# Streaming(DataStream API)

原文参考:

<https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/datastream_api.html#collection-data-sources>

## Overview

* + 1. Flink DataStream Api 编程指南

在Flink中的DataStream 程序在数据流(data streams)上实现了各种转换(transformation)操作(如,filter,updating,state,window,aggregating 等)。Data Streams 可以从各种数据源(message queue,socket,fiels 等)中被创建。产生的结果可以输出到各种sink(目的地)，比如将它写入到数据文件或一些标准的输出当中。Flink 程序可以在各种环境中运行，如 standlone ,嵌入到其他程序中。Flink能在本地的JVM中执行，也可以在集群中运行（yarn）.

flink Api的基本概念请参照 [basic concepts](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/api_concepts.html)

为了创建你自己的Flink DataStream 程序，我们鼓励你一开始使用 [anatomy of a Flink Program](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/api_concepts.html#anatomy-of-a-flink-program) 并逐步的添加 [stream transformations](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/operators/index.html). 下面的章节将为添加一些operations(翻者注:Flink 中的任何的transformations)和高级特性做一些引用说明

|  |
| --- |
| Example 程序案例 |
| Data Source 数据源 |
| DataStream transformation |
| Data sink 数据输出 |
| Iterations |
| Execution Parametes 执行参数 |
| Fault Tolerance(故障容错) |
| Controlling Latency (延迟控制) |
| Debugging |
| Local Execution Envionment |
| Collection Data Sources |
| Iterator Data Sink |
| Where to go next (下一站)? |

* + 1. Example Program

下面的代码是一段完整的基于窗口的 word count 应用的例子，单词的数量来源于一个5秒窗口的socket . 你可以复制到本地并运行它。

|  |
| --- |
| Java 代码片段 |
| **public class** WindowWordCount {  **public static void** main(String[] args) **throws** Exception {   StreamExecutionEnvironment env = StreamExecutionEnvironment.*getExecutionEnvironment*();   DataStream<Tuple2<String, Integer>> dataStream = env  .socketTextStream(**"localhost"**, 9999)  .flatMap(**new** Splitter())  .keyBy(0)  .timeWindow(Time.*seconds*(5))  .sum(1);   dataStream.print();   env.execute(**"Window WordCount"**);  }   **public static class** Splitter **implements** FlatMapFunction<String, Tuple2<String, Integer>> {  @Override  **public void** flatMap(String sentence, Collector<Tuple2<String, Integer>> out) **throws** Exception {  **for** (String word: sentence.split(**" "**)) {  out.collect(**new** Tuple2<String, Integer>(word, 1));  }  }  } } |

|  |
| --- |
| Scala 代码片段 |
| import org.apache.flink.streaming.api.scala.\_ import org.apache.flink.streaming.api.windowing.time.Time /\*\*  \* Created by yuanhailong on 2018/9/19.  \*/ object WindowWordCount {  def main(args: Array[String]) {   val env = StreamExecutionEnvironment.getExecutionEnvironment  val text = env.socketTextStream("localhost", 9999)   val counts = text.flatMap { \_.toLowerCase.split("\\W+") filter { \_.nonEmpty } }  .map { (\_, 1) }  .keyBy(0)  .timeWindow(Time.seconds(5))  .sum(1)   counts.print()   env.execute("Window Stream WordCount")  } } |

为了运行这个例子，首先你需要启动在命令行终端用netcat 启动一个输入流：

|  |
| --- |
| nc –lk 9999 |

只要输入一些词就会返回一些新的单词。这些词将会成为word count 程序的输入。如果你想看到的结果大于1 。你只要重复的输入5秒钟之内相同的词即可。（如果你的输入不够快，你可以增加窗口大小）

* + 1. Data Sources [数据源]

数据源表示你的程序从哪里读取数据。通过StreamExecutionEnvironment.addSource(sourceFunction). 你能添加数据源到你的程序中。Flink 实现了几种数据源函数(function) ,但你可以通过实现SourceFunction 自定义数据源[翻者注：SourceFunction并行度1]。如果你想要实现多个并行度的数据源函数你可以通过实现ParallelSourceFunction 接口或者扩展RichParallelSourceFunction 。

有一些预先定义的数据源来源于StreamExecutionEnvironment。

file-based[基于文件的]:

* readTextFile(path)：读取文本文件，file 遵循TextInputFormat 规范，文本文件中的数据每一行作为一个字符串返回。
* readFile(fileinputFormat,path): 通过指定文件的输入格式来读取数据文件
* readFile(fileInputFormat, path, watchType, interval, pathFilter) ：这个方法的调用实际是通过上面两个方法中的一个来实现的。它使用给定的fileInputFormat读取指定路径下面的文件。根据提供的watchType. 数据源周期性（根据interval ms）的监控Path路径下的新数据(FileProcessingMode.PROCESS\_CONTINUOUSLY)。或者仅处理一次当前路径下面的数据然后退出（FileProcessingMode.PROCESS\_ONCE）。使用pathFilter排除不需要处理的数据。

IMPLEMENTATION（实现）：

在内部，Flink 将读数据程序划分为两个子任务(sub-task) ,也就是目录监控和数据读取。每个子任务通过独立的条目实现。监控是通过并行度为1的任务实现的。然而数据读取时通过多个任务并行实现的。并行度等于Job任务的并行度。目录监控任务去监控目录（根据watchType 周期性的监控或仅读取一次），找到文件，切割文件，并切割文件到下游readers . readers将读取实际的数据。每个切割的文件仅被一个readers 读取。然而一个readers 可以读取多个文件。

IMPORTANT NOTES（特别注意）:

1. 如果watchType 被设置为 FileProcessingMode.PROCESS\_CONTINUOUSLY。当files被修改的时候，它的整个内容将会被重新处理。这就会破坏“exactly-once”的语义，当追加数据到文件的末尾将导致所有的数据都会被重新处理。
2. 如果watchType 被设置为FileProcessingMode.PROCESS\_ONCE. 数据源只会被扫描一次然后退出，无需等待readers完成文件内容的读取[这里指的是监控服务]。当然readers 会持续读取文件内容直到文件内容读取完成.关闭source 将会导致此后的信息不会再有检查点。这将导致在节点失败后恢复变慢，因为Job需要从上一个检查点恢复

Socked-Based:

* socketTextStream: 从Socket中读取数据。通过指定分隔符切割数据

Collection-Based:

* fromCollection(Seq): 从java 的  Java.util.Collection 中创建data stream,集合中所有的元素必须具备相同的数据类型
* fromCollection(Iterator)：从Iterator中创建data stream. 该类指定迭代器返回的元素的数据类型。
* fromElements(elements: \_\*) : 从一系列的对象中创建data stream. 所有的对象必须具备相同的类型
* fromParallelCollection(SplittableIterator):从Iterator中创建data stream. 该类指定迭代器返回的元素的数据类型。
* generateSequence(from, to) ：在给定的范围类生成一系列的数字
  + 1. DataStream transformations

参见 [operators](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/operators/index.html)

* + 1. Data Sinks

Data sinks 消费 DataStream中的数据并将数据输出到files,socket，其他额外系统或print。 Flink 有多种输出格式它封装了DataStream上的背后的多种operators

* writeAsText() / TextOutputFormat: 写元素一行作为一个String . 这个Strings 通过调用每个元素的toString() 方法来获取。
* writeAsCsv(...) / CsvOutputFormat: 用逗号分隔value写元组(tuple). Row 和field分隔符可配置。Value通过调用toString（） 方法来获取。
* print() / printToErr()：打印每个元素toString()的value到标准输出。
* writeUsingOutputFormat() / FileOutputFormat：方法和自定义文件输出的基础类。支持自定义的对象到字节的转换
* writeToSocket：根据SerializationSchema 写元素到Socket
* addSink : 调用自定义的sink 函数。Flink 自带了多重sink 函数（如Apache kafka）

注意，在DataStrem上的WriteAs\*()方法主要是为了调试的目的。他们不会参加flink的chekpoint操作。这就意味着它使用的是“at-least-once”语义。数据如何刷写到目标系统依赖于实现的OutputFormat. 这就意味着不是发送到目标系统的数据会立即展现出来。当然，在失败的场景中，这些数据可能会丢失。

为了可靠性，stream exactly-once 传递到文件系统，可以使用flink-connector-filesystem。

* + 1. Iterations

Iterative streaming(迭代流)程序实现一个step 函数，并将其嵌入到IterativeStream中。由于一个DataStream程序可能永远都不会结束，因此没有最大的迭代次数。相反，你需要指定那些stream需要返回到iteration并且通过split或filter transformation 指定那些需要输出到下游。在这里，我们有一个iteration例子。代码的主体部分是一个简单的map 转换 ，并通过返回的元素区分不同的元素返回给下游。

|  |
| --- |
| **val** iteratedStream **=** someDataStream**.**iterate**(**  iteration => {  val iterationBody = iteration.map(/\* this is executed many times \*/)  (iterationBody.filter(/\* one part of the stream \*/), iterationBody.filter(/\* some other part of the stream \*/))  }) |

