Apache Flink 1.2.0正式发布及其功能介绍

大家期待已久的Apache Flink 1.2.0今天终于正式发布了。本版本一共解决了650个issues，详细的列表参见这里。Apache Flink 1.2.0是1.x.y系列的第三个主要版本；其API和其他1.x.y版本使用@Public标注的API是兼容的，推荐所有用户升级到此版本。更多关于Apache Flink 1.2.0新功能可以参见Apache Flink 1.2.0新功能概述

Dynamic Scaling / Key Groups

Flink streaming job 现在支持通过从带有不同并行度的保持点(savepoint)恢复来修改作业的并行度。整个作业的并行度和操作符(operator)的并行度的修改都是支持的。在StreamExecution Environment中，用户可以通过设置被称为“max parallelism”的参数为每个作业进行参数配置，这个参数决定了并行度的上限。在默认情况下，这个参数值的设置规则如下：
a、128：所有的并行度 128

下面内置函数和操作符支持rescaling：
a、Window operator
b、Rolling/Bucketing sink
c、Kafka consumers
d、Continuous File Processing source

write-ahead log Cassandra sink和CEP操作符目前不支持rescalable。使用keyed state接口的用户可以在不改变代码的情况下使用动态缩放(dynamic scaling)。

Rescalable Non-Partitioned State

作为动态扩展工作的一部分，社区同时为诸如没有使用keyed state而使用了运算符状态(operator state)的Kafka consumer添加了rescalable non-partitioned state功能。

在重新缩放(rescaling)的情况下，运算符状态(operator state)需要在并行消费者实例之间重新分配。在Kafka consumer例子里面，需要重新分配已经分配的分区和偏移量。

ProcessFunction

ProcessFunction函数是低级别的流处理操作，可访问所有（非循环）流应用程序的基本构建块(basic building blocks)，比如：
a、Events (stream elements)
b、State (fault tolerant, consistent)
c、Timers (event time and processing time)
ProcessFunction可以看作是一个可访问keyed state和定时器(timers)的FlatMapFunction。详细的文档请参见：https://ci.apache.org/projects/flink/flink-docs-release-1.2/dev/stream/process_function.html

异步I/O(Async I/O)

Flink现在有一个专用的异步I/O操作符，用于异步和以检查点方式进行阻塞调用。比如有许多Flink应用程序需要为流中每个元素查询外部数据存储区；为了避免因为外部系统的调用而降低流速，异步I/O操作符允许请求重叠。

使用Apache Mesos运行Flink

大家应该知道，直到Apache Flink 1.1.x，内置支持的集群管理主要包括：Standalone和Flink on Yarn。但是我们也都知道，Apache Mesos也是一款很不错的开源分布式资源管理框架；不过高兴的是，在Apache Flink 1.2.0，我们可以直接在Apache Mesos运行Flink！感谢EMC公司的贡献！

安全数据访问

Flink现在能够使用Kerberos对外部服务进行身份验证，例如Zookeeper，Kafka，HDFS和YARN；同时增加了对线上加密的实验支持。

可查询状态(Queryable State)

这个实验功能允许用户查询某个算子的当前状态。比如，如果你有个flatMap()算子为每个key进行聚合运算，可查询状态允许你随时连接到TaskManager来检索当前的聚合值。关于Queryable State详细的设计动机可以参见：/archives/1969.html。

向后兼容的保存点

Flink 1.2.0允许用户从1.1.4版本的保存点恢复，这使我们可以直接升级Flink版本，而不会丢失应用程序的状态；以下的内置算子是向后兼容的：
a、Window operator

b、 Rolling/Bucketing sink
c、 Kafka consumers
d、 Continuous File Processing source

Table API & SQL

Flink 1.2.0显著扩展了Flink的 Table API和SQL的性能、稳定性以及覆盖范围，而且支持了批处理和流处理的表。

在流处理方面，社区添加了tumbling, sliding以及session group-window aggregations功能，比如

table.window(Session withGap 10.minutes on 'rowtime as 'w)。

SQL支持更多的内置函数和算子，比如：EXISTS, VALUES, LIMIT, CURRENT_DATE, INITCAP, NULLIF；

Table API & SQL支持更多的数据类型，并支持更好的集成；

用户可以自定义的scalar functions，比如：

table.select('uid, parse('field) as 'parsed).join(split('parsed) as 'atom)

其他改进

a、 Metrics in Flink web interface
b、 支持Kafka 0.10
c、 Evictor Semantics