Apache Spark 1.5重要的修改和Bug修复

Apache Spark 1.5版本目前正在社区投票中，相信到9月初应该会发布。这里先剧透一下Apache Spark 1.5版本的一些重要的修改和Bug修复。Apache Spark 1.5有来自220多位贡献者的1000多个commits。这里仅仅是列出重要的修改和Bug修复，详细的还请参见Apache JIRA changelog.

微信扫一扫，加关注
即可及时了解Spark、Hadoop或者Hbase等相关的文章
欢迎关注微信公共帐号：iteblog_hadoop

过往记忆博客(http://www.iteblog.com)专注于Hadoop、Spark、Flume、Hbase等技术的博客，欢迎关注。

Hadoop、Hive、Hbase、Flume等交流群：138615359和149892483

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

一、RDD/DataFrame/SQL APIs

- 全新的UDAF接口
- DataFrame hints for broadcast join
- expr function 来将SQL表达式转换成DataFrame的列
- 改进支持NaN
- StructType现在支持ordering
- TimestampType的精度减少到1us
- 多达100个新的内置表达式，其中包括date/time, string, math
- memory and local disk only checkpointing

二、DataFrame/SQL Backend Execution

- Code generation默认情况下开启
- 使用缓存友好的算法和外部算法来提升join, aggregation, shuffle, sorting
- 提升window function性能
- 为DF/SQL执行计划提供更好的metrics instrumentation和reporting

三、 Data Sources, Hive, Hadoop, Mesos and Cluster Management

- 动态资源分配现在支持所有的资源管理系统(Mesos, YARN, Standalone)
- 提升对Mesos的支持(framework authentication, roles, dynamic allocation, constraints)
- 提升对YARN的支持(dynamic allocation with preferred locations)
- 提升对Hive的支持 (metastore partition pruning, metastore connectivity to 0.13 to 1.2, internal Hive upgrade to 1.2)
- 在metastore中支持持久化与Hive兼容的数据格式
- JSON数据源支持数据分区
- 对Parquet文件支持提升(upgrade to 1.7, predicate pushdown, faster metadata discovery and schema merging, support reading non-standard legacy Parquet files generated by other libraries)
- 动态分区插入现在更加快速并且容错性更好
- DataSourceRegister接口为外部数据源提供了短名

四、SparkR

- 在YARN集群上运行R
- GLMs with R formula, binomial/Gaussian families, and elastic-net regularization
- 提供更好的错误提示信息
- 为DataFrame函数提供了别名，使得那些函数更符合R语言风格

五、Streaming

- 为突发的输入流提供了Backpressure（Backpressure for handling bursty input streams.）
- 为Python提供了更多的流数据源(Kafka offsets, Kinesis, MQTT, Flume)
- 为Python streaming 提供了更多的机器学习算法 (K-Means, linear regression, logistic regression)
- 内置就可以可靠的支持Kinesis stream
- 输入元数据诸如Kafka offsets现在可以在UI界面是看到详情
- 在集群上更好地进行负载均衡和调度receivers
- 在WEB ui界面里面引入了streaming storage

六、Machine Learning and Advanced Analytics

- Feature transformers: CountVectorizer, Discrete Cosine transformation, MinMaxScaler, NGram, PCA, RFormula, StopWordsRemover, and VectorSlicer.
- Estimators under pipeline APIs: naive Bayes, k-means, and isotonic regression.
- 算法: multilayer perceptron classifier, PrefixSpan for sequential pattern mining, association rule generation, 1-sample Kolmogorov-Smirnov test.
- 提升现有的算法: LDA, trees/ensembles, GMMs
- 为GraphX提供更加高效的Pregel API实现
- Model summary for linear and logistic regression.
- Python API: distributed matrices, streaming k-means and linear models, LDA, power iteration

clustering, etc.
- Tuning and evaluation: train-validation split and multiclass classification evaluator.
- Documentation: document the release version of public API methods

本博客文章除特别声明，全部都是原创！
原创文章版权归过往记忆大数据（过往记忆）所有，未经许可不得转载。
本文链接: 【】（）