例如： 这里有一个程序冲一个整数中持续减1，直到它等于0

|  |
| --- |
| val someIntegers: DataStream[Long] = env.generateSequence(0, 1000)  val iteratedStream = someIntegers.iterate(  iteration => {  val minusOne = iteration.map( v => v - 1)  val stillGreaterThanZero = minusOne.filter (\_ > 0)  val lessThanZero = minusOne.filter(\_ <= 0)  (stillGreaterThanZero, lessThanZero)  }  ) |

* + 1. Execution Parameters

StreamExecutionEnvironment  包含ExecutionConfig  ，ExecutionConfig 允许为Flink运行时设置一些配置参数。

更多的参数参见[execution configuration](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/execution_configuration.html) 。这些参数属于DataStream API:

* setAutoWatermarkInterval(long millseconds):设置watermark发射的频率。通过getAutoWatermarkInterval可以得到当前的watermark的value.
  + - 1. Fault Tolerance(故障容错)

[State & Checkpointing](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/checkpointing.html) 描述了如何启用Flink的checkpoint 机制。

* + - 1. Controlling Latency

默认情况下，数据元素在网络上不是一对一的传输(如果这样将会导致不必要的网络延迟)而是先缓存起来。缓存（在两台机器上实际传输的对象）的大小在flink的配置文件中能被配置。为了更好的吞吐量这往往是一个好方法，但是当数据不足够快的时候会导致一定的数据延迟。为了控制吞吐量和延迟。在execution 环境上你可以通过env.setBufferTimeout(timeoutMillis)设置缓存等待被填满的最大等待时间。这样即使缓存区没有被填满也会被自动发送。这个timeout的默认值时100 ms

Usage:

|  |
| --- |
| val env: LocalStreamEnvironment = StreamExecutionEnvironment.createLocalEnvironment  env.setBufferTimeout(timeoutMillis)  env.generateSequence(1,10).map(myMap).setBufferTimeout(timeoutMillis) |

为了最大的吞吐量。set setBufferTimeout(-1)，这样会移除timeout并且只有当缓存区填满的时候才能被发送。为了最小的延迟，设置timeout = 0 来关闭缓存。Timeout=0 应该要去避免，因为这会引起服务性能下降。

* + 1. Debugging

在分布式集群中运行分布式程序之前，一个很好的办法是确定实现的算法按照期待的方式运行。因此，实现数据分析的程序通常是一个结果检查，调试，改善提高的过程。

Flink 在IDE内通过本地调试的方式提供了数据分析程序开发处理的特性，注入测试，收集数据。本小节将给一些提示如何开发Flink程序。

* + - 1. Local Execution Enviroment

LocalStreamEnvironment 在同一的JVM内启动内创建Flink System.如果你是在IDE里面启动LocalEnvironment 。你可以在你的代码中打断点这样就很容易去调试了。

|  |
| --- |
| val env = StreamExecutionEnvironment.createLocalEnvironment()  val lines = env.addSource(/\* some source \*/)  // build your program  env.execute() |

* + - 1. Collection Data Sources

Flink 为方便调试通过java collections 提供了一些特殊的数据源,一旦程序通过测试，source 和 sink 很容易被替换。

|  |
| --- |
| val env = StreamExecutionEnvironment.createLocalEnvironment()  // Create a DataStream from a list of elements  val myInts = env.fromElements(1, 2, 3, 4, 5)  // Create a DataStream from any Collection  val data: Seq[(String, Int)] = ...  val myTuples = env.fromCollection(data)  // Create a DataStream from an Iterator  val longIt: Iterator[Long] = ...  val myLongs = env.fromCollection(longIt) |

* + - 1. Iterator Data Sink

Flink 为调试和测试的目的提供了收集DataStream 结果的sink .可以像下面这样使用：

|  |
| --- |
| import org.apache.flink.streaming.experimental.DataStreamUtils  import scala.collection.JavaConverters.asScalaIteratorConverter  val myResult: DataStream[(String, Int)] = ...  val myOutput: Iterator[(String, Int)] = DataStreamUtils.collect(myResult.javaStream).asScala |

注：在flink 1.5.0 中 flink-streaming-contrib 被移除了。使用flink-streaming-java 和flink-streaming-scala 来替代

* + 1. Where to go next（下一步）?
* [Operators](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/operators/index.html): stream operators 规范说明
* [Event Time](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/event_time.html)：介绍flink的时间概念
* [State & Fault Tolerance](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/index.html):解释如何开发有状态的应用
* [Connectors](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/connectors/index.html)：描述有效的输入输出Connectors

## Event Time

|  |
| --- |
| **Event Time / Process Time / Ingestion time** |
| Setting a Time Characteristic |
| **Event Time and Watermarks [事件时间和水位]** |
| Watermarks in parallel Streams [并行流中的水位] |
| Late Elements [延迟元素] |
| Debugging Watermarks |

* + 1. Event Time / Process time / Ingestion Time

Flink 在流式程序中支持不同类型的时间概念:

* Process time:Porcess time 是执行各自operation时候机器系统时间

当streaming 程序采用processing time运行时，所有基于时间的操作（如 windows） 都将使用机器系统时钟来运行各自的operator. 每小时处理时间窗口将包含在系统时钟指向下一个整点时间之间的所有记录.例如，如果一个应用在早上9：15开始，那么第一个小时处理的时间窗口将包含上午9：15到10:00的数据。下一个窗口是10：到11:00

Processing time 是最简单的时间概念，它不需要在streams和机器之间协调。它提供了最高吞吐量以及最低的延迟。然而，在分布式和异步的环境中他不能保证正确性，因为它容易受到记录到达系统的影响，以及系统内部和外部的影响。

* Event time：Event time 的时间是在每台生产设备上发生的时间【既日志记录产生时间】。这个时间在到达Flink之前被嵌入到了日志里面，并且每条日志都能够抽取event timestamp. 在 event time 中，数据处理的时间依赖于数据本身。Event time的程序必须指定如何获取 Event Time Watermarks ,这是event time 内部信号处理的一种机制。Watermark 机制将在后面的章节描述[below](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/event_time.html#event-time-and-watermarks).

在一个非常完美的世界中，event time 程序产生一个始终一致和确定性的结果，而不用去理会event何时到达以及它的顺序。然而，除非event time确定数据是按次序到达的，否则，在等待event时，会导致事件处理的延迟。由于不可能无限的等待下去，这就限制了应用程序的event time

假设所有的event都到达了，event time 操作处理会如预期一样，并即使在event乱序和有有延迟的情况下都能产生正确和一致的结果。例如，一个小时的时间窗口将包含该时间内所有数据而不用管他们以什么顺序到达和何时到达(更多信息查看 [late events](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/event_time.html#late-elements))

注意：有时当event time 处理实时数据时，为了保证能够及时正确的处理他们会使用processing time 来进行操作。

* Ingestion time：Ingestion time 是event 进入flink的时间，每行记录在source operator 获得source的当前时间作为timestamp

Ingestion time 位于Event time 和 processing Time 之间 。对比processing time Ingestion time更加昂贵，但是他给定一个可预测的结果，因为Ingestion time给定了一个更加稳定的时间（在source一次性分配），不同的window操作会有相同的时间戳，然而在processing time中每个不同的window 操作都可能在相同的记录上分配不同的时间。

对比Event Time ，Ingestion time 不能处理乱序或延迟的event，但是程序不需要指定watermark.

实际上，Ingestion time 趋向于event time，只不过自动的分配了timestamp和watermark.



* + - 1. Setting a Time Characteristic

Flink DataStream 程序的开始通常都是设置一个时间特性。

下面是Flink程序在一个小时的窗口内做聚合的例子。

|  |
| --- |
| val env = StreamExecutionEnvironment.getExecutionEnvironment  env.setStreamTimeCharacteristic(TimeCharacteristic.ProcessingTime)  // alternatively:  // env.setStreamTimeCharacteristic(TimeCharacteristic.IngestionTime)  // env.setStreamTimeCharacteristic(TimeCharacteristic.EventTime)  val stream: DataStream[MyEvent] = env.addSource(new FlinkKafkaConsumer09[MyEvent](topic, schema, props))  stream  .keyBy( \_.getUser )  .timeWindow(Time.hours(1))  .reduce( (a, b) => a.add(b) )  .addSink(...) |

注：为了在这个案例中运行event time例子，程序需要为数据指定event time 和 watermark。或者程序在Source后注入一个 Time Assigner 和 Watermark。

下面的小节将描述时间戳和watermark背后的机制，。至于在Flink DataStream API中如何使用assignment 和watermark，请查阅 [Generating Timestamps / Watermarks](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/event_timestamps_watermarks.html).

