Apache Beam 0.5.0正式发布

今天，Apache Beam 0.5.0 发布了，此版本通过新的State API添加对状态管道的支持，并通过新的Timer API添加对计时器的支持。此外，该版本还为Elasticsearch和MQ Telemetry Transport（MQTT）添加了新的IO连接器，以及常见的一些错误修复和改进。对于此版本中的所有主要更改，请参阅release notes。

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

Apache Beam是一个开源的数据处理编程库，由Google贡献给Apache的项目，前不久刚刚成为Apache TLP项目。它提供了一个高级的、统一的编程模型，允许我们通过构建Pipeline的方式实现批量、流数据处理，并且构建好的Pipeline能够运行在底层不同的执行引擎上。主要目标是统一批处理和流处理的编程范式，为无限，乱序，web-scale的数据集处理提供简单灵活，功能丰富以及表达能力十分强大的SDK。Apache Beam 希望基于 Beam 开发的数据处理程序可以执行在任意的分布式计算引擎上。

下载Apache Beam 0.5.0

在Maven中引用

```
<dependency>
  <groupId>org.apache.beam</groupId>
  <artifactId>beam-sdks-java-core</artifactId>
  <version>0.5.0</version>
</dependency>
```

```
<dependency>
  <groupId>org.apache.beam</groupId>
  <artifactId>beam-runners-direct-java</artifactId>
  <version>0.5.0</version>
  <scope>runtime</scope>
</dependency>
```

完整的邮件信息

> The Apache Beam community is pleased to announce the availability of the 0.5.0 release.
>
> Apache Beam is a unified programming model for both batch and streaming data processing, enabling efficient execution across diverse distributed execution engines and providing extensibility points for connecting to different technologies and user communities.
>
> This release adds support for stateful pipelines via the new State API, and timers via the new Timer API. Additionally, the release adds new IO connectors for Elasticsearch and MQ Telemetry Transport (MQTT), along with a usual batch of bug fixes and improvements. For all major changes in this release, please refer to the release notes [2].
>
> The 0.5.0 release is now the recommended version; we encourage everyone to upgrade from any earlier releases.
>
> We thank all users and contributors who have helped make this release possible. If you haven't already, we'd like to invite you to join us, as we work towards our first release with API stability.
>
> - Davor Bonaci, on behalf of the Apache Beam community.
>
> [1] https://beam.apache.org/get-started/downloads/
> [2] https://issues.apache.org/jira/secure/ReleaseNote.jspa?projectId=12319527&version=12338859

Release Notes

Bug

- [BEAM-560] - In JAXBCoder, use a pair of ThreadLocals to cache Marshaller/Unmarshaller
- [BEAM-647] - Fault-tolerant sideInputs via Broadcast variables
- [BEAM-853] - Force streaming execution on batch pipelines for testing.
- [BEAM-932] - Findbugs doesn't pass in Spark runner
- [BEAM-979] - ConcurrentModificationException exception after hours of running
- [BEAM-1023] - Add test coverage for BigQueryIO.Write in streaming mode
- [BEAM-1097] - Dataflow error message for non-existing gcpTempLocation is misleading
- [BEAM-1136] - Empty string value should be allowed for ValueProvider<String>
- [BEAM-1144] - Spark runner fails to deserialize MicrobatchSource in cluster mode
- [BEAM-1165] - Unexpected file created when checking dependencies on clean repo
- [BEAM-1177] - Input DStream "bundles" should be in serialized form and include relevant metadata.
- [BEAM-1203] - Exception when running apex runner in non embedded mode
- [BEAM-1207] - Incompatible httpclient version being used with apex runner in YARN mode
- [BEAM-1214] - fail to run on SparkRunner with VerifyError
- [BEAM-1217] - Some examples fail to run due to private / public options mismatch.
- [BEAM-1229] - flink KafkaIOExamples submit error
- [BEAM-1230] - Typo in the documentation of the Window class
- [BEAM-1235] - BigQueryIO doesn't show the load job error to the user
- [BEAM-1248] - Combine with side inputs API should match ParDo
- [BEAM-1249] - Flatten with heterongeneous coders does not have a RunnableOnService test
- [BEAM-1250] - Remove leaf when materializing PCollection to avoid re-evaluation.
- [BEAM-1255] - java.io.NotSerializableException in flink on UnboundedSource
- [BEAM-1258] - BigQueryIO.Write: CREATE_IF_NEEDED and per-window tables is broken
- [BEAM-1273] - Error with FlinkPipelineOptions serialization after setStateBackend
- [BEAM-1292] - PubSubIO: throws error when configured with topic
- [BEAM-1326] - WindowedWordCountIT generated output location can easily collide
- [BEAM-1370] - AfterWatermarkEarlyAndLate does not invoke the onMerge of the early trigger

Improvement

- [BEAM-298] - Make TestPipeline implement the TestRule interface
- [BEAM-370] - Remove the .named() methods from PTransforms and sub-classes
- [BEAM-708] - Migrate BoundedReadFromUnboundedSource to use AutoValue to reduce boilerplate
- [BEAM-716] - Migrate JmsIO to use AutoValue to reduce boilerplate
- [BEAM-757] - The SparkRunner should utilize the SDK's DoFnRunner instead of writing it's own.
- [BEAM-807] - [SparkRunner] Replace OldDoFn with DoFn
- [BEAM-814] - Improve performance when staging files
- [BEAM-921] - Register Coders and Sources to serialize with JavaSerializer
- [BEAM-974] - Adds attributes support to PubsubIO

- [BEAM-1137] - Empty string values should be allowed for ValueProvider of all supported types (Collection, Array, Enum)
- [BEAM-1145] - Remove classifier from shaded spark runner artifact
- [BEAM-1146] - Decrease spark runner startup overhead
- [BEAM-1176] - Make our test suites use @Rule TestPipeline
- [BEAM-1186] - Migrate the remaining tests to use TestPipeline as a JUnit rule.
- [BEAM-1201] - Remove producesSortedKeys from BoundedSource
- [BEAM-1223] - Replace public constructors with static factory methods for Sum.[*]Fn classes
- [BEAM-1225] - Add a ToString transform in Java SDK
- [BEAM-1260] - PAssert should capture the assertion site
- [BEAM-1266] - Use full windowed value representations within Dataflow job representation
- [BEAM-1291] - KafkaIO: don't log warning in offset fetcher while closing the reader.
- [BEAM-1302] - Improve warning messages in BigQueryServicesImpl.

New Feature

- [BEAM-85] - PAssert needs sanity check that it's used correctly
- [BEAM-425] - Create Elasticsearch IO
- [BEAM-606] - Create MqttIO
- [BEAM-1038] - Support for new State API in DataflowRunner
- [BEAM-1117] - Support for new Timer API in Direct runner