kafka 流操作中的producer 支持缓存以实现低延迟；
- Additional Stateful APIs: [flat]MapGroupsWithState 操作支持复杂的状态处理以及超时处理；
- Run Once Triggers：详情：[Running Streaming Jobs Once a Day For 10x Cost Savings](#)

SQL 和 Core APIs

自从 Spark 2.0 发布，Spark 已经成为大数据领域中功能最丰富并且符合标准的SQL查询引擎之一。它可以连接各种数据源，并且可以在这些数据上执行 SQL-2003 标准语句，包括分析函数以及子查询。Spark 2.2 还添加了许多 SQL 新功能，包括：

- API 更新: 统一了数据源和hive serde表的 CREATE TABLE 语法；SQL查询支持广播提示（broadcast hints）比如BROADCAST, BROADCASTJOIN, 以及 MAPJOIN；
- 总体性能和稳定性:
 - filter、join、aggregate、project 以及 limit/sample 操作支持基于成本优化器的基数统计（Cost-based optimizer cardinality estimation）；
 - 使用星型启发式（star-schema heuristics）来提升 TPC-DS 性能；
 - CSV 和 JSON 文件 listing/IO 性能提升；
 - HiveUDAFFunction 支持部分集合；
 - 引入基于JVM对象的聚合运算符
- 其他值得关注的改变:
 - 支持解析多行的JSON 和 CSV 文件
 - 分析分区表的命令

MLlib 和 SparkR

Spark 2.2.0 的最后一大变化主要集中在高级分析，MLlib 和 GraphX 添加了以下的新算法：

- 局部敏感哈希（Locality Sensitive Hashing）
- 多级逻辑回归（Multiclass Logistic Regression）
- 个性化PageRank（Personalized PageRank）

Spark 2.2.0还在 SparkR 中添加了以下分布式算法：

- 交替最小二乘 (ALS , Alternating Least Squares)
- 保序回归 (Isotonic Regression)
- 多层感知分类器 (Multilayer Perceptron Classifier)
- 随机森林 (Random Forest)
- 高斯混合模型 (Gaussian Mixture Model)
- 线性判别式分析 (Linear Discriminant Analysis, LDA)
- 多级逻辑回归 (Multiclass Logistic Regression)
- 梯度提升树 (Gradient Boosted Trees)
- Structured Streaming API 支持 R 语言
- R 中支持 to_json, from_json
- 支持Multi-column approxQuantile

随着这些算法的增加，SparkR已经成为 R 中最全面的分布式机器学习库。

本文翻译至：[Introducing Apache Spark 2.2](#)



优秀人才不缺工作机会，只缺适合自己的好机会。但是他们往往没有精力从海量机会中找到最适合的那个。

100offer 会对平台上的人才和企业进行严格筛选，让「最好的人才」和「最好的公司」相遇。注册 100offer，谈谈你对下一份工作的期待。一周内，收到 5-10 个满足你要求的好机会！

本博客文章除特别声明，全部都是原创！

禁止个人和公司转载本文、谢谢理解：过往记忆 (<https://www.iteblog.com/>)
本文链接: 【】 ()