* + 1. Event Time and Watermarks

注:Flink 从DataStream 模型实现了多种技术。 event time 和watermarks 更好的介绍，请参阅：

* [Streaming 101](https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101)
* The [Dataflow Model paper](https://research.google.com/pubs/archive/43864.pdf)

流处理器为支持event time 需要有一个方法去测量event time的进度。例如，当event时间超过一小时后，构建每小时窗口的窗口operator需要得到通知,因此Opertor 能关闭window的处理。

Event time的处理独立于processing time . 例如，一个operator的event time 时间可能稍微落后于processing time(考虑event的延迟)，而两者以相同的速度运行。另一方面，另外一个streaming 程序通过快速的消费缓存在kafka topic(或其他消息队列)的数据可能在几秒钟之内就能处理数周的event time 的数据。

在flink机制中监测event time的进度是通过watermark来实现。Watermark flow携带timestamp作为data stream 的一部分。Watermark 定义了在stream中到达的时间t ,这意味着从带有timestamp的stream中不应有更多的元素

下面的图片这展示了一个带时间戳的event stream和水位线。在这个例子中时间是有序的（通过各自的timestamp）,这意味着在stream中watamark较容易周期性标记



Watermark 更重要的是为无序的streams,正如下面阐述的一样，event不是有序的（通过timestamp） ,在一个通常的watermark定义了stream中的一个点，所有的event 在某个确定的timestamp都应该到达。一旦watermark到达operator，operator可以将其内部event time 时钟 提前到 watermark值。



* + - 1. Watermarks in parallel Streams

Watermarks是在source function 生成或在其后指定的，source function的每个并行子任务通常都生成各自的watermark. watermark在特定的并行source上定义event time .

当watermarks通过stream程序流动时，它将为其后续operator在下游生成一个新的watermarks。

有些operators消费多个input stream, 比如 union ,keyBy ,partition 等函数。这些operators的当前event time 是他们input stream 最小的event time .就像input streams 更新它的event time 一样。

下面的图显示了event 和 watermark 通过并行stream 流动的示意图 。operator 跟踪event time.



注：kafka 支持每个分区的watermark，更多参见[here](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/event_timestamps_watermarks.html#timestamps-per-kafka-partition).

* + - 1. Late Elements

有可能某些元素将违反水位的条件，这意味着watermark发生后，还会有部分stream 元素。实际上，在真实的世界中，某些元素就是延迟到达的，我们无法指定某个事件时间戳的发生时间。此外，即使延迟的元素是有界的，延迟太多通常是不被建议的，因为在window操作中会引起更多的延迟。

基于这个原因，streaming 能够显示的指定一些延迟的元素，late element的element是指在系统事件时间时钟(由watermark发出信号)已经超过late element时间戳后到达的element,更多资料参见[Allowed Lateness](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/operators/windows.html#allowed-lateness)

* + - 1. Debugging Watermarks

参见[Debugging Windows & Event Time](https://ci.apache.org/projects/flink/flink-docs-release-1.6/monitoring/debugging_event_time.html)

## State & Fault Tolerance（状态&容错）

* + 1. Overview(概览)

有状态的function和operator通过处理独立的element/event 存储数据，这使得状态成为了构建任何复杂operator的关键组件。

例如：

* 当application搜索某个event模型时，状态将存储到目前为止一系列的事件
* 当每分钟/小时/天发生一个聚合事件时, 状态将保留这些被悬置(还没有到达聚合时间)聚合数据
* 当在一个数据流上训练一个机器学习模型时，状态将保存当前版本的模型参数。
* 当历史数据需要被管理时，状态允许有效的接受过去发生过的event

Flink系统需要知道状态，以便能够使用[checkpoints](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/checkpointing.html) 标记状态容错以及使用streaming application 的[savepoints](https://ci.apache.org/projects/flink/flink-docs-release-1.6/ops/state/savepoints.html)

关于Flink的状态它允许你在应用中进行调整，这意味着你在重新划分并行度的时Flink会管理好状态

Flink [queryable state](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/queryable_state.html)  特征允许你在运行期间接收外部状态。

使用状态进行工作，对 [Flink’s state backends](https://ci.apache.org/projects/flink/flink-docs-release-1.6/ops/state/state_backends.html)[状态恢复]非常有用，Flink 通过指定状态如何存储以及存储在哪里提供了多种状态恢复机制。State可以存储在本地JVM的堆内和堆外内存。Flink管理应用的state, 这意味着flink为了允许能够管理非常大的状态需要自己处理内存管理（状态也可能被溢写到磁盘）。在无需改变任何应用代码的情况下可以配置State backends。

* + 1. Working with state(状态处理)

这篇文档将解析在开发flink应用的时如何使用flink的状态。

|  |
| --- |
| **Keyed State And Operator State** |
| Keyed State |
| Operator State |
| **Raw and Manager State** |
| **Using Managed Keyed State** |
| State Time-to-Live (TTL) |
| State in the Scala DataStream Api |
| **Using Managed Operator State** |
| Stateful Source Functions |

* + - 1. Keyed State And Operator State

在Flink的状态中有两个基本的状态概念 Keyed State 和 Operator State

* + - * 1. Keyed State

Keyed State 状态总是和key相关的并且他只能使用在KeyedStream的function和Operator上。

你可以将Keyed State 看做是一个已经分区分片并的Operator State，它的每个key只有一个状态分区。每个Keyed State 逻辑绑定到唯一的一组< parallel-operator-instance, key >, 由于每个key “属于” 一个keyed operator的并行实例，我们可以简单的认为<operator,key>

Keyed State 进一步组织为所谓的Key Groups，Key Group是 通过Flink 重新分配Keyed State 的原子单元。Key group 数量和定义的最大并行度相同。在执行过程中，每keyed operator 的并行实例都使用一个或多个Key groups

* + - * 1. Operator State

Operator 状态(非 keyed state)，每个operator 状态绑定到一个并行操作实例。在Flink的状态使用中 [Kafka Connector](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/connectors/kafka.html)是一个很好的例子。Kafka的消费者实例的每个分区维护了topic分区和偏移量的映射作为它的Operator State

当并行性发生变化时，operator status 接口支持在并行操作符实例之间重新分配状态。他们之间可以有不同的scema

* + - 1. Raw和 Managed 状态

Keyed State 和 Operator State 存在两种形式 Managed 和 Raw。

Managed State 是 被Flink 运行时管理的数据结构。例如哈希表或RocksDB. 例如”ValueState”,”ListState” 等。Flink 运行时对状态编码并写入到checkpoints

Raw State 是 Operator 保留在它们自己的数据结构内的状态。当发生checkpoint的时候Operator只需要将一系列的字节写入到checkpoint中。对于state的数据结构Flink一无所知，它只仅能看到原始的字节。

所有的datastream 函数都能使用Managed State, 但是raw state 接口只有当实现了Opertator的时才能被使用。我们推荐使用Managed State。因为使用Managed State当Flink的并行度发生变化的时候Flink能自动的重新分布state和更好内存管理。

注意：如果你的managed state 需要自定义序列化逻辑，为了消除兼容性请查阅[corresponding guide](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/custom_serialization.html) 。 Flink的序列化在默认情况下不需要做特殊处理。

* + - 1. Using Managed Keyed Status

Managed keyed stated 接口提供了对不同类型状态的访问，这些状态的作用域是当前输入元素的key . 这意味着这些状态的key 仅能使用在keyedStream上。KeyStream可以通过stream.keyBy()创建。

现在，我将首先探讨可用状态的不同类型，然后我将看下如何在在程序中使用它们。有效状态原语是：

* ValueState<T>:它保留了一个可以被更新和获取的一个value(输入元素key的作用于上面已经提到了，因此这里每个key可能拥有一个value.可以使用update(T)设置该值以及 T value（） 方法获取value
* ListState<T>：他保留了一组元素。你可以追加元素和在当前所有存储的元素上获得一个Iterable 。 添加元素可以使用add(T) 和 addAll(List(T))方法。通过使用terable<T> get() 获取这个迭代器（Iterable ）。你也可以使用update(List(T))方法覆盖已经存在的List.
* ReducingState<T>: 表示一个所有的value聚合后添加到State的单个value. 这个接口类似于ListState，只不过使用add(T)添加的元素通过使用一个特殊的ReduceFunction进行聚合。
* AggregatingState<IN, OUT>：表示所有的value聚合后添加到state的一个value. 对比ReducingState 。 聚合的类型可能和被添加到状态的元素类型不同。这个接口和ListState一样，只不过使用add(IN)添加的元素通过使用一个特殊的AggregateFunction进行聚合。
* FoldingState<T, ACC>:示所有的value聚合后添加到state的一个value. 对比ReducingState. 聚合的类型可能和被添加到状态的元素类型不同，只不过使用add(IN)添加的元素通过使用一个特殊的FoldFunction进行聚合。
* MapState<UK, UV>:它保留了一个list的映射，你可以设置键值对到到state中以及从当前存储的state mapping 中获取一个Iterable 。 Mapping的添加可以通过put(UK, UV) 和 putAll(Map<UK, UV>)方法。Value 通过一个相关的key获取。Iterable 的mappings，keys和values 可以通过entries(), keys()和 values()来分别获取。

所有的类型都可以通过一个clear()方法来清空当前Active key的state.

注意：FoldingState 和FoldingStateDescriptor 在Flink1.4中已经过时，在未来的版本中将会被彻底移除。请使用AggregatingState和AggregatingStateDescriptor替代。

需要牢记的是这些state object（状态对象）只能和state交互。状态不一定存储在内存，他有可能存储在磁盘或者其他地方。第二个要记住的是从你从state中获得的value 取决于输入元素的key. 所以你获得的value可能会在不同的调用中不相同。

通过创建StateDescriptor 可以获得state的句柄(handle). StateDescriptor保留状态的名字（后面你会看到，你创建了若干state,由于他们都会有唯一的名称所以你能引用它们），状态保存的值类型(value type) ，还可能有一些用户函数，比如ReduceFunction 。 根据你想获取的状态类型，你可以要么创建一个ValueStateDescriptor，ListStateDescriptor，ReducingStateDescriptor，FoldingStateDescriptor 要么创建MapStateDescriptor。

使用RuntimeContext来访问State. 因此他只可能在 rich function中。关于rich function 的更多信息请查看[here](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/api_concepts.html#rich-functions) 。这里我们有一个简短的例子。RichFunction中有有效的RuntimeContext为访问state有这些方法：

* ValueState<T> getState(ValueStateDescriptor<T>)
* ReducingState<T> getReducingState(ReducingStateDescriptor<T>)
* ListState<T> getListState(ListStateDescriptor<T>)
* AggregatingState<IN, OUT> getAggregatingState(AggregatingState<IN, OUT>)
* FoldingState<T, ACC> getFoldingState(FoldingStateDescriptor<T, ACC>)
* MapState<UK, UV> getMapState(MapStateDescriptor<UK, UV>)

这是一个FlatMapFunction  的例子并展示了各个部分如何在组合在一起的。

|  |
| --- |
| Java 代码 |
| */\*\*  \* Created by yuanhailong on 2018/9/25.  \*/* **public class** CountWindowAverage **extends** RichFlatMapFunction<Tuple2<Long,Long>,Tuple2<Long,Long>> {  */\*\*  \* ValueState 句柄 第一个元素是数量 第二个元素是求和。  \*/* **private transient** ValueState<Tuple2<Long, Long>> **sum**;   @Override  **public void** flatMap(Tuple2<Long, Long> input, Collector<Tuple2<Long, Long>> out) **throws** Exception {  *//访问 state* Tuple2<Long, Long> currentSum = **sum**.value();   *// 更新数量* currentSum.**f0** += 1;   *//第二个元素求和* currentSum.**f1** += input.**f1**;   *//更新状态* **sum**.update(currentSum);   *//如果状态达到2 ，发送平均值并清空* **if**(currentSum.**f0**>=2){  out.collect(**new** Tuple2<>(input.**f0**,currentSum.**f1**/currentSum.**f0**));  **sum**.clear();  }  }   @Override  **public void** open(Configuration parameters) **throws** Exception {  ValueStateDescriptor<Tuple2<Long,Long>> descriptor=  **new** ValueStateDescriptor<Tuple2<Long, Long>>(  "average",  TypeInformation.of(new TypeHint<Tuple2<Long, Long>>() {})，  Tuple2.of(0L,0L)  );   **sum**=getRuntimeContext().getState(descriptor);  } }  env.fromElements(Tuple2.of(1L, 3L), Tuple2.of(1L, 5L), Tuple2.of(1L, 7L), Tuple2.of(1L, 4L), Tuple2.of(1L, 2L))  .keyBy(0)  .flatMap(new CountWindowAverage())  .print(); |

|  |
| --- |
| Scala 代码 |
| class CountWindowAverage extends RichFlatMapFunction[(Long, Long), (Long, Long)] {  private var sum: ValueState[(Long, Long)] = \_  override def flatMap(input: (Long, Long), out: Collector[(Long, Long)]): Unit = {  // access the state value  val tmpCurrentSum = sum.value  // If it hasn't been used before, it will be null  val currentSum = if (tmpCurrentSum != null) {  tmpCurrentSum  } else {  (0L, 0L)  }  // update the count  val newSum = (currentSum.\_1 + 1, currentSum.\_2 + input.\_2)  // update the state  sum.update(newSum)  // if the count reaches 2, emit the average and clear the state  if (newSum.\_1 >= 2) {  out.collect((input.\_1, newSum.\_2 / newSum.\_1))  sum.clear()  }  }  override def open(parameters: Configuration): Unit = {  sum = getRuntimeContext.getState(  new ValueStateDescriptor[(Long, Long)]("average", createTypeInformation[(Long, Long)])  )  }  }  object ExampleCountWindowAverage extends App {  val env = StreamExecutionEnvironment.getExecutionEnvironment  env.fromCollection(List(  (1L, 3L),  (1L, 5L),  (1L, 7L),  (1L, 4L),  (1L, 2L)  )).keyBy(\_.\_1)  .flatMap(new CountWindowAverage())  .print()  // the printed output will be (1,4) and (1,5)  env.execute("ExampleManagedState")  } |

这个例子实现了一个很简单的窗口计数器。我们的key是元组(tuple)的第一个字段，函数存储count和运行的sum到ValueState中。一旦count到达2时它将发送一个平均值并清空状态。所以他就会又从0开始。注：如果我们的元组（tuple）第一个字段不同的value 它将保存不同的状态

* + - 1. State Time-To-Live(TTL)

任何类型的Keyed State都能够分配Time-To-Live(TTL). 如果说配置了TTL并且State已经过期，stored value 将被清空，下面详细讨论

所有state集合类型支持每个条目级别的TTLs.这就意味着List和Map元素的TTLs是相互独立的。

为了使用state TTL 必须首先构建StateTtlConfig 配置对象。通过传递配置，TTL的功能可以在任何state descriptor 中启用。

|  |
| --- |
| Java 代码 |
| import org.apache.flink.api.common.state.StateTtlConfig**;**  import org.apache.flink.api.common.state.ValueStateDescriptor**;**  import org.apache.flink.api.common.time.Time**;**  StateTtlConfig ttlConfig **=** StateTtlConfig  **.**newBuilder**(**Time**.**seconds**(**1**))**  **.**setUpdateType**(**StateTtlConfig**.**UpdateType**.**OnCreateAndWrite**)**  **.**setStateVisibility**(**StateTtlConfig**.**StateVisibility**.**NeverReturnExpired**)**  **.**build**();**    ValueStateDescriptor**<**String**>** stateDescriptor **=** **new** ValueStateDescriptor**<>(**"text state"**,** String**.**class**);**  stateDescriptor**.**enableTimeToLive**(**ttlConfig**);** |

|  |
| --- |
| Scala 代码 |
| **import** org.apache.flink.api.common.state.StateTtlConfig  **import** org.apache.flink.api.common.state.ValueStateDescriptor  **import** org.apache.flink.api.common.time.Time  **val** ttlConfig **=** **StateTtlConfig**  **.**newBuilder**(Time.**seconds**(**1**))**  **.**setUpdateType**(StateTtlConfig.UpdateType.OnCreateAndWrite)**  **.**setStateVisibility**(StateTtlConfig.StateVisibility.NeverReturnExpired)**  **.**build    **val** stateDescriptor **=** **new** **ValueStateDescriptor[String](**"text state"**,** classOf**[String])**  stateDescriptor**.**enableTimeToLive**(**ttlConfig**)** |

这个配置有若干个选项值得考虑：

第一个方法的参数newBuilder是强制的，它设置TTL.

Update 类型在更新TTL时配置（默认是OnCreateAndWrite）

* StateTtlConfig.UpdateType.OnCreateAndWrite ： 仅创建和写入的时候访问
* StateTtlConfig.UpdateType.OnReadAndWrite ： 包含读取

状态的可见性配置是指过期的value如果在读取的时候没有被清除是否返回（默认NeverReturnExpired）

* StateTtlConfig.StateVisibility.NeverReturnExpired：永不返回过期值
* StateTtlConfig.StateVisibility.ReturnExpiredIfNotCleanedUp：如果有效返回过期value

这个例子中NeverReturnExpired，过期状态的行为就好像它已经不存在了，即使它仍然需要被移除。这个选项在TTL发生后访问变得不在有效特别有用。

另外一个选项ReturnExpiredIfNotCleanedUp 在清空之前允许返回过期数据。

**Notes:**

* State 背后存储了用户最后一次修改的时间戳，这意味着启用这个特征是增量消耗存储的状态。Heap State 在内存中额外存储了一个引用用户对象和原始long 值得java对象。RocksDB状态后端为每个存储value、list或map添加8字节
* 目前仅支持processing time 的TTL
* 之前没有配置TTL尝试恢复State,将导致兼容性失败和StateMigrationException
* TTL配置不是checkpoint和savepoint的一部分而是Flink如何在当前运行的job中如何处理的一种方式。
* 如果用户序列化可以处理Null值那么使用TTL的map state可以处理null 值。如果不支持Null值可以采用NullableSerializer包装，代价是1个字节的额外消费。
  + - * 1. 清空过期状态

目前，过期的值只有在显示的通过调用ValueState.value()读取时才会被删除。

注意:这意味着默认清空下过期的value 如果没有被读取是不会被移除的。可能导致state的增长。这在未来发布版本中被修改。

此外,你可以在获得完整的快照时候激活清空操作这样将减少state大小。 local state在目前的实现中不会被删除。但是当它从上一个快照恢复的时候不会包含过期的数据。在StateTtlConfig可以配：

|  |
| --- |
| Java 代码 |
| import org.apache.flink.api.common.state.StateTtlConfig**;**  import org.apache.flink.api.common.time.Time**;**  StateTtlConfig ttlConfig **=** StateTtlConfig  **.**newBuilder**(**Time**.**seconds**(**1**))**  **.**cleanupFullSnapshot**()**  **.**build**();** |

|  |
| --- |
| Scala 代码 |
| **import** org.apache.flink.api.common.state.StateTtlConfig  **import** org.apache.flink.api.common.time.Time  **val** ttlConfig **=** **StateTtlConfig**  **.**newBuilder**(Time.**seconds**(**1**))**  **.**cleanupFullSnapshot  **.**build |

这个选项不适合在rockdbs state 增量的checkpoint

* + - 1. State In Scala Data Stream Api

除了上面的描述的接口外，Scala Api 在KeyedStream单个ValueState 的map和flatmap函数上有一些快捷方法。这个用户函数获得当前ValueState状态的value.并且他必须返回一个更新的value用户更新state.

|  |
| --- |
| **val** stream**:** **DataStream[(String**, **Int)]** **=** **...**  **val** counts**:** **DataStream[(String**, **Int)]** **=** stream  **.**keyBy**(\_.**\_1**)**  **.**mapWithState**((**in**:** **(String,** **Int),** count**:** **Option[Int])** **=>**  count **match** **{**  **case** **Some(**c**)** **=>** **(** **(**in**.**\_1**,** c**),** **Some(**c **+** in**.**\_2**)** **)**  **case** **None** **=>** **(** **(**in**.**\_1**,** 0**),** **Some(**in**.**\_2**)** **)**  **})** |

* + - 1. Using Managed State Operator

为了使用managed operator state ，一个有状态的函数要么实现CheckpointedFunction接口要么实现ListCheckpointed<T extends Serializable>  接口

* + - * 1. CheckpointedFunction

CheckpointedFunction接口通过重分配不同的schema提供对non-keyed state的访问。实现它必须要实现两个方法

|  |
| --- |
| **void** **snapshotState(**FunctionSnapshotContext context**)** **throws** Exception**;**  **void** **initializeState(**FunctionInitializationContext context**)** **throws** Exception**;** |

无论一个checkpoint合适执行，snapshotState（）方法总是会被调用（既只要执行快照就会执行该方法）。initializeState（） 是每次用户定义的函数初始化的时候被调用或从之前的checkpoint中恢复的时候进行调用。因此，initializeState() 不仅是不同的状态类型被初始化的地方而且也包含了逻辑的恢复。

目前，list-style managed operator state是被支持的。这个state期望是一个序列化的List对象，彼此相互独立，因而在重新调整时可以重新分配state. 换句话说，这些对象是可以重新分配non-keyed state的最佳粒度。根据状态可访问方法，定义了以下重分配方案：

* **Even-split redistribution：**每个operator返回一组state元素，所有列表的整个状态在逻辑上是串联的。在恢复/重分配，该列表被均匀划分为与并行operator相同数量的子列表。每个operator都会得到一个子列表，这个字列表可能是空的或者包含一个或者多个元素。作为一个例子，如果一个包含了element1和element2的operator的checkpoint state 使用的并行度为1 ，那么当增加并行度到2的时，element1 可能在operator第0个实例中而element2可能在实例1中。
* **Union redistribution：**每个operator返回一组state元素，所有列表的整个状态在逻辑上是串联的。在恢复/重分配，每个operator都将获得一个完整的一组state element.

下面的例子是一个有状态的SinkFunction 。 在发送他们到外部之前使用CheckpointedFunction 缓存元素。这里演示了基本的 even-split redistribution 列表状态。

|  |
| --- |
| Java 代码 |
| **package** com.feicheng.java.flink.office.example;  **import** org.apache.flink.api.common.state.ListState; **import** org.apache.flink.api.common.state.ListStateDescriptor; **import** org.apache.flink.api.common.typeinfo.TypeHint; **import** org.apache.flink.api.common.typeinfo.TypeInformation; **import** org.apache.flink.api.java.tuple.Tuple2; **import** org.apache.flink.runtime.state.FunctionInitializationContext; **import** org.apache.flink.runtime.state.FunctionSnapshotContext; **import** org.apache.flink.streaming.api.checkpoint.CheckpointedFunction; **import** org.apache.flink.streaming.api.functions.sink.SinkFunction;  **import** java.util.ArrayList; **import** java.util.List;  */\*\*  \* Created by yuanhailong on 2018/9/26.  \*/* **public class** BufferingSink **implements** SinkFunction<Tuple2<String,Integer>> ,CheckpointedFunction{   **private final int threshold**;   **private transient** ListState<Tuple2<String,Integer>> **checkpointedState**;   **private** List<Tuple2<String, Integer>> **bufferedElements**;    **public** BufferingSink(**int** threshold){  **this**.**threshold**=threshold;  **this**.**bufferedElements** = **new** ArrayList<>();  }   @Override  **public void** invoke(Tuple2<String, Integer> value, Context context) **throws** Exception {  **bufferedElements**.add(value);  **if**(**bufferedElements**.size()==**this**.**threshold**){  **for**(Tuple2<String,Integer> element:**bufferedElements**){  *//****TODO 发送到sink*** }  }  }   @Override  *//只要发生快照就会执行该方法* **public void** snapshotState(FunctionSnapshotContext functionSnapshotContext) **throws** Exception {  *//快照繁盛的时候清空之前的快照信息，添加现有的数据到快照* **checkpointedState**.clear();  **for** (Tuple2<String, Integer> element : **bufferedElements**) {  **checkpointedState**.add(element);  }  }   @Override  *//初始化，快照恢复的时候回执行该方法* **public void** initializeState(FunctionInitializationContext context) **throws** Exception {  ListStateDescriptor<Tuple2<String, Integer>> descriptor =  **new** ListStateDescriptor<Tuple2<String, Integer>>(  **"buffered-elements"**,*//名称必须全局唯一* TypeInformation.*of*(**new** TypeHint<Tuple2<String, Integer>>() {}));   **checkpointedState** = context.getOperatorStateStore().getListState(descriptor);    **if**(context.isRestored()){  **for** (Tuple2<String, Integer> element : **checkpointedState**.get()) {  **bufferedElements**.add(element);  }  }  } } |

initializeState 获取到一个FunctionInitializationContext参数。它是用于初始化non-keyed state的“容器”。是Liststate 类型检查点存储的位置。

注：状态初始化类似于keyed state ,使用StateDescriptor 包含状态的名字以及和value有关的一些类型信息。

|  |
| --- |
| ListStateDescriptor<Tuple2<String, Integer>> descriptor =  **new** ListStateDescriptor<Tuple2<String, Integer>>(  **"buffered-elements"**,*//名称必须全局唯一* TypeInformation.*of*(**new** TypeHint<Tuple2<String, Integer>>() {}));   **checkpointedState** = context.getOperatorStateStore().getListState(descriptor); |

State 访问方法的命名包含了重分配模式以及它的数据结构。例如，为了使用list state 在restore上使用union redistribution scheme[详见上文描述] 。通过使用getUnionListState(descriptor) 方法获得state. 如果名字不包含重分配模式那么使用getListState(descriptor) ，这意味着只是简单的使用even-split redistribution 方案。

Container 被初始化之后，我们使用context的isRestored() 方法去检查在失败后是否恢复成功。如果这个值等于true，那么将使用恢复逻辑进行恢复。

正如在BufferingSink 的代码片段中展现的一样，ListState 恢复 state初始化期间保留在类变量中为将来在snapshotState（）中使用。ListState 对象被清空包含之前的checkpoint，然后填充我们想要checkpoint的元素到ListState 。

* + - * 1. ListCheckpointed

ListCheckpointed结果是CheckpointedFunction具有更多限制的变体，在restore上仅支持使用了even-split 重分配方案的list-style 状态。它也必须实现两个方法：

|  |
| --- |
| List**<**T**>** **snapshotState(long** checkpointId**,** **long** timestamp**)** **throws** Exception**;**  **void** **restoreState(**List**<**T**>** state**)** **throws** Exception**;** |

snapshotState() 要返回一个List 对象为checkpoint 和restoreState 处理list上的恢复。如果state没有重分区。在snapshotState() 你总是可以返回Collections.singletonList(MY\_STATE)。

* + - 1. Stateful Source Functions

对比其他operator有状态的Sources必须要更多的关注。为了让state的更新和输出集合原子化（在失败/恢复中保证exactly-once 语义），用户必须从source’s 上下文（context）中获得锁。

|  |
| --- |
| Java 代码 |
| **package** com.feicheng.java.flink.office.example;  **import** org.apache.flink.streaming.api.checkpoint.ListCheckpointed; **import** org.apache.flink.streaming.api.functions.source.RichParallelSourceFunction;  **import** java.util.Collections; **import** java.util.List;  */\*\*  \* Created by yuanhailong on 2018/9/26.  \*/* **public class** CounterSource **extends** RichParallelSourceFunction<Long> **implements** ListCheckpointed<Long> {  */\*\* current offset for exactly once semantics \*/* **private** Long **offset**;   */\*\* flag for job cancellation \*/* **private volatile boolean isRunning** = **true**;   @Override  **public void** run(SourceContext<Long> ctx) **throws** Exception {  **final** Object lock = ctx.getCheckpointLock();  **while** (**isRunning**) {  *// output and state update are atomic* **synchronized** (lock) {  ctx.collect(**offset**);  **offset** += 1;  }  }  }   @Override  **public void** cancel() {  **isRunning** = **false**;  }   @Override  **public** List<Long> snapshotState(**long** l, **long** l1) **throws** Exception {  **return** Collections.*singletonList*(**offset**);  }   @Override  **public void** restoreState(List<Long> state) **throws** Exception {  **for** (Long s : state)  **offset** = s;  } } |

|  |
| --- |
| Scala 代码 |
| **class** **CounterSource**  **extends** **RichParallelSourceFunction[Long]**  **with** **ListCheckpointed[Long]** **{**  @volatile  **private** **var** isRunning **=** **true**  **private** **var** offset **=** 0L  **override** **def** run**(**ctx**:** **SourceFunction.SourceContext[Long]):** **Unit** **=** **{**  **val** lock **=** ctx**.**getCheckpointLock  **while** **(**isRunning**)** **{**  *// output and state update are atomic*  lock**.**synchronized**({**  ctx**.**collect**(**offset**)**  offset **+=** 1  **})**  **}**  **}**  **override** **def** cancel**():** **Unit** **=** isRunning **=** **false**  **override** **def** restoreState**(**state**:** **util.List[Long]):** **Unit** **=**  **for** **(**s **<-** state**)** **{**  offset **=** s  **}**  **override** **def** snapshotState**(**checkpointId**:** **Long,** timestamp**:** **Long):** **util.List[Long]** **=**  **Collections.**singletonList**(**offset**)**  **}** |

当Flink完全了解checkpoint时，可能需要这些信息和外部世界沟通。请查阅 org.apache.flink.runtime.state.CheckpointListener 接口。

* + 1. The BroadCast State Pattern(广播状态模型)

|  |
| --- |
| **Provided APIs** |
| BroadcastProcessFunction 和 KeyedBrodcastProcssFunction |
| **Import Considerations** |

[Working with State](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/state.html)  介绍了operator 要么均匀的分布并行operator任务，要么统一的使用整个状态来恢复并行的任务。

支持的第三种operator类型是Broadcast state. Broadcast state 用于支持某些输入来源于一个stream并且这个stream被下游任务都使用的场景。Broadcast state存储在本地并且用于处理在其它stream上的每个元素。我们可以把它想象为一个低吞吐量的流，它包含了一些规则，我们想从其它stream上过来的元素都使用这个规则进行评估。Broadcast state 和其它Operator state 有一些不同：

* Map 格式
* 它只有在一个输入是 broadcasted stream 和 non-broadcasted 的特殊operator上才有效
* Operator可以使用不同的名字拥有多个broadcast states
  + - 1. Provided Apis

为了说明提供的APis，在呈现所有功能前我们将使用一个例子。当我们运行这个例子的时候。我将使用一个具有不同颜色和形状的stream对象的例子。并且我们想找到具有相同颜色的对象，他们遵循一定的模式比如矩形或者三角形。我们假设这个有趣的模型随着时间而变化。

在这个例子中，第一个Stream的element是一个包含了color和shape属性的Item对象。其他stream包含了Rules

从Item 流开始，由于想要相同color的对，我们只需要通过color作为它的key。这将确保相同具备相同color的元素在相同的物理机上。

|  |
| --- |
| *// key the shapes by color*  KeyedStream**<**Item**,** Color**>** colorPartitionedStream **=** shapeStream  **.**keyBy**(new** KeySelector**<**Shape**,** Color**>(){...});** |

接下来是rules，stream包含这些rules并将它们广播到所有的下游任务。并且这些任务将这些规则保留在本地。因此rules可以对所有进来的Item进行评估。下面的代码片段：

1. 广播所有的rules
2. 使用MapStateDescriptor，它将在rules被存储的地方创建broadcast state

|  |
| --- |
| *// a map descriptor to store the name of the rule (string) and the rule itself.*  MapStateDescriptor**<**String**,** Rule**>** ruleStateDescriptor **=** **new** MapStateDescriptor**<>(**  "RulesBroadcastState"**,**  BasicTypeInfo**.**STRING\_TYPE\_INFO**,**  TypeInformation**.**of**(new** TypeHint**<**Rule**>()** **{}));**  *// broadcast the rules and create the broadcast state*  BroadcastStream**<**Rule**>** ruleBroadcastStream **=** ruleStream  **.**broadcast**(**ruleStateDescriptor**);** |

最后，为了评估从Item Stream进来的元素的规则。我需要

1. 连接（connect）两个stream
2. 指定匹配逻辑

用BroadcastStream 连接一个Stream(keyed 或 non-keyed) 可以通过在non-broadcast stream上通过调用connect()方法来完成。它将返回一个BroadcastConnectedStream。 在这个上面我们可用一个CoProcessFunction调用process() 方法。 这个函数将包含我们的匹配逻辑。函数精确类型依赖于non-broadcasted stream的类型。

* 如果是keyed ，那么这个函数是KeyedBroadcastProcessFunction
* 如果是**non-keyed，那么这个函数是**BroadcastProcessFunction

考虑到我们的non-broadcast是keyed ,下面的代码片段包含上面的调用：

注意：connect要在 non-broadcasted stream上调用。使用BroadcastStream 作为一个参数

|  |
| --- |
| DataStream**<**Match**>** output **=** colorPartitionedStream  **.**connect**(**ruleBroadcastStream**)**  **.**process**(**  *// type arguments in our KeyedBroadcastProcessFunction represent:*  *// 1. the key of the keyed stream*  *// 2. the type of elements in the non-broadcast side*  *// 3. the type of elements in the broadcast side*  *// 4. the type of the result, here a string*  **new** KeyedBroadcastProcessFunction**<**Color**,** Item**,** Rule**,** String**>()** **{**  *// my matching logic*  **}**  **)** |

* + - * 1. BroadcastProcessFunction和KeyedBroadcastProcessFunction

就像CoProcessFunction一样，这些函数有两个方法可以去实现。processBroadcastElement() 代表了在broadcast stream中处理进来的数据而processElement（） 在non-broadcast stream中使用。方法的完整签名如下：

|  |
| --- |
| **public** **abstract** **class** **BroadcastProcessFunction<**IN1**,** IN2**,** OUT**>** **extends** BaseBroadcastProcessFunction **{**  **public** **abstract** **void** **processElement(**IN1 value**,** ReadOnlyContext ctx**,** Collector**<**OUT**>** out**)** **throws** Exception**;**  **public** **abstract** **void** **processBroadcastElement(**IN2 value**,** Context ctx**,** Collector**<**OUT**>** out**)** **throws** Exception**;**  **}** |

|  |
| --- |
| **public** **abstract** **class** **KeyedBroadcastProcessFunction<**KS**,** IN1**,** IN2**,** OUT**>** **{**  **public** **abstract** **void** **processElement(**IN1 value**,** ReadOnlyContext ctx**,** Collector**<**OUT**>** out**)** **throws** Exception**;**  **public** **abstract** **void** **processBroadcastElement(**IN2 value**,** Context ctx**,** Collector**<**OUT**>** out**)** **throws** Exception**;**  **public** **void** **onTimer(long** timestamp**,** OnTimerContext ctx**,** Collector**<**OUT**>** out**)** **throws** Exception**;**  **}** |

首先我们需要注意的是为了能够在broadcast 端能处理元素，两个函数都必须实现processBroadcastElement() 方法，为了能在 non-broadcasted端处理元素实现processElement()方法。

这两个方法在他们提供的context有一些差异。non-broadcast端是ReadOnlyContext而broadcasted 是Context。

这两个context:

1. 给定一个访问broadcast state 方法ctx.getBroadcastState(MapStateDescriptor<K, V> stateDescriptor)
2. 允许查询每个元素的时间戳：ctx.timestamp()
3. 获取当前水位：ctx.currentWatermark()
4. 获取当前processing time : ctx.currentProcessingTime()
5. 发送元素到输出端： ctx.output(OutputTag<X> outputTag, X value)

在getBroadcastState()中的stateDescriptor 应该在上面的.broadcast(ruleStateDescriptor) 中能唯一区分。

不同之处在于每种方法对broadcast state的访问类型。Broadcast是读写访问，non-broadcast只有只读访问。这是应为在Flink中没有交叉任务。所以，为了保证我们的operator的所有并行实例有相同的访问控制，我将读写访问仅给了broadcast端，在所有任务上他们都会看到相同的元素。忽略掉rules规则被破坏一致性导致的不一致性。通常这也是非常难以去调试的结果。

注意：逻辑上实现processBroadcast（）必须在所有的并行实例上有相同的确定性行为。

最后，KeyedBroadcastProcessFunction 是在keyed stream上的一个operator,它公开了一些在BroadcastProcessFunction不是有效的功能。它们是:

* 在processElement  方法中的ReadOnlyContext给了一个访问Flink底层timer（定时器） 服务，当timer(定时器)服务被触发时，onTimer()通过OnTimerContext 被调用， OnTimerContext 和ReadOnlyContext 有相同的功能并增加了：
  + 能够询问定时器触发是event time 或processing time
  + 查询与定时器相关的key

这和KeyedProcessFunction的onTimer() 方法是一致的：

* Context 的processBroadcastElement()方法包含applyToKeyedState(StateDescriptor<S, VS> stateDescriptor, KeyedStateFunction<KS, S> function) 方法，这个方法允许注册一个KeyedStateFunction 到所有key的所有state和提供的stateDescriptor关联。

注：由于没有key 和broadcasted 元素关联，因此注册定时器只可能在KeyedBroadcastProcessFunction的processElement方法上而不能在processBroadcastElement方法上

回到我们最初的例子，KeyedBroadcastProcessFunction 看上去像下面这样：

|  |
| --- |
| **new** KeyedBroadcastProcessFunction**<**Color**,** Item**,** Rule**,** String**>()** **{**  *// store partial matches, i.e. first elements of the pair waiting for their second element*  *// we keep a list as we may have many first elements waiting*  **private** **final** MapStateDescriptor**<**String**,** List**<**Item**>>** mapStateDesc **=**  **new** MapStateDescriptor**<>(**  "items"**,**  BasicTypeInfo**.**STRING\_TYPE\_INFO**,**  **new** ListTypeInfo**<>(**Item**.**class**));**  *// identical to our ruleStateDescriptor above*  **private** **final** MapStateDescriptor**<**String**,** Rule**>** ruleStateDescriptor **=**  **new** MapStateDescriptor**<>(**  "RulesBroadcastState"**,**  BasicTypeInfo**.**STRING\_TYPE\_INFO**,**  TypeInformation**.**of**(new** TypeHint**<**Rule**>()** **{}));**  @Override  **public** **void** **processBroadcastElement(**Rule value**,**  Context ctx**,**  Collector**<**String**>** out**)** **throws** Exception **{**  ctx**.**getBroadcastState**(**ruleStateDescriptor**).**put**(**value**.**name**,** value**);**  **}**  @Override  **public** **void** **processElement(**Item value**,**  ReadOnlyContext ctx**,**  Collector**<**String**>** out**)** **throws** Exception **{**  **final** MapState**<**String**,** List**<**Item**>>** state **=** getRuntimeContext**().**getMapState**(**mapStateDesc**);**  **final** Shape shape **=** value**.**getShape**();**  **for** **(**Map**.**Entry**<**String**,** Rule**>** entry:  ctx**.**getBroadcastState**(**ruleStateDescriptor**).**immutableEntries**())** **{**  **final** String ruleName **=** entry**.**getKey**();**  **final** Rule rule **=** entry**.**getValue**();**  List**<**Item**>** stored **=** state**.**get**(**ruleName**);**  **if** **(**stored **==** **null)** **{**  stored **=** **new** ArrayList**<>();**  **}**  **if** **(**shape **==** rule**.**second **&&** **!**stored**.**isEmpty**())** **{**  **for** **(**Item i **:** stored**)** **{**  out**.**collect**(**"MATCH: " **+** i **+** " - " **+** value**);**  **}**  stored**.**clear**();**  **}**  *// there is no else{} to cover if rule.first == rule.second*  **if** **(**shape**.**equals**(**rule**.**first**))** **{**  stored**.**add**(**value**);**  **}**  **if** **(**stored**.**isEmpty**())** **{**  state**.**remove**(**ruleName**);**  **}** **else** **{**  state**.**put**(**ruleName**,** stored**);**  **}**  **}**  **}**  **}** |

* + - 1. 重要考虑(important Considerations)

Api描述完成 后，下面聚焦于在使用broadcast state状态时需要重点记住的几点：

* 没有交叉任务交流：如上所述，这是为什么只能在BroadcastProcessFunction 的broadcast端修改broadcast state内容的原因。此外，用户需要确保所有进来的每个元素以相同的方式修改broadcast state. 否则，不同的任务可能有不同的内容从而导致结果不一致。
* Broadcast State的event顺序因不同的交叉任务而不同：虽然广播一个流的元素可以保证所有元素(最终)都将到达所有下游任务，但是元素可能以不同的顺序到达每个任务。因此状态的更新必须不能依赖进来event的顺序。
* 所有任务都checkpoint 他们的broadcast state:尽管当checkpoint 发生时所有的元素在他们的broadcast state中都有相同的元素，但是所有的任务checkpoint他们的broadcast state而不仅仅是他们中的一个。这样的设计是为了避免在恢复期间从同一个文件中读取。
* **No RocksDB state backend： Broadcast 保存在内存，必须有相应的内存供给去保留所有的operator state**
  + 1. Checkpointing(检查点)

|  |
| --- |
| **Prequisites(预置条件)** |
| **Enableing and Configuring Checkpointing** |
| Related Config Options |
| **Selecting a State Backend** |
| **State Checkpoint in Iterative Jobs** |
| **Restart Strategies** |

在Flink中的每个function和Operator 都能拥有状态（详见[working with state](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/state.html)） 。有状态的函数在处理独立的element/event时存储数据，是的state成为许多关键operator的一个关键构建块。

为了让state可以容错，Flink需要将state checkpoint(检查点)。 Checkpoints 允许Flink在Stream中恢复存储的数据和位置从而给应用程序当失败恢复的时候具备相同的语义。

[documentation on streaming fault tolerance](https://ci.apache.org/projects/flink/flink-docs-release-1.6/internals/stream_checkpointing.html)  详细的描述了Flink背后的容错机制。

* + - 1. **Prequisites(预置条件)**

Flink的checkpoint机制需要和持久化的stream和state进行交互。一般情况下，必须:

* 持久化的数据可在一定持续的时间对数据进行重放。比如消息队列（apache kafka,RabbitMQ 等）或文件系统（HDFS,S3,GFS,NFS,CEPH）
* 状态的持久化存储，通常是分布式文件系统（HDFS,S3…）
  + - 1. **Enableing and Configuring Checkpointing（启用&配置checkpoint）**

Checkpoint在默认情况下是被禁用的。为了启用checkpoint,在StreamExecutionEnvironment上通过调用enableCheckpointing（n） ，n代表checkpoint的频率，单位是毫秒。

Checkpoint的一些其他参数包含：

* exactly-once vs. at-least-once:在调用enableCheckpointing（n,mode）方法的时候你可以选择他们其中的一个mode。在大多数应用中exactly-once是更好的选择。At-least-once相对而言会有更低的延迟。
* checkpoint timeout:如果在checkpoint在这个时间内没有完成，则会退出checkpoint.
* minimum time between checkpoints：为了让streaming应用在checkpoint之间取得一定的进展，一种可能的定义是在checkpoint之间需要花费多少时间。如果这个value设置为5000，下一个checkpoint将在上一个checkpoint完成后的5秒钟内进行启动，而不用例会checkpoint的持续时间和频率。注checkpoint的时间不能比这个参数小。
* number of concurrent checkpoints：默认情况下当一个checkpoint在处理的时候系统不会触发另一个checkpoint.这样可以确保这个拓扑不会在checkpoint上花费太多时间。但是多个检查点是有可能的，一个有趣的现象处理的管道总会有一些延迟（例如函数从外部系统获取数据）但是我们想要在上面重新处理一下小的失败。

这个选项在checkpoint的minimum time 定义的时候不可用。

* externalized checkpoints：你可以配置定期性的将chekpoint持久化到外部系统，checkpoint写入元数据到外部存储系统，当Job失败的时候不会自动清理checkpoint.这样如果Job失败了你就会有一个可恢复的checkpoint. 更多详情参见[deployment notes on externalized checkpoints](https://ci.apache.org/projects/flink/flink-docs-release-1.6/ops/state/checkpoints.html#externalized-checkpoints).
* fail/continue task on checkpoint errors:如果在执行任务的checkpoint过程中发生错误决定是否任务就失败了，这是默认行为。然而，当他被禁用的时候，任务将通知checkpoint协处理器，并继续运行。

|  |
| --- |
| Java 代码 |
| StreamExecutionEnvironment env **=** StreamExecutionEnvironment**.**getExecutionEnvironment**();**  *// start a checkpoint every 1000 ms*  env**.**enableCheckpointing**(**1000**);**  *// advanced options:*  *// set mode to exactly-once (this is the default)*  env**.**getCheckpointConfig**().**setCheckpointingMode**(**CheckpointingMode**.**EXACTLY\_ONCE**);**  *// make sure 500 ms of progress happen between checkpoints*  env**.**getCheckpointConfig**().**setMinPauseBetweenCheckpoints**(**500**);**  *// checkpoints have to complete within one minute, or are discarded*  env**.**getCheckpointConfig**().**setCheckpointTimeout**(**60000**);**  *// allow only one checkpoint to be in progress at the same time*  env**.**getCheckpointConfig**().**setMaxConcurrentCheckpoints**(**1**);**  *// enable externalized checkpoints which are retained after job cancellation*  env**.**getCheckpointConfig**().**enableExternalizedCheckpoints**(**ExternalizedCheckpointCleanup**.**RETAIN\_ON\_CANCELLATION**);** |

* + - * 1. Related Config Options(相关配置项)

更多参数或默认值在 conf/flink-conf.yaml中进行设置（配置指南查看[*configuration*](https://ci.apache.org/projects/flink/flink-docs-release-1.6/ops/config.html)）

|  |  |  |
| --- | --- | --- |
| 配置项 | 默认值 | 描述 |
| state.backend | 无 | 用于存储和checkpoint的状态 |
| state.backend.async | true | 状态备份是否应该使用异步快照的方式，这个是可配置的。某些状态备份可能不支持异步快照，也有些仅支持异步快照，有些可能直接忽略这个选项。 |
| state.backend.fs.memory-threshold | 1024 | 状态数据文件大小的最小值。所有比他小的state chunks 都被存储在根检查点的元数据文件中。 |
| state.backend.incremental | false | 状态备份是否应该创建增量检查点。增量checkpoint值存储当前checkpoint一些有差异的部分，而不是完整的checkpoint。有些state backend可能不支持增量checkpoint，那么他们会直接会略掉这个选项 |
| state.backend.local-recovery | false |  |
| state.checkpoints.dir | 无 | 这个目录用于配置flink支持的文件系统来存储checkpoint的数据文件和元数据。 |
| state.checkpoints.num-retained | 1 | 需要被保留的已经完成了的checkpoint的最大值 |
| state.savepoints.dir | 无 | Savepoint的目录，调用savepoint命令时将状态写入到这里 |
| taskmanager.state.local.root-dirs | 无 |  |

* + - 1. **Selecting a State Backend（状态备份）**

Flink的 [checkpointing mechanism](https://ci.apache.org/projects/flink/flink-docs-release-1.6/internals/stream_checkpointing.html) 在计时器和有状态的操作（包含connect,window 和 [user-defined state](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/state.html)）中存储所有的一致快照。Checkpoint被存储位置（JobManager 内存，文件系统，数据库等）依赖于**State Backend** 的配置。

默认情况下，state保存在TaskManagers 的内存中，checkpoint保存在JobManager的内存中。大的state更适合持久化，Flink支持各种方法将state和checkpoint存储在其他系统中。通过StreamExecutionEnvironment.setStateBackend(…)可以选择state存储的后端备份系统。

有效的状态备份的更多详细信息可查看[state backends](https://ci.apache.org/projects/flink/flink-docs-release-1.6/ops/state/state_backends.html)

* + - 1. **State Checkpoint in Iterative Jobs（迭代JOB中的State 检查点）**

Flink目前仅支持为没有迭代的JOB提供处理的正确性保证。在迭代的job上启用checkpoint将引起一些异常。为了在迭代的程序中强制启用checkpoint用户需要启动一个标识：env.enableCheckpointing(interval, force = true).

* + - 1. **Restart Strategies(重启策略)**

在失败的情况下为了控制job如何重启Flink提供了不同的重启策略。详情查阅 [Restart Strategies](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/restart_strategies.html).

* + 1. Queryable State(可查询状态)

|  |
| --- |
| **Architecture(架构)** |
| **Activating Queryable state(激活可查询的状态)** |
| **Making State Queryable** |
| Queryable State Stream |
| Managed keyed State |
| **Querying State** |
| Example |
| **Configuration** |
| State Server |
| Proxy |
| **Limitations** |

注意：Queryable state 客户端API 目前还在进行中，它提供的接口不能保证稳定性。在即将发布的Flink版本中有可能中断客户端Api的修改。

简单来说，该特性将Flink管理的Keyed state(见 [Working with State](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/state.html))发布到外部世界并且允许用户查询来自于Flink外部的Job状态。在一些场景中，queryable state 通常和外部系统有一些分布式的operations/transactions，比如key-value存储，在实践中这通常是系统的瓶颈。此外，这个特性在调试中可能特别有用。

注意：当查询一个state对象时，该对象可以在不进行任何异步和复制的情况下从并发线程中访问。这是一种设计上的选择，上面的任何一种情况都可能导致Job的延迟而这是我们想要避免的。使用java 堆空间(MemoryStateBackend , FsStateBackend)的任何state backends 当获取value的时候都不会复制而是直接访问存储的引用value. 由于read-modify-write模式是不安全的，因此他可能导致在并发修改过程中某些queryable state服务的失败。但是RocksDBStateBackend 是安全的。

* + - 1. **Architecture(架构)**

前面描述了如何使用Queryable State,简要的描述他的实体和组成是非常有用的。Queryable State 特性包含了三个主要的条目：

1. QueryableStateClient，运行在Flink 集群外部并提交用户的查询
2. QueryableStateClientProxy，运行在每个taskmanager上（Flink集群内部）并负责接收客户端请求，并将结果返回给客户端
3. QueryableStateServer 运行在每个taskmanager上并负责服务本地存储状态

client 连接到其中的一个代理并发送请求用一个特殊的key【k】管理状态。在[Working with State](https://ci.apache.org/projects/flink/flink-docs-release-1.6/dev/stream/state/state.html)已经知道，keyed state 被组织在Key Groups，*并且每个taskmanager都将分配一些key groups.*

* + - 1. **Activating Queryable State**
      2. **Making State Queryable**
      3. Querying state
      4. Configuration（配置）
      5. Limitaions(限制)
    1. State Backends(状态恢复)
    2. Custom Serialization

## State & Fault Tolerance

## State & Fault Tolerance

# Batch (DataSet